Hello class!

This is perhaps my finest work I have done for Stat 101/Stat 102. I say it's the finest because **one**, I have done this on Microsoft Word (and oh, the unbearable pain to type math equations on Word!) and **two**, I have actually PROOFREAD the guide to make it more pleasurable for you to read. The work took me 20 hours STRAIGHT (from April 27 around 12:00PM and to 7:30AM April 28…in three hours, I have a review session to give…). The point (and yes, I like to be obnoxious): please thank me at some point in your lives (i.e. when you become a CEO of FORTUNE 500 company and decide that donating your company profits to your past TA can be a LEGAL way to not pay taxes). Food is my preferred method of donation at this point in time and will probably remain that way for a long time (with 95% confidence)

Now, on to business. This is a nine-page final review guide that summarizes roughly 14 weeks of Stat 102. It is meant to be complete, but not comprehensive. That is, if you haven't read much of the lecture slides/attended lectures/etc. this review guide may not be for you as some of the concepts discussed may be foreign. I suggest going back to the lecture slides to learn the material AND then coming back to this guide. For those who attended lectures/read the lecture slides/did the homework independently, this final review guide will be equivalent to 93 Christmas presents as I have made every attempt to consolidate and distill material for your reading/reviewing/copying on to your cheat sheet pleasure.

As always, please send me an e-mail if you see any errors so I can correct them ASAP AND check webCafe regularly for updates to this document. Special thanks to all my students during office hours for making corrections/comments (or lack of) to the rough draft of this. Another gigantic thanks to your professors, Professor Brown and Professor Zhao, for making, I would say, very clear lecture slides for your TAs to read and distill from. Finally, thanks to all my students in the last two semesters for a terrific, wonderful year (and the $500 award too). I will certainly not forget the times you fed me (or thought about feeding me at some point).

Yours truly,

Hyunseung

Updates:

April 28, 2011@10:18PM: Made correction to false positive and false negative in ROC curve.

May 1, 2011 @ 12:08PM: Added extra conditions for prediction intervals in time series and regression models.

May 3, 2011 @ 5:15PM: Changed typos on degrees of freedom for F and minor spelling typos

May 4,2011 @ 3:22PM: Changed correction to normality and trade-off between specificity and sensitivity (Shikha)

May 5, 2011 @ 8:07PM: FINAL REVIEW GUIDE WITH ALL THE UPDATES. Changed Tukey Kramer and explained in detail the CI/PI chart. Good luck!

July 5, 2013: Updated some errors with the degrees of freedom and the Anova table.

REGRESSION:

Model Description and Assumptions

Suppose we collect Y and $X_1,\dots,X_p$. We are interested in building a model that explains the relationship between Y and $X_1,\dots,X_p$. Depending on what type of variable Y and X are (continuous or discrete/categorical), we can build different models listed below. Subscript k denotes the kth observation.

Model Name	**Model Formulation**	**Type of X**
Simple linear regression (SR)	$Y_k = \beta_0 + \beta_1 X_{1,k} + \varepsilon_k$	One X ($X_{1,k}$ represents the kth person's X_1): Continuous X
Multiple linear regression (MR)	$Y_k = \beta_0 + \beta_1 X_{1,k} + \dots + \beta_p X_{p,k} + \varepsilon_k$	p Xs ($X_{j,k}$ represents the kth person's X_j) Continuous Xs
One-Way ANOVA	$Y_{i,k} = \mu + \alpha_i + \varepsilon_{i,k}$	One group (α_i is the effect of the ith factor): Categorical Xs coded as α (e.g. α_i = ith factor in category X)
Two-Way ANOVA	$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \varepsilon_{i,j,k}$	Two groups (β_j is the effect of the jth factor) Categorical Xs coded as α, β
Two-Way ANOVA with Interactions	$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \Upsilon_{i,j} + \varepsilon_{i,j,k}$	Two groups ($\Upsilon_{i,j}$ is the interaction between ith and jth factor) Categorical Xs coded as α, β, γ
ANCOVA*	$Y_{i,j,k} = \beta_0 + \beta_1 X_{1,i} + \dots + \beta_p X_{p,i} +$ $\mu + \alpha_i + \beta_j + \Upsilon_{i,j} +$ $\beta_{\alpha i} X_{1,i} + \dots + \beta_{\alpha i} X_{p,i} +$ $\beta_{\beta j} X_{1,i} + \dots + \beta_{\beta j} X_{p,i} + \varepsilon_{i,j,k}$	One or two groups AND p Xs: Categorical and Continuous Xs
Logistic Regression (LR)	$P(Y_k = 0 \mid X_{1,k}, \dots, X_{p,k}) = \frac{e^{\theta}}{1 + e^{\theta}}$	Y must be discrete/categorical (1 or 0) θ can be any model formulation above. For example, $\theta = \beta_0 + \beta_1 X_{1,i} + \dots + \beta_p X_{p,i} + \mu + \alpha_i + \beta_j + \Upsilon_{i,j} + \beta_{\alpha i} X_{1,i} + \dots + \beta_{\alpha i} X_{p,i} + \beta_{\beta j} X_{1,i} + \dots + \beta_{\beta j} X_{p,I}$ (aka ANCOVA)

*Note 1: β_i is the slope WITHOUT interaction and $\beta_{\alpha i}$ or $\beta_{\beta j}$ are slopes WITH interaction. For example, $\beta_{\alpha i}$ is represents the interaction between α_i and the corresponding X next to $\beta_{\alpha i}$

*Note2: Polynomial regression is MR except there is generally one X_1, along with its squares/cubic/etc. (X_1^2 or X_1^3)

The unknown parameters in all models are fit by using a **least square method**, except logistic regression where it uses a **maximum likelihood** algorithm. Through these methods, we get estimates of our unknown parameters in our model.

Parameter Estimates

The estimates are presented to you via the Parameter Estimates table (for simple and multiple regression) or the Expanded Parameter Estimates table (for one-way, two-way, and ANCOVA). Both tables are identical except for a few things.

Term	**Estimate**	**Std Error**	**t-Ratio OR ChiSquare (LR)**	**Prob > |t| = p-value** **Prob > Chisq = p-value**
Intercept	SR: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ One/Two Way: $\hat{\mu}$ Others: $\hat{\beta}_0$	SR: $SE(\hat{\beta}_0) = \widehat{\sigma_e}\sqrt{\left\{\frac{1}{n} + \frac{\bar{x}}{(n-1)s_x^2}\right\}}$ One/Two Way: $SE(\hat{\mu})$ Others: $SE(\hat{\beta}_0)$	SR: $t_{slope}^2 = F_{full} = F_{reduced}$ MR: $t_{slope}^2 = F_{reduced}$ ALL (except LR):	Testing Framework: $H_0: term = 0$ $H_a: term \neq 0$ *(controlling for other terms)* *(DF of critical t = DFE_{full})* SR: p-value of t_{slope} = p-value of F_{full}
Slope for X_j with and without interaction	SR: $\hat{\beta}_1 = corr\frac{s_y}{s_x}$ Others: $\hat{\beta}_j$ or $\hat{\beta}_{\alpha i}$ or $\hat{\beta}_{\beta j}$	SR: SE($\hat{\beta}_1$) = $\hat{\sigma}_e\sqrt{\frac{1}{(n-1)s_x^2}}$ Others: $SE(\hat{\beta}_j)$ or $SE(\hat{\beta}_{\alpha i})$ or $SE(\hat{\beta}_{\beta j})$		
α or β or γ	$\hat{\alpha}_i$ or $\hat{\beta}_j$ or $\hat{\gamma}_{i,j}$	$SE(\hat{\alpha}_i)$ or $SE(\hat{\beta}_j)$ or $SE(\hat{\gamma}_{i,j})$		

			$t = \frac{Estimate}{Std\ Error}$	MR, and ANCOVA (only the slopes without interaction): p-value of t_{slope} = p-value of $F_{-source}$

*Note 1: In One-Way, Two-Way, and ANCOVA, all α, β, γ, and the interaction terms with the numerical Xs, β_α and β_β, have to add up to zero (e.g $\sum \hat{\alpha}_i = 0, \sum\hat{\beta}_j = 0, \sum\hat{\gamma}_{i,j} = 0, \sum\hat{\beta}_{\alpha i} = 0$, and $\sum\hat{\beta}_{\beta j} = 0$).

*Note 2: Under a balanced design (aka when there are same number of people in factor per group), adding the interaction term, $\hat{\gamma}_{i,j}$, to your two-way model will not change the estimates of $\hat{\alpha}_i$ or $\hat{\beta}_j$

*Note 3: LR, Logistic Regression, tests using the Wald Test

Each term in our estimates table have certain interpretations. A slope term in simple and multiple linear regressions is interpreted as an increase in one unit of X leads to $\hat{\beta}_j$ change in Y, controlling for others. The α or β or γ term in ANOVA/ANCOVAs is interpreted to be the effect of some factor in a group. For example, α_i is the effect of factor i in one group and $\gamma_{i,j}$ is the interactive effect of factor i and j from two different groups. Finally, in ANCOVA, $\hat{\beta}_{\alpha i}$ and $\hat{\beta}_{\beta j}$ have interpretations that combines that of regression and ANOVA. For example, $\hat{\beta}_{\alpha i}$ is interpreted an effect of the i^{th} factor on the X where the effect is defined to be an increase in one unit of X leads to a $\hat{\beta}_{\alpha i}$ change in Y.

Model Testing and Analysis

Once we have estimates for our unknown parameters for our model, statisticians want to test a variety of things about our model. In particular, here are a couple of useful things to test.

1. **Is the entire model we proposed useful**?
   ANS: USE Full F-Test in Full ANOVA Table

Source	**DF**	**Sum of Squares OR -LogLikelihood (LR)**	**Mean Squares OR Not Applicable (LR)**	**F-Ratio** $= \boldsymbol{F_{full}}$ **OR ChiSquare (LR)**
Model OR Difference (LR)	$DFR_{full} =$ Simple and Multiple: p One-Way ANOVA*: I – 1, I = # of factors Two-Way ANOVA*: (I – 1) + (J -1) ANCOVA: ANOVA DF + Multiple Reg DF + (# of factors in interaction -1) LR: Same as what $\theta$is	All but LR: $SSM_{full} = \sum(\bar{Y} - \hat{Y}_k)^2$ LR: $DIFF$	$MSR_{full} = \frac{SSM_{full}}{DFR_{full}}$	All but LR: $F_{full} = \frac{MSR_{full}}{MSE_{full}}$ LR: $2 * DIFF$
Error OR Full (LR)	All but LR: $DFE_{full} = DFT - DFR_{full}$ Alternatively, to figure out the DFE_{full} you can just count the number of parameters being estimated.	All but LR: $SSE_{full} = \sum(Y_k - \hat{Y}_k)^2$ LR: $FULL$	$MSE_{full} = \frac{SSE_{full}}{DFE_{full}}$	**Prob > F = p-value OR** **Prob > ChiSq (LR)** H_0: all parameters = 0 (except intercept) H_a: at least one parameter $\neq 0$ (DF of critical F = DFR_{full}, DFE_{full})
Total OR Reduced (LR)	All but LR: DFT = n-1	All but LR: $SST = \sum(Y_k - \bar{Y})^2$ $= SSE_{full} + SSM_{full}$ LR: $REDUCED = DIFF + FULL$		

*Note 1: I = # of factors in one group and J = # of factors in second group

*Note 2: LR uses a Likelihood Ratio Test

*Note 3: the subscript *full* represents the MSE/MSR/DFE/etc. for a full model with all the covariates included.

2. **Which part of the model is most useful? Is each term in our model useful**?

ANS: USE Partial F-Test in Effects Table or Parameter Estimates Table

Source	**DF = All the DFs add up to DFR_{full}**	**Sum of Squares (SS)**	**F-Ratio = $F_{-source}$ OR L-R ChiSquare (LR)**	**Prob > F = p-value OR Prob > ChiSq (LR)**
X OR group (with all the factors)	Simple and Multiple: 1 One-Way or Two-Way ANOVA: (# of factors -1) ANOVA with Interaction: (# of factors -1) ANCOVA*: (I*J-I-J+1) LR: DF of chosen θ	One/Two Way (balanced): $SS = SSR_{source}$ One/Two/Interaction (balanced): All SSs have to add up to SSR_{full} ALL: $SS = SSE_{-source} - SSE_{full}$	All but LR: $F_{-source} = \frac{\frac{SS}{DF}}{MSE_{full}}$ LR: $\chi^2 = 2 * SS$	H_0: all parameters associated with this source is zero H_a: at least one of the parameter $\neq 0$ (DF of critical F = DF,DFE$_{full}$)

*Note 1: Remember this one by using the fact the DFs have to add up to the DFR

*Note 2: Notation $SSE_{-source}$ is the SSE of a model without (aka minus or '-') that source/X/group and SSE_{source} is the SSE of the model ONLY with that source/X/group

*Note 3: LR uses a Likelihood Ratio Test, but the values may be different between the full model ANOVA and it.

ONE-SIDED TEST or Non-Zero Two-sided Test: If we want to test whether there is a one-sided difference (aka $H_0: parameter = \mu_0 \; vs. H_a: parameter < \mu_0$, where μ_0 can be any value (most of the time 0)) OR if we want to test $H_0: parameter = \mu_0 \; vs. H_a: parameter \neq \mu_0$, (where μ_0 is something but zero), we can only do so on parameters listed on the Parameter Estimates table. The t-statistic is $t_{obs} = \frac{Estimate - \mu_0}{SE(Estimate)}$, the critical t value's degrees of freedom is the same as that listed in on this review guide, and p-value is $P(t_{obs} < t_{crit})$. Note that the direction in the probability expression is the same as that in the alternative hypothesis.

FOR SLOPE TERMS WITHOUT INTERACTION ONLY: When you use the parameter estimates table, notice that the hypothesis for the p-value is the same as the hypothesis for the partial F in the Effects table (aka the p-value is the same)

FOR TESTING PRESENCE OF INTERACTION: In addition to the Effects Table, you can test this graphically by looking at interactive plots. Basically, if the lines on the interactive plots intersect, then there is reason to believe an interaction term is significant (aka the p-value should be low). Each point on the line represents $\hat{y}$ given certain Xs/groups/factors.

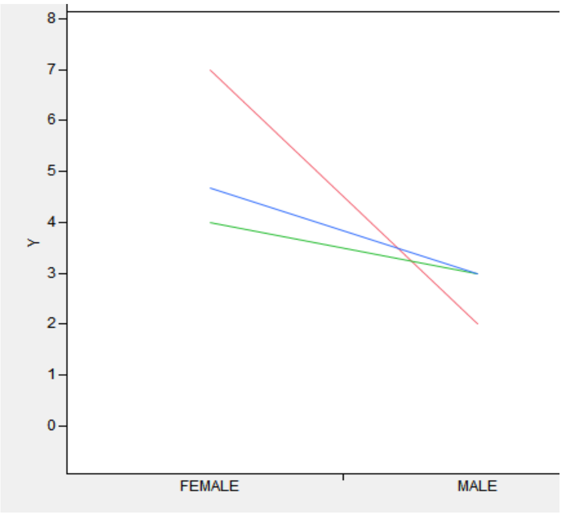


Figure 1: Interaction plot for Two-Way ANOVA. Green = Frosh, Red = Soph, Blue= Juniors

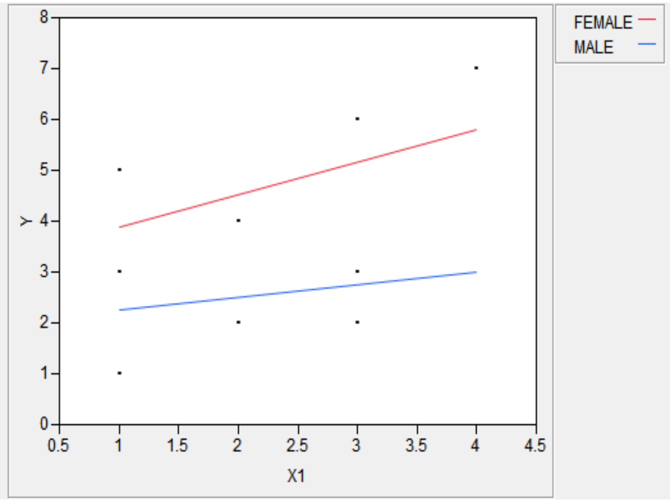


Figure 2: Interaction plot for ANCOVA where the X-axis is continuous.

3. **(ANOVA only) Based on this model, are there any differences between groups? If there are, which group is better? Controlling for other variables, are certain groups better than others?**

ANS: USE confidence intervals/Effects Table/Full ANOVA and Tukey/Kramer

Depending on the number of differences you are testing and whether you are controlling for other variables, you have a different set of hypothesis. However, answering this question all leads to looking at the same tables/confidence intervals; they just have different interpretations

a. If you want to test whether a specific, SINGLE pair of groups is different from each other, you are testing one single hypothesis and thus, you need to use Student t-interval

b. If you want to test whether MULTILPE pairs of groups are different from each other (e.g. you want to know whether one group is better than the rest, which requires you to test that group vs. all the other groups), you are testing multiple hypothesis and thus, you need to use Tukey/Kramer

c. If you are testing for (a) or (b), but controlling for some other variable (e.g. different categorical variable or continuous X), then depending on how many pairs you want to compare, you either choose Student t (one pair) or Tukey-Kramer (multiple pairs)

d. If you are testing (a),(b), or (c), but doing a one-sided test instead of a two-sided test (e.g. $H_0: \mu_A = \mu_B\ vs. H_a: \mu_A < \mu_B$, you have to construct your own t-statistic and obtain the critical value. The t-statistic is $t_{obs} = \frac{\bar{y}_A - \bar{y}_B}{\hat{\sigma}_e\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}}$, the critical t value's degrees of freedom is DFE_{full}, and the p-value is $P(t_{obs} < t_{crit})$.

Note that the direction in the probability expression is the same as that in the alternative hypothesis.

Level	-Level	Difference	Lower CL	Upper CL	p-value
Factor A	Factor B	$\bar{y}_A - \bar{y}_B$	Student t interval (single hypo.) $\bar{y}_A - \bar{y}_B \pm t_{DFE_{full},1-\frac{\alpha}{2}}\hat{\sigma}_e\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}$ Tukey/Kramer (multiple hypo.) $\bar{y}_A - \bar{y}_B \pm q^*\hat{\sigma}_e\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}$ *Formulas do not apply if you have case c.		Single Hypothesis: (type I error control) $H_0: \mu_A = \mu_B$ $H_a: \mu_A \neq \mu_B$ (if the confidence interval does not contain zero, reject null) Multiple Hypothesis: (FWER control) $H_0: \mu_A = \mu_B$ $H_a: \mu_A \neq \mu_B$ (technically, you have a family of the hypothesis above)

*Note: FWER = Family-wise Error Rate (aka you are controlling for type I error for many hypothesis at once)

You can also use the familiar “Letter” tests. If the levels are not connected by the same letter, they are different.

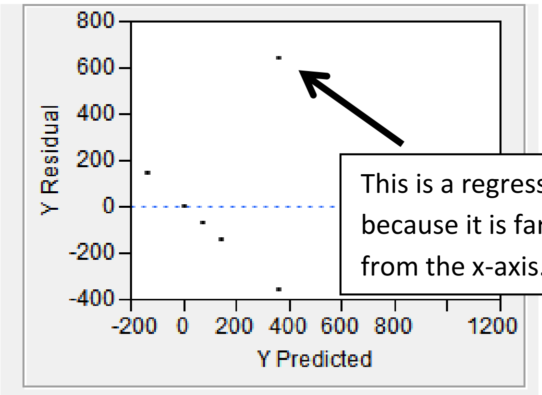
Level				**Mean**
Frosh	A			$\bar{y}_{Frosh}$
Soph	A	B		$\bar{y}_{Soph}$
Junior		B		$\bar{y}_{Junior}$
Senior			C	$\bar{y}_{Senior}$

Example: Here, Soph is statistically similar to Frosh and Junior. But, Soph is different from Seniors

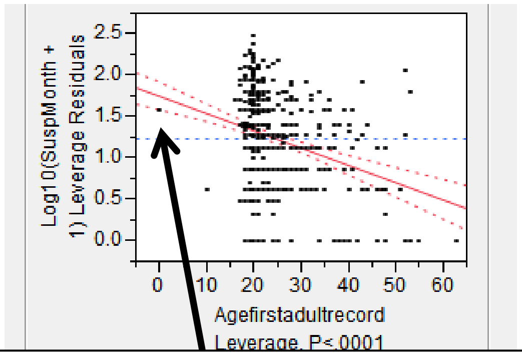
4. **(For Every Model Except Logistic Regression) Does the model follow the three assumptions?**
All the models above are built under the assumption that (i) the model is **linear** (ii) the errors are **independent and identically distributed** (aka **homoscedastic**) (iii) the errors are **normally distributed**. You can test these model assumptions visually by using the scatterplot, residual plots, and normal quantile (aka QQ) plots.
Linearity:
   a. SIMPLE LINEAR REGRESSION: Use scatterplot of X and Y and see if it looks ‘curvy’
   b. FOR ALL MODELS: Use residual plot and check to see whether there are ‘curvy’ patterns along the horizontal x-axis.
   c. IF THERE IS A PROBLEM WITH LINEARITY: Transform Xs. Fix this after heteroscedasticity

   IID errors (in particular, homoscedasticity):
   a. SIMPLE LINEAR REGRESSION: Use scatterplot of X and Y and see if the data points spread/shrinks along the fitted line as X increases
   b. ONE-WAY ANOVA: Look at the scatterplot of each factor and Y and check whether the points spread equally across factors.
   c. FOR ALL MODELS: Use residual plot and check to see whether the spreading around the horizontal x-axis is roughly identical across the x-axis.
   d. IF THERE IS A PROBLEM WITH HOMSCEDASTICITY: Transform Ys. Fix this first

   Normally distributed errors:
   a. FOR ALL MODELS: look at the qq plot and check to see if points are inside the ‘butterfly’
   b. IF THERE IS A PROBLEM WITH NORMALITY: Transform Ys. Fix this after linearity.
5. **(For Every Model Except Logistic Regression) Are there outliers/influential/leverage points in your data set**?
<u>ANS</u>: Use scatterplots, residual plots, or leverage plots
Regression Outlier: Must be “far away” from the fitted line (in SR) or “far away” from the x-axis in residual plots where “far away” is defined by some constant times RMSE (i.e. $\hat{\sigma}_e$) away from the line/x-axis
Leverage points: Must be an outlier in the Xs.
Influential point: Must be a leverage point AND if the point is taken out, the slope changes dramatically. Generally this means that the point is not on the ‘line’, but it is very close to it.



This is a regression outlier because it is far away from the x-axis.



This is a potential influential point. If this were on the red line, then it would not be influential, but a leverage point

6. **Are your Xs (continuous) collinear**?
<u>ANS</u>: Use correlation between Xs.

If there are lots of Xs that are highly collinear in your regression, the estimates of the slope, the p-value associated with it, and the interpretation may be difficult. To avoid collinearity, you check to see whether the Xs are correlated with each other and exclude all but one variable.

7. **Which model is the best model**?
   ANS: Use Stepwise Regression (forward/backward/all-subsets).
   Stepwise regression picks the 'best' (and usually this means a smaller) model given a set of Xs. The way JMP chooses which Xs are 'best' and should be included is roughly based on partial F-tests.

Prediction and Confidence Intervals

Once we tested our model to our satisfaction, we want to use the model to make predictions and create confidence/prediction intervals for our estimated terms or for a new $\hat{y}$. To help us make this, we need a couple information from JMP.

Summary of Fit (All Models)	
RSquare OR **RSquare U (LR)**	SR: $R^2 = corr^2$ All except LR: $R^2 = \frac{SSR_{full}}{SST_{full}} = 1 - \frac{SSE_{full}}{SST_{full}}$
Root Mean Square Error	All except LR: $RMSE = \hat{\sigma}_e$ $= \sqrt{MSE_{full}}$
Mean of Response	$\bar{y}$
Observations	n

Means for One-Way ANOVA	
Level	Factor A
Number	n_A= sample size of factor A
Mean	$\bar{y}_A$ = mean of factor A
Std Error	$\frac{\hat{\sigma}_e}{\sqrt{n_A}}$
Lower and Upper 95%	$\bar{y}_A \pm t_{DFE_{full,1-\frac{\alpha}{2}}} \frac{\hat{\sigma}_e}{\sqrt{n_A}}$

Point Estimates: Given values of Xs, what is the best estimate for Y? OR What is the difference between these two Ys?

This is a basic plug-and-chug exercise. Plug in your values for X into the model we fitted and get the estimate $\hat{y}$. For example, in ANCOVA, suppose we are measuring the relationship between GPA (Y) and gender ($\alpha_{male}, \alpha_{female}$), male and female, and Age (X_1), and interaction between Age and gender and we build a model of the following type: $\hat{y}_k = \hat{\beta}_0 + \hat{\alpha}_i + \hat{\beta}_1 X_{1,k} + \hat{\beta}_{\alpha i} X_{1,k}$. The estimated GPA for a female student who is 19 year old is $\hat{y} = \hat{\beta}_0 + \alpha_{female} + \hat{\beta}_1(19) + \hat{\beta}_{female}(19)$. Notice that we DO NOT add both α_{male} and α_{female}.

Confidence Intervals(CI) /Prediction Intervals(PI): All confidence/prediction intervals (except Tukey-Kramers) have the following structure (note that Δ replaces DFE_{full} when we do unpooled confidence intervals):

$$\text{point estimate } \pm t_{DFE_{full} or\ \Delta, 1-\frac{\alpha}{2}} SE(point\ est.)$$

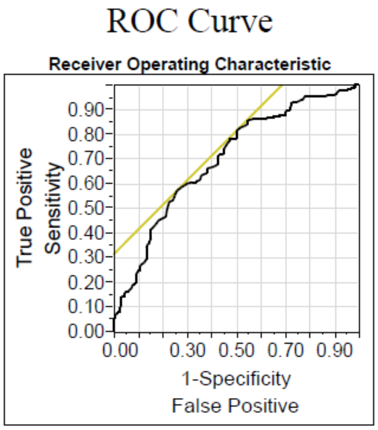
Since the point estimate is generally easy to obtain, the trick is finding the SE, the standard error. The decision tree below should help you find the correct SE. If you face a problem in the final that does not follow the decision tree below, the next best (and the method that ALWAYS works) is to actually compute $SE(point\ est.) = \sqrt{Var(point\ est.)}$ using Stat 101's Expectation/Variance formulas (e.g. sum of variances is the variance of the sums if independent, etc.). For example, Var(X – Y) = Var(X) + Var(Y), if X and Y are independent. Generally, this formula is used to obtain the SE for differences between two groups.

(LR ONLY) Classification Rules and ROC Curves: In logistic regression, our point estimate is a probability. Suppose, based on this probability you want to assign 1/0s, like the actual Ys. The most natural way to do this is assign a cut-off value of 0.5 and say that Y = 0 when $P(Y_k = 0 \mid X_{1,k}, \ldots, X_{p,k})$ > 0.5 and 1 otherwise. However, depending on the problem type,

we may want different cut-off values. To help us better pick our cut-off values, there are a couple of terms and tools we should know.

	$\hat{Y} = 0$ (Predicted Ys)	$\hat{Y} = 1$ (Predicted Ys)
Y = 0 (Actual Y values)	Sensitivity =True Positive Rate = $\frac{\# \, of \, \widehat{Ys} = 0}{\# \, Ys = 0}$	1 – Sensitivity = False Negative Rate
Y =1 (Actual Y values)	1 – Specificity = False Positive Rate	Specificity = True Negative Rate = $\frac{\# \, of \, \hat{Y}s=1}{\# \, of \, Ys=1}$

There is always a trade-off between sensitivity and specificity. For example, if we choose a cut-off such that all of our prediction would be assigned as $\hat{Y} = 1$, then our specificity will increase, but our sensitivity will decrease. To describe this trade-off, there is a curve known as the ROC curve

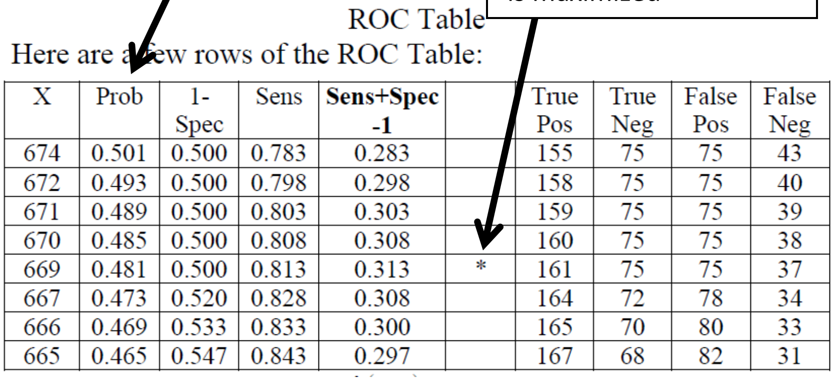


Using PASTDUE='0' to be the positive level

Area Under Curve = 0.69823

This is the cut-off value

This is the point where Sensitivity + Specificity is maximized



ROC Table

Here are a few rows of the ROC Table:

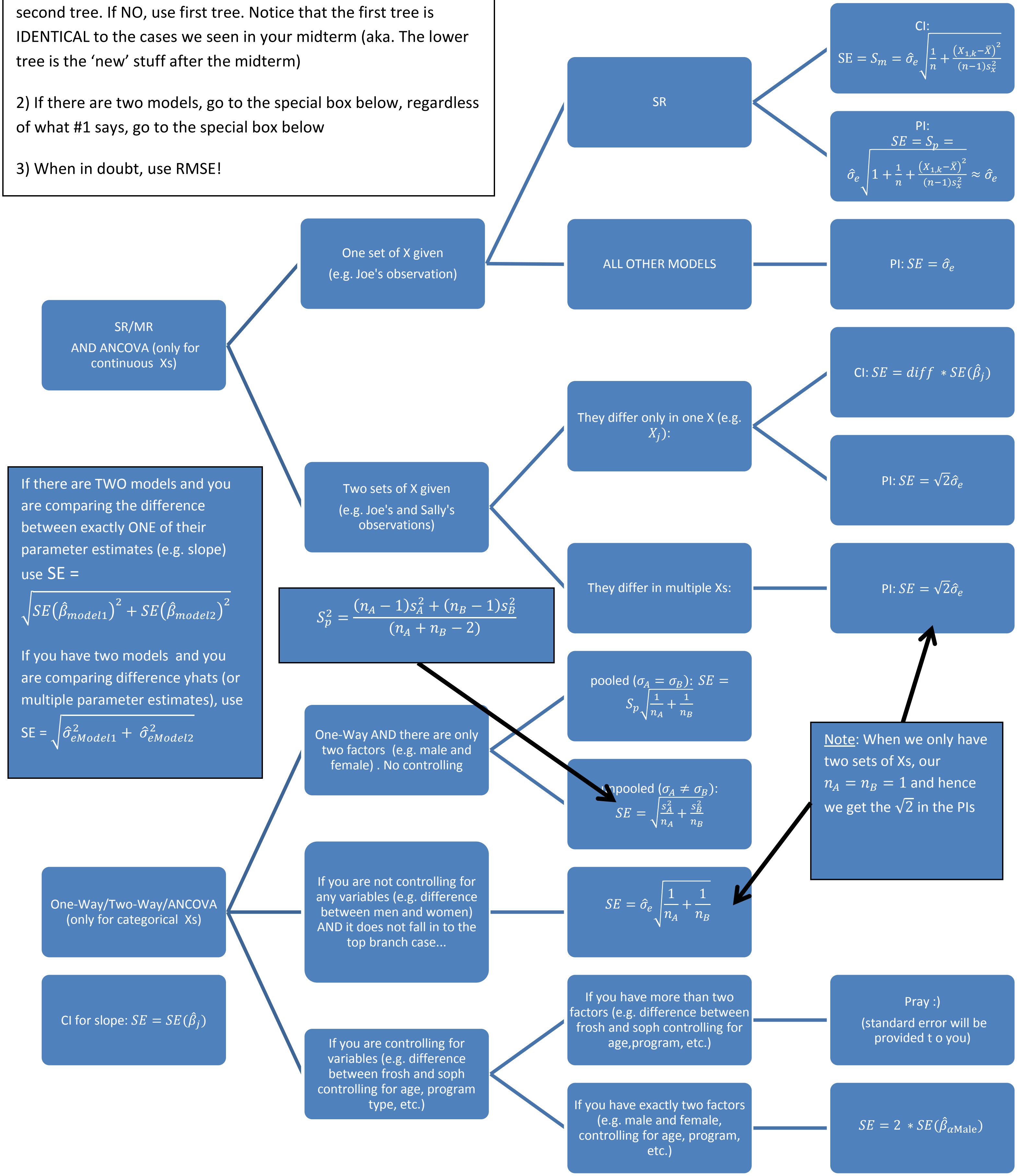
X	Prob	1-Spec	Sens	**Sens+Spec -1**		True Pos	True Neg	False Pos	False Neg
674	0.501	0.500	0.783	0.283		155	75	75	43
672	0.493	0.500	0.798	0.298		158	75	75	40
671	0.489	0.500	0.803	0.303		159	75	75	39
670	0.485	0.500	0.808	0.308		160	75	75	38
669	0.481	0.500	0.813	0.313	*	161	75	75	37
667	0.473	0.520	0.828	0.308		164	72	78	34
666	0.469	0.533	0.833	0.300		165	70	80	33
665	0.465	0.547	0.843	0.297		167	68	82	31

General Rule of Thumb for Identifying SEs:

1) Is the point estimate comparing two groups? If YES, use the second tree. If NO, use first tree. Notice that the first tree is IDENTICAL to the cases we seen in your midterm (aka. The lower tree is the 'new' stuff after the midterm)

2) If there are two models, go to the special box below, regardless of what #1 says, go to the special box below

3) When in doubt, use RMSE!



SR/MR AND ANCOVA (only for continuous Xs)

One set of X given (e.g. Joe's observation)

SR

CI: $\mathrm{SE} = S_m = \hat{\sigma}_e \sqrt{\frac{1}{n} + \frac{\left(X_{1,k} - \bar{X}\right)^2}{(n-1)s_x^2}}$

PI: $SE = S_p = \hat{\sigma}_e \sqrt{1 + \frac{1}{n} + \frac{\left(X_{1,k} - \bar{X}\right)^2}{(n-1)s_x^2}} \approx \hat{\sigma}_e$

ALL OTHER MODELS

PI: $SE = \hat{\sigma}_e$

Two sets of X given (e.g. Joe's and Sally's observations)

They differ only in one X (e.g. X_j):

CI: $SE = diff * SE(\hat{\beta}_j)$

PI: $SE = \sqrt{2}\hat{\sigma}_e$

They differ in multiple Xs:

PI: $SE = \sqrt{2}\hat{\sigma}_e$

Note: When we only have two sets of Xs, our $n_A = n_B = 1$ and hence we get the $\sqrt{2}$ in the PIs

If there are TWO models and you are comparing the difference between exactly ONE of their parameter estimates (e.g. slope) use SE = $\sqrt{SE\left(\hat{\beta}_{model1}\right)^2 + SE\left(\hat{\beta}_{model2}\right)^2}$

If you have two models and you are comparing difference yhats (or multiple parameter estimates), use SE = $\sqrt{\hat{\sigma}_{eModel1}^2 + \hat{\sigma}_{eModel2}^2}$

$$S_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{(n_A + n_B - 2)}$$

One-Way/Two-Way/ANCOVA (only for categorical Xs)

One-Way AND there are only two factors (e.g. male and female) . No controlling

pooled ($\sigma_A = \sigma_B$): $SE = S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$

unpooled ($\sigma_A \neq \sigma_B$): $SE = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$

If you are not controlling for any variables (e.g. difference between men and women) AND it does not fall in to the top branch case...

$SE = \hat{\sigma}_e\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$

If you are controlling for variables (e.g. difference between frosh and soph controlling for age, program type, etc.)

If you have more than two factors (e.g. difference between frosh and soph controlling for age,program, etc.)

Pray :)

(standard error will be provided t o you)

If you have exactly two factors (e.g. male and female, controlling for age, program, etc.)

$SE = 2 * SE(\hat{\beta}_{\alpha\mathrm{Male}})$

CI for slope: $SE = SE(\hat{\beta}_j)$

TIME SERIES (SLIGHT DEPARTURE FROM REGRESSION):

Model Description

Time series is probably the most intuitive thing to understand in this course. We measure something in Y in time and we want to build a model that would allow us to forecast into the future. When we look at a time-series plot, we have to be mindful of three things:

1. Trend {T(t)}: Is there a linear/quadratic/cubic/etc trend?
   For example, the time series could exhibit a quadratic trend, $T(t) = \beta_0 + \beta_1 t + \beta_2 t^2$
2. Seasonality {S(t)}: Are there any seasonal patterns in the time series data (e.g. cyclical deviations in temperature).
   S(t) is generally a categorical variable and it's very similar to One-Way ANOVA's α_i. For example, if we think there is a summer effect and a winter effect, then $S(t) = S_{summer}\ or\ S_{winter}$ (aka α_{summer} or α_{winter})
3. Lag{E(t)}: Is there a correlation between the past observation and the current observation? If so, how much?
   In almost all cases, the previous observation, Y_{t-1}, is highly correlated with the present observation, Y_t, and we must take this into account. This is generally represented by a lag term, $E(t) = \beta_1 Y_{t-1}$ or more generally, $E(t) = Lag(Y_t, 1)$. If we believe the current observation is highly correlated with more than one, say Y_{t-2} and Y_{t-1}, then $E(t) = Lag(Y_t, 2)$. These E(t) terms are called AR(1) and AR(2), respectively.

Depending on the presence/absence of these factors in a time-series plot, we create the following model

$$Y_t = T(t) + S(t) + E(t) \qquad \text{e.g. } Y_t = \beta_0 + \beta_1 t + S_{summer/winter} + \beta_2 Y_{t-1} + \epsilon_t$$

where ϵ_t is i.i.d. normally distributed (aka assumption 2 and 3 of regression). As always, we find the unknown parameter values by the Least Squares Method:

Parameter Estimates and Testing:

Since we use the same method as a regression (except LR) to get our analysis, our parameter estimates, Full ANOVA Table, and Effects Table, all have the same meaning/interpretation/testing as ANCOVA (SWEET!). One key thing to note, though, is that if we only have E(t) in our model and $E(t) = Lag(Y_t, 1)$, then $R = \hat{\rho} = corr(Y_t, Y_{t-1}) \approx \hat{\beta}_1$ where ρ is called the autocorrelation term. This means that when we are testing whether $H_0: \beta_1 = 0\ vs.\ H_a: \beta_1 \neq 0$, we are essentially testing $H_0: \rho = 0\ vs.\ H_a: \rho \neq 0$ (or that Y_t are dependent from each other).

Prediction:

There are only two scenarios for time-series prediction. For CI/PI for time series, use RMSE for the standard error portion and the critical t value to have DFE_{full} degrees of freedom.

1. If our model only contains E(t) and $E(t) = Lag(Y_t, 1)$, then the long-run average of Y is $\hat{\mu} = \frac{\hat{\beta}_0}{1-\hat{\beta}_1}$ and the prediction at r steps in the future from time T is $\hat{Y}_{T+r} = \hat{\mu} + \hat{\beta}_1^r\ (Y_T - \hat{\mu})$
2. If our model contains E(t), S(t), and T(t), then the prediction at r steps in the future from time T is $\hat{Y}_{T+r} = T + S(T+R) + E(T+R)$. For example, taking the model example we have above, $\hat{Y}_{T+r} = \hat{\beta}_0 + \hat{\beta}_1(T+r) + S(T+r, season\ at\ time\ T+r) + \hat{\beta}_2 Y_{T+R-1}$. We may or may not have Y_{T+R-1} in our data. If we do not, we would have to estimate Y_{T+R-1} by repeating the procedure and plug in this estimate to the equation. Note that we can repeat the procedure until we find data for our lag term.