Leyli Garryyeva
Professor deCamp

Assignment 7

CSCI 688
December 30, 2019

Ch.8 - Regression Models for Quantitative & Qualitative Predictors: 8, 12, 15, 19, 34.

Problem 8.8

Refer to **Commercial properties** Problems 6.18 and 7.7. The vacancy rate predictor (X_3) does not appear to be needed when property age (X_1), operating expenses and taxes (X_2), and total square footage (X_4) are included in the model as predictors of rental rates (Y).

Regression Analysis: Y.Rental Rates versus x1, x1^2, ... es.Taxes, X4.Size

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	145.023	36.256	30.10	0.000
x1	1	61.144	61.144	50.77	0.000
x1^2	1	7.115	7.115	5.91	0.017
X2.Expenses.Taxes	1	34.350	34.350	28.52	0.000
X4.Size	1	48.583	48.583	40.34	0.000
Error	76	91.535	1.204		
Total	80	236.558			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.09745	61.31%	59.27%	54.93%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	10.189	0.671	15.19	0.000	
x1	-0.1818	0.0255	-7.13	0.000	1.90
x1^2	0.01415	0.00582	2.43	0.017	1.61
X2.Expenses.Taxes	0.3140	0.0588	5.34	0.000	1.53
X4.Size	0.000008	0.000001	6.35	0.000	1.27

Regression Equation

Y.Rental Rates = 10.189 - 0.1818 x1 + 0.01415 x1^2 + 0.3140 X2.Expenses.Taxes
+ 0.000008 X4.Size

Figure 1: Fitted regression model output

a. The age of the property (X_1) appears to exhibit some curvature when plotted against the rental rates (Y). Fit a polynomial regression model with centered property age (x_1), the square of centered property age (x_1^2), operating expenses and taxes (X_2), and total square footage (X_4). Plot the Y observations against the fitted values. Does the response function provide a good fit?

   Based on the Figure 2, the response function provides a good fit.

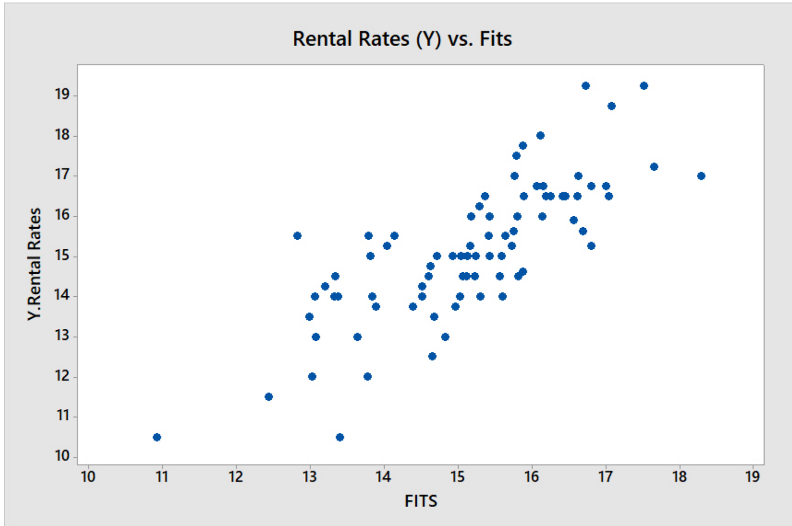

Figure 2: Y observations vs. the fitted values

b. Calculate R_a^2. What information does this measure provide?
From the output in Figure 1, $\boxed{R_a^2 = 59.27\%}$. Or,

$$R_a^2 = 1 - \frac{n-1}{n-p} \cdot \frac{SSE}{SSTO} = 1 - \frac{80 * 91.535}{76 * 236.588} = 59.27\%$$

The variation in the Rental Rates (Y) is reduced by about 59.27% with the use of the given set of X variables, when the sums of squares are divided by their degrees of freedom.

c. Test whether or not the square of centered property age (x_1^2) can be dropped from the model; use $\alpha = 0.05$. State the alternatives, decision rule, and conclusion. What is the $p-$value of the test?

$$\begin{aligned} H_0 &: \beta_2 = 0 \\ H_a &: \beta_2 \neq 0 \end{aligned} \tag{1}$$

Decision Rule:

$$\begin{aligned} &\text{If } |t^*| \leq t(1-\alpha/2; n-p), \text{ conclude } H_0 \\ &\text{If } |t^*| > t(1-\alpha/2; n-p), \text{ conclude } H_a \end{aligned} \tag{2}$$

From the output in Figure 1: $t^* = 2.43$ and p-value $= 0.017$. Also,
$t^* = \frac{b_1}{s\{b_1\}} = \frac{0.01415}{0.00582} = 2.49$

p-value $= 0.01$

$t(1-\alpha/2; n-p) = t(0.975, 76) = 1.99$

Then, $|t^*| > 1.99$ and we conclude H_a that the square of centered property age (x_1^2) cannot be dropped from the model with p-value = 0.01 (< 0.05).

d. Estimate the mean rental rate when $X_1 = 8$, $X_2 = 16$, and $X_4 = 250{,}000$; use a 95 percent confidence interval. Interpret your interval.

$\bar{X} = 7.864$; $x_1 = X_1 - \bar{X} = 8 - 7.864 = 0.136$; $x_1^2 = 0.0185$

$\boxed{17.25 \pm 0.744 = [16.506, 17.995]}$

e. Express the fitted response function obtained in part (a) in the original X variables.

$$\begin{aligned} b_0^{'} &= b_0 - b_1\bar{X} + b_{11}\hat{X} = 10.189 + (0.1818 \cdot 7.86) + (0.01415 \cdot 7.86^2) = 12.49 \\ b_1^{'} &= b_1 - 2b_{11}\hat{X} = -0.1818 - 2(0.01415 \cdot 7.86) = -0.40 \\ b_{11}^{'} &= b_{11} = 0.01415 \end{aligned} \tag{3}$$

$\boxed{\hat{Y} = 12.49 - 0.40X_1 + 0.01415X_1^2 + 0.3140X_2 + 0.000008X_4}$

Problem 8.12

A student who used a regression model that included indicator variables was upset when receiving only the following output on the multiple regression printout: XTRANSPOSE X SINGULAR. What is a likely source of the difficulty?

This would happen if $(X'X)^{-1}$ does not exist given $X'X$-matrix is not of full rank or singular with the determinant $D = 0$ (ref. p.190), making it non-invertible.The singular design matrix is created if instead of using $c - 1$ indicator variables c variables are used.

Problem 8.15

Refer to **Copier maintenance** Problem 1.20. The users of the copiers are either training institutions that use a small model, or business firms that use a large, commercial model. An analyst at Tri-City wishes to fit a regression model including both number of copiers serviced (X_1) and type of copier (X_2) as predictor variables and estimate the effect of copier model (S-small, L-large) on number of minutes spent on the service call. Records show that the models serviced in the 45 calls were:

i:	1	2	3	...	43	44	45
X_{i2}:	S	L	L	...	L	L	L

Assume that regression model (8.33) is appropriate, and let $X_2 = 1$ if small model and 0 if large, commercial model.

a. Explain the meaning of all regression coefficients in the model.
X_1 - number of copiers serviced;
X_2 is an indicator variable where: $X_2 = 1$ if copier is small & $X_2 = 0$ if copier is large.

The response function for this regression model is:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

For a large copier $X_2 = 0$ the response function becomes $E\{Y\} = \beta_0 + \beta_1 X_1$ (large copier), which would be a straight line, with Y intercept β_0 and slope β_1.

For a small copier $X_2 = 1$ with response function $E\{Y\} = (\beta_0 + \beta_2) + \beta_1 X_1$ (small copier), Y intercept $\beta_0 + \beta_2$ and slope β_1.

Overall, β_0 is the intercept for large copiers, β_1 is the increase in service time per copier serviced; β_2 is the increase in service time over large copiers.

b. Fit the regression model and state the estimated regression function.
$\boxed{Y = -0.92 + 15.046X_1 + 0.76X_2}$.

Regression Analysis: Y versus X1.CopiersServed, X2.ModelType

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	76966.5	38483.2	473.94	0.000
X1.CopiersServed	1	76560.5	76560.5	942.88	0.000
X2.ModelType	1	6.0	6.0	0.07	0.786
Error	42	3410.3	81.2		
Lack-of-Fit	14	1089.0	77.8	0.94	0.533
Pure Error	28	2321.3	82.9		
Total	44	80376.8			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
9.01101	95.76%	95.56%	95.05%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0.92	3.10	-0.30	0.767	
X1.CopiersServed	15.046	0.490	30.71	0.000	1.01
X2.ModelType	0.76	2.78	0.27	0.786	1.01

Regression Equation

Y = -0.92 + 15.046 X1.CopiersServed + 0.76 X2.ModelType

Figure 3: Fitted regression model output

c. Estimate the effect of copier model on mean service time with a 95 percent confidence interval. Interpret your interval estimate.
β_2 measure the effect of copier model on the service time, then we need to find a 95 percent confidence interval for β_2. With $t(0.975; 42) = 2.02$ the confidence interval for β_2 is:

$0.76 \pm 2.02 \cdot 2.78 = 0.76 \pm 5.62$ or $-4.86 \leq \beta_2 \leq 6.38$

With 95% confidence, we conclude that on average small copiers' service times tends to be varying from around 5 minutes less than the large copiers or over 6 minutes later than the large copiers.

d. Why would the analyst wish to include X_1, number of copiers, in the regression model when interest is in estimating the effect of type of copier model on service time?

The analyst should include X_1, number of copiers, in the regression model even if his interest is in estimating the effect of type of copier model on service time. Because X_2 on its own is not a good enough predictor variable since having multiple small copiers might have the same effect as having one large copier (control number of copiers). Then, omitting X_1 would lead to an incorrect regression output. If we conduct a test whether the X_1 can be dropped at $\alpha = 0.05$, we would get $t^* = 30.71, t(0.95, 42) = 1.68$, which would lead to a conclusion that the X_1 predictor cannot be dropped.

e. Obtain the residuals and plot them against X_1X_2. Is there any indication that an interaction term in the regression model would be helpful?
Based on the Figure 4, there is no evidence of unequal error variances or inadequacies. There is a small upward slope and adding an interaction term could help improve the model.

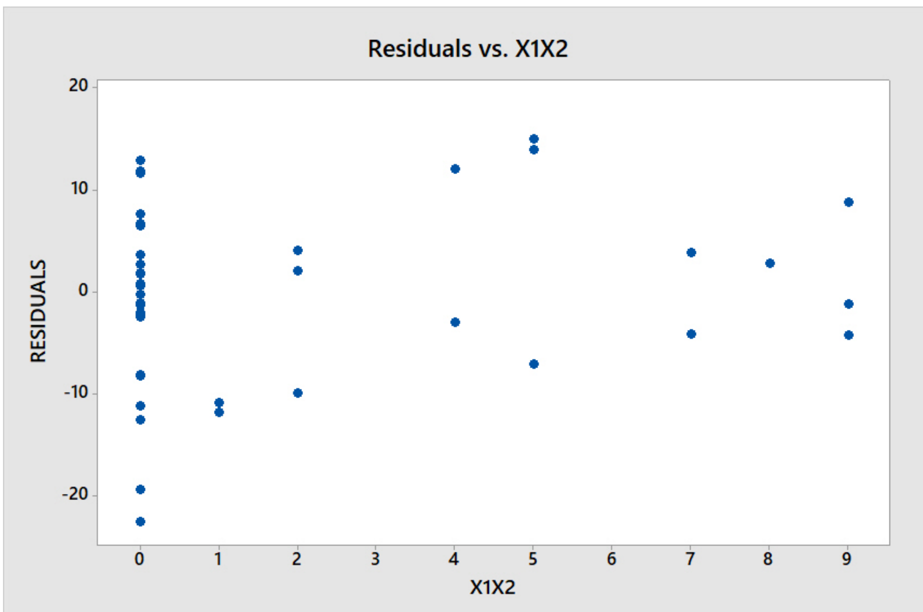

Figure 4: Residual Plot against X_1X_2 interaction term

Problem 8.19

Refer to **Copier maintenance** Problems 1.20 and 8.15.

a. Fit regression model (8.49) and state the estimated regression function.
$\boxed{Y = 2.81 + 14.339X_1 - 8.14X_2 + 1.777X_1X_2}$

Regression Analysis: Y versus X1.CopiersServed, X2.ModelType, X1X2

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	77222.4	25740.8	334.57	0.000
X1.CopiersServed	1	41887.5	41887.5	544.44	0.000
X2.ModelType	1	163.8	163.8	2.13	0.152
X1X2	1	255.9	255.9	3.33	0.075
Error	41	3154.4	76.9		
Lack-of-Fit	13	833.1	64.1	0.77	0.680
Pure Error	28	2321.3	82.9		
Total	44	80376.8			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
8.77140	96.08%	95.79%	95.21%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.81	3.65	0.77	0.445	
X1.CopiersServed	14.339	0.615	23.33	0.000	1.67
X2.ModelType	-8.14	5.58	-1.46	0.152	4.28
X1X2	1.777	0.975	1.82	0.075	4.70

Regression Equation

Y = 2.81 + 14.339 X1.CopiersServed - 8.14 X2.ModelType + 1.777 X1X2

Figure 5: Fitted regression model output

b. Test whether the interaction term can be dropped from the model; at $\alpha = 0.10$. State the alternatives, decision rule, & conclusion. What is the $P-$value of the test? If the interaction term cannot be dropped from the model describe the nature of the interaction effect.

$$\begin{aligned} H_0 &: \beta_3 = 0 \\ H_a &: \beta_3 \neq 0 \end{aligned} \tag{4}$$

Decision Rule:

$$\begin{aligned} &\text{If } |t^*| \leq t(1-\alpha/2; n-p), \text{ conclude } H_0 \\ &\text{If } |t^*| > t(1-\alpha/2; n-p), \text{ conclude } H_a \end{aligned} \tag{5}$$

From the output in Figure 1: $t^* = 1.82$ and p-value $= 0.075$. Also,
$t(1-\alpha/2; n-p) = t(0.95, 41) = 1.68$

Then, $|t^*| > 1.68$ and we conclude H_a that the interaction term (X_1X_2) cannot be dropped from the model with p-value = 0.075 (< 0.10).

Problem 8.34

In a regression study, three types of banks were involved, namely, commercial, mutual savings, and savings and loan. Consider the following system of indicator variables for type of bank:

Type of Bank	X_2	X_3
Commercial	1	0
Mutual savings	0	1
Savings and loan	-1	-1

a. Develop a first-order linear regression model for relating last year's profit or loss (Y) to size of bank (X_1) and type of bank (X_2, X_3).

$$\boxed{E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_i}$$

b. State the response functions for the three types of banks.

Type of Bank	E{Y}
Commercial	$(\beta_0 + \beta_2) + \beta_1 X_1$
Mutual savings	$(\beta_0 + \beta_3) + \beta_1 X_1$
Savings and loan	$(\beta_0 - \beta_2 - \beta_3) + \beta_1 X_1$

c. Interpret each of the following quantities:

(1) β_2 *indicates how much higher or lower the response function for* `Commercial` *bank type is than for* `Savings and loan` *bank for any given size of bank. (Difference of the* `Commercial` *bank type from Average* β_0*)*

(2) β_3 *indicates how much higher or lower the response function for* `Savings and loan` *bank type is than for* `Commercial` *bank for any given size of bank.*

(3) -β_2 - β_3 *indicates how much higher or lower the response function for* `Commercial` *bank type is than for* `Mutual savings` *bank types for any given size of bank (forces* `Savings and loan` *into the sum of the* `Commercial` *and* `Mutual savings` *types in the opposite direction).*