Day By Day Notes for MATH 385/585

Fall 2011

Day 1

Activity: Go over syllabus. Take roll. Guess some lines.

Goals: Review course objectives: To model linear relationships, interpret the model estimates, and explore the various uses of regression. Introduce Least Squares. Introduce Excel.

Most if not all of you are taking this course as a mathematics elective. The topic of regression is quite useful in fitting equations to data. Such equations can then be used with some degree of predictability; researchers will be able to accurately describe and predict future observations. Regression techniques are a basic part of many software packages, including MINITAB and Excel, both of which we will explore.

I believe to be successful in this course, you must actually **read** the text (and these notes) carefully, and work problems. The most important thing is to **engage** yourself in the material. However, our class activities will sometimes be unrelated to the homework you practice and/or turn in for the homework portion of your grade; instead they will be for **understanding** of the underlying principles. For example, we will do simulations of the regression model. This is something you would never do in practice, but which I think will demonstrate several lessons for us. In these notes, I will try to point out to you when we're doing something to gain **understanding**, and when we're doing something to gain **skills**.

I believe you get out of something what you put into it. Very rarely will someone fail a class by attending every day, doing all the assignments, and working many practice problems; typically people fail by not applying themselves enough - either through missing classes, or by not allocating enough time for the material. Obviously I cannot tell you how much time to spend each week on this class; you must all find the right balance for you and your life's priorities. One last piece of advice: don't procrastinate. I believe statistics is learned best by daily exposure. Cramming for exams may get you a passing grade, but you are only cheating yourself out of understanding and learning.

In these notes, I will put the daily **task** in gray background.

Today I would like to explore the mathematical idea of Least Squares. With this technique, an equation is "fitted" to data in such a way that the squared errors between the data and the fits is as small as possible. Some history:

> In 1795, Carl Friedrich Gauss, at the age of 18, is credited with developing the fundamentals of the basis for least-squares analysis. However, as with many of his discoveries, he did not publish them. The strength of his method was demonstrated in 1801, when it was used to predict the future location of the newly discovered asteroid Ceres.
>
> On January 1st, 1801, the Italian astronomer Giuseppe Piazzi had discovered the asteroid Ceres and had been able to track its path for 40 days before it was lost in the glare of the

sun. Based on this data, it was desired to determine the location of Ceres after it emerged from behind the sun without solving the complicated Kepler's nonlinear equations of planetary motion. The only predictions that successfully allowed the German astronomer Franz Xaver von Zach to relocate Ceres were those performed by the 24-year-old Gauss using least-squares analysis. However, Gauss did not publish the method until 1809, when it appeared in volume two of his work on celestial mechanics, *Theoria Motus Corporum Coelestium in sectionibus conicis solem ambientium*.

The idea of least-squares analysis was independently formulated by the Frenchman Adrien-Marie Legendre in 1805 and the American Robert Adrain in 1808.

(Taken from http://en.wikipedia.org/wiki/Least_squares.)

I want each of us to guess a good fit to some data I will supply, and we will then use the computer to assess which of us made good guesses. I am going to begin using Excel, but later we will use MINITAB quite a bit. However, I like to begin with Excel because we can see dynamically what effect our changes have.

In the spreadsheet, I will use each of your guesses to calculate a fit for each point, and from that we will find the error. The sum of the squared errors will be our measure of goodness; small sums mean close (and therefore good) fits.

Skills: (In these notes, each day I will identify **skills** I believe you should have after working the day's activity, reading the appropriate sections of the text, and practicing exercises in the text.

- **Understand the definition of Least Squares. Least Squares** is a mathematical concept of goodness concerning data and an equation describing the data. Each data value has a "fit" from the model, and the "best fitting equation" is the one that makes the total sum of the squared deviations from the fitted model as small as possible.

- **Know how to input the formulas in a spreadsheet (or by hand) to assess the goodness of a model.** We usually put our data into columns when we use a spreadsheet. Additional columns needed to calculate Least Squares are "fit", "error", and "squared error". The sum of the squared error column is the measure for how well a model is fitting.

- **Realize that the idea of Least Squares is not tied to the notion of Linear Models.** The model that we fit can be any calculable equation. It is **common** to use models that are linear in the parameters, but is not necessary.

Reading: (The reading mentioned in these notes refers to what reading you should do for the **next** day's material.)

Sections 1.1 to 1.5.

Day 2

Activity: Simulate the basic regression model.

The model we will begin using is the **basic regression model**, also called **simple linear regression**. It has one **independent**, or **predictor** variable, one **dependent**, or **response**

variable, several parameters, and a random error term. Notice the model is **linear** in the parameters because they do not appear multiplied together or with any exponents. This idea becomes very important in Chapter 5 when we use matrices. The error term helps explain unexplained variation, sometimes called "white noise". These errors may be due to other unmeasured variables, or perhaps to just randomness that we cannot explain. One of our tasks in upcoming sessions is to try to determine if the data we've collected matches the model we've selected. Then we will be paying very close attention to **all** of our assumptions in this basic model.

The alternate model on page 12 is a **centered** model. We will want to use this model after we learn about multicollinearity (Day 23). The important feature of this model is that it yields the same fitted values, and is thus an **equivalent model**. It is important for you to be able to show the equivalence of the two models algebraically, i.e. prove it is true.

I think the best way to understand the model presented on page 9 (equation 1.1) is to use a computer to simulate responses from it. We can see the **true** line on a graph, and how data are scattered around the line.

In my simulation I will assume the errors have a normal distribution, but that is not required for estimation purposes. When we apply **statistical inference**, however, we **will** require that assumption. In my spreadsheet, we will be able to control which error distribution we use. One of our goals is to see if different distributions create different views.

Goals: Understand the notation and ideas of the basic linear model.

Skills:

- **Understand each term in the basic regression model.** Understanding regression begins with understanding the model we are posing. You need to know what **parameters** are, and how they differ from **random error**. You should be able to recite the model we use from memory.
- **Know how to simulate the basic regression model.** From our class demonstration, you should be able to produce a simulation yourself, using a spreadsheet or other computer program, such as MINITAB. While I haven't gone through the MINITAB commands, if you would like to use that program to do simulations, I can help you outside of class.
- **Understand the alternative "centered" model.** In some cases we want to use transformed data instead of raw data. The resulting model produces identical fitted values and is thus an equivalent model. However, the parameters we use are different. In a sense, parameters are merely a convenience for us to describe a model, and are **not** unique.

Reading: Sections 1.6 to 1.8 (first part).

Day 3

Activity: Estimation of Parameters.

Today we will use Least Squares, and some calculus, to derive the estimates for the simple linear regression model. We will encounter *Non-Linear* regression later (Chapter 13, Day 37),

but today I will introduce it with a separate model, and we will see how the calculus approach takes us only so far.

To minimize the sums of the squared errors, we will treat the data as fixed, and the parameters as variables. Then, as you know from Calculus I, we find where the derivatives are zero to locate the extrema, in this case the minimum(s). For Simple Linear Regression, it turns out we can do these results with simple algebra. Later on, with more variables (Chapter 5), we will have to use linear algebra instead.

The calculus we do today will involve partial derivatives; for those who haven't had Calculus III yet, fortunately these derivatives are not any trickier than regular Calculus I derivatives. The key is to think of the **other** variable not being looked at as a fixed constant. Once we have found the derivatives, we set them to zero, **simultaneously**, and solve. In general, this step is quite difficult. For the case of Simple Linear Regression, it turns out to be quite straightforward. The key is that the derivative of squared functions are linear functions (the power rule).

Today is the first day we formally see the **residuals**, the deviations from the fitted values. Residual analysis is quite important as it is our chief tool for assessing model adequacy. We will come back to them later (Chapter 3). For now, what you need to know about them is their definition and some basic algebraic facts about them, in particular that they sum to zero.

Our last result today involves the error variance, σ^2. The reasoning behind our estimate of the variance, which is called the **mean square error** or MSE, is beyond our abilities; you would use techniques from Math 401. The **idea** is understandable, but requires that you understand the difference between the residuals e_i and the error terms ε_i. The residuals are calculated from data values; the error terms are unobservable terms in the model, the result of a random selection from a distribution. The key result is that if the model is correct, then they have the same normal distribution, so that the variance of the residuals ought to be the same as the variance of the error terms. There is also one more additional complication: degrees of freedom. Again, using Math 401 results, we find estimates for variances by dividing **sums of squares** by **degrees of freedom**. The ANOVA results from Day 6 will shed additional light on this situation.

Goals: Know how the estimates of the basic model are found.

Skills:

- **Know how calculus is used to derive the Least Squares estimates.** Because Least Squares is an optimization, we can use basic calculus results to derive the answers. The key idea that makes the solution feasible is that the derivative of squaring is linear, and we know much about solving systems of linear equations. Once we have the partial derivatives for all parameters in the model, we simultaneously set them equal to zero and solve.

- **Know the formulas for the model estimates for slope and intercept.** While I'm not a fan of memorizing results, it **will** be helpful to know at least the form of the least squares estimates.

- **Know the definition and simple results for residuals.** The residuals are the deviations of the data from the model. We can also think of them as the "error" in the fit. You should know the

formula for them as well as simple facts about them, such as their sum is zero and their sum of squares gives SSE.

- **Know the best estimate for the model variance.** Because the residuals behave similarly to the model error terms, their **mean square**, or sample variance, estimates the model variance. The key idea will come up again: variance is estimated using a ratio of sums of squares and degrees of freedom. See Day 6.

- **Know what the Estimation of Mean Response is and how to calculate it.** Often we want to estimate a particular point on the regression line. This is a **mean response**, and should not be confused with a **prediction** of a new observation. Basically, to estimate a mean response for a particular x-value, substitute that new x-value in the fitted equation.

Reading: Sections 2.1 to 2.3.

Day 4

Activity: Inference on slope and intercept using a simulation.

I believe the best way to see how the distribution results work is to conduct a simulation. We will do problem 2.66 on page 98 to demonstrate how sampling distributions work. The idea is to generate many, many samples of results and then calculate the estimates for each sample. If we look at appropriate graphs (like a histogram) we can compare the theoretical results with the simulated results.

In addition to the simulation, we will also look at the theoretical results. The key on page 42 is that the equations can be rewritten as linear combinations of the y-values. I don't expect you to memorize the details, but you should know the results: the least squares estimates have normal distributions. Therefore, to derive the means and variances we need for the confidence intervals and tests, we use the linear combination formulas from Math 301.

Fortunately, in MINITAB, the calculations are mostly done for us; it is just a matter of interpreting the outputs, which we will spend time doing in class. For hypothesis tests, you must know which null hypothesis is being tested, and how to interpret the P-value. The confidence interval needs to be calculated manually (from the basic output). Of course, it is up to you to learn how to use the software yourself.

Goals: Know the basic confidence interval and hypothesis test results for the slope and intercept.

Skills:

- **Know that least squares estimates are linear combinations of the *y*-values.** Once we write the least squares estimates as linear combinations of the y-values, the linear combination formulas can be used to calculate the mean and standard deviation of the estimates. You do not need to memorize the particular weights in the formulas, but you should be able to follow the algebra on page 42.

- **Know that least squares estimates are tested with the *t*-test.** We know the estimates are linear combinations. We also know that MSE is an estimate of σ^2. Using this information,

and results from Math 301 and Math 401, we can test the estimates using the *t*-distribution results.

- **Understand how simulation can be used to observe sampling distributions.** Using the spreadsheet in class, you should understand what we mean by the sampling distributions of the least squares estimates. In particular, you need to have a solid understanding of the mean and standard deviation. If our model is correct, we can predict how variable the fitted line can be, and from that information we can assess the fitness of our model.

Reading: Sections 2.4 to 2.6.

Day 5

Activity: Interval Estimates.

In addition to point estimates, in statistics we typically use *interval* estimates. Using our simulation from Day 4, we can investigate how variable the interval estimates are, and we can observe how the confidence coefficient is interpreted. The chief idea is that the interval estimate contains a notion of the sampling variability along with it. The confidence coefficient represents the chance that the interval contains the true parameter, in this case the value on the regression line. Our simulation should show us this.

It is important that you recognize the two types of intervals we are generally interested in: estimation of a mean response, and prediction of a new observation. With the mean response we are estimating where the true regression line falls. With the prediction interval, we are recognizing two sources of variation: the sampling variation, and the randomness inherent in the model, encompassed in the error terms. The formula (on page 59) shows us these two sources of variation, and is thus much wider than the interval for the estimation of a mean response. I will refer back to these simulations often throughout the rest of the course.

We can construct a confidence band by applying our mean response results for all possible *x*-values. This yields hyperbolas (see page 62). The key difference between this band and an individual confidence interval is that the band's confidence coefficient is a **family** confidence coefficient. The proper interpretation is the coefficient is the chance that the true regression line lies entirely within the band.

Goals: Use the Math 301 results on confidence intervals to estimate the parameters with intervals.

Skills:

- **Know how to use the confidence interval results from Math 301.** In Math 301 we used the normal curve to estimate a parameter, giving a range of values between which we believed the parameter was. The interval was composed of a lower value and an upper value, and a confidence coefficient, which represented the chance that the random interval contained the parameter.

- **Know how a prediction interval differs from a confidence interval for a mean response.** To estimate a new response, we must not only account for the variation inherent in the model (the ε error terms) but also the uncertainty in our estimates themselves. Thus the prediction interval is wider, having two sources of variation.

- **Know how to construct and interpret a confidence band for the regression line.** A typical confidence interval estimates a single parameter or combination of parameters. A regression band around the line estimates the region where the line may completely fall. In general, due to this global sort of coverage, it must be a larger (wider) interval.

Reading: Section 2.7.

Day 6

Activity: ANOVA. **Homework 1 due today.**

The ANOVA table is a convenient way to summarize the information we have about the least squares estimates. We will examine the details today. Some of what we will do is algebraic. The key result is that the sums of squares we are interested in are **additive**. The sums of squares decompose into two orthogonal components, which means their sums of squares are additive (the cross product term sums to zero). The degrees of freedom associated with these sums of squares are also additive. We will be unable to prove this fact, as it relates to advanced linear algebra. It happens that degrees of freedom can be conveniently associated with parameters estimated, although that appears to be a mysterious explanation.

The last column in the ANOVA table is the Mean Square column. You are already familiar with this idea from calculating the sample variance, where you took a sum of squared deviations and divided by one less than the sample size. That calculation was really a "sum of squares" divided by a "degrees of freedom". In regression, these Mean Squares, under suitable conditions and hypotheses, have a chi-squared distribution. The ratio of two independent chi-squares divided by their degrees of freedom has an F distribution.

Sometimes an additional column is included in the table, representing the Expected Mean Square. Developing these formulas requires much more mathematical statistics than we have so far, so we will accept these formulas from the text on faith. I have found the greatest use for these EMS's is in choosing the proper test statistic in experimental designs, one of the topics of Math 386, taught in the spring.

With only one independent variable, the ANOVA table shows us nothing we didn't already have with our standard *t*-tests and intervals. However, when we proceed to more than one independent variable, the ANOVA approach is vital as the *t*-procedures will be inadequate.

The main test we perform is whether there is a relationship present or not. This is most easily phrased by equating the slope to zero. Again, we already have a *t*-test for that, but there is a corresponding *F*-test too. One difference between the two is that the *t*-test is a little more flexible in that we can test one-sided alternate hypotheses, whereas with the *F*-test we **must** use the two-sided alternate.

Goals: Understand the details of the ANOVA table, including the *F*-test.

Skills:

- **Know the layout and the relationships of the ANOVA table.** The sums of squares and degrees of freedom in an ANOVA table are additive. We can show the sums add using algebra; the degrees of freedom require advanced linear algebra. However, we can relate the

degrees of freedom to estimation of parameters. The ratio of the two is a Mean Square, and ratios of Mean Squares form the basis of our F-tests.

- **Understand the components broken down by the Sums of Squares decomposition.** The sketch on page 64 helps remind us what quantities are involved in the sums of squares. The key notion is that we have different estimates involved, and the deviations from these estimates are the components in the sums of squares. Caution: not all sums of squares will add in this way; we must also check the orthogonality. Fortunately, for the regression results we encounter, the sums of squares always decompose.

- **Know the *F*-test for testing the slope is zero.** While we already have a test for the slope being zero, the t-test, we will not be able to get by with t-tests when we introduce more than one independent variable. One important restriction is that the F-test can only test the two-sided alternate hypothesis.

Reading: Section 2.8.

Day 7

Activity: GLM and R^2.

One convenient way of testing regression models is to test individual parameters, as we have been doing. For example, our basic test is whether the slope is zero or not. Another approach is to fit different models, and compare the SSE for each model. This new approach is called the **General Linear Models** approach, or GLM. We will see more of this approach later, in Chapter 7, but it is appropriate for us to see it now too.

The GLM test is another F-test, so we need a ratio of mean squares. The difference in this test is that one of the mean squares is found by subtraction. We follow the steps on the bottom of page 73. I will demonstrate using various null hypotheses, including $\beta_1 = 0$, $\beta_1 = 2$, and $\beta_1 = \beta_0$.

We have one last detail before we continue on to diagnostics and model checking. One common measure of the goodness of a model is the value of R^2. This value is simply the percentage of the total variation accounted for by the regression line. Note the misconceptions on page 75. It is important that we don't misuse this measure. It says what it says, nothing more. Its best use is in comparing different models.

Goals: Introduce the General Linear Models approach to testing hypotheses. Explore the goodness of fit measure R^2.

Skills:

- **Know the strategy behind the General Linear Models approach.** For simple hypotheses, like the slope is equal to a constant, we can use the GLM approach for testing. The key is to fit two models, and compare the SSE's appropriately.

- **Know the details of using the GLM approach.** After fitting the two models, we compare the MSE's in a new F-test. An important detail to worry about is that some models have to have the y-values transformed according to the null hypothesis. See the class notes.

- **Know the calculation and interpretation of R^2.** R^2 is a measure of the goodness of a model. It is the fraction of explained variation, as compared to the overall variation in the *y*-values.

Reading: Sections 3.1 to 3.6.

Day 8

Activity: Residuals I.

So far we have only discussed fitting a model. However, we must also check to make sure we have a reasonable model, and that all the assumptions of our model seem reasonable. Our chief tool for making these assessments is **residual analysis**. The behavior of the residuals, the deviations from the model fit, tells us a lot about the effectiveness of the model. We can use them to test for normality, for goodness of fit, for influential outliers, and other departures from the model.

We first will examine the formula for residuals and see what we can deduce about their behavior. For example, are they a linear combination of the *y*-values, as the slope and intercept were? We will take some time today to try our hand at algebraic manipulation to see if we can answer that question.

Next, we will check for goodness of fit by looking at plots of residuals versus independent variables, both those included in the model and those not yet included in the model. If our residual plots show any patterns, we have evidence of "lack of fit", or in the case of variables not yet included in the model, evidence of missing variables. What we're looking for is a random scatter of points. If we see patterns, such as increasing spread, or organized clustering, we suspect we have a not-so-perfect model. Sometimes we can correct the defect with remedial measures, which we will pursue on Days 10 and 11. Be cautious with your interpretations of these plots. It is tempting to say that something is a pattern when it really is just the result of randomness. Of course, this is somewhat of a judgment call. The more you study regression and use it in real world data, the better you will be at the **art** of model fitting.

Goals: Introduce residuals as a diagnostic tool.

Skills:

- **Know the definition of residuals.** The departures of the data from our model are the residuals. Each data value produces one residual, and they are measured in the same units as the *y*-values. We have encountered residuals before; they are the items being squared and summed in the least squares exercise.

- **Know how residuals are (roughly) distributed.** Because the residuals can be written as linear combinations of the *y*-values, we know they have normal distributions. Unfortunately, they aren't distributed with the same variance; their variance depends on their distance from *x*-bar. However, we can use MSE as an approximate variance.

- **Know about basic residual plots.** Our chief diagnostic tool will be residual plots. If we plot the residuals against the independent variable, we can see if we have lack of fit, or non-constant variance, or extreme outliers. We can also plot residuals against variables not already in the model to see if those variables would help explain variation.

Reading: Sections 3.1 to 3.6.

Day 9

Activity: Residuals II.

Continuing our exploration of residuals:

We check for normality of errors by looking at probability plots, histograms, etc and comparing them to the corresponding normal curve plots. If we use normal probability plots, we can use the Looney table, Table B.6. There are other tests for normality that we can discuss, but the Looney table is the easiest to use. The details of constructing a normal probability plot are on page 111. Essentially we are comparing the actual data with where data of that rank (3^{rd} smallest, 4^{th} smallest, etc) would fall if the data were truly normally distributed. If the data is normally distributed, this plot will be linear. To use the Looney table, we calculate a correlation coefficient for the normal probability plot. If it is large, the data looks normal.

Another simple test we can perform is the Brown-Forsythe Test. We assume in our model that the variance of the error terms is constant, i.e., not dependent on the value of x. An easy way to check this is to compare the spread of the residuals for the residuals associated with small x-values to the spread of the residuals for the residuals associated with large x-values. The details of the test use the 2-sample t-test: we first calculate the absolute size of the residuals (about their median) in each half and then perform a pooled 2-sample t-test on these absolute residual deviations.

Goals: Know how to create and interpret a probability plot. Understand the Brown-Forsythe test.

Skills:

- **Know the steps needed to create a normal probability plot.** We have several options to creating a normal probability plot. Of course, we could use software, such as on the TI-83 or in MINITAB. But you will not be able to use the tests from the TI-83. Therefore, you should know how to create one yourself. Basically you are going to translate each rank using an inverse normal calculation, (3.6) on page 111. Then plot these inverses versus the data.

- **Know how to use a normal probability plot to detect normality.** If the data is normally distributed, the normal probability plot should be a straight line. The Looney and Gulledge Table (B.6) on page 1329 gives us the critical values of the correlation coefficient for assessing whether the observed line is close enough to straight. If the observed correlation is high enough, we conclude that normality is plausible.

- **Know the details of the Brown-Forsythe Test.** The Brown-Forsythe test helps us determine if the variance is constant. We split the data set into two parts, based on the independent variable, and find the size of the residuals in each half. The test statistic is a modified two-sample t-test, based on the absolute size of the residuals around their respective medians.

Reading: Section 3.7.

Day 10

Activity: Lack of Fit.

When we have at least one x-value that has more than one observation, we can calculate a standard deviation for that "internal error". Using partitioning similar to that on Day 6, we can formulate a test to check for one kind of "lack of fit". We are able to estimate a "true" variance for the model, by pooling together all the variance estimates from all the unique x-values. If we compare this value to the MSE from the linear model, we have the basis for a test.

The lack-of-fit test is a GLM test. The Full model has a mean for each unique x-value. The Reduced model is the standard linear model we've been using. If we have c x-values with repeated observations, then our "pure error" sums of squares has $n - c$ degrees of freedom. (We lose one degree of freedom for each unique mean we have to estimate to calculate the pure error sum of squares.) The standard linear model has $n - 2$ degrees of freedom. The details of the test are on page 123.

A few comments on this technique: we need only have **one** x-value with repeats. The idea is that we get an estimate of the variance that is independent of the linear model. If we have no replicates, we can use **near** replicates. These require judgment; if we choose cases too far from each other, our estimate of the variance may be too large. If we choose too few cases, the degrees of freedom may be too small to be useful.

Goals: See how to use internal variances to check for lack of fit.

Skills:

- **Know how to set up the Lack of Fit testing procedure.** To setup the ANOVA table for the Lack of Fit test, we must calculate the variances of each unique x-value. (Note that the variance is zero for x-values with no replicates.) We pool all the variance estimates together, weighting by the degrees of freedom. The difference between the linear model SSE and this new **pure** error sum of squares (the numerator of the pooled variance) is the numerator for our F-test. The pooled variance, MSPE, is the denominator.

- **Know the details of when the Lack of Fit test can be used.** When the null hypothesis is true, the test statistic (3.25) on page 124 has an F distribution. So, small values of F support the null hypothesis that the linear model is an appropriate model. The alternate is that some **other** model is appropriate. Notice we are not specifying what the other model is if we reject the null. There are $n - c$ degrees of freedom for MSPE, so we need at least one repeated x-value. But note that not **every** x-value need be repeated.

- **Near replicates require judgment.** In some data sets, and with the judgment of the user, we can use clusters of points as "near replicates", pretending they are repeated x-values for purposes of calculating MSPE. Of course the further apart the real x-values are, the less likely that particular estimate of the variance is to be correct. One must use good judgment as to what constitutes "close enough".

Reading: Sections 3.8 to 3.9.

Day 11

Activity: Transformations.

After we uncover departures from the model, we often use remedial measures to correct the model. The most common of these is a transformation of the response variable, but sometimes transforming the x-values can be effective.

We can use several prototype plots to help with our choice. A few examples are on pages 130 and 132. But the extent of the curvature may be such that even these transformations are insufficient. However they are often a good first attempt at remediating the lack of fit.

Another procedure to select a transformation is the Box-Cox procedure. What we seek is a transformation of the Y variable that corrects the non-constant variance as well as the non-normal errors, if needed. The transformation is given on page 135, as well as the formulas for finding the optimal power (3.36). We will use Excel to implement these, but one could also have MINITAB perform the operations. The key is that we fit many models and choose the one that makes SSE small.

Goals: Investigate using transformations to improve the model fit.

Skills:

- **Be familiar with the prototype plots for making transformations.** The prototypes on page 130 show us suggestions for transforming the x-variables to correct certain types of monotonic lack of fit. The ones on page 132 give us an idea of transformations for the y-variable to correct some types of non-constant variance. However, we may not find a suitable transformation just using these diagrams.

- **Know the Box-Cox transformation procedure.** To find a reasonable transformation of the response variable, the Box-Cox procedure finds a suitable power for the transformation. Because of differences in scale, and after a suitable transformation, we can compare SSE's for a variety of powers. The one that minimizes SSE is the most reasonable transformation, which will often correct non-constant variance and lack of fit problems as well.

Reading: Sections 4.1 to 4.3.

Day 12

Activity: Simultaneous Inference. **Homework 2 due today.**

Due to the nature of the relationship between the least squares estimates, it is inappropriate to make inferences about them separately. When one increases, the other is likely to decrease. Simply making two interval estimates would be too conservative, yielding a higher confidence coefficient than is appropriate. Today we will explore a more efficient method, making use of the correlation between the two estimates.

First, let's look at the two estimates jointly. Equation (4.5) on page 157 shows us the relationship between the two estimates. We can see this correlation from our simulation from Day 4. Notice that if the x-bar is zero, the two estimates are uncorrelated, and because they

have normal distributions, this makes them independent. Our text does not have a method to construct exact confidence bounds when the two estimates are correlated, so I will show you a technique from an earlier edition of the text. Because the distributions are bivariate normal, we construct ellipses around the estimate, lying on a tilt, wide or narrow according to the correlation.

Here are the details of the method. We need to have the two estimates, b_0 and b_1, the sample mean, the sample sums of squares, MSE, and F. We then assemble using the following formula: $b_1 + \frac{n\bar{x}}{\Sigma x^2}(b_0 - \beta_0) \pm \sqrt{\frac{2(MSE)F}{\Sigma x^2} - \frac{n(\Sigma x^2 - n\bar{x}^2)(b_0 - \beta_0)^2}{(\Sigma x^2)^2}}$. Note that we are plotting this equation in the β_0/β_1 axis system. The interpretation of this region is analogous to the interpretation of a confidence interval: there is a 95% chance that the constructed region captures the true parameters. We will compare this approach to the Bonferroni approach next.

The exact joint confidence interval above is complicated at best, and is much worse for the multiple regression coming up later. A much simpler approach is to use the Bonferroni inequality (4.2) on page 155, which basically allows us to apportion the error probability into parts. For example, if we make 2 inference statements and want a 95% chance that **both** statements are true simultaneously, we could use 97.5% confidence levels on each statement. (We have divided up the 5% error into two equal parts in this case.) The joint confidence is a lower bound, so we have **at least** a 95% chance that both statements are correct. The **real** advantage of the Bonferroni method is that it extends easily to multiple regression.

Goals: Explore the important concept of simultaneous inference.

Skills:

- **Recognize the issue of making multiple inference statements.** When we make several confidence statements, the chances that they are **all** correct at once gets exponentially smaller as more and more statements are made. We have several approaches to dealing with this. One is to use the actual joint distributions of the estimates, but this approach is often quite complicated. Another approach is to use probability statements to produce conservative **families** of confidence intervals, such as with the Bonferroni method.

- **Know the Bonferroni method.** The Bonferroni inequality shows us that the probability in a confidence family can be apportioned equally among the individual intervals. The family confidence coefficient is then **at least** as high as desired. The main usefulness of this technique is that the statements being made can be of any sort, from any analyses, as long as we divide up the error probability appropriately.

- **Know of the true joint inference statements.** Because the least squares estimates follow the normal distribution, and we can calculate their correlation, we can form ellipses around the estimates that capture the desired amount of the probability distribution. Also, note that if the mean of the independent variable is zero, the estimates are uncorrelated, and the resulting "ellipse" is really a circle.

Reading: Sections 5.1 to 5.7.

Day 13

Activity: Review.

Goals: Know everything about simple linear regression!

Reading: Chapters 1 to 4.

Day 14

Activity: Exam 1. This first exam will cover simple linear regression, including estimates, inference, diagnostics, and remedial measures. Some of the questions may be multiple choice. Others may require you to show your worked out solution. Don't forget to review your class notes and these notes.

Day 15

Activity: Introduction to Matrices.

The simple linear regression model can be written with matrices. This forms the basis for using matrix algebra to perform statistical regression. We will take a few days to go through the algebra involved, building up to being able to write all of our results so far using the much simpler matrix algebra equations.

The most important operation of matrix algebra that we will use is matrix multiplication. You are already familiar with this operation, as it is part of the way your GPA is calculated, a sum of products. ("This times that" plus "this times that" plus etc.) Basically, any sum of products of two items (sometimes using a 1 as one of the multiplicands) can be written as a matrix product. If we have several such products, as is the case in regression, we simple add more rows to one of the matrices involved. We will setup the basic equations from regression to demonstrate the matrix multiplication.

Another useful application of matrix multiplication is the product of a vector and its transpose. This creates a 1 by 1 matrix of a sum of squares, which as you can imagine, is important to us in creating the ANOVA table, and other statistics we've been using. We will also look at the important class of quadratic forms, which also produce scalars that are sums of squares.

I will outline the idea behind the least squares estimates, without using calculus, but the result can also be derived that way. The chief operation involved is matrix inversion, which we will look at briefly today. However, naturally in practice we will use computer software for the actual calculations. It is still important for our analyses to know exactly how matrix inversion is performed. We will go through a few examples to illustrate. I especially want to do an example where the determinant is close to zero.

Goals: Begin the transition from regular algebra to matrix algebra.

Skills:

- **Understand matrix multiplication.** A sum of binary products can be written as the matrix product of two vectors, one a row vector and one a column vector. With many such

combinations, we have what is called **matrix multiplication**. Only two **conformable** matrices can be multiplied together, as the binary products **must** have the same number of elements or they cannot be paired together. Also, matrix multiplication is **not** commutative. Order is very important, and we must take great care with our algebraic operations such as transposing.

- **Be able to write the Linear Regression model using matrices.** Using matrices, one for the data and one for the parameters, we can write the basic regression equation. We also need a vector for the error term, but due to calculus results we often only concentrate only on the X matrix, the actual data collected.

- **Know the important class of matrices that produce a scalar that is a sum of squares.** When we multiply a row vector and its transpose, the corresponding column vector, we see that we are really squaring each term and then summing. Thus for example $Y'Y$ is the sum of the squared y-values. As you have seen, our ANOVA table uses sums of squares as the basis for our inferences, so this feature of matrices is quite important to us.

- **Know of matrix inverses and how they're used in regression.** As we will see from Day 16, the least squares results involve matrix inversion. You should know the definition of a matrix inverse, and how to calculate one, or at least how to verify that a given matrix is the inverse of another. Further, you should understand some of the difficulties in calculating an inverse, and you should also know that not all inverses exist. This situation is analogous to dividing by zero.

Reading: Sections 5.8 to 5.13.

Day 16

Activity: Regression Matrices.

We will continue with our matrices, and find out how much easier it is to write the regression results with matrix algebra. I also want to try a little matrix calculus results, to demonstrate the complexity of matrix algebra.

We will begin by looking at the residuals formula. Then we will find the sum of the squared residuals using our matrix multiplication trick, and using calculus results, derive a solution to the least squares problem. The solution is $\hat{B} = (X'X)^{-1}X'Y$. The most important feature of this equation is that the estimates are linear combinations of the observations. Revisiting the residuals formulas, we discover $e = Y - \hat{Y} = Y - X\mathrm{B} = (I - H)Y$, where $H = X(X'X)^{-1}X'$.

Now that we have expressed our primary formulas using matrices, we would like to be able to derive distributions. To do that, we first need to know the rules. From Math 301, we know these facts: $E(aX + b) = aE(X) + b$ and $Var(aX + b) = a^2Var(X)$. When we have more than one variable, though, the formulas get a bit more complicated. We have a new idea, called covariance, which is closely related to correlation. An important consequence of covariance and correlation is that if two variables are independent, their correlation (and also their covariance) is zero.

In class we will derive the expectation and variance formulas for matrices. The conclusion is quite similar to the Math 301 material: $E(AX + B) = AE(X) + B$ and

$Var(AX + B) = AVar(X)A'$. applying these results to our least squares estimates, and to our residual formula, gives us the necessary background for inference. We will explore this further on Day 18.

Goals: Know how matrices are used in regression formulas.

Skills:

- **Know the form of the Least Squares Formula.** Using matrix calculus, we can find the least squares estimates. What we are trying to do is the same as before: minimize the sums of the squared errors. Using matrix notation, we have $Q = (Y - XB)'(Y - XB)$, which is a 1x1 matrix, a scalar.

- **Be able to derive the least squares estimates.** Using Q, you should be able to use calculus and find the derivatives with respect to B, and then solve using matrix algebra. The result is $\hat{B} = (X'X)^{-1}X'Y$, and the important feature is that the least squares estimates are linear combinations of the independent variables.

- **Know the residuals formula.** Using the least squares formula, the residuals can be written as $e = Y - \hat{Y} = Y - X\mathrm{B} = (I - H)Y$. In particular, note that the residuals are a linear function of the observations. We make use of this fact when we determine distributional results.

- **Know the variance results of a random vector.** Know how to take a matrix equation and derive its mean and variance. In general, we have $E(AX + B) = AE(X) + B$ and $Var(AX + B) = AVar(X)A'$.

Reading: Sections 6.1 to 6.2.

Day 17

Activity: Multiple Regression Models.

Today we begin **real** regression, using more than one independent variable. We will discover more complicated models, due to the multiple dimensions. Our first look at multiple regression will be at the various different techniques used, including:

1) More than one independent variable.

2) Indicator variables.

3) Polynomial regression.

4) Inherently linear models.

5) Interaction variables.

Each of these situations will be examined in later chapters, but we can see their common features using our regression results.

Goals: Introduction to the various multiple regression models we will encounter in later chapters.

Skills:

- **Be able to write the regression models for the five situations listed above.** Knowing how each situation uses matrices is the key to writing multiple linear regression models. You must be able to perform matrix multiplication to achieve the proper models. See your class notes.

- **Understand the real definition of Linear.** We might naively believe that "linear" means "straight line". In fact, for our purposes in this course on "linear models" it means that the model can be written using matrices. In particular, polynomials qualify as linear models under this definition. A common explanation of this idea is to say the model is "linear in the parameters".

- **Know the full statement of the Multiple Linear Regression model.** Using matrices, you should be able to fully detail the multiple linear regression model, including error term assumptions. We typically assume normal errors, which lead to the usual *t*- and *F*-tests and other inferences.

Reading: Sections 6.3 to 6.6.

Day 18

Activity: Inference.

We will revisit the matrix results from Chapter 5, but with emphasis on the inference results. Specifically, we will look at the ANOVA table and the *t*-tests and *F*-test.

Goals: See how we use matrices with the inference results we need.

Skills:

- **Know the sums of squares formulas using matrices.** The three Sums of Squares formulas are presented on page 225. You should be able to write and manipulate them. Don't worry about the E(MSR) formula details; the important fact with E(MSR) is that under the null hypothesis, it should be the same as E(MSE).

- **Know the matrix results for inference.** Using our rules from Day 16, you should understand what the means and variances are for the least squares estimates, the hat matrix, and the residuals. Because we use the normal assumption for our errors, you should understand how to use the means and variances to create *t*-tests and intervals.

Reading: Section 6.7.

Day 19

Activity: Intervals.

We will revisit the interval formulas we saw earlier, including the prediction interval. Once again, the expectation and variance rules guide us. The details are on page 229.

Goals: Use the matrix results with the interval formulas.

Skills:

- **Know the matrix representation of the inference intervals.** Using our expectation and variance rules you should be able to understand the confidence intervals for mean response and for predicting a new value. For a confidence interval for the mean response, remember we are trying to guess the true mean value for that set of x-values whereas when we predict a new value we also incorporate the uncertainty in the individual observations, measured by σ.

Reading: Section 6.8.

Day 20

Activity: Diagnostics.

The chief diagnostic tool for multiple regression is once again the residual plot. However, we have many more options now as to what independent variable to plot the residuals against. In addition, we often want to see the relationships **between** the independent variables.

We can also modify the other techniques we saw earlier, such as the Brown-Forsythe test or the Lack of Fit test.

Goals: Begin the residual analysis for multiple dimensions.

Skills:

- **Know the uses of the scatter plot matrix.** The scatter plot matrix shows us the simple two-way scatter plots for all the variables in our database, but in one display. We can see at a glance which variables are most correlated with the response variable, and also with the other independent variables.

- **Know how to modify the earlier diagnostics for use in multiple regression.** We can modify our earlier techniques in multiple dimensions. These include the Brown-Forsythe test, the lack of fit test, and the Box-Cox procedure. The Brown-Forsythe test is used on each x-variable in turn. The lack of fit test can be used wherever we have multiple responses; the addition independent variables have no effect on the details of the test. The Box-Cox procedure is modified by using all the independent variables, instead of just one.

Reading: Section 7.1.

Day 21

Activity: Extra SS. **Homework 3 due today.**

Today we revisit the important class of tests for comparing hierarchical models. In these models, we have a "Full model" and a "Reduced model", as we saw in the lack of fit tests from Day 10. The Full model has the most parameters, and our assumption is that it is an adequately fitted model. In other words, the Full model is assumed to be "correct". The Reduced model has one or more of the independent variables from the Full model removed. Statistical theory shows that under normal errors models, the sums of squares for these two models are additive, but we require a different interpretation for the extra sum of squares.

The order the variables enter the model is critical. As variables are added, we add successive "extra sums of squares". The increase lets us know how useful the added variables are in explaining variation. If the extra sums of squares is a small value, this tells us the new variable doesn't substantially help explain variation. It is important to realize that the extra sum of squares is **not** a measure of the importance of the added variable alone. It is only the effect of that new variable **given the other variables are already part of the model**.

Goals: Explore how the order of the variables in the analysis gives different results.

Skills:

- **Recognize the problem of the non-additivity of single-variable sums of squares with two independent variables.** When we add a second variable to a regression model, we explain more variation. However, the additional variation explained is not typically equal to the variation explained by using just the new variable alone. This non-additivity occurs whenever the independent variables are correlated among themselves.

- **Know the definitions of the Extra Sums of Squares.** The notation we use for the extra sums of squares involves keeping track of what is in the model and what has just been added. It is important to be able to tell these two sets of variables apart.

- **Know where to find the Extra Sums of Squares on MINITAB.** In MINITAB, we look for sequential sums of squares. The important feature to look for is the fact that the extra sums of squares add up to the total sum of squares for the regression model.

Reading: Sections 7.2 to 7.3.

Day 22

Activity: GLM Tests.

Using our familiar GLM *F*-tests for sums of squares, we can test whether the most recently added variables have zero value. We can also test other sorts of models, such as parameters being equal, or function of other parameters.

In each use of the GLM tests, we must make sure that the full model contains all the reduced model variables. The extra sums of squares facts only work in the case of nested models. An equivalent interpretation is that the hierarchy holds when we set parameters equal to constants and then consider the reduced model. (Technically, it is an issue with column spaces of the matrices involved, but we will not get into those details.)

Today we will go over several types of tests that can be performed, including the interesting examples of setting parameters equal to constants or each other.

Goals: Revisit the GLM procedures, and see how they extend to multiple regression.

Skills:

- **Know the GLM procedure for testing in multiple regression models.** once a model has been established as a subset of another model, the GLM tests proceed in the same way as we

have seen already: a Reduced model MSE is compared to a Full model MSE, and the F-test is used.

Reading: Sections 7.5 to 7.6.

Day 23

Activity: Computational Problems and Multicollinearity.

Sometimes, it is computationally hard to avoid round off errors in multiple regression. This may be due to variables being measured on different scales, but most often it is due to high correlations among the independent variables. We can use the correlation transformation to combat this problem.

The main source of computational difficulty in multiple regression is multicollinearity. When the independent variables are correlated, we get the problem we saw on Day 21 with the Extra Sums of Squares not adding together. If the variables are **highly** correlated, it becomes harder and harder to estimate with any precision. One way to think about this is to imagine trying to balance a board on two points close together, compared to two points far apart. Multicollinearity makes two highly correlated variables appear as only one variable, and therefore it is like an underdetermined system of equations; with **perfect** correlation, there are an infinite number of solutions.

We have several ways of trying to deal with this, although we won't pursue them in detail. One simple way is to remove from the analysis any variables that are highly correlated with other variables. The reasoning is that if the two variables are highly correlated, then just knowing one of the pair is sufficient; being highly correlated means knowing one value is like knowing the other value also. In Chapter 11, there is a technique to deal with multicollinearity, called ridge regression, but we will not cover it. Of course, you may want to read about it anyway.

Goals: Explore situations where the $X'X$ matrix is nearly singular, and some cures for it.

Skills:

- **Know why we have computational difficulties in the $X'X$ matrix.** By examining some simple matrix inversions, we can see exactly when we have multicollinearity problems.
- **Understand the use of the correlation transformation.** By first standardizing our variables through the correlation transformation, we can sometimes substantially decrease the computational difficulties.
- **Understand the effects of multicollinearity.** When multicollinearity is present, we can often have a large estimate for the error variance, which results in unreliable estimates for the regression parameters.

Reading: Section 8.1.

Day 24

Activity: Polynomial Models.

As we have seen, straight line models are sometimes insufficient. It seems natural to try other fairly simple models, such as polynomials. Because polynomials are still linear in their parameters, our multiple regression techniques apply. There are complications, however. First of all, we may have polynomials in more than one variable. The order of the polynomial is the largest summed powers of independent variables, including interaction terms. For example, $z = x^2 + 3xy + 3x + 2y^2$ is a second order two-dimensional polynomial because the sum of the exponents of both x and y in any term does not exceed two.

We can use our GLM techniques for testing purposes, but with polynomials, it makes sense to include all lower order terms in a higher power polynomial. In addition, we find multicollinearity problems if we don't center the independent variables. When we are interested in inferences for the original variables, as opposed to the centered variables, we can use Chapter 5 matrix techniques to calculate the transformed variances. The purpose for centering is to reduce computational problems due to multicollinearity.

Today we will experimentally look at the multicollinearity problem and explore inferences using polynomial models.

Goals: Adapt polynomials to our linear algebra approach to regression.

Skills:

- **Create the model for polynomial regression models with any number of predictor variables.** To make a polynomial model with matrices, we simply add columns to the X matrix corresponding to the powers of the original independent variable(s). Adding these columns is done **after** centering.
- **Understand why we center the independent variables before fitting a model.** If a polynomial model is not centered around the average for each independent variable, typically large correlations result between the various powers of the independent variable. Often these correlations will be substantially reduced if each independent variable (and therefore the successive powers) are centered around their means before including them in the model. Because we are using a linear transformation of the parameters (see p. 299), all inference procedures can be modified with the appropriate matrix of coefficients.

Reading: Section 8.1.

Day 25

Activity: Interactions I.

We will look at two types of interaction models: linear models with interaction, and curvilinear models with interaction. There are many types of interaction models. The key idea is that the response surface is not parallel. One way of modeling a curved surface is to include a cross-product term. Today we will explore what adding a cross-product term does to the shape of a linear model.

When we have additive independent variables, the response surface is a plane. We can see this in 3-dimensions if we plot carefully, either with contours or with cross sections. With higher dimensions, we have to trust the intuitions we learn with two and three dimensions.

What I want to do today is sketch some response surfaces, and see the effects of changing parameter values. Later, when we fit models to data, I think it will help our interpretations if we understand what the relative sizes of the parameter estimates mean.

1) $E(Y) = 10 + 1X + 2Y$

2) $E(Y) = 10 + 1X + 2Y + 3XY$

3) $E(Y) = 10 + 1X + 2Y - 3XY$

For each of these functions, we'll produce cross sections, and contours, to see which plots help us see the surface adequately. The website "Wolfram Alpha" will help us greatly to produce and understand these surfaces.

Goals: Understand the effects of interaction in linear models.

Skills:

- **Use a contour plot or cross sections to explore a non-linear relationship.** Through the use of contour plots (holding the dependent variable constant and graphing the resulting set of solutions) or cross sections (holding one of the independent variables constant and graphing the resulting set of solutions), you should be able to visualize a three-dimensional surface.
- **Know how to include an interaction term in a linear regression.** To include an interaction term in a regression model, we use the product of two independent variables as a new variable. The resulting effect on the surface is a sort of "twisting" effect. Along any cross section, the surface is still linear. However, the slope of this surface is a function of the particular cross section used.

Reading: Section 8.2.

Day 26

Activity: Interactions II.

The other type of interaction term involves non-linear polynomial models. We will repeat the sort of activity from Day 25 but with polynomial models today.

1) $E(Y) = 10 + 1X - 1X^2 + 2Y + 2Y^2$

2) $E(Y) = 10 + 1X - 1X^2 + 2Y + 2Y^2 + 3XY$

3) $E(Y) = 10 + 1X - 1X^2 + 2Y + 2Y^2 - 3XY$

For each of these functions, we'll produce cross sections, and contours, to see which plots help us see the surface adequately. Once again, Wolfram Alpha will help us visualize our surfaces.

Goals: Relate interaction effects between linear and polynomial models.

Skills:

- **Know how the interaction term affects polynomial models.** In a polynomial model, the interaction is formed in the same way, but causes interesting effects to the resulting surface. In the case of a parabaloid surface, the interaction term rotates the ellipse. For other polynomial surfaces, a similar effect to the linear case occurs, a sort of "twisting", although in the higher dimensions this twisting is more difficult to summarize.

Reading: Sections 8.3 to 8.7.

Day 27

Activity: Dummy Variables I.

A very useful technique in regression is the use of indicator variables, or dummy variables. We will explore a variety of uses of indicator variables, including different intercepts, different slopes, jumps, piecewise continuity, multiple levels, and interference variables. The important feature of an indicator variable is that there are two possible levels, 0 and 1. A simple use of one is when we have two treatments or categories that our observations fall into.

Different intercepts: If we include just the dummy variable as another variable, we get different intercepts for our surfaces, but parallel slopes.

Different slopes: If we include a cross product of the dummy variable and another independent variable, we get different slopes, but the same intercept.

Jumps: We must work a little bit to get a jump of a particular height at a predetermined point. We will derive the proper function in class.

Goals: Introduce qualitative variables and some of their uses.

Skills:

- **Be able to use qualitative variables.** Indicator variables are variables with binary outcomes, the simplest of which are the values 0 and 1. You should be able to use them in our models to reflect parallel slopes, different slopes with the same intercept, discontinuities, and piecewise linear functions (which we will discuss on Day 28).

- **Know how to make a model with different intercepts.** By including an indicator variable in the linear model, we create two parallel lines, one for each of the two binary categories. The indicator variable coefficient describes the difference between the two intercepts.

- **Know how to make a model with different slopes.** By including an interaction between the independent variable and the indicator variable in the linear model, we create two lines, one for each of the two binary categories, with the same intercept but different slopes. The indicator variable coefficient describes the difference between the two slopes.

Reading: Sections 8.3 to 8.7.

Day 28

Activity: Dummy Variables II. **Homework 4 due today.**

Piecewise Continuous: To make sure the functions intersect at a predetermined point, we have to adjust the slopes **and** intercepts in just the right way. The key is to have the point where the two functions intercept fall on both the line before the point and the line after the point.

Multiple Levels: If we have several categories, we can create one less indicator variable than categories to accommodate our variable.

Interference variables: One way to deal with outliers without removing them completely from the analysis, which can be a burden, is to include an interference variable. Essentially this variable uses one degree of freedom, associated with the removed data point, and estimates a parameter perfectly matching the residual.

Goals: Continue qualitative variables.

Skills:

- **Know how to make a model with piecewise continuity.** We can derive this model by including both a dummy variable and an interaction with the dummy variable. However, adding both new variables does not require the continuity at the predetermined point. To maintain the continuity, we must describe the dependence between the two new parameters. Depending on our parameterization, the new coefficient will either be the change in the intercept (less interpretable) or the change in the slope (more interpretable).
- **Know how to make a model with multiple levels.** To account for a categorical variable with c different levels, we create $c - 1$ dummy variables. Note that if we created c dummy variables, one for each level of the categorical variable, they would add together to form a column of 1's. Thus we can only use $c - 1$ of them.
- **Know how to make a model with interference variables.** We can create a variable with is all zeroes except for the potential outlying case. This dummy variable is called an interference variable and will fit the potential outlying case exactly, without affecting how the model fits the rest of the cases. It will therefore be equivalent (in terms of estimates and inferences) to having removed the case from the dataset.

Reading: Sections 9.1 to 9.5.

Day 29

Activity: Review.

Goals: Know everything.

Reading: Chapters 5 to 8.

Day 30

Activity: Exam 2. This second exam will cover simple linear regression using matrices, including polynomial models, interaction models, and uses of indicator variables. Some of the questions may be multiple choice. Others may require you to show your worked out solution. Don't forget to review your class notes and these notes.

Day 31

Activity: Model Building.

Given a set of predictor variables, some fundamental questions are "Which variables are useful in predicting the y-values?" and "Which are unnecessary?" Another way of looking at it is which sets of variables make adequate models? We have two main techniques for model building: best subsets, and stepwise procedures.

We will first look at the stepwise procedures.

Forward. With this technique, we begin by examining all regressions using just one variable at a time. Then we pick the one variable with the best value of R^2. We now consider all models using the chosen variable and one other, again choosing the model with the highest R^2. Caution: we do not include a variable if its t-test (Extra Sum of Squares test) isn't significant.

Backward. With this technique, we start by including **all** the available independent variables in the model. Then we successively drop variables whose t-test is insignificant, until all remaining variables have significant values.

Forward Stepwise (Backward Stepwise): With these techniques, we begin just as in the Forward (or Backward) technique. However, at every step we check to make sure all the variables in the model are still significant (or no variable not included is significant). Those that are not significant may be dropped from the model (or those that are significant can be added to the model). The process then continues with this new set of variables as the starting point.

We will use MINITAB today to explore these stepwise procedures to build a model. There is no one technique that is best for all situations. Judgment on the part of the analyst is a necessary part of model building. My personal preference is a backwards stepwise technique, because including all variables to begin with is a stronger starting point than building up. The danger is that the forward technique could possibly end up excluding a strong variable, due to multicollinearity issues. The Best Subsets methods we will look at next should be used in conjunction with the automatic search techniques we explored today.

Goals: Introduce model building using stepwise procedures.

Skills:

- **Know how to use the forward selection technique, and its limitations.** Beginning with one variable models, the best single predictor is chosen. Using the best single predictor in each model, the best two variable model is selected. This process is continues until no more variables have significant t-values. Due to multicollinearity, important variables may not have

been added to the final model, and variables included in the model may have insignificant extra sum of squares *t*-values, when all other variables in the model are taken into effect.

- **Know how to use the backward elimination technique, and its limitations.** Starting with all predictor variables in the model, insignificantly contributing variables are removed, one by one. When no more variables are insignificant, the procedure stops. Again, due to multicollinearity, there is no assurance that the best set of variables is chosen. However, at least no variables in the final model have insignificant *t*-values, so this procedure is perhaps better than the forward procedure.

- **Know how to use the stepwise technique, and its limitations.** Using a stepwise approach, either forward or backward, each stage is followed by another look at the variables dropped or added so far. In the forward stepwise procedure, if a previously added variable is now marginally insignificant (small extra sum of squares) it is removed from the model. Similarly, in the backward technique, if a previously discarded variable is now significant, it is added back into the model. While stepwise procedures are better than forward or backward only procedures, there is still no guarantee that the best subset of variables will be chosen.

Reading: Sections 9.4 to 9.6.

Day 32

Activity: Best Subsets.

While stepwise procedures lead us to reasonable models, they won't always find all the good models. For that we need another search technique. The most popular is the Best Subsets routine. This method finds best models using a variety of goodness criteria. Before we examine models with this technique, we will look at the various criteria available for assessing model adequacy.

As a first option, we might think of using R^2. But a little thought convinces us that this measure won't work, as it **always** increases when we add more variables. We have an adjustment to R^2 that should help, as it accounts for the number of predictors in the model. Equation 9.4 in the text shows us that this method is equivalent to using MSE, which is an estimate for the model error variance.

Another popular choice is Mallow's C_p. The details of this technique are beyond our mathematical background, but the essence is that unbiased models have a value of C_p near p. Values above p indicate a model that is biased, indicating the mean values of the model are different from the true means.

In practice, we often find conflicting models using our different criteria. Also, we must closely examine the final model chosen, as there may be variables included that have small *t*-values, for example.

Goals: Use the Best Subsets routine to build models.

Skills:

- **Know the criteria used to select models with Best Subsets.** MINITAB uses two main measures: Adjusted R^2 and Mallow's C_p. Each measure slightly different features of the model. Maximizing Adjusted R^2 is equivalent to minimizing MSE, the estimated error variance. Mallow's C_p measures the total "size" of all fitted values.

- **Know the limitations of the models selected with Best Subsets.** Just as with stepwise procedures, we have no guarantee that multicollinearity hasn't created odd subsets of variables. It is possible that a "best" subset has variables with small *t*-values. The automatic procedures are neither foolproof nor guaranteed. They must be used with caution and good judgment.

Reading: Sections 10.1 to 10.2.

Day 33

Activity: Diagnostics.

We have several tools available to us to assess the adequacy of a multiple regression model. Just as for simple linear regression, we will use the residuals from our model to assess model adequacy. We begin by using some graphic methods to determine whether variables not presently included in the model are needed. The main tool is the Added-Variable plot. This is a view of the residuals formed when regressing our response variable on the variables already in the model plotted against the residuals formed when regressing our new variable on the variables already in the model. From such a plot we should be able to notice whether the new variable helps our model, and also what the nature of the (marginal) relationship is. As the authors point out, though, we must be very cautious using this tool, as it is highly dependent on the current model being appropriate.

We finish today by looking at the residuals in more detail than we did on Day 8, where we tried to write the residuals as linear combinations of the *y*-values. Using matrix notation, we can accomplish this task with no trouble.

The key result is that the variance of the residuals themselves involves the hat matrix that we saw on Day 16. Another feature to notice is the correlation among the residuals. However, these correlations are generally small for large data sets.

If we standardize using the adjusted variances, we have **internally studentized residuals**. Another modification is to use the residual for each point developed from the model excluding just that data point. These are called **deleted residuals**, and standardizing them yields **studentized deleted residuals**.

Goals: Introduce diagnostics for multiple regression models.

Skills:

- **Added-Variable plots.** Plotting the residuals from predicting the response variable in a regression model versus the residuals from predicting a variable not included in the model should give us information and insight about the marginal relationship to the new variable. If

the plot shows any trend, we should consider including the new variable. The shape of the plot will further guide us as to the relationship to be used. The example in the text on page 387 points out the danger of just looking at residuals. In that example, the residual plot makes the relationship appear strongly quadratic, but using just a linear term explains the majority of the error, without the unnecessary complication of a quadratic term.

- **Studentized residuals.** By using matrix notation, we see that the residuals are a linear combination of the *x*-values in our model, but they do not have the same variances. One logical fix to this problem is to standardize (or studentize) the residuals by dividing by the square root of the variance so that all residuals now have variance 1.

- **Studentized deleted residuals.** One drawback to residuals is that they tend to be small when a data point is far from the center of a data set. The outlying data point tends to pull the regression surface close to the outlier. Therefore the residual will not be large, and the outlier will remain undetected. One refinement to the residual calculation is to calculate a regression fit **without** the potential outlier and see how close the point is to the fit. If we also standardize, these residuals, they are called **studentized deleted residuals**. If the outlying data point produces a large studentized deleted residual, then we have evidence that the point is inconsistent with the other data, and worthy of further attention.

Reading: Section 10.3.

Day 34

Activity: *X* Outliers. **Homework 5 due today.**

The material from Day 33 can be used to detect outliers in the dependent variable, but caution must be used. It is quite possible for an outlier to dominate the fitting process, if it is far from other data points in the **independent** variables. It would be helpful if we had a measure of how far data points are from each other in the *x*-values. Fortunately, the hat matrix again helps us out.

We know that the variance of the residuals is derived from the diagonal elements of the hat matrix. Thus when the variance is small, we have no difficulty estimating the response surface at that point. This means the point is close to the center of the data, in a multivariate sense. Conversely, when the variance is large, the *y*-values significantly control where the fitted value falls, indicating a difficult point to estimate. This generally occurs on the edges of a data set. Our measure for distance from the center is simply the diagonal elements from the hat matrix, and we call this measure **leverage**.

I would like to explore the hat matrix today, along with various outliers to see how this measure works in practice. Then we will try it on a complicated data set with several independent variables. In all cases, we're looking for **extreme** leverages.

Goals: Use the hat matrix to detect outliers in the independent variables.

Skills:

- **Hat Matrix.** We have seen the hat matrix is $H = X(X'X)^{-1}X'$. Because this is a function of X alone, we can use it to describe the independent values in a data set. On Day 34 we explored outliers in the y dimension. The hat matrix is used to detect outliers in the x dimensions.

- **Leverage.** The diagonal elements of the hat matrix tell us important information about the influence of the individual cases. Cases with large leverage generally are far from the center of a data set, and cases with low leverage generally are near the center of a data set. We often consider leverage to be large when it exceeds $2p/n$, and we consider high leverage cases to be **influential** points.

Reading: Section 10.4.

Day 35

Activity: *Y* Outliers.

So far, we have used only residuals to detect outliers in the dependent variable. But, as we have seen, outliers in the independent variables yield very small residuals, due to their leverage. We need some measures that detect outliers that are not solely based on residuals. Today we will explore three candidates. Df Fits, Cook's distance, and DF Betas.

DF Fits measures how different the fitted values are when a case is excluded from a model. If a case has a large value for Df Fits, that means the model is quite different when including that point in the analysis.

Cook's distance is a combination of leverage and residuals. It measures the effect of a case not just on its own residuals, but on all residuals simultaneously.

Df Betas is a measure of how much the coefficients in the model change when excluding a particular data value. If Df Betas is large, we have evidence that the data point is far from the other data points, either due to being an outlier in the independent variables or in the dependent variable.

Goals: Explore further outlier detection statistics.

Skills:

- **Df Fits.** Df Fits is the difference in the fitted value from a regression using all cases and a regression using all but the current case. If Df Fits is large for a particular data point, we have some evidence that the point is different from the other points. The authors suggest considering a case an outlier if the Df Fits values exceeds 1 from smaller data sets and $2\sqrt{p/n}$ for larger data sets. The formula for Df Fits is a product of the studentized deleted residual and a factor involving leverage. High residuals and high leverage tend to make Df Fits larger.

- **Cook's Distance.** Df Fits measures how much one case affects the fitted value. Cook's distance measures how much **all** residuals have changed by deleting a single case. Again, outliers will have large values of Cook's distance. Our authors consider Cook's distance large if it exceeds the median from the F distribution with p and $n - p$ degrees of freedom. The

formula for Cook's distance also shows it to be a product of the residual and a factor involving leverage. As with Df Fits, high residuals and high leverage tend to make large Cook's distances.

- **Df Betas.** Another way to describe the influence a case has on the regression fit is to see how much the coefficients change when removing the case. Our authors consider Df Betas to be large when they exceed 1 for smaller data sets and $2/\sqrt{n}$ for larger data sets.

Reading: Sections 11.4 and 9.6.

Day 36

Activity: Trees.

Today we will look at fitting a model in a totally different way. Using each variable in turn, we will split the data set into two groups, and measure the internal variance of the response variable to see how effective such a split is. I will show you a spreadsheet approach to doing tree regression. My formula will calculate the standard deviation of the two splits on the data, for each possible split. Then we repeat on each subset we create. The end result is a series of binary splits, which we organize into a tree. Working our way down the tree to a final node, we classify each case and assign it a predicted value. Because we are not fitting an equation, but creating a series of nodes or clusters, this technique is categorized as a **non-parametric** procedure. The diagrams and charts on pages 454 and 455 demonstrate the tree creating process.

Because we are looking so closely at our data, and not requiring any structure in a variable, such as a linear equation or even any formula, we should be very careful not to **overfit** a model to our data. The best way to avoid this is to keep a test data set aside, chosen randomly, from which to verify our model. This is referred to as **data splitting**. We can measure how effective a model is by first fitting it on the development cases or **training sample** and then verifying its predictive ability on the other cases, the **prediction set**. We use MSPR, comparable to MSE, to evaluate how good a model is.

Goals: Examine an alternate approach to making a model, using a classification tree.

Skills:

- **Trees Structure.** We can use a binary tree to classify cases into various nodes. Using an appropriate splitting criterion, the data set is broken into two subgroups. Each subgroup is further divided until a stopping criterion is met. The final set of subgroups constitutes the classification scheme. Each subgroup's average is then the assigned fitted value for all cases in the subgroup. A goodness-of-fit, like MSE, can then be measured.

- **Splitting criterion.** A reasonable criterion to use to split cases into subgroups is the overall MSE. For two groups, this amounts to making the individual subgroup standard deviations as small as possible. One drawback of using a classification scheme is the discrete nature of the resulting rules. The obvious advantage is not having to specify the form of the model beforehand.

- **Data Splitting.** With any form of model fitting, we would like a method to establish model replicability. One way to accomplish this is to set aside some of the sample cases to be used as a "verification" or "predication" data set. The remainder of the cases, the "training set", are used to develop the model, and then the predication data set is used to assess how well the model performs. If too many variables are used, the model may be "overfit", in which case the model will not perform well using the prediction data set. If too few variables are used, the model may not be very useful in practice. The art of model fitting is creating an effective model somewhere between these two extremes.

Reading: Sections 13.1 to 13.2.

Day 37

Activity: Non-Linear Regression I.

Often we want to fit a model that is non-linear, because in nature, not every response surface is linear or can be transformed into a linear surface. We will use a Taylor series expansion in our estimates. Basically we are trying to find places where the derivatives are all simultaneously zero. The spreadsheet I will use in class should do this for us, but I think it is important that you know what steps the program takes.

As a first step into non-linear regression, I would like to take us through the steps we would take if we were trying the straightforward calculus approach. You should understand where our technique fails, and why we cannot derive a general solution. We will be able to produce a few equations that are useful, but in general we will not be able to find a complete solution as we were able to do with linear regression.

We have several options at this point. We can go ahead and numerically search for parameter values that minimize the squared errors, but without using any calculus results. This is what Excel does when we use the "optimal" solutions.

Another option we have is to use the Gauss-Newton method, which is a systematic search for optimal answers, using more complicated calculus results. Specifically, the G-N method resembles the Newton method you may have seen in introductory calculus. Using an iterative approach, we calculate the derivative and project linearly to the "axis", generating a new update to the parameter estimates. Then we repeat the procedure, until we zero in on a place where the estimates change very little, and hopefully we are at a relative minimum. However, just as in the one-dimensional Newton method, we have no guarantee of that our relative minimum is actually a **global** minimum.

Goals: Introduce non-linear regression and its complexities.

Skills:

- **Forms of Non-Linear Models.** As we have seen, the linear model can be written using matrix algebra. Non-linear models cannot be written as matrix products. We can still try using calculus methods on these models. However, the resulting systems of equations (the normal equations) are not generally solvable with closed form methods.

- **Least Squares Approach.** By taking first derivatives of the non-linear model and setting them to zero, we can set up the equations needed to find the critical point minimizing SSE. However, as we have seen, these equations are generally not solvable in close form. If we are able, we can iteratively find solutions.

- **Gauss-Newton Method.** The Gauss-Newton method is a way to iteratively find a solution to a non-linear regression model. The method requires a reasonable starting point, but once begun, the method should converge to a solution relatively quickly.

Reading: Sections 13.3 to 13.4.

Day 38

Activity: Non-Linear Regression II.

While using the first option from Day 37 may be much simpler (we **are** after all using the built-in search routines of Excel to "derive" the solutions) we do not have the luxury of knowing the uncertainty in our estimates. The Gauss-Newton method has been studied sufficiently that we now know techniques to estimate variances of our estimates. They are based on large sample theory, so there is some caution to using it indiscriminately. See the guidelines on pages 528 and 529. Essentially, $E\{\mathbf{g}\} = \gamma$ and $s^2\{\mathbf{g}\} = MSE(\mathbf{D}'\mathbf{D})^{-1}$.

Goals: Explore inference techniques for non-linear regression.

Skills:

- **Error Variance Estimate.** Because we are using a linear approximation to a multivariate surface, we can use our usual matrix results to estimate the variance-covariance structure. The estimates are unbiased, and with the variance estimates we can thus construct confidence intervals and hypothesis tests.

Reading: Sections 14.2 to 14.3.

Day 39

Activity: Logistic Regression.

We have considered binary independent variables, but we have always had a continuous response variable. Today we will explore a model for a binary response variable. Before we try logistic regression, we will attempt to impose our usual linear model, and discover the reasons it is inadequate.

The most common model used with the binary response variable is the logistic model, which has two parameters. One controls the horizontal asymptote, and the other two control the steepness of the curve. The form of the model is $E(Y_i) = \dfrac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$. We will take a few minutes to do some calculus on the function, then we will try our hand at fitting a model numerically using our G-N spreadsheet.

Goals: Examine the case of a qualitative response variable, using the logistic function as a model.

Skills:

- **Binary Response Variable.** When the response variable requires a yes/no answer, our usual linear regression models are inadequate, as nothing in the model specification accounts for 0/1 y-values.

- **Logistic Form.** The usual form of the logistic function is $E(Y_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$. Using this form, the main parameter of interest is β_1, which can be interpreted as the log of an odds ratio.

- **Non-Linear Fit.** Description.

Reading: Section 14.5.

Day 40

Activity: Logistic Inference. **Homework 6 due today.**

Our last topic will be the inferences we can make using logistic regression. We have two options: the Wald test, which is based on the large sample theory of maximum likelihood estimates and is used for single parameters at a time, and the Likelihood Ratio test which is used for sets of variables.

Goals: Perform inference for logistic regression.

Skills:

- **Intervals and Tests: Wald.** The kind of test we performed for non-linear regression is also referred to as the Wald test. We calculate the variance based on the non-linear estimates and use them in t- or z- procedures.

- **Intervals and Tests: Likelihood Ratio.** Using Extra Sums of Squares procedures we can test whether several parameters are significant. The details of the test involve the χ^2 distribution.

Reading: Chapters 9 to 11 and 13 to 14.

Day 41

Activity: Unit 3 Review.

Goals: Know everything.

Reading: Chapters 9 to 11 and 13 to 14.

Day 42

Activity: Exam 3. This last exam covers model building, advanced diagnostics, and non-linear regression. Some of the questions may be multiple choice. Others may require you to show your worked out solution. Don't forget to review your class notes and these notes.

Return to Chris' Homepage

Return to UW Oshkosh Homepage

Managed by: chris edwards
Last updated August 1, 2011