**Qualitative Independent Variables**

In our treatment of both the simple and multiple regression we have used independent variables, which were measurable quantities such as advertizing dollars, certain short-term interest rates etc. In certain applications, we may wish to include variables that are not quantitative but qualitative. For instance, in a model designed to explore the relation between total cholesterol and incidences of heart disease we may wish to include gender to see if sex plays a role in the response variable. In other words, we may wish to determine if heart disease strikes men and women differently as a result of high levels of cholesterol. Or in the example of predicting long term interest rates based on federal funds rate and 3-month treasure bill rate, we may suspect that investment managers’ opinions about the state of the economy (either favorable or unfavorable) may have a bearing on the response variable.

These qualitative factors can be included in a multiple regression by use of dummy (or indicator) variables. These variables, sometimes referred to as binary, can assume either the value of *one* or *zero.*  In the case of cholesterol, for instance, we may code males in our sample as one and females as zero; in the second example one may code favorable as one and unfavorable as zero. Which of the qualitative level is coded one and which as zero is completely arbitrary. Although the examples we used involve qualitative variables with two levels (male vs. female and favorable vs. unfavorable), they can also be used to represent qualitative factors that can assume more than two levels. A qualitative variable that can assume *q* distinct levels can be represented by (*q* - 1) dummy variables. Suppose in the advertising example we suspect the demand is seasonal and we would like to take the seasonality into consideration. The four seasons, in addition to the only quantitative variable, *X*1 (advertizing), can be represented by three dummy variables: *X*2 = 1 if winter, 0 otherwise; *X*3 = 1 if spring, 0 otherwise; *X*4 = 1 if summer, zero otherwise. For a given observation, if all three dummy variables are coded as zero, it must be fall because the observation has to be in one of the four seasons. The level that is implicitly used as default (fall, in this example) is referred to as the ‘base case’; the level chosen as the ‘base case’ is completely arbitrary. Notice, we could have just as well chosen winter, summer or spring as the base case without affecting the results. With the seasons represented as another explanatory factor, the model becomes:

*Y* = *A* + *B*1*X*1 + *B*2*X*2 + *B*3*X*3+ *B*4*X*4 + *ε*

This model can be estimated as a multiple regression and the results are interpreted as previously discussed.

**Example:** In the model we used to predict long-term interest rates based on the *Fed Funds rate* and *three-month treasury bill* *rate*, let’s say we also have accumulated information regarding the opinions of investment managers regarding the economy at the time of the observation, categorized as either “favorable” or “unfavorable”. We can introduce this factor by coding a binary variable *X*3 = 1 if ‘favorable,’ and 0 if ‘unfavorable;’ notice that this implies ‘unfavorable’ is the base case.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year** | **Y** | **X1** | **X2** | **X3** |
| 1980 | 11.43 | 13.35 | 11.39 | 0 |
| 1981 | 13.92 | 16.39 | 14.04 | 0 |
| 1982 | 13.01 | 12.24 | 10.6 | 1 |
| 1983 | 11.1 | 9.09 | 8.62 | 0 |
| 1984 | 12.46 | 10.23 | 9.54 | 1 |
| 1985 | 10.62 | 8.1 | 7.47 | 1 |
| 1986 | 7.67 | 6.8 | 5.97 | 0 |
| 1987 | 8.39 | 6.66 | 5.78 | 1 |
| 1988 | 8.85 | 7.57 | 6.67 | 1 |
| 1989 | 8.49 | 9.21 | 8.11 | 0 |
| 1990 | 8.55 | 8.1 | 7.5 | 0 |
| 1991 | 7.86 | 5.69 | 5.38 | 0 |
| 1992 | 7.01 | 3.52 | 3.43 | 1 |
| 1993 | 5.87 | 3.02 | 3 | 0 |
| 1994 | 7.69 | 4.21 | 4.25 | 1 |
| 1995 | 6.57 | 5.83 | 5.49 | 0 |

Estimating the true model *Y* = *A* + *B*1*X*1 + *B*2*X*2 + *B*3*X*3 + *ε*  by OLS we obtain:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* |  |  |  |  |  |  |  |
| Multiple R | 0.980585 |  |  |  |  |  |  |  |
| R Square | 0.961548 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.951935 |  |  |  |  |  |  |  |
| Standard Error | 0.532881 |  |  |  |  |  |  |  |
| Observations | 16 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 3 | 85.2098 | 28.40327 | 100.0249 | 9.32E-09 |  |  |  |
| Residual | 12 | 3.407545 | 0.283962 |  |  |  |  |  |
| Total | 15 | 88.61734 |   |   |   |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* |  |
| Intercept | 2.089921 | 0.51292 | 4.074554 | 0.001541 | 0.972364 | 3.207478 | 0.972364 |  |
| X1 | -1.28108 | 0.460842 | -2.77988 | 0.016654 | -2.28517 | -0.277 | -2.28517 |  |
| X2 | 2.329628 | 0.557213 | 4.180855 | 0.001275 | 1.115564 | 3.543691 | 1.115564 |  |
| x3 | 1.354212 | 0.271919 | 4.980206 | 0.00032 | 0.761752 | 1.946673 | 0.761752 |  |

Notice that with the addition of the new binary variable we have improved the ‘fit’: the error of estimate *se* is smaller and both *r*2 and adjusted *r*2 are larger. Notice also that the new qualitative variable is highly significant as judged by its p-value, which is 0.0032. This means had the true value of *B*3 been zero (opinions did not matter) a *b*3 value as large as 1. 354212 or larger would be very unlikely to occur. The model estimates that a favorable opinion of the investment managers increases the yield on ten-year treasuries by 1.35%. Methods of inferences (*i.e.*, testing hypotheses, stating confidence intervals, etc.) on *B*3 could be done exactly the same way as for qualitative independent variables.

A dummy variable operates by actually modifying the intercept. To see this, consider the fact that for all cases in which the opinion has been unfavorable, the *X*3 term falls out and the model is:

Y = 2.0899 -1.28108*X*1 +2.3296*X*2;

While for all cases with a favorable opinion, because *X*3 =1, the model becomes:

 Y = 2.089921 - 1.28108*X*1 + 2.3296*X*2+ 1.354212 or

 Y = 3.444133 - 1.28108*X*1 + 2.3296*X*2

As you can see, the two models are identical except for their intercept. Geometrically, this means that this model can be represented in *Y*, *X*1 and the *X*2 space by one plane for the ‘favorable’ cases and another for the ‘unfavorable’ cases, whose heights (*i.e.*, *Y* values) differ by the fixed quantity of 1.13542.

**Second Order Regression Models**

 We have seen, while discussing multiple regression models, that by adding additional independent (explanatory) variables into a regression model we can reduce the standard deviation of the response (dependent) variable and thus increase its predictive power. Intuitively, this is by virtue of the fact that the new variables explain some of the hitherto unexplained deviations of the observed (*Yi*) versus the estimated () *Y* values. One can also use the power of multiple regression to improve the fit without introducing any new independent variables, but by allowing a more complex relationship (*e.g.,* non- linear) between the existing independent variables and the response variable. In this note, instead of identifying new explanatory variables, we are going to introduce some extensions to the linear model that allow “curvature” in the response surface. These extensions use the existing independent variables to derive new variables (such as *X*2 or *X*1*X*2) and thus increase the number of factors in the regression model. Therefore here is a caution before we proceed.

As we mentioned before, it is a mathematical fact that the more independent variables used in a regression model, the smaller the deviations between the actual and the predicted response variables and thus higher *r*2. In fact, if you have *n* observations, *n-*1 independent variables (even if they may have no bearing on the dependent variable) will result in a perfect fit with no deviations. If you have one independent variable, in a two dimensional *XY* space you can draw a line that passes perfectly through *n* = 2 sample points. Likewise with two independent variables, a perfect plane can be fitted to three observations. The more variables are used, the more degrees of freedom are lost however. If *n-*1 independent variables are used to fit a model to a data of *n* observations, the result is 0 degrees of freedom, making the model completely useless for predicting the dependent variable. Therefore we need to be judicious when deciding whether to include an additional factor (either a new variable or a derived variable such as the square of an existing one) as an independent variable, because there is a “cost.” Recall that the “adjusted *r*2” tries to account for this cost. The principle of *parsimony* requires that only those variables that theoretically explain significant variability in the response variable should be used, but those that have insignificant or no explanatory power should be left out to preserve degrees of freedom. The contribution of an independent variable should be judged on the basis of its effect on the adjusted *r*2: In general, if the addition of some new variable increases the adjusted *r*2, it is worth to have that variable in the model.

**I. Second order Models**

Look at the plot of the sample observations of income versus consumption below.

 

It is fairly obvious that the relationship between consumption (dependent variable) and income (independent variable) is not best described by a straight line. As economic theory predicts, it appears that as income increases the rate of increase in consumption tapers off; as your income increases you tend to consume more but the increase in your consumption begins to slow. The linear model *Y = A + BX + ε* would yield a poor fit to this data, because it essentially would be forcing a round peg in a square hole. Obviously a better fitting model would allow a curvature in the response surface. In this case the addition of a square term as *Y = A + B*1*X + B*2*X2 + ε*, may capture the apparent non-linearity. We can estimate this model by adding the *X*2 values as a second “independent” variable and run it as a multiple regression. The parameter *B*2 is called the rate of curvature and its significance can be tested in the usual way. That is one can test the null hypothesis Ho: *B*2 = 0 versus *B*2 < 0 (or *B*2 > 0, or B ≠ 0), using the *t*-statistic. If *B*2 is significant and < 0 then as shown in the above example the curvature is concave--the impact of the independent variable on the dependent variable decreases as *X* increases. If, on the other hand, *B*2 is significant and > 0, we have a convex relationship, where the rate of change of the dependent variable strengthens as *X* increases. Obviously, if we can not reject the null hypothesis, the relationship is linear with no significant curvature. Mathematically, this conclusion is based on the sign of the derivative of *Y* with respect to *X*, which is *B*1 + 2*B*2*X*. A negative *B*2 will reduce the derivative as *X* increases and vice versa. Notice further that in a model in which *B*1 is zero implies a U shape relationship if *B*2 > 0 (inverted U if *B*2 < 0). Finally even higher order models can be constructed by including a cubic term, fourth power term etc.

Example: assume *X* is the household weekly income and *Y* the weekly consumption

|  |  |
| --- | --- |
| Y | X |
| 252.71 | 300 |
|  271.81 | 350 |
| 333.73 | 450 |
| 238.08 | 235 |
| 361.16 | 1020 |
| 383.60 | 880 |
| 359.20 | 567 |
| 209.23 | 230 |
| 324.49 | 470 |
| 344.93 | 905 |
| 297.71 | 468 |
| 367.60 | 750 |

 

 The first order model *Y = A + BX + ε* is estimated as

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Regression Statistics* |  |  |  |  |  |
| Multiple R | 0.856799 |  |  |  |  |  |
| R Square | 0.734104 |  |  |  |  |  |
| Adjusted R Square | 0.707515 |  |  |  |  |  |
| Standard Error | 30.91203 |  |  |  |  |  |
| Observations | 12 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 1 | 26381.61 | 26381.61 | 27.60872 | 0.000371 |  |
| Residual | 10 | 9555.536 | 955.5536 |  |  |  |
| Total | 11 | 35937.15 |   |   |   |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 213.1951 | 20.81785 | 10.24098 | 1.28E-06 | 166.81007 | 259.58019 |
| X | 0.179006 | 0.034068 | 5.2544 | 0.000371 | 0.1030984 | 0.2549145 |

The coefficient of determination of .734 indicated a good fit with 73.4% of the observed differences in consumption being attributed to variations in income. The independent variable, income is highly significant with a *p*-value of .00037 (in other words we would be able to reject the null hypothesis that *B* = 0 at any level of significance > 0.00037). The error of estimate, *se* is about 31. However, from the graph of the points above it is apparent that the fit may be improved by a second-order model which includes *X*2 as another independent variable:

|  |  |  |
| --- | --- | --- |
| Y | X | X2 |
|  252.71 | 300 | 90000 |
| 271.81 | 350 | 122500 |
| 333.73 | 450 | 202500 |
| 238.08 | 235 | 55225 |
| 361.16 | 1020 | 1040400 |
| 383.60 | 880 | 774400 |
| 359.20 | 567 | 321489 |
| 209.23 | 230 | 52900 |
| 324.49 | 470 | 220900 |
| 344.93 | 905 | 819025 |
| 297.71 | 468 | 219024 |
| 367.60 | 750 | 562500 |

*Y = A + B*1*X + B*2*X2 + ε*  is estimated below

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Regression Statistics* |  |  |  |  |  |
| Multiple R | 0.968032 |  |  |  |  |  |
| R Square | 0.937087 |  |  |  |  |  |
| Adjusted R Square | 0.923106 |  |  |  |  |  |
| Standard Error | 15.84973 |  |  |  |  |  |
| Observations | 12 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 2 | 33676.22 | 16838.11 | 67.02694 | 3.93E-06 |  |
| Residual | 9 | 2260.927 | 251.2141 |  |  |  |
| Total | 11 | 35937.15 |   |   |   |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 75.98443 | 27.60975 | 2.752087 | 0.0224 | 13.526787 | 138.44207 |
| X | 0.735177 | 0.104679 | 7.023129 | 6.17E-05 | 0.4983753 | 0.9719781 |
| X2 | -0.00045 | 8.44E-05 | -5.38864 | 0.000439 | -0.0006458 | -0.0002639 |

Compared to the first-order model this is a much better model of the consumption as a function of income. Coefficient of determination is now about 93.7%, both *B*1 and *B*2 highly significant. Negative *B*2 indicates that the relationship between income and consumption moderates as income increases. Caution: for sufficiently high income levels an increase in income may reduce consumption *i.e*., the slope may become negative. Remember, however that predictions of the dependent variable for values of the independent variable outside of the range in the sample (here from 230 to 1020) will give misleading results.

**II. Interaction Models**

Consider the linear model with two independent variables

*Y = A + B*1*X*1 *+ B*2*X2*+ *ε*. Say *Y* is compensation ($000), *X*1 education and *X*2 experience of bank tellers both in years. Suppose the estimated linear model is *Y = 42 + 4.86X*1 *+ 3.02X2*+ *ε*. We can examine the relationship between pay (*Y*) and education (*X*1) for any fixed value of experience (*X*2).  For instance, for *X*2 = 1 or 2, *i.e*., we are looking for the impact of education for the population of all tellers with one versus two years of experience. Substituting 1 for *X*2, the equation becomes *Y* = 45.02 + 4*.86X*1 + *ε,* whilefor *X*2 = 2, it is *Y* = 48.04 + 4*.86X*1 + *ε.*  Therefore in this model, regardless of the experience (*X*2), pay tends to increase by 4.86 ($000) for every additional year of education (*X*1). This relationship between *Y* and *X*1 for various levels of *X*2 (1, 2, and 3 years) can be graphed as follows.



The slope of the line does not change as *X*2 changes, only the intercept changes. In this type of a relationship *X*1 and *X*2 are said not to interact—in the sense that the impact of education on pay remains at 4.86 ($000) per year regardless of the experience. It is plausible however to suspect that the impact of education on pay for those with little experience might be stronger than those with a lot of experience. Namely we might think that the impact of education on pay might moderate as experience increases and we may want our model to reflect this possibility as graphed below:



For a person with little experience (*X*2 = 1) the rate of increase in pay as education increases is stronger (the line is steeper) than for a more experienced person (*X*2 = 3). In a model that allows this type of relationship, *X*1 and *X*2 are said to interact. We can model the interaction by including a term *X*1*X*2 in the model. With this term included the model becomes *Y = A + B*1*X*1 *+ B*2*X2*+ *B*3 *X*1 *X2 + ε.*  This model can be estimated and the significance of the interaction term *X*1 *X*2 can be questioned by testing Ho: *B*3 = 0 versus H1: *B*3 ≠ 0 (or *B*3 > 0, or *B*3 < 0) using student’s *t*. If *B*3 < 0 and significant, then the interaction is negative. This means that as one of the independent variable’s value increases, the effect of the other variable on the dependent variable moderates. This is the case in the above example. If however, *B*3 > 0 and significant, the interaction is positive and the two variables reinforce one another. This conclusion is reached by examining the derivatives of *Y* with respect to *X*1 and *X*2. *dY/dX*1 = *B*1 + *B*3 *X2* and *dY/dX2* = *B*2 + *B*3 *X*1. If *B*3 > 0 and significant, either derivative will be larger (mutually reinforcing) as the other variable increases and *vice versa*.

Example; *Y* is pay ($000), *X*1 education (yrs) and *X*2 experience (yrs)

|  |  |  |  |
| --- | --- | --- | --- |
| Y | X1 | X2 | X1X2 |
| 53 | 1 | 3 | 3 |
| 64.2 | 2 | 2 | 4 |
| 42.8 | 1 | 2 | 2 |
| 66.4 | 4 | 1 | 4 |
| 81.5 | 5 | 4 | 20 |
| 63.8 | 2 | 3 | 6 |
| 66.2 | 1 | 5 | 5 |
| 57.2 | 3 | 2 | 6 |
| 77.8 | 6 | 3 | 18 |
| 97.5 | 8 | 6 | 48 |
| 84.3 | 4 | 8 | 32 |
| 68.1 | 3 | 2 | 6 |

The first order model is *Y = A + B*1*X*1 *+ B*2*X2*+ *ε* (disregarding the *X*1*X*2 term) and estimated as

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Regression Statistics* |  |  |  |  |  |
| Multiple R | 0.944482 |  |  |  |  |  |
| R Square | 0.892047 |  |  |  |  |  |
| Adjusted R Square | 0.868057 |  |  |  |  |  |
| Standard Error | 5.393216 |  |  |  |  |  |
| Observations | 12 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 2 | 2163.166 | 1081.583 | 37.18469 | 4.46E-05 |  |
| Residual | 9 | 261.781 | 29.08678 |  |  |  |
| Total | 11 | 2424.947 |   |   |   |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 42.04253 | 3.542825 | 11.86695 | 8.47E-07 | 34.0281 | 50.05696 |
| X1 | 4.859923 | 0.795379 | 6.110198 | 0.000177 | 3.060651 | 6.659194 |
| X2 | 3.021774 | 0.861268 | 3.508518 | 0.006634 | 1.073451 | 4.970097 |

The regression is highly significant (p-value 4.46E-05), coefficient of determination is better than 89%, both *X*1 and *X*2 are significant. Estimate of the standard deviation of pay is about $5,393. Suspecting significant interaction between education and experience and estimating the interactive model: *Y = A + B*1*X*1 *+ B*2*X2*+ *B*3 *X*1 *X2 + ε* yields:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Regression Statistics* |  |  |  |  |  |
| Multiple R | 0.951435 |  |  |  |  |  |
| R Square | 0.905228 |  |  |  |  |  |
| Adjusted R Square | 0.869689 |  |  |  |  |  |
| Standard Error | 5.35976 |  |  |  |  |  |
| Observations | 12 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 3 | 2195.13 | 731.7101 | 25.47114 | 0.000191 |  |
| Residual | 8 | 229.8162 | 28.72703 |  |  |  |
| Total | 11 | 2424.947 |   |   |   |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 34.24439 | 8.188269 | 4.182128 | 0.003071 | 15.36221 | 53.12657 |
| X1 | 7.177251 | 2.334712 | 3.074149 | 0.015252 | 1.793396 | 12.56111 |
| X2 | 5.100153 | 2.14819 | 2.374162 | 0.044953 | 0.146417 | 10.05389 |
| X1X2 | -0.54759 | 0.519117 | -1.05485 | 0.322307 | -1.74468 | 0.649496 |

Is this a better fit? The answer is found by testing the null hypothesis: Ho: *B*3 = 0 versus *B*3 < 0. We can not reject the null hypothesis (there is no significant interaction) at even a modest level of significance of α = .10 (p-value is .322). Despite the fact that the coefficient of determination improved to 90.5% and standard deviation is reduced by a small amount (due to another independent variable which cost a degree of freedom), there is no compelling evidence that the inclusion of the interaction improves the model.

 An interesting application of interaction among variables is when one of the variables suspected to interact with another happens to be a qualitative variable such as gender. Suppose in the above example we differentiate between male and female observations by coding a new dummy variable, *X*3 (1 for males and 0 for females). We can add an interaction term *B*4 *X*1*X*3 to the model to investigate whether or not the length of education affects pay for males differently than it does for females. In this extended model the derivative of *Y* with respect to education is *B*1 + *B*4*X*3. If *B*4 is significant then we can conclude that the impact of education on pay for males (*X*3 = 1) is *B*1 + *B*4 while for females (*X*3 = 0) is simply *B*1. Further, if *B*4 > 0 then education impacts male pay more strongly than it does female pay and vice versa.

**III. General second order model**

Suppose we have two independent variables to use for predicting the value of a dependent variable. A complete second order model can be formed by including both the squared variables as well as the interaction term as follows:

*Y* = *A + B*1*X*1 *+ B*2*X*2 *+ B*3*X*1*X*2 *+ B*4*X*1*2 +**B*5*X*2*2* *+ ε.*  One way to test the appropriateness of this complex model compared to the simpler alternative first-order model

*Y* = *A + B*1*X*1 *+ B*2*X*2 *+ ε* is to use the ordinary *t-* test the significance of *B*3, *B*4 or *B*5 one at a time. However, this will not always give a reliable diagnosis. To see why not, suppose for a moment that none of *B*3, *B*4 and *B*5 is significant. If we test each of these null hypotheses individually (that *Bi* = 0) at *α* = .05, there will be 95% chance we’ll make the correct decision for *B*3 (that it is zero); 95% chance with respect to *B*4 and 95% chance with respect to *B*5. Thus the probability of correctly finding all of the second order terms insignificant (*i.e*., *B*3 =*B*4 = *B*5 =0) will be .953 = .857 leading to a type I error (probability of rejecting the null when it is true) of about 14.3%. Obviously the more additional terms we test the larger this error will be.

**Partial F test**

 To avoid this we need to test the contribution of these second order terms collectively as

 Ho: *B*3 = *B*4 = *B*5 = 0

 H1: at least one is not zero

Notice how similar this is to the F-test used for the general significance of the entire multiple regression model. As you may guess, the appropriate test statistic for this test is the *F* distribution and the test is called the partial *F*-test for we are testing a subset of all the parameters, and not all of them. Let us refer to the simpler model

*Y* = *A + B*1*X*1 *+ B*2*X*2 *+ ε* as the reduced model (as opposed to the complete model). For a general case let *g* to denote the number of *B* parameters in the reduced model (in our case *g* = 2) and *k* to denote the number of *B* parameters in the complete model (5 here). Let SSER and SSEC be the sum of squared errors for the reduced and the complete models respectively given in the Excel output for the two models. Then the test statistic for the partial *F* test is given by: with *k - g* degrees of freedom for the numerator and *n - k -* 1 degrees of freedom for the denominator, where *n* is the sample size as before. If the computed test statistic exceeds the critical *F* (for the appropriate *α* with *k - g* and *n - k -* 1 degrees of freedom) the null is rejected and the significant contribution of the square terms and the interaction term to the predictive power of the model is acknowledged. To conduct this test in order to choose between the simpler (parsimonious) and the more complex model we need to estimate both models first and then do the partial *F*-test to choose between them.

Example

In the previous example of *Y* = pay; *X*1 = education and *X*2 = experience we can construct the complete model as *Y* = *A + B*1*X*1 *+ B*2*X*2 *+ B*3*X*1*X*2 *+ B*4*X*1*2 +**B*5*X*2*2* *+ ε* and define the model *Y* = *A + B*1*X*1 *+ B*2*X*2 *+ ε*  as the reduced model. The data to estimate both models is:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Y | X1 | X2 | X1X2 | X12  | X22 |
| 53 | 1 | 3 | 3 | 1 | 9 |
| 64.2 | 2 | 2 | 4 | 4 | 4 |
| 42.8 | 1 | 2 | 2 | 1 | 4 |
| 66.4 | 4 | 1 | 4 | 16 | 1 |
| 81.5 | 5 | 4 | 20 | 25 | 16 |
| 63.8 | 2 | 3 | 6 | 4 | 9 |
| 66.2 | 1 | 5 | 5 | 1 | 25 |
| 57.2 | 3 | 2 | 6 | 9 | 4 |
| 77.8 | 6 | 3 | 18 | 36 | 9 |
| 97.5 | 8 | 6 | 48 | 64 | 36 |
| 84.3 | 4 | 8 | 32 | 16 | 64 |
| 68.1 | 3 | 2 | 6 | 9 | 4 |

We had already estimate the first order model *Y* = *A + B*1*X*1 *+ B*2*X*2 *+ ε* above.

The estimate of *Y* = *A + B*1*X*1 *+ B*2*X*2 *+ B*3*X*1*X*2 *+ B*4*X*1*2 +**B*5*X*2*2* *+ ε*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Regression Statistics* |  |  |  |  |  |
| Multiple R | 0.954852 |  |  |  |  |  |
| R Square | 0.911742 |  |  |  |  |  |
| Adjusted R Square | 0.838194 |  |  |  |  |  |
| Standard Error | 5.972436 |  |  |  |  |  |
| Observations | 12 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 5 | 2210.927 | 442.1853 | 12.39656 | 0.004072 |  |
| Residual | 6 | 214.02 | 35.67 |  |  |  |
| Total | 11 | 2424.947 |   |   |   |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 29.03369 | 12.49285 | 2.324025 | 0.059122 | -1.53521 | 59.60259 |
| X1 | 8.691706 | 3.630475 | 2.394096 | 0.053725 | -0.19175 | 17.57516 |
| X2 | 7.108731 | 4.789042 | 1.484374 | 0.188244 | -4.60963 | 18.8271 |
| X1X2 | -0.18994 | 0.823781 | -0.23057 | 0.825313 | -2.20565 | 1.825784 |
| X12  | -0.37711 | 0.62401 | -0.60433 | 0.567755 | -1.90401 | 1.149787 |
| X22 | -0.35318 | 0.581485 | -0.60737 | 0.565865 | -1.77602 | 1.069665 |

The coefficient of determination is marginally better for the complete model, 91.1% versus 89.2% (*i.e*., the complete model accounts for 91.1% of the variation in the pay, the parsimonious model accounts for 89.2%). However, the complete model is not a better model-- none of the *Bs* appears to be significant. This happens as a result of two factors. First, since the sample size is relatively small the complete model has very few degrees of freedom (six as opposed to nine for the reduced model); and second, since the derived variables *X*1*X*2 *, X*1*2* and *X*2*2* are mathematically related to the original variables *X*1 and *X*2 , in general, second order models tend to be susceptible to multi-co linearity.

Although with these observations we can easily see that the parsimonious model is superior, let’s do the formal partial *F*-test to verify that the second order terms are not contributing significantly to the power of the model in predicting pay based on education and experience.

SSE*C* = 214.02; SSE*R* = 261.78; *k* =5; *g* = 2; *n* = 12

 Ho: *B*3 = *B*4 = *B*5 = 0

 H1: at least one is not zero

 *α = .*05

 = 

The critical *F* value with (*k-g* = 3 and (*k-n*-1= 6) degrees of freedom for *α = .*05 is 4.757055 therefore, as we suspected, we can not reject the null hypothesis that the interaction and square terms are all insignificant.

 The use of the partial *F*-test is not confined to test the significance of the interaction and/or squared terms; it can be used to choose between any two alternative models in which one model contains all the *B* parameters of the other model and then some.