UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Statistical Machine Learning I

International Undergraduate Summer Enrichment Program (IUSEP)

Linglong Kong

Department of Mathematical and Statistical Sciences
University of Alberta

July 18, 2016

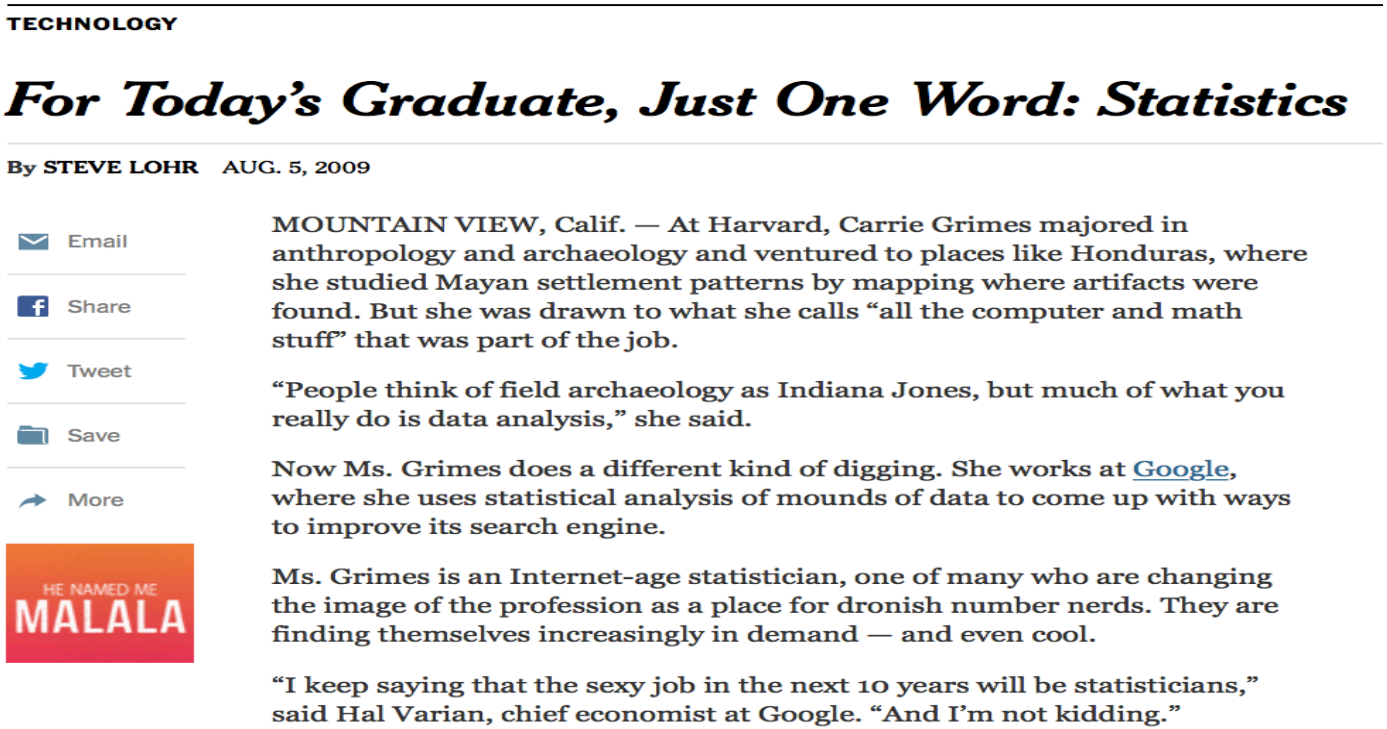
Outline

Stigler's seven pillars of statistical wisdom

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- What is statistics - It is what statisticians do
- Stigler's seven pillars of statistical wisdom
  - Aggregation
  - The law of diminishing information
  - Likelihood
  - Intercomparison
  - Regression and multivariate analysis
  - Design
  - Models and Residuals
- *http://blogs.sas.com/content/iml/2014/08/05/stiglers-seven-pillars-of-statistical-wisdom/*
- Stigler's law of eponymy: No scientific discovery is named after its original discoverer. by Robert K. Merton (Matthew effect)

Statistics



TECHNOLOGY

For Today's Graduate, Just One Word: Statistics

By STEVE LOHR AUG. 5, 2009

Email
Share
Tweet
Save
More

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.

"People think of field archaeology as Indiana Jones, but much of what you really do is data analysis," she said.

Now Ms. Grimes does a different kind of digging. She works at Google, where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

Ms. Grimes is an Internet-age statistician, one of many who are changing the image of the profession as a place for dronish number nerds. They are finding themselves increasingly in demand — and even cool.

"I keep saying that the sexy job in the next 10 years will be statisticians," said Hal Varian, chief economist at Google. "And I'm not kidding."

- Quote of the Day, New York Times, August 5, 2009
  "I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding." HAL VARIAN, chief economist at Google.

Machine Learning

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- Wikipedia: Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence.
- Machine learning is closely related to computational statistics; a discipline that aims at the design of algorithms for implementing statistical methods on computers.
- Machine learning and pattern recognition *can be viewed as two facets of the same field*.
- Machine learning tasks are typically classified into three broad categories, supervised learning, unsupervised learning, and reinforcement learning.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Alphago

- Artificial intelligence pioneered by University of Alberta graduates masters Chinese board game
- Augment Monte Carlo Search Tree (MCST) with deep neural networks

Statistical Machine Learning

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- This courses is not exactly statistics, nor exactly machine learning.
- So what do we do in this course? Statistical machine learning!
- Statistical machine learning merges statistics with the computational sciences - computer science, systems science and optimization. `http://www.stat.berkeley.edu/~statlearning/`.
- Statistical machine learning emphasizes models and their interpretability, and precision and uncertainty.

Supervised Learning

- Data: response Y and covariate X.
- In the regression problem, Y is quantitative (e.g. price and blood pressure).
- In the classification problem, Y takes categorical data (e.g. survived/died, digits $0-9$).
- In regression, techniques include linear regression, model selection, nonlinear regression, ...
- In classification, techniques include logistic regression, linear and quadratic discriminant analysis, support vector machine, ...
- There are many other supervised learning methods, like tree-based methods, Ensembles (Bagging, Boosting, Random forests), and so on.

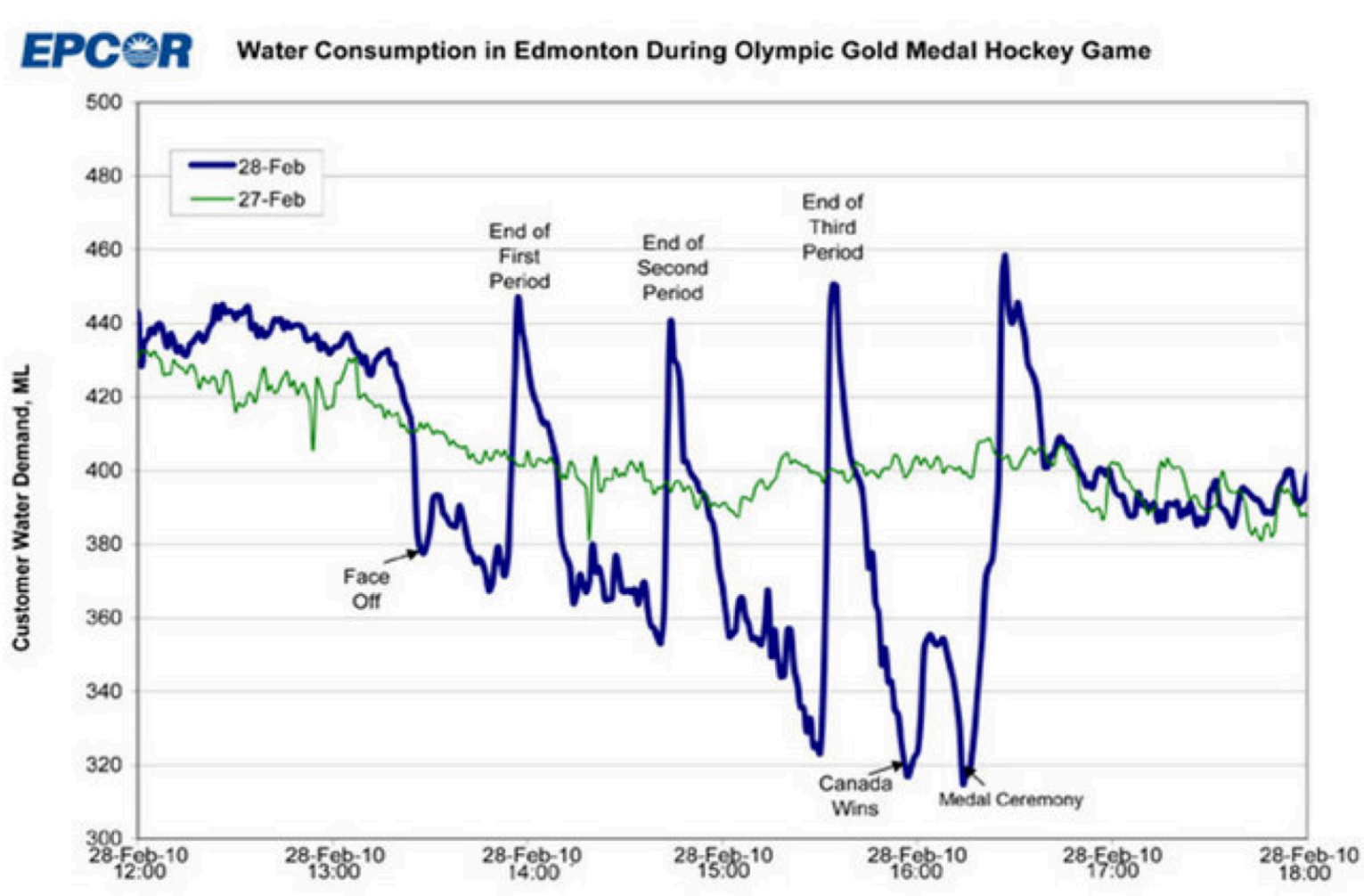
Unsupervised Learning

- No response, just a set of covariates.
- objective is more fuzzy - find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- Difficult to know how well your are doing.
- Different from supervised learning, but can be useful as a pre-processing step for supervised learning.
- Methods include cluster analysis, principal component analysis, independent component analysis, factor analysis, canonical correlation analysis, ...

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Seeing the data

- They say a picture is worth 1000 (10000) words



Vancouver 2010 final Canada vs. USA

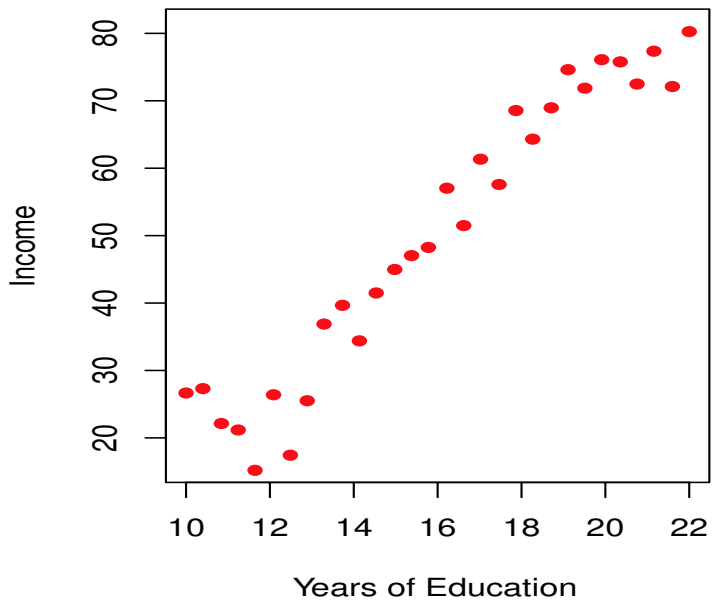
Statistical Machine Learning

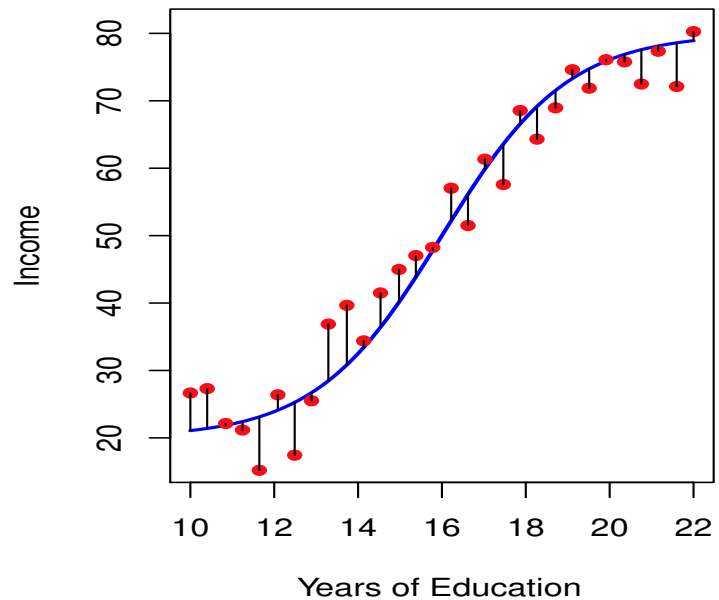
- Given response Y_i and covariates $\boldsymbol{X}_i = (x_{1i}, x_{2i}, \cdots, x_{pi})^T$, we model the relationship

$$Y_i = f(\boldsymbol{X}_i) + \varepsilon_i,$$

where f is an unknown function and ε is random error with mean zero.

- A Simple example





Estimate or learn the relationship

- Statistical machine learning is to estimate the relationship f, or using data to learn f. Why?
- To make prediction for the response Y for a new value of $\boldsymbol{X}$;
- To make inference on the relationship between Y and $\boldsymbol{X}$, say, which x actually affect Y, positive or negative, linearly or more complicated.
- Prediction Interested in predicting how much money an individual will donate based on observations from 90,000 people on which we have recorded over 400 different characteristics.
- Inference Wish to predict median house price based on 14 variables. Probably want to understand which factors have the biggest effect on the response and how big the effect is.

Estimate or learn the relationship

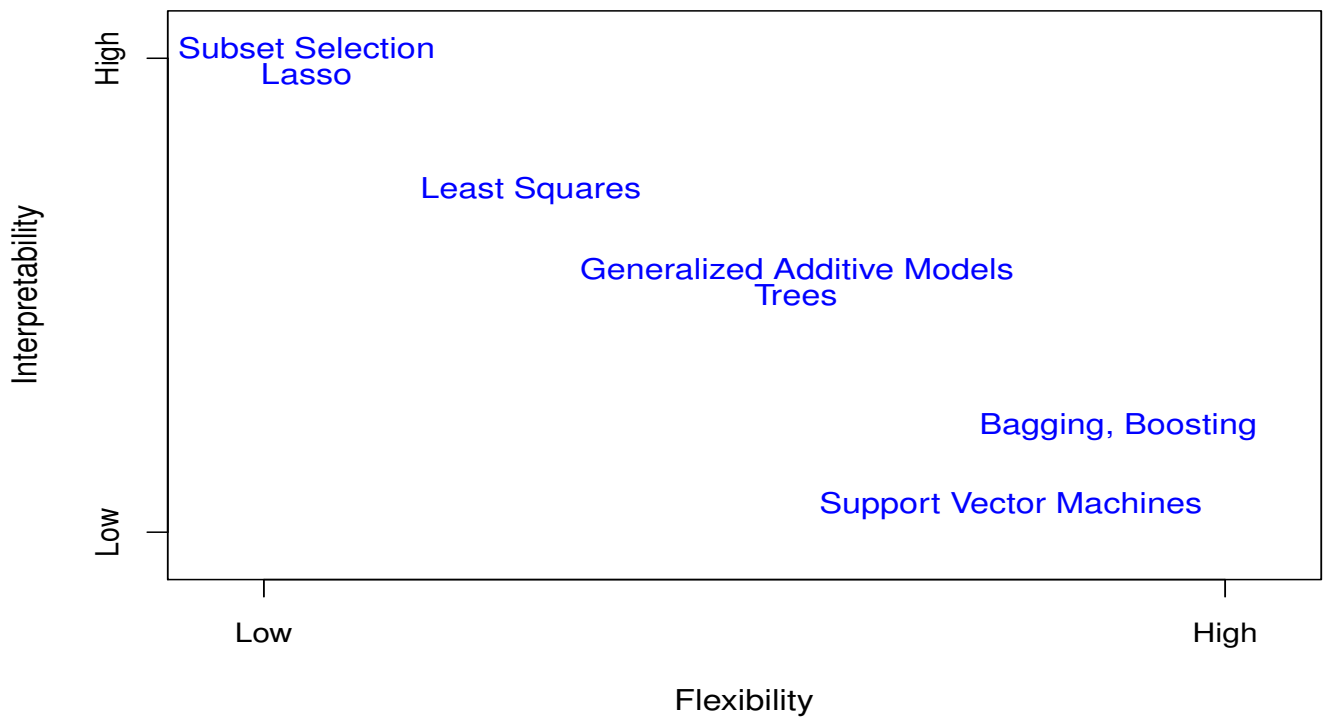
- How estimate or learn f?
- Parametric methods say, linear regression (Chapter 3)

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi},$$

  by certain loss function, e.g. ordinary least squares (OLS).
- Nonparametric methods, say, spline expansion (Chapter 5) and kernel smoothing (Chapter 6) methods.
- Nonparametric methods are more felxible but need more data to obtain an accurate estimation.

Tradeoff between accuracy and interpretability

- The simpler, the better - parsimony or Occam's razor.
- A simple method is much easier to interpret, e.g. linear regression model.
- A simple model is possible to achieve more accurate prediction without overfitting. It seems counter intuitive though.



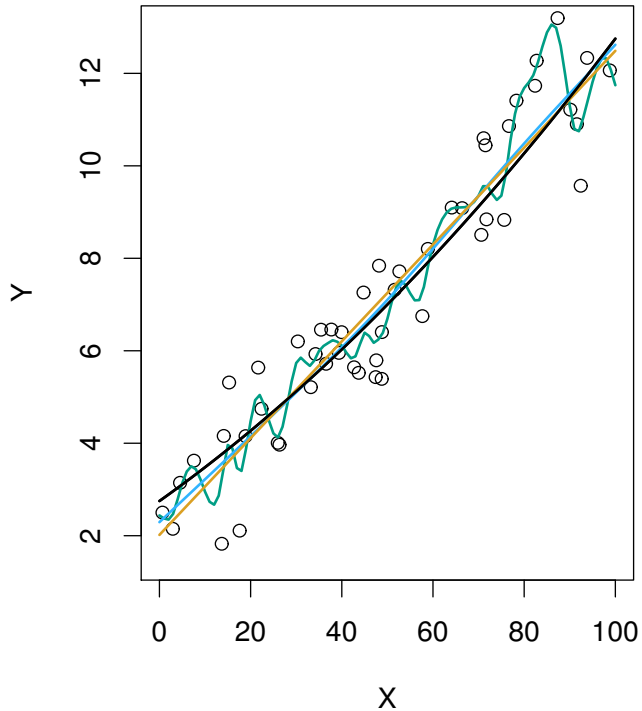
Quality of fit

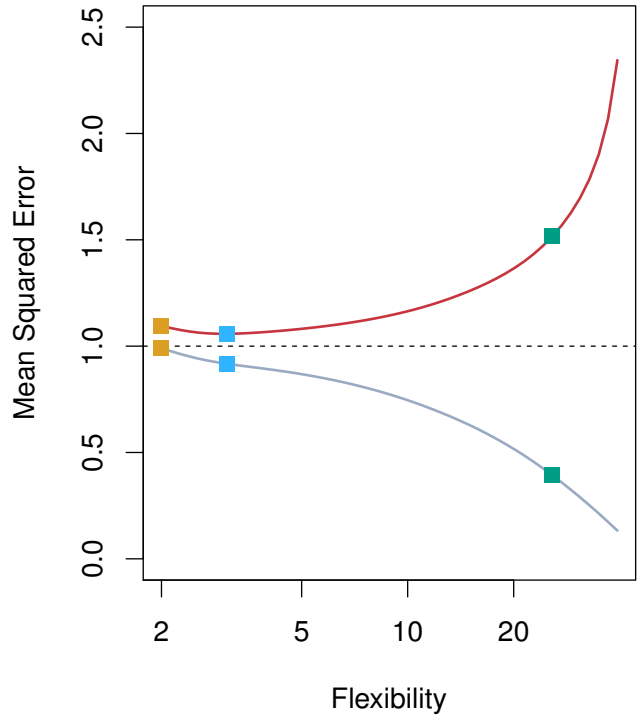
- A common measure of accuracy is the mean squared error (MSE),

$$\text{MSE} = 1/n \sum_i \left(Y_i - \hat{Y}_i\right)^2,$$

  where $\hat{Y}_i$ is the prediction using the training data.
- In general, we minimize MSE and care how the method works for new data, we call it test data.
- More flexible models could have lower MSE for training data but higher test MSE.

Levels of flexibility





- Black - Truth; Orange - Linear Estimate; Blue - smoothing spline; Green - smoothing spline (more flexible)
- RED - Test MSE; Grey - Training MSE; Dashed - Minimum possible test MSE (irreducible error)

Bias and Variance tradeoff

- There are always two competing forces that govern the choice of learning method i.e. bias and variance.
- Bias refers to the error that is introduced by modeling a real life problem (that is usually extremely complicated) by a much simpler problem.
- The more flexible/complex a method is the less bias it will generally have.
- Variance refers to how much your estimate for f would change by if you had a different training data set.
- Generally, the more flexible a method is the more variance it has.

Bias and Variance tradeoff

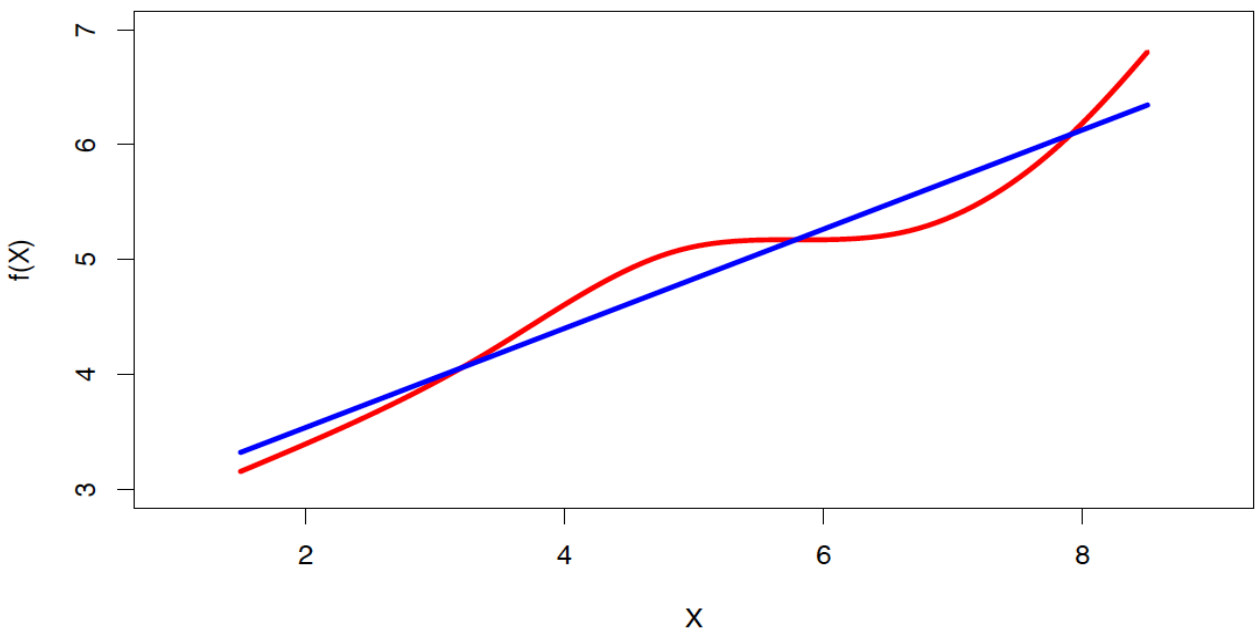
- For a new observation Y at $\boldsymbol{X} = \boldsymbol{X}_0$, the expected MSE is

$$E\left[\left(Y - \hat{Y}|\boldsymbol{X}_0\right)^2\right] = E\left[\left(f(\boldsymbol{X}_0) + \varepsilon - \hat{f}(\boldsymbol{X}_0)\right)^2\right] = \text{Bias}^2\left[\hat{f}(\boldsymbol{X}_0)\right] + \text{Var}\left[\hat{f}(\boldsymbol{X}_0)\right] + Var[\varepsilon].$$

- What this means is that as a method gets more complex the bias will decrease and the variance will increase but expected test MSE may go up or down!

Simple Linear Regression

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on $X_1, X_2, \cdots, X_p$ is linear.
- True regression functions are never linear! although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.



Simple Linear Regression

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- Simple Linear Regression Model (SLR) has the form of

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

where β_0 and β_1 are two unknown parameters (coefficients), called intercept and *slope*, respectively, and ε is the error term.

- Given the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, the estimated regression line is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x.$$

- For $X = x$, we predict Y by $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, where the hat symbol denotes an estimated value.

Estimate the parameters

- Let (y_i, x_i) be the i-th observation and $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, we call $e_i = y_i - \hat{y}_i$ the ith residual.
- To estimate the parameters, we minimized the residual sums of squares (RSS),

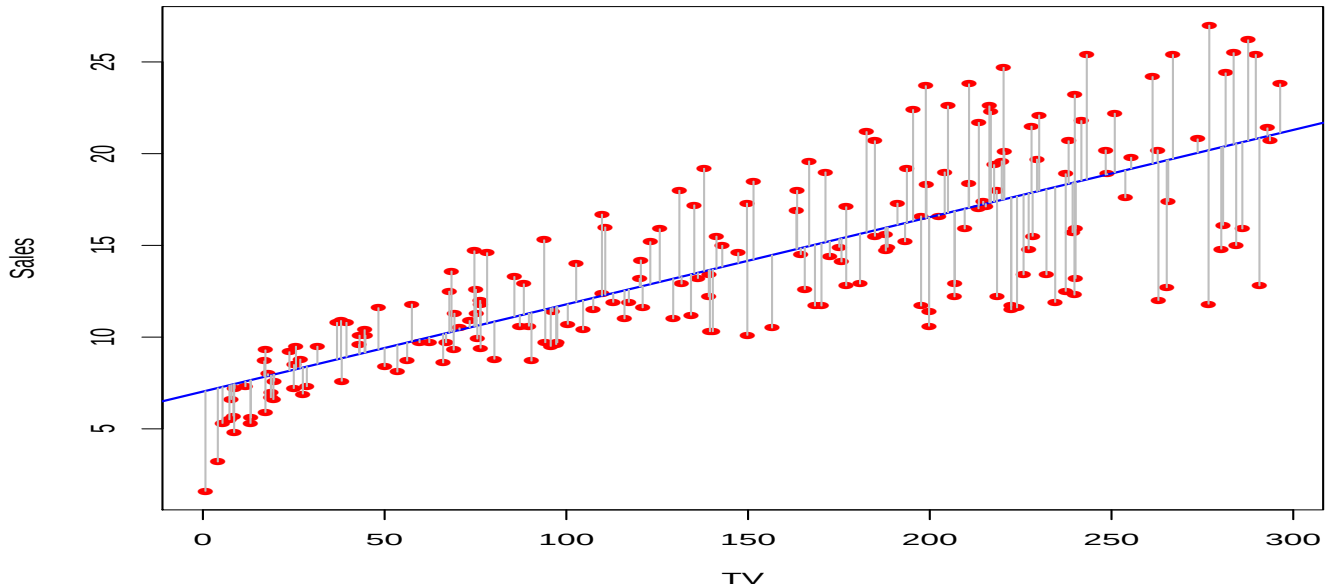
$$\text{RSS} = \sum_i e_i^2 = \sum_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \right)^2.$$

- Denote $\bar{y} = \sum_i y_i / n$ and $\bar{x} = \sum_i x_i / n$. The minimized values are

$$\hat{\beta}_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \left(r \frac{\sqrt{\sum_i (y_i - \bar{y})^2}}{\sqrt{\sum_i (x_i - \bar{x})^2}} \right),$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Example



- Advertising data: the least square fit for the regression of `sales` and `TV`.
- Each grey line segment represents an error, and the fit makes a compromise by averaging their squares.
- In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Assess the coefficient estimates

- The standard error of an estimator reflects how it varies under repeated sampling.

$$\text{SE}(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum(x_i - \bar{x})^2}}, \qquad \text{SE}(\hat{\beta}_0) = \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right)},$$

  where $\sigma^2 = \text{Var}(\varepsilon)$.

- A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.
- It has the form

$$\hat{\beta}_1 \pm 2 \cdot \text{SE}(\hat{\beta}_1).$$

- For the advertising data, the 95% confidence interval for β_1 is $[0.042, 0.053]$, which means, there is approximately 95% chance this interval contains the true value of β_1 (under a scenario where we got repeated samples like the present sample).

Hypothesis testing

- Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of

  H_0: There is no relationship between X and Y versus the alternative hypothesis

  H_A: There is some relationship between X and Y.
- Mathematically, we test

$$H_0 : \beta_1 = 0 \text{ versus } H_A : \beta_1 \neq 0,$$

since if $\beta_0 = 0$ then the model reduces to $Y = \beta_0 + \varepsilon$, and X is not associated with Y.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Hypothesis testing

- To test the null hypothesis, we compute a t-statistics,

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}\left(\hat{\beta}_1\right)}.$$

- This statistics follows t_{n-2} under the null hypothesis $\beta_1 = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the p-value.
- Results for the advertising data

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Measure of fit

- We compute the Residual Standard Error

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_i (y_i - \hat{y}_i)^2},$$

  where the residual sum-of-squares is $\text{RSS} = \sum_i (y_i - \hat{y}_i)^2$.
- R-squared or fraction of variance explained is

$$\text{R}^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}},$$

  where $\text{TSS} = \sum_i (y_i - \bar{y})^2$ is the total sum of squares.
- It can be shown that in this simple linear regression setting that $\text{R}^2 = r^2$, where r is the correlation between Y and X:

$$r = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_i (y_i - \bar{y})^2}\sqrt{\sum_i (x_i - \bar{x})^2}} = \left(\hat{\beta}_1 \frac{\sqrt{\sum_i (x_i - \bar{x})^2}}{\sqrt{\sum_i (y_i - \bar{y})^2}}\right).$$

R code

```
> TVadData = read.csv('... Advertising.csv')
> attach(TVadData)
> TVadlm = lm(Sales~TV)
> summary(TVadlm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- Multiple Linear Regression has more than one covariates,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

  where usually $\varepsilon \sim N(0, \sigma^2)$.
- We interpret β_j as the average effect on Y of a one unit increase in X_j, while holding all the other covariates fixed.
- In the advertising example, the model becomes

$$\text{Sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{Radio} + \beta_3 \times \text{Newspaper} + \varepsilon.$$

Coefficient Interpretation

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- The ideal scenario is when the predictors are uncorrelated — a balanced design.
  - Each coefficient can be estimated and tested separately.
  - Interpretations such as a unit change in X_j is associated with a β_j change in Y , while all the other variables stay fixed, are possible.
- Correlations amongst predictors cause problems.
  - The variance of all coefficient tends to increase, sometimes dramatically.
  - Interpretations become hazardous — when X_j changes, everything else changes.
- Claims of causality should be avoided for observational data.

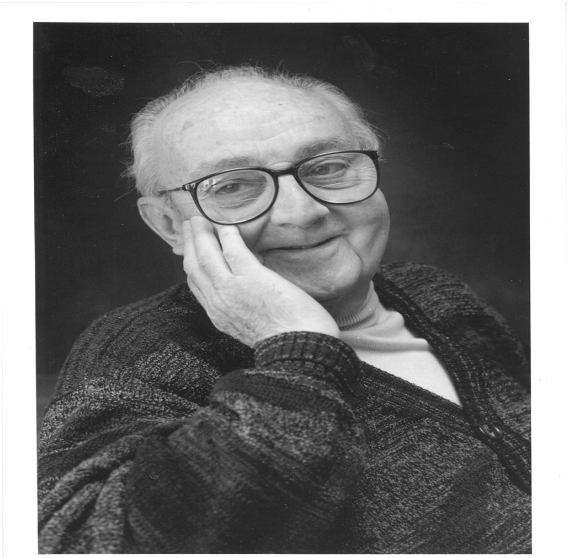
The woes of regression coefficients

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Data Analysis and Regression, Mosteller and Tukey 1977

- A regression coefficient β_j estimates the expected change in Y per unit change in X_j, with all other predictors held fixed. But predictors usually change together!
- Example: Y total amount of change in your pocket; $X_1 = \#$ of coins; $X_2 = \#$ of pennies, nickels and dimes. By itself, regression coefficient of Y on X_2 will be > 0. But how about with X_1 in model?
- $Y =$ number of tackles by a football player in a season; W and H are his weight and height. Fitted regression model is $Y = \beta_0 + 0.50W - 0.10H$. How do we interpret $\hat{\beta}_2 < 0$?

Two quotes by famous Statisticians



1919 - 2013 (aged 93)

- Essentially, all models are wrong, but some are useful.
  George Box
- The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively.
  Fred Mosteller and John Tukey, paraphrasing George Box

Coefficient estimation

- Given the estimates $\hat{\beta}_0, \hat{\beta}_1, \cdots,$ and $\hat{\beta}_p$, the estimated regression line is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p.$$

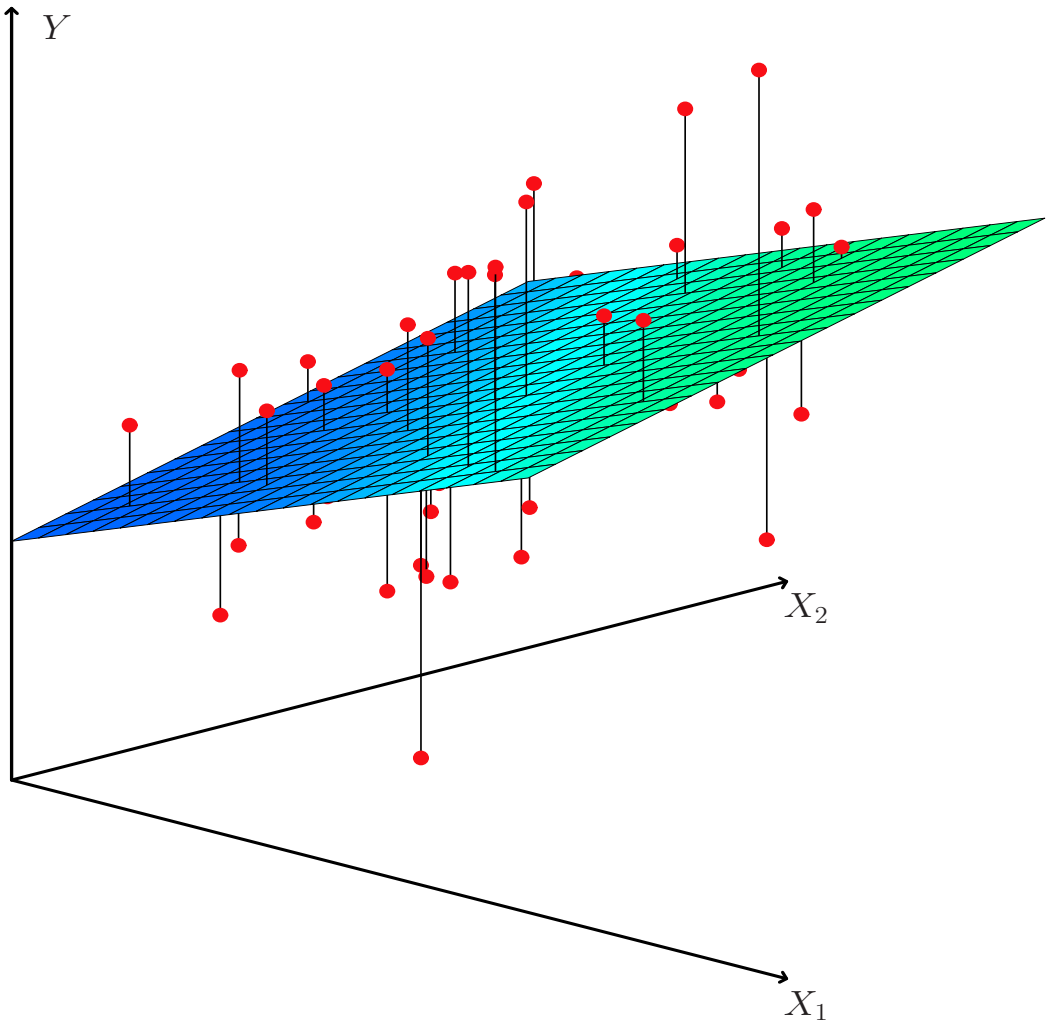
- We estimate all the coefficients $\beta_i, i = 0, 1, \cdots, p$ as the values that minimize the sum of squared residuals

$$\text{RSS} = \sum_i (y_i - \hat{y}_i)^2,$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$ is the predicted values.

- This is done using standard statistical software. The values $\hat{\beta}_0, \hat{\beta}_1, \cdots,$ and $\hat{\beta}_p$ that minimize RSS are the multiple least squares regression coefficient estimates.

Estimation Example



Inference

- Is at least one predictor useful?

$$\mathrm{F} = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)} \sim F_{p,n-p-1}.$$

- What about an individual coefficient, say if β_i useful?

$$t = \frac{\hat{\beta}_i - 0}{\mathrm{SE}\left(\hat{\beta}_i\right)} \sim t_{n-p-1}.$$

- For given $x_1, \cdots, x_p$, what is the prediction interval (PI) of the corresponding y?
- What about the estimation interval (CI) of y?
- What is the difference — PI, individual and CI, average, PI wider than CI.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Advertising example

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16

> predict(TVadlm, newdata, interval="c", level=0.95)
       fit      lwr      upr
1 20.52397 19.99627 21.05168
> predict(TVadlm, newdata, interval="p", level=0.95)
       fit      lwr      upr
1 20.52397 17.15828 23.88967
```

Indicator Variables

- Some predictors are not quantitative but are qualitative, taking a discrete set of values.
- These are also called categorical predictors or factor variables.
- Example: investigate difference in credit card balance between males and females, ignoring the other variables. We create a new variable,

$$x_i = \begin{cases} 1 & \text{if } i\text{-th person is female,} \\ 0 & \text{if } i\text{-th person is male} \end{cases}.$$

- Resulting model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{-th person is female,} \\ \beta_0 + \varepsilon_i & \text{if } i\text{-th person is male} \end{cases}.$$

- Interpretation and more than two levels (categories)?

Indicator Variables

- In general, if we have k levels, we need $(k-1)$ indicator variables.
- For example, we have 3 levels — $A, B,$ and C for a covariate x,

$$x_A = \begin{cases} 1 & \text{if } x \text{ is A,} \\ 0 & \text{if } x \text{ is not A} \end{cases}; \quad x_B = \begin{cases} 1 & \text{if } x \text{ is B,} \\ 0 & \text{if } x \text{ is not B} \end{cases}.$$

- If x is C, then $x_A = x_B = 0$. We call C as baseline.
- β_A is the contrast between A and C and β_B is the contrast between B and C.

Why Model Selection

- In many situations, many predictors are available. Some times, the number of predictors is even larger than the number of observations ($p > n$). We follow Occam's razor (aka Ockham's razor), the law of parsimony, economy, or succinctness, to include only the important predictors.
- The model will become simpler and easier to interpret (unimportant predictors are eliminated).
- Cost of prediction is reduced-there are fewer variables to measure.
- Accuracy of predicting new values of y may improve.
- Recall MSE(prediction) = Bias(prediction)^2 + Var (prediction).
- Variable selection is a trade off between the bias and variance.

How to select model in Linear Regression

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

- Subset Selection. We identify a subset of the p predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables. Best subset and stepwise model selection.
- Shrinkage. We fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as regularization) has the effect of reducing variance and can also perform variable selection.
- Dimension Reduction. We project the p predictors into a M-dimensional subspace, where $M < p$. This is achieved by computing M different linear combinations, or projections, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares.

Best subset selection

- Fit all possible models $(2^p - 1)$ and select a single best model from according certain criteria.
- Possible criteria include adjusted R^2, cross-validated prediction error, C_p, AIC, or BIC.
- We consider the adjusted R^2 statistics

$$R^2_{adj} = 1 - \frac{SSE/(n-q-1)}{SST/(n-1)},$$

where q is the number of predictors in the model.

- Adjusted R^2 criterion: we pick the best model by maximizing the adjusted R^2 over all $2^p - 1$ models.
- R^2 is suitable for selecting the best model as it always select the largest model to have smallest training error while we need to have small testing error.

AIC Criterion

- The AIC statistics for a model is defined as

$$AIC = -2l(y) + 2(q+1) \overset{LM}{=} n\log(SSE/n) + 2(q+1),$$

  where $l(y)$ is log-likelihood of y and q is the number of predictors in the model.
- The first part of AIC statistic decreases as the number of predictors in the model q increases.
- The second part increases as q increases. This part is to penalize larger models.
- The AIC statistics is not necessary to decrease or increase as q increases.
- AIC criterion: pick the best model by minimizing AIC criterion over all models.

BIC Criterion

- The BIC statistics for a model is defined as

$$BIC = -2l(y) + \log(n)(q+1) \overset{LM}{=} n\log(SSE/n) + \log(n)(q+1),$$

  where $l(y)$ is log-likelihood of y and q is the number of predictors in the model.
- Similar to AIC statistics, the BIC statistics adds the second part to penalize larger models.
- BIC criterion: pick the best model by minimizing BIC criterion over all models.
- The only difference between AIC and BIC is the coefficient for the second part.
- The BIC criterion can guarantee that we can pick all the important predictors as $n \longrightarrow \infty$, while the AIC criterion cannot.

Cross-Validation

- The idea of cross-validation (CV) criterion is to find a model which minimizes the prediction/testing error.
- For $i = 1, \ldots, n$, delete the i-th observation from the data and the linear regression model. Let $\hat{\boldsymbol{\beta}}_{-i}$ denote the LSE for $\boldsymbol{\beta}$. Predict y_i using $\hat{y}_{-i} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{-i}$.
- CV criterion: pick the best model by minimizing the $\text{CV} = \sum_{i=1}^{n}(y_i - \hat{y}_{-i})^2$ statistics over all the models.
- We did not use y_i to get $\hat{\boldsymbol{\beta}}_{-i}$ and we predict y_i as if it were new "observation".
- So CV statistics is simplified to

$$\text{CV} = \sum_{i=1}^{n}\left(\frac{r_i}{1-h_{ii}}\right)^2,$$

where h_{ii} is the ii-th element of the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

Mallow's C_p Statistic

- The C_p statistics is another statistic which penalizes larger model. In the original definition, p is the number of predictors in the model. Unfortunately, we use q to denote the number of predictors. In the following we use the notation C_q instead.
- The C_q statistics for a given model is defined as

$$C_q = \frac{SSE(q)}{SSE(p)/(n-p-1)} - (n - 2(q+1)).$$

- It can be shown that $C_q \approx q + 1$, if all the important predictors are in the model.
- C_q criterion: pick the model such that C_q is close to $q + 1$ and also q is small (we like simpler model).
- In linear model, under Gaussian error assumption C_p criterion is equivalent to AIC.

Backward Elimination

- Backward elimination starts with all p predictors in the model. Delete the least significant predictor.
- Fit the model containing all the p predictors $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$ and for each predictor calculate the p-value of the single F-test. Other criteria, say, AIC, BIC, C_p, apply as well.
- Check whether the p-values for all the p predictors are smaller than α, called alpha to drop.
- If yes, stop the algorithm and all the p predictors are treated as important.
- If not, delete the least significant variable, i.e., the variable with the largest p-value and repeat checking.

Forward Selection

- Forward Selection starts with no predictor in the model. Pick the most significant predictor.
- Fit p simple linear regression models

$$y = \beta_0 + \beta_1 x_j, \;\; j = 1, \ldots, p.$$

For each predictor, we calculate the p-value of the single F-test for the hypothesis $H_0 : \; \beta_1 = 0$. Other criteria, say, AIC, BIC, C_p, apply as well.

- Choose the most significant predictor, denoted by $x_{(1)}$ such that the p-value of the F-test statistic for the hypothesis $H_0 : \; \beta_1 = 0$ is smallest.
- If the p-value for the most significant predictor is larger than α (alpha to enter). We stop and no predictor is needed.
- If not, the most significant predictor is added in the model and we repeat choosing.

Stepwise selection

- A disadvantage of backward elimination is that once a predictor is removed, the algorithm does not allow it to be reconsidered.
- Similarly, with forward selection once a predictor is in the model, its usefulness is not re-assessed at later steps.
- Stepwise selection, a hybrid of the backward elimination and the forward selection, allows the predictors enter and leave the model several times.
- Forward stage: Do Forward Selection until stop.
- Backward stage: Do Backward Elimination until stop.
- Continue until no predictor can be added and no predictor can be removed according to the specified α to enter and α to drop.

Summary and Remark

- Install software **R**, if necessary, play demos, browse documentation.
- In my opinion, the best way to learn in this course is to try everything in **R**.
- Once it works, then think why, and how to write it in your own way.