DSCC/CSC/STAT 462 Assignment 5

Due December 1, 2022 by 11:59 p.m.

Daxiang Na

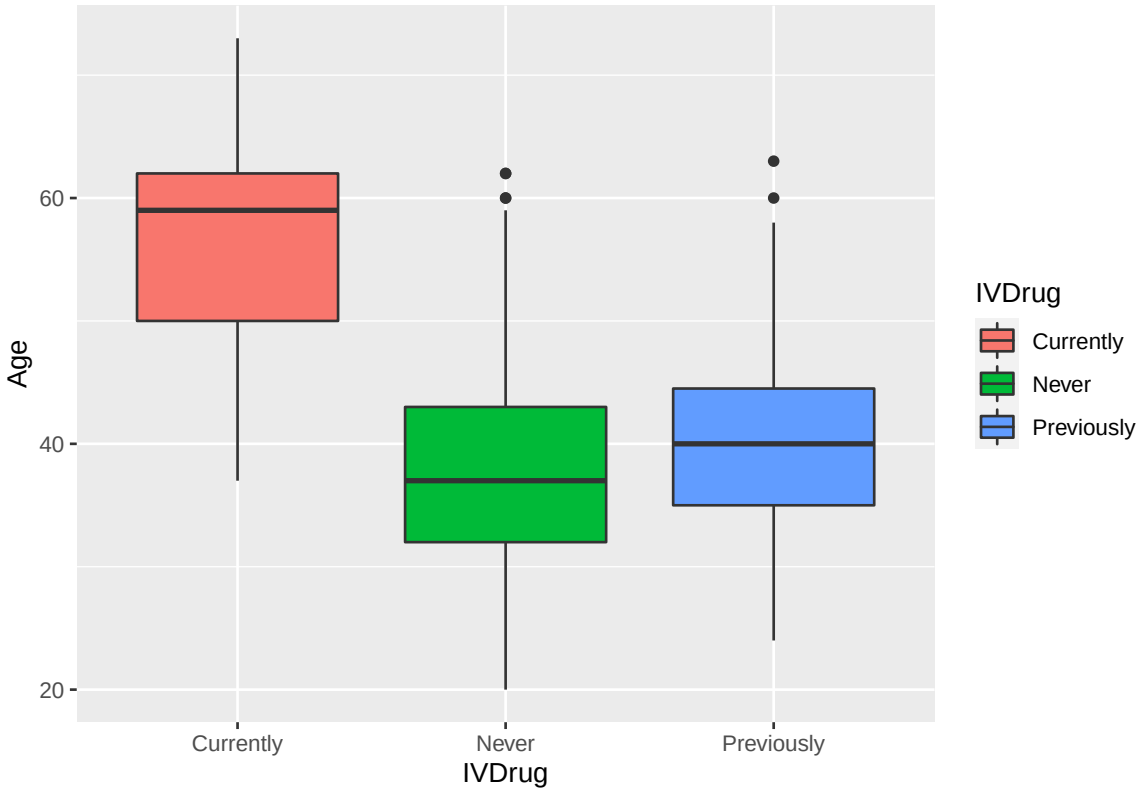
2022.12.05

Please complete this assignment using `RMarkdown`, and submit the knitted PDF. *For all hypothesis tests, state the hypotheses, report the test statistic and p-value, and comment on the results in the context of the problem.*

1. The dataset "actg.csv" contains data on subjects who were enrolled in a HIV clinical trial. For this dataset, we are focusing on variables `IVDrug` and `Age`. `IVDrug` is a categorical variable that indicates whether each subject never, previously, or currently uses IV drugs. `Age` lists the subject's age in year. Thus, we have age for subjects in three treatment groups. Assume that the samples were collected independently and come from normally distributed populations.

    a. Create side-by-side boxplots of age by IV use group. Does the equal variance assumption seem reasonable?

```
df <- read.csv("actg.csv")
p <- ggplot(data = df, aes(x = IVDrug, y = Age))
p + geom_boxplot(aes(fill = IVDrug))
```



b. Construct an ANOVA table to test at the $\alpha = 0.05$ significance level whether the average age in these three groups is different.

```
n1 <- df %>%
    filter(IVDrug == "Currently") %>%
    nrow()
n2 <- df %>%
    filter(IVDrug == "Never") %>%
    nrow()
n3 <- df %>%
    filter(IVDrug == "Previously") %>%
    nrow()
n <- n1 + n2 + n3
s <- df %>%
    group_by(IVDrug) %>%
    summarise(sd = sd(Age), mean = mean(Age))
s1 <- pull(s[1, "sd"])
m1 <- pull(s[1, "mean"])
s2 <- pull(s[2, "sd"])
m2 <- pull(s[2, "mean"])
s3 <- pull(s[3, "sd"])
m3 <- pull(s[3, "mean"])
m <- (n1 * m1 + n2 * m2 + n3 * m3)/n
```

```
SSE <- (n1 - 1) * s1^2 + (n2 - 1) * s2^2 + (n3 - 1) * s3^3
sw <- SSE/(n - 3)
SSB <- n1 * (m1 - m)^2 + n2 * (m2 - m)^2 + n3 * (m3 - m)^2
sb <- SSB/(3 - 1)
f <- sb/sw
p <- 1 - pf(f, (3 - 1), (n - 3))
```

```
sw
```

```
## [1] 110.0596
```

```
sb
```

```
## [1] 4601.691
```

```
m
```

```
## [1] 38.61251
```

```
f
```

```
## [1] 41.8109
```

```
p
```

```
## [1] 0
```

Answer: $s_w = 110.0596$, $s_b = 4601.691$, $\bar{x} = 38.61251$, $F = \frac{s_b^2}{s_w^2} = 41.8109$, $p = 0 < \alpha = 0.05$. We reject the null hypothesis and conclude that the equal variance assumption is not reasonable.

c. Further explore the results using a Bonferroni multiple comparison procedure with overall familywise error rate of $\alpha_{FWE} = 0.05$. For this part, use the `pairwise.t.test()` function with `p.adj="bonferroni"`.

```
pairwise.t.test(df$Age, df$IVDrug, p.adj = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  df$Age and df$IVDrug
##
##            Currently Never
## Never      <2e-16    -
## Previously <2e-16    0.0036
##
## P value adjustment method: bonferroni
```

d. As an alternative approach to Bonferroni's multiple comparison adjustment, we will explore explore Scheffe's method for multiple comparisons. Scheffe's method is a more general method that does not depend on the number of comparisons

being made (whereas Bonferroni directly adjust for that in doing α/k as the significance level). With Scheffe's method, the experimentwise error rate is α (i.e. the overall significance level), and it ensure that the probability of declaring at least one false significant comparison is at most α. This method is preferable to Bonferroni particularly in cases when you are looking at many comparisons. For more information about Scheffe's method I would recommend taking a look at the video here: https://www.youtube.com/watch?v=l6i0xhnIzzk. Essentially, though, this is just another approach for testing for multiple comparisons. **For this question, implement Scheffe's method to test for significant pairwise differences at the experimentwise error rate of** $\alpha = 0.05$, and interpret the results. To do Scheffe's method, you first need to go to the R Console (i.e. not directly in your .RMD document) and install the `DescTools` package. Then, in your .RMD document, load the package into R using `library(DescTools)`. The `ScheffeTest()` function will allow you to perform Scheffe's method of multiple comparisons. Look at `help(ScheffeTest)` for more information, but generally, you will put your `aov` object into the `ScheffeTest()` function, and it will look for significant pairwise comparisons.

```
model1 <- aov(Age ~ IVDrug, data = df)
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: Age
##             Df Sum Sq Mean Sq F value    Pr(>F)
## IVDrug       2   9203  4601.7  72.452 < 2.2e-16 ***
## Residuals 1148  72914    63.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ScheffeTest(model1, g = IVDrug)
```

```
##
##   Posthoc multiple comparisons of means: Scheffe Test
##     95% family-wise confidence level
##
## $IVDrug
##                              diff      lwr.ci     upr.ci   pval
## Never-Currently      -18.010745 -21.7564656 -14.265023 <2e-16 ***
## Previously-Currently -15.901836 -19.8714375 -11.932234 <2e-16 ***
## Previously-Never       2.108909   0.5165364   3.701281 0.0053 **
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: for all of the pairs (Never-Currently, Previously-Currently and Previously-Never), adjusted p values are lower than 0.05. We rejected the null hypothesis and

conclude that there is significant difference between each pairs.

The rest of the questions in this homework assignment rely on the following premise. Suppose a nonprofit organization is looking to analyze trends among its donors, and they have collected data in "donors.csv". In particular, suppose they have a sample of 94 donors who gave a gift in response to a recent solicitation. Suppose you have been hired by the organization to analyze trends among the donors. We ultimately are interested in predicting donation amount based on other factors. Begin by predicting donation amount (`amount`) based on the income a donor is making (`income`).

2. Let's begin with exploratory analysis.

   a. State the regression assumptions in terms of the variables we are investigating.

   Answer: the donation amount and the income a donor is making are correlated.

   b. What is the Pearson correlation between the donation amount and income?

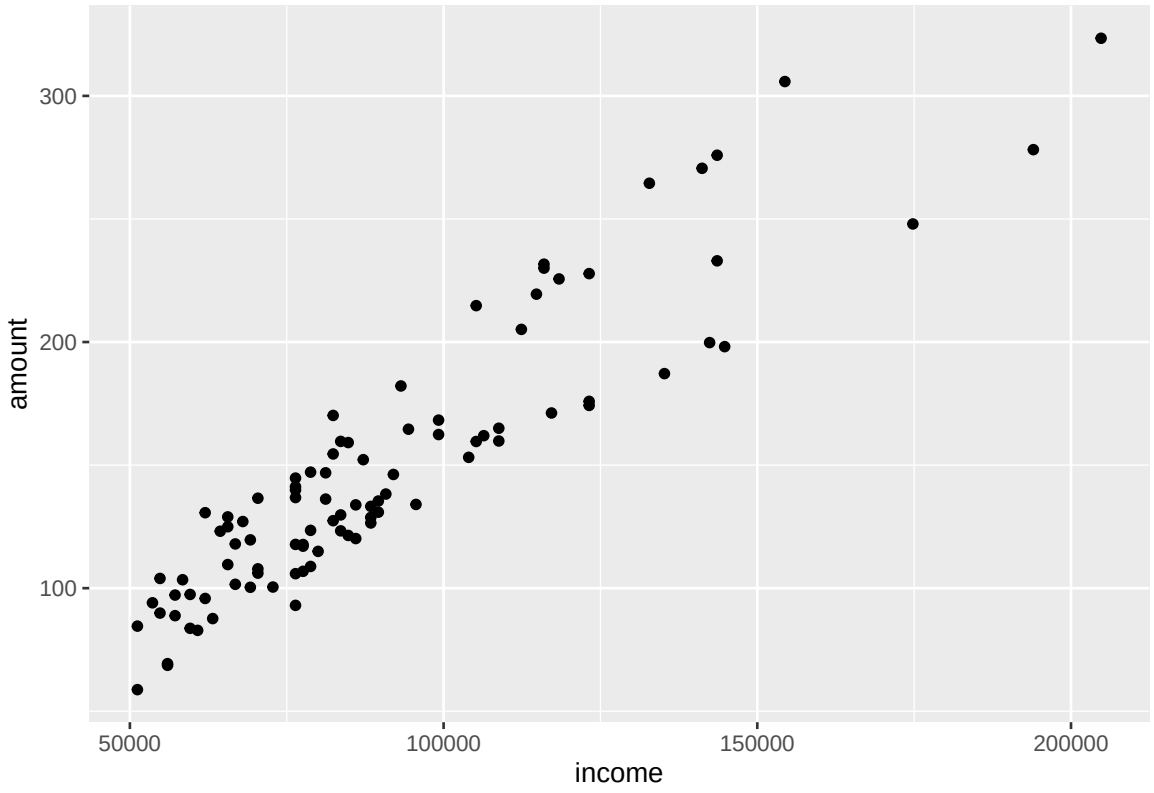
```
rm(list = ls())
df <- read.csv("donors.csv")
r <- cor(df$income, df$amount, method = "pearson")
r
```

```
## [1] 0.9185
```

   Answer: the Pearson correlation between the donation amount and income is 0.9185.

   c. Construct a scatterplot of donation amount over income. Comment on whether linear regression seems appropriate.

```
p <- ggplot(data = df, aes(x = income, y = amount))
p + geom_point()
```

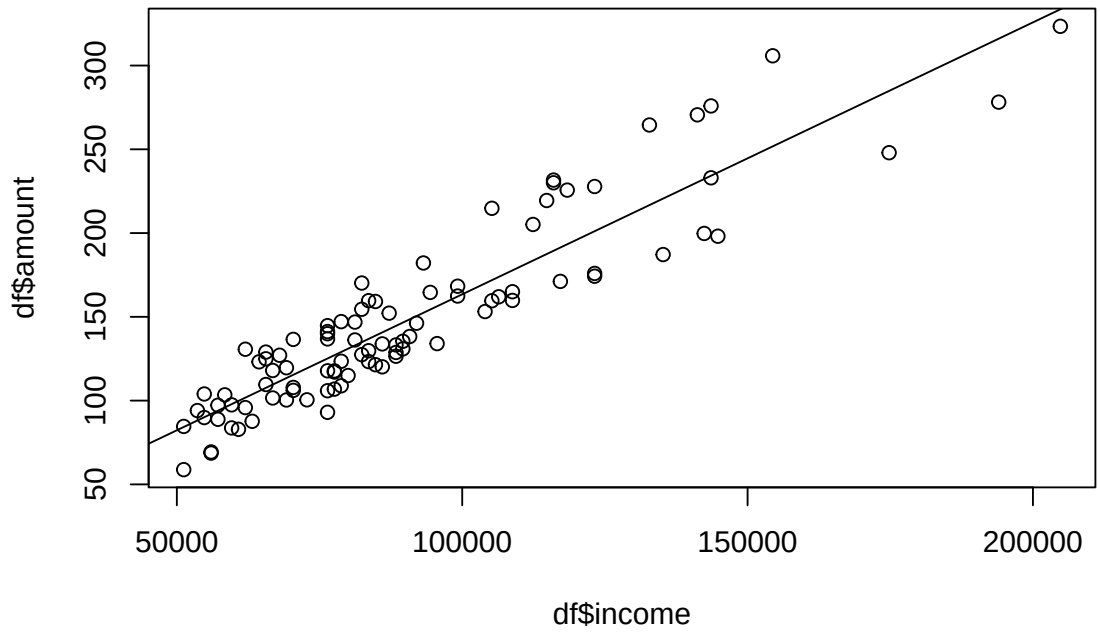


Answer: from the scatter plot, linear regression appears to be very appropriate.

3. Now, let's move onto simple linear regression.

   a. Using a simple linear regression analysis, calculate and report the prediction equation for donation amount as a function of income.

```
model1 <- lm(formula = amount ~ income, data = df)
plot(df$income, df$amount)
abline(model1)
```




```
summary(model1)
```

```
##
## Call:
## lm(formula = amount ~ income, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -38.03 -16.05  -5.74  15.96  54.08
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.129e+00  6.958e+00   0.162    0.871
## income      1.623e-03  7.285e-05  22.280   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.72 on 92 degrees of freedom
## Multiple R-squared:  0.8436, Adjusted R-squared:  0.8419
## F-statistic: 496.4 on 1 and 92 DF,  p-value: < 2.2e-16
```

Answer: the prediction equation for donation amount as a function of income is $amount = 1.623 * 10^{-3} * income + 1.129$

b. What is the estimated mean donation amount for a donor with an income of $92,000?

```
f <- function(x) {
    1.623 * 10^(-3) * x + 1.129
}
y1 <- f(92000)
y1
```

```
## [1] 150.445
```

Answer: the estimated mean donation amount for a donor with an income of $92,000 is 150.445.

c. What is the estimated mean donation amount for a donor with an income of $1,000,000?

```
y2 <- f(1e+06)
y2
```

```
## [1] 1624.129
```

Answer: the estimated mean donation amount for a donor with an income of $1,000,000 is 1624.129.

d. Do you feel equally comfortable making predictions for the previous two questions? Explain.

Answer: I feel more comfortable with making prediction in c. than d., that is because $92,000 is within the range of our data by which we fit the simple linear mode. $1,000,000 is way out of range of our dataset (largest income is around $200,000).

e. Create a 90% confidence interval for average donation amount for a donor with an income of $86,000. Interpret the results. You may find the `predict()` function useful.

```
predict(model1, newdata = data.frame(income = 86000), interval = "confidence",
    level = 0.9)
```

```
##       fit      lwr      upr
## 1 140.721 136.9608 144.4811
```

Answer: the 90% confidence interval for average donation amount for a donor with an income of $86,000 is (136.9608, 144.4811).

f. Create a 90% prediction interval for average donation amount for a donor with an income of $86,000. Interpret the results.

```
predict(model1, newdata = data.frame(income = 86000), interval = "prediction",
    level = 0.9)
```

```
##       fit      lwr      upr
```

```
## 1 140.721 104.4402 177.0018
```

Answer: I am 90% confident that the interval (104.4402, 177.0018) contains the true donation amount for a donor with an income of $86,000.

g. Construct a 95% confidence interval for the slope of the estimated regression equation and interpret the results. You may find the `confint()` function useful.

```
confint(model1, level = 0.95)
```

```
##                      2.5 %       97.5 %
## (Intercept) -12.689492122 14.947354136
## income        0.001478471  0.001767856
```

Answer: I am 95% confident that the interval (0.001478471, 0.001767856) contains the true income slope.

h. Test the hypothesis that $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ at the $\alpha = 0.05$ significance level. What conclusion do you reach in the context of the problem? Report the test statistic. You can refer to the summary of your linear regression model to answer this question.

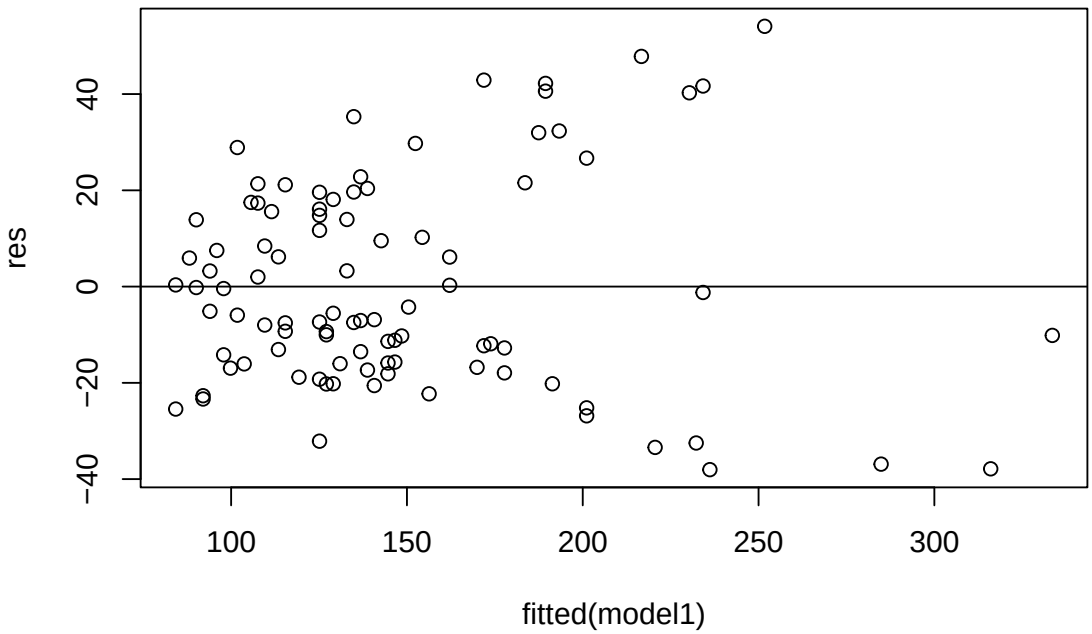
Answer: The test statistic is 22.280. Given $p < 2 * 10^{-6} < \alpha = 0.05$, we reject the null hypothesis and conclude that the β_1 is significantly different from 0.

i. What is the value of the coefficient of determination? Interpret this result in the context of the question.

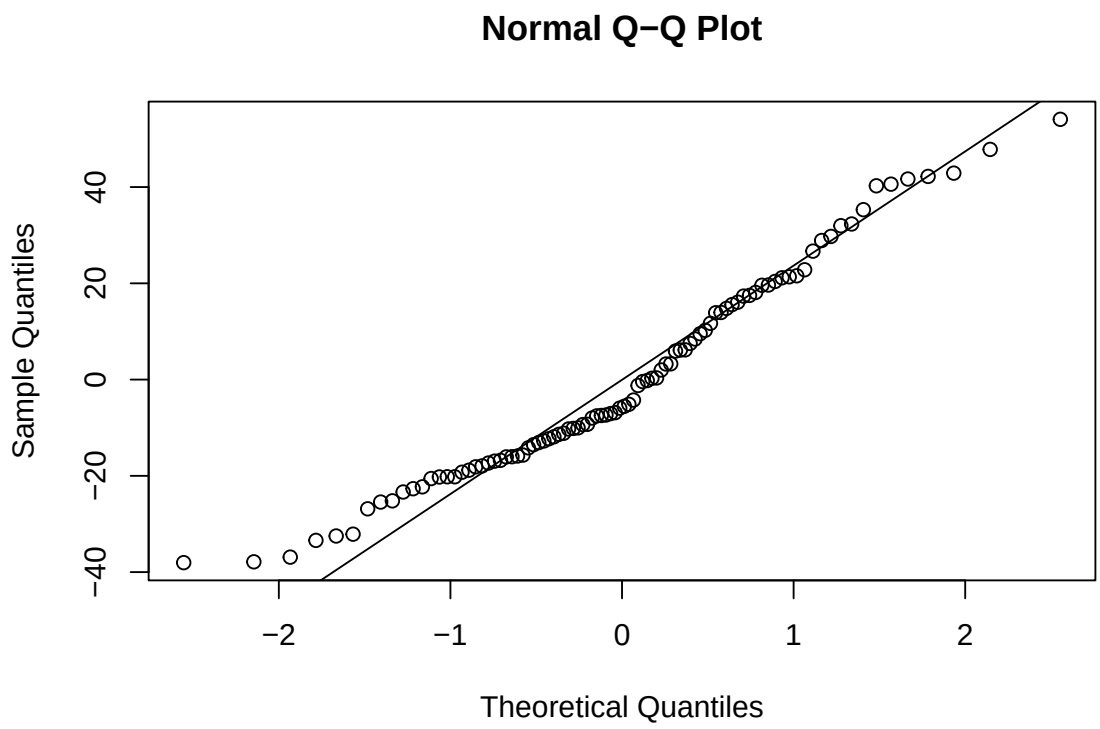
The coefficient of determination is 0.8436. This means 84.36% of donation amount variability can be explained by its linear relationship with the income of donor.

j. Construct diagnostic plots (including a residual plot and a normal qq plot). Comment on the fit of the model with respect to the model assumptions from part a.

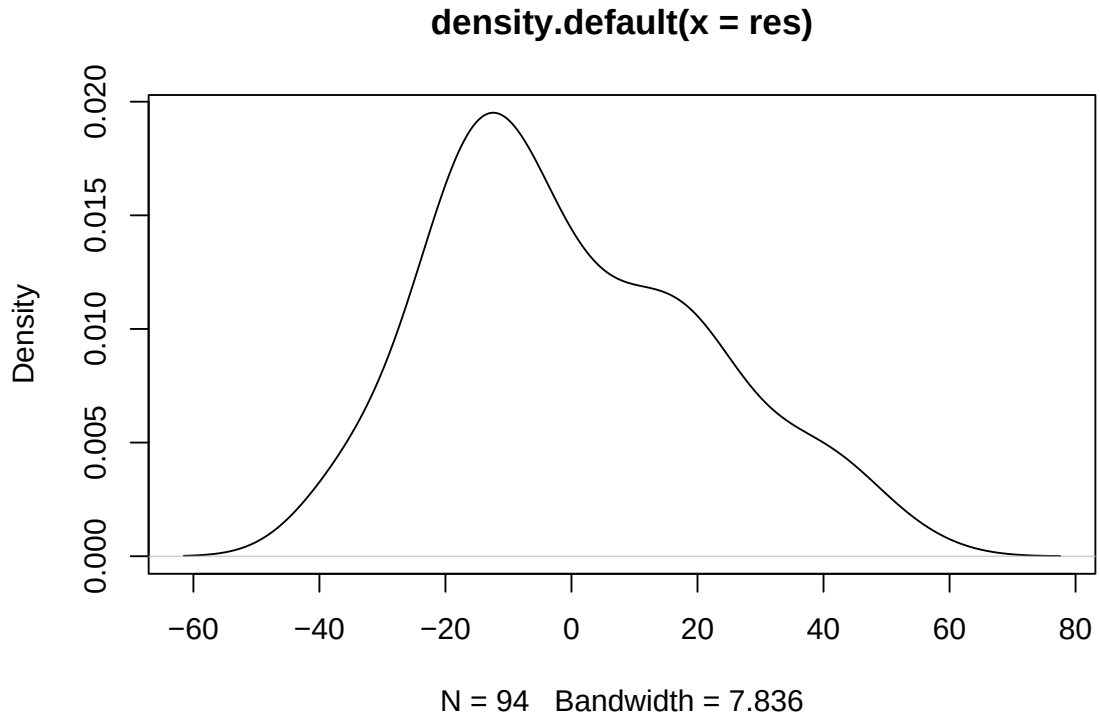
```
res <- resid(model1)
plot(fitted(model1), res)
abline(0, 0)
```




```
qqnorm(res)
qqline(res)
```




```
plot(density(res))
```



Answer: the residual plot appears to show pattern when fitted values are large. On the Q-Q plot, the points tend to stray away from the line near the tails. Those indicates that our linear line may not be too appropriate for modeling the data.

4. Now, let's move onto multiple linear regression.

   a. Calculate the prediction equation for donation amount as a function of income, age, and sex. Report the prediction equation.

```
model2 <- lm(amount ~ income + age + sex, data = df)
summary(model2)
```

```
##
## Call:
## lm(formula = amount ~ income + age + sex, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.316 -14.255  -1.613  13.330  52.723
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.566e+01  1.324e+01   3.449 0.000859 ***
## income       1.636e-03  6.632e-05  24.669  < 2e-16 ***
## age         -1.003e+00  2.336e-01  -4.292 4.45e-05 ***
## sexmale      4.913e+00  4.199e+00   1.170 0.245100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.67 on 90 degrees of freedom
## Multiple R-squared:  0.8745, Adjusted R-squared:  0.8703
## F-statistic: 209.1 on 3 and 90 DF,  p-value: < 2.2e-16
```

Answer: the prediction equation is $amount = 1.636 * 10^{-3} * income - 1.003 * age + 4.913 * sex + 45.66$ (male = 1, female = 0).

   b. Use an F test at the $\alpha = 0.05$ significance level to determine if `age` and/or `sex` significantly add to the model, once we account for `income`. Make sure to state your hypotheses, report the test statistic, and interpret the results in the context of the problem. For this part, you may want to run `anova` with both linear regression models you have previously made; it should look something like `anova(lm1, lm2)`.

```
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: amount ~ income
## Model 2: amount ~ income + age + sex
```

```
##   Res.Df   RSS Df Sum of Sq      F   Pr(>F)
## 1     92 43392
## 2     90 34826  2    8565.5 11.068 5.04e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: Hypotheses: H_0: age and sex do not significantly add to the model, H_1: at least one of those factors (age or sex) adds to the model once we account for income. The test statistics F is 11.068, p value $= 5.04 * 10^{-5} < \alpha = 0.05$. We reject the null hypothesis and conclude that `age` and/or `sex` significantly add to the model once we account for `income`.

Short Answers:

- About how long did this assignment take you? Did you feel it was too long, too short, or reasonable?

  About 2hrs. It is reasonable.

- Who, if anyone, did you work with on this assignment?

  I work with my classmates.

- What questions do you have relating to any of the material we have covered so far in class?

  No.