Announcement

# How to Read the Output From Multiple Linear Regression Analyses

Here's a typical piece of output from a multiple linear regression of homocysteine (LHCY) on vitamin B12 (LB12) and folate as measured by the CLC method (LCLC). That is, vitamin B12 and CLC are being used to predict homocysteine. A (common) logarithmic transformation had been applied to all variables prior to formal analysis, hence the initial L in each variable name, but that detail is of no concern here.

Dependent Variable: LHCY

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Prob>F |
|---|---|---|---|---|---|
| Model | 2 | 0.47066 | 0.23533 | 8.205 | 0.0004 |
| Error | 233 | 6.68271 | 0.02868 | | |
| C Total | 235 | 7.15337 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.16936 | R-square | 0.0658 |
| Dep Mean | 1.14711 | Adj R-sq | 0.0578 |
| C.V. | 14.76360 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| |
|---|---|---|---|---|---|
| INTERCEP | 1 | 1.570602 | 0.15467199 | 10.154 | 0.0001 |
| LCLC | 1 | -0.082103 | 0.03381570 | -2.428 | 0.0159 |
| LB12 | 1 | -0.136784 | 0.06442935 | -2.123 | 0.0348 |

**Parameter Estimates**.

The column labeled **Variable** should be self-explanatory. It contains the names of the predictor variables which label each row of output.

**DF** stands for **degrees of freedom**. For the moment, all entries will be 1. Degrees of freedom will be discussed in detail later.

The **Parameter Estimates** are the regression coefficients. The regression equation is

```
LHCY = 1.570602 - 0.082103 LCLC - 0.136784 LB12
```

To find the predicted homocysteine level of someone with a CLC of 12.3 and B12 of 300, we begin by taking logarithms. Log(12.3)=1.0899 and log(300)=2.4771. We then calculate

```
LHCY = 1.570602 - 0.082103 1.0899 - 0.136784 2.4771
     = 1.1423
```

Homocysteine is the anti-logarithm of this value, that is, $10^{1.1423} = 13.88$.

The **Standard Errors** are the standard errors of the regression coefficients. They can be used for hypothesis testing and constructing confidence intervals. For example, confidence intervals for LCLC are constructed as (-0.082103± k 0.03381570), where k is the appropriate constant depending on the level of confidence desired. For example, for 95% confidence intervals based on large samples, k would be 1.96.

The **T** statistic tests the hypothesis that a population regression coefficient is 0 **WHEN THE OTHER PREDICTORS ARE IN THE MODEL**. It is the ratio of the sample regression coefficient to its

standard error. The statistic has the form (estimate - hypothesized value) / SE. Since the hypothesized value is 0, the statistic reduces to Estimate/SE. If, for some reason, we wished to test the hypothesis that the coefficient for LCLC was -0.100, we could calculate the statistic (-0.082103-(-0.10))/0.03381570.

**Prob > |T|** labels the **P values** or the **observed significance levels** for the t statistics. The degrees of freedom used to calculate the P values is given by the Error DF from the ANOVA table. The P values tell us whether a variable has statistically significant predictive capability in the presence of the other variables, that is, whether it adds something to the equation. In some circumstances, a nonsignificant P value might be used to determine whether to remove a variable from a model without significantly reducing the model's predictive capability. For example, if one variable has a nonsignificant P value, we can say that it does not have predictive capability in the presence of the others,remove it, and refit the model without it. These P values should not be used to eliminate more than one variable at a time, however. A variable that does not have predictive capability in the presence of the other predictors may have predictive capability when some of those predictors are removed from the model.

## The Analysis of Variance Table

The **Analysis of Variance** table is also known as the **ANOVA table** (for ANalysis Of VAriance). There is variability in the response variable. It is the uncertainty that would be present if one had to predict individual responses without any other information. The best one could do is predict each observation to be equal to the sample mean. The amount of uncertainty or variability can be measured by the Total Sum of Squares, which is the numerator of the sample variance. The ANOVA table partitions this variability into two parts. One portion is fitted by (many incorrectly say "explained by") the model. It's the reduction in uncertainty that occurs when the regression model is used to predict the responses. The remaining portion is the uncertainty that remains even after the model is used. The model is considered to be statistically significant if it can account for a large amount of variability in the response.

The column labeled **Source** has three rows, one for total variability and one for each of the two pieces that the total is divided into--**Model,** which is sometimes called **Regression**, and **Error**, sometimes called **Residual**. The **C** in **C Total** stands for **corrected**. Some programs ignore the **C** and label this **Total**. The C Total **Sum of Squares** and **Degrees of Freedom** will be the sum of Model and Error.

**Sums of Squares:** The total amount of variability in the response can be written $\Sigma$(y-ybar)², where ybar is the sample mean. (The "Corrected" in "C Total" refers to subtracting the sample mean before squaring.) If we were asked to make a prediction without any other information, the best we can do, in a certain sense, is the sample mean. The amount of variation in the data that can't be accounted for by this simple method of prediction is given by the Total Sum of Squares.

When the regression model is used for prediction, the amount of uncertainty that remains is the variability about the regression line, $\Sigma$(y-yhat)². This is the Error sum of squares. The difference between the Total sum of squares and the Error sum of squares is the Model Sum of Squares, which happens to be equal to $\Sigma$(yhat-ybar)².

Each sum of squares has corresponding degrees of freedom (DF) associated with it. Total df is one less than the number of observations, n-1. The Model df is the number of independent variables in the model, p. The Error df is the difference between the Total df (n-1) and the Model df (p), that is, n-p-1.

The **Mean Squares** are the Sums of Squares divided by the corresponding degrees of freedom.

The **F Value** or **F ratio** is the test statistic used to decide whether the model as a whole has statistically significant predictive capability, that is, whether the regression SS is big enough, considering the number of variables needed to achieve it. **F** is the ratio of the Model Mean Square to the Error Mean Square. Under the null hypothesis that the model has no predictive capability--that is,

that all population regression coefficients are 0 simultaneously--the F statistic follows an F distribution with *p* numerator degrees of freedom and *n-p-1* denominator degrees of freedom. The null hypothesis is rejected if the F ratio is large. Some analysts recommend ignoring the P values for the individual regression coefficients if the overall F ratio is not statistically significant, because of the problems caused by multiple testing. I tend to agree with this recommendation with one important exception. If the purpose of the analysis is to examine a particular regression coefficient after adjusting for the effects of other variables, I would ignore everything but the regression coefficient under study. For example, if in order to see whether dietary fiber has an effect on cholesterol, a multiple regression equation is fitted to predict cholesterol levels from dietary fiber along with all other known or suspected determinants of cholesterol, I would focus on the regression coefficient for fiber regardless of the overall F ratio. (This isn't quite true. I would certainly wonder why the overall F ratio was not statistically significant if I'm using the known predictors, but I hope you get the idea. If the focus of a study is a particular regression coefficient, it gets most of the attention and everything else is secondary.)

The **Root Mean Square Error** (also known as **the standard error of the estimate**) is the square root of the Residual Mean Square. It is the standard deviation of the data about the regression line, rather than about the sample mean.

**$R^2$** is the squared multiple correlation coefficient. It is also called the **Coefficient of Determination**. $R^2$ is the ratio of the Regression sum of squares to the Total sum of squares, RegSS/TotSS. It is the proportion of the variability in the response that is fitted by the model. Since the Total SS is the sum of the Regression and Residual Sums of squares, $R^2$ can be rewritten as (TotSS-ResSS)/TotSS = 1-ResSS/TotSS. Some call $R^2$ *the proportion of the variance explained by the model*. I don't like the use of the word *explained* because it implies causality. However, the phrase is firmly entrenched in the literature. If a model has perfect predictability, $R^2=1$. If a model has no predictive capability, $R^2=0$. (In practice, $R^2$ is never observed to be exactly 0 the same way the difference between the means of two samples drawn from the same population is never exactly 0.) R, the multiple correlation coefficient and square root of $R^2$, is the correlation between the observed values (y), and the predicted values (yhat).

As additional variables are added to a regression equation, $R^2$ increases even when the new variables have no real predictive capability. The **adjusted-$R^2$** is an $R^2$-like measure that avoids this difficulty. When variables are added to the equation, adj-$R^2$ doesn't increase unless the new variables have additional predictive capability. Where $R^2$ is 1 - ResSS/TotSS , we have
adj $R^2$ = 1 - (ResSS/ResDF)/(TotSS/(n-1)), that is, it is 1 minus the ratio of (the square of the standard error of the estimate) to (the sample variance of the response). Additional variables with no explanatory capability will increase the Regression SS (and reduce the Residual SS) slightly, except in the unlikely event that the sample partial correlation is *exactly* 0. However, they won't tend to decrease the standard error of the estimate because the reduction in Residual SS will be accompanied by a decrease in Residual DF. If the additional variable has no predictive capability, these two reductions will cancel each other out.

---



Last modified: 05/23/2012 08:23:08.