DS705 Advanced Modeling Tools Homework

Load packages

```
library(tidyverse)
```

```
## Warning: package ’tidyr’ was built under R version 4.1.2

## Warning: package ’readr’ was built under R version 4.1.2

## Warning: package ’dplyr’ was built under R version 4.1.2
```

```
library(broom)
```

```
## Warning: package ’broom’ was built under R version 4.1.2
```

```
library(car)
```

```
## Warning: package ’carData’ was built under R version 4.1.2
```

How to complete the homework.

If you’ve made it this far, then you’ve already downloaded and unzipped the HW packet for this week. We suggest that you keep all of the materials, including this .rmd file, for the week in one folder. It will help to set the working directory to the folder that contains the HW materials. You can do this by opening the rmd file in an RStudio editor window and then using the menu commands Session -> Set Working Directory -> To Source File Location.

You are free to add R code and type answers in the designated spaces throughout this document. At the end of the week, you’ll input your answers to the Canvas quiz associated with this homework.

Exercise 1

The Long Term Resource Monitoring (LTRM) project has been conducting research and monitoring on the Upper Mississippi River System since 1986.

Here is a bit about the LTRM from their website: "Fishes of the Upper Mississippi River System have recreational and commercial value, conservation potential, and can be used to assess the ecological integrity of the aquatic ecosystem.

The objective of the standardized monitoring is to quantify the status and trends of fish populations and communities and identify relations with various other ecological attributes. The findings can be used to address fisheries management concerns.

The Long Term Resource Monitoring element uses a multigear and multihabitat sampling design to collect fish data in six study pools/reaches."

The file fish_data.rda is available in the homework download packet and in the DS705data package. It contains length and weight measurements for a sample of 50 fish of four difference species of interest: BHMW = Bullhead minnow, BKCP = Black crappie, BLGL = Bluegill, BWFN = Bowfin. Additionally, the file includes the date that the fish was observed.

Question 1

It is standard practice to plot and model the relationship between fish weight and length using a logarithm transformation on both variables. To see the reason for the use of these transformations, we are going to start by fitting a model without any transformation.

Fit a model for predicting weight based on length and species. Include the length by species interaction term. Create and inspect the 4-pack of model diagnostic plots.

What conditions fail based on the 4-pack of plots? Select all that apply.

Answer 1

- L = Linear

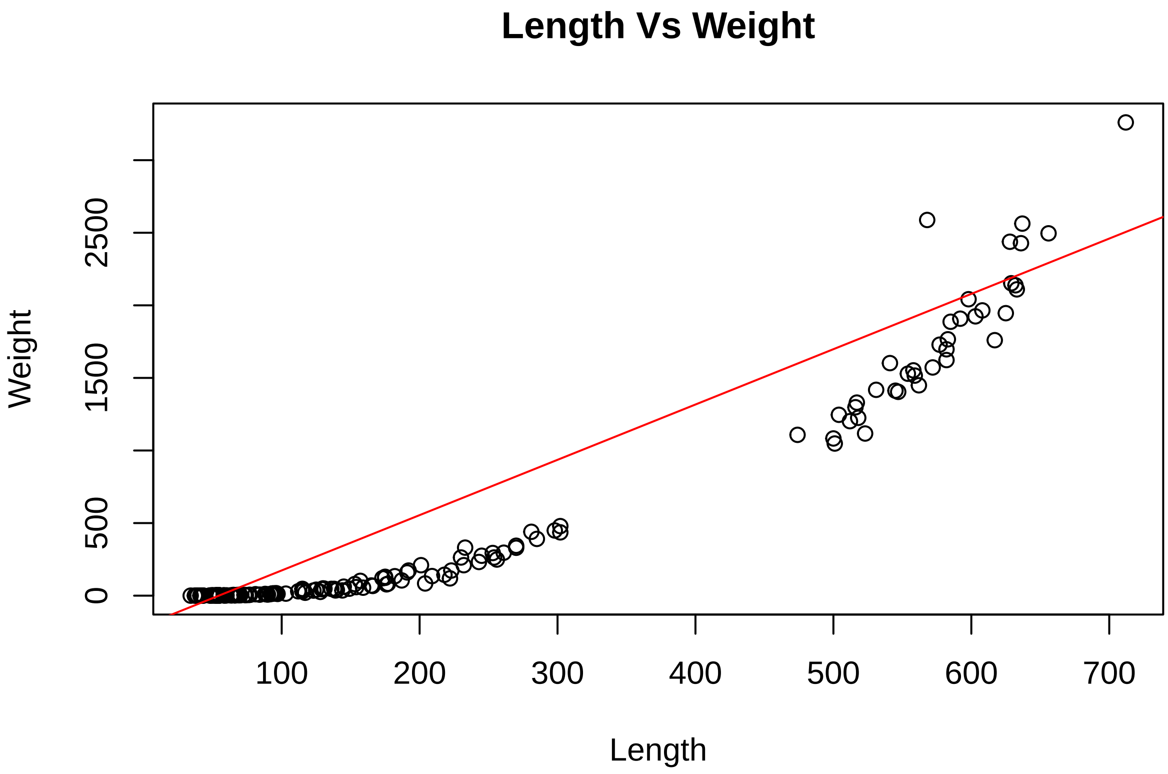
- I = Independent errors
- E = Equal variance of errors

```
load("fish_data.rda")

glimpse(fish_data)
```

```
## Rows: 160
## Columns: 4
## $ length   <dbl> 74, 66, 63, 42, 66, 42, 64, 39, 73, 69, 53, 55, 48, 48, 67, 3~
## $ weight   <dbl> 4.00, 3.00, 2.00, 0.67, 3.00, 0.67, 2.00, 0.53, 5.00, 3.00, 1~
## $ fishcode <chr> "BHMW", "BHMW", "BHMW", "BHMW", "BHMW", "BHMW", "BHMW", "BHMW~
## $ fdate    <date> 2020-07-23, 2020-10-06, 2020-09-29, 2020-07-14, 2019-08-28, ~
```

```
fish_data <-  fish_data %>% na.omit()

plot(fish_data$length, fish_data$weight,
     main='Length Vs Weight',
     xlab='Length', ylab='Weight')

abline(lm(weight ~ length + fishcode,data=fish_data),col='red')
```

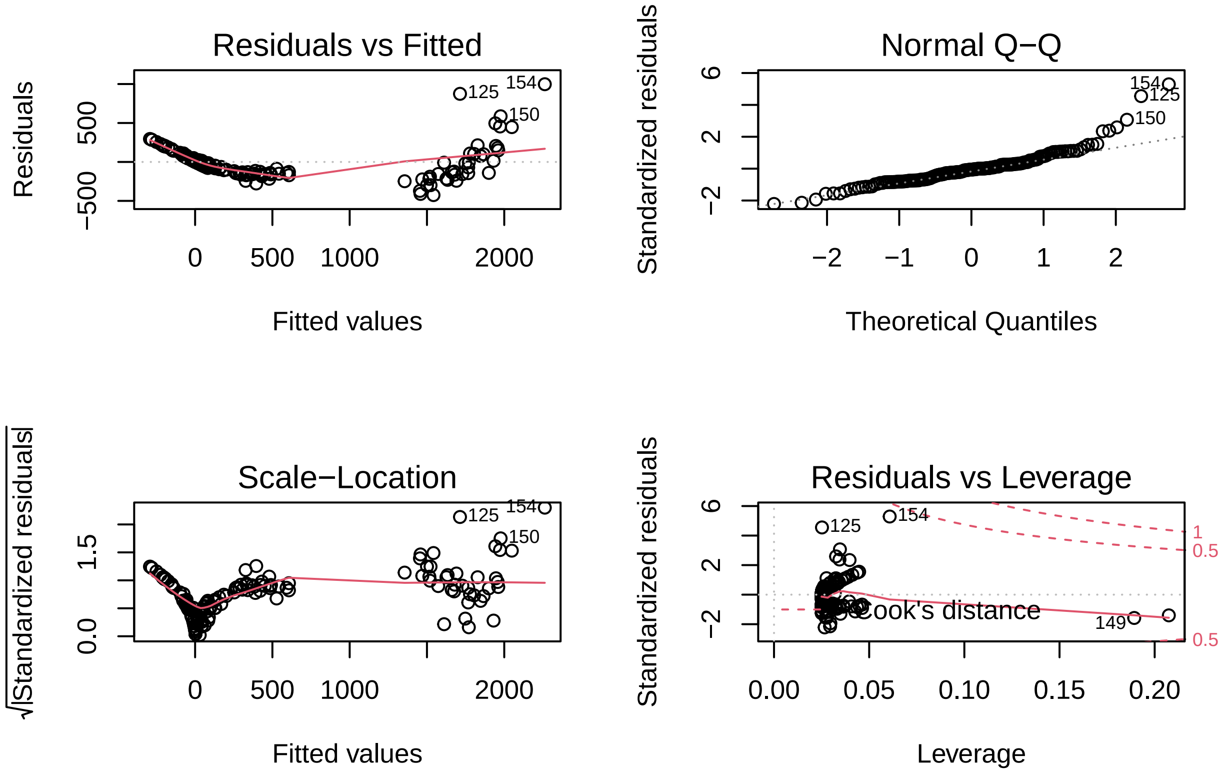
```
## Warning in abline(lm(weight ~ length + fishcode, data = fish_data), col =
## "red"): only using the first two of 5 regression coefficients
```



```r
summary(lm(weight ~ length + fishcode,data=fish_data))
```

```
##
## Call:
## lm(formula = weight ~ length + fishcode, data = fish_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -425.33 -130.68  -11.77   62.17  997.53
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -207.224     33.377  -6.209 4.68e-09 ***
## length          3.810      0.236  16.143  < 2e-16 ***
## fishcodeBKCP -336.333     52.827  -6.367 2.08e-09 ***
## fishcodeBLGL -189.619     46.181  -4.106 6.50e-05 ***
## fishcodeBWFN -243.222    126.054  -1.930   0.0555 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 194.6 on 155 degrees of freedom
## Multiple R-squared:  0.9367, Adjusted R-squared:  0.9351
## F-statistic: 573.5 on 4 and 155 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(lm(weight ~ length +fishcode,data=fish_data))
```



```r
par(mfrow=c(1,1))
```

Question 2

Refit the model from question 1 with a log-transformation applied to both weight and length. Include the log(length) by species interaction term. Again, inspect the diagnostic plots. Call this model fit MOD_interact as we will refer to it later.

Have model conditions improved in comparison to the model based on non-transformed data?

Answer 2

Options:

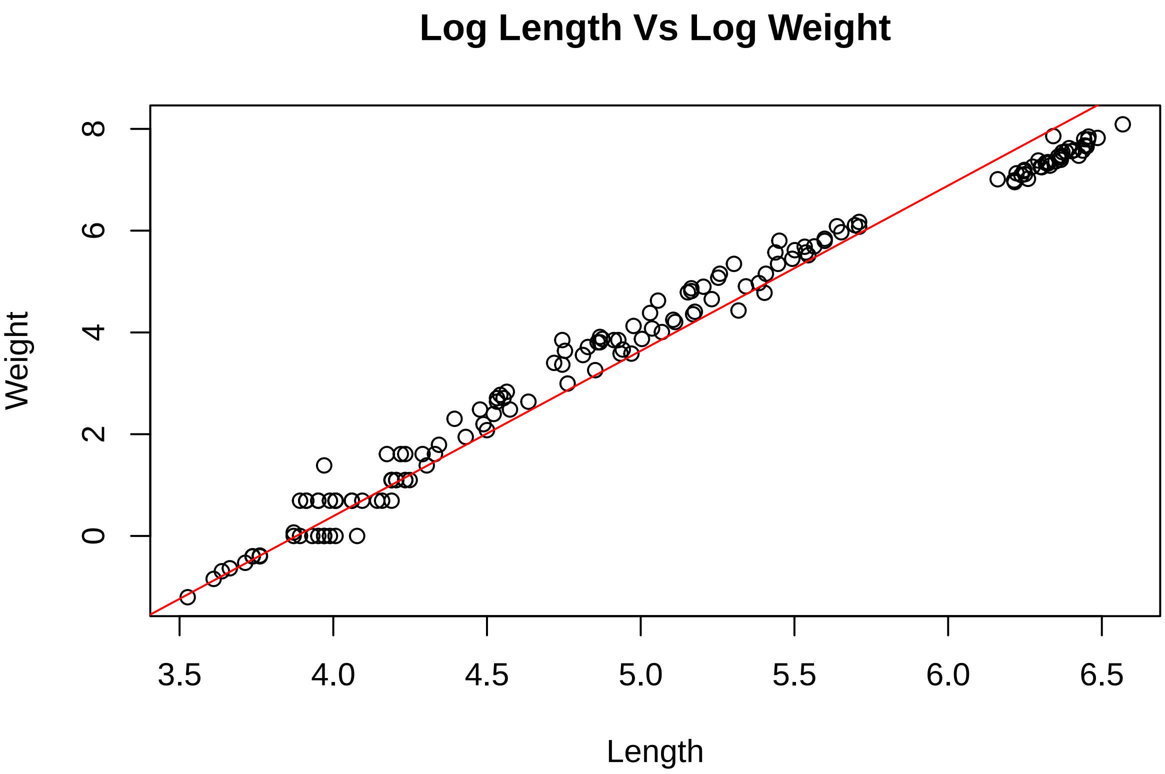
- Yes, there is improvement but there are still some concerns with normality and outliers

```r
fish_data <-  fish_data %>% mutate(log_length = log(length), log_weight = log(weight) )
glimpse(fish_data)
```

```
## Rows: 160
## Columns: 6
## $ length     <dbl> 74, 66, 63, 42, 66, 42, 64, 39, 73, 69, 53, 55, 48, 48, 67,~
## $ weight     <dbl> 4.00, 3.00, 2.00, 0.67, 3.00, 0.67, 2.00, 0.53, 5.00, 3.00,~
## $ fishcode   <chr> "BHMW", "BHMW", "BHMW", "BHMW", "BHMW", "BHMW", "BHMW", "BH~
## $ fdate      <date> 2020-07-23, 2020-10-06, 2020-09-29, 2020-07-14, 2019-08-28~
## $ log_length <dbl> 4.304065, 4.189655, 4.143135, 3.737670, 4.189655, 3.737670,~
## $ log_weight <dbl> 1.38629436, 1.09861229, 0.69314718, -0.40047757, 1.09861229~
```

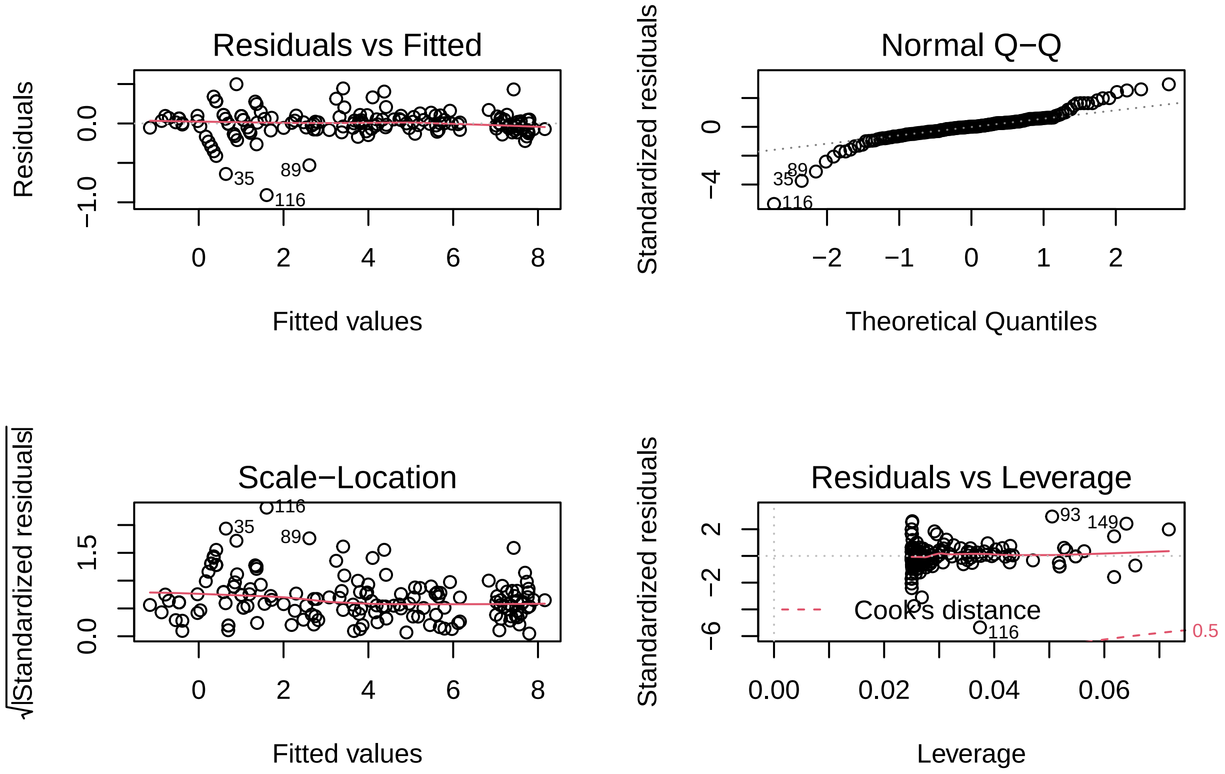
```
plot(fish_data$log_length, fish_data$log_weight,
     main='Log Length Vs Log Weight',
     xlab='Length', ylab='Weight')

abline(lm(log_weight ~ log_length + fishcode,data=fish_data),col='red')
```

```
## Warning in abline(lm(log_weight ~ log_length + fishcode, data = fish_data), :
## only using the first two of 5 regression coefficients
```



```
MOD_interact   <- lm(log_weight ~ log_length + fishcode,data=fish_data)
```

```
par(mfrow=c(2,2))
plot(MOD_interact)
```



```
par(mfrow=c(1,1))
```

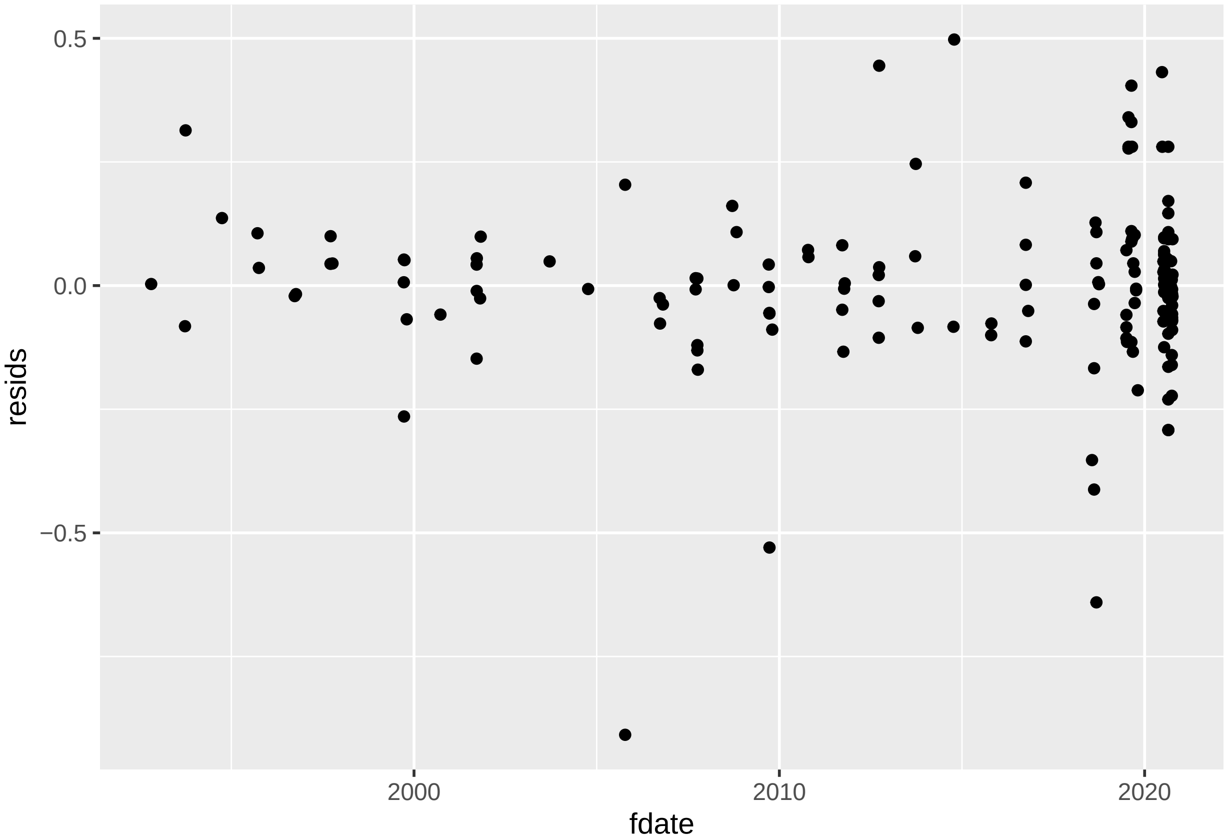
Question 3

Extract the residuals from your fit model using the residuals() function. Create a scatter plot of your model residuals (y) over time (x=fdate) to assess the assumed independence of model errors. Distinct patterns in the residuals over time (e.g. residuals trending upward over time) suggest a violation of the assumption of independent errors in linear regression. Do you see any clear patterns in your plot of the residuals over time?

Answer 3

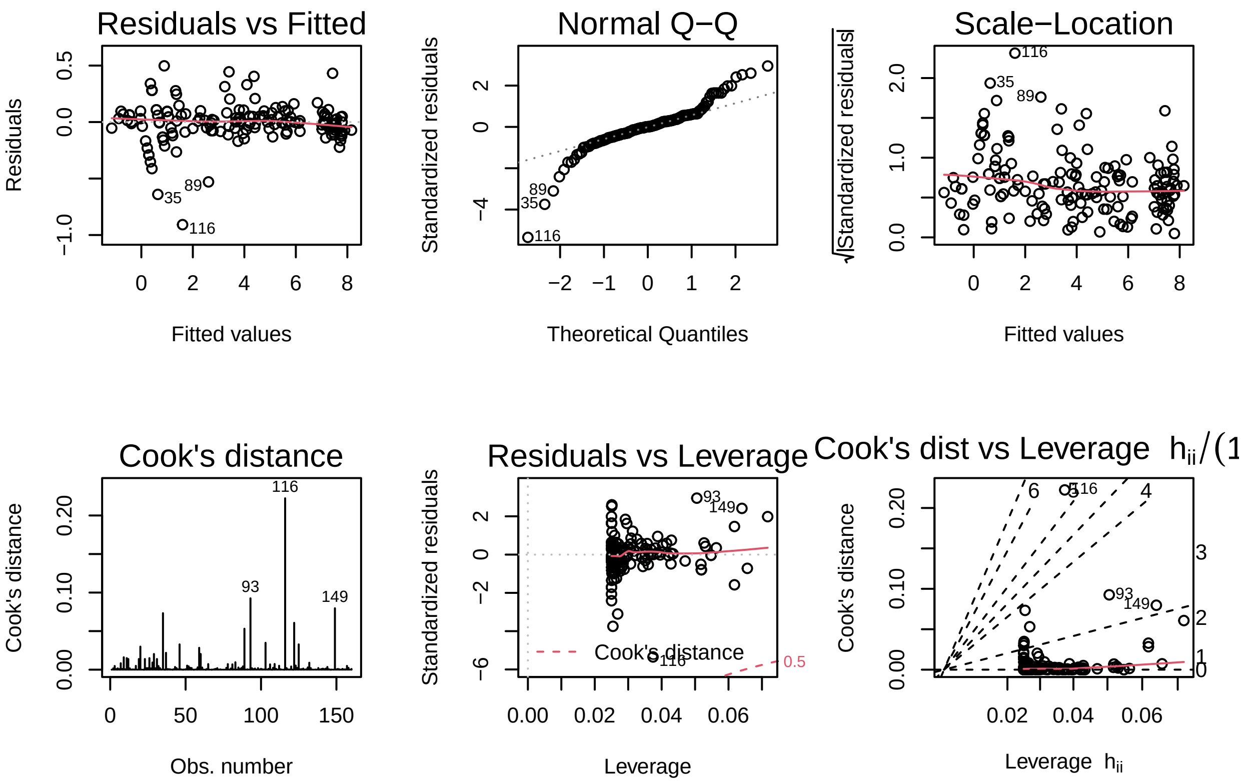
- No

```
# Here is possible code for computing model residuals and mutating them onto the fish_data
fish_data = fish_data %>% mutate(resids = residuals(lm(log_weight ~ log_length +fishcode,data=fish_data
library(ggplot2)
ggplot(fish_data, aes(x = fdate, y = resids)) +
    geom_point()
```



The model is decent, but could be better. There are some outliers that negatively effect the residuals.

```
## Residual analysis for assumptions (independence, linearity, normality, and homoscedasticity)
layout(matrix(1:6, byrow = T, ncol = 3))
plot(MOD_interact, which = 1:6)
```

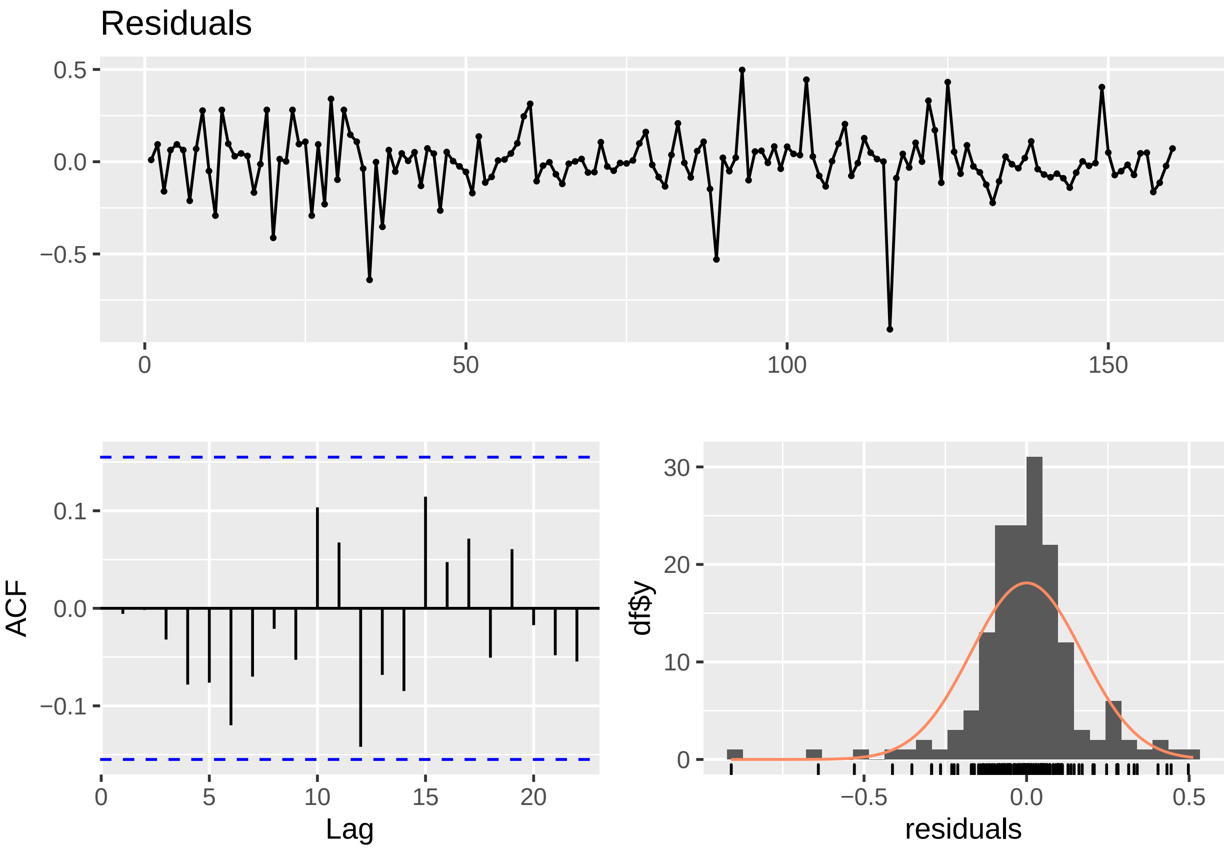


```r
library(fpp2)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

## -- Attaching packages --------------------------------------------- fpp2 2.4 --

## v forecast  8.16      v expsmooth 2.3
## v fma       2.4

## Warning: package 'forecast' was built under R version 4.1.2

## -- Conflicts ------------------------------------------------- fpp2_conflicts --
## x car::some() masks purrr::some()
```

```r
checkresiduals(fish_data$resids)
```

```
## Warning in modeldf.default(object): Could not find appropriate degrees of
## freedom for this model.
```



Question 4

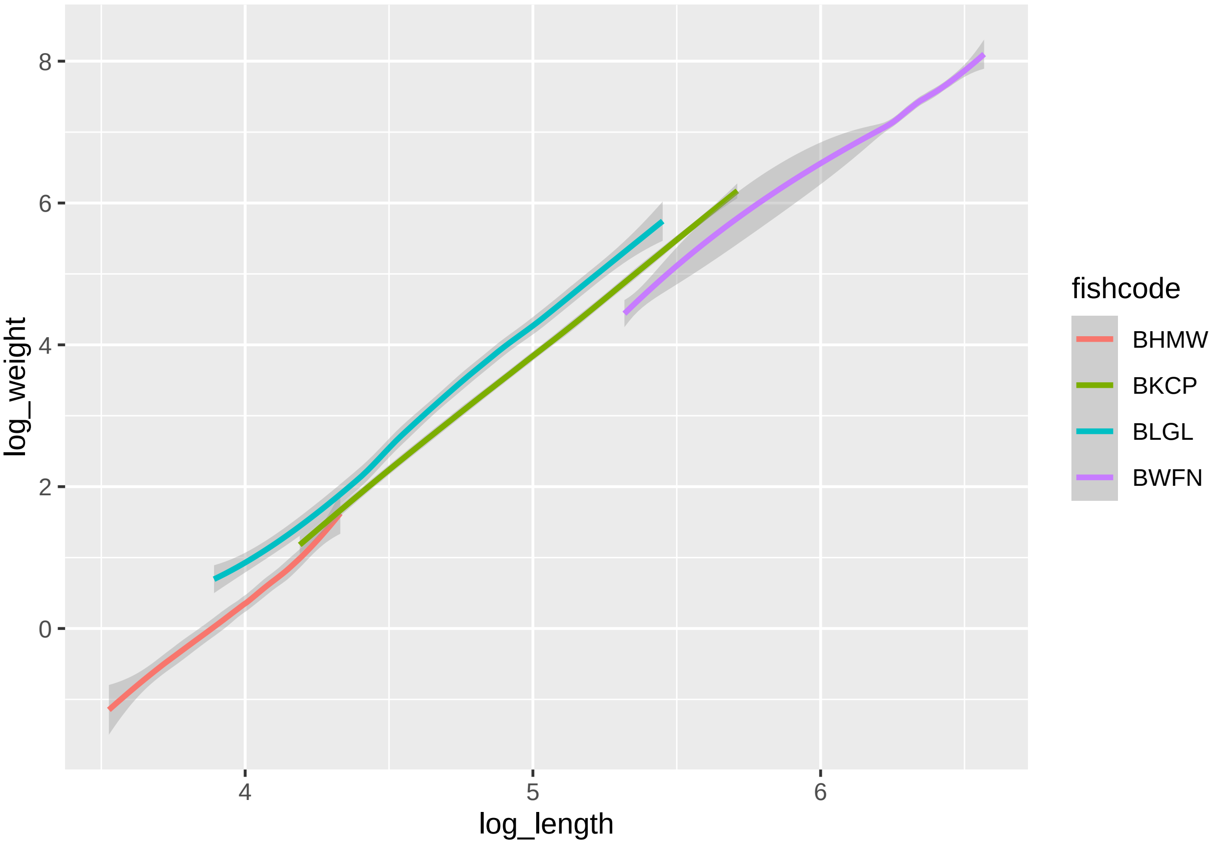
Make a plot that visualizes the interaction: weight on the y-axis (log-scale), length on the x-axis (log-scale), and fishcode showing the categorical predictor variable with a separate line fit for each category. You can do this with geom_abline or with geom_smooth with color/group set in the aes.

What best describes the nature of this interaction?

Answer 4

- The slope for BWFN is slightly shallower than for the other 3 species (interaction present)

```r
#unique(fish_data$fishcode)

ggplot(fish_data, aes(log_length, log_weight, colour = fishcode)) +
  geom_smooth()
```



Question 5

Fit the model with no interaction term. Call this model MOD_additive.

Based on AIC values, which model is better?

Answer 5

- MOD_interact

```
MOD_additive <- lm(log_weight ~ log_length,  data=fish_data)
summary(MOD_additive)
```

```
##
## Call:
## lm(formula = log_weight ~ log_length, data = fish_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95647 -0.31998 -0.05429  0.31626  0.90647
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.18462    0.16976  -65.89   <2e-16 ***
## log_length    2.97755    0.03326   89.51   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3844 on 158 degrees of freedom
## Multiple R-squared:  0.9807, Adjusted R-squared:  0.9805
## F-statistic:  8012 on 1 and 158 DF,  p-value: < 2.2e-16
```

```r
AIC(MOD_additive)
```

```
## [1] 152.1253
```

```r
AIC(MOD_interact)
```

```
## [1] -100.2225
```

The basic idea of AIC is to penalize the inclusion of additional variables to a model. It adds a penalty that increases the error when including additional terms. The lower the AIC, the better the model.

Question 6

Based on adjusted R^2 values, which model is better?

Answer 6

- Essentially no difference between the two models (adjusted R2 is identical to 3 decimal places)

```r
summary(MOD_interact)
```

```
##
## Call:
## lm(formula = log_weight ~ log_length + fishcode, data = fish_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90831 -0.06853  0.00153  0.06480  0.49754
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -12.60721    0.15434 -81.683  < 2e-16 ***
## log_length     3.24896    0.03813  85.205  < 2e-16 ***
## fishcodeBKCP   0.21409    0.05768   3.712 0.000286 ***
## fishcodeBLGL   0.59666    0.04728  12.619  < 2e-16 ***
## fishcodeBWFN  -0.57103    0.09642  -5.922 1.98e-08 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1731 on 155 degrees of freedom
## Multiple R-squared:  0.9962, Adjusted R-squared:  0.9961
## F-statistic: 1.003e+04 on 4 and 155 DF,  p-value: < 2.2e-16
```

```r
summary(MOD_additive)
```

```
##
## Call:
## lm(formula = log_weight ~ log_length, data = fish_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95647 -0.31998 -0.05429  0.31626  0.90647
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.18462    0.16976  -65.89   <2e-16 ***
## log_length    2.97755    0.03326   89.51   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3844 on 158 degrees of freedom
## Multiple R-squared:  0.9807, Adjusted R-squared:  0.9805
## F-statistic:  8012 on 1 and 158 DF,  p-value: < 2.2e-16
```

Question 7

Carry out a comparative ANOVA test H0: MOD_additive versus H1: MOD_interact. What is the associated p-value?

Answer 7

p-value = 0.00000000000000022

```r
anova(MOD_interact,MOD_additive,test="Chisq")
```

```
## Analysis of Variance Table
##
## Model 1: log_weight ~ log_length + fishcode
## Model 2: log_weight ~ log_length
##   Res.Df     RSS Df Sum of Sq  Pr(>Chi)
## 1    155  4.6455
## 2    158 23.3496 -3   -18.704 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Question 8

Using the interaction model, what is the predicted weight of a blue gill (BLGL) with a length of 100? Be mindful of the log transformation that the model has applied to length and weight.

Answer 8

Predicted weight = 1.082296

```
log(predict(MOD_interact, newdata=data.frame(log_length =c(log(100)), fishcode =c("BLGL")), type="respo
```

```
##        1
## 1.082296
```

Exercise 2

For this exercise, we are going to look at salaries for college teachers and how they relate to age, gender, and highest degree completed. We have access to a random sample of college teachers taken from the 2010 American Community Survey (ACS) 1-year Public Use Microdata Sample (PUMS). The file **salary_gender.csv** in the HW download packet contains 100 observations on the following 4 variables.

- Salary: Annual salary in $1,000's
- Gender 0=female or 1=male
- Age Age in years
- PhD 1=have PhD or 0=no PhD

Read in the associated data file and convert Gender and PhD to be factors.

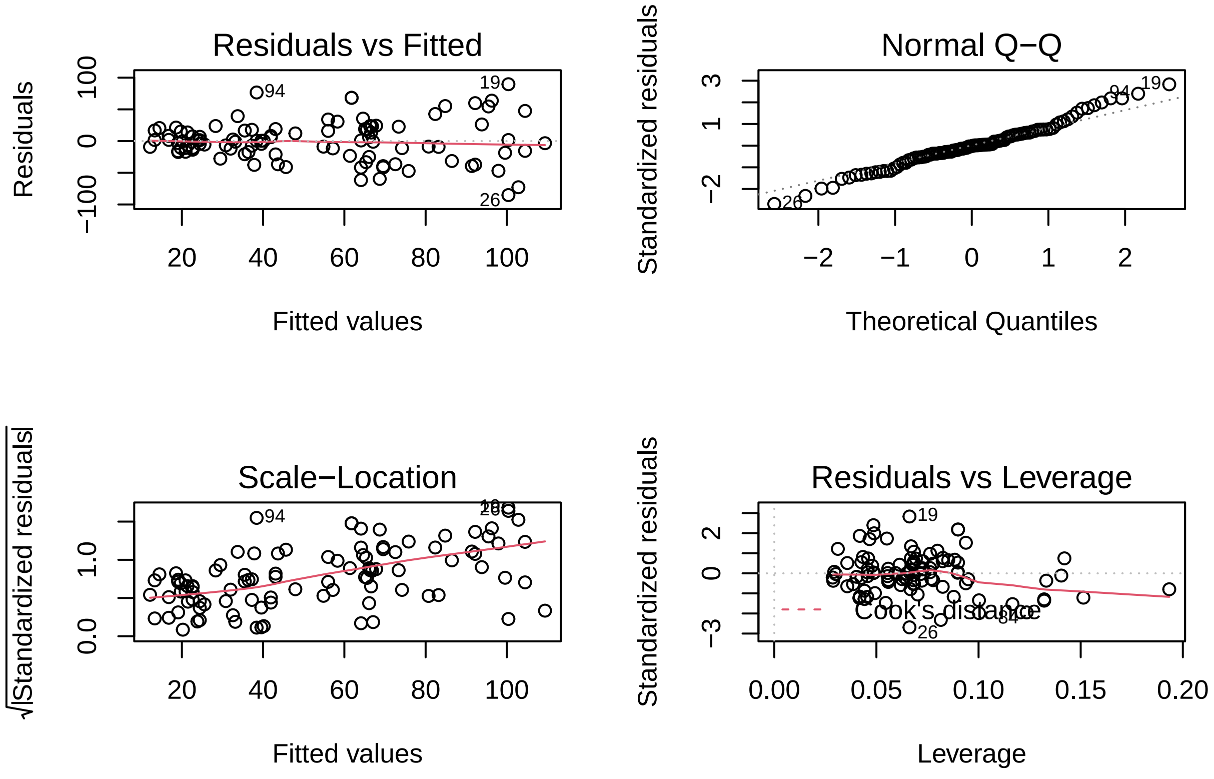
```
salary_gender = read_csv("salary_gender.csv")
salary_gender = salary_gender %>% mutate(Gender = factor(Gender), PhD = factor(PhD))
glimpse(salary_gender)
```

```
## Rows: 100
## Columns: 4
## $ Salary <dbl> 140.0, 30.0, 35.1, 30.0, 80.0, 30.0, 60.0, 31.1, 125.0, 51.0, 3~
## $ Gender <fct> 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, ~
## $ Age    <dbl> 47, 65, 56, 23, 53, 27, 53, 30, 44, 63, 22, 59, 60, 28, 65, 25,~
## $ PhD    <fct> 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, ~
```

Question 9

Begin by fitting a linear model that uses Salary as the response variable and Age, Gender, PhD, and all associated two-term interactions as predictor variables. Plot the diagnostic 4-pack to assess the model conditions.

```
model_salary <-  lm(Salary ~ Age + Gender + PhD + Age*Gender + Age*PhD + Gender*PhD,  data=salary_gende

par(mfrow=c(2,2))
plot(model_salary)
```



```
par(mfrow=c(1,1))
```

Based on the plot of residuals versus fitted values (1st plot in the 4-pack), what best describes the pattern of residual variance associated with this model:

Answer 9

- The variance of the residuals increases as fitted values increase

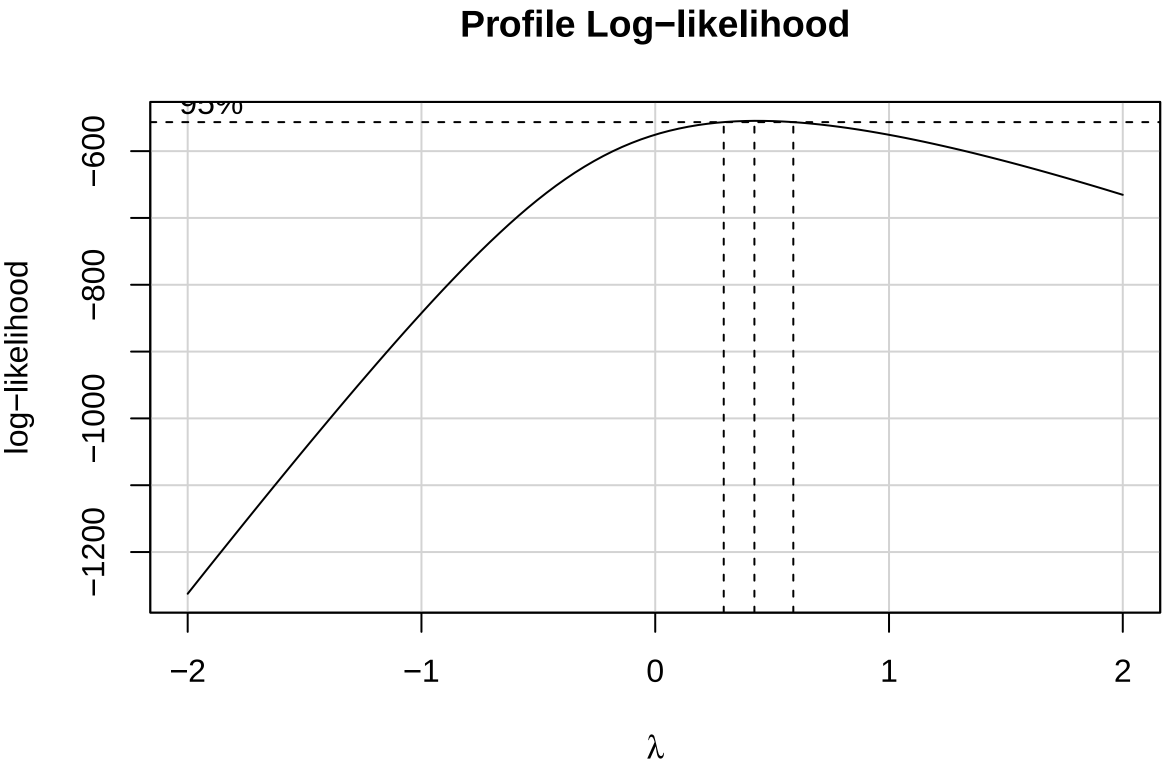
---

Question 10

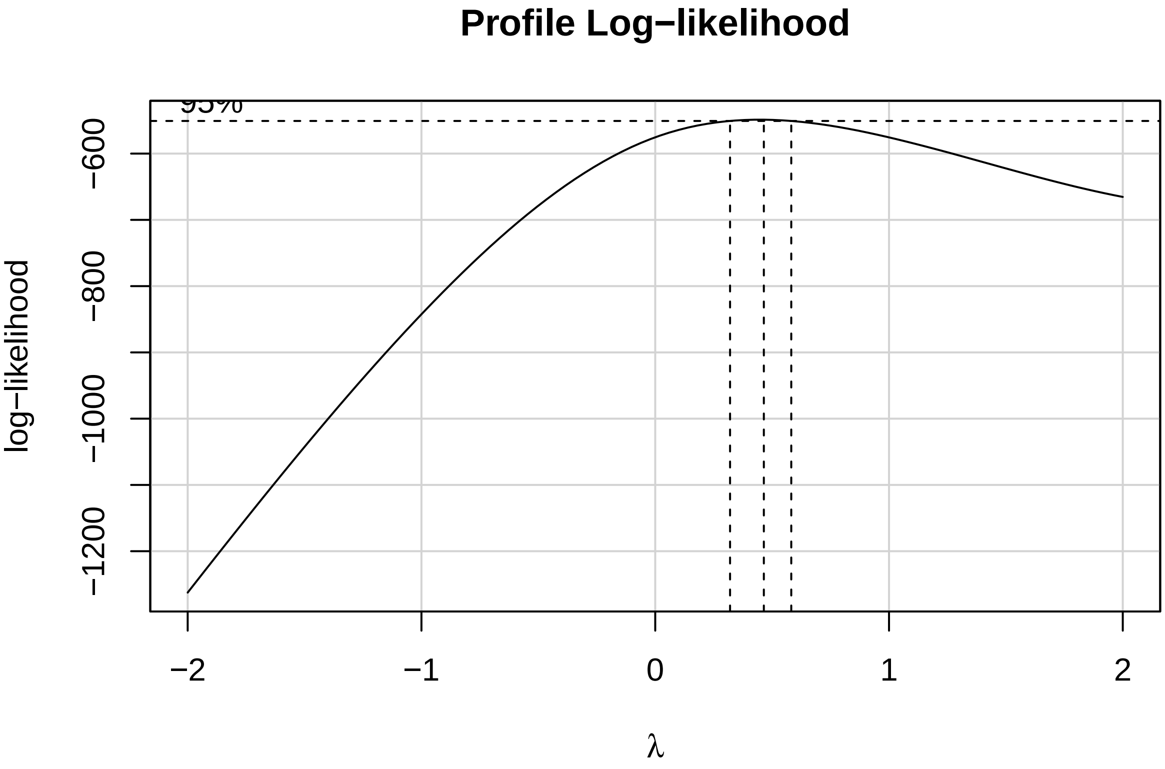
When model conditions fall short, we have found that the use of transformations can help to correct these short-comings. The boxCox function from the car package (Companion for Applied Regression) is one of many tools that may be used to help find a suitable transformation for the response variable in a linear model. To use this function, simply type boxCox(my.model) where my.model is a linear model object. The results will be a plot of possible power transformations (lambda) on the x-axis and corresponding log-likelihood values on the y-axis. The idea of the Box-Cox method is to select a value of λ that maximizes the log-likelihood. Vertical dashed lines on the plot highlight a window of plausible λ values.

Apply the boxCox function to the model that you fit for question 8. What transformation for Salary is suggested by the Box-Cox method?

```
library(car)
boxCox( model_salary )
```



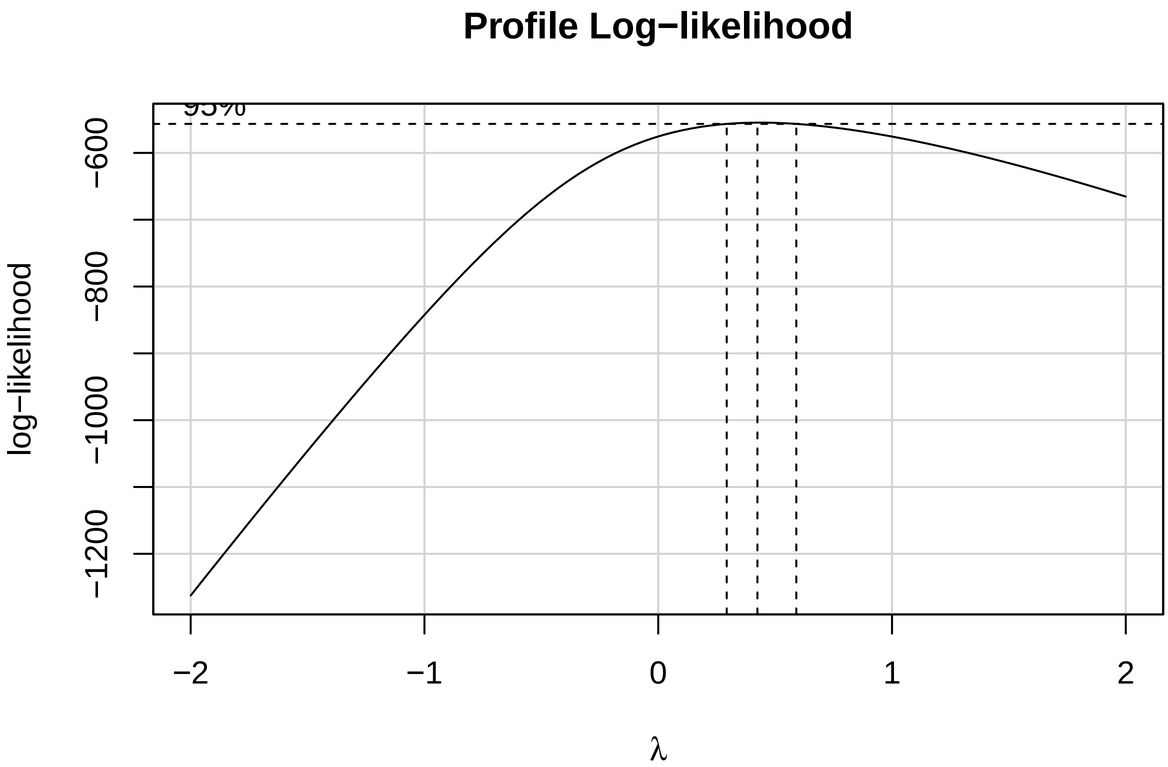
```
boxCox(model_salary, lambda = seq(-2,2))
```



The 95% confidence interval of lambda is between about 0.3 and 0.6

We can calculate the exact value

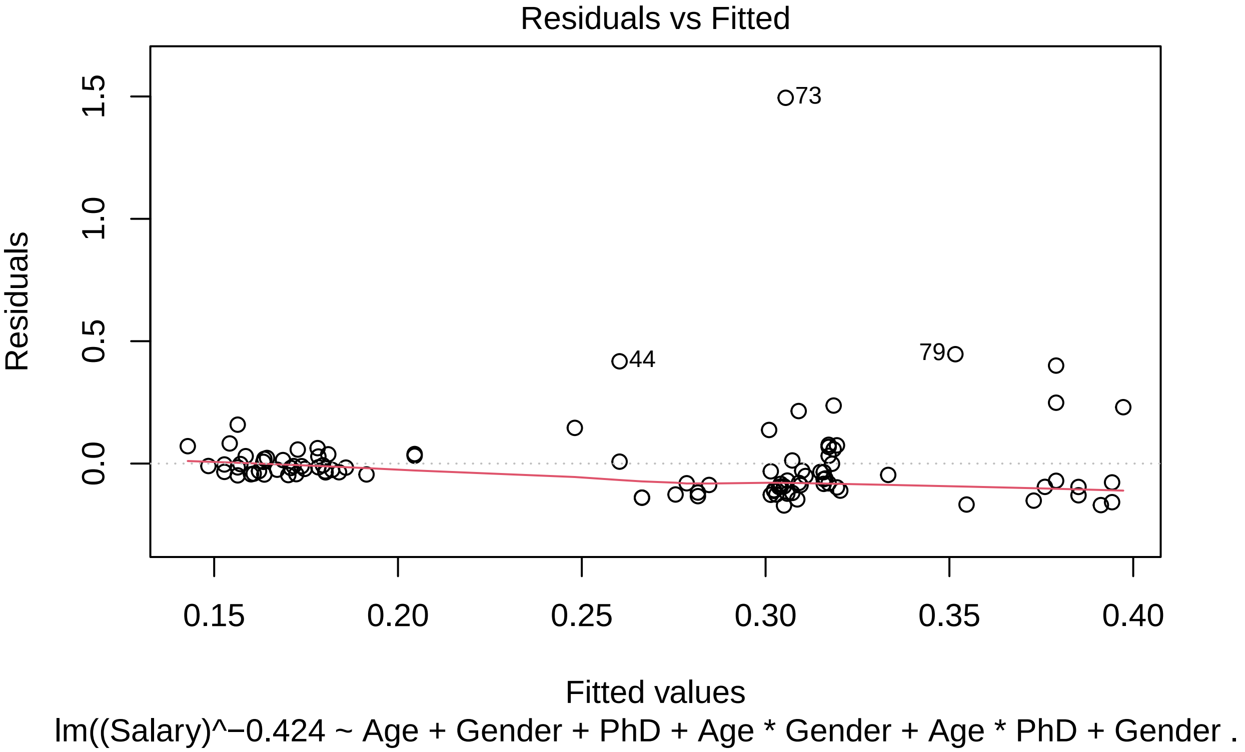
```
bc<-boxCox(model_salary)
```

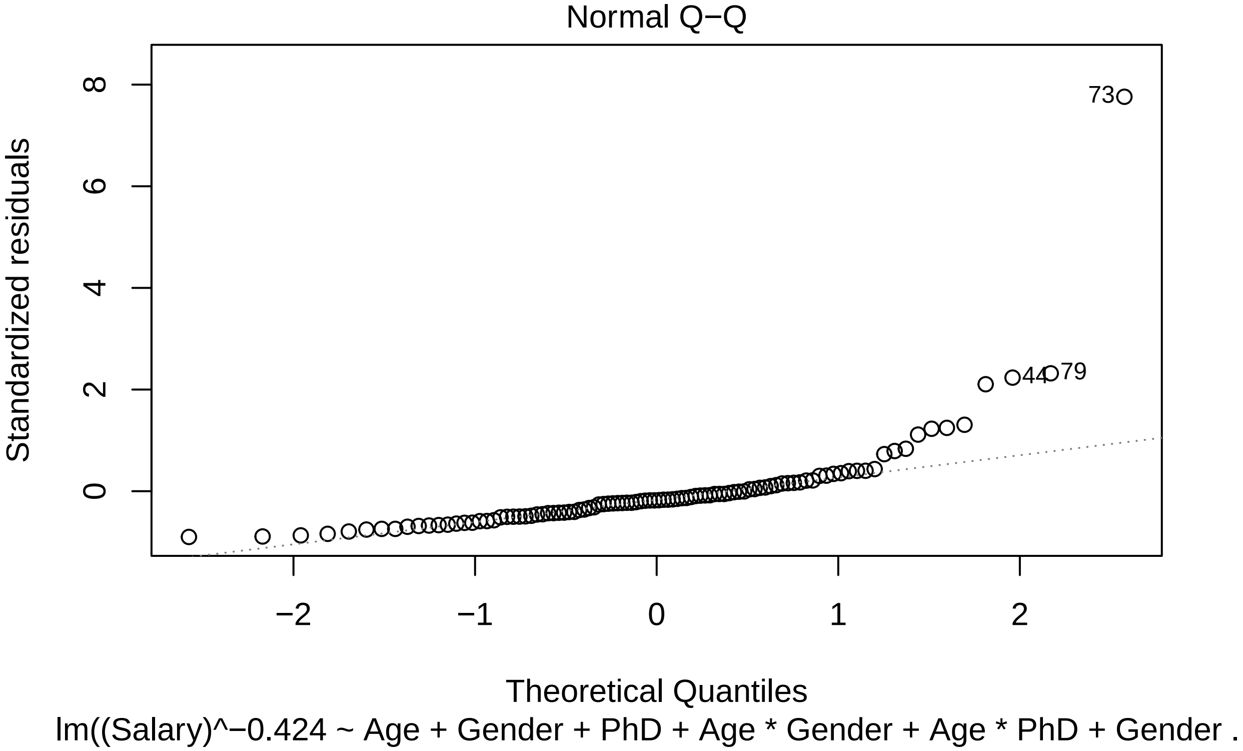


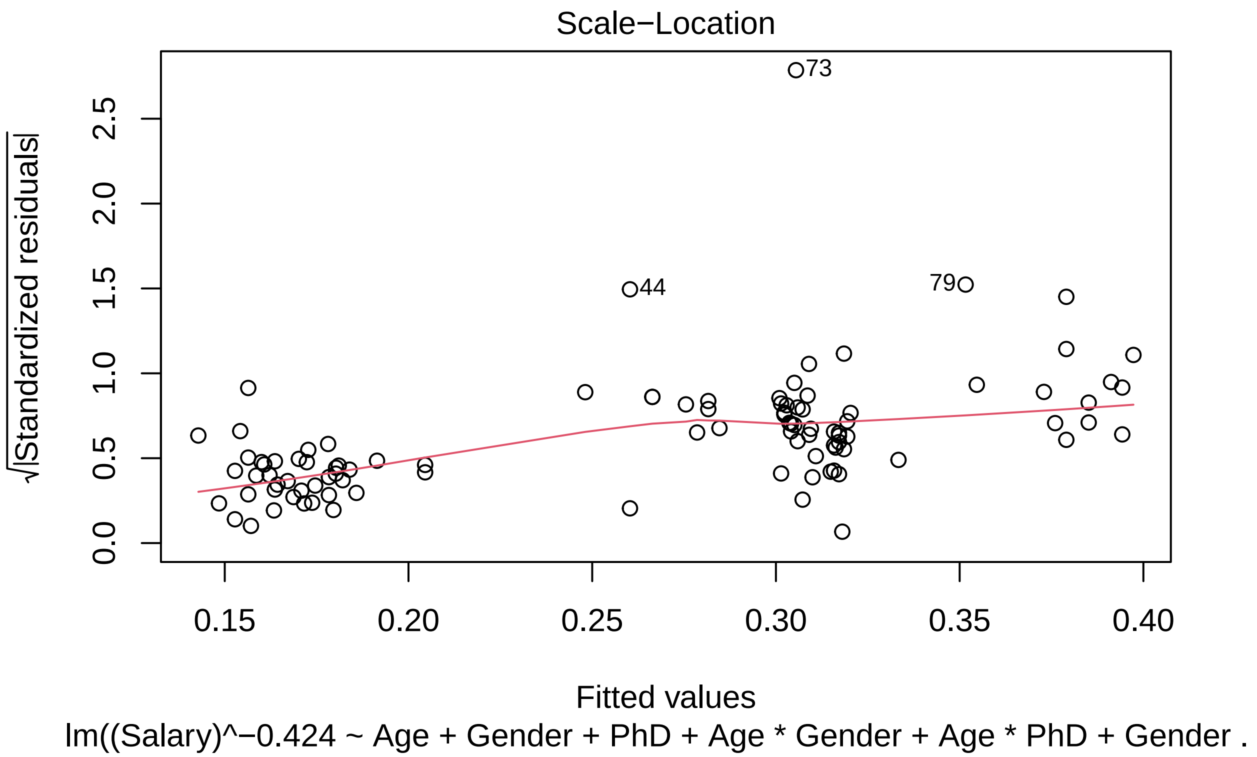
```
bc$x[which(bc$y==max(bc$y))]
```

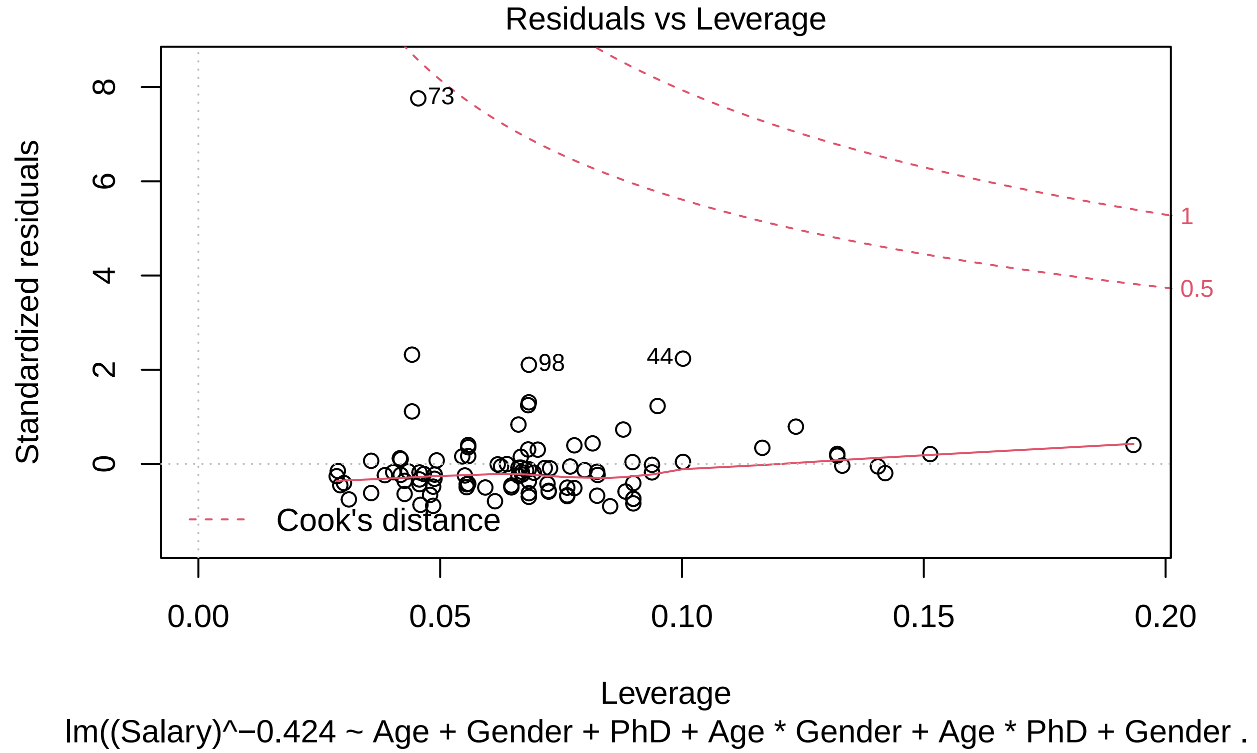
```
## [1] 0.4242424
```

```
fullmodel.bc <- lm((Salary)^-0.424~Age + Gender + PhD + Age*Gender + Age*PhD + Gender*PhD,  data=salary
plot(fullmodel.bc)
```

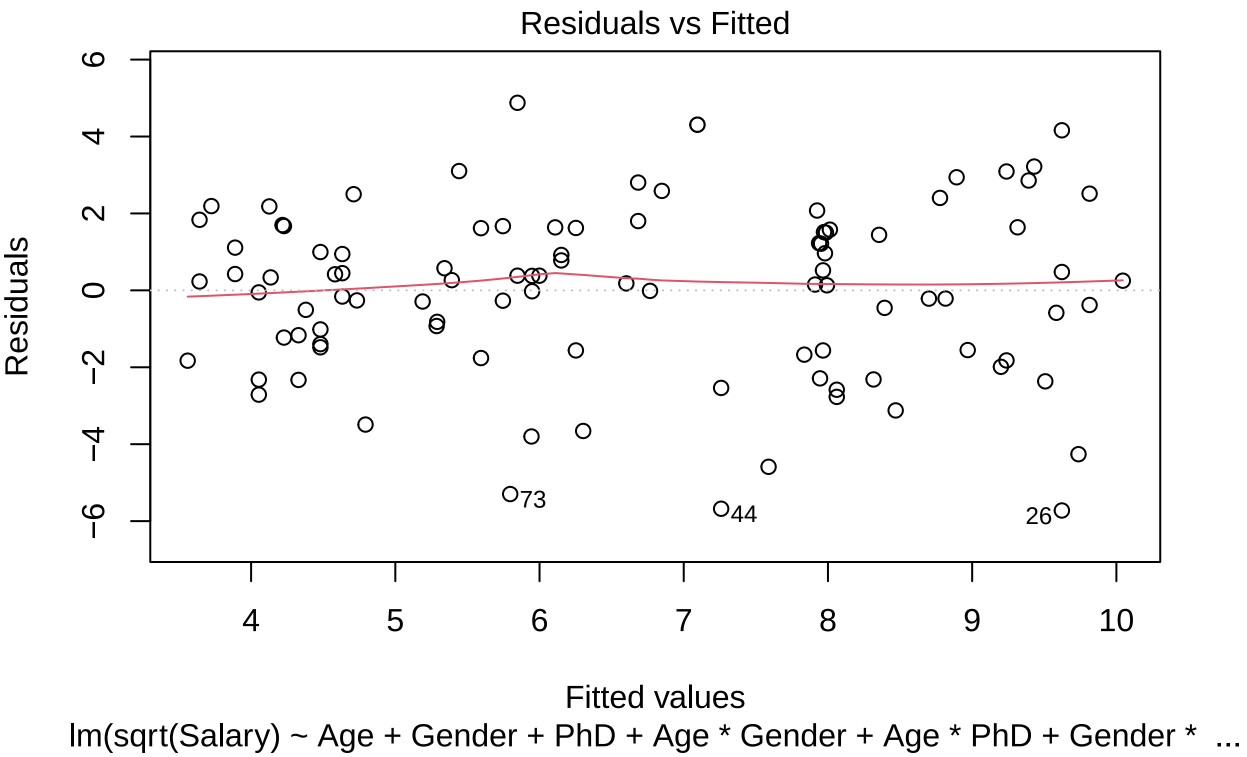

Residuals vs Fitted
Residuals
1.5
1.0
0.5
0.0
73
44
79
0.15
0.20
0.25
0.30
0.35
0.40
Fitted values
lm((Salary)^−0.424 ~ Age + Gender + PhD + Age * Gender + Age * PhD + Gender .

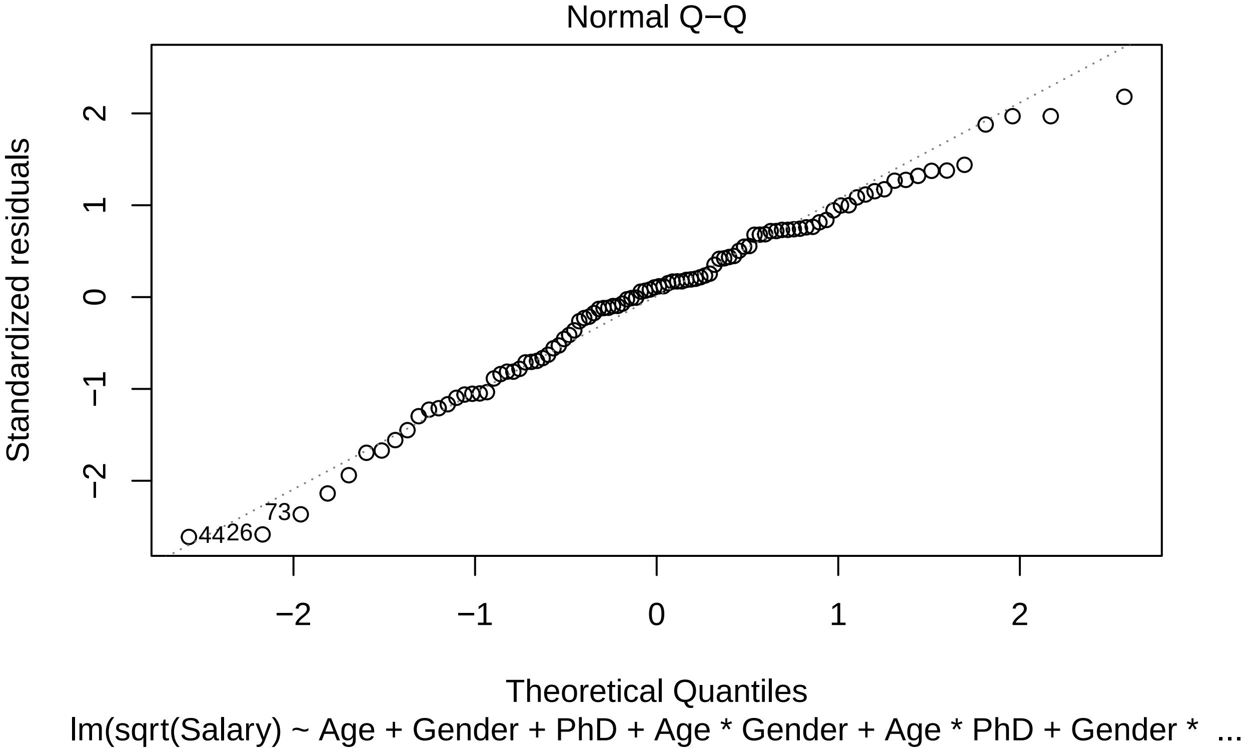

Normal Q−Q
Standardized residuals
8
6
4
2
0
73
44
79
−2
−1
0
1
2
Theoretical Quantiles
lm((Salary)^−0.424 ~ Age + Gender + PhD + Age * Gender + Age * PhD + Gender .

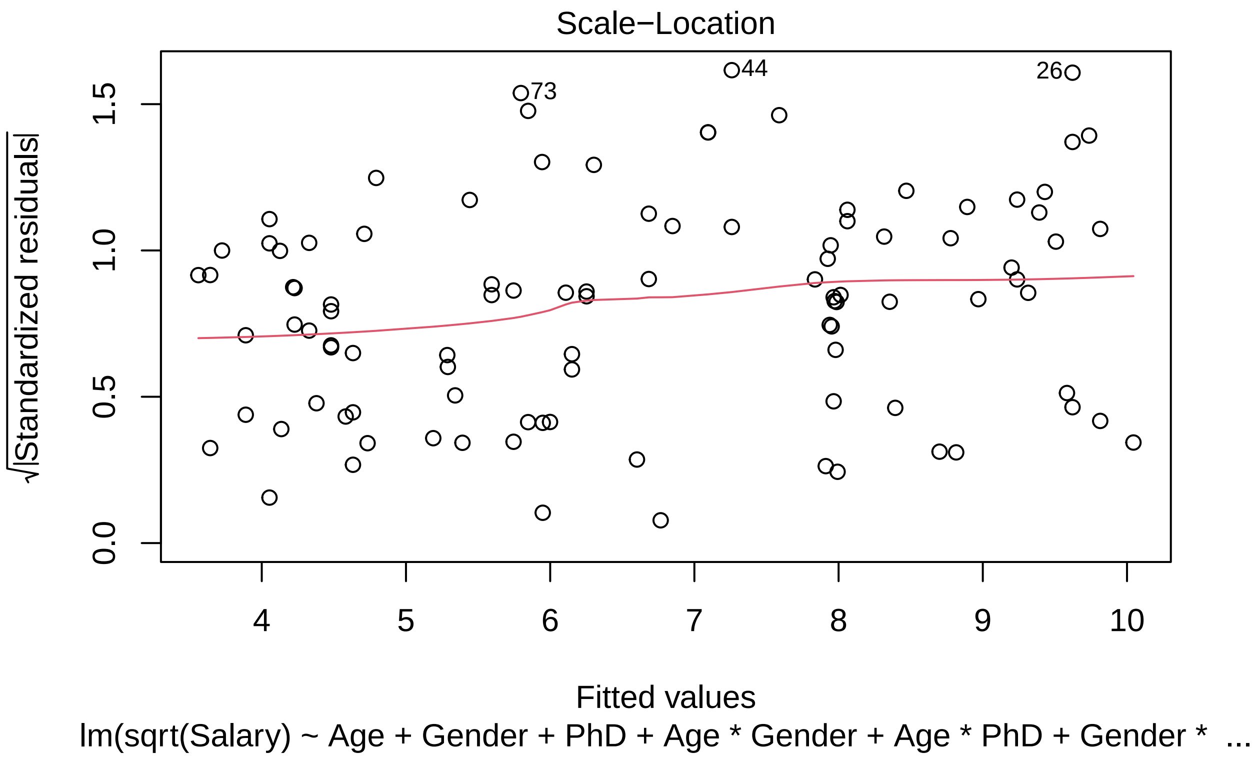


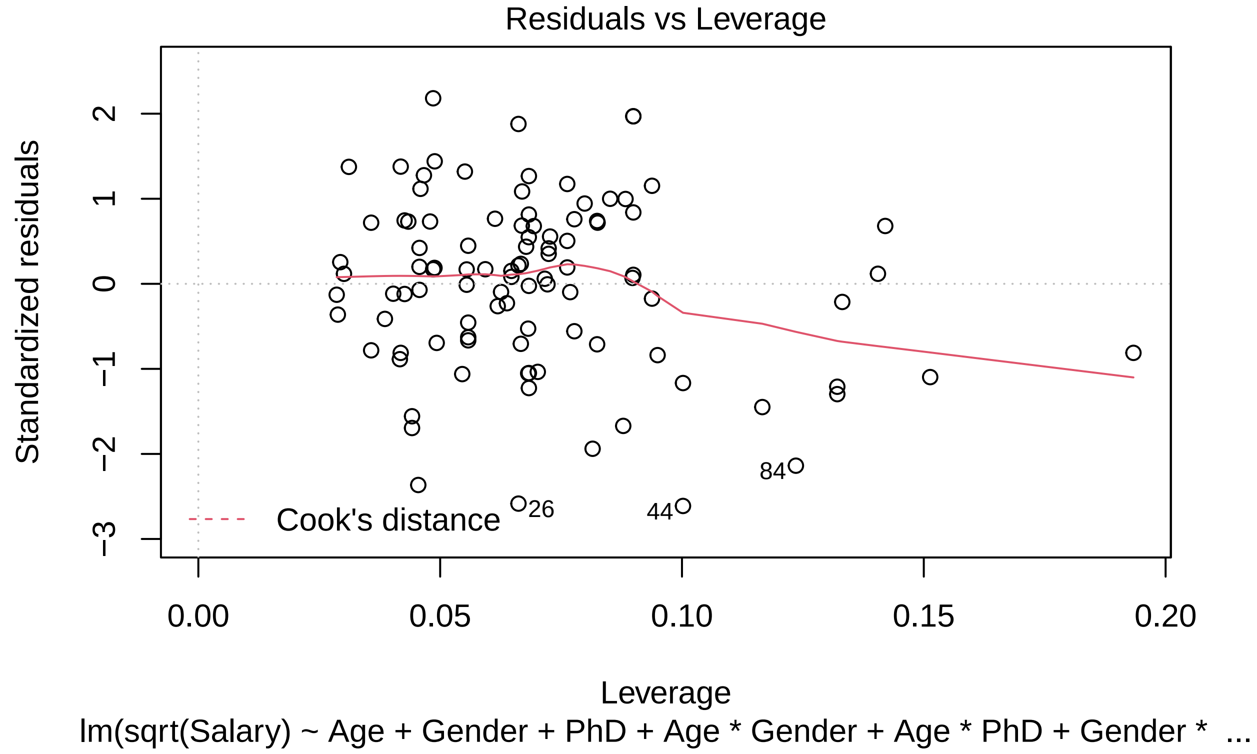


```
fullmodel.sqrt <- lm(sqrt(Salary)~Age + Gender + PhD + Age*Gender + Age*PhD + Gender*PhD,  data=salary_
plot(fullmodel.sqrt)
```


Residuals vs Fitted
Residuals
6
4
2
0
−2
−4
−6
73
44
26
4
5
6
7
8
9
10
Fitted values
lm(sqrt(Salary) ~ Age + Gender + PhD + Age * Gender + Age * PhD + Gender * ...


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
73
44
26
−2
−1
0
1
2
Theoretical Quantiles
lm(sqrt(Salary) ~ Age + Gender + PhD + Age * Gender + Age * PhD + Gender * ...





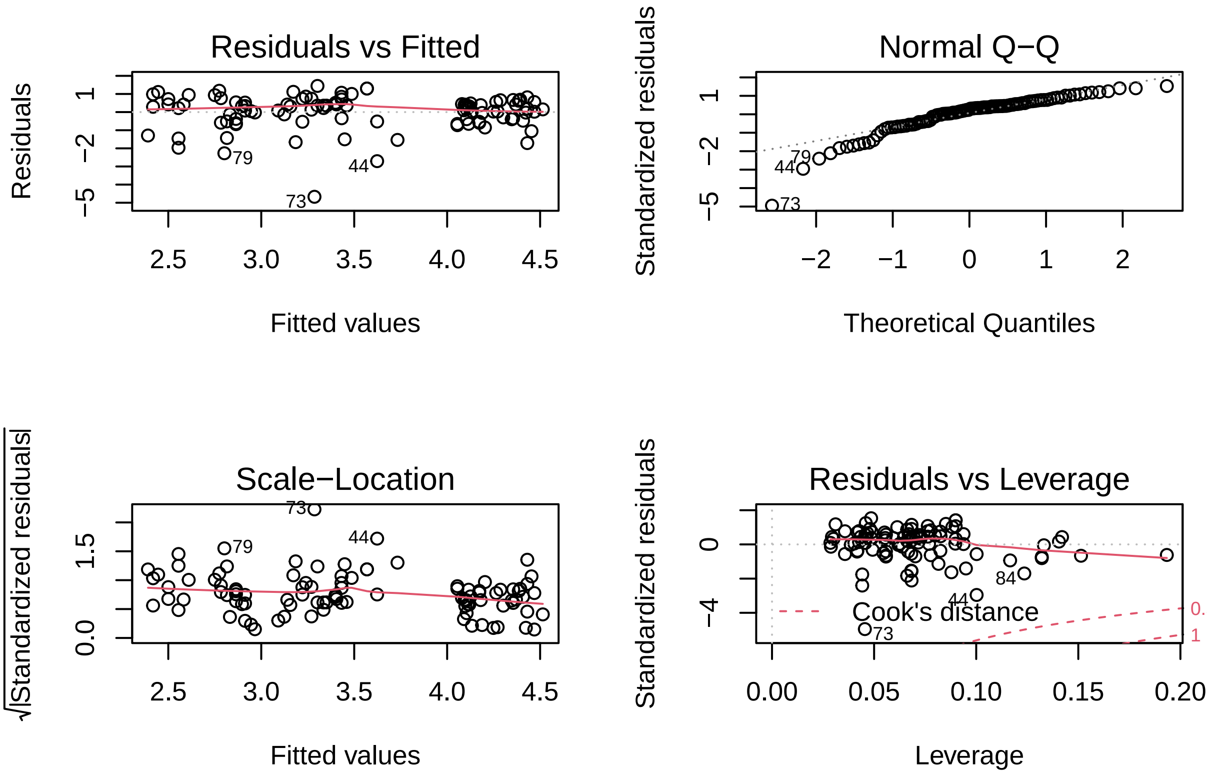
Answer 10

- $\lambda = 1/2$ or sqrt(Salary)

Question 11

Fit a model with the power transformation suggested by the boxCox function (your response to question 10) applied to Salary. Keep the same set of predictor variables as in question 9; Age, Gender, PhD and all possible two-term interactions. Plot the diagnostic 4-pack to assess the model conditions.

```r
par(mfrow=c(2,2))
plot(lm(log(Salary) ~ Age + Gender + PhD + Age*Gender + Age*PhD + Gender*PhD,  data=salary_gender))
```



```r
par(mfrow=c(1,1))
```

Has the equal variance condition improved on the transformed Salary when compared to the original Salary scale?

Answer 11

Yes

Question 12

Use Lasso with cross-validation to fit a model with same response variable (transformed Salary) and potential predictor variables (Age, Gender, PhD and all possible two-term interactions) as used in question 11. Use the full data set for this.

Using the lambda value that gives the minimum mean cross-validated error (lambda.min), select all model terms have non-zero coefficients in the Lasso model.

Answer 12

Age
Gender1 PhD1
Age:Gender1 Age:PhD1 Gender1:PhD1

```
library(glmnet)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-3
```

```
x = model.matrix(Salary~., salary_gender)[,-1]
y = salary_gender %>%
  select(Salary) %>%
  unlist() %>%
  as.numeric()
grid = 10^seq(10, -2, length = 100)
lasso_mod = glmnet(x,
                   y,
                   alpha = 1,
                   lambda = grid)



set.seed(1)
cv.out <-  cv.glmnet(x, y, alpha = 1)

bestlam <-  cv.out$lambda.min
bestlam
```

```
## [1] 0.1274824
```

```
#lasso_mod$lambda

coef(cv.out, cv.out$lambda.min)
```

```
## 4 x 1 sparse Matrix of class "dgCMatrix"
##                     s1
## (Intercept) -6.5295149
## Gender1     10.8939099
## Age          0.8417732
## PhD1        36.2685476
```

Question 13

Here are a few interpretation sentences based on the Lasso model fit.

- Assuming that age is fixed, the expected gap in college teacher salary between a female with a PhD and a female without a PhD is $1.69 thousand.
- Still assuming that age is fixed, the expected gap in salary between a male with PhD and a male without a PhD is $(1.69 + .28) thousand or approximately $2 thousand.
- For each additional year of age, we expect that college teacher salary will increase on average by $30 with gender and PhD status held constant.

There is a **major** error that is common to all three of these interpretation sentences. Explain what is incorrect about all of these statements. (This is an open response question so it will not be auto-graded by Canvas.)

Answer 13

Type your explanation in Canvas quiz LASSO (a penalized estimation method) aims at estimating the same quantities (model coefficients) as, say, OLS maximum likelihood (an unpenalized method). The model is the same, and the interpretation remains the same. The numerical values from LASSO will normally differ from those from OLS maximum likelihood: some will be closer to zero, others will be exactly zero. If a sensible amount of penalization has been applied, the LASSO estimates will lie closer to the true values than the OLS maximum likelihood estimates, which is a desirable result.

There is no inherent problem with that, but you could use LASSO not only for feature selection but also for coefficient estimation. As I mention above, LASSO estimates may be more accurate than, say, OLS maximum likelihood estimates.

Question 14

We now want to apply step-wise model selection algorithms (forward, backward, and both) for this context of predicting Salary based on Age, Gender, and PhD. We will use the transformed Salary as our response variable. We will look for an optimal model between the simple intercept-only model (Y~1) and the full model that uses the predictor variables from question 10 (Age, Gender, PhD and all possible two-term interactions). Use the step function three times: once with direction = "backward", once with direction = "forward", and once with direction = "both". You should find that all three algorithms land on the same final model. What terms are included in the model selected by step?

Answer 14

Age
Gender1 PhD1
Age:Gender1 Age:PhD1 Gender1:PhD1

```
step(lm(log(Salary) ~ Age + Gender + PhD ,  data=salary_gender),direction="backward")
```

```
## Start:  AIC=-3.71
## log(Salary) ~ Age + Gender + PhD
##
##          Df Sum of Sq     RSS     AIC
## - Gender  1    0.0173  88.967 -5.6904
## <none>                 88.950 -3.7098
## - Age     1    6.5043  95.454  1.3476
## - PhD     1   20.6792 109.629 15.1932
##
## Step:  AIC=-5.69
## log(Salary) ~ Age + PhD
##
##        Df Sum of Sq     RSS     AIC
## <none>               88.967 -5.6904
## - Age   1     6.505  95.472 -0.6336
## - PhD   1    20.915 109.882 13.4235
```

```
##
## Call:
## lm(formula = log(Salary) ~ Age + PhD, data = salary_gender)
##
## Coefficients:
## (Intercept)          Age         PhD1
##     2.29065      0.01794      1.00217
```

```
step(lm(log(Salary) ~ Age + Gender + PhD ,  data=salary_gender),direction="forward")
```

```
## Start:  AIC=-3.71
## log(Salary) ~ Age + Gender + PhD
```

```
##
## Call:
## lm(formula = log(Salary) ~ Age + Gender + PhD, data = salary_gender)
##
## Coefficients:
## (Intercept)          Age      Gender1         PhD1
##     2.29801      0.01804     -0.02690      1.00620
```

```
step(lm(log(Salary) ~ Age + Gender + PhD ,  data=salary_gender),direction="both")
```

```
## Start:  AIC=-3.71
## log(Salary) ~ Age + Gender + PhD
##
##          Df Sum of Sq     RSS     AIC
## - Gender  1    0.0173  88.967 -5.6904
## <none>                 88.950 -3.7098
## - Age     1    6.5043  95.454  1.3476
## - PhD     1   20.6792 109.629 15.1932
##
## Step:  AIC=-5.69
## log(Salary) ~ Age + PhD
##
```

```
##          Df Sum of Sq     RSS     AIC
## <none>                 88.967 -5.6904
## + Gender  1    0.0173  88.950 -3.7098
## - Age     1    6.5050  95.472 -0.6336
## - PhD     1   20.9148 109.882 13.4235
```

```
## 
## Call:
## lm(formula = log(Salary) ~ Age + PhD, data = salary_gender)
## 
## Coefficients:
## (Intercept)          Age         PhD1  
##     2.29065      0.01794      1.00217
```

```
step(model_salary,direction="backward")
```

```
## Start:  AIC=704.57
## Salary ~ Age + Gender + PhD + Age * Gender + Age * PhD + Gender * 
##     PhD
## 
##              Df Sum of Sq    RSS    AIC
## - Age:PhD     1    397.97 100194 702.97
## - Gender:PhD  1    861.31 100657 703.43
## - Age:Gender  1   1465.11 101261 704.03
## <none>                     99796 704.57
## 
## Step:  AIC=702.97
## Salary ~ Age + Gender + PhD + Age:Gender + Gender:PhD
## 
##              Df Sum of Sq    RSS    AIC
## - Gender:PhD  1    808.88 101003 701.77
## - Age:Gender  1   1181.26 101375 702.14
## <none>                    100194 702.97
## 
## Step:  AIC=701.77
## Salary ~ Age + Gender + PhD + Age:Gender
## 
##              Df Sum of Sq    RSS    AIC
## <none>                    101003 701.77
## - Age:Gender  1    2159.1 103162 701.89
## - PhD         1   27354.6 128357 723.74
```

```
## 
## Call:
## lm(formula = Salary ~ Age + Gender + PhD + Age:Gender, data = salary_gender)
## 
## Coefficients:
## (Intercept)          Age      Gender1         PhD1  Age:Gender1  
##      9.9381       0.4661     -18.2701      36.6009       0.6325
```

```
step(model_salary,direction="forward")
```

```
## Start:  AIC=704.57
## Salary ~ Age + Gender + PhD + Age * Gender + Age * PhD + Gender *
##     PhD
```

```
##
## Call:
## lm(formula = Salary ~ Age + Gender + PhD + Age * Gender + Age *
##     PhD + Gender * PhD, data = salary_gender)
##
## Coefficients:
##  (Intercept)           Age       Gender1          PhD1   Age:Gender1
##       6.8956        0.5834      -20.1322       46.5398        0.5707
##     Age:PhD1  Gender1:PhD1
##      -0.3354       13.0493
```

```
step(model_salary,direction="both")
```

```
## Start:  AIC=704.57
## Salary ~ Age + Gender + PhD + Age * Gender + Age * PhD + Gender *
##     PhD
##
##              Df Sum of Sq    RSS    AIC
## - Age:PhD     1    397.97 100194 702.97
## - Gender:PhD  1    861.31 100657 703.43
## - Age:Gender  1   1465.11 101261 704.03
## <none>                     99796 704.57
##
## Step:  AIC=702.97
## Salary ~ Age + Gender + PhD + Age:Gender + Gender:PhD
##
##              Df Sum of Sq    RSS    AIC
## - Gender:PhD  1    808.88 101003 701.77
## - Age:Gender  1   1181.26 101375 702.14
## <none>                    100194 702.97
## + Age:PhD     1    397.97  99796 704.57
##
## Step:  AIC=701.77
## Salary ~ Age + Gender + PhD + Age:Gender
##
##              Df Sum of Sq    RSS    AIC
## <none>                    101003 701.77
## - Age:Gender  1    2159.1 103162 701.89
## + Gender:PhD  1     808.9 100194 702.97
## + Age:PhD     1     345.5 100657 703.43
## - PhD         1   27354.6 128357 723.74
```

```
##
## Call:
## lm(formula = Salary ~ Age + Gender + PhD + Age:Gender, data = salary_gender)
##
## Coefficients:
## (Intercept)          Age      Gender1         PhD1  Age:Gender1
##      9.9381       0.4661     -18.2701      36.6009       0.6325
```

Question 15

This example is interesting and to some degree incomplete. The question of whether or not Gender is associated with notable differences in Salary is very relevant. The model settled on using Lasso included a term connected to Gender. The model that the step function led us to did not include a term related to Gender. If you try splitting the data into training/test sets, you will find that the choice of terms for the trained model will vary between random splits, confusing the story to an even greater degree. This is a place where visualizations, descriptive statistics, along with statistical models could work together to tell a more complete story.

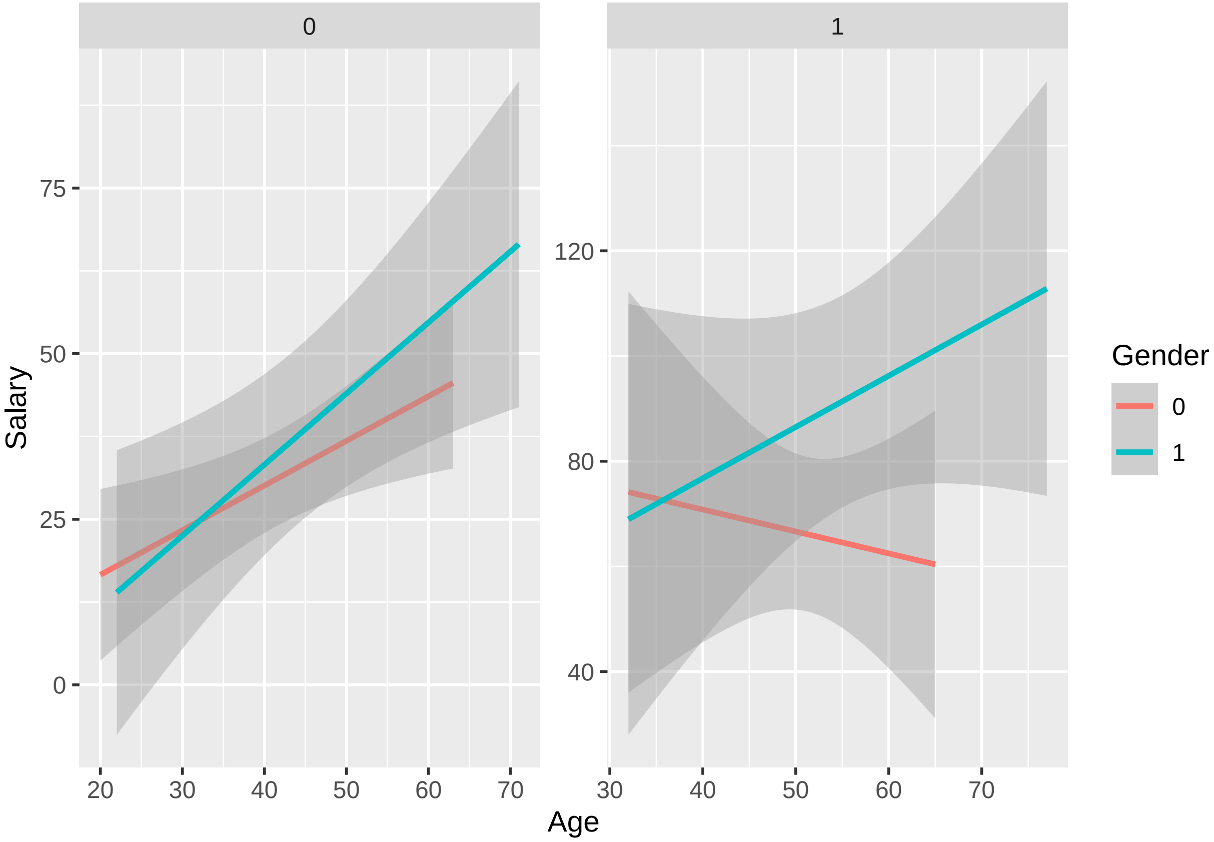
Make a scatter plot that displays Salary on the y-axis, Age on the x-axis, uses different colors for Gender, and facets by PhD (facet_wrap). Additionally, use geom_smooth with method="lm" to overlay trend lines for each PhD by Gender combination.

Which one subgroup has a negatively sloping trend line on your plot?

Answer 15

- Women with a PhD

```
ggplot(salary_gender, aes(Age, Salary,  colour = Gender)) +
  geom_smooth(method="lm") +
  facet_wrap(vars(PhD), scales = "free")
```


Salary
Age
Gender