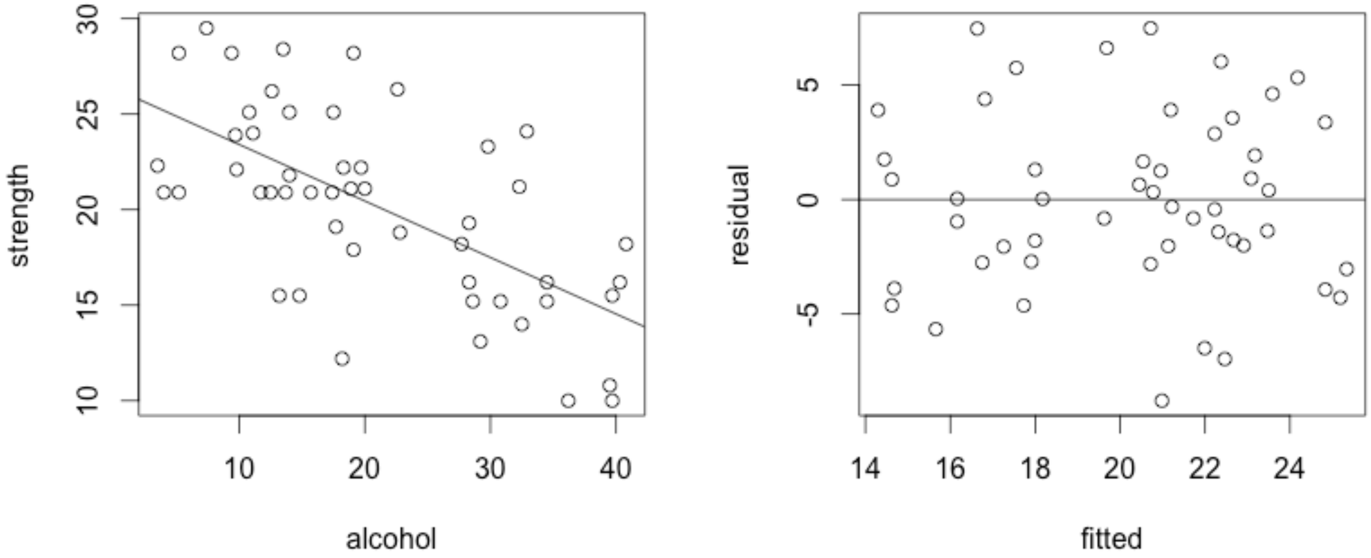
# Homework 1 – Due Jan 13th @ 11pm

- You may type or handwrite your homework.
- No specific format is required as far as the graders are able to find your answers.
- Submit your HW in a PDF format. The submission link is in Week 1 module.
- All homework assignment will be graded out of 20.

## 1. Review on simple linear regression

Suppose that some researchers were interested in predicting muscle strength with alcohol consumption with a linear model. The researchers measured the total lifetime consumption of alcohol (in gallons) on a random sample of 50 alcoholic men. They also measured the strength of the deltoid muscle (in grade) in each person's non-dominant arm.



| | Alcohol consumption (gallon) | Strength of deltoid muscle (grade) |
|---|---|---|
| mean | 20.97 | 20.16 |
| standard deviation | 10.84 | 5.00 |
| $R^2$ | 0.4117 | |

(1) From the scatterplot, which of following best describes the relationship between the alcohol consumption and strength of deltoid muscle?

A. They have nearly perfect negative linear relationship.
B. They have moderate negative linear relationship.
C. They have positive nonlinear relationship.
D. They have relatively strong positive linear relationship.

(2) Find a correlation coefficient of the two variables.

$r = \sqrt{0.4117} = \pm 0.6416$ (relationship is negative) → $-0.6416$

(3) Find the slope and interpret it.

Slope: $r \cdot \frac{S_y}{S_x} = -0.6416 \cdot \frac{5}{10.84} \approx -0.2959$

For an increase of 1 gallon in alcohol consumption, we expect a decrease of 0.2959 in deltoid muscle strength on average

(4) Which of the following is the best interpretation of this value of $R^2$?

A. About 41% of observations can be accurately predicted by the model.
B. About 41% of observations fall on the regression line.
C. About 41% of the total variability in strength of deltoid muscle is explained by the model.
D. About 41% of the total variability in alcohol consumption is explained by the strength of the deltoid muscle.

(5) We plot the residuals. Which of the following best interprets the plot?

A. It does have a curve pattern and it could be a problem.
B. It does not have any apparent pattern and it could be a problem.
C. It does not have any apparent pattern and it implies the good fit of the model.
D. It does have fan-shaped pattern and it implies the good fit of the model.

(6) What is the approximate predicted strength of deltoid muscle when one consumes 35 gallons of alcohol throughout the lifetime? Show your work.

mean(y) − (slope · mean(x)) = intercept

$20.16 - (-0.2959 \cdot 20.47) = 26.365$

$y = -0.2959x + 26.365$ $y = -10.3565 + 26.365$

$y = -0.2959(35) + 26.365$ $y = 16$

## 2. Review on Hypothesis testing

The average number of close friends for all people living in the U.S. is 5.7. An investigator claims that the mean number of close friends for introverts will be significantly less than the mean of the population, 5.7. The mean number of close friends for a sample of 36 introverts is 5.1, and the standard deviation is 1.4.

(1) State the null and alternative hypotheses.

$H_0: \mu = 5.7$

$H_A: \mu < 5.7$

(2) Calculate the test statistic. Show your work.

$$T = \frac{\text{mean (Sample)} - \text{mean (population)}}{\text{sd( sample)} / \sqrt{\text{sample size}}} \qquad T = \frac{5.1 - 5.7}{1.4/\sqrt{36}} = \frac{-0.6}{1.4/6} = -2.5714$$

(3) Find the p-value. run \`pt( -2.5714, df = 36-1)' in R

$P(t < -2.5714) = 0.007$

(4) Make a conclusion for the test. Use α=0.05.

The p-value is < 0.05 so we reject the null hypothesis. There's sufficient hypothesis suggesting the average number of close friends is < 5.7

(5) Construct a 95% confidence interval for the average number of close friends. (for introverts, assuming random sampling)

run \`qt(0.025, 36-1)' in R for t-score

$$5.1 \pm \text{new t-score} \cdot \frac{sd(x)}{\sqrt{\text{sample size}}}$$

$$5.1 \pm 2.0301 \cdot \frac{1.4}{\sqrt{36}} = 4.626,\ 5.574$$

(6) Interpret the interval in the context of the study.

We're 95% confident that the average number of close friends for introverts falls between 4.629 and 5.574

## 3. Review on R programming - see the next page

- Download the *heart.csv* data from Week 1 in Bruinlearn.
- Copy and paste your R commands, plots, and analysis results when you answer the questions.

(1) Conduct simple linear regression using MaxHR as outcome variable and Chol as a predictor.

- Report the summary of your linear model, i.e. write a regression equation.
- Interpret the slope and the y-intercept in the context of data.

(2) Create a scatter plot for the Chol vs MaxHR then plot the least square regression line on the same graph (use: abline command)

(3) Report R-square and create a residual plot. How do you assess the goodness-of-fit of the model?

## 4. Pre-Course Survey: (1pt)

Copy the following statement if you have completed the Pre-Course Survey in Week 1 module:
"I have completed the Pre-Course Survey."

I have completed the Pre-Course Survey

# Stats 101A Homework 1 Problem 3

Damien Ha

01/13/2023

## Contents

## Loading Data

```r
heart <- read.csv("Heart.csv")
heart = subset(heart, select = -c(X) )
```

## Part 1

Conduct simple linear regression using MaxHR as outcome variable and Chol as a predictor.

- Report the summary of your linear model, i.e. write a regression equation.
- Interpret the slope and the y-intercept in the context of data.

```r
model <- lm(MaxHR ~ Chol, data = heart)
summary(model)
```

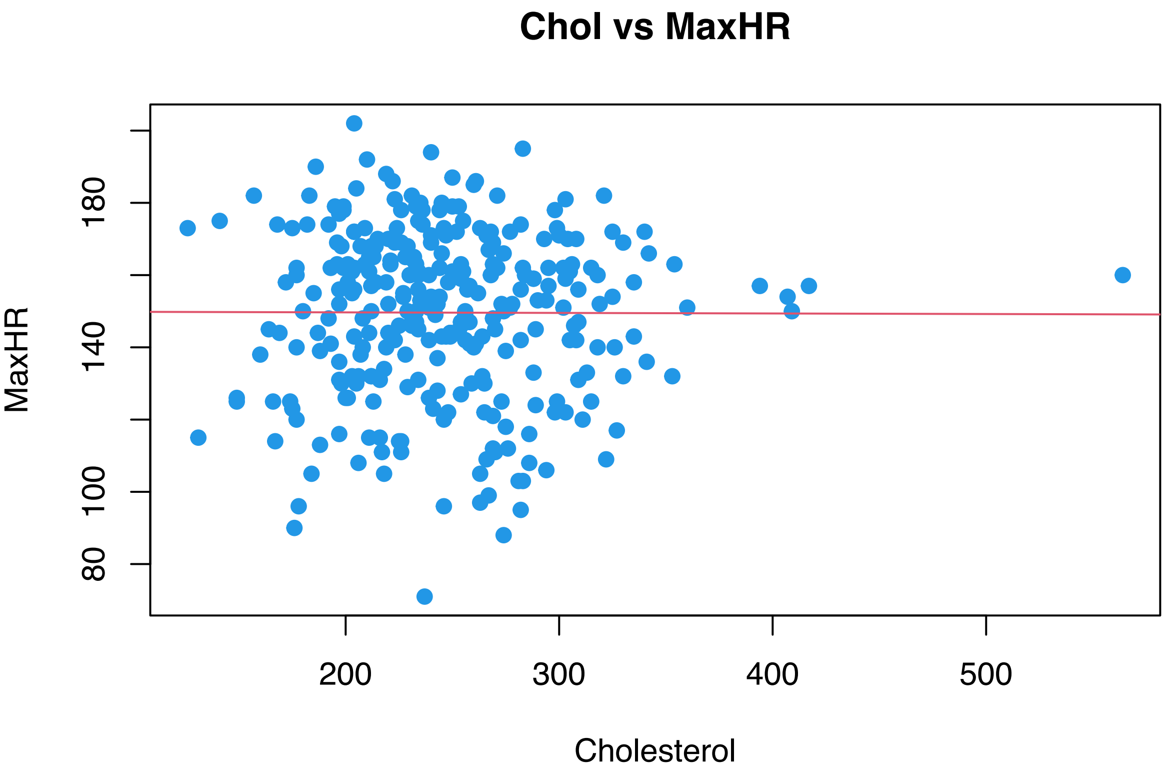
```
##
## Call:
## lm(formula = MaxHR ~ Chol, data = heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -78.622 -16.079   3.375  16.412  52.328
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 149.981292   6.418400   23.37   <2e-16 ***
## Chol         -0.001516   0.025465   -0.06    0.953
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.91 on 301 degrees of freedom
## Multiple R-squared:  1.178e-05,  Adjusted R-squared:  -0.00331
## F-statistic: 0.003545 on 1 and 301 DF,  p-value: 0.9526
```

The slope is -0.0015 and the intercept is 149.98. For an increase of 1 unit in cholesterol, we espect a decrease of 0.0015 units in MaxHR on average. At a cholesterol level of 0, we expect a MaxHR of 149.98 on average.

## Part 2

Create a scatter plot for the Chol vs MaxHR then plot the least square regression line on the same graph

```
plot(heart$Chol, heart$MaxHR,
     xlab = "Cholesterol", ylab = "MaxHR", main = "Chol vs MaxHR",
     pch = 19, col = 4)
abline(model, col = 2)
```



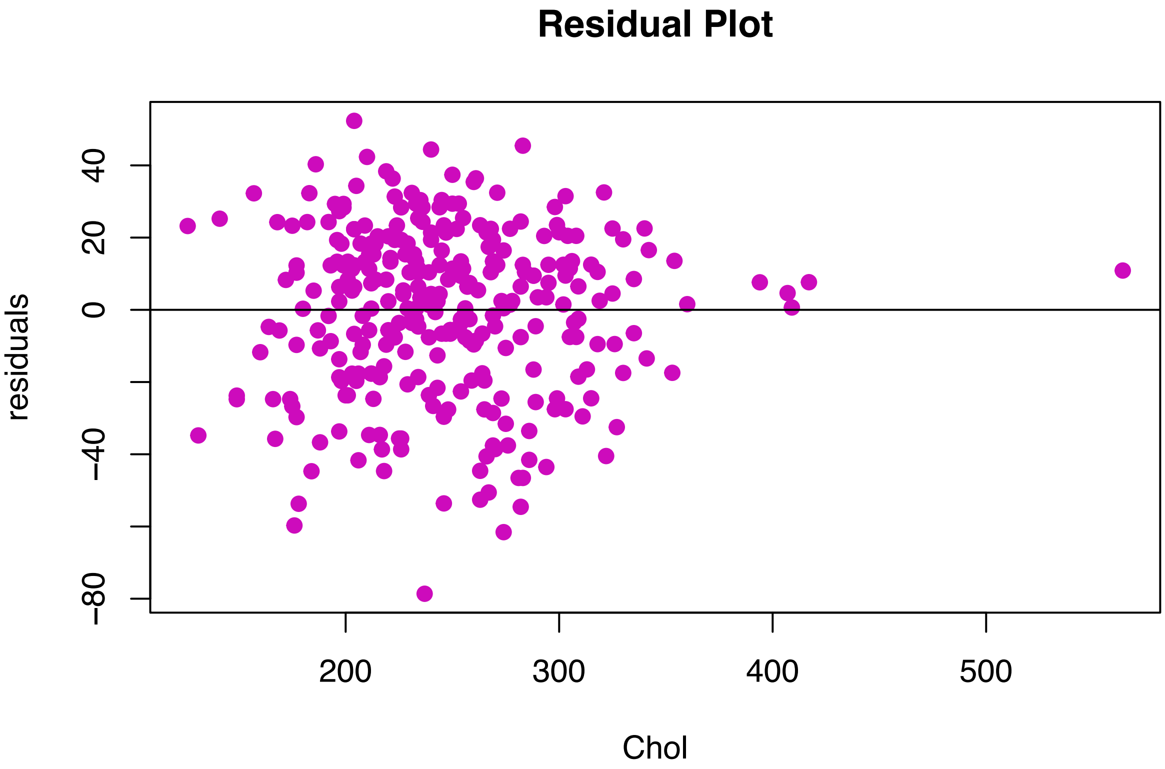
## Part 3

Report R-square and create a residual plot. How do you assess the goodness-of-fit of the model?

```
summary(model)$r.squared
```

```
## [1] 1.177747e-05
```

```
plot(heart$Chol, model$residuals,
     xlab = "Chol", ylab = "residuals", main = "Residual Plot",
     pch = 19, col = 6)
abline(0,0)
```



From the low R-squared value, it seems that MaxHR is not explained very much by the linear model. The residual plot does seem to be random without any patterns, but the data points in the scatterplot are not clustered very closely to the line. The model does not have a very good fit.