MTH 325, Exam #2, 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 **325exam2data.xlsx** posted in Blackboard. There are multiple sheets in this file. Save them to separate dataframes. 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.

1. On Sheet 4 is data on employees including gender, age, prior experience, beta experience, education and annual salary (eliminate Employee). Use this data to predict annual salary using gender and education (convert these to factors) using a generalized linear model (ANOVA). Identify main effects and if any interaction term is significant. Be prepared to write the equation of the model and discuss diagnostics such as residual plots.

2. Using the data on Sheet 1, (after eliminating the Boiler number column) create a logistic regression model that predicts Drum Type from Boiler Capacity and ~~Design Pressure~~ Worker Hours. Plot the graph(s). Create appropriate exploratory graphs. Create appropriate diagnostic plots, and a confusion matrix.

3. Create a graph of the data on Sheet 2 with Average Monthly Temperature on the horizontal axis, and Average Monthly Bill on the vertical axis. Create a nonlinear model for the data by transforming variables. Plot the resulting model. Create appropriate diagnostic plots. Bonus points for comparing your model to any other nonlinear model (splines, LOESS or Gaussian). Predict the monthly bill if the average monthly temperature was 92.

4. On Sheet 4 is employee data. Eliminate the Employee column. Gender is already encoded as a binary dummy variable. You'll need to encode the Education variable as separate dummy variables. The rest of the variables are numerical. Use LASSO regression on data to find a model of best fit. Salary Compare the resulting model to a model using a linear model with the same variables. Prepare appropriate diagnostics and diagnostic graphs.

MTH 325, Exam #2, 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 employee salaries in our dataset.

1. Write the model from your two-way ANOVA or glm model including the interaction term. Be sure to explain which level is considered the default in your two variables.

Salary = −11488·Gender − 1628 Associates + 31348 Bachelors
+ 39505 Masters + 87903 Doctorate + 46691

Male = Gender 0
HS = Education 0

interaction terms were rejected as not significant

2. Briefly describe any boxplots, residual plots or normal plots you created to verify your model.

Some possible outliers, but otherwise appears mostly normal

Use the work you did at home to answer these questions about boilers in our dataset.

3. Write the equation of your logistic model below. You can write it in the form $\ln\left(\frac{p}{1-p}\right) = linear\ model$.

$\ln\left(\frac{P}{1-P}\right) = 3.465\times10^{-5}$ Boiler Capacity $- 5.343\times10^{-3}$ Worker Hours $+ 1.068\times10^{1}$

4. Interpret the slope (of Work Hours) in the context of the problem.

as worker hours increase, the odds of a mud type drum decreases

5. Explain the meaning of the null and residual deviance for your model in this context.

null deviance 49.461
residual deviance = 10.770

how errors are measured in logistic models, similar to standar error

6. What is the accuracy of your model?

88.89%

7. Does your confusion matrix suggest any potential problems with the data? Could masking or bias be a potential issue?

accuracy pretty good
errors are not high in either direction
masking does not seem present.

Use the data on electric bills to answer the following questions.

8. Describe the type of non-linear (parametric) model that would seem appropriate for this data. Why? Write the equation of your model.

using quadratic model

AvgBill = 0.09 Temp^2 − 11.73 Temp + 449.45

9. What is the R^2 value for your model?

89.03%

10. What is the residual standard error of your model?

9.106

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

few points make them hard to interpret but residual might warrant trying higher degree polynomial

12. (Bonus) Describe nonparametric (other nonlinear) model and compare it to your polynomial model. Describe any advantages or disadvantages to this model. Which model did you use and why?

I used Loess (your answers may vary) because easy to plot in ggplot. more flexible than quadratic.

Standard error is smaller, residual plots look a little better

Use the employee data to answer the following questions.

13. Write the equation of your LASSO model below.

Salary = 63633.76 Doctorate + 25281.82 Masters + 13849.37 Bachelors − 6461.20 Associates + 2601.10 Beta + 2940.51 Prior − 26.11 Age − 8082.64 Gender + 19366.19

14. Write the equation of the linear model with the same variables below.

Salary = 64685.48 Doctorate + 26225.8 Masters + 14763.40 Bachelors + −5696.71 Associates + 2606.36 Beta + 2949.23 Prior − 35.52 Age − 8105.39 Gender + 18782.22

15. Compare the coefficients in your two models. How do they differ?

They are similar

education & experience coeff are larger in the linear model

others are smaller

16. Are any of the retained variables in your model unable to pass a hypothesis test for the coefficient in the linear model? Explain how you would handle this in an analysis.

Age does not pass significance
I would drop it for parsimony

17. Even though the Education was encoded as an ordinal variable, why should we not analyze them in the model this way?

because the relationship is not linear
jump from category to category may not be equal

Part II:

18. Describe at least two reasons why someone might want to create a 2^p factorial design experiment.

to test whether a variable has any change over a largish range to be included in later tests

19. Describe one reason why we might want to recode a continuous variable as discrete (or categorical)?

answers may vary - apply ANOVA model
income may be sensitive. might be more willing to give range than exact # ; look at groups

20. Describe how k-fold cross validation works in validating a model.

data divided in k groups. one group is for test, K-1 groups for training
repeated K times for each subgroup. results averaged
helps to avoid risk of unlucky splits

21. Describe supervised learning (in machine learning) and give an example of a machine learning algorithm that implements this learning method.

Supervised learning includes both regression and classification
where outcomes are known and used in model selection

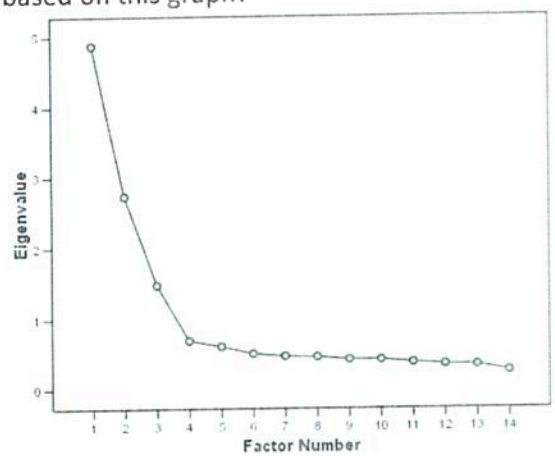
22. Describe how LOESS regression works in general terms.

a percentage of the data nearest a point is chosen and model w/ a low degree polynomial (typically 2). combination of points used changes as scans across.

23. How does adding a penalty improve model selection in regression? What is a potential disadvantage?

penalty is a trade off between bias and variance
as one increases, the other decreases and vice versa

24. An example of a scree plot is shown below. How many factors should be selected for the model based on this graph?



3 (or 4)

25. Describe one advantage and one disadvantage of ensemble methods in machine learning.

advantage - can be very robust and applied to multiple model types
less risk of overfitting

disadvantage - computationally complex

26. Spline regression is used for nonlinear regression modeling. What is one advantage and one disadvantage of this regression method?

more flexible than polynomials, can be written as a piecewise function

differs by type of spline, usually applied, may not be continuous

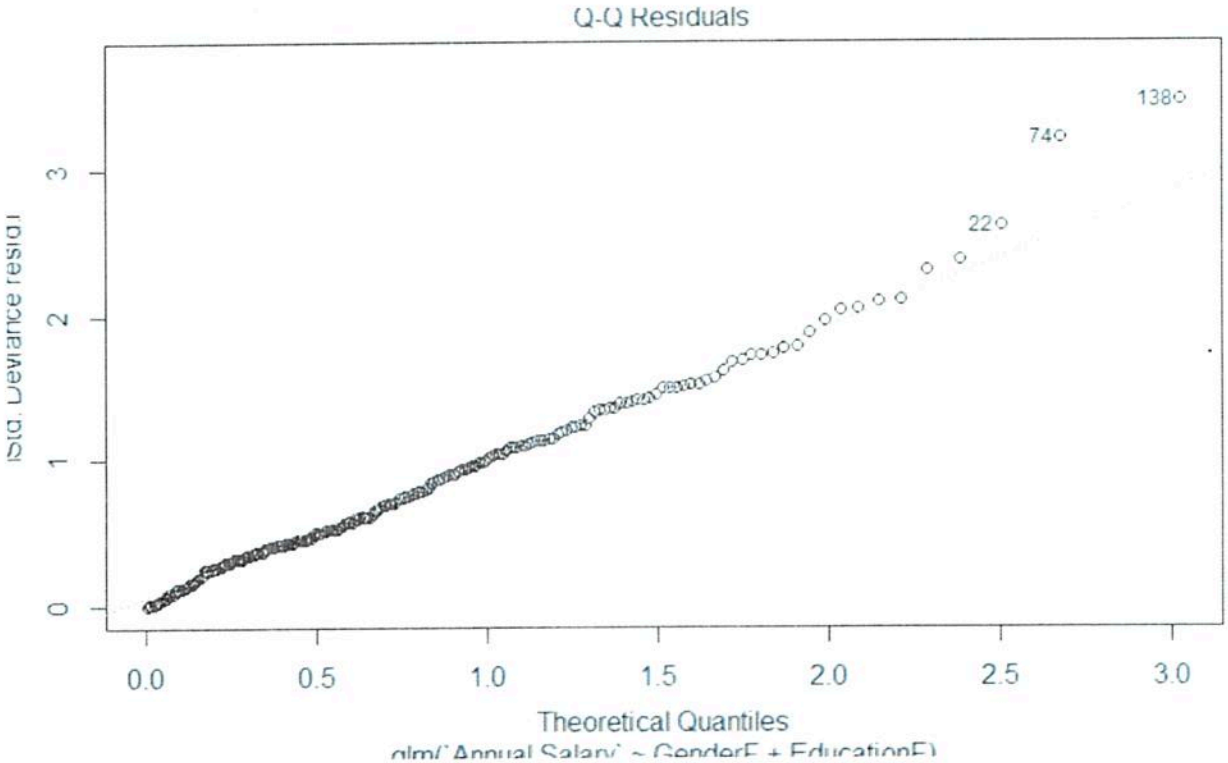
(answers will vary)

MTH 325 Exam #2 At-home Analysis

No interaction

As glm:

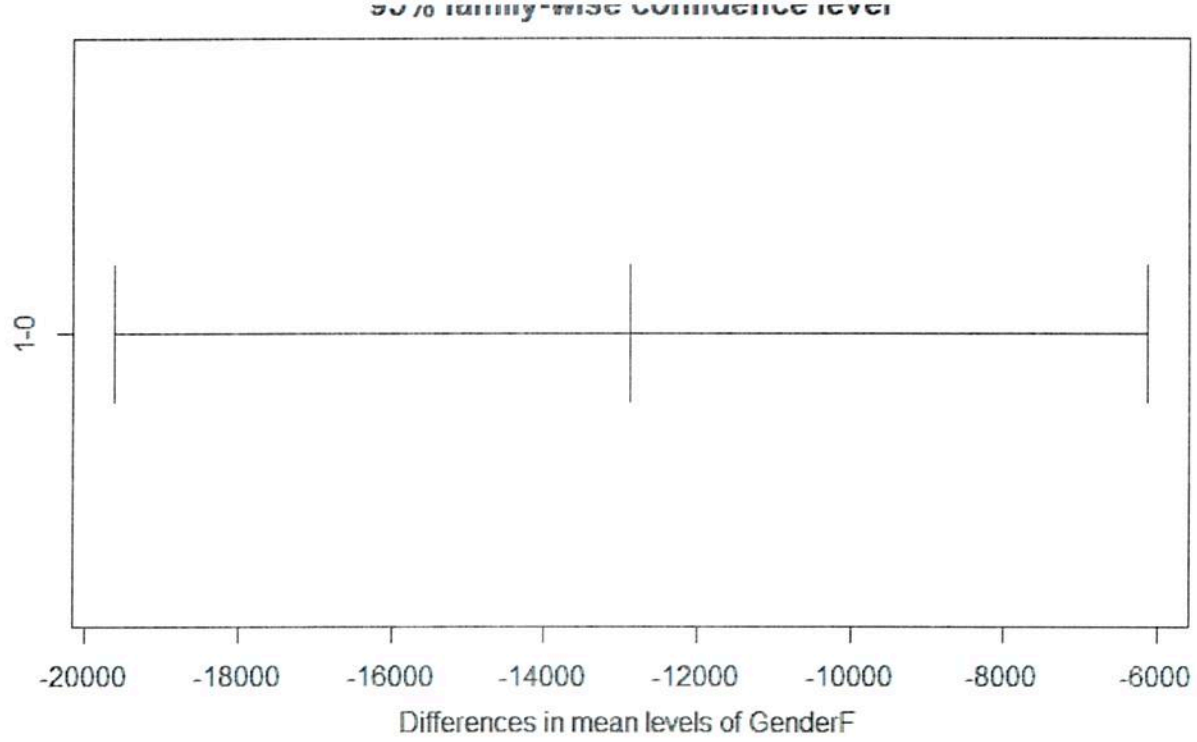
```
Call:
glm(formula = `Annual Salary` ~ GenderF + EducationF, data = data4)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    46691       8105   5.761 3.16e-08 ***
GenderF1      -11488       3448  -3.332 0.001028 **
EducationF2    -1628       9462  -0.172 0.863540
EducationF4    31368       7975   3.933 0.000116 ***
EducationF6    39505       8387   4.710 4.65e-06 ***
EducationF8    87903      11105   7.916 1.72e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 580799319)

    Null deviance: 1.8579e+11  on 203  degrees of freedom
Residual deviance: 1.1500e+11  on 198  degrees of freedom
AIC: 4703.5

Number of Fisher Scoring iterations: 2

(Intercept)    GenderF1 EducationF2 EducationF4 EducationF6 EducationF8
  46690.785  -11488.482   -1628.427   31367.521   39505.078   87902.816
```

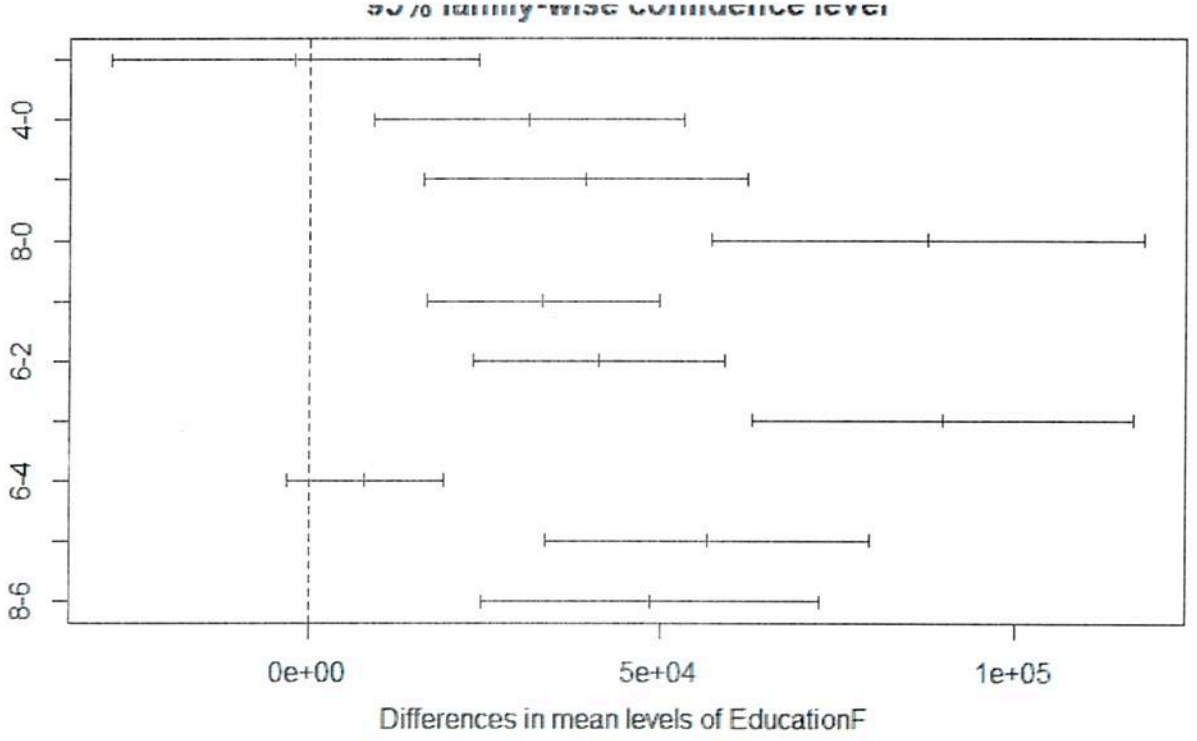


As ANOVA:

```
             Df     Sum Sq    Mean Sq F value    Pr(>F)
GenderF       1 8.219e+09 8.219e+09   14.15 0.000222 ***
EducationF    4 6.257e+10 1.564e+10   26.93  < 2e-16 ***
Residuals   198 1.150e+11 5.808e+08
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```





With interaction

As glm

```
Call:
glm(formula = `Annual Salary` ~ GenderF * EducationF, data = data4)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	77800	16843	4.619	7e-06	***
GenderF1	-50375	18831	-2.675	0.00811	**
EducationF2	-29256	18620	-1.571	0.11777	
EducationF4	-3998	17205	-0.232	0.81650	
EducationF6	11175	17530	0.637	0.52457	
EducationF8	65675	20628	3.184	0.00169	**
GenderF1:EducationF2	32271	21780	1.482	0.14005	
GenderF1:EducationF4	45980	19369	2.374	0.01858	*
GenderF1:EducationF6	33637	19981	1.683	0.09390	.
GenderF1:EducationF8	22900	24696	0.927	0.35494	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 567343365)

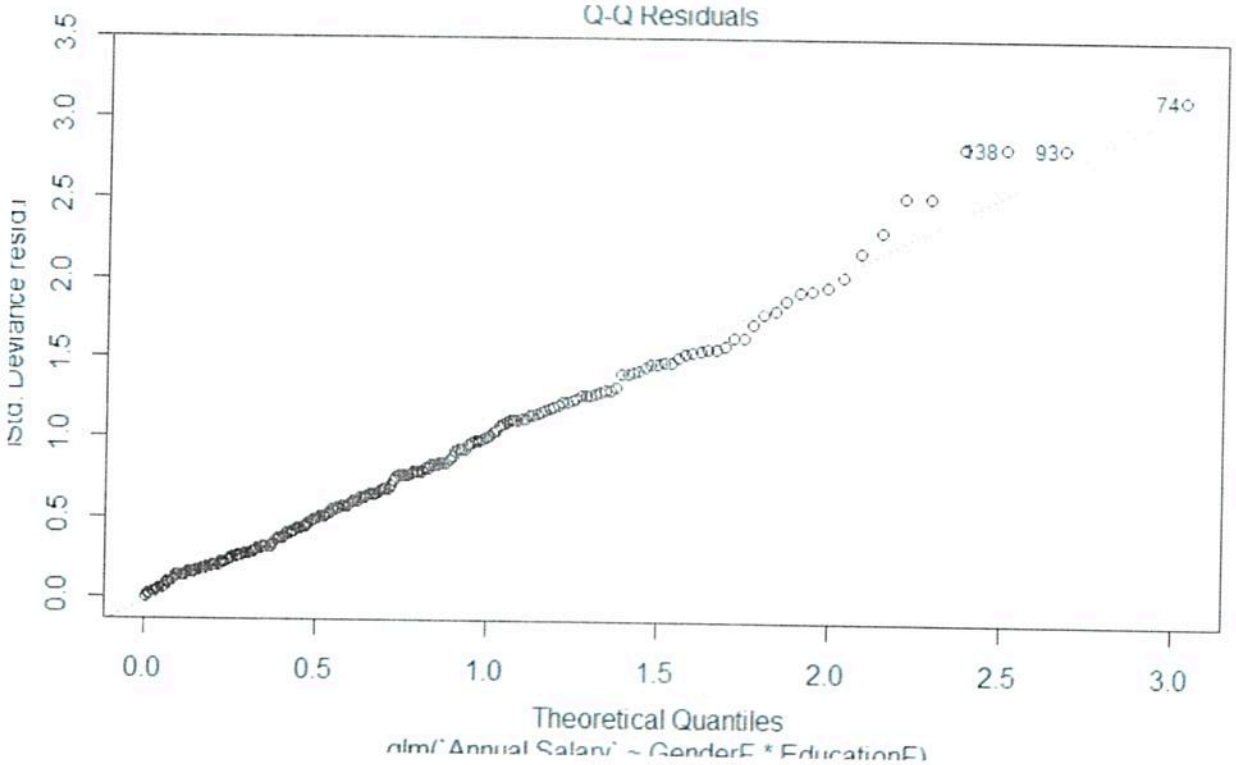
Null deviance: 1.8579e+11 on 203 degrees of freedom
Residual deviance: 1.1006e+11 on 194 degrees of freedom
AIC: 4702.6

Number of Fisher Scoring iterations: 2

(Intercept)	GenderF1	EducationF2	EducationF4
77800.000	-50375.000	-29255.556	-3997.826

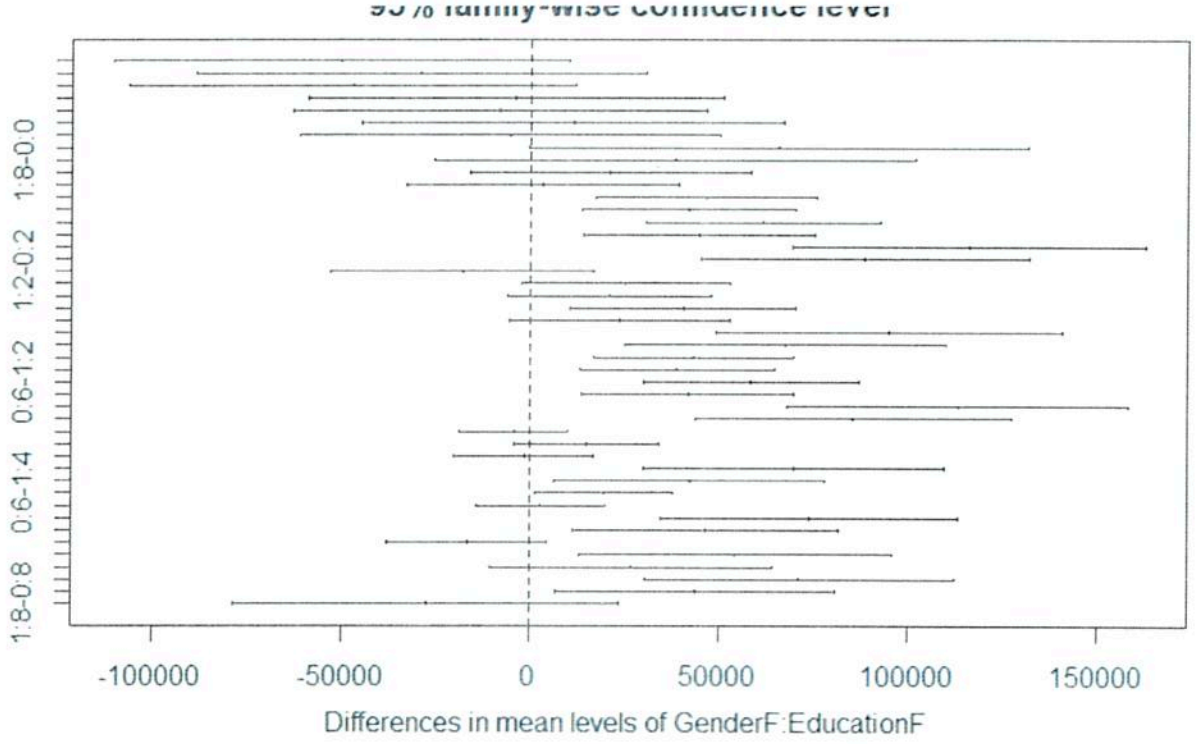
EducationF6	EducationF8	GenderF1:EducationF2	GenderF1:EducationF4
11175.000	65675.000	32270.556	45980.072

GenderF1:EducationF6	GenderF1:EducationF8
33637.037	22900.000



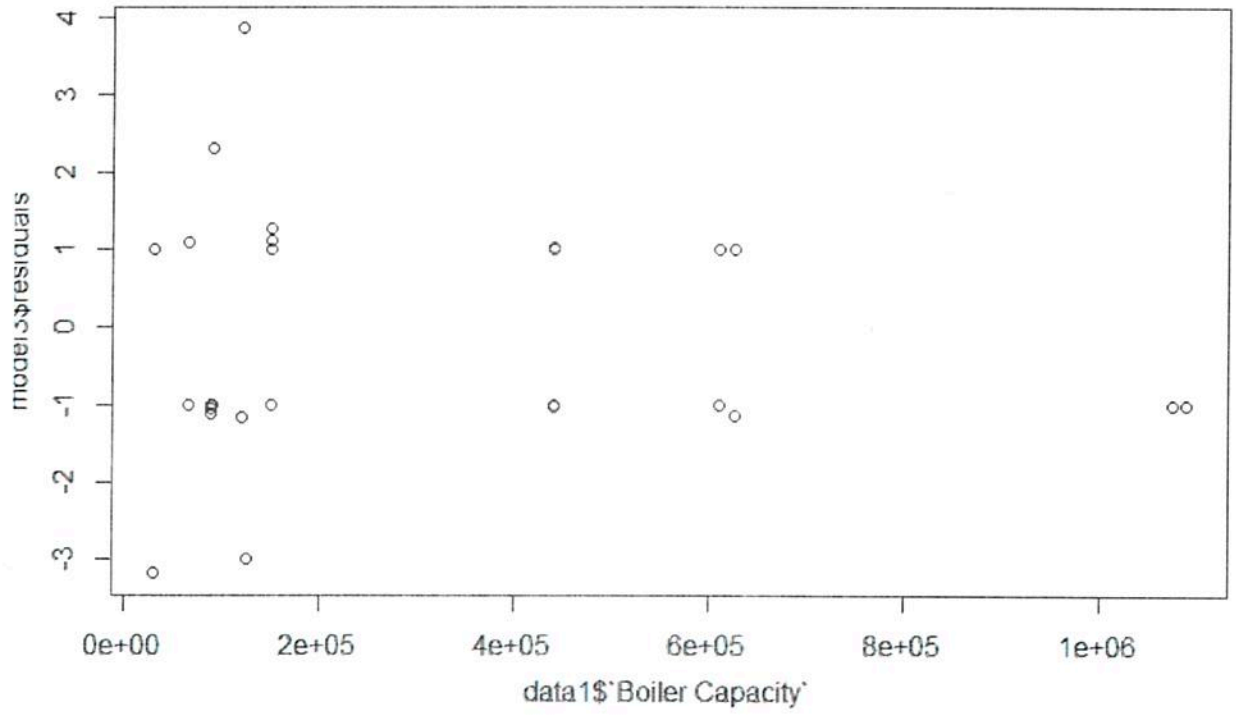
As ANOVA:

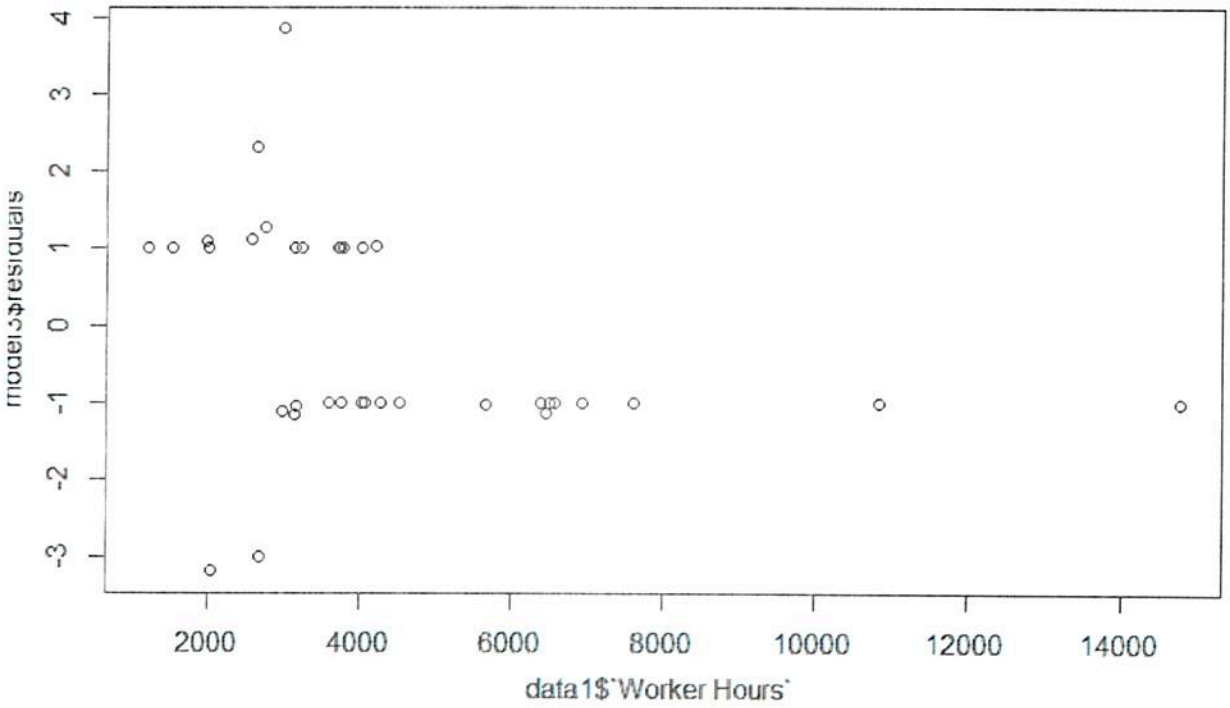
```
                    Df    Sum Sq   Mean Sq F value   Pr(>F)
GenderF              1 8.219e+09 8.219e+09  14.486 0.000189 ***
EducationF           4 6.257e+10 1.564e+10  27.572  < 2e-16 ***
GenderF:EducationF   4 4.934e+09 1.233e+09   2.174 0.073375 .
Residuals          194 1.101e+11 5.673e+08
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Logistic model:

```
Call:
glm(formula = DType ~ `Boiler Capacity` + `Worker Hours`, family = "binomial"
,
    data = data1)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.068e+01  4.750e+00   2.250   0.0245 *
`Boiler Capacity` 3.465e-05  1.758e-05   1.972   0.0487 *
`Worker Hours`   -5.343e-03  2.397e-03  -2.229   0.0258 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 49.461  on 35  degrees of freedom
Residual deviance: 10.770  on 33  degrees of freedom
AIC: 16.77

Number of Fisher Scoring iterations: 8
```





Confusion Matrix and Statistics

	FALSE	TRUE
FALSE	18	2
TRUE	2	14

Accuracy : 0.8889

```
                95% CI : (0.7394, 0.9689)
    No Information Rate : 0.5556
    P-Value [Acc > NIR] : 1.823e-05

                  Kappa : 0.775

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9000
            Specificity : 0.8750
         Pos Pred Value : 0.9000
         Neg Pred Value : 0.8750
             Prevalence : 0.5556
         Detection Rate : 0.5000
   Detection Prevalence : 0.5556
      Balanced Accuracy : 0.8875

       'Positive' Class : FALSE
```

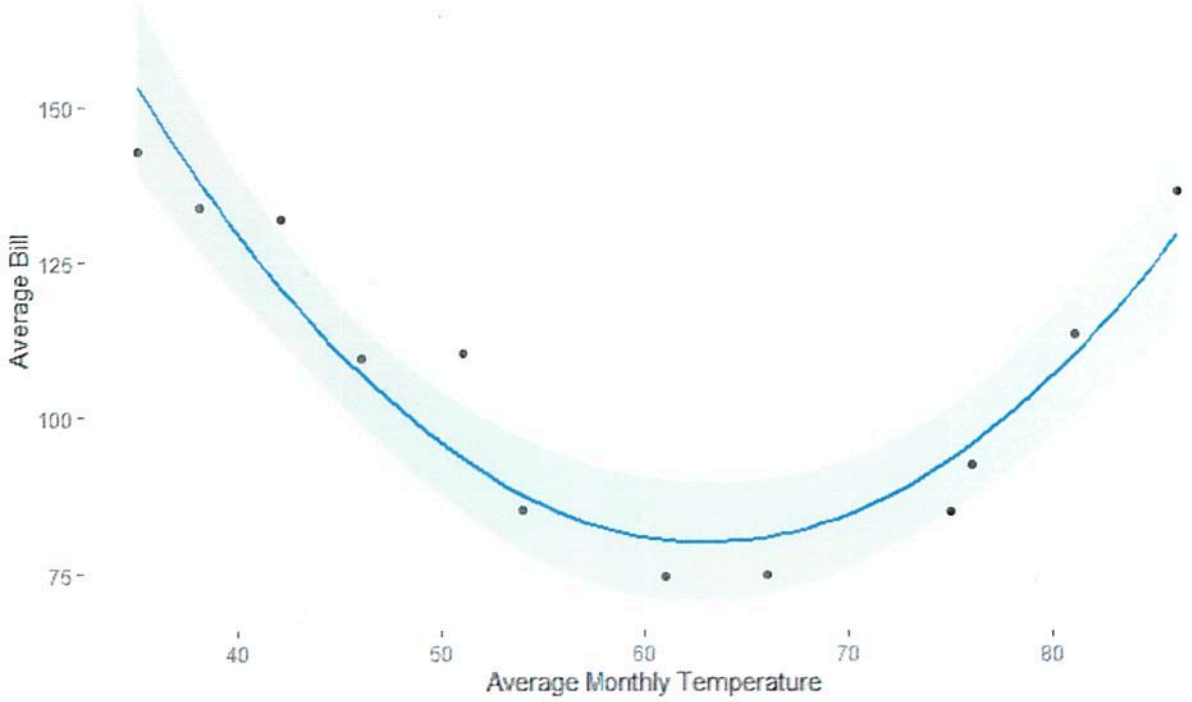
Nonlinear model

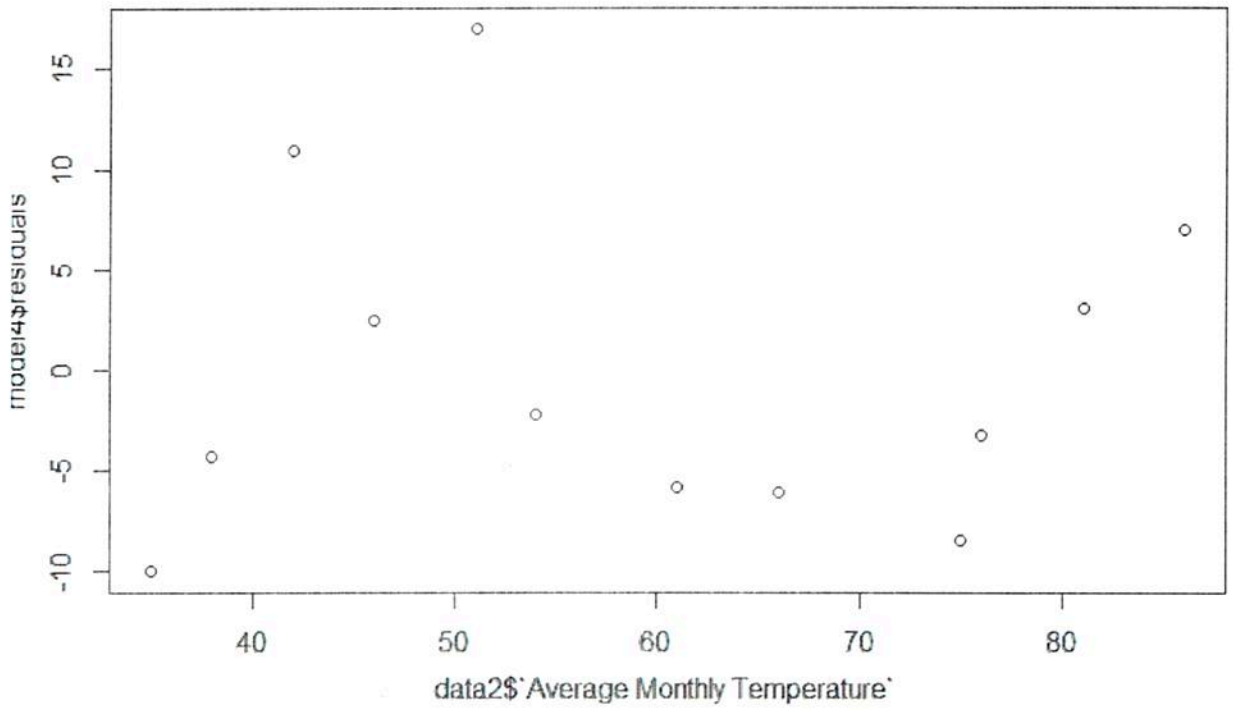
```
Call:
lm(formula = `Average Bill` ~ `Average Monthly Temperature` +
    I(`Average Monthly Temperature`^2), data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-10.029  -5.871  -2.778   4.033  16.944

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     449.44983   40.69830  11.043 1.56e-06 ***
`Average Monthly Temperature`   -11.73135    1.43332  -8.185 1.84e-05 ***
I(`Average Monthly Temperature`^2)   0.09311    0.01184   7.867 2.53e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.106 on 9 degrees of freedom
Multiple R-squared:  0.8903,   Adjusted R-squared:  0.8659
F-statistic: 36.52 on 2 and 9 DF,  p-value: 4.798e-05
```

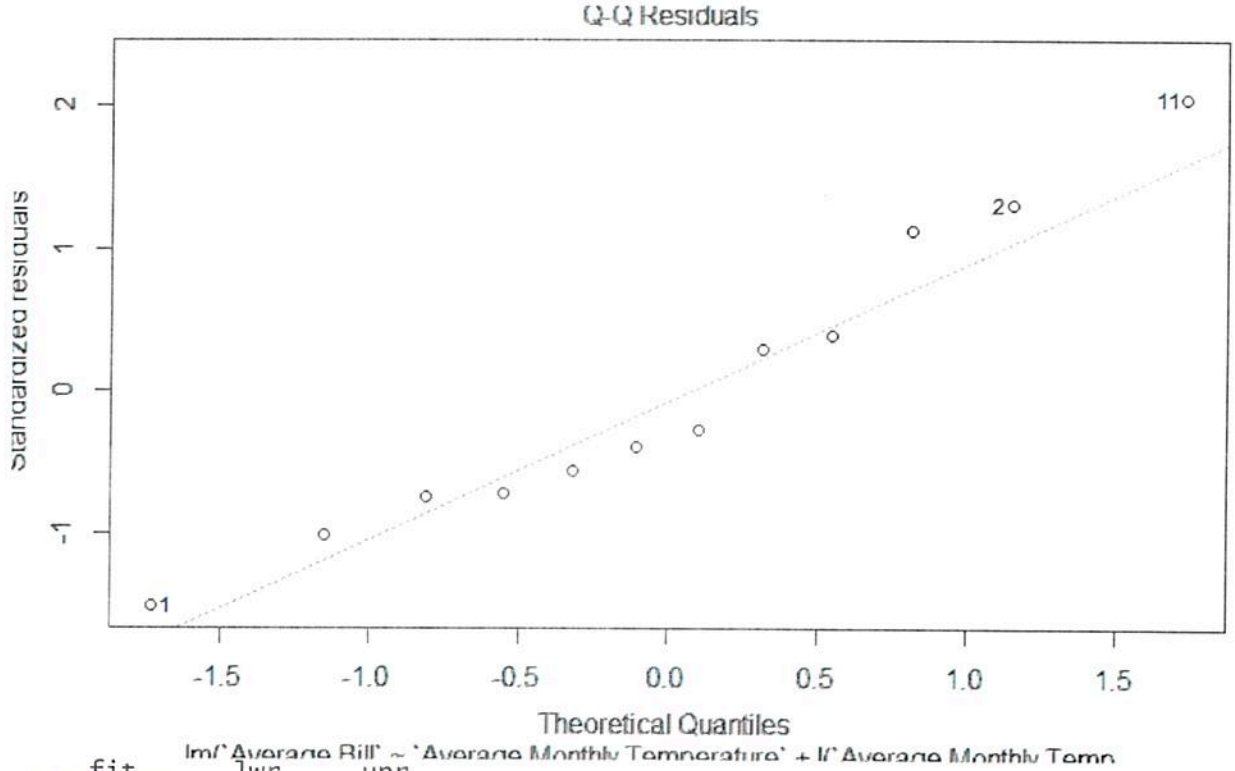
Vs. poly version

```
Call:
lm(formula = `Average Bill` ~ poly(`Average Monthly Temperature`,
    2), data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-10.029  -5.871  -2.778   4.033  16.944

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                            107.408      2.629  40.859 1.57e-11
***
poly(`Average Monthly Temperature`, 2)1  -30.409      9.106  -3.339  0.00867
**
poly(`Average Monthly Temperature`, 2)2   71.637      9.106   7.867 2.53e-05
***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.106 on 9 degrees of freedom
Multiple R-squared:  0.8903,   Adjusted R-squared:  0.8659
```

F-statistic: 36.52 on 2 and 9 DF, p-value: 4.798e-05



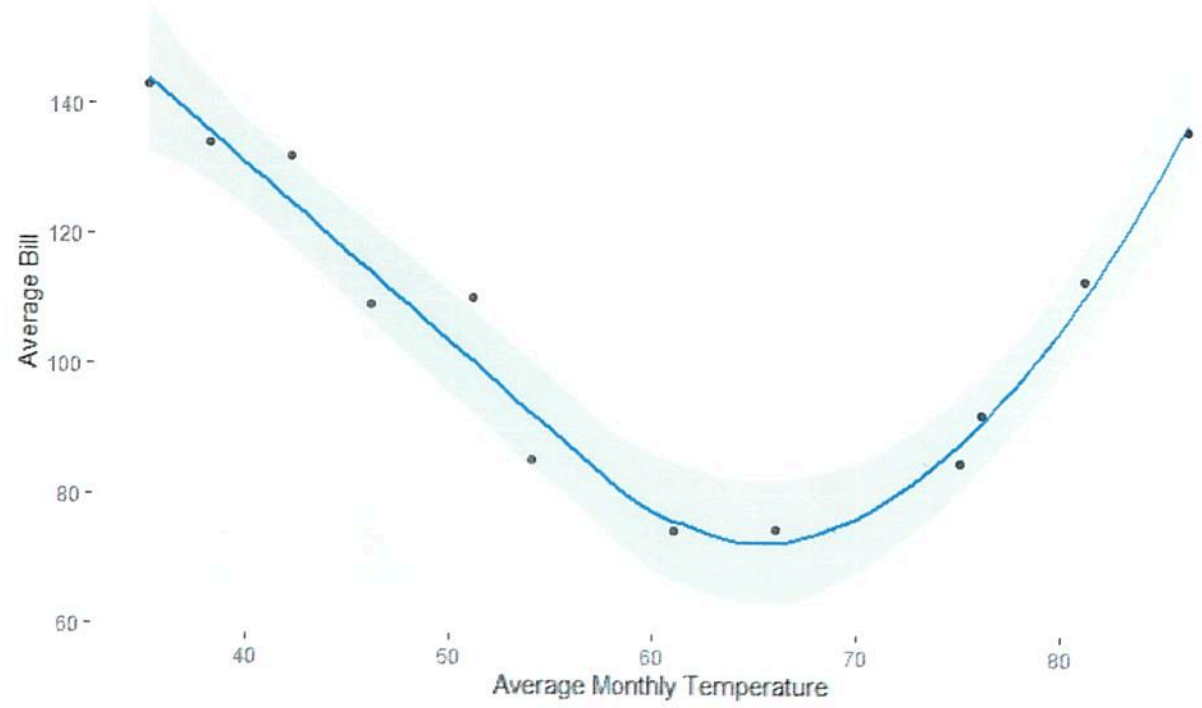




lm(`Average Bill` ~ `Average Monthly Temperature` + I(`Average Monthly Temp

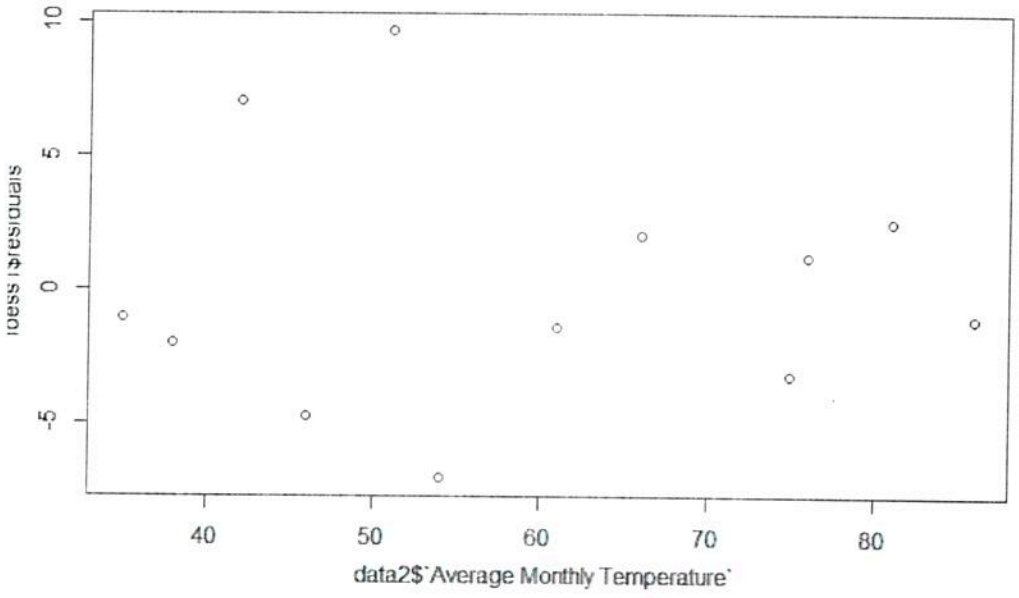
```
      fit      lwr      upr
1 158.2278 126.9777 189.478
```

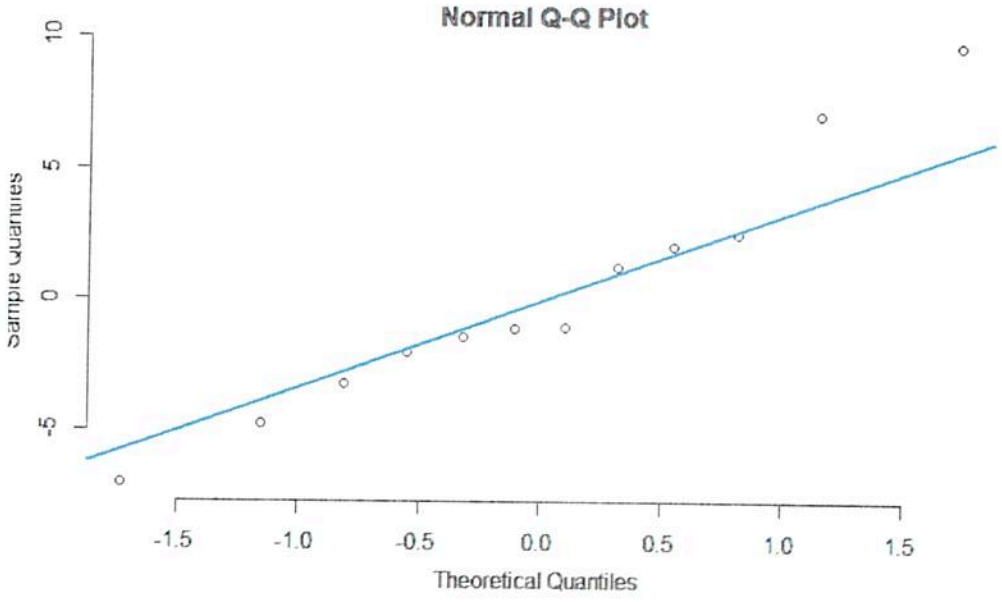
Loess:




```
Call:
loess(formula = data2$`Average Bill` ~ data2$`Average Monthly Temperature`)
```

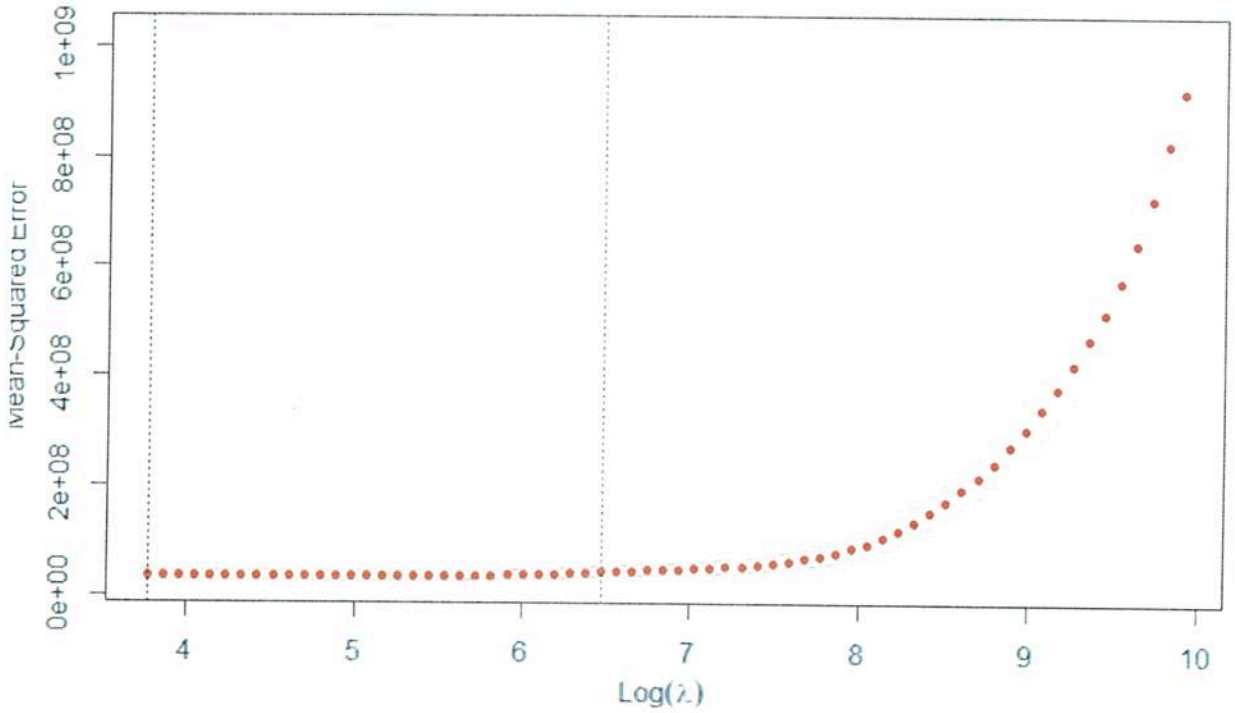
```
Number of Observations: 12
Equivalent Number of Parameters: 4.34
Residual Standard Error: 5.996
Trace of smoother matrix: 4.78  (exact)

Control settings:
  span      :  0.75
  degree    :  2
  family    :  gaussian
  surface   :  interpolate        cell = 0.2
  normalize:  TRUE
 parametric:  FALSE
drop.square:  FALSE
```





Loess will not extrapolate outside the range of the original data

LASSO:

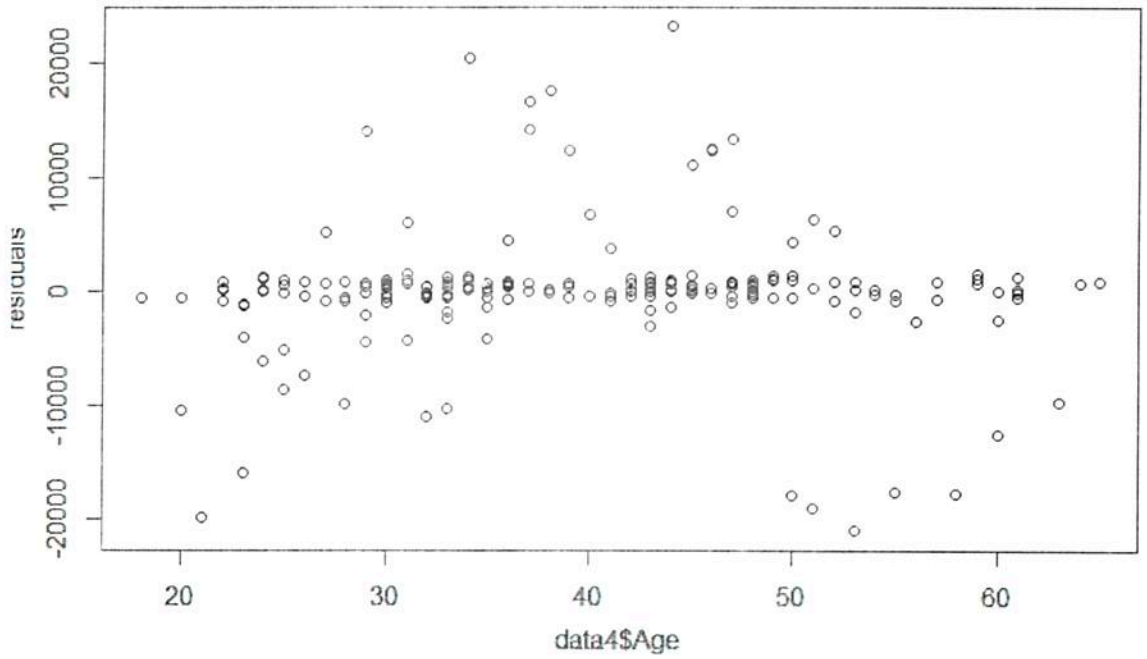


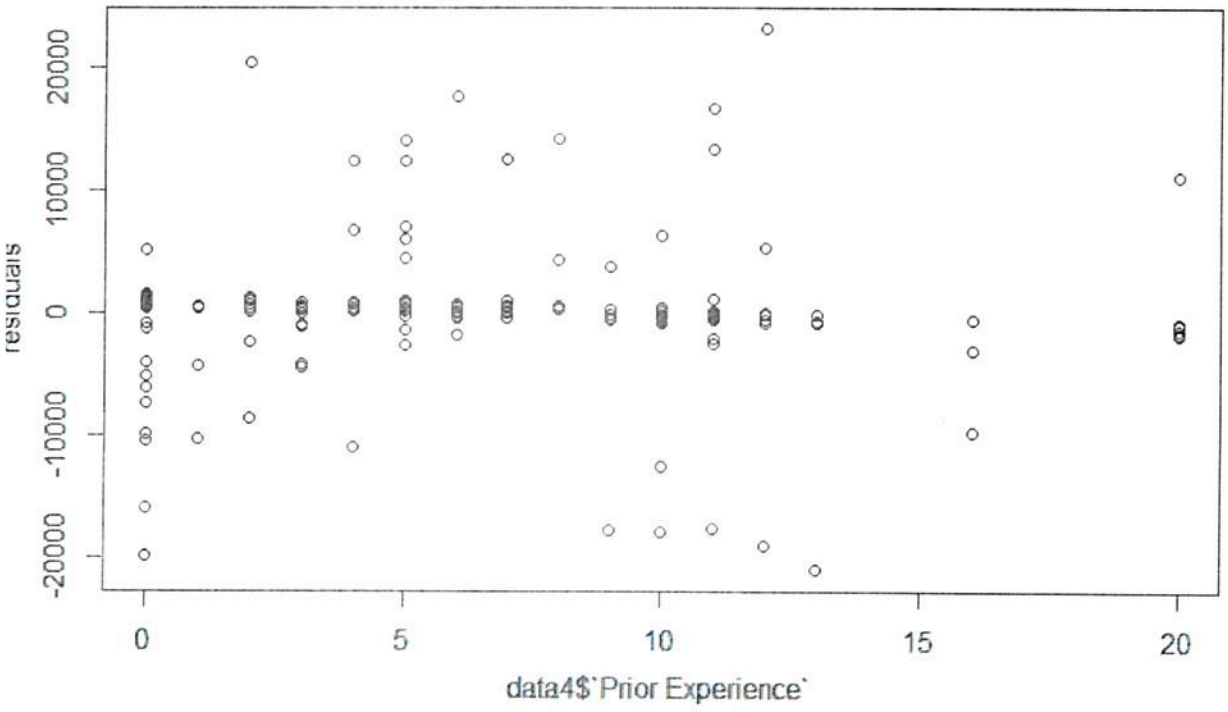

```
best_lambda
[1] 43.54158

9 x 1 sparse Matrix of class "dgCMatrix"
                          s0
(Intercept)       19366.19333
Gender            -8082.68835
Age                 -26.10891
Prior Experience   2940.51022
Beta Experience    2601.10385
Associates        -6461.19774
Bachelors         13849.39023
Masters           25281.81692
Doctorate         63633.76154

rsq
[1] 0.9652331
```

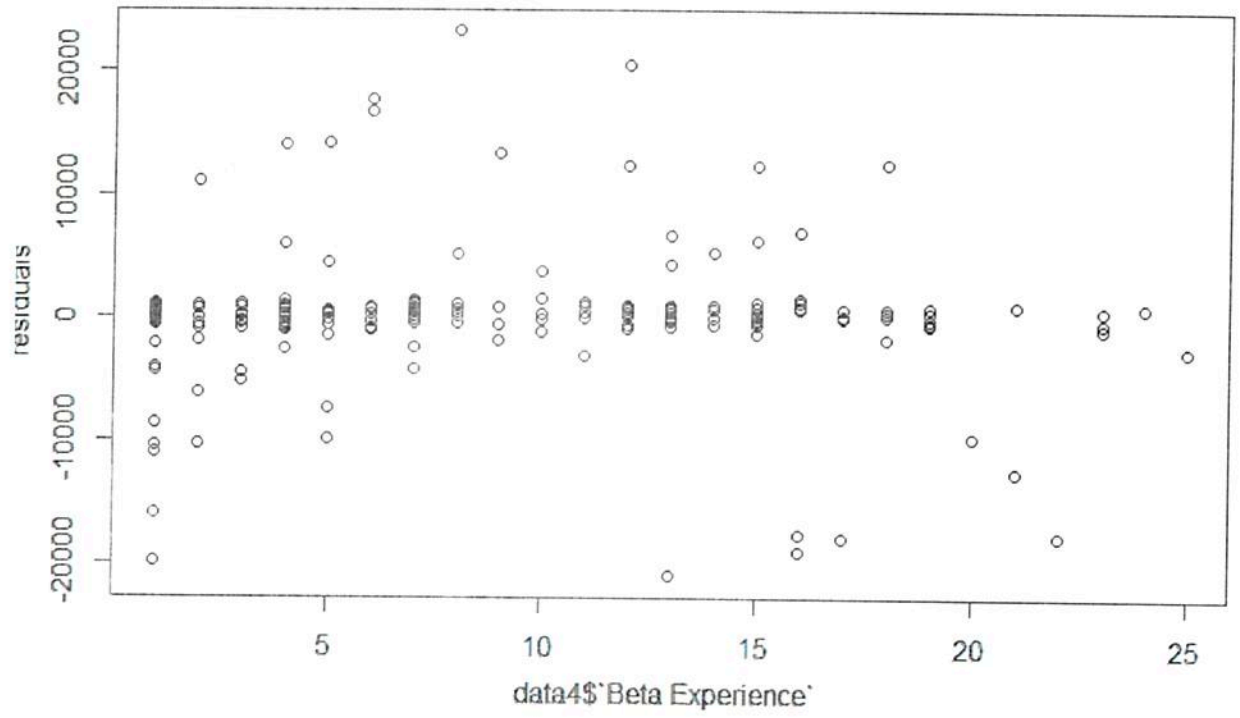
Linear model with same variables:

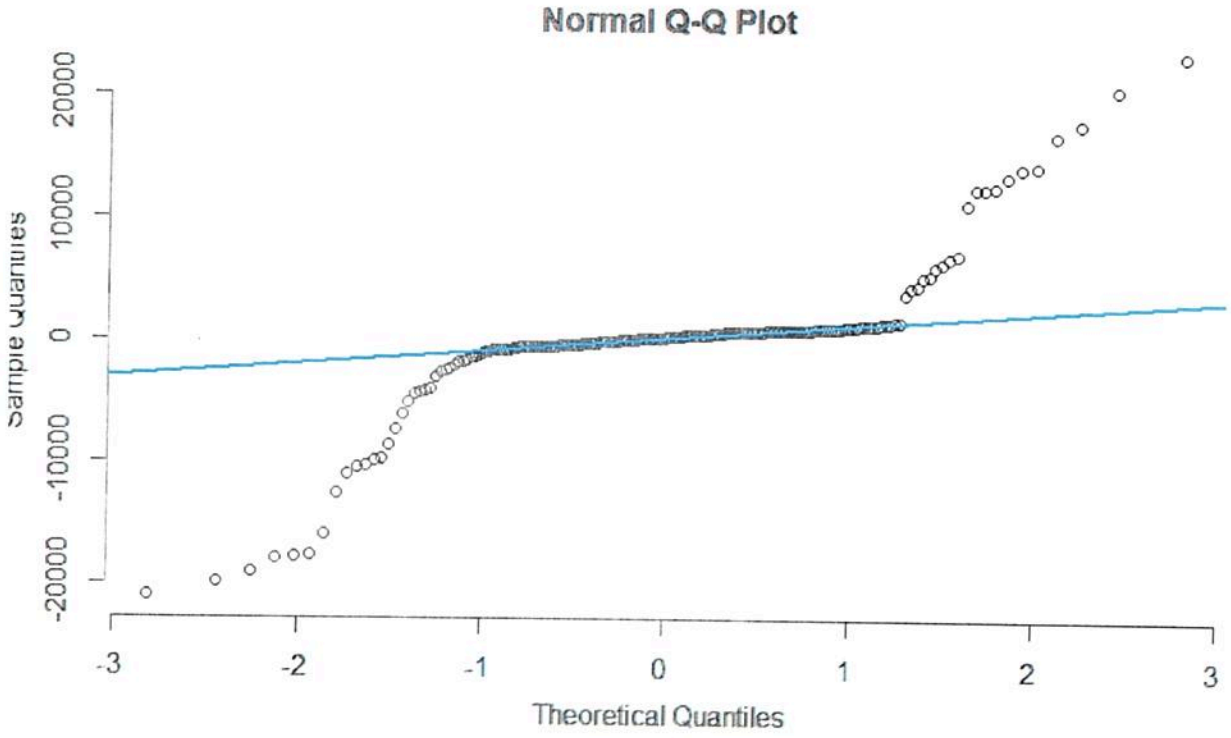
```
Call:
lm(formula = `Annual Salary` ~ ., data = data4)

Residuals:
     Min       1Q   Median       3Q      Max
-23286.6   -858.5   -254.0    442.7  20952.1

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       18782.22    2229.31   8.425 7.81e-15 ***
Gender            -8105.39     832.30  -9.739  < 2e-16 ***
Age                 -35.52      40.84  -0.870   0.3854
`Prior Experience` 2949.23      88.79  33.217  < 2e-16 ***
`Beta Experience`  2606.36      68.58  38.007  < 2e-16 ***
```

```
Associates          -5696.71    2303.14  -2.473   0.0142 *
Bachelors           14763.40    1947.67   7.580 1.37e-12 ***
Masters             26225.80    2035.54  12.884  < 2e-16 ***
Doctorate           64685.48    2692.14  24.028  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5751 on 195 degrees of freedom
Multiple R-squared:  0.9653,   Adjusted R-squared:  0.9639
F-statistic: 677.8 on 8 and 195 DF,  p-value: < 2.2e-16
```






residuals
20000
10000
0
-10000
-20000
5
10
15
20
25
data4$`Beta Experience`


Normal Q-Q Plot
Sample Quantiles
20000
10000
0
-10000
-20000
-3
-2
-1
0
1
2
3
Theoretical Quantiles