MTH 325, Exam #2, At Home, Spring 2023 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 1 is data on boilers including boiler type, drum type, design pressure, capacity and worker hours. Use this data to predict worker hours using boiler type and drum type 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.

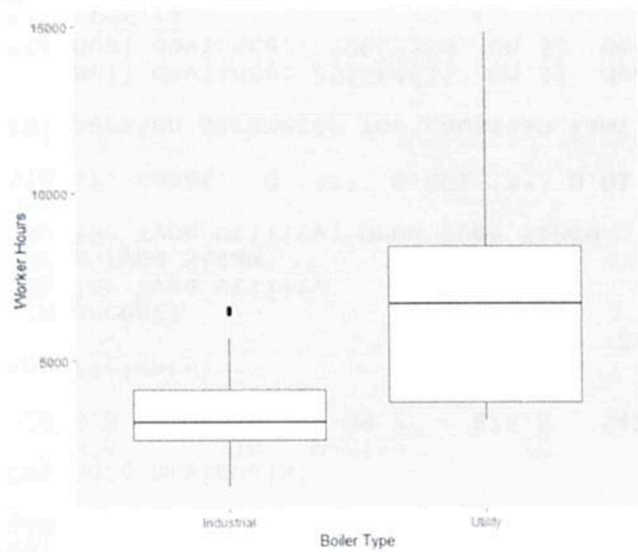
2. Using the same data on Sheet 1, (after eliminating the Boiler number column) create a logistic regression model that predicts Boiler Type from Worker Hours. Plot the graph. Create appropriate exploratory graphs. Create appropriate diagnostic plots, and a confusion matrix.

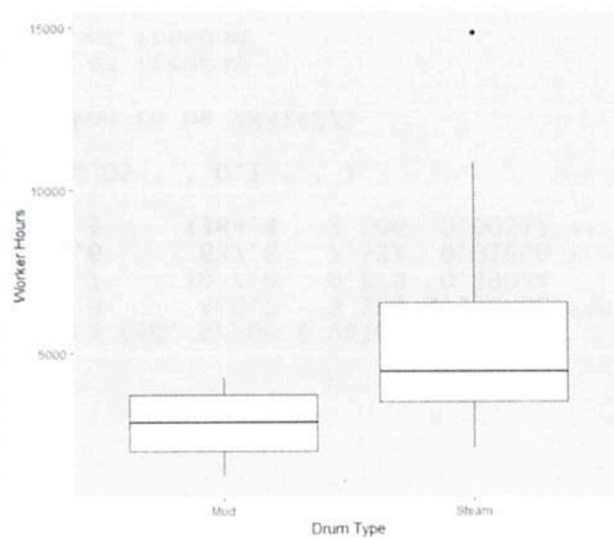
3. Create a graph of the data on Sheet 2 with Average Monthly Temperature on the horizontal axis, and Average Monthly Bill on the vertical axis. Create a nonlinear model for the data by transforming variables. Plot the resulting model. Create appropriate diagnostic plots. Bonus points for comparing your model to a LOESS model.

4. On Sheet 3 is employee data. Eliminate the Employee column. Gender is already encoded as a binary dummy variable. You'll need to encode the Department variable as separate dummy variables. The rest of the variables are numerical. Use LASSO regression on data to find a model of best fit. Compare the resulting model to a model using a linear model with the same variables. Prepare appropriate diagnostics and diagnostic graphs.

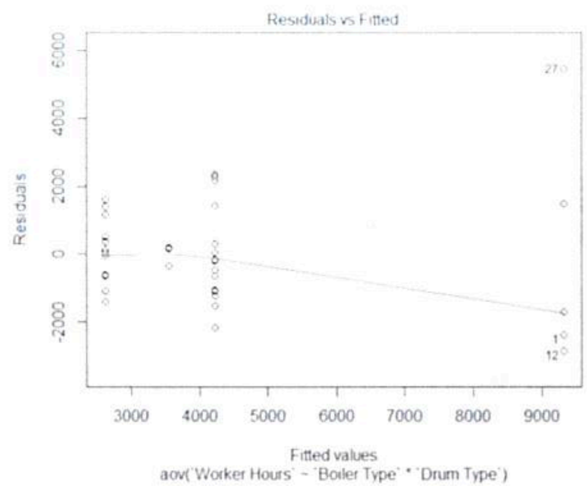
MTH 325 Exam #2 At Home analysis

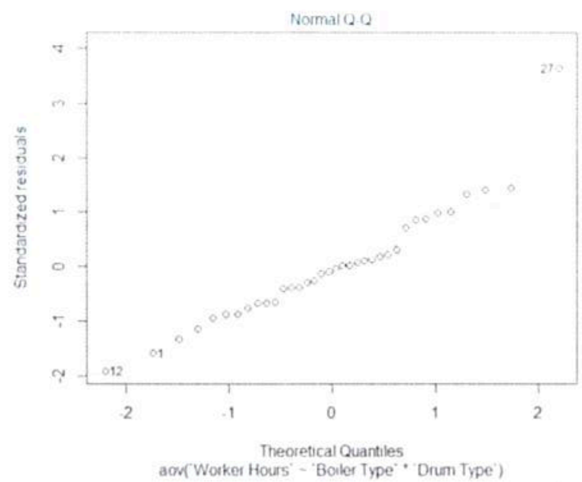
1.



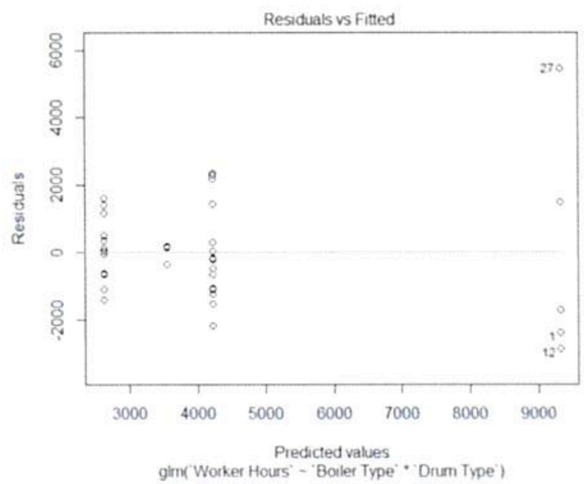


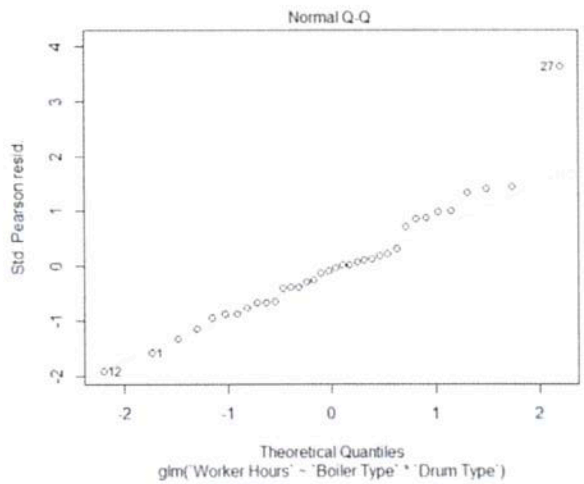

```
                         Df   Sum Sq  Mean Sq F value    Pr(>F)    
`Boiler Type`             1 84390625 84390625  29.805 5.22e-06 ***
`Drum Type`               1 55084671 55084671  19.455 0.000109 ***
`Boiler Type`:`Drum Type` 1 25583814 25583814   9.036 0.005116 ** 
Residuals                32 90605504  2831422                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                (Intercept)                         `Boiler Type`Utility 
                  2607.9231                                     937.7436 
           `Drum Type`Steam `Boiler Type`Utility:`Drum Type`Steam 
                  1613.6103                                    4161.5231 
```






```
Call:
glm(formula = `Worker Hours` ~ `Boiler Type` * `Drum Type`, data = data1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2866.8  -1086.6    -99.2    675.8   5470.2

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                              2607.9      466.7   5.588 3.59e-06 ***
`Boiler Type`Utility                      937.7     1077.8   0.870  0.39074
`Drum Type`Steam                         1613.6      637.6   2.531  0.01650 *
`Boiler Type`Utility:`Drum Type`Steam    4161.5     1384.4   3.006  0.00512 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 2831422)

    Null deviance: 255664615  on 35  degrees of freedom
Residual deviance:  90605504  on 32  degrees of freedom
AIC: 642.75

Number of Fisher Scoring iterations: 2
```

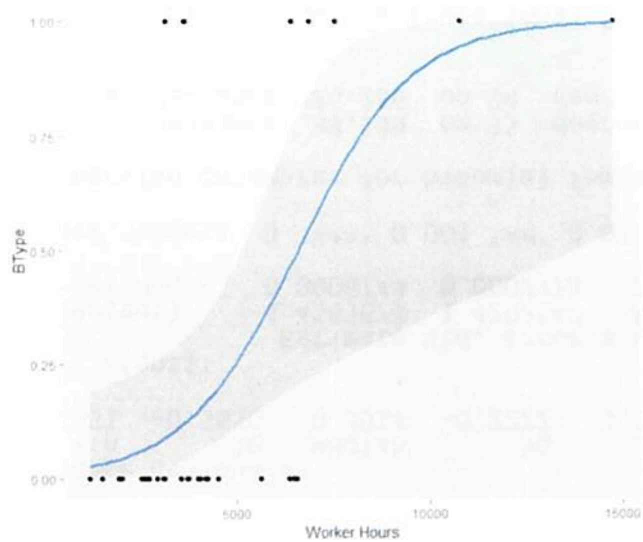


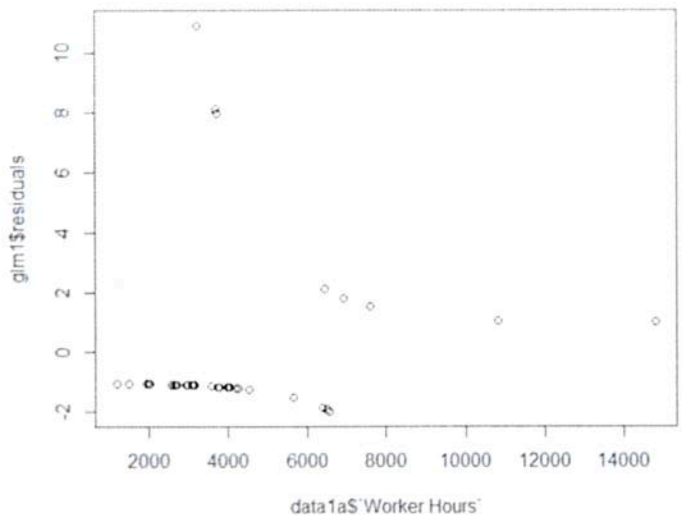


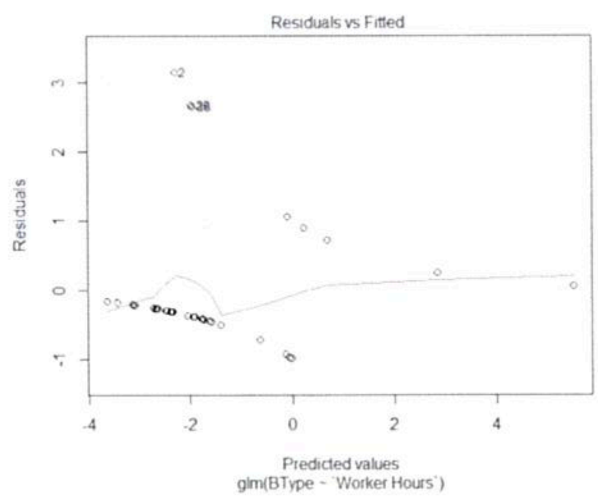
2.

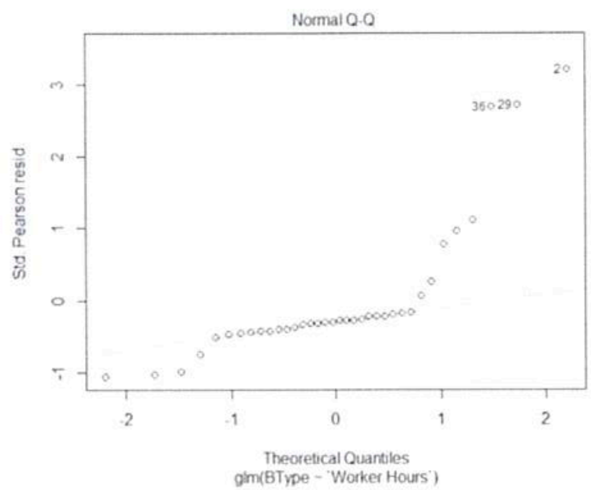
```
Call:
glm(formula = BType ~ `Worker Hours`, family = binomial, data = data1a)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1651  -0.5656  -0.3914  -0.2272   2.1851

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.4565673  1.4202147  -3.138   0.0017 **
`Worker Hours`  0.0006744  0.0002723   2.477   0.0132 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38.139  on 35  degrees of freedom
Residual deviance: 26.286  on 34  degrees of freedom
AIC: 30.286

Number of Fisher Scoring iterations: 5
```

Note, including other variables creates p-values =1, and z-scores near zero.

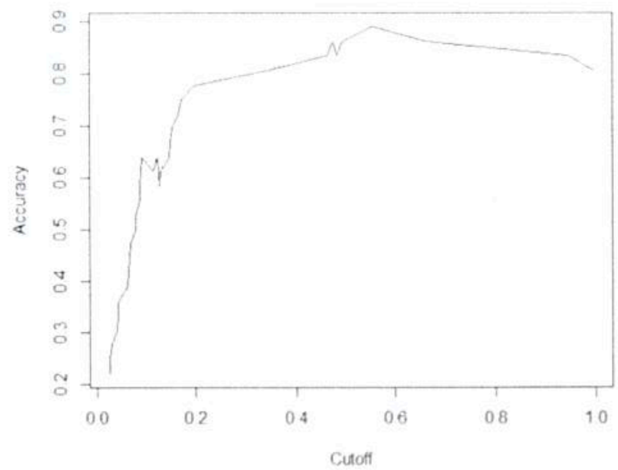

BType
Worker Hours


glm1$residuals
data1a$`Worker Hours`

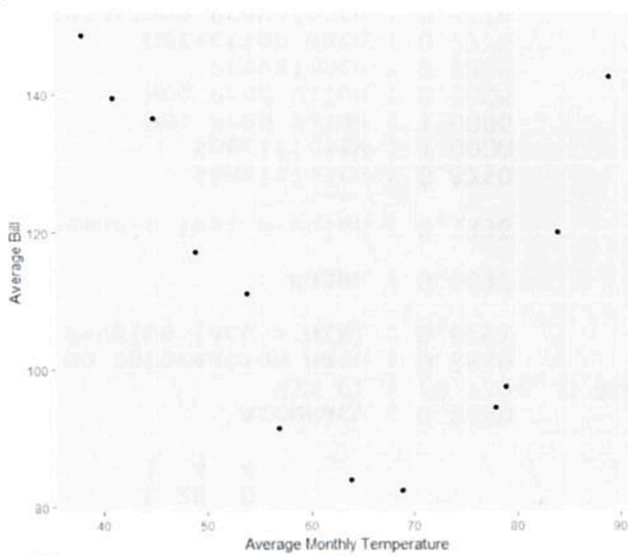

Residuals vs Fitted
Residuals
Predicted values
glm(BType ~ `Worker Hours`)

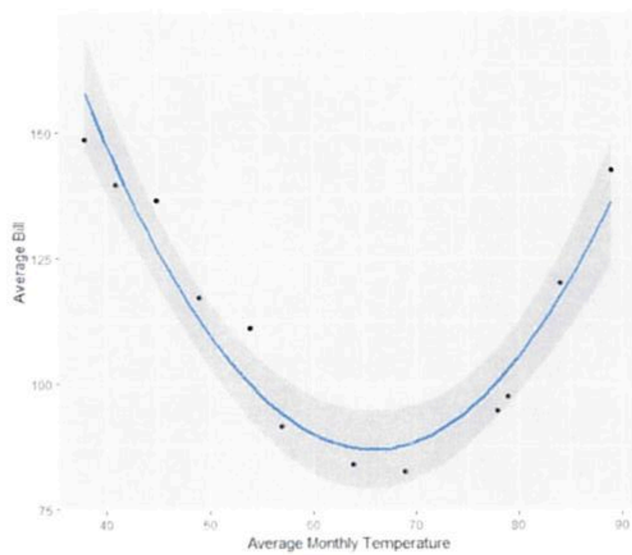

Normal Q-Q
Std. Pearson resid.
Theoretical Quantiles
glm(BType ~ `Worker Hours`)

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 28  0
         1  4  4

               Accuracy : 0.8889
                 95% CI : (0.7394, 0.9689)
    No Information Rate : 0.8889
    P-Value [Acc > NIR] : 0.6291

                  Kappa : 0.6087

 Mcnemar's Test P-Value : 0.1336

            Sensitivity : 0.8750
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.5000
             Prevalence : 0.8889
         Detection Rate : 0.7778
   Detection Prevalence : 0.7778
      Balanced Accuracy : 0.9375

       'Positive' Class : 0
```



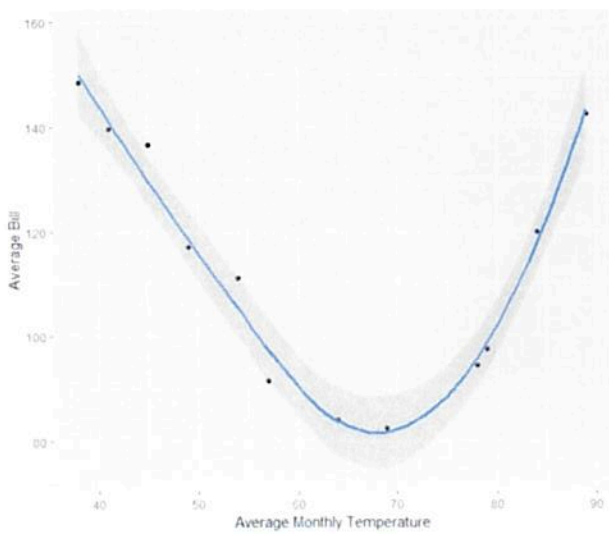
3.

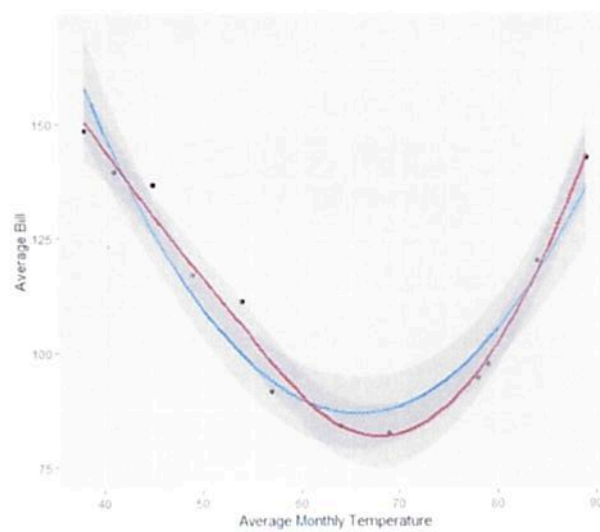





```
Call:
lm(formula = `Average Bill` ~ `Average Monthly Temperature` +
```

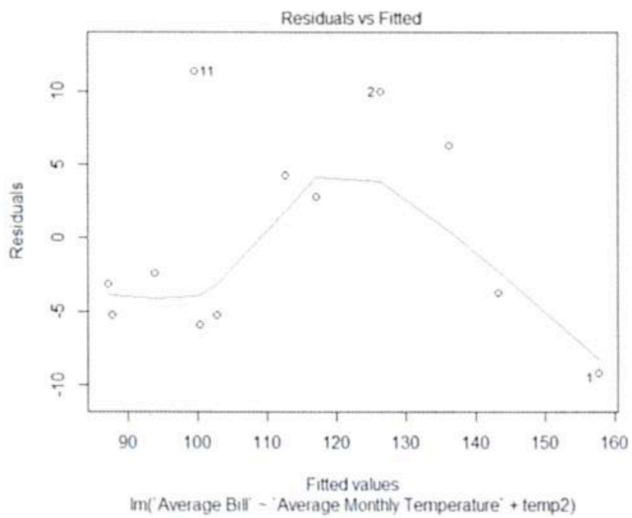
```
    temp2, data = data2)

Residuals:
   Min     1Q Median     3Q    Max
-9.219 -5.239 -2.744  4.811 11.428

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   484.107572  36.956206  13.099 3.64e-07 ***
`Average Monthly Temperature` -12.076035   1.233305  -9.792 4.26e-06 ***
temp2                           0.091760   0.009706   9.454 5.70e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.468 on 9 degrees of freedom
Multiple R-squared:  0.9199,	Adjusted R-squared:  0.9021
F-statistic: 51.66 on 2 and 9 DF,  p-value: 1.167e-05
```

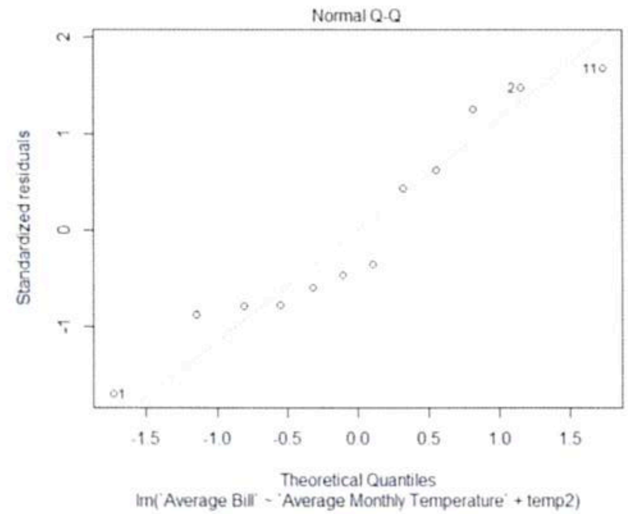


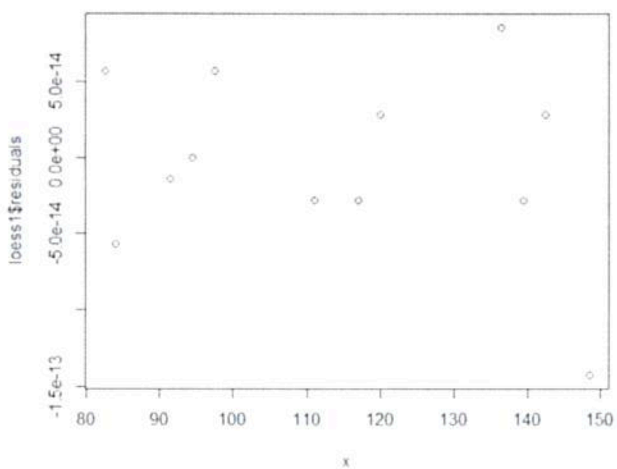


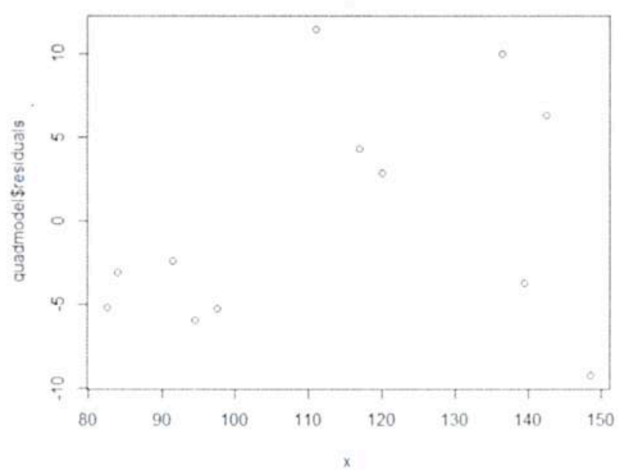

```
Call:
loess(formula = y ~ x)

Number of Observations: 12
Equivalent Number of Parameters: 4.45
Residual Standard Error: 7.892e-14
Trace of smoother matrix: 4.9  (exact)
```

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





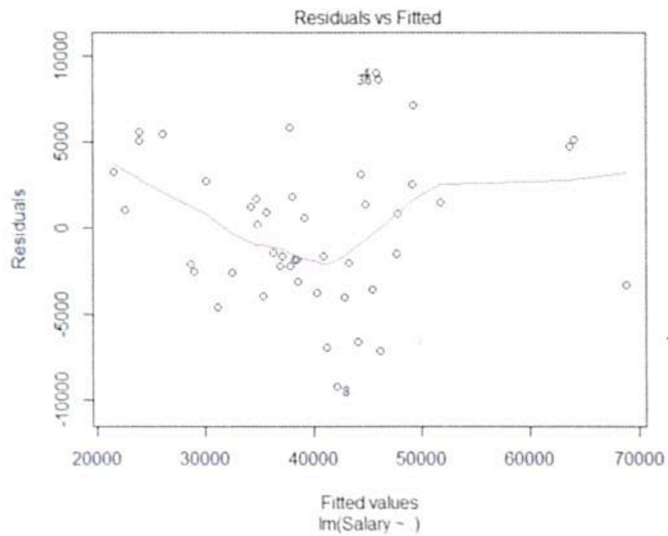




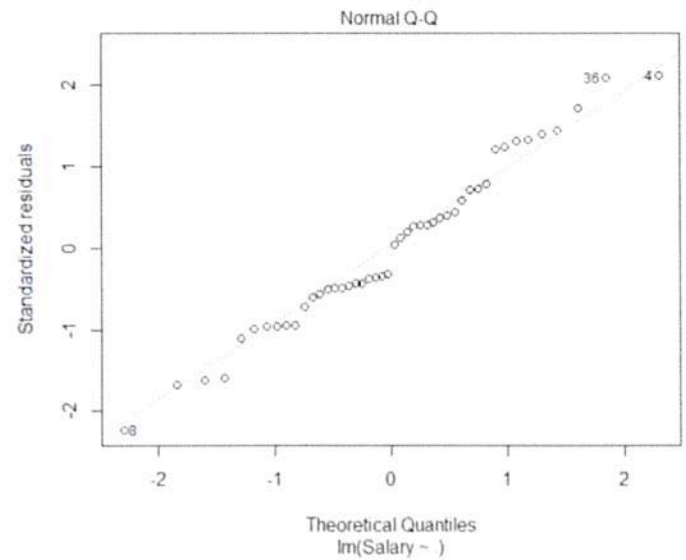
4.

```
9 x 1 sparse Matrix of class "dgCMatrix"
                                  s0
(Intercept)              19400.09357
```

```
`Years Previous Experience`    -68.05517
`Years Employed`               707.67883
`Years Education`             1543.66760
Gender                       -1993.19030
`Number Supervised`            129.00709
Dept2                         8296.70232
Dept3                         4912.70929
Dept4                         7978.53047

Call:
lm(formula = Salary ~ ., data = data3)

Residuals:
    Min      1Q  Median      3Q     Max 
-9197.7 -2541.2  -610.1  2687.0  9024.4 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 19313.19    2518.57   7.668 3.72e-09 ***
`Years Previous Experience`   -72.80     198.39  -0.367 0.715738    
`Years Employed`              709.45     120.96   5.865 9.56e-07 ***
`Years Education`            1544.52     338.22   4.567 5.33e-05 ***
Gender                      -2040.25    1448.97  -1.408 0.167458    
`Number Supervised`           130.17      81.68   1.594 0.119522    
Dept2                        8455.63    2288.73   3.694 0.000709 ***
Dept3                        5049.08    2333.00   2.164 0.036975 *  
Dept4                        8096.05    1830.64   4.423 8.26e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4650 on 37 degrees of freedom
Multiple R-squared:  0.8531,	Adjusted R-squared:  0.8213 
F-statistic: 26.85 on 8 and 37 DF,  p-value: 3.573e-13
```



Residuals vs Fitted
Residuals
Fitted values
lm(Salary ~)



Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Salary ~)

MTH 325, Exam #2, Spring 2023 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 boilers 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 binary variables.

$\hat{y} = 2607.92 + 937(\text{Utility}) + 1613(\text{Steam}) + 4161(\text{Utility} * \text{Steam})$

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

answers may vary
boxplots appear to show a difference, coeffs significant in ANOVA
normality of residuals is mostly ok w/ one outlier

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.*

$\ln\left(\frac{p}{1-p}\right) = -4.4565673 + 0.0006744(\text{WorkerHours})$

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

the log odds varies by positive 0.00067 for each additional worker hour (more hours increases likelihood of the boiler being type 1)

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

the null deviance compares intercept only model to a "perfect" model. residual deviance uses the full model compared to a "perfect model". The residual deviance is lower which means Worker Hours improves the prediction.

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?

the # of objects in class 1 is much smaller and there are some resulting unbalanced predictions.

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.

quadratic / polynomial model. it is u-shaped

Monthly Bill = 484.1 − 12 temp + 0.09 temp^2

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

91.99%

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

7.468

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

they seem okay plotted against temps
possibly problematic against y
outliers are not strong

12. (Bonus) Describe the LOESS model and compare it to your polynomial model. Describe any advantages or disadvantages to this model.

Loess model has a much smaller residual standard error

Use the employee data to answer the following questions.

13. Write the equation of your LASSO model below.

Salary = $\hat{y}$ = 19,400 − 68 (yrs exp) + 707 (empl) + 1543 (ed) − 1993 (gender)
+ 129 (#Sup) + 8296 Dept2 + 4912 Dept3 + 7978 Dept4

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

Salary = $\hat{y}$ = 19,313 − 72 (yrs exp) + 709 (emp) + 1544 (ed) − 2040 Gender
+ 130 (# Sup) + 8455 Dept2 + 5049 Dept3 + 8096 Dept 4

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

they are very similar but not identical

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.

Yrs Previous Experience, Gender, Number Supervised
do not pass significance. I would remove them
in favor of a simpler model. P-values are not close to 5%.

17. Even though the departments were encoded as ordinal variables, why could be not analyze them in the model this way?

they are not ordered and the diffs between them is not constant.
dept2 & 4 are more similar than 3.

Part II:

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

usually this is done to test effect of multiple variables in
a preliminary way before do more levels in a later analysis

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

we may be interested in looking at groups of values rather than continuous changes, such as generations in age or income levels

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

the data is split into K parts w/ 1 part as test and k-1 parts as training to test the type of model making predictions, k-times. Results are averaged

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

unsupervised learning is a type of learning where y-labels/classes are unknown, but the model seeks to find patterns.

KNN, SVM, LDA are all examples of unsupervised learning

22. Describe how spline regression works in general terms.

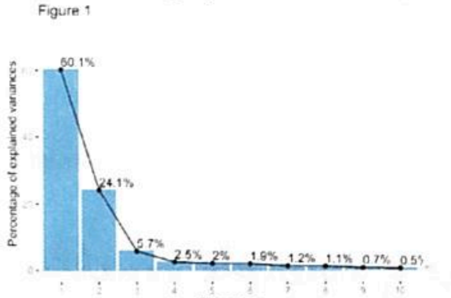
split data into sections, such as quantiles, and create a piecewise regression on each segment.

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

penalties can be used to reduce overfitting, or to enforce additional requirements like continuity in piecewise regression.
it can lead to harder to interpret models.

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

Figure 1



at least 2
no more than 3

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

Can produce more accurate models w/ less overfitting but can be computationally expensive and difficult to interpret

26. Gaussian process regression is especially useful for uncertainty quantification. What is one disadvantage of this regression method?

it is computionally expensive and can be difficult to use w/ large numbers of variables/observations