STAT 1293 - Final Exam

Gordon Lu

7/29/2020

Problem 1: Inference for proportions (50 points)

1a) Are the guidelines for the use of the large-sample confidence interval satisfied? (3 points)

Solution:

Yes. First observing that our data comes from a SRS, we can assume the outcomes are unbiased and random. Additionally, notice how each individual's response can be summarized with a binary outcome. That being, either "Having at least one credit card" or "Having no credit cards". Then, note how the response of one individual does not impact the response of another, thus individuals are independent. Lastly, note that `np` $\geq 10 = 1430(\frac{1087}{1430}) \approx 1087 \geq 10$ and $n(1-p) \geq 10 = 1430(\frac{343}{1430}) \approx 343 \geq 10$. Thus, we can conclude that the sample size if sufficiently large, and importantly we can trust the confidence interval.

1b) Give a 90% confidence interval for the proportion of all college students who have at least one credit card. Make sure you use R. Include your R code and output in your answer. Don't use continuity correction. (6 points)

Solution:

```
prop.test(1087, 1430, conf.level = 0.90, correct = FALSE)$conf.int
```

```
## [1] 0.7410865 0.7782107
## attr(,"conf.level")
## [1] 0.9
```

1c) Does the data provide sufficient evidence that more than 75% of college students have at least one credit card? Use R to find the p-value. Write down the 4 steps of the hypothesis test. Report the chi-square statistic as the test statistics. Let $\alpha = 0.10$ Don't use continuity correction. (14 points)

Solution:

```
prop.test(1087, 1430, 0.75, conf.level = 0.90, alternative = "greater", correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
```

```
## data:  1087 out of 1430, null probability 0.75
## X-squared = 0.78415, df = 1, p-value = 0.1879
## alternative hypothesis: true p is greater than 0.75
## 90 percent confidence interval:
##  0.7453758 1.0000000
## sample estimates:
##         p
## 0.7601399
```

4-Step H.T.

Hypothesis: $H_0 : p = 0.75$

$H_a : p > 0.75$

χ^2 statistic: $z_0 = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = 0.78415$

P-value: $P(Z > z_0) = 0.1879$

Decision: Since p > 0.10, we fail to reject the null hypothesis.

Conclusion: Thus, we do not have sufficient evidence to claim that more than 75% of college studenmts have at least one credit card.

1d) Are the guidelines for the use of the large-sample significance test satisfied? (3 points)

Solution:

Yes. First observing that our data comes from a SRS, we can assume the outcomes are unbiased and random. Additionally, notice how each individual's response can be summarized with a binary outcome. That being, either "Having at least one credit card" or "Having no credit cards". Then, note how the response of one individual does not impact the response of another, thus individuals are independent. Lastly, note that np $\geq 10 = 1430(\frac{1087}{1430}) \approx 1087 \geq 10$ and $n(1-p) \geq 10 = 1430(\frac{343}{1430}) \approx 343 \geq 10$. Thus, we can conclude that the sample size if sufficiently large, and importantly that the sampling distribution of p is normally distributed. Thus, it is appropriate to use a large-sample significance test.

Problem 2: Inference for two proportions (24 points)

2a) Are the guidelines for the use of the large-sample confidence interval satisfied? (4 points)

Solution:

Yes. We have the following populations: Those who report that they stress about their health, which is 358, and those who reported that they did not stress about their health, which is 851. Now, for the 358 students who reported that they stressed about their health, 29.9 said that they were exergamers, meaning the probability of success for the first population is 29.9. It is important to note that whether or not a student is an exergamer does not impact another student's result. Therefore for the first population, the students are independent of one another. Additionally, notice that $0.299 \times 358 \approx 107 \geq 10$, and $(1 - 0.299) \times 358 \approx 251 \geq 10$, therefore the sample sizes for population 1 is sufficiently large, and thus the sampling distribution for population 1 is approximately normally distributed. The same approach can be used for population 2. Note that among the 851 students who reported that they did not stress about their health, 20.8 were exergamers, meaning the probability of success for the second population is 20.8. It is also important to note that whether or not a student is an exergamer does not impact another student's result. Therefore, for the second population, the students are independent of one another as well. Additionally, notice that $0.208 \times 851 \approx 177 \geq 10$, and $(1 - 0.208) \approx 674 \geq 10$, therefore the sample size for population 2 is suffiently large, and importantly we can trust the confidence interval.

2b) Find a 95% confidence interval for the difference in proportions of exergamers. Define the difference be $p_s - p_n$ which means proportion of exergamers among the stressed minus that among the non-stressed. Don't use continuity correction. (6 points)

Solution:

```
prop.test(c(107, 177), c(358, 851), correct = FALSE, conf.level = 0.95)$conf.int
```

```
## [1] 0.03619142 0.14559275
## attr(,"conf.level")
## [1] 0.95
```

2c) Does the data provide sufficient evidence that people who are stressed are more likely to play exergames? Conduct a 4-step hypothesis test at $\alpha = 0.05$. Don't use continuity correction. (14 points)

Solution:

```
prop.test(c(107, 177), c(358, 851), correct = FALSE, conf.level = 0.95, alternative = "greater")
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(107, 177) out of c(358, 851)
## X-squared = 11.583, df = 1, p-value = 0.0003327
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.04498584 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.2988827 0.2079906
```

4-Step H.T.

Hypothesis: $H_0 : p_{stressed-and-exergamer} = p_{not-stressed-and-exergamer}$

$H_a : p_{stressed-and-exergamer} > p_{not-stressed-and-exergamer}$

χ^2 statistic: $z_0 = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}} = 11.583$

P-value: $P(Z > z_0) = 0.0003327$

Decision: Since p < 0.05, we have sufficient evidence to reject the null hypothesis.

Conclusion: Therefore, we have sufficient evidence to conclude that people who are stressed are more likely to play exergames.

Problem 3: Two-way table and chi-squares test (38 points)

3a) Create a two-way table as above. make sure column and row names are defined. (8 points)

Solution:

```
matA <- matrix(c(8,15,13,14,19,15,15,4,7,3,1,4), 4, 3, byrow = T)
colnames(matA) <- c("Psychology", "Biology", "Other")
rownames(matA) <- c("A", "B", "C", "D-F")
matA
```

```
##     Psychology Biology Other
## A            8      15    13
## B           14      19    15
## C           15       4     7
## D-F          3       1     4
```

3b) Is there any relationship between grade and major? Conduct a chi-square test at level 0.05. (30 points)

Part I: R code and output (6 points)

Solution:

```
chisq.test(matA, correct = F)
```

```
## Warning in chisq.test(matA, correct = F): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  matA
## X-squared = 12.183, df = 6, p-value = 0.058
```

Part II: Create a table combing rows "C" and "D-F"

```
matB<- matrix(c(matA["A",] ,matA["B",], matA["C",] + matA["D-F",]), 3, 3, byrow = T)
colnames(matB) <- c("Psychology", "Biology", "Other")
rownames(matB) <- c("A", "B", "C-F")
matB
```

```
##     Psychology Biology Other
## A            8      15    13
## B           14      19    15
## C-F         18       5    11
```

Part III: Conduct a chi-square test based on the new table at 0.05 level. (6 points)

Solution:

```
chisq.test(matB, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  matB
## X-squared = 10.446, df = 4, p-value = 0.03354
```

Part IV: Based on the R output, write down the hypotheses, report the test statistic and the p-value. (6 points)

4-Step H.T.

Hypothesis: $H_0 : P(Major|Grade) = P(Major|Grade^C)$

$H_a : P(Major|Grade) \neq P(Major|Grade^C)$

χ^2 statistic: 10.446

P-value: $P(\chi_2^2 > 10.446) =$ `0.03354`

Part 5: Based on the R output, write down the decision and conclusion. (6 points)

Decision: Since p < 0.05, we have sufficient evidence to reject the null hypothesis.

Conclusion: Therefore, we have sufficient evidence to claim that grade and major are related.

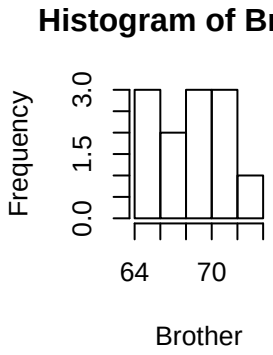
Problem 4: Linear Regression (82 points)

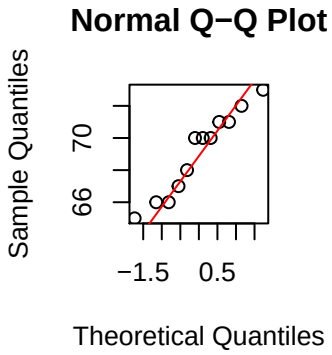
Part I: Data Exploration I (Single Variables) (30 points)

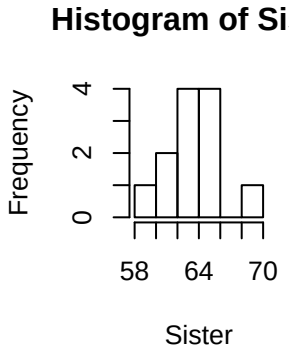
1a) Create histograms and Q-Q plots for both "Brother" and "Sister". Any obvious deviation from normality? Put the 4 figures in one 2-by-2 panel. (10 points)

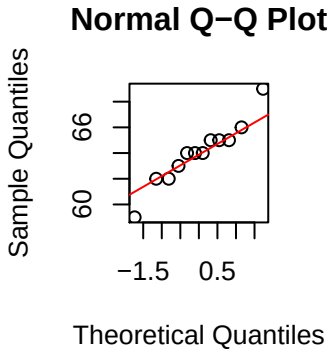
Solution:

```
brosis <- read.table("C:/Users/gordo/Desktop/brosis.txt", header = TRUE) #read in brosis.txt
attach(brosis)

par(mfrow = c(2,2), pty = "s")

hist(Brother)
qqnorm(Brother)
qqline(Brother, col = 2)

hist(Sister)
qqnorm(Sister)
qqline(Sister, col = 2)
```









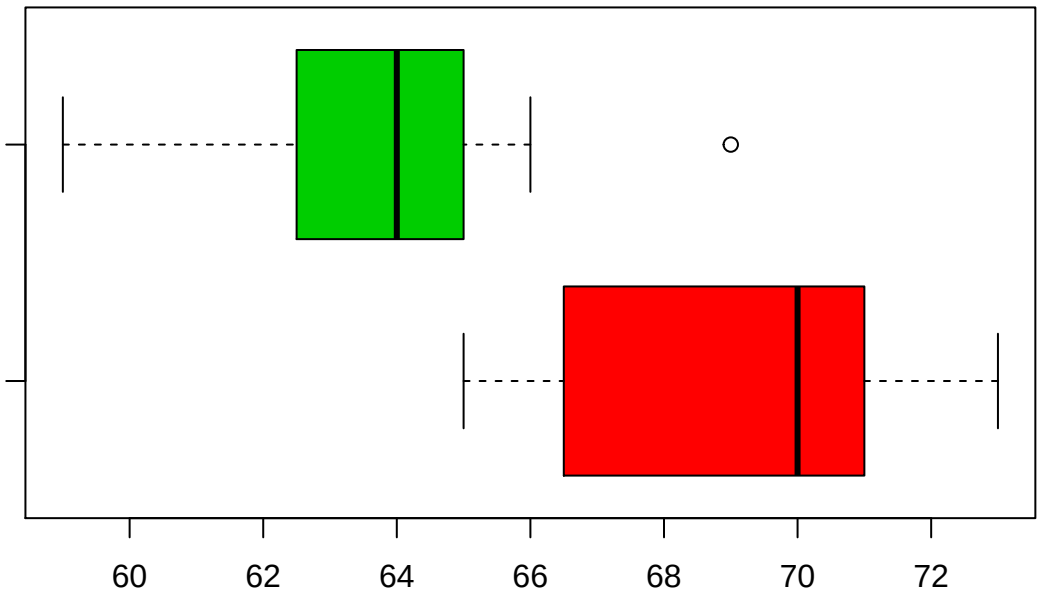
Answer:

No, there does not appear to be any obvious deviation from normality.

1b) Create a side-by-side boxplot. (6 points)

Solution:

```
#Brother is bottom plot, sister is one above it!
boxplot(Brother, Sister, col = c(2,3), outlier = TRUE, horizontal = TRUE)
```



1c) Calculate the five number summary of the two groups. (10 points)

Solution:

```
summary(brosis)
```

```
##     Brother          Sister
##  Min.   :65.00   Min.   :59.00
##  1st Qu.:66.75   1st Qu.:62.75
##  Median :70.00   Median :64.00
##  Mean   :69.08   Mean   :64.00
##  3rd Qu.:71.00   3rd Qu.:65.00
##  Max.   :73.00   Max.   :69.00
```

1d) Calculate the mean and the standard deviation for each group. (4 points)

Solution:

```
mean(Brother)
```

```
## [1] 69.08333
```

```r
sd(Brother)
```

```
## [1] 2.609714
```

```r
mean(Sister)
```

```
## [1] 64
```

```r
sd(Sister)
```

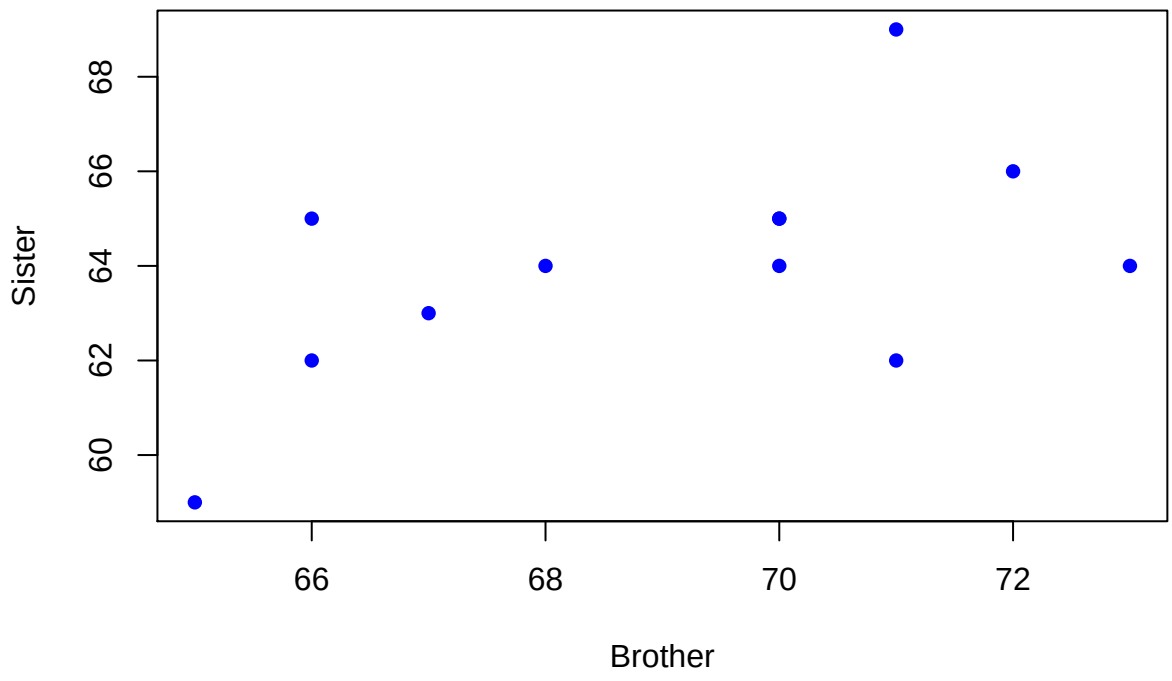
```
## [1] 2.44949
```

Part II: Data Exploration II (Relationship) (16 points)

2a) Create a scatterplot and analyze it in terms of form, direction, and strength. (10 points)

Solution:

```r
plot(Brother, Sister, pch = 16, col = 4)
```



Analysis of the scatterplot:

Form: Appears to be linear.

Direction: Appears to be positive.

Strength: Appears to be moderate.

2b) Calculate the correlation between "Brother" and "Sister". Is the value of correlation r consistent with what you have observed in Part a? (6 points)

Solution:

```
cor(Brother, Sister)
```

```
## [1] 0.5546301
```

Yes, since $r > 0$, the relationship is positive. Also, since $|r|$ is close to 0.5, there is a moderate linear relationship.

Part III: Fit a least-squares regression model. Use "Brother" as the explanatory variable and "Sister" as the response variable. (36 points)

3a) Find the intercept and the slope manually using the formulas of $\hat{\beta}_0$ and $\hat{\beta}_1$ in the lecture slides. (8 points)

Solution:

```
b1 <- cov(Brother, Sister)/var(Brother) #slope
b1
```

```
## [1] 0.5205784
```

```
b0 <- mean(Sister)-b1*mean(Brother) #intercept
b0
```

```
## [1] 28.03671
```

3b) Using R function lm to find the least-squares line. (4 points)

Solution:

```
fit <- lm(Sister ~ Brother)
fit
```

```
##
## Call:
## lm(formula = Sister ~ Brother)
##
## Coefficients:
## (Intercept)      Brother
##     28.0367       0.5206
```

3c) Write down the regression equation explicitly. (4 points)

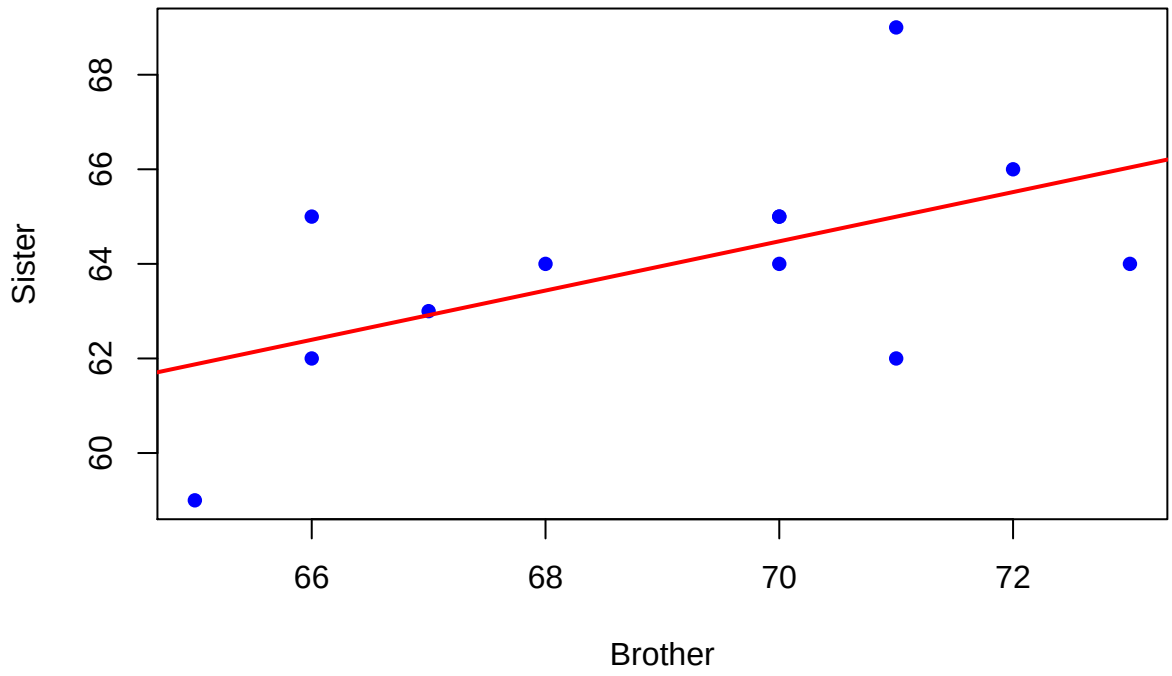
Solution:

The regression equation is: $\hat{Sister}_i = 28.0367 + 0.5206(Brother_i)$

3d) Superimpose the regression line on the scatterplot. (4 points)

Solution:

```
plot(Brother, Sister, pch = 16, col = 4)
abline(fit, col = 2, lwd = 2)
```



3e) Interpret the slope b_1 in context. (4 points)

Solution:

For every additional unit of height for a Brother, the Sister's height will increase by 0.5206 inches.

3f) Suppose a brother's height is 69 inches. What is the predicted height of his sister's height? (4 points)

Solution:

```
b_val <- 69
predicted <- predict(fit, data.frame(Brother = b_val))
predicted
```

```
##        1
## 63.95662
```

The predicted height of his sister's height is `63.95662` inches.

3g) Suppose the actual height of the sister of the 69-inch brother is 60 inches. What is the prediction error (residual)? (4 points)

Solution:

```
actual <- 60

residual <- actual - predicted

residual
```

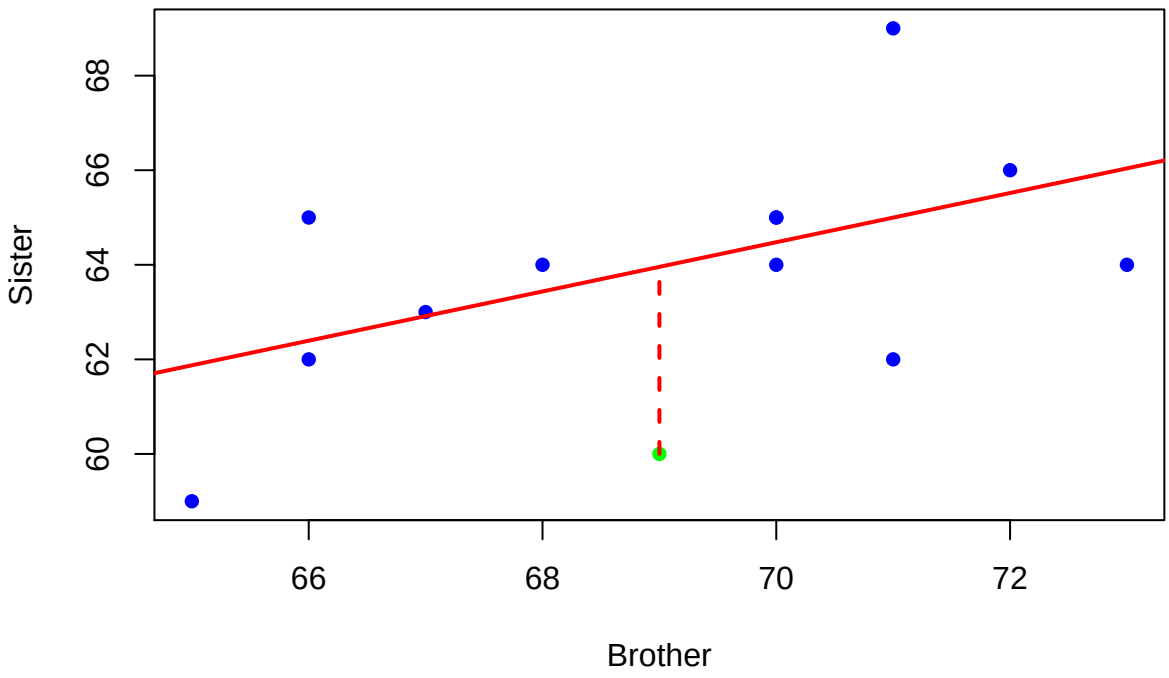
```
##         1
## -3.956618
```

The residual is: `-3.956618`.

3h) Suppose the actual height of the sister of the 69-inch brother is 60 inches. What is the prediction error (residual)? (4 points)

Solution:

```
plot(Brother, Sister, pch = 16, col = 4)
abline(fit, col = 2, lwd = 2)
points(x = b_val, y = actual, col = "green", pch = 16)
segments(b_val, actual, b_val, predicted, col = 2, lwd = 2, lty = 2)
```



Problem 5: Inference for Regression (80 points)

5a) Based on the least-squares regression model in Problem 3, estimate the standard deviation term σ of Y (or ϵ) in the model. (10 points)

Part I: Calculate $s(\sqrt{MSE})$ manually first. (6 points)

Solution:

```
SSE <- sum(resid(fit)^2)
n <- dim(brosis)[1]
MSE <- SSE/(n-2)

s <- sqrt(MSE)
s
```

```
## [1] 2.137696
```

The standard deviation term of the model is: `2.137696`.

Part II: Use R function `summary()` to extract the details of the regression model. Report the estimate of σ from the output. (4 points)

Solution:

```
summary(fit)
```

```
##
## Call:
## lm(formula = Sister ~ Brother)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9978 -0.8676  0.2831  0.5331  4.0022
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.0367    17.0732   1.642   0.1316
## Brother       0.5206     0.2470   2.108   0.0613 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.138 on 10 degrees of freedom
## Multiple R-squared:  0.3076, Adjusted R-squared:  0.2384
## F-statistic: 4.443 on 1 and 10 DF,  p-value: 0.06127
```

The residual standard error is: `2.138`.

5b) Calculate a 90% confidence interval for the slope β_1. (10 points)

Part I: Manually calculate the lower and upper bound. (6 points)

Solution:

```
t <- qt(.95, n-2)
b1 <- summary(fit)$coefficients[2,1]
se_b1 <- summary(fit)$coefficients[2,2]
LL <- b1 - t*se_b1
UL <- b1 + t*se_b1
c(LL, UL)
```

```
## [1] 0.07294199 0.96821485
```

The confidence interval is: `(0.07294199, 0.96821485)`.

Part II: Calculate a 90% confidence interval for the slope β_1 using the function confint. (4 points)

Solution:

```
confint(fit, "Brother", level = 0.90)
```

```
##                5 %      95 %
## Brother 0.07294199 0.9682148
```

The confidence interval is: `(0.07294199, 0.9682148)`.

5c) Conduct a hypothesis test about the linear relationship between "Brother" and "Sister" at level 0.10. Based on the R output, write down the four steps of it. (8 points)

Solution:

```
summary(fit)
```

```
##
## Call:
## lm(formula = Sister ~ Brother)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9978 -0.8676  0.2831  0.5331  4.0022
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.0367    17.0732   1.642   0.1316
## Brother       0.5206     0.2470   2.108   0.0613 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.138 on 10 degrees of freedom
## Multiple R-squared:  0.3076, Adjusted R-squared:  0.2384
## F-statistic: 4.443 on 1 and 10 DF,  p-value: 0.06127
```

4-Step H.T.

Hypothesis: $H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

t statistic: $t_0 = \frac{b_1}{se(b_1)} = 2.108$

P-value: $2P(t_{n-2} < |t_0|) =$ `0.0613`.

Decision: Since p < 0.10, we have sufficient evidence to reject the null hypothesis.

Conclusion: Therefore, we have sufficient evidence to claim that there is a linear relationship between the heights of Brothers and Sisters.

5d) Given a new observation of brother's height, 69 inches, what is the mean value of the sisters' height? Calculate a 95% confidence interval for the mean response using R function. (6 points)

Solution:

```
new_brother <- data.frame(Brother = 69)
predict(fit, new_brother, level = 0.95, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 63.95662 62.58087 65.33237
```

The mean of the sisters' height is: `63.95662`, and the confidence interval is: `(62.58087, 65.33237)`.

5e) Given a new observation of brother's height, 69 inches, what is the predicted value of a single observation of sister's height? Calculate a 95% prediction interval for the single observation using R function. (6 points)

Solution:

```
predict(fit, new_brother, level = 0.95, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 63.95662 58.99883 68.91441
```

The predicted value of a single observation of sister's height given that the new observation of brother's height is 69 inches is: `63.95662`.

5f) Conduct a utility test on the model at 0.10 level. (12 points)

Part I: Pull out the ANOVA table using the anova() function. (2 points)

Solution:

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: Sister
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Brother    1 20.303 20.3026  4.4428 0.06127 .
## Residuals 10 45.697  4.5697
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Part II: Based on the ANOVA table, calculate the F statistic manually using the following formula. Show your code and output. (6 points)

Solution:

```
f_stat <- (20.303/1)/(45.697/10)
f_stat
```

```
## [1] 4.442961
```

Part III: Calculate the p-value of the F statistic. (4 points)

Solution:

```r
p_val <- 1 - pf(f_stat, 1, 10)
p_val
```

```
## [1] 0.06126503
```

5g) Coefficients of determination R^2 (12 points)

Part I: Calculate the coefficients of determination R^2 manually based on the ANOVA table. (4 points)

Solution:

```r
SSR <- 20.303
SSE <- 45.697
SST <- SSR + SSE
R_sq <- SSR/SST
R_sq
```

```
## [1] 0.3076212
```

The coefficient of determination of the model is: `0.3076212`.

Part II: Verify that R^2 is equal to the square of the correlation coefficient, r. (4 points)

Solution:

```r
r <- cor(Brother, Sister)
r^2; R_sq
```

```
## [1] 0.3076145
```

```
## [1] 0.3076212
```

Although there may be a slight deviation in values, this is likely due to a rounding error. Thus, we can say that the R^2 value is equal to the square of the correlation coefficient, `r`.

Part III: Carefully interpret the value of R^2 in context. (4 points)

Solution:

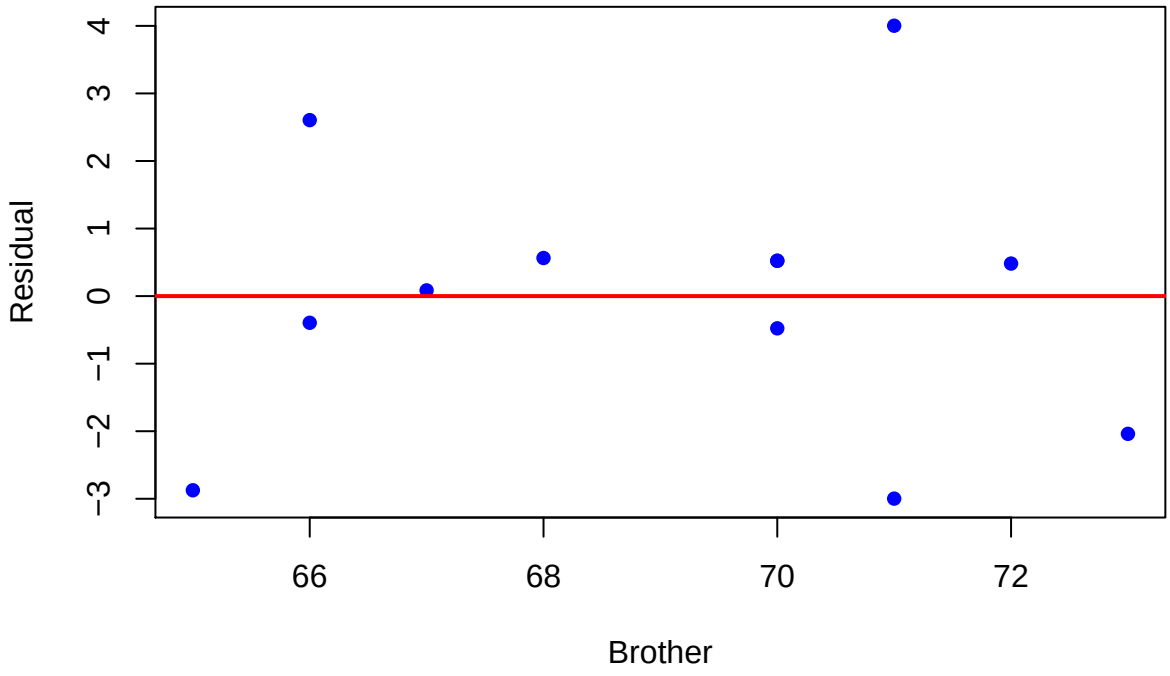
Only `0.3076212` of the variation in Sister's heights are explained by Brother's heights.

5h) Residuals (16 points)

Part I: Create a residual plot for the model. Add a red reference line (y=0) to the plot. (6 points)

Solution:

```r
plot(Brother, resid(fit), pch = 16, col = 4, ylab = "Residual")
abline(h = 0, lwd = 2, col = 2)
```



Part II: Any concern about the residual plot? (2 points)

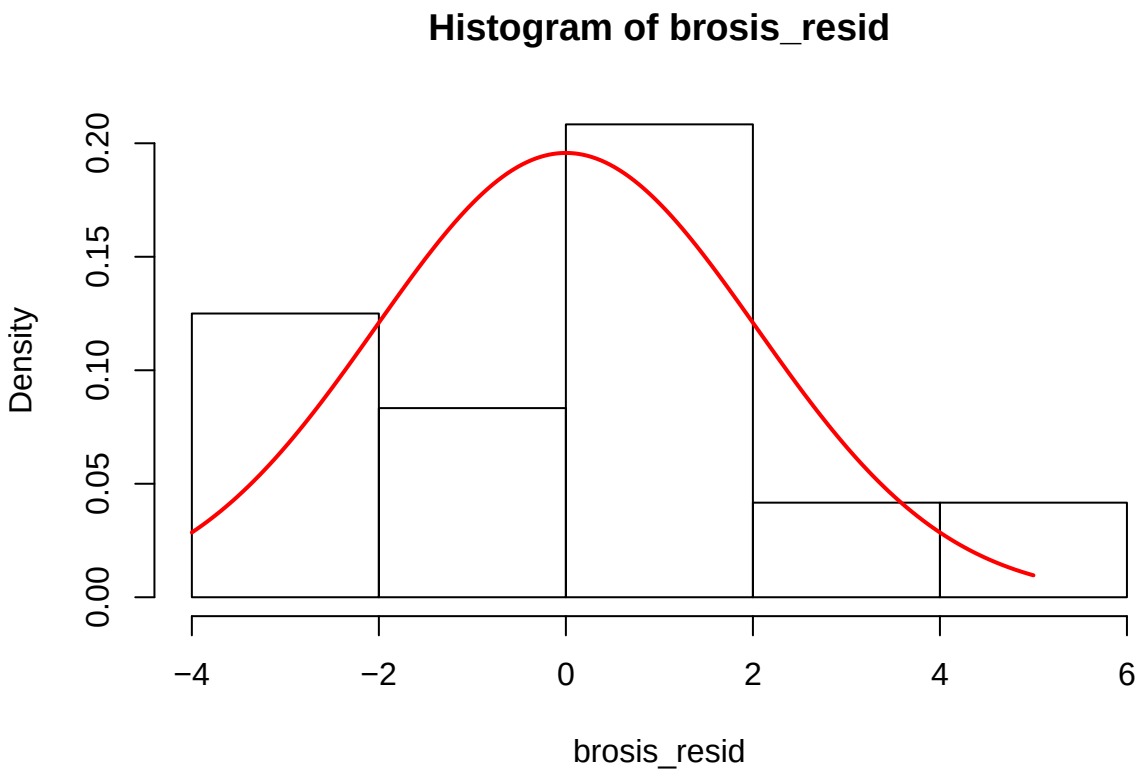
Solution:

No, there do not appear to be any concerns with the residual plot. It looks like an unstructured horizontal band, and the points are roughly symmetric about 0. There's also no curve pattern observed.

Part III: Create a histogram for the residual plot. Overlay a normal density curve. Any severe deviation from normality? (4 points)

Solution:

```r
# par(mfrow = c(1,2), pty = "s")
brosis_resid <- resid(fit)
hist(brosis_resid, freq=F)
lines(seq(-4, 5, 0.01), dnorm(seq(-4,5,0.01), mean(brosis_resid), sd(brosis_resid)), col = 2, lwd = 2)
```



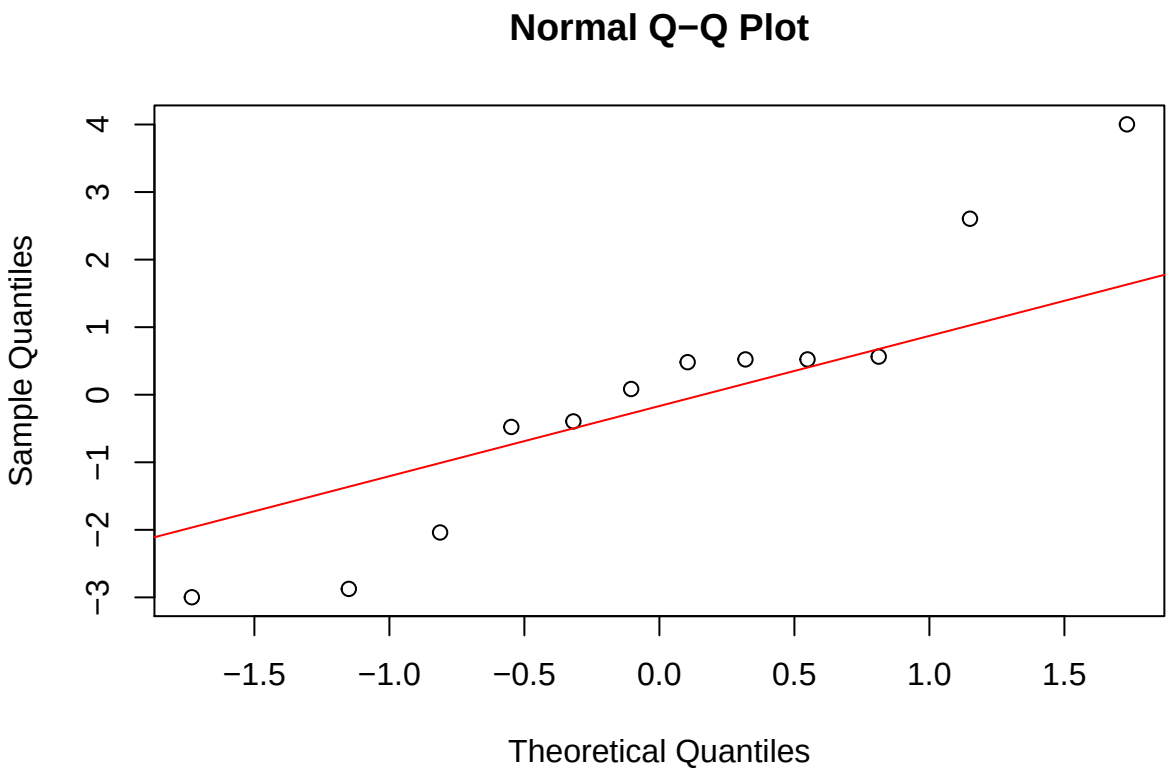

```
# qqnorm(brosis_resid)
# qqline(brosis_resid, col = 2)
```

Yes, it does not appear to look that normal.

Part IV: Create a Q-Q plot for the residuals. Any severe deviation from normality? (4 points)

Solution:

```
qqnorm(brosis_resid)
qqline(brosis_resid, col = 2)
```



Yes, there are a fair bit of points that the reference lines does not capture, indicating deviations from normality.