(possibly incomplete list of)

Topics Covered in Chapters 7 and 8

Chapter 7 – Simple Linear Regression

The regression model equation (as given on page 290)
The four assumptions for linear regression (as given on pages 291-292)
How to visualize the linear regression model (figure 7.2 on page 292)
Independent vs. dependent variable
Regression and Causality
Regression and Extrapolation

Difference between the model parameters and their estimates ($\beta_0, \beta_1, \varepsilon$ vs. $\hat{\beta}_0, \hat{\beta}_1, e$)

Regression to the mean and why we "look at slices of *x*"
The estimates of β_0 and β_1 are gotten by minimizing the sum of the squared residuals (SSE)

The analysis of variance table and $\sqrt{MSE}$

- TSS is the error/variation when we assume H_0 is true (the error in *y* without taking *x* into account)
- SSE is the error/variation when we use the line we got from the data
- SSR = TSS - SSE is the amount of error we explained by using the regression line
- the MS are like variances
- TMS (which doesn't show up on the table) is the variance of y, or what the variance of the errors would be for a flat regression line
- MSE is the estimate of the variance of the errors ($\hat{\sigma}_\varepsilon^2 = s_e^2$) when the regression line is used
- because we assumed the residuals are normal we can use the MS to do an F test

How to find the mean squares if you are given the sum of squares and degrees of freedom
How the degrees of freedom are found

- for total and error it is the sample size minus the number of parameters estimated
- for the regression it is df for the total - the df for the error

F = MSR/MSE
When we accept or reject $\beta_1 = 0$ based on the ANOVA table

Interpreting $\hat{\beta}_1$, $\sqrt{MSE}$, and the p-value

The t-test and confidence interval for β_1 (and how to find $\sqrt{MSE/S_{xx}}$ from the SAS output)

The predicted value for *y* at a given *x*
The confidence interval for the mean response / regression line / $\mu_{y|x}$ (page 304-305)
The prediction interval for a new observation / $y_{y|x}$ (page 304-305)

What the different parts of the standard error for $\hat{\mu}_{y|x}$ and $\hat{y}_{y|x}$ are due to

Why isn't the variance for the predicted value of y given x just the MSE?
The correlation as a single number summary of regression:

- *r* is the "slope of the regression line after adjusting for the scale of *y* and *x*"
- r^2 = coefficient of determination = percent of variation in *y* explained by the regression
- larger r^2 implies larger F statistic

A feeling for what r^2 value different data sets will have by looking at the *y* vs. *x* plot
How to use the residual vs. predicted plot to check that the linear form is appropriate (mean of errors is zero) and that the errors have constant variance (both, at each level of the independent variable)
How to use the Q-Q plots to check that the errors are approximately normally distributed.
That linear regression is robust, and what that means
Reading the SAS output

NOT: The test and confidence interval for the correlation coefficient on page 318

Chapter 8 - Multiple Regression

Uses m independent variables instead of 1, so that there are m+1 βs.
The coefficients in multiple regression reflect the change in the value of the dependent variable when one of the independent variables changes and the rest stay the same. This is sometimes impossible.

That the ANOVA table, $\sqrt{MS_{res}}$, and r^2 work the same as in simple linear regression

The hypotheses that the ANOVA table, Type I Tests, and Type III Tests test (in the supplement to 8.3)

Using the logarithm or square root of the dependent variable to stabilize variance (not in text)
Taking logarithm of x and logarithm of y changes the linear regression into a multiplicative one (pg. 375-378)
Tranformations of y and x for curved simple linear regression (not in text)

Multicollinearity is when several of the predictor variables are highly correlated. This can lead to: 1) parameter estimates being unintuitively negative, 2) concluding a variable doesn't influence the dependent variable when it actually does. Note that multicollinearity doesn't violate any of the assumptions though. The tests actually work, they just don't answer the question we "want them to".
VIF can be used to measure the multicollinearity of a variable. No variance inflation is a value of 1. A VIF of 10 or more is commonly taken to mean that there is severe multicollinearity. In reality much smaller values can be associated with trouble.
Why getting more data can help reduce multicolinearity
Why scaling things appropriately (for rate of inflation, population etc.) can help reduce multicolinearity

Why r^2 is bad for comparing models with different numbers of x variables and why we need to have an adjusted r^2
Using Mallow's C_p to remove some variables: If the model with all of the variables is good, then any model with C_p near p+1 or less (p = # terms left in the model) should be unbiased (e.g. an ok choice)
Might then choose one with as few variables as possible if concerned with being easy to explain, or choose one with the biggest adjusted r^2 or smallest $\sqrt{MSE}$ if you want to predict accurately, and you might have to keep certain variables in or out for "political" reasons.

Potential = Leverage = Hat Diagonal = h_{ii} measures the potential of the observation to change the regression based only on the x values. That is, it indicates how similar the observations x values are to the other observations. It indicates a possibly troublesome x value if it is much larger than the other values. Rule of thumb is to be concerned if $> 2(m+1)/n$ (H in SAS)
Externally Studentized Residuals - the residuals rescaled to take into account the MSE so that they don't depend on the scale of y, and so that they are calculated without the observation in question. Should be distributed similarly to a standard normal distribution, so that values larger than 2 should be uncommon, and larger than 3 should be rare. (RT in SAS)
Influence – DFITTS combines the leverage and studentized residual and gives a measure of how much removing the observation would change the model. Values Greater than $2\sqrt{(m+1)/n}$ are typically taken to indicate removal will have a significant effect. (F in SAS)
Why a group of 2 or more points might be outliers, but not show up as having leverage or influence
What you can do if you have an outlier

NOT: The material in the text for 8.2 and 8.3 (as opposed to the supplement which you do need to know!)
NOT: The partial correlation on page 364-365.
NOT: Polynomial regression on page 370-374, although its an interesting topic and you should read about it!