Chapter 6 - Multiple Regression : 6.18, 19, 20, 21, 25, 26.

Problem 6.18

6.18. Commercial properties. A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data below are taken from 81 suburban commercial properties that are the newest, best located, most attractive, and expensive for five specific geographic areas. Shown here are:
the age (X1),
operating expenses and taxes (X2),
vacancy rates (X3),
total square footage (X4),
and rental rates (Y)

a. Prepare a stem-and-leaf plot for each predictor variable. What information do these plots provide?

Stem-and-leaf of X1 N = 81

16	0	0111111111111111
39	0	2222222222333333333333
(5)	0	44444
37	0	6
36	0	8
35	1	11
33	1	22333
28	1	4444455555555
15	1	6666666667
5	1	888
2	2	00

Leaf Unit = 1

Stem-and-leaf of X2 N = 81

1	3	0
4	4	089
11	5	0023579
14	6	126
19	7	03999
28	8	011234569
36	9	01444678
(15)	10	013333555666777
30	11	1223335556679
17	12	000011466777899
2	13	2
1	14	6

Leaf Unit = 0.1

Stem-and-leaf of X3 N = 81

(48)	0	000000000000000000000000000000023333333333444444
33	0	5555556678889
20	1	0234444
13	1	69
11	2	12234
6	2	77
4	3	3
3	3	
3	4	
3	4	
3	5	
3	5	7
2	6	0
1	6	
1	7	3

Leaf Unit = 0.01

Stem-and-leaf of X4 N = 81

13	0	2333333344444
29	0	5555666777888999
(16)	1	0000001222222333
36	1	568888
30	2	0001123333444
17	2	5567788999
7	3	14
5	3	56
3	4	23
1	4	8

Leaf Unit = 10000

Figure 1: Stem-and-leaf plot for the Predictor variables X1 and X2

The stem-and-leaf plots show that all of the X values are skewed and may not be normally distributed with $X3$ having potential outliers.

b. Obtain the scatter plot matrix and the correlation matrix. Interpret these and state your principal findings.

The output shown below indicates that `Vacancy Rates` variable is negatively correlated with `Age` and `Expenses & Taxes`. `Rental Rates (Y)` is negatively correlated with `Age` and positively correlated with `Expenses & Taxes` and `Total Square Footage`. `Age` is positively correlated with `Expenses & Taxes`.

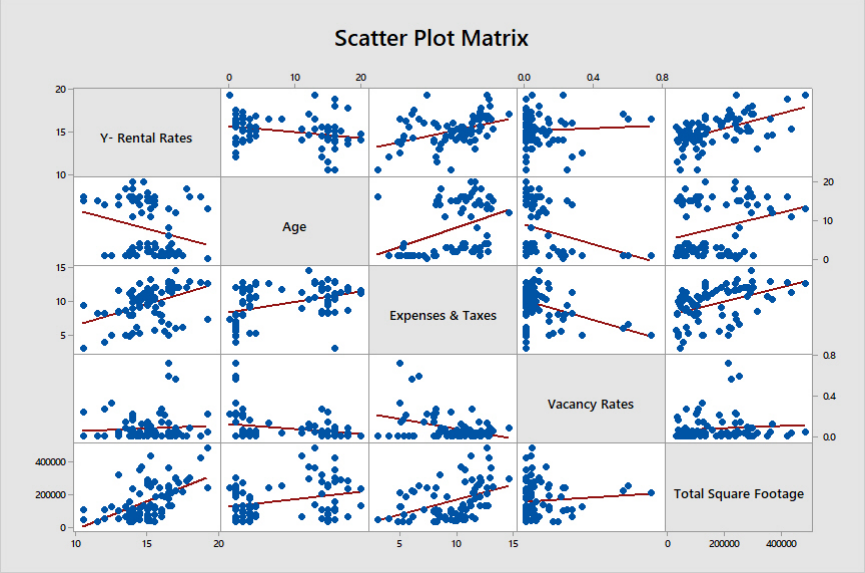


Figure 2: The scatter plot matrix

Correlations

	Y- Rental Rates	Age	Expenses & Taxes	Vacancy Rates
Age	-0.250			
	0.024			
Expenses & Taxes	0.414	0.389		
	0.000	0.000		
Vacancy Rates	0.067	-0.253	-0.380	
	0.555	0.023	0.000	
Total Square Foo	0.535	0.289	0.441	0.081
	0.000	0.009	0.000	0.474

Cell Contents
Pearson correlation
P-Value

Figure 3: The Correlation Matrix

c. Fit regression model (6.5) for four predictor variables to the data. State the estimated regression function.

$$\boxed{Y = 12.201 - 0.1420\ X1 + 0.2820\ X2 + 0.62\ X3 + 0.000008\ X4}$$

d. Obtain the residuals and prepare a box plot of the residuals. Does the distribution appear to be fairly symmetrical?

The distribution is somewhat symmetrical with slight skewness towards the bottom (below zero).

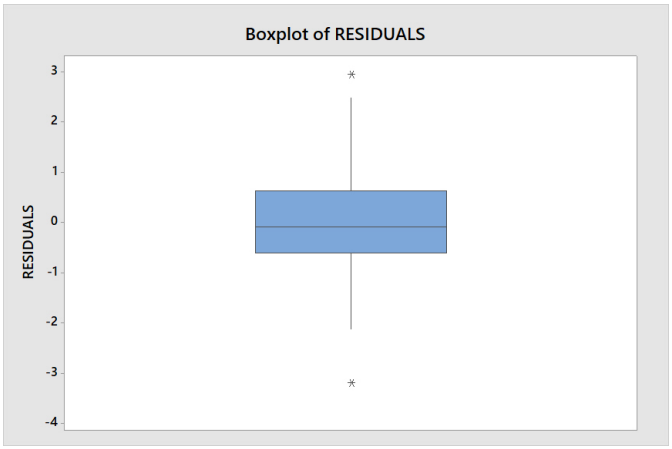
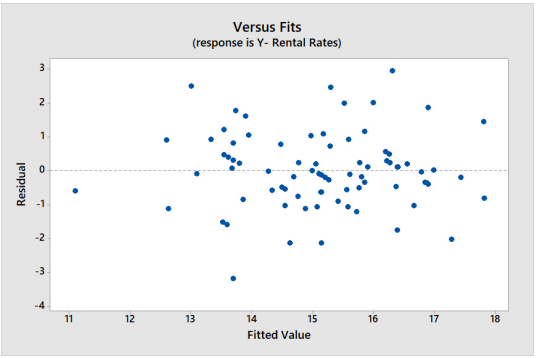


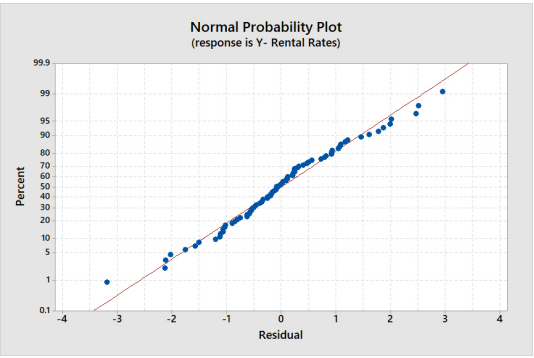
Figure 4: The Correlation Matrix

e. Plot the residuals against $\hat{Y}$, each predictor variable, and each two-factor interaction term on separate graphs. Also prepare a normal probability plot. Analyze your plots and summarize your findings.

From the outputs below, Figure 5a shows constant variance assumption of the residuals holds and Figure 5b indicates normality of the residual distribution.



(a) Residuals vs Fits

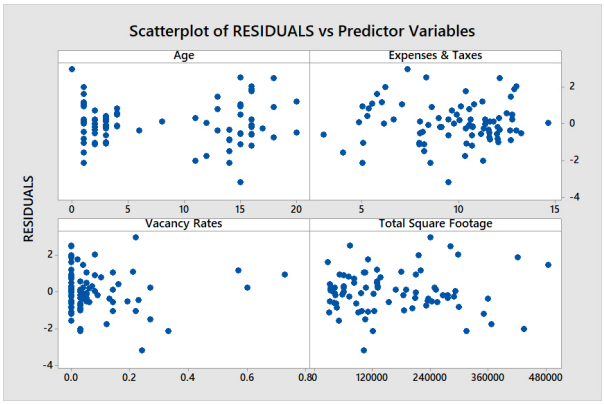


(b) Normal Probability Plot

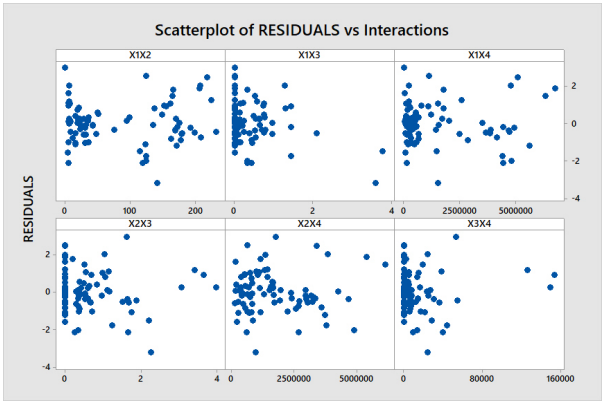
Figure 5: Diagnostic Plots - Commercial Properties

From the Figure 6a `Vacancy Rates` is non-random as well as the interactions of this variable with other variables. All other variables and their interactions show that there is no relationship between the residuals these variables/interactions.

f. Can you conduct a formal test for lack of fit here?
No (because there are no replicates).



(a) Residuals vs Predictor Variables



(b) Two factor interactions

Figure 6: Diagnostic Plots - Commercial Properties

g. Divide the 81 cases into two groups. placing the 40 cases with the smallest fitted values $\hat{Y}_i$ into group 1 and the remaining cases into group 2. Conduct the Brown-Forsythe test for constancy of the error variance, using $\alpha = .05$. State the decision rule and conclusion.

Decision Rule for $\alpha = 0.05$ with t(0.975, 79) = 1.99:

$$\begin{aligned} |t^*_{BF}| &<= 1.99, \text{ conclude the error variance is constant} \\ |t^*_{BF}| &> 1.99, \text{ conclude the error variance is not constant} \end{aligned} \tag{1}$$

Based on the calculations in Excel:

$\bar{d}_1 = 20.309$
$\bar{d}_2 = 1.287E - 05$
$s^2 = 0.5412$
$\frac{1}{n_1} + \frac{1}{n_2} = 0.04939$
$t^*_{BF} = \frac{\bar{d}_1 + \bar{d}_2}{\sqrt{s^2 * \frac{1}{n_1} + \frac{1}{n_2}}} = 0.55$

$\boxed{t^*_{BF} < 1.99,}$ conclude that the error variance is constant & doesn't vary with the level of X (p-value = 0.582).

Problem 6.19

Refer to **Commercial properties** Problem 6.18. Assume that regression model (6.5) for four predictor variables with independent normal error terms is appropriate.

Regression Analysis: Y versus X1, X2, X3, X4

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	4	138.327	58.47%	138.327	34.5817	26.76	0.000
X1	1	14.819	6.26%	57.243	57.2428	44.29	0.000
X2	1	72.802	30.78%	25.759	25.7590	19.93	0.000
X3	1	8.381	3.54%	0.420	0.4197	0.32	0.570
X4	1	42.325	17.89%	42.325	42.3250	32.75	0.000
Error	76	98.231	41.53%	98.231	1.2925		
Total	80	236.558	100.00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
1.13689	58.47%	56.29%	114.278	51.69%

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	12.201	0.578	(11.049, 13.352)	21.11	0.000	
X1	-0.1420	0.0213	(-0.1845, -0.0995)	-6.65	0.000	1.24
X2	0.2820	0.0632	(0.1562, 0.4078)	4.46	0.000	1.65
X3	0.62	1.09	(-1.55, 2.78)	0.57	0.570	1.32
X4	0.000008	0.000001	(0.000005, 0.000011)	5.72	0.000	1.41

Regression Equation

Y = 12.201 - 0.1420 X1 + 0.2820 X2 + 0.62 X3 + 0.000008 X4

Figure 7: The Regression Analysis Output

a. Test whether there is a regression relation; use $\alpha = 0.05$. State the alternatives. decision rule, and conclusion. What does your test imply about β_1, β_2, β_3, β_4? What is the $P-$value of the test?

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \\ H_a &: \text{ not all } \beta_k = 0 \end{aligned} \tag{2}$$

$$\begin{aligned} F^* \leq F(1-\alpha; p-1, n-p), \text{ conclude } H_0 \\ \text{Otherwise, conclude } H_a \end{aligned} \tag{3}$$

$F^* = 26.76$
$p-$value $= 0.000$
$F(1-\alpha; p-1, n-p) = F(0.95, 4, 76) = 2.492$
$\boxed{F^* > 2.492}$, conclude H_a that not all $\beta_k = 0$ (there is a relation).

b. Estimate β_1, β_2, β_3, β_4 jointly by the Bonferroni procedure, using a 95 percent family confidence coefficient. Interpret your results.

$B = t(1 - \alpha/2g; n - p) = t(1 - 0.05/8; 81 - 5) = t(0.99375; 76) = 2.5585$
$b_k \pm Bs\{b_k\}$

$$
\begin{aligned}
b_1 \pm Bs\{b_1\} &= -0.1420 \pm 2.5585 * 0.0213 = \boxed{-0.1420 \pm 0.054449 = [-0.1965, -0.0875]} \\
b_2 \pm Bs\{b_2\} &= 0.2820 \pm 2.5585 * 0.0632 = \boxed{0.2820 \pm 0.1616872 = [0.1203, 0.4437]} \\
b_3 \pm Bs\{b_3\} &= 0.62 \pm 2.5585 * 1.09 = \boxed{0.62 \pm 2.7888 = [-2.1688, 3.4088]} \\
b_4 \pm Bs\{b_4\} &= 8 \cdot 10^{-6} \pm 2.5585 \cdot 10^{-6} = \\
&= \boxed{8 \cdot 10^{-6} \pm 2.5585 \cdot 10^{-6} == [5.4415 \cdot 10^{-6}, 1.0559 \cdot 10^{-6}]}
\end{aligned}
\tag{4}
$$

Bonferroni procedure gives wider confidence interval for the β's compared to the individual 95% intervals shown in the Figure 7, because here we are estimating the family confidence interval which requires 0.9873% for individual confidence intervals.

c. Calculate R^2 and interpret this measure.

Based on the Minitab output $\boxed{R^2 = 58.47\%}$, which implies that adding the set of X (Age, Expenses & Taxes, Vacancy Rates, Total Square Footage) values in the model reduces the total variation in Y by 58.47%.

Problem 6.20

Refer to **Commercial properties** Problem 6.18. Assume that regression model (6.5) for four predictor variables with independent normal error terms is appropriate. The researcher wishes to obtain simultaneous interval estimates of the mean rental rates for four typical properties specified as follows (see Table 1):

Table 1

	1	2	3	4
X_1	5.0	6.0	14.0	12.0
X_2	8.25	8.50	11.50	10.25
X_3	0	0.23	0.11	0
X_4	250,000	270,000	300,000	310,000

Obtain the family of estimates using a 95 percent family confidence coefficient. Employ the most efficient procedure.
$W = \sqrt{pF(1 - \alpha; p, n - p} = \sqrt{5F(0.95, 5, 76)} = 2.179$
$B = t(1 - \alpha/2g; n - p) = t(1 - 0.05/8; 76) = t(0.9938; 76) = 2.5585$
Because $W < B$, Working-Hotelling procedure will give tighter confidence interval and would be a better procedure to use.

Table 2

W = 2.179	1	2	3	4
$\hat{Y}_h$	15.8175	16.0486	15.9242	15.8675
$s\{\hat{Y}_h\}$	0.2780832	0.2359255	0.2221593	0.2591281
$\hat{Y}_h \pm Ws\{\hat{Y}_h\}$	15.8175 ± 0.6059	16.0486 ± 0.5141	15.9242 ± 0.4841	15.8675 ± 0.5646
	[15.2115, 16.4234]	[15.5345 , 16.5627]	[15.4401 , 16.4083]	[15.3029 , 16.4321]

Problem 6.21

Refer to **Commercial properties** Problem 6.18. Assume that regression model (6.5) for four predictor variables with independent normal error terms is appropriate. Three properties with the following characteristics did not have any rental information available.

	1	2	3
X_1	4.0	6.0	12.0
X_2	10.0	11.5	12.5
X_3	0.10	0	0.32
X_4	80,000	120,000	340,000

Develop separate prediction intervals for the rental rates of these properties, using a 95 percent statement confidence coefficient in each case. Can the rental rates of these three properties be predicted fairly precisely? What is the family confidence level for the set of three predictions?
$\hat{Y}_h \pm t(1-\alpha/2; n-p)s\{pred\}$, where $s\{pred\} = \sqrt{MSE + s\{\hat{Y}_h\}}$ & $t(1-0.025; 76) = 1.992$

	1	2	3
$s\{pred\}$	1.118	1.118	1.161
$\hat{Y}_h$	15.15	15.54	16.91
$\hat{Y}_h \pm t(0.975; 76)s\{pred\}$	15.15 ± 2.226	15.54 ± 2.228	16.91 ± 2.312
	[12.92, 17.37]	[13.31, 17.77]	[14.6, 19.23]

The rental rates of the given three properties cannot be predicted fairly precisely, because the predicted values are within too wide of a confidence interval.
$B = t(1-\alpha/12; 76) = t(0.9917, 76) = 2.448$
$S = \sqrt{gF(1-\alpha; g, n-p)} = \sqrt{3F(0.95; 3, 76)} = \sqrt{3 * 2.725} = 2.859$
$B < S$ use Bonferroni procedure. See the prediction intervals below:

Bonferroni	15.15 ± 2.737	15.54 ± 2.737	16.91 ± 2.842
	[12.41, 17.89]	[13.81, 18.28]	[14.07, 19.75]

Problem 6.25

An analyst wanted to fit the regression model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, i = 1, ..., n$, by the method of least squares when it is known that $\beta_2 = 4$. How can the analyst obtain the desired fit by using a multiple regression computer program?

Transform $Y_i = \beta_0 + \beta_1 X_{i1} + 4X_{i2} + \beta_3 X_{i3} + \epsilon_i$ to
$Y_i^* = Y_i - 4X_{i2} = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \epsilon_i$.

Problem 6.26

For regression model (6.1), show that the coefficient of simple determination between Y_i and $\hat{Y}_i$ equals the coefficient of multiple determination R^2.
$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$; $\sum \hat{Y}_i e_i = 0$; $\sum X_i e_i = 0$; $\sum e_i = 0$;

NTS: $SSE(\hat{Y}) = SSE(X_1 X_2)$ or $\sum (Y_i - \hat{Y}_i^*)^2 = \sum (Y_i - \hat{Y})^2$, where $\hat{Y}_i^* = b_0^* + b_1^* \hat{Y}_i$ is the fits for $Y|\hat{Y}$ with $\hat{Y}_i$ being the fits for $Y|X_1 X_2$.

Prep Work:
$b_1 = \frac{S_{xy}}{S_{xx}} => b_1^* = \frac{S_{\hat{y}y}}{S_{\hat{y}\hat{y}}}$

$$
\begin{aligned}
b_1^* = \frac{S_{\hat{y}y}}{S_{\hat{y}\hat{y}}} = \frac{\sum (\hat{Y}_i - \bar{Y})(Y_i - \bar{Y})}{\sum (\hat{Y}_i - \bar{Y})^2} = \frac{\sum (\hat{Y}_i - \bar{Y}) \left[(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \right]}{\sum (\hat{Y}_i - \bar{Y})^2} = \\
= \frac{\sum (\hat{Y}_i - \bar{Y}) \left[e_i + (\hat{Y}_i - \bar{Y}) \right]}{\sum (\hat{Y}_i - \bar{Y})^2} = \frac{\sum (\hat{Y}_i - \bar{Y}) e_i + \sum (\hat{Y}_i - \bar{Y})^2}{\sum (\hat{Y}_i - \bar{Y})^2} = 0 + 1 = 1
\end{aligned} \tag{5}
$$

$b_0 = \bar{Y} - b_1 \bar{X} => b_0^* = \bar{Y} - b_1^* \bar{\hat{Y}} = \bar{Y} - \bar{\hat{Y}} = 0$
$\hat{Y}_i^* = b_0^* + b_1^* \hat{Y}_i = 0 + 1 \cdot \hat{Y}_i = \hat{Y}_i \qquad => \hat{Y}_i^* = \hat{Y}_i$
$\boxed{\therefore SSE(\hat{Y}) = \sum (Y_i - \hat{Y}_i^*)^2 = \sum (Y_i - \hat{Y}_i)^2 = SSE(X_1 X_2)}$

A R Code:

Problem 6.20 Calculations:

```
# fit regression model
comm_fit <- lm(Y.Rental.Rates ~. , data = comm_data)
summary(comm_fit)

# get variance-covariance matrix
varcovar <- as.matrix(vcov(summary(comm_fit)))

# store X's as cases and add 1's to the first column as X_0
case1 <- c(1, 5, 8.25, 0,  250000)
case2 <- c(1, 6, 8.5, 0.23, 270000)
case3 <- c(1,14, 11.5, 0.11, 300000)
case4 <- c(1, 12, 10.25, 0, 310000)

s1 <- sqrt(colSums(case1*varcovar%*%case1))
s2 <- sqrt(colSums(case2*varcovar%*%case2))
s3 <- sqrt(colSums(case3*varcovar%*%case3))
s4 <- sqrt(colSums(case4*varcovar%*%case4))

# store betas as matrix
beta <- as.matrix(coef(comm_fit))

# calculate Y_hats
Y_hat1 = t(case1)%*%beta
Y_hat2 = t(case2)%*%beta
Y_hat3 = t(case3)%*%beta
Y_hat4 = t(case4)%*%beta
```

Problem 6.21 Calculations:

```
mse <- mean(comm_fit$residuals^2)
# store X's as cases and add 1's to the first column as X_0
case1 <- c(1,4, 10, 0.10, 80000)
case2 <- c(1, 6, 11.5, 0, 120000)
case3 <- c(1, 12.0, 12.5, 0.32, 340000)

s1 <- sqrt(mse + colSums(case1*varcovar%*%case1))
s2 <- sqrt(mse + colSums(case2*varcovar%*%case2))
s3 <- sqrt(mse + colSums(case3*varcovar%*%case3))

Y_hat1 = t(case1)%*%beta
Y_hat2 = t(case2)%*%beta
Y_hat3 = t(case3)%*%beta
```