Statlab Workshop

Introduction to Regression and Data Analysis

with

Dan Campbell

and

Sherlock Campbell

October 28, 2008

I. The basics

A. Types of variables

Your variables may take several forms, and it will be important later that you are aware of, and understand, the nature of your variables. The following variables are those which you are most likely to encounter in your research.

- Categorical variables

  Such variables include anything that is "qualitative" or otherwise not amenable to actual quantification. There are a few subclasses of such variables.

  - Dummy variables take only two possible values, 0 and 1. They signify conceptual opposites: war vs. peace, fixed exchange rate vs. floating exchange rate, etc.
  - Nominal variables can range over any number of non-negative integers. They signify conceptual categories that have no inherent relationship to one another: red vs. green vs. black, Christian vs. Jewish vs. Muslim, etc.
  - Ordinal variables are like nominal variables, only there is an ordered relationship among them: no vs. maybe vs. yes, etc.

- Numerical variables

  Such variables describe data that can be readily quantified. Like categorical variables, there are a few relevant subclasses of numerical variables.

  - Continuous variables can appear as fractions; in reality, they can have an infinite number of values. Examples include temperature, GDP, etc.
  - Discrete variables can only take the form of whole numbers. Most often, these appear as count variables, signifying the number of times that something occurred: the number of firms invested in a country, the number of hate crimes committed in a county, etc.

A useful starting point is to get a handle on your variables. How many are there? Are they qualitative or quantitative? If they are quantitative, are they discrete or continuous?

Another useful practice is to explore how your data are distributed. Do your variables all cluster around the same value, or do you have a large amount of variation in your variables? Are they normally distributed? Plots are extremely useful at this introductory stage of data analysis – histograms for single variables, scatter plots for pairs of continuous variables, or box-and-whisker plots for a continuous variable vs. a categorical variable. This preliminary data analysis will help you decide upon the appropriate tool for your data.

If you are interested in whether one variable differs among possible groups, for instance, then regression isn't necessarily the best way to answer that question. Often you can find your answer by doing a t-test or an ANOVA. The flow chart shows you the types of questions you should ask yourselves to determine what type of analysis you should perform. Regression will be the focus of this workshop, because it is very commonly used and is quite versatile, but if you need information or assistance with any other type of analysis, the consultants at the Statlab are here to help.

II. Regression: An Introduction:

A. What is regression?

Regression is a statistical technique to determine the linear relationship between two or more variables. Regression is primarily used for prediction and causal inference.

In its simplest (bivariate) form, regression shows the relationship between one independent variable (X) and a dependent variable (Y), as in the formula below:

$$Y = \beta_0 + \beta_1 X + u$$

The magnitude and direction of that relation are given by the slope parameter (β_1), and the status of the dependent variable when the independent variable is absent is given by the intercept parameter (β_0). An error term (u) captures the amount of variation not predicted by the slope and intercept terms. The regression coefficient (R^2) shows how well the values fit the data.

Regression thus shows us how variation in one variable co-occurs with variation in another. What regression cannot show is causation; causation is only demonstrated analytically, through substantive theory. For example, a regression with shoe size as an independent variable and foot size as a dependent variable would show a very high regression coefficient and highly significant parameter estimates, but we should not conclude that higher shoe size causes higher foot size. All that the mathematics can tell us is whether or not they are correlated, and if so, by how much.

It is important to recognize that regression analysis is fundamentally different from ascertaining the correlations among different variables. Correlation determines the strength of the relationship between variables, while regression attempts to describe that relationship between these variables in more detail.

B. The linear regression model (LRM)

The simple (or bivariate) LRM model is designed to study the relationship between a *pair* of variables that appear in a data set. The multiple LRM is designed to study the relationship between one variable and several of other variables.

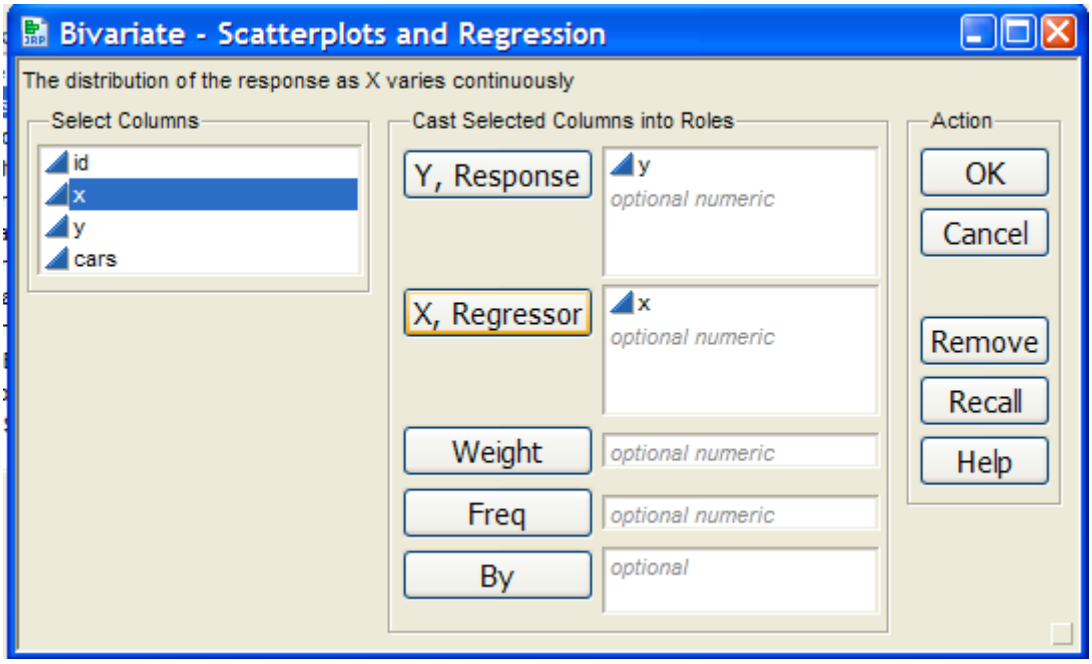
In both cases, the sample is considered a random sample from some population. The two variables, X and Y, are two measured outcomes for each observation in the data set. For

example, let's say that we had data on the prices of homes on sale and the actual number of sales of homes:

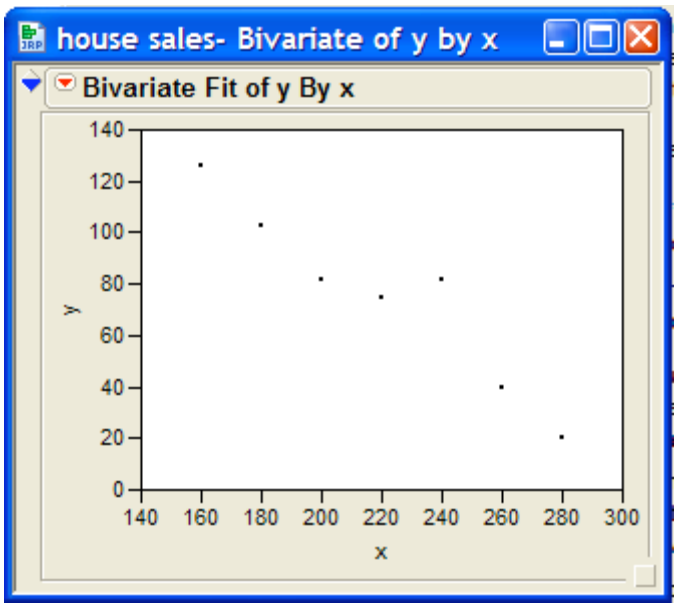
Price(thousands of $) Sales of new homes

x	y
160	126
180	103
200	82
220	75
240	82
260	40
280	20

This data is found in the file "house sales.jmp".
We want to know the relationship between X and Y. Well, what does our data look like? We will use the program JMP (pronounced 'jump') for our analyses today. Start JMP, look in the JMP Starter window and click on the "Open Data Table" button. Navigate to the file and open it. In the JMP Starter, click on "Basic" in the category list on the left. Now click on "Bivariate" in the lower section of the window.



Click on 'y' in the left window and then click the "Y, Response" button. Put 'x' in the "X, Regressor" box, as illustrated above. Now click "OK" to display the following scatterplot.



We need to specify the population regression function, the model we specify to study the relationship between X and Y.

This is written in any number of ways, but we will specify it as:

$$Y = \beta_1 + \beta_2 X + u$$

where

- **Y** is an observed random variable (also called the response variable or the left-hand side variable).
- **X** is an observed non-random or conditioning variable (also called the predictor or right-hand side variable).
- β_1 is an unknown population parameter, known as the constant or intercept term.
- β_2 is an unknown population parameter, known as the coefficient or slope parameter.
- **u** is an unobserved random variable, known as the error or disturbance term.

Once we have specified our model, we can accomplish 2 things:

- Estimation: How do we get "good" estimates of β_1 and β_2? What assumptions about the LRM make a given estimator a good one?
- Inference: What can we infer about β_1 and β_2 from sample information? That is, how do we form confidence intervals for β_1 and β_2 and/or test hypotheses about them?

The answer to these questions depends upon the assumptions that the linear regression model makes about the variables.

The Ordinary Least Squres (OLS) regression procedure will compute the values of the parameters β_1 and β_2 (the intercept and slope) that best fit the observations.

Obviously, no straight line can exactly run through all of the points. The vertical distance between each observation and the line that fits "best"—the regression line—is called the residual or error. The OLS procedure calculates our parameter values by minimizing the sum of the squared errors for all observations.

Why OLS? It has some very nice mathematical properties, and it is compatible with Normally distributed errors, a very common situation in practice. However, it requires certain assumptions to be valid.

C. Assumptions of the linear regression model

1. The proposed linear model is the correct model.
   Violations: Omitted variables, nonlinear effects of X on Y (e.g., area of circle = π*radius^2)
2. The mean of the error term (i.e. the unobservable variable) does not depend on the observed X variables.
3. The error terms are uncorrelated with each other and exhibit constant variance that does not depend on the observed X variables.
   Violations: Variance increases as X or Y increases.
   Errors are positive or negative in bunches – called heteroskedasticity.
4. No independent variable exactly predicts another.
   Violations: Including monthly precipitation for 12 months, and annual precipitation in the same model.
5. Independent variables are either random or fixed in repeated sampling

If the five assumptions listed above are met, then the Gauss-Markov Theorem states that the Ordinary Least Squares regression estimator of the coefficients of the model is the Best Linear Unbiased Estimator of the effect of X on Y. Essentially this means that it is the most accurate estimate of the effect of X on Y.

III. Deriving OLS estimators

The point of the regression equation is to find the best fitting line relating the variables to one another. In this enterprise, we wish to minimize the sum of the squared deviations (residuals) from this line. OLS will do this better than any other process as long as these conditions are met.

The best fit line associated with the n points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ has the form

$$y = mx + b$$

where

$$\text{slope} = m = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$\text{intercept} = b = \frac{\sum y - m(\sum x)}{n}$$

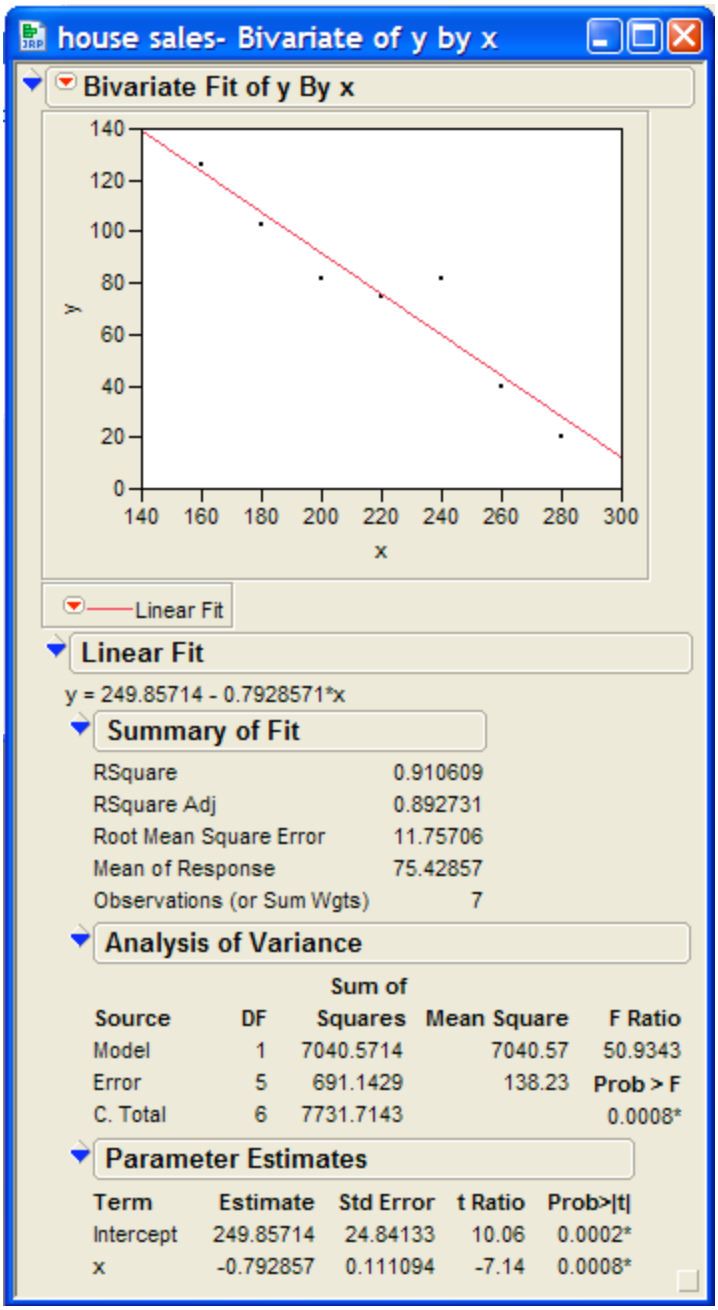
So we can take our data from above and substitute in to find our parameters:

Price(thousands of $) X	Sales of new homes y	xy	x^2
160	126	20,160	25,600
180	103	18,540	32,400
200	82	16,400	40,000
220	75	16,500	48,400
240	82	19,680	57,600
260	40	10,400	67,600
280	20	5,600	78,400
Sum 1540	528	107280	350000

$$\text{Slope} = m = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} = \frac{7(107280) - (1540)*(528)}{7(350000) - (1540)^2} = \frac{-62160}{78400} = -0.79286$$

$$\text{Intercept} = b = \frac{\sum y - m(\sum x)}{n} = \frac{528 - 0.79286(1540)}{7} = 249.8571$$

To have JMP calculate this for you, click on the small red triangle just to the left of "Bivariate Fit of y By x". From the drop down menu, choose "Fit Line" to produce the following output.



Thus our least squares line is

$$y = 249.85714 - 0.7928571x$$

IV. Interpreting data

Coefficient: (Parameter Estimate) for every one unit increase in the price of a house, -.793 fewer houses are sold.

Std Error: if the regression were performed repeatedly on different datasets (that contained the same variables), this would represent the standard deviation of the estimated coefficients

t-Ratio: the coefficient divided by the standard error, which tells us how large the coefficient is relative to how much it varies in repeated sampling. If the coefficient varies a lot in repeated sampling, then its t-statistic will be smaller, and if it varies little in repeated sampling, then its t-statistic will be larger.

Prob>|t|: the p-value is the result of the test of the following null hypothesis: in repeated sampling, the mean of the estimated coefficient is zero. E.g., if p = 0.001, the probability of observing an estimate of β that is at least as extreme as the observed estimate is 0.001, if the true value of β is zero.

In general, a p-value less than some threshold α, like 0.05 or 0.01, will mean that the coefficient is "statistically significant."

Confidence interval: the 95% confidence interval is the set of values that lie within 1.96 standard deviations of the estimated β

Rsquare: this statistic represents the proportion of variation in the dependent variable that is explained by the model (the remainder represents the proportion that is present in the error). It is also the square of the correlation coefficient (hence the name).

V. Multivariate regression models.

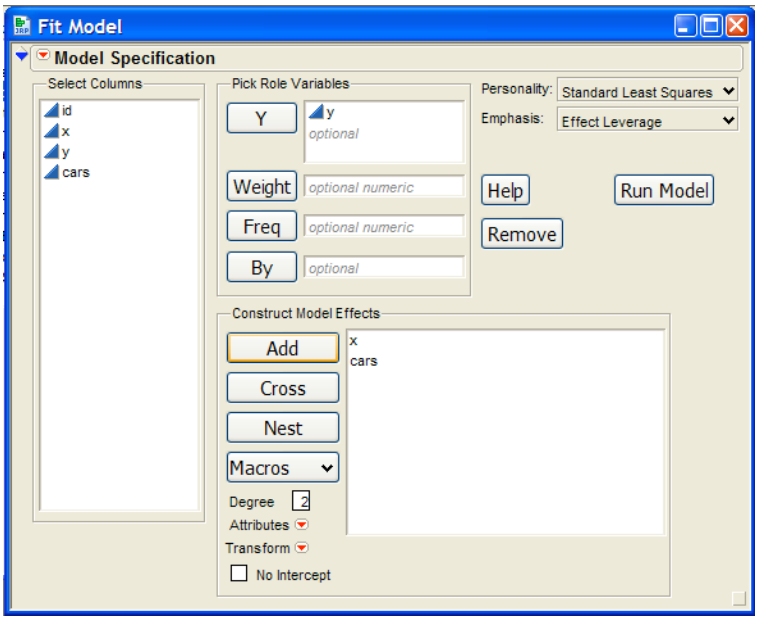
Now let's do it again for a multivariate regression, where we add in number of red cars in neighborhood as an additional predictor of housing sales:

Price(thousands of $)	Sales of new homes	Number of red cars
x	**y**	**cars**
160	126	0
180	103	9
200	82	19
220	75	5
240	82	25
260	40	1
280	20	20

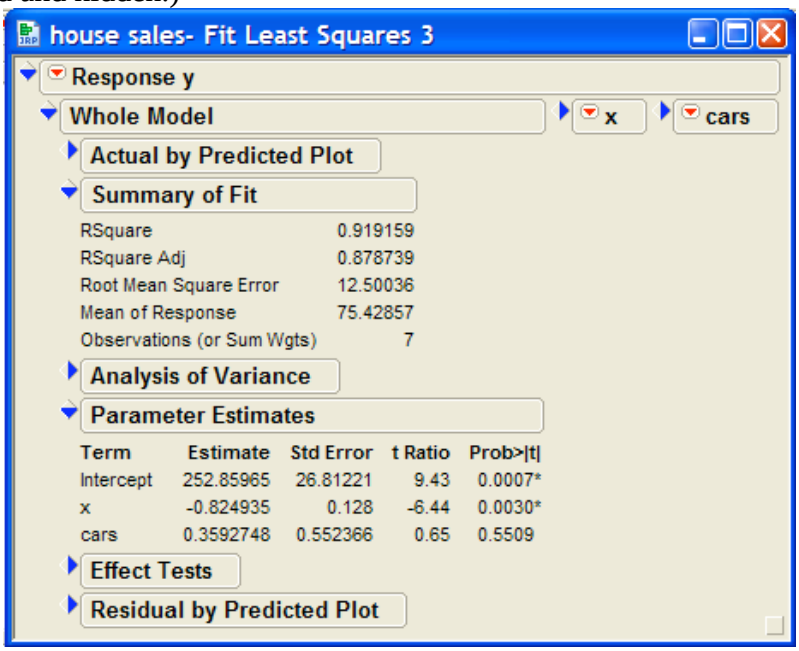
Our general regression equation this time is this:

$$Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + u$$

To calculate a multiple regression, go the JMP Starter window and select the “Model” category. Click on the top button, “Fit Model” for the following dialog:



Click on ‘y’ in the left window, then on the “Y” button. Now click on ‘x’ and then “Add” in the “Construct Model Effects” section. Repeat with‘cars’ then click on “Run Model” to produce the following window. (I’ve hidden a few results to save space. Click on the blue triangles to toggle between displayed and hidden.)



The regression equation we get from this model is

$$y = 252.85965 - .824935\ x_1 + .3592748\ x_2$$

In this case, our regression model says that the number of houses sold is a linear function of both the house price and the number of red cars in the neighborhood. The coefficient of each predictor variable is the effect of that variable, for a given value of the other.

Let's look at our t-Ratio on red cars. We can see that the t is quite low, and its associated p-value is much larger than 0.05. Thus, we cannot reject the null hypothesis that this coefficient is equal to 0. Since a coefficient of 0 essentially erases the contribution of that variable in the regression equation, we might be better off leaving it out of our model entirely.

Be careful, though – there is a difference between statistical significance and actual significance. Statistical significance tells us how sure we are about the coefficient of the model, based on the data. Actual significance tells us the importance of the variable – a coefficient for X could be very close to 0 with very high statistical significance, but that could mean it contributes very little to Y. Conversely, a variable could be very important to the model, but its statistical significance could be quite low because you don't have enough data.

What happened to our Rsquare in this case? It increased from the previous model, which is good, right? We explained more of the variation in Y through adding this variable. However, you can ALWAYS increase the R^2 by adding more and more variables, but this increase might be so small that increasing the complexity of the model isn't worth it. Testing the significance of each coefficient through p-values is the way to decide if a predictor variable should be included or not.

V. When linear regression does not work

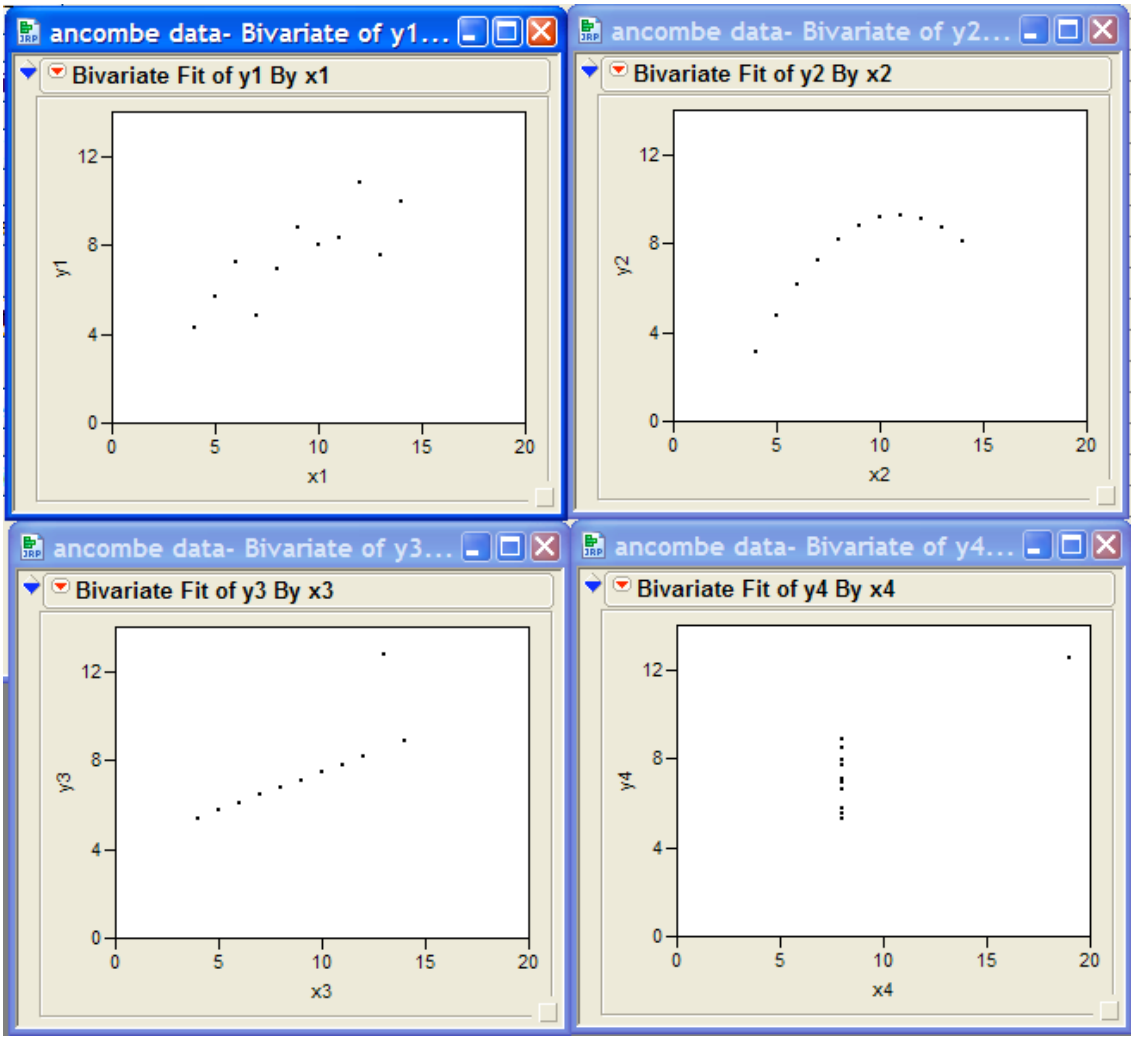
Linear regression will fail to give good results when any of the assumptions are not valid. Some violations are more serious problems for linear regression than others, though.

Consider an instance where you have more independent variables than observations. In such a case, in order to run linear regression, you must simply gather more observations. Similarly, in a case where you have two variables that are very highly correlated (say, GDP per capita and GDP), you may omit one of these variables from your regression equation. If the expected value of your disturbance term is not zero, then there is another independent variable that is systematically determining your dependent variable. Finally, in a case where your theory indicates that you need a number of independent variables, you may not have access to all of them. In this case, to run the linear regression, you must either find alternate measures of your independent variable, or find another way to investigate your research question.

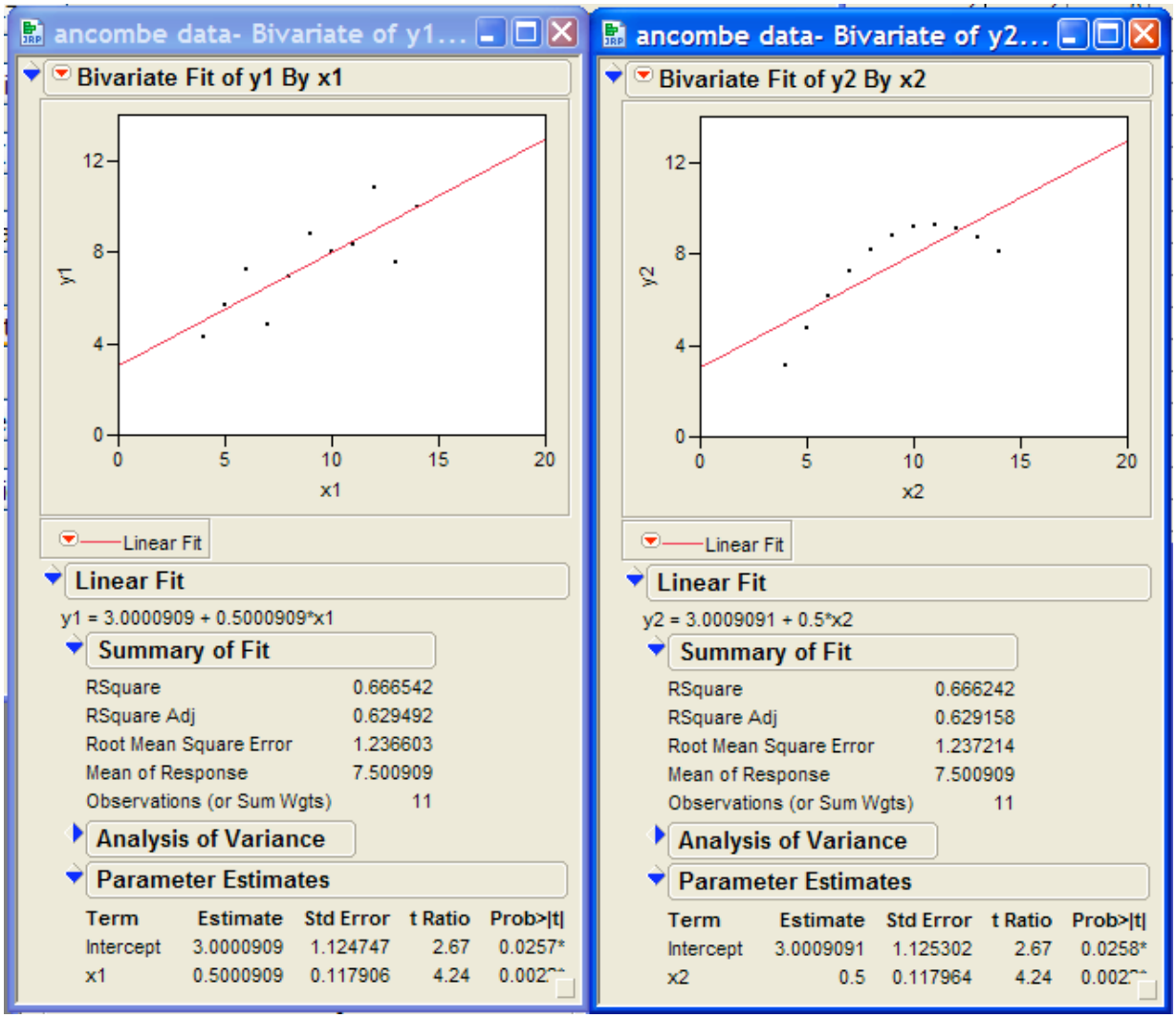
Other violations of the regression assumptions are addressed below. In all of these cases, be aware that Ordinary Least Squares regression, as we have discussed it today, gives biased estimates of parameters and/or standard errors.

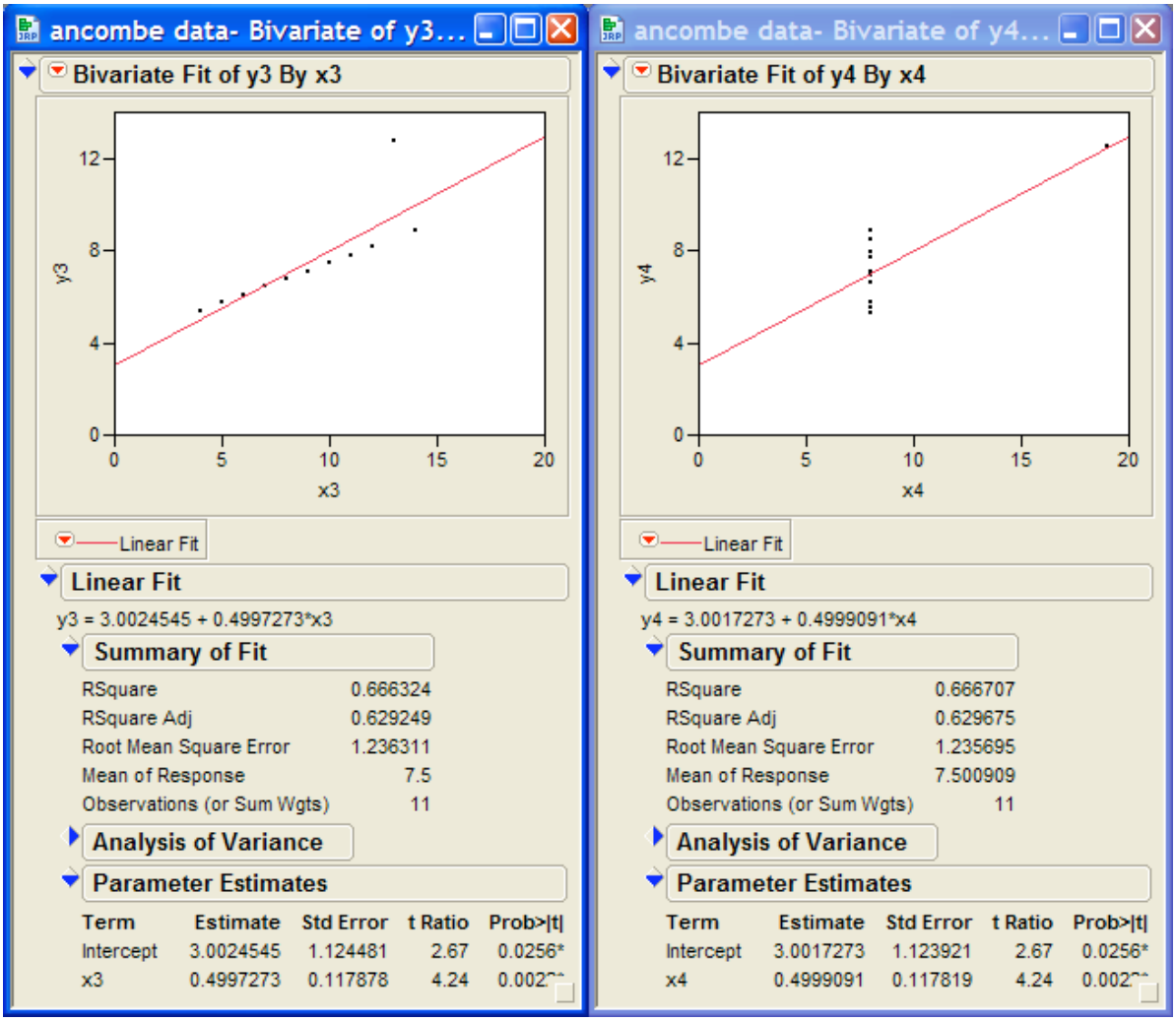
- Non-linear relationship between X and Y: use Non-linear Least Squares or Maximum Likelihood Estimation
- Endogeneity: use Two-Stage Least Squares or a lagged dependent variable
- Autoregression: use a Time-Series Estimator (ARIMA)
- Serial Correlation: use Generalized Least Squares
- Heteroskedasticity: use Generalized Least Squares

How will you know if the assumptions are violated? You usually will NOT find the answer from looking at your regression results alone. Consider the four scatter plots below:



Clearly they show widely different relationships between x and y. However, in each of these four cases, the regression results are exactly the same: the same best fit line, the same t-statistics, the same R^2.





In the first case, the assumptions are satisfied, and linear regression does what we would expect it to. In the second case, we clearly have a non-linear (in fact, a quadratic) relationship. In the third and fourth cases, we have heteroskedastic errors: the errors are much greater for some values of x than for others. The point is, the assumptions are crucial, because they determine *how well the regression model fits the data*. If you don't look closely at your data and make sure a linear regression is appropriate for it, then the regression results will be meaningless.

VI. Additional resources

If you believe that the nature of your data will force you to use a more sophisticated estimation technique than Ordinary Least Squares, you should first consult the resources listed on the Statlab Web Page at http://statlab.stat.yale.edu/links/index.jsp. You may also find that Google searches of these regression techniques often find simple tutorials for the methods that you must use, complete with help on estimation, programming, and interpretation. Finally, you should also be aware that the Statlab consultants are available throughout regular Statlab hours to answer your questions (see the areas of expertise for Statlab consultants at http://statlab.stat.yale.edu/people/showConsultants.jsp).

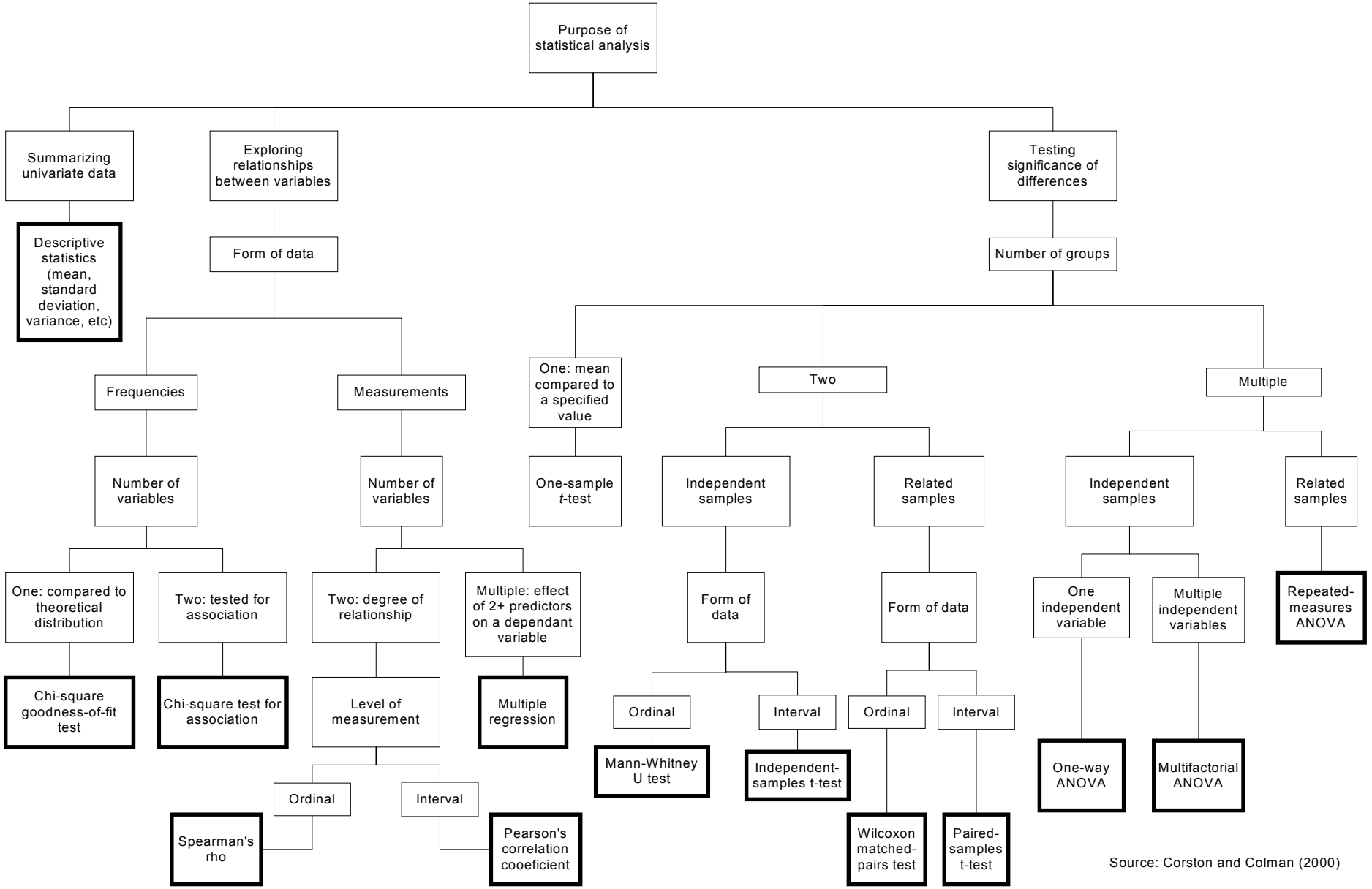


Figure 9. Choosing an appropriate statistical procedure