Ch. 5 and 13: Linear Regressions

Part I: (mostly review, hw 9)

In a regression analysis, we use independent variables/predictors (x) to try to predict the dependent variable/response (y).

From constructing a scatterplot, you can determine if there is a positive or negative relationship between the predictor and the response.

A negative slope means as the value in one variable increases, the value in the other decreases. If there is a positive slope, then as values in one variable increases, then the value in the other variable also increases.

We can summarize the relationship between the predictor and response with a linear regression model:

$$Y = \alpha + \beta x + e$$

β is the true slope, α is the true y-intercept, e is the error term

*Note that the Greek letters here still mean "true value", which means that we will have formulas to predict those parameters and do hypothesis testing!

Assumptions:
- $e_i \sim N(0, \sigma^2)$
- All e_i are independent of one another

Implications:
- Y is a random variable that follows a normal distribution with $E(Y_i) = \alpha + \beta x_i$ and $Var(Y_i) = \sigma^2$.
- All of the Y_i, Y_j are independent.

With these assumptions, we have a way of predicting α and β by a method of **"least squares"** and getting a predicted linear model:

$$\hat{y} = a + bx$$

*Note that this is a and b, which are estimated values, whereas α and β are the true parameters of the linear regression. And $\hat{y}$ is the predict response.

$$b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$a = \bar{y} - b\bar{x}$$

For a set of n pairs of values (x_i, y_i)

$$S_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}$$

$$S_{yy} = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}$$

$$S_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n}$$

A residual ($\hat{e}_i$) is the difference between the true observed y and the predicted value $\hat{y}$:

$$\hat{e}_i = y - \hat{y}$$

Recall the assumption that the error term is normally distributed with a mean of 0 and variance σ^2. We can find an estimate of this variance by finding the **mean square error** (MSE):

$$\text{MSE} = \frac{\text{SSE}}{n-1}$$

Sum of Squares Error (SSE) $= \sum_i(\hat{e}_i^2) = \sum_i(y_i - \hat{y}_i)^2$

Sum of Squares Regression (SSR) $= \sum_i(\hat{y}_i - \bar{y})^2$

Sum of Squares Total (SST) $= \text{SSR} + \text{SSE} = \sum_i(y_i - \bar{y})^2$

These values (along with some other relevant numbers) can be organized in an ANOVA table.

ANOVA Table:

Source	df	Sum of Squares	Mean Square	*F* statistic
Regression	1	SSR	$MSR = \frac{SSR}{1}$	$\frac{MSR}{MSE}$
Error	$n-2$	SSE	$MSE = \frac{SSE}{n-2}$	
Total	$n-1$	$SSTo$		

The standard error of residuals, denoted s_e is equivalent to $\sqrt{\text{MSE}}$. Also, $\hat{\sigma}^2 = \text{MSE}$.

Some examples

1. Given three points: (1,3), (2, 4), (4, 6); calculate S_{xx} and S_{yy}

2. Fill in the blanks of the ANOVA table.

Source	Df	SS	MS	F
Regression				93.44
Error	12		13.4	
Total	14	2672		

3. Calculate s_e, which is the standard error of residuals.

4. Calculate $\hat{\sigma}^2$.

Answers:

1. $s_{xx} = 14/3, s_{yy} = 14/3$

2. There's more than one approach to solve this problem, here's one way:
   Degrees of freedom for regression is $14 - 12 = 2$, which is equal to the number of independent variables in a model. You've probably seen $df_R = 1$, because there is only one predictor
   $\frac{SSE}{df} = MSE$, so $SSE = MSE * df = 13.4 * 12 = 160.8$
   Next, $SSR + SSE = SST$ by definition, so $SSR = SST - SSE = 2672 - 160.8 = 2511.2$
   Since $MSR = SSR/df = 2511.2/2 = 1255.6$
   Lastly, we can check that our answer is correct by calculating the F-stat with our numbers: $F = \frac{MSR}{MSE} = \frac{1255.6}{13.4} = 93.7$, which is close enough to the value in the table (93.44), small differences can be due to rounding.

3. $s_e = \sqrt{MSE} = \sqrt{13.4} = 3.66$

4. The predictor for variance is the MSE.
   $\hat{\sigma}^2 = MSE = 13.4$

Part II: Inferences and residual analysis

First, we have **hypothesis testing**. Similar to previous chapters, follow the 5-steps:

(1) Null hypothesis: $H_0 : \beta = 0$

(2) Alternative: $H_1 : \beta \neq 0$

(3) T-stat: $t_{obs} = \frac{b}{\sqrt{\frac{MSE}{S_{xx}}}}$

The denominator is the formula for s_b (estimate of the standard error of the slope)

(4) Rejection region: find the critical value $t_{\alpha/2}$ that has $n - 2$ degrees of freedom

(5) Decision: Reject H_0 if $|t_{obs}| > t_{\alpha/2}$

Confidence interval:

$$b \pm t_{\alpha/2}\sqrt{\frac{MSE}{S_{xx}}}$$

Residual analysis:

We can look at residuals to assess our original assumptions. Specifically a graph with the residuals on the y-axis and the predicted values on the x-axis. If we see no relationship, then our assumptions are likely true for the data. You don't want to see fanning either.

We can detect an outlier by looking at **leverage**.

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

Rule of thumb: High leverage if h_{ii} is greater than **2 times** the average leverage, (average leverage = **(k+1)/n** , k = number of predictors, you will usually see in this class 2/n since k = 1.)

Or we can look at the **standardized residuals**.

$$r_i = \frac{\hat{e}_i}{\sqrt{MSE(1 - h_{ii})}}$$

Rule of thumb: Outlier if $r_i > 2$ or $r_i < -2$. You may be asked to identify the outliers on a plot of standardized residuals.