Learning to Use Regression Analysis
Studenmund Chapter 3
Regression analysis can be done well or it can be done poorly.

In your own work, if you understand the proper techniques for different situation, the potential difficulties and how to correct for them and the proper interpretation of the results, you may freely choose a type of analysis that suits your intentions.

More importantly, if you understand all these things, you may be able to successfully attack analyses whose results or intentions you oppose.

Steps in Applied Regression Analysis
These steps are important in doing responsible analysis. More importantly (as above) they also provide a good framework for asking questions about analyses you read to determine if the work was done responsibly.

1. Review the literature and develop the theoretical model
The best way to do this is to find some recent papers or articles written on the topic. This will let you know what models and approaches other people have used and what problems they may have encountered.

One good starting point for academic topics is the "EconLit" database (available through the library web site as one of the available databases. It contains articles from numerous social science journals and is a very good starting point for background work. Of course, there are lots of other sources you might use to find articles related to what you want to do.

2. Specify the model: Select the independent variables and the functional form
For example, if you think about the quantity of gasoline demanded in an area, what variables would you include and what specific mathematical relationship should be used? There are a lot of possible options, some of which are better than others.

One option in this case might be

\[ Q = \beta_0 + \beta_1 \text{Price} + \beta_2 \text{Pop} \]

In this case, the coefficient on Price, \(\beta_1\), would be the slope of the estimated demand curve.

Believe it or not, a good model might be
\[ \ln(Q/\text{Pop}) = \beta_0 + \beta_1 \ln\text{Price} \]

In this case, the coefficient on Price, \( \beta_1 \), would be the estimated price elasticity of demand.

Studenmund offers some good guidelines (I believe it’s in Chapter 6) for choosing explanatory or independent variables.

We will talk about various mathematical forms that you might use and what interpretations they carry with them.

3. Hypothesize the expected signs of the coefficients
From the previous gasoline example, what sign should each coefficient have? This is a good way to determine if something is seriously wrong with the analysis. If a coefficient has a sign opposite of what is expected this invites further consideration. Remember, though, that while the analysis may be wrong, it could also be the prior expectations that are incorrect. This always makes for the best studies.
It may well be the case that there is no expected sign for a coefficient. Indeed, the point of the study may be to determine what the likely sign of a coefficient is. Can you think of examples of this?

That is, can you think of situations where you would get some data and estimate a model to see if the impact of an explanatory factor is positive or negative?

4. Collect the data
Hopefully, another person will have been kind enough to collect and clean the data that you’ll be using. However, there may well be occasions in which you’ll have to get your own data.

Use good judgment here. More data is always better, but bad data is a horrible curse. You do, however, need to have more observations than coefficients to be estimated. The difference between these is the number of degrees of freedom in your regression.

Be careful in quantifying variables which may be originally qualitative. What are some examples of variables which are originally qualitative and how could they be used in regression analysis?

Also, be careful about allocating resources between the often opposing tasks of getting a large sample and getting a good random sample.
5. **Estimate and evaluate the equation**
Because you've done a good job collecting the data and using previous work to formulate the model, including the variables to be included and the functional relationship, this should take about five minutes of your time. That's my story and I'm sticking to it.

In reality, as you will all soon find out, there will be about a million possible models that will suggest themselves and you’ll really, really want to run them all (plus a few more) to satisfy your nearly feline level of intellectual curiosity.

The important thing is that, in the end, you can reasonably defend the model whose results you chose to report over other models.

Good defenses for a chosen model include
1. Compared to other possible models, it had good explanatory power as described by the adjusted $R^2$.
2. A large number of the explanatory variables have significant coefficients with the expected signs.
3. Other people have modeled the same phenomenon using this model.

6. **Document the results**
Write clearly, check your spelling, have a friend or colleague read your work before submitting it and get their opinion, etc. What seems very clear to you may be confusing to people less intimately involved in the project.

Use examples supplied by the class to illustrate these steps.

**Two added topics**

1. **Lagged variables**
Lagged variables come up when you’re working with time series data, data collected on particular variables over a number of periods (years, usually).

For example, if you’re looking at the relationship between the unemployment rate and the inflation rate in the U.S., you would use information on these two variables collected over a number of years. This is time series data.

Alternatively, you might look at rainfall and water consumption in a community over a period of time to see if they are related. These would be time series data.

A lagged value is the value of a variable from a previous period, usually the period just prior.
For example, if you believed that a lower inflation rate resulted from higher unemployment, you might estimate the equation

\[ I_t = \beta_0 + \beta_1 U_t + \varepsilon_t \]

and see what the estimated coefficient on the unemployment rate in period t, \( U_t \), is.

However, if you believe that inflation depends not on what unemployment is now, but rather on what it was last period (in the previous quarter or the previous year) you might regress inflation on lagged unemployment, or the unemployment rate from the previous period

\[ I_t = \beta_0 + \beta_1 U_{t-1} + \varepsilon_t \]

In a spreadsheet, the numbers would look like this:

<table>
<thead>
<tr>
<th>Year</th>
<th>Inflation</th>
<th>Unemployment Rate</th>
<th>Lagged Unemployment Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1989</td>
<td>5.2%</td>
<td>3.2%</td>
<td></td>
</tr>
<tr>
<td>1990</td>
<td>5.0%</td>
<td>4.6%</td>
<td>3.2%</td>
</tr>
<tr>
<td>1991</td>
<td>4.8%</td>
<td>4.9%</td>
<td>4.6%</td>
</tr>
<tr>
<td>1992</td>
<td>4.6%</td>
<td>5.9%</td>
<td>4.9%</td>
</tr>
<tr>
<td>1993</td>
<td>4.9%</td>
<td>4.5%</td>
<td>5.9%</td>
</tr>
</tbody>
</table>

There are situations in which the value of an explanatory variable from a previous period can be useful in predicting or forecasting the value of a dependent variable today. More valuable, if you know the value of the explanatory variables now, you can predict the value of a dependent variable next period.

Examples include questions of serial correlation. These may be "hot hand" studies in sports (especially basketball) or about changes in the stock market. Does success on one try predict success on a subsequent try? Does an increase on one day predict anything about the next day?

A popular story here has to do with crop prices. Farmers may make planting decisions based on crop prices at the time of planting or from the previous harvest.

Examples from the class
2. Dummy variables
A dummy variable is used for including qualitative variables in a regression. A dummy variable generally takes the value 1 if a condition is satisfied and 0 if a condition isn't satisfied. This allows estimation of the effect on the dependent variable of the condition being satisfied.

For example, if you are looking at personal income and include a dummy variable which takes the value 1 if the individual is a male and 0 if the individual is a female, then the estimated coefficient on this variable is the expected difference in a person's income if they are a man rather than a woman, other things held constant.

What other variables would you include in this regression and why is this one coefficient of interest?

If you were to regress Personal Income on Age and Gender, the equation to be estimated might be

\[ PI_i = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Male}_i + \varepsilon_i \]

The coefficient \( \beta_1 \) would be the slope of the line or the average difference that a year of age makes in terms of personal income.

The coefficient on Male, \( \beta_2 \), is a shift in the intercept for men. The possible results are drawn below:

![Diagram of lines representing Personal Income (PI) vs. Age for Men and Women, showing the intercepts and slopes.](image-url)
Alternatively, you could add an interactive term that would allow the effect of age on men’s and women’s incomes to be different. In this case, the model would be

\[ P_i = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Male}_i + \beta_3 \text{Male}_i \times \text{Age}_i + \epsilon_i \]

The picture of a possible result is:

If there are several mutually exclusive characteristics that you want to include as explanatory variables, each may have its own dummy variable but one of these must be excluded from the analysis.

For example, if you do the above analysis of personal income and decide to include some measure of education there are several ways this could be done. Including the number of years of education ignores non-linear effects of diplomas. You could have dummy variables for some high school, high school diploma, some college, college diploma, some graduate school, masters degree and Ph.D. This would be seven dummy variables, only six of which could be included in the regression. If the first (some high school) is excluded, the interpretation of each of the other dummies would be the differential income to be expected if a person goes from having some high school to the indicated level of education, holding the other characteristics constant.
If you did this regression and found the results:

\[ \text{PI} = 14398 + 4569\text{HS} + 4834\text{SC} + 8993\text{CD} + 6734\text{SGS} + 15982\text{MD} - 9875\text{PHD} \]

where the dummies are
- HS = high school
- SC = some college
- CD = college degree
- SGS = some graduate school
- MD = medical degree
- PHD = Ph.D.

The interpretation would be that a person who had some high school would have an estimated income of 14398, a person who completed high school would make an estimated $4,569 more than a person with some high school and a person with a Ph.D. would make an estimated $9,875 less than a person with some high school.

You might also have a dummy dependent variable. That is, you might be trying to model a situation in which an event either did or did not occur as a result of a number of explanatory factors.

For example, if you are looking at police pursuit information, you might be interested in factors related to a pursuit ending in a crash. So, with data on pursuits from a particular department, you might use as your dependent variable a dummy taking the value 1 if the pursuit ended in a crash and 0 if it did not. The explanatory variables could be nice and quantitative (years of experience of the pursuing officer) or they could also be dummy variables (a male dummy, for example, taking the value 1 if the pursuing officer was male and 0 if not).

Techniques for this sort of analysis (situations with a dummy dependent variable) are discussed in Chapter 13 of Studenmund.