Stat 3301: Homework 6

Due by date and time specified on Carmen

Jane Weissberg (weissberg.11)

Setup:

```r
knitr::opts_chunk$set(echo = TRUE)
library(alr4)
library(tidyverse)
library(readr)
```

Instructions

- Replace "FirstName LastName (name.n)" above with your information.
- Provide your solutions below in the spaces marked "Solution:".
- Include any R code that you use to answer the questions; if a numeric answer is required, show how you calculated it in R. I have set the global option `echo = TRUE` to make sure the R code is displayed.
- Knit this document to HTML and upload both the HTML file and your completed Rmd file to Carmen
- Make sure your solutions are clean and easy-to-read by
  - formatting all plots to be appropriately sized, with appropriate axis labels.
  - only including R code that is necessary to answer the questions below.
  - only including R output that is necessary to answer the questions below (avoiding lengthy output).
  - providing short written answers explaining your work, and writing in complete sentences.
- Data files mentioned below are from the `alr4` package unless specified otherwise.

Question 1 Problem 2.20.1 and 2.20.2 from Textbook,Weisberg: **Old Faithful** (Data file: `oldfaith`).

Note: the question asks for a confidence interval, but think carefully about the quantity you are estimating and what type of interval would be appropriate.

Solutions

- **Part 1:**

```r
data('oldfaith')
model <- lm(Interval ~ Duration, data = oldfaith)
coef(model)
```

```
## (Intercept)    Duration
##  33.9878076   0.1768629
```

$\hat{y} = 33.9878 + 0.1769x$ $\hat{y}$ represents the predicted interval. 33.9878 is the predicted intercept, meaning when the duration of the current eruption is 0 seconds, the interval or the time until the next eruption is predicted to be 33.9878 minutes. 0.1769 is the predicted slope. This means that every time the duration of the current eruption increases by 1, the predicted time until the next eruption increases by 0.1769.

- **Part 2:**

```r
new_duration <- data.frame(Duration = 250)
predicted_wait <- predict(model, newdata = new_duration, interval = "confidence", level = 0.95)
predicted_wait
```

```
##        fit      lwr      upr
## 1 78.20354 77.36915 79.03794
```

The 95% confidence interval for the time the individual will have to wait for the next eruption when the last eruption was 250 seconds long is (77.369, 79.0379).

Question 2 This question relates to the **salary** data set which can be found in `alr4` package. Simply load the package and access salary date set.

This data set about the salary of faculty members in a small Midwestern college in the early 1980s.

This data frame contains the following columns:

- degree: Factor with levels "PhD" or "Masters"
- rank: Factor, "Asst", "Assoc" or "Prof"
- sex: Factor, "Male" or "Female"
- year: Years in current rank
- ysdeg: Years since highest degree earned
- salary: dollars per year

You do not need to filter the data set by Male and use the **salary** data as it is.

Here consider `salary` is the response variable and `year` is the predictor variable for the analysis.

1. Calculate the total sum of squares, the regression sum of squares and the residual sum of squares for the fitted regression model. Use these to calculate the coefficient of determination (R^2) for the model. What do these values tell you about the strength of the linear relationship between `salary` and `year`?

2. Use appropriate residual plots to assess whether the fitted mean function is appropriate. Do the plots indicate any lack of fit?

3. Use appropriate residual plots to assess the assumption that the variance of employer salary is the same at all years.
4. Use appropriate residual plots to assess the assumption that the error terms in the model are normally distributed.

   (Make your best judgment on these plots to answer the questions)

Solutions

- **Part 1:**

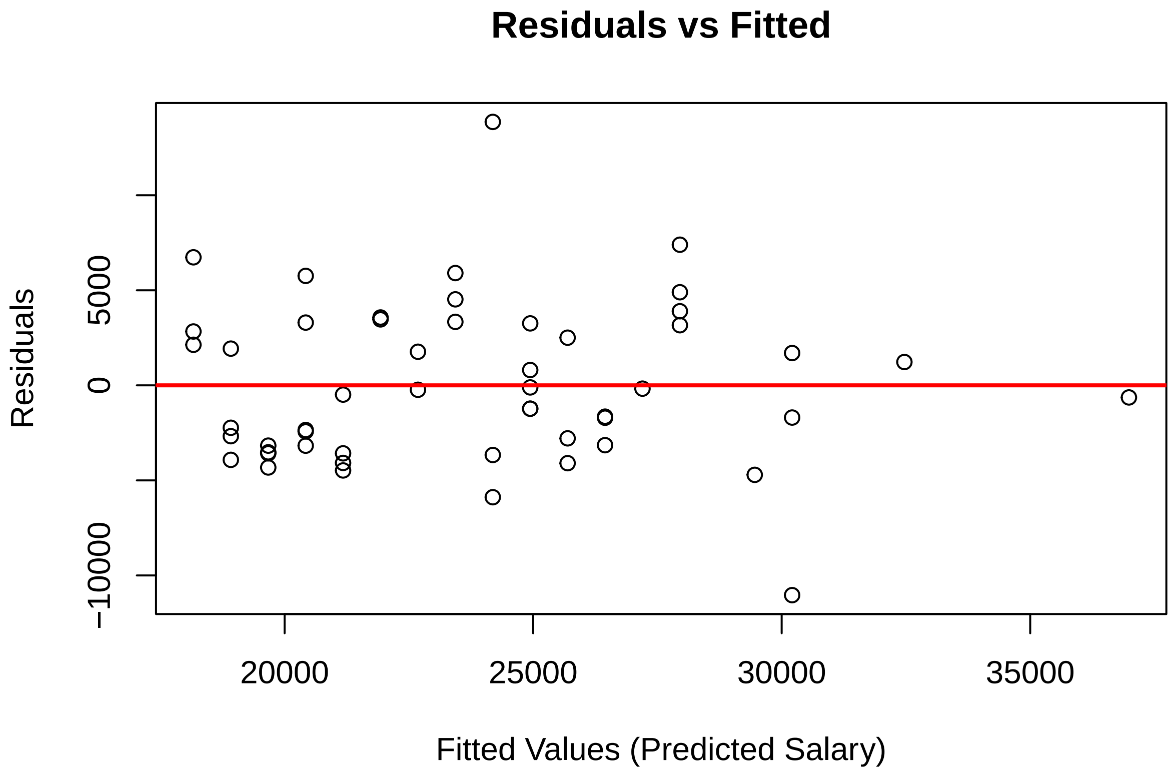
```
data(salary)

# Fit a simple linear regression model: salary ~ year
fit <- lm(salary ~ year, data = salary)

# Get the fitted values and residuals
fitted_values <- fitted(fit)
residuals <- resid(fit)

# Total sum of squares (TSS)
TSS <- sum((salary$salary - mean(salary$salary))^2)

# Regression sum of squares (SSreg)
SSreg <- sum((fitted_values - mean(salary$salary))^2)

# Residual sum of squares (RSS)
RSS <- sum(residuals^2)

# Coefficient of determination (R^2)
R_squared <- SSreg / TSS

# Output the results
cat("Total Sum of Squares = ", TSS,
"Regression Sum of Squares = ", SSreg,
"Residual Sum of Squares = ", RSS,
"R Squared = ", R_squared)
```

```
## Total Sum of Squares =  1785729858 Regression Sum of Squares =  876680907 Residual Sum of Squares =
```

0.490937 is the proportion of variability in salary that is explained by years in rank. There is a lot of variability not explained by the model.

- **Part 2:**

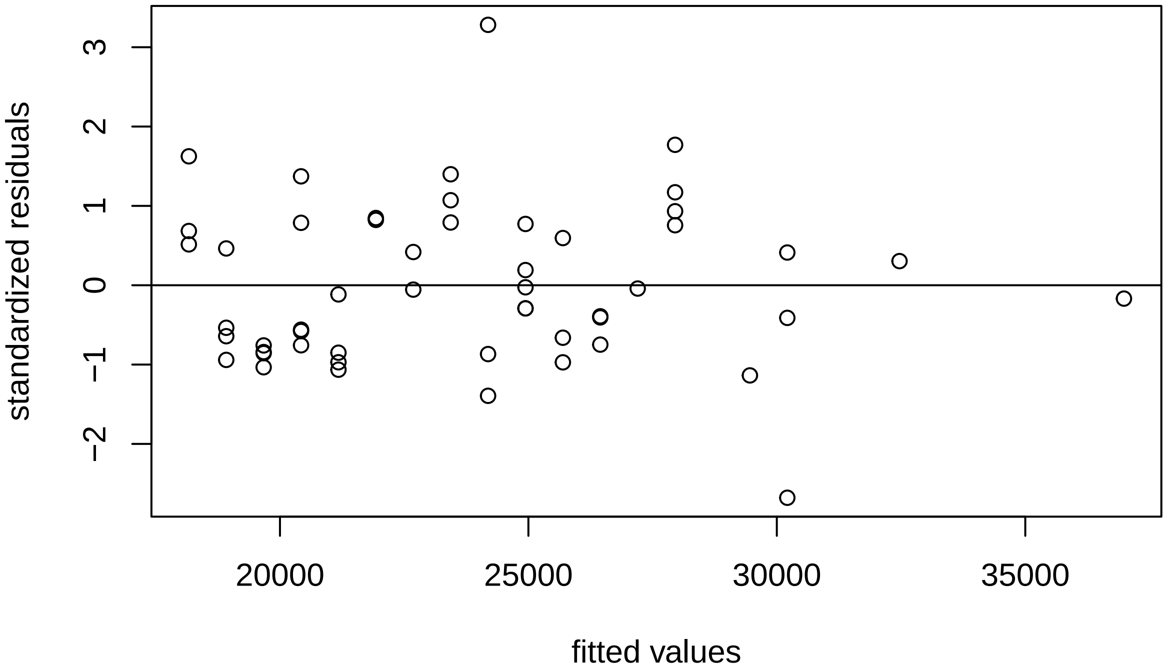
```
plot(fitted_values, residuals,
     xlab = "Fitted Values (Predicted Salary)",
     ylab = "Residuals",
     main = "Residuals vs Fitted")
abline(h = 0, col = "red", lwd = 2)
```



The plot does not indicate any lack of fit because there is no clear pattern.

- **Part 3:**

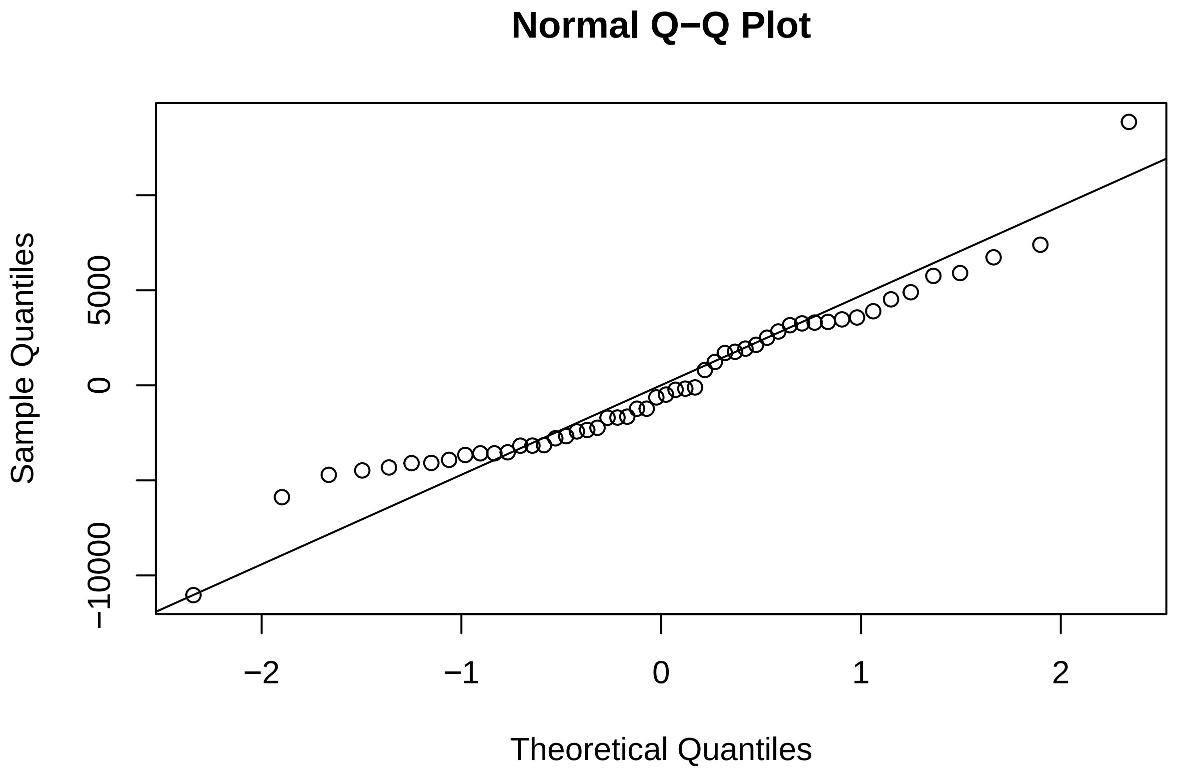
```
plot(fitted_values, rstandard(fit), xlab = "fitted values",
ylab = "standardized residuals", main = ""); abline(h = 0)
```



It does not seem like there is constant variance for employer salary at all years.

- **Part 4:**

```
qqnorm(residuals); qqline(residuals)
```



The error terms in the model seem to be normally distributed.