Topic 4 – Simple Linear Regression

Basics/Review

Topics 2 & 3 explored relationships between a categorical (explanatory) variable and a numerical (response) variable. Suppose instead a relationship between two numerical variables. Specifically, a response and *exactly* one explanatory variable (hence the "simple").

A function is a rule that defines how numbers are changed (or mapped) to produce other numbers.

e.g. the natural log transformation had the following function:
$Z = \ln(Y)$

Z is a function of Y. It tells us how to change a number in Y to produce a number in Z. For example, the number '10' in Y is $\ln(10) = 2.30$ in Z.

Linear (Straight Line) Functions

A function mapping variable X to variable Y is linear if it can be written in the form

$Y = \beta_0 + \beta_1 X$ where β_0 is the intercept and β_1 is the slope.

Recall previous forms $y = a + bx$ or $y = ax + b$.

The Simple Linear Regression Model

If we assume linearity between response, Y, and explanatory variable, X, we may want to write the model as

$Y = \beta_0 + \beta_1 X$ (or, equivalently, $(Y \mid X)$ = "Y given X" $= \beta_0 + \beta_1 X$)

Suppose Y is weight and X is height. According to the model above, all subjects with the same height have the exact same weight. There is no account for any deviation of error from the model. To make the model realistic, we must allow for error or variability in the response at fixed values of the explanatory variable. Consider the following model for the response, Y,

(1) $Y = \beta_0 + \beta_1 X + \varepsilon$ (or, equivalently, $(Y \mid X)$ = "Y given X" $= \beta_0 + \beta_1 X + \varepsilon$)

where ε is a random error term. We assume, at any fixed value of X, $E(\varepsilon) = \mu_\varepsilon = 0$.

This new model contains a predictable part (a linear function of X) and an unpredictable part (some random error). The predictable part is the *expected* response at X, where the expected response is the average response. That is,

(2) $E(Y) = \mu_Y = \beta_0 + \beta_1 X$ (or, equivalently, $E(Y \mid X) = \mu(Y \mid X) = \beta_0 + \beta_1 X$)

Thus, the linear relationship is not between X and Y, but between X and μ_Y. For illustration, substituting (2) into (1) gives

$Y = \mu(Y \mid X) + \varepsilon$

Hence, for some fixed X, every response is equal to the average of all responses plus some error.

Estimating the Model

Observing data allows estimation of β_0 and β_1, thus giving estimated regression equation,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$ (Note: The '^' denotes an estimate.)

This is also called the prediction equation; in other words, based on observed data, the predicted value of Y (given X) is $\hat{Y}$. Again, the expected response (given X) is based on averaging, so our estimated model can be (equivalently) written as

$$\hat{Y} = \hat{\mu}(Y \mid X) = \hat{\beta}_0 + \hat{\beta}_1 X$$

Thus, the predicted value of Y (given X) is equivalent to the mean response and both of these terms come from calculation of the estimated regression equation.

Estimation and Prediction

Once the model is estimated, we can use it to estimate mean responses or predict single responses at certain values of the explanatory variable. Thus, for some value of X, say x, the estimated mean response or predicted single response is

$$\hat{y} = \hat{\mu}(Y \mid X = x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Again, the estimated mean response is the same as the predicted response. The estimated response is also called the fitted value.

Examples:

1. An ecologist wants to see if dimensions of bears are related.

The explanatory variable is X = height.
The response variable is Y = weight.
Suppose it is estimated that

$$\hat{\mu}(Weight \mid Height) = -316.215 + 7.085 Height$$

Then, the average weight for a bear that is 6 feet tall (72 inches) is
-316.215 + 7.085(72) = 193.905 lbs.

2. The same ecologist thinks that bear survival requires a minimum 50% meat intake in its weekly diet. They predict meat intake percentage by foraging index (*FI*).

Suppose it is estimated that

$$\hat{\mu}(Intake\% \mid FI) = 0.00933 + 0.190 FI$$

Thus, a foraging index of 2.5 yields a meat intake percentage of
0.00933 + 0.190(2.5) = 0.484 = 48.4%
This bear would not survive, based on this model. In order to survive, the bear must have an FI of at least 2.58.

Fitted Values, Residuals and Least Squares Estimation

The difference between the observed response and its fitted response (or estimated value) is called its residual, which we will denote as res_i.

$$res_i = Y_i - \hat{Y}_i = Y_i - \hat{\mu}(Y_i \mid X_i) = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

The method of least-squares regression finds the line (equivalent to finding the slope and intercept in the linear relationship) that minimizes the residual sum of squares denoted by *SS*(Residuals). That is, choose $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

$$SS(\text{Residuals}) = \sum_{i=1}^{n} (res_i)^2 = \sum_{i=1}^{n} \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\right)^2$$

Solving for $\hat{\beta}_0$ and $\hat{\beta}_1$ (by using calculus), we find

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \qquad \text{and} \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$\text{where } S_{XX} = \sum_{i=1}^{n} (X_i - \overline{X})^2 \text{ and } S_{XY} = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

Variability of the Response

We have dealt with 2 of 3 parts of the right side of (1). The third relates to the variability in the response at fixed *X*, or the error variability. By assuming the variability is constant (independent of *X*), we denote it as

$$Var(\varepsilon) = \sigma_\varepsilon^2$$

We must estimate the variance as the sum of squared deviations between each observation and its estimated value; thus, we look for *SS*(Residuals). Moreover, the value requires weighting by degrees of freedom such that

> Degrees of freedom associated with a set of residuals =
> (# of observations) – (# of parameters in the model for the means) = $n - 2$.
>
> Why $n - 2$? The number of parameters in the model is 2 (β_0 and β_1).
>
> Thus, we estimate $\sigma_\varepsilon^2 \approx \hat{\sigma}_\varepsilon^2 = s_\varepsilon^2 = \dfrac{SS(\text{Residuals})}{n-2} = MSE$

Inference for Regression Coefficients

When the SLR model holds, expectation properties show that

$$E(\hat{\beta}_1) = \beta_1 \text{ and } E(\hat{\beta}_0) = \beta_0$$

Thus, the above estimators are *unbiased*. Also,

$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \quad \text{and} \quad V(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]$$

Thus, the estimated standard error is

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \qquad \text{and} \qquad se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]}$$

One result is that the (conditional) dist'n of $\hat{\beta}_1$ is $N(\beta_1, \sigma^2 / S_{xx})$. The second is testing the significance of the slope. For H_0: $\beta_1 = 0$, the test statistic is

$$T_0 = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} \sim t_{n-2}$$

Note that $\hat{\beta}_1$ is independent of $\hat{\sigma}^2$, allowing for the simplicity of the standard error. Also, a similar test exists for the intercept. More importantly, H_0: $\beta_1 = 0$ can test for the *significance of regression*.

For sources of variability in the SLR model, the *ANOVA identity* is

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

An ESS *F*-test will also test for regression (see p. 209) between the reduced model (Equal-Means model) and the full model (SLR model). Thus, $F_0 = t_0^2$ (ONLY for SLR).

If the observations are normally and independently distributed, the CI for β_1 is

$$\hat{\beta}_1 \pm t_{\alpha/2,\, n-2}\ se(\hat{\beta}_1)$$

while the CI for β_0 is computed in a similar fashion.

Example:
a) Construct a 95% CI for β_1.

$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = 0.058$ (from SPSS) $\qquad t_{n\text{-}2,\alpha/2} = t_{33,0.025} \approx t_{30,0.025} = 2.042$

$\hat{\beta}_1 \pm t_{\alpha/2,\, n-2}\ se(\hat{\beta}_1) = 1.077 \pm (2.042)(0.058) = 1.077 \pm 0.118 = (0.959, 1.195)$

b) Is there clear evidence of a positive association between the two variables?
H_0: $\beta_1 \leq 0$ $\qquad$ H_a: $\beta_1 > 0$

$$t_0 = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{1.077 - 0}{0.058} = 18.569$$

In the *t*-table, the corresponding range of *p*-values is (0.0005, 0). Test is one-tailed, so the range is correct in the current form. Thus, there is convincing evidence against H_0 and there is a positive association between midterm and final marks. (Interval agrees.)

The "Effect" of an Explanatory Variable on the Mean Response
What is the effect of *X* on the mean of *Y*?

Effects of Explanatory Variables:

1. *For Numerical Explanatory Variables*: The effect of a numerical explanatory variable on the mean response is measured as the change in the mean response that is associated with a one-unit increase in that variable holding all other explanatory variables fixed.

2. *For Categorical Explanatory Variables*: There are potentially several effects for a categorical variable. There is an effect for each pair of levels, say level 1 vs. level 2, of the categorical variable. This effect of level 1 vs. level 2 on the mean response is measured as the difference in the mean response at level 1 vs. the mean response at level 2.

SLR: Consider an SLR model for now. What is the effect of X (a numerical explanatory variable) on the mean response Y?

By definition, the effect X on μ_Y for a SLR model is:
$\mu(Y \mid X+1) - \mu(Y \mid X) = (\beta_0 + \beta_1(X+1)) - (\beta_0 + \beta_1 X) = \beta_1$

Thus, as we increase the explanatory variable X by one unit, the mean of Y changes by β_1, the slope. Or, the effect on the μ_Y is β_1.

If the slope is zero ($\beta_1 = 0$), then X has no effect on the mean of Y. If the slope is positive, then X has a positive effect on the mean of Y. If the slope is negative, then X has a negative effect on the mean of Y.

Then testing

H_0: there is no linear relationship between X and the mean of Y

is equivalent to

H_0: the effect of X on the mean of Y is zero (or $\beta_1 = 0$).

Inference: In a randomized experiment, we can make a causal relationship between the two variables. In observational studies, we cannot make causal inferences. We can still, however, discuss the effect (by the definition above) of the explanatory variables on the response.

Inference for the Mean Response and the Predicted Response

- CI for the mean value of y corresponding to x^*

  $\hat{\mu}_{Y|x^*} \pm t_{\alpha/2,\, n-2}\ se(\hat{\mu}_{Y|x^*})$

  $$se(\hat{\mu}_{Y|x^*}) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

  Notice how the standard error increases with $(x^* - \bar{x})^2$. Why? (Being further away from the center of the x values denotes a "less precise" estimate.)

- Prediction interval for an individual value of y corresponding to x^*

  $\hat{y} \pm t_{\alpha/2,\, n-2}\ se(\hat{y})$

  $$se(\hat{y}) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$$

  Why is the PI wider than the CI? (Considering individual value, not mean.)

Example:
a) Give a 95% CI for *oneweek* when *initial* = 95 people. Compare with 95% PI.

$$se(\hat{\mu}_{Y|x^*}) = \hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x^*-\bar{x})^2}{(n-1)s_x^2}} = (11.466)\sqrt{\frac{1}{35}+\frac{(95-98.202)^2}{(35-1)(33.630)^2}} = 1.947$$

$\hat{\mu}_{Y|x^*} = \hat{\beta}_0 + \hat{\beta}_1 x^* = 16.863 + 1.077(95) = 119.178$ $\qquad t_{n\text{-}2,\alpha/2} = t_{33,\alpha/2} \approx 2.042$

$$\hat{\mu}_{Y|x^*} \pm t_{\alpha/2,\,33}\, se(\hat{\mu}_{Y|x^*}) = 119.178 \pm (2.042)(1.947) = 119.178 \pm 3.976 = (115.202,\ 123.154)$$

$$se(\hat{y}) = \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{(n-1)s_x^2}} = (11.466)\sqrt{1+\frac{1}{35}+\frac{(95-98.202)^2}{(35-1)(33.630)^2}} = 11.630$$

$$\hat{y} \pm t_{\alpha/2,\,n-2}\ se(\hat{y}) = 119.178 \pm (2.042)(11.630) = 119.178 \pm 23.749$$
$$= (95.429,\ 142.927)$$

b) Give a 95% CI for β_0. Is this result consistent with $\beta_0 = 0$? (Intercept at origin)

This requires that $x^* = 0$. Thus, $se(\hat{\mu}_{Y|x^*}) = (11.466)\sqrt{\frac{1}{35}+\frac{(0-98.202)^2}{(35-1)(33.630)^2}} = 6.060.$

$$\hat{\mu}_{Y|x^*} \pm t_{\alpha/2,\,33}\ se(\hat{\mu}_{Y|x^*}) = 16.863 \pm (2.042)(6.060)$$
$$= 16.863 \pm 12.375 = (4.488,\ 29.238)$$

Correlation
Def'n: (Pearson's) sample correlation coefficient r_{XY} is given by

$$r_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_x}\right)\left(\frac{Y_i-\bar{Y}}{s_y}\right)$$
$$= \frac{1}{n-1}\sum z_x z_y$$
$$= \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}}$$

(Example graphs of correlation: 1. strong positive linear; 2. weak positive linear; 3. negative linear; 4. no pattern; 5. parabola; 6. exponential)

Properties of r_{XY}:

- A measure of the LINEAR relationship between two variables.
- $-1 \le r_{XY} \le 1$
- The magnitude of r_{XY} measures the strength of the relationship:
  - If $r_{XY} = \pm 1$, then the points follow a straight line.
  - If $r_{XY} = 0$, then the pattern of scatter suggest no linear relationship.
- The sign of r_{XY} indicates the nature of the relationship:
  - Positive association if $r_{XY} > 0$,

- Negative association if $r_{XY} < 0$.
- Sign of r_{XY} = sign of β_1.
- The two variables x and y play symmetric roles.
- Location and scale invariance (unitless).
- Correlation does not imply causation (or vice versa).
- We can have $r_{XY} = 0$, even when the data reveal a strong nonlinear relationship.
  - e.g. $y = x^2$

R-squared: The Coefficient of Determination

Properties of R^2:

Only in SLR

- $R^2 = 100\%\left(\dfrac{SS(\text{Total}) - SS(\text{Residual})}{SS(\text{Total})}\right) = r_{YX}^2$, $\quad SS(\text{Total}) = S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- The formula above represents the coefficient of determination.
- R^2 measures the proportion of variation in Y explained by X.
- $0 \le R^2 \le 1$.
- $R^2 = 1$ when $SS(\text{Residual}) = 0$, which only happens when all points fall exactly on the line. Thus, there is no error in using x to predict y and, also, $r_{XY} = \pm 1$.
- $R^2 = 0$ when $SS(\text{Residual}) = SS(\text{Total})$, which happens when the regression line and $\bar{y}$ give the same predictions. Thus, the slope $\hat{\beta}_1 = 0$.
- The closer R^2 is to 1, the stronger the association.

Interpolation and Extrapolation

The SLR model allows for inferences about any mean response within the range of X. Estimating values of Y outside the observed values of X is called extrapolation. This process is dangerous because the linear relationship may not hold outside this range. On the other hand, interpolation involves estimating responses at values of X not in the data but within the range of the observed values of X; this process is quite useful.

Causation

Recall causal inferences from Topic 1. The presence of an association between two variables does NOT imply that the X causes the specific response. **Association does NOT imply causation**. As a result, observational studies permit only an association between the two variables. Only a randomized experiment may infer a causal relationship.

Assumptions and Robustness for Inferential Tools in Linear Regression

Recall the model: $Y = \beta_0 + \beta_1 X + \varepsilon$

1. Linearity: There is a real linear relationship between the *mean* response and the explanatory variable. In other words, the expected error is zero. $E(\varepsilon) = \mu_\varepsilon = 0$.

   Checking: In a scatterplot of Y vs. X, look for a linear pattern!

   Robustness: If no linear relationship exists in the data, don't use linear regression!

Problem?: Consider transformation. If problem only in linearity (not in constant variance), try transforming X. If both problems exist, transform Y first.

2. Constant Variance: The spread of the responses about the line is the same at all levels of the explanatory variable. In other words, the variability of the errors at any fixed X is the same for all values of X. At any fixed X, $Var(\varepsilon) = \sigma_\varepsilon^2$.

   Checking: In a scatterplot of "residuals" vs. "fitted values", look for an equal spread about the horizontal line at zero. Checking for roughly the same spread about the line across all values of X on the original scatterplot works, as well. Usually, the residual plot is better.

   Robustness: The consequence for violating this assumption is critical. When there is lack of equal variability, the resulting standard errors inaccurately estimate their respective parameters, thus confidence intervals and hypothesis tests can be misleading.

   Problem?: Consider transformation of Y.

3. Normality: The sub-populations of responses, at different values of X, all have a normal distribution. Equivalently, at any fixed X, the errors are normally distributed, or centered on the line. At any fixed X, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$.

   Checking: Normal Probability Plots. Look for an increasing linear pattern.

   Robustness: Standard errors are robust to non-normal distributions (t-tools). The consequences of violating this assumption are usually minor. The only situation of large concern is when the distributions have long tails, outliers are present, and/or sample sizes are small. In terms of prediction, normality is critical.

   Problem?: Consider transformation of Y, only if the problem is very serious.

4. Independence: All responses are independent of each other.

   Checking: Evaluate study design (as per Topic 2).

   Robustness: Lack of independence causes inaccurate standard errors.

   Problem?: Account for cluster dependence through analysis of variance or multiple regression. Time series analysis may help serial dependence.

Illustration of Interpretation of Model Effects in SLR after Log Transformation

Recall the *effect* of a numerical explanatory variable on the mean response is measured as the change in the mean response that is associated with a 1-unit increase in that variable, holding all other variables fixed.

SLR model of Blood Pressure on Age: The effect of age on mean blood pressure is measured as the change in the mean blood pressure associated with a 1-unit (one year) increase in age. This effect is measured as β_1 in the model below.

$\mu(bp \mid age) = \beta_0 + \beta_1 age$

This is known as an additive effect.

IN GENERAL, for Y vs. X:

Additive change of k units in X ($X + k$) → *Additive* change of $k\beta_1$ in the mean of Y.

Suppose (in separate circumstances) that the following natural log transformations were required. Interpretation of the model effect on the original scale will follow.

i) a natural log transformation was used on the response variable only. (ln Y vs. X)
ii) a natural log transformation was used on the predictor variable only. (Y vs. ln X)
iii) a natural log transformation was used on both variables. (ln Y vs. ln X)

Case i): (ln Y vs. X)

Model: $\mu(\ln(bp) \mid age) = \beta_0 + \beta_1 age$ → $\hat{\mu}(\ln(bp) \mid age) = 4.481 + 0.010 age$

The additive effect of *age* on $\mu_{\ln(bp)}$ is β_1.

IN GENERAL, for ln Y vs. X:

Additive change of k units in X → *Multiplicative* change of $e^{k\beta_1}$ in the median of Y.

For example, the additive effect of 1-year ($k = 1$) in *age* is associated with a multiplicative effect of $e^{1(\beta_1)} = e^{\beta_1}$ on *Median*(*bp*). It is estimated that a 1-year increase in *age* is associated with a multiplicative change of $e^{0.010} = 1.01$ in *Median*(*bp*). In other words, the median *bp* at *age* + 1 is estimated to be 1.01 times (1% higher than) the median *bp* at *age*.

More examples:

- It is estimated that a 5-year increase in *age* is associated with a multiplicative change of $e^{5(0.010)} = 1.051$ in *Median*(*bp*). In other words, the median *bp* at *age* + 5 is estimated to be 1.051 times (5.1% higher than) the median *bp* at *age*.
- A 10-year increase in *age* is associated with a multiplicative change of $e^{10(0.010)} = 1.105$ (a 10.5% increase) in *Median*(*bp*).

- The average blood pressure of 55-year-olds will be 22% higher than the average blood pressure of 35-year-olds ($e^{20(0.010)} = 1.22$).

Case ii): (Y vs. ln X)

Model: $\mu(bp \mid \ln(age)) = \beta_0 + \beta_1 \ln(age)$ → $\hat{\mu}(bp \mid \ln(age)) = -81.784 + 58.967\ln(age)$

The additive effect of ln(*age*) on μ_{bp} is β_1.

IN GENERAL, for Y vs. ln X:

Multiplicative change by a factor of k in X ($X \times k$)	→	*Additive* change of $\beta_1 \ln(k)$ in the mean of Y.

For example, a multiplicative change in *age* by a factor of k is associated with an additive change of ($\beta_1 \ln(k)$) in μ_{bp}. It is estimated that a multiplicative change in *age* by a factor of k is associated with an additive change of $58.967\log(k)$ in μ_{bp}.

More examples:

- It is estimated that aging from 40 to 50 ($k = 50/40 = 1.25$) is associated with an increase in μ_{bp} of $58.967(\ln(1.25)) = 13.16$.
- It is estimated that aging from 28 to 35 ($k = 35/28 = 1.25$) is associated with an increase in μ_{bp} of $58.967(\ln(1.25)) = 13.16$.
- It is estimated that aging from 30 to 50 ($k = 50/30 = 1.67$) is associated with an increase in μ_{bp} of $58.967(\ln(5/3)) = 30.12$.

Case iii): (ln Y vs. ln X)

Model: $\mu(\ln(bp) \mid \ln(age)) = \beta_0 + \beta_1 \ln(age)$ → $\hat{\mu}(\ln(bp) \mid \ln(age)) = 3.332 + 0.426\ln(age)$

The additive effect of ln(*age*) on $\mu_{\ln(bp)}$ is β_1.

IN GENERAL, for ln Y vs. ln X:

Multiplicative change by a factor of k in X ($X \times k$)	→	*Multiplicative* change of k^{β_1} in the median of Y.

For example, a multiplicative change in *age* by a factor of k is associated with a multiplicative change of k^{β_1} in *Median*(*bp*). It is estimated that a multiplicative change in *age* by a factor of k is associated with a multiplicative change of $k^{0.426}$ in *Median*(*bp*).

More examples:

- It is estimated that aging from 40 to 50 ($k = 1.25$) is associated with a 9.97% increase (a multiplicative change of $1.25^{0.426} = 1.0997$) in *Median*(*bp*).
- It is estimated that aging from 35 to 55 ($k = 1.57$) is associated with a 21.2% increase (a multiplicative change of $1.57^{0.426} = 1.212$) in *Median*(*bp*).