دانشکده مهندسی صنایع
دکتری مهندسی صنایع

عنوان درس:
مدلسازی داده محور

بر اساس کتاب:

Business Analytics

Jeffrey D. Camm et al.

فصل هفتم:

Regression

Regression

- **After estimating the relationship between advertising expenditures and sales, a marketing manager might attempt to predict sales for a given level of advertising expenditures.**
- **A public utility might use the relationship between the daily high temperature and the demand for electricity to predict electricity usage on the basis of next month's anticipated daily high temperatures.**
- Sometimes managers will rely on intuition to judge how two variables are related.

Regression analysis:

- A statistical procedure to develop an equation showing how the variables are related if data can be obtained.

Dependent variable, or response: The variable being predicted.

Independent variables, or predictor variables: The variables being used to predict dependent variable

1

Regression

Simple Linear Regression Model:

- To develop better work schedules for Butler Trucking Company, the managers want to estimate the total daily travel times for their drivers.
- The managers believe that the total daily travel times (y) are closely related to the number of miles traveled in making the daily deliveries (x).

SIMPLE LINEAR REGRESSION MODEL

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

- β_0 and β_1 are population parameters that describe the y-intercept and slope of the line relating y and x.
- The error term ε accounts for the variability in y that cannot be explained by the linear relationship.
- Assumption: ε is a normally distributed variable with a mean of zero and constant variance for all observations.

ESTIMATED SIMPLE LINEAR REGRESSION EQUATION

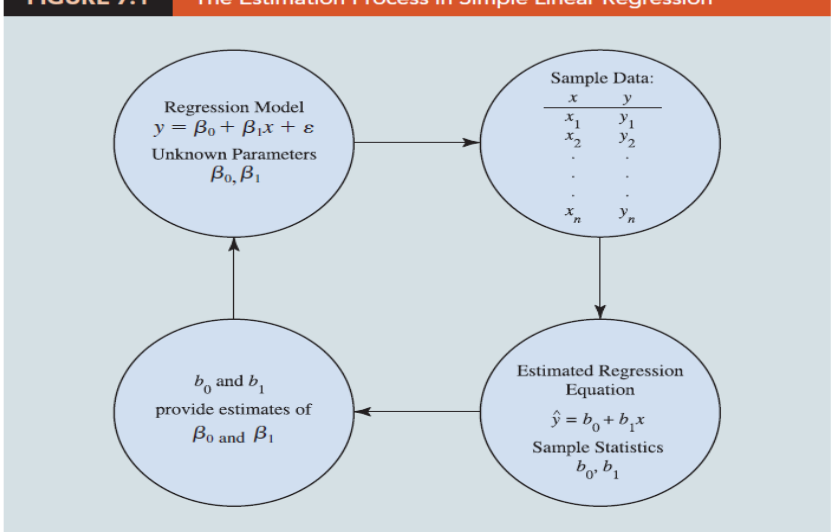
$$\hat{y} = b_0 + b_1 x$$

2

Regression

- In practice, β_0 and β_1 are not known and must be estimated using sample data.

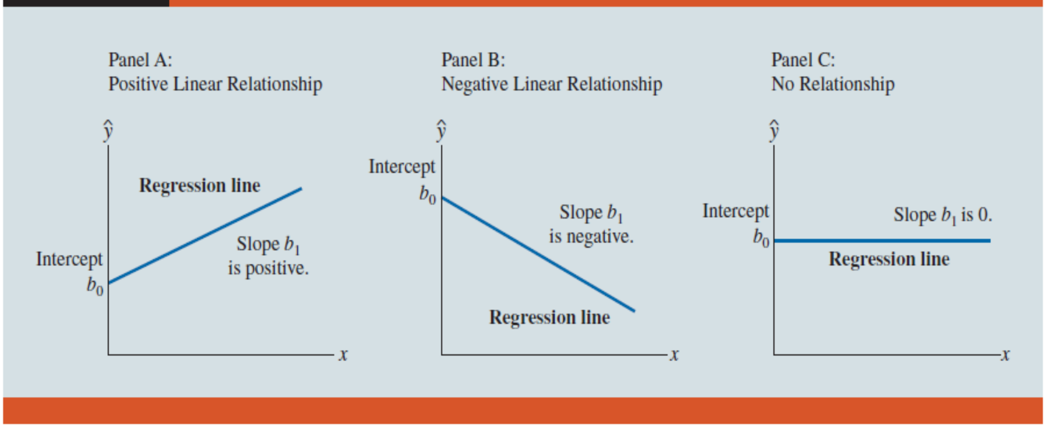
FIGURE 7.1 The Estimation Process in Simple Linear Regression



3

Regression

FIGURE 7.2 Possible Regression Lines in Simple Linear Regression



4

Regression

Least squares method:

A procedure for using sample data to find the estimated regression equation.

- It minimizes the sum of squares of the deviations between observed and predicted values of y.

$E(y|x)$= The mean value of y for a given value of x.

$\hat{y}$ =The point estimator of E(y|x).

- For the ith driving assignment in sample, x_i is the miles traveled and y_i is the travel time (in hours).

TABLE 7.1 Miles Traveled and Travel Time for 10 Butler Trucking Company Driving Assignments

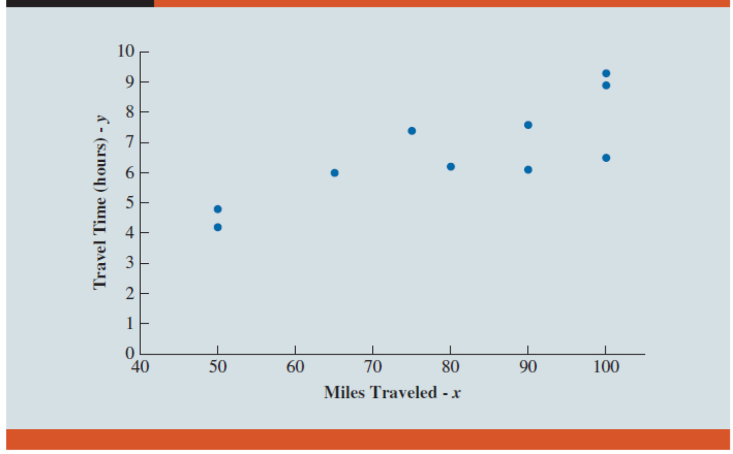
Driving Assignment *i*	x = Miles Traveled	y = Travel Time (hours)
1	100	9.3
2	50	4.8
3	50	8.9
4	100	6.5
5	50	4.2
6	80	6.2
7	75	7.4
8	65	6.0
9	90	7.6
10	90	6.1

5

Regression

- Longer travel times coincide with more miles traveled.
- The relationship appears to be approximated by a straight line;
- A positive linear relationship

FIGURE 7.3 Scatter Chart of Miles Traveled and Travel Time for Sample of 10 Butler Trucking Company Driving Assignments



6

Regression

- For ith driving assignment, the estimated regression equation provides:

$$\hat{y}_i = b_0 + b_1 x_i$$

where

$\hat{y}_i$ = predicted travel time (in hours) for the i^{th} driving assignment
b_0 = the y-intercept of the estimated regression line
b_1 = the slope of the estimated regression line
x_i = miles traveled for the i^{th} driving assignment

LEAST SQUARES EQUATION

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n} (y_1 - b_0 - b_1 x_1)^2$$

where

y_i = observed value of the dependent variable for the i^{th} observation
$\hat{y}_i$ = predicted value of the dependent variable for the i^{th} observation
n = total number of observations

7

Regression

Residual:

$$e_i = y_i - \hat{y}_i$$

$$\min \sum_{i=1}^{n} e_i^2$$

Least Squares Estimates of the Regression Parameters:

SLOPE EQUATION

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

y-INTERCEPT EQUATION

$$b_0 = \bar{y} - b_1\bar{x}$$

where

x_i = value of the independent variable for the i^{th} observation
y_i = value of the dependent variable for the i^{th} observation
$\bar{x}$ = mean value for the independent variable
$\bar{y}$ = mean value for the dependent variable
n = total number of observations

8

Regression

- The regression model is valid only over the experimental region, which is the range of values of the independent variables in the data used to estimate the model.

Extrapolation: Predicting the value of dependent variable outside the experimental region which is a risky task and should be avoided if possible.

- No empirical evidence that the relationship between y and x holds true outside the range of x values in the data used to estimate the relationship.
- For Butler Trucking, any prediction of the travel time for a driving distance less than 50 miles or greater than 100 miles is not a reliable estimate.

$$\hat{y}_i = 1.2739 + 0.0678(100) = 8.0539$$

$$e_1 = y_1 - \hat{y}_i = 9.3 - 8.0539 = 1.2461$$

9

Regression

TABLE 7.2 Predicted Travel Time and Residuals for 10 Butler Trucking Company Driving Assignments

Driving Assignment i	x = Miles Traveled	y = Travel Time (hours)	$\hat{y}_i = b_0 + b_1x_i$	$e_i = y_i - \hat{y}_i$	e_i^2
1	100	9.3	8.0565	1.2435	1.5463
2	50	4.8	4.6652	0.1348	0.0182
3	100	8.9	8.0565	0.8435	0.7115
4	100	6.5	8.0565	−1.5565	2.4227
5	50	4.2	4.6652	−0.4652	0.2164
6	80	6.2	6.7000	−0.5000	0.2500
7	75	7.4	6.3609	1.0391	1.0797
8	65	6.0	5.6826	0.3174	0.1007
9	90	7.6	7.3783	0.2217	0.0492
10	90	6.1	7.3783	−1.2783	1.6341
	Totals	67.0	67.0000	0.0000	8.0288

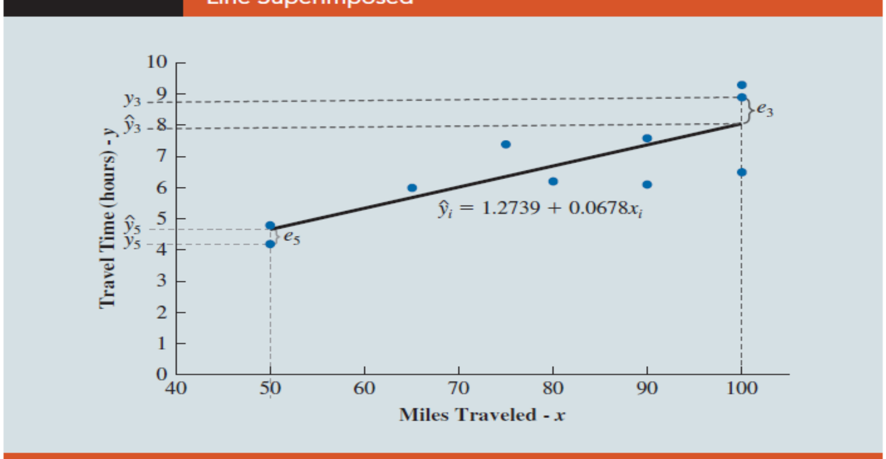
Three points are always true for a simple linear regression:

- The sum of predicted values $\hat{y}_i$ is equal to the sum of the values of the dependent variable y.
- The sum of the residuals e_i is 0.
- The sum of the squared residuals e_i^2 is minimized.

10

Regression

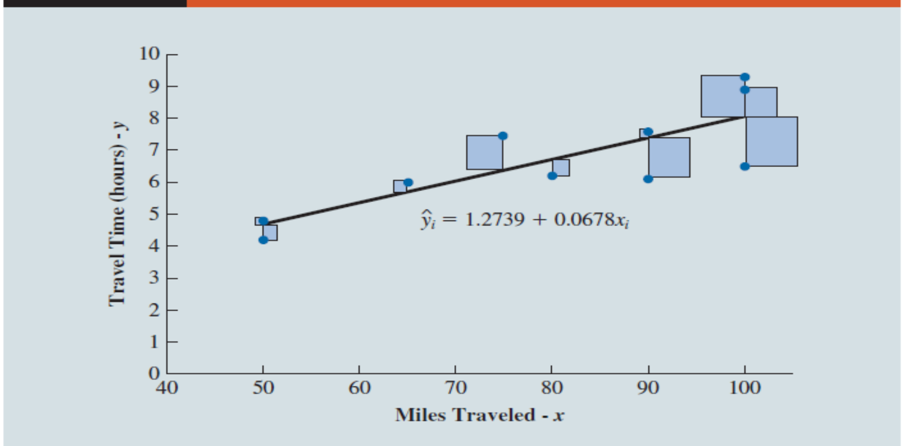
FIGURE 7.4 Scatter Chart of Miles Traveled and Travel Time for Butler Trucking Company Driving Assignments with Regression Line Superimposed



11

Regression

FIGURE 7.5 A Geometric Interpretation of the Least Squares Method



12

Regression

Assessing the Fit of the Simple Linear Regression Model:

SUM OF SQUARES DUE TO ERROR

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

SSE:

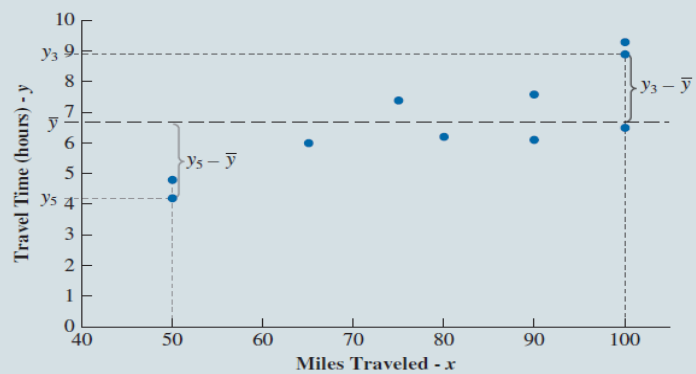
- A measure of the error as a result of using the estimated regression equation to predict the values of the dependent variable in the sample.

13

Regression

If we want to predict travel time without knowing the miles traveled, we use $\bar{y} = 6.7$ as a predictor of y.

FIGURE 7.7 The Sample Mean $\bar{y}$ as a Predictor of Travel Time for Butler Trucking Company



14

Regression

TABLE 7.3 Calculations for the Sum of Squares Total for the Butler Trucking Simple Linear Regression

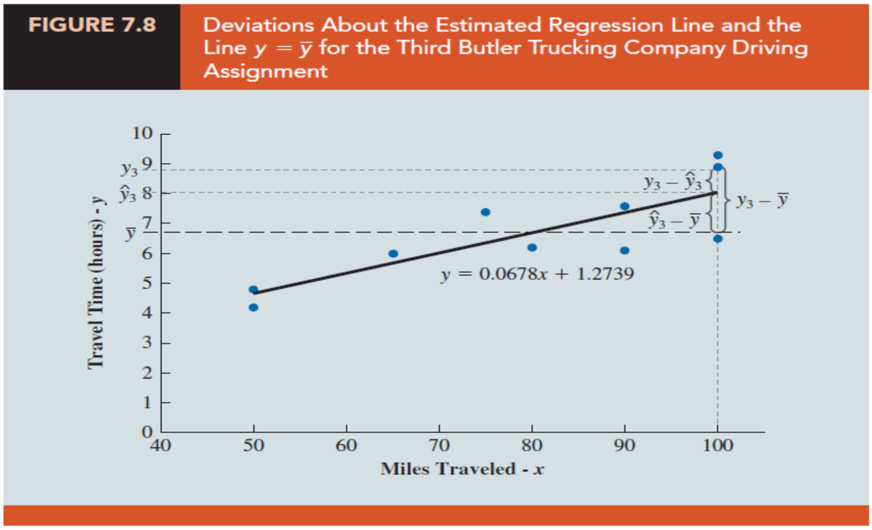
Driving Assignment i	x = Miles Traveled	y = Travel Time (hours)	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	100	9.3	2.6	6.76
2	50	4.8	−1.9	3.61
3	100	8.9	2.2	4.84
4	100	6.5	−0.2	0.04
5	50	4.2	−2.5	6.25
6	80	6.2	−0.5	0.25
7	75	7.4	0.7	0.49
8	65	6.0	−0.7	0.49
9	90	7.6	0.9	0.81
10	90	6.1	−0.6	0.36
	Totals	67.0	0	23.9

TOTAL SUM OF SQUARES, SST

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

15

Regression



16

Regression

SUM OF SQUARES DUE TO REGRESSION, SSR

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$SST = SSR + SSE$$

where

SST = total sum of squares
SSR = sum of squares due to regression
SSE = sum of squares due to error

COEFFICIENT OF DETERMINATION

$$r^2 = \frac{SSR}{SST}$$

17

Regression

For the Butler Trucking Company, the value of the coefficient of determination is

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{15.8712}{23.9} = 0.6641$$

r^2 is the percentage of SST that can be explained using the estimated regression equation.

66.41% of the variability in travel time can be explained by the linear relationship between.

18

Regression

The Multiple Regression Model:

MULTIPLE REGRESSION MODEL

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \varepsilon$$

- $\beta_0, \beta_1, \beta_2, \ldots, \beta_q$: The population parameters
- Error term e is a normally distributed variable with a mean of zero and a constant variance across all observations.
- β_j: The change in the mean value of y that corresponds to a one-unit increase in x_j, holding all other independent variables constant.

ESTIMATED MULTIPLE REGRESSION EQUATION

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_q x_q$$

where

$$b_0, b_1, b_2, \ldots, b_q = \text{the point estimates of } \beta_0, \beta_1, \beta_2, \ldots, \beta_q$$
$$\hat{y} = \text{estimated mean value of } y \text{ given values for } x_1, \ldots, x_q$$

19

Regression

Least Squares Method and Multiple Regression:

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^{n} (y_i - b_0 - b_1 x_1 - \cdots - b_q x_q)^2 = \min \sum_{i=1}^{n} e_i^2$$

Butler Trucking Company and Multiple Regression:

- With r^2=0.6641, 33.59% of the variability in sample travel times remains unexplained.
- Butler's managers felt that the number of deliveries made on a driving assignment also contributed to the total travel time.

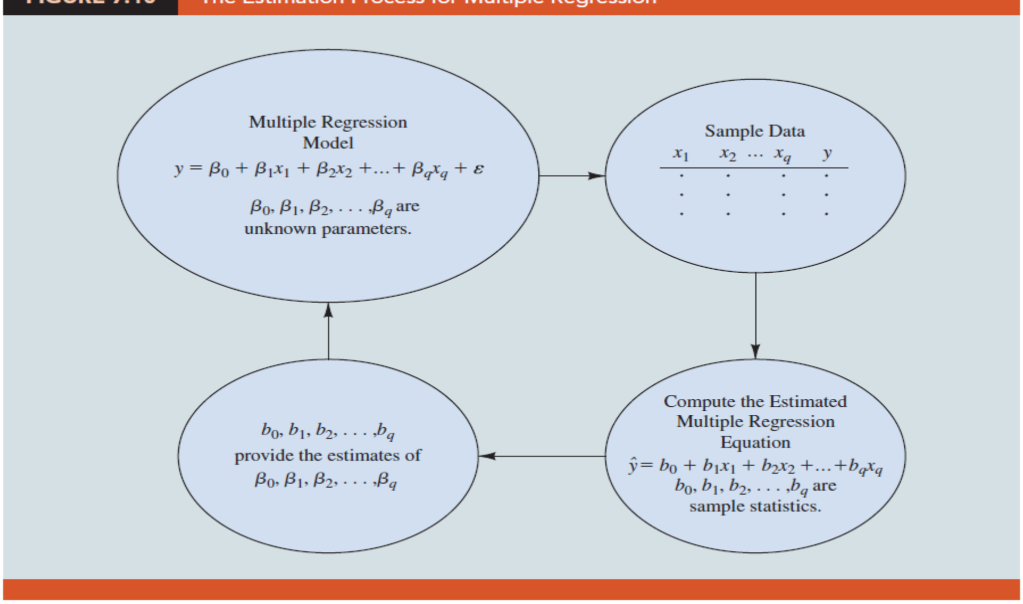
$$\hat{y} = 0.1273 = 0.0672x_1 + 0.6900x_2$$

- For a fixed number of deliveries, the mean travel time will increase by 0.0672 hours when the distance traveled increases by 1 mile.
- For a fixed distance traveled, the mean travel time will increase by 0.69 hours when the number of deliveries increases by 1 delivery.
- r^2=0.8173.

20

Regression

FIGURE 7.10 The Estimation Process for Multiple Regression



21

Regression

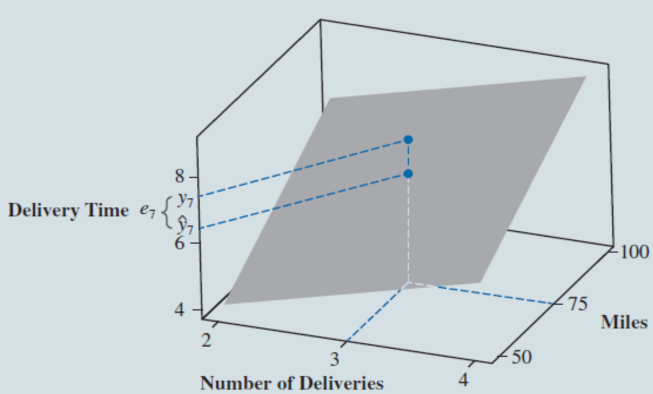
FIGURE 7.13 Excel Regression Output for the Butler Trucking Company with Miles and Deliveries as Independent Variables

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.90407397							
5	R Square	0.817349743							
6	Adjusted R Square	0.816119775							
7	Standard Error	0.829967216							
8	Observations	300							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	915.5160626	457.7580313	664.5292419	2.2419E-110			
13	Residual	297	204.5871374	0.68884558					
14	Total	299	1120.1032						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	0.127337137	0.20520348	0.620540826	0.53537766	−0.276499931	0.531174204	−0.404649592	0.659323866
18	Miles	0.067181742	0.002454979	27.36551071	3.5398E-83	0.062350385	0.072013099	0.06081725	0.073546235
19	Deliveries	0.68999828	0.029521057	23.37308852	2.84826E-69	0.631901326	0.748095234	0.613465414	0.766531147

22

Regression

FIGURE 7.14 Graph of the Regression Equation for Multiple Regression Analysis with Two Independent Variables



23

Regression

F test for testing the null hypothesis that multiple regression parameters are all equal to zero.

$$H_0:\ \beta_1 = \beta_2 = \cdots = \beta_q = 0$$
$$H_a:\ \exists j = 1,2,\ldots,q;\ \beta_j = 0$$

Test Statistics:

$$F = \frac{\frac{SSR}{q}}{\frac{SSE}{n-q-1}} \approx F_{q,n-q-1}$$

24

Regression

Inference and Regression:

- The statistics $b_0, b_1, b_2, \ldots, b_q$ as random variables are point estimators of $\beta_0, \beta_1, \beta_2, \ldots, \beta_q$.
- $\hat{y}$ is a point estimator of $E(y \mid x_1, x_2, \ldots, x_q)$, the conditional mean of y given values of x1, x2, . . . , xq.
- Different samples will result in different values of $b_0, b_1, b_2, \ldots, b_q$.
- If $b_0, b_1, b_2, \ldots, b_q$ change relatively little from sample to sample, they have low variability and more reliability.
- If $b_0, b_1, b_2, \ldots, b_q$ change dramatically from sample to sample, they have high variability and less reliability.
- How confident can we be $b_0, b_1, b_2, \ldots, b_q$ for the Butler Trucking multiple regression model?
- Do they have little variation and so are relatively reliable, or do they have so much variation that they have little meaning?

25

Regression

Statistical inference:

- The process of making estimates and drawing conclusions about one or more characteristics of a population through the analysis of sample data.
- The regression parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_q$.
- The mean value and/or predicted value of y for specific values of independent variables $x_1, x_2, \ldots, x_q$.

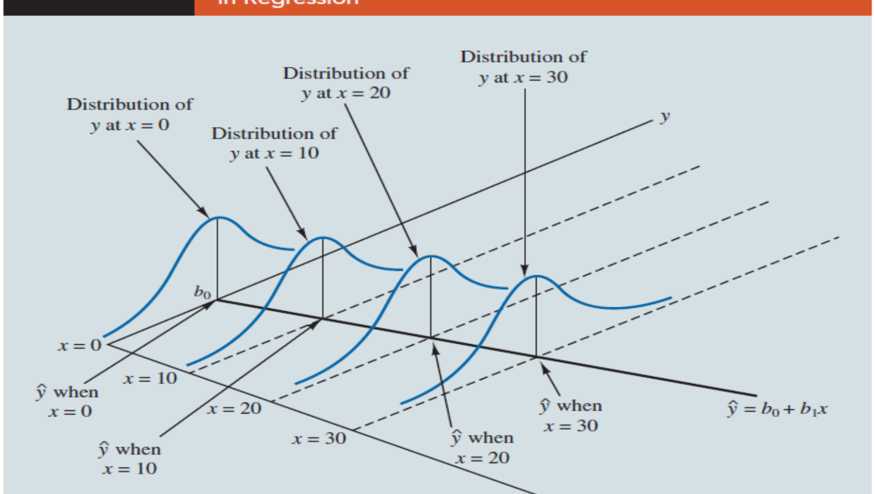
Conditions Necessary for Valid Inference in the Least Squares:

1. For any given combination of $x_1, x_2, \ldots, x_q$, the population of potential error terms ε is normally distributed with a mean of 0 and a constant variance. The regression estimates are unbiased
2. The values of ε are statistically independent.

- Simple scatter charts of the residuals versus the predicted values of y and the residuals versus x are an used to assess whether these conditions are violated.

Regression

FIGURE 7.15 Illustration of the Conditions for Valid Inference in Regression

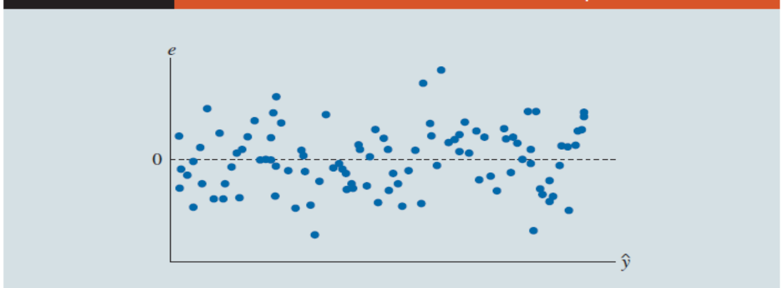


Note: The distribution of y has the same shape at each x value.

Regression

- The center of the residuals should be approximately zero
- The errors should be symmetrically distributed with values near zero occurring more frequently than values that differ greatly from zero.
- A pattern in the residuals such as this gives us little reason to doubt the validity of inferences made on the regression that generated the residuals.

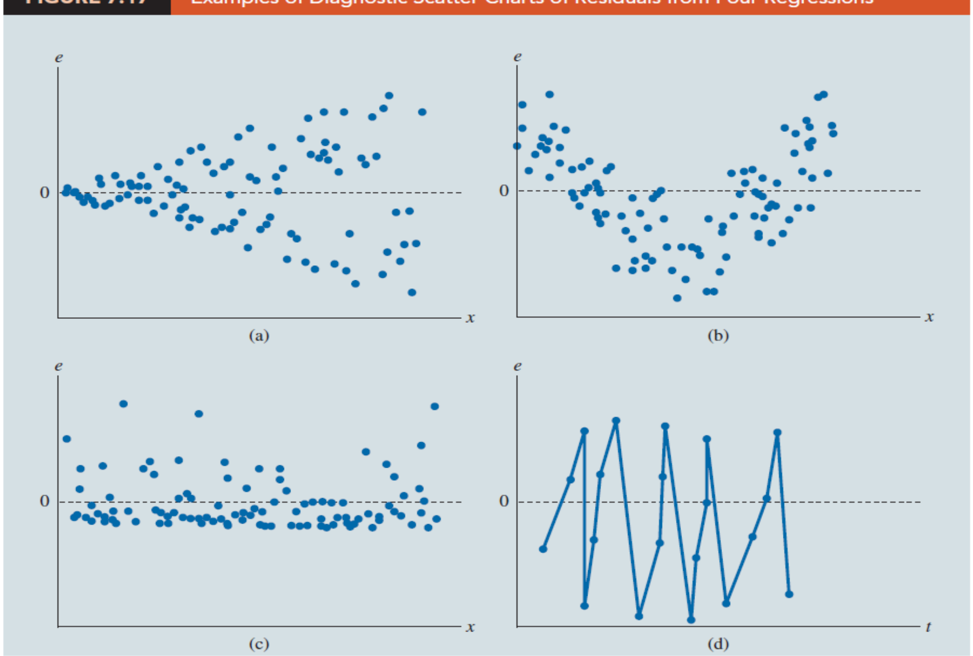
FIGURE 7.16 Example of a Random Error Pattern in a Scatter Chart of Residuals and Predicted Values of the Dependent Variable



28

FIGURE 7.17 Examples of Diagnostic Scatter Charts of Residuals from Four Regressions

Violation of at least one condition



29

Regression

(a) The variation in residuals increases as x increases, (residuals do not have a constant variance).

(b) The residuals are positive for small and large values of x but are negative for moderate values of x. The model underpredicts y for small and large values of x and overpredicts y for intermediate values of x. The regression model does not adequately capture the relationship between x and y.

(c) The residuals are not symmetrically distributed around 0. The residuals are not normally distributed.

(d) The residuals are plotted over time t as an independent variable. A distinct pattern across every set of four residuals. The residuals are not independent. Perhaps, we have collected quarterly data.

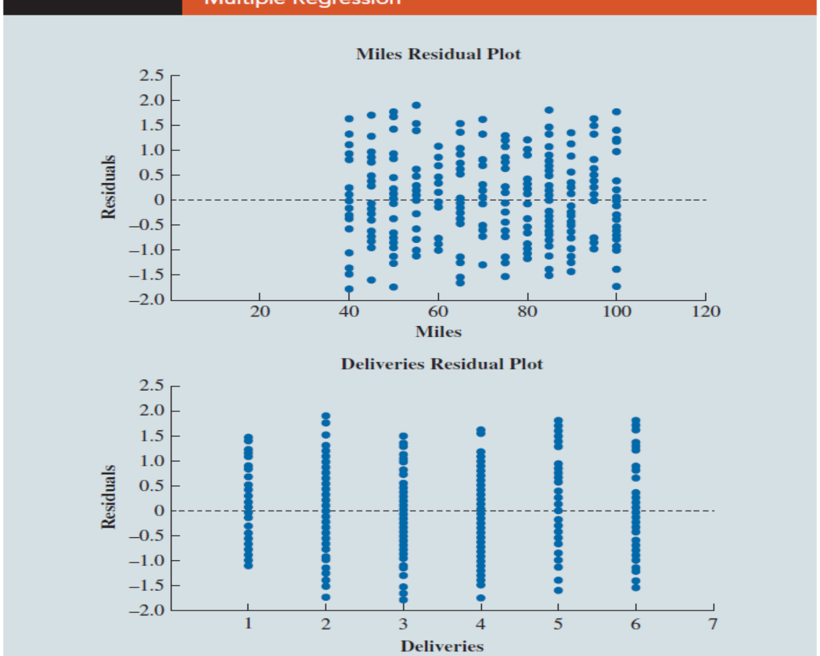
The residuals violate conditions either because:

(1) An important independent variable has been omitted or

(2) The functional form of the model is inadequate to explain the relationship.

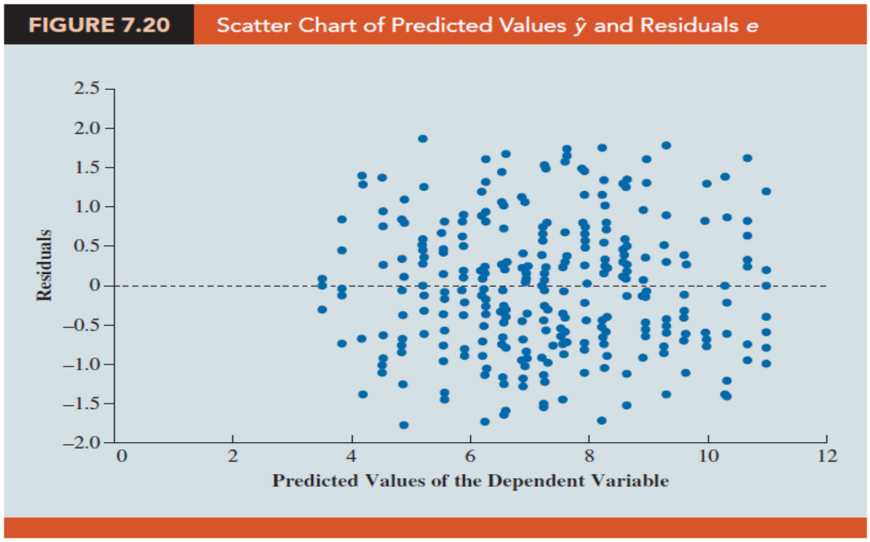
30

FIGURE 7.18 Excel Residual Plots for the Butler Trucking Company Multiple Regression



31

Regression



32

Regression

Testing Individual Regression Parameters:

- When regression model satisfies the necessary conditions, we can begin testing hypotheses and building confidence intervals.

$$H_0: \beta_j = 0$$
$$H_a: \beta_j \neq 0$$
$$t = \frac{b_j}{s_{b_j}}$$

- As the magnitude of t increases in any direction, we are more likely to reject the null hypothesis.
- Rejecting the null hypothesis means that a relationship exists between y and x_j.
- Smaller p-values indicate stronger evidence against the null hypothesis (i.e., stronger evidence of a relationship between x_j and y).
- The null hypothesis is rejected when p-value is smaller than predetermined level of significance (usually 0.05 or 0.01).

Confidence interval for a regression parameter: $b_j \pm t_{\alpha/2} s_{b_j}$

If the confidence interval does not contain zero, the null hypothesis is rejected at the level of significance.

33

Regression

Addressing non-significant independent variables:

- If we do not reject the null hypothesis, the question of how to handle the corresponding x is raised.
- Do we use the model with the non-significant x, or do we rerun without the non-significant x and use the new result?
- If practical experience dictates that the non-significant x has a relationship with y, the x should remain in the model.
- If the model sufficiently explains the y without the non-significant x, rerun the regression without the non-significant x.

Note:

- The estimates of the other regression coefficients and their p-values may change considerably when we remove the non-significant x.

34

Regression

- In Butler Trucking multiple regression model, the p-value for b_0 is 0.5354 (Not statistically significant).
- Should we remove the y-intercept?
- This will force the y-intercept to go through the origin. However, this can substantially alter the estimated slopes and result in a less effective and less accurate regression.
- Regression through the origin should not be forced.
- If there are strong a priori reasons for that, collect data for which the values of x are at or near zero to empirically validate this belief and avoid extrapolation.
- If data is not obtainable, then forcing the y-intercept to be zero may be a necessary action, although it results in extrapolation.

35

Regression

Multi-collinearity:

- The correlation among the independent variables
- Most independent variables in a multiple regression problem are correlated with one another to some degree.
- In Butler Trucking, we could compute the sample correlation coefficient $r(x_1, x_2)$ to determine the extent to which these two variables are related.
- $r(x_1, x_2)=0.16$. Some degree of linear association between x_1 and x_2.
- Let x2 denote the number of gallons of gasoline consumed.
- x1 (the miles traveled) and x2 are now related. Logically, x1 and x2 are highly correlated and multi-collinearity is present in the model.

36

Regression

FIGURE 7.21 Excel Regression Output for the Butler Trucking Company with Miles and Gasoline Consumption as Independent Variables

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.69406354							
5	R Square	0.481724198							
6	Adjusted R Square	0.478234125							
7	Standard Error	1.398077545							
8	Observations	300							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	539.5808158	269.7904079	138.0269794	4.09542E-43			
13	Residual	297	580.5223842	1.954620822					
14	Total	299	1120.1032						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	2.493095385	0.33669895	7.404523781	1.36703E-12	1.830477398	3.155713373	1.620208758	3.365982013
18	Miles	0.074701825	0.014274552	5.233216928	3.15444E-07	0.046609743	0.102793908	0.037695279	0.111708371
19	Gasoline Consumption	–0.067506102	0.152707928	–0.442060235	0.658767336	–0.368032789	0.233020584	–0.463398955	0.328386751

37

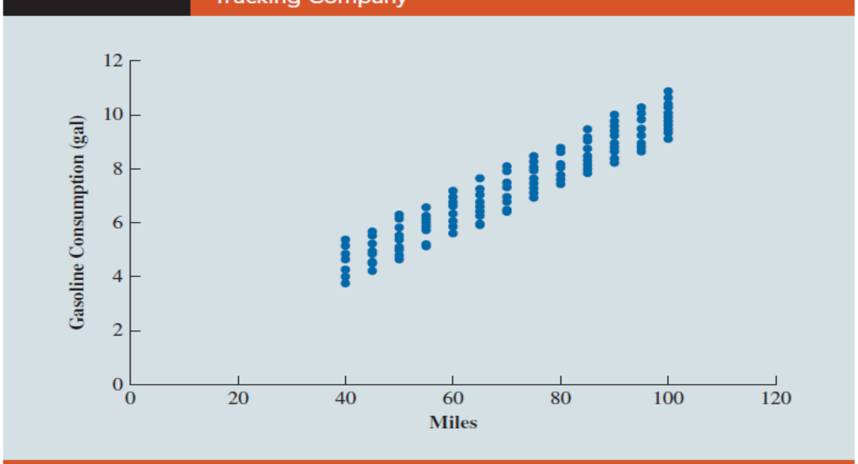
Regression

- When we conduct a t test to determine whether β1 is equal to zero, p-value is 3.1544E-07 which means travel time is related to miles traveled.
- When we conduct a t test to determine whether β2 is equal to zero, p-value is 0.6588.
- Does this mean that travel time is not related to gasoline consumption? Not necessarily.
- It probably means that with x_1 in the model, x_2 does not make a significant marginal contribution to predicting y.
- If we know the miles traveled, we do not gain much new useful information in predicting driving time by also knowing the amount of gasoline consumed.

38

Regression

FIGURE 7.22 Scatter Chart of Miles and Gasoline Consumed for Butler Trucking Company



39

Regression

The difficulty caused by multi-collinearity:

- A parameter associated with a multi-collinear independent variable is not significantly different from zero when the independent variable actually has a strong relationship with the dependent variable.
- This problem is avoided when there is little correlation among the independent variables.

Common rule-of-thumb test:

- Multi-collinearity is a potential problem if the absolute value of the sample correlation coefficient exceeds 0.7 for any two of the independent variables.
- Multi-collinearity increases the standard deviation of $b_0, b_1, \ldots, b_q$ and $\hat{y}$
- Inference based on these estimates is less precise than it should be.
- Confidence intervals for $b_0, b_1, \ldots, b_q$ and $\hat{y}$ are wider than they should be.
- We are less likely to reject the null hypothesis. x_j is not related to y while they in fact are related.
- If the primary objective is inference, avoid including highly correlated independent variables.
- If the primary objective is prediction, then multi-collinearity is not a concern.

Regression

Categorical Independent Variables:

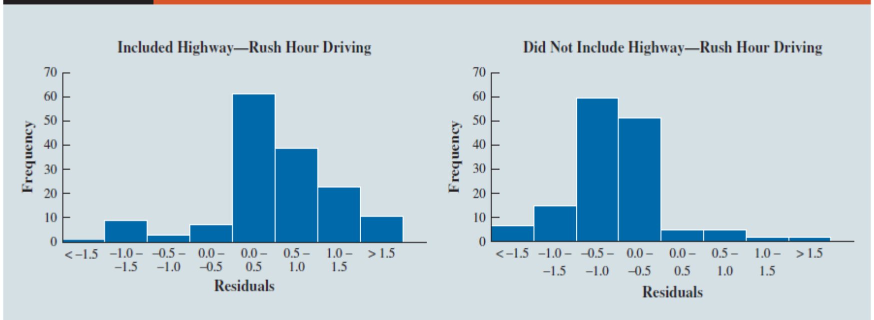
- Sometimes, we must work with categorical independent variables such as marital status (married, single), method of payment (cash, credit card, check), ...
- Butler driving assignments require the driver to travel on a congested segment of a highway during the afternoon rush hour.
- This factor may contribute substantially to variability in the travel times across driving assignments.
- How do we incorporate into a regression model information on which driving assignments include travel on a congested segment of a highway during the afternoon rush hour?
- **Dummy variable:** If an assignment includes in the model the travel on the congested segment of a highway during the rush hours

$$x_3 = \begin{cases} 0 \text{ if an assignment did not include travel on the congested segment of highway during afternoon rush hour} \\ 1 \text{ if an assignment included travel on the congested segment of highway during afternoon rush hour} \end{cases}$$

Regression

FIGURE 7.23 Histograms of the Residuals for Driving Assignments That Included Travel on a Congested Segment of a Highway During the Afternoon Rush Hour and Residuals for Driving Assignments That Did Not

Positive residuals- Under-predicting



Negative residu Under-predictir

The dummy variable could potentially explain a substantial proportion of the variance in travel time that is unexplained by the current model

We add x_3 to the current Butler Trucking multiple regression model. $\hat{y} = -0.3302 + 0.0672x_1 + 0.6735x_2 + 0.9980x_3$

42

FIGURE 7.24 Excel Data and Output for Butler Trucking with Miles Traveled (x_1), Number of Deliveries (x_2), and the Highway Rush Hour Dummy Variable (x_3) as the Independent Variables

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.940107228							
5	R Square	0.8838016							
6	Adjusted R Square	0.882623914							
7	Standard Error	0.663106426							
8	Observations	300							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	3	989.9490008	329.9830003	750.455757	5.7766E–138			
13	Residual	296	130.1541992	0.439710132					
14	Total	299	1120.1032						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	–0.330229304	0.167677925	–1.969426232	0.04983651	–0.66022126	–0.000237349	–0.764941128	0.104482519
18	Miles	0.067220302	0.00196142	34.27125147	4.7852E-105	0.063360208	0.071080397	0.062135243	0.072305362
19	Deliveries	0.67351584	0.023619993	28.51465081	6.74797E-87	0.627031441	0.720000239	0.612280051	0.734751629
20	Highway	0.9980033	0.076706582	13.0106605	6.49817E-31	0.847043924	1.148962677	0.799138374	1.196868226

43

Regression

- $r^2 = 0.8838$: The regression model explains approximately 88.4% of the variability in travel time for the driving assignments in the sample.
- Using a dummy variable provides two estimated regression equations to predict the travel time.
- 1. One that corresponds to driving assignments that include travel on the congested segment of highway during the afternoon rush hour period
- 2. One that corresponds to driving assignments that do not include such travel.

$$\hat{y} = -0.3302 + 0.0672x_1 + 0.6735x_2 + 0.9980(0)$$
$$= -0.3302 + 0.0672x_1 + 0.6735x_2$$

In the case that when $x_3 = 1$, we have

$$\hat{y} = -0.3302 + 0.0672x_1 + 0.6735x_2 + 0.9980(1)$$
$$= 0.6678 + 0.0672x_1 + 0.6735x_2$$

44

Regression

More Complex Categorical Variables:

- If a categorical variable has k levels, k – 1 dummy variables are required, each dummy variable is 0 or 1.
- A manufacturer of vending machines organized the sales territories into three regions: A, B, and C.
- The managers want to use regression to predict the number of vending machines sold per week.
- Several independent variables (the number of sales personnel, advertising expenditures, etc.).
- Sales region is also an important factor in predicting the number of units sold.

Region	x_1	x_2
A	0	0
B	1	0
C	0	1

$$x_1 = \begin{cases} 1 \text{ if sales Region B} \\ 0 \text{ otherwise} \end{cases} \quad x_2 = \begin{cases} 1 \text{ if sales Region C} \\ 0 \text{ otherwise} \end{cases}$$

45

Regression

The regression equation relating the estimated mean number of units sold to the dummy variables is written as

$$\hat{y} = b_0 + b_1x_1 + b_2x_2$$

Observations corresponding to Region A correspond to $x_1 = 0$, $x_2 = 0$, so the estimated mean number of units sold in Region A is

$$\hat{y} = b_0 + b_1(0) + b_2(0) = b_0$$

Observations corresponding to Region B are coded $x_1 = 1$, $x_2 = 0$, so the estimated mean number of units sold in Region B is

$$\hat{y} = b_0 + b_1(1) + b_2(0) = b_0 + b_1$$

Observations corresponding to Region C are coded $x_1 = 0$, $x_2 = 1$, so the estimated mean number of units sold in Region C is

$$\hat{y} = b_0 + b_1(0) + b_2(1) = b_0 + b_2$$

b0: the estimated mean sales for Region A,

b1: the estimated difference between Region B and Region A,

b2: the estimated difference between Region C and Region A.

Regression

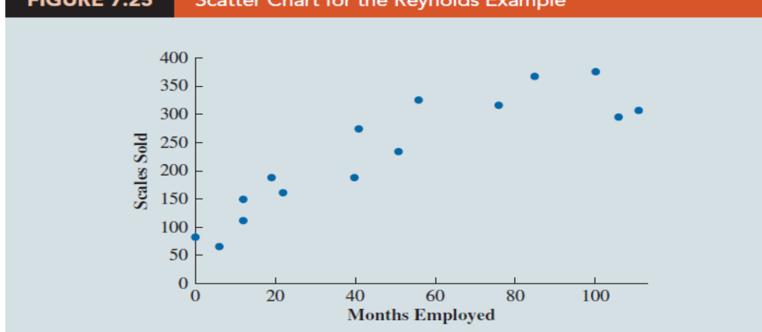
Modeling Nonlinear Relationships:

Reynolds, Inc., a manufacturer of industrial scales and laboratory equipment.

- Managers want to investigate the relationship between length of employment of their salespeople and the number of electronic laboratory scales sold.

Sales = 113.7453 + 2.3675 Months Employed

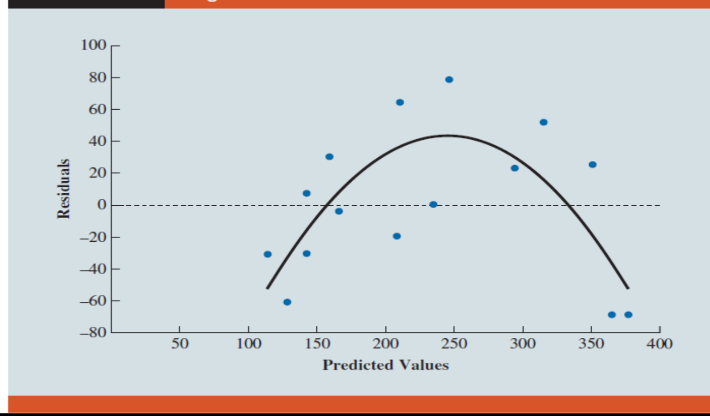
FIGURE 7.25 Scatter Chart for the Reynolds Example



Regression

- The relationship is significant (p-value = 9.3954E-06 for the t test that $\beta_1 = 0$) and a linear relationship explains a high percentage of the variability in sales $r^2 = 0.7901$.
- A pattern in the scatter chart of residuals against the predicted values of y that a curvilinear relationship may be a better fit.

FIGURE 7.27 Scatter Chart of the Residuals and Predicted Values of the Dependent Variable for the Reynolds Simple Linear Regression

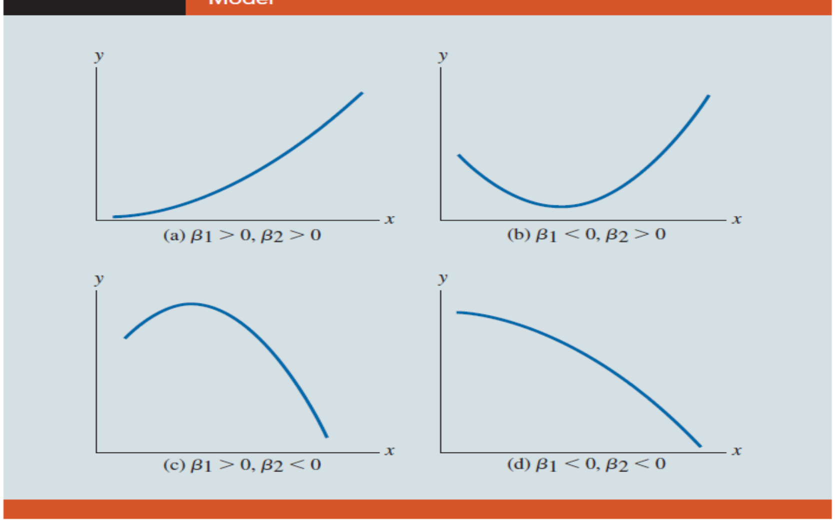


48

Regression

Quadratic Regression Models: $\hat{y} = b_0 + b_1x_1 + b_2x_1^2$

FIGURE 7.28 Relationships That Can Be Fit with a Quadratic Regression Model



49

Regression

Sales = 61.4299 + 5.8198 Months Employed − 0.0310 MonthsSq

FIGURE 7.29 Excel Data for the Reynolds Quadratic Regression Model

	A	B	C
1	Months Employed	MonthsSq	Scales Sold
2	41	1,681	275
3	106	11,236	296
4	76	5,776	317
5	100	10,000	376
6	22	484	162
7	12	144	150
8	85	7,225	367
9	111	12,321	308
10	40	1,600	189
11	51	2,601	235
12	0	0	83
13	12	144	112
14	6	36	67
15	56	3,136	325
16	19	361	189

50

Regression

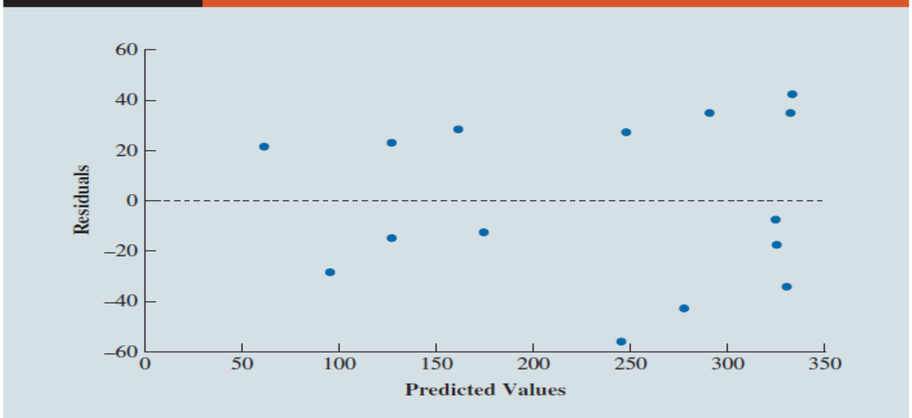
FIGURE 7.30 Excel Output for the Reynolds Quadratic Regression Model

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.949361402							
5	R Square	0.901287072							
6	Adjusted R Square	0.884834917							
7	Standard Error	34.61481184							
8	Observations	15							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	131278.711	65639.35548	54.78231208	9.25218E-07			
13	Residual	12	14378.22238	1198.185199					
14	Total	14	145656.9333						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	61.42993467	20.57433536	2.985755485	0.011363561	16.60230882	106.2575605	−1.415187222	124.2750566
18	Months Employed	5.819796648	0.969766536	6.001234761	6.20497E-05	3.706856877	7.93273642	2.857606371	8.781986926
19	MonthsSq	−0.031009589	0.008436087	−3.675826286	0.003172962	−0.049390243	−0.012628935	−0.05677795	−0.005241228

51

Regression

FIGURE 7.31 Scatter Chart of the Residuals and Predicted Values of the Dependent Variable for the Reynolds Quadratic Regression Model



52

Regression

Piecewise Linear Regression Models:

- As an alternative to a quadratic regression model
- Below some values of Months Employed, the relationship between Months Employed and Sales appears to be positive and linear
- The relationship between Months Employed and Sales appears to be negative and linear for the remaining observations.

Knot, or breakpoint:

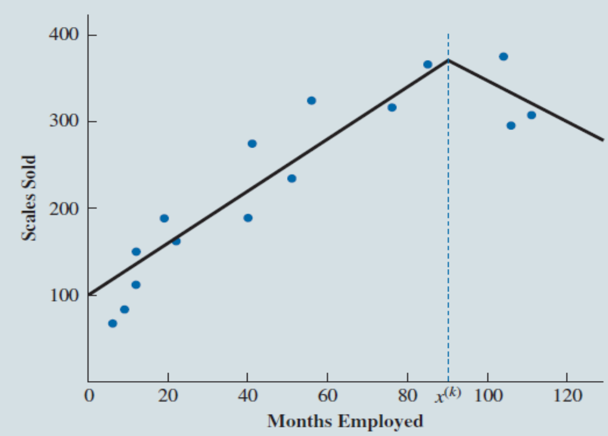
- The value of the independent variable Months Employed at which the relationship between Months Employed and Sales changes.

53

Regression

The knot: approximately 90 months.

FIGURE 7.32 Possible Position of Knot $x^{(k)}$



54

Regression

- A dummy variable that is zero for any observation for which the value of Months Employed is less than or equal to knot:

$$x_k = \begin{cases} 0 \text{ if } x_1 \le x^{(k)} \\ 1 \text{ if } x_1 > x^{(k)} \end{cases}$$

x_1 = Months
$x^{(k)}$ = the value of the knot (90 months for the Reynolds example)
x_k = the knot dummy variable

$$\hat{y} = b_0 + b_1 x_1 + b_2(x_1 - x^{(k)})x_k$$

$$\hat{y} = 87.2172 + 3.4094x_1 - 7.8726(x_1 - 90)x_k$$

- P-value corresponding to the t statistic for knot term (p-value=0.0014) is less than 0.05
- Adding the knot to the model with Months Employed as the independent variable is significant.
- A salesperson's sales are expected to increase by 3.4094 electronic laboratory scales for each month of employment until 90 months.
- The salesperson's sales are expected to decrease by 4.4632 electronic laboratory scales for each additional month of employment.

55

Regression

Interaction Between Independent Variables:

Often the relationship between the dependent variable and one independent variable is different at various values of a second independent variable.

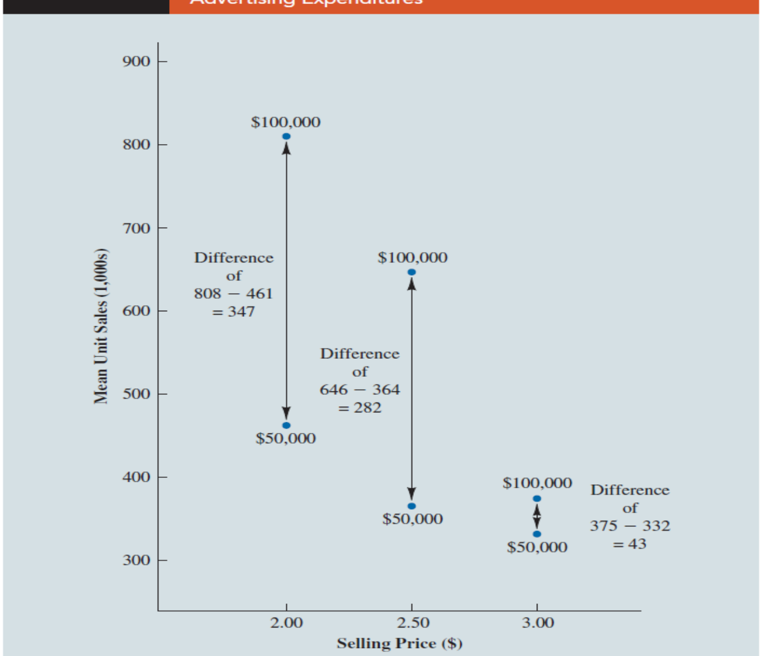
$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2$$

Tyler Personal Care

- Two factors believed to have the most influence on sales are unit selling price and advertising expenditure.
- To investigate the effects of these two variables on sales, prices of $2.00, $2.50, and $3.00 were paired with advertising expenditures of $50,000 and $100,000 in 24 test markets.

56

FIGURE 7.34 Mean Unit Sales (1,000s) as a Function of Selling Price and Advertising Expenditures



57

Regression

- The difference between mean sales for advertising expenditures of $50,000 and mean sales for advertising expenditures of $100,000 depends on the price of the product.
- At higher selling prices, the effect of increased advertising expenditure diminishes.
- Evidence of interaction between the price and advertising expenditure.

$$y = \text{Unit Sales (1000s)}$$
$$x_1 = \text{Price (\$)}$$
$$x_2 = \text{Adverstising Expenditure (\$1000s)}$$

$$\text{Sales} = -275.8333 + 175\,\text{Price} + 19.68\,\text{Advertising} - 6.08\,\text{Price* Advertising}$$

- p-value to the t test for Price*Advertising is 8.6772E-10 meaning interaction is significant.
- The relationship between advertising expenditure and sales depends on the price.
- The relationship between price and sales depends on advertising expenditure.

58

Regression

- How can price have a positive estimated regression coefficient?
- With the exception of luxury goods, we expect sales to decrease as price increases.
- This model can make sense if we work through the interpretation of the interaction.
- The relationship between Price and Sales is different at various values of Advertising Expenditure.
- The relationship between Advertising Expenditure and Sales is different at various values of Price.

$$\text{Sales After \$1 Price Increase} = -275.8333 + 175\,(\text{Price} + 1) + 19.68\,\text{Advertising} - 6.08\,(\text{Price} + 1) * \text{Advertising}$$

$$\text{Sales After \$1 Price Increase} - \text{Sales Before \$1 Price Increase} = 175 - 6.08 * \text{Advertising Expenditure}$$

- The change in the predicted value of sales when Price increases by $1 depends on advertising expenditure.

59

Regression

FIGURE 7.35 Excel Output for the Tyler Personal Care Linear Regression Model with Interaction

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.988993815							
5	R Square	0.978108766							
6	Adjusted R Square	0.974825081							
7	Standard Error	28.17386496							
8	Observations	24							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	3	709316	236438.6667	297.8692	9.25881E-17			
13	Residual	20	15875	793.7666667					
14	Total	23	5191.3333						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	–275.8333333	112.8421033	–2.444418575	0.023898351	–511.2178361	–40.44883053	–596.9074508	45.24078413
18	Price	175	44.54679188	3.928453489	0.0008316	82.07702045	267.9229796	48.24924412	301.7507559
19	Advertising Expenditure ($1,000s)	19.68	1.42735225	13.78776683	1.1263E-11	16.70259538	22.65740462	15.61869796	23.74130204
20	Price*Advertising	–6.08	0.563477299	–10.79014187	8.67721E-10	–7.255393049	–4.904606951	–7.683284335	–4.476715665

Regression

- If Advertising Expenditures is $50,000 when price is $2.00, we estimate sales:

$$\text{Sales} = -275.8333 + 175(2) + 19.68(50) - 6.08(2)(50) = 450.1667, \text{or } 450{,}167 \text{ units}$$

- At the same level of Advertising Expenditures ($50,000) when price is $3.00, we estimate sales:

$$\text{Sales} = -275.8333 + 175(3) + 19.68(50) - 6.08(3)(50) = 321.1667, \text{or } 321{,}167 \text{ units}$$

- When Advertising Expenditures is $50,000, a change in price from $2.00 to $3.00 results in a 450,167 - 321,167 =129,000 unit decrease in estimated sales.
- If Advertising Expenditures is $100,000 when price is $2.00, we estimate sales:

$$\text{Sales} = -275.8333 + 175(2) + 19.68(100) - 6.08(2)(100) = 826.1667, \text{or } 826{,}167 \text{ units}$$

- At the same level of Advertising Expenditures ($100,000) when price is $3.00, we estimate sales:

$$\text{Sales} = -275.8333 + 175(3) + 19.68(100) - 6.08(3)(100) = 393.1667, \text{or } 393{,}167 \text{ units}$$

- When Advertising Expenditures is $100,000, a change in price from $2.00 to $3.00 results in a 826,167 - 393,167 = 433,000 unit decrease in estimated sales.
- When Tyler spends more on advertising, its sales are more sensitive to changes in price.

Regression

- The relationship between Advertising Expenditure and Sales is different at various values of Price:

$$\text{Sales After \$1K Advertising Increase} = -275.8333 + 175\ \text{Price} + 19.68\ (\text{Advertising} + 1) - 6.08\ \text{Price} * (\text{Advertising} + 1)$$

$$\text{Sales After \$1K Advertising Increase} - \text{Sales Before \$1K Advertising Increase} = 19.68 - 6.08\ \text{Price}$$

- The change in the predicted value of the dependent variable that occurs when Advertising Expenditure increases by $1,000 depends on the price.
- If Price is $2.00 when Advertising Expenditure is $50,000, we estimate sales:

$$\text{Sales} = -275.8333 + 175(2) + 19.68(50) - 6.08(2)(50) = 450.1667, \text{or } 450{,}167 \text{ units}$$

- At the same level of Price ($2.00) when Advertising Expenditure is $100,000, we estimate sales:

$$\text{Sales} = -275.8333 + 175(2) + 19.68(100) - 6.08(2)(100) = 826.1667, \text{or } 826{,}167 \text{ units}$$

- When Price is $2.00, a change in Advertising Expenditures from $50,000 to $100,000 results in a 826,167 - 450,167 = 376,000 unit increase in estimated sales.

62

Regression

- If Price is $3.00 when Advertising Expenditure is 50,000, we estimate sales:

$$\text{Sales} = -275.8333 + 175(3) + 19.68(50) - 6.08(3)(50) = 321.1667, \text{or } 321{,}167 \text{ units}$$

- At the same level of Price ($3.00) when Advertising Expenditure is $100,000, we estimate sales:

$$\text{Sales} = -275.8333 + 175(3) + 19.68(100) - 6.08(3)(100) = 393.1667, \text{or } 393{,}167 \text{ units}$$

- When Price is $3.00, a change in Advertising Expenditure from $50,000 to $100,000 results in a 393.167 - 321,167 = 72,000 unit increase in estimated sales.
- When the price of Tyler's product is high, its sales are less sensitive to changes in advertising expenditure.

63

Regression

- For the Butler Trucking, suppose the relationship between miles traveled and travel time differs for driving assignments that included travel on a congested segment of a highway and those did not.
- We could create a new variable for the interaction between miles traveled and the dummy variable ($x_4 = x_1 x_3$) and estimate the following model:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_4$$

If a driving assignment does not include travel on a congested segment of a highway, $x_4 = x_1 * x_3 = x_1 * 0 = 0$ and the regression model is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

If a driving assignment does include travel on a congested segment of a highway, $x_4 = x_1 * x_3 = x_1 * 1 = x_1$ and the regression model is

$$\begin{aligned}\hat{y} &= b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1(1) \\ &= b_0 + (b_1 + b_3) x_1 + b_2 x_2\end{aligned}$$

- We can combine a quadratic effect with interaction to produce a second-order polynomial model with interaction between the two independent variables.

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 x_2^2 + b_5 x_1 x_2$$

64

Regression

Model Fitting:

Variable Selection Procedures:

- When there are many independent variables to consider, special procedures are sometimes employed to select the independent variables to include in the regression model.

Backward elimination:

- Begin with all of the independent variables.
- At each step, backward elimination considers the removal of an independent variable.
- A criterion is to remove least significant independent variable among those currently are not significant at a specified level of significance.
- Refit the regression model with the remaining independent variables.
- The procedure stops when all independent variables in the model are significant at a specified level.

65

Regression

Forward selection:

- Begin with none of the independent variables.
- At each step, forward selection considers the addition of an independent variable.
- A criterion is to add any the most significant independent variable currently significant at a specified level.
- Refit the regression model with the additional independent variable.
- The procedure stops when all the independent variables not in the model would not be significant at a specified level of significance if included in the model.

66

Regression

Stepwise selection:

- Begin with none of the independent variables.
- A criterion for independent variables to enter the model and a criterion for independent variables to remain in the model.
- A criterion adds the most significant variable and removes the least significant variable at each iteration.
- First, the most significant independent variable is added to the empty model if its level of significance satisfies the entering threshold.
- Each subsequent step involves two intermediate steps: (1) The most significant independent variable not in the model is added if its significance satisfies the threshold. (2) The least significant independent variable in the model is removed if its level of significance fails to satisfy the threshold.
- The procedure stops when no independent variable not currently in the model is significant and all independent variables currently in the model are significant.

67

Regression

Best subsets procedure:

- Simple linear regressions for each of the independent variables are generated
- The multiple regressions with all combinations of two or more independent variables are generated.
- Once a regression model has been generated for every possible subset of the independent variables, the entire collection of regression models are compared and evaluated by the analyst.

Use your own judgment and intuition about your data to refine the results of these algorithms.

Regression

Overfitting:

- If we attempt to fit a model too closely to the sample data, it does not accurately reflect the population, the model is said to have been overfit.
- The use of complex functional forms or independent variables that do not have meaningful relationships with the dependent variable.
- An overfit model can be misleading with regard to its predictive capability and its interpretation.
- Overfitting is difficult to detect and avoid, but to mitigate this problem:

(1) Use only independent variables that you expect to have real and meaningful relationships with the dependent variable.

(2) Use complex models, such as quadratic and piecewise linear, only when you have a reason.

(3) Do not let software dictate your model.

(4) Use iterative procedures, such as stepwise and best-subsets, for guidance not to generate final model.

(5) Use your own judgment and intuition about your data and to refine your model.

Regression

Cross-validation:

- If you have access to sufficient data, assess your model on data other than the sample data.

Holdout method:

- The sample data are randomly divided into mutually exclusive training and validation sets.

Training set:

- The data set used to build the candidate models that appear to make practical sense.

Validation set:

- The set of data used to compare model performances and ultimately select a model for predicting values of the dependent variable.

70

Regression

- We might randomly select half of data for use in developing regression models.
- Then we use the remaining half of data as a validation set to assess and compare the models' performances and ultimately select the model that minimizes some measure of overall error.
- Results of a holdout sample can vary greatly depending on which observations are randomly selected for the training set, the number of observations in the sample, and the number of observations that are randomly selected for the training and validation sets.

71

Regression

k-fold cross-validation:

- The sample data set is randomly divided into k equal-sized, mutually exclusive subsets called folds, and k iterations are executed.
- For each iteration, a different subset is designated as the validation set and the remaining k – 1 subsets are combined and designated as the training set.
- The model is estimated using the respective training set and evaluated using the respective validation set.
- The results of k iterations are combined and evaluated.
- A common choice for the number of folds is k = 10.
- The k-fold cross-validation method is more complex and time-consuming, but the results are less sensitive to how the observations are randomly assigned to the training validation sets.

72

Regression

Leave-one-out cross-validation:

- For a sample of n observations, an iteration consists of estimating the model on n – 1 observations and evaluating the model on the single observation that was omitted from the training data.
- This procedure is repeated for n total iterations so that the model is trained on each possible combination of n - 1 observations and evaluated on the single remaining observation in each case.

73

Regression

Big Data and Regression:

Inference and Very Large Samples:

- A credit card company with a very large database of its customers when they apply for credit cards.
- The customer records include information on the customer's annual household income, number of years of post-high school education, and number of members of the customer's household.
- In a second database, the company has records of the credit card charges accrued by each customer over the past year.
- A data analyst links these two databases to create one data set of all relevant information for a sample of 5,000 customers.

74

Regression

Inference and Very Large Samples:

- The file contains these data, split into a training set of 3,000 observations and a validation set of 2,000 observations.
- The company has decided to apply multiple regression to develop a model for predicting annual credit card charges for its new applicants (y).
- The independent variables are the customer's annual household income (x_1), number of members of the household (x_2), and number of years of post-high school education (x_3).

75

Regression

FIGURE 7.36 Excel Regression Output for Credit Card Company Example

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.602663145							
5	R Square	0.363202867							
6	Adjusted R Square	0.362565219							
7	Standard Error	4834.449957							
8	Observations	3000							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	3	39937797910	13312599303	569.5983495	6.5207E-293			
13	Residual	2996	70022231537	23371906.39					
14	Total	2999	1.0996E+11						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	2119.600282	333.0922952	6.363402314	2.27497E-10	1466.487528	2772.713036	1261.064442	2978.136122
18	Annual Income ($1000)	121.3384676	3.165148859	38.33578544	5.4905E-262	115.1323826	127.5445525	113.1803871	129.496548
19	Household Size	528.0996852	42.84154037	12.32681366	4.29401E-34	444.097873	612.1014973	417.6768433	638.522527
20	Years of Post-High School Education	-535.3593516	58.5960221	-9.136445316	1.15792E-19	-650.2518601	-420.4668432	-686.3889184	-384.3297849

76

Regression

- Coefficient of determination: 0.3632 indicating that this model explains approximately 36% of the variation in credit card charges accrued by the customers.
- P-value for each test of the individual regression parameters is also very small indicating that the estimated slopes associated with the dependent variables are all highly significant.
- For a fixed number of household members and number of years of post-high school education, accrued credit card charges increase by $121.34 when a customer's annual household income increases by $1,000.
- For a fixed annual household income and number of years of post-high school education, accrued credit card charges increase by $528.10 when a customer's household increases by one member.
- For a fixed annual household income and number of household members, accrued credit card charges decrease by $535.36 when a customer's number of years of post-high school education increases by one year.

77

Regression

- The small p-values associated with a model that is fit on an extremely large sample do not imply that an extremely large sample solves all problems.
- Virtually all relationships between independent variables and the dependent variable will be statistically significant if the sample size is sufficiently large.
- How much does sample size matter?

78

Regression

TABLE 7.4 Regression Parameter Estimates and the Corresponding *p* values for 10 Multiple Regression Models, Each Estimated on 50 Observations from the *LargeCredit* Data

Observations	b_0	*p* value	b_1	*p* value	b_2	*p* value	b_3	*p* value
1–50	−805.152	0.7814	154.488	1.45E-06	234.664	0.5489	207.828	0.6721
5–100	894.407	0.6796	125.343	2.23E-07	822.675	0.0070	−355.585	0.3553
101–150	−2,191.590	0.4869	155.187	3.56E-07	674.961	0.0501	−25.309	0.9560
151–200	2,294.023	0.3445	114.734	1.26E-04	297.011	0.3700	−537.063	0.2205
201–250	8,994.040	0.0289	103.378	6.89E-04	−489.932	0.2270	−375.601	0.5261
251–300	7,265.471	0.0234	73.207	1.02E-02	−77.874	0.8409	−405.195	0.4060
301–350	2,147.906	0.5236	117.500	1.88E-04	390.447	0.3053	−374.799	0.4696
351–400	−504.532	0.8380	118.926	8.54E-07	798.499	0.0112	45.259	0.9209
401–450	1,587.067	0.5123	81.532	5.06E-04	1,267.041	0.0004	−891.118	0.0359
451–500	−315.945	0.9048	148.860	1.07E-05	1,000.243	0.0053	−974.791	0.0420
Mean	1,936.567		119.316		491.773		−368.637	

79

Regression

- The individual values of the estimated regression parameters in the regressions based on 50 observations show a great deal of variation.
- In these 10 regressions, the estimated values of b_0 range from 22,191.590 to 8,994.040, the estimated values of b_1 range from 73.207 to 155.187, the estimated values of b_2 range from 2489.932 to 1,267.041, and the estimated values of b_3 range from 2974.791 to 207.828.
- p-values based on 50 observations are substantially larger than p-values based on 3,000 observations.
- suppose the credit card company also has a separate database of information on shopping and lifestyle characteristics that it has collected from its customers during a recent Internet survey.
- To increase the variation in the dependent variable explained by the model, we decided to include the number of hours per week spent watching television (x_4).

80

Regression

FIGURE 7.37 Excel Regression Output for Credit Card Company Example after Adding Number of Hours per Week Spent Watching Television

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.603724482							
5	R Square	0.36448325							
6	Adjusted R Square	0.36363448							
7	Standard Error	4830.393498							
8	Observations	3000							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	4	40078588918	10019647230	429.4250838	8.3277E-293			
13	Residual	2995	69881440529	23332701.35					
14	Total	2999	1.0996E+11						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	1712.552073	371.7837807	4.606311953	4.26973E-06	983.5746542	2441.529492	754.2898349	2670.814311
18	Annual Income ($1000)	121.6120724	3.164453912	38.43066631	4.943E-263	115.4073492	127.8167955	113.4557814	129.7683633
19	Household Size	531.213362	42.82435656	12.40446803	1.71315E-34	447.2452317	615.1814922	420.8347874	641.5919365
20	Years of Post-High School Education	-539.8345703	58.57519443	-9.216095235	5.64208E-20	-654.6862563	-424.9828843	-690.8104864	-388.8586541
21	Hours Per Week Watching Television	12.55178379	5.109759992	2.456433142	0.014088759	2.532789303	22.57077828	-0.618478873	25.72204645

81

Regression

- Coefficient of determination: 0.3645 indicating the addition of new independent variable increased the explained variation in sample values of accrued credit card charges by less than 1%.
- The estimated regression parameter for x_4 is 12.55.
- A 1-hour increase coincides with an increase of $12.55 in credit card charges accrued by each customer over the past year.
- The p-value associated with this estimate is 0.014 meaning that there is a relationship between x_4 and y.
- When the model is based on a very large sample, almost all relationships will be significant whether they are real or not
- on a very large sample, Statistical significance does not necessarily imply that a relationship is meaningful or useful.

82

Regression

Model Selection:

- when dealing with a sufficiently large sample, the p-value of every independent variable will be small, and variable selection procedures may suggest models with most or all the variables.
- If developing a regression model to make future predictions, the selection of the independent variables to include in the regression model should be based on the predictive accuracy on observations that have not been used to train the model.

Model A: y with three independent variables x_1, x_2, and x_3

Model B: y with four independent variables x_1, x_2, x_3, and x_4

83

Regression

- Compare the models based on predictive accuracy on the 2,000 observations in the validation set.
- For the first observation in the validation set (account number 18572870).

$$\hat{y}_1^A = 2119.60 + 121.33(50.2) + 528.10(5) - 525.36(1) = \$10,315.93$$

$$\hat{y}_1^B = 1712.55 + 121.61(50.2) + 531.21(5) - 539.89(1) + 12.55(4) = \$9,983.92$$

- Account number 18572870 has actual annual charges of $5,472.51
- Model A's prediction has a squared error of $(5,472.51-10,315.93)^2$ = 23,458,721
- Model B's prediction has a squared error of $(5,472.51 - 9,983.92)^2$ = 20,352,797.
- Model A's predictions are slightly more accurate than Model B's predictions on the validation set, as measured by squared error.

84

Regression

FIGURE 7.37 Excel Regression Output for Credit Card Company Example after Adding Number of Hours per Week Spent Watching Television

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.603724482							
5	R Square	0.36448325							
6	Adjusted R Square	0.36363448							
7	Standard Error	4830.393498							
8	Observations	3000							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	4	40078588918	10019647230	429.4250838	8.3277E-293			
13	Residual	2995	69881440529	23332701.35					
14	Total	2999	1.0996E+11						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
17	Intercept	1712.552073	371.7837807	4.606311953	4.26973E-06	983.5746542	2441.529492	754.2898349	2670.814311
18	Annual Income ($1000)	121.6120724	3.164453912	38.43066631	4.943E-263	115.4073492	127.8167955	113.4557814	129.7683633
19	Household Size	531.213362	42.82435656	12.40446803	1.71315E-34	447.2452317	615.1814922	420.8347874	641.5919365
20	Years of Post-High School Education	-539.8345703	58.57519443	-9.216095235	5.64208E-20	-654.6862563	-424.9828843	-690.8104864	-388.8586541
21	Hours Per Week Watching Television	12.55178379	5.109759992	2.456433142	0.014088759	2.532789303	22.57077828	-0.618478873	25.72204645

85

Regression

Prediction with Regression:

Butler Trucking Company

Multiple regression equation based on the 300 past routes using Miles (x1) and Deliveries (x2) as the independent variables to estimate travel time (y) for a driving assignment:

$$\hat{y} = 0.1273 + 0.0672x_1 + 0.6900x_2$$

TABLE 7.5 Predicted Values and 95% Confidence Intervals and Prediction Intervals for 10 New Butler Trucking Routes

Assignment	Miles	Deliveries	Predicted Value	95% CI Half-Width(+/−)	95% PI Half-Width(+/−)
301	105	3	9.25	0.193	1.645
302	60	4	6.92	0.112	1.637
303	95	5	9.96	0.173	1.642
304	100	1	7.54	0.225	1.649
305	40	3	4.88	0.177	1.643
306	80	3	7.57	0.108	1.637
307	65	4	7.25	0.103	1.637
308	55	3	5.89	0.124	1.638
309	95	2	7.89	0.175	1.643
310	95	3	8.58	0.154	1.641

86

Regression

Confidence interval:

- An interval estimate of the mean y value given values of the independent variables.

Prediction interval:

- An interval estimate of an individual y value given values of the independent variables.

The general form for the confidence interval on the mean y value given values of $x_1, x_2, \ldots, x_2$ is

$$\hat{y} \pm t_{\alpha/2} s_{\hat{y}}$$

The prediction interval on the individual y value given values of $x_1, x_2, \ldots, x_q$ is

$$\hat{y} \pm t_{\alpha/2} \sqrt{s_{\hat{y}}^2 + \frac{SSE}{n - q - 1}}$$

87