STAT 3340 Assignment 5, Fall 2024 - due Sunday, November 24 at 11:59 PM

Your name here

Banner: B00??????

==

1. The data set "fish'' has data on fish lengths, age and water temperature.

The following reads the data, centres the age variable by subtracting its mean, and calculates the square of the centred age variable

```
fish=read.csv("http://chase.mathstat.dal.ca/~bsmith/stat3340/Data/fish.csv",header=T)
age=fish$age
age=age-mean(age)
age2=age^2
length=fish$length
temp=fish$temp
```

The following fits the linear model $length = \beta_0 + \beta_1 temp + \epsilon$ and displays the summary output.

```
lm1=lm(length~temp)
summary(lm1)
```

```
##
## Call:
## lm(formula = length ~ temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2904.5  -898.5   233.7  1060.5  1520.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6517.03    2716.43   2.399   0.0216 *
## temp         -119.70      96.98  -1.234   0.2249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1339 on 37 degrees of freedom
## Multiple R-squared:  0.03955,    Adjusted R-squared:  0.01359
## F-statistic: 1.523 on 1 and 37 DF,  p-value: 0.2249
```

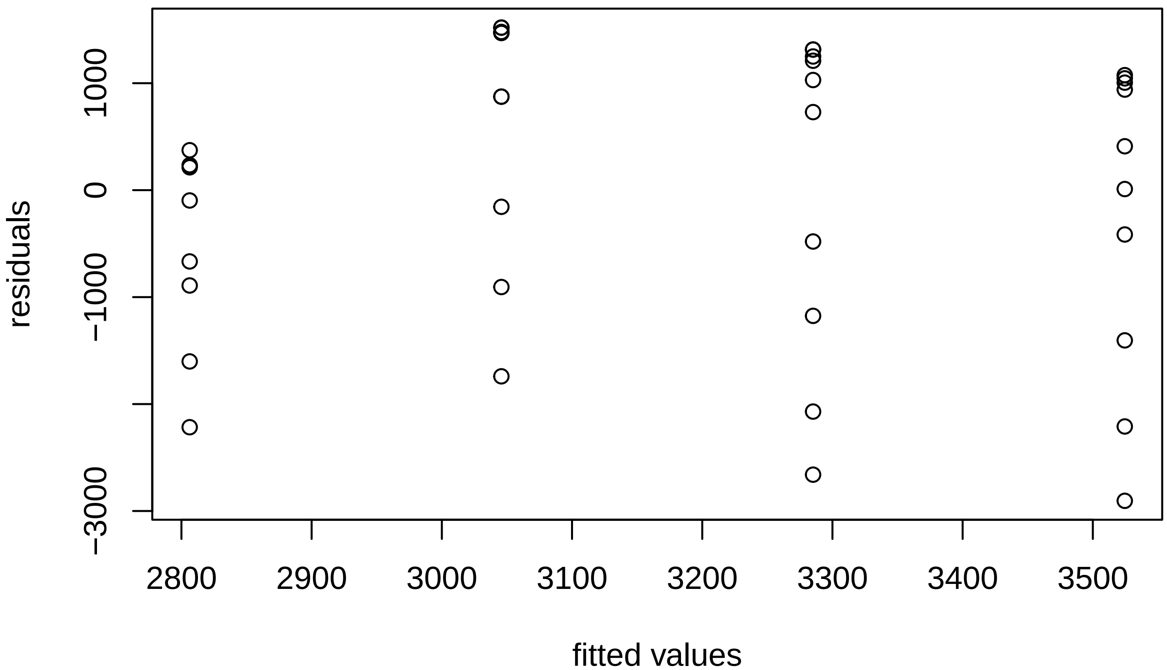
```
anova(lm1)
```

```
## Analysis of Variance Table
##
## Response: length
##           Df   Sum Sq Mean Sq F value Pr(>F)
```

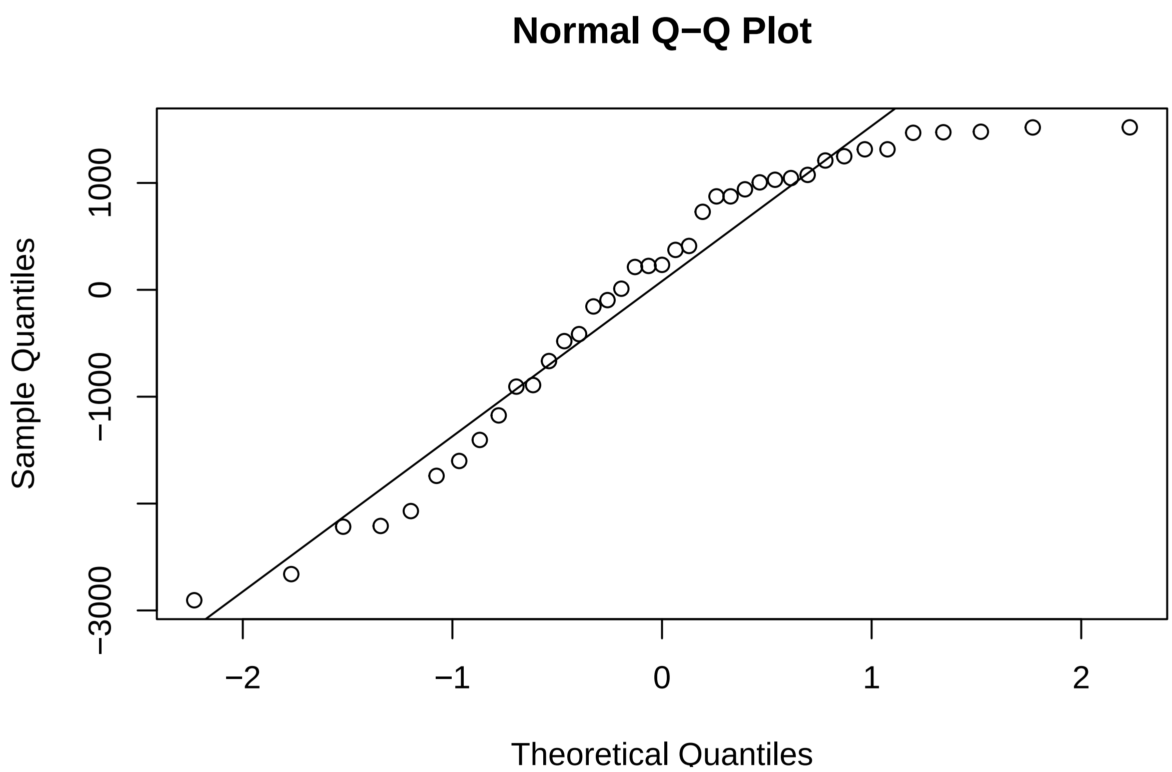
```
## temp       1  2733359 2733359  1.5235 0.2249
## Residuals 37 66384142 1794166
```

Following are a plot of residuals vs fitted values, and a normal probability plot of the residuals.

```
fit1=fitted(lm1)
e1=residuals(lm1)
plot(fit1,e1,xlab="fitted values",ylab="residuals")
```




```
qqnorm(e1)
qqline(e1)
```



- 1a) Comment briefly on the plots. Do one or more of the assumptions of the linear model appear to be violated? Which one(s)?
- 1b) Following is an added variable plot which helps to decide whether age should be added to the model, and to determine the functional form of age to use - eg. linear, quadratic, cubic ... The points on the plot are coloured according to the value of temp.

```
lm2=lm(age~temp)
summary(lm2)
```

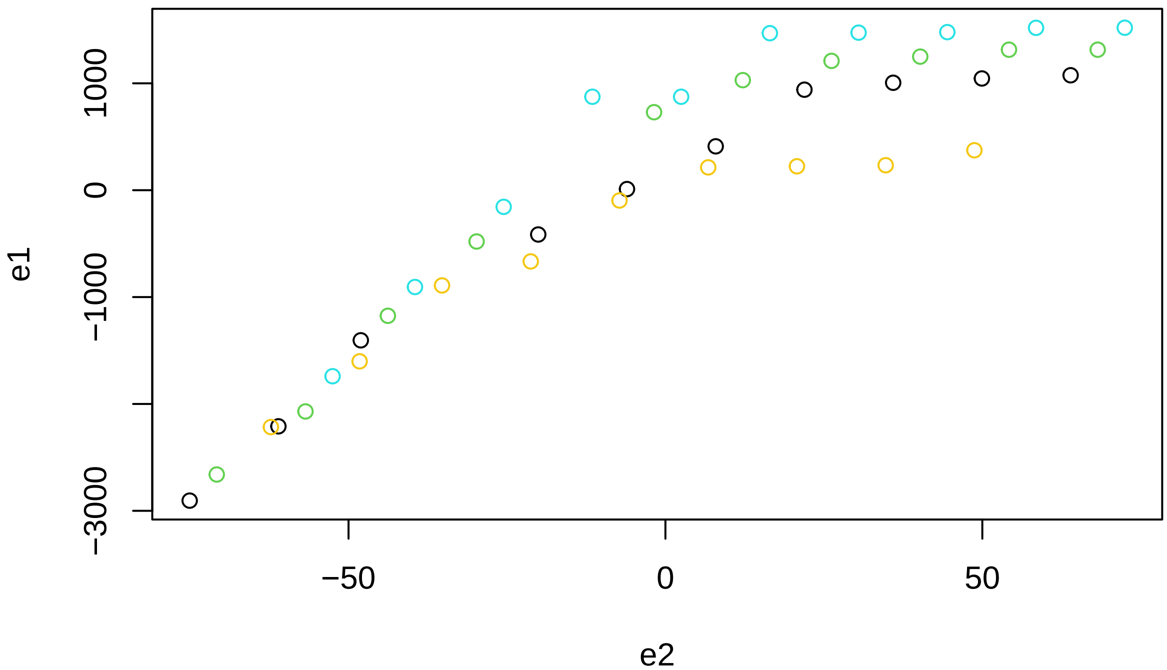
```
##
## Call:
## lm(formula = age ~ temp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -75.06 -37.38   2.48  35.34  72.48
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   59.652     88.581   0.673    0.505
## temp          -2.136      3.162  -0.676    0.504
##
## Residual standard error: 43.68 on 37 degrees of freedom
## Multiple R-squared:  0.01218,    Adjusted R-squared:  -0.01451
## F-statistic: 0.4563 on 1 and 37 DF,  p-value: 0.5035
```

```
anova(lm2)
```

```
## Analysis of Variance Table
##
## Response: age
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## temp       1    871  870.62  0.4563 0.5035
## Residuals 37  70591 1907.87
```

```r
e2=residuals(lm2)
plot(e2,e1,col=temp)
```



Which functional form of age seems more appropriate, a linear or a quadratic term?

2. In class we talked about how we can consider regression of y on X_1 and X_2 to be the result of three regressions. In this question we apply this approach where y is length,

- 2a) $lm1$ contains the result of regressing length on temp, with the residuals stored in e1.
- 2b) $lm2$ contains the result of regressing age on temp, with the residuals stored in e2.
- 2c) Regress the residuals $e1$ on the residuals $e2$. Do not include an intercept. Use the formula $lm(e1 \sim e2 - 1)$. Print the *summary* and *anova* outputs.

```r
#lm3=lm( ...
#summary(lm3)
#anova(lm3)
```

- 2d) Fit the model including *age* and *temperature*, and show the *summary* and *anova* outputs.

```r
#lmfull=lm( ...
#summary(lmfull)
#anova(lmfull)
```

- 2e) Show that the coefficient of *age* in $lmfull$ is is the same as that in the regression of $e1$ on $e2$. Ans: The coefficient of *age* equals ??? in both cases.
- 2f) Use {Step 3} in the notes to show that the intercept and the coefficient of *temp* in the $lmfull$ fit are the same as those reconstructed from the three stage regression process.
  (This is what we did in class with the tree data. That is, substitute for e_1 and e_2 in the equation $e_1 = \alpha e_2$, where α is the coefficient from the 3rd regression. Isolate length on the left hand side, and calculate the regression coefficients on the right hand side.)

- 2g) Show that the residual sum of squares from the third regression equals that of the *lm* fit to the full model. Ans: The error SS equals ??? in both cases.
- 2h) Show that $SSR(\beta_2|\beta_1)$, the extra regression sum of squares explained by *age* is the same in the third regression as in the *anova* output for the full model. Ans: the regression sum of squares is ??? in both cases.

3. It is apparent from the added variable plot in 1b that a nonlinear term in age should be added.

- 3a) The following fits the model $y = \beta_0 + \beta_1 temp + \beta_2 age + \beta_3 age^2 + e$, evaluates the fitted values and the residuals, plots residuals (on y axis) vs fitted values (on x axis), and shows a normal QQ plot of the residuals. **Comment on the plots, and in particular, whether any of the assumptions of the regression analysis appear to be violated.**

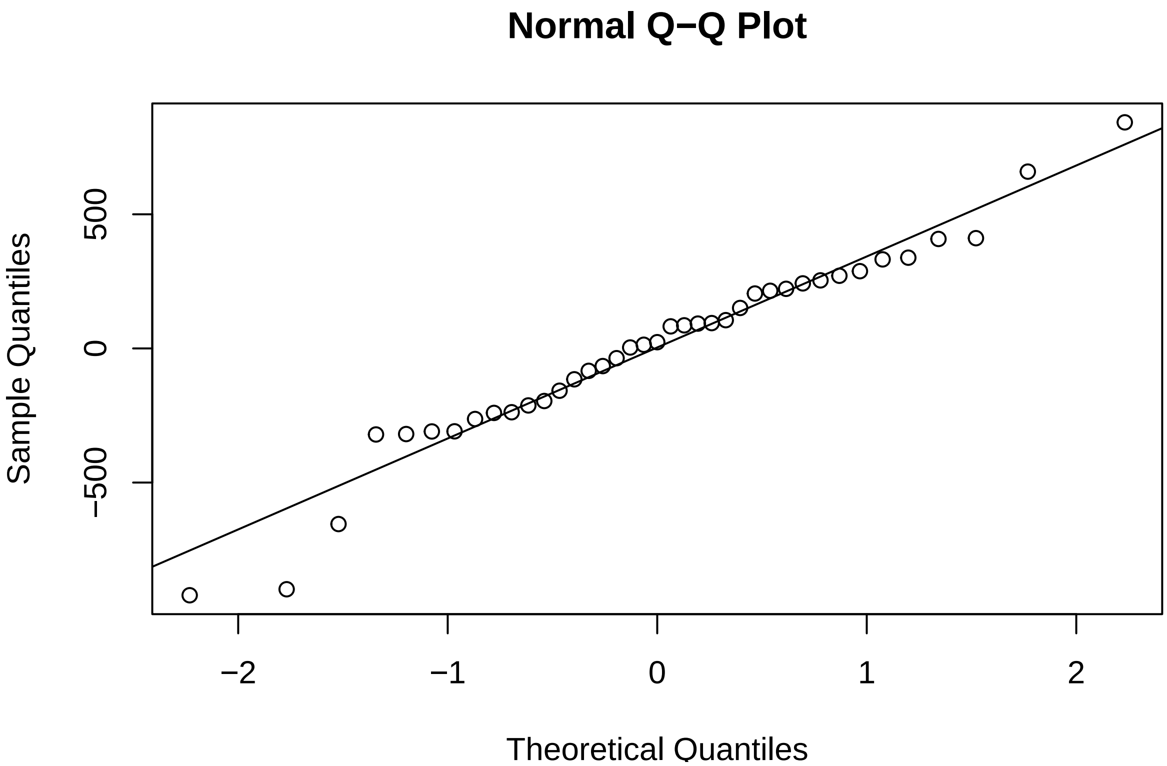
```
# enter your work here
lmbig=lm(length~temp+age+age2)
summary(lmbig)
```

```
##
## Call:
## lm(formula = length ~ temp + age + age2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -920.13 -225.21   23.29  232.36  842.79
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5988.70675  791.92102   7.562 7.29e-09 ***
## temp         -85.57812   27.80535  -3.078  0.00404 **
## age           27.89209    1.42082  19.631  < 2e-16 ***
## age2          -0.23165    0.03732  -6.207 4.12e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 377.3 on 35 degrees of freedom
## Multiple R-squared:  0.9279, Adjusted R-squared:  0.9217
## F-statistic: 150.1 on 3 and 35 DF,  p-value: < 2.2e-16
```

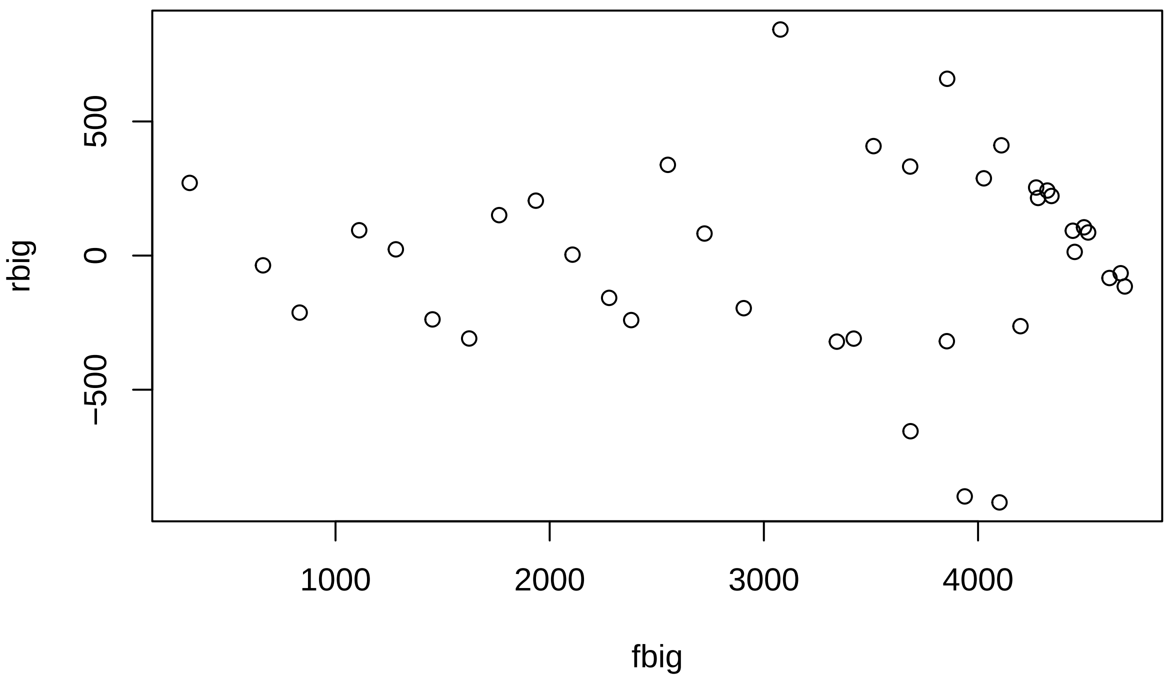
```
anova(lmbig)
```

```
## Analysis of Variance Table
##
## Response: length
##           Df   Sum Sq  Mean Sq F value    Pr(>F)
## temp       1  2733359  2733359  19.197 0.0001022 ***
## age        1 55914633 55914633 392.695 < 2.2e-16 ***
## age2       1  5485961  5485961  38.529 4.123e-07 ***
## Residuals 35  4983548   142387
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fbig=fitted(lmbig)
rbig=residuals(lmbig)
qqnorm(rbig); qqline(rbig)
```




```r
plot(fbig,rbig)
```



- 3b) now do the same for the model

$$y = \beta_0 + \beta_1 temp + \beta_2 age + \beta_3 age^2 + \beta_4 temp \times age + \beta_5 temp \times age^2 + e$$

which includes the interaction of age and temperature, and the interaction of age^2 and temperature. That is using the R code "lm(length~ temp+age+age2+temp:age + temp:age2)".

- 3c) Test the hypothesis $H_0 : \beta_4 = \beta_5 = 0$. Report the observed value of F, the numerator and denominator degrees of freedom, and the p-value.