This is not an exhaustive list of things to know for the final! It's a collection of stuff I found in previous finals that caught my eye for one reason or another. Maybe I found it difficult compared to the rest, maybe I found it to be a relatively obscure piece of information, or maybe I'm just weird. So caveat emptor.

Final 2015

Question 3e We want to do an F test but only for variables d_1 and d_2, not for the overall regression. Thus we are testing

$$H_0 : \beta_{d1} = \beta_{d2} = 0,$$

$$H_A : \text{at least one of } \beta_{d1}, \beta_{d2} \neq 0.$$

The unrestricted model has $RSS_u = 46491.431$; the restricted model has $RSS_r = 46635.671$; the number of things being estimated in the unrestricted model is $k = 4$; the number of restrictions being tested is $q = 2$; and the sample size of $n = 48$. Therefore the F statistic is

$$F = \frac{(RSS_r - RSS_u)/q}{RSS_u/(n-k)} = \frac{(46635.671 - 46491.431)/2}{46491.431/44} \approx 0.068.$$

Under the null, F here is distributed according to $F_{q,n-k}$. Therefore the critical value we use is $F_{0.05;2,44} = 3.209278$. Our value of F is less than the critical value, which means we fail to reject the null. In other words, the dummies d_1 and d_2 are jointly statistically insignificant at 5% significance.

Multiple Choice 3. Because $d\ln(x)/dx = 1/x$, it follows that

$$\Delta \ln(x) \approx d\ln(x) = \frac{dx}{x} \approx \frac{\Delta x}{x}$$

Multiply both sides by 100 and the RHS becomes the percentage change in x, that is,

$$100 \times \Delta \ln(x) \approx \%\Delta x.$$

We are told that $\%\Delta x = 10$, and therefore $\Delta \ln(x) \approx 10/100 = 0.1$.

Other Random Stuff

Standard Error of Regression with k Estimators. Also known as the **standard error of the residuals** or the **Root MSE (mean squared error)**,

$$s_e = \sqrt{\frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

Note that the sum is really just RSS.

R^2 and Adjusted R^2. We can write R^2 as

$$R^2 = 1 - \frac{RSS}{TSS}.$$

When more regressors are added, TSS is the same but RSS can only decrease. Thus R^2 cannot decrease, even if irrelevant regressors are added. That's a flaw, so use adjusted R^2 instead,

$$\bar{R}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)} = 1 - \frac{s_e^2}{s_y^2} = R^2 - \frac{k-1}{n-k}(1-R^2).$$

Multicollinearity. If we can write one regressor as a linear combination of other regressors, then we have *perfect multicollinearity*. In such a case, we cannot estimate coefficients for each regressor. (The model is not *identified*). This means there's not enough variation in the data, or the model is poorly specified. Examples of linear combinations are $x_1 = 2x_2$ and $x_2 = 1 - x_1 - 5x_3$. The intuition, considering the latter example, is that x_2 contains no additional information that isn't already provided by x_1 and x_3, so it just kind of gets in the way and messes things up.

Multicollinearity occurs when you have high (but not perfect) correlation between two or more regressors. Multicollinearity reduces the precision of the estimate coefficients i.e. bigger standard errors of the coefficient estimates, and makes the estimates very sensitive to minor changes in the model, and we are less likely to get statistically significant results (since larger standard errors imply t statistics closer to zero). We can still do the typical OLS stuff in this case, though.

Population Assumptions for Multivariate OLS Regression.

1) $y = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$ (true population model)
2) $E[u_i | x_{2i}, \ldots, x_{ki}] = 0$ (zero conditional mean)
3) $\text{Var}(u_i | x_{2i}, \ldots, x_{ki}) = \sigma^2$ (homoskedasticity)
4) errors for different observations are independent (yeah)

Assumptions 1 and 2 imply unbiased coefficients and conditional mean

$$E[y_i|x_{2i}, \ldots, x_{ki}] = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_k x_k.$$

Assumptions 3 and 4 in addition imply coefficients are consistent and that we can use the default standard errors. If errors are normally distributed, then inference is exactly $t(n-1)$ distributed. As n blows up to infinity, $t(n-1)$ goes standard normal. In practice, we use $t(n-1)$ as an approximation anyway.

Assumptions 1-4 also imply OLS is the best linear unbiased estimator.

Dummy Variable Trap. The dummy variable trap can be avoided by including all indicator variables but dropping the intercept. This renders R^2 meaningless, however, so we usually omit one category form the regression.

Retransformation Bias. If we do OLS estimation on dependent variable $\ln(y)$, it can give rise to unbiased prediction of $\ln(y)$, assuming assumption 1-2 hold. But it gives a *biased* prediction for y itself. In other words, $\hat{y}_i = e^{\widehat{\ln(y_i)}}$ is biased.