(possibly incomplete list of)

Topics Covered in Chapters 1 to 6

Chapter 2 – Simple Linear Regression

The regression model equation (as given on page 14)
The four assumptions for linear regression (as given on page 14 or in class)
Independent vs. dependent variable
Regression and Causality (e.g. storks bringing babies example)
Regression and Extrapolation
The estimates of β_0 and β_1 are gotten by minimizing the sum of the squared residuals
The analysis of variance table
    SS_{tot} is the error when we assume H_0 is true (the error in Y without taking X into account)
    SS_{res} (=SSE) is the error when we use the line we got from the data
    $SS_{reg} = SS_{tot} - SS_{res}$ is the amount of error we explained by using the regression line
    the MS are like variances
    MS_{tot} is the estimated error variance (s_y^2) when $\beta_1 = 0$
    MS_{res} (=MSE) is the estimated error variance ($s_{y|x}^2$) when the regression line is used
    because we assumed the residuals are normal we can use the MS to do an F test
Finding the mean squares 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
$F = MS_{reg}/ MS_{res}$
When we accept or reject $\beta_1 = 0$ based on the ANOVA table
How we find the SS for the ANOVA table
How the MS relate to the variances of the errors
The t-test and confidence interval for β_1 (and how to find s_{b1} from the SAS output)
The predicted value of y given x
The confidence interval for the line of means (for the regression line) (page 37-38)
The confidence interval for a new observation (for predicting an individual) (page 38-39)
What the different parts of the variances s_{Yhat} and s_{Ynew} 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: larger r^2 implies larger F and smaller $\sqrt{MS_{res}}$,
    r^2 = coefficient of determination = percent of variation in the dependent variable that is explained by the regression using the independent variable, and r is the slope of the regression line after adjusting for the scale of y and x.
How to use the residual vs. predicted plot to check that the linear form is mean of the
    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 will still generally perform well if there are small violations of
    the assumptions.

NOT: Comparing two slopes (pg. 27-28)

Chapter 3 - Multiple Regression

Uses k independent variables instead of 1, so that there are k+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)

NOT: Dummy Variables, Matrix Notation, Polynomial Regression, or Interactions (pg. 73-106)

Chapter 4 – Outliers and Transformations

Using the logarithm of the dependent variable to stabilize variance
Taking logarithm of x and logarithm of y changes the linear regression into a multiplicative one
The common transformations to fix the case of nonlinearity (the diagram put up in class)
What you can do if you have an outlier
Potential = Leverage = Hat Diagonal - h_{ii} measures the potential of the observation to change the regression based only on the *x* values. 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(k+1)/n (H)
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 <u>similarly</u> to a standard normal distribution, so that values larger than 2 should be uncommon, and larger than 3 should be rare. (RT)
Influence – Cook's D combines the leverage and (internally) studentized residual and gives a measure of how muchremoving the observation would change the model. Greater than 1 are possibly worth some additional attention, and > 4 seriously affect the model. (D)
Why a group of 2 or more points might be outliers, but not show up as having leverage or influence

Chapter 5 – Multicolinearity

Multicollinearity is when several of the predictor variables are highly correlated. This can lead to: 1) parameters 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 4 should be of concern, and a VIF of 10 or more indicates that there is very severe multicollinearity.
Why getting more data can help reduce multicolinearity (its in the readings!)

NOT: Centering (203-210), Principal Components (219-237)

Chapter 6 – Model Selection

Why we need to have an adjusted R^2
Using Mallow's C_p to select a good model. To avoid bias, want C_p to be near k+1 for k = # terms left in.
That a good regression model depends on more than having good fit statistics. It has to be simple enough to understand, and ii has to make sense to the users.

NOT: Press (251-252)