Hyunseung Kang
STAT 102 Review Guide
Mar. 19, 2011

March 19, 2011: Thanks to Eliza and Jennie, fixed typo on S_m and S_p in simple linear regression as well as the typo on α on the prediction interval for difference in only ONE x

March 20, 2011: Thanks to Tianpu, fixed general spelling errors + fixed formula for test statistic for paired difference.

March 20, 2011: Thanks to Jacob for the typo of critical t approximation of 1.96

March 20, 2011: Thanks to Adam, corrected typo on MSE for simple linear regression

March 20, 2011: Added Effects Table under multiple linear regression. Added section about corrections to previous years midterm solutions. I'll punish the TAs from previous years for writing terrible solutions (lol. j/k...they're just TAs...have sympathy for them too :))

March 21, 2011: Made additional correction notes on previous years midterm solutions.

March 21, 2011: Thanks to Eliza and Jennie, corrected typo on the point estimate of prediction and confidence interval given one X in the multiple regression section.

March 21, 2011: Thanks to Adam, corrected typo on confidence interval for pooled case in group mean comparison

March 21, 2011: Thanks to Bob and Mac, reworded the explanation about standard deviation of paired group mean test.

March 21, 2011: Added question on FAQ about clarifying the terms standard error and standard deviation. Added more updates from e-mail I sent out to class.

March 21, 2011: Thanks to You Jin, fixed degrees of freedom for paired group mean test from $n-2$ to $n-1$

March 22, 2011: Thanks to Erika, corrected the α value for the critical t in multiple regression t-test from α to $\alpha/2$

March 22, 2011: Thanks to Billy, corrected the MSE term for ANOVA in multiple regression from $K-2$ to $n-K-1$.

March 22, 2011: Thanks to Liz, corrected degrees of freedom of one-sided paired group mean test from $n-2$ to $n-1$

Based on comments from last semester's Stat 101, many students requested that I formally type out a review sheet/guide before each exam that's short, concise, and up to the point. Also, many students found practice questions we made for Stat 101 final last year to be tremendously helpful in learning to compute power and Type I error (even though none of the concepts showed up on the final exam). I hope that this document achieves both goals.

The review guide first starts out with FAQs about concepts and the course material. Afterwards, there is a concise, 6∼7-page guide which contains formulas, concepts, and rudimentary question-solving strategies. Each section has corresponding lecture numbers, book chapters, homework number, and quizzes. Finally, the guide ends with practice midterm questions listed by topic, which will be VERY HELPFUL. Also at the end are additional problems that I believe have a 95% chance on showing up on the exam (based on fitting a "mental regression" :) between the previous three examinations for this course) and corrections to previous midterms.

All parts, except for the practice question parts, are complete (well, some parts are scanty, especially the part on multiple regression and I doubt I will make it super-complete by this Tuesday (I have a life outside of 102 you know...)). I hope you find it useful. As always (despite how annoying this is...), please check webCafe often for updates. Also, let me know if there are any errors on this document so I can fix it ASAP.

1. *When do I construct a confidence interval and/or prediction interval? What are the formulas for those?*

   Ah!!! This is probably the most unnecessarily confusing part of this course. It mainly stems from identifying what interval to make (and in essence, what SE to use). I looked through lecture notes, quizzes, homework, and your practice midterm and here's my attempt at this clarification. I also provided citations in case if you wanted more practice with these type of questions.

   First, what key words/phrases should I look for in a question? I Regardless of whether you are doing simple regression or multiple regression, you get the **prediction interval** when the question states

   (a) "predict the selling price of an *individual* car" (lecture 7, slide 4)

   (b) "*individual* confidence intervals are " (lecture 12, slide 9)

   (c) "provide a 95% *prediction interval* for the price " (homework 3, question 4 & Practice Multiple Regression Questions, Analysis III, Question (a) & Practice Multiple Regression Questions, Question 3fii)

   (d) "should be *predicted* with 95% confidence to be between" (quiz 2, section 2)

(e) "*predict* with 95% confidence that the difference" (quiz 2, section 3)

(f) "with 95 % confidence, *the prediction lies*..." (practice quiz 2, version 1, question 2)

(g) "give a 95% confidence interval for the difference in price asked by two cars" (2008 Midterm, Question 2a)

(h) "give a 95% confidence interval for the price ... for *a particular car*"(2008 Midterm, Question 2f)

(i) "give a 90% confidence interval for *this prediction*"(2009 Midterm, Question 3b & 2010 Midterm, Question 1d)

(j) "*John* wants to buy *a ... house*...Find a 95% confidence interval for the price of the house" (2010 Midterm, Part I, question 2d & 2010 Midterm, Part II, question a)

(k) "Ren is interested in buying *all* five (identically built) house...Give a 95% confidence interval for the total price of the five house " (2010 Midterm, Part I, question 2f)

(l) "Stimpy wants to buy *5* independent, unrelated houses...Give a 95% confidence interval for the total amount...'(2010 Midterm, Part I, question 2g)'

Note that key terms have been italicized, although some may be missing because I don't even have the slightest clue on what interval to construct.

You get the **confidence interval** when the question states

(a) "estimate of the *average* selling price of all of these cars" (lecture 7, slide 4)

(b) "*mean* confidence intervals are " (lecture 12, slide 9)

(c) "give a 95% confidence interval to" (homework 3, question 4)

(d) "with 95% confidence, we can state that the *average* difference in" (practice quiz 2, version 2, question 1)

(e) "95% chance that the true value (of the predicted value) will be within' (practice quiz 2, version 3, question 4)'

(f) "Find the 95% confidence interval for the *slope of the regression equation*" (2009 Midterm, question 3a & 2010 Midterm, question 3b & Practice Multiple Regression Questions, Analysis I, Question 1b)

(g) "...they were below the 2.5 percentile of *all* scores" (2010 Midterm, question 1e)

(h) "Find a 95% interval for the *average* price of a..."(2010 Midterm, Part I, question 2c)

(i) "Find a 95% confidence interval for the *average* difference in price between a..."(2010 Midterm, Part I, question 2e & Practice Multiple Regression Questions, 2d)

The whole point of all this: Besides the obvious clues (e.g."get a prediction interval"), I think the key terms seem to be **average** and ***individual***. If you see the term "average" anywhere, get **confidence intervals**. If you see the term "individual", go for ***prediction intervals***!

Second, how do you choose the standard error? Do you multiply by the amount of change in x? What do you do? Well, thankfully, the hard part is over. As long as you know what interval you need to make (confidence or prediction), you can use the formulas given in the course summary (or the e-mail) I wrote below. There are a couple of cases, depending on how many Xs are varying at once, but it's relatively straightforward to use (or I hope so).

2. *What is the difference between standard deviation and standard error?*

Standard error is a subset of standard deviation. In the end, both are identical from a conceptual standpoint (i.e. they explain the variation of a histogram or data point). But, standard error is often used when we want to know the variation around a single random point or a estimator (e.g. sample mean) while standard deviation is often used when you want to describe variation in a histogram. In the case below, the quantity of interest, the sample mean, has $\frac{sd}{\sqrt{n}}$ variation around the sample mean. The quantity, $\frac{sd}{\sqrt{n}}$ is often referred to as the standard error.

In various questions throughout the previous midterms, sometimes standard deviation is used while other times, standard error is used. While there is no clear-cut rule, you can be certain that if your quantity of interest is the sample mean value, then the term "standard error" and you would use $\frac{sd(x)}{\sqrt{n}} = Sd(\bar{x})$. On the other hand, if you are only interested in the variation of the distribution of x, you would use $Sd(x)$. Note that if you only have one data point and you're interested in where this data point lies on the distribution (e.g. Midterm 2010, question 1e), you only use $Sd(x)$ or you can think of it as the case when $n = 1$ in the $\frac{sd(x)}{\sqrt{n}}$, although this interpretation is not correct.

Another explanation (from the e-mail I sent out)

Basically, you can think of standard error as another flavor of standard deviation. Conceptually and (I'll regret saying this because you'll get ultra-confused) mathematically, they are identical; they both measure variation of something. They are all square root of the variance of something (i.e. sqrt(Var(something))).

For example, the variation around the sample mean (i.e.$\bar{X} = \frac{1}{n}\sum_{n=1}^{n} X_i$ (assuming we have iid Xs (i.e. $Var(X1) = ... = Var(Xn)$)...which is almost always the case with

sampling where we take iid samples from the population)

$$\sqrt{Var(Xbar)} = \sqrt{\frac{1}{n^2}Var(X_1 + ... + X_n)} = \sqrt{\frac{1}{n^2} * n * Var(X_1)} = \sqrt{\frac{1}{n}} * SD(X1)$$

The derivation uses Stat 101 principals which you should be familiar with, I hope...**

Another example of a variation around a sample of people's height. We collect n people's height as $X_1, ..., X_n$ and want to measure the variation; note that we sample, again, in an iid fashion, which is almost always the case in sampling. Hence, our variation around a sample of people's heights would be just $\sqrt{Var(X_1)} = SD(X_1)$. Note $Var(X_1) = Var(X_2) = ... = Var(X_n)$ because the samples were obtained iid and hence, they would have the same variation...in fact, let's define $Var(X) = Var(X_1) = ... = Var(X_n)$ (and this is what is often done in textbooks/lectures/exams/ because they don't want to write out all $X_1, X_2, ..., X_n$). Further note that this quantity, $\sqrt{Var(X1)}$, is the standard deviation of the histogram that you would find in JMP.

Hence, depending on what your "something" is, you either use just the SD of that something or $\frac{SD}{\sqrt{n}}$. In problems dealing with difference between two groups (Midterm 2008), you use the $\sqrt{\frac{1}{n}} * SD(X)$ because that "something" is the mean sample difference. In Midterm 2010 about the principle and getting the z-score of whether the principle is correct or not, you divide by the sd because that "something" is just his school (or in terms of random variables, just $X_1 ... \sqrt{Var(X1)} = Sd(X1)$)

However, in most of statistics and in your previous midterms, the term standard error is used to explain the variation around a particular estimate (e.g. sample mean, sample difference of mean, estimate of a slope, etc.). For example, standard error of the mean refers to, in essence, the standard deviation (aka the variation) of the sample mean. On the other hand, the term standard deviation is often used to explain variation of a distribution (e.g. distribution of heights, distribution of differences between paired groups, distribution of scores of 40 schools, etc.).

Finally, so what to use Sd or $\frac{SD}{\sqrt{n}}$? Well, as stated before, it depends on that "something". It's difficult to give you a rule or a case-by-case scenarios because there really isn't one. Sd and $\frac{SD}{\sqrt{n}}$ both measure variation and BOTH ARE Standard deviation. But the first one is the standard deviation of some sample (e.g. heights, weights, GPA) while the other one is the standard deviation of the SAMPLE MEAN.

1 Group Mean Comparisons

Lecture 1-3, Book 2.7-2.9, Homework 1, Midterm 2008 Question 1. Suppose we have group A and B (or labeled 1 and 2). We want to know whether the groups share the same mean based on some measurement (X) we made. There are three possible ways to test this, depending on the assumptions we make. For each test, we have a test statistic and we decide to reject the null (aka that the mean of the groups are the same) by comparing the test statistic with a critical value (usually 2 or 1.96).

1. **When $\sigma_A = \sigma_B$ (or "pooled")**, we conduct either a one-sided or a two-sided test. In both cases, the test statistic is $t = \frac{\bar{X}_A - \bar{X}_B - \mu_0}{\sqrt{s_p^2\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}}$ where $\bar{X}$ is the sample mean from group A or B, n_1 and n_2 are the sample sizes for groups A and B, respectively, and $s_p^2 = \frac{(n_A-1)s_A^2+(n_B-1)s_B^2}{(n_A+n_B-2)}$. Here, s_A and s_B stand for standard deviations for groups A and B. For μ_0, see (a) for details. You generally use this case when they specifically tell you that the σs are equal to each other. Notice how the case for (2) and (3) are almost identical to case for (1) when you do the two-sided and one-sided tests.

   (a) <u>One-sided case</u>: your hypothesis is $H_0 : \mu_A - \mu_B = \mu_0$ and $H_a : \mu_A - \mu_B > \mu_0$ or $H_a : \mu_A - \mu_B < \mu_0$ depending on what side you're looking for. $\boldsymbol{\mu_0}$ *stands for the amount of difference you want to test. For example, if we want to test to see if the difference is greater than two,* $\mu_0 = 2$ *(with* $>$ *sign). If we want to test that there is just a difference,* $\mu_0 = 0$. If the alternative is $>$, you reject the null when the test statistic is greater than t_α with $n-2$ degrees of freedom (note that the t_α should be positive). If the alternative is $<$, you reject the null when the test statistic is less than t_α with $n-2$ degrees of freedom (note that the t_α should be negative) OR when the p-value is less than α. If $n-2 \geq 30$, you can approximate the t_α with a z value at α. If $\alpha = 0.05$ and $n-2 \geq 30$, we have $t_\alpha \approx 1.65$. You generally know the problem is one-sided when they ask whether the mean of one group will be bigger than the other group.

   (b) <u>Two-sided case</u>: your hypothesis is $H_0 : \mu_A - \mu_B = \mu_0$ and $H_a : \mu_A - \mu_B \neq \mu_0$. You decide to reject the null when the *absolute value* of the test statistic is greater than $t_{\alpha/2}$ with $n-2$ degrees of freedom OR when the p-value is less than α (NOT $\alpha/2$). If $n-2 \geq 30$, you can approximate $t_{\alpha/2}$ with z value at $\alpha/2$. If $\alpha = 0.05$ and $n-2 \geq 30$, we have $t_{\alpha/2} \approx 2$ or 1.96. You generally know the problem is two-sided when they ask whether the mean of two groups will be different from each other.

   Alternatively, you can perform a two-sided test by constructing **confidence intervals for the pooled case** at level α. The confidence interval for this case is $\bar{X}_A - \bar{X}_B \pm t_{\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}$ where $t_{\alpha/2}$ has $n-2$ degrees of freedom. If the confidence interval does not contain μ_0, you can reject the null; otherwise you retain the null.

2. **When $\sigma_A \neq \sigma_B$ (or "unpooled")**, we conduct either a one-sided or a two-sided test. In both cases, the test statistic is $t = \frac{\bar{X}_A - \bar{X}_B - \mu_0}{\sqrt{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)}}$ where we use the same notation as the case for "pooled". You almost always use this case unless you are told to use (1)

   (a) <u>One-sided case</u>: everything (the hypothesis, when to reject, when to use one-sided, the approximation of t_α by 1.65, etc.) is the same as (1) except that the degrees of freedom on t is $\Delta = \frac{(s_A^2/n_A + s_B^2/n_B)^2}{(s_A^2/n_A)^2/(n_A-1) + (s_B^2/n_B)^2/(n_B-1)}$.

   (b) <u>Two-sided case</u>: everything (the hypothesis, when to reject, when to use two-sided, the approximation of $t_{\alpha/2}$ by 1.96 or 2, etc.) is the same as (1) except that the degrees of freedom on t is $\Delta = \frac{(s_A^2/n_A + s_B^2/n_B)^2}{(s_A^2/n_A)^2/(n_A-1) + (s_B^2/n_B)^2/(n_B-1)}$.

   Alternatively, you can perform a two-sided test by constructing **confidence intervals for the unpooled case** at level α. The confidence interval for this case is $\bar{X}_A - \bar{X}_B \pm t_{\alpha/2}\sqrt{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)}$ where $t_{\alpha/2}$ has Δ degrees of freedom. If the confidence interval does not contain μ_0, you can reject the null; otherwise you retain the null.

3. When the **groups are paired**, we conduct either a one-sided or a two-sided test. In both cases, the test statistic is $t = \frac{\bar{X}_{diff} - \mu_0}{s_{diff}}$ where $\bar{X}_{diff}$ is the average of the difference between pairs of measurements (e.g. if $X_{i,A}$ and $X_{i,B}$ are "paired" with each other, it is the average of the differences $X_{1,A} - X_{1,B}, X_{2,A} - X_{2,B}, ..., X_{n,A} - X_{n,B}$) and the s_{diff} is the standard deviation of $\bar{X}_diff$ (or standard error, as notated on Midterm 2008 Q1b). Only pair the observations from two groups when you are told to do so or if it makes sense to do so. For example, if you are comparing male and female salary, it make sense to compare by pairing each female by a male who have the same job title. Also, ignore pairs with missing values.

   (a) <u>One-sided case</u>: everything (when to reject, when to use one-sided, the approximation of t_α by 1.65, etc.) is the same as (1) except (i) the degrees of freedom on t is $n-1$ where n stands for the number of paired differences (n may be different from (1) if we ignored a few observations because they were missing when we paired them) and (ii) the hypothesis is that $H_0 : \mu_{diff} = \mu_0$ and $H_a : \mu_{diff} < \mu_0$ or $H_a : \mu_{diff} > \mu_0$, depending on which "side" of the test you want.

   (b) <u>Two-sided case</u>: everything (the hypothesis, when to reject, when to use two-sided, the approximation of $t_{\alpha/2}$ by 1.96 or 2, etc.) is the same as (1) except (i) the degrees of freedom on t is $n-1$ where n means the same thing as (i) for this case and (ii) the hypothesis is that $H_a : \mu_{diff} = \mu_0$ and $H_a : \mu_{diff} \neq \mu_0$

   Alternatively, you can perform a two-sided test by constructing **confidence intervals for the paired case** at level α. The confidence interval for this case is

$\bar{X}_{diff} \pm t_{\alpha/2} s_{diff}$ where $t_{\alpha/2}$ has $n-1$ degrees of freedom. If the confidence interval does not contain μ_0, you can reject the null; otherwise you retain the null.

2 Simple Regression

Lecture 4-9, Book 3.1-3.5,3.7,6.1-6.7, Homework 2 and 3, Quiz 1 Suppose we take two types of measurements, y_i and x_i, and we want to build a model that describes the relationship between them. In the framework of simple linear regression, for each observation i, we model these measurements as $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where ϵ_i is independent and identically distribute as a normal distribution with mean 0 and variance σ^2 the i represent the ith observation; note that under this model, the conditional mean of y given a particular x_i is $\mu_{y|x_i} = \beta_0 + \beta_1 x_i$. Now, we find out what β_0 and β_1 is (aka "fit" the model) by using a least squares fit algorithm.

2.1 Estimates, ANOVA, and Summary of Fit

Note that *corr* below refers to the correlation between X and Y. Also, for the ANOVA table, $K = 1$ for simple regression

Summary of Fit	Values
RSquare	$R^2 = \frac{SSR}{SST} = (corr)^2$
Root Mean Square Error	$s_e = \sqrt{MSE} = S_y\sqrt{1 - corr^2}$ = Standard deviation of residuals
Mean of Response	$\bar{y}$
Observation	n

ANOVA Table	DF	SS	MS = SS/DF
Model	K	$SSR = \sum_{i=1}^{n}(\bar{y} - \hat{y}_i)^2$	$MSR = \frac{SSR}{K}$
Error	$n - K - 1$	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$MSE = \frac{SSE}{n-K-1}$
C. Total	$n - 1$	$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = SSE + SSR$	NA

The **F-statistic (i.e the "F-ratio" or the "F-test")** is $F = \frac{MSR}{MSE} = (t_{\hat{\beta}_1})^2$ with 1 and $n-2$ degrees of freedom; note that the F-test always has two degrees of freedom. The test resulting from this statistic is of the form $H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$ (aka whether the slope term is necessarily or not in the regression OR whether X has any predictive power for Y) and you reject the null if the $F > F_\alpha$ where F_α is the critical value of F given 1 and $n-2$ degrees of freedom. JMP also gives you a p-value = $P(F > F_\alpha)$ and you can reject the null if the p-value is less than α.

Parameters	Estimate	Std Error = Std Dev of Estimate	t Ratio (or test)	$Prob > \|t\|$
Intercept (in units of y)	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$	$SE(\hat{\beta}_0) = s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}$	$t_{\hat{\beta}_0} = \frac{\hat{\beta}_0}{SE(\hat{\beta}_0)}$	p-value of test for intercept
x or slope (in units of y over x)	$\hat{\beta}_1 = \frac{s_y}{s_x} corr$	$SE(\hat{\beta}_1) = s_e \sqrt{\frac{1}{(n-1)s_x^2}}$	$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$	p-value of test for slope

2.2 Hypothesis Testing, Confidence Intervals, and Prediction Intervals

If you wanted to do a **one-sided test on** β_i (either slope or intercept), you need to construct your own t-test. The hypothesis would be $H_0 : \beta_i = 0$ and $H_a : \beta_i > \mu_0$ or $H_a : \beta_i < \mu_0$ depending on the context of the question. The t-statistic to do both hypothesis would be $t = \frac{\hat{\beta}_i - \mu_0}{SE(\hat{\beta})_i}$. For <-sided test, you reject the null if $t < t_\alpha$ (t_α should be negative and have $n-2$ degrees of freedom). For >-sided test, you reject the null if $t > t_\alpha$ (t_α should be positive and have $n-2$ degrees of freedom). Or, for both tests, you can reject the null if the p-value $= P(t > t_\alpha)$ or $P(t < t_\alpha)$, depending on the $<, >$ sign, is less than α

If you want to do a **two-sided test on** β_i (either slope or intercept), the hypothesis would be $H_0 : \beta_i = \mu_0$ and $H_a : \beta_i \neq \mu_0$; note that the case when $\mu_0 = 0$ is printed in JMP outputs. The t-statistic (or t-Ratio in JMP output) is $t = \frac{\hat{\beta}_i - \mu_0}{SE(\hat{\beta})_i}$. You reject the null if the *absolute value* of the t-statistic is greater than $t_{\alpha/2}$ where $t_{\alpha/2}$ is the critical value with $n-2$ degrees of freedom (this value should be positive). Alternatively, you can reject the null if the p-value $= P(|t| > t_{\alpha/2})$ is less than α. Or, you can construct a confidence interval and see if the interval contains μ_0 or not. Generally, if we want to see if x has any predictive power for y, we do a t-test of the slope.

To construct **confidence intervals around** β_i (either slope or intercept), it is $\hat{\beta}_i \pm t_{\alpha/2} SE(\hat{\beta}_i)$ where the $t_{\alpha/2}$ has $n-2$ degrees of freedom .

Suppose we have a particular value of x_i and we want to make a prediction interval and a confidence interval. The formula for the **confidence interval given** x_i would be $\hat{y}_i \pm t_{\alpha/2} s_m$ where $t_{\alpha/2}$ would have $n-2$ degrees of freedom and $s_m = s_e \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}}$. The formula for the **prediction interval given** x_i would be $\hat{y}_i \pm t_{\alpha/2} s_p$ where $s_p^2 = s_m^2 + s_e^2 \approx s_e^2 = RMSE^2$ or $s_p = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}}$.

2.3 Model Assumptions and Diagnostics

There are three assumption to the model in the linear regression:

1. **Linearity:** Is x and y linearly related? We can check this by looking at the scatterplot of y and x and seeing whether there is any non-linear pattern OR looking at the residual plot and seeing whether there is any non-linear pattern there.

2. **Homoscedasticity:** Are the variances for each observation constant? We can check this by looking at the scatterplot of y and x and seeing whether there is any "spreading" around the fitted line OR looking at the residual plot and seeing whether there is any "spreading" along the x-axis.

3. **Normality:** Are the error terms/residuals normally distributed? We can check this by looking at the residual plot and seeing whether there is uniform spreading of the data points around the x-axis OR looking at the normal quantile plot (i.e. the QQ plot) of your residuals and seeing whether your data points lie inside the red butterfly lines OR looking the histogram of your residuals and seeing whether they are normally distributed around 0

In addition, there are **outliers** and **influential points** that may or may affect the performance of your regression. **Outliers** are points that lie far from the fitted line (in a vertical direction); in some, but not all cases, outliers are "outliers" in the y-direction. **Influential points** are points that, once removed, significantly alters the slope and/or intercept of the fitted line; in almost all cases, influential points are outliers in the x-direction. Easy ways to detect both are looking at scatterplots of y and x with the fitted line and looking for these points :)

3 Multiple Linear Regression

Lecture 10-14, Book 4.1-4.5,5.2,8.1-8.4, Homework 4, Quiz 2 Unlike simple regression, in multiple regression we have multiple different types of measurements of $x_{i,1}, x_{i,2}, ..., x_{i,K}$ where each $x_{i,j}$ represents the ith observation of a j type measurement. For example, we might have a data set where the ith person's SAT Verbal (i.e. $x_{i,1}$) and SAT Math (i.e. $x_{i,2}$) are measured to predict his/her GPA (i.e y_i). Under multiple linear regression, our model is $y_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_k x_{i,K} + \epsilon_i$ where ϵ_i is independent and identically distribute as a normal distribution with mean 0 and variance σ^2 the i represent the ith observation; note that under this model, the conditional mean of y given a particular x_i is $\mu_{y|x_i} = \beta_0 + \beta_1 x_{i,1} + ... + \beta_k x_{i,K}$. Now, onto what happens after we fit using least squares.

3.1 Estimates, ANOVA, Summary of Fit

Summary of Fit	Values
RSquare	$R^2 = \frac{SSR}{SST}$
Root Mean Square Error	$s_e = \sqrt{MSE}$= Standard deviation of residuals
Mean of Response	$\bar{y}$
Observation	n

ANOVA Table	DF	SS	MS = SS/DF
Model	K	$SSR = \sum_{i=1}^{n}(\bar{y} - \hat{y}_i)^2$	$MSR = \frac{SSR}{K}$
Error	$n-K-1$	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$MSE = \frac{SSE}{n-K-1}$
C. Total	$n-1$	$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = SSE + SSR$	NA

The **F-statistic (i.e the "F-ratio" or the "Full F test")** is $F = \frac{MSR}{MSE}$ with K and $n-K-1$ degrees of freedom. The test resulting from this statistic is of the form $H_0 : \beta_1 = ... = \beta_K = 0$ and H_a: at least one of β_i is not zero (aka or whether all the Xs have any predictive power for Y) and you reject the null if the $F > F_\alpha$ where F_α is the critical value of F given K and $n-K-1$ degrees of freedom. JMP also gives you a p-value = $P(F > F_\alpha)$ and you can reject the null if the p-value is less than α. You use this when you want to test the overall performance of the regression OR if you don't know some Xs and you built the model with the known Xs to see if the known Xs are any good for the regression.

Effects Test	DF	Sum of Squares = $SSE_{removed} - SSE_{allvar.}$	F Ratio	Prob
Variable to remove(e.g. one X)	1	(see equation above)	$\frac{SSE_{removed} - SSE_{full}}{MSE_{full}}$	p-val

Parameters	Estimate	Std Error	t Ratio (or test)	$Prob > \|t\|$
Intercept, X_1,..., or X_K	$\hat{\beta}_i$	$SE(\hat{\beta}_i)$	$t_{\beta_i} = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$	p-values for that slope/intercept

3.2 Hypothesis Testing

T-test: Used if you want to test whether a chosen variable, let's say β_i is significant controlling for other βs. It tests the hypothesis $H_0 : \beta_i = 0$ and $H_a : \beta_i \neq 0$. You reject the null if the t Ratio is greater than $t_{\alpha/2,n-K-1}$ where $n-K-1$ represents the degrees of freedom OR if the p-value is less than α.

Partial F -test: Used if you have a sub-model that has less parameters than the full model (and it comes from the same data set). The hypothesis is $H_0 : \beta$ not in the full model is zero and H_a : at least one of the β not in the full model is non-zero. The test statistic is $F = \frac{(SSE_{reduced} - SSE_{full})/(K-L)}{MSE_{full}}$ where L is the DF for the reduced model (under "Model") with degrees of freedom $K-L$ and $n-K-1$ (aka the DFs for the numerator and the denominator). You reject the null if F is greater than $F_{K-L,n-K-1}$ (the subscripts represent

the degrees of freedom) or if the p-value is less than α
Relationship between t ratio and F test: $t^2_{\beta_{removed}} = \frac{(SSE_{removed} - SSE_{full})/(K-L)}{MSE_{full}} = F$. $SSE_{removed}$ is the SSE obtained from an ANOVA table with the variable removed from the full regression (i.e. run Fit Model with that variable removed and use the ANOVA table you get from that Fit Model run). This relation ONLY WORKS if you remove only ONE variable from the full regression. The degrees of freedom for F is 1 and $n-K-1$ and the degrees of freedom for t is $n-K-1$.

3.3 Confidence Intervals, Prediction Intervals, and CI/PI for Difference Between Two Predictions

Confidence intervals for slopes ONLY: $\hat{\beta}_i \pm t_{\alpha/2,n-K-1} SE(\hat{\beta}_i)$
Prediction interval given ONE set of x values: $\hat{y} \pm t_{\alpha/2,n-K-1} Sp(x)$ where $Sp(x)$is the Sp value you find for that x and $\hat{y}$ is the predicted value given x. You can replace $Sp(x)$ by $RMSE$ as an approximation
Confidence interval given ONE set of x values: $\hat{y} \pm t_{\alpha/2,n-K-1} Sm(x)$. You CANNOT replace $Sm(x)$ by RMSE.

We construct an example for the next cases easier to follow. Consider a regression where we want to measure how physically attractive a TA is (Y) based on Age (x_1) and Number of Students in Review Sessions (x_2). Suppose TA Hyunseung (X_A) is 22 years old and has 150 students during review sessions and TA Bob (X_B) is 22 years old and has 10 students during review sessions. Here, the two TAs only differ in the number of students in the review session. In the equations below, $\hat{\beta}_i$ would refer to the slope coefficient for the number of students in review sessions. Also note that the "center" of confidence or prediction interval can either be the multiple of the slope by the amount of difference, $\Delta\hat{\beta}_i$, or $\hat{Y}_A - \hat{Y}_B$ where $\hat{Y}_A$ are predictions for person A.

Confidence intervals for difference in only ONE β_i by Δ amount: $\Delta(\hat{\beta}_i \pm t_{\alpha/2,n-K-1} SE(\hat{\beta}_i))$ OR (slightly inaccurate one) $\Delta(\hat{\beta}_i \pm t_{\alpha/2,n-K-1}\sqrt{Sm(A)^2 + Sm(B)^2})$ where $Sm(A)$ represents the Sm for person A
Prediction intervals for difference in only ONE β_i by Δ amount:
$\Delta\hat{\beta}_i \pm t_{\alpha/2,n-K-1}\sqrt{Sp(X_A)^2 + Sp(X_B)^2}$ where $Sp(X_A)$ is the Sp value obtained from person A OR (if we approximate all Sps by RMSE) $\Delta\hat{\beta}_i \pm t_{\alpha/2,n-K-1}\sqrt{2}RMSE$. THIS IS NOT A TYPO!

Now consider a case where we have TA Hyunseung (X_A) who is 22 years old and has 150 students during review sessions and TA Bob (X_B) who is 50 years old and has 10 students during review sessions. Here, the two TAs differ in multiple Xs.

Confidence intervals for difference in TWO predicted values through MULTIPLE Xs: $(\hat{Y}_A - \hat{Y}_B) \pm t_{\alpha/2,n-k-1}\sqrt{Sm(A)^2 + Sm(B)^2}$. RARELY USED!

Prediction intervals for difference in TWO predicted values through MULTIPLE Xs: $(\hat{Y}_A - \hat{Y}_B) \pm t_{\alpha/2,n-k-1}\sqrt{Sp(A)^2 + Sp(B)^2}$ OR (slightly inaccurate one) $(\hat{Y}_A - \hat{Y}_B) \pm t_{\alpha/2,n-k-1}\sqrt{2}RMSE$. IF you don't have Sps use the inaccurate one!

3.4 Stepwise Regression

The whole point: To pick out the "best" Xs for Y (aka to pick the best linear model for Y)

Forward: Choose the next variable that has the highest improvement in SSE

Backward: Choose the next variable that has the least amount of decrease in SSE (i.e. take out the lowest t-ratio)

3.5 Polynomial Regression

Mechanics: Basically a regression with x and x^2. Interpretation, confidence intervals, hypothesis testing, etc. are identical.

3.6 Regression Diagnostics

Mechanics: The same as simple regression, except you're stuck with residual plots. But, you look for the same kind of non-linearity trends, spreading trends, etc.

4 Questions Organized by Topic, Practice Questions, and Corrections to Previous Midterms

Here is a list of questions by topic. You don't have to be a genius to see a pattern developing in these categories and what would be expected on your upcoming exam.

4.1 Midterm Practice Questions by Topic

1. Hard questions (in my opinion): Midterm 2009 (Q3e; Q4d)
2. **Stat 101 Stuff**: Midterm 2009 (Q2a; Q3e; Q4a,b,d)
   (a) Testing mean of a population: Midterm 2009 (Q2a; 3e)
   (b) Making confidence intervals for a mean: Midterm 2009 (Q4a)
   (c) Central Limit Theorem: Midterm 2009 (Q4b,d)
3. **Group Mean Comparisons**: Midterm 2008 (Q1), Midterm 2009 (Q4c), Midterm 2010 (Q1a)
   (a) Pooled and Unpooled, Hypothesis Testing: Midterm 2008 (Q1a), Midterm 2009 (Q4c)
   (b) Paired, Hypothesis Testing: Midterm 2008 (Q1b), Midterm 2010 (Q1a)
4. **Simple Linear Regression**: Midterm 2008 (Q2, Q3), Midterm 2009 (Q1, Q2b-d, Q3), Midterm 2010 (Q1b-e;Q2;Q3), Multiple Regression Questions (Q1A1)
   (a) doing regression by hand (calculating the $\hat{\beta}$s by hand): Midterm 2008 (Q3a), Midterm 2009 (Q2b,c,d), Midterm 2010 (Q2P1b;Q3f)
   (b) hypothesis testing of slope/intercept/F-test: Midterm 2008 (Q3b), Midterm 2009 (Q1d, Q2b), Midterm 2010 (Q1c,d;Q3c,d), Multiple Regression Questions (Q1A1a)
   (c) questions specifically dealing with transformations (logs): Midterm 2009 (Q1a,b), Midterm2010 (Q2P2a;Q3)
   (d) prediction, prediction interval of a single observation, interpreting prediction: Midterm 2008 (Q2a,b,f(with logs); Q3c,d,e), Midterm 2009 (Q3b,c,d), Midterm 2010 (Q1d;Q2P1h;Q2P2a)
   (e) prediction interval of difference between two observations: Midterm 2008 (Q2d)
   (f) confidence interval of a prediction: Midterm 2010(Q1P1d)
   (g) confidence interval of slope or difference between two observations: Midterm 2008 (Q2c), Midterm 2009 (Q3a), Midterm 2010 (Q1P1e;Q3b), Multiple Regression Questions (Q2A1b)
   (h) filling in blank ANOVA/Estimate/Summary of Fit tables using previous ANOVA/Estimate tal Midterm 2008 (Q2e), Midterm 2009 (Q1c, Q2d), Midterm 2010 (Q1b; Q2P1a,b;Q3a)

(i) checking regression assumptions and doing diagnostics: Midterm 2008 (Q2g), Midterm 2010 (Q2P1i;Q2P2b)

(j) curveball questions (fundamental understanding of Stat 101 and Mixing it with Regression): Midterm 2009 (Q3e), Midterm 2010 (Q1e;Q2P1f,g;Q3e)

5. **Multiple Linear Regression**: Multiple Regression Questions (Q1A2-4;Q2-3)

   (a) t-test: (Q1A2b)

   (b) full F test, partial F test: (Q1A2a)

   (c) Polynomial Regression: (Q3a-e)

   (d) confidence interval of slope or difference between observations: (Q2d)

   (e) prediction, prediction interval of a single observation: (Q1A3a,b;Q2b)

   (f) comparing multiple ANOVA tables + Stepwise-esque Regression: (Q1A2a; Q1A4a,b)

   (g) filling in missing ANOVA tables: (Q2a)

   (h) general knowledge regression, assumptions about regression: (Q2c)

4.2 Practice Questions (a work in progress)

Based on looking at what questions transferred from year to year in the practice midterms and what was emphasized in both Professor Brown and Zhao's lectures, here is what I predict to be on your next midterm exam

1. Questions on analyzing ANOVA tables from two or more analysis.

2. Suppose we have the following print-out from JMP

4.3 Correction to Midterms and Practice Questions

1. **Mult Regression Questions**

   (a) (Thanks to Eliza and Jennie): Q1, Analysis II, part a, the F should have the second degrees of freedom be 106, not 108

   (b) (Thanks to Adam, Nicole, Eric, Lauren, and Zhang): Q1, Analysis II, part b, the t-ratio that they substitute should be -2.44, not -0.304.

   (c) (Thanks to Eliza): Q2e, the prediction should have $(20 - 27.0754)^2$ in the expression, not 20^2

2. **Midterm 2010**

   (a) (Thanks to Akshay): Q1a can either be interpreted as a one-sided test or a two-sided test. We will make this clear on the exam.

(b) (Thanks to Liz): Q1c. The t-statistic is 0.79, not 0.74.

(c) (Thanks to Eliza and Jennie): Q2, part 1, part d. The unnecessarily derivation is not required for full credit.

(d) (Thanks to Adam): Q2c and b. The point estimate is wrong. It should be 528.5, not 478.5

(e) (Thanks to Helen): Q2d. The $(n-1)$ term in the denominator of S_p should be 93, not 96.

(f) (Thanks to Shikha and Robert): Q3e. The p-value is incorrect. The p-value should be 0.036, not 0.0034.

3. **Midterm 2009**

(a)