Chapter 12 – Simple Linear Regression

Notation:

- bivariate sample: $\{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \}$
- sample means: $\bar{x}$, $\bar{y}$
- sample std dev.: s_x, s_y
- sums of squares and cross-products:

$$S_{XX} = \sum (x_i - \bar{x})^2 = \sum x^2 - \frac{\left(\sum x\right)^2}{n} = (n-1)s_x^2$$

$$S_{YY} = \sum (y_i - \bar{y})^2 = \sum y^2 - \frac{\left(\sum y\right)^2}{n} = (n-1)s_y^2$$

$$S_{XY} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}$$

Terminology:

x	y
Explanatory variable	Response variable
Independent variable	Dependent variable
Predictor variable	Predicted variable

Ex12.1) Four variables of current Oilers roster: height, weight, jersey, age

- which relationships might be valid?
- how can we describe the relationship between any pair?
- how do we use the description to make predictions?
- how do we quantify errors in estimates and predictions?

Simple Linear Regression (SLR) model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad E(Y_i \mid x_i) = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n$$

- $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are the observed data.
- $\epsilon_1, \ldots, \epsilon_n$ are unobserved "errors", assumed to be a random sample from $N(0, \sigma^2)$.
- $\beta_0, \beta_1, x_1, \ldots, x_n$ are assumed to be fixed; $y_1, \ldots, y_n$ are random variables.
- β_0, β_1, σ are unknown parameters.
  - β_0 is the "population" intercept.
  - β_1 is the *average* change in y associated with a 1-unit increase in x.
  - σ determines the extent to which points deviate from the line
- The conditional distribution of y_i given x_i is $N(\beta_0 + \beta x_i, \sigma^2)$.
  - $E(Y \mid x) = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x + E(\epsilon) = \beta_0 + \beta_1 x$
  - $V(Y \mid x) = V(\beta_0 + \beta_1 x + \epsilon) = V(\beta_0 + \beta_1 x) + V(\epsilon) = 0 + \sigma^2$
- Basic Assumptions of the SLR Model
  - The distribution of ϵ at any x has a mean of zero ($\mu_\epsilon = 0$ to aid linearity).
  - The std. dev. of ϵ is the same for any x (i.e. it's constant).
  - The distribution of ϵ at any x is normal.
  - The random deviations $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ associated with different observations are independent of one another.

12.2 Fitting the Regression Line

Least squares estimation of β_0 and β_1:

For any given line $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, the value $\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$ represents the vertical deviation of the point from the line. We want to choose (β_0, β_1) to minimize the sum of squared deviations (hence "least squares"):

$$\sum (y_i - \beta_0 - \beta_1 x_i)^2$$

Using calculus, the corresponding solution becomes

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \qquad \text{and} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Subsequently, the estimated regression line is: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. We interpret $\hat{y}$ as:

- estimator of $\beta_0 + \beta_1 x$ = the conditional mean of y given x
- predictor of new individual y values given x

NOTE: Extrapolation can be dangerous.

Ex12.2) Choosing to predict final pctg from midterm pctg (both vars. are continuous)

x = midterm percentage (in %), y = final percentage (in %)

Via calculation, $n = 99$, $\bar{x} = 70.20$, $\bar{y} = 59.33$,

$$S_{XX} = 20845.96,\ S_{YY} = 23600.42,\ S_{XY} = 12680.25$$

a) Determine the estimated regression line.

b) Estimate final percentage when midterm percentage is 50%.

c) Estimate final percentage when midterm percentage is zero.

(Excel example shown.)

Estimating error:

- Predicted or fitted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.
- Residuals: $e_i = y_i - \hat{y}_i = y_i - \beta_0 - \beta_1 x_i$.
- Residual plots are often used as a diagnostic tool. Plot of x vs. residuals. (plots drawn in class and shown with Excel example)
- Residual sum of squares (formula not in textbook):

$$SSE = \sum (y_i - \hat{y}_i)^2 = S_{YY} - \hat{\beta}_1 S_{XY}$$

- Estimate the error variance σ^2 by

$$\hat{\sigma}^2 = s_e^2 = \frac{SSE}{n-2} = \frac{\sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i}{n-2}$$

- Warning: Do NOT confuse s_e with s_y or s_x.
- Why divide by $n - 2$? (estimation of β_0 and β_1 is a loss of 2 degrees of freedom)

Ex12.3) Estimate σ.

<u>12.3 Inferences on the Slope</u>

When the 4 assumptions of the SLR model are satisfied, then $E(\hat{\beta}_1) = \beta_1$.
Also,

$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \qquad \rightarrow \qquad S.E.(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

Since $\hat{\beta}_1$ is a linear combination of normally distributed random variables Y_i, then

$$\hat{\beta}_1 \sim N(\beta_1,\ \sigma^2 / S_{xx}).$$

Testing the significance of the slope creates H_0: $\beta_1 = b_1$, such that the test statistic is

$$t_0 = \frac{\hat{\beta}_1 - b_1}{S.E.(\hat{\beta}_1)} \sim t_{n-2}$$

Both two-sided and one-sided tests are possible. More importantly, H_0: $\beta_1 = 0$ can test for the *significance of regression*.

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

$$\hat{\beta}_1 \pm t_{\alpha/2,\, n-2} \times S.E.(\hat{\beta}_1)$$

Ex12.4) a) Construct a 95% CI for β_1.

b) Is there evidence that final percentage increases as midterm percentage increases?

12.6 ANOVA for SLR

A method called ANalysis Of VAriance (ANOVA) can also test significance of regression. For sources of variability in the SLR model, the *ANOVA identity* is

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$$SST = SSR + SSE$$

where SST is the total corrected sum of squares and SSR is the regression sum of squares. This latter term summarizes how much less error there is in predicting y using the regression line compared to using $\bar{y}$. Due to the presence of "squares", an appropriate test requires a ratio of values. Thus, for H_0: $\beta_1 = 0$, the test statistic is

$$F_0 = \frac{SSR/1}{SSE/(n-2)} = \frac{MSR}{MSE} \sim F_{1,\,n-2}$$

Def'n: The *F*-distribution has the following properties:

1. It is continuous and skewed to the right.
2. It has two parameters: v_1 for the numerator and v_2 for the denominator.
3. The units of an F distribution are nonnegative.

ANOVA Table

Source	**SS**	**df**	**MS**	***F***	***p*-value**
Regression	SSR	1	MSR	MSR/MSE	$P(F_0 > F_{1,\,n-2})$
Error	SSE	$n-2$	MSE		
Total	$\boldsymbol{SST}$	$\boldsymbol{n-1}$			

Note: MS denotes mean squares, and, always, MS = SS/df for a particular row.
Also, $\hat{\sigma}^2 = MSE$ and $F = t^2$.

(Excel example shown with full hypothesis test carried out in class.)

12.4/12.5 Inferences based on the estimated regression line

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

  $$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,\,n-2} \times \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}}$$

  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^*

  $$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,\,n-2} \times \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}}$$

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

Ex12.5)

a) Give a 95% CI for FPct when MPct = 75%. Compare with 95% PI.

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

12.6/12.9 "R-squared"/Correlation

- R^2 = coefficient of determination
  = the proportion of variance of y explained by regression on x
  $= \dfrac{SSR}{SST} = 1 - \dfrac{SSE}{SST}$ $= R^2$ = squared correlation coefficient
- The coefficient of determination is called R^2 in the context of multiple linear regression (several predictors), but we have $R^2 = r^2$ in the context of simple linear regression (one predictor).

Def'n: (Pearson's) sample correlation coefficient r is given by

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$= \frac{1}{n-1}\sum z_x z_y$$

$$= \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}}$$

(Example graphs to show correlation were drawn in class: 1. strong positive linear; 2. weak positive linear; 3. negative linear; 4. no pattern; 5. parabola; 6. exponential)

Properties of r:

- A measure of the LINEAR relationship between two variables.
- The population correlation is denoted by ρ (rho).
- $-1 \le r \le 1$ and $-1 \le \rho \le 1$
- The magnitude of r measures the strength of the relationship:
  - If $r = \pm 1$, then the points follow a straight line.
  - If $r = 0$, then the pattern of scatter suggest no linear relationship.
- The sign of r indicates the nature of the relationship:
  - Positive association if $r > 0$,
  - Negative association if $r < 0$.
- Sign of r = sign of β_1.
- The two variables x and y play symmetric roles.
- Location and scale invariance (unitless).
- We can have $r = 0$, even when the data reveal a strong nonlinear relationship.
- Correlation does not imply causation (or vice versa).
- Correlation is sensitive to outliers.

Ex12.6)
a) What proportion of the variance of final percentage is explained by midterm percentage?

b) What is the correlation between final percentage and midterm percentage?

12.7 – Residual Analysis (discussed back in 12.2; consider also "normality plot" of e_i)