FI=NOT0106

MORE ON SIMPLE REGRESSION:

EXPLANATION OF R^2 , SE OF B, SIGNIFICANCE TESTING.

OVERVIEW

In these notes we examine output from simple regression. We discuss measures of explained and unexplained variance, and statistical significance tests of the b weight and Rsquared and (1-Rsquared). We discuss the assumptions behind significance testing.

STARTING WITH ASSUMPTIONS AND MODEL FOR HYPOTHESIS TESTING

We have been talking about how the purpose of significance testing is to make inferences about population parameters, and how we get to population parameters by means of sampling distributions. You can have sampling distributions of R squareds, of means, or of b weights, or of other parameters.

Let's begin by fully stating our assumptions, and the specific model that we wish to test.

In these last few notes we have been working with an example where

X=AVGPAY and Y=PROPRA85.

Assumptions

We want to test the following null hypothesis if we focus on the B weight:

A. in the population of states the slope of PROPRA85 on AVGPAY=0

NOTE - if are testing the Rsquared, this assumption becomes

In the larger population Rsquared = 0; i.e., that the predictor(s) do not explain any of the variance in the outcome.

This assumption is the only assumption that is in question. Here are the other assumptions that are not in question; they canNOT be empirically tested.

B. We assume that the 50 states for which we have data here represent a random sample from the [fictional] population of 500 states. Cases have been randomly and independently sampled. This is what we mean by the assumption of random sampling.

I know this assumption seems strange because in fact we are working with all 50 US states. But recall some of our earlier tests. For example, in the z test we had a sample of southern states and we were testing to see if the 16 were in fact a random sample from the larger population of 50.

Now that we are working with all 50 states, for our statistical tests to make any sense we must assume -- and this is the assumption we test -- that the 50 states represent a random sample from a larger population of states (500, 1,000, 2,000, pick whatever number you want) where average B = 0

In the case of the nationally representative sample from the Guns in America survey, we would be assuming that the sample represented a random sample of all US households. Since you are working with an extract that includes only gun owners, you would be assuming more specifically that the extract is a representative random sample of all US gunowning households.

C.We assume at least an interval level of measurement. (By the way, what levels of measurement do we have here?)
D.We also make assumptions about error terms in regression. (See Hamilton p. 31). Hamilton calls these i.i.d. errors).

What this means is

- "errors have identical distributions, with 0 mean and comparable variances" for every range of X

- "errors are independent; unrelated to X variables or to the errors of other cases" [this is something you can verify through some special plots]

- "errors are normally distributed" [this is something SPSS can give you information on, and you also can verify through some special plots]

These assumptions about error terms are extremely important in regression models. If they are violated bad stuff [the technical term] can happen. We will spend some time talking about how to verify these assumptions.

E.We assume <u>linearity</u> - that the X-Y relationship is linear, not curvilinear. (We will talk later about how to test this assumption.)
F.That in the population the variables are distributed in a multivariate normal fashion. Each single variable has approximates a normal population, and each pair of variables has a normal bivariate distribution. (For an example see Blalock p. 389)
G.Equal variances of Y within each level of X; this is called <u>homoscedasticity</u>.

<u>Rest of the Model</u>

In addition to the <u>assumptions</u>, we need to complete the model

H.We select an alpha level of $p < .10$. We select this alpha level in order to have an acceptable level of statistical power (1-Type II error rate) given our small sample of 50 cases.
I.We need to decide if we want to do a one tail or a two-tailed alternate hypothesis.

We opt for a one-tailed alternate hypothesis (H1). We expect that states with higher average pay will have higher property crime rates. Our rationale is as follows. Given Felson and Cohen's routine activity theory of crime, we expect that in states where average pay is higher there will be more opportunities for burglary and larceny. So our alternate or research hypothesis is $B > 0$

NOTE - if we are doing a statistical test of Rsquared it is a "no tail" test because Rsquared is <u>always positive</u>, and we are using an F test which is always positive. Explained variance can never be lower than 0 unless some strange things are happening with adjusted Rsquared.

An aside ....

WHAT DOES AN F TEST DO? An F-test is part of an “analysis of variance” or ANOVA. For more details see the STATSOFT weblinks. Basically you are looking at the ratio of two mean squares: the mean squares associated with differences BETWEEN different groups with different scores on the predictor, and the differences WITHIN groups with different scores on the predictor.

The focus is on the outcome variable. That is where we are calculating the mean squares and sums of squares from.

Sums of squares (SS in Hamilton) refers, in the outcome variable, to the sum of squared differences around the mean outcome score. (This is variance before we divide by n of cases). Mean squares (MS) refers to the SS after it has been divided by the appropriate degrees of freedom. See comments in NOT0105 and below about degrees of freedom.

Note that we get the actual B weight with our sample. When we do a statistical test of a B weight we are testing if in the larger population from which the sample was drawn B>0. (See Hamilton p. 42). Stated more technically, we are seeing how likely it is that the B weight obtained belongs on a sampling distribution of B weights where the average B weight=0, for samples n=50. (Again we assume repeated random samples. We have obtained just one of many possible samples. And for each of those samples we could place the resulting B weight on the sampling distribution of B weights. That is why the assumption of repeated random samples is important.) How likely is it that the B weight we get in our actual sample belongs on this sampling distribution of B weights, with mean=0, for n=50?

For the Rsquared when we do a statistical test of Rsquared we are testing to see if in the larger population Rsquared = 0.

LET'S LOOK AT SOME OUTPUT: SOME OF THE BASIC FEATURES

```
* * * * M U L T I P L E R E G R E S S I O N * * * *
Equation Number 1 Dependent Variable.. PROPRA85 PROPERTY CRIME
RATE 1985

Block Number 1. Method: Enter AVGPAY

Variable(s) Entered on Step Number

1.. AVGPAY AVERAGE ANNUAL PAY PER WORKER 1983

Multiple R .46669

R Square .21780

Adjusted R Square .20150

Standard Error 911.49191

Analysis of Variance
```

```
DF Sum of Squares Mean Square

Regression 1 11104017.41392 11104017.41392

Residual 48 39879239.86608 830817.49721

F = 13.36517 Signif F = .0006

------------------ Variables in the Equation ------------------

Variable B SE B Beta T Sig T

AVGPAY .206358 .056446 .466688 3.656 .0006

(Constant) 876.530150 955.057691 .918 .3633

End Block Number 1 All requested variables entered.
```

Multiple R, Rsquared, and Tests

MULTIPLE R represents the multiple correlation between all the predictors and the outcome. When you add in all the predictors, and take into account their overlap, how much do they correlate, **as a group**, with the outcome? Since we are dealing only with one predictor this is the same as the 0-order correlation between X and Y (r). But if you have more than one predictor you will see that sometimes the multiple correlation can be substantially larger. Even though 0 order correlations can be negative, multiple R is always positive.

When you square this term you get MULTIPLE Rsquared. This is also called the coefficient of determination. It tells you the proportion of the Y variance explained by the X variable(s). See Hamilton pp. 38-42.

Rsquared = explained Y variance / total Y variance

Rsquared = variance $Y_{predicted}$ / variance Y_{actual}

Explained variance can range from 0 percent of Y_{actual}'s variance to 100 percent of it. Since it is a squared term it is always positive.

Sums of Squares (SS)

You can see that the last table in the output refers to sums of squares. These are sums of squared differences around the mean.

To get the total SS (TSS) you take all the actual Y scores, compute difference scores [from the mean], square, and add up.

After you have done a regression and have obtained a multiple Rsquared, you can further divide the SS in Y. It now has two portions. First, there is the **explained SS.** This is simply, across all $Y_{predicted}$ cases,

Sum [($Y_{predicted}$ - Y_{mean}) squared] HAMILTON CALLS THIS ESS

Second, there is the residual SS. This is simply, across all Y residual cases

Sum **[($Y_{residual}$) squared]** HAMILTON CALLS THIS RSS

In other words, it is the sum of the squared error terms or the sum of the squared residuals. As Hamilton points out, R squared = ESS/TSS; and (1 - R squared) = RSS/TSS

(1 - R squared) is also called the **coefficient of alienation**; it is the portion of Y's variance that is not explained by the predictors.

SIGNIFICANCE TESTING: R SQUARED

First you do the stat test to see if R squared > 0.

When we have only one regressor or predictor the multiple R squared is the same as the Pearson correlation squared (r squared). We will soon be dealing with more than one predictor, and you will see that the multiple R squared becomes something different; it will refer to the portion of the Y variance that can be explained by all of the predictor or regressor variables, controlling for the overlap between the different predictors.

We will start with more general tests. More specifically, we start with the test of the multiple R squared. (See Darlington p. 113 for the logic of starting with the more general tests).

If your multiple R squared turns out to be significant, allowing you to reject the null hypothesis at the specified alpha level, you can then go on to interpret more specific parts of the results, such as the b weights.

You want to test for the significance of the R squared first, because this is the more general test. If it is not significant, then you do not have as much of a "right" to interpret individual B weights.

The test for R squared is an F test for an analysis of variance. What it asks is: for each level of X, is the mean level of Y equal?

Imagine that scores on your X variable were divided up into four different categories: high, medium high, medium low, and low. What the F test tells you is if the mean scores on Y significantly increase or decrease as you change levels of X. In terms of the F-test, this is your alternate hypothesis, where ---- represents the mean score in a group:

	X VARIABLE RANGE			
	low	medium low	medium high	high
mean of Y	X X	X X X	X X X X	X X X X X X

This is the null hypothesis you hope to reject:

	X VARIABLE RANGE			
	low	medium low	medium high	high
mean of Y	X X X	X X X	X X X	X X X

The F test is the ratio of the average ESS to the average RSS. The average ESS = ESS divided by the appropriate number of degrees of freedom. Since you were only using one predictor to get ESS, df= P = 1 where P = number of predictors. The average ESS is also called the mean square. It is the average squared difference of $Y_{predicted}$ scores.

Stated differently: The significance test for the multiple R squared is a ratio of the average explained sum of squared deviations about the mean to the average residual sums of squared deviations about the mean.

The average RSS = RSS divided by the appropriate degrees of freedom where

df = n - p - 1 where n is the number of cases in the sample, and p = the number of predictors. It is the average square of the residuals.

We call degrees of freedom in the numerator df1, and degrees of freedom in the denominator df2.

$$F = \frac{(ESS/df1)}{(RSS/df2)}$$

Let's say our in our model alpha=.10. We need to find our F critical. We go to table A.4.3 on p.353. We use the lookup values for df1=1, df2=40.

What is our $F_{critical}$ value? $F > 2.84$. We see from the output table above that

$F_{obtained} = 13.365$, $p < .001$.

Therefore we are in the critical region and we reject the null hypothesis that in the population of (e.g.,) 750 states all of the predictors explain 0 percent of the outcome variance. In short there is a connection between our predictor(s) and our outcome, with the predictor(s), average state level pay, explaining 21.8% of the variance in state level property crime rates in 1985.

NOTE THAT THIS F TEST IS NON-DIRECTIONAL. It tells us whether or not overlap exists, but does not tell us more than that. We need to look to specific tests of specific b weights to see about directionality.

A note on degrees of freedom

These different statistical tests will make use of the idea of degrees of freedom or df. When you are looking at a sample statistic you often take into account the degrees of freedom in order to arrive at an unbiased population parameter based on that statistic. "The number of degrees of freedom is equal to the number of quantities that are unknown minus the number of independent equations linking these unknowns." For example in computing a standard deviation you must make use of the sample mean. You have "lost" one degree of freedom in computing that mean. To get unbiased estimates -- i.e., population estimates -- quantities are often divided by their degrees of freedom rather than N.

R squared. When you are testing the significance of the multiple R squared, your degrees of freedom are P for the numerator, where P = the number of regressors (predictors), and N - P - 1 for the denominator. Notice that putting these two together you come up with N - 1 total degrees of freedom. See Darlington Table 5.3. (p. 119).

If you take a sum of squares and divide it by its df you get a mean square. The mean square is an unbiased estimate. (Remember: so we can get back to the population, or, more specifically, the sampling distribution. Here we are getting back to the sampling distribution of mean squares) For the numerator the mean square is an average SS for each regressor. For the denominator the mean square is an average residual SS for each case.

SS / df = MS

The F ratio = MS regression / MS error and is tested with (P, and (N-P-1)) degrees of freedom for the numerator and denominator. This is the F ratio you see in the output table on SPSS.

B WEIGHT

Each sample unstandardized slope is an unbiased estimator of the population slope (Darlington p. 125). Our null hypothesis is that:

$\mathbf{B_{population} = 0.}$

The statistical test to test the significance of a B weight is a t test. You want to see if the b weight is significantly larger or smaller than the null hypothesis of an average B weight of 0 on the sampling distribution of B weights. You are asking: could the B weight that I observe in my regression come from a sampling distribution of B weights where the n of cases is 50, the mean B weight is 0, and where deviations around 0 emerge simply due to sampling fluctuations?

The t test, like the F test, is a ratio of (effect / spread).

$$t = \frac{B_{sample} - B_{population}}{\text{se of B}}$$

but since we are assuming that $B_{population} = 0$, this last term in the denominator drops out and we have

$$t(df) = \frac{B_{sample}}{\text{se of B}}$$

The df for the t test = N - P - 1.

If the t ratio exceeds the critical value required given a specified alpha level, and a specified df, then you can reject the null hypothesis at that alpha level.

T tests can be one tailed or two tailed. The probability levels reported in SPSS are always two tailed. So they always presuppose the alternate hypothesis that b <> 0.

If you want to use a 1 tailed test (either b > 0, or b < 0), you can cut the p level shown in half. Thus if I am doing a one tailed test (b > 0) and the probability level shown for my t value is .08, my real probability level is .04, which is less than .05, which allows me to reject my null hypothesis if I had set my alpha level at .05.

Under "COEFFICIENT" in THE SPSS table you see the constant (A) 876.53. You also see the slope (B), .206. Notice that it is positive. This unstandardized regression coefficient tells us that for every additional dollar in average pay in a state, the property crime rate goes up one fifth of a crime/100,000 population.

This is the result for your sample. If you want to go further, and decide if the slope is significantly different from 0 in the <u>population</u> of (e.g.,) 823 states, you need to look at the results of the t test.

Notice also under STD ERROR there is a standard error for A and B. Let's ignore for now the first one and concentrate on SE for B.

This is a measure of the dispersion of B under the assumptions we are working with. It is the standard error of B, or the standard deviation on the sampling distribution of B weights. You assume that the mean of the B weights in the population, and in the sampling distribution of B weights, = 0.

Since we do not know what is happening in the population, we estimate se of B. (See Hamilton p. 43). This is the formula for <u>bivariate</u> or <u>simple</u> regression.

$$\textbf{SE B=} \quad \frac{s_e}{\textbf{Sq. rt TSS}}$$

It is the standard deviation of the residuals divided by the square root of the total sum of squares.

In other words you are standardizing the standard deviation of the residuals by dividing through by the total sum of squares. Naturally, the larger TSS, the larger you would expect s of error terms to be.

Note how s of e will be smaller the larger your R squared is, all else equal. As R squared gets bigger your error terms get smaller, and therefore have a smaller standard deviation and smaller variance and smaller RSS.

So what is happening in our output here? First, recall our model. We set alpha = .10 and chose a 1 tailed test, given our rationale. We look up $t_{critical}$ for df=n-P-1 = 50-1-1 = 48. Using df=40 we see $t_{critical}$ = +1.303. We see that $t_{obtained}$ = +3.656, exceeding our critical value, putting us well into the region for rejecting the null hypothesis. How unlikely is it that this sample B could reside on a sampling distribution of B weights where the mean B = 0? SPSS tells us p < .001. But since that is two-tailed we know that p < .0005 in our one tailed test.

So we reject the null hypothesis that B=0 in the population of (e.g.,) 823 states. It appears that states with higher average pay have higher reported property crime rates. These results would appear to support Cohen and Felson's routine activity model of property crime. More property crimes per capita occur in states with higher income levels perhaps because there are more items to be burgled and stolen. An alternate explanation could rely on perceived inequality. Perceived inequality will be greater among poor populations if they reside in states where the average pay is higher. This greater perceived inequality may more strongly motivate people to burgle and steal from those better off. What additional variables would you need to do a test to distinguish between these two possible explanations of the result obtained?

SOME MORE FEATURES OF PRINTOUT: ADJUSTED R SQUARED AND SEE

Adjusted R squared

Note that your multiple R squared is not an unbiased estimator of the population R squared. Your adjusted R squared gives you an unbiased estimator of the explained variance of Y.

adj. R squared = R squared - [[P (1 - R squared) / [N - P - 1]]

= .218 - [[1 *(1 - .218) / [48]]

= .2017

So you would say that in the population of (e.g.,) 823 states AVGPAY explains 20.17% of the variance in 1985 reported 1985 property crime rate.

Notice from the above form of the equation that as you have more predictors (P increases), your adjusted R squared will be smaller relative to your R squared. Notice also that all else equal as you N of cases increases your adjusted R squared, relative to the R squared, will shrink less.

You should not confuse the adjusted R squared with the shrunken R squared or the cross validated R squared. The latter is more concerned with how well you could do predicting a second sample using an equation derived from a first sample.

SEE

You also see something called SEE: standard error of the estimate.

"The true residual standard deviation, the square root of the residual variance, is estimated by the square root of the MSE; this statistic is called the standard error of the estimate" (SEE) (Darlington, p. 120)

Residual SD = Sq. Rt. Residual VAR = Sq. Rt. MSE = SEE

squaring all sides of this would give you:

Residual VAR = MSE = SEE squared

SPSS tells you SEE=911.492. If we square this we get 830817.67. You find this under the analysis of variance for MEAN SQUARE-RESIDUAL. This is MSE, the mean of your squared error term, or the variance of your residuals.

So:

MSE = variance of residuals

SEE = unbiased estimate of standard deviation of the residuals

Glossary

(1-Rsquared) (coefficient of alienation; unexplained or residual variance)

adjusted Rsquared

ESS

F test of Rsquared

homoscedasticity

i.i.d. error terms

Rsquared (coefficient of determination; explained variance) residual

RSS

s of e

se of B

SEE

t-test of B

TSS