Overview of Regression

Most of the information below is based on materials from Greenlaw (2006) and Gujiarti (2002). Some of this is review from Assignments #3, #4, #7 and other supplements. Please review those assignments before conducting your own regression analysis.

Steps in Regression Analysis

1. *State the hypothesis*
   In the last assignment, you proposed a regression equation. Now, using this regression equation, you should clearly state your hypothesis (or hypotheses) that answer your research question. Remember, a hypothesis has a null and alternative hypothesis. These are based on your economic model.

2. *Test the hypothesis*
   Estimate the regression equation and conduct the hypothesis tests you stated above.

3. *Interpret the regression results*
   - Do the estimated coefficients (e.g., signs on the coefficients) conform with your hypotheses? Are they consistent with the real world and economic theory/your economic model?
     - For example, economic theory predicts that an increase in interest rates reduces the demand for housing, leading to a decrease in the equilibrium price. This means that in a regression equation with housing price as the dependent variable, we'd expect the coefficient on the interest rate to be negative.
     - If your coefficients are the wrong sign, but not statistically significant (one-tail test), then you can use this to defend your results. If the coefficient is not statistically significant (one-tailed test), then we are unsure it differs from 0 (with the wrong sign).
     - If your coefficients are the wrong sign, but are statistically significant (one-tail test), then it gets trickier. Either you have a problem with your regression (see common problems below), have misinterpreted how the economic theory relates to your regression equation, or you've found evidence that the theory is incorrect. Researchers are reluctant to go with the last of these explanations, especially for generally accepted theories, so be careful.

   - Are the coefficient estimates statistically significant (e.g., statistically different from zero)?
     - You can evaluate this by looking at the t-statistics/p-values generated from the regression output.

   - Are the estimates economically significant? This usually is a matter of magnitude. For example, you may find an estimated coefficient is statistically significant, but has a very small effect on the dependent variable.
     - This is more subjective as it is really a matter of whether the magnitude of the coefficient is significant.
     - For example, consider two estimated coefficients for the interest rate (referring to the housing example above): b1 = -0.01 vs. b1= -1.2. Both of these estimates might be statistically significant, but differ in their effect on the dependent variable (housing price). According to the first regression, a one percentage point increase in the interest rate is associated with a $0.01 decrease in housing price. In the second one, a one percentage point increase in the interest rate is associated with a $1.20 decrease in housing price.

   - How much of the variation in your dependent variable does the regression explain?
     - This is evaluated looking at the R-squared of the regression. Specifically, this measures what percent of the variation (changes) in the dependent variable are explained by all of the explanatory variables combined.
     - In cross section data, the R-squared is typically low (10-50%). In time series data, the R-squared is typically much higher (60-99%), especially when the researcher includes a trend variable (as most time series following a long-run trend).

4. *Check for and correct common problems*
   - **Specification error**
     - Omitted variable bias
       If you've excluded an explanatory variable that is relevant for explaining your dependent variable, then the estimated coefficients from your regression are biased. The regression you estimate could mistakenly attribute the effect of the omitted variable to one of your included explanatory variables.

     - Inclusion of irrelevant variables
       This is not as much of a problem as leaving out a relevant variable (see above); explaining why researchers err on the side of caution and include more explanatory variables rather than fewer. If you include variables that are not relevant, it makes it more difficult for you to estimate the effects of each explanatory variable on your dependent variable. Essentially, this irrelevant variables create noise that makes your regression less precise – possible leading to large standard errors and explanatory variables are not statistically significant.

     - Functional form
       It may be that the correct form for your regression equation is nonlinear. To correct for this problem, you may be able to think of a way to transform the data in some way that generates a linear relationship. This is why researchers working with time series often take the natural log of variables in the regression – it presumes that the variable in question grows exponentially. Consider the following Cobb-Douglas production function: $Y = K^{\alpha} L^{\beta}$

       If I estimate the following regression: $Y_t = \beta_0 + \beta_1 K_t + \beta_2 L_t$

       This would lead to specification error because the theory says the relationship is nonlinear. Alternatively, I could take a log transformation of the Cobb-Douglas function:

       $$\ln(Y) = \alpha \ln(K) + \beta \ln(L)$$

       Now, I have a linear relationship that I can estimate using linear regression.

   - **Simultaneous equations bias**
     OLS assumes that the explanatory variables are determined outside of the model – e.g., that they are exogenous. If you include explanatory variables that are endogenous in the model, the coefficient estimated will be biased. A common example of this is demand and supply. Consider the housing price example from above. If we express housing price as a function of the number of houses sold (e.g., include it as an explanatory variable), the coefficient estimate is biased. This is because the quantity of homes sold is a function of the price – that is, it is determined endogenously.

     Correcting for simultaneous equations bias usually involves more advanced regression techniques. If you suspect you have this problem in your regression, simply identify it as a problem and don't give too much weight to the estimated coefficient for this variable in your regression.

   - **Multicollinerity**
     This problem arises when the explanatory variables are highly correlated with one another, making it difficult to disentangle the effects of the individual variables on the dependent variable. This problem usually shows itself by having a high R-squared and few individual explanatory variables that are statistically significant. Note, this would imply that the explanatory variables are collectively important, but not individually important.

Another sign of multicollinearity is that small changes in the data sample lead to large changes in the coefficient estimates. You can check for this problem by looking at the correlation matrix (correlation table we looked at before) for the variables.

This problem doesn't lead to any bias in you estimates, but it can make it difficult for you to test hypotheses. If you have two variables that measure similar things (e.g., two different interest rates), then you may want to drop on of these explanatory variables.

- **Heteroskedasticity**
  This is a problem that arises almost exclusively with cross section data. Heteroskedasticity means "unequal variance". Specifically, it refers to the variance of the residuals from your regression. The variance of the residuals should not be related to the explanatory variables (in a systematic way). A common example is consumer spending (C) and income (Y). If we run the following regression:

  $$C_t = \beta_0 + \beta_1 Y_t$$

  The residuals from this regression are likely to be heteroskedastic. Specifically, the residuals will tend to be more variable as spending and income increase. That is, higher income individuals may spend a large or small portion of their income on consumption (leading to large errors). Lower income individuals tend to vary less in their consumption patterns (primarily purchasing necessities).

  You can diagnose this problem by creating a scatter plot of the squared residuals (genr r2=resid^2) against the explanatory variable (or predicted dependent variable) that is suspected of being the source of the problem. Or, you can use the more formal White test (built into EViews).

  This problem leads to biased standard errors – therefore making hypothesis testing not possible. You can often correct for this problem by transforming the variables in the regression.

- **Autocorrelation**
  This is a problem that arises almost exclusively with time series data. This problem arises when the residuals are dependent on time (e.g., correlated with itself over time). This leads to many of the same problems as heteroskedasticity. Your hypothesis tests will be based on biased standard errors if you do not correct this problem. This is why researchers often include a trend or lagged dependent variable, OR transform the data in some way to avoid this problem.

  Note, autocorrelation could be a sign of omitted variable bias (an explanatory variable that would explain the behavior of the dependent variable over time is excluded, creating a time-dependent error).

  You can diagnose this problem by creating a line graph of the residuals (generated automatically in the EViews regression output). Or, you can use the Durbin-Watson statistic. The rule of thumb is that you want the D-W statistic to be as close to 2 as possible (it ranges between 0 and 4; between 0 and 2 indicates positive autocorrelation and between 2 and 4 indicates negative autocorrelation). You can do a more formal test using built in commands in EViews.

5. *Evaluate the test results*
   - Overall, do the regression results support your hypotheses? Remind the reader about your research question, the implications of the model, and whether your empirical findings support your hypotheses.