

Chapter 18

Model Building

Regression Analysis

Regression analysis is one of the most powerful and commonly used techniques in statistics; it allows us to create mathematical models that *realistically* describe relationships between the dependent variable and independent variables.

We've seen it used for linear models using interval data, but regression analysis can also be used for:

- non-linear (polynomial) models, and
- models that include *nominal* independent variables.

Polynomial Models

Previously we looked at this multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

(it's considered linear or first-order since the exponent on each of the x_i 's is 1)

The independent variables may be *functions* of a smaller number of predictor variables; polynomial models fall into this category. If there is *one predictor value* (\mathbf{x}) we have:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots + \beta_P x^P + \varepsilon$$

Polynomial Models

$$\textcircled{1} \quad y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots + \beta_p x^p + \varepsilon$$

$$\textcircled{2} \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_p x_p + \varepsilon$$

Technically, equation $\textcircled{1}$ is a multiple regression model with \mathbf{p} independent variables (x_1, x_2, \dots, x_p). Since $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$, \dots , $x_p = x^p$, it's based on one predictor value (x).

\mathbf{p} is the **order** of the equation; we'll focus equations of order $p = 1, 2$, and 3 .

First Order Model

When $p = 1$, we have our simple linear regression model:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

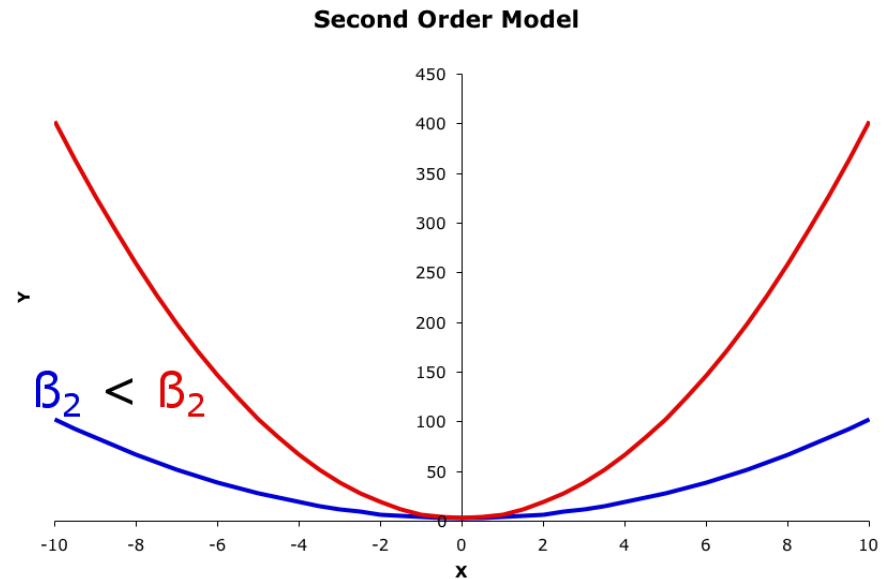
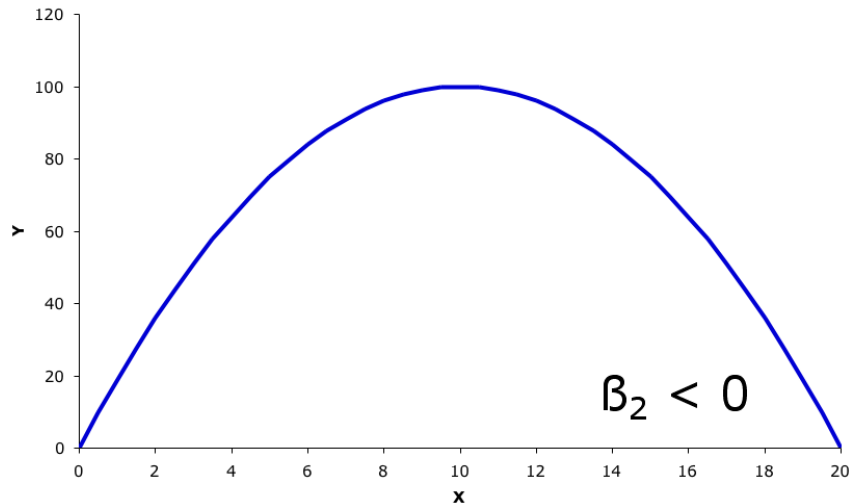
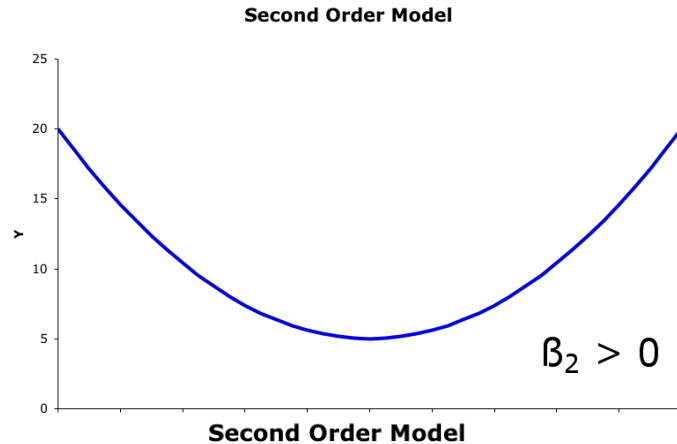
That is, we believe there is a *straight-line relationship* between the dependent and independent variables over the range of the values of x :



Second Order Model

When $p = 2$, the polynomial model is a parabola:

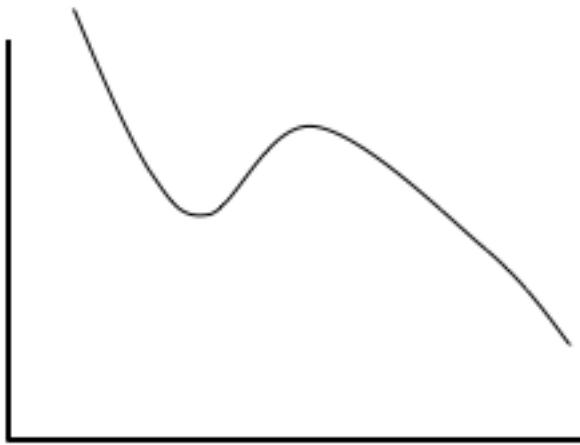
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$



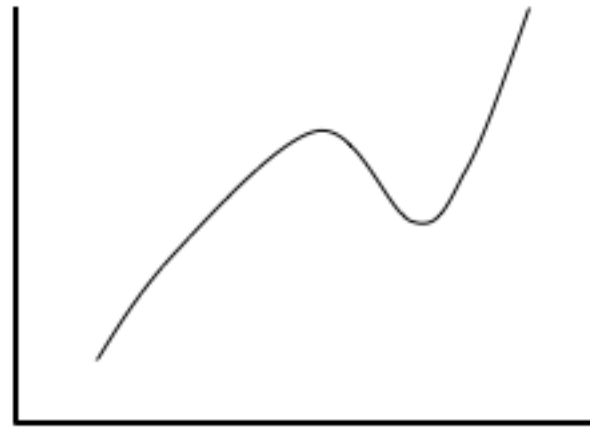
Third Order Model

When $p = 3$, our third order model looks like:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$



$$\beta_3 < 0$$



$$\beta_3 > 0$$

Polynomial Models: 2 Predictor Variables

Perhaps we suspect that there are two predictor variables (x_1 & x_2) which influence the dependent variable:

First order model (no interaction):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

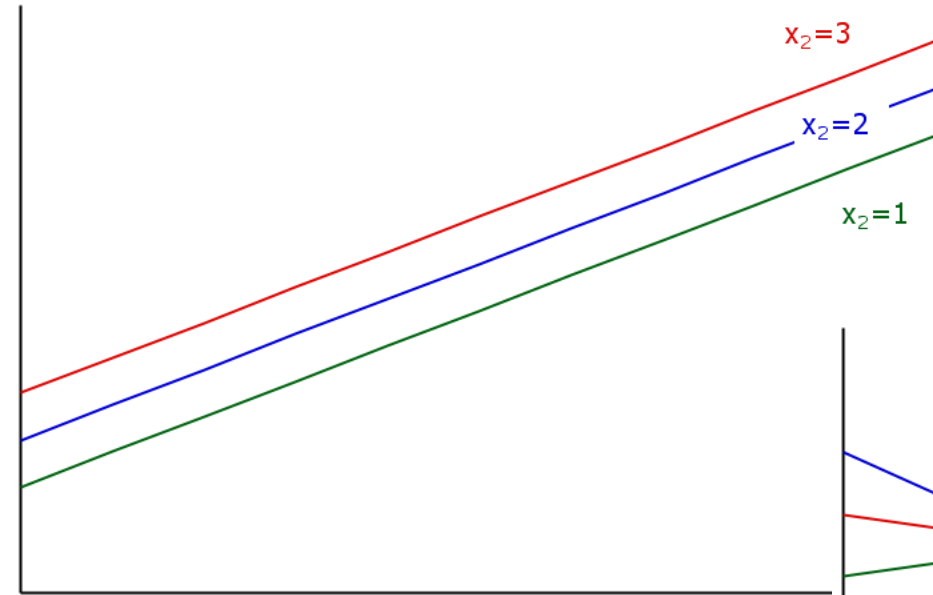
First order model (*with interaction*):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \boxed{\beta_3 x_1 x_2} + \varepsilon$$

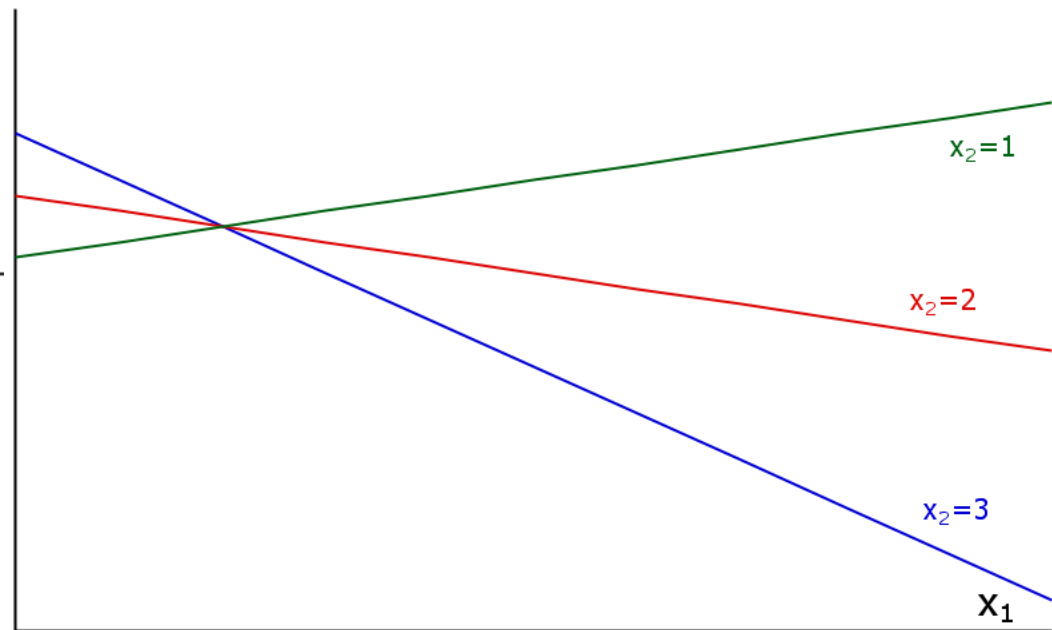
Polynomial Models: 2 Predictor Variables

First order models, 2 predictors, without & with interaction:

First Order Model, 2 Ind. Vars., No Interaction

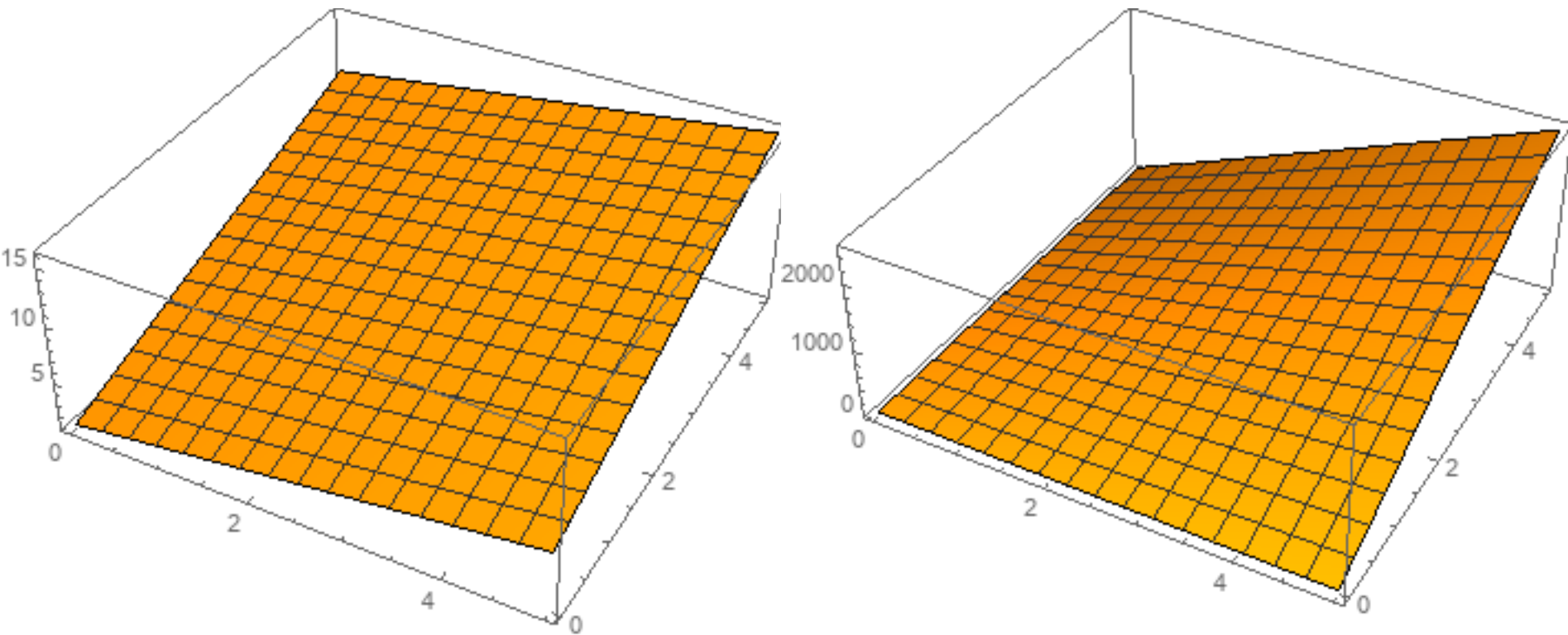


First Order Model WITH Interaction



Polynomial Models: 2 Predictor Variables

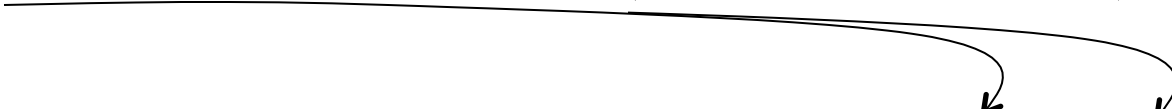
First order models, 2 predictors, without & with interaction:



Polynomial Models: 2 Predictor Variables

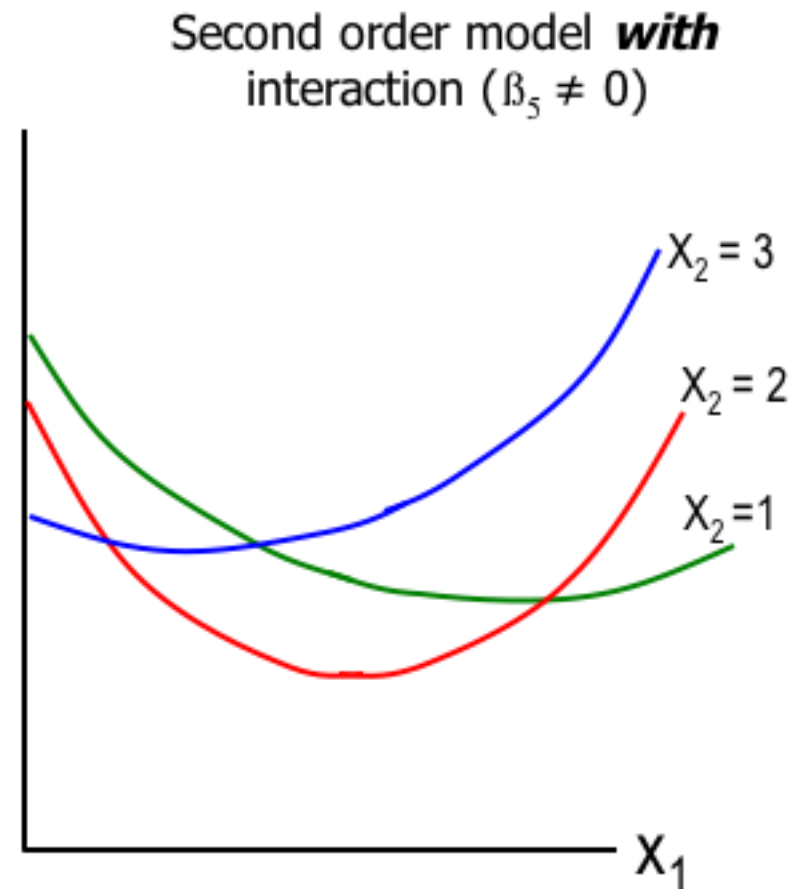
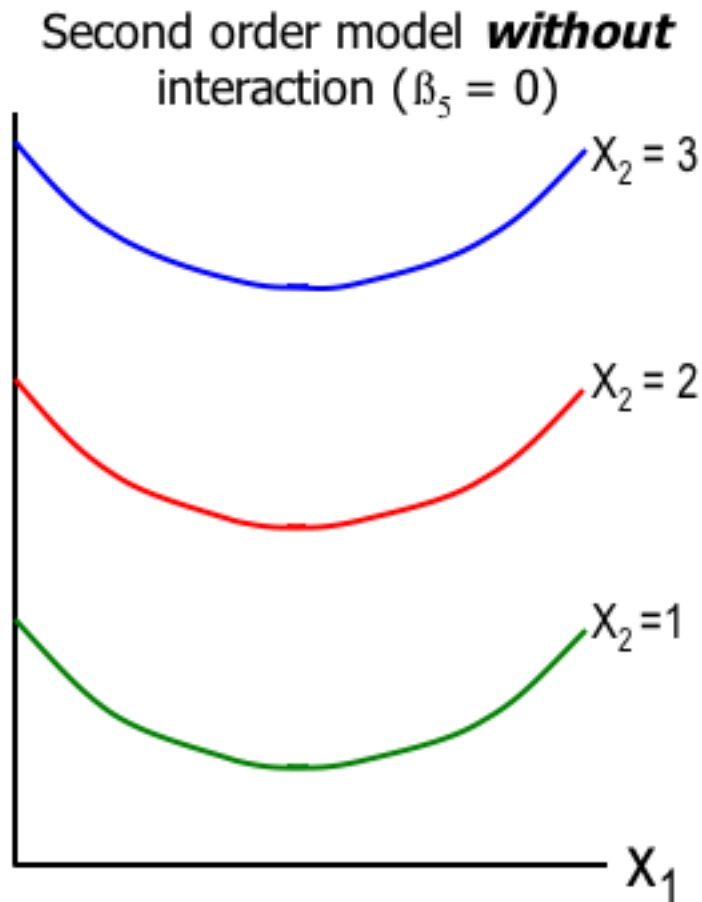
If we believe that a quadratic relationship exists between y and each of x_1 and x_2 , ***and*** that the predictor variables ***interact*** in their effect on y , we can use this model:

Second order model (in two variables) WITH ***interaction***:


$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \boxed{\beta_5 x_1 x_2} + \varepsilon$$

Polynomial Models: 2 Predictor Variables

2nd order models, 2 predictors, without & with interaction:



Selecting a Model

One predictor variable, or two (or more)?

First order? Second order? Higher order?

With interaction? Without?

How do we choose the right model??

Use our knowledge of the variables involved to build an initial model.

Test that model using statistical techniques.

If required, modify our model and re-test...

Example 18.1

We've been asked to come up with a regression model for a fast food restaurant. We know our primary market is middle-income adults and their children, particularly those between the ages of 5 and 12.

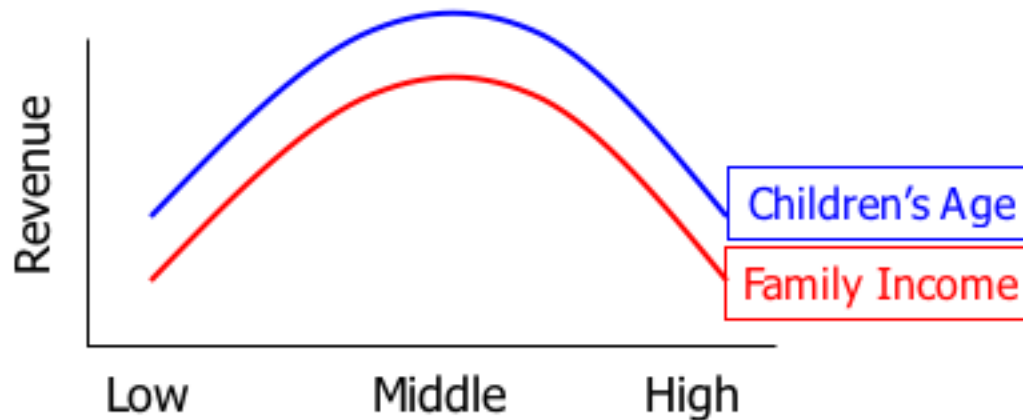
Dependent variable — restaurant revenue (gross or net)

Predictor variables — family income, age of children

Is the relationship first order? quadratic?...

Example 18.1

The relationship between the dependent variable (revenue) and each predictor variable is probably **quadratic**.



Members of low or high income households are less likely to eat at this chain's restaurants, since **the restaurants attract mostly middle-income customers**.

Neighborhoods where the mean age of children is either quite low or quite high are also less likely to eat there vs. the **families with children in the 5-to-12 year range**.

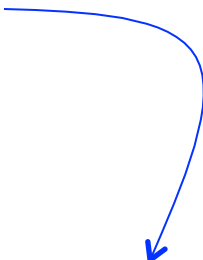
Seems reasonable?

Example 18.1

Should we include the interaction term in our model?

*When in doubt, it is probably best to **include** it.*

Our model then, is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon$$


Where y = annual gross sales

x_1 = median annual household income*

x_2 = mean age of children*

*in the neighborhood

Example 18.2

Our fast food restaurant research department selected 25 locations at random and gathered data on revenues, household income, and ages of neighborhood children.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon$$

	A	B	C	D	E	F
1	Revenue	Income	Age	Income sq	Age sq	(Income)(Age)
2	1128	23.5	10.5	552.25	110.25	246.75
3	1005	17.6	7.2	309.76	51.84	126.72
25	1233	24.3	8.3	590.49	68.89	201.69
26	950	17.8	6.1	316.84	37.21	108.58
27						

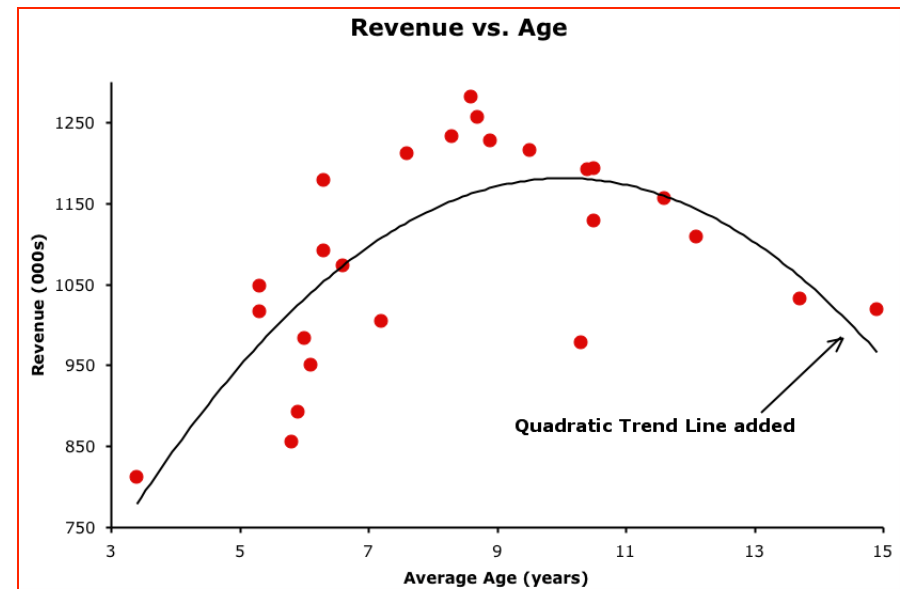
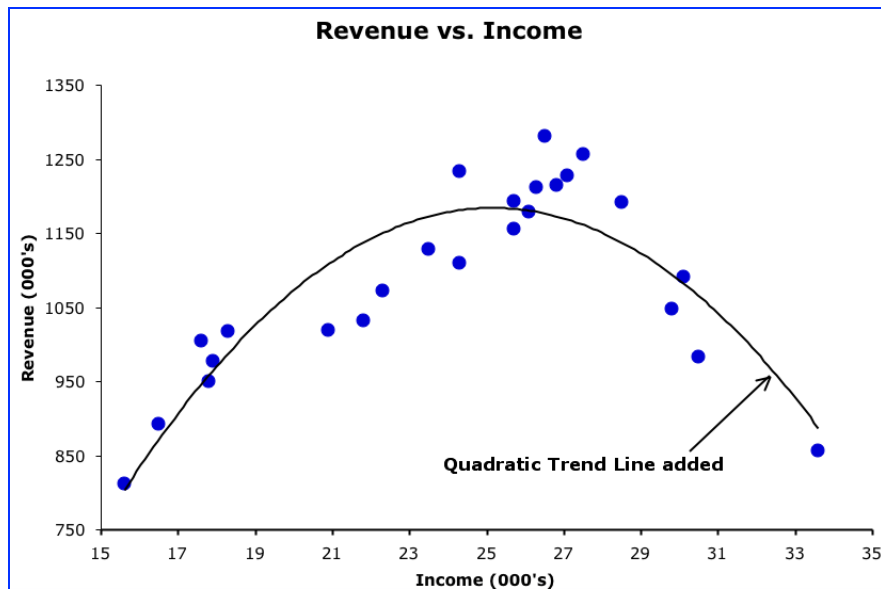
Collected Data

Calculated Data

Xm18-02

Example 18.2

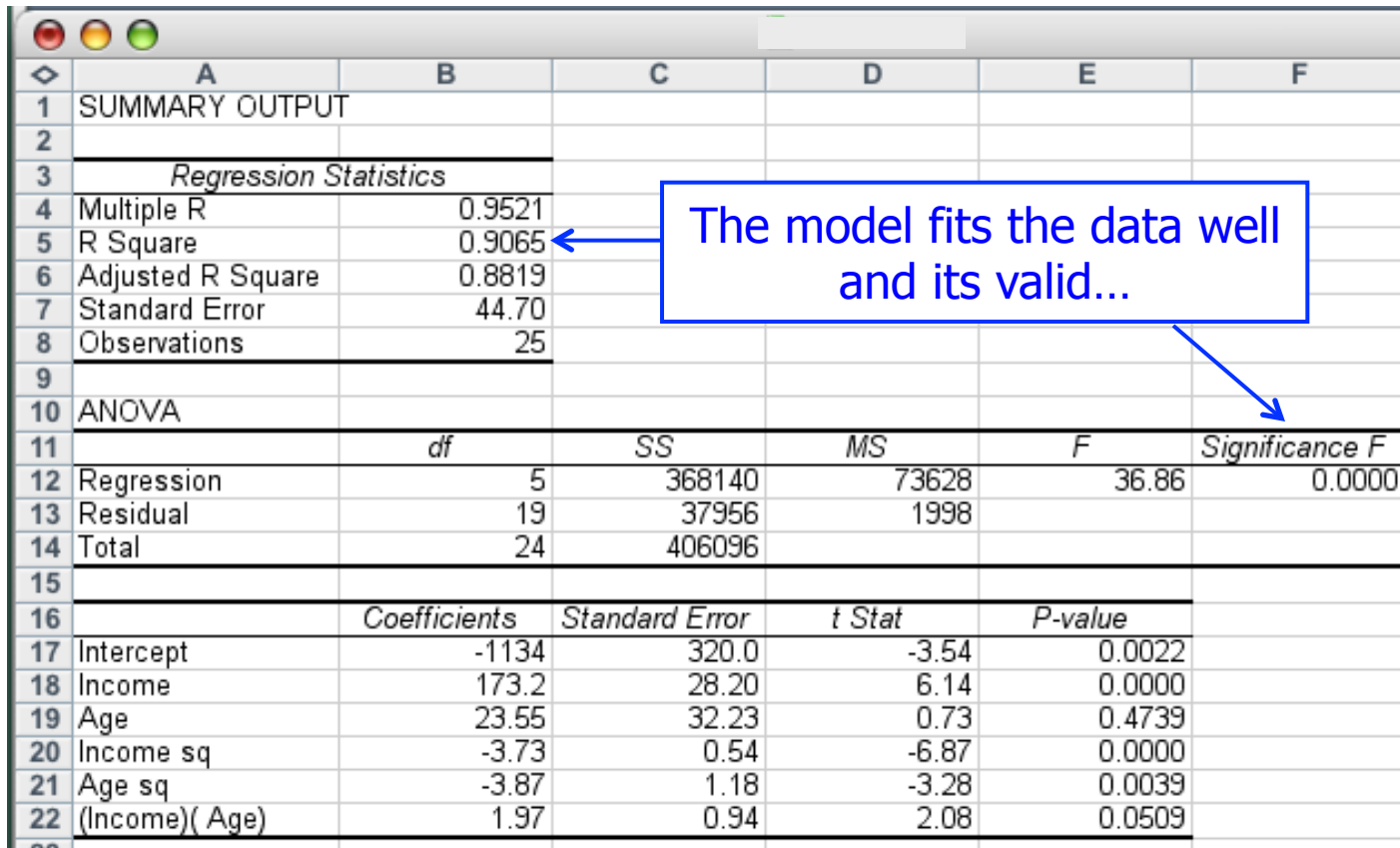
You can take the original data collected (revenues, household income, and age) and plot **y vs. x_1** and **y vs. x_2** to get a feel for the data; trend lines were added for clarity...



Example 18.2

INTERPRET

Checking the regression tool's output...



The image shows a screenshot of a regression output window. A blue callout box with the text "The model fits the data well and its valid..." has two arrows pointing to the "R Square" value (0.9065) in the "Regression Statistics" section and the "Significance F" value (0.0000) in the "ANOVA" table.

	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	<i>Regression Statistics</i>					
4	Multiple R	0.9521				
5	R Square	0.9065				
6	Adjusted R Square	0.8819				
7	Standard Error	44.70				
8	Observations	25				
9						
10	<i>ANOVA</i>					
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
12	Regression	5	368140	73628	36.86	0.0000
13	Residual	19	37956	1998		
14	Total	24	406096			
15						
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	
17	Intercept	-1134	320.0	-3.54	0.0022	
18	Income	173.2	28.20	6.14	0.0000	
19	Age	23.55	32.23	0.73	0.4739	
20	Income sq	-3.73	0.54	-6.87	0.0000	
21	Age sq	-3.87	1.18	-3.28	0.0039	
22	(Income)(Age)	1.97	0.94	2.08	0.0509	

Nominal Independent Variables

Thus far in our regression analysis, we've only considered variables that are *interval*. Often however, we need to consider *nominal data* in our analysis.

For example, our earlier example regarding the market for used cars focused only on mileage. Perhaps color is an important factor. How can we model this new variable?

Indicator Variables

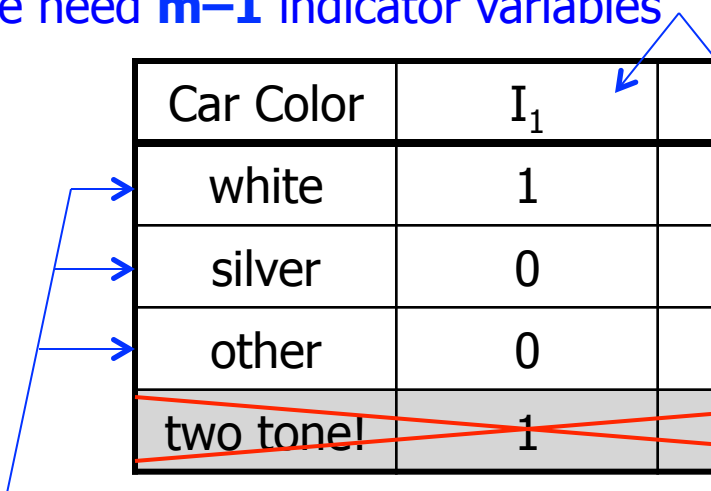
An *indicator variable* (also called a *dummy variable*) is a variable that can assume *either one of only two values* (usually 0 and 1).

A value of one usually indicates the existence of a certain condition, while a value of zero usually indicates that the condition does not hold.

we need $m-1$ indicator variables

$$I_1 = \begin{cases} 0 & \text{if color **not** white} \\ 1 & \text{if color is white} \end{cases}$$

$$I_2 = \begin{cases} 0 & \text{if color **not** silver} \\ 1 & \text{if color is silver} \end{cases}$$



Car Color	I_1	I_2
white	1	0
silver	0	1
other	0	0
two tone!	1	1

to represent m categories...

Interpreting Indicator Variable Coefficients

After performing our regression analysis:

		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
16					
17	Intercept	16.837	0.197	85.42	0.0000
18	Odometer	-0.0591	0.0051	-11.67	0.0000
19	I-1	0.0911	0.0729	1.25	0.2143
20	I-2	0.3304	0.0816	4.05	0.0001
21					

we have this regression equation...

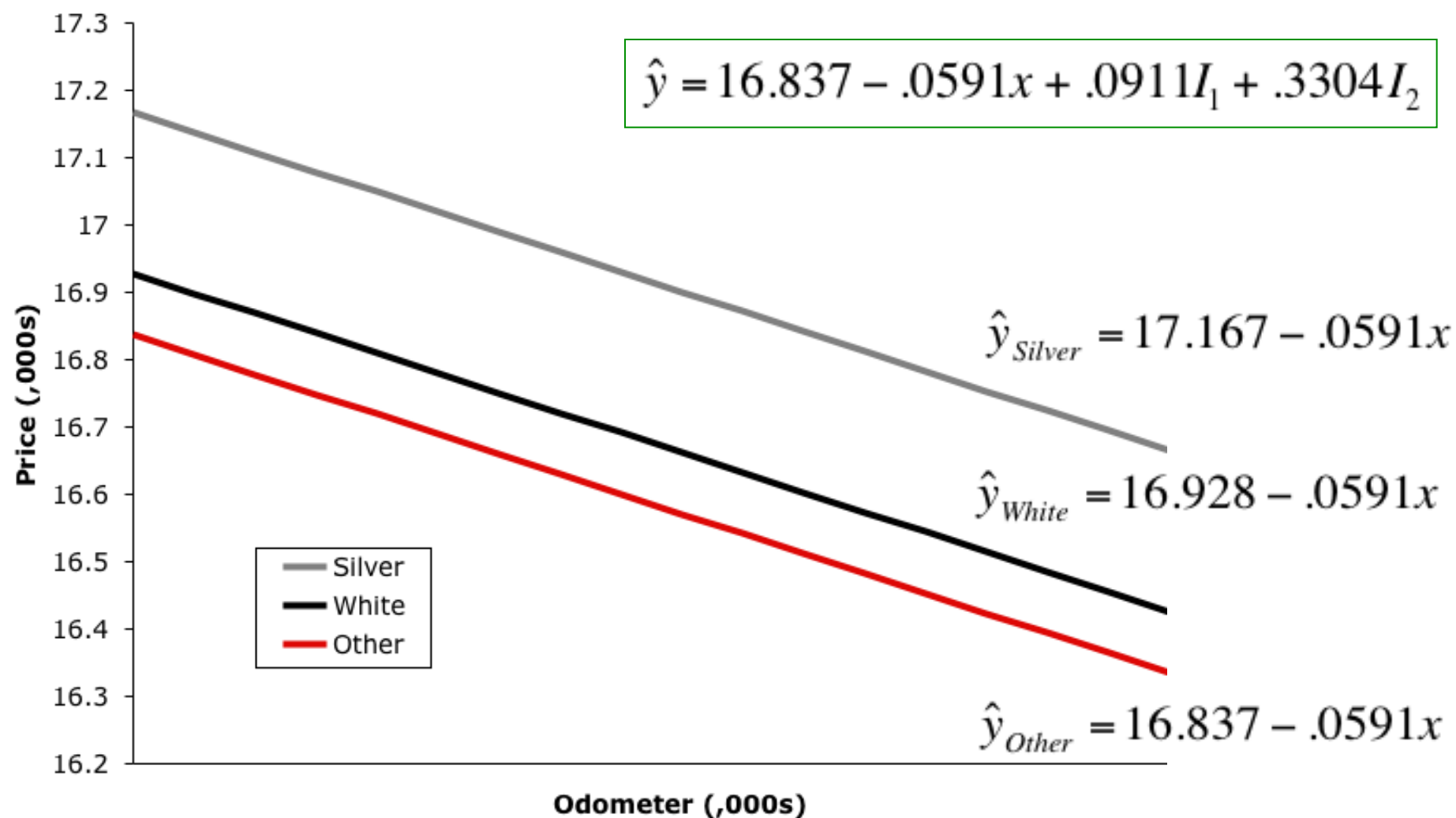
$$\hat{y} = 16.837 - .0591x + .0911I_1 + .3304I_2$$

Thus, the price diminishes with additional mileage (x)

a white car sells for \$91.10 more than other colors (I_1)

a silver car fetches \$330.40 more than other colors (I_2)

Graphically



Testing the Coefficients

To test the coefficient of I_1 , we use these hypotheses...

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

β_0

β_1

β_2

β_3

		Coefficients	Standard Error	t Stat	P-value
16	Intercept	16.837	0.197	85.42	0.0000
17	Odometer	-0.0591	0.0051	-11.67	0.0000
18	I-1	0.0911	0.0729	1.25	0.2143
19	I-2	0.3304	0.0816	4.05	0.0001
20					

There is **insufficient evidence** to infer that in the population of 3-year-old **white** Tauruses with the same odometer reading have a **different** selling price than do Tauruses in the "other" color category...

Testing the Coefficients

To test the coefficient of I_2 , we use these hypotheses...

$$H_0: \beta_3 = 0$$

$$H_1: \beta_3 \neq 0$$

$H_1: \beta_3 \neq 0$

		Coefficients	Standard Error	t Stat	P-value	
β_0	16					
β_1	17	Intercept	16.837	0.197	85.42	0.0000
	18	Odometer	-0.0591	0.0051	-11.67	0.0000
β_2	19	I-1	0.0911	0.0729	1.25	0.2143
β_3	20	I-2	0.3304	0.0816	4.05	0.0001

We can conclude that there **are differences** in auction selling prices between all 3-year-old **silver**-colored Tauruses and the “other” color category with the same odometer readings

Nominal Independent Variables; Example: MBA Program Admission (MBA II)

- Recall: The Dean wanted to evaluate applications for the MBA program by predicting future performance of the applicants.
- The following three predictors were suggested:
 - Undergraduate GPA
 - GMAT score
 - Years of work experience
- It is now believed that the type of undergraduate degree should be included in the model.

Note: The undergraduate degree is nominal data.

Nominal Independent Variables; Example: MBA Program Admission (II)

$$I_1 = \begin{array}{l} 1 \text{ if B.A.} \\ 0 \text{ otherwise} \end{array}$$

$$I_2 = \begin{array}{l} 1 \text{ if B.B.A} \\ 0 \text{ otherwise} \end{array}$$

$$I_3 = \begin{array}{l} 1 \text{ if B.Sc. or B.Eng.} \\ 0 \text{ otherwise} \end{array}$$

The category “Other group” is defined by:

$$I_1 = 0; I_2 = 0; I_3 = 0$$

Nominal Independent Variables; Example: MBA Program Admission (II)

MBA-II

SUMMARY OUTPUT					
<i>Regression Statistics</i>					
Multiple R	0.7461				
R Square	0.5566				
Adjusted R Square	0.5242				
Standard Error	0.729				
Observations	89				
ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	6	54.75	9.13	17.16	0.0000
Residual	82	43.62	0.53		
Total	88	98.37			
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	
Intercept	0.19	1.41	0.13	0.8930	
UnderGPA	-0.0061	0.114	-0.05	0.9577	
GMAT	0.0128	0.0014	9.43	0.0000	
Work	0.098	0.030	3.24	0.0017	
I-1	-0.34	0.22	-1.54	0.1269	
I-2	0.71	0.24	2.93	0.0043	
I-3	0.03	0.21	0.17	0.8684	

Applications in Human Resources Management: Pay-Equity

- Pay-equity can be handled in two different forms:
 - Equal pay for equal work
 - Equal pay for work of equal value.
- Regression analysis is extensively employed in cases of equal pay for equal work.

Human Resources Management: Pay-Equity

- Example 18.3 (Xm18-03)
 - Is there sex discrimination against female managers in a large firm?
 - A random sample of 100 managers was selected and data were collected as follows:
 - Annual salary
 - Years of education
 - Years of experience
 - Gender

Human Resources Management: Pay-Equity

- Solution
 - Construct the following multiple regression model:
$$y = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Experience} + \beta_3 \text{Gender} + \varepsilon$$
 - Note the nature of the variables:
 - Education – Interval
 - Experience – Interval
 - Gender – Nominal (Gender = 1 if male; =0 otherwise).

Human Resources Management:

Pay-Equity

- Solution – Continued (Xm18-03)

SUMMARY OUTPUT					
<i>Regression Statistics</i>					
Multiple R	0.8326				
R Square	0.6932				
Adjusted R Square	0.6836				
Standard Error	16274				
Observations	100				
ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	3	57434095083	19144698361	72.29	0.0000
Residual	96	25424794888	264841613.4		
Total	99	82858889971			
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	
Intercept	-5835.1	16082.8	-0.36	0.7175	
Education	2118.9	1018.5	2.08	0.0401	
Experience	4099.3	317.2	12.92	0.0000	
Gender	1851.0	3703.1	0.50	0.6183	

Analysis and Interpretation

- The model fits the data quite well.
- The model is very useful.
- Experience is a variable strongly related to salary.
- There is no evidence of sex discrimination.

Human Resources Management:

Pay-Equity

- Solution – Continued (Xm18-03)

SUMMARY OUTPUT					
<i>Regression Statistics</i>					
Multiple R	0.8326				
R Square	0.6932				
Adjusted R Square	0.6836				
Standard Error	16274				
Observations	100				
ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	3	57434095083	19144698361	72.29	0.0000
Residual	96	25424794888	264841613		
Total	99	82858889971			
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	
Intercept	-5835.1	16082.8	-0.36	0.7175	
Education	2118.9	1018.5	2.08	0.0401	
Experience	4099.3	317.2	12.92	0.0000	
Gender	1851.0	3703.1	0.50	0.6183	

Analysis and Interpretation

- Further studying the data we find:
Average experience (years) for women is 12.
Average experience (years) for men is 17
- Average salary for female manager is \$76,189
Average salary for male manager is \$97,832

Stepwise Regression

- Multicollinearity may prevent the study of the relationship between dependent and independent variables.
- The correlation matrix may fail to detect multicollinearity because variables may relate to one another in various ways.
- To reduce multicollinearity we can use stepwise regression.
- In stepwise regression variables are added to or deleted from the model one at a time, based on their contribution to the current model.

Variable-Selection Procedures

- Stepwise Regression

- At each iteration, the first consideration is to see whether the least significant variable currently in the model can be removed because its F value, F_{MIN} , is less than the user-specified or default F value, F_{REMOVE} . (That is, check the P -value of t -tests for the slope parameters.)

- If no variable can be removed, the procedure checks to see whether the most significant variable not in the model can be added because its F value, F_{MAX} , is greater than the user-specified or default F value, F_{ENTER} . (That is, check the P -value of t -tests for the slope parameters.)

- If no variable can be removed and no variable can be added, the procedure stops.

Variable-Selection Procedures

- For example
- The first thing we need to do is set a significance level for deciding when to enter a predictor into the stepwise model. We'll call this the **Alpha-to-Enter** significance level and will denote it as α_E .
- We also need to set a significance level for deciding when to remove a predictor from the stepwise model. We'll call this the **Alpha-to-Remove** significance level and will denote it as α_R .

Variable-Selection Procedures

- For example
- Alpha-to-Enter significance level will typically be greater than the usual 0.05 level so that it is not too difficult to enter predictors into the model.
Many software packages — Minitab included — set this significance level by default to $\alpha_E = 0.15$.
- Alpha-to-Remove significance level will typically be greater than the usual 0.05 level so that it is not too easy to remove predictors from the model.
Again, many software packages — Minitab included — set this significance level by default to $\alpha_R = 0.15$.

Variable-Selection Procedures

- **Step #1.**
- Once we've specified the starting significance levels, then we:
- Fit each of the one-predictor models — that is, regress y on x_1 , regress y on x_2 , ..., and regress y on x_p .
- Of those predictors whose t -test P -value is less than $\alpha_E = 0.15$, the first predictor put in the stepwise model is the predictor that has the smallest t -test P -value.
- If no predictor has a t -test P -value less than $\alpha_E = 0.15$, stop.

Variable-Selection Procedures

- **Step #2.**
- Suppose x_1 had the smallest t -test P -value below $\alpha_E = 0.15$ and therefore was deemed the "best" single predictor arising from the the first step.
- Now, fit each of the two-predictor models that include x_1 as a predictor — that is, regress y on x_1 and x_2 , regress y on x_1 and x_3 , ..., and regress y on x_1 and x_p .
- Of those predictors whose t -test P -value is less than $\alpha_E = 0.15$, the second predictor put in the stepwise model is the predictor that has the smallest t -test P -value.

Variable-Selection Procedures

- If no predictor has a t -test P -value less than $\alpha_E = 0.15$, stop. The model with the one predictor obtained from the first step is your final model.
- But, suppose instead that x_2 was deemed the "best" second predictor and it is therefore entered into the stepwise model.
- Now, since x_1 was the first predictor in the model, step back and see if entering x_2 into the stepwise model somehow affected the significance of the x_1 predictor. That is, check the t -test P -value for testing $\beta_1 = 0$. If the t -test P -value for $\beta_1 = 0$ has become not significant — that is, the P -value is greater than $\alpha_R = 0.15$ — remove x_1 from the stepwise model.

Variable-Selection Procedures

- **Step #3.**
- Suppose both x_1 and x_2 made it into the two-predictor stepwise model and remained there.
- Now, fit each of the three-predictor models that include x_1 and x_2 as predictors — that is, regress y on x_1, x_2 , and x_3 , regress y on x_1, x_2 , and x_4 , ..., and regress y on x_1, x_2 , and x_p .
- Of those predictors whose t -test P -value is less than $\alpha_E = 0.15$, the third predictor put in the stepwise model is the predictor that has the smallest t -test P -value.
- If no predictor has a t -test P -value less than $\alpha_E = 0.15$, stop. The model containing the two predictors obtained from the second step is your final model.

Variable-Selection Procedures

- But, suppose instead that x_3 was deemed the "best" third predictor and it is therefore entered into the stepwise model.
- Now, since x_1 and x_2 were the first predictors in the model, step back and see if entering x_3 into the stepwise model somehow affected the significance of the x_1 and x_2 predictors. That is, check the t -test P -values for testing $\beta_1 = 0$ and $\beta_2 = 0$. If the t -test P -value for either $\beta_1 = 0$ or $\beta_2 = 0$ has become not significant — that is, the P -value is greater than $\alpha_R = 0.15$ — remove the predictor from the stepwise model.

Variable-Selection Procedures

- **Stopping the procedure** - Continue the steps as described above until adding an additional predictor does not yield a t -test P -value below $\alpha_E = 0.15$.

Variable-Selection Procedures

- **Cautions!**
- The final model is not guaranteed to be optimal in any specified sense.
- The procedure yields a single final model, although there are often several equally good models.
- Stepwise regression does not take into account a researcher's knowledge about the predictors. It may be necessary to force the procedure to include important predictors.
- One should not over-interpret the order in which predictors are entered into the model.

Variable-Selection Procedures

- **Cautions!**
- One should not jump to the conclusion that all the important predictor variables for predicting y have been identified, or that all the unimportant predictor variables have been eliminated.
- Many t -tests for testing $\beta_k = 0$ are conducted in a stepwise regression procedure. The probability is therefore high that we included some unimportant predictors or excluded some important predictors.
- It's for all of these reasons that one should be careful not to overuse or overstate the results of any stepwise regression procedure.

Variable-Selection Procedures

- Example: IQ vs brain size and body size
 - The first predictor entered into the stepwise model is **Brain**. The P -value for testing $\beta_{\text{Brain}} = 0$ is 0.019. The estimate S is 21.2, the R^2 -value is 14.27%, and the adjusted R^2 -value is 11.89.
 - The second and final predictor entered into the stepwise model is **Height**. The P -value for testing $\beta_{\text{Brain}} = 0$ is 0.001. The P -value for testing $\beta_{\text{Height}} = 0$ is 0.009. The estimate S is 19.5, the R^2 -value is 29.49%, and the adjusted R^2 -value is 25.46%.
 - At no step is a predictor removed from the stepwise model.

Variable-Selection Procedures

- Forward Selection

- This procedure is similar to stepwise-regression, but does not permit a variable to be deleted.

- This forward-selection procedure starts with no independent variables.

- It adds variables one at a time as long as a significant reduction in the error sum of squares (SSE) can be achieved.

Variable-Selection Procedures

- Backward Elimination

- This procedure begins with a model that includes all the independent variables the modeler wants considered.

- It then attempts to delete one variable at a time by determining whether the least significant variable currently in the model can be removed because its F value, F_{MIN} , is less than the user-specified or default F value, F_{REMOVE} .

- Once a variable has been removed from the model it cannot reenter at a subsequent step.

Variable-Selection Procedures

- Best-Subsets Regression
 - The three preceding procedures are one-variable-at-a-time methods offering no guarantee that the best model for a given number of variables will be found.
 - Some software packages include best-subsets regression that enables the user to find, given a specified number of independent variables, the “best” regression model.
 - Minitab output identifies the two best one-variable estimated regression equations, the two best two-variable equations, and so on.

Example: PGA Tour Data

The Professional Golfers Association keeps a variety of statistics regarding performance measures. Data include the average driving distance, percentage of drives that land in the fairway, percentage of greens hit in regulation, average number of putts, percentage of sand saves, and average score.

The variable names and definitions are shown on the next slide.

Example: PGA Tour Data

- Variable Names and Definitions

Drive: average length of a drive in yards

Fair: percentage of drives that land in the fairway

Green: percentage of greens hit in regulation (a par-3 green is “hit in regulation” if the player’s first shot lands on the green)

Putt: average number of putts for greens that have been hit in regulation

Sand: percentage of sand saves (landing in a sand trap and still scoring par or better)

Score: average score for an 18-hole round

Example: PGA Tour Data

- Sample Data

<u>Drive</u>	<u>Fair</u>	<u>Green</u>	<u>Putt</u>	<u>Sand</u>	<u>Score</u>
277.6	.681	.667	1.768	.550	69.10
259.6	.691	.665	1.810	.536	71.09
269.1	.657	.649	1.747	.472	70.12
267.0	.689	.673	1.763	.672	69.88
267.3	.581	.637	1.781	.521	70.71
255.6	.778	.674	1.791	.455	69.76
272.9	.615	.667	1.780	.476	70.19
265.4	.718	.699	1.790	.551	69.73

Example: PGA Tour Data

- Sample Data (continued)

<u>Drive</u>	<u>Fair</u>	<u>Green</u>	<u>Putt</u>	<u>Sand</u>	<u>Score</u>
272.6	.660	.672	1.803	.431	69.97
263.9	.668	.669	1.774	.493	70.33
267.0	.686	.687	1.809	.492	70.32
266.0	.681	.670	1.765	.599	70.09
258.1	.695	.641	1.784	.500	70.46
255.6	.792	.672	1.752	.603	69.49
261.3	.740	.702	1.813	.529	69.88
262.2	.721	.662	1.754	.576	70.27

Example: PGA Tour Data

- Sample Data (continued)

<u>Drive</u>	<u>Fair</u>	<u>Green</u>	<u>Putt</u>	<u>Sand</u>	<u>Score</u>
260.5	.703	.623	1.782	.567	70.72
271.3	.671	.666	1.783	.492	70.30
263.3	.714	.687	1.796	.468	69.91
276.6	.634	.643	1.776	.541	70.69
252.1	.726	.639	1.788	.493	70.59
263.0	.687	.675	1.786	.486	70.20
263.0	.639	.647	1.760	.374	70.81
253.5	.732	.693	1.797	.518	70.26
266.2	.681	.657	1.812	.472	70.96

Example: PGA Tour Data

- Sample Correlation Coefficients

	<u>Score</u>	<u>Drive</u>	<u>Fair</u>	<u>Green</u>	<u>Putt</u>
Drive	-.154				
Fair	-.427	-.679			
Green	-.556	-.045	.421		
Putt	.258	-.139	.101	.354	
Sand	-.278	-.024	.265	.083	-.296

Example: PGA Tour Data

- Best Subsets Regression of SCORE

Vars	R-sq	R-sq(a)	C-p	s	D	F	G	P	S
1	30.9	27.9	26.9	.39685			X		
1	18.2	14.6	35.7	.43183		X			
2	54.7	50.5	12.4	.32872	X	X			
2	54.6	50.5	12.5	.32891			X	X	
3	60.7	55.1	10.2	.31318	X	X		X	
3	59.1	53.3	11.4	.31957	X	X	X		
4	72.2	66.8	4.2	.26913	X	X	X	X	
4	60.9	53.1	12.1	.32011	X	X		X	X
5	72.6	65.4	6.0	.27499	X	X	X	X	X

Example: PGA Tour Data

- Minitab Output

The regression equation

$$\text{Score} = 74.678 - .0398(\text{Drive}) - 6.686(\text{Fair}) \\ - 10.342(\text{Green}) + 9.858(\text{Putt})$$

Predictor	Coef	Stdev	t-ratio	p
Constant	74.678	6.952	10.74	.000
Drive	-.0398	.01235	-3.22	.004
Fair	-6.686	1.939	-3.45	.003
Green	-10.342	3.561	-2.90	.009
Putt	9.858	3.180	3.10	.006

s = .2691

R-sq = 72.4%

R-sq(adj) = 66.8%

Example: PGA Tour Data

- Minitab Output

Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	4	3.79469	.94867	13.10	.000
Error	20	1.44865	.07243		
Total	24	5.24334			

Peruvian Blood Pressure Data

Variables possibly relating to blood pressures of Peruvians who have moved from rural high altitude areas to urban lower altitude areas.

- $n=39$
- Response (y): systolic blood pressure
- Potential predictor (x_1): age
- Potential predictor (x_2): years in urban area
- Potential predictor (x_3): fraction of life in urban area
- Potential predictor (x_4): weight (kg)
- Potential predictor (x_5): height (mm)
- Potential predictor (x_6): chin skinfold
- Potential predictor (x_7): forearm skinfold
- Potential predictor (x_8): calf skinfold
- Potential predictor (x_9): resting pulse rate

Peruvian Blood Pressure Data

Result

- p -values for the variables **Height**, **Chin**, **Forearm**, **Calf**, and **Pulse** are not at a statistically significant level.
- These individual tests are affected by correlations amongst the x -variables, so we will use the **Variable-Selection Procedures** to see whether it is reasonable to declare that all five non-significant variables can be dropped from the model.

Model Building

Here is a procedure for building a mathematical model:

① Identify the dependent variable; what is it we wish to predict? Don't forget the variable's unit of measure.

② List potential predictors; how would changes in predictors change the dependent variable? Be selective; go with the *fewest* independent variables required. Be aware of the effects of multicollinearity.

③ Gather the data; at *least* six observations for each independent variable used in the equation.

Model Building

- ④ **Identify several possible models;** formulate first- and second- order models with and without interaction. Draw scatter diagrams.
- ⑤ **Use statistical software to estimate the models.**
- ⑥ **Determine whether the required conditions are satisfied;** if not, attempt to correct the problem.
- ⑦ **Use your judgment and the statistical output to select the best model!**