Urban Design and Planning
URBDP 520 Quantitative Methods in Urban Design and Planning

Lecture Notes 8: Regression Analysis
These notes refer to a couple of books which you may want to purchase if you wind up doing a lot of regressions. One is Using Econometrics: A Practical Guide by A.H. Studenmund and the other is A Guide to Econometrics by Peter Kennedy.

Introduction
Regression analysis is a method of statistically estimating a mathematical relationship between one *dependent variable* and one or more *explanatory* or *independent variables*.

For example, if you believe that the price of a house in a particular neighborhood is dependent on its square footage, the number of bedrooms and the number of bathrooms, you might want to estimate the model

$$PRICE = \beta_0 + \beta_1 SQFT + \beta_2 BEDROOM + \beta_3 BATHROOM$$

In which the price of the house is the dependent variable and the square footage, number of bedrooms and number of bathrooms are explanatory variables.

Where PRICE is the price of the house, SQFT is the number of square feet, BEDROOM is the number of bedrooms and BATHROOM is the number of bathrooms.

The data necessary for doing this would be a number of observations of either house sales prices or assessed values and the square footage and number of bedrooms and bathrooms for those houses. For example:

PRICE	SQFT	BEDROOM	BATHROOM
$345,000	1700	4	2
$280,000	1400	3	1.5
$459,500	1800	4	3
$240,000	1450	2	2

More observations are always better, but at the very least you would like for the number of observations minus the number of explanatory variables to be greater than thirty or forty. In this case there were three explanatory variables, so you would like to have at least thirty five observations.

Student Examples of Interesting Regression Possibilities

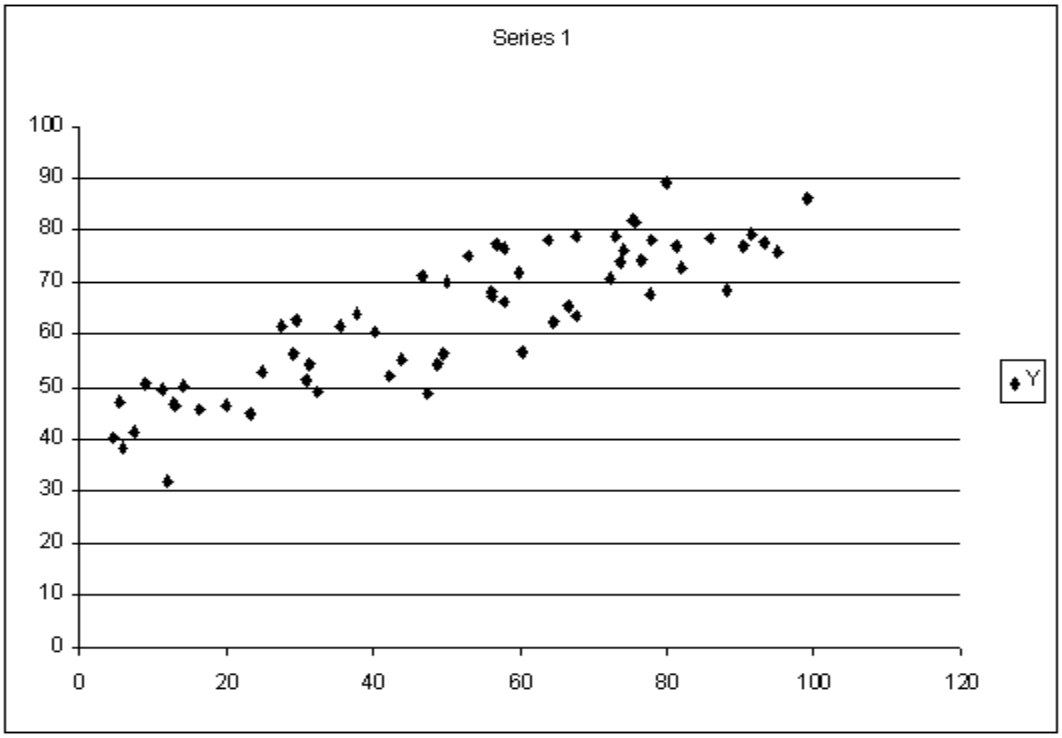
A Two-dimensional Example
Imagine that you have information on income and consumption spending. You believe that consumption spending is dependent on income and you graph the two with the dependent variable on the vertical axis and the explanatory or independent variable on the horizontal axis.

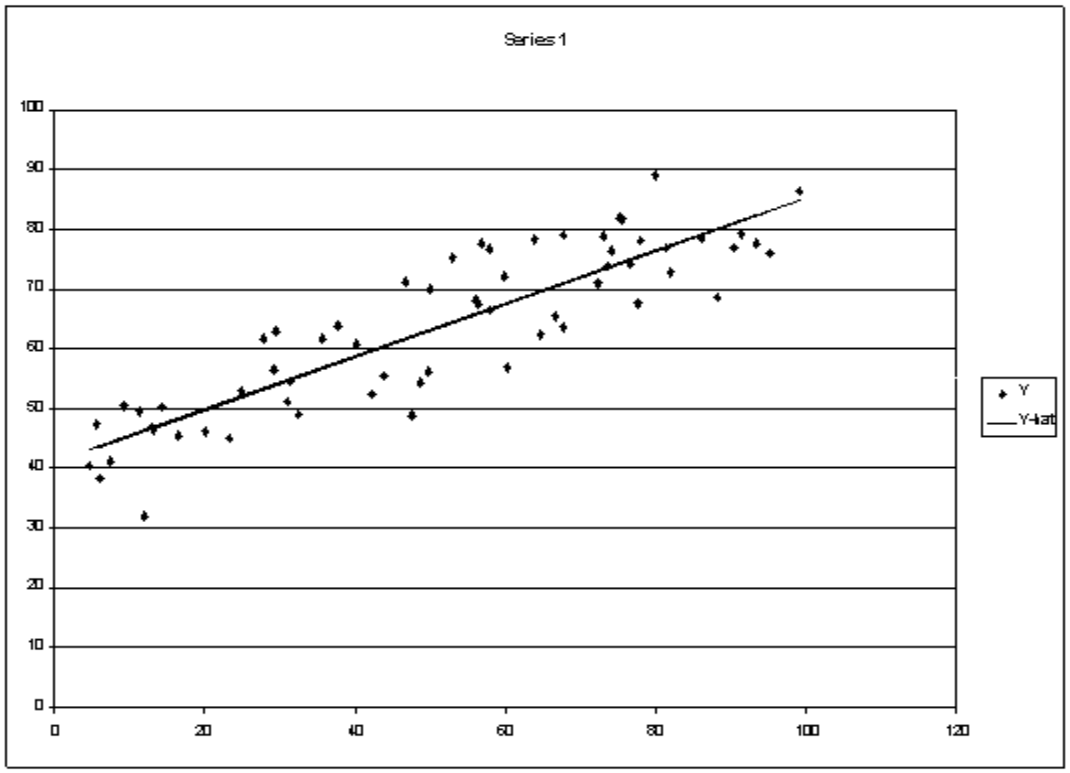
The goal of doing a regression is to draw the best possible straight line through the data points to describe the relationship between the two variables. If this line is upward sloping the data suggest a positive relationship. If this line is downward sloping the data suggest a negative relationship. If this line is horizontal, the data suggest no relationship... maybe.

You should be careful about using the results of the regression to make statements about what happens at levels outside of what has been observed. For example, if you've observed income levels between $10,000 and $30,000 and you use this data to estimate a relationship between income and consumption, you should be careful about using this relationship to predict the level of consumption consistent with income of, say, $60,000.

Linear Equations
The regression analysis we'll look at will estimate linear relationships between the dependent variable and the independent or explanatory variables.

Basic linear relationship: $Y = b_0 + b_1X$
Y is a function of X.
If X increases by one unit, Y increases by b_1 units. A graph of a linear function.





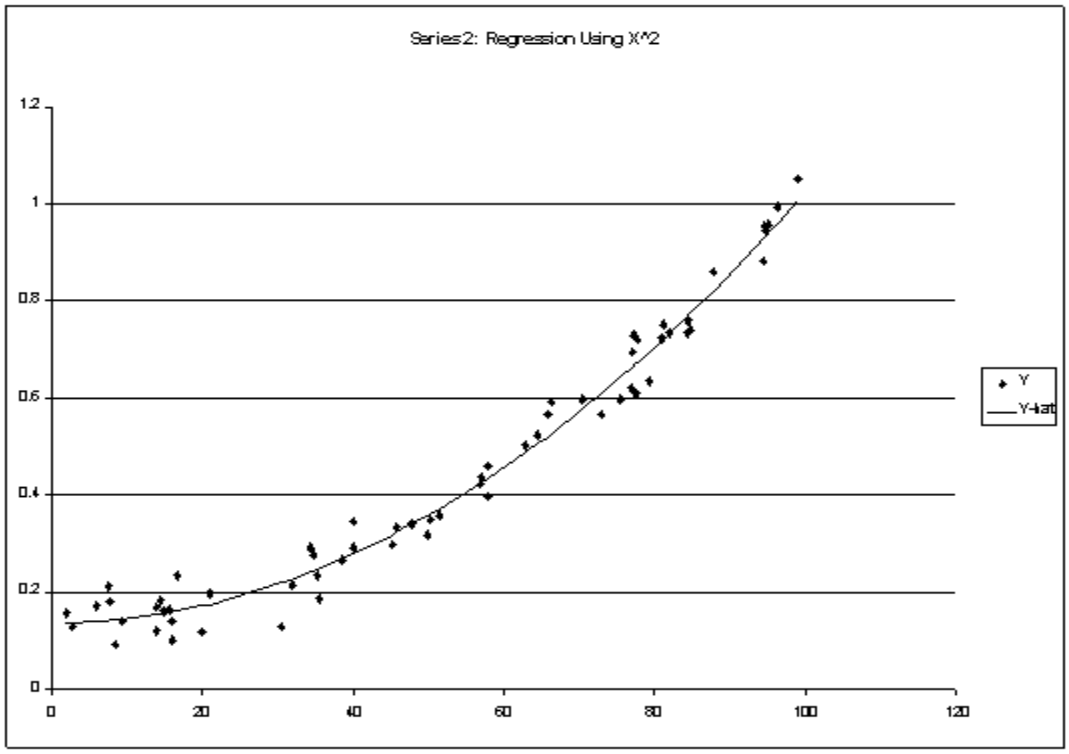
For example, if you are looking at income and consumption: $C = b_0 + b_1 I$
Consumption (C) is a function of income (I) If you do a regression here, you might find the estimated values: $C = 8{,}435 + 0.63I$ What do these estimated coefficients mean?

The $8,435 is the amount of consumption spending that the model predicts for a person with no income. The 0.63 says that, for an average person in the sample, an increase in income of $1.00 will result in an increase in consumption spending of $0.63.

Interpretation of Coefficients from Student Examples

Some Non-linear Stuff
Now, what if you graph the points and the relationship is clearly not a straight line? If we try to estimate a straight line (linear) relationship between the explanatory and the dependent variable it will be inherently wrong. Unfortunately, we can't estimate non-linear stuff. What to do?

What if the relationship is positive and seems to get steeper and steeper? We can estimate a relationship of the form: $Y_i = b_0 + b_1 X_i + b_2 X_i^2 + \varepsilon_i$ That is, we can use both X and X^2 as explanatory variables in a linear equation.



Ordinary Least Squares

Ordinary least squares (OLS) is a mathematical technique used to estimate a relationship between different variables. The most simple version of this relationship is:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

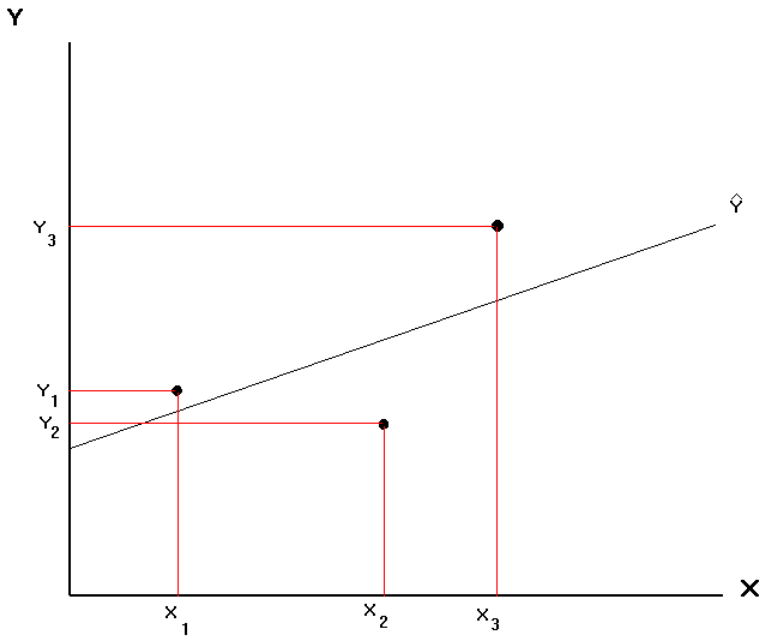
The result of this estimation procedure is estimates of the coefficients, b_0 and b_1 which are called b_0-hat and b_1-hat. These coefficients are used to generate estimates of the dependent variable which are called Y_i-hat and we can say that

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$$

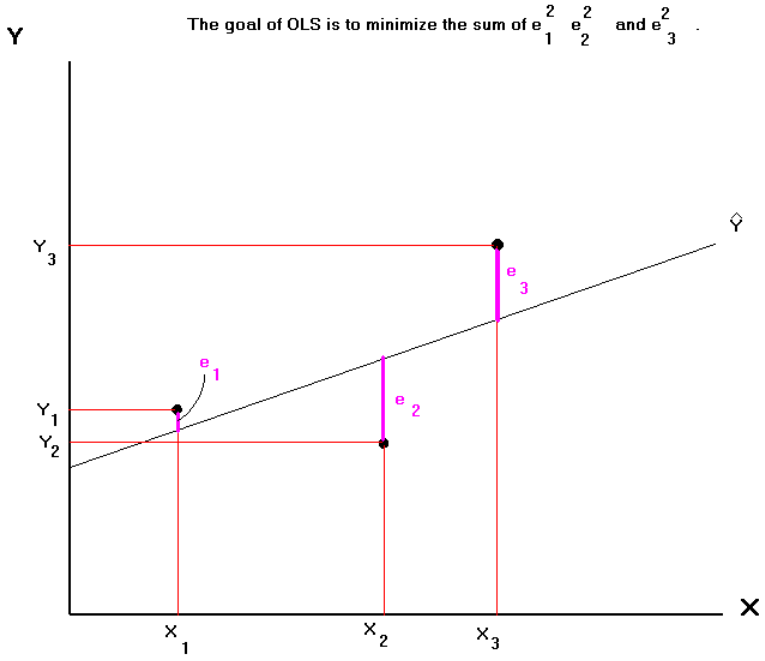
The difference between the actual value of Y_i and its estimated value, $\hat{Y}_i$ is equal to e_i, the error term. This can be written as:

$$Y_i = \hat{b}_0 + \hat{b}_1 X_i + e_i$$
$$Y_i - \hat{Y}_i = e_i$$

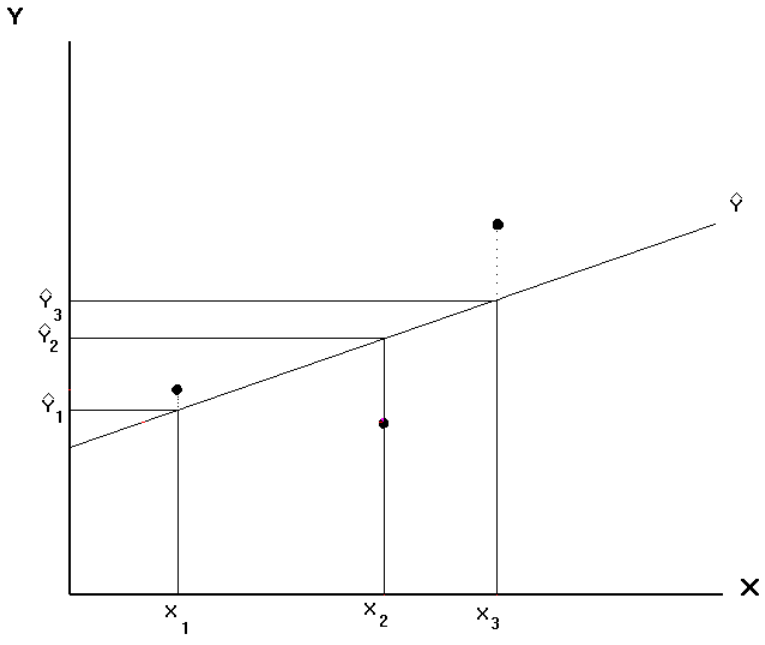
So, what do these things look like on a graph?


Y
Ŷ
Y3
Y1
Y2
X
X1
X2
X3

What do e_i look like?



What do Yi-hat look like?



The line relating X and Y that is calculated by OLS is good because it minimizes the sum of the squared errors. That is it minimizes:

$$\sum e_i^2$$

It is equivalent to say that it minimizes:

$$\sum \left(Y_i - \hat{Y}_i\right)^2$$

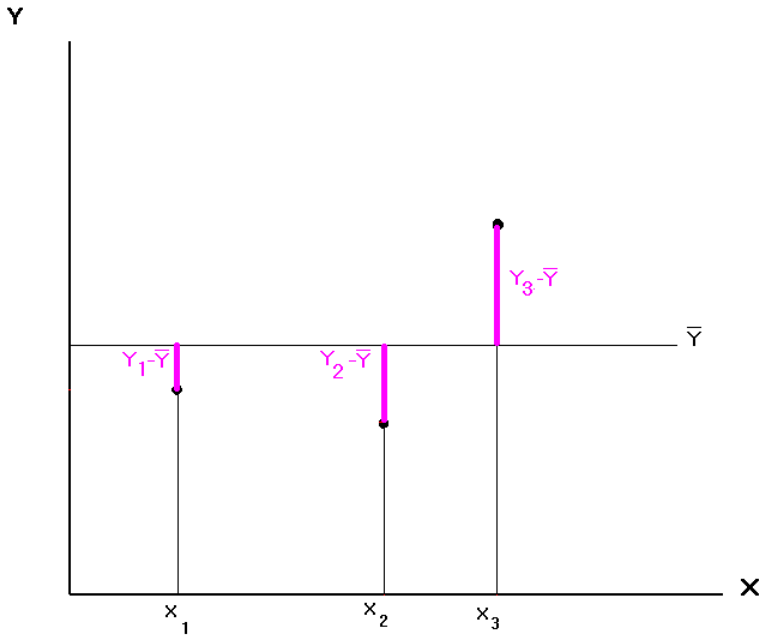
This has three good characteristics:
1. The regression line goes through the point $\left(\overline{X}, \overline{Y}\right)$ which is the mean of the data
2. The sum of the errors or residuals is zero
3. OLS gives the "best" estimation, depending on some conditions and definitions

Definitions
Standard Error of the Estimate (SEE)

$$SEE = \left[\frac{\sum e_i^2}{n-2}\right]^{1/2}$$

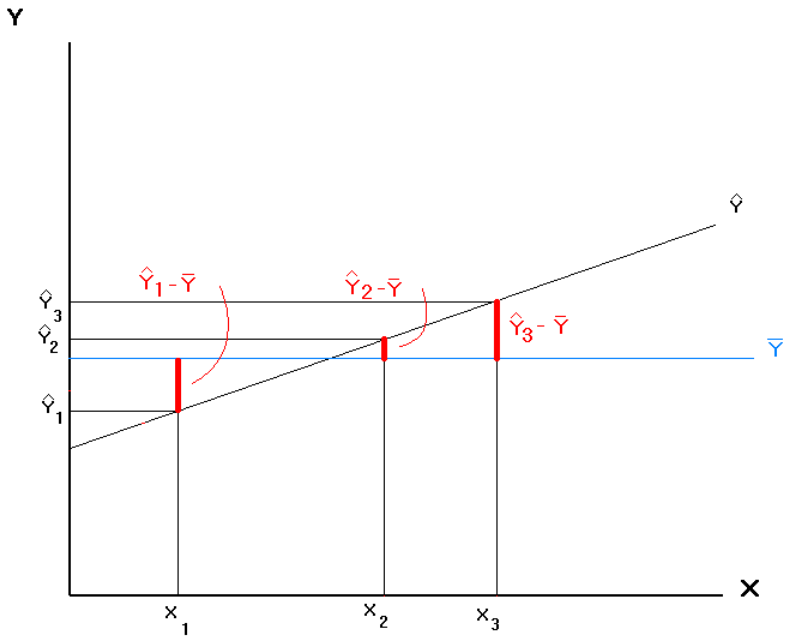
Total Sum of Squares (TSS)

$$TSS = \sum \left(Y_i - \overline{Y}\right)^2$$



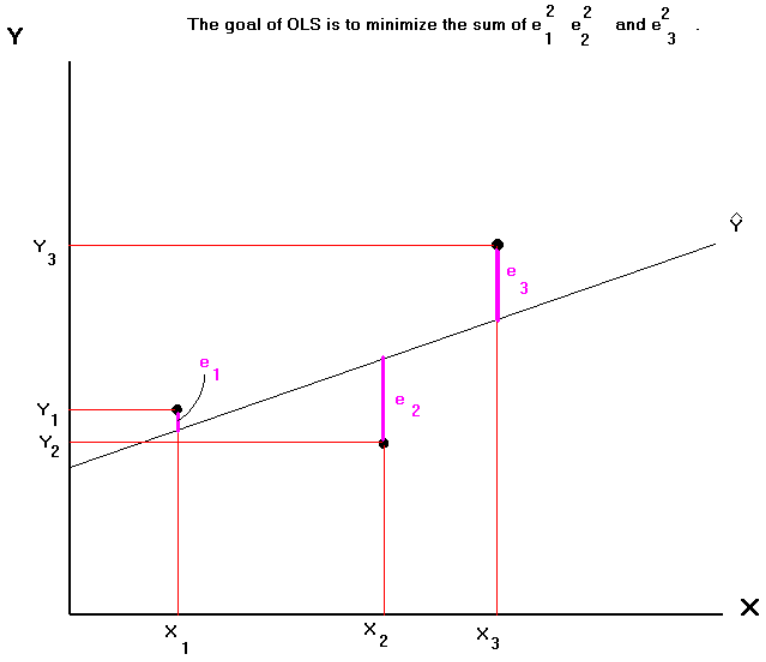
Explained Sum of Squares (ESS)

$$ESS = \sum \left(\hat{Y}_i - \bar{Y}\right)^2$$



Residual Sum of Squares (RSS)

$$RSS = \sum \left(e_i^2\right) = \sum \left(\hat{Y}_i - Y_i\right)^2$$



To put this all together: TSS = ESS + RSS

R^2

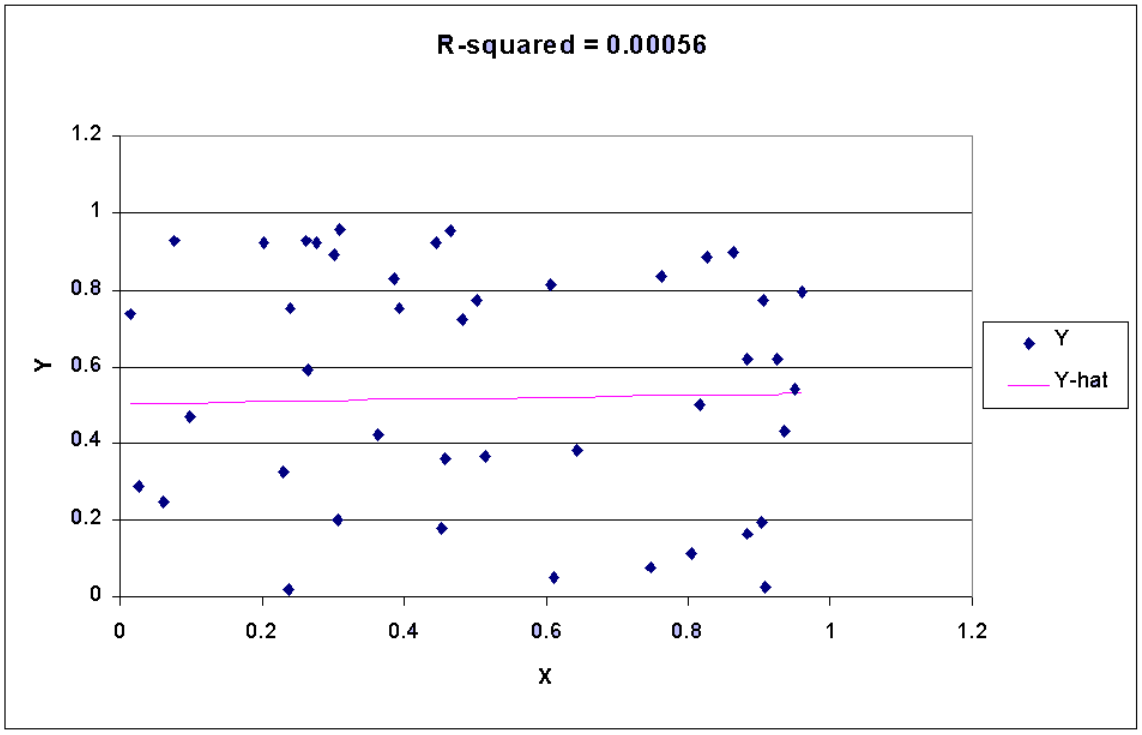
How much of the variation in the dependent variable is explained by the model? This is given by R^2, which is ratio of explained sum of squares to total sum of squares, or:

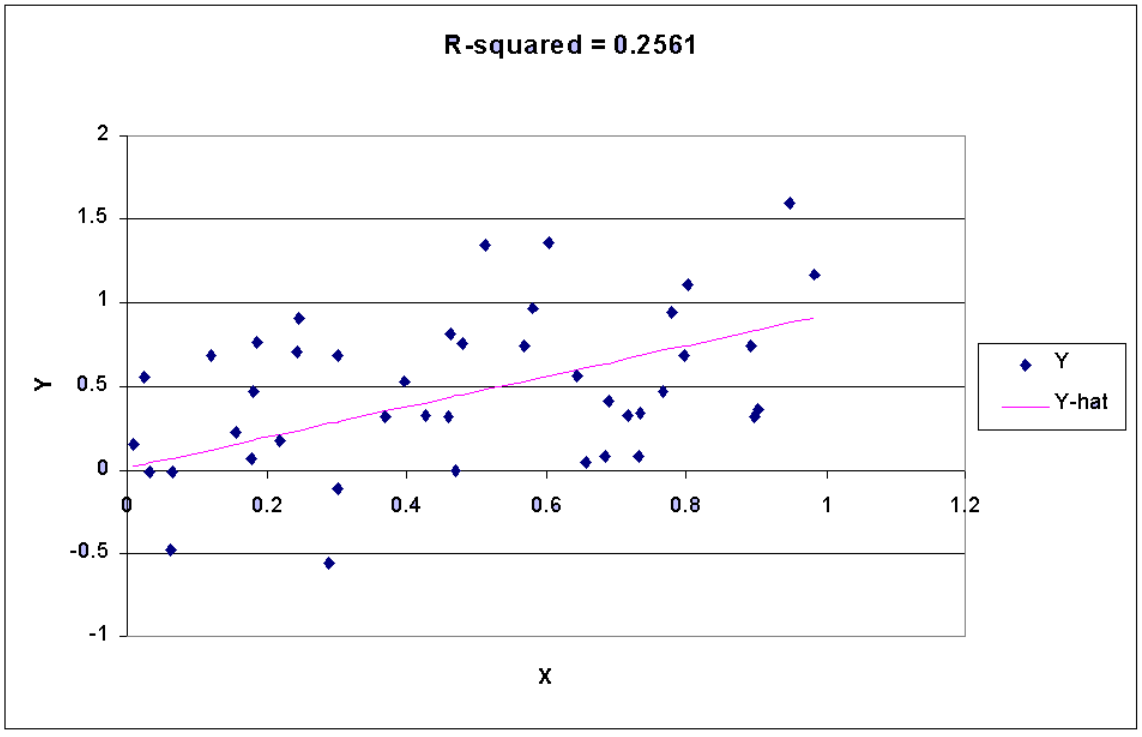
$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

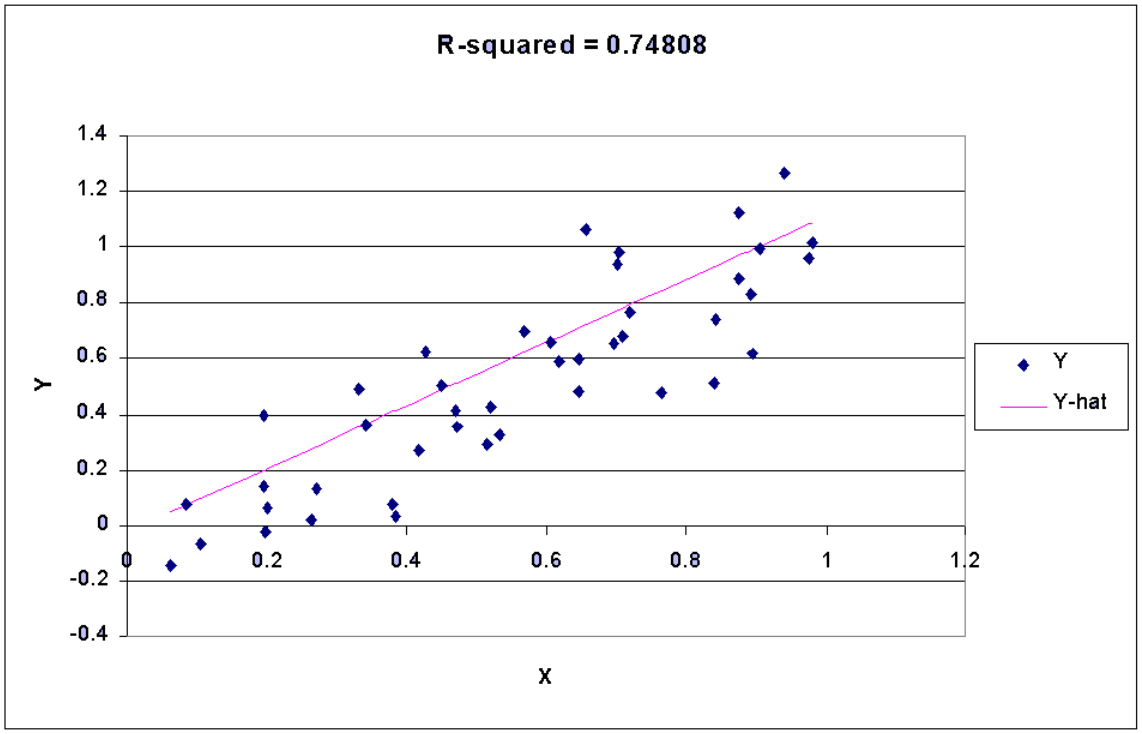
A higher R^2 means that the model being estimated explains a higher level of variation in the dependent variable.

If the R^2 is zero, then the model offers no information about the dependent variable and the best prediction you can make about the value of the dependent variable is its mean. The "explanatory" variables really offer no explanatory power whatsoever.

What does this look like?


R-squared = 0.00056
1.2
1
0.8
0.6
0.4
0.2
0
Y
0
0.2
0.4
0.6
0.8
1
1.2
X
Y
Y-hat


R-squared = 0.2561
2
1.5
1
0.5
0
-0.5
-1
Y
0
0.2
0.4
0.6
0.8
1
1.2
X
Y
Y-hat



Ordinary least squares (OLS) is a very good way of estimating a linear relationship between a dependent variable and some independent or explanatory variables. In some sense, it is the best method of doing this. For it to be "best", however, there are seven assumptions which must be satisfied. These assumptions are somewhat technical, but here is an attempt to explain what each means and how it affects regression results.

I. The regression model is linear in the coefficients, is correctly specified and has an additive error term. This assumption has three parts. Let's look at them each in turn. First, the model must be linear in the coefficients. This means that the process that is actually occurring in the real world is described by a relationship of the form

$$Yi = b0 + b1X1i + b2X2i + b3X3i + ... + ei$$

or that the actual relationship can be rewritten in this form by, for example, taking logs. This part is best considered in conjunction with the second.

Second, the model is correctly specified. Combining these first two assumptions, we see that we need to know the actual process through which the dependent value is determined and that relationship has to be linear or some derivation of linear. While this may be possible for simple physical processes, it is virtually impossible that the decision making process of any rational person operating in a complex environment can be correctly modeled by a linear equation. What must be believed instead is that the model is sufficiently close to the actual process that the difference isn't important. In a practical

sense, you can defend a specification by saying that it is a standard specification used in looking at the situation being examined. If you find other studies, papers or reports that have used a particular model, that may be an acceptable reason for using it. The opposite also holds. On the other hand, if you use a model which no one else has used, the results may argue for rejection of other models in favor of yours.

Third, the error term is additive. This means that the term ei is simply added to the predicted value so that

$$Yi\text{-}hat = Yi + ei$$

rather than being, for example,

$$Yi = Yi\text{-}hat * ei.$$

This is probably no more suspicious that the assumption that you have the correct model and that it is linear, but by looking at the residuals (ei) it can be determined whether or not this may be true. If the model is not correct, we may have problems such as those we've seen in class where the correct model is, say, curved but the estimated model is a straight line.

II. The error term has a zero population mean. This means that the expected value of ei is zero (E[ei]=0). Because Yi-hat = Yi + ei, this means that: E[Yi-hat] = E[Yi + ei] E[Yi-hat] = E[Yi] + E[ei] E[Yi-hat] = Yi + 0 So, the expected value of Yi-hat is equal to the actual value of Yi. There is nothing earthshaking about this.

III. All explanatory variables are uncorrelated with the error term. This means that the error term is not likely to be larger or smaller, positive or negative when any of the explanatory variables are larger or smaller. If this was not true, then you might know, for example, that the error term is larger when one of the explanatory variables is larger and smaller when the explanatory variable is smaller. If this was true, the model could be improved based on the value of that explanatory variable. Whether or not this condition is satisfied can be investigated in two ways. The residuals (the differences between the actual value of the dependent variable and the predicted values) can be graphed against the various explanatory variables. You can also calculate correlation coefficients between the residuals and the various explanatory variables. There should be no discernable patterns in the graphs and only very small correlation coefficients.

IV. Observations of the error term are uncorrelated with each other (no serial correlation). If you're looking at time series data (data collected from the same source in a number of different periods) the error term (Yi - Yi-hat) in one period should not have any relation to the error term from the previous period. A way to investigate this is to graph the error terms over time and see if there are any patterns or long runs of either positive or negative values.

V. The error term has a constant variance (no heteroskedasticity). This means that the errors aren't more spread out for some of the observations than for others. It's tough to describe, but there's a good picture of it on Studenmund, p. 99. Basically, if you graph the squared residuals against all the explanatory variables, the size of the residuals shouldn't depend on the value of the explanatory variables. An example of a case in which heteroskedasticity might be a problem is in modeling house prices as a function of the house characteristics. There might be larger variance for the error term for more expensive houses and smaller variance for less expensive houses. A 95% confidence interval for the true value of a house with an estimated value of $40,000 might be [$38,000, $42,000] while the same interval for a house with an estimated value of $2,000,000 might be [$1,900,000, $2,100,000]. Kennedy (pp. 118-21) has a very good discussion about the consequences of heteroskedasticity, methods of testing for it and a somewhat vague description of how to correct for it. Kennedy offers four methods of detecting heteroskedasticity:
1. Visual inspection of the residuals
2. The Goldfeld-Quandt test
3. The Breusch-Pagan test
4. The White test
The first of these is within your power to do in Excel. The others should be available options in any good software package. To deal with heteroskedasticity, you need to run a weighted least squares (rather than an ordinary least squares) regression.

VI. No explanatory variable is a perfect linear function of any other explanatory variable(s) (no perfect multicollinearity). This means that you can include as explanatory variables X and X^2 but you cannot include, for example, temperature in degrees Fahrenheit (F) and in degrees Celsius (C) because F = 32 + 1.8C, that is, Celsius is a linear function of Fahrenheit. This is why you must exclude one of the dummy variables if there is an exhaustive list of them. If, for example, you have dummy variables for male (M) and female (F) subjects and there are no other genders, then for each observation M + F = 1 or F = 1 - M or M = 1 - F. Because these two variables are linear functions of each other, one of them must be excluded. One way to see if this might be a problem is to generate a matrix of correlation coefficients for all the explanatory variables and the dependent variable. This won't tell you if a large number of variables are linearly related (such as dummy variables for a person's home state, for example) but it will tell you if two variables are linearly related. Kennedy (pp. 183-89) has a very good section on multicollinearity. A fun quote from this section: The OLS estimator in the presence of multicollinearity remains unbiased and in fact is still BLUE. The R^2 statistic is unaffected. In fact, since all the CLR assumptions are (strictly speaking) still met, the OLS estimator retains all its desirable properties, as noted in chapter 3. The major undesirable consequence of multicollinearity is that the variances of the OLS estimates of the parameters of the collinear variables are quite large. These high variances arise because in the presence of multicollinearity the OLS estimating procedure is not given enough independent variation in a variable to calculate with confidence the effect it has on the dependent variable. Possible remedies are later suggested by Kennedy.

VII. The error term is normally distributed This is important when generating confidence intervals and doing hypothesis testing in small samples but is less important as sample sizes increases.

Multivariate Regression

There aren't a lot of interesting questions to be answered using one explanatory variable. It's more fun to look at a number of explanatory variables through multivariate regressions, or regression on multiple variables. The explicit model is:

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + ... + b_KX_{Ki} + \varepsilon_i$$

Each of the coefficients, b_j, is the partial derivative of the dependent variable, Y_i with respect to the explanatory variable X_{ji}. That is, the coefficient is the expected change in Y resulting from a one unit change in X holding the other variables constant.

A Beef Example (Studenmund, p. 44)

Consider the following estimated model:

$$\hat{B}_t = 37.53 - 0.88P_t + 11.9Yd_t$$

where

B_t = the per capita consumption of beef in year t in pounds per person
P_t = the price of beef in year t (in cents per pound)
Yd_t = the per capita disposable income in year t (in thousands of dollars)

Questions

1. What is the interpretation of the coefficient on P_t?
2. What is the interpretation of the coefficient on Yd_t?
3. Is beef a normal good?
4. Do these estimated coefficients conform to the law of demand?
5. According to the model, what would happen to per capita consumption of beef if the price rose by $0.02/pound?
6. According to the model, what would happen to per capita consumption of beef if per capita disposable income rose by $2,000?
7. According to the model, what would happen to per capita consumption of beef if the price of beef doubled?
8. According to the model, what would happen to per capita consumption of beef if the price rose by $0.50? Do you believe this result? Explain the problem and offer a solution.
9. How would the coefficient estimates change if beef consumption was expressed in kilograms per person? If the price was expressed in dollars per pound?

Adjusted R^2

Now, an ideal model will have a lot of explanatory power. This means that ESS/TSS = R^2 should be as close as possible to 1. However, adding more and more variables to the equation to be estimated will never decrease the R^2 and will often increase it. The result is

that a model with many unrelated explanatory variables will appear to have very high explanatory power, but this high R^2 will merely be a result of spurious correlation.

To correct for this, we calculate the adjusted R^2.

$$\text{Adj. } R^2 = R^2 - \left(1 - R^2\right)\frac{K}{n - K - 1}$$

where

n = the number of observations in the data set
K = the number of slope coefficients estimated

For example, in the beef regression above, n would be the number of years in which data were gathered and K would be two because one coefficient was estimated for price and one was estimated for income.

Adding a variable to a regression, even if it has nothing to do with the dependent variable, is likely to increase R^2 but will also increase K and may decrease the adjusted R^2.

A good general rule to use in choosing between models is to choose the model with a higher adjusted R^2. If you are considering adding a new variable to a regression, see if it increases the adjusted R^2. If it does, then it should perhaps be added.

This, however, is a rule which should be applied only carefully, as the example in Studenmund 2.5 explains.

The best way to choose variables for a regression is to research the dependent variable and, on the basis of your understanding of the variable, decide which variables should be included prior to doing any regressions. Your model should be theoretically sound and you should be able to offer a good explanation for your inclusion of each explanatory variable.

Degrees of Freedom

The number of degrees of freedom in a regression is equal to the difference between the number of observations in the data set (n) minus the number of coefficients to be estimated (K+1).

It must be the case that n-K-1>=0.

To see why, consider the case of a univariate regression with one data point.

$Y_i = b_0 + b_1 X_i$

Because there are two coefficients to be estimated (b_0 and b_1) we have K+1=1+1=2. So, we must have at least two points to estimate a line here. This can be seen quite easily graphically.

1. What if you try to estimate an OLS relationship with one observation?

2. How can you estimate an OLS relationship with one observation? What do you need to assume?

Tough Questions You Can Ask

Studenmund (p. 49) offers several questions you can and should ask when reading a report involving an OLS regressions.

1. Is the equation supported by sound theory?

At one level, ask yourself if the included explanatory variables make sense and if there are other variables you believe should be included. It may be that there is a very good reason for exclusion of some variables, but you should ask.

2. How well does the regression as a whole fit the data?

This relates to the R^2. A low R^2 doesn't necessarily mean you should condemn the model, but it should raise some flags about what the results should be used for. It should be understood that if the R^2 is low, the predictive power for any one observation may be very low, although it might be good for a large number of observations. Similarly, a very high R^2 might suggest something suspicious.

3. Is the data set reasonably large and accurate?

The number of observations is important, but even more important is the number of degrees of freedom. Further, you should ask yourself if all of the variables seem quantifiable and measurable and how accurately they may have been measured.

4. Is OLS the best estimator to be used for this equation?

There are some other options we will discuss, although they are all basically variations on the OLS theme.

5. How well do the estimated coefficients correspond to the expectations developed by the researcher before the data were collected?

Look for weird signs on the estimated coefficients. If quantity demanded is positively related to price, for example, there is something suspicious going on and you should ask questions.

6. Are all the obviously important variables included in the equation?

7. Has the most theoretically logical functional form been used?

Consider the beef example. Does the prediction about the effect of a $0.50/pound increase in the price of beef make sense? A different model may be appropriate here. Explanatory variables may need to be raised to some power to more accurately describe how they affect the dependent variable.

8. Does the regression appear to be free of major econometric problems?

We'll discuss some more of these. The one we have talked about that you should remember is endogeneity. That is, can you imagine a model in which one of the explanatory variables is explained by other explanatory variables?

Answers to Questions on the Beef Model

1. What is the interpretation of the coefficient on P_t?

This is the change in annual, per capita beef consumption in pounds resulting from a $0.01 increase in the price of beef.

2. What is the interpretation of the coefficient on Yd_t?

This is the change in annual, per capita beef consumption in pounds resulting from a $1,000 increase in per capita income.

3. Is beef a normal good?

Yes, the data seem to suggest that it is because the coefficient on income is positive, so as income rises (holding other things constant), so will beef consumption.
4. Do these estimated coefficients conform to the law of demand?
Yes they do. The estimated coefficient on price is negative, suggesting that as the price of beef rises (holding other things constant) the quantity demanded will fall.
5. According to the model, what would happen to per capita consumption of beef if the price rose by $0.02/pound?
If the price of beef rises by $0.02/pound, the model predicts that annual per capita consumption will fall by 2*0.88=1.76 pounds.
6. According to the model, what would happen to per capita consumption of beef if per capita disposable income rose by $2,000?
If per capita disposable income rose by $2,000, the model predicts that annual per capita beef consumption would rise by 11.9*2=23.8 pounds.
7. According to the model, what would happen to per capita consumption of beef if the price of beef doubled?
We can't tell without knowing the current price of beef because we don't know the amount of the increase in cents per pound.
8. According to the model, what would happen to per capita consumption of beef if the price rose by $0.50? Do you believe this result? Explain the problem and offer a solution.
If the price of beef rose by $0.50/pound, annual per capita consumption would fall by 50*0.88=44 pounds. This seems a bit drastic and I would suggest a model in which beef consumption was maybe related to the square root of the price.
9. How would the coefficient estimates change if beef consumption was expressed in kilograms per person?
Because there are 2.2 pounds per kilogram, each coefficient would be divided by 2.2. For example, to get a one kilogram increase in per capita consumption, the necessary price decrease would have to be 2.2 times what is needed to get a one pound increase.
If the price was expressed in dollars per pound?
If the price was expressed in dollars per pound the coefficient on P_t would be multiplied by 100. In any case, a one dollar increase in the price of beef will decrease annual per capita consumption by 88 pounds.

Four Important Specification Criteria

Studenmund offers four rules for determining if a variable should be added to a regression equation.
The first is the most important; the others are supplemental. If you're considering adding a theoretically justifiable variable to your equation, do a regression without it in the equation and then with it in the equation. If these rules are satisfied, the variable should be included.
1. Theory: Is the variable's place in the equation unambiguous and theoretically sound?
2. t-test: Is the variable's estimated coefficient significant in the expected direction?
3. Adj. R2: Does the overall fit of the equation (adjusted for the degrees of freedom) improve when the variable is added to the equation?
4. Bias: Do other variables' coefficients change significantly when the variable is added to the equation?

Some Practices to Avoid and/or Understand

1. Data Mining: Data mining is the estimation of many (or all) possible regression equations with no regard for theoretical justification in a blind attempt to get the desired results. Remember, a level of significance of 5% in a hypothesis test means that even if an explanatory variable has no influence on a dependent variable, 5% of the time it will appear to have influence. If you try twenty different models which have no real explanatory power, the expected number of models which will appear to have significant explanatory power (at a 5% level of significance) is one.
2. Stepwise Regressions: This is the process of allowing a computer package to choose the explanatory variable which has the greatest explanatory power, then having chosen the first, chooses a second explanatory variable which adds the most explanatory power from those remaining, and so on. There are actually procedures written into some statistical packages which do this automatically if asked. Computers are dumb machines and know not what they do, but the software packages have this features because there are equally dumb researchers out there who want to use this procedure. If someone seriously presents results from a stepwise regression, you should taunt them about their lack of an underlying theory until they cry.
3. Sequential Specification Searches: This is the process of starting with the variables you know should be included and then trying others about which you are less sure. This isn't necessarily a bad idea, but there is always a concern about what reasons the researchers may have had for reporting the results that they did. If a number of specifications are tried, all of their results should be either presented or, at least, mentioned in the final report. If, for example, a large number of the secondary explanatory variables had little or no effect on the estimated coefficients and explanatory power, this could be discussed in a footnote or an appendix.
4. Relying on t-test Results: Problems with multicollinearity and omitted variable bias can make t-tests unreliable indicators of which variables should be included or excluded.
5. Scanning: As far as I can tell, this refers to data mining one data set to find a good specification and then estimating that model using a different data set. Please note that this requires two distinct data sets.
6. Sensitivity Analysis: This is the practice of estimating your preferred specification, determining which results are important, and then estimating some slight variations of the specification to see if the important results are preserved or if they disappear. If the important result(s) persist across slight changes in the model, these results are said to be robust to slight changes in the model. If these important results disappear when the model is changed slightly, they may be artificial products of a particular specification and do not accurately reflect a relationship in the data. Presentation of the results from several different specifications can clarify the robustness of important regression results.

A fun question to ask someone presenting results is, "Are your important results robust to changes in model specification?"