Applied Regression Analysis

Course notes for STAT 378/502

Adam B Kashlak
Mathematical & Statistical Sciences
University of Alberta
Edmonton, Canada, T6G 2G1

November 19, 2018

Contents

Preface

> I never felt such a glow of loyalty and respect towards the sovereignty and magnificent sway of mathematical analysis as when his answer reached me confirming, by purely mathematical reasoning, my various and laborious statistical conclusions.
>
> *Regression towards Mediocrity in Hereditary Stature*
> Sir Francis Galton, FRS (1886)

The following are lecture notes originally produced for an upper level undergraduate course on linear regression at the University of Alberta in the fall of 2017. Regression is one of the main, if not the primary, workhorses of statistical inference. Hence, I do hope you will find these notes useful in learning about regression.

The goal is to begin with the standard development of ordinary least squares in the multiple regression setting, then to move onto a discussion of model assumptions and issues that can arise in practice, and finally to discuss some specific instances of generalized linear models (GLMs) without delving into GLMs in full generality. Of course, what follows is by no means a unique exposition but is mostly derived from three main sources: the text, *Linear Regression Analysis*, by Montgomery, Peck, and Vining; the course notes of Dr. Linglong Kong who lectured this same course in 2013; whatever remains inside my brain from supervising (TAing) undergraduate statistics courses at the University of Cambridge during my PhD years.

Adam B Kashlak
Edmonton, Canada
August 2017

Chapter 1

Ordinary Least Squares

Introduction

Linear regression begins with the simple but profound idea that some observed *response* variable, $Y \in \mathbb{R}$, is a function of p input or *regressor* variables $x_1, \ldots, x_p$ with the addition of some unknown noise variable ε. Namely,

$$Y = f(x_1, \ldots, x_p) + \varepsilon$$

where the noise is generally assumed to have mean zero and finite variance. In this setting, Y is usually considered to be a random variable while the x_i are considered fixed. Hence, the expected value of Y is in terms of the unknown function f and the regressors:

$$\mathrm{E}\,(Y|x_1, \ldots, x_p) = f(x_1, \ldots, x_p).$$

While f can be considered to be in some very general classes of functions, we begin with the standard linear setting. Let $\beta_0, \beta_1, \ldots, \beta_p \in \mathbb{R}$. Then, the *multiple regression model* is

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \beta^{\mathrm{T}} X$$

where $\beta = (\beta_0, \ldots, \beta_p)^{\mathrm{T}}$ and $X = (1, x_1, \ldots, x_p)^{\mathrm{T}}$. The *simple regression model* is a submodel of the above where $p = 1$, which is

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon,$$

and will be treated concurrently with multiple regression.

In the statistics setting, the parameter vector $\beta \in \mathbb{R}^p$ is unknown. The analyst observes multiple replications of regressor and response pairs, $(X_1, Y_1), \ldots, (X_n, Y_n)$ where n is the *sample size*, and wishes to choose a "best" estimate for β based on these n observations. This setup can be concisely written in a vector-matrix form as

$$Y = X\beta + \varepsilon \tag{1.0.1}$$

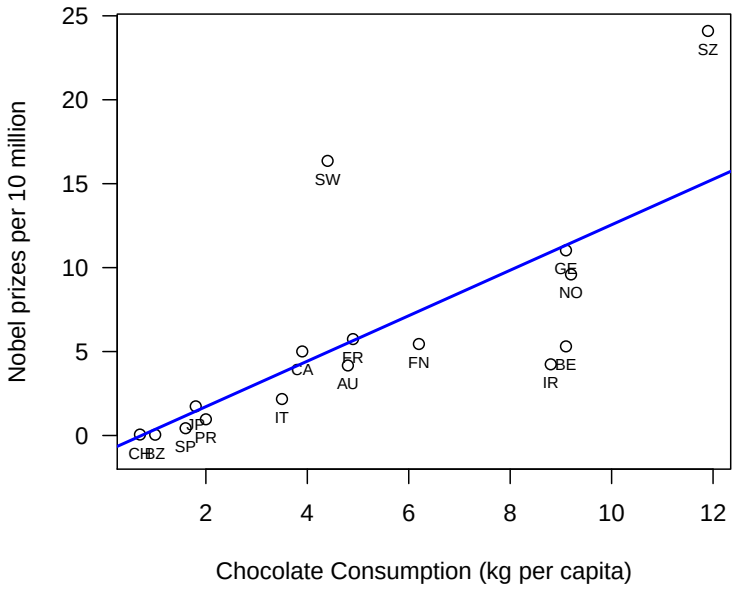

Figure 1.1: Comparing chocolate consumption in kilograms per capita to number of Nobel prizes received per 10 million citizens.

where

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ 1 & x_{2,1} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Note that $Y, \varepsilon \in \mathbb{R}^n$, $\beta \in \mathbb{R}^{p+1}$, and $X \in \mathbb{R}^{n \times p+1}$.

As Y is a random variable, we can compute its mean vector and covariance matrix as follows:

$$\mathrm{E}Y = \mathrm{E}\left(X\beta + \varepsilon\right) = X\beta$$

and

$$\mathrm{Var}\left(Y\right) = \mathrm{E}\left((Y - X\beta)(Y - X\beta)^{\mathrm{T}}\right) = \mathrm{E}\left(\varepsilon\varepsilon^{\mathrm{T}}\right) = \mathrm{Var}\left(\varepsilon\right) = \sigma^2 I_n.$$

An example of a linear regression from a study from the New England Journal of Medicine can be found in Figure 1.1. This study highlights the correlation between chocolate consumption and Nobel prizes received in 16 different countries.

Table 1.1: The four variables in the linear regression model of Equation 1.0.1 split between whether they are fixed or random variables and between whether or not the analyst knows their value.

	Known	Unknown
Fixed	X	β
Random	Y	ε

Definitions

Before continuing, we require the following collection of terminology.

The *response* Y and the *regressors* X were already introduced above. These elements comprise the *observed data* in our regression. The *noise* or *error* variable is ε. The entries in this vector are usually considered to be independent and identically distributed (iid) random variables with mean zero and finite variance $\sigma^2 < \infty$. Very often, this vector will be assumed to have a multivariate normal distribution: $\varepsilon \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 I_n\right)$ where I_n is the $n \times n$ identity matrix. The variance σ^2 is also generally considered to be unknown to the analyst.

The unknown vector β is our *parameter* vector. Eventually, we will construct an *estimator* $\hat{\beta}$ from the observed data. Given such an estimator, the *fitted values* are $\hat{Y} := X\hat{\beta}$. These values are what the model believes are the expected values at each regressor.

Given the fitted values, the *residuals* are $r = Y - \hat{Y}$ which is a vector with entries $r_i = Y_i - \hat{Y}_i$. This is the difference between the observed response and the expected response of our model. The residuals are of critical importance to testing how good our model is and will reappear in most subsequent sections.

Lastly, there is the concept of sum of squares. Letting $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$ be the sample mean for Y, the *total sum of squares* is $SS_{\text{tot}} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$, which can be thought of as the total variation of the responses. This can be decomposed into a sum of the *explained sum of squares* and the *residual sum of squares* as follows:

$$SS_{\text{tot}} = SS_{\text{exp}} + SS_{\text{res}} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

The explained sum of squares can be thought of as the amount of variation explained by the model while the residual sum of squares can be thought of as a measure of the variation that is not yet contained in the model. The sum of squares gives us an expression for the so called *coefficient of determination*, $R^2 = SS_{\text{exp}}/SS_{\text{tot}} = 1 - SS_{\text{res}}/SS_{\text{tot}} \in [0, 1]$, which is treated as a measure of what percentage of the variation is explained by the given model.

1.1 Point Estimation

In the ordinary least squares setting, the our choice of estimator is

$$\hat{\beta} = \underset{\tilde{\beta} \in \mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} (Y_i - X_{i,\cdot} \cdot \tilde{\beta})^2 \tag{1.1.1}$$

where $X_{i,\cdot}$ is the ith row of the matrix X. In the simple regression setting, this reduces to

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\tilde{\beta}_0, \tilde{\beta}_1) \in \mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} (Y_i - (\tilde{\beta}_0 + \tilde{\beta}_1 x_1))^2.$$

Note that this is equivalent to choosing a $\hat{\beta}$ to minimize the sum of the squared residuals.

It is perfectly reasonable to consider other criterion beyond minimizing the sum of squared residuals. However, this approach results in an estimator with many nice properties. Most notably is the Gauss-Markov theorem:

Theorem 1.1.1 (Gauss-Markov Theorem)**.** *Given the regression setting from Equation 1.0.1 and that for the errors,* $\mathrm{E}\varepsilon_i = 0$ *for* $i = 1, \ldots, n$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ *for* $i = 1, \ldots, n$, *and* $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ *for* $i \neq j$, *then the least squares estimator results in the minimal variance over all linear unbiased estimators. (This is sometimes referred to as the "Best Linear Unbiased Estimator" or BLUE)*

Hence, it can be shown that the estimator is unbiased, $\mathrm{E}\hat{\beta} = \beta$ and that the constructed least squares line passes through the centre of the data in the sense that the sum of the residuals is zero, $\sum_{i=1}^{n} r_i = 0$ and that $\bar{Y} = \hat{\beta}\bar{X}$ where $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$ is the sample mean of the Y_i and where $\bar{X}$ is the vector of column means of the matrix X.

1.1.1 Derivation of the OLS estimator[1]

The goal is to derive an explicit solution to Equation 1.1.1. First, consider the following partial derivative:

$$\begin{aligned}
\frac{\partial}{\partial \hat{\beta}_k} \sum_{i=1}^{n} (Y_i - X_{i,\cdot} \cdot \hat{\beta})^2 &= -2 \sum_{i=1}^{n} (Y_i - X_{i,\cdot} \cdot \hat{\beta}) X_{i,k} \\
&= -2 \sum_{i=1}^{n} (Y_i - \textstyle\sum_{j=1}^{p} X_{i,j} \hat{\beta}_j) X_{i,k} \\
&= -2 \sum_{i=1}^{n} Y_i X_{i,k} + 2 \sum_{i=1}^{n} \sum_{j=1}^{p} X_{i,j} X_{i,k} \hat{\beta}_j
\end{aligned}$$

[1] See Montgomery, Peck, Vining Sections 2.2.1 and 3.2.1 for simple and multiple regression, respectively.

The above is the kth entry in the vector $\nabla \sum_{i=1}^{n}(Y_i - X_{i,\cdot}\hat{\beta})^2$. Hence,

$$\nabla \sum_{i=1}^{n}(Y_i - X_{i,\cdot}\hat{\beta})^2 = -2X^{\mathrm{T}}Y + 2X^{\mathrm{T}}X\hat{\beta}.$$

Setting this equal to zero results in a critical point at

$$X^{\mathrm{T}}Y = X^{\mathrm{T}}X\hat{\beta}$$

or $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$ assuming $X^{\mathrm{T}}X$ is invertible. Revisiting the terminology in the above definitions sections gives the following:

Least Squares Estimator:	$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$
Fitted Values:	$\hat{Y} = X\hat{\beta} = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$
Residuals:	$r = Y - \hat{Y} = (I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})Y$

In the case that $n > p$, the matrix $P_X := X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is a rank $p+1$ projection matrix. Similarly, $I_n - P_X$ is the complementary rank $n-p-1$ projection matrix. Intuitively, this implies that the fitted values are the projection on the observed values onto a p-dimensional subspace while the residuals arise from a projection onto the orthogonal subspace. As a result, it can be shown that $\operatorname{cov}\left(\hat{Y}, r\right) = 0$.

Now that we have an explicit expression for the least squares estimator $\hat{\beta}$, we can show that it is unbiased.

$$\mathrm{E}\hat{\beta} = \mathrm{E}\left((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y\right) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{E}Y = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X\beta = \beta.$$

Following that, we can compute its variance.

$$\begin{aligned}
\operatorname{Var}\left(\hat{\beta}\right) &= \mathrm{E}\left((\hat{\beta} - \beta)(\hat{\beta} - \beta)^{\mathrm{T}}\right) \\
&= \mathrm{E}\left(\hat{\beta}\hat{\beta}^{\mathrm{T}}\right) - \beta\beta^{\mathrm{T}} \\
&= \mathrm{E}\left((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y)^{\mathrm{T}}\right) - \beta\beta^{\mathrm{T}} \\
&= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{E}\left(YY^{\mathrm{T}}\right)X(X^{\mathrm{T}}X)^{-1} - \beta\beta^{\mathrm{T}} \\
&= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(\sigma^2 I_n + X\beta\beta^{\mathrm{T}}X^{\mathrm{T}})X(X^{\mathrm{T}}X)^{-1} - \beta\beta^{\mathrm{T}} \\
&= \sigma^2(X^{\mathrm{T}}X)^{-1}.
\end{aligned}$$

Thus far, we have only assumed that ε is a random vector with iid entries with mean zero and variance σ^2. If in addition, we assumed that ε has a *normal* or *Gaussian* distribution, then

$$\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right), \quad Y \sim \mathcal{N}\left(X\beta, \sigma^2 I_n\right), \text{ and } \hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2(X^{\mathrm{T}}X)^{-1}\right).$$

Furthermore, with a little work, one can show that for the fitted values and residuals also have normal distributions in this setting:

$$\hat{Y} \sim \mathcal{N}\left(X\hat{\beta}, \sigma^2 P_X\right), \quad \text{and} \quad r \sim \mathcal{N}\left(0, \sigma^2(I_n - P_X)\right).$$

Notice that the two above covariance matrices are not generally of full rank. This assumption that the errors follow a normal distribution is a very common assumption to make in practice.

1.1.2 Maximum likelihood estimate under normality[2]

In the previous section, the OLS estimator is derived by minimizing the sum of the squared errors. Now, given the additional assumption that the errors have a normal distribution, we can compute an alternative estimator for β: the maximum likelihood estimate (MLE). We can also use this to simultaneously compute the MLE for σ^2.

From above we have that $Y \sim \mathcal{N}\left(X\beta, \sigma^2 I_n\right)$, and hence the likelihood is

$$L(\beta, \sigma^2; X, Y) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(Y - X\beta)^{\mathrm{T}}(Y - X\beta)\right).$$

The log likelihood is then

$$\begin{aligned}\ell(\beta, \sigma^2; X, Y) = \log L(\beta, \sigma^2; X, Y) = \\ = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(Y - X\beta)^{\mathrm{T}}(Y - X\beta).\end{aligned}$$

This implies that the MLE for β comes from solving

$$0 = \frac{\partial \ell}{\partial \beta} = \frac{\partial}{\partial \beta}(Y - X\beta)^{\mathrm{T}}(Y - X\beta),$$

which is solved by the OLS estimator from above. Hence, the MLE under normality is the least squares estimator.

For the variance term σ^2, the MLE is similarly found by solving

$$0 = \frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2}(\sigma^2)^{-1} + \frac{(\sigma^2)^{-2}}{2}(Y - X\beta)^{\mathrm{T}}(Y - X\beta).$$

This occurs for $\hat{\sigma}^2 = n^{-1}(Y - X\hat{\beta})^{\mathrm{T}}(Y - X\hat{\beta})$, which is just the average sum of squares of the residuals: $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n} r_i^2$. However, this is a biased estimator of the variance as the residuals are not independent and have a degenerate covariance matrix of rank $n - p - 1$. Intuitively, this implies that the sum of squared residuals has $n - p - 1$ degrees of freedom resulting in

$$\frac{SS_{\text{res}}}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n} r_i^2 \sim \chi^2\left(n - p - 1\right)$$

[2] See Montgomery, Peck, Vining Section 3.2.6.

and the unbiased estimator of σ^2 being $SS_{\text{res}}/(n-p-1) = \hat{\sigma}^2(n/(n-p-1))$. For a more precise explanation of where this comes from, see Cochran's theorem (https://en.wikipedia.org/wiki/Cochran%27s_theorem), which is beyond the scope of this course.

Chocolate-Nobel Data

Running a regression in R on the chocolate consumption vs Nobel prize data from Figure 1.1 results in

$$\hat{\beta} = \begin{pmatrix} -0.9910 \\ 1.3545 \end{pmatrix}.$$

1.1.3 Proof of the Gauss-Markov Theorem

Proof. Any linear estimator can be written as AY for some non-random matrix $A \in \mathbb{R}^{(p+1)\times n}$. We can in turn write $A = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} + D$ for some matrix $D \in \mathbb{R}^{(p+1)\times n}$. Then, as

$$\begin{aligned} \mathrm{E}(AY) &= AX\beta \\ &= \left[(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} + D\right] X\beta \\ &= \beta + DX\beta, \end{aligned}$$

the unbiased condition implies that $DX\beta = 0$ for any $\beta \in \mathbb{R}^{p+1}$ and hence that $DX = 0$.

Next, we compute the variance of the arbitrary linear unbiased estimator to get

$$\begin{aligned} \mathrm{Var}(AY) &= A\mathrm{Var}(Y)A^{\mathrm{T}} \\ &= \sigma^2\left[(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} + D\right]\left[(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} + D\right]^{\mathrm{T}} \\ &= \sigma^2\left[(X^{\mathrm{T}}X)^{-1} + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}D^{\mathrm{T}} + DX(X^{\mathrm{T}}X)^{-1} + DD^{\mathrm{T}}\right] \\ &= \sigma^2\left[(X^{\mathrm{T}}X)^{-1} + DD^{\mathrm{T}}\right]. \end{aligned}$$

Hence, to minimize the variance, we must minimize DD^{T} as DD^{T} is necessarily a positive semi-definite matrix. This is achieved by setting $D = 0$ and arriving at $(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$ having minimial variance. □

Remark 1.1.2. *Note that DD^{T} is positive semi-definite for any choice of D as for any $w \in \mathbb{R}^{p+1}$, we have*

$$w^{\mathrm{T}}(DD^{\mathrm{T}})w = (D^{\mathrm{T}}w)^{\mathrm{T}}(Dw) = \|Dw\|_2 \geq 0.$$

Remark 1.1.3. *While $\hat{\beta}$ has minimal variance over all unbiased estimators, we can lessen the variance further if we allow for biased estimators. This is considered in many more advanced regression methods such as ridge regression.*

1.2 Hypothesis Testing

1.2.1 Goodness of fit[3]

We now have a model for our data, and in some sense, this model is optimal as it minimizes the squared errors. However, even being optimal, we are still interested in knowing whether or not this is a good model for our data. This is a question of *goodness of fit*.

The first question to ask is, do any of the regressors provide information about the response in the linear model framework? This can be written mathematically as

$$H_0 : \beta_1 = \ldots = \beta_p = 0 \qquad H_1 : \exists i \geq 1 \ s.t. \ \beta_i \neq 0, \tag{1.2.1}$$

which is asking is there at least one β_i that we can claim is non-zero and hence implies that the regressor x_i has some nontrivial influence over y.

To test this hypothesis, we revisit the explained and residual sums of squares introduced in the Definitions section. Specifically, we already have that $SS_{\text{res}}/\sigma^2 \sim \chi^2(n-p-1)$ from above. Similarly, $SS_{\text{exp}}/\sigma^2 \sim \chi^2(p)$ under the null hypothesis where $\beta_1 = \ldots = \beta_p = 0$, and hence any variation in those terms should be pure noise. Lastly, it can be demonstrated that SS_{res} and SS_{exp} are independent random variables, which intuitively follows from the orthogonality of the fitted values and the errors.

The usual test statistic for Hypothesis 1.2.1 is

$$\frac{SS_{\text{exp}}/p}{SS_{\text{res}}/(n-p-1)} \sim F(p, n-p-1),$$

which leads to an *F test*. If the test statistic is large, then the explained variation is larger than the noise resulting in a small p-value and a rejection of the null hypothesis.

F test on Chocolate-Nobel data

From the final line of the R output from the `summary()` command, we have a test statistic value of 15.45 with degrees of freedom 1 and 14. This results in a very small p-value of 0.001508.

If you were to run the regression in R without the intercept term, which is fixing $\beta_0 = 0$, then the result is $\hat{\beta}_1 = 1.22$, a value for the test statistic for the F test of 44.24, now with degrees of freedom 1 and 15, and an even smaller p-value of 7.7×10^{-06}.

[3] See Montgomery, Peck, Vining Sections 2.3.2 and 3.3.1 for simple and multiple regression, respectively.

1.2.2 Regression coefficients[4]

Given that the previous F test results in a significant p-value, the subsequent question is to ask which of the p regressors are significant? Hence, we have the following hypotheses for $j = 0, 1, \ldots, p$.

$$H_{0,j} : \beta_j = 0 \qquad H_{1,j} : \beta_j \neq 0.$$

Each individual $\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2 (X^{\mathrm{T}}X)^{-1}_{j,j}\right)$ where $(X^{\mathrm{T}}X)^{-1}_{j,j}$ is the jth entry in the diagonal of $(X^{\mathrm{T}}X)^{-1}_{j,j}$.[5] Thus, under the null hypothesis that $\beta_j = 0$, we have that

$$\hat{\beta}_j / \sqrt{\sigma^2 (X^{\mathrm{T}}X)^{-1}_{j,j}} \sim \mathcal{N}(0, 1).$$

However, we cannot perform a z test as σ^2 is unknown. To rectify this, the unbiased estimator for σ^2 is used in its place resulting in

$$\frac{\hat{\beta}_j}{\sqrt{(X^{\mathrm{T}}X)^{-1}_{j,j} SS_{\mathrm{res}}/(n-p-1)}} \sim t(n-p-1),$$

and a t test can be performed. If the value of the test statistic is large, then there may be sufficient evidence to reject the null that $\beta_j = 0$. The denominator is often referred to as the *standard error*. To simplify future formulae, this will be denoted as $\mathrm{se}(\beta_j)$.

It is worth noting that this test looks for significant influence of the jth regressor on the response given all of the other regressors. Hence, it quantifies the marginal as opposed to the absolute effect of that variable on the model. These ideas will be investigated further when discussing variable selection in Section 3.2. However, as a quick word of caution, when p hypothesis tests are performed, the analyst needs to consider multiple testing corrections.

t test on Chocolate-Nobel data

The R commands `lm()` and `summary()` will return a table of regression coefficients, t test statistics and p-values associated with each coefficient. For the Chocolate-Nobel prize data, the table looks like

	Estimate	Std. Error	t value	Pr(> |t|)	
(Intercept)	-0.9910	2.1327	-0.465	0.64932	
choco	1.3545	0.3446	3.931	0.00151	⋆⋆

[4] See Montgomery, Peck, Vining Sections 2.3.1 and 3.3.2 for simple and multiple regression, respectively.

[5] We will index the entries of the matrix from $0, 1, \ldots, p$ to conform with the indexing of the β's. Note that this is a $(p+1) \times (p+1)$ matrix.

1.2.3 Partial F-test[6]

In the previous two sections, we first tested as to whether or not there exists at least one β_j, $j = 1, \ldots, p$, that is non-zero. Then, we tested whether or not a specific β_j is non-zero. The next logical question is whether or not some collection of β_j's of size strictly between 1 and p has a non-zero element. That is, for a fixed q,

$$H_0 : \beta_{p-q+1} = \ldots = \beta_p = 0 \qquad H_1 : \exists i \geq p - q + 1 \ s.t. \ \beta_i \neq 0. \tag{1.2.2}$$

Here, we are comparing two different models, which are the partial and full models, respectively,

$$Y = \beta_{0:p-q} X + \varepsilon, \quad \text{and} \quad Y = \beta_{0:p-q} X + \beta_{p-q+1:p} X + \varepsilon,$$

and want to know whether the final q regressors add any significant explanation to our model given the other $p - q$. For the above notation,

$$\beta_{i:j} = (0, \ldots, 0, \beta_i, \beta_{i+1}, \ldots, \beta_j, 0, \ldots, 0)^{\mathrm{T}}.$$

To run the hypothesis test 1.2.2, we would have to compute the least squares estimator in the partial model, $\hat{\beta}_{1:p-q}$, and the standard least squares estimator in the full model, $\hat{\beta}$. Then, we will have to compute the additional explained sum of squares gained from adding the q extra regressors to our model, which is

$$SS_{\exp}(\beta_{p-q+1:p} | \beta_{1:p-q}) = SS_{\exp}(\beta) - SS_{\exp}(\beta_{1:p-q}),$$

the explained sum of squares from the full model minus the explained sum of squares from the partial model.

Similarly to the full F-test from above, we have under the null hypothesis that $SS_{\exp}(\beta_{p-q+1:p} | \beta_{1:p-q}) / \sigma^2 \sim \chi^2 (q)$. Hence,

$$\frac{SS_{\exp}(\beta_{p-q+1:p} | \beta_{1:p-q}) / q}{SS_{\mathrm{res}} / (n - p - 1)} \sim F (q, n - p - 1),$$

so if this test statistic is large, then we have evidence to suggestion that at least one of the additional q regressors adds some explanatory power to our model.

1.3 Confidence Intervals[7]

Confidence intervals play a complementary role with hypothesis testing. From the development of the above test for an individual β_j, we have that

$$\frac{\hat{\beta}_j - \beta_j}{\mathrm{se}(\beta_j)} \sim t (n - p - 1),$$

[6] See Montgomery, Peck, Vining the latter half of Section 3.3.2.

[7] See Montgomery, Peck, Vining Sections 2.4.1 and 3.4.1 for simple and multiple regression, respectively.

Hence, a $1-\alpha$ confidence interval for the parameter β_j is

$$\hat{\beta}_j - t_{\alpha/2,n-p-1}\text{se}(\beta_j) \leq \beta \leq \hat{\beta}_j + t_{\alpha/2,n-p-1}\text{se}(\beta_j)$$

where $t_{\alpha/2,n-p-1} \in \mathbb{R}^+$ is such that $\text{P}\left(T \leq t_{\alpha/2,n-p-1}\right) = \alpha/2$ when $T \sim t\left(n-p-1\right)$.

While the above can be used to produce a confidence interval for each individual parameter, combining these intervals will not result in a $1-\alpha$ confidence set for the entire parameter vector. To construct such a confidence region, a little more care is required.[8] Also, we will construct a confidence set for the entire vector $(\beta_0, \beta_1, \ldots, \beta_p)$, which results in $p+1$ degrees of freedom in what follows. As $\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2(X^{\text{T}}X)^{-1}\right)$ we have that

$$\sigma^{-2}(\hat{\beta}-\beta)^{\text{T}}X^{\text{T}}X(\hat{\beta}-\beta) \sim \chi^2\left(p+1\right).$$

From before, we have that $SS_{\text{res}}/\sigma^2 \sim \chi^2\left(n-p-1\right)$. Hence

$$\frac{(\hat{\beta}-\beta)^{\text{T}}X^{\text{T}}X(\hat{\beta}-\beta)/(p+1)}{SS_{\text{res}}/(n-p-1)} \sim F\left(p+1, n-p-1\right).$$

Thus, a $1-\alpha$ confidence ellipsoid can be constructed as

$$\frac{(\hat{\beta}-\beta)^{\text{T}}X^{\text{T}}X(\hat{\beta}-\beta)/(p+1)}{SS_{\text{res}}/(n-p-1)} \leq F_{\alpha,p+1,n-p-1}.$$

A 95% and a 99% confidence ellipsoid for the Chocolate-Nobel prize data is displayed in Figure 1.2. Notice that both ellipses contain $\hat{\beta}_0 = 0$ which had a t statistic p-value of 0.649. Meanwhile neither contain $\hat{\beta}_1 = 0$ whose p-value was the very significant 0.0015. The confidence ellipses were plotted with help from the R library `ellipse`.

1.4 Prediction Intervals

1.4.1 Prediction for an expected observation[9]

Given the least squares model, the analyst may be interested in estimating the expected value of Y have some specific input $x = (1, x_1, \ldots, x_p)$. Our new random variable is $\hat{Y}_0 = \hat{\beta} \cdot X$ where X is fixed and $\hat{\beta}$ is random. Of course, the expected value is just

$$\text{E}\left(\hat{Y}_0 \middle| X = x\right) = \text{E}\hat{\beta} \cdot x = \beta_0 + \sum_{i=1}^{p} \beta_i x_i.$$

[8] See Montgomery, Peck, Vining Section 3.4.3

[9] See Montgomery, Peck, Vining Sections 2.4.2 and 3.4.2 for simple and multiple regression, respectively.

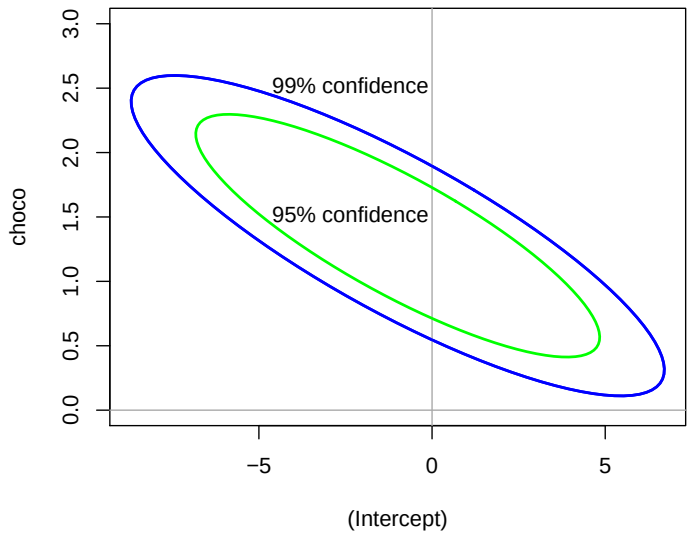

Figure 1.2: Confidence ellipses at the 95% and 99% level for the chocolate consumption vs Nobel prizes data.

To find a $1-\alpha$ interval estimate for $\hat{Y}_0$ at $X = x$, recall once again that $\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 (X^{\mathrm{T}}X)^{-1}\right)$. Thus,

$$\hat{Y}_0 | (X = x) \sim \mathcal{N}\left(\beta \cdot x, \sigma^2 x^{\mathrm{T}} (X^{\mathrm{T}}X)^{-1} x\right).$$

Hence,

$$\frac{\hat{\beta} \cdot x - \mathrm{E}\left(\hat{Y}_0 \middle| X = x\right)}{\sqrt{\sigma^2 x^{\mathrm{T}} (X^{\mathrm{T}}X)^{-1} x}} \sim \mathcal{N}(0,1),$$

and

$$\frac{\hat{\beta} \cdot x - \mathrm{E}\left(\hat{Y}_0 \middle| X = x\right)}{\sqrt{(SS_{\mathrm{res}}/(n-p-1)) x^{\mathrm{T}} (X^{\mathrm{T}}X)^{-1} x}} \sim t(n-p-1),$$

which results in the following $1-\alpha$ confidence interval:

$$\hat{\beta} \cdot x - t_{\alpha/2, n-p-1} \sqrt{\frac{SS_{\mathrm{res}}}{n-p-1} x^{\mathrm{T}} (X^{\mathrm{T}}X)^{-1} x} \leq$$
$$\leq \mathrm{E}\left(\hat{Y}_0 \middle| X = x\right) = \beta \cdot x \leq$$
$$\leq \hat{\beta} \cdot x + t_{\alpha/2, n-p-1} \sqrt{\frac{SS_{\mathrm{res}}}{n-p-1} x^{\mathrm{T}} (X^{\mathrm{T}}X)^{-1} x}.$$

1.4.2 Prediction for a new observation[10]

In the previous subsection, we asked for a confidence interval for the expected value of the response given a new vector of regressors, which was a confidence interval for $\mathrm{E}\left(\hat{Y}_0 \middle| X = x\right) = \beta \cdot x$ based on $\hat{\beta} \cdot x$. Now, we want to determine a confidence interval for the future response given a vector of regressors. That is, we want an interval for $Y_0 = \beta \cdot x + \varepsilon_0 \sim \mathcal{N}\left(\beta \cdot x, \sigma^2\right)$, but, as usual, β unknown. To circumvent this, note that

$$Y_0 - \hat{Y}_0 = (\beta \cdot x + \varepsilon_0) - \hat{\beta} \cdot x \sim \mathcal{N}\left(0, \sigma^2(1 + x^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x)\right),$$

because the variances of ε_0 and $\hat{\beta} \cdot x$ sum as these are independent random variables. Hence, applying the usual rearrangement of terms and replacement of σ^2 with $SS_{\mathrm{res}}/(n-p-1)$ results in

$$\hat{\beta} \cdot x - t_{\alpha/2,n-p-1}\sqrt{\frac{(1 + x^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x)SS_{\mathrm{res}}}{n-p-1}} \leq Y_0 \leq$$
$$\leq \hat{\beta} \cdot x + t_{\alpha/2,n-p-1}\sqrt{\frac{(1 + x^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x)SS_{\mathrm{res}}}{n-p-1}}.$$

Intervals for the Chocolate-Nobel prize data for both the expected mean and for a new observation are displayed in Figure 1.3.

1.5 Indicator Variables and ANOVA[11]

1.5.1 Indicator Variables[12]

Thus far, we have considered models of the form

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

where the regressors $x_1, \ldots, x_p \in \mathbb{R}$ can take on any real value. However, very often in practice, we have regressors that take on categorical values. For example, male vs female, employed vs unemployed, treatment vs placebo, Edmonton vs Calgary, etc. When there is a binary choice as in these examples, we can choose one category to correspond to zero and the other category to correspond to one.

As an example, consider

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{1.5.1}$$

[10] See Montgomery, Peck, Vining Section 3.5.
[11] See Montgomery, Peck, and Vining Chapter 8.
[12] See Montgomery, Peck, and Vining Section 8.1 and 8.2.

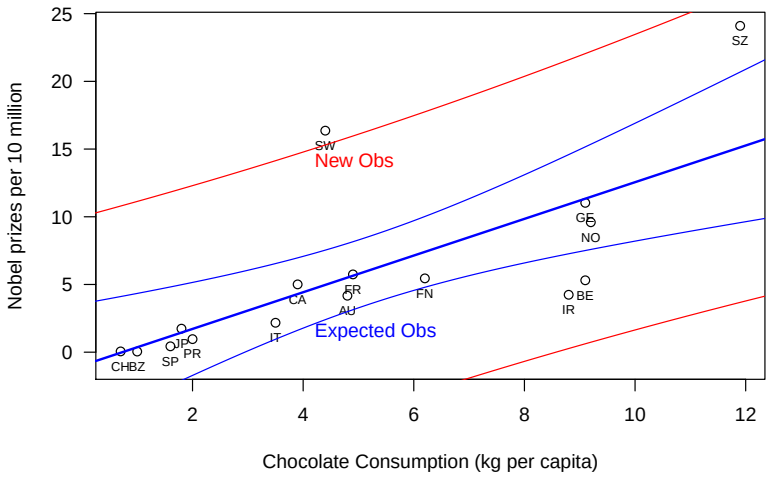

Figure 1.3: A 95% confidence interval (in blue) for the expected value and a 95% prediction interval (in read) for a new observation for the chocolate consumption vs Nobel prizes data.

where $x_1 \in \mathbb{R}$ and $x_2 \in \{0, 1\}$. Then, we effectively have two models:

$$y = \beta_0 + \beta_1 x_1 + 0 + \varepsilon$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 + \varepsilon = (\beta_0 + \beta_2) + \beta_2 x_1 + \varepsilon.$$

What we have is two models with the same slope β_1 but with two different intercepts β_0 and $\beta_0 + \beta_2$, which are two parallel lines.

Remark 1.5.1. *A first thought is to merely split the data and train two separate models. However, we want to use the entire dataset at once specifically to estimate the common slope β_1 with as much accuracy as possible.*

While the range of the regressors has changed, we will fit the least squares estimate to the model precisely as before. Now, considering the model 1.5.1, assume that we have m samples with $x_2 = 0$ and n samples with $x_2 = 1$. Our design matrix takes on a new form:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_m \\ Y_{m+1} \\ \vdots \\ Y_{m+n} \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & x_m & 0 \\ 1 & x_{m+1} & 1 \\ \vdots & \vdots & \vdots \\ 1 & x_{m+n} & 1 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{m+n} \end{pmatrix}.$$

However, the least squares estimate is computed as before as $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$. Furthermore, we can perform hypothesis tests on the fitted model such as

$$H_0 : \beta_2 = 0 \quad H_1 : \beta_2 \neq 0,$$

which is equivalently asking whether or not the regression lines have the same intercept.

Models can be expanded to include multiple indicator variables as long as the matrix $X^{\mathrm{T}}X$ is still invertible. For example, let's suppose we want to look at wages in Alberta with respect to age but partitioned for male vs female and for Edmonton vs Calgary. Then, the model would look like

$$(\text{wage}) = \beta_0 + \beta_1(\text{age}) + \beta_2(\text{Is male?}) + \beta_3(\text{Is from Edmonton?}).$$

In the silly case that our data only consisted of men from Calgary and women from Edmonton, then the final regressor is redundant and $X^{\mathrm{T}}X$ will not be invertible. While this extreme case should not occur, it is possible to have an imbalance in the categories, which we will discuss later.

1.5.2 ANOVA[13]

ANOVA, or the Analysis of Variance, is a slightly overloaded term in statistics. We already considered ANOVA tables when comparing nested models in the hypothesis tests of Section 1.2. However, ANOVA can also be used in the setting of the so-called *One-Way Analysis of Variance*[14]. In this case, we want to compare k samples for equality of the means. For example, we take height measurements from randomly selected citizens from different countries and ask whether or not there is significant evidence to reject the claim that all nations have roughly the same height distribution.

The reason for discussing ANOVA in these notes is that it can be written in a linear regression context as follows. Imagine that we have k different groups of observations with sample sizes n_j, $j = 1, \ldots, k$ for each group. Let $y_{i,j}$ be the ith observation from the jth group where $i \in \{1, \ldots, n_j\}$ and $j \in \{1, \ldots, k\}$. The model is

$$y_{i,j} = \mu_j + \varepsilon_{i,j},$$

which is each observation is just some group mean, μ_j, with the addition of random noise.

From here, one can show that the fitted values as just $\hat{y}_{i,j} = n_j^{-1} \sum_{l=1}^{n_j} y_{l,j}$, which is the jth sample mean. Then, an F-test can be performed similar to that of Section 1.2 to test

$$H_0 : \mu_1 = \ldots = \mu_k \quad H_1 : \exists j_1 \neq j_2 \; s.t. \; \mu_{j_1} \neq \mu_{j_2}.$$

[13] See Montgomery, Peck, and Vining Section 8.3.

[14] https://en.wikipedia.org/wiki/One-way_analysis_of_variance

To reformulate the model to align with our F-test from before, we rewrite it as a linear regression with indicator variables for the regressors

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-1} x_{k-1} + \varepsilon_{i,j}$$

with $\beta_0 = \mu_k$ and $\beta_j = \mu_j - \mu_k$ for $j = 1, \ldots, k-1$. Then, we can test for whether or not there exists at least one $\beta_j \neq 0$ for $j = 1, \ldots, k-1$. Here, the degrees of freedom for the explained sum of squares is $k-1$ and the degrees of freedom for the residual sum of squares is $N - (k-1) - 1 = N - k$ with $N = \sum_{j=1}^{k} n_j$. If all of the n_j are equal, this reduces to $k(n-1)$.

In this case, the vector Y and the design matrix X will take on the form

$$Y = \begin{pmatrix} y_{1,1} \\ y_{2,1} \\ \vdots \\ y_{n_1,1} \\ y_{1,2} \\ \vdots \\ y_{n_k,k} \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 1 & \ldots & 0 \\ 1 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \ldots & 1 \\ 1 & 0 & 0 & \ldots & 1 \\ 1 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \ldots & 0 \\ 1 & 0 & 0 & \ldots & 0 \end{pmatrix}.$$

Chapter 2

Model Assumptions

Introduction

Thus far, we have considered linear regression given some key assumptions. Namely, for models of the form

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

we assume that the noise or errors ε have zero mean, constant finite variance, and are uncorrelated. Furthermore, in order to perform hypothesis tests–F test, t test, partial F-tests–and to construct confidence and prediction intervals as we did in the previous chapter, we further assumes that ε has a normal distribution.

In this chapter, we will consider deviations from these assumptions, which will lead to questions such as

1. What happens if the variance of ε is not constant?
2. What happens if ε has heavier tails than those of a normal distribution?
3. What effect can outliers have on our model?
4. How can we transform our data to correct for some of these deviations?
5. What happens if the true model is not linear?

2.1 Plotting Residuals[1]

In the last chapter, we constructed the residuals as follows. Assume we have a sample of n observations (Y_i, X_i) where $Y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^p$ with p being the number of regressors and $p < n$. The model, as before, is

$$Y = X\beta + \varepsilon$$

[1] See Montgomery, Peck, Vining Section 4.2.

where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times(p+1)}$, $\beta \in \mathbb{R}^{p+1}$, and $\varepsilon \in \mathbb{R}^n$. The least squares estimator is $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$ and the vector of residuals is thus

$$r = (I - P)Y = (I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})Y.$$

We can use the residuals to look for problems in our data with respect to deviations from the assumptions. But first, they should be normalized in some way. As we know from the previous chapter, the covariance of the residuals is $(I - P)\sigma^2$ and that $SS_{\text{res}}/(n - p - 1)$ is an unbiased estimator for the unknown variance σ^2. This implies that while the errors ε_i are assumed to be uncorrelated, the residuals are, in fact, correlated. Explicitly,

$$\text{Var}\,(r_i) = (1 - P_{i,i})\sigma^2, \text{ and } \text{cov}\,(r_i, r_j) = -P_{i,j}\sigma^2 \text{ for } i \neq j$$

where $P_{i,j}$ is the (i, j)th entry of the matrix P. Hence, a standard normalization technique is to write

$$s_i = \frac{r_i}{\sqrt{(1 - P_{i,i})SS_{\text{res}}/(n - p - 1)}},$$

which are denoted as the *studentized residuals*. For a linear model fit in `R` by the function `lm()`, we can extract the residuals with `resid()` and extract the studentized residuals with `rstudent()`.

2.1.1 Plotting Residuals

Studentized Residuals

A plot of studentized residuals from a simple linear regression is displayed in Figure 2.1. Generally, abnormally large studentized residuals indicate that an observation may be an outlier.

Remark 2.1.1. *There are other types of residuals that can be computed such as standardized residuals, PRESS residuals, and externally studentized residuals. These are also used to look for outliers.*

Residuals vs Fitted Values

The residuals and the fitted values as necessarily uncorrelated. That is, $\text{cov}\left(r, \hat{Y}\right) = 0$. However, if ε is not normally distributed, then they may not be independent. Plotting the fitted values against the residuals can give useful diagnostic information about your data.

Figure 2.2 gives four examples of plots of the residuals against the fitted values. The first plot in the top left came from a simple linear regression model where all of the standard assumptions are met. The second plot in the top right came from a

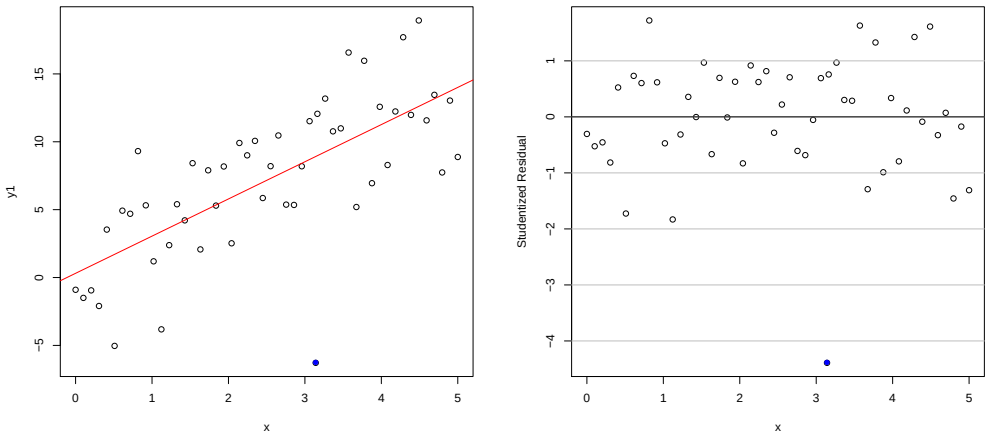

Figure 2.1: On the left, a plot of some data with an outlier labelled in blue and a least squares regression line in red. On the right, the studentized residuals for each of the 51 data points.

simple linear regression but with errors $\varepsilon_i \sim \mathcal{N}\left(0, \sigma_i^2\right)$ where σ_i^2 was increasing in i. Hence, the plot has an expanding look to it. The third plot in the bottom left came from a simple linear regression with the addition of a quadratic term. Fitting a model without the quadratic term still yielded significant test statistics, but failed to account for the nonlinear interaction between x and y. The final plot in the bottom right came from a simple linear regression where the errors were correlated. Specifically, $\operatorname{cov}\left(\varepsilon_i, \varepsilon_j\right) = \min\{x_i, x_j\}$.[2]

Normal Q-Q plots

Another type of plot that can offer insight into your data is the so-called Normal Q-Q plot. This tool plots the studentized residuals against the *quantiles* of a standard normal distribution.

Definition 2.1.2 (Quantile Function)**.** *The quantile function is the inverse of the cumulative distribution function. That is, in the case of the normal distribution, let* $Z \sim \mathcal{N}(0,1)$*. Then the CDF is* $\Phi(z) = \mathrm{P}(Z < z) \in (0,1)$ *for* $z \in \mathbb{R}$*. The quantile function is* $\Phi^{-1}(t) \in [-\infty, \infty]$ *for* $t \in [0,1]$*.*[3]

For a normal Q-Q plot, let $s_1, \ldots, s_n$ be the **ordered** studentized residuals, so that $s_1 \leq \ldots \leq s_n$. The theoretical quantiles are denoted $q_1, \ldots, q_n$ where

[2] Fun fact: This is actually the covariance of Brownian motion.
[3] For more details, see https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot

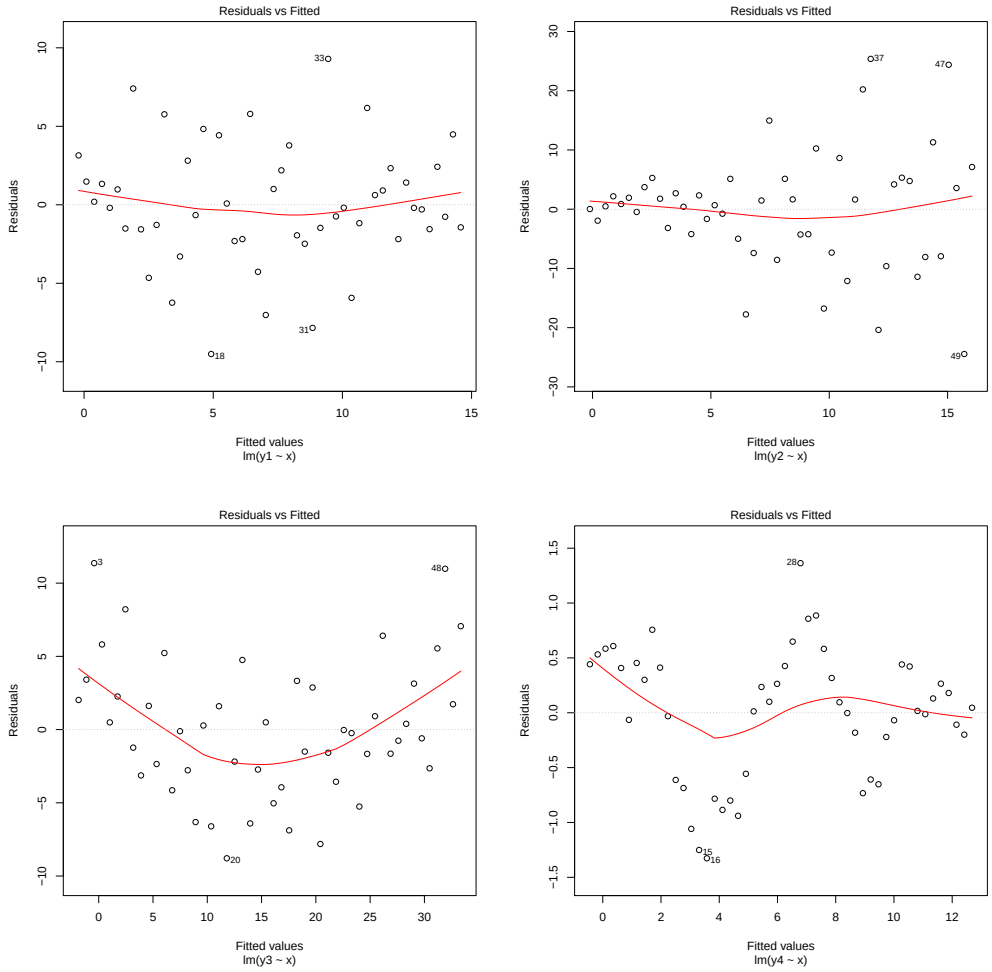

Figure 2.2: For examples of plotting residuals vs fitted values. Top left is when assumptions hold. Top right has a non-constant variance. Bottom left has nonlinear terms. Bottom right has correlated errors.

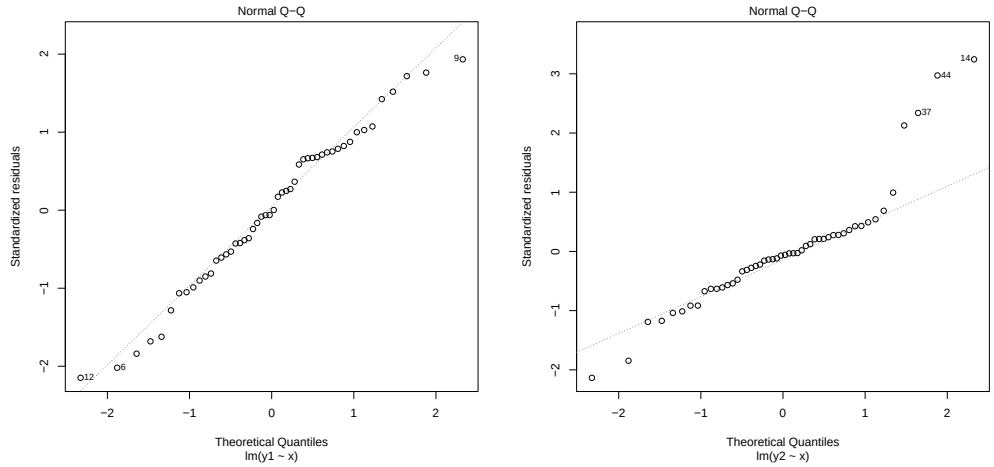

Figure 2.3: On the left, a normal Q-Q plot for data from a simple linear regression with ε having a normal distribution. On the left, a normal Q-Q plot for data from a simple linear regression with ε having a t distribution.

$q_i = \Phi^{-1}(i/(n+1))$. In R, a slightly different formula is used. Figure 2.3 compares two normal Q-Q plots. On the left, the errors are normally distributed and the ordered residuals roughly follow the gray line. On the right, the errors have a t distribution which has heavier tails than expected. Hence, we see deviation from the gray line for the extreme residuals.

Remark 2.1.3. *As noted in Montgomery, Peck, & Vining, these are not always easy to interpret and often fail to capture non-normality in the model. However, they seem to be very popular nonetheless.*

2.2 Transformations[4]

Often, data does not follow all of the assumptions of the ordinary least squares regression model. However, it is often possible to transform data in order to correct for this deviations. We will consider such methods in the following subsections. One dissatisfying remark about such methods is that they often are applied "empirically", which less euphemistically means in an add-hoc way. Sometimes there may be genuine information suggesting certain methods for transforming data. Other times, transforms are chosen because they seem to work.

[4] See Montgomery, Peck, Vining Chapter 5.

2.2.1 Variance Stabilizing[5]

One of the major requirements of the least squares model is that the variance of the errors is constant, which is that $\text{Var}(y_i) = \text{Var}(\varepsilon_i) = \sigma^2$ for $i = 1, \ldots, n$. Mainly, when σ^2 is non-constant in y, problems can occur. Our goal is thus to find some transformation $T(y)$ so that $\text{Var}(T(y_i))$ is constant for all $i = 1, \ldots, n$.

Such a transformation $T(\cdot)$ can be determined through a tool known as the *delta method*,[6] which is beyond the scope of these notes. However, we will consider a simplified version for our purposes. For simplicity of notation, we write $\text{E}Y = \mu$ instead of $\text{E}Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$. Furthermore, assume that T is twice differentiable.[7] Then, Taylor's theorem says that

$$T(Y) = T(\mu) + T'(\mu)(Y - \mu) + \frac{T''(\xi)}{2}(Y - \mu)^2$$

for some ξ between Y and μ. We can, with a little hand-waving, ignore the higher order remainder term and just write

$$T(Y) \approx T(\mu) + T'(\mu)(Y - \mu),$$

which implies that

$$\text{E}T(Y) \approx T(\mu) \quad \text{and} \quad \text{Var}(T(Y)) \approx T'(\mu)^2 \text{Var}(Y).$$

We want a transformation such that $\text{Var}(T(Y)) = 1$ is constant. Meanwhile, we assume that the variance of Y is a function of the mean μ, which is $\text{Var}(Y) = s(\mu)^2$. Hence, we need to solve

$$1 = T'(\mu)^2 s(\mu)^2 \quad \text{or} \quad T(\mu) = \int \frac{1}{s(\mu)} d\mu.$$

Example 2.2.1. *For a trivial example, assume that $s(\mu) = \sigma$ is already constant. Then,*

$$T(\mu) = \int \frac{1}{\sigma} d\mu = \mu/\sigma.$$

Thus, T is just scaling by σ to achieve a unit variance.

Example 2.2.2. *Now, consider the nontrivial example with $s(\mu) = \sqrt{\mu}$, which is that the variance of Y is a linear function of μ. Then,*

$$T(\mu) = \int \frac{1}{\sqrt{\mu}} d\mu = 2\sqrt{\mu}.$$

This is the square root transformation, which is applied to, for example, Poisson data. The coefficient of 2 in the above derivation can be dropped as we are not concerned with the scaling.

[5] See Montgomery, Peck, Vining Section 5.2.

[6] See Chapter 3, Asymptotic Statistics, A W van der Vaart or https://en.wikipedia.org/wiki/Delta_method.

[7] We only require the second derivative for a more pleasant exposition.

Given that we have found a suitable transform, we can then apply it to the data $y_1, \ldots, y_n$ to get a model of the form

$$T(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

where the variance of ε is constant–i.e not a function of the regressors.

2.2.2 Linearization[8]

Another key assumption is that the relationship between the regressors and response, x and y, is linear. If there is reason to believe that

$$y = f(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon)$$

and that $f(\cdot)$ is invertible, then we can rewrite this model as

$$y' = f^{-1}(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

and apply our usual linear regression tools.

An example from the textbook, which is also quite common in practice, is to assume that

$$y = c\mathrm{e}^{\beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon}$$

and apply a logarithm to transform to

$$y' = \log y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

where $\beta_0 = \log c$. Furthermore, if ε has a normal distribution then e^{ε} has a log-normal distribution. This model is particularly useful when one is dealing with exponential growth in some population.

Remark 2.2.3. *Linearization by applying a function to y looks very similar to the variance stabilizing transforms of the previous section. In fact, such transforms have an effect on both the linearity and the variance of the model and should be used with care. Often non-linear methods are preferred.*

Sometimes it is beneficial to transform the regressors, x's, as well. As we are not treating them as random variables, there are less problems to consider.

$$y = \beta_0 + \beta_1 f(x) + \varepsilon \implies y = \beta_0 + \beta_1 x' + \varepsilon, \quad x = f^{-1}(x')$$

Examples include

$$y = \beta_0 + \beta_1 \log x + \varepsilon \implies y = \beta_0 + \beta_1 x' + \varepsilon, \quad x = \mathrm{e}^{x'}$$
$$y = \beta_0 + \beta_1 x^2 + \varepsilon \implies y = \beta_0 + \beta_1 x' + \varepsilon, \quad x = \sqrt{x'}$$

This second example can be alternatively dealt with by employing polynomial regression to be discussed in a subsequent section.

[8] See Montgomery, Peck, Vining Section 5.3.

2.2.3 Box-Cox and the power transform[9]

In short, Box-Cox is a family of transforms parametrized by some $\lambda \in \mathbb{R}$, which can be optimized via maximum likelihood. Specifically, for $y_i > 0$, we aim to choose the best transform of the form

$$y_i \to y_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y_i & \lambda = 0 \end{cases}$$

by maximizing the likelihood as we did in Chapter 1, but with parameters β, σ^2, and λ.

To do this, we assume that the transformed variables follow all of the usual least squares regression assumptions and hence have a joint normal distribution with

$$f(Y^{(\lambda)}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_{i,\cdot}\beta)^2\right).$$

Transforming $Y \to Y^{(\lambda)}$ is a change of variables with Jacobian

$$\prod_{i=1}^{n} \frac{dy_i}{dy_i}^{(\lambda)} = \prod_{i=1}^{n} y_i^{\lambda-1}.$$

Hence, the likelihood function in terms of X and y is

$$L(\beta, \sigma^2, \lambda | y, X) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^\lambda - X_{i,\cdot}\beta)^2\right) \prod_{i=1}^{n} y_i^{\lambda-1}$$

with log likelihood

$$\log L = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_{i,\cdot}\beta)^2 + (\lambda - 1)\sum_{i=1}^{n}\log y_i.$$

From here, the MLEs for β and σ^2 are solved for as before but are now in terms of the transformed $Y^{(\lambda)}$.

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y^{(\lambda)}$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i^{(\lambda)} - X_{i,\cdot}\hat{\beta})^2 = \frac{SS_{\text{res}}^{(\lambda)}}{n}.$$

Plugging these into the log likelihood gives and replacing all of the constants with some C gives

$$\begin{aligned} \log L &= -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{n}{2} + (\lambda - 1)\sum_{i=1}^{n}\log y_i \\ &= C - \frac{n}{2}\log\hat{\sigma}^2 + \log\left((\prod y_i)^{\lambda-1}\right). \end{aligned}$$

[9] See Montgomery, Peck, Vining Section 5.4.

Defining the geometric mean of the y_i to be $\gamma = (\prod_{i=1}^n y_i)^{1/n}$, we have

$$\begin{aligned}\log L &= C - \frac{n}{2}\log\hat{\sigma}^2 + \frac{n}{2}\log\left(\gamma^{2(\lambda-1)}\right) \\ &= C - \frac{n}{2}\log\left(\frac{\hat{\sigma}^2}{\gamma^{2(\lambda-1)}}\right).\end{aligned}$$

Considering the term inside the log, we have that it is just the residual sum of squares from the least squares regression

$$\frac{Y^{(\lambda)}}{\gamma^{\lambda-1}} = X\theta + \varepsilon$$

where $\theta \in \mathbb{R}^{p+1}$ is a transformed version of the original β. Hence, we can choose $\hat{\lambda}$ by maximizing the log likelihood above, which is equivalent to minimizing the residual sum of squares for this new model. This can be calculated numerically in statistical programs like R.

2.2.4 Cars Data[10]

To test some of these linearization techniques, we consider the cars dataset, which is included in the standard distribution of `R`. It consists of 50 observations of a car's speed and a car's stopping distance. The goal is to model and predict the stopping distance given the speed of the car. Such a study could be used, for example, for influencing speed limits and other road rules for the sake of public safety. The observed speeds range from 4 to 25 mph.

We can first fit a simple regression to the data with `lm( dist~speed, data=cars)`. This results in a significant p-value, an $R^2 = 0.651$, and an estimated model of

$$(\text{dist}) = -17.6 + 3.9(\text{speed}) + \varepsilon.$$

If we wanted to extrapolate a bit, we can use this model to predict the stopping distance for a speed of 50 mph, which is 179 feet.

We could stop here and be happy with a significant fit. However, looking at the data, there seems to be a nonlinear relationship between speed and stopping distance. Hence, we could try to fit a model with the response being the square root of the stopping distance: `lm( sqrt(dist)~speed, data=cars)`. Doing so results in

$$\sqrt{(\text{dist})} = 1.28 + 0.32(\text{speed}) + \varepsilon.$$

In this case, we similarly get a significant p-value and a slightly higher $R^2 = 0.709$. The prediction for the stopping distance for a speed of 50 mph is now the much higher 302 feet.

[10] https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/cars.html

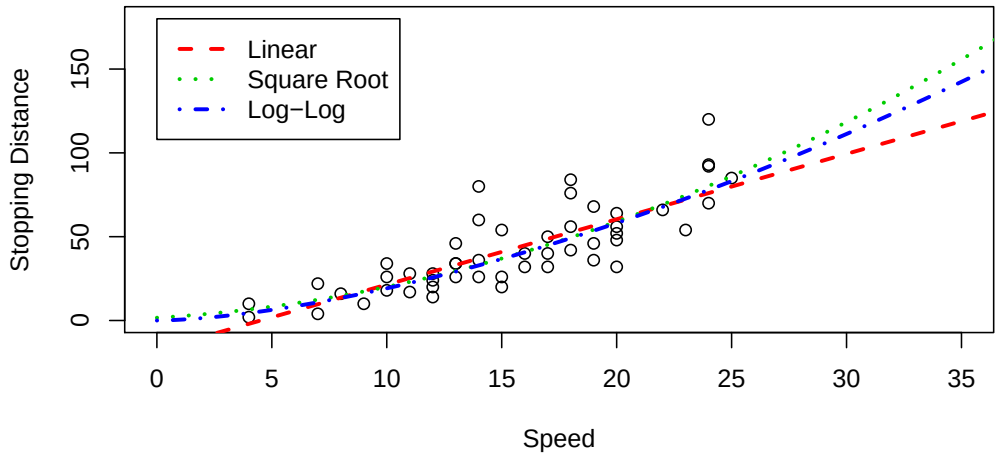

Figure 2.4: A plot of the cars dataset with three different regressions fit to it: simple regression; square root transform; log-log transform.

We can further apply a log-log transform with `lm( log(dist)~log(speed), data=cars)`. This results in an even higher $R^2 = 0.733$. The fitted model is

$$\log(y) = -0.73 + 1.6 \log(x) + \varepsilon,$$

and the predicted stopping distance for a car travelling at 50 mph is 254 feet. Note that the square root transform is modelling $y \propto x^2$ whereas the log-log transform is modelling $y \propto x^{1.6}$, which is a slower rate of increase.

The three models considered are plotted in Figure 2.4. As we move away from the range of the regressors–i.e 4 to 25 mph–the models begin to diverge making extrapolating a bit dangerous.

2.3 Polynomial Regression[11]

In the following subsections, we will only consider models with a single regressor $x \in \mathbb{R}$ until Section 2.3.3 where we will consider polynomials in multiple variables $x_1, \ldots, x_p$.

One of the examples of atypical residual bahaviour from Figure 2.2 is occurs when there is an unaccounted for quadratic term in the model such as trying to fit a

[11] See Montgomery, Peck, Vining Chapter 7

model of the form

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

to data generated by

$$y = \beta_0 + \beta_1 x_1^2 + \varepsilon.$$

If this is suspected, we can look at the residuals and try to transform x, such as $x \to \sqrt{x}$, in order to put this back into the linear model framework. However, we can also just fit a polynomial model to our data.

Consider the setting where we observe n data pairs $(y_i, x_i) \in \mathbb{R}^2$. We can then attempt to fit a model of the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p + \varepsilon$$

to the observations. The $n \times p$ design matrix[12] will look like

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^p \\ 1 & x_2 & x_2^2 & \ldots & x_2^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^p \end{pmatrix}$$

and the parameters can be estimated as usual: $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$.

Remark 2.3.1. *While the columns of X are, in general, linearly independent, as p gets larger, many problems can occur. In particular, the columns become near linearly dependent resulting in instability when computing $(X^{\mathrm{T}}X)^{-1}$.*

2.3.1 Model Problems[13]

Polynomial regression is very powerful for modelling, but can also lead to very erroneous results if used incorrectly. What follows are some potential issues to take into consideration.

Overfitting

As a general rule, the degree of the polynomial model should be kept as low as possible. High order polynomials can be misleading as they can often fit the data quite well. In fact, given n data points, it is possible to fit an $n-1$ degree polynomial that passes through each data point. In this extreme case, all of the residuals would be zero, but we would never expect this to be the correct model for the data.

[12] This matrix arises in Linear Algebra as the Vandermonde matrix https://en.wikipedia.org/wiki/Vandermonde_matrix.

[13] See Montgomery, Peck, Vining Section 7.2.1

Problems can occur even when p is much smaller than n. As an example, two models were fit to $n = 50$ data points generated from the model

$$y = 3x^2 + \varepsilon$$

with $\varepsilon \sim \mathcal{N}(0, 1)$. The first was a cubic model. The second was a degree 20 model. The first regression resulted in three significant regressors. Note that in these two cases, orthogonal polynomials were used to maintain numerical stability.

	Estimate	Std. Error	t value	Pr(>|t|)	
Intercept	4.0856	0.1498	27.276	< 2e-16	***
Degree 1	25.0368	1.0592	23.638	< 2e-16	***
Degree 2	4.6890	1.0592	4.427	5.84e-05	***

The second regression resulted in five significant regressors.

	Estimate	Std. Error	t value	Pr(>|t|)	
Intercept	4.08562	0.13253	30.829	< 2e-16	***
Degree 1	25.03676	0.93709	26.717	< 2e-16	***
Degree 2	4.68903	0.93709	5.004	2.51e-05	***
Degree 15	-2.79288	0.93709	-2.980	0.00578	**
Degree 17	-2.67700	0.93709	-2.857	0.00784	**

Furthermore, the ANOVA table seems to indicate that the addition of these variables may be useful.

```
Model 1:  y ~ poly(x, 3)
Model 2:  y ~ poly(x, 20)
   Res.Df   RSS    Df  Sum of Sq   F      Pr(>F)
1   46     51.606
2   29     25.466  17   26.139    1.751  0.08954  .
```

However, plotting the data in Figure 2.5 with the two different regression lines demonstrates the overfitting problem with the very high order second model.

Extrapolating

Figure 2.5 also indicates problems that can occur when extrapolating with polynomial models. When high degrees are present, the best fit curve can change directions quickly and even make impossible or illogical predictions.

Beyond that, even when the fitted polynomial models all trend in the same general direction, polynomials of different degree will diverge quickly from one another. Consider Figure 2.6 where the data was generated from

$$y = 2x + 3x^3 + \varepsilon.$$

All four models displayed generated very significant F tests. However, each one will give very different answers to predicting the value of y when $x = 5$.

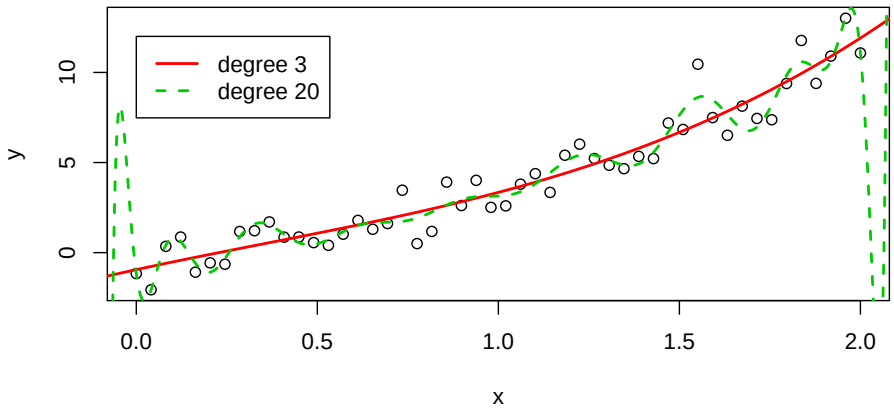

Figure 2.5: A degree 3 and a degree 20 model fit to the same data.

Hierarchy

An hierarchical polynomial model is one such that if it contains a term of degree k, then it will contain all terms of order $i = 0, 1 \ldots, k-1$ as well. In practice, it is not strictly necessary to do this. However, doing so will maintain invariance to linear shifts in the data.

Consider the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$. If we were to shift the values of x by some constant a, then

$$\begin{aligned} y &= \beta_0 + \beta_1(x + a) + \varepsilon \\ &= (\beta_0 + a\beta_1) + \beta_1 x + \varepsilon \\ &= \beta_0' + \beta_1 x + \varepsilon \end{aligned}$$

and we still have the same model but with a modified intercept term.

Now consider the polynomial regression model $y = \beta_0 + \beta_2 x^2 + \varepsilon$. If we were to similarly shift the values of x by some constant a, then

$$\begin{aligned} y &= \beta_0 + \beta_2(x + a)^2 + \varepsilon \\ &= (\beta_0 + a^2\beta_2) + 2\beta_2 a x + \beta_2 x^2 + \varepsilon \\ &= \beta_0' + \beta_1' x + \beta_2 x^2 + \varepsilon. \end{aligned}$$

Now, our model has a linear term, that is $\beta_1' x$, in it, which was not there before.

In general, for a degree p model, if x is shifted by some constant a, then

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x^i \implies y = \beta_0' + \sum_{i=1}^{p} \beta_i' x^i.$$

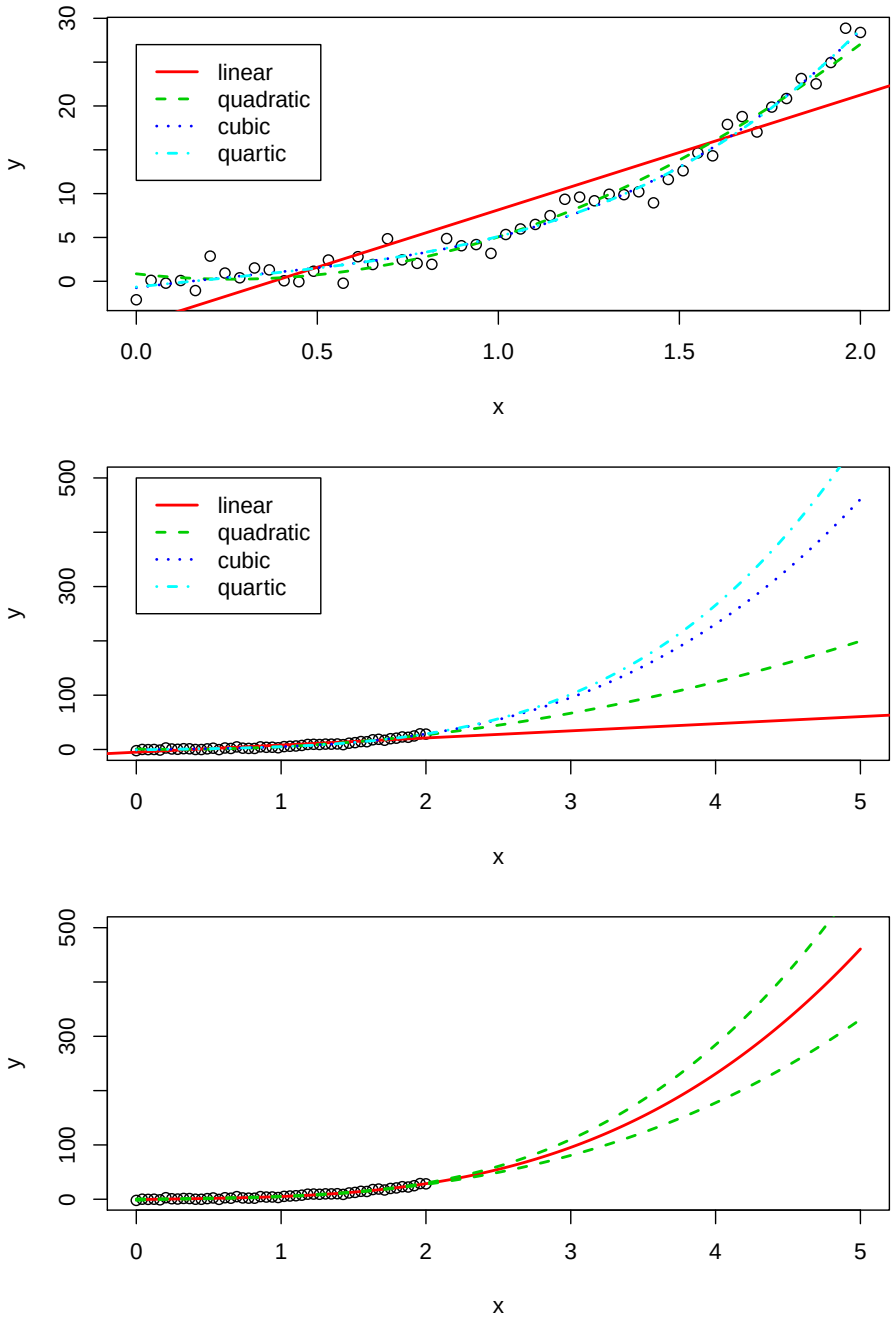

Figure 2.6: Top: four different regressions fit to the data of increasing degree. Middle: the problem of extrapolating with higher degree polynomial regressions. Bottom: a confidence interval for the mean of the cubic model.

Thus, the model is invariant under linear translation.

2.3.2 Piecewise Polynomials[14]

While a single high degree polynomial can be fit very closely to the observed data, there exist the already discussed problems of overfitting and subsequent extrapolation. Hence, an alternative to capture the behaviour of highly nonlinear data is to apply a piecewise polynomial model, which is often referred to as a spline model.

To begin, assume that the observed regressors take values in the interval $[a, b]$. Then, we partition the interval with $k+1$ *knots* by $a = t_1 < t_2 < \ldots < t_{k+1} = b$. The general spline model of order p takes on the form

$$y = \sum_{j=1}^{k} \beta_{0,j} \mathbf{1}[x \geq t_j] + \sum_{i=1}^{p} \sum_{j=1}^{k} \beta_{i,j} (x - t_j)_+^i \tag{2.3.1}$$

where $\mathbf{1}[x > t_j]$ is the indicator function that takes on a value of 0 when $x < t_j$ and a value of 1 otherwise, and where $(x - t_j)_+^i = (x - t_j)^i \mathbf{1}[x > t_j]$.

Remark 2.3.2. *Equation 2.3.1 is equivalent to fitting a separate pth order polynomial to the data in each interval $[t_j, t_{j+1}]$. While doing so does result in many parameters to estimate, it will allow us to perform hypothesis tests for continuity and differentiability of the data, which we will discuss in the next section.*

Submodels of 2.3.1 have a collection of interesting properties. If the coefficients $\beta_{0,j}$ are set to 0, then we will fit a model that ensures continuity at the knots. For example, the model

$$y = \beta_{0,1} + \sum_{j=1}^{k} \beta_{1,j} (x - t_j)_+$$

is a piecewise linear model with continuity at the knots. Conversely, a model of the form

$$y = \sum_{j=1}^{k} \beta_{0,j} \mathbf{1}[x \geq t_j]$$

fits a piecewise constant model and can be used to look for change points in an otherwise constant process.

In practice, spline models with only cubic terms are generally preferred as they have a high enough order to ensure a good amount of smoothness–i.e. the curves are twice continuously differentiable–but generally do not lead to overfitting of the data.

[14] See Montgomery, Peck, Vining Section 7.2.2

An Example

Consider a model of the form

$$y = 2 + 3x - 4x^5 + x^7 + \varepsilon$$

with $\varepsilon \sim \mathcal{N}(0, 4)$ with $n = 50$ observations with regressor $x \in [0, 2]$. A degree 4 and degree 7 polynomial regression were fit to the data, and the results are displayed in the top plot of Figure 2.7. These deviate quickly from the true curve outside of the range of the data. The middle plot of Figure 2.7 fits a piecewise linear spline with 6 knots–i.e. $k = 5$. The model is of the form

$$y = \beta_{0,1} + \beta_{1,1}(x - 0.0)_+ + \beta_{1,2}(x - 0.4)_+ + \ldots + \beta_{1,5}(x - 1.6)_+.$$

The bottom plot of Figure 2.7 fits a spline with only piecewise cubic terms and a spline with cubic and all lower order terms. The only cubic model provides a very reasonable approximation to the data. The full spline model becomes a bit crazy.

Hypothesis testing for spline models

It is useful to consider what the hypothesis tests mean in the context of spline models. For example, consider fitting the piecewise constant model to some data:

$$y = \sum_{j=1}^{k} \beta_{0,j} \mathbf{1}[x \geq t_j].$$

The usual F-test will consider the hypotheses[15]

$$H_0 : \beta_{0,2} = \ldots = \beta_{0,k} = 0 \qquad H_1 : \exists i \in \{2, 3, \ldots, k\} \text{ s.t. } \beta_{0,i} \neq 0.$$

This hypothesis is asking whether or not we believe the mean of the observations changes as x increases.

We can also compare two different spline models with a partial F-test. For example,

$$\begin{aligned} &\text{Model 1:} && y = \beta_{0,1} + \textstyle\sum_{j=1}^{k} \beta_{3,j}(x - t_j)_+^3 \\ &\text{Model 2:} && y = \beta_{0,1} + \textstyle\sum_{j=2}^{k} \beta_{0,j}\mathbf{1}[x \geq t_j] + \sum_{j=1}^{k} \beta_{3,j}(x - t_j)_+^3 \end{aligned}$$

The partial F-test between models 1 and 2 asks whether or not the addition of the piecewise constant terms adds any explanatory power to our model. Equivalently, it is testing for whether or not the model has any discontinuities in it. Similarly, first order terms can be used to test for differentiability.

Instead of constructing a bigger model by adding polynomial terms of different orders, it is also possible to increase the number of knots in the model. Similar to all of the past examples, we will require that the new set of knots contains the old set–that is, $\{t_1, \ldots, t_k\} \subset \{s_1, \ldots, s_m\}$. This requirement ensures that our models are nested and thus can be compared in an ANOVA table.

[15] Note that $\beta_{0,1}$ is the overall intercept term in this model.

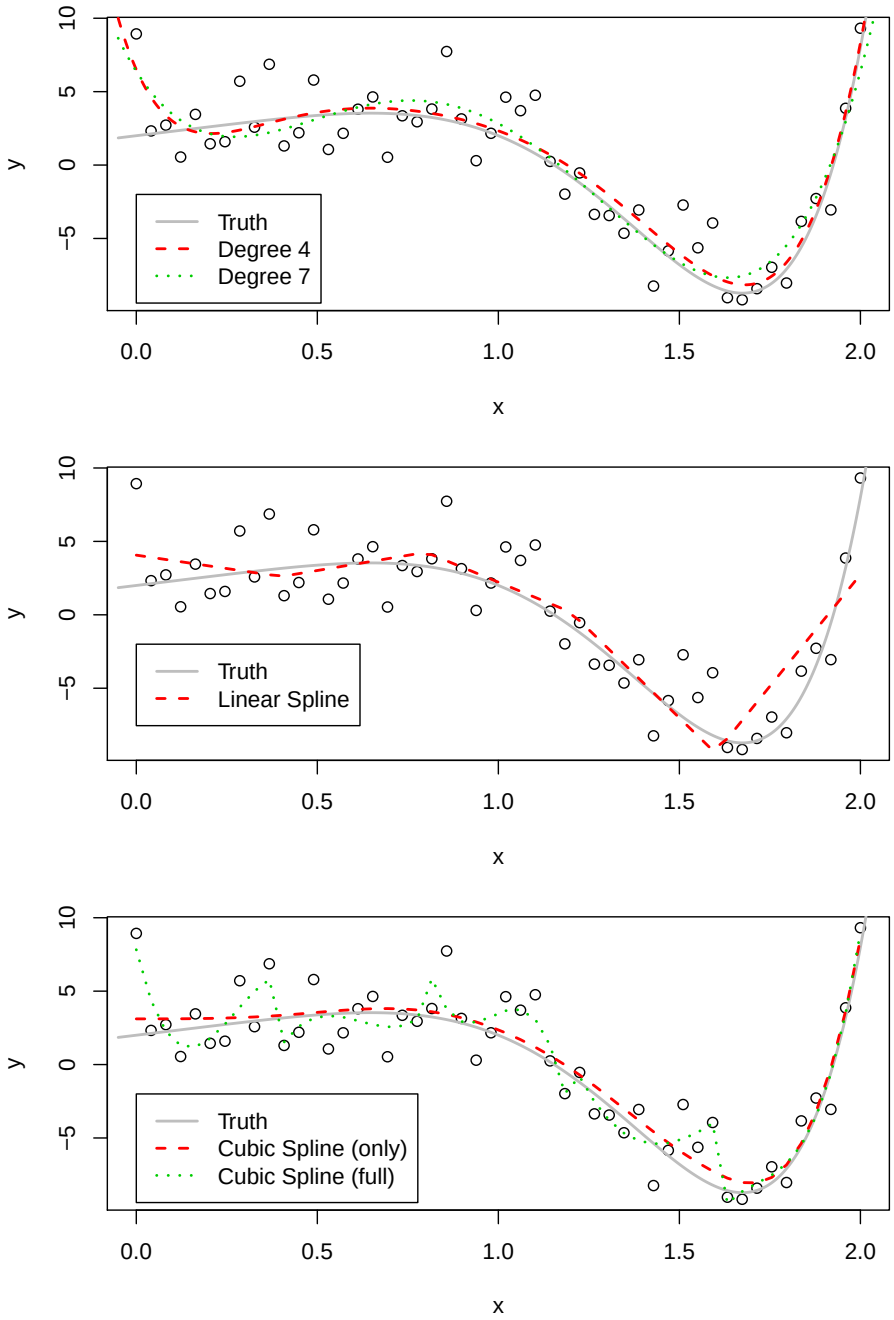

Figure 2.7: Top: A degree 4 and a degree 7 polynomial fit to data generated by $y = 2 + 3x - 4x^5 + x^7 + \varepsilon$. Middle: A linear spline model. Bottom: A spline model with only cubic terms and another including all lower order terms.

B-Splines

In what we considered above, the spline models were comprised of polynomials with supports of the form $[t_j, \infty)$ for knots $t_1 < \ldots < t_k$. However, there are many different families of splines with different desirable properties. Some such families are the B-splines, Non-uniform rational B-splines (NURBS), box splines, Bézier splines, and many others. Here we will briefly consider the family of B-splines due to their simplicity and popularity. Note that in the spline literature, sometimes the knots are referred to as control points.

The ultimate goal of the of the B-splines is to construct a polynomial basis where the constituent polynomials have finite support. Specifically, for an interval $[a, b]$ and knots $a = t_1 < t_2 < \ldots < t_k < t_{k+1} = b$, the constant polynomials will have support on two knots such as $[t_1, t_2]$, the linear terms on three knots such as $[t_1, t_3]$, and so on up to degree p terms, which require $p + 2$ knots.

B-splines can be defined recursively starting with the constant, or degree 0, terms:

$$B_{j,0}(x) = \mathbf{1}_{x \in [t_j, t_{j+1}]},$$

which takes a value of 1 on the interval $[t_j, t_{j+1}]$ and is 0 elsewhere. From here, the higher order terms can be written as

$$B_{j,i}(x) = \left(\frac{x - t_j}{t_{j+i} - t_j} \right) B_{j,i-1}(x) + \left(\frac{t_{j+i+1} - x}{t_{j+i+1} - t_{j+1}} \right) B_{j+1,i-1}(x).$$

For example, with knots $\{0, 1, 2, 3\}$, we have

$$\begin{array}{lll}
\text{Constant:} & B_{j,0} = \mathbf{1}_{x \in [j, j+1]} & j = 0, 1, 2 \\
\text{Linear:} & B_{j,1} = \begin{cases} (x - j), & x \in [j, j+1] \\ (j + 2 - x), & x \in [j+1, j+2] \end{cases} & j = 0, 1 \\
\text{Quadratic:} & B_{j,2} = \begin{cases} x^2/2, & x \in [0, 1] \\ (-2x^2 + 6x - 3)/2, & x \in [1, 2] \\ (3 - x)^2/2, & x \in [2, 3] \end{cases} & j = 0.
\end{array}$$

The linear and quadratic splines are displayed in Figure 2.8.

Just as we fit a spline model above, we could use the linear regression tools to fit a B-spline model of the form

$$y = \sum_{i=0}^{p} \sum_{j=1}^{k-i} \beta_{i,j} B_{j,i}(x).$$

Here, we require that $k > p$ as we will at least need knots $t_1, \ldots, t_{p+1}$ for a pth degree polynomial. The total number of terms in the regression, which is number of parameters $\beta_{i,j}$ to estimate, is

$$k + (k - 1) + \ldots + (k - p).$$

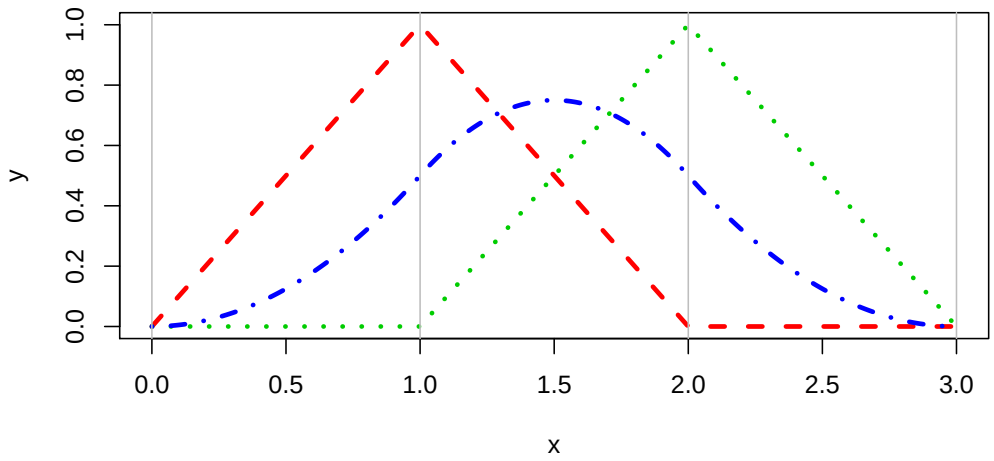

Figure 2.8: An example of linear and quadratic B-splines.

2.3.3 Interacting Regressors[16]

Thus far in this section, we have considered only polynomial models with a single regressor. However, it is certainly possible to fit a polynomial model for more than one regressor such as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \beta_{22} x_2^2 + \varepsilon.$$

Here we have linear and quadratic terms for x_1 and x_2 as well as an interaction term $\beta_{12} x_1 x_2$.

Fitting such models to the data follows from what we did for single variable polynomials. In this case, the number of interaction terms can grow quite large in practice. With k regressors, $x_1, \ldots, x_k$, there will be $\binom{k}{p}$ interaction terms of degree p assuming $p < k$. This leads to the topic of Response Surface Methodology,[17] which is a subtopic of the field of experimental design.

We will not consider this topic further in these notes. However, it is worth noting that in `R`, it is possible to fit a linear regression with interaction terms such as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3 + \varepsilon$$

with the simple syntax `lm( y~x1*x2*x3 )` where the symbol `*` replaces the usual `+` from before.

[16] See Montgomery, Peck, Vining Section 7.4

[17] https://en.wikipedia.org/wiki/Response_surface_methodology

2.4 Influence and Leverage[18]

The overall intuition for this section is that each observation does not have an equal influence on the estimation of $\hat{\beta}$. If a given observed regressor x lies far from the other observed values, it can have a strong effect on the least squares regression line. The goal of this section is to identify such points or subset of points that have a large influence on the regression.

If we use the `R` command `lm()` to fit a linear model to some data, then we can use the command `influence.measures()` to compute an array of diagnostic metrics for each observation to test its influence on the regression. The function `influence.measures()` computes DFBETAS, DFFITS, covariance ratios, Cook's distances and the diagonal elements of the so-called hat matrix. We will look at each of these in the following subsections.

Remark 2.4.1 (Warning)**.** *You may notice that the word "heuristic" appears often in the following subsections when it comes to identifying observations with significant influence. Ultimately, these are rough guidelines based on the intuition of past statisticians and should not be taken as strict rules.*

2.4.1 The Hat Matrix[19]

The projection or "hat" matrix, $P = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, directly measures the influence of one point on another. This is because under the usual model assumptions, the fitted values have a covariance matrix $\mathrm{var}(\hat{Y}) = \sigma^2 P$. Hence, the i, jth entry in P is a measure of the covariance between the fitted values $\hat{Y}_i$ and $\hat{Y}_j$.

The function `influence.measures()` reports the diagonal entries of the matrix P. Heuristically, any entries that are much larger than the rest will have a strong influence on the regression. More precisely, we know that for a linear model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

we have that $\mathrm{rank}(P) = p + 1$. Hence, $\mathrm{trace}(P) = p + 1$ where the trace of a matrix is the sum of the diagonal entries. Thus, as P is an $n \times n$ matrix, we roughly expect the diagonal entries to be approximately $(p + 1)/n$. Large deviations from this value should be investigated. For example, Montgomery, Peck, & Vining recommend looking at observations with $P_{i,i} > 2(p + 1)/n$.

Remark 2.4.2. *The ith diagonal entry $P_{i,i}$ is referred to as the leverage of the ith observation in Montgomery, Peck, & Vining. However, in `R`, leverage is $P_{i,i}/(1-P_{i,i})$. This sometimes referred to as the leverage factor.*

[18] See Montgomery, Peck, Vining Chapter 6
[19] See Montgomery, Peck, Vining Section 6.2

2.4.2 Cook's D[20]

Cook's D or distance computes the distance between the vector of estimated parameters on all n data points, $\hat{\beta}$, and the vector of estimated parameters on $n-1$ data points, $\hat{\beta}_{(i)}$ where the ith observation has been removed. Intuitively, if the ith observation has a lot of influence on the estimation of β, then the distance between $\hat{\beta}$ and $\hat{\beta}_{(i)}$ should be large.

For a linear model with $p+1$ parameters and a sample size of n observations, the usual form for Cook's D is

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^{\mathrm{T}} X^{\mathrm{T}} X (\hat{\beta}_{(i)} - \hat{\beta})/(p+1)}{SS_{\mathrm{res}}/(n-p-1)}.$$

This is very similar to the confidence ellipsoid from Section 1.3. However, this is not an usual F statistic, so we do not compute a p-value as we have done before. Instead, some heuristics are used to determine what a large value is. Some authors suggest looking for D_i greater than 1 or greater than $4/n$.[21]

Cook's D can be written in different forms. One, in terms of the diagonal entries of P, is

$$D_i = \frac{s_i^2}{p+1}\left(\frac{P_{i,i}}{1-P_{i,i}}\right)$$

where s_i is the ith studentized residual. Another form of this measure compares the distance between the usual fitted values on all of the data $\hat{Y} = X\hat{\beta}$ and the fitted values based on all but the ith observation, $\hat{Y}_{(i)} = X\hat{\beta}_{(i)}$. That is,

$$D_i = \frac{(\hat{Y}_{(i)} - \hat{Y})^{\mathrm{T}} (\hat{Y}_{(i)} - \hat{Y})/(p+1)}{SS_{\mathrm{res}}/(n-p-1)}$$

Note that like $\hat{Y}$, the vector $\hat{Y}_{(i)} \in \mathbb{R}^n$. The ith entry in the vector $\hat{Y}_{(i)}$ is the predicted value of y given x_i.

2.4.3 DFBETAS[22]

The intuition behind DFBETAS is similar to that for Cook's D. In this case, we consider the normalized difference between $\hat{\beta}$ and $\hat{\beta}_{(i)}$. What results is an $n \times (p+1)$ matrix whose ith row is

$$\mathrm{DFBETAS}_i = \frac{\hat{\beta} - \hat{\beta}_{(i)}}{\sqrt{(X^{\mathrm{T}}X)^{-1}_{i,i} SS_{\mathrm{res}(i)}/(n-p-2)}} \in \mathbb{R}^{p+1}$$

[20] See Montgomery, Peck, Vining Section 6.3

[21] https://en.wikipedia.org/wiki/Cook%27s_distance

[22] See Montgomery, Peck, Vining Section 6.4

where $SS_{\text{res}(i)}$ is the sum of the squared residuals for the model fit after removing the ith data point and where $(X^{\mathrm{T}}X)^{-1}_{i,i}$ is the ith diagonal entry of the matrix $(X^{\mathrm{T}}X)^{-1}$. The recommended heuristic is to consider the ith observation as an influential point if the i, jth entry of DFBETAS has a magnitude greater than $2/\sqrt{n}$.

2.4.4 DFFITS[23]

The DFFITS value is very similar to the previously discussed DFBETAS. In this case, we are concerned with by how much the fitted values change when the ith observation is removed. Explicitly,

$$\text{DFFIT} = \frac{\hat{Y} - \hat{Y}_{(i)}}{\sqrt{(X^{\mathrm{T}}X)^{-1}_{i,i} SS_{\text{res}(i)}/(n-p-2)}} \in \mathbb{R}^n.$$

The claim is that DFFIT is effected by both leverage and prediction error. The heuristic is to investigate any observation with DFFIT greater in magnitude than $2\sqrt{(p+1)/n}$.

2.4.5 Covariance Ratios[24]

The covariance ratio whether the precision of the model increases or decreases when the ith observation is included. This measure is based on the idea that a small value for $\det\left[(X^{\mathrm{T}}X)^{-1}SS_{\text{res}}\right]$ indicates high precision in our model. Hence, the covariance ratio considers a ratio of determinants

$$\begin{aligned}\text{covratio}_i = \frac{\det\left[(X^{\mathrm{T}}_{(i)}X_{(i)})^{-1}SS_{\text{res}(i)}/(n-p-2)\right]}{\det\left[(X^{\mathrm{T}}X)^{-1}SS_{\text{res}}/(n-p-1)\right]} = \\ = \left(\frac{SS_{\text{res}(i)}/(n-p-2)}{SS_{\text{res}}/(n-p-1)}\right)^{p+1}\left(\frac{1}{1-P_{i,i}}\right).\end{aligned}$$

If the value is greater than 1, then inclusion of the ith point has increased the model's precision. If it is less than 1, then the precision has decreased. The suggested heuristic threshold for this measure of influence is when the value is greater than $1+3(p+1)/n$ or less than $1-3(p+1)/n$. Though, it is noted that this only is valid for large enough sample sizes.

2.4.6 Influence Measures: An Example

To test these different measures, we create a dataset of $n = 50$ observations from the model

$$y = 3 + 2x + \varepsilon$$

[23] See Montgomery, Peck, Vining Section 6.4
[24] See Montgomery, Peck, Vining Section 6.5

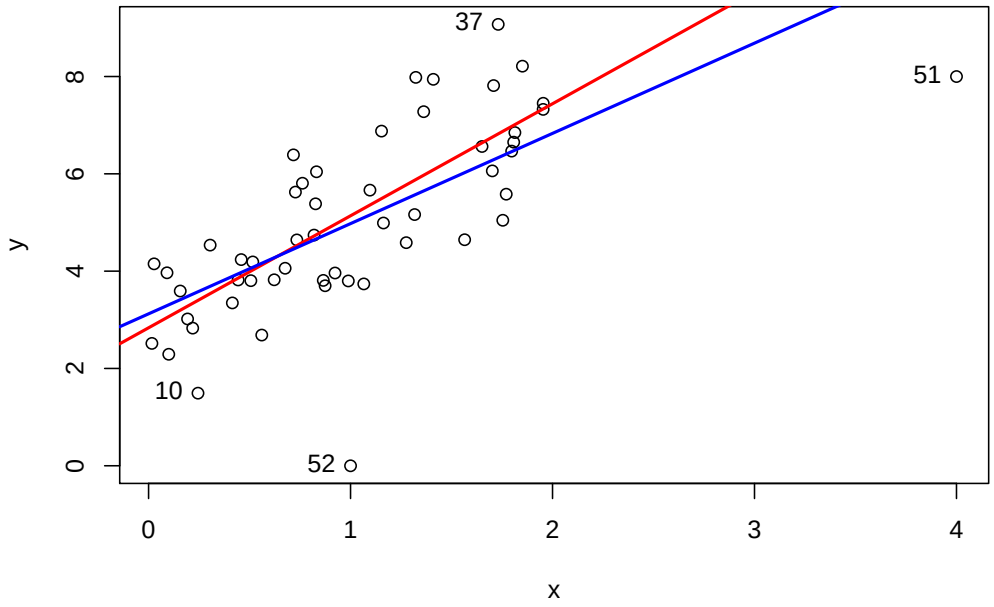

Figure 2.9: A simple regression in red fit to the original 50 data points, and a simple regression in blue fit to the original 50 and 2 anomalous data points.

where $\varepsilon \sim \mathcal{N}(0,1)$ and $x \in [0,2]$. To this dataset, we add 2 anomalous points at $(4,8)$ and at $(1,0)$. Thus, we fit a simple linear regression to the original 50 data points and also to the new set of 52 data points resulting in the red and blue lines, respectively, in Figure 2.9.

We can use the `R` function `influence.measures()` to compute a matrix containing the DFBETAS, DFFITS, covariance ratios, Cook's D, and leverage for each data point. Applying the recommended thresholds in the previous sections results in the following extreme points, which are labelled in Figure 2.9:

Measure	DF β_0	DF β_1	DFFITS	
Extremes	10, 51, 52	37, 51	37, 51, 52	
Measure	Cov Ratio $>$	Cov Ratio $<$	Cook's D	Leverage
Extremes	52	51	37, 51, 52	51

The interpretation of the table is that points...

- 10, 51, 52 have a strong influence on the estimation of $\hat{\beta}_0$;
- 37, 51 have a strong influence on the estimation of $\hat{\beta}_1$;

- 37, 51, 52 have a strong influence on the estimation of $\hat{Y}$;
- 52 significantly increases the precision of the model;
- 51 significantly decreases the precision of the model;
- 37, 51, 52 have large Cook's D values;
- 51 has a large leverage value.

Note that points 10 and 37 are just part of the randomly generated data while points 51 and 52 were purposefully added to be anomalous. Hence, just because an observation is beyond one of these thresholds does not necessarily imply that it lies outside of the model.

2.5 Weighted Least Squares

A key assumption of the Gauss-Markov theorem is that $\varepsilon_i = \sigma^2$ for all $i = 1, \ldots, n$. What happens when $\varepsilon_i = \sigma_i^2$–i.e. when the variance can differ for each observation? Normalizing the errors ε_i can be done by

$$\frac{\varepsilon_i}{\sigma_i} = \frac{1}{\sigma_{i,i}}(Y_i - X_{i,.}\beta) \sim \mathcal{N}(0,1).$$

In Chapter 1, we computed the least squares estimator as the vector $\hat{\beta}$ such that

$$\hat{\beta} = \underset{\tilde{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} (Y_i - X_{i,.}\tilde{\beta})^2$$

Now, we will solve the slighly modified equation

$$\hat{\beta} = \underset{\tilde{\beta} \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \frac{(Y_i - X_{i,.}\tilde{\beta})^2}{\sigma_i^2}.$$

In this setting, dividing by σ_i^2 is the "weight" that gives this method the name weighted least squares.

Proceeding as in chapter 1, we take a derivative with respect to the jth $\tilde{\beta}_j$ to get

$$\begin{aligned}
\frac{\partial}{\partial \tilde{\beta}_j} \sum_{i=1}^{n} \frac{(Y_i - X_{i,.}\tilde{\beta})^2}{\sigma_i^2} &= 2 \sum_{i=1}^{n} \frac{(Y_i - X_{i,.}\tilde{\beta})}{\sigma_i^2} X_{i,j} \\
&= 2 \sum_{i=1}^{n} \frac{Y_i X_{i,j}}{\sigma_i^2} - 2 \sum_{i=1}^{n} \frac{X_{i,.}\tilde{\beta} X_{i,j}}{\sigma_i^2} \\
&= 2 \sum_{i=1}^{n} Y_i' X_{i,j}' - 2 \sum_{i=1}^{n} X_{i,.}'\tilde{\beta} X_{i,j}'
\end{aligned}$$

where $Y'_i = Y_i/\sigma_i$ and $X'_{i,j} = X_{i,j}/\sigma_i$. Hence, the least squares estimator is as before

$$\begin{aligned}\hat{\beta} &= (X'^{\mathrm{T}}X')^{-1}X'^{\mathrm{T}}Y' \\ &= (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}WY\end{aligned}$$

where $W \in \mathbb{R}^{n\times n}$ is the diagonal matrix with entries $W_{i,i} = \sigma_i^2$.

In practise, we do not know the values for σ_i^2. Methods to find a good matrix of weights W exist such as Iteratively Reweighted Least Squares,[25] which is equivalent to finding the estimator

$$\hat{\beta} = \underset{\tilde{\beta}\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} |Y_i - X_{i,\cdot}\tilde{\beta}|^q$$

for some $q \in [1, \infty)$.

[25] https://en.wikipedia.org/wiki/Iteratively_reweighted_least_squares

Chapter 3

Model Building

Introduction

The following sections will discuss various topics regarding constructing a good representative regression model for your data. Three main topics will be considered/

First, multicollinearity deals with the problem of linear relationships between your regressions. We already know that we require the columns of the design matrix to be linearly independent in order to solve for the least squares estimate. However, it is possible to have near dependencies among the columns. This can lead to numerical stability issues and unnecessary redundancy among the regressors.

Second, there are many different variable selection techniques in existence. Given a large number of regressors to can be included in a model, the question is, which should and which should not be included? We will discuss various techniques such as forward and backward selection as well as different tools for comparing models.

Third, penalized regression will be discussed. This section introduces two modern and quite powerful approaches to linear regression: ridge regression from the 1970's and LASSO from the 1990's. Both arise from modifying how we estimate the parameter vector $\hat{\beta}$. Up until now, we have chosen $\hat{\beta}$ to minimize the sum of the squared error. Now, we will add a penalty term to this optimization problem, which will encourage choices of $\hat{\beta}$ with small-in-magnitude or just zero entries.

3.1 Multicollinearity[1]

The concept of multicollinearity is intuitively simple. Say we have a model of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

[1] See Montgomery, Peck, Vining Section 9.3

This results in a design matrix of the form

$$X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} \end{pmatrix}$$

Then, we can consider a new model of the form

$$x_2 = \alpha_0 + \alpha_1 x_1 + \varepsilon.$$

If this simple regression has a strong fit,[2] then the addition of the regressor x_2 to the original model is unnecessary as almost all of the explanatory information provided by x_2 with regards to predicting y is already provided by x_1. Hence, the inclusion of x_2 in our model is superfluous.

Taking a more mathematical approach, it can be shown that such near linear dependencies lead to a very high variance for the least squares estimator $\hat{\beta}$. Furthermore, the magnitude of the vector is much larger than it should be.

Assuming that the errors have a covariance matrix $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$, then we have from before that

$$\mathrm{Var}\left(\hat{\beta}\right) = \mathrm{Var}\left((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y\right) = \\ = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{Var}(Y)\,X(X^{\mathrm{T}}X)^{-1} = \sigma^2(X^{\mathrm{T}}X)^{-1}.$$

With some effort, it can be shown that the diagonal entries of the matrix $(X^{\mathrm{T}}X)^{-1}$ are equal to $(1-R_0^2)^{-1}, \ldots, (1-R_p^2)^{-1}$ where R_j^2 is the coefficient of determination for the model

$$x_j = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_{j-1}x_{j-1} + \alpha_{j+1}x_{j+1} + \ldots + \alpha_p x_p + \varepsilon,$$

which is trying to predict the jth regressor by the other $p-1$ regressors. If the remaining regressors are good predictors for x_j, then the value R_j^2 will be close to 1. Hence,

$$\mathrm{Var}\left(\hat{\beta}_j\right) = \frac{\sigma^2}{1-R_j^2}$$

will be very large.

Furthermore, this implies that the expected Euclidean distance between $\hat{\beta}$ and β will be quite large as well. Indeed, we have

$$\mathrm{E}\left((\hat{\beta}-\beta)^{\mathrm{T}}(\hat{\beta}-\beta)\right) = \sum_{i=0}^{p}\mathrm{E}\left(\hat{\beta}_i-\beta_i\right)^2 = \sum_{i=0}^{p}\mathrm{Var}\left(\hat{\beta}_i\right) = \sigma^2\mathrm{tr}\left((X^{\mathrm{T}}X)^{-1}\right)$$

[2] A significant F-test or R^2 value, for example.

where $\operatorname{tr}(\cdot)$ denotes the trace of a matrix–i.e. the sum of the diagonal entries. Hence, if at least one of the R_j^2 is close to 1, then the expected distance from our estimator to the true β will be quite large.

The trace of a matrix is also equal to the sum of its eigenvalues. Hence, if we denote the eigenvalues of $X^{\mathrm{T}}X$ by $\lambda_1, \ldots, \lambda_{p+1}$, then

$$\operatorname{tr}\left((X^{\mathrm{T}}X)^{-1}\right) = \sum_{i=1}^{p+1} \lambda_i^{-1}.$$

Hence, an equivalent condition to check for multicollinearity is the presence of eigenvalues of $X^{\mathrm{T}}X$ very close to zero, which would make the above sum very large.

3.1.1 Identifying Multicollinearity[3]

To identify the presence of multicollinearity in our linear regression, there are many measures to consider.

We already established that near linear dependencies will result in large values for the diagonal entries of $(X^{\mathrm{T}}X)^{-1}$. These values are known as the *Variance Inflation Factors* and sometimes written as $\mathrm{VIF}_i = (1 - R_i^2)^{-1}$.

An interesting interpretation of the VIF is in terms of confidence intervals. Recall that for β_j, we can construct a $1-\alpha$ confidence interval as

$$-t_{\alpha/2,n-p-1}\sqrt{(X^{\mathrm{T}}X)^{-1}_{j,j}\frac{SS_{\mathrm{res}}}{n-p-1}} \leq \beta_j - \hat{\beta}_j \leq t_{\alpha/2,n-p-1}\sqrt{(X^{\mathrm{T}}X)^{-1}_{j,j}\frac{SS_{\mathrm{res}}}{n-p-1}}.$$

If all $i \neq j$ regressors are orthogonal to the jth regressor, then $R_j^2 = 0$ and the term $(X^{\mathrm{T}}X)^{-1}_{j,j} = 1$. Under multicollinearity, $(X^{\mathrm{T}}X)^{-1}_{j,j} \gg 1$. Hence, the confidence interval is expanded by a factor of $\sqrt{(X^{\mathrm{T}}X)^{-1}_{j,j}}$ when the regressors are not orthogonal.

We can alternatively examine the eigenvalues of the matrix $X^{\mathrm{T}}X$. Recall that finding the least squares estimator is equivalent to solving a system of linear equations of the form

$$X^{\mathrm{T}}X\hat{\beta} = X^{\mathrm{T}}Y.$$

To measure to stability of a solution to a system of equations to small perturbations, a term referred to as the *condition number* is used.[4] It is

$$\kappa = \lambda_{\max}/\lambda_{\min}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximal and minimal eigenvalues, respectively. According to Montgomery, Peck, & Vining, values of κ less than 100 are not significant whereas values greater than 1000 indicate severe multicollinearity.

[3] See Montgomery, Peck, Vining Sections 9.4.2 & 9.4.3

[4] This term arises in more generality in numerical analysis https://en.wikipedia.org/wiki/Condition_number

If the minimal eigenvalue is very small, we can use the corresponding eigenvector to understand the nature of the linear dependency. That is, consider the eigenvector $u = (u_0, u_1, \ldots, u_p)$ for the matrix $X^{\mathrm{T}}X$ corresponding to the eigenvalue $\lambda_{\min}$. Recall that this implies that

$$(X^{\mathrm{T}}X)u = \lambda_{\min}u \approx 0,$$

which is approximately zero because $\lambda_{\min}$ is close to zero. Hence, for regressors $1, x_1, \ldots, x_p$,

$$u_0 + u_1x_1 + \ldots + u_px_p \approx 0.$$

Thus, we can use the eigenvectors with small eigenvalues to get a linear relationship between the regressors.

Remark 3.1.1. *If you are familiar with the concept of the Singular Value Decomposition, then you could alternatively consider the ratio between the maximal and minimal singular values of the design matrix X.*[5] *Furthermore, you can also analyze the singular vectors instead of the eigen vectors.*

3.1.2 Correcting Multicollinearity

Ideally, we would design a model such that the columns of the design matrix X are linearly independent. Of course, in practise, this is often not achievable. When confronted with real world data, there are still some options available.

First, the regressors can be *respecified*. That is, if x_1 and x_2 are near linearly related, then instead of including both terms in the model, we can include a single combination term like x_1x_2 or $(x_1 + x_2)/2$. Second, one of the two variables can be dropped from the model, which will be discussed below when we consider variable selection in Section 3.2.

More sophisticated solutions to this problem include penalized regression techniques, which we will discuss in Section 3.3. Also, principal components regression[6] and partial least squares are two other methods that can be applied to deal with multicollinear data.

Remark 3.1.2. *A common thread among all of these alternatives is that they result in a biased estimate for β unlike the usual least squares estimator. Often in statistics, we begin with unbiased estimators, but can often achieve a better estimator by adding a small amount of bias. This is the so-called bias-variance tradeoff.*[7]

[5] https://en.wikipedia.org/wiki/Singular-value_decomposition

[6] See Montgomery, Peck, Vining Sections 9.5.4 for more on PC regression

[7] https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff

3.2 Variable Selection[8]

In general, if we have p regressors, we may want to build a model consisting only of the best regressors for modelling the response variable. In some sense, we could compare all possible subset models. However, there are many issues with this, which we will address in the following subsections. First, what are the effects of removing regressors from your model? Second, how do we compare models if they are not nested? Third, there are 2^p possible models to consider. Exhaustively fitting and comparing all of these models may be computational impractical or impossible. Hence, how do we find a good subset of the regressors?

3.2.1 Subset Models[9]

What happens to the model when we remove some regressors? Assume we have a sample of n observations and $p+q$ regressors and want to remove q of them. The full model would be

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p+q} x_{p+q} + \varepsilon.$$

This can be written in terms of the design matrix and partitioned over the two sets of regressors as

$$\begin{aligned} Y &= X\beta + \varepsilon \\ &= X_p \beta_p + X_q \beta_q + \varepsilon \end{aligned}$$

where $X_p \in \mathbb{R}^{n\times p}$, $X_q \in \mathbb{R}^{n\times q}$, $\beta_p \in \mathbb{R}^p$, $\beta_q \in \mathbb{R}^q$, and

$$X = \begin{pmatrix} X_p & X_q \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_p \\ \beta_q \end{pmatrix}$$

We have two models to compare. The first is the full model, $Y = X\beta + \varepsilon$, where we denote the least squares estimator as $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$ as usual with components

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_p \\ \hat{\beta}_q \end{pmatrix}.$$

The second is the reduced model obtained by deleting q regressors: $Y = X_p\beta_p + \varepsilon$. The least squares estimator for this model will be denoted as $\tilde{\beta}_p = (X_p^{\mathrm{T}}X_p)^{-1}X_p^{\mathrm{T}}Y$

[8] See Montgomery, Peck, Vining Section 10.1.3
[9] See Montgomery, Peck, Vining Section 10.1.2

Bias may increase

The first concern with the reduced model is that the estimator $\tilde{\beta}_p$ can be biased as

$$\mathrm{E}\tilde{\beta}_p = (X_p^{\mathrm{T}}X_p)^{-1}X_p^{\mathrm{T}}\mathrm{E}Y = (X_p^{\mathrm{T}}X_p)^{-1}X_p^{\mathrm{T}}(X_p\beta_p + X_q\beta_q) = \\ = \beta_p + (X_p^{\mathrm{T}}X_p)^{-1}X_p^{\mathrm{T}}X_q\beta_q = \beta_p + A\beta_q.$$

Hence, our reduced estimator is only unbiased in two cases. Case one is when $A = 0$, which occurs if the p regressors and q regressors are orthogonal resulting in $X_p^{\mathrm{T}}X_q = 0$. Case two is when $\beta_q = 0$, which occurs if those regressors have no effect on the given response. If neither of these cases occurs, then $A\beta_q \neq 0$ and represents the bias in our estimator $\tilde{\beta}_p$. Note that the matrix A is referred to as the *alias* matrix.

Variance may decrease

While deleting regressors can result in the addition of bias to our estimate, it can also result in a reduction in the variance of our estimator. Namely,

$$\mathrm{Var}\left(\tilde{\beta}_p\right) = \sigma^2(X_p^{\mathrm{T}}X_p)^{-1}, \text{ while}$$
$$\mathrm{Var}\left(\hat{\beta}_p\right) = \sigma^2(X_p^{\mathrm{T}}X_p)^{-1} + \sigma^2 A[{X_q}^{\mathrm{T}}(I - P_p)X_q]^{-1}A^{\mathrm{T}},$$

where $P_p = X_p(X_p^{\mathrm{T}}X_p)^{-1}X_p^{\mathrm{T}}$.[10] The matrix $A[{X_q}^{\mathrm{T}}(I - P_p)X_q]^{-1}A^{\mathrm{T}}$ is symmetric positive semi-definite, so the variance for $\hat{\beta}_p$ can only be larger than $\tilde{\beta}_p$.

MSE may or may not improve

Generally in statistics, when deciding whether or not the increase in the bias is worth the decrease in the variance, we consider the change in the *mean squared error* (MSE) of our estimate. This is,

$$\begin{aligned}\mathrm{MSE}(\tilde{\beta}_p) &= \mathrm{E}\left((\tilde{\beta}_p - \beta_p)(\tilde{\beta}_p - \beta_p)^{\mathrm{T}}\right) \\ &= \mathrm{E}\left((\tilde{\beta}_p - \mathrm{E}\tilde{\beta}_p + \mathrm{E}\tilde{\beta}_p - \beta_p)(\tilde{\beta}_p - \mathrm{E}\tilde{\beta}_p + \mathrm{E}\tilde{\beta}_p - \beta_p)^{\mathrm{T}}\right) \\ &= \mathrm{var}(\tilde{\beta}_p) + \mathrm{bias}(\tilde{\beta}_p)^2 \\ &= \sigma^2(X_p^{\mathrm{T}}X_p)^{-1} + A\beta_q\beta_q^{\mathrm{T}}A^{\mathrm{T}}.\end{aligned}$$

For the full model,

$$\mathrm{MSE}(\hat{\beta}_p) = \mathrm{var}(\hat{\beta}_p) = \sigma^2(X_p^{\mathrm{T}}X_p)^{-1} + \sigma^2 A[X_q^{\mathrm{T}}(I - P_p)X_q]^{-1}A^{\mathrm{T}},$$

If $\mathrm{MSE}(\hat{\beta}_p) - \mathrm{MSE}(\tilde{\beta}_p)$ is positive semi-definite, then the mean squared error has decreased upon the removal of the regressors in X_q.

[10] This expression can be derived via the formula for inverting a block matrix https://en.wikipedia.org/wiki/Invertible_matrix#Blockwise_inversion

3.2.2 Model Comparison[11]

We have already compared models in Chapter 1 with the partial F-test. However, for that test to make sense, we require the models to be nested–i.e. the larger model must contain all of the parameters of the smaller model. But, given a model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon,$$

we may want to compare two different subset models that are not nested. Hence, we have some different measures to consider.

Note that ideally, we would compare all possible subset models. However, given p regressors, there are 2^p different models to consider, which will often be computationally infeasible. Hence, in Section 3.2.3, we will consider two approaches to model selection that avoid this combinatorial problem.

Remark 3.2.1. *To avoid confusion and awkward notation, assume that all subset models will always contain the intercept term* β_0

Residual Sum of Squares

For two subset models with p_1 and p_2 regressors, respectively, with $p_1 < p$ and $p_2 < p$, we can compare the mean residual sum of squares for each

$$\frac{SS_{\text{res}}(p_1)}{n - p_1 - 1} \quad \text{vs} \quad \frac{SS_{\text{res}}(p_2)}{n - p_2 - 1}$$

and choose the model with the smaller value.

We know from before that the mean of the residual sum of squares for the full model, $SS_{\text{res}}/(n-p-1)$, is an unbiased estimator for σ^2. Similar to the calculations in the previous section, we can show that

$$\mathrm{E}\left(\frac{SS_{\text{res}}(p_1)}{n - p_1 - 1}\right) \geq \sigma^2 \quad \text{and} \quad \mathrm{E}\left(\frac{SS_{\text{res}}(p_2)}{n - p_2 - 1}\right) \geq \sigma^2,$$

which is that these estimators for subset models are upwardly biased.

Mallows' C_p

We can also compare different models by computing Mallows' C_p. The goal of this value is to choose the model the minimizes the mean squared prediction error, which is

$$MSPE = \sum_{i=1}^{n} \frac{\mathrm{E}\left(\tilde{y}_i - \mathrm{E}y_i\right)^2}{\sigma^2}$$

where $\tilde{y}_i$ is the ith fitted value of the submodel and $\mathrm{E}y_i$ is the ith fitted value of the true model. Furthermore, let $\hat{y}_i$ be the ith fitted value for the full model. This is the

[11] See Montgomery, Peck, Vining Section 10.1.3

expected squared difference between what the submodel predicts and what the real value is. As usual with mean squared errors in statistics, we rewrite this in terms of the variance plus the squared bias, which is

$$\begin{aligned} MSPE &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[\mathrm{E}\left(\tilde{y}_i - \mathrm{E}\tilde{y}_i + \mathrm{E}\tilde{y}_i - \mathrm{E}y_i\right)^2\right] \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[\mathrm{E}\left(\tilde{y}_i - \mathrm{E}\tilde{y}_i\right)^2 + \left(\mathrm{E}\tilde{y}_i - \mathrm{E}y_i\right)^2\right] \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[\mathrm{Var}\left(\tilde{y}_i\right) + \mathrm{bias}(\tilde{y}_i)^2\right] \end{aligned}$$

Recall that the variance of the fitted values for the full model is $\mathrm{Var}\left(\hat{y}\right) = \sigma^2 P_x$ where $P_x = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. For a submodel with $p_1 < p$ regressors and design matrix X_{p_1}, we get the similar $\mathrm{Var}\left(\tilde{y}\right) = \sigma^2 X_{p_1}(X_{p_1}^{\mathrm{T}}X_{p_1})^{-1}X_{p_1}^{\mathrm{T}}$. As $X_{p_1}(X_{p_1}^{\mathrm{T}}X_{p_1})^{-1}X_{p_1}^{\mathrm{T}}$ is a rank $p_1 + 1$ projection matrix, we have that

$$\sum_{i=1}^{n}\mathrm{Var}\left(\tilde{y}_i\right) = \sigma^2\mathrm{tr}\left(X_{p_1}(X_{p_1}^{\mathrm{T}}X_{p_1})^{-1}X_{p_1}^{\mathrm{T}}\right) = \sigma^2(p_1+1).$$

For the bias term, consider the expected residual sum of squares for the submodel:

$$\begin{aligned} \mathrm{E}\left(SS_{\mathrm{res}}(p_1)\right) &= \mathrm{E}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \\ &= \mathrm{E}\sum_{i=1}^{n}(y_i - \mathrm{E}\tilde{y}_i + \mathrm{E}\tilde{y}_i - \mathrm{E}y_i + \mathrm{E}y_i - \tilde{y}_i)^2 \\ &= \sum_{i=1}^{n}\left[\mathrm{Var}\left(\tilde{r}_i\right) + \left(\mathrm{E}\tilde{y}_i - \mathrm{E}y_i\right)^2\right] \\ &= (n - p_1 - 1)\sigma^2 + \sum_{i=1}^{n}\mathrm{bias}(\tilde{y}_i)^2. \end{aligned}$$

Hence, rearranging the terms above gives

$$\sum_{i=1}^{n}\mathrm{bias}(\tilde{y}_i)^2 = \mathrm{E}\left(SS_{\mathrm{res}}(p_1)\right) - (n - p_1 - 1).$$

Combining the bias and the variance terms derived above results in Mallows' C_p statistic for a submodel with $p_1 < p$ regressors:

$$C_{p_1} = \frac{\mathrm{E}\left(SS_{\mathrm{res}}(p_1)\right)}{\sigma^2} - n + 2p_1 + 2 \approx \frac{SS_{\mathrm{res}}(p_1)}{SS_{\mathrm{res}}/(n-p-1)} - n + 2p_1 + 2.$$

Here, we estimate $\mathrm{E}\left(SS_{\mathrm{res}}(p_1)\right)$ by $SS_{\mathrm{res}}(p_1)$ and estimate σ^2 by $SS_{\mathrm{res}}/(n-p-1)$.

Remark 3.2.2. *Note that if we compute Mallows' C_p for the full model, we get*

$$C_p = \frac{SS_{res}}{SS_{res}/(n-p-1)} - n + 2p + 2 = p + 1.$$

Hence, Mallows' C_p in this case is just the number of parameters in the model. In general, we want to find submodels with C_p value smaller than $p + 1$.

Information Criteria

Information criteria are concerned with quantifying the amount of information in a model. With such a measure, we can choose a model that optimizes this measurement. A main requirement for these methods is that the response y is the same. Hence, we should not use the measures below when comparing transformed models–e.g. different linearized models–without the necessary modifications.

The first such measure is the Akaike Information Criterion or AIC, which is a measure of the entropy of a model. Its general definition is

$$\text{AIC} = -2\log(\text{Likelihood}) + 2(\#\text{ parameters})$$

where p is the number of parameters in the model. This can be thought of a measurement of how much information is lost when modelling complex data with a p parameter model. Hence, the model with the minimal AIC will be optimal in some sense.

In our least squares regression case with normally distributed errors,

$$\text{AIC} = n\log(SS_{\text{res}}/n) + 2(p+1)$$

where $p+1$ is for the p regressors and 1 intercept term. Thus, adding more regressors will decrease SS_{res} but will increase p. The goal is to find a model with the minimal AIC. This can be shown to give the same ordering as Mallows' C_p when the errors are normally distributed.

The second such measure is the closely related Bayesian Information Criterion or BIC, which, in general, is

$$\text{BIC} = -2\log(\text{Likelihood}) + (\#\text{ parameters})\log n.$$

In the linear regression setting with normally distributed errors,

$$\text{BIC} = n\log(SS_{\text{res}}/n) + (p+1)\log n.$$

Remark 3.2.3. *Using AIC versus using BIC for model selection can sometimes result in different final choices. In some cases, one may be preferred, but often both can be tried and discrepancies, if they exist, can be reported.*

There are also other information criterion that are not as common in practise such as the Deviation Information Criterion (DIC) and the Focused Information Criterion (FIC).

3.2.3 Forward and Backward Selection[12]

Ideally, we choose a measure for model selection from the previous section and then compare all possible models. However, for p possible regressors, this will result in 2^p models to check, which may be computationally infeasible. Hence, there are iterative approaches that can be effective.

Forward selection is the process of starting with the constant model

$$y = \beta_0 + \varepsilon$$

and choosing the best of the p regressors with respect to the model selection criterion. This gives

$$y = \beta_0 + \beta_1 x_1 + \varepsilon.$$

This process will continue to add terms to the model as long as it results in an improvement in the criterion. For example, computing the AIC at each step.

Backwards selection is the reverse of forward selection. In this case, the algorithm begins with the full model,

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon,$$

and iteratively removes the regressor that gives the biggest improvement in the model selection criterion. If the best choice is to remove no regressors, then the process terminates.

A third option is *stepwise selection*, which incorporates both forward and backward steps. In this case, we begin with the constant model as in forward selection. However, at every step, we choose either to add a new regressor to our model or remove one that is already in the model depending on which choice improves the criterion the most.

Variable Selection Example

Consider the same example as in the spline section where $x \in [0, 2]$ and

$$y = 2 + 3x - 4x^5 + x^7 + \varepsilon$$

with a sample of $n = 50$ observations. We can fit two regression models, an empty and a saturated model, respectively,

$$y = \beta_0 + \varepsilon \quad \text{and} \quad y = \beta_0 + \sum_{i=1}^{7} \beta_i x^i + \varepsilon.$$

and use the `R` function `step( )` to choose a best model with respect to AIC.

[12] See Montgomery, Peck, Vining Section 10.2.2

For our simulated data, the backward selection method, `step( md1, direction='backward' )`, decided that the best AIC of 78.154 is achieved by not removing any regressors from the model.

The forward selection method,

```
step( md0, direction='forward', scope=list(lower=md0,upper=md1) ),
```

sequentially added three regressors to the model as follows:

Step	Model	AIC
0	$y = \beta_0$	142.5
1	$y = \beta_0 + \beta_1 x_1$	118.8
2	$y = \beta_0 + \beta_1 x_1 + \beta_7 x^7$	109.8
3	$y = \beta_0 + \beta_1 x_1 + \beta_5 x^5 + \beta_7 x^7$	73.4

This method was able to achieve a better AIC value than the backward selection method.

The stepwise selection method,

```
step( md0, direction='both', scope=list(upper=md1) )
```

runs very similarly to the forward method, but makes one deletion in the final step:

Step	Model	AIC
0	$y = \beta_0$	142.5
1	$y = \beta_0 + \beta_1 x_1$	118.8
2	$y = \beta_0 + \beta_1 x_1 + \beta_7 x^7$	109.8
3	$y = \beta_0 + \beta_1 x_1 + \beta_5 x^5 + \beta_7 x^7$	73.4
4	$y = \beta_0 + \beta_5 x^5 + \beta_7 x^7$	71.5

This method has more choices than forward selection, which can only grow the model, and hence is able to achieve an even better AIC value.

The `R` package `leaps` is able to consider all possible subset models. Running this process on the same data yields the top three best models in descending order as

Model 1	$y = \beta_0 + \beta_5 x^5 + \beta_7 x^7$
Model 2	$y = \beta_0 + \beta_6 x^6 + \beta_7 x^7$
Model 3	$y = \beta_0 + \beta_5 x^5 + \beta_6 x^6$

Hence, in this case, the stepwise method was able to find the optimal model.

Remark 3.2.4. *The VIF for x^5 versus x^7 in this example is 42.43, which is quite large. Hence, model selection did not eliminate all of the multicollinearity present in the model.*

3.3 Penalized Regressions

No matter how we design our model, thus far we have always computed the least squares estimator, $\hat{\beta}$, by minimizing the sum of squared errors

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - X_i^{\mathrm{T}} \beta)^2 \right\}.$$

This is an unbiased estimator for β. However, as we have seen previously, the variance of this estimator can be quite large. Hence, we *shrink* the estimator towards zero adding bias but decreasing the variance.[13] In the context of regression, we add a penalty term to the above minimization to get a new estimator

$$\hat{\beta}^{\mathrm{pen}} = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - X_i^{\mathrm{T}} \beta)^2 + \mathrm{penalty}(\beta) \right\},$$

which increases as β increases thus attempting to enforce smaller choices for the estimated parameters. We will consider some different types of penalized regression.[14]

Remark 3.3.1. *We generally do not want to penalize the intercept term β_0. Often to account for this, the regressors and response are centred–i.e. Y is replaced with $Y - \bar{Y}$ and each X_j is replaced with $X_j - \bar{X}_j$ for $j = 1, \ldots, p$–in order to set the intercept term to zero.*

3.3.1 Ridge Regression[15]

The first method we consider is ridge regression, which arose in statistics in the 1970's,[16] but similar techniques arise in other areas of computational mathematics. In short, a quadratic penalty is applied to the least squares estimator resulting in

$$\hat{\beta}_{\lambda}^{\mathrm{R}} = \underset{\beta \in \mathbb{R}^{p}}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - X_i^{\mathrm{T}} \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

for any $\lambda \geq 0$. When $\lambda = 0$, we have the usual least squares estimator. As λ grows, the β's are more strongly penalized.

To solve for $\hat{\beta}_{\lambda}^{\mathrm{R}}$, we proceed as before with the least squares estimator $\hat{\beta}$ by setting the partial derivatives equal to zero

$$0 = \frac{\partial}{\partial \beta_k} \left\{ \sum_{i=1}^{n} (y_i - \beta_1 x_{i,1} - \ldots - \beta_p x_{i,p})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

[13] General idea of shrinkage is attributed to Stein (1956) https://en.wikipedia.org/wiki/James%E2%80%93Stein_estimator

[14] In R, the `glmnet` package has a lot of functionality to fit different types of penalized general linear models http://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html

[15] See Montgomery, Peck, Vining Section 9.5.3

[16] Hoerl, A.E.; R.W. Kennard (1970).

This results in the system of equations

$$X^{\mathrm{T}}Y - (X^{\mathrm{T}}X)\hat{\beta}_{\lambda}^{\mathrm{R}} - \lambda\hat{\beta}_{\lambda}^{\mathrm{R}} = 0$$

with the ridge estimator being $\hat{\beta}_{\lambda}^{\mathrm{R}} = (X^{\mathrm{T}}X + \lambda I_n)^{-1}X^{\mathrm{T}}Y$.

The matrix $X^{\mathrm{T}}X$ is positive semi-definite even when $p > n$–i.e. the number of parameters exceeds the sample size. Hence, any positive value λ will make $X^{\mathrm{T}}X + \lambda I_n$ invertible as it adds the positive constant λ to all of the eigenvalues. Increasing the value of λ will increase the numerical stability of the estimator–i.e. decrease the condition number of the matrix. Furthermore, it will decrease the variance of the estimator while increasing the bias. It can also be shown that the bias of $\hat{\beta}_{\lambda}^{\mathrm{R}}$ is

$$\mathrm{E}\hat{\beta}_{\lambda}^{\mathrm{R}} - \beta = -\lambda(X^{\mathrm{T}}X + \lambda I_n)^{-1}\beta,$$

which implies that the estimator does, in fact, shrink towards zero as λ increases.

3.3.2 Best Subset Regression

Another type of penalty related to the variable selection techniques from the previous section is the Best Subset Regression approach, which counts the number of non-zero β's and adds a larger penalty as more terms are included in the model. The optimization looks like

$$\hat{\beta}_{\lambda}^{\mathrm{B}} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \left\{ \sum_{i=1}^{n}(y_i - X_i^{\mathrm{T}}\beta)^2 + \lambda \sum_{j=1}^{p} \mathbf{1}[\beta_j \neq 0] \right\}.$$

The main problem with this method is that the optimization is non-convex and becomes severely difficult to compute in practice. This is why the forwards and backwards selection methods are used for variable selection.

3.3.3 LASSO

The last method we consider is the Least Absolute Shrinkage and Selection Operator, which is commonly referred to as just LASSO. This was introduced by Tibshirani (1996) and has since been applied to countless areas of statistics. The form is quite similar to ridge regression with one small but profound modification,

$$\hat{\beta}_{\lambda}^{\mathrm{L}} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \left\{ \sum_{i=1}^{n}(y_i - X_i^{\mathrm{T}}\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

which is that the penalty term is now the sum of the absolute values instead of a sum of squares.

The main reason for why this technique is popular is that it combines shrinkage methods like ridge regression with variable selection and still results in a convex optimization problem. Delving into the properties of this estimator requires convex analysis and will be left for future investigations.

3.3.4 Elastic Net[17]

The elastic net regularization method combines both ridge and lasso regression into one methodology. Here, we include a penalty term for each of the two methods:

$$\hat{\beta}_{\lambda}^{\text{EN}} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - X_i^{\text{T}} \beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}.$$

This method has two tuning parameters $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$. In the `R` library `glmnet`, a *mixing* parameter α and a *scale* parameter λ is specified to get

$$\hat{\beta}_{\lambda}^{\text{EN}} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - X_i^{\text{T}} \beta)^2 + \lambda \sum_{j=1}^{p} \left[\alpha |\beta_j| + \frac{1-\alpha}{2} \beta_j^2 \right] \right\}.$$

The intuition behind this approach is to combine the strengths of both ridge and lasso regression. Namely, ridge regression shrinks the coefficients towards zero reducing the variance while lasso selects a subset of the parameters to remain in the model.

3.3.5 Penalized Regression: An Example

Consider a sample of size $n = 100$ generated by the model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{50} x_{50} + \varepsilon$$

where $\beta_{27} = 2$, $\beta_{34} = -2$, all other $\beta_i = 0$, and $\varepsilon \sim \mathcal{N}(0, 16)$. Even though only two of the regressors have any effect on the response y, feeding all 50 regressors into `R`'s `lm()` function can result in many false positives as in the following table where ten of the regressors have a moderate to very significant p-value.

	Estimate	Std. Error	t value	Pr(>|t|)	
x_4	-1.01	0.42	-2.42	0.019	*
x_{16}	-1.54	0.56	-2.75	0.008	**
x_{18}	1.16	0.53	2.20	0.033	*
x_{19}	1.15	0.45	2.55	0.014	*
x_{22}	-1.58	0.50	-3.17	0.003	**
x_{24}	-1.01	0.60	-1.69	0.097	.
x_{27}	1.52	0.46	3.27	0.002	**
x_{29}	-1.02	0.50	-2.04	0.047	*
x_{32}	0.95	0.49	1.93	0.060	.
x_{34}	-2.44	0.45	-5.47	0.0000015	***

We could attempt one of the stepwise variable selection procedures from the previous section. Running backwards and forwards selection results in the following regressors being retained in the model.

[17] https://en.wikipedia.org/wiki/Elastic_net_regularization

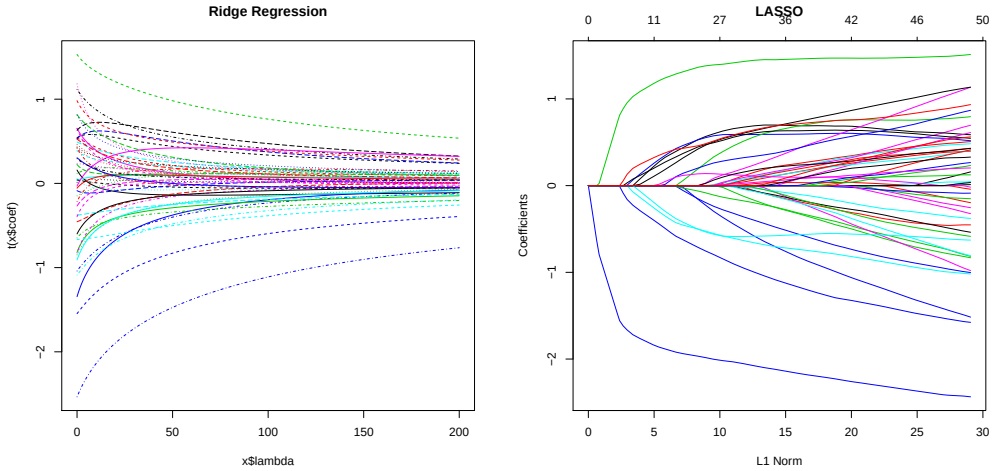

Figure 3.1: Plots of the so-called Ridge and LASSO paths.

Method	Regressors Kept
Backward	$x_2,\ x_3,\ x_4,\ x_6,\ x_7,\ x_{10},\ x_{15},\ x_{16},\ x_{18},\ x_{19}$
	$x_{22},\ x_{25},\ x_{27},\ x_{28},\ x_{29},\ x_{32},\ x_{34},\ x_{43},\ x_{47}$
Forward	$x_4,\ x_7,\ x_{10},\ x_{15},\ x_{16},\ x_{18},\ x_{19},\ x_{25},\ x_{27},\ x_{28}$
	$x_{29},\ x_{32},\ x_{34},\ x_{43},\ x_{47}$

Hence, both of these procedures retain too many regressors in the final model. The stepwise selection method was also run, but returned results equivalent to forward selection.

Applying ridge regression to this dataset will result in all 50 of the estimated parameters being shrunk towards zero. The plot on the left hand side of Figure 3.1 demonstrates this behaviour. The vertical axis corresponds to the values of $\beta_1, \ldots, \beta_{50}$. The horizontal axis corresponds to increasing values of the penalization parameter λ. As λ increases, the estimates for the β's tend towards zero. Hence we see all 50 of the curves bending towards the zero.

Apply LASSO to the dataset results in a different set of paths from the ridge regression. The right hand plot in Figure 3.1 displays the LASSO paths. In this case, the horizontal axis corresponds to some K such that $\|\hat{\beta}^{\mathrm{L}}_{\lambda}\|_1 < K$, which is equivalent to adding the penalty term $\lambda \sum_{j=1}^{p} |\beta_j|$. As this bound K grows, more variables will enter the model. The blue and green lines represent the regressors x_{34} and x_{27}, which are the first two terms to enter the model. Eventually, as the penalty is relaxed, many more terms begin to enter the model. Hence, choosing a suitable K, or equivalently λ, is a critical problem for this method.

Chapter 4

Generalized Linear Models

Introduction

A generalized linear model (GLM) extends the usual linear model to cases where the errors ε have proposed distributions that are very different than the normal distribution. In this case, we assume that ε comes from an exponential family[1], and the form of the model is

$$g\left(\mathrm{E}y\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p.$$

Here, $g(\cdot)$ is referred to as the link function. As these are still parametric models, the parameters $\beta_0, \ldots, \beta_p$ are often solved for via maximum likelihood. These models can be fit in `R` via the `glm()` function. While such models can be studied in full generality, for this course, we will only consider two specific GLMs: the logistic regression and the Poisson regression.

4.1 Logistic Regression[2]

One of the most useful GLMs is the logistic regression. This is applied in the case of a binary response–i.e when $y \in \{0, 1\}$. This could, for example, be used for diagnosis of a disease where $x_1, \ldots, x_p$ are predictors and y corresponds to the presence or absence of the disease.

The usual setup is to treat the observed responses $y_i \in \{0, 1\}$ as Bernoulli (π_i) random variables, which is

$$\mathrm{P}\left(y_i = 1\right) = \pi_i \quad \text{and} \quad \mathrm{P}\left(y_i = 0\right) = 1 - \pi_i.$$

This implies that the mean is $\mathrm{E}y_i = \pi_i$ and the variance is $\mathrm{Var}\left(y_i\right) = \pi_i(1 - \pi_i)$. Hence, the variance is a function of the mean, which violates the assumption of constant variance in the Gauss-Markov theorem.

[1] https://en.wikipedia.org/wiki/Exponential_family

[2] See Montgomery, Peck, Vining Section 13.2.

The goal is then to model $\mathrm{E}y_i = \pi_i$ as a function of $x_i^{\mathrm{T}}\beta$. While different link functions are possible,[3] the standard link function chosen is the logistic response function–sometimes referred to as the logit function–which is

$$\mathrm{E}y_i = \pi_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1 + \exp(x_i^{\mathrm{T}}\beta)}$$

and results in an S-shaped curve. Rearranging the equation results in

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^{\mathrm{T}}\beta = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p}$$

where the ratio $\pi_i/(1-\pi_i)$ is the odds ratio or simply the odds. Futhermore, the entire response $\log\left(\pi_i/(1-\pi_i)\right)$ is often referred to as the *log odds*.

The estimator $\hat{\beta}$ can be computed numerically via maximum likelihood as we assume the underlying distribution of the data. Hence, we also get fitted values of the form

$$\hat{y}_i = \hat{\pi}_i = \frac{\exp x_i^{\mathrm{T}}\hat{\beta}}{1 + \exp x_i^{\mathrm{T}}\hat{\beta}}.$$

However, as noted already, the variance is not constant. Hence, for residuals of the form $y_i - \hat{\pi}_i$ to be useful, we will need to normalize them. The *Pearson residual* for the logistic regression is

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}.$$

4.1.1 Binomial responses

In some cases, multiple observations can be made for each value of predictors x. For example, $x_i \in \mathbb{R}^+$ could correspond to a dosage of medication and $y_i \in \{0, 1, \ldots, m_i\}$ could correspond to the number of the m_i subjects treated with dosage x_i that are cured of whatever ailment the medication was supposed to treat.

In this setting, we can treat the observed values $\pi_i = y_i/m_i$. Often when fitting a logistic regression model to binomial data, the data points are weighted with respect to the number of observations m_i at each regressor x_i. That is, if m_1 is very large and m_2 is small, then the estimation of $\pi_1 = y_1/m_1$ is more accurate–i.e. lower variance–than $\pi_2 = y_2/m_2$.

[3] See, for example, probit regression https://en.wikipedia.org/wiki/Probit_model

4.1.2 Testing model fit

Estimation of the parameters β is achieved by finding the maximum likelihood estimator. The log likelihood in the logistic regression with Bernoulli data is

$$\begin{aligned}\log L(\beta) &= \log \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \\ &= \sum_{i=1}^{n} \left[y_i \log \pi_i + (1-y_i)\log(1-\pi_i)\right] \\ &= \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i)\right] \\ &= \sum_{i=1}^{n} \left[y_i x_i^{\mathrm{T}}\beta - \log(1+\mathrm{e}^{x_i^{\mathrm{T}}\beta})\right].\end{aligned}$$

The MLE $\hat{\beta}$ is solved for numerically.

Beyond finding the MLEs, the log likelihood is also used to test for the goodness-of-fit of the regression model. This comes from a result known as Wilks' theorem.

Theorem 4.1.1 (Wilks' Theorem)**.** *For two statistical models with parameter spaces Θ_0 and Θ_1 such that $\Theta_0 \subset \Theta_1$[4] and likelihoods $L_0(\theta)$ and $L_1(\theta)$, then -2 times the log likelihood ratio has an asymptotic chi-squared distribution with degrees of freedom equal to the difference in the dimensionality of the parameter spaces. That is,*

$$-2\log(LR) = -2\log\left(\frac{\sup_{\theta\in\Theta_0} L_0(\theta)}{\sup_{\theta\in\Theta_1} L_1(\theta)}\right) \xrightarrow{d} \chi^2\left(|\Theta_1| - |\Theta_0|\right).$$

In the case of the logistic regression, we can use this result to construct an analogue to the F-test when the errors have a normal distribution. That is, we can compute the likelihood of the constant model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 \tag{4.1.1}$$

and the likelihood of the full model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p. \tag{4.1.2}$$

and claim from Wilks' theorem that $-2\log(LR)$ is approximately distributed $\chi^2(p)$ assuming the sample size is large enough.

[4] i.e. the models are nested

In R, the `glm()` function does not return a p-value as the `lm()` function does for the F test. Instead, it provides two values, the *null* and *residual deviances*, which are respectively

$$\text{null deviance} = -2\log\left(\frac{L(\text{null model})}{L(\text{saturated model})}\right), \text{ and}$$
$$\text{residual deviance} = -2\log\left(\frac{L(\text{full model})}{L(\text{saturated model})}\right).$$

Here, the null model is from Equation 4.1.1 and the full model is from Equation 4.1.2. The saturated model is the extreme model with $p = n$ parameters, which perfectly models the data. It is, in some sense, the largest possible model used as a baseline.

In practice, the deviances can be thought of like the residual sum of squares from the ordinary linear regression setting. If the decrease is significant, then the regressors have predictive power over the response. Furthermore,

$$(\text{null deviance}) - (\text{residual deviance}) = -2\log\left(\frac{L(\text{null model})}{L(\text{full model})}\right),$$

which has an asymptotic $\chi^2(p)$ distribution. Such a goodness of fit test can be run in R using `anova( model, test="LRT" )`. Note that there are other goodness-of-fit tests possible for such models.

4.1.3 Logistic Regression: An Example

A sample of size $n = 21$ was randomly generated with $x_i \in [-1, 1]$ and y_i such that

$$y_i \sim \text{Bernoulli}\left(\frac{e^{2x_i}}{1 + e^{2x_i}}\right).$$

A logistic regression was fit to the data in R using the `glm()` function with the `family=binomial(logit)` argument to specify that the distribution is Bernoulli–i.e. binomial with $n = 1$–and that we want a logistic link function.

The result from `summary()` is the model

$$\log\left(\frac{\pi}{1-\pi}\right) = -0.4 + 2.16x$$

with a p-value of 0.03 for $\hat{\beta}_1 = 2.16$. A plot of the true and predicted curves is displayed in Figure 4.1.

We could consider the alternative case were

$$y_i \sim \text{Binomial}\left(m_i, \frac{e^{2x_i}}{1 + e^{2x_i}}\right)$$

for some $m_i \in \mathbb{Z}^+$, which is m_i Bernoulli observations at each x_i. In this case, we have a much larger dataset, which results in the better fitting curve displayed in the bottom plot of Figure 4.1.

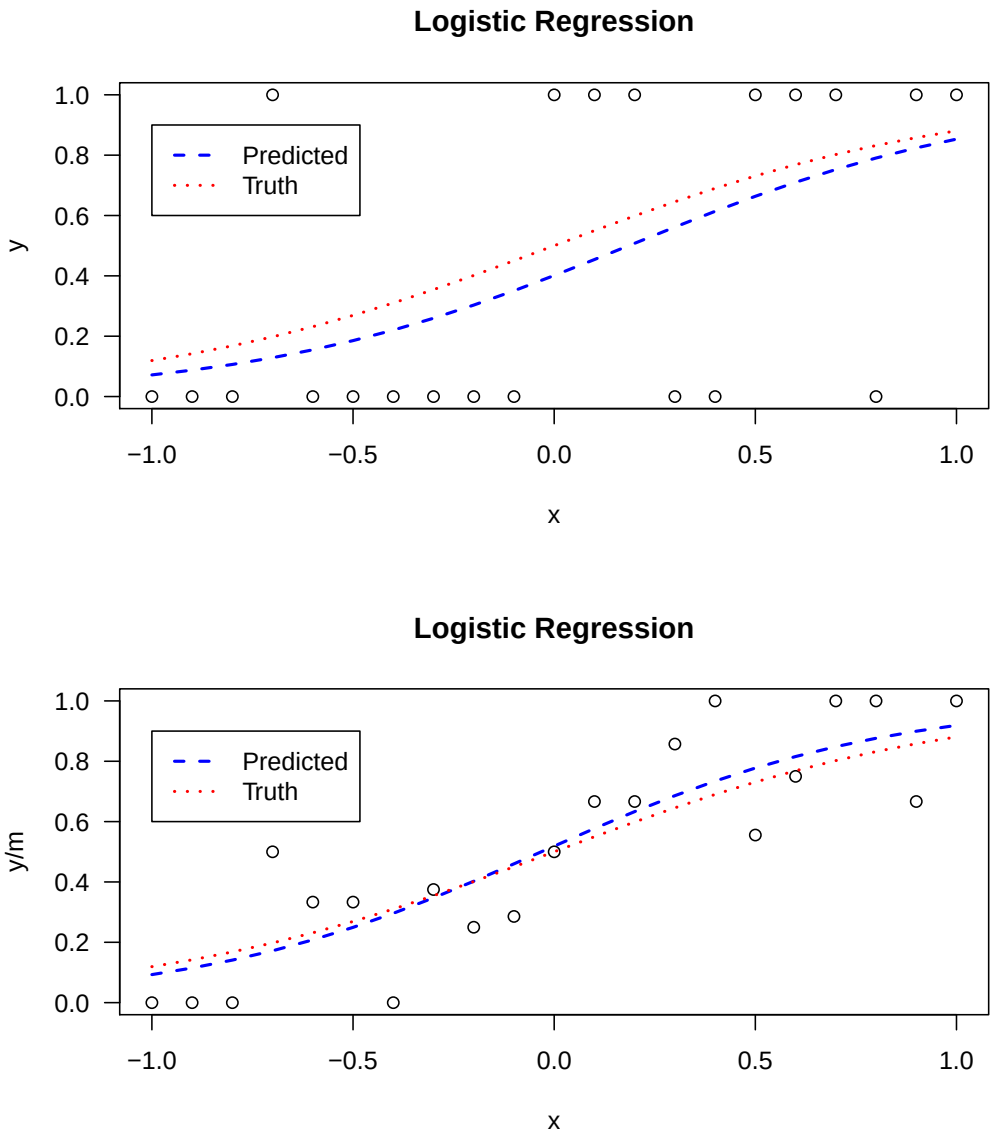

Figure 4.1: Plots of the 21 data points, the true Bernoulli probabilities in red, and the predicted probabilities in blue. Top plot, $y_i \sim \text{Bernoulli}(\pi_i)$. Bottom plot, $y_i \sim \text{Binomial}(m_i, p_i)$.

4.2 Poisson Regression[5]

The Poisson regression is another very useful GLM to know, which models counts or occurrences of rare events. Examples include modelling the number of defects in a manufacturing line or predicting lightning strikes across a large swath of land.

The Poisson distribution with parameter $\lambda > 0$ has PDF $f(x) = \mathrm{e}^{-\lambda}\lambda^x/x!$. This can often be thought of as a limiting case of the the binomial distribution as $n \to \infty$ and $p \to 0$ such that $np \to \lambda$. If we want a linear model for a response variable y such that y_i has a Poisson (λ_i) distribution, then, similarly to the logistic regression, we apply a link function. In this case, one of the most common link functions used is the log. The resulting model is

$$\log \mathrm{E}y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

where $\mathrm{E}y_i = \lambda_i$.

Similarly to the logistic regression case–and to GLMS in general–the parameters β are estimated via maximum likelihood based on the assumption that y has a Poisson distribution. Furthermore, the deviances can be computed and a likelihood ratio test can be run to assess the fit of the model. The log likelihood for a Poisson regression is

$$\begin{aligned}
\log L(\beta) &= \log \prod_{i=1}^{n} \frac{\mathrm{e}^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \\
&= \sum_{i=1}^{n} \left[-\lambda_i + y_i \log \lambda_i - \log(y_i!)\right] \\
&= \sum_{i=1}^{n} \left[-\mathrm{e}^{x_i^{\mathrm{T}}\beta} + y_i x_i^{\mathrm{T}}\beta - \log(y_i!)\right]
\end{aligned}$$

4.2.1 Poisson Regression: An Example

As an example, consider a sample of size $n = 21$ generated randomly by

$$y_i \sim \text{Poisson}\left(\mathrm{e}^{2x_i}\right).$$

A Poisson regression with log link can be fit to the data using `R`'s `glm()` function with the argument `family=poisson(link=log)`.

The result from `summary()` is

$$\log(\lambda) = 0.16 + 1.95x$$

with a p-value of 10^{-8} for $\hat{\beta}_1 = 1.95$. A plot of the true and predicted curves is displayed in Figure 4.2.

[5] See Montgomery, Peck, Vining Section 13.3.

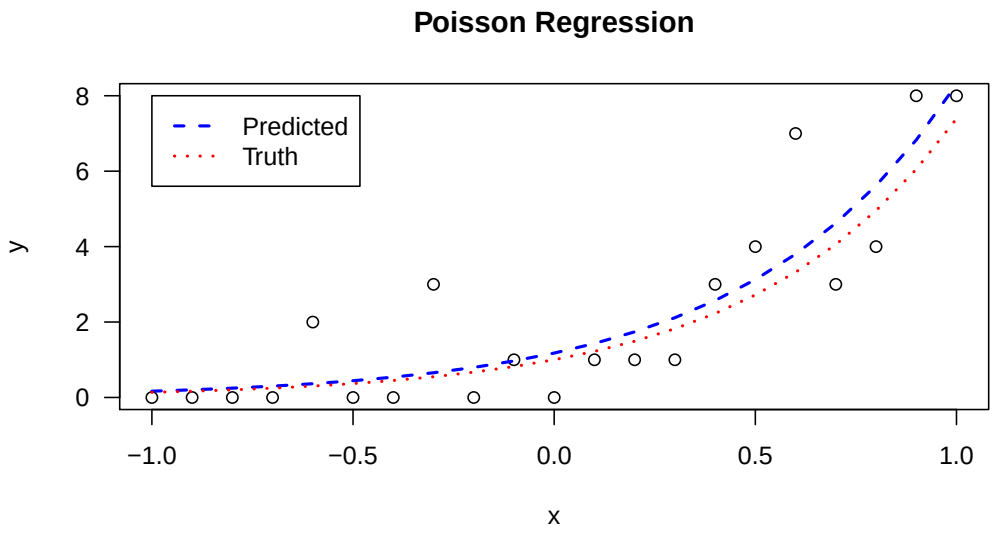

Figure 4.2: Poisson regression (blue) and true rate parameter (red) fit to $n = 21$ data points.

Appendix A

Distributions

Introduction

The following is a short overview of some of the most common probability distributions encountered in the context of linear regression.

A.1 Normal distribution

The normal or Gaussian distribution is of principal importance in probability and statistics. In its univariate form, we say that $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}^+$ if it has the following probability density function (pdf):

$$f_{\mathcal{N}(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

Such a random variable X can be centred and scaled into a standardized form as $(X-\mu)/\sigma \sim \mathcal{N}(0,1)$. Furthermore, the normal distribution is very stable in the sense that the sum of a collection of independent normal random variables has a normal distribution as well as scaling a normal random variable by some real valued scalar.

Part of its importance stems from the central limit theorem, which in its simplest form states that the standardized sum of a collection on independent random variables converges in distribution to a standard normal random variable. However, its worth noting that there are many other central limit theorems in existence.

Theorem A.1.1 (Central Limit Theorem)**.** *Let $Y_1, \ldots, Y_n$ be a sample of n iid random variables with mean $\mu \in \mathbb{R}$ and finite variance $\sigma^2 < \infty$. Letting $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$ be the sample mean, then*

$$\sqrt{n}(\bar{Y} - \mu) \xrightarrow{d} \mathcal{N}(0,1)$$

where $\xrightarrow{d}$ denotes convergence in distribution.

The normal distribution can be extended to the multivariate normal distribution for a vector $X \in \mathbb{R}^p$ where we write $X \sim \mathcal{N}(\mu, \Sigma)$. Here, $\mu \in \mathbb{R}^p$ is the mean vector while $\Sigma \in \mathbb{R}^{p \times p}$ is a $p \times p$ matrix that is symmetric and positive semi-definite. This distribution also has elliptical symmetry. The form of the pdf, assuming Σ is positive definite, is

$$f_{\mathcal{N}(\mu,\Sigma)}(x) = \frac{1}{\sqrt{2\pi \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right)$$

Otherwise, X will have a degenerate normal distribution, which is still normal but supported on a subspace of dimension less than p.

The multivariate normal (MVN) distribution has some very nice characterizations. A vector $X = (X_1, \ldots, X_p)$ is MVN if and only if every linear combination of the X_i is univariate normal. That is, for all $a \in \mathbb{R}^p$, $a \cdot X \sim \mathcal{N}\left(\tilde{\mu}, \tilde{\sigma}^2\right)$. Hence, it is worth emphasizing that if X is a vector that is marginally normal (i.e. every component X_i is univariate normal), then the joint distribution is **not** necessarily multivariate normal.

In general if two random variables $X, Y \in \mathbb{R}$ are independent, then we have that $\operatorname{cov}(X, Y) = 0$. However, the reverse implication is not necessarily true. One exception to this is that if (X, Y) is MVN then if $\operatorname{cov}(X, Y) = 0$ then X and Y are independent.

A.2 Chi-Squared distribution

The chi-squared distribution arises throughout the field of statistics usually in the context of goodness-of-fit testing. Let $Z \sim \mathcal{N}(0, 1)$, then $Z \sim \chi^2(1)$, which is said to be chi-squared with one degree of freedom. Furthermore, chi-squared random variables are additive in the sense that if $X \sim \chi^2(\nu)$ and $Y \sim \chi^2(\eta)$ are independent, then $X + Y \sim \chi^2(\nu + \eta)$. A chi-squared random variable is supported on the positive real line and the pdf for $X \sim \chi^2(\nu)$ with $\nu > 0$ is

$$f_{\chi^2(\nu)}(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} \mathrm{e}^{-x/2}$$

Its mean and variance are ν and 2ν, respectively. Note that while very often the degrees of freedom parameter is a positive integer, it can take on any positive real value and still be valid. In fact, the chi-squared distribution is a specific type of the more general gamma distribution, which will not be discussed in these notes.

This random variable is quite useful in the context of linear regression with respect to how it leads to the t and F distributions discussed in the following subsections. However, it also arises in many other areas of statistics including Wilks' Theorem regarding the asymptotic distribution of the log likelihood ratio, goodness-of-fit in multinomial hypothesis testing, and testing for independence in a contingency table.

A.3 t distribution

The t distribution is sometimes referred to as Student's t distribution in recognition of its, at the time, anonymous progenitor William Sealy Gosset working at the Guinness brewery. It most notably arises in the context of estimation of a normal random variable when the variance is unknown as occurs frequently in these lecture notes.

Let $Z \sim \mathcal{N}(0,1)$ and let $V \sim \chi^2(\nu)$ be independent random variables. Then, we say that

$$T = \frac{Z}{\sqrt{V/\nu}} \sim t(\nu)$$

has a t distribution with ν degrees of freedom. Such a distribution can be thought of as a heavier tailed version of the standard normal distribution. In fact, $t(1)$ is the Cauchy distribution with pdf

$$f_{t(1)}(x) = (\pi(1+x^2))^{-1}$$

while $T \sim t(\nu)$ converges to a standard normal distribution as $\nu \to \infty$.

A noteworthy property of a t distributed random variable is that it will only have moments up to but not including order ν. That is, for $T \sim t(\nu)$, $\mathrm{E}T^k < \infty$ for $k < \nu$. For $k \geq \nu$, the moments do not exist.

A.4 F distribution

The F distribution arises often in the context of linear regression when a comparison is made between two sources of variation. Let $X \sim \chi^2(\nu)$ and $Y \sim \chi^2(\eta)$ be independent random variables, then we write that

$$F = \frac{X/\nu}{Y/\eta} \sim F(\nu, \eta)$$

has an F distribution with degrees of freedom ν and η. Due to this form of the distribution, if $F \sim F(\nu,\eta)$, then $F^{-1} \sim F(\eta,\nu)$. The F distribution is supported on the positive real line. The F and t distributions are related by the fact that if $T \sim t(\nu)$, then $T^2 \sim F(1,\nu)$.

Appendix B

Some Linear Algebra

To successfully understand linear regression, we will require some basic notations regrading the manipulation of vectors and matrices. First, we will denote a n dimensional vector as $Y \in \mathbb{R}^n$ and an $n \times p$ matrix as $X \in \mathbb{R}^{n \times p}$. The following is a list of definitions and results:

1. For a matrix $A \in \mathbb{R}^{n \times p}$ with row i and column j entry denoted $a_{i,j}$, then the transpose of A is $A^{\mathrm{T}} \in \mathbb{R}^{p \times n}$ with row i and column j entry $a_{j,i}$. That is, the indices have swapped.

2. For matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, we have that $(AB)^{\mathrm{T}} = B^{\mathrm{T}} A^{\mathrm{T}}$

3. For an invertible matrix $A \in \mathbb{R}^{n \times n}$, we have that $\left(A^{-1}\right)^{\mathrm{T}} = (A^{\mathrm{T}})^{-1}$.

4. A matrix A is *square* if the number of rows equals the number of columns. That is, $A \in \mathbb{R}^{n \times n}$.

5. A square matrix $A \in \mathbb{R}^{n \times n}$ is *symmetric* if $A = A^{\mathrm{T}}$.

6. A symmetric matrix $A \in \mathbb{R}^{n \times n}$ necessarily has real eigenvalues.

7. A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is *positive definite* if for all $x \in \mathbb{R}^n$ with $x \neq 0$, we have that $x^{\mathrm{T}} A x > 0$.

8. A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is also positive definite if all of its eigenvalues are positive real valued.

9. A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is *positive semi-definite* (also non-negative definite) if for all $x \in \mathbb{R}^n$ with $x \neq 0$, we have that $x^{\mathrm{T}} A x \geq 0$. Or alternatively, all of the eigenvalues are non-negative real valued.

10. Covariance matrices are always positive semi-definite. If a covariance matrix has some zero valued eigenvalues, then it is called *degenerate*.

11. If $X, Y \in \mathbb{R}^n$ are random vectors, then

$$\operatorname{cov}(X, Y) = \mathrm{E}\left((X - \mathrm{E}X)(Y - \mathrm{E}Y)^{\mathrm{T}}\right) \in \mathbb{R}^{n \times n}.$$

12. If $X, Y \in \mathbb{R}^n$ are random vectors and $A, B \in \mathbb{R}^{m \times n}$ are non-random real valued matrices, then

$$\operatorname{cov}(AX, BY) = A\operatorname{cov}(X, Y) B^{\mathrm{T}} \in \mathbb{R}^{m \times m}.$$

13. If $Y \in \mathbb{R}^n$ is multivariate normal–i.e. $Y \sim \mathcal{N}(\mu, \Sigma)$–and $A \in \mathbb{R}^{m \times n}$ then AY is also multivariate normal with $AY \sim \mathcal{N}\left(A\mu, A\Sigma A^{\mathrm{T}}\right)$.

14. A square matrix $A \in \mathbb{R}^{n \times n}$ is *idempotent* if $A^2 = A$.