MTH 325, Exam #1, At Home, Spring 2024 Name KEY

**Instructions**: This exam is in two parts: Part I is to be completed partly at home using the materials posted in the course for the at-home portion and you will answer questions about that work during the in-class portion of the exam; Part II is to be completed entirely in class. You may not use cell phones, and you may only access internet resources you are specifically directed to use.

At home, prepare for questions in Part I using R. Open the data file entitled **325exam1data.xlsx** posted in Blackboard. Complete the calculations noted below. You will be asked for additional analysis and interpretation of this data in the in-class portion of the test. Print out the results of your analysis and code, and bring the pages with you to the exam. You will submit all this work along with the in-class exam.

Use the data on employment experience at the Beta Technology company to complete the following tasks.

1. Import the data in the file into R and removed the Employee column (it is not a variable).
2. Create a correlation table of the variables. Make a correlation plot (type is of your choice), or a pairplot.
3. Create a simple linear regression model between number supervised and salary. Create appropriate graphs for diagnostic testing of assumptions, and identify potential outliers.
4. Create a multiple variable model of salary using all available variables. Use appropriate automated selection techniques. Compare the result to manual backward selection. In your backward selection, stop only when all the coefficients are significant at the 0.05 level.
5. Construct diagnostic plots for your machine selected model and your manually selected model (these may be the same). Identify any potential problems with model assumptions, outliers and influential points.
6. Construct a confidence interval for the ~~gender~~ Department variable coefficient.
7. Construct a 95% prediction interval for the salary of an employee with gender 1, 4 years of education, 15 years of previous experience, 15 years employed, department 3 and supervises 5 people.

MTH 325, Exam #1, Spring 2024 Name KEY

Instructions: Answer each question thoroughly. For questions in Part 1, use the work you did at home to answer the questions. Be sure to answer each part of each question. In Part 2, report exact answers unless directed to round.

Part I:
Use the work you did at home to answer these questions about tax paid and the neighborhoods in our dataset.

1. Based on your correlation table, identify the variable that has the highest correlation with Salary. What is the correlation value?

Years Education 0.777

2. Based on your correlation table (or graphs), which variables (other than Salary) appear to have potential collinearity problems?

Years Education and Years Employed have a conclation of 0.6
highest among independent variables

3. What is the simple linear regression equation your found relating Number Supervised to Annual Salary?

Salary = 563.9 x Number Supervised + 36,615.6

4. Interpret the slope in the context of the problem.

for each additional employee supervised, salary goes up by an average of $563.90

5. What percent of the variability in Salary can be explained by the relationship with Number Supervised?

27.45%

6. Compare your machine found model with your final backwards selection model. Describe any differences in your models (variables included), any errors generated in the selection, etc.

machine selected model (Stepwise selection) kept #Supervised while backward selection eliminated it.

7. Answer this question and the remaining questions in Part 1 using the **backward selection model** you found by hand. Write the equation of your model that describes your multiple regression model.

Salary = 2278 * Department + 1871.3 years Education
+ 648.2 years Employed + 17168.4

8. Construct a prediction interval for an employee with gender 1, 4 years of education, 15 years of previous experience, 15 years employed, department 3 and supervises 5 people.

machine model (30,787.15, 51,633.24)
vs
backward selection (30,989.75, 51,576.18)

9. Interpret the meaning of the Department coefficient in the context of the problem.

for each unit increase in department number, salary increases by $2278. (this is questionable due to category)

10. Construct a confidence interval for the Department variable coefficient.

(1037.7378, 3518,3407
or
[919, 99239, 3392,6130)

11. Test your model assumptions using your residual plots and other diagnostic plots. Do they appear to be approximately satisfied? Identify any potential outliers.

there are some outliers in number supervised

QQ plot of residuals also shows some potential outliers

12. Based on your best model, interpret the meaning of the $R^2$ value.

80.4% of the variability in salary can be explained by the (81.37%) relationship w/ years education, years employed, department (and number supervised)

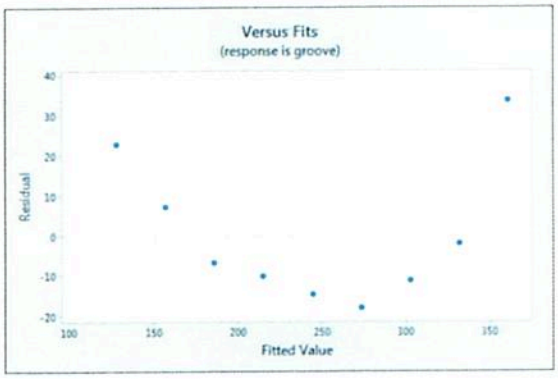
13. Are there any potential problems with treating department like a numerical variable in this context? Explain.

Yes. even if they are sorted by order in which salary is affected, the relationship may not be linear since it's really a category

Part II:

14. Examine the residual plots below. Identify any problems associated with each plot in terms of potential problems for the standard assumptions made for a linear regression model.

a.

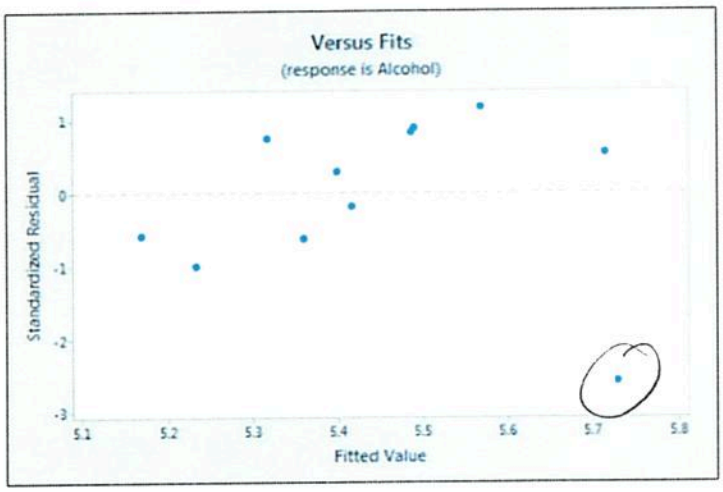



nonlinear

b.


non-constant variance heteroscedashic

c.




outlier

15. State the null and alternative hypothesis for a multiple regression model (for the full model and any other tests conducted to assess model quality).

$H_0: \beta_i = 0 \ \forall \beta_i \quad H_a: \beta_i \neq 0$ for some $i$ full model

individual coeff.

$H_0: \beta_i = 0, \quad H_a: \beta_i \neq 0$

16. Recall that $Cov(X,Y) = E(XY) - E(X)E(Y)$. For the probability density function $f(x,y) = \frac{5}{384}x^4(x+\sqrt{y}), x \in [0,2], y \in [0,4]$, find the covariance.

$$E(X) = \int_0^2\int_0^4 \frac{5}{384} x \cdot x^4(x+\sqrt{y})\,dy\,dx = \frac{5}{384}\cdot\frac{8192}{63} = \frac{320}{189}$$

$$E(Y) = \int_0^2\int_0^4 \frac{5}{384} y\, x^4(x+\sqrt{y})\,dy\,dx = \frac{5}{384}\cdot\frac{12544}{75} = \frac{98}{45}$$

$$E(XY) = \int_0^2\int_0^4 \frac{5}{384} xy\cdot x^4(x+\sqrt{y})\,dy\,dx = \frac{5}{384}\cdot\frac{29696}{105} = \frac{232}{63}$$

$$E(XY) - E(X)E(Y) = \frac{232}{63} - \frac{320}{189}\cdot\frac{98}{45} = \frac{232}{63} - \frac{896}{243} \approx -0.0047$$

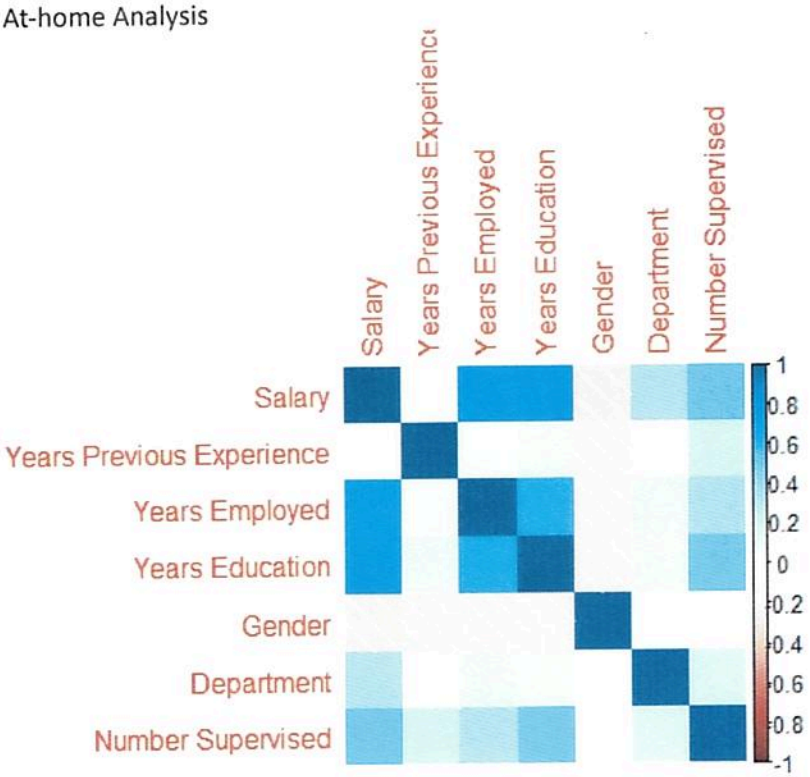
17. Consider the small data set $\{(10,1),(8,3),(4,7)\}$. Find the value of the regression coefficients for $y = \beta_0 + \beta_1 x$, using the normal equation $(A^TA)^{-1}A^TY = B$. Write the coefficients you find in the equation.

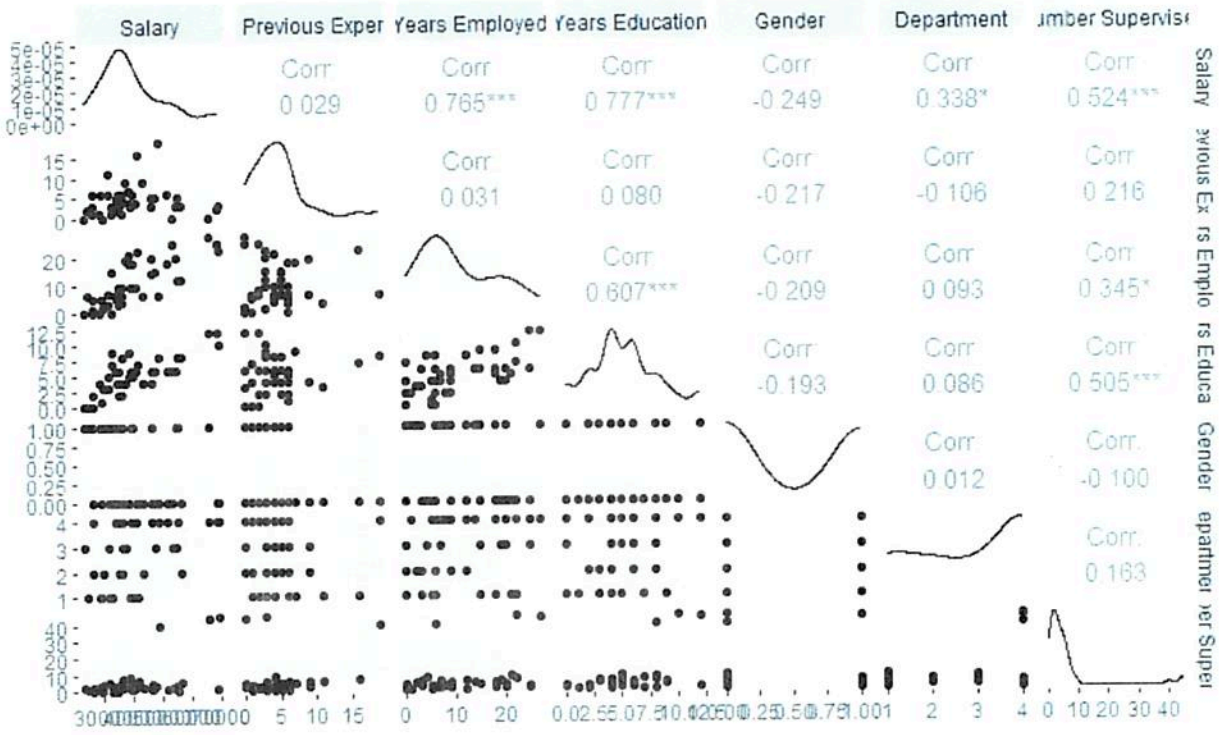
$$A = \begin{bmatrix} 10 & 1 \\ 8 & 1 \\ 4 & 1 \end{bmatrix} \qquad A^TA = \begin{bmatrix} 10 & 8 & 4 \\ 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} 10 & 1 \\ 8 & 1 \\ 4 & 1 \end{bmatrix} = \begin{bmatrix} 180 & 22 \\ 22 & 3 \end{bmatrix} \qquad Y = \begin{bmatrix} 1 \\ 3 \\ 7 \end{bmatrix}$$

$$A^TY = \begin{bmatrix} 10 & 8 & 4 \\ 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 3 \\ 7 \end{bmatrix} = \begin{bmatrix} 62 \\ 11 \end{bmatrix}$$

$$B = \begin{bmatrix} -1 \\ 11 \end{bmatrix} \qquad y = -x + 11$$

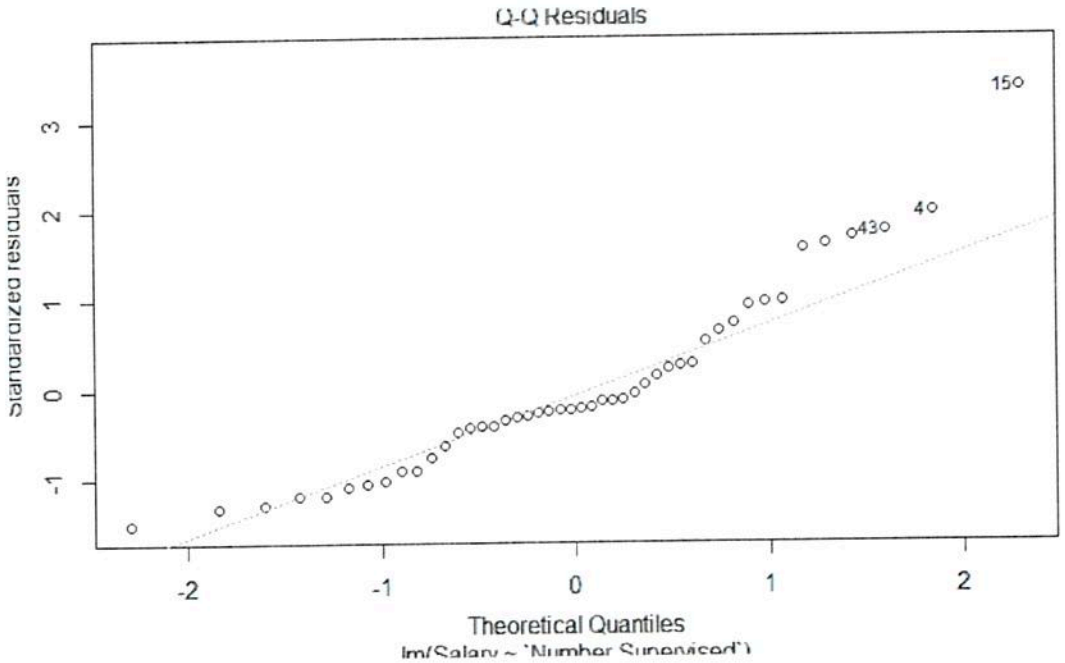
MTH 325 Exam #1 At-home Analysis

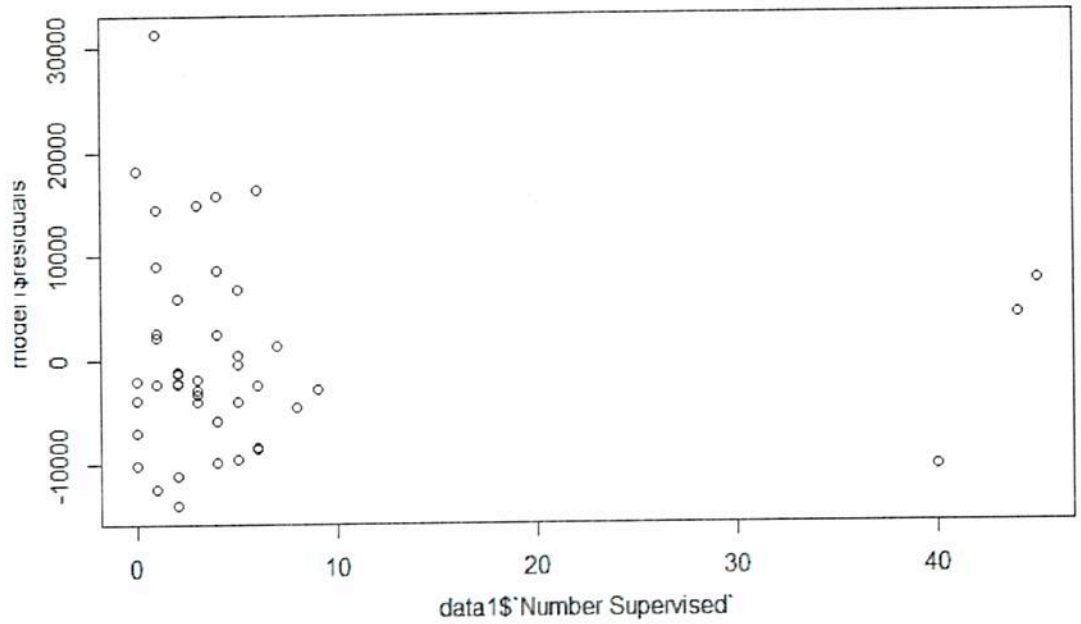


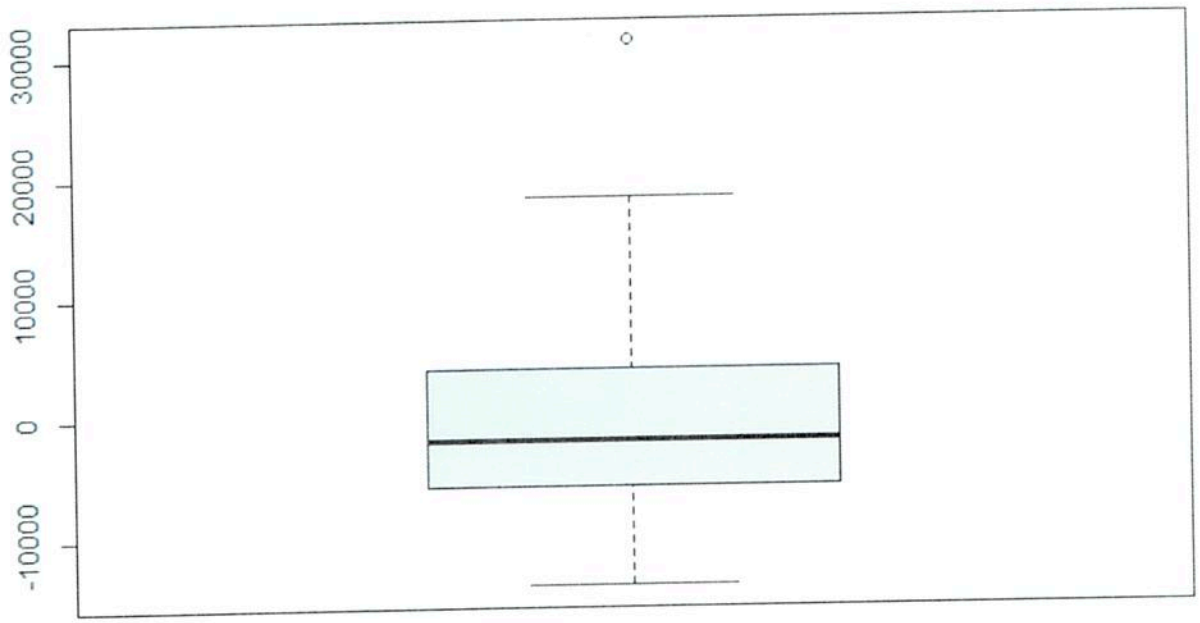




```
Call:
lm(formula = Salary ~ `Number Supervised`, data = data1)

Residuals:
```

```
   Min     1Q Median     3Q    Max
-14089  -5601  -2048   3686  31246

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          36615.6     1603.5   22.84  < 2e-16 ***
`Number Supervised`    563.9      138.2    4.08 0.000186 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9475 on 44 degrees of freedom
Multiple R-squared:  0.2745,   Adjusted R-squared:  0.258
F-statistic: 16.65 on 1 and 44 DF,  p-value: 0.0001863
```






```
    15
31245.48
```

All variables

```
Call:
lm(formula = Salary ~ ., data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-11368.4  -2673.6    420.3   2190.9  11476.2

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               19589.47    2862.64   6.843 3.52e-08 ***
`Years Previous Experience`  -106.55     213.08  -0.500   0.6199
`Years Employed`              621.06     125.41   4.952 1.46e-05 ***
`Years Education`            1631.83     362.76   4.498 6.01e-05 ***
Gender                      -1654.07    1558.11  -1.062   0.2950
Department                   2134.29     624.77   3.416   0.0015 **
`Number Supervised`           134.01      88.14   1.520   0.1365
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5022 on 39 degrees of freedom
Multiple R-squared:  0.8193,   Adjusted R-squared:  0.7915
F-statistic: 29.47 on 6 and 39 DF,  p-value: 4.895e-13
```

Eliminating Years Previous Experience:

```
Call:
lm(formula = Salary ~ `Years Employed` + `Years Education` +
    Gender + Department + `Number Supervised`, data = data1)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-11257.8  -2569.3    228.2   2335.5  11744.9

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         18888.05    2471.90   7.641 2.44e-09 ***
`Years Employed`      624.48     124.05   5.034 1.06e-05 ***
`Years Education`    1637.45     359.17   4.559 4.77e-05 ***
Gender              -1488.31    1508.10  -0.987  0.32964
Department           2178.10     612.77   3.555  0.00099 ***
`Number Supervised`   123.91      84.98   1.458  0.15264
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4975 on 40 degrees of freedom
Multiple R-squared:  0.8181,   Adjusted R-squared:  0.7954
F-statistic: 35.99 on 5 and 40 DF,  p-value: 8.666e-14
```

Now eliminate Gender

```
Call:
lm(formula = Salary ~ `Years Employed` + `Years Education` +
    Department + `Number Supervised`, data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-12251.8  -3070.5     62.9   2452.6  12332.4

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         17945.73    2279.31   7.873 9.92e-10 ***
`Years Employed`      639.17     123.11   5.192 6.04e-06 ***
`Years Education`    1665.10     357.96   4.652 3.41e-05 ***
Department           2156.30     612.17   3.522  0.00107 **
`Number Supervised`   124.07      84.96   1.460  0.15182
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4973 on 41 degrees of freedom
Multiple R-squared:  0.8137,   Adjusted R-squared:  0.7955
F-statistic: 44.77 on 4 and 41 DF,  p-value: 1.932e-14
```

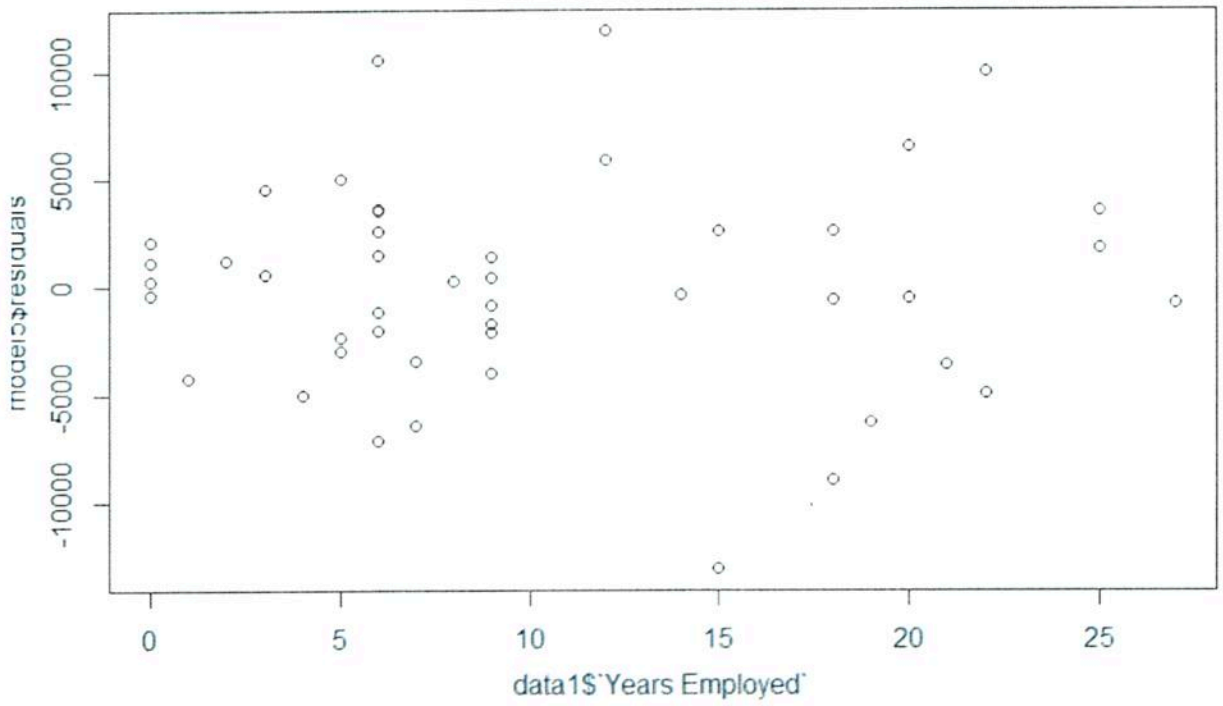
Now eliminate Number Supervised

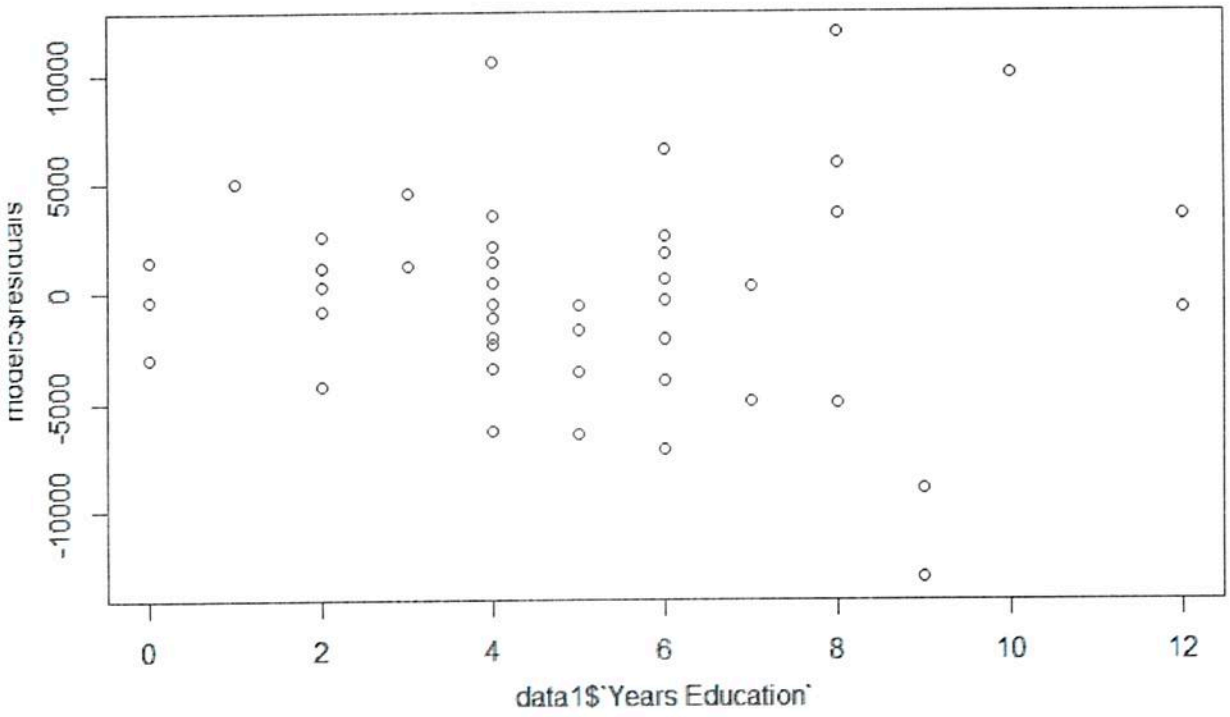
```
Call:
lm(formula = Salary ~ `Years Employed` + `Years Education` +
    Department, data = data1)

Residuals:
   Min     1Q Median     3Q    Max
-13090  -2824    -56   2562  11853

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       17168.4     2246.0   7.644 1.77e-09 ***
`Years Employed`    648.2      124.6   5.202 5.52e-06 ***
`Years Education`  1871.3      333.3   5.614 1.43e-06 ***
Department         2278.0      614.6   3.707 0.000609 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5040 on 42 degrees of freedom
Multiple R-squared:  0.804,    Adjusted R-squared:   0.79
F-statistic: 57.44 on 3 and 42 DF,  p-value: 6.489e-15
```

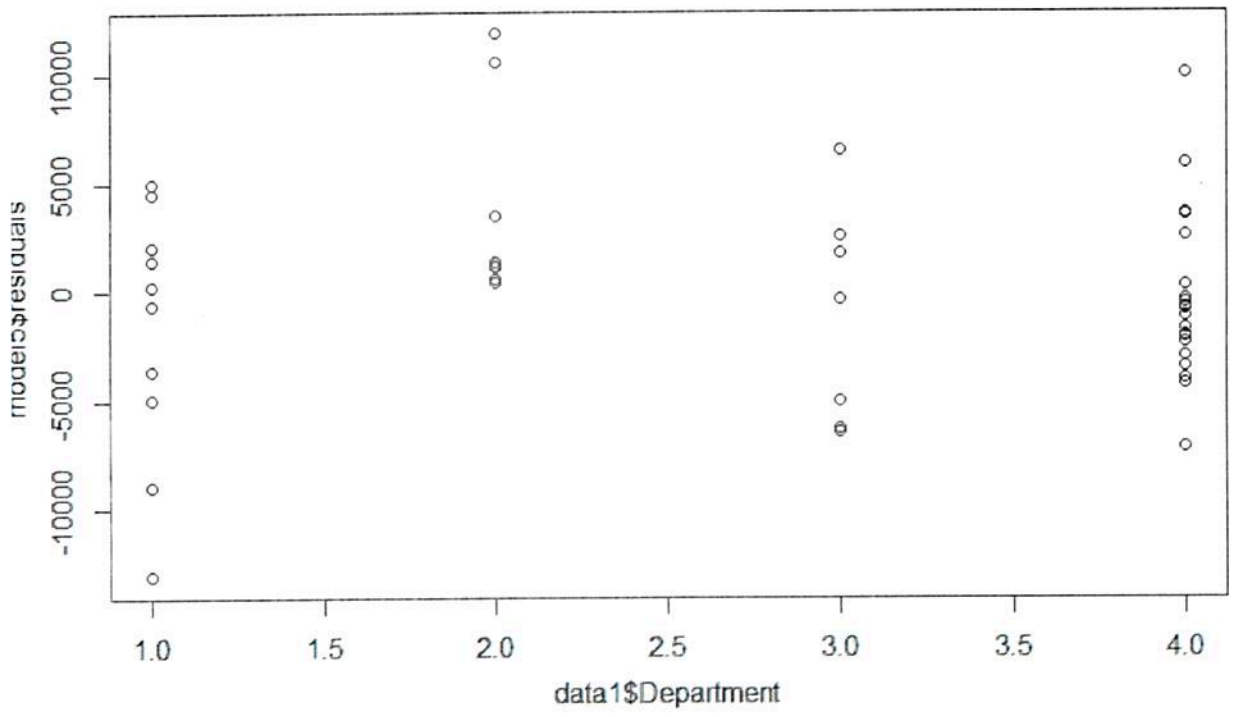
This is the final version for backward selection.

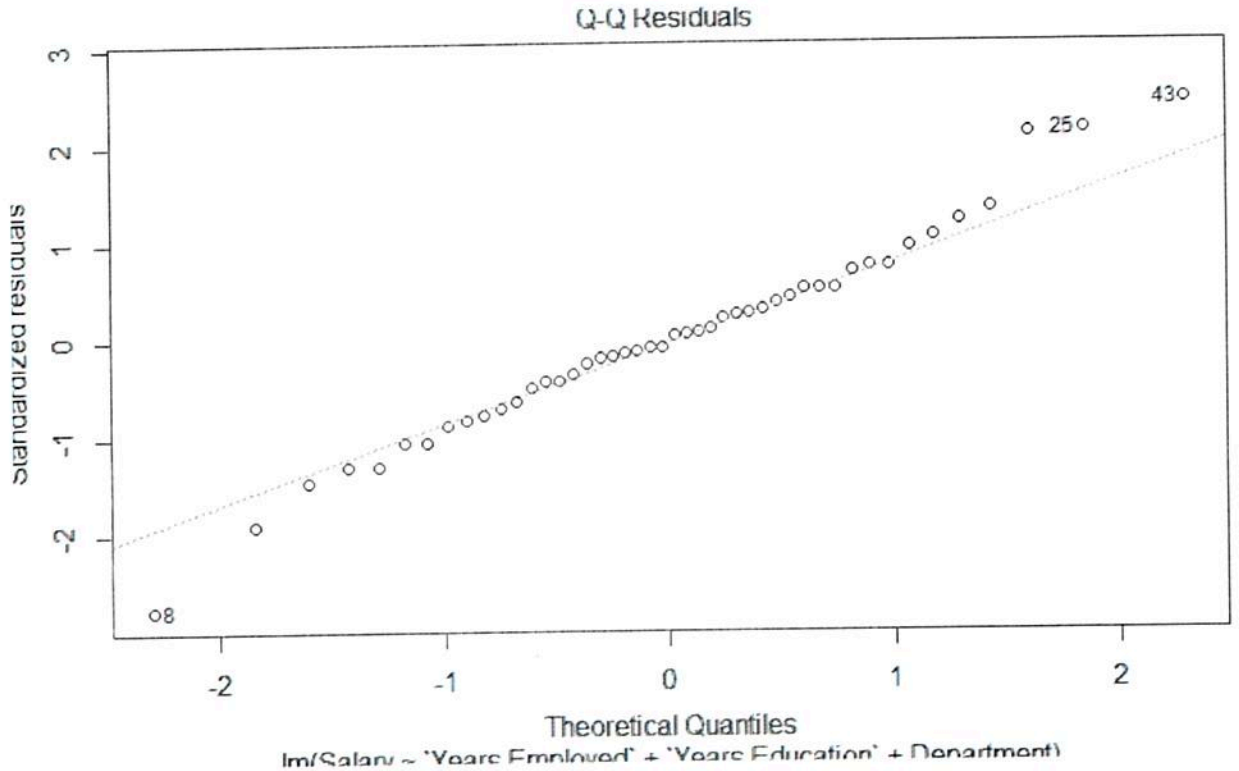
Stepwise regression, in both directions, using AIC for best model:

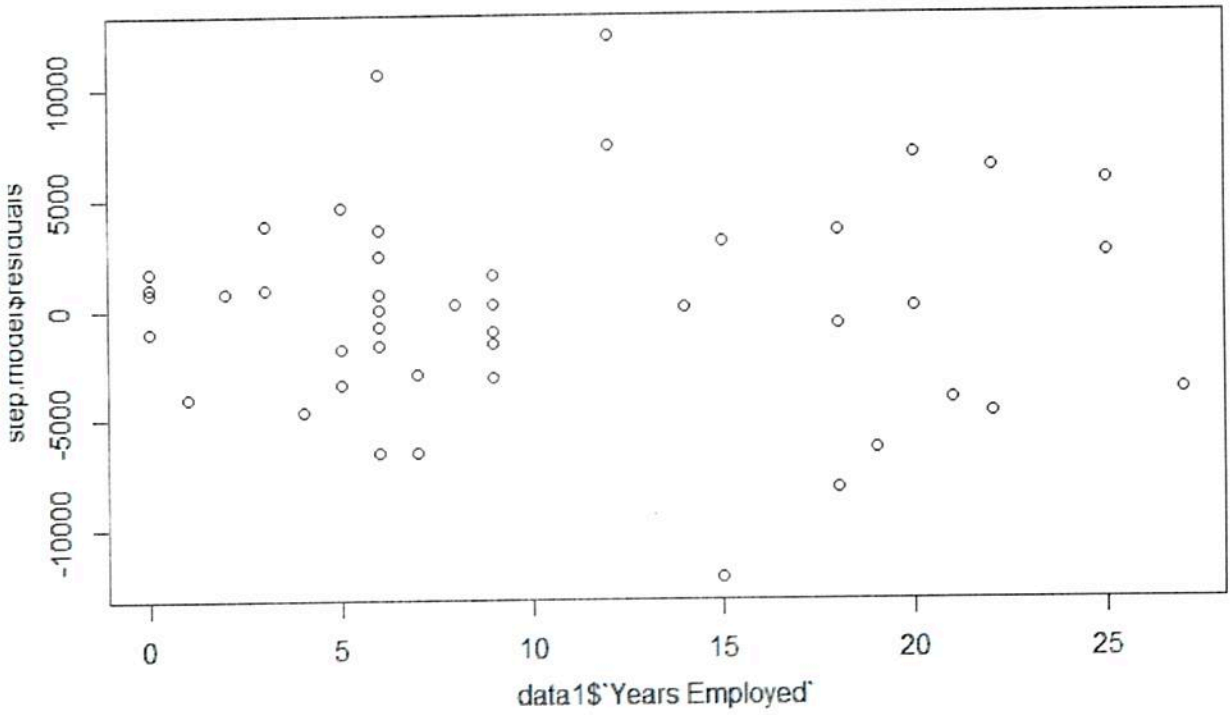
```
Call:
lm(formula = Salary ~ `Years Employed` + `Years Education` +
    Department + `Number Supervised`, data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-12251.8  -3070.5     62.9   2452.6  12332.4

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        17945.73    2279.31   7.873 9.92e-10 ***
`Years Employed`     639.17     123.11   5.192 6.04e-06 ***
`Years Education`   1665.10     357.96   4.652 3.41e-05 ***
Department          2156.30     612.17   3.522  0.00107 **
`Number Supervised`  124.07      84.96   1.460  0.15182
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4973 on 41 degrees of freedom
Multiple R-squared:  0.8137,   Adjusted R-squared:  0.7955
F-statistic: 44.77 on 4 and 41 DF,  p-value: 1.932e-14
```

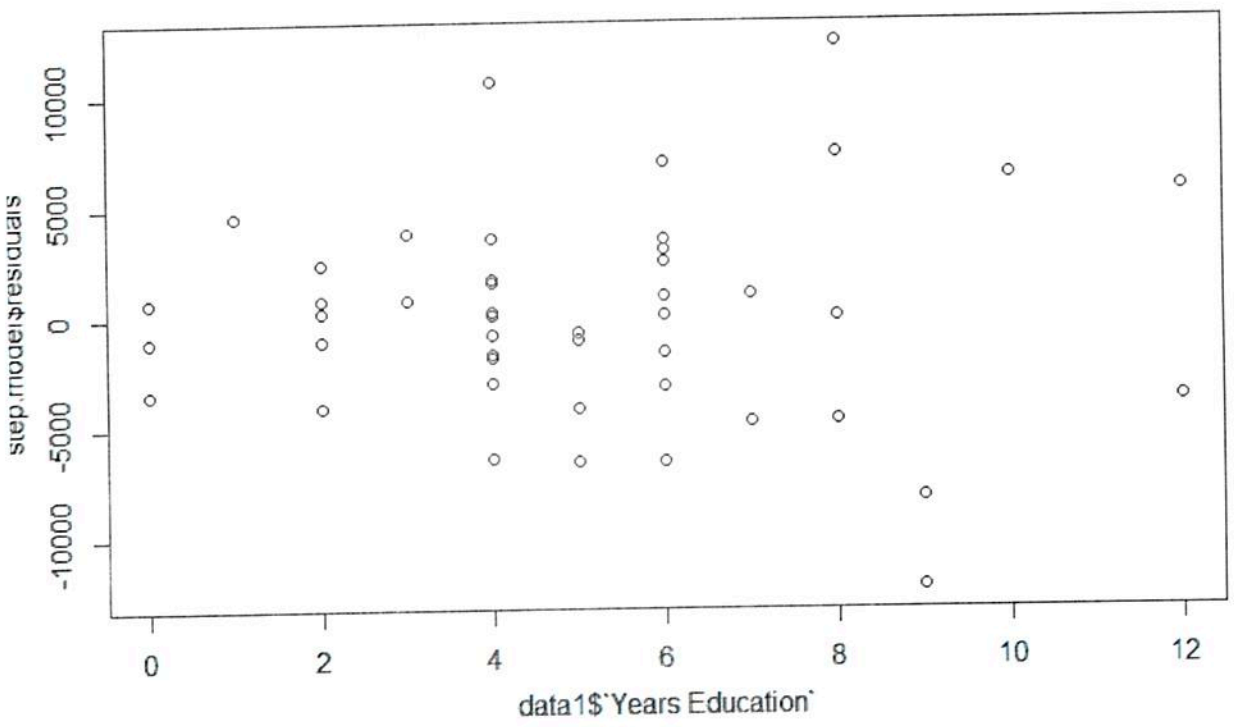


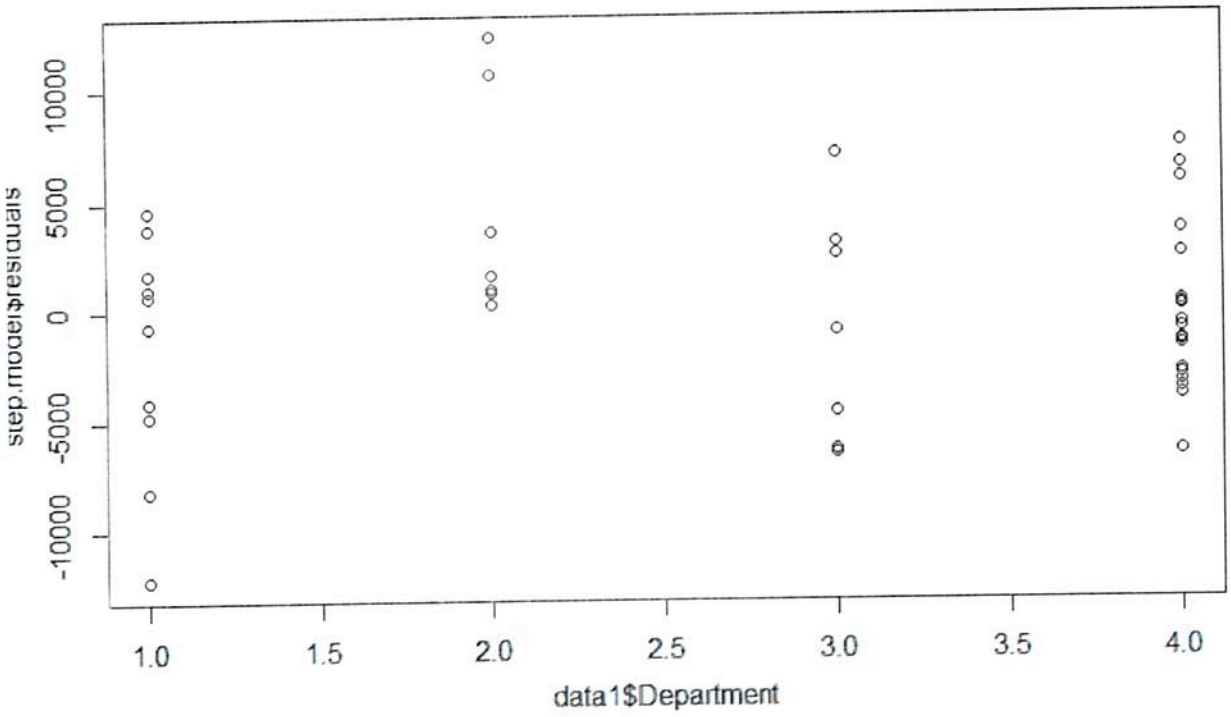

model3$residuals
10000
5000
0
-5000
-10000
0
2
4
6
8
10
12
data1$`Years Education`

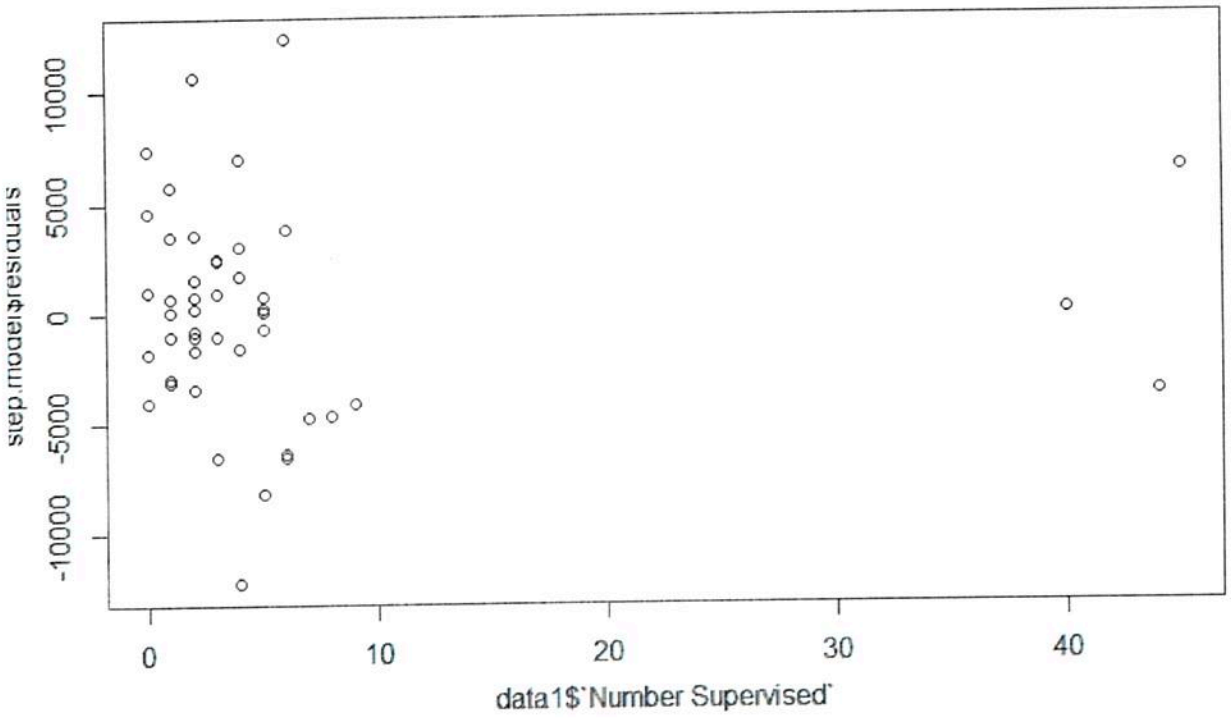

model3$residuals
10000
5000
0
-5000
-10000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
data1$Department

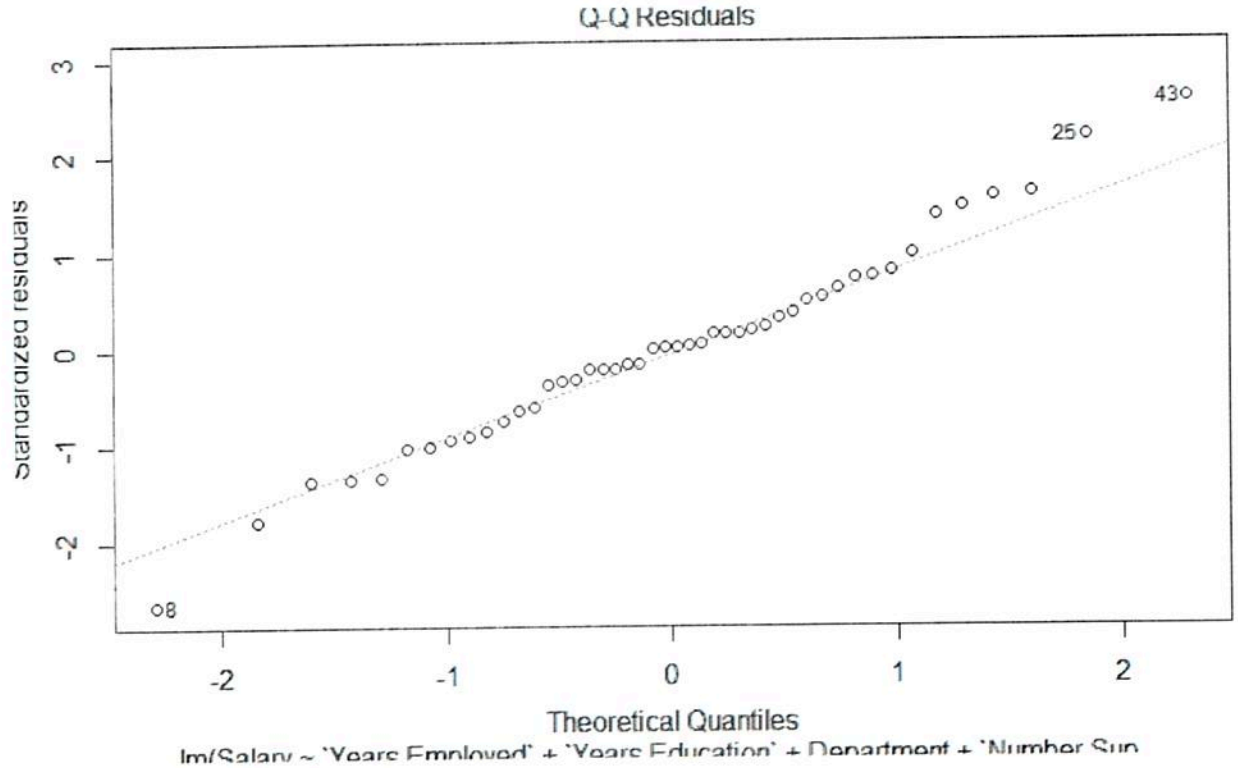

Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Salary ~ `Years Employed` + `Years Education` + Department)
430
250
8


step.model$residuals
data1$`Years Employed`


step.model$residuals
10000
5000
0
-5000
-10000
0
2
4
6
8
10
12
data1$`Years Education`


step.model$residuals
10000
5000
0
-5000
-10000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
data1$Department




| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | 12635.8935 | 21701.0030 |
| `Years Employed` | 396.7024 | 899.6295 |
| `Years Education` | 1198.5710 | 2543.9989 |
| Department | 1037.7378 | 3518.3407 |

```
                          2.5 %     97.5 %
(Intercept)         13342.57698 22548.8789
`Years Employed`      390.54230   887.8038
`Years Education`     942.18662  2388.0125
Department            919.99239  3392.6130
`Number Supervised`   -47.50853   295.6418

predicted_salary1
[1] 41210.2
> lower_limit1
[1] 30787.15
> upper_limit1
[1] 51633.24

predicted_salary2
[1] 41282.96
> lower_limit2
[1] 30989.75
> upper_limit2
[1] 51576.18
```