MULTILINEAR REGRESSION

Maurice J. Dupré
Department of Mathematics
New Orleans, LA 70118
email: mdupre@tulane.edu
28 November 2012

Abstract. We give a simple discussion of regression, multilinear regression and diagnositic anaylsis using the EXCEL SUMMARY OUTPUT.

1. REGRESSION IN A NUTSHELL

Suppose that we have an algebra of unknowns, denoted $\boldsymbol{\mathcal{A}}$, that we have a subset $\mathcal{K}$ of $\boldsymbol{\mathcal{A}}$ consisting of unknowns that we feel we know a lot about or can reasonably determine or easily observe, and suppose that Y is an unknown which we are interested in determining. If we are given the values of the unknowns in $\mathcal{K}$, can we use that information to help us guess the value of Y? For instance, if I am trying to guess your age, might some other information be of help? Given this extra information, how do we make the best use of it? The idea in **Regression** is to try to choose the unknown in $\mathcal{K}$ which is "closest" to Y and use its value as a guess for the value of Y. If Y_R in $\mathcal{K}$ is the chosen unknown to use in place of Y, here, we call Y_R a **Regression Model** for Y. We call the difference $Y - Y_R$ the **Error** or **Residual** and denote it by Y_E, so

$$Y = Y_R + Y_E.$$

The next question here is how to choose Y_R among all the unknowns in $\mathcal{K}$? The simplest criteria is to minimize the expected squared error $E(Y_E^2)$, which is called the **Method of Least Squares**.

2. INTRODUCTION TO BASICS OF REGRESSION

In more detail, the basic idea of **Regression** is to attempt to predict the value of an unknown Y from knowledge of the values of unkowns $X_1,\ X_2,\ X_3, ..., X_k$. We should then look for a function of k real variables, say f, so that knowing x_j is the value of the unknown X_j for each $j \leq k$ would then lead us to guess the value y for the unknown Y, where the number y is calculated using the function f as

$$y = f(x_1, x_2, x_3, ..., x_k).$$

In this setting, we refer to Y as the **Dependent or Objective Variable** and each X_j is called an **Independent or Explanatory Variable**.

Of course the first question here is naturally: **How do we choose the function f?**

In many cases, we have some experience with the unknowns which may dictate at least a specific form for the function f. The function f would then be built with various algebraic expressions involving the symbols $x_1, x_2, x_3, ..., x_k$, as well as other parameters (numbers to be chosen), say $\beta_0, \beta_1, \beta_2, \beta_3, ..., \beta_l$. We then really have a function f of $k + l$ variables $x_1, ..., x_k, \beta_0, \beta_1, ..., \beta_l$. This means our guess y for the value of Y is

$$y = f(x_1, x_2, x_3, ..., x_k, \beta_0, \beta_1, \beta_2, ..., \beta_l).$$

In effect, we are really forming the new unknown Y_R, called the **Regression Model** where

$$Y_R = f(X_1, X_2, X_3, ..., X_k, \beta_0, \beta_1, \beta_2, ..., \beta_l),$$

and all the various possible choices for the parameters $\beta_0, \beta_1, \beta_2, \beta_3, ..., \beta_l$ lead to all the various possible regression models forming the set $\mathcal{K}$ of unknowns which are the candidates for the best regression model.

Our next problem is then to try and choose the parameter values $\beta_0, \beta_1, \beta_2, ..., \beta_l$ so as to "minimize" the error $= Y - y$. At least we would like to have for any $x_1, x_2, ..., x_k$, that

$$y = E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3, ..., X_k = x_k).$$

But

$$E(Y_R|X_1 = x_1, X_2 = x_2, X_3 = x_3, ..., X_k = x_k) = f(x_1, x_2, x_3, ..., x_k, \beta_0, \beta_1, \beta_2, ..., \beta_l) = y,$$

so this means that we need

$$E(Y_R) = E(Y).$$

But, with a little more careful thought here, we realize the sense in which this error is to be minimized involves a choice as to how to define a "distance" between Y and Y_R. The simplest most obvious choice here is to use the **Expected Squared Error**, so we would choose the parameters $\beta_0, \beta_1, ..., \beta_l$ so as to minimize

$$\sigma_e^2 = E([Y - Y_R]^2).$$

Let us denote the error in our model by Y_E, that is,

$$Y_E = Y - Y_R,$$

so

$$Y = Y_R + Y_E,$$

and

$$\text{if } E(Y_R) = E(Y), \text{ then } E(Y_E) = 0 \text{ and } \sigma_e^2 = E(Y_E^2).$$

Consider now, that if we notice some correlation of Y_E and Y_R, we would seek to include that correlated part of the error somehow with our model, so it is fairly clear that an optimal regression model would have Y_E and Y_R uncorrelated. This means that

$$\sigma_Y^2 = \sigma_e^2 + \sigma_{Y_R}^2.$$

In practice, we may not know enough about the unknowns to be able to directly calculate σ_e, but as with anything involving expectation our recourse is to use sample data from a sample of some reasonable size, denoted n. In this case, we have our unknowns actually being replaced by unknowns on a sample space S of n equally likely outcomes, so the set of all unknowns on this sample space is denoted $\mathbb{R}^n$, as it is an $n-$dimensional space. Notice that the result of taking a sample of observations of Y is a list of n numerical values, and the space of such lists is $n-$dimensional. In effect, the sampling replaces Y by a list $\mathbf{y}$ and each of the unknowns X_j is replaced by a list $\mathbf{x}_j$, which results in Y_R being replaced by a list of numbers $\mathbf{y}_R$ and the process of minimizing the expected squared error becomes replaced by the problem of minimizing the squared distance between $\mathbf{y}$ and $\mathbf{y}_R$ in $\mathbb{R}^n$. Solving this should result in specific values $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_l$, which for a large sample size n we would expect should be reasonably close to the actual optimal values $\beta_0, \beta_1, \beta_2, ..., \beta_l$. We then use the sample regression model

$$\hat{Y}_R = f(X_1, X_2, X_3, ..., X_k, \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, ..., \hat{\beta}_l),$$

as that is certainly the best we can do based only on the data.

The notational convention used here is very common in statistics, that is if ξ is a population parameter, then when we estimate it using sample data, its estimate would be denoted $\hat{\xi}$. For instance, if we use μ_Y

to denote the expected value of Y, then the sample mean $\bar{y}$ may also be denoted by $\hat{\mu}_Y$, and the sample variance s_y^2 may also be denoted by $\hat{\sigma}_Y^2$, since s_y^2 is an unbiased estimator of σ_Y^2.

3. BASICS OF MULTILINEAR REGRESSION

At this point, we have decided that the regression model is to be evaluated by the criteria of minimizing the expected squared error, so finally, we must actually make the choice of the form of f. A particularly simple choice is called the **Multilinear Model** formed by choosing f to be the simple **multilinear function**, with $k = l$, and

$$y = f(x_1, x_2, x_3, ..., x_k, \beta_0, \beta_1, \beta_2, \beta_3, ..., \beta_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k.$$

This means that our regression model Y_R is simply

$$Y_R = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k.$$

In this case we can see that as the constants $\beta_0, \beta_1, \beta_2, ..., \beta_k$ vary over $\mathbb{R}^{k+1}$, that the resulting possible regression models vary throughout an m dimensional linear space, $\mathcal{K}$, where $dim(\mathcal{K}) = m \leq k+1$, with the most likely situation being $m = k+1$. This means that the optimal choice for the regression model is the point of this space closest to Y. This is found by simply dropping a perpendicular to the linear space of possible models, and this means that for the resulting optimal regression model, the error or residual Y_E is perpendicular to $\mathcal{K}$. Since the linear space $\mathcal{K}$ of possible models includes the constants, as well as the regression model Y_R, in particular, this certainly guarantees that Y_E and Y_R are perpendicular, and that $E(Y_E) = 0$, so

$$E(Y) = E(Y_R), \text{ and } \sigma_Y^2 = \sigma_e^2 + \sigma_{Y_R}^2.$$

Likewise, regarding the sample data, in $\mathbb{R}^{n+1}$, the same would apply, namely, the linear space of possible sample regression models vary over $k+1$ dimensions inside $\mathbb{R}^{n+1}$, and likewise the optimal sample regression model is chosen so that $\mathbf{y}_R$ and $\mathbf{y}_E$ are perpendicular in $\mathbb{R}^{n+1}$ which guarantees that $\mathbf{y}$ and $\mathbf{y}_R$ have the same sample mean and that the corresponding sample errors called **Residuals** have sample mean zero.

We should notice here the significance of the coefficient parameters $\beta_0, \beta_1, \beta_2, \beta_3, ..., \beta_k$. In the expression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k,$$

If we set $x_1 = x_2 = x_3 = ... = x_k = 0$, then the resulting value of y is $y = \beta_0$, and we call β_0 the **Regression Intercept**. On the other hand, for any $x_1, x_2, x_3, ..., x_k$ if a particular variable, say x_j is increased by one unit, then the value of y increases by β_j. Thus, β_j is the **Rate of change** of y with respect to change in x_j, or the increase in y per unit increase in x_j.

Given that the methods of linear algebra can be effectively used to calculate the regression model Y_R, our next job is to assess the utility of the regression model. We define the **Multilinear Correlation** as the correlation of Y and Y_R, which we denote by ρ. The **Coefficient of Determination** is the square of the multilinear correlation, ρ^2. Think now in terms of simple linear regression where we have only a single explanatory variable which is Y_R itself. We know that

$$\sigma_{Y_R}^2 = \rho^2 \sigma_Y^2,$$

so the equation

$$\sigma_Y^2 = \sigma_{Y_R}^2 + \sigma_e^2$$

gives

$$\sigma_Y^2 = \rho^2 \sigma_Y^2 + \sigma_e^2$$

which we can easily solve for ρ^2 giving

$$\rho^2 = 1 - \frac{\sigma_e^2}{\sigma_Y^2}.$$

Returning now to the sample data and the resulting estimated regression model $\hat{Y}_R$, we see that we would estimate σ_Y^2 as the sample variance s_y^2 and we also need to estimate σ_e^2 from the sample data. To get an idea how to proceed here, let us first assume that we are dealing with a population of size N and that our sample in fact consists of the whole population, so that $n = N$. Then, the variance of the residual is simply the sum of the squared errors divided by N. We denote the sum of the squared errors for the sample data by SSE. We denote the sum of the squared deviations of the sample data $\mathbf{y}$ of values of Y from the sample mean $\bar{\mathbf{y}}$ by SSY, so we likewise denote by SSR the sum of the squared deviations of the sample regression model from the sample mean. Thus,

$$\sigma_Y^2 = \frac{SSY}{N},\ \sigma_{Y_R}^2 = \frac{SSR}{N},\ \text{and}\ \sigma_e^2 = \frac{SSE}{N},\ \text{in case}\ n = N.$$

and therefore the coefficient of determination is, after cancelling factors of N,

$$\rho^2 = 1 - \frac{SSE}{SSY}.$$

We can also note here that as $\sigma_Y^2 = \sigma_{Y_R}^2 + \sigma_e^2$, multiplying by N gives

$$SSY = SSR + SSE.$$

But this last equation must also be true even if $n < N$, since the result applies to the case where the sample of size n is taken as the whole population-anything true for any whole population must be true for any sample treated as a whole population, and therefore also, for any sample data,

$$1 - \frac{SSE}{SSY} = \frac{SSY - SSE}{SSY} = \frac{SSR}{SSY}.$$

However, for samples of size $n < N$, we have

$$s_y^2 = \frac{SSY}{n-1}\ \text{best estimates}\ \sigma_Y^2,$$

because the number of **Degrees of Freedom** in the SSY is $n - 1$. Likewise, the degrees of freedom in SSE can be shown to be $n - (k + 1)$ because the computation of the SSE involves all the $k + 1$ estimated regression coefficients and each replacement of a true regression coefficient by a sample estimate forces the loss of one degree of freedom. Thus we write

$$df(SSY) = n - 1\ \text{and}\ df(SSE) = n - (k + 1).$$

Since the sample regression model and the sample residuals are perpendicular and sum to the sample data for the objective variable, it follows that

$$df(SSY) = df(SSR) + df(SSE),\ \text{and therefore}\ df(SSR) = k,$$

that is, the SSR has degrees of freedom equal to the number of explanatory variables. Thus our best estimates from the sample data are

$$\hat{\sigma}_Y^2 = \frac{SSY}{df(SSY)},\ \hat{\sigma}_R^2 = \frac{SSR}{df(SSR)},\ \text{and}\ \hat{\sigma}_e^2 = \frac{SSE}{df(SSE)}.$$

In multilinear regression and in other statistical applications involving the analysis of variance (ANOVA), we must deal with various sums of squares, as here we have three different sums of squares (SS), and generally, each will have its associated degrees of freedom. The variance estimate is then obtained by dividing the SS by its degrees of freedom, the result being termed the mean sum of squares, denoted MS. For instance, we would write

$$MSY = \frac{SSY}{df(SSY)},\ MSR = \frac{SSR}{df(SSR)},\ \text{and } MSE = \frac{SSE}{df(SSE)}.$$

With this notation, commonly used in computer readouts for ANOVA, then

$$\hat{\sigma}_Y^2 = MSY,\ \hat{\sigma}_R^2 = MSR,\ \text{and } \hat{\sigma}_e^2 = MSE.$$

It is customary to denote the sample estimate of ρ in multlinear regression by R(adjusted), that is, the estimate of the coefficient of determination ρ^2 from the data is

$$\hat{\rho}^2 = R^2(\text{adjusted}) = 1 - \frac{MSE}{MSY} \text{ so the correlation estimate is } \hat{\rho} = R(\text{adjusted}),$$

whereas the sloppier estimate of the coefficient of determination given by

$$R^2 = \frac{SSR}{SSY}$$

is the "unadjusted" R^2, which is in effect treating the sample as if it is the whole population. Thus, for very large samples, the adjusted R-square is usually very close to the R-square, so for large samples, it is enough to examine R-square as the estimate of the true coefficient of determination, ρ^2.

Keep in mind that the adjusted R^2 is the estimate of the coefficient of determination from the sample data which tells us the fraction of variation in Y that is accounted for with the regression model $\hat{Y}_R$, and this is a number between 0 and 1. Therefore, we want the coefficient of determination and likewise the adjusted R^2 to be a number near 1, that is the closer the coefficient of determination is to 1, the more variation in Y is being captured by the variation in the regression model, and therefore the better the regression model. Likewise, the unadjusted R^2 is hopefully near 1. Of course, "near" 1 in certain situations might be as low as only 0.3, the specifics of the application really dictate what is needed here.

But, beyond the coefficient of determination, we also have the regression coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$, which are themselves actually functions of the sample data, and therefore are themselves actually random variables which can be standardized. Of course the variance of $\hat{\beta}_j$ can only be estimated from the sample data, and that estimate actually involves all the estimated values $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$, which means the appropriate sum of squares also has $n - (k+1)$ degrees of freedom. That is, for normally distributed unknowns,

$$\frac{s_{\hat{\beta}_j}^2}{\sigma_{\hat{\beta}_j}^2} \text{ has the } \chi_d^2, \text{ distribution for } d = n - (k+1) \text{ degrees of freedom.}$$

The result is that given a hypothetical value b_j for β_j, then $\hat{\beta}_j$ itself can be standardized using the sample data to have the student$-t$ distribution for $n - (k+1)$ degrees of freedom using the formula

$$t_{data} = \frac{\hat{\beta}_j - b_j}{s_{\hat{\beta}_j}},\ df = n - (k+1).$$

Thus to form a confidence interval for β_j with confidence $= C$, first find the proper $t-$critical value for this level of confidence, denoted t_C and $n - (k+1)$ degrees of freedom, and then the margin of error, ME is simply

$$ME = [t_C] s_{\hat{\beta}_j}$$

which results in the confidence interval

$$\beta_j = \hat{\beta}_j \pm ME = \hat{\beta}_j \pm [t_C] s_{\hat{\beta}_j}.$$

Likewise, to perform the two-tail hypothesis test

$$H_0 : \beta_j = b_j \text{ versus } H_a : \beta_j \neq b_j,$$

we have the P-Value of the data is simply

$$\text{P-Value} = 2P(t \geq |t_{data}| \text{ Given } df = n - (k+1)),$$

and we can easily make the obvious modifications here to perform one-tail tests.

4. EXCEL SUMMARY OUTPUT FOR MULTILINEAR REGRESSION

When performing multilinear regression using EXCEL, we begin by labeling the columns with the names of the various variables in cells of an upper row, with the dependent variable in the left most column. Then simply put the data for each variable in the column beginning with the cell underneath the variable name. Make sure each row contains the data for a single member of the population. Then clicking the DATA tab, if the Data Analysis Toolpak has been added on to the EXCEL software package, then there will appear a tab button labelled DATA ANALYSIS. Clicking the DATA ANALYSIS button then brings up a drop-down menu from which you highlight and click on REGRESSION. This brings up a dialog box in which to enter the data and make choices for the output, and the first time you do this, just check all the boxes. That way you get everything and you can experiment to see what you can use. To enter the data, begin by clicking the button next to the space marked DEPENDENT VARIABLE and then highlight the dependent variable column INCLUDING THE LABEL for the dependent variable. Then for the INDEPENDENT VARIABLE highlight the whole rectangle of cells including the column headings for all columns of explanatory variables. Then just hit the enter button and the EXCEL software will produce a new worksheet. You will find on this worksheet a table labeled SUMMARY OUTPUT and underneath a table labelled ANOVA and underneath that a table labelled RESIDUAL OUTPUT. The first thing to notice at the top of the SUMMARY OUTPUT table is MULTIPLE R which is the value of R computed from the data, and underneath appear R-square and Adjusted R-square, the standard error, and Observations (=n, the sample size). Next in the table labelled ANOVA which is an acronym for ANALYSIS OF VARIANCE that is ANalysis Of VAriance, appears the more detailed results used in computing R and R^2, namely, the column giving the degrees of freedom for the Regression, the Residual, and the Total, the SS or Sum of Squares of each, and then the MS which is SS/df. The next column has the $F-$ratio

$$F = \frac{MSR}{MSE}.$$

Clearly we want this $F-$ratio to be large, as ideally, MSE is near zero and MSR is close to σ_Y^2, and SSR is nearly all of SSY. On the other hand, for a regression model with little value most of the variation of Y will be in the residuals that is the errors, so MSR will be small and the denominator MSE will be large making the $F-$ratio have a value near zero. The $F-$statistic can be looked up in standard tables to see how good the model is. However, the last column of the ANOVA table has the significance F, so you do not even need the $F-$statistical table. Thus, the *Significance* F in the table is actually the significance of the data or equivalently, the P-Value of the data as evidence against the null hypothesis that all regression coefficients are actually zero.

Underneath the table with the degrees of freedom is another table for the Regression Coefficients which has no label but its rows to the left are labelled with Intercept followed by the names of the explanatory variables used. The next column then has the sample values for all the regression coefficients followed by the column giving the standard error for each regression coefficient followed by the $t-$statistic using the hypothetical value zero for the coefficient followed by the P-Value for the null hypothesis that the coefficient is actually zero, followed by lower and upper boundaries for the 95 percent confidence interval for the true value of the regression coefficient. Thus to find the margin of error, ME for a confidence interval with confidence C, we simply look up the appropriate $t-$critical value just as for any other confidence interval, using $df = n - (k+1)$, and then

$$ME = t \cdot s,$$

where s is the standard error in the SUMMARY OUTPUT table for the given regression coefficient.

Regarding the P-Values in the regression coefficient table, since the null hypothesis is that the coefficient is zero, any time we see a P-Value above 0.05, we should be suspicious that the particular coefficient might actually be zero so that the particular variable is actually not explaining any of the variation in the dependent variable. Such variables with high P-Value are good candidates for being eliminated from the model.

Finally, the last table of the regression output is the table labelled RESIDUAL OUTPUT. The first column simply lists the numbers 1 through n, and the second column actually gives the values of the dependent variable predicted by the regression model followed by the column of residuals, that is, the differences between the values predicted by the model and the values of the dependent variable actually in the data (the left most column in the original data). Since the expected residual is zero and the sample mean residual is zero, it follows that the standard residual is simply the residual divided by the square root of MSE:

$$t_j = \frac{(y_j - \hat{y_j})}{\sqrt{MSE}},\ j \leq n.$$

However, the values reported by the EXCEL SUMMARY OUTPUT compute this using $n-1$ degrees of freedom which is incorrect, as the correct number of degrees of freedom is $n-(k+1)$. For large samples this does not make a big difference, but we need to keep in mind that this part of the summary output is not correct.

5. SUMMARY OUTPUTS FOR MULTILINEAR REGRESSION

Besides EXCEL, there are a number of computer software packages and online tools for doing statistical analysis ranging from simple tools that are free online to industry standard packages such as SAS and S. In case of doing multilinear regression quite often you are simply faced with understanding the summary output as a real statistician has done most of the work of handling the actual data electronically. These summary outputs all look a little different but have a common general form. Usually the first thing near the beginning will be the values some or all of R, R^2, and R^2_a. You will see "R-square" or"R-sq" for R^2 and "R-square adjusted" or "R-sq-adj" or any other obvious reference for R^2_a. That is the first thing to look for. You want these near 1.

Next you will probably see the value of n indicated nearby as "sample size" maybe or some obvious reference to the size of the sample. The various sums of squares are usually given in a table similar in form to that for the EXCEL ANOVA. That is you will see a row for the errors which might be indicated by the word errors or the word residual(s), a row for the model which might be indicated by the word model or the word regression, and usually last, a row for the objective, typically indicated as "total". Across the top of such a small table you will usually see column labels, SS for sum of squares, maybe DF for degrees of freedom or some other obvious abbreviation for degrees of freedom, a column labeled MS for mean sum of squares, a column labelled F, and a column labelled p-value or maybe simply "p". In the row for the model under the SS appears the sum of squared deviations the model makes from the sample mean of the objective variable data, whereas in the row for the errors under SS appears the sum of squared errors the model makes from the actual sample values of the objective variable, and then for the row indicated by total you will notice the sum of the other two rows, as we know $SSR + SSE = SST$. Likewise, in the column giving the degrees of freedom, you will notice the value in the model row gives simply the number of predictor variables, as we know that is its number of degrees of freedom. Then for the degrees of freedom in the error row you will see the difference between the total degrees of freedom, $n-1$, which is in the row for the totals, as we know the degrees of freedom for the errors plus the number of degrees of freedom for the model is the total number of degrees of freedom. Then under the MS column you will typically see the SS divided by the DF for each row giving the corresponding mean sum of squares. The MST is often left out because the main objective is usually to compute F which is MSR/MSE. So the column labelled F usually simply has that quotient reported giving the value of the $F-$statistic.

Now, keep in mind, that if the model is totally meaningless, the data will still lead to a regression model, so we should first look to test this overall null hypothesis that the model is of no help. Well, if we assume the model is no help, then the scattering of values of the model is just more errors, so the SSR and SSE would have the same expected values and thus calling this σ^2_e, on dividing through the equation $SSR+SSE = SST$

by σ_e^2 we get a sum of chi-squares giving the chi-square for the total, and thus the MSR/MSE under the null hypothesis does in fact have the $F(d_R, d_E)-$distribution. Here I use d_E for the degrees of freedom in SSE and d_R for the degrees of freedom in SSR. Of course, then d_R is simply the number of explanatory variables. Therefore, under the null hypothesis that the model is useless, both numerator and denominator of the $F-$ ratio

$$F = \frac{MSR}{MSE}$$

are estimating the same σ_e^2 and so this ratio should end up near 1 indicating a useless model. On the other hand, for a good model, the MSR should be large capturing the σ_Y^2 and the denominator should be near σ_e^2 which for a good model would be small leading to a large value for F. Thus the overall p-value of the model as indicated by the data is

$$P(F(d_R, d_E) > F)$$

which you could look up in a table for the $F-$ distribution, but this p-value is typically reported in the last column for your convenience.

Of course, in any model, some of the explanatory variables may individually be very good while others are relatively useless, so in addition to the tabulated sums of squares and the $F-$ratio there will be a table where the rows are labelled by the various explanatory variables, followed by a column of their coefficient values in the model followed by a column of standard errors for those coefficients computed from the data, and as in the EXCEL summary output, typically the p-value for the null hypothesis that the indicated coefficient could in fact simply be zero. Thus, whenever we see a p-value above our working level of significance, we are well to consider eliminating that ”explanatory” variable as it may be actually useless in the model. But of course, if we do that, then the data must have that column in the data eliminated and the whole multilinear regression computation done over again in order to get the proper coefficient calculations for the remaining explanatory variables and see the new p-values. Thus eliminating any explanatory variables leads to redoing the regression calculation and getting new regression coefficients for the remaining explanatory variables.

Now, once the model is evaluated statistically as described here, of course the main use of the model is to make actual predictions of the objective variable when the explanatory variables are given specific values, so this you should try doing. You will notice that in the outputs the values of coefficients are usually given to many decimal places. This is because the regression computations involve calculation related to inversion which become very sensitive to slight changes. For instance you can use a calculator to see that slight changes in the values of MSR and MSE can lead to major variations in the $F-$ratio.

Some summary outputs many not include the adjusted R-square, but only the value of R^2. In fact, the three numbers R^2, R_a^2, and F are related and it might be useful to note some of the relations which are following. For degrees of freedom,

$$d_R = k,\ d_T = n-1,\ d_E = d_T - d_R = n-k-1,\ d_T = d_R + d_E.$$

$$\text{Set } D = \frac{d_E}{d_R}. \text{ Then } \frac{d_T}{d_R} = 1 + D \text{ and } \frac{d_T}{d_{E.}} = 1 + D^{-1}.$$

$$SST = SSR + SSE,\ SSR = SST - SSE,$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, \text{ so } 1 - R^2 = \frac{SSE}{SST}, \text{ and } \frac{SSR}{SSE} = \frac{R^2}{1-R^2}.$$

$$R_a^2 = 1 - \frac{MSE}{MST}, \text{ so } 1 - R_a^2 = \frac{MSE}{MST} = \frac{d_T}{d_E} \cdot \frac{SSE}{SST} = (1 + D^{-1})(1 - R^2).$$

Also,

$$d_T \cdot (1 - R^2) = \frac{SSE}{MST} = d_E \cdot \frac{MSE}{MST} = d_E \cdot (1 - R_a^2),$$

so

$$d_T \cdot (1 - R^2) = d_E \cdot (1 - R_a^2),$$

and therefore,

$$R_a^2 = 1 - \frac{d_T}{d_E} \cdot (1 - R^2) = \frac{d_T \cdot R^2 - d_R}{d_E}.$$

Likewise,

$$1 - R^2 = \frac{d_E}{d_T} \cdot (1 - R_a^2) \text{ and } R^2 = 1 - \frac{d_E}{d_T} \cdot (1 - R_a^2) = \frac{d_R + d_E \cdot R_a^2}{d_T}.$$

Therefore,

$$\frac{R^2}{1 - R^2} = \frac{d_R + d_E \cdot R_a^2}{d_E \cdot (1 - R_a^2)} = \frac{1 + DR_a^2}{D(1 - R_a^2)}.$$

Now, we have

$$F = \frac{MSR}{MSE} = D \cdot \frac{SSR}{SSE} = D \cdot \frac{R^2}{1 - R^2} = \frac{1 + DR_a^2}{(1 - R_a^2)}.$$

When we solve the equation for R^2 in terms of F, we easily find

$$R^2 = \frac{F}{D + F},$$

whereas when we solve for R_a^2 in terms of F, we find

$$R_a^2 = \frac{F - 1}{D + F},$$

which means in particular that

$$R^2 - R_a^2 = (D + F)^{-1} \text{ and } F = (R^2 - R_a^2)^{-1} - D.$$

It should be noted that because of the inversions here that these last four equations though simple are very sensitive to inputs, especially when R^2 and R_a^2 are nearly the same as happens in examples where both are nearly 1.

DEPARTMENT OF MATHEMATICS, TULANE UNIVERSTIY, NEW ORLEANS, LA 70118
Email address: `mdupre@tulane.edu`