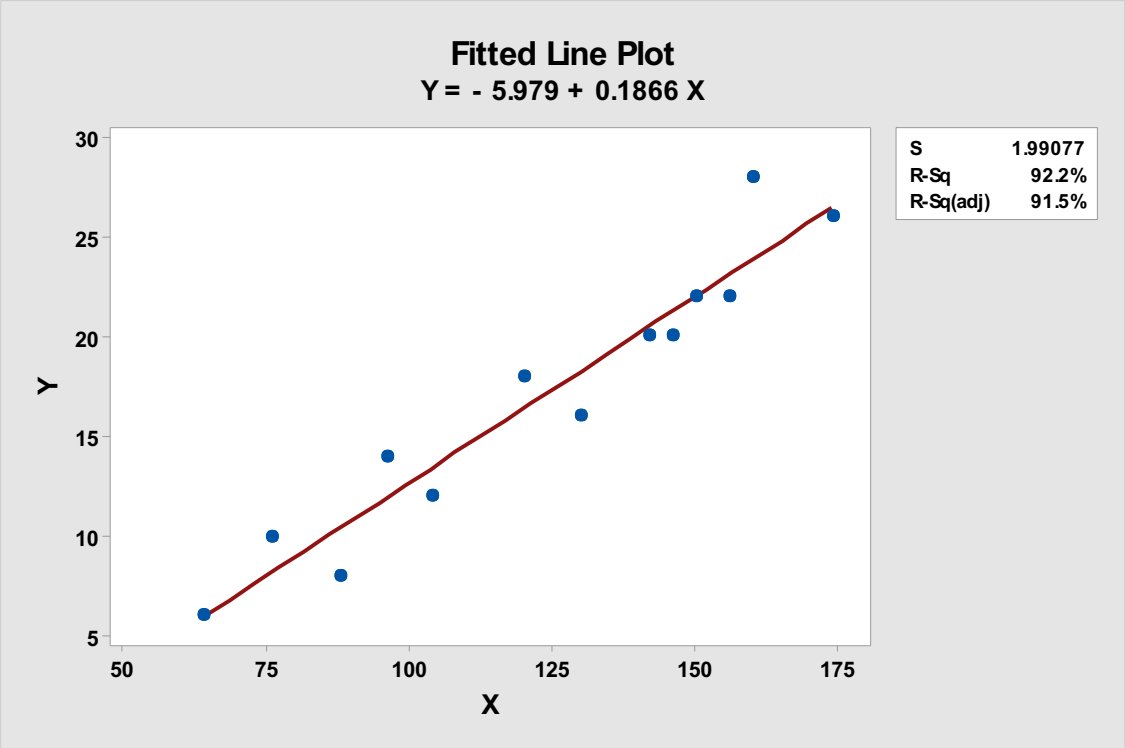
# ADM 2304
# Applications of Statistical Methods in Business
# Assignment Four

**Question 1**

**(a)**




## Correlation: X, Y

### Correlation

*Pearson correlation of X and Y = 0.960064*
*P-Value = <0.0001*

$$z_{x_i} = \frac{x_i - \bar{x}}{s_x} \qquad z_{y_i} = \frac{y_i - \bar{y}}{s_y}$$

$$r = \frac{\sum_i z_{x_i} z_{y_i}}{n-1} = 0.960064$$

The graph shows that there is a positive correlation between X and Y. The correlation coefficient is significant because it is very high and it is close to 1 as it is 0.9601.

**(b)**

| Variable | N | N* | Mean | SE Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|---|---|
| Y | 13 | 0 | 17.077 | 1.889 | 6.813 | 6.000 | 11.000 | 18.000 | 22.000 | 28.000 |
| X | 13 | 0 | 123.538 | 9.720 | 35.044 | 64.000 | 92.000 | 130.000 | 153.000 | 174.000 |

# Simple Regression: Y versus X

## Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 1 | 513.328 | 513.328 | 129.53 | <0.0001 |
| Error | 11 | 43.595 | 3.963 | | |
| Total | 12 | 556.923 | | | |

## Model Summary

| S | R-sq | R-sq(adj) |
|---|---|---|
| 1.99077 | 92.17% | 91.46% |

## Coefficients

| Term | Coef | SE Coef | T-Value | P-Value |
|---|---|---|---|---|
| Constant | -5.979 | 2.100 | -2.85 | 0.0159 |
| X | 0.18663 | 0.01640 | 11.38 | <0.0001 |

## Regression Equation

Y = –5.979 + 0.18663 X

## Fits and Diagnostics for Unusual Observations

| Obs | Y | Fit | Resid | Std Resid | |
|---|---|---|---|---|---|
| 12 | 28 | 23.8819 | 4.11813 | 2.27 | R |

*R Large residual*

The 'full' regression equation:

$\hat{Y}_v = b_0 + b_1 X_v$

$b_1 = \frac{r(s_y)}{s_x} = \frac{0.9600(6.813)}{35.044} = 0.1866$

$b_0 = \bar{y} - b_1\bar{x} = 17.077 - 0.1866(123.538) = -5.9797$

$\hat{Y} = -5.979 + 0.1866X$

**(c)**

$$R^2 = \frac{SSR}{TTS} = \frac{513.328}{556.923} = 0.921722 = \ 92.1722\%$$

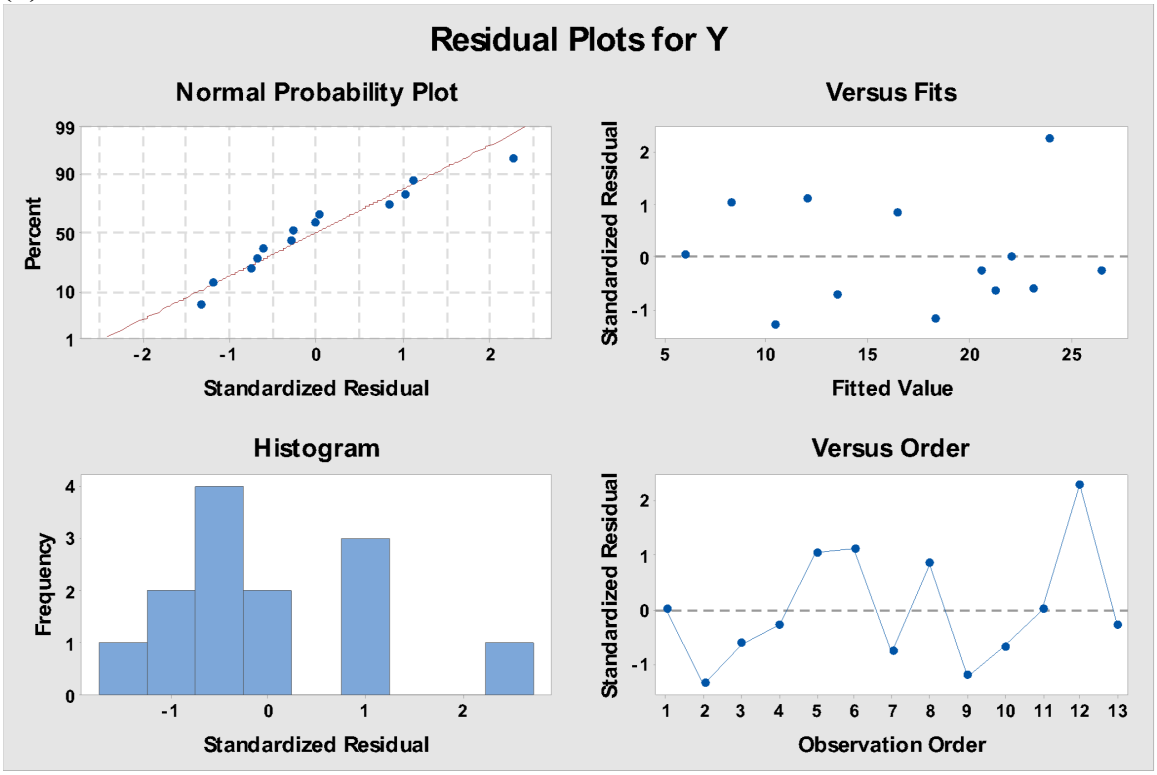
$$R^2(adj) = 1 - \frac{MSE}{MST} = 1 - \frac{3.963}{\left(\frac{556.923}{12}\right)} = 0.914609 = 91.4609\%$$

Based on the $R^2$ there is a 92.1722% variation in the 'Y' data explained by the regression. The $R^2$ is significant because it shows how close the data are to the fitted regression line, the higher the $R^2$ the better the model.

The relation between the correlation coefficient and the coefficient of determination is; the coefficient of determination ($R^2$) is the square of the correlation coefficient (r).

$r^2 = (0.9601)^2 = 0.921792 = R^2$

**(d)**




Based on the normality plot the data is reasonably normally distributed because they lie reasonably near or on the line. The versus fits shows no pattern as they are randomly placed. Therefore, the regression is significant

**(e)**
The unusual observation 12 was dropped to get the 'reduced' regression equation.

# Simple Regression: Y_1_1 versus X_1_1

## Method

Rows unused 1

## Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 1 | 404.434 | 404.434 | 174.08 | <0.0001 |
| Error | 10 | 23.233 | 2.323 | | |
| Total | 11 | 427.667 | | | |

## Model Summary

| S | R-sq | R-sq(adj) |
|---|---|---|
| 1.52422 | 94.57% | 94.02% |

## Coefficients

| Term | Coef | SE Coef | T-Value | P-Value |
|---|---|---|---|---|
| Constant | -4.849 | 1.652 | -2.93 | 0.0149 |
| X_1_1 | 0.17440 | 0.01322 | 13.19 | <0.0001 |

## Regression Equation

$Y\_1\_1 = -4.849 + 0.17440\ X\_1\_1$

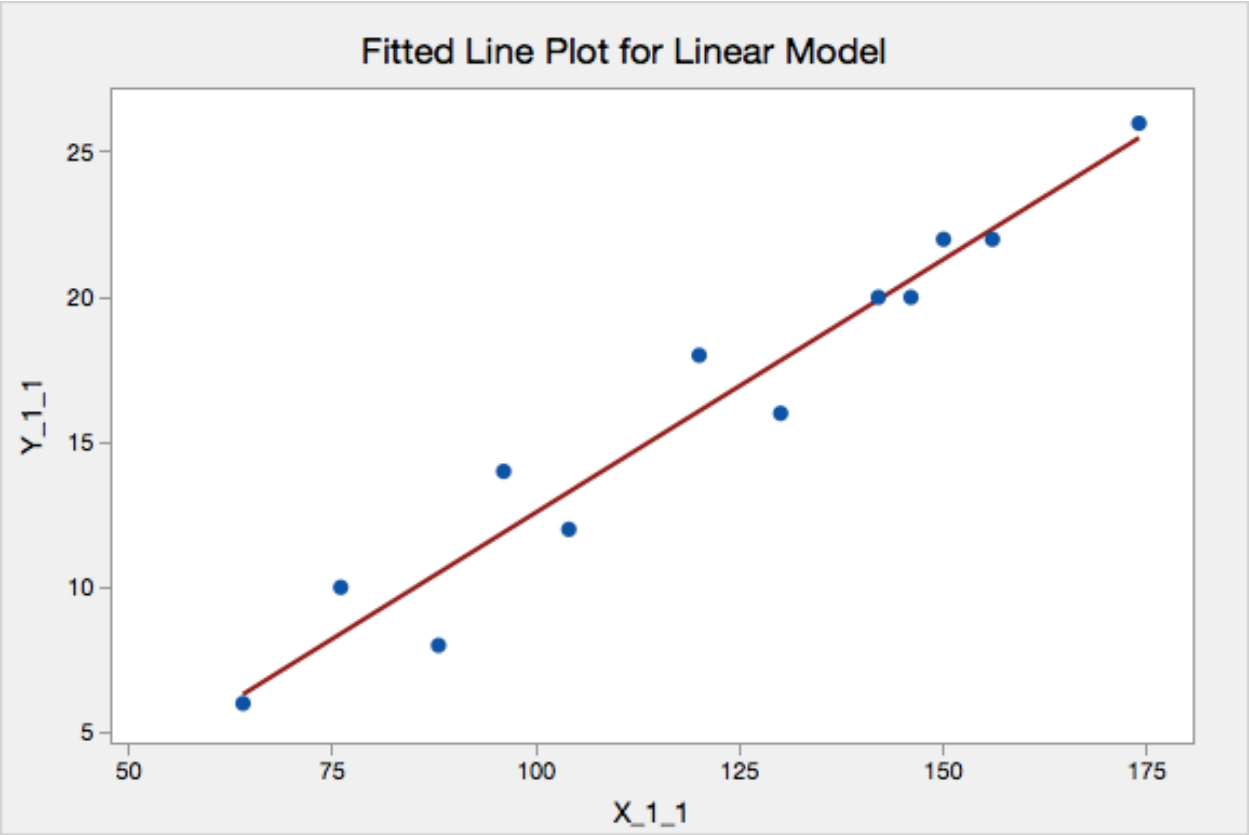
$The\ 'reduced' regression\ equation: \hat{Y} = -4.849\ + 0.1744X$

**Fitted Line: Y_1 versus X_1**

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{23.233}{427.667} = 0.945675 = \ 94.5675\%$$

$$R^2(adj) = 1 - \frac{MSE}{MST} = 1 - \frac{2.323}{\left(\frac{427.667}{11}\right)} = 0.940250 = \ 94.0250\%$$

Based on the $R^2$ there is a 94.5675% variation in the 'Y' data explained by the regression. The $R^2$ is significant because it shows how close the data are to the fitted regression line, the higher the $R^2$the better the model.




There is a difference in the quality of this 'reduced' regression equation and the 'full' regression equation because since the unusual observation was dropped the $R^2$ improved and the higher the $R^2$ the better the regression model. There the reduced regression equation is better than the full regression equation

**Question 2**

**(a)**
**LS=0.05**
**S1**
$H_0: \beta_1 = 0$
$H_a: \beta_1 \neq 0$

**S2**

$$SE(b_1) = \frac{s_e}{s_x\sqrt{n-1}} = \frac{1.9908}{35.044\sqrt{13-1}} = 0.0164$$

$$t_{Calc}(n-2) = \frac{b_1 - \beta_1}{SE(b_1)}$$

$$t_{Calc}(13-2) = \frac{0.1866 - 0}{0.0164} = 11.3780$$

**S3**

Inverse of the Cumulative Probability

| P(X ≤ x) | x |
|---|---|
| 0.975000 | 2.20099 |

$$t_{crit} = t_{\frac{\alpha}{2}}(n-2) = t_{\frac{0.05}{2}}(13-2) = t_{0.025}(11) = 2.2001$$

**S4**

Since $\{|t_{calc}| = 11.3780\} > \{t_{crit} = 2.2001\}$

⬇

Reject $H_0$

**S5**

$$\textbf{p} - \textbf{val} = P[t(df_E) > |t_{calc}|] \times \#of\ tails$$
$$= p[t_{0.025}(11) > |11.3780|] \times 2$$
$$= 0.0000$$

Since $\{p - val = 0.0000\} < \{\alpha = 0.05\}$

⬇

Reject $H_0$

Since the p-val is less than the level of significance, it is concluded that there is sufficient evidence to suggest that $\beta_1 \neq 0$, meaning X is significant.

**(b)**

**S1**

$H_0$: $Regression\ is\ not\ significant$
$H_a$: $regression\ is\ significant$

**S2**

$$MSR = \frac{SSR}{df_R} = \frac{513.328}{1} = 513.328$$
$$MSE = \frac{SSE}{df_E} = \frac{43.595}{11} = 3.9632$$
$$F_{calc} = \frac{MSR}{MSE} = \frac{513.328}{3.9632} = 129.5236$$

**S3**

Inverse of the Cumulative Probability

| P(X ≤ x) | x |
|---|---|
| 0.950000 | 4.84434 |

$F_{crit} = F_{0.05}(df_{num} = 1, df_{Denum} = 11)$
$F_{crit} = F_{0.05}\ (1, 11) = 4.8443$

**S4**

Since $\{F_{calc} = 129.5236\} > F_{crit} = 4.8443$

⬇

Reject $H_o$

**S5**

Since $\{p - val = 0.0000\} < \{\alpha = 0.05\}$

⬇

Reject $H_o$

Since the p-val is less than the level of significance we reject $H_o$. It can be concluded that there is sufficient evidence to show that the regression is significant.

**(c)**

Prediction for Y

Regression Equation

Y = −5.979 + 0.18663 X

Settings

| Variable | Setting |
|---|---|
| X | 170 |

Prediction

| Fit | SE Fit | 99% CI | 99% PI |
|---|---|---|---|
| 25.7482 | 0.9409 | (22.8258, 28.6706) | (18.9094, 32.5870) |

**99% confidence interval for the 'Percentage Market Share' if 'Monthly AD Expenses' is $170 in thousands**

Inverse of the Cumulative Probability

| P(X ≤ x) | x |
|---|---|
| 0.995000 | 2.78744 |

$t^*(n-2) = t^*(13-2) = 3.1058$
$\hat{y}_v = b_0 + b_1 x_v$
$Y = -5.979 + 0.1866X$
$\hat{y}_v = -5.9797 + 0.1866(170) = 25.743$

$$SE(\hat{\mu}_v) = \sqrt{SE^2(b_1)(x_v - \bar{x})^2 + \frac{s_e^2}{n}}$$

$$SE(\hat{\mu}_v) = \sqrt{0.0164^2(170 - 123.538)^2 + \frac{1.9908^2}{13}} = \sqrt{0.885477} = 0.9410$$

$\hat{y}_v \pm t^*(n-2)SE(\hat{\mu}_v)$
$25.743 \pm 3.1058(0.9410)$
$25.743 \pm 2.9226$
$(22.8204, 28.6656)$
Based on the 99% confidence interval, with 99% confidence we can say the expected percentage market share or mean is between $(22.8204, 28.6656)$.

**(d)**

**99% Prediction interval for the 'Percentage Market Share' if 'Monthly AD Expenses' is $170 in thousands**

$\hat{y}_v \pm t^*(n-2)SE(\hat{y}_v)$

$SE(\hat{y}_v) = s_e^2 + SE^2(\hat{\mu}_v)$
$SE^2(\hat{y}_v) = 1.9908^2 + 0.9410^2 = 4.8488$
$SE(\hat{y}_v) = \sqrt{4.8488} = 2.2020$

$\hat{y}_v \pm t^*(n-2)SE(\hat{y}_v)$
$25.743 \pm 3.1058(2.2020)$
$25.743 \pm 6.8390$
$(18.904, 32.5820)$

Based on the 99% Predicted interval, with 99% confidence we can say the expected percentage market share or mean is between $(18.904, 32.5820)$

**(e)**

The prediction interval is broader than the confidence interval because a confidence interval is an interval of a parameter which is constant, and it is meant to find the confidence interval of the mean or expected predicted value; while a prediction interval is meant to find the confidence interval of the single or next predicted value.

**Question 3**

**(a)**

## Simple Regression: SPrice versus SqFt

### Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 1 | 1048996 | 1048996 | 123.97 | <0.0001 |
| Error | 27 | 228463 | 8462 | | |
| Total | 28 | 1277459 | | | |

### Model Summary

| S | R-sq | R-sq(adj) |
|---|---|---|
| 91.9869 | 82.12% | 81.45% |

### Coefficients

| Term | Coef | SE Coef | T-Value | P-Value |
|---|---|---|---|---|
| Constant | 38.30 | 63.52 | 0.60 | 0.5515 |
| SqFt | 23.349 | 2.097 | 11.13 | <0.0001 |

### Regression Equation

SPrice = 38.30 + 23.349 SqFt

### Fits and Diagnostics for Unusual Observations

| Obs | SPrice | Fit | Resid | Std Resid | | |
|---|---|---|---|---|---|---|
| 2 | 157 | 411.89 | -254.885 | -2.96 | R | |
| 29 | 1345 | 1205.75 | 139.248 | 1.76 | | X |

*R Large residual*
*X Unusual X*

# Simple Regression: SPrice versus NumFlrs

## Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 1 | 133319 | 133319 | 3.15 | 0.0874 |
| Error | 27 | 1144140 | 42376 | | |
| Total | 28 | 1277459 | | | |

## Model Summary

| S | R-sq | R-sq(adj) |
|---|---|---|
| 205.853 | 10.44% | 7.12% |

## Coefficients

| Term | Coef | SE Coef | T-Value | P-Value |
|---|---|---|---|---|
| Constant | 526.7 | 115.2 | 4.57 | <0.0001 |
| NumFlrs | 139.74 | 78.78 | 1.77 | 0.0874 |

## Regression Equation

SPrice = 526.7 + 139.74 NumFlrs

## Fits and Diagnostics for Unusual Observations

| Obs | SPrice | Fit | Resid | Std Resid | |
|---|---|---|---|---|---|
| 2 | 157 | 666.444 | -509.444 | -2.55 | R |
| 28 | 1099.5 | 666.444 | 433.056 | 2.16 | R |
| 29 | 1345 | 806.182 | 538.818 | 2.75 | R |

*R Large residual*

# Simple Regression: SPrice versus Baths

## Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 1 | 638728 | 638728 | 27.00 | <0.0001 |
| Error | 27 | 638731 | 23657 | | |
| Total | 28 | 1277459 | | | |

## Model Summary

| S | R-sq | R-sq(adj) |
|---|---|---|
| 153.807 | 50.00% | 48.15% |

## Coefficients

| Term | Coef | SE Coef | T-Value | P-Value |
|---|---|---|---|---|
| Constant | -10.7 | 143.4 | -0.07 | 0.9413 |
| Baths | 223.94 | 43.10 | 5.20 | <0.0001 |

## Regression Equation

SPrice = −10.7 + 223.94 Baths

## Fits and Diagnostics for Unusual Observations

| Obs | SPrice | Fit | Resid | Std Resid | |
|---|---|---|---|---|---|
| 1 | 345 | 325.242 | 19.758 | 0.15 | X |
| 29 | 1345 | 997.053 | 347.947 | 2.46 | R |

*R Large residual*
*X Unusual X*

Simple Linear Regression Equations:

Selling Price Vs Area in Square Feet*100

$\hat{y}_v = b_0 + b_1 x_v$

$\hat{Y}_i = 38.30 + 23.349 X_i$

$SPrice = 38.30 + 23.349\ SqFt$

Selling Price Vs Number of Floors

$\hat{y}_v = b_0 + b_1 x_v$

$\hat{Y}_i = 526.7 + 139.74 X_i$

$SPrice = 526.7 + 139.74\ NumFlrs$

Selling Price VS Number of Bedroom

$\hat{y}_v = b_0 + b_1 x_v$

$\hat{Y}_i = -14.0 + 113.14 X_i$

$SPrice = -14.0 + 113.14\ BedRms$

Selling Price VS Number of Bathroom

$\hat{y}_v = b_0 + b_1 x_v$

$\hat{Y}_i = -10.7 + 223.94 X_i$

$SPrice = -10.7 + 223.94 Baths$

Rank of the four Simple Linear Regression equations in 'Descending Order' using the $R^2(adj)$ criteria

1. Selling Price VS Area in Square Feet*100 ($R^2(adj) = 81.45\%$)
2. Selling Price VS Number of Bathroom($R^2(adj) = 48.15\%$)
3. Selling Price VS Number of Bedroom($R^2(adj) = 43.31\%$)
4. Selling Price VS Number of Floors($R^2(adj) = 7.12\%$)

**(b)**

## Multiple Regression: SPrice versus SqFt, NumFlrs, BdRms, Baths

### Regression Equation

SPrice = −65.25 + 20.055 SqFt − 141.22 NumFlrs − 13.89 BdRms + 148.60 Baths

### Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 4 | 1119011 | 279753 | 42.37 | <0.0001 |
| Error | 24 | 158448 | 6602 | | |
| Total | 28 | 1277459 | | | |

### Model Summary

| S | R-sq | R-sq(adj) |
|---|---|---|
| 81.2526 | 87.60% | 85.53% |

### Coefficients

| Term | Coef | SE Coef | 95% CI | T-Value | P-Value |
|---|---|---|---|---|---|
| Constant | -65.25 | 94.19 | (-259.65, 129.15) | -0.69 | 0.4951 |
| SqFt | 20.055 | 3.034 | (13.792, 26.318) | 6.61 | <0.0001 |
| NumFlrs | -141.22 | 56.16 | (-257.13, -25.31) | -2.51 | 0.0190 |
| BdRms | -13.89 | 19.96 | (-55.08, 27.30) | -0.70 | 0.4931 |
| Baths | 148.60 | 47.68 | (50.20, 247.01) | 3.12 | 0.0047 |

### Fits and Diagnostics for Unusual Observations

| Obs | SPrice | Fit | Resid | Std Resid | |
|---|---|---|---|---|---|
| 1 | 345 | 201.523 | 143.477 | 2.26 | R |
| 2 | 157 | 356.044 | -199.044 | -2.76 | R |

*R Large residual*

**The Multiple Linear Regression Equation for all the independent/explanatory variables.**

$Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i}$

$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i} + b_4 X_{4i}$

$\hat{Y}_i = -65.25 + 20.055 X_{1i} - 141.22 X_{2i} - 13.89 X_{3i} + 148.60 X_{4i}$

$SPrice = -65.25 + 20.055\ SqFt - 141.22\ NumFlrs - 13.89\ BdRms + 148.60\ Baths$

**(c)**

$$R^2 = \frac{SSR}{SST} = \frac{1119011}{1277459} = 0.875966 = \ 87.5966\%$$

$$R^2(adj) = 1 - \frac{MSE}{MST} = 1 - \frac{6602}{\left(\frac{1277459}{28}\right)} = 0.855294\ 0r\ 85.5294\%$$

The Quality of the regression is good because the of the model is p-val is small and the $R^2$ is significant because there is 87.5966% variation in the response data explained by the regression.

**(d)**

The numerical criterion used is the P-vals using level of significance of 5% sqft ($\beta_1$)and bathrooms ($\beta_4$) and Number of floors ($\beta_2$) are significant because their p-val is less than the level of significance unlike bedrooms ($\beta_3$) and the constant ($\beta_0$) which is not significant because its p-val is higher than the level of significance.

Coefficients

| Term | Coef | SE Coef | 95% CI | T-Value | P-Value |
|---|---|---|---|---|---|
| Constant | -65.25 | 94.19 | (-259.65, 129.15) | -0.69 | 0.4951 |
| SqFt | 20.055 | 3.034 | (13.792, 26.318) | 6.61 | <0.0001 |
| NumFlrs | -141.22 | 56.16 | (-257.13, -25.31) | -2.51 | 0.0190 |
| BdRms | -13.89 | 19.96 | (-55.08, 27.30) | -0.70 | 0.4931 |
| Baths | 148.60 | 47.68 | (50.20, 247.01) | 3.12 | 0.0047 |

**(e)**
The worst performing independent variable would be Bedrooms.

## Regression Analysis: SPrice versus SqFt, NumFlrs, Baths

```
Analysis of Variance

Source         DF   Adj SS  Adj MS  F-Value  P-Value
Regression      3  1115813  371938    57.52    0.000
  SqFt          1   365714  365714    56.56    0.000
  NumFlrs       1    42858   42858     6.63    0.016
  Baths         1    65116   65116    10.07    0.004
Error          25   161646    6466
  Lack-of-Fit  16   137032    8564     3.13    0.044
  Pure Error    9    24615    2735
Total          28  1277459


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
80.4105  87.35%     85.83%      80.38%


Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant  -102.2     77.0    -1.33    0.197
SqFt       18.90     2.51     7.52    0.000  1.88
NumFlrs   -119.9     46.6    -2.57    0.016  2.29
Baths      133.6     42.1     3.17    0.004  3.49


Regression Equation

SPrice = -102.2 + 18.90 SqFt - 119.9 NumFlrs + 133.6 Baths


Fits and Diagnostics for Unusual Observations

Obs  SPrice     Fit   Resid  Std Resid
  1   345.0   205.2   139.8       2.22  R
  2   157.0   347.6  -190.6      -2.64  R
 29  1345.0  1204.3   140.7       2.04  R

R  Large residual
```

The new best MLR equations would be:

$Y = \beta_0 + \beta_1 + \beta_2 + \beta_3$

$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i}$

$\hat{Y}_i = -102.20 + 18.90 X_{1i} - 119.90 X_{2i} + 133.60 X_{3i}$

$SPrice = -102.20 + 18.90\ SqFt - 119.90\ NumFlrs + 133.60\ Baths$

**(f)**

$$R^2 = \frac{SSR}{SST} = \frac{1115813}{1277459} = 0.873463 = \ 87.3463\%$$

$$R^2(adj) = 1 - \frac{MSE}{MST} = 1 - \frac{6466}{\left(\frac{1277459}{28}\right)} = 0.858275\ 0r\ 85.8275\%$$

The Quality of the regression is good because the of the model is p-val is small and the $R^2$ is significant because there is $87.3463\%$ variation in the response data explained by the regression
Best Subset:

Best Subsets Regression: SPrice versus SqFt, NumFlrs, BdRms, Baths

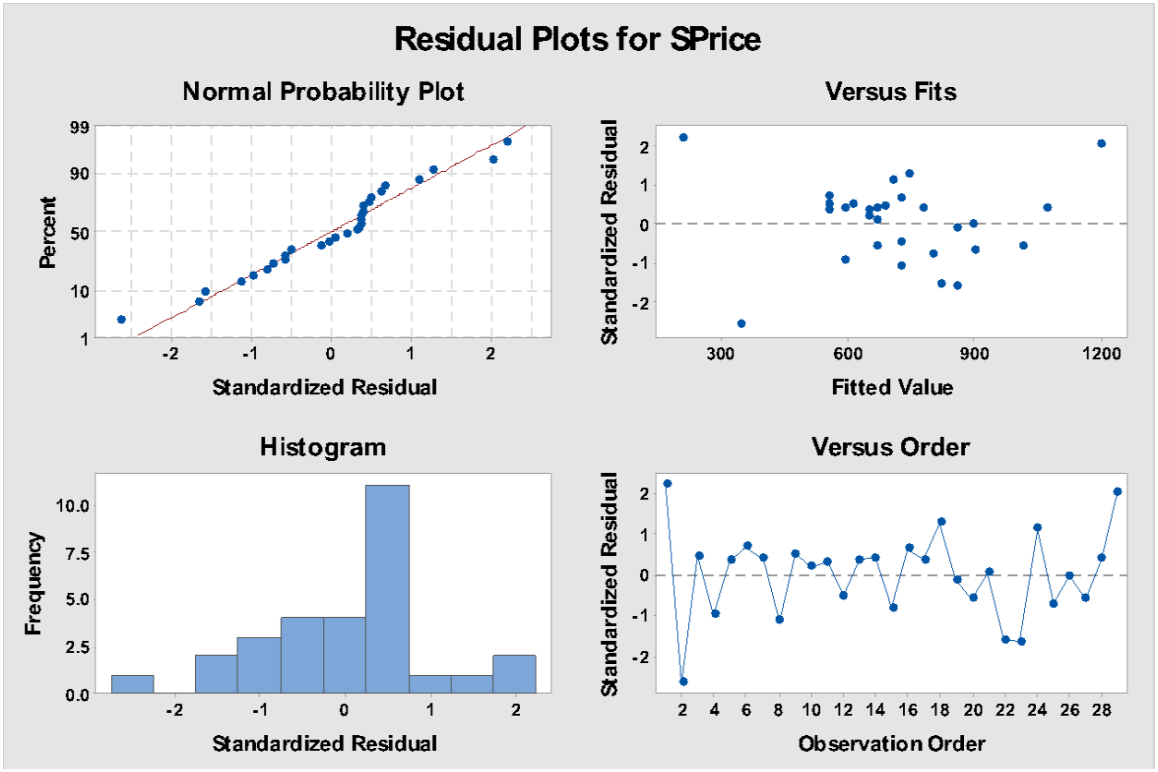
Model Summary

| Number of Predictors | R-sq | R-sq(adj) | R-sq(pred) | Mallows' Cp | S | SqFt | NumFlrs | BdRms | Baths |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 82.1 | 81.5 | 77.8 | 9.6 | 91.99 | X | | | |
| 1 | 50.0 | 48.1 | 39.6 | 71.7 | 153.8 | | | | X |
| 2 | 84.0 | 82.8 | 78.8 | 8.0 | 88.69 | X | | | X |
| 2 | 82.6 | 81.2 | 76.2 | 10.8 | 92.58 | X | | X | |
| 3 | 87.3 | 85.8 | 80.4 | 3.5 | 80.41 | X | X | | X |
| 3 | 84.3 | 82.4 | 77.2 | 9.3 | 89.49 | X | | X | X |
| 4 | 87.6 | 85.5 | 79.5 | 5.0 | 81.25 | X | X | X | X |

In this case the 'Best' Regression equation when the worst performing independent variable, bedrooms, is drop is the best because it has the highest $R^2(adj)\ which\ is\ 85.8\%$ which is the better regression equation. It also has the lowest $s_e$ which makes it the best regression equation. Therefore, the SPrice and sqft, NumFlrs, and Baths is the indeed the 'Best' multiple regression equation.

**(g)**

In this case the 3 variable Regression is the best because it has the higher




Based on the normality plot the data is reasonably normally distributed because they lie reasonably near or on the line, based on the versus fits all the points are reasonably within $\pm 2$ and the points seem to be randomly scattered on the plot and based on the versus order graph there is no clear pattern indicating reasonably random residuals and independent.

**(h)**

Based on the $R^2(adj)$ the 'Best' MLR equation has a higher $R^2(adj)$ than the 'Full' MLR equation making it the better regression model. The Best MLR also has a lowers standard deviation than the Full MLR equation, making it the better model The 'Best' MLR equation is better than the 'Full' MLR equation with 4 independent variables because the Correlation coefficient is higher in the best MLR equation.

**(i)**

## Regression Analysis: SPrice_1 versus SqFt_1, NumFlrs_1, BdRms_1, Baths_1

```
Method
Rows unused  1
Analysis of Variance

Source         DF   Adj SS  Adj MS  F-Value  P-Value
Regression      4  1007419  251855    46.41    0.000
  SqFt_1        1   300163  300163    55.31    0.000
  NumFlrs_1     1    65682   65682    12.10    0.002
  BdRms_1       1     5096    5096     0.94    0.343
  Baths_1       1    96110   96110    17.71    0.000
Error          23   124821    5427
  Lack-of-Fit  17   108415    6377     2.33    0.150
  Pure Error    6    16405    2734
Total          27  1132240

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
73.6681  88.98%     87.06%      81.88%


Coefficients

Term         Coef  SE Coef  T-Value  P-Value   VIF
Constant   -185.8     98.2    -1.89    0.071
SqFt_1      20.50     2.76     7.44    0.000  2.27
NumFlrs_1  -189.6     54.5    -3.48    0.002  3.66
BdRms_1     -17.6     18.2    -0.97    0.343  2.36
Baths_1     206.9     49.2     4.21    0.000  4.24


Regression Equation

SPrice_1 = -185.8 + 20.50 SqFt_1 - 189.6 NumFlrs_1 - 17.6 BdRms_1 + 206.9 Baths_1


Fits and Diagnostics for Unusual Observations

Obs  SPrice_1    Fit   Resid  Std Resid
  2     157.0  296.1  -139.1      -2.29  R

R  Large residual
```

## Residual Plots for SPrice_1

Observation 1 was dropped.
The new equation would be:
$\hat{Y}_i = -185.8 + 20.50X_{1i} - 189.6X_{2i} - 17.6X_{3i} + 206.9X_{4i}$
$SPrice = -185.8 + 20.50SqFt - 189.6NumFlrs - 17.6BdRms + 206.9\ Baths$

The new equation is better than the 'full' MLR equation because the $R^2(adj)$ of the new equation is the higher than the full MLR equation making it the better regression model.

**Question 4**

$\hat{Y}_i = -102.20 + 18.90X_{1i} - 119.90X_{2i} + 133.60X_{3i}$

**(a)**

**LS=0.05**

**S1**

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

**S2**

$SE(b_1) = 2.513$

$n = 29, k = 28$

$$t_{Calc}(n - (k+1)) = \frac{b_1 - \beta_1}{SE(b_1)}$$

$$t_{Calc}(29 - (3+1)) = t_{Calc}(25) = \frac{18.90 - 0}{2.513} = 7.5201$$

**S3**

Inverse of the Cumulative Probability

| P(X ≤ x) | x |
|---|---|
| 0.975000 | 2.05954 |

$t_{crit} = t_{\frac{\alpha}{2}}(25) = t_{\frac{0.05}{2}}(25) = t_{0.025}(25) = 2.0595$

**S4**

Since $\{|t_{calc}| = 7.5201\} > \{t_{crit} = 2.0595\}$

⬇

Reject $H_0$

**S5**

$\mathbf{p - val} = p[t(df) > |t_{calc}|] \times \#of\ tails$

$= p[t_{0.025}(25) > |7.5201|] \times 2$

$= 0.0000$

Since $\{p - val = 0.0000\} < \{\alpha = 0.05\}$

⬇

Reject $H_0$

Since the p-val is less than the level of significance, it is concluded that there is sufficient evidence to suggest that $\beta_1 \neq 0$. There is a relationship between selling price and SqFt

**(b)**

**S1**

$H_0: Regression\ is\ not\ significant$

$H_a: regression\ is\ significant$

**S2**

$$MSR = \frac{SSR}{df_R} = \frac{1115813}{3} = 371937.6667$$

$$MSE = \frac{SSE}{df_E} = \frac{161646}{25} = 6465.84$$

$$F_{calc}(k, n-(k+1) = F_{calc}(3,25) = \frac{MSR}{MSE} = \frac{38604.3333}{6465.84} = 57.5235$$

**S3**

Inverse of the Cumulative Probability

| P(X ≤ x) | x |
|---|---|
| 0.950000 | 2.99124 |

$F_{crit} = F_{0.05}(df_{num} = 3, df_{Denum} = 25)$

$F_{crit} = F_{0.05}\ (3, 25) = 2.9912$

**S4**

Since $\{F_{calc} = 57.5235\} > F_{crit} = 2.9912$

⬇

Reject $H_o$

**S5**

Since $\{p - val = 0.0000\} < \{\alpha = 0.05\}$

⬇

Reject $H_o$

Since the p-val is less than the level of significance we reject $H_o$. It can be concluded that there is sufficient evidence to show that the regression is significant

**(c)**

**99% Confidence interval**

**40 Square feet, 2 floors, 5 bedrooms, 3 bathrooms**
**Prediction for SPrice**

```
Regression Equation

SPrice = -102.2 + 18.90 SqFt - 119.9 NumFlrs + 133.6 Baths


Variable  Setting
SqFt           40
NumFlrs         2
Baths           3


    Fit   SE Fit        99% CI              99% PI
814.877  55.0857  (661.329, 968.425)  (543.187, 1086.57)  X

X denotes an unusual point relative to predictor levels used to fit the model.
```

$Y = \beta_0 + \beta_1 + \beta_2 + \beta_3 +$
$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i}$
$\hat{Y}_i = -102.20 + 18.90X_{1i} - 119.90X_{2i} + 133.60X_{3i}$
$SPrice = -102.20 + 18.90SqFt - 119.90\ NumFlrs + 133.60\ Baths$
$SPrice = -102.20 + 18.90(40) - 119.90(2) + 133.60(3)$
$SPrice = -102.20 + 756 - 239.80 + 400.80 = 814.80$
$SE(\hat{\mu}_v) = 55.09$
$t_{\frac{\alpha}{2}}(df_E) = t_{\frac{0.01}{2}}(25) = 3.078$
$\hat{y}_v \pm t_{\frac{\alpha}{2}}(df_E)SE(\hat{\mu}_v)$
$\hat{y}_v \pm t_{\frac{0.01}{2}}(25)SE(\hat{\mu}_v)$
$814.80 \pm (2.7874)(55.09)$
$814.80 \pm 153.5579$
$(661.2421, 968.3579)$

Based on the 99% confidence interval, with 99% confidence we can say the expected selling price or mean is between $(661.2421, 968.3579)$.

**(d)**
**99% Prediction Interval**
$\hat{y}_v \pm t^*(n-2)SE(\hat{y}_v)$

$SE(\hat{y}_v) = s_e^2 + SE^2(\hat{\mu}_v)$
$SE^2(\hat{y}_v) = 80.4105^2 + 55.09^2 = 9500.7566$
$SE(\hat{y}_v) = \sqrt{9500.7566} = 97.4718$

$\hat{y}_v \pm t^*(n-2)SE(\hat{y}_v)$
$814.80 \pm (2.7874)(97.4718)$
$814.80 \pm 271.6930$
$814.80 \pm 271.6930$
$(543.107, 1086.493)$

Based on the 99% Predicted interval, with 99% confidence we can say the predicted value selling price or mean is between $(543.107, 1086.493)$.

**(e)**

$VIF = \frac{1}{(1-R_j^2)}$,

This number is calculated by using X1 as the response variable regressed on X2 and X3, where you will be given the $R^2$ of X1 which will then be used in the VIF formula to get VIF. The significance of this VIF number is since it is less than 5 and greater than 1 the multicollinearity may exist.

Model Summary

| S | R-sq | R-sq(adj) |
|---|---|---|
| 6.27560 | 46.78% | 42.69% |

$R^2 = 46.78\%$

$$VIF = \frac{1}{(1-R^2)} = 1.879$$

$1 = 1.879(1-R^2)$
$1 = 1.879 - 1.879R^2$
$1 - 1.879 = -1.879R^2$
$-0.879 = -1.879R^2$

$$\frac{-0.879}{-1.879} = R^2$$

$R^2 = 46.7802\%$

$$VIF = \frac{1}{(1-0.4678)} = 1.879$$

**(f)**

They would be indicator variables which are qualitative attributes. These categories should be coded as binary variables which can only be either '1' or '0'.
For example, in the case:
$X_i = Indicator\ Variables\ (binary), of\ 'Poor'$
$X_{1i}$ ='1' if poor,'0' if otherwise
$X_{2i}$ ='1' if middle class, '0' if otherwise
$X_{3i}$ ='1' if Upscale, '0' if otherwise