Chapter 10

Is the Model Any Good?[1]

In the last chapter we built regression models that measured the effects of several explanatory variables on a dependent variable. For example, how educational background, prior experience, years with a company, job level, or gender affect salary. We determined how each explanatory variable, whether numerical or categorical, expressed its effect on salary through its coefficient in the regression equation. The process of building such a model is a statistical one; that is, it involves determining a best-fit equation by calculating how much of the total variation is accounted for by the model. This calculation, in turn, is based on certain probabilistic assumptions concerning how the data is distributed. The first section of this chapter concerns how confident we can be that the coefficients of our explanatory variables are trustworthy. This is critically important if we are to make decisions based on our understanding of what a model seems to be telling us. We need criteria to determine which explanatory variables are truly significant in affecting the dependent variable–and which are not–if our model is to be at all useful. This section helps us to separate the wheat from the chaff.

The second section of this chapter furthers the process of building more complex and accurate models from several explanatory variables by considering how interactions between the variables themselves might have an effect on the dependent variable. That is, some of these variables might express their effects on the dependent variable in combination with other explanatory variables. In fact, there are even cases in which an explanatory variable appears to have a significant effect only when it is combined with one or more other explanatory variables. For example, it may be that employees' gender by itself has no significant effect on salary, but gender together with job level might have a negative impact on salary. That is, the negative effect of gender on salary only has a significant impact when the employee is a female in a higher-level position: the well-known "glass-ceiling" effect. This section, then, concerns not only the effects of several individual explanatory variables on a dependent variable, but also the effects of pairs of them on the dependent variable. You will learn in this chapter how to create multiple regression models with interaction variables built from both numerical and categorical explanatory variables and assess their significance. You will learn how to analyze and interpret these often complex models.

- *As a result of this chapter, students will learn*

- √ How to determine the trustworthiness of the coefficients of a regression equation
- √ How to determine which coefficients should be kept in a model and which should not
- √ How to interpret models with complex interaction terms involving both numerical and categorical variables

- *As a result of this chapter, students will be able to*
  - √ To determine with 95% confidence the range of values within which regressions coefficients fall
  - √ Create interaction terms
  - √ Identify the reference categories of interaction variables
  - √ Use StatPro's interaction routine to construct dummy variable for interaction variables
  - √ Construct a model using interaction terms
  - √ How to use StatPro's stepwise regression routine to build complex models

10.1 Which coefficients are trustworthy?

In the last chapter, several regression models of EnPact's employee salary structure were developed in order to determine if female employees earn less than their male counterparts. These models indicate that females do earn less than their male counterparts, often many thousands of dollars a year less, depending on which variables are used in the models. As EnPact's Human Resources Director, you are aware that if females do indeed earn substantially less than males, say $5000 a year, then EnPact could be liable for a potentially ruinous multi-million dollar law suit. But to what degree can you be confident that these models are indeed producing accurate results?

We will answer this question and related questions in this chapter, but first we need some concepts.

Suppose we have a regression equation with two explanatory variables, X_1 and X_2, and their coefficients, and , respectively:

$$\text{dependent variable} = \text{constant} + B_1 \times X_1 + B_2 \times X_2$$

If one of the coefficients is zero, say B_1, then X_1 makes no contribution to the dependent variable no matter what value it takes on because $0 \times X_1 = 0$ and the equation reduces to

$$\text{dependent variable} = \text{constant} + B_2 \times X_2$$

In this case, X_1 is said to be insignificant.

Just because a coefficient is nonzero, however, does not mean that the variable is necessarily significant. A statistician would warn us that regression coefficients are only estimates[2] and that some of them, in fact, should–or rather could–be zero. The question is, then, can we identify which variables could possibly have zero coefficients and thus be eliminated from our analysis because they are insignificant? The answer is: not with 100% certainty–but we can be 95% confident as to which variables are significant and which are not. When statisticians use the phrase, "95% confident," they mean that 95% of the time we will be able to correctly identify whether a particular variable is or is not significant.

We need to understand two formulations concerning what it means to say that a variable is significant:

1. A variable is significant if we are 95% confident that its coefficient is nonzero is equivalent to saying

2. A variable is significant if there is less than a 5% chance that its coefficient is zero.

Both of these perspectives concerning the significance of a variable are given to us in regression output and provide slightly different information.

[2]Remember: the data we are working with is a *sample* rather than the entire population. If we sample the data again, we would get different values for the coefficients in the regression model.

10.1.1 Definitions and Formulas

p-value The probability that a particular regression coefficient is zero. When p is small, say less than .05, there is only a 5% chance or less than the coefficient is zero.

Significant variable or coefficient A variable or a coefficient of a variable is significant when its p-value is less than .05. That is, there is less than a 5% chance that the coefficient is zero.

Insignificant variable or coefficient A variable or a coefficient of a variable is insignificant when its p-value is greater than .05. That is, there is more than a 5% chance that the coefficient is zero. As a general rule (there are exceptions), when a variable is found to be insignificant in a particular model, it should not be included in future models.

95% confidence interval The interval in which we can be 95% certain that a coefficient will lie, meaning that the coefficient will lie in this interval 95% of the time.

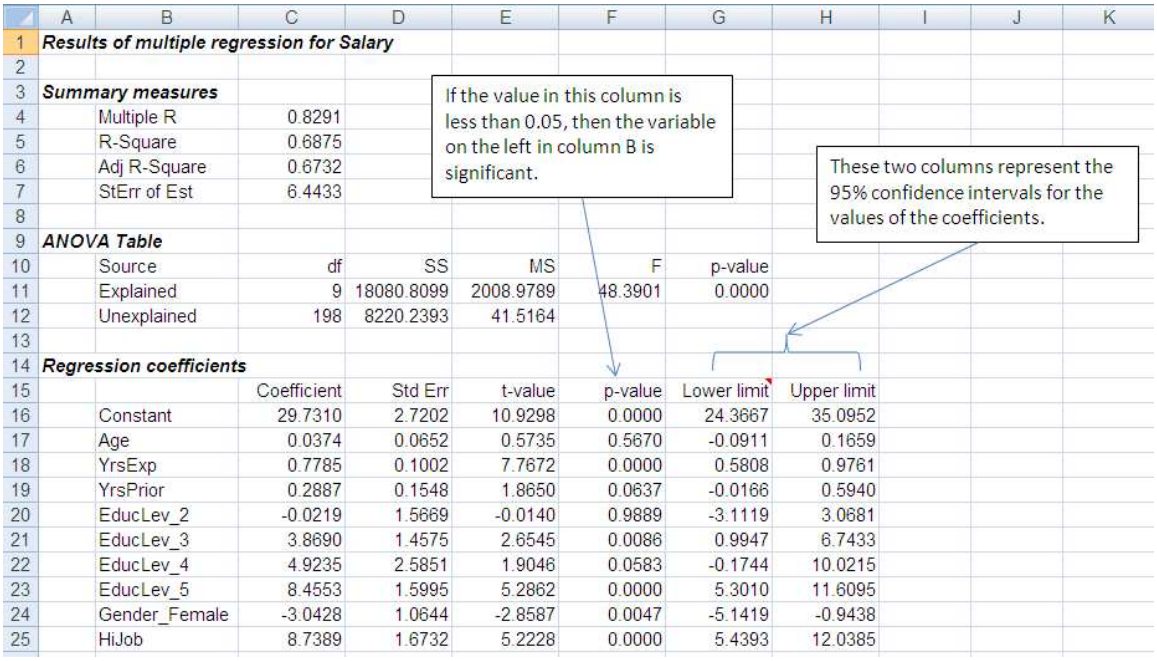
Principle of parsimony Equivalent to K.I.S.S. If we have a choice between two models, we should choose the simpler or smaller model of the two, provided that it does reasonably as well as the larger, more complicated model. (This principle is also known as Occam's Razor: Things should not be multiplied without reason.)

10.1.2 Worked Examples

Example 10.1. Determining significance of a variable from a confidence interval We look to the last three columns of the "Regression coefficients" block in the Excel spreadsheet below to determine if a variable is significant. This data is shown in file C10 Enpact Data.xls. The variable HiJob is a dummy variable that is 1 if the employee's job grade is 5 or 6.

We can be 95% confident that the coefficient of a variable, say Age, lies somewhere between the lower-limit number, -.0911, and the upper-limit number, .1659. Since the lower limit is negative and the upper limit is positive, the coefficient, given as .0374, could very well be 0. This means that the variable is insignificant. On the other hand, if the signs of the lower and upper limits are the same, then we can be 95% confident that the associated variable (or the constant in the case of the first row) is not zero and is therefore significant at a 95% level of confidence. For example, we can be 95% confident that the variable YrsExp is significant and that its coefficient lies somewhere between .5808 and .9761.

Example 10.2. Determining the significance of a variable from a p-value We can determine if a variable, say HiJob, is significant by examining the p-value of its coefficient (third column from the right in the regression output.) Since its p-value, .0000, is less than .05, we can expect that its coefficient, 8.7389, will be zero less than 5% of the time. This means that we can expect the coefficient will not be zero 95% of the time and therefore the variable is significant at a 95% level of confidence. On the other hand, the p-value of the coefficient of the Age variable is .5670, which is greater than .05. This says that the



Results of multiple regression for Salary							
Summary measures							
	Multiple R	0.8291					
	R-Square	0.6875					
	Adj R-Square	0.6732					
	StErr of Est	6.4433					
ANOVA Table							
	Source	df	SS	MS	F	p-value	
	Explained	9	18080.8099	2008.9789	48.3901	0.0000	
	Unexplained	198	8220.2393	41.5164			
Regression coefficients							
		Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
	Constant	29.7310	2.7202	10.9298	0.0000	24.3667	35.0952
	Age	0.0374	0.0652	0.5735	0.5670	-0.0911	0.1659
	YrsExp	0.7785	0.1002	7.7672	0.0000	0.5808	0.9761
	YrsPrior	0.2887	0.1548	1.8650	0.0637	-0.0166	0.5940
	EducLev_2	-0.0219	1.5669	-0.0140	0.9889	-3.1119	3.0681
	EducLev_3	3.8690	1.4575	2.6545	0.0086	0.9947	6.7433
	EducLev_4	4.9235	2.5851	1.9046	0.0583	-0.1744	10.0215
	EducLev_5	8.4553	1.5995	5.2862	0.0000	5.3010	11.6095
	Gender_Female	-3.0428	1.0644	-2.8587	0.0047	-5.1419	-0.9438
	HiJob	8.7389	1.6732	5.2228	0.0000	5.4393	12.0385

Figure 10.1: Multiple regression results with p-values and confidence intervals highlighted.

Age variable is insignificant because we cannot be confident that its coefficient, 0.0374, is nonzero less than 5% of the time.

Example 10.3. The relative advantages of using confidence intervals vs p-values
A confidence interval not only tells us whether a variable is significant or not, it also gives us a range of values within which we can be 95% confident that the coefficient will lie. A p-value only tells us whether a variable is significant or not. On the other hand, the eye can scan a single column of p-values for significance much quicker and readily than it can scan two columns of numbers looking for a sign change across them.

Example 10.4. Refining your model
The presence of insignificant variables in a model is usually a cause for concern. The reason is this: the presence of insignificant variables raises the model's R^2 by introducing information in which we should not have confidence. In other words, insignificant variables inflate the model's R^2 so that it is not a reliable indicator of how well the model fits the data. This means that we could be basing our inferences and decisions on a faulty model, which, in turn, could lead to disastrous consequences.

To avoid the problem of producing an untrustworthy model, we rerun the regression routine after leaving out all the insignificant variables. Our new reduced model will now be built with significant explanatory variables, each of which has passed the 95% confidence test.

After dropping the insignificant variables from the model displayed in example 1 (page 288), our reduced model will now be based on the following significant variables: YrsExp,

HiJob, GenderFemale, EducLevel3, EducLevel4, and EducLevel5. The resulting reduced model is shown below:

	A	B	C	D	E	F	G	H
1	***Results of multiple regression for Salary***							
2								
3	***Summary measures***							
4		Multiple R	0.8246					
5		R-Square	0.6799					
6		Adj R-Square	0.6704					
7		StErr of Est	6.4716					
8								
9	***ANOVA Table***							
10		Source	df	SS	MS	F	p-value	
11		Explained	6	17882.9017	2980.4836	71.1650	0.0000	
12		Unexplained	201	8418.1475	41.8813			
13								
14	***Regression coefficients***							
15			Coefficient	Std Err	t-value	p-value	Lower limit	Upper limit
16		Constant	32.1066	1.4074	22.8121	0.0000	29.3313	34.8818
17		YrsExp	0.7897	0.0770	10.2544	0.0000	0.6378	0.9415
18		EducLev_3	3.2902	1.1663	2.8210	0.0053	0.9904	5.5900
19		EducLev_4	4.3646	2.4630	1.7721	0.0779	-0.4920	9.2212
20		EducLev_5	7.8043	1.3451	5.8019	0.0000	5.1519	10.4568
21		Gender_Female	-2.9490	1.0662	-2.7660	0.0062	-5.0513	-0.8467
22		HiJob	9.0251	1.6740	5.3912	0.0000	5.7241	12.3260

Figure 10.2: Regression output for Enpact data after insignificant variables are dropped.

Notice that the R^2 of our reduced model, 0.8246, is smaller than the R^2 of the original full model, but only by .0045. For all practical purposes, the R^2 of the original model and the reduced model are the nearly identical. Similarly, the S_e of the reduced model, 6.4716, is larger than the S_e of the full model, but only by .0283, which again, for all practical purposes, is nearly identical. Other models, however, may show much larger differences between the R^2 and S_e of the full model and a reduced one.

This example illustrates another principle of good modeling practice: the principle of parsimony. The principle of parsimony can be thought of as a principle of simplicity. If a smaller set of explanatory variables produces a model that fits the data almost as well as a model with a larger set of explanatory variables - and with almost the same standard error - it is usually preferable to use the model with the smaller number of explanatory variables. As we shall see, each explanatory variable in a model comes with a price, not only in terms of increasing the unwieldiness of the model, but more importantly in terms of understanding or explaining how the particular variable affects the dependent variable.

Also notice that one of the variables in the original example 1 (page 288), the variable EducLevel4, was on the border between being significant and not. Its value of 0.0583 is right about equal to the cutoff of 0.05. Because the p-values change dramatically as variables are eliminated from the model, it is important to leave such borderline variables in the model at first and see if they become more significant. In this case, the p-value got larger when we

eliminated some of the variables; in the reduced model, it is definitely not significant at a p-value of 0.0779. In fact, because of the way p-values change as the variables are eliminated, it is always best to eliminate one variable at a time, making a new model as each of the variables is dropped and re-assessing which variables are significant. Often, a variable that began an insignificant can become significant.

Summary: Refining a model is both an art and a science. The general procedure is:

1. Run a full model with all the explanatory variables
2. Determine the significant explanatory variable from the results of the full model
3. Run a reduced model with the variables from 2.
4. With the principle of parsimony in mind, run models built on various subsets of significant (or nearly significant) explanatory variables until you obtain a model that you are satisfied gives the best fit to the data with the fewest explanatory variables.

10.1.3 Exploration 10A: Building a Trustworthy Model at EnPact

1. Run a full model using StatPro's multiple regression routine (don't use the StatPro's stepwise procedure just yet) with all the explanatory variables, both numerical and categorical, of the EnPact data found in "C11 EnPact Data.xls." Be sure to create dummy variables of the categorical data first. And while the Job Grade and Education Level variables are ordinal, they are categorical and should be treated as such. Enter your results in the chart below.

2. Select the significant variables from the output of the full model regression in Part 1 and run the reduced model. Record your results in the chart below.

3. Rerun Part 1 using StatPro's stepwise regression procedure (see the How to Guide in chapter 9A). Enter your results in the chart below.

	Model	R^2	Adj R^2	S_e	List of significant variables
Part 1	Full Model				
Part 2	Reduced Model				
Part 3	StatPro Stepwise regression				

4. What do you observe about your results from Parts 2 and 3? How do you account for this?

5. Write down what you think is the most suitable model and defend your choice.

6. Interpret your model.

10.1.4 How To Guide

Using a VLOOKUP table

In doing some tasks, we find that we need some way to use different information depending on the result of some number. For example, in calculating employee pay, different job types might have different, standardized pay rates at our company. Wouldn't it be nice if Excel could figure it out from the information given and calculate the pay rate correctly? Using a lookup table, in this case a VLOOKUP table, Excel can.

If you open the file "C10 HowTo.xls" you'll see an example. Shown below is an image of the screen illustrating a sample employee database. This database contains information on each employee: hours worked that week, job type, and years of experience.

	A	B	C	D	E	F	G	H	I
1	Employee	Hours	JobType	YrsExp	GrossPay				
2	1	36	5	9	$ 1,184.40		JobType	BasePay	Raise
3	2	44	2	12	$ 741.40		1	$ 6.00	$ 0.55
4	3	20	4	13	$ 562.00		2	$ 7.25	$ 0.80
5	4	16	4	5	$ 296.00		3	$ 8.25	$ 0.95
6	5	40	5	11	$ 1,484.00		4	$ 12.50	$ 1.20
7	6	36	3	13	$ 741.60		5	$ 14.00	$ 2.10
8	7	38	1	12	$ 478.80				
9	8	42	3	2	$ 426.30				
10	9	35	3	11	$ 654.50				
11	10	27	2	11	$ 433.35				
12	11	5	2	1	$ 40.25				
13	12	18	1	9	$ 197.10				
14	13	40	4	3	$ 644.00				
15	14	37	1	13	$ 486.55				
16	15	41	3	11	$ 766.70				

Figure 10.3: Employee database illustrating use of VLOOKUP tables.

Off to the right of database, in cells G2:I7 is the lookup table. (Normally, one would put this on a different sheet of the workbook and name the entire range of cells to make it easier to reference, but for this example, we wanted to keep it easy to visualize.) Now we want Excel to take the employees hours and multiply it by the correct hourly rate, based on the job type and the years of experience. This hourly pay rate will be something like

(Base Pay Rate) + (Years Experience)*(Annual Raise)

But Excel must use the Job Type to determine both the base pay rate and annual raise. To do this, we use VLOOKUP:

```
=VLOOKUP(Lookup Value, Lookup range, Column, [range lookup])
```

So, we can find the base hourly rate for employee 1 by looking up his/her job type (cell C2) in the lookup table ($G$3:$I$7 - the absolute reference is a MUST here!) and using the information in column 2 of the table. To find the annual raise, we perform the same lookup, but instead of returning the information in column 2, we want the information in column 3. Thus, we can compute employee 1's pay by the following formula (shown in text and Excel notation to make it easier to read).

Pay = (Hours Worked) * ((Base Pay Rate) + (Years Experience)*(Annual Raise))

```
E2 = B2*(VLOOKUP(C2,$G$3:$I$7,2) + D2*VLOOKUP(C2,$G$3:$I$7,3))
```

Copying this formula to the cells in E3:E16 will compute each employee's pay, using the correct job type to calculate the pay rate. One could also use this to calculate the taxes based on the number of dependents declared on W4 forms, or practically anything.

IMPORTANT TIP: Lookup tables must be organized a certain way. Excel always uses the leftmost column of the table to match with the LookupValue in the formula, so be sure this is the way it is organized. It is also vital that the table be sorted in ascending order by the first column. If it is not sorted, Excel cannot find the proper match, and you will see an error in the calculation.

NICE FEATURE: Lookup tables don't have to return numbers; they can return any type of data. And, they don't require an exact match. If you have a range of possible values that should return a certain result, then just put the lower end of each range in the left column.

10.2 More Complexity with Interaction Terms

We are becoming aware that gender may have a significant impact on employees' salaries at EnPact. But is its impact isolated from that of the other variables that affect salary? Is it possible that the variable GenderFemale, for example, is somehow implicated in the impact that some other variable, say YrsExp, has on salary? If so, then a portion of the magnitude of the coefficient of YrsExp (the measurable effect of experience on salary) should actually be attributed to gender. Or, to put it another way, some of the effect of gender on salary is lost to experience. This means that our regression model is not measuring the true effect that gender has on salary. In addition, our understanding of the nature of any alleged discrimination at EnPact would be greatly increased if we could not only measure the effect that gender by itself makes on salary, but also measure the effect that the interplay or interaction between gender and years of experience makes on employees' salaries. Similarly, it would also be informative to learn, for example, that gender does not play a role in how some other variable, say education, affects salary.

These kinds of combined effects can be captured in regression models by forming new variables called interaction variables (or terms), which are created by taking the product of two variables that we believe have a combined effect on the dependent variable. The first entry in a column of data for an interaction variable $X_1 \times X_2$ is the product of the first entry of X_1 with the first entry of X_2. The second entry of $X_1 \times X_2$ is the product of the second entry of X_1 with the second entry of X_2, etc. When the interaction variables along with the original variables are submitted to StatPro's regression routine, its computational procedure makes no distinction between variables that are interaction variables and those which are not. When StatPro computes regression coefficients for any set of variables, it treats all columns of data with names at their heads the same, whether those names are GenderFemale, YrsExper, or GenderFemale*YrsExp. StatPro has a convenient routine for creating interaction terms under its Data Utilities menu.

The following is an example of a regression model containing interaction variables:

$$\begin{aligned}\text{Salary} \;=\; & 25 + 1.2 * \text{YrsExp} - 2.4 * \text{GenderFemale} - .80 * \text{GenderFemale*YrsExp} \\ & +1.30 * \text{GenderFemale*EducLev3} - .42 * \text{GenderFemale*EducLev6}\end{aligned}$$

Things to know about interaction terms when building models:

1. Variables that were significant before the introduction of interaction variables may become insignificant in subsequent models containing the interaction variables

2. The reverse can also occur. That is, variables that have been insignificant may become significant when combined in new interaction terms.

10.2.1 Definitions and Formulas

Interaction variable The product of two variables, say Female and Age, that constitutes a new variable and that captures, if it proves to be significant, the combined effect of the two original variables. An interaction variable is formed by multiplying the

corresponding cells of the two variables and placing the resulting products in a new column, usually denoted, for example, by Female × Age.

Interaction terms can be created from any two variables. Most commonly, though, they are created from interacting either two categorical variables, or a categorical variable and a numerical variable. Interaction variables created from two numerical variables really lead us away from linear models for the data and create one type of quadratic model (See chapter 13).

Base Variable These are the original "uninteracted" variables from which the interaction terms were created.

10.2.2 Worked Examples

Example 10.5. Creating and interpreting interaction terms from the EnPact data

An interaction term can be created from a numerical variable and a categorical variable:

Variable Type	Variable Name	Categories
The numerical variable	Age	N/A
The categorical variable	EducLev	EducLev1, EducLev2, EducLev3, EducLev4, EducLev5
The interaction variable	Age*EducLev	Age* EducLev1, Age* EducLev2, Age* EducLev3 Age* EducLev4, Age* EducLev5

We will interpret a rather simple model built on Age, EducLev3 and Age × EducLev3 where EducLev1 indicates a high-school grad and has been chosen as the reference category for the categorical variable EducLev, and EducLev3 indicates a college grad.

Model: Salary = 12 + .56*Age + 5.2*EducLev3 + .22*Age* EducLev3

Interpretation: When EducLev3 has the value 1, a college graduate is indicated. After substituting 1 for EducLev3 in the model equation, we have

Salary =12 + .56*Age + 5.2*1 + .22*Age* 1

After combing the Age terms, we have a college grad's salary:

Salary = 17.2 + .78*Age (1)

When EducLev3 has the value 0, a high-school graduate is indicated. After substituting 0 for EducLev3 in the model equation, we have

Salary =12 + .56*Age + 5.2*0 + .22*Age* 0

Simplifying, we have a high-school grad's salary:

Salary =12 + .56*Age (2)

Comparing equations (1) and (2), we see that a college grad receives a bonus of $5200 (17.2-12=5.2) for having a college degree plus an additional $220 (.78-.56=.220) for each year that he or she has lived compared to a high-school grad of the same age. At age 30, for example, a high-school grad earns $28,800 whereas a 30-year old college grad earns $40,600. At age 60, they earn $45,600 and $64,000, respectively.

Example 10.6. An interaction terms created from two categorical variables

Suppose we have the variables Gender and EducLev from the previous example, and we plan to construct an interaction term using these variables.

Gender:	GenderFemale, GenderMale
	Reference category: GenderMale
EducLev:	EducLev1, EducLev2, EducLev3, EducLev4, EducLev5
	Reference category: EducLev1

There are 2x5, or 10, interaction terms involved in the interaction variable Gender*Ed. Not all 10 can be submitted to StatPro's regression routine, however. Only those interaction terms that do not contain a reference for either variable may be submitted to the regression routine. The following interaction terms are the only ones that may be submitted to StatPro's regression routine:

EducLev2*GenderFemale
EducLev3*GenderFemale
EducLev4*GenderFemale
EducLev5*GenderFemale

The other interaction terms cannot be submitted to StatPro's regression routine because each contains either one or both of the reference categories (in bold) from which they are created: **EdLev1*** **GenderMale**, **EducLev1***GenderFemale , EducLev2***GenderMale**, EducLev3 * **GenderMale**, EducLev4 * **GenderMale**, EducLev5* **GenderMale**. This means that each of these is a reference category for the interaction variable EducLev*Gender.

We will interpret a modification of the models built above based on the variables Age, EducLev3, Age* EducLev3, GenderFemale and EducLev3*GenderFemale.

$$\begin{aligned}\text{Model: Salary} &= 13 + .52 * \text{Age} + 5.8 * \text{EducLev3} + .21 * \text{Age} * \text{EducLev3} \\ &\quad +4.1 * \text{GenderFemale} - 2.5 * \text{EducLev3*GenderFemale}\end{aligned}$$

Interpretation: If GenderFemale = 0 and EducLev3 = 1, we have a male college graduate. Substituting these values in the model equation, we have

Salary = 13 + .52*Age + 5.8*1 + .21*Age* 1 + 4.1*0 - 2.5*1*0

Combining the constants and the Age terms, we have the equation for a male college graduate

$$\text{Salary} = 18.8 + .73*\text{Age} \quad (3)$$

If GenderFemale = 1 and EducLev3 = 1, we have a female college graduate. Substituting these values in the model equation, we have

$$\text{Salary} = 13 + .52*\text{Age} + 5.8*1 + .21*\text{Age}* 1 + 4.1*1 - 2.5*1*1 \quad (4)$$

In equation (4) we see that a female receives \$4100 more than a male on the basis of gender alone. But she will receive \$2500 less than a male if she has a college degree. Simplifying (4), we have the equation for a female college graduate:

$$\text{Salary} = 20.4 + .73*\text{Age} \quad (5)$$

Comparing (3) and (5), we see that a female college graduate earns on the average of \$1600 (20.4-18.8) more than a male college graduate. The difference is larger, however, for high school graduates (EducLev3 = 0). In this case, female high-school graduates earn \$4100 a year more than male graduates. For example, comparing the salaries of 25-year old high school graduates, we have:

$$\begin{array}{ll} \text{Female:} & \text{Salary} = 13 + .52 * 25 + 5.8 * 0 + .21 * 25 * 0 + 4.1 * 1 - 2.5 * 0 * 1 \\ & = \$30{,}100 \\ \text{Male:} & \text{Salary} = 13 + .52 * 25 + 5.8 * 0 + .21 * 25 * 0 + 4.1 * 0 - 2.5 * 0 * 0 \\ & = \$26{,}000 \end{array}$$

Example 10.7. Simplifying variables in the EnPact data
When we introduce interaction variables into the EnPact gender discrimination study, we find that if we use the given variable names as they are found in C11 EnPact.xls StatPro will create interaction variable names that are too long to be completely viewed in its multiple regression routine window. In addition, when we interact categorical variables with other variables, particularly other categorical variables, the number of possible models from which we must find an optimal model increases greatly, depending on the number of categories involved in creating the interaction terms. There are situations, therefore, in which we have to not only shorten variable names but also combine certain categories together in a meaningful way in order to reduce the number of models we have to analyze. We illustrate how to do this with the EnPact data spreadsheet:

1. Shorten the variable name "EducLev" to "Ed" by retyping directly in cell B3

2. At the top of a blank column just to the right of the Salary column, type the variable name "Female" (do not use quotes). This variable will be a discrete numerical variable with values 0 and 1 to indicate the employee's gender. If Female has value 1, we have a female employee, whereas if Female has value 0 we have a male. We do this by placing the following conditional statement in the first data cell of our new Female variable: =IF(F4="Female",1,0). Then we sweep down the column.

3. Following the directions under Option #1 of the How to Guide for Section 9B, use StatPro to generate one dummy variable based on the categorical variable JobGrade that is coded. Use a Cutoff value of 4 and check "Greater than" in the Dummy variable definition window. This will create a discrete numerical variable called JobGradeGT4. See figure 10.4. Change the name JobGradeGE4 to HiJob as shown in figure 10.5. HiJob has value 1 if JobGrade is 5 or 6 (this designates a higher level job) and has value 0 if JobGrade is 1, 2, 3, or 4 (this designates a lower job level).

4. Convert "Ed" to a set of dummy variables, Ed1, Ed2, Ed3, and so forth. See figure 10.6.

	A	B	C	D	E	F	G	H	I	J
3	Employee	Ed	JobGrade	Age	YrsExp	Gender	YrsPrior	Salary	Female	JobGrade_GT4
4	1	3	1	26	3	Male	1	35.4	0	0
5	2	1	1	38	14	Female	1	41.6	1	0
6	3	1	1	35	12	Female	0	35.8	1	0
7	4	2	1	40	8	Female	7	34.1	1	0
8	5	3	1	28	3	Male	0	31.9	0	0
9	6	3	1	24	3	Female	0	33.1	1	0
10	7	3	1	27	4	Female	0	32.8	1	0

Figure 10.4: Steps 1, 2, and 3 of example 7 illustrated.

	A	B	C	D	E	F	G	H	I	J
3	Employee	Ed	JobGrade	Age	YrsExp	Gender	YrsPrior	Salary	Female	HiJob
4	1	3	1	26	3	Male	1	35.4	0	0
5	2	1	1	38	14	Female	1	41.6	1	0
6	3	1	1	35	12	Female	0	35.8	1	0
7	4	2	1	40	8	Female	7	34.1	1	0
8	5	3	1	28	3	Male	0	31.9	0	0
9	6	3	1	24	3	Female	0	33.1	1	0
10	7	3	1	27	4	Female	0	32.8	1	0

Figure 10.5: Step 3 of example 7 completed.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
3	Employee	Ed	JobGrade	Age	YrsExp	Gender	YrsPrior	Salary	Female	HiJob	Ed_1	Ed_2	Ed_3	Ed_4	Ed_5
4	1	3	1	26	3	Male	1	35.4	0	0	0	0	1	0	0
5	2	1	1	38	14	Female	1	41.6	1	0	1	0	0	0	0
6	3	1	1	35	12	Female	0	35.8	1	0	1	0	0	0	0
7	4	2	1	40	8	Female	7	34.1	1	0	0	1	0	0	0
8	5	3	1	28	3	Male	0	31.9	0	0	0	0	1	0	0
9	6	3	1	24	3	Female	0	33.1	1	0	0	0	1	0	0

Figure 10.6: Step 4 of example 7

10.2.3 Exploration 10B: Complex Gender Interactions at EnPact

Part 1. Simplify the variables in the EnPact data file (C11 EnPact Data.xls) until your data spreadsheet looks like the spreadsheet in Step 2 of example 7 (page 298). By this simplification of our data, we now have only one categorical variable, Ed, with 5 categories. Female and HiJob are now discrete numerical variables with values 0 or 1. This is important to know when we create interaction terms in the next part. We will use Ed1, high-school graduate, as the reference category when we begin building our models.

Part 2. Use StatPro to create the following interaction variables (see How to Guide for this section): YrsExp*HiJob, Female* YrsExp, Female* YrsPrior, Female* HiJob, Female*Ed. As we noted in the above step, the only variable that you will check as categorical in StatPro's routine for creating interaction variables is Ed. Moreover, when you select your variables for regression analysis, do not select Female*Ed1 since it is the reference category for the Female*Ed categorical variable.

Part 3. StatPro's stepwise regression routine to create a regression model using the following variables and interaction variables: Base Variables: YrsExp, YrsPrior, Female, HiJob, Ed2, Ed3, Ed4, Ed5 Numerical-Categorical Interactions: YrsExp*HiJob Female* YrsExp, Female* Age, Female* YrsPrior Categorical-Categorical Interactions: Female* HiJob, Female* Ed2, Female* Ed3, Female* Ed4, Female* Ed5

Part 4. Explain what goes into determining salary at EnPact and what role gender plays in the salary structure in terms of experience, education and job level. Then give a thumbnail description of life at EnPact for women.

10.2.4 How To Guide

Creating interaction terms with StatPro

1. Go to StatPro/Data utilities/Create Interaction Variable(s)
2. Click past the next window
3. Select the two variables you wish to interact. Remember to hold the control key in order to select variables that are not listed next to each other on the list. Click OK.

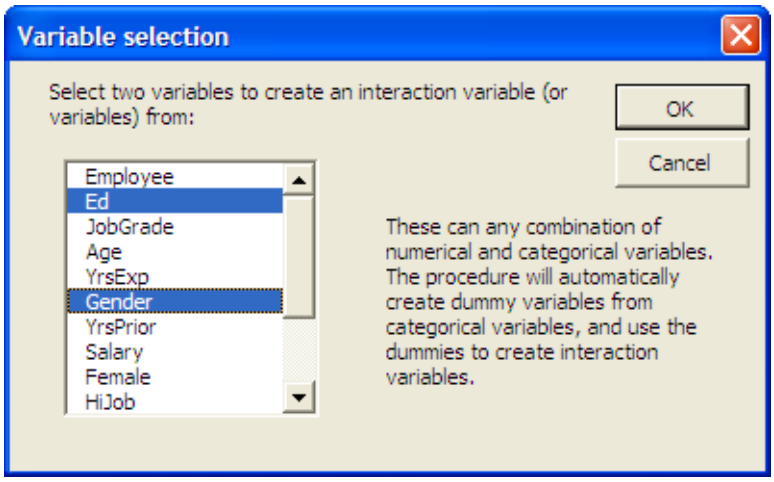


Figure 10.7: Selecting variables to create interaction terms with StatPro.

4. Check any of the two variables that happen to be categorical variables. Do NOT check numerical variables, even if they were originally created from categorical variables. Click OK.

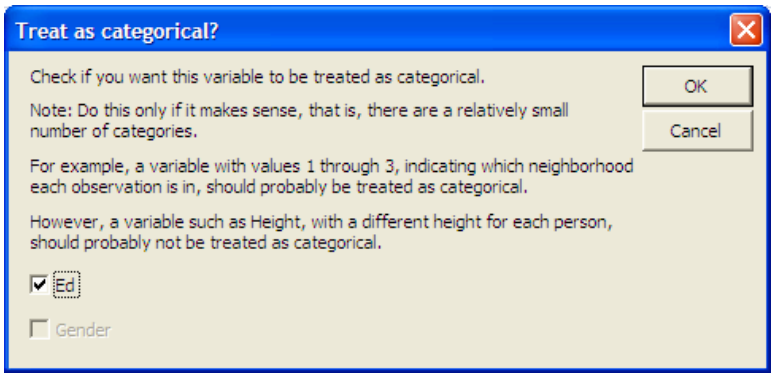


Figure 10.8: Ensuring that StatPro treats the Education Level (Ed) as a categorical variable.

You can see a portion below of the interaction terms that StatPro has created.

	P	Q	R	S	T	U	V
3	Ed_1*Gender_Female	Ed_1*Gender_Male	Ed_2*Gender_Female	Ed_2*Gender_Male	Ed_3*Gender_Female	Ed_3*Gender_Male	Ed_4*Gender_Female
4	0	0	0	0	0	1	0
5	1	0	0	0	0	0	0
6	1	0	0	0	0	0	0
7	0	0	1	0	0	0	0
8	0	0	0	0	0	1	0
9	0	0	0	0	1	0	0
10	0	0	0	0	1	0	0

Figure 10.9: A portion of the new interaction terms created by StatPro after the above steps.

10.3 Homework

10.3.1 Mechanics and Techniques Problems

10.1. Bring up the Excel file C10 Laptops.xls.

1. Change the variable names "Manufacturer" to "Manu" so that interaction terms will be short enough to view in StatPro regression windows.
2. Form dummy variables for the categorical variable Manu.
3. Create interaction terms for Manu*Wt

10.2. Submit the following variables to StatPro's multiple regression routine with Price as the dependent variable:

1. The numerical variable Weight
2. The dummy variables for the categorical variables Manu
3. The dummy variables for the interaction terms for Manu*Wt
4. Let Sony be the reference category for Manu. Reminder: this choice of reference category for Manu automatically determines the reference category for Manu*Wt.

10.3.2 Application and Reasoning Problems

10.3. Write the regression equation for Price.

1. What is the predicted price of each of the following types of laptops?

Model of Laptop	Equation to Predict Price
Sony	
Compaq	
Hp	
Toshiba	

2. Explain how a computer brand's weight affects its price. Do heavier computer brands cost more? Or less?

10.4. Interpret the following model related to the laptop prices in the previous example:

$$\text{Price} = 560 + 115\text{*Wt} + 230\text{* ManuToshiba*Wt}$$

10.3.3 Memo Problem

To: Analysis Staff
From: Project Management Director
Date: May 27, 2008
Re: New Truck Contract

As you know, we have been doing some work for Ms. Mini Driver, the Director of Operations at MetroArea Trucking, on how location affects the maintenance expenses for the trucks in the fleet. We have received an additional contract to further analyze the fleet's maintenance expenses. Ms. Mini Driver would like us to analyze the entire truck data set (see attachment), which includes last year's maintenance expense, the mileage, age, and type of truck, as well as the location (based either in city or out of city) of where the truck is based. Ms. Mini Driver wants us to provide her with an analysis of what factors affect maintenance expenses and how much each affects the expenses.

I'd like you to develop your own optimal regression model by choosing your own variables and going through your own model-refining process before seeing what StatPro's stepwise regression routine produces for an optimal model. This process should give you a better feel for how the variables contribute to the maintenance expense, which should be helpful when you interpret your models.

1. Start with a full model without any interaction terms and record your findings in the chart below. I would like you to begin this way because there are situations when interaction terms aren't really worth their trouble, whereas in others they are.
2. Run the reduced model with the significant variables that you get from the full model, again without any interaction terms. Record your findings in the chart.
3. Start over with a full model with all interaction terms. Record your findings.
4. Run a reduced model with the significant variables only. Record your findings.
5. Now run a full model with all interaction terms using StatPro's stepwise regression routine. Record your findings.
6. Write a memo to me stating what you think the model should be and why, including a description of how you went about finding your model. Be sure to include your supporting evidence (you will find the chart helpful here). Comment on the quality of your model and then interpret your model, explaining which variables significantly affect maintenance expenses and how much each affects the expenses.

Attachment: Data file "C10 Truck.XLS"

Model	R^2	Adj R^2	S_e	List of significant variables
Full Model With no Interactions				
Reduced Model with no interactions				
Full Model with all interactions				
Reduced Model with significant interactions				
StatPro Stepwise regression				