# STAT 501 – Mid-Term Exam 2 Fall 2015

**Instructions**: Use Word to type your answers within this document. Then, submit your answers in the appropriate dropbox in ANGEL by the due date and **within 3 hours of downloading the exam**. The point distribution is located next to each question.

---

1. (**6x2 = 12 points**) State which of the following statements is TRUE and which is FALSE. For the statements that are false, explain why they are false.

   a) The sum of leverages adds to p, the # of regression coefficients in the model (including the intercept). **TRUE.**
   b) Removing an outlier in a regression analysis will result in narrower confidence intervals. **TRUE.**
   c) Since leverages depend only on the predictors, removal of an outlier (unusual Y value) has no impact on the leverages. **FALSE. Removal of an outlier can change the leverage values since removal of an outlier leads to removing one row in the design matrix.**
   d) In a simple linear regression (SLR) model, if a log transformation is performed on X to remedy some non-linearity, the mean value of Y is bound to change. **FALSE. Y is a variable observed independently of X and hence its mean value will not be affected by X at all.**
   e) In model selection, the highest adjusted $R^2$-value and the smallest S-value criteria always yield the same "best" models. **TRUE.**
   f) Regression models with different responses, but the same predictor X matrix, will have the same leverage values. **TRUE.**

2. (**3+3+4+4+3+3 = 20 points**) ) Open the "Math Scores" dataset. The dataset consists of Math scores (Math) for 14 students with information about their undergraduate major(Major) and weekly hours of study (Hours). Your goal is to fit a regression model to express the dependence of Y (Math) on X (Hours) and Major.

   a) Clearly define a set of indicator variables that could be used in a regression model to represent the qualitative variable Major. *[Hint: Think carefully about the number of indicator variables needed given the number of levels of Degree and use "Engineering" as the reference level.]*

      **Let T1 = 1 if the major is Science and 0 otherwise, T2 = 1 if the major is History and 0 otherwise.**

   b) Write a population multiple linear regression equation for predicting the math scores in terms of Hours and Major. Since one's major could impact the dependence of Y on X, the model should contain an interaction effect between Hours and Major, together with their main effects. *[Hint: Your equation should include Y, X, the indicator variables you defined in part (a), interaction terms, and population regression coefficients (β's).]*

      **$E(Y) = \beta_0 + \beta_1 X + \beta_2 T1 + \beta_3 T2 + \beta_4 X*T1 + \beta_5 X*T2$.**

c) Conduct a hypothesis test at significance level 0.05 to determine if the average math score increase due to one hour of extra study per week differs by major (i.e., test if the slopes for two or more Major categories differ). Write out the null and alternative hypotheses, the test statistic, the p-value, and the conclusion. *[Minitab v17: Select Math as the Response, Hours as the Continuous predictor, Major as the categorical predictor, click "Model," select both Hours and Major together in the Predictors box and click the Add button next to "Interactions through order 2." Minitab v16: Create interaction terms using Calc > Calculator before fitting the regression model. Consder p < .05 to be statistically significant ]*

***$H_0$: $\beta_4 = \beta_5 = 0$ vs Ha: at least one of $\beta_4, \beta_5 \neq 0$***

Analysis of Variance

| Source | DF | Adj SS | Adj MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 5 | 8419.22 | 1683.84 | 72.57 | 0.000 |
| Hours | 1 | 2310.40 | 2310.40 | 99.57 | 0.000 |
| Major | 2 | 238.43 | 119.22 | 5.14 | 0.037 |
| Hours*Major | 2 | 0.22 | 0.11 | 0.00 | 0.995 |
| Error | 8 | 185.63 | 23.20 | | |
| Lack-of-Fit | 7 | 185.13 | 26.45 | 52.90 | 0.105 |
| Pure Error | 1 | 0.50 | 0.50 | | |
| Total | 13 | 8604.86 | | | |

***Since F-stat=0.00 with p=0.995 > 0.05 we fail to reject $H_0$ and conclude that the interaction effect is not statistically significant. Therefore, the interaction effect terms can be dropped from the model.***

d) Write a new population regression equation based on your conclusion to part (c). Fit this model and conduct two separate hypothesis tests for whether the mean math score for a fixed number of weekly study hours differs by major. For each test, write out the null and alternative hypotheses, the test statistic, the p-value, and the conclusion.

***The newly proposed population regression model is:***

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 T1 + \beta_3 T2$$

***$H_0$: $\beta_2 = 0$ vs 0 Ha: $\beta_2 \neq 0$***

Coefficients

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | 25.05 | 3.19 | 7.86 | 0.000 | |
| Hours | 15.115 | 0.845 | 17.88 | 0.000 | 1.11 |
| Major | | | | | |
| History | -16.78 | 3.01 | -5.57 | 0.000 | 1.40 |
| Science | -22.40 | 2.73 | -8.22 | 0.000 | 1.29 |

**Since t-stat=-8.22 with p=0.000 < 0.05 we reject H0 and conclude that $\beta_2$ is non-zero, which indicates that the average Math scores of Science and Engineering majors differ .**

**Similarly, for *$H_0$: $\beta_3 = 0$ vs 0 Ha: $\beta_3 \neq 0$, we reject H0 since t-stat=-5.57 with p=0.000 < 0.05 and conclude that the* average Math scores of History and Engineering majors differ.**

e) Based on your conclusion to part (d), write three fitted sample regression equations that can be used to predict the Math scores for each major. *[Hint: Your equations should include number values, not β's.]*

```
Regression Equation

Major
Engineering  Math = 25.05 + 15.115 Hours

History      Math = 8.27 + 15.115 Hours

Science      Math = 2.65 + 15.115 Hours
```

f) Based on one of the equations from part (e), predict the math score of a History major who studies hours per week. *[Hint: A point estimate is sufficient.]*

**History Math = 8.27 + 15.115(4) = 69.16**

3. (**3+2+3+2 = 10 points**) The following Minitab output resulted from a multiple linear regression model fit to response variable, *Y*, and predictor terms, $X_1$, $X_2$, *and* $X_1X_2$:

```
Coefficients

Term        Coef  SE Coef  T-Value  P-Value
Constant    4.49     1.89     2.37    0.022
X1         0.759    0.374     2.03    0.048
X2         0.965    0.426     2.26    0.028
X1*X2     0.1742   0.0821     2.12    0.039
```

a) Conduct a hypothesis test for whether the interaction term, $X_1X_2$, can be dropped from the model. Write out the population model, null and alternative hypotheses, the test statistic, the p-value, and the conclusion.

***$E(Y) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3 X_1*X_2$***
***H0: $\beta_3 = 0$ vs Ha: $\beta_3 \neq 0$***
***The t-stat=2.12 with p-value=0.039 < 0.05 allows us to reject H0 and conclude that the interaction effect is statistically significant. Therefore, this term should not be dropped from the model.***

b) Based on your conclusion to part (b), write the fitted sample regression equation.

$\hat{Y} = 4.49 + 0.759\, X_1 + 0.965\, X_2 + 0.1742\, X_1 * X_2$

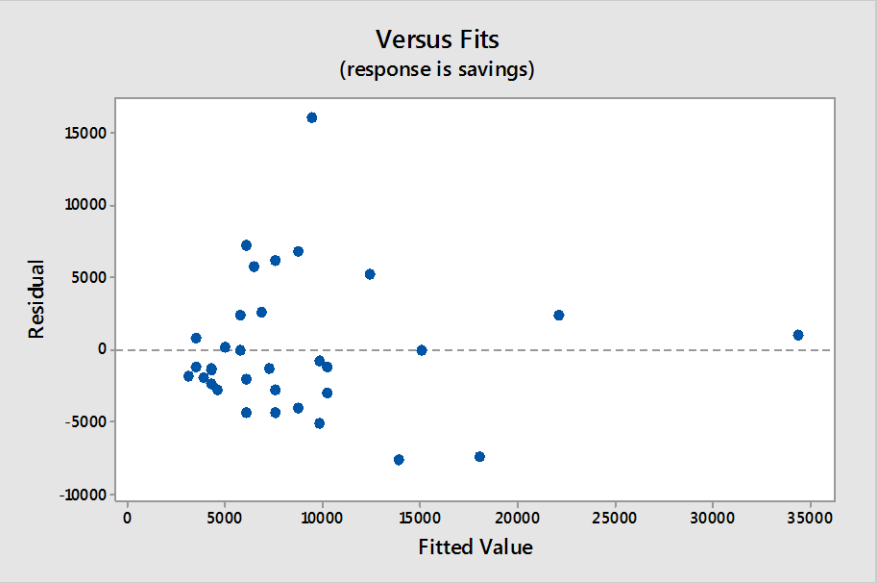
c) State whether the following statements are supported by the Minitab output. (simply write "yes" or "no" for each statement).
   i. $X_1$ and $X_2$ are positively associated. **No.**
   ii. $Y$ and $X_1$ are positively associated for fixed values of $X_2$ between 0 and 10. **Yes.**
   iii. The linear association between $Y$ and $X_1$ increases as $X_2$ increases. **Yes.**

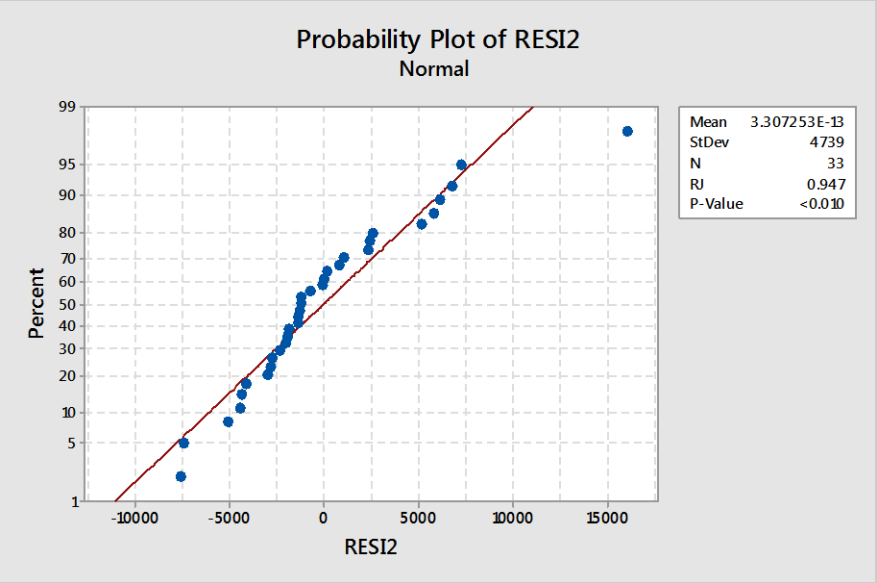
d) Use the fitted equation in part (b) to predict $Y$ for an observation with $X_1 = 6$ and $X_2 = 5$. *[Hint: A point estimate is sufficient.]*

$\hat{Y} = 4.49 + 0.759\,(6) + 0.965\,(5) + 0.1742\,(6)(5) = 19.095.$

4. (**6+2+6+3+3 = 20 points**) The dataset "Savings" contains savings of 33 individuals along with their age. It is apparent that Y = Savings (in $) has a positive association with X = Age (in years). An appropriate regression model relating Savings to Age could be useful for predicting savings based on age. The most straightforward approach would be to fit a simple linear regression (SLR) model for Y vs X, provided that the LINE assumptions are satisfied. *[Consult "Worked Examples Using Minitab" in the Online Notes for help with any Minitab procedures.]*

a) Fit an SLR model for Y vs X and perform a residual plot analysis to determine if the LINE assumptions are satisfied. Include a numerical test when checking for normality (use the Ryan Joiner test in Minitab). Discuss your findings and include any relevant graphs.



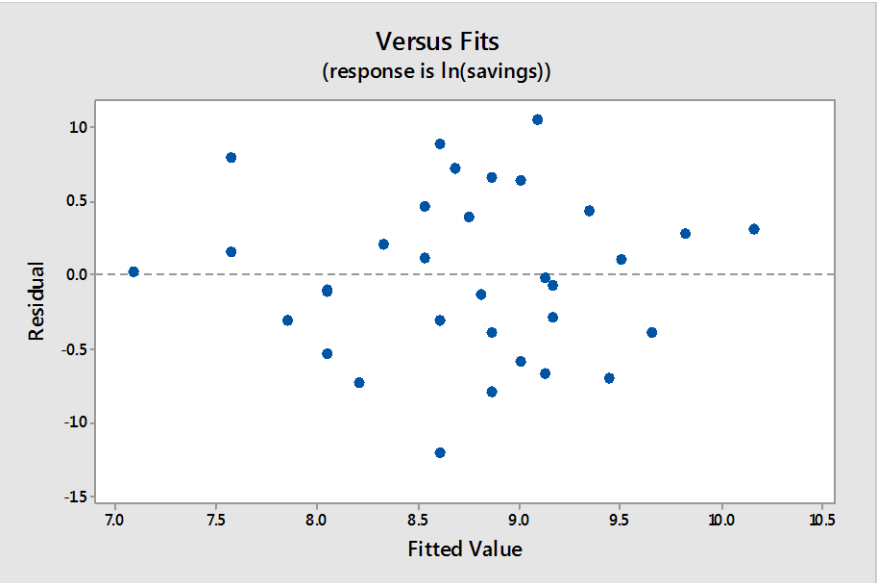


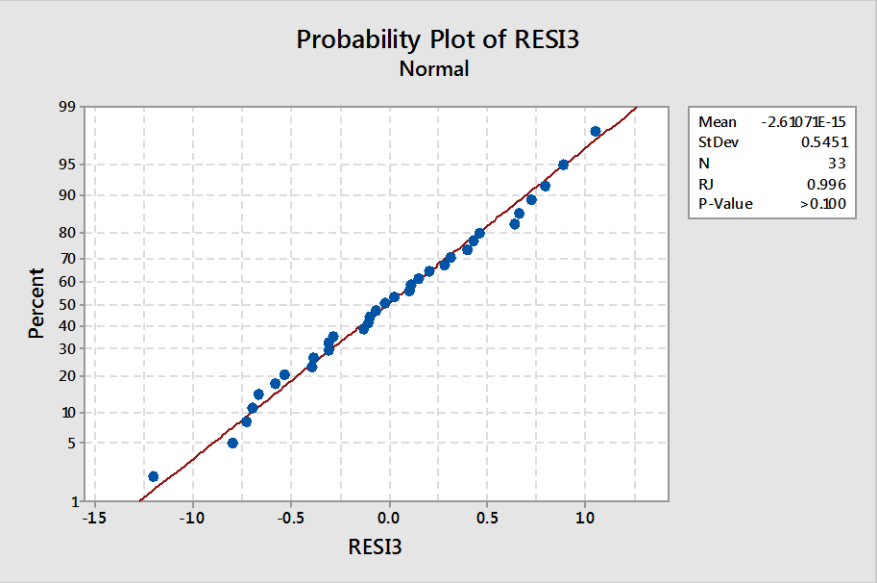

***RJ test has a p-value < 0.01 indicates that residuals are not normal. Also, residuals vs predicted plot indicates unequal variances and possible curvature with at least one outlier.***

b) Based on your conclusion in part (a), determine if any transformations are suggested for X and/or Y. *[Hint: You should find that both X and Y need to be transformed.]*

***Given the manner in which the LINE conditions are not satisfied, a log transformation on both X and Y may be suitable.***

c) Fit an SLR model for the transformed variable(s) and comment on this model's validity with supporting statements, numerical tests and/or plots.






***The RJ test now has a p-value > 0.1 suggesting normality. Also, residuals vs fits has a more random pattern around zero except possibly for the outliers.***

d) Use Minitab to compute a 95% confidence interval for the mean amount of savings (in $) expected for 40 year-olds based on the fitted model in part (c). *[Hint: Remember to take into account the transformations to X and Y.]*

```
Variable  Setting
ln(age)    3.6889

    Fit     SE Fit         95% CI              95% PI
9.64044  0.162996  (9.30800, 9.97287)  (8.46301, 10.8179)
```

***Note: ln(40) = 3.6889***

***A 95% CI for mean ln(savings) of the 40 year olds = (9.30800, 9.97287). Therefore, a 95% CI for means savings is ($11026, $21437).***

e) Use Minitab to compute a 95% prediction interval for the amount of savings (in $) predicted for a randomly selected 40 year-old based on the fitted model in part (c). *[Hint: Remember to take into account the transformations to X and Y.]*

***A 95% PI for ln(savings) of a 40 year old = (8.46301, 10.8179). Therefore, a 95% PI for savings is ($4736, $49906).***

5. (**7X2 =14 points**) The table below was obtained from the Best Subsets regression procedure for the "Infection Risk" dataset.

```
Response is InfctRsk
```

| Vars | R-Sq | R-Sq (adj) | R-Sq (pred) | Mallows Cp | S | Stay | Culture | Xrays | Census | Nurses |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 35.5 | 34.8 | 30.2 | | 1.1351 | X | | | | |
| 1 | 34.7 | 34.0 | 30.7 | 53.2 | 1.1428 | | X | | | |
| 2 | 53.0 | 52.0 | 48.3 | 14.0 | 0.97380 | X | X | | | |
| 2 | 46.3 | 45.1 | 40.9 | 29.2 | 1.0415 | | X | | X | |
| 3 | 57.0 | 55.6 | 51.4 | 7.1 | 0.93657 | X | X | | | X |
| 3 | 56.0 | 54.6 | 49.5 | 9.4 | 0.94740 | X | X | | X | |
| **4** | **59.3** | **57.5** | **53.1** | **4.0** | **0.91622** | **X** | **X** | **X** | | **X** |
| 4 | 58.7 | 56.9 | 52.0 | 5.4 | 0.92323 | X | X | X | X | |
| 5 | 59.3 | 57.1 | 51.0 | 6.0 | 0.92120 | X | X | X | X | X |

a) Calculate SSTO using information in the table.

***R-Sq(adj)= 1- (n-1)MSE/(SSTO) → SSTO = [(n-1)MSE]/( 1- R-sq(adj))***
***You may use any model***
***Let us use the 'Best' model from part a.***
***SSTO = [(n-1)MSE]/( 1- R-sq(adj))***
***=96*(.91622^2)/(1-.575)=189.6 [Minitab output: 189.842]***

b) Calculate the Mallows $C_p$ for the model with Stay as the only predictor.

$MSE_p = 1.2885$ , $MSE_{all} = 0.84860944$

$$C_p = p + \frac{(MSE_p - MSE_{all})(n-p)}{MSE_{all}} = 2 + 49.24480 = 51.24480$$

c) Based on the criteria listed in the table above select what you believe to be the "Best" model and write down its population regression equation. Support your answer.

***First choose models such that $C_p$ close to p and are < p to exclude any models that provide biased predicted responses. There is only one model that satisfies this criterion: the one that has predictors Stay, Culture, Xrays and Nurses ($C_p$ =4.0 < 5). Also this is the model which has the highest R-sq(adj) (equivalently the lowest S value ensuring lowest variance of the predicted response).***

d) Would you consider this model to yield an unbiased predicted response? Support your answer.

***Yes, since $C_p$ = 4.0 < 5 = p.***

e) Name a model in the table that may yield a biased predicted response. Support your answer.

***The top model that contains only Stay as a predictor. $C_p$ =51.32 > 2 indicates substantial bias in the predicted response.***

f) Use Minitab's Backward Elimination procedure on this dataset and write down the fitted sample regression equation for the resulting "best" model. Use $\alpha_r = 0.15$ and the Minitab v17 command sequence: Stat > Regression > Regression > Fit regression model > Stepwise (select Backward Elimination for Method). For Minitab v16 use Stat > Regression > Stepwise.

| | ------Step 1----- | | ------Step 2------ | |
|---|---|---|---|---|
| | Coef | P | Coef | P |
| Constant | -0.318 | | -0.331 | |
| Stay | 0.2435 | 0.000 | 0.2464 | 0.000 |
| Cultures | 0.0477 | 0.000 | 0.04751 | 0.000 |
| Xrays | 0.01323 | 0.027 | 0.01315 | 0.025 |
| Census | 0.00016 | 0.922 | | |
| Nurses | 0.00202 | 0.239 | 0.002170 | 0.004 |
| | | | | |
| S | | 0.921196 | | 0.916224 |
| R-sq | | 59.32% | | 59.32% |
| R-sq(adj) | | 57.09% | | 57.55% |
| R-sq(pred) | | 50.97% | | 53.12% |
| Mallows' Cp | | 6.00 | | 4.01 |

α to remove = 0.15

***This procedure also chooses Stay, Cultures, Xray and Nurses as predictors to be included in the model and is exactly the same "Best model" to that selected by the Best Subsets procedure.***

***Fitted Regression Equation***

***InfctRsk = -0.331 + 0.2464 Stay + 0.04751 Cultures + 0.01315 Xrays + 0.002170 Nurses***

g) State any extra useful information provided by the Backward Elimination output that is not available in the Best Subsets table above.

***In addition to the fitted regression equation, the Backward Elimination provides, for example, the t-test p-values that can be used to rank the predictors. As clear from the output below, the contribution of the predictor Xrays is the least significant.***

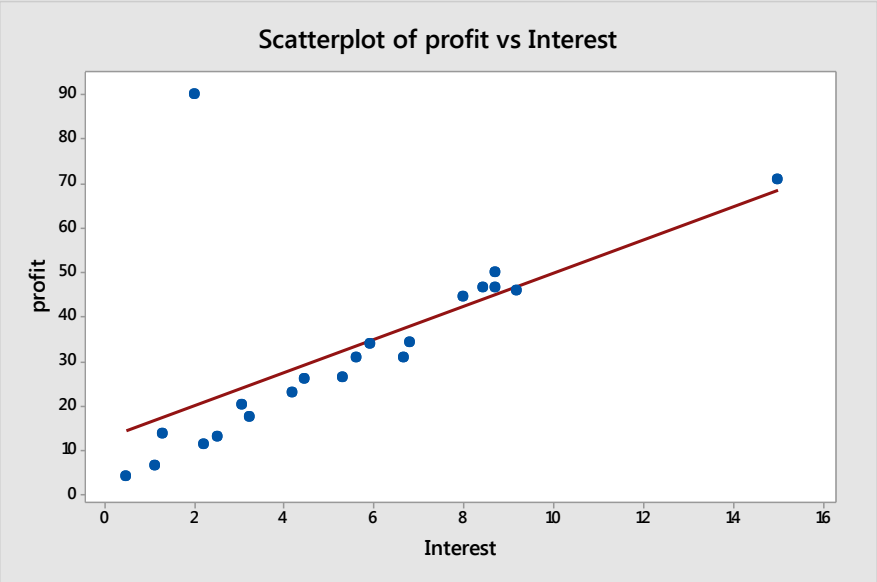
**Coefficients**

| Term | Coef | SE Coef | T-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | -0.331 | 0.551 | -0.60 | 0.550 | |
| Stay | 0.2464 | 0.0570 | 4.33 | 0.000 | 1.35 |
| Cultures | 0.04751 | 0.00985 | 4.82 | 0.000 | 1.26 |

| | | | | | |
|---|---|---|---|---|---|
| **Xrays** | 0.01315 | 0.00578 | 2.28 | 0.025 | 1.36 |
| **Nurses** | 0.002170 | 0.000727 | 2.99 | 0.004 | 1.14 |

6. (4X3 = 12 points) Open the "Profits" dataset in the Lesson 12 folder. The data indicate a positive linear association between interest rates and broker profits. The data are to be primarily used to obtain a regression model and compute confidence/prediction intervals.

   a. Detect any outliers, extreme X values and/or influential data points after fitting an SLR model for Y = profits and X = interest rate. Support your answer by means of scatter plots and quantitative diagnostic measures.




***Outlier threshold: |studentized residuals| > 3***
***Leverage threshold= 3 (p/n) = 3 (2/21) = 0.286***
***Influential data point: Cooks D > 0.5, 1 or more***

| *Interest* | *profit* | *SRES1* | *HI1* | *COOK1* |
|---|---|---|---|---|
| *2* | *90* | *4.30291* | *0.092813* | *0.947119* |
| *15* | *71* | *0.18312* | *0.418035* | *0.012044* |

***Outliers: (2,90)***
***Influential data points: none***
***Extreme X value: (15,71)***

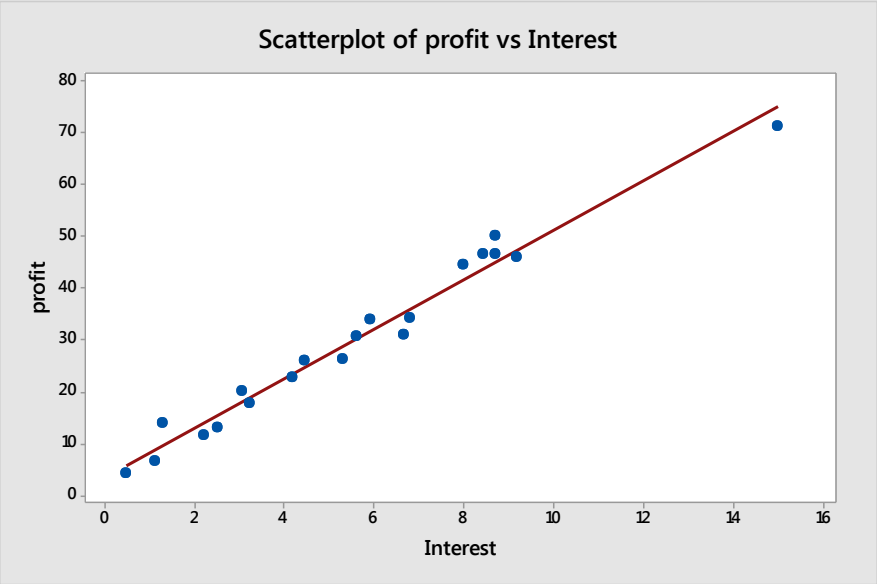
**Regression Analysis: profit versus Interest**

Analysis of Variance

| Source | DF | Seq SS | Seq MS | F-Value | P-Value |
|---|---|---|---|---|---|
| Regression | 1 | 3491 | 3491.3 | 12.00 | 0.003 |
| Interest | 1 | 3491 | 3491.3 | 12.00 | 0.003 |

```
Error        19    5528   290.9
Total        20    9019

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
17.0569  38.71%     35.48%      25.41%

Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant  12.63     6.88     1.84    0.082
Interest   3.73     1.08     3.46    0.003  1.00

Regression Equation
profit = 12.63 + 3.73 Interest
```

b. Repeat your regression analysis after deleting any outliers (unusual Y values).




```
Analysis of Variance

Source       DF  Seq SS   Seq MS  F-Value  P-Value
Regression    1  5425.2  5425.24   692.14    0.000
  Interest    1  5425.2  5425.24   692.14    0.000
Error        18   141.1     7.84
Total        19  5566.3

Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
2.79972  97.47%     97.32%      96.45%
```

```
Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant    3.41     1.18     2.88    0.010
Interest   4.767    0.181    26.31    0.000  1.00

Regression Equation

profit = 3.41 + 4.767 Interest
```

c. Compare the results of your regression analyses and plots obtained from parts a and b.

***The y-intercept and slope estimates are substantially different, especially in the case of the y-intercept. It is interesting to observe that with the outlier, the y-intercept is much higher than once it was removed, even though the p-values hint of lesser significance in the former case. The first plot clearly shows how the line has been unduly 'lifted up', thus overestimating the y-intercept. The second plot clearly shows a better fit.***

d. In the context of this problem, comment on any detrimental effects if outliers were not removed.

***It is apparent that the S value (sqrt of MSE) is substantially larger (17.06 vs 2.80) when the outlier was included resulting in much higher standard errors for the model coefficients. This will certainly widen their confidence intervals. The Rsq(adj) comparison (from 35.48% to 97.32%) also clearly points out another adverse effect of not removing an outlier.***

7. (**2+4+3+3 = 12 points**) There is evidence in general that broker profits increase as interest rates increase while # of sales decline as interest rates increase. The tables below give results from two SLR regression analyses carried out using n = 5 observations.

a) Fill in the two blanks in the tables below. *[There are different ways to approach this. One is to use formulas and properties to calculate the various measures – this will get you the most points. The other is to use Minitab to find the missing numbers – this can be used to check your calculations but will get you less points if you don't also show how to calculate the numbers using formulas and properties.]*

| Observation # | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Interest rate (X) | 8.7 | 9.2 | 4.0 | 15.1 | 18.0 |
| Profit in thousands (Y) | 46.5 | 45.8 | 75.0 | 10.1 | 71.0 |
| Leverage | 0.243 | 0.226 | 0.597 | 0.336 | 0.598 |

***0.598* in the first table is because the leverages add to p=2.**

| Observation # | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Interest rate (X) | 8.7 | 9.2 | 4.0 | 15.1 | 18.0 |
| # of sales (Y) | 20 | 15 | 30 | 10 | 11 |
| Fitted values | 20.277 | 19.608 | 26.564 | 11.715 | 7.836 |
| Cook's Distances | 0.001 | 0.261 | 1.415 | 0.073 | 1.206 |

***Formula for Cook's D:***

$$D_i = \frac{(y_i - \hat{y}_i)^2}{p \times MSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

***To find MSE, use observation 4 (any observation except observation 5 is okay): MSE= (1.715^2)/(2*0.073)*(0.336/(0.664^2)) = 15.3525.***
***(Note the leverages remain the same since the model is the same)***

***Cook's D for obs 5 = (3.164^2)/(2*15.3525)*(0.598/(0.402^2)) = 1.206 (Minitab result: 1.201).***

b) Which observation(s), if any, are extreme X values in predicting the broker profits? Justify your answer.

***Extreme X value threshold: 3(p/n) = 1.2.***
***There are no extreme X values since no leverage exceeds 1.2.***

c) Which observation(s), if any, would you consider to be influential data point(s) in predicting the # of sales? Justify your answer.

***Threshold for Cook's D > 1.***
***Observations 3 and 5 are influential data points.***