STA 2210 Final Review Exercises

The following data represents the results of a survey of general health for a random sample of 20000 American adults (we used this data during Lab 1).

Data: source("http://www.openintro.org/stat/data/cdc.R")

Directions: Use statistical inference to investigate the 3 research questions below. Include all of the following in your presentation.

1. Formally check if the conditions for inference are satisfied.
2. Perform a hypothesis test using a 0.0001 level of significance.
3. Interpret the decision of the test in the context. In your interpretation, also describe the type of error that may have been made, and what that error means in the context.

Research Questions:

**Q1**. Is there a relationship between gender and perception of one’s general health?

Chi Squared of Independence/ Association

table(cdc$gender, cdc$genhlth)

excellent very good good fair poor

m 2298 3382 2722 884 283

f 2359 3590 2953 1135 394

a) Conditions: 1) Independence: Sample size < 10% of the population and the sample is randomly generated, so independence is satisfied. 2) The sample size of each cell’s expected value is greater than or equal to 5.

chisq <- chisq.test(table(cdc$gender, cdc$genhlth))

> chisq$expected

excellent very good good fair poor

m 2228.142 3335.753 2715.204 965.9905 323.9106

f 2428.858 3636.247 2959.796 1053.0095 353.0894

Both conditions are checked off.

b) Hypotheses:

Ho: There is no relation between gender and perception of one’s general health.

Ha: There is relation between gender and perception of one’s general health.

chisq

Pearson's Chi-squared test

data: table(cdc$gender, cdc$genhlth)

X-squared = 28.712, df = 4, p-value = 8.945e-06

P value < 0.0001, so we reject the null hypothesis.

c) Since we reject the null hypothesis, we can conclude with the alternate hypothesis and say that there is a relationship between gender and one’s perceived general health. A Type 1 error is possible if we reject the null hypothesis, and it may have been true.

**Q2**. Do heavier people tend to have a larger desired weight?

Linear Regression

m1 <- lm(wtdesire ~ weight, data = cdc)

>

> summary(m1)

Call:

lm(formula = wtdesire ~ weight, data = cdc)

Residuals:

Min 1Q Median 3Q Max

-167.98 -9.32 0.08 11.51 518.31

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 46.664015 0.590782 78.99 <2e-16 \*\*\*

weight 0.639014 0.003388 188.59 <2e-16 \*\*\*

---

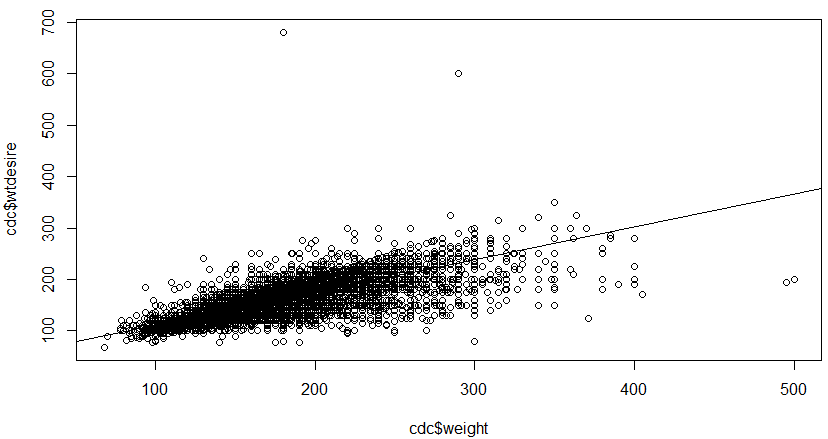
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.21 on 19998 degrees of freedom

Multiple R-squared: 0.6401, Adjusted R-squared: 0.6401

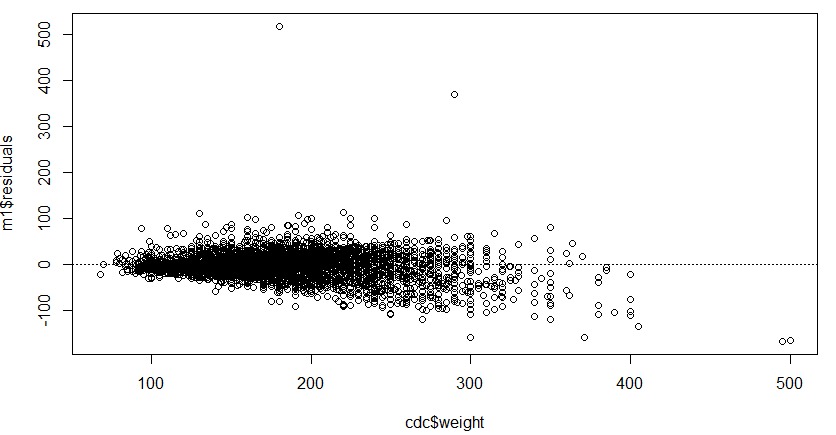
F-statistic: 3.556e+04 on 1 and 19998 DF, p-value: < 2.2e-16

plot(cdc$weight, cdc$wtdesire)

abline(m1)

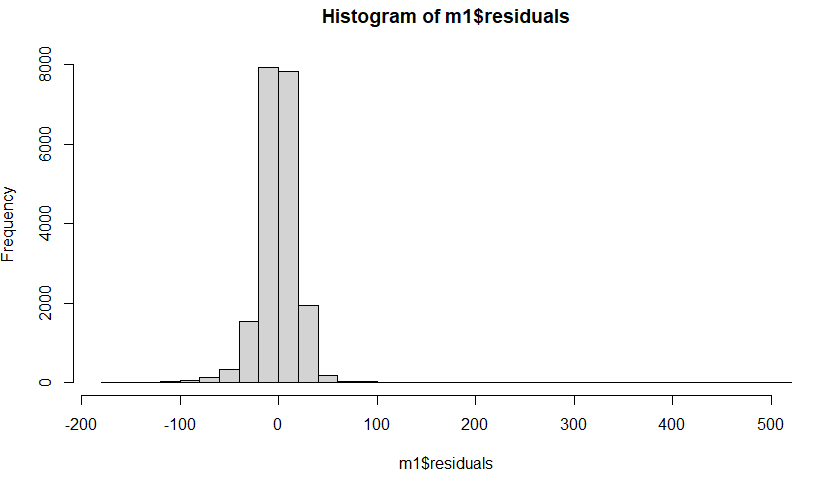
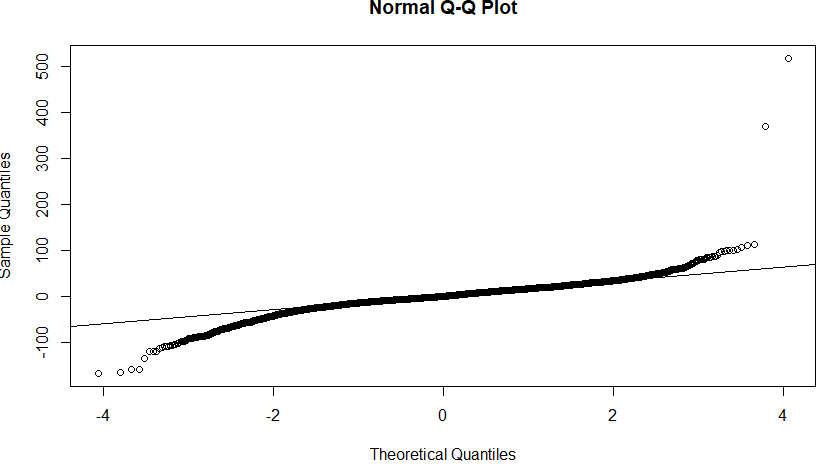
plot(m1$residuals ~ cdc$weight)

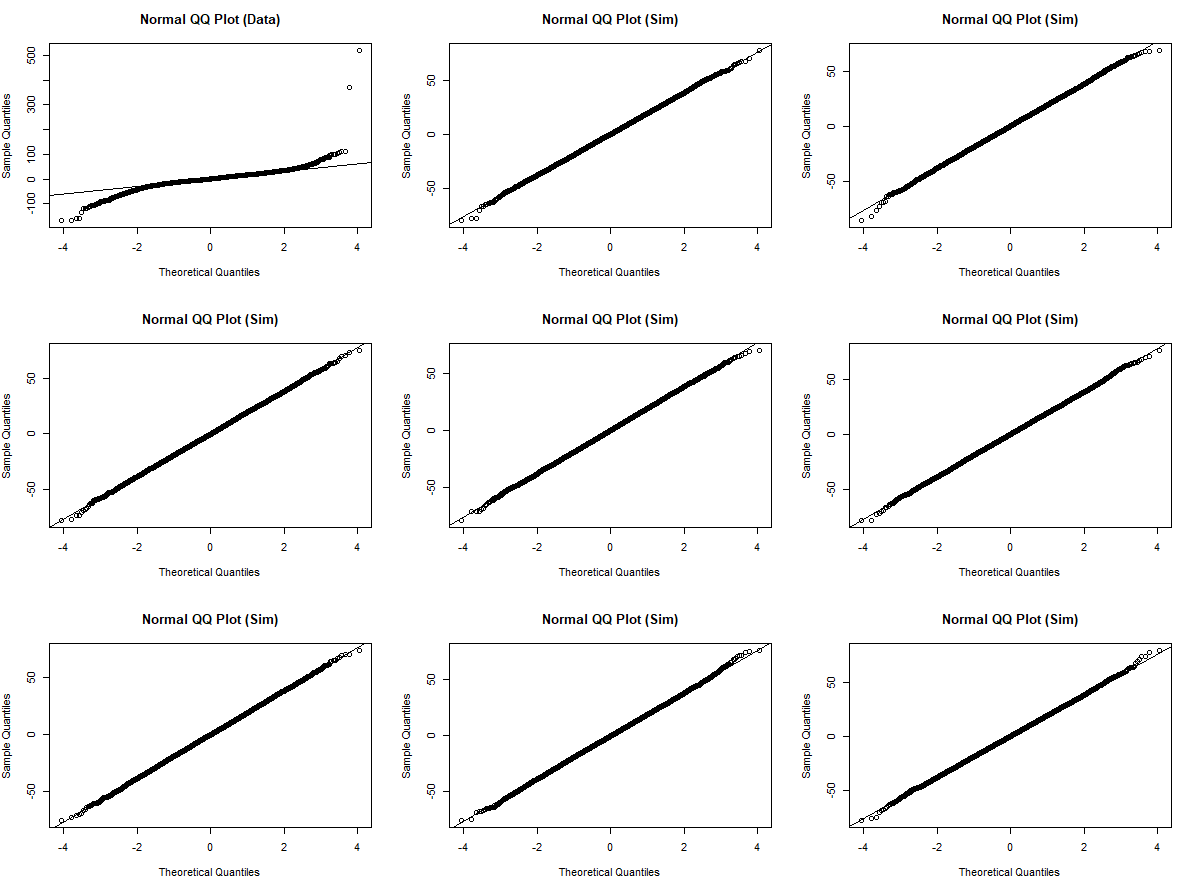
> abline(h = 0, lty = 3)



hist(m1$residuals, breaks = 30)

> qqnorm(m1$residuals)

> qqline(m1$residuals)

Qqnormsim(m1$residuals)

a) Conditions:

1) Linearity: There appears to be a slightly linear pattern for the residuals plot at y=0 for the relationship between weight and desired weight, so I would say that linearity is not satisfied.

2) Nearly Normal Residuals: Looking at the histogram and qq plot of the data, the histogram appears properly normally distributed with a center around 0 and solid symmetry. The qq plot, though, appears normal in the center, but the tails veer off the line a little, indicating a little left skew, and a slight right skew. These could be attributed to bad data and outliers, though, and the simulated qq plots also give these skews at the tails. Since the histogram is normal and the qq plot is nearly normal with small left skew, we can say that nearly normal residuals is checked off, but we exercise caution with the test.

3) Constant Variability: The amount of overestimate and underestimate needs to be nearly-evenly spread out above and under the best fit line for the scatterplot, which it appears to do until the outliers are met. The average amount of error for the residuals appears similar for both sides until we get into the higher weights, so constant variability appears to be met.

At least one condition is not met, so we can say that a linear model would not be appropriate for this relationship. However, I will move on with the linear test and assume the conditions are true.

b) Hypotheses:

Ho: Heavier people do not tend to have a larger desired weight (Beta = 0).

Ha: Heavier people do tend to have a larger desired weight (Beta =/= 0).

m1 <- lm(wtdesire ~ weight, data = cdc)

>

> summary(m1)

Call:

lm(formula = wtdesire ~ weight, data = cdc)

Residuals:

Min 1Q Median 3Q Max

-167.98 -9.32 0.08 11.51 518.31

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 46.664015 0.590782 78.99 <2e-16 \*\*\*

weight 0.639014 0.003388 188.59 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.21 on 19998 degrees of freedom

Multiple R-squared: 0.6401, Adjusted R-squared: 0.6401

F-statistic: 3.556e+04 on 1 and 19998 DF, p-value: < 2.2e-16

Equation: 46.66 + 0.639(weight)

R-squared: 64.01

P value = < 2.2e-16

Since the p value is less than 0.0001, we can reject the null hypothesis.

c) If the conditions were met, we would reject the null hypothesis and conclude that heavier people tend to desire a larger weight. We may have committed a Type 1 error, where we rejected the null hypothesis when it may have been true. In this case, the error could be that we said heavier people tend to desire a larger weight, when the case is that heavier people do not desire a larger weight.

Since the conditions were not all met, this conclusion does not reflect the true relationship of the variables we are looking at, weight and desired weight. A linear model was not appropriate to do a hypothesis test on this data, so this conclusion is not trustworthy to see if heavier weight is an indicator of larger desired weight.

**Q3**. Is there a difference in the average weight of those who exercise and those who do not exercise?

Two-sided T Test

by(cdc$weight, cdc$exerany, length)

cdc$exerany: 0 = 5086

------------------------------------------------------------------------------------------

cdc$exerany: 1 = 14914

by(cdc$weight, cdc$exerany, summary)

cdc$exerany: 0

Min. 1st Qu. Median Mean 3rd Qu. Max.

78.0 140.0 165.0 171.6 195.0 400.0

------------------------------------------------------------------------------------------

cdc$exerany: 1

Min. 1st Qu. Median Mean 3rd Qu. Max.

68 140 165 169 190 500

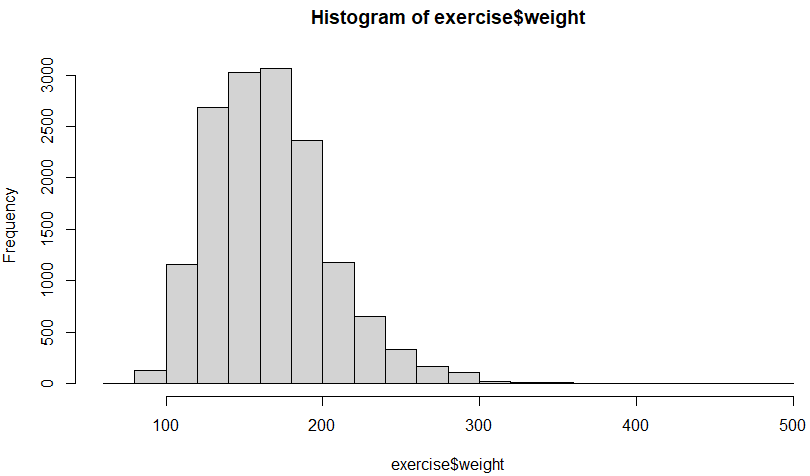
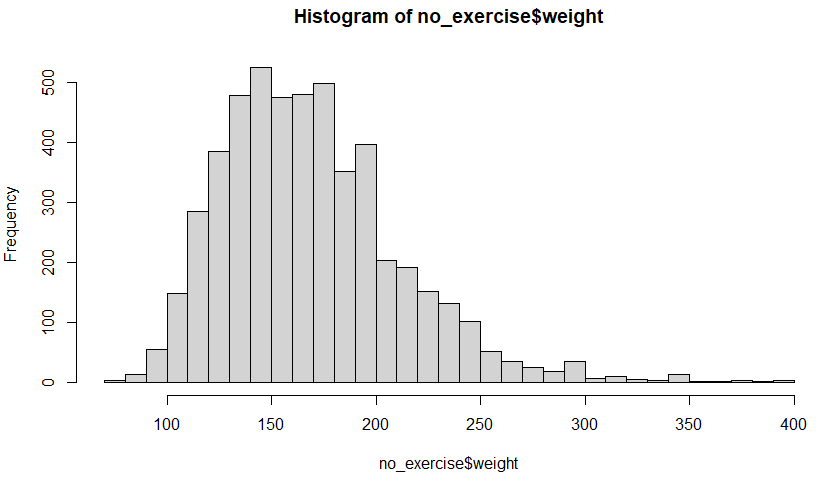
a) Conditions:

1) Independence: For independence, we need to see that the sample sizes are less than 10% of the population and randomly sampled. We see that there are 20,000 cases, which is less than 10% of the population and they are randomly sampled. These cases all check off, and independence is met.

2) Not Paired: We see that the two groups are not paired, because those that exercise are not related to those that do not exercise, so this case is checked.

3) Normality:

hist(exercise$weight, breaks = 25)

hist(no\_exercise$weight, breaks = 25)

Looking at both histograms, we need to see if they are both nearly normally distributed. Both graphs do have nearly normal distributions, with a slight left skew for both. Based on this, we can say normality is met.

Both conditions are met, so we can move forward with the hypothesis test.

b) Hypotheses:

Ho: There is no difference in average weight for those who exercise and those who do not.

(mu\_exercise = mu\_noexercise)

Ha: There is a difference in average weight for those who exercise and those who do not.

(mu\_exercise =/= mu\_noexercise)

inference(y = cdc$weight, x = cdc$exerany, est = "mean", type = "ht", null = 0, alternative = "twosided", method = "theoretical")

Response variable: numerical, Explanatory variable: categorical

Difference between two means

Summary statistics:

n\_0 = 5086, mean\_0 = 171.5722, sd\_0 = 43.4829

n\_1 = 14914, mean\_1 = 169.0387, sd\_1 = 38.8333

Observed difference between means (0-1) = 2.5335

H0: mu\_0 - mu\_1 = 0

HA: mu\_0 - mu\_1 != 0

Standard error = 0.688

Test statistic: Z = 3.684

p-value = 2e-04

**(CAN ALSO USE t.test(cdc$weight ~ cdc$exerany) WHICH GAVES THE SAME ANSWER).**

The p value is .0002, which is greater than the alpha value of 0.0001. In this case, we fail to reject the null hypothesis.

c) We fail to reject the null hypothesis, so we conclude that there is no difference in average weight between those who exercise and those who do not exercise. We may have made a Type 2 error, where we failed to reject the null hypothesis when we should have rejected it. This means we said there was no difference in average weight between those who exercise and those who do not, when there is a difference.