Lecture 20

Serial Correlation

The following are some sketchy notes about serial correlation. Causation versus association is included in this lecture although it might be best included elsewhere.

The Durbin-Watson statistic

The Durbin-Watson statistic is the standard method for detecting serial correlation in regression data. Positive serial correlation means that positive residuals tend to lead to positive adjacent residuals and negative residuals tend to lead to negative adjacent residuals. Adjacent here means in the order in which the data was collected, usually a time order. If the data is sorted by predictor variable, positive serial correlation could be a result of curvature in the model. It is not, however, the appropriate test for curvature. You should only worry about positive serial correlation when your data was collected in order and you know that order.

The statistic is:

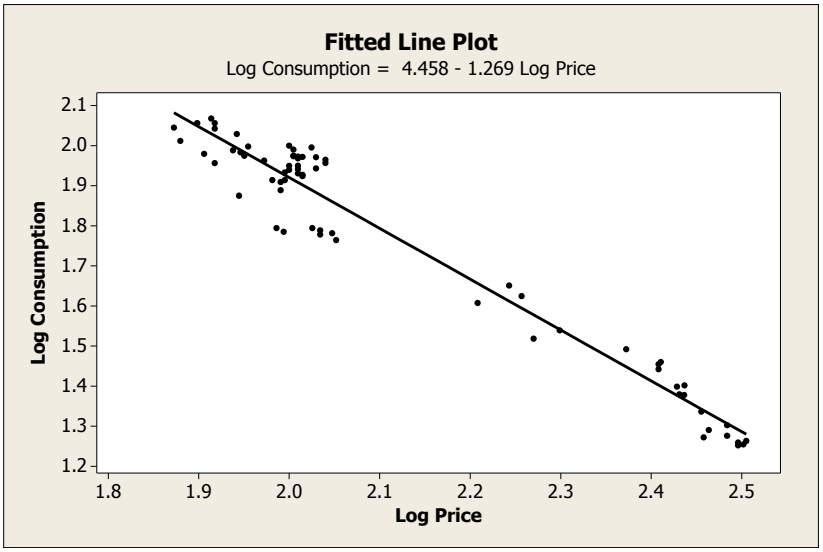
$$d = \frac{\sum_{t=2}^{n} \left(\hat{\epsilon}_t - \hat{\epsilon}_{t-1}\right)^2}{\sum_{t=1}^{n} \hat{\epsilon}_t^2}$$

The expected value of d when there is no correlation amongst the residuals is 2. If d is substantially lower than 2, that is a sign of positive serial correlation - the most likely possibility and the one with the most serious consequences. If you do not correct for positive serial correlation, you are stating more confidence in your results than you really have. This is called anti-conservatism in statistics and statisticians regard it as a serious error. If d is substantially higher than 2, that is a sign of negative serial correlation, but this is a rarer and less of a problem.

Serial correlation is an issue when you take your measurements in order of, say, increasing X value. The most common example of this is for data collected over time.

The tables for determining whether a particular value of the Durbin-Watson statistic is statistically significant or not is located at: http://www.jstor.org/stable/2332325 Unfortunately, there are not clear-cutoffs for significance of the Durbin-Watson statistic. There is a value, d_L so that if $d < d_L$, then you have significant positive serial correlation. There is another value, d_U so that, if $2 > d > d_U$, there is no evidence of positive serial correlation. If $d_L < d < d_U$ you are in a grey area. To check for negative serial correlation, replace the value of d you obtained from the data ($d > 2$) with $4 - d$ and use the same values of d_L and d_U to determine significance/insignificance.

Example: Using the data from the Durbin and Watson paper, originally from Priest, we can consider the regression of (log of) consumption of spirits on the (log of) the relative price (relative to real income, as in a cost of living adjustment). The data are available on the WEB page for this class. The base for these logarithms is 10, by the way.



We see that the data do look like they follow a line. The best fitting (linear) model is

$$\text{Log Consumption} = 4.458 - 1.269\text{Log Price}$$

which translates to the following model on the scale of the original data (the logarithm here is base 10):

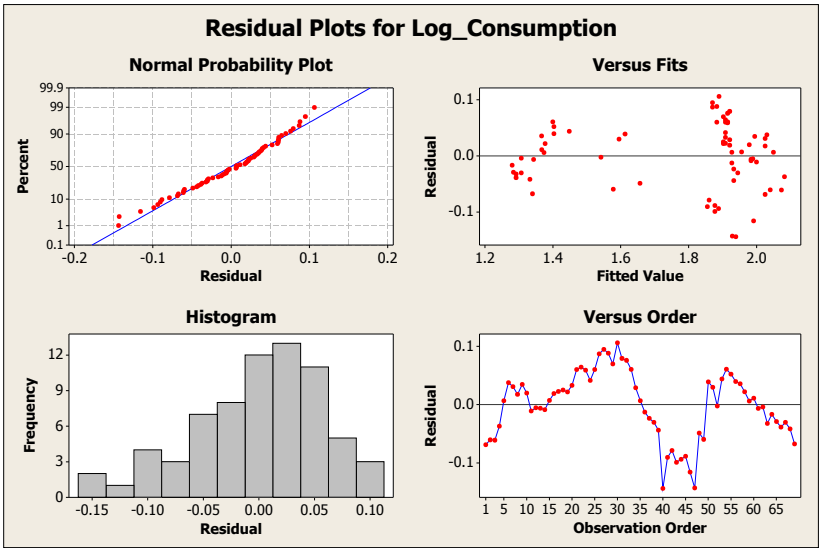
$$\text{Consumption} = \frac{28713.1}{(\text{Price})^{1.27}}$$

Minitab spits out a value for the Durbin Watson statistic, but offers no guidance as to whether it is significant or not. To determine significance, we need to go to Durbin and Watson's tables.

We see that for around 68 data points, the table looks like:

n	$k=1$		$k=2$		$k=3$	
	d_L	d_U	d_L	d_U	d_L	d_U
65	1.57	1.63	1.54	1.66	1.50	1.70
70	1.58	1.64	1.55	1.67	1.52	1.70

We see that for our example, d is much smaller than d_L so we clearly see that there is positive serial correlation. We will look at plots of the residuals in class to see this as well, such as the following:



The regression line itself is still valid, but the estimates of the slope, intercept, confidence, and prediction bands are all much more variable than is being reported without taking the serial correlation into account.

Causation versus Association

So far, we have seen several linear (or allometric) relationships: Spirit Consumption $= \dfrac{28713.1}{(\text{Price})^{1.27}}$, Heart Rate $= \dfrac{365.75}{(\text{Egg Mass})^{0.09396}}$, FEMUR(normal) $= -9.91 + 0.883$BPD(normal) Which ones are causal? Can we tell?

The regression says nothing about cause and effect. It only says that there is an association and it appears to be linear (or linear in the transformed variables). Causation is best determined in the lab where one can manipulate the predictor at will. In these equations, there may be a third variable, a confounding variable, that is controlling both the predictor and the response. There may even be multiple confounding variables. We will talk about "controlling" for other variables when we get to including multiple predictors in two lectures.

References and Readings

[1] J. Durbin and G. S. Watson. Testing for serial correlation in least squares regression. II. *Biometrika*, 38:159–177, 1951.

Exercises for Lecture 20

1. –

2. –