# Exam I Review

No notes allowed. You won't need them. You will definitely want to bring a calculator.

As for the overall format/feel of the test: I plan to give you a good bit of R output and plots to ask you about including plots of residuals vs fits, qqplots, partial regression plots, partial residual plots, square-roots of standardized residuals vs. fits, standardized residuals vs. leverage, Cook's distance vs. half-normal quantiles, etc., etc., etc. (see pages 74, 75, 76, 78, 79, 80, 82, 85, 87, 90, 91, 93 of Faraway's book). I will also give you a lot of output from procedures I run in R, such as that from the lm() function, ad.test(), shapiro.test(), anova(), leveneTest(), bartlett.test(), durbinWatsonTest(), pureErrorAnova(), etc. etc. etc.

With the above in mind, below are some things I plan to focus on. I apologize for the redundancy.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Given some data (predictor variable values and response variable values), be able to

1. Write down the X matrix. This is also called the model matrix. There are 1s in the first column and each row contains the observed variable values for an observation. The number of rows equals the number of observations, and the number of columns equals the number of beta values to be predicted.
2. Write down the Y matrix. This is a column vector containing the responses.
3. Write the equation for obtaining the bs: b = (X^T X) ^-1 X^T Y

Given x and y data (predictor and response observations), be able to calculate the mean and standard deviation of the x values, the mean and standard deviation of the y-values, r, and use these things to calculate b1 and b0 for simple linear regression (remember the formulas: b1 = r * s_y / s_x; b0 = y_bar – b1*x_bar).

Be able to describe what a p-value is/does for hypothesis tests in general in your own words. Here are my words: The p-value for any hypothesis test is the chance
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

that you'd see a test statistic as strange or stranger than what you actually saw if the null hypothesis is actually true.

Know and be able to use the formula for r = correlation coefficient in the case of simple linear regression: r = 1/(n-1) * sum (z_x * z_y)

Know that in simple linear regression, if the data are standardized, r = slope of the regression line for the standardized data and the regression line runs right through the origin.

Know the assumptions for simple linear regression model and how to check on them. You should be able to list them. Same for the general linear model.

Given output from the lm() function with R, be able to judge whether certain parameters could be discarded based on their t-test p-values.

Be able to ascertain possible collinearity of predictor variables from a correlation matrix.

Given a linear model and a new data point, be able to sub this point back into the model to obtain a prediction value.

Know the matrix equations for computing b, fits, the vector of residuals, and the hat matrix. I will ask you to write them down: b = (X^T X) ^-1 X^T Y; H = X (X^T X) ^-1 X^T; Xb = hat(Y) = H Y;
e = Y – hat(Y) = Y – HY = Y(I-H)

Know that the hat matrix is idempotent (and know what idempotent means.) H^k = H, k in {1, 2, 3, …}

Given a regression model, be able to calculate a predicted value of the response given values for the predictor variables.

Anything from our unit on matrices is fair game- eigenvalues, eigenvectors, linear independence, positive semidefiniteness (If for all vectors v, if v^T A v is at least 0, A is positive semidefinite), etc.

Know how to interpret CIs for mean responses and the betas, as well as prediction intervals for responses.

Be able to apply the Bonferonni adjustment to obtain families of confidence intervals for multiple parameters/mean responses/etc. For example, if I give you the family confidence level, be able to provide the individual error rate. Know why we perform the Bonferroni adjustment. How to adjust with Bonferroni? Here's an example. If the family confidence level is .90 and you want to make n confidence intervals, the individual confidence level is 1 – (1-.90)/n.

For the general linear model, what are R^2 and R^2_adj? Know their differences, and basic definitions/formulas, practical uses. R^2 = 1 – SSE/SSTO = SSR/SSTO. R^2_adjusted = ???? Why use it instead of R^2???

Be able to interpret output from lm(), anova(), lack-of-fit ANOVA (pureErrorAnova()). Given output for condition numbers and variance inflation factors, be able to locate them and interpret their meanings. This means you will need to know rules of thumb (see your text) for determining abnormal condition numbers and variance inflation factors.

Be able to fill out a lack-or-fit ANOVA table (like in your homework). Definitely be able to get those p-values, too. You can use your calculators: p-value = Fcdf(t.s., "infinity", numerator df, denominator df)

Know what SSE, SSR, and SSTO, SSPE, SSLF are.

Be able to interpret partial regression and partial residual plots.

Know when a transformation of the response might be useful, and some common transformations (like sqrt(), Box-Cox, log, etc.).

What do shapiro.test() and ad.test() do? What does qqplot() do? (not ggplot… but qqplot) What does levene.test() do?

What does the Durbin-Watson test do?

Be able to look at plots of residuals vs. fits to check for homo/heteroscedasticity and linearity/nonlinearity, etc.

Be able to look at qq-plots (normal probability plots) to assess normality.

Know the definitions of and differences between outliers, leverages, and influential observations, and know how to check for these things.

Explain why it might be good to plot the sqrt(abs(residuals)) vs. fits, and get a regression model.

What's the formula for the H matrix and why do we call it the "hat" matrix? What is the significance of the values along the diagonal and how are they used? What does their sum equal?

What's Cook's distance? What's it used for?

Collinearity- know how to get and interpret the condition numbers, correlations, and VIFs.

Weighted least-squares, generalized least squares, testing for lack of fit, robust regression

Know that "large" p-values for the predictor variables and moderate to large r^2 value can be a sign of collinearity.

Given some graphs, be able to spot leverage points, influential observations, and outliers. Know the differences. Also, know that it is possible for a regression outlier to not present itself as an outlier in each predictor direction by themselves.