STAT 3340 Assignment 5 solutions, Fall 2024

(out of 25 points)

==========

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)
```

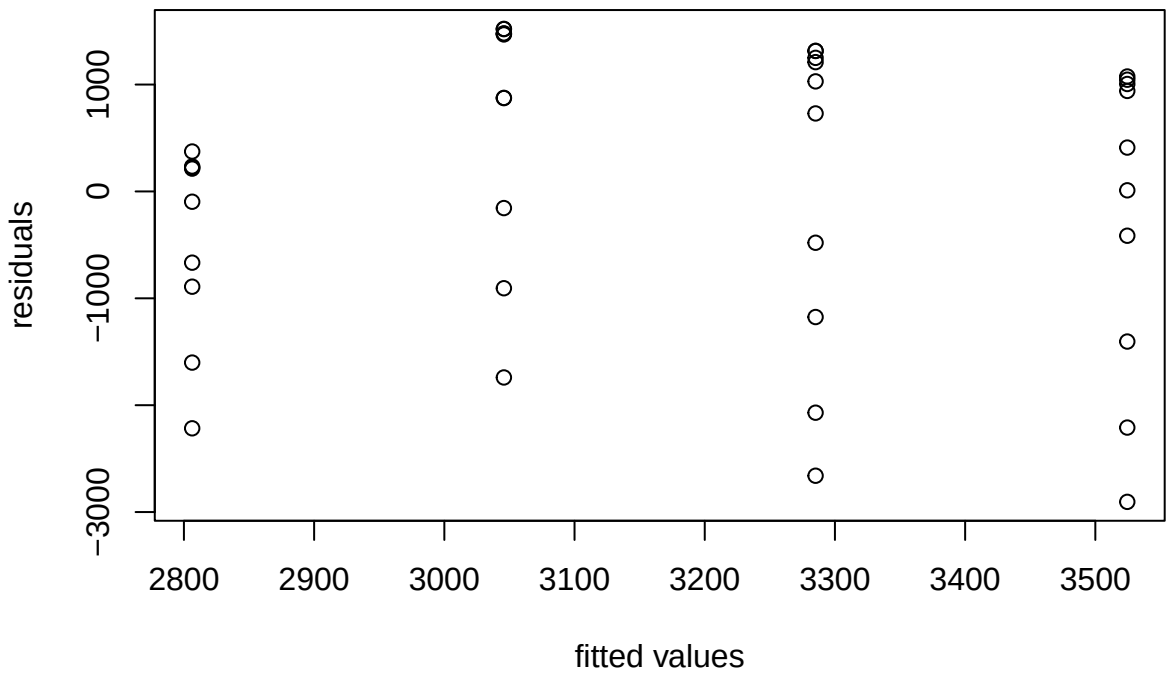
```
##
## Call:
## lm(formula = length ~ temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2904.5  -898.5   233.7  1060.5  1520.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6517.03    2716.43   2.399   0.0216 *
## temp         -119.70      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)
```

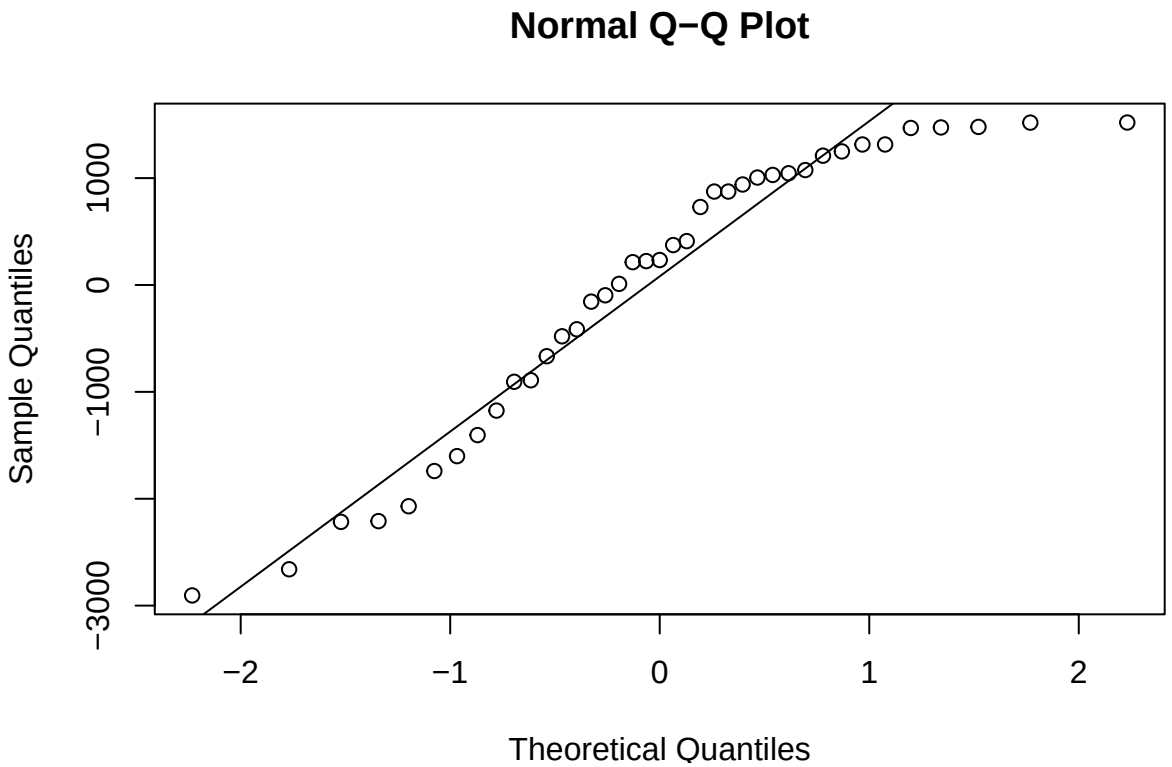
```
## Analysis of Variance Table
##
## Response: length
##           Df   Sum Sq Mean Sq F value Pr(>F)
## temp       1  2733359 2733359  1.5235 0.2249
## Residuals 37 66384142 1794166
```

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

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




```r
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)?

`Yes, it appears that the assumption of normality is violated.  (2 points for any reasonable answer.)`

- 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.

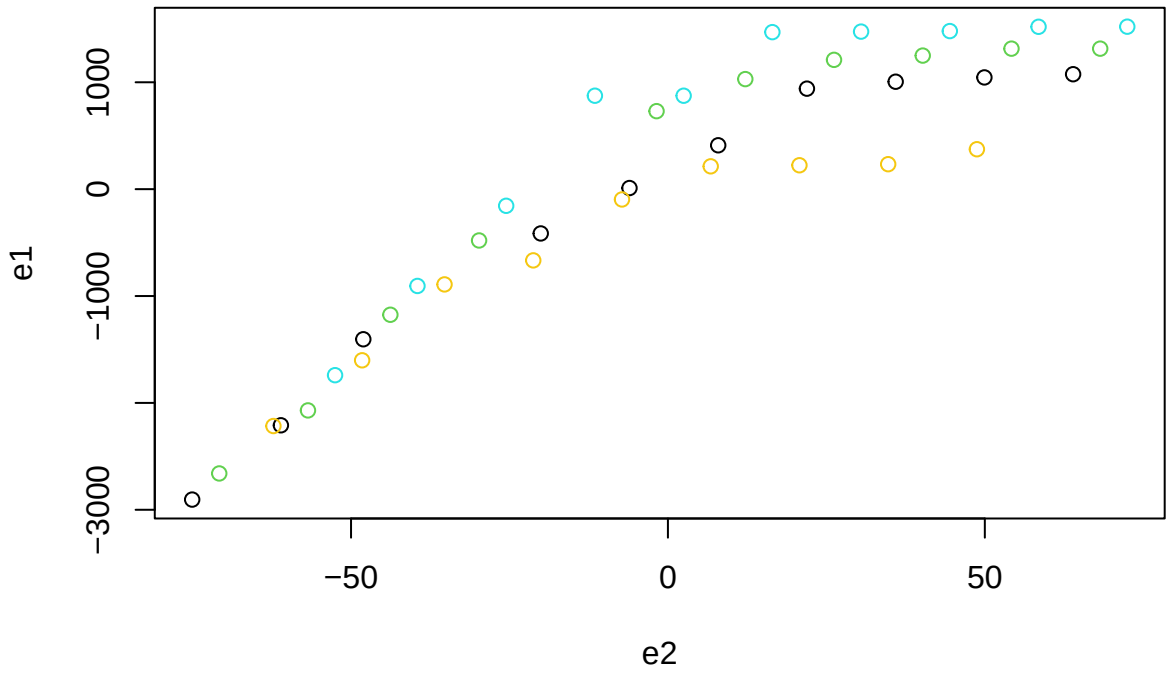
```
lm2=lm(age~temp)
summary(lm2)
```

```
##
## 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
```

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



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

A quadratic model appears more appropriate than linear. (2 points)

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, X_1 is temp, and X_2 is age.

- 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. (3 points)

```
lm3=lm( e1~e2-1)
summary(lm3)
```

```
##
## Call:
## lm(formula = e1 ~ e2 - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -998.43 -412.42   23.62  275.27 1198.48
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## e2   28.144      1.976   14.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 524.9 on 38 degrees of freedom
## Multiple R-squared:  0.8423, Adjusted R-squared:  0.8381
## F-statistic: 202.9 on 1 and 38 DF,  p-value: < 2.2e-16
```

```
anova(lm3)
```

```
## Analysis of Variance Table
```

```
##
## Response: e1
##           Df   Sum Sq  Mean Sq F value    Pr(>F)
## e2         1 55914633 55914633  202.95 < 2.2e-16 ***
## Residuals 38 10469509   275513
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

```
lmfull=lm(length~temp+age)
summary(lmfull)
```

```
##
## Call:
## lm(formula = length ~ temp + age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -998.43 -412.42   23.62  275.27 1198.48
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4838.19    1100.33   4.397 9.32e-05 ***
## temp          -59.58      39.28  -1.517    0.138
## age            28.14       2.03  13.866 5.22e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 539.3 on 36 degrees of freedom
## Multiple R-squared:  0.8485, Adjusted R-squared:  0.8401
## F-statistic: 100.8 on 2 and 36 DF,  p-value: 1.762e-15
```

```
anova(lmfull)
```

```
## Analysis of Variance Table
##
## Response: length
##           Df   Sum Sq  Mean Sq  F value    Pr(>F)
## temp       1  2733359  2733359   9.3988  0.004102 **
## age        1 55914633 55914633 192.2656 5.217e-16 ***
## Residuals 36 10469509   290820
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 2e) Show that the coefficient of *age* in $lmfull$ is is the same as that in the regression of $e1$ on $e2$. Answer: From the outputs above, the coefficient of *age* equals 28.14 in both cases. (2 points)
- 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. (4 points, two for each of the estimated intercept and coefficient of *temp*.)

(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.)

In general we solve something like:

$(length - \hat{\theta}_0 - \hat{\theta}_1 temp) = \hat{\alpha}(age - \hat{\gamma}_0 - \hat{\gamma}_1 temp)$

Then isolate *length* on the left of =, and move everything else to the right, and group the intercept terms, and the terms multiplying temp to get:

$length = \hat{\theta}_0 + \hat{\theta}_1 temp + \hat{\alpha}(age - \hat{\gamma}_0 - \hat{\gamma}_1 temp)$.

The intercept term in this equation is

$\hat{\theta}_0 - \hat{\alpha}\hat{\gamma}_0 = 6517.03 - 28.14(59.652) \approx 4838.42$ and the coefficient of "temp" is $\hat{\theta}_1 - \hat{\alpha}\hat{\gamma}_1 = -119.7 - 28.14(-2.136) \approx -59.59$

which, apart from rounding error, agrees with the estimates $\hat{\beta}_0 = 4838.19$ and $\hat{\beta}_1 = -59.58$ from the regression lm(length~temp+age).

- 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 10469509 in both cases. (2 points)
- 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 55914633 in both cases. (2 points)

3. It is apparent from the added variable plot in 1b that a quadratic 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.

```
lmbig=lm(length~temp+age+age2)
summary(lmbig)
```

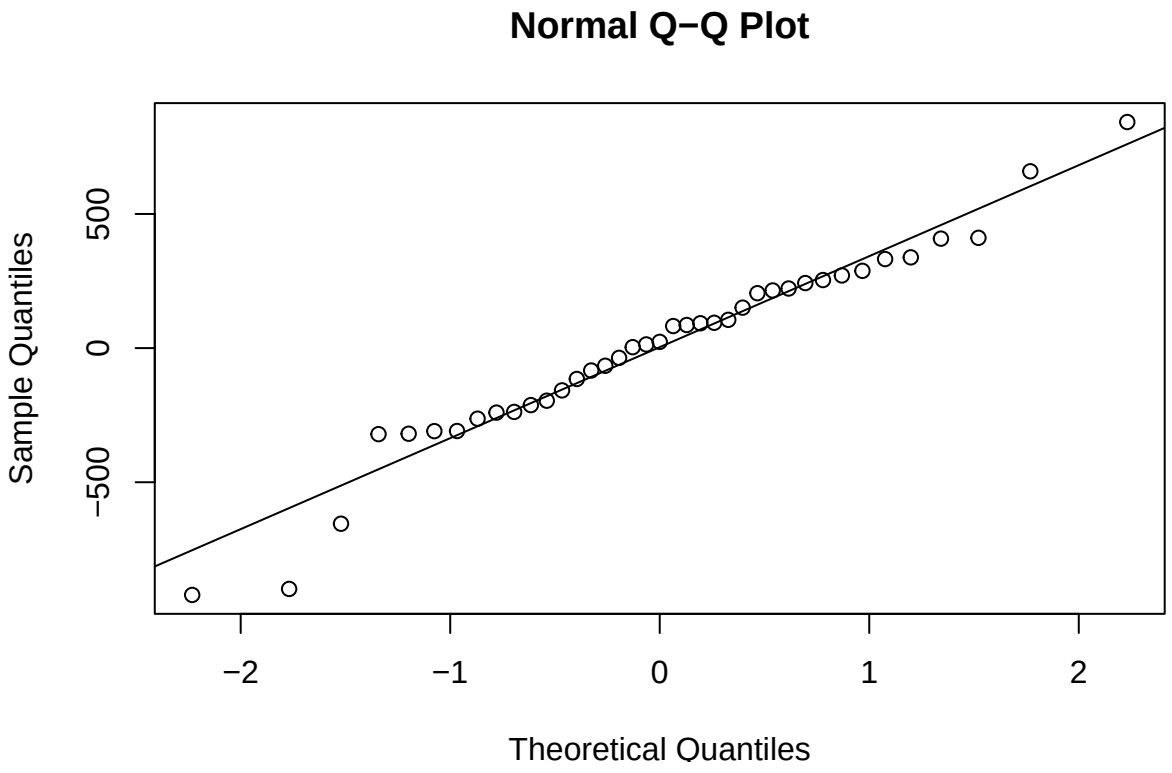
```
##
## Call:
## lm(formula = length ~ temp + age + age2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -920.13 -225.21   23.29  232.36  842.79
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5988.70675  791.92102   7.562 7.29e-09 ***
## temp         -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)
```

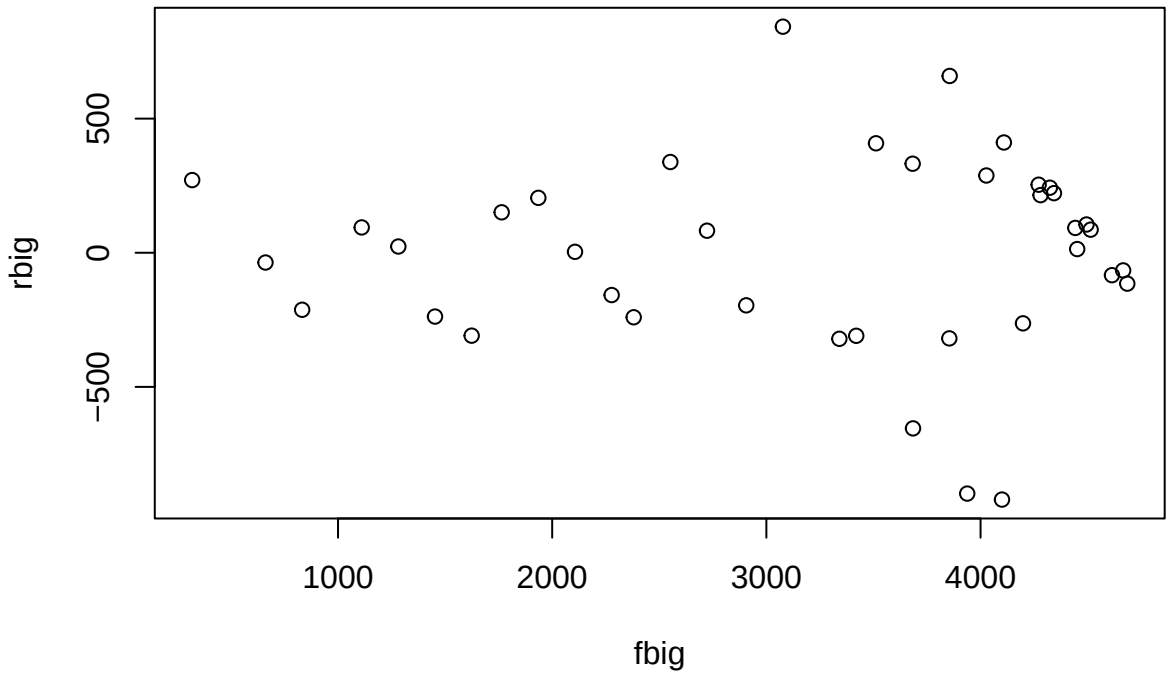
```
## Analysis of Variance Table
##
## Response: length
##           Df   Sum Sq  Mean Sq F value    Pr(>F)
## temp       1  2733359  2733359  19.197 0.0001022 ***
```

```
## 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)
```




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



There is a suggestion of increased variability of the residuals with increasing fitted values.

(2 points for any reasonable statement.)

- 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)".

```
lmbig2=lm(length~temp+age+age2+temp:age+temp:age2)
summary(lmbig2)
```

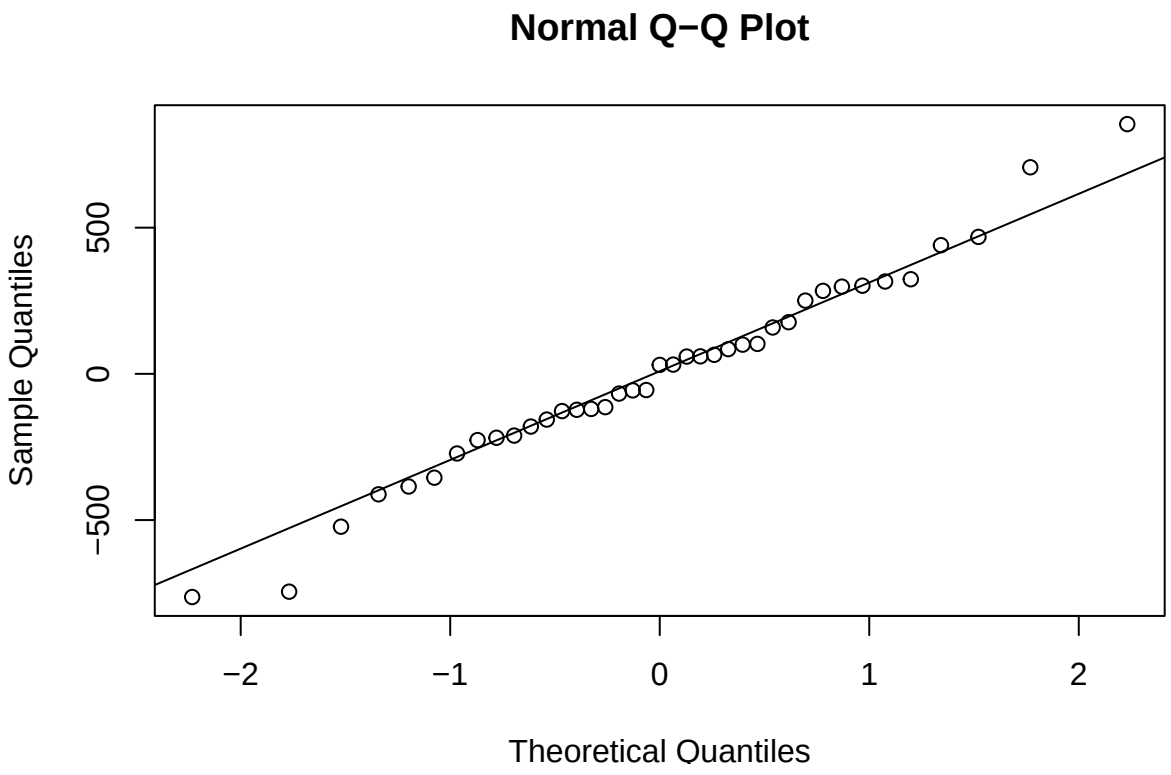
```
##
## Call:
## lm(formula = length ~ temp + age + age2 + temp:age + temp:age2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -763.18 -195.62   30.63  213.64  854.07
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6855.89495 1115.57962   6.146 6.29e-07 ***
## temp        -116.57534   39.58389  -2.945  0.00588 **
## age           55.92284   18.50538   3.022  0.00483 **
## age2          -0.64588    0.47385  -1.363  0.18209
## temp:age      -1.01115    0.66996  -1.509  0.14075
## temp:age2      0.01475    0.01714   0.861  0.39558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 367.9 on 33 degrees of freedom
```

```
## Multiple R-squared:  0.9354, Adjusted R-squared:  0.9256
## F-statistic: 95.56 on 5 and 33 DF,  p-value: < 2.2e-16
```

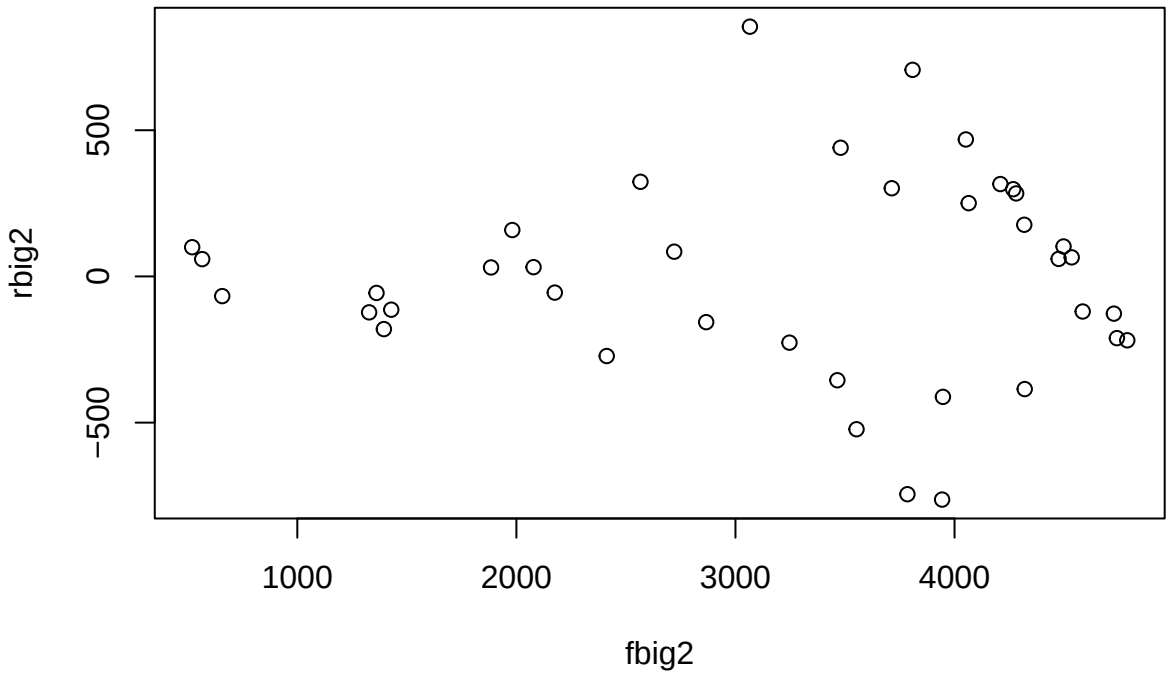
```
anova(lmbig2)
```

```
## Analysis of Variance Table
##
## Response: length
##           Df   Sum Sq  Mean Sq  F value    Pr(>F)
## temp       1  2733359  2733359  20.2001 8.113e-05 ***
## age        1 55914633 55914633 413.2217 < 2.2e-16 ***
## age2       1  5485961  5485961  40.5425 3.286e-07 ***
## temp:age   1   417935   417935   3.0886   0.08812 .
## temp:age2  1   100255   100255   0.7409   0.39558
## Residuals 33  4465358   135314
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fbig2=fitted(lmbig2)
rbig2=residuals(lmbig2)
qqnorm(rbig2); qqline(rbig2)
```




```
plot(fbig2,rbig2)
```



This has not improved the residual plots. The QQ plot has a slight suggestion of non-normality, and the plot of residuals vs fitted values still suggests non-constant variance.

(2 points for plots, and 2 points for any reasonable statement regarding the results.)