Simple Linear Regression - Solutions

Chapter 6, Lab 1

OpenIntro Biostatistics

Topics

- Examine visual relationships between numerical variables
- Estimate a regression line using least squares
- Interpret a least squares regression line
- Evaluate model fit
- Perform statistical inference with regression
- Prediction with the least squares regression line
- Check the necessary conditions for regression

This lab introduces simple linear regression.

The material in this lab corresponds to Sections 6.1 - 6.4 of *OpenIntro Biostatistics*.

Background

Triglycerides levels provide strong indication of heart health. Triglycerides are a type of lipid stored in your blood derived from unused calories. When your body needs energy between meals hormones release triglycerides. In some cases, people who eat excess high carbohydrate foods (more than the body requires) have high triglyceride levels. This condition is known as hypertriglyceridemia.

A lipid panel or blood test for cholesterol can easily determine a person's triglyceride level. The following thresholds show generally accepted standards for certain triglyceride levels (mg/dL).

- Normal: < 150 mg/dL
- Moderate: 150 to 199 mg/dL
- High: 200 to 499 mg/dL
- Very High: > 500 mg/dL

A patient with high and very high triglyceride levels are at extreme risk for arteriosclerosis (hardening of the arteries or thickening of the artery walls). This condition increases the risk of stroke and heart and vascular failure. In addtion to heart disease, high triglycerides contribute to obesity, pancreatitis, hypothyroidism, and diabetes.

Prevention or at least mitigation of hypertriglyceridemia can be handled with multiple methods. Maintaining a health lifestyle, including regular exercise, proper diet, and limited alcohol consumption helps ensure triglyceride levels are managed appropriately. In some cases living a healthy lifestyle is not enough and medication, such as Lipitor or Crestor (statins) must also be taken to manage triglyceride levels.

The tri dataset provides a list of de-identified patients who visited their primary care physician in the past month. For each patient, data elements including their triglyceride level, age, waist circumference measurement, gender, and patient disease indicators were recorded.

The tri dataset can be found in the oibiostat package. For more information, use the help(tri) command.

```
library(oibiostat)
data("tri")
```

Exploratory Data Analysis

1. Explore the tri dataset. What is the dimensions, data types, etc.?

```
str(tri)
```

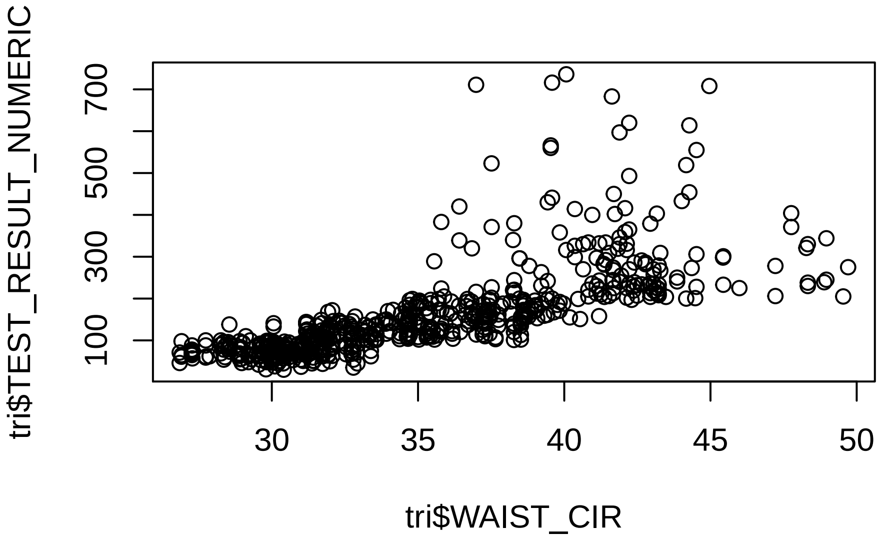
```
## Classes 'tbl_df', 'tbl' and 'data.frame':    592 obs. of  18 variables:
##  $ MEDICAL_RECORD_NUMBER      : chr  "45644" "108976" "493556" "857332" ...
##  $ TEST_CODE_NAME             : chr  "LAB134 - TRIGLYCERIDES" "LAB134 - TRIGLYCERIDES" "LAB134 - TRIGL
##  $ TEST_RESULT_NUMERIC        : num  73 274 199 450 141 142 208 331 144 298 ...
##  $ SEX_NAME                   : Factor w/ 2 levels "FEMALE","MALE": 2 1 2 1 1 2 2 2 1 2 ...
##  $ SMOKING_STATUS             : Factor w/ 5 levels "CURRENT EVERY DAY SMOKER",..: 5 5 5 4 4 5 4 5 5 5 .
##  $ ASTHMA_IND                 : Factor w/ 3 levels "1","2","NULL": 3 3 3 3 3 3 3 3 1 3 ...
##  $ COPD_IND                   : Factor w/ 3 levels "1","2","NULL": 3 3 3 3 3 3 3 3 3 3 ...
##  $ CHRONIC_CARE_IND           : Factor w/ 3 levels "1","2","NULL": 1 1 2 1 1 1 3 3 1 2 ...
##  $ CHRONIC_KIDNEY_DISEASE_IND : Factor w/ 3 levels "1","2","NULL": 1 1 3 3 3 3 3 3 3 3 ...
##  $ CORONARY_ARTERY_DISEASE_IND: Factor w/ 3 levels "1","2","NULL": 1 3 2 1 3 3 3 1 3 1 ...
##  $ DEPRESSION_IND             : Factor w/ 3 levels "1","2","NULL": 1 2 2 1 3 3 3 3 3 1 ...
##  $ DIABETES_IND               : Factor w/ 3 levels "1","2","NULL": 1 3 1 1 1 3 3 1 1 1 ...
##  $ HEART_FAILURE_IND          : Factor w/ 3 levels "1","2","NULL": 3 3 1 3 3 3 3 3 3 3 ...
##  $ HIV_IND                    : Factor w/ 2 levels "1","NULL": 2 2 2 2 2 2 2 2 2 2 ...
##  $ HYPERTENSION_IND           : Factor w/ 3 levels "1","2","NULL": 1 1 1 1 1 1 3 1 1 1 ...
##  $ OBESITY_IND                : Factor w/ 3 levels "1","2","NULL": 3 3 1 3 1 3 3 1 3 3 ...
##  $ WAIST_CIR                  : num  30.7 41.7 36.7 41.7 31.7 ...
##  $ AGE                        : num  71.9 71 64.3 65.7 71.9 ...
```

```
dim(tri)
```

```
## [1] 592  18
```

2. Visually inspect the relationship between a patient's waist circumference measurement, WAIST_CIR and their associated triglyceride level, TEST_RESULT_NUMERIC. What type of relationship exists?

```r
plot(tri$WAIST_CIR, tri$TEST_RESULT_NUMERIC)
```



A positive and roughly linear relationship exists between triglyceride level and waist circumference measurement

3. What is the strength of the linear relationship between a patient's waist circumference measurement, `WAIST_CIR` and their associated triglyceride level, `TEST_RESULT_NUMERIC`. Does this indicate larger waist circumference measurements cause higher triglyceride levels? Explain.

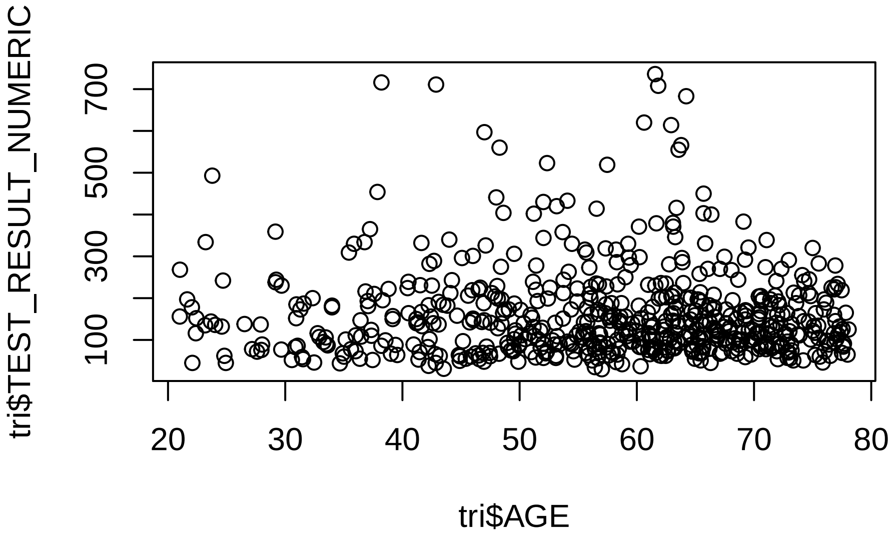
```r
cor(tri$WAIST_CIR, tri$TEST_RESULT_NUMERIC)
```

```
## [1] 0.7191177
```

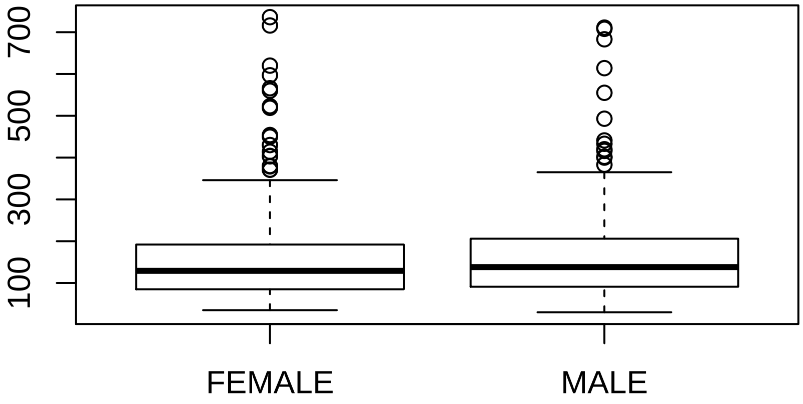
There is a strong linear relationship according to the correlation coefficient (0.719) between waist circumference and triglyceride levels. The correlation may be strong, indicating association, but this is not enough to demonstrate causation

4. Explore the relationships between the remaining variables, `AGE`, `SEX_NAME`, and other disease indicators in relationship to the target variable, `TEST_RESULT_NUMERIC`. Is a scatterplot appropriate to evaluate the relationship between `SEX_NAME` and `TEST_RESULT_NUMERIC`? If not, use the appropriate visual tool to assess if there is any difference in triglyceride levels across genders.

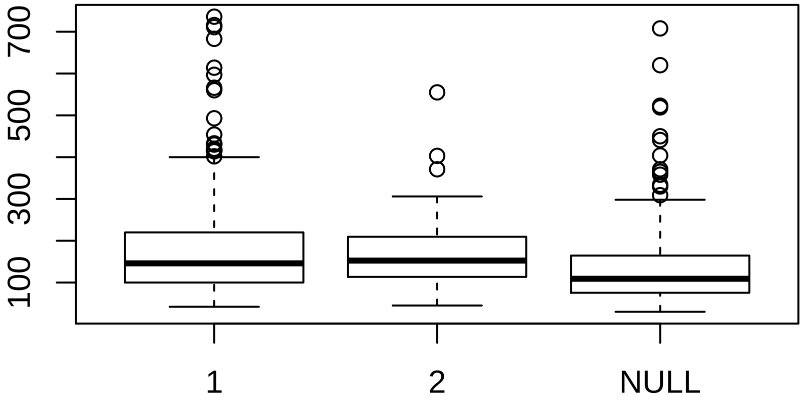
```r
plot(tri$AGE, tri$TEST_RESULT_NUMERIC)
```



```
boxplot(tri$TEST_RESULT_NUMERIC ~ tri$SEX_NAME)
```



```
boxplot(tri$TEST_RESULT_NUMERIC ~ tri$OBESITY_IND)
```



Since gender and other disease indicators are categorical variables (factors) using a scatterplot to evaluate the relationship is not appropriate. A boxplot across each factor level would be the best way to display such relationships.There appears to a difference between triglyceride levels among genders, but the difference may not be statistically significant.

5. Based on your exploratory data analysis, which of the variables examined best explains the

variability in TEST_RESULT_NUMERIC?

Waist circumference and obesity indicator

Least Squares Estimation

In this section we will fit two regression lines using the least squares approach. First, we are interested in examining the relationship between a patient's waist circumference measurement, WAIST_CIR and their associated triglyceride level, TEST_RESULT_NUMERIC. Secondly, we are interested in examining the relationship between a patient's gender, SEX_NAME and their associated triglyceride level, TEST_RESULT_NUMERIC.

Determine the best fitting linear regression equations using the following procedures.

Triglyceride Level ~ Waist Circumference

1. Calculate the slope, b_0 and y-intercept, b_1 using the following expressions.

$$b_1 = r\frac{s_y}{s_x}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

```
corr.coef = cor(tri$WAIST_CIR, tri$TEST_RESULT_NUMERIC)
s.y = sd(tri$TEST_RESULT_NUMERIC)
s.x = sd(tri$WAIST_CIR)
xbar.y = mean(tri$TEST_RESULT_NUMERIC)
xbar.x = mean(tri$WAIST_CIR)

b1 = corr.coef * (s.y / s.x)
b1
```

```
## [1] 15.69962
```

```
b0 = xbar.y - b1 * xbar.x
b0
```

```
## [1] -391.5859
```

2. Now, use R to calculate the coefficients of the least squares line using the lm() function. Compare your results.

```
fit1 = lm(TEST_RESULT_NUMERIC ~ WAIST_CIR, data = tri)
fit1$coefficients
```

```
## (Intercept)   WAIST_CIR
##  -391.58590    15.69962
```

Triglyceride Level ~ Gender

Refer to *Question 4* in the Exploratory Data Analysis section above, specifically the summary statistics and visualizations created to infer the relationship between triglyceride levels, `TEST_RESULT_NUMERIC` and gender, `SEX_NAME`.

3. Does the observed means for triglyceride levels by gender appear to be widely different? Calculate the two sample means of triglyceride level for each gender (if not determined earlier). Calculate the difference in both sample means.

```
xbar.male = mean(subset(tri, tri$SEX_NAME == "MALE")$TEST_RESULT_NUMERIC)
xbar.female = mean(subset(tri, tri$SEX_NAME == "FEMALE")$TEST_RESULT_NUMERIC)
diff = xbar.male - xbar.female
diff
```

```
## [1] 7.106309
```

4. Now, build a linear regression model using R's `lm()` function to explain triglyceride levels, `TEST_RESULT_NUMERIC` using gender, `SEX_NAME` as your *only* explanatory variable.

```
fit2 = lm(TEST_RESULT_NUMERIC ~ SEX_NAME, data = tri)
fit2$coefficients
```

```
##  (Intercept) SEX_NAMEMALE
##   159.038095     7.106309
```

5. Comment on the relationship between the estimated coefficients for β_0 and β_1 and the sample means of triglyceride levels calculated for each gender in Question 1. What does the difference in sample means of triglyceride levels for each gender equal in terms of estimated coefficients?

The β_0 equals the sample mean triglyceride level for females. The β_1 equals the difference in sample means of triglyceride levels between males and females.

Interval Estimation

Just as we constructed confidence intervals for parameters such as μ, we can also calculate confidence intervals for regression parameters, β. We will discuss two approaches to calculating confidence intervals for regression parameters in the model, *Triglyceride Level ~ Waist Circumference*.

1. Calculate the 95% confidence interval for the slope, β_1 using the following expression. Recall the expression for mean squared error, $MSE = \frac{1}{n-2}\sum(y_i - \hat{y}_i)^2$

$$b_1 \pm t_{\frac{\alpha}{2},n-2}\sqrt{\frac{MSE}{\sum(x_i-\bar{x})^2}}$$

```
n = nrow(tri)
cc = 0.95
t = qt((1 - cc) / 2, df = n - 2, lower.tail = TRUE)
hat.y = b0 + b1 * tri$WAIST_CIR
```

```
mse = (1 / (n - 2)) * sum((tri$TEST_RESULT_NUMERIC - hat.y) ^ 2)

b1 + c(1, -1) * t * sqrt(mse / sum((tri$WAIST_CIR - xbar.x) ^ 2))
```

```
## [1] 14.47298 16.92626
```

2. Now, use R to calculate the 95% confidence interval for the slope, β_1 using `confint()` function. Compare your results.

```
confint(fit1, level = 0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept) -435.30489 -347.86692
## WAIST_CIR     14.47298   16.92626
```

Interpretation

Triglyceride Level ~ Waist Circumference

1. Interpret b_1, b_0, and the 95% confidence interval for β_1 in the context of the research question. Does, b_0 have any contextual meaning?

b_1: For each additional inch change in waist circumference, on average you would expect to observe a 15.69 mg/dL change in triglyceride level, holding all else constant. b_0: Does not have any contextual meaning.

Triglyceride Level ~ Gender

2. Interpret b_1 and b_0 in the context of the research question. Does, b_0 have any contextual meaning?

b_1: The average difference in triglyceride levels for males and females. b_0: If a patient is a Female, the average triglyceride level is 159.04.

Model Fit

In this section we will evaluate the fit of both models estimated previously.

Triglyceride Level ~ Waist Circumference

1. Calculate the strength of the association, r between a patient's waist circumference measurement, `WAIST_CIR` and their triglyceride level, `TEST_RESULT_NUMERIC`.

```
corr.coef
```

```
## [1] 0.7191177
```

2. What is the proportion of variability in the response, triglyceride levels, explained by the model? Calculate using the following expression. Does the model, including only the explanatory variable, waist circumference, do a fair job at explaining the variability in triglyceride levels according to R^2?

$$R^2 = \frac{s_y^2 - s_{residuals}^2}{s_y^2}$$

```
r2 = (var(tri$TEST_RESULT_NUMERIC) - var(fit1$residuals)) /
      var(tri$TEST_RESULT_NUMERIC)
r2
```

```
## [1] 0.5171303
```

```
# or

corr.coef ^ 2
```

```
## [1] 0.5171303
```

The model using a patient's waist circumference explains approximately 52 percent of the variation observed triglyceride levels. For the model only including this one variable this is a good contribution in terms of explanatory power

Triglyceride Level ~ Gender

3. What is the proportion of variability in the response, triglyceride levels, explained by the model? Calculate using the expression offered in *Question 2*. Does the model, including only the explanatory variable, gender, do a fair job at explaining the variability in triglyceride levels according to R^2?

```
r2 = (var(tri$TEST_RESULT_NUMERIC) - var(fit2$residuals)) /
      var(tri$TEST_RESULT_NUMERIC)
r2
```

```
## [1] 0.001041977
```

The model using a patient's gender explains approximately 0.1 percent of the variation observed triglyceride levels. For the model only including this one variable this is a poor contribution in terms of explanatory power

Statistical Inference

Since the least squares method offers a way to estimate population parameters, we are essentially performing statistical inference, where the slope of the least squares line, b_1 is an estimate of the population slope, β_1 and the intercept of the least squares line, b_0 is an estimate of the population intercept, β_0.

In regresssion the typical hypothesis for β_1 is as follows:

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

where under the null hypothesis, H_0 the explanatory and response variables are not associated and under the alternative hypothesis the explanatory and response variables are associated. The test statistic, $t-statistic$ is provided below:

$$t = \frac{b_1}{s.e.(b_1)}$$

where the $t-statistic$ follows a t distribution with $df = n-2$.

Triglyceride Level ~ Waist Circumference

1. Setup the appropriate hypothesis test for $\beta_{Waist-Circumference}$ and provide the test statistic and $p-value$ for the testing outcome. Rule on the test and provide your interpretation in the context of the research hypothesis.

```
t.stat = b1 / sqrt(mse / sum((tri$WAIST_CIR - xbar.x) ^ 2))
t.stat
```

```
## [1] 25.13686
```

```
p = 2 * pt(t.stat, df = n - 2, lower.tail = FALSE)
p
```

```
## [1] 2.447292e-95
```

2. Use the R function, summary(model) to confirm your $t-statistic$ and $p-value$ for $\beta_{Waist-Circumference}$.

```
summary(fit1)
```

```
##
## Call:
## lm(formula = TEST_RESULT_NUMERIC ~ WAIST_CIR, data = tri)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -181.17  -38.94  -10.20   16.74  521.93
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -391.5859    22.2603  -17.59   <2e-16 ***
## WAIST_CIR     15.6996     0.6246   25.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76.46 on 590 degrees of freedom
## Multiple R-squared:  0.5171, Adjusted R-squared:  0.5163
## F-statistic: 631.9 on 1 and 590 DF,  p-value: < 2.2e-16
```

Triglyceride Level ~ Gender

3. Use the R function, summary(model) to determine the $t-statistic$ and $p-value$. Also, use the R function t.test() to evaluate the difference between triglyceride levels according to gender. Do you see any similiarities between the test statistics in the regression output and t test output?

```r
summary(fit2)
```

```
##
## Call:
## lm(formula = TEST_RESULT_NUMERIC ~ SEX_NAME, data = tri)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -136.14  -75.14  -28.14   37.88  576.96
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  159.038      6.196  25.666   <2e-16 ***
## SEX_NAMEMALE   7.106      9.059   0.784    0.433
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 110 on 590 degrees of freedom
## Multiple R-squared:  0.001042,   Adjusted R-squared:  -0.0006512
## F-statistic: 0.6154 on 1 and 590 DF,  p-value: 0.4331
```

```r
t.test(tri$TEST_RESULT_NUMERIC ~ tri$SEX_NAME)
```

```
##
##  Welch Two Sample t-test
##
## data:  tri$TEST_RESULT_NUMERIC by tri$SEX_NAME
## t = -0.78423, df = 579.64, p-value = 0.4332
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -24.90362  10.69100
## sample estimates:
## mean in group FEMALE   mean in group MALE
##             159.0381             166.1444
```

The test statistic and p-value for the estimated linear regression and t test are the same. As expected...

Prediction

In addition to performing inference about the association between the response and explanatory variables, we can use the least squares line to predict responses according to values of the explanatory variable that fall within the scope of the model. For example one might like to know the predicted triglyceride level for a patient with a waist circumference, $x_{Waist-Circumference}$ or the predicted triglyceride level for a patient that is x_{Male}.

Triglyceride Level ~ Waist Circumference

1. Use your estimated simple linear regression model for *Triglyceride Level ~ Waist Circumference* to predict the triglyceride level of a patient whos waist circumference measures 32 inches.

```
b0 + b1 * 32
```

```
## [1] 110.8019
```

2. Now use the R function `predict()` to confirm your answer in part 1 and provide a *prediction* interval.

```
new.df = data.frame(WAIST_CIR = 32)
predict(fit1, newdata = new.df, interval = "predict")
```

```
##        fit       lwr     upr
## 1 110.8019 -39.54717 261.151
```

Triglyceride Level ~ Gender

3. Use your estimated simple linear regression model *Triglyceride Level ~ Gender* to predict the triglyceride level of a patient whos is 'Female'.

```
159.038 + 7.106 * 0
```

```
## [1] 159.038
```

4. Now use the R function `predict()` to confirm your answer in part 3 and provide a *prediction* interval.

```
new.df1 = data.frame(SEX_NAME = "FEMALE")
predict(fit2, newdata = new.df1, interval = "predict")
```

```
##        fit       lwr      upr
## 1 159.0381 -57.29612 375.3723
```

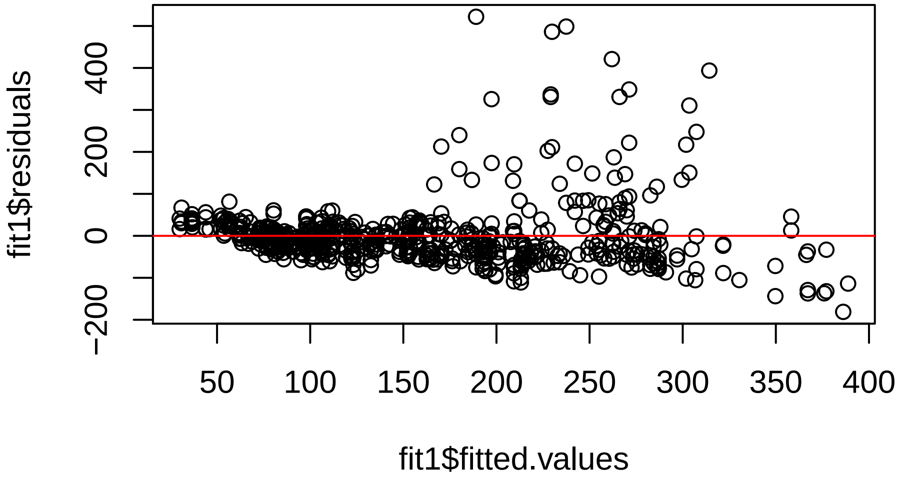
Checking Model Assumptions for Regression

In order to use linear regression to better understand the relationship between two variables, the following conditions need to validated. You will need to verify each is met or call out concerns based on your observations of the residual plots, normal probability plots, and histogram of your residuals. In this section we will only use the *Triglyceride Level ~ Waist Circumference* simple linear regression model.

- **Linearity**: data show a linear trend in change for response y as a function of predictor x.
- **Constant variability**: the variability of the response variable about the line remains roughly constant as the predictor variable changes.
- **Independent observations**: the (x, y) pairs are independent; i.e., values of one pair provide no information about other pairs.
- **Approximate normality of residuals**: $e_i = y_i - \hat{y}_i$ where e_i are normally distributed under $N(0, \sigma)$.

1. Create a residual plot where the predicted values, $\hat{y}$ are on the *x-axis* and the residuals are on the *y-axis*. Use the following R function to add a horizontal line to your residual plot, *abline(h = 0, col = "red")*. What do you observe regarding conditions 1 and 2?

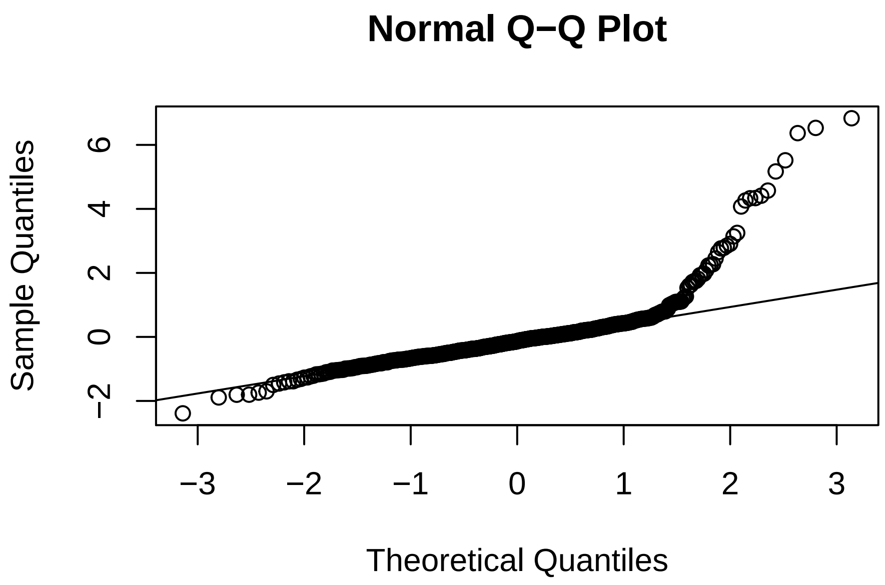
```
plot(fit1$fitted.values, fit1$residuals)
abline(h = 0, col = "red")
```



There are concerns with both linearity and constant variance according to the residual plot.

2. Create a normal probability plot where the theoretical quantiles for a normal distribution are on the *x-axis* and the observed quantiles from the data are on the *y-axis*, and a histogram of the residuals. Observe the tails of the distribution in the normal probability to monitor for departures from normality. Is the histogram of the residuals nearly normal? Does them confirm or deny assumption 4?

```
fit1.std.res = rstandard(fit1)
qqnorm(fit1.std.res)
qqline(fit1.std.res)
```



There are major concerns with linearity based on the upper tail departing from its expected behavior.

3. Use a common sense approach paired with your understanding of the study to validate assumption 3. Explain your reasoning.

Since each patient is independent of the next, the independence criteria should easily be satisfied.

4. Does your investigation of the conditions for linear regression provide pause for concern? If so, what might you suggest to correct such issues and continue using linear regression as your choice for model building? Hint: *Transformations, weighted regression, leverage values/outlier identification, etc.* Explain.

There are quite a few concerns with using simple linear regression to model this relation. 3 out of the 4 conditions for linear regression are violated. One might suggest using a log-linear or log-log model to best address the modeling.