3301HW4

Joshua O’Neill

2024-03-10

R Markdown

1. In this problem, we will analyze the relationship between wine consumption and mortality from heart disease. To do so, we will use data collected from 18 countries:

• mortality, the mortality rate from heart disease per thousand;

• consumption, the average per capita consumption of wine in liters.

These data are available for download on canvas in the file “wine.txt”. Once you’ve set the working directory, you may read the data into R and look at its first six rows using

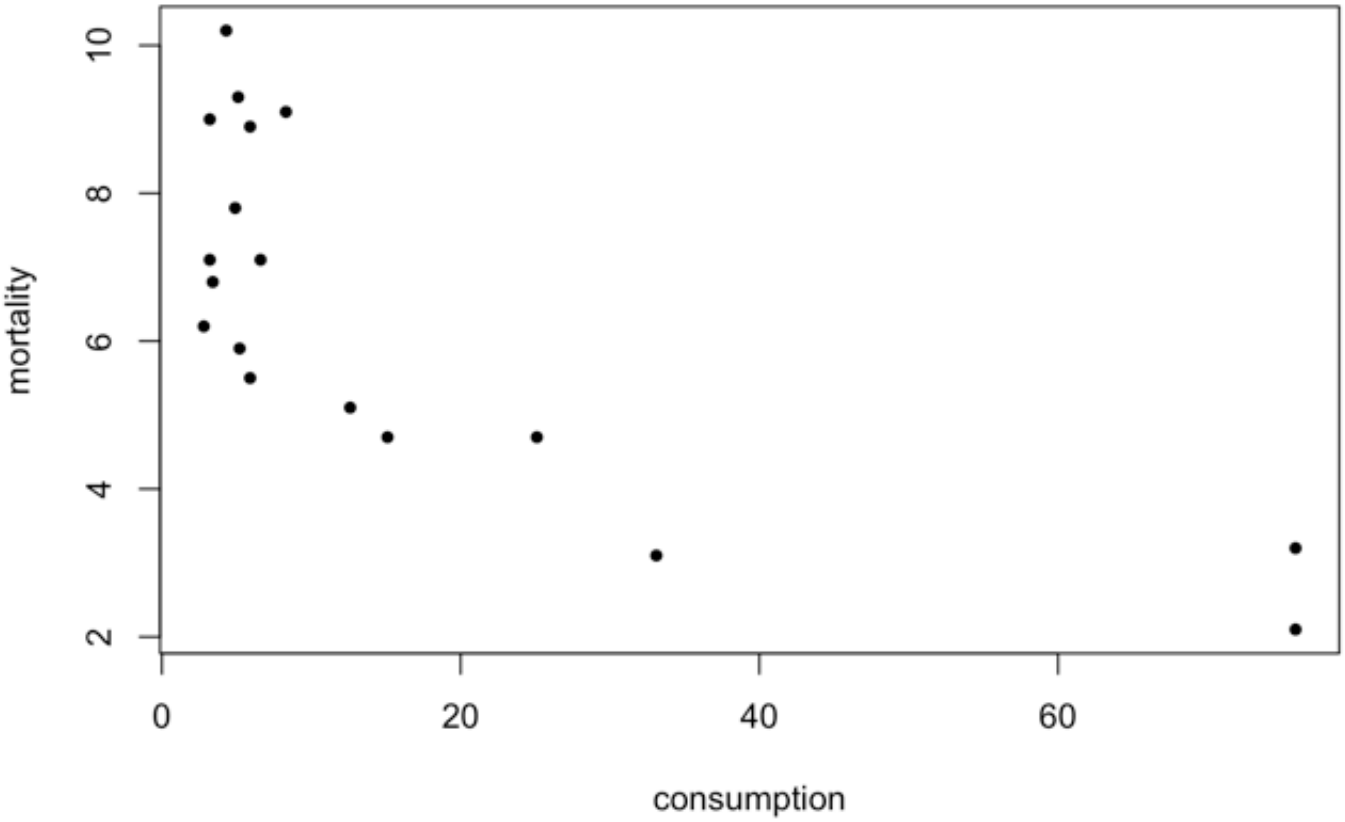
```
wine <- read.table("wine.txt", header=TRUE)
head(wine)
```

```
##   consumption mortality  country
## 1         2.8       6.2   Norway
## 2         3.2       9.0 Scotland
## 3         3.2       7.1  England
## 4         3.4       6.8  Ireland
## 5         4.3      10.2  Finland
## 6         4.9       7.8   Canada
```

Our goal is to study the relationship between mortality and consumption.

(a) (1pt) Produce a scatter plot of mortality versus consumption. Based on this scatter plot, would a linear regression model with mortality as the response and consumption as the predictor be appropriate? Explain.

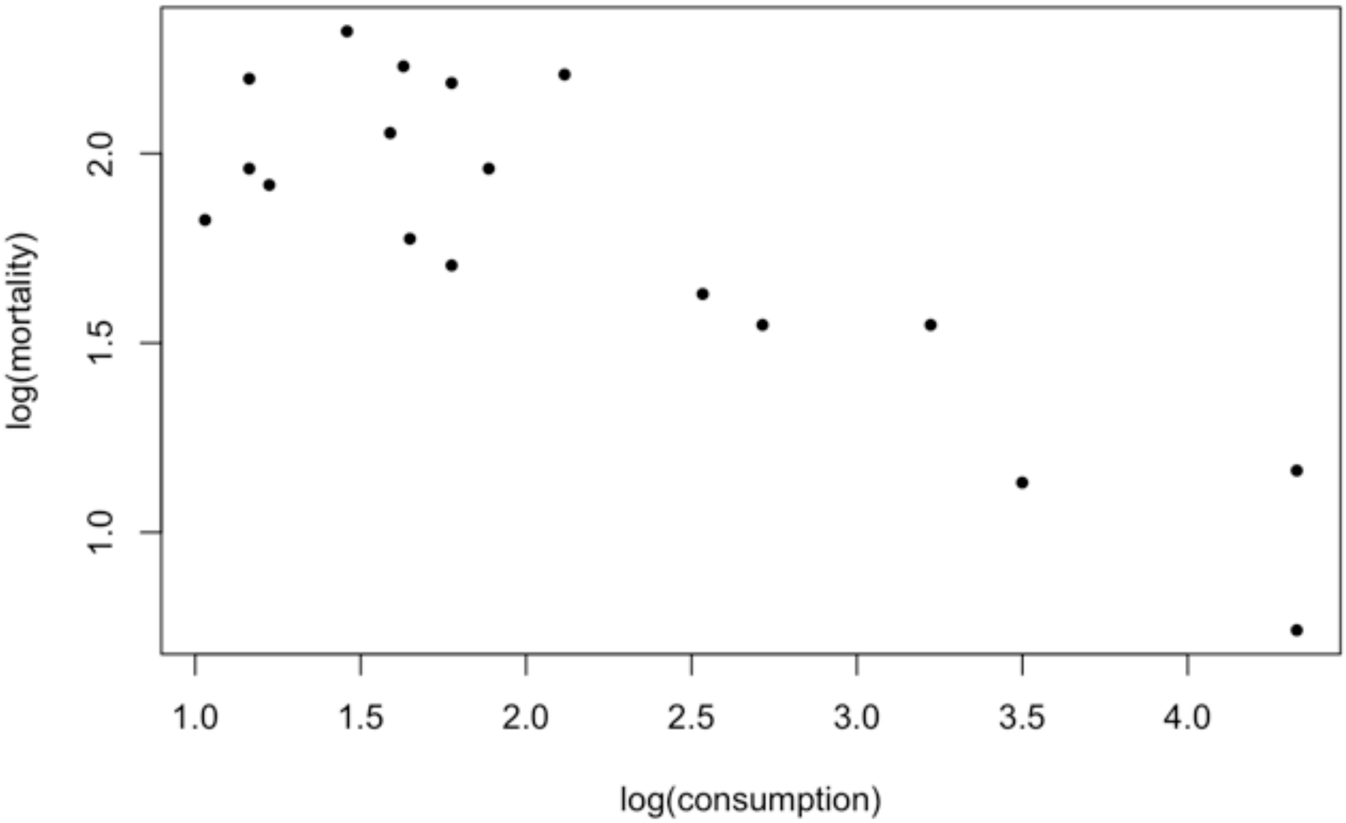
```
plot(mortality ~ consumption, pch=20, data=wine)
```



It does not appear that a linear regression model would be appropriate, since the dots instead seem arranged along a curve. Said in mathematical terms, the best slope of a line connecting the dots would be steeper when consumption is low, and flatter when consumption is high. A linear line has a constant slope, so this approximation does not seem appropriate.

(b) (1pt) Instead of predicting mortality with consumption, we will consider predicting log(mortality) with log(consumption), where log() is the natural logarithm. In R, this is simply the function log. Produce a scatter plot of log(mortality) versus log(consumption). Relative to their original scales, do you think linear regression model with log(mortality) as the response and log(consumption) as predictor is more appropriate?

```
plot(log(mortality) ~ log(consumption), pch=20, data=wine)
```



A linear regression model accounting for the relationship between log(mortality) and log(consumption) seems significantly more appropriate, as the path the dots take from one side of the chart to the other does not seem be curvy, implying that an ideal regression line would have a constant slope.

(c) (2pts) We will fit a linear regression model to these data where the response is log(mortality) and the predictor is log(consumption). What does this model assume about these data? Be as specific as possible and define every symbol that you use.

This model assumes that log(consumption) plotted against log(mortality) produces a roughly linear relationship that can be approximated well by a line, which assumes that the data is related via their logs. Specifically, it assumes that log(mortality) is a realization of log(consumption)(beta)=epsilon, where beta is a p X 1 matrix, and epsilon is a n X 1 matrix, where each value in the epsilon matrix ~Norm(0,σ^2), where σ is some unknown constant value.

(d) (2pts) Use R to compute estimates of the unknown parameters in the linear regression model described in part 1c. Report these estimates. You may use lm, or you may compute the estimates “by-hand”.

```
ones.vector <- rep(1,times=length(wine$consumption))
mortality.X <- cbind(ones.vector,log(wine$consumption))
mortality.y <- log(wine$mortality)
mortality.beta.cap <- qr.solve(crossprod(mortality.X),crossprod(mortality.X,mortality
.y))
print(mortality.beta.cap)
```

```
##                    [,1]
## ones.vector  2.5555519
##             -0.3555959
```

2.5555519 is the intercept of the fitted line; β_1

-0.3555959 is the slope of the fitted line; β_2

(e) From 1d, you should have the OLS estimate of the regression coefficients from your model in 1c.

i. (2pts) If it makes sense, interpret the estimated intercept, ˆβ_1, in the context of the problem.

β_1 = 2.5555519 means, in the context of the problem, that log(mortality)=2.5555519 when log(consumption)=0 is what our model predicts.Note that log(1)=0, so the above is equivalent to saying, when consumption=1, log(mortality)=2.5555519. Finally, we can get that mortality=e^2.5555519=12.8784, when consumption = 1, as our model's prediction.

ii. (2pts) Based on your fitted model, how does a unit increase in log(consumption) affect log(mortality)? Be as specific as possible.

A unit increase in log(consumption) decreases log(mortality) by 0.3555959, based on the model.

iii. (1pt EC) Describe how a unit increase in consumption affects mortality. It may be helpful to refer to the discussion in part 1g.

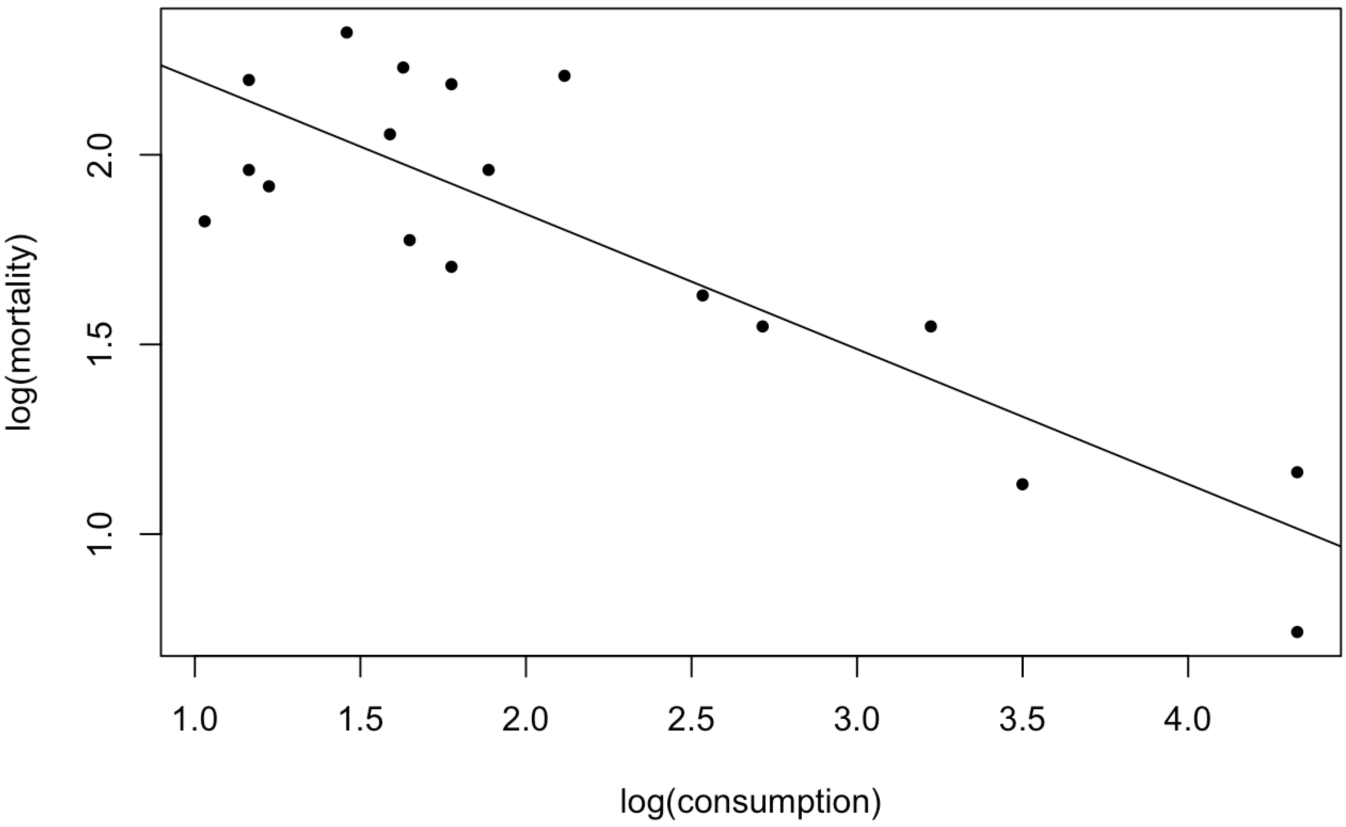
According to our model,

$$\log(\text{mortality}) = 2.5555519 - 0.3555959 * \log(\text{consumption})$$
$$\text{mortality} = e^{2.5555519 - 0.3555959 * \log(\text{consumption})}$$
$$\text{mortality} = e^{2.5555519} * e^{-0.3555959 * \log(\text{consumption})}$$
$$\text{mortality} = e^{2.5555519} * e^{\log(\text{consumption}^{-0.3555959})}$$
$$\text{mortality} = e^{2.5555519} * \text{consumption}^{-0.3555959}$$
$$\text{Thus, a unit increase in consumption will effect mortality like-so:}$$
$$e^{2.5555519} * (\text{consumption} + 1)^{-0.3555959} - (e^{2.5555519} * (\text{consumption})^{-0.3555959})$$
$$e^{2.5555519} * ((\text{consumption} + 1)^{-0.3555959} - (\text{consumption})^{-0.3555959})$$

Practically, this means that a unit increase in consumption will have a different change in mortality, depending on what the current wine consumption is prior to the increment. That precise difference involves taking the difference of the current consumption taken to a negative power and the current wine consumption plus one, all taken a negative power, then multiplying the result by a constant.

(f) (2pts) Plot the fitted model's estimate of the mean log(mortality) as the log(consumption) varies from to 1.0 to 4.5 over the scatter plot of log(mortality) versus log(consumption).

```
plot(log(mortality) ~ log(consumption), pch=20, data=wine)
abline(2.5555519,-0.355959)
```



(g) (3pts) Plot the model's predicted values of mortality as consumption varies from to 3.0 to 75.0 over the scatter plot of mortality versus consumption. You should note that if

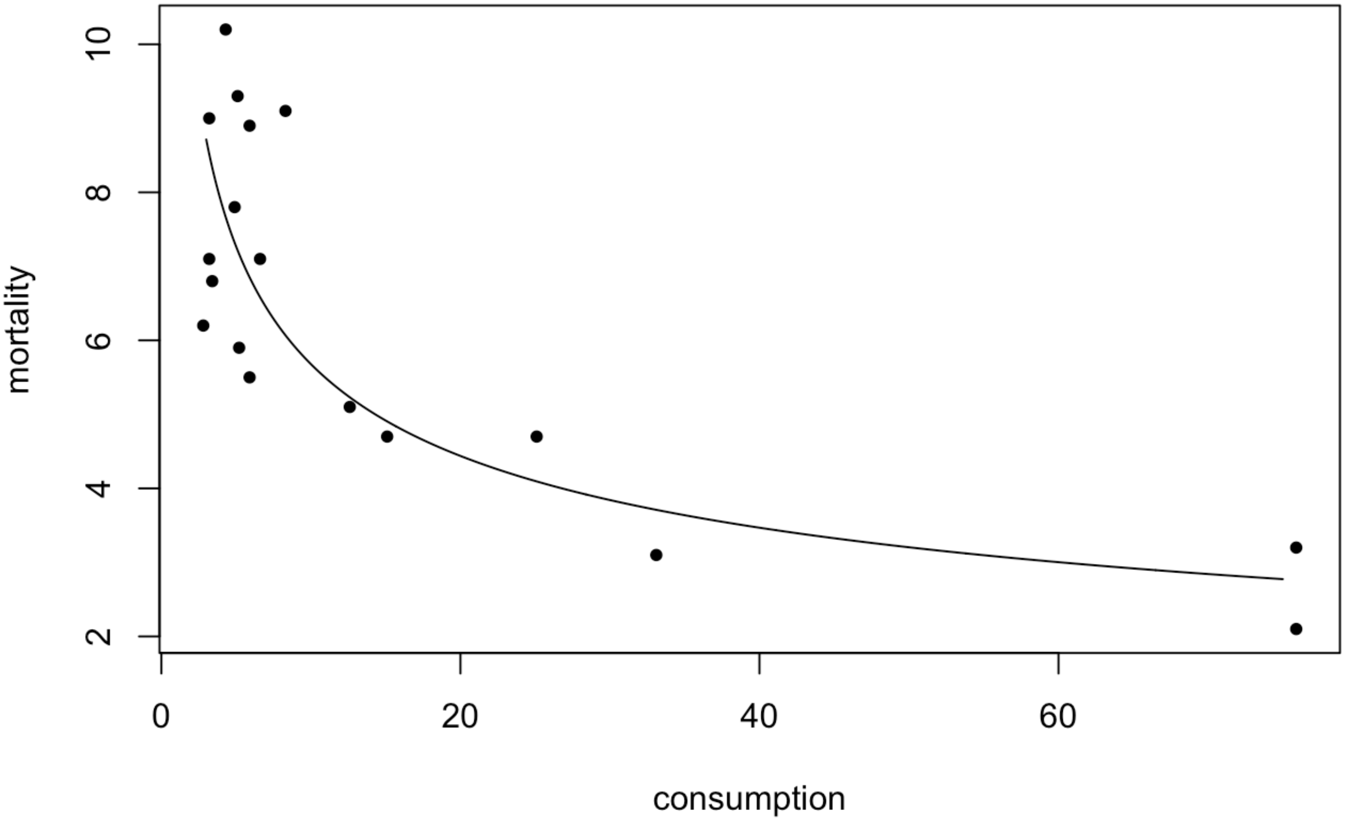
$$\log(\text{mortality}) = \beta_1 + \beta_2 \log(\text{consumption}) + \epsilon$$

for mean zero random error ε, then

$$\text{mortality} = e^{\beta_1 + \beta_2 \log(\text{consumption}) + \epsilon}$$

Hence, based on our model, a reasonable prediction of mortality at consumption = x would be exp{β1 + β2 log(x)}.

```
plot(mortality ~ consumption, pch=20, data=wine)
x.axis <- seq(3,75,0.1)
y.axis <- exp(2.5555519-0.3555959*log(x.axis))
lines(x.axis,y.axis)
```



(h) (2pts) Based on your fitted model, what do you predict to be the mortality rate of a country with per capita wine consumption of 15.5 liters?

Based on earlier computation, mortality, by our model, is about:

$$\text{mortality} = e^{2.5555519} * \text{consumption}^{-0.3555959}$$

Thus when consumption = 15.5, we can solve for mortality

$$\text{mortality} = e^{2.5555519} * 15.5^{-0.3555959}$$

```
exp(2.5555519)*15.5^(-0.3555959)
```

```
## [1] 4.859426
```

Therefore, we have 4.859426 as our prediction of a mortality rate for a country that, per capita, consumes 15.5 liters of wine on average.

Testing:

```
wine.summary <- lm(I(log(mortality)) ~ I(log(consumption)),wine)
show(wine.summary)
```

```
##
## Call:
## lm(formula = I(log(mortality)) ~ I(log(consumption)), data = wine)
##
## Coefficients:
##         (Intercept)  I(log(consumption))
##              2.5556              -0.3556
```

###2. Researchers at the University of Minnesota are interested in studying how a sustained cardiovascular exercise program reduces the resting heart rate (RHR) of untrained individiuals. To answer this question, they recruited 35 subjects who had never performed any type of cardiovascular exericse and assigned them to perform either 0.5, 1.0, 1.5, 2.0, or 2.5 hours of low-intensity cardiovascular exercise per week. The data from the study can be downloaded using

```
rhrdat <- readRDS("RHR.RDS")
head(rhrdat)
```

```
##        RHRdec exercise age
## 1 -0.07485900      0.5  24
## 2 -0.22325231      1.5  26
## 3 -0.20845607      2.5  30
## 4 -0.06535626      0.5  37
## 5 -0.14499982      1.5  22
## 6 -0.13974863      2.5  37
```

###The data contain three variables

###• RHRdec: the decrease in resting heart rate after three months of the protocol, computed as

$$\text{RHRdec} = \frac{\text{Month 3 RHR} - \text{Initial RHR}}{\text{Initial RHR}}$$

###so that a value of –0.1 can be interpreted as a 10% decrease in RHR.

###• exercise: the number of hours of exercise per week during the three month study

###• age: the subject's age at the start of the study

(a) (2pts) By hand (i.e., not using the lm function in R), fit the regression model with RHRdec as response and both exercise and age as predictors. Print your estimate of the regression coefficients.

```
rhr.X <- cbind(1,rhrdat$exercise,rhrdat$age)
rhr.y <- rhrdat$RHRdec
rhr.beta.cap <- qr.solve(crossprod(rhr.X),crossprod(rhr.X,rhr.y))
print(rhr.beta.cap)
```

```
##              [,1]
## [1,] -0.208160127
## [2,] -0.038776177
## [3,]  0.004359802
```

β_1 = -0.208160127 = data intercept

β_2 = -0.038776177 = weight of exercise

β_3 = 0.004359802 = weight of age

(b) (2pts) Interpret the estimate of β_2, the regression coefficient corresponding to exercise, in the context of the problem.

For every additional weekly hour that a participant exercises for, after three months, they will experience an additional 3.8776177% decrease in RHR.

c) (1pt) Does it make sense to interpret the estimate of the intercept? If so, do it. If not, explain why not.

It does not, because β_1 = -0.208160127 literally means that if a person was 0 years old, and exercised for 0 hours a week, they would experience approximately a 20.8160127% decrease in RHR, which makes very little sense. Most 0 year olds cannot walk, or even crawl, so the idea of them doing cardio is simply absurd and dismissible as a notion. Hence, this intercept has no inherent meaning.

Note. For the rest of the problem, you may use lm in R.

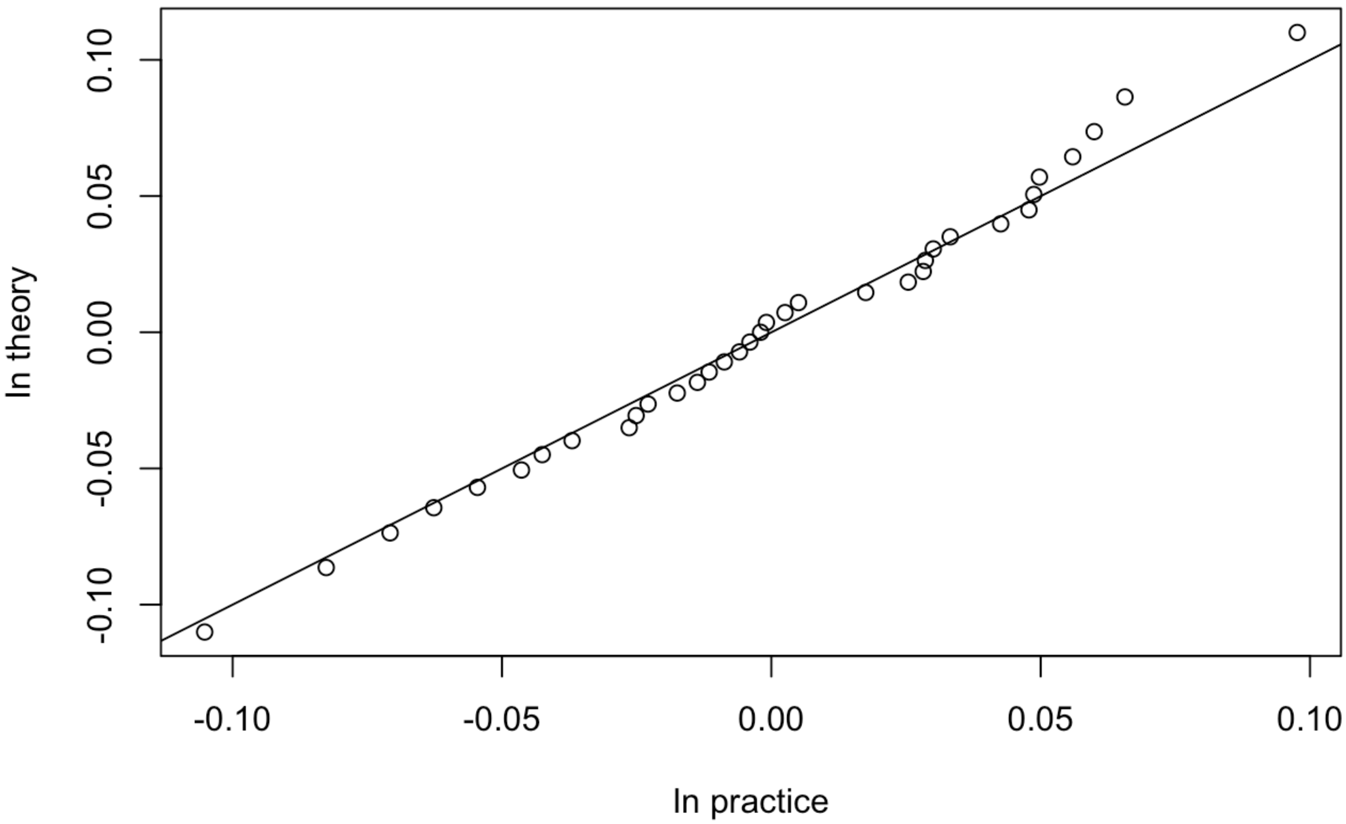
d) Perform a hypothesis test for the effect of age. That is, formally test

$$H_0 : \beta_3 = 0 \text{ vs } H_a : \beta_3 \neq 0$$

To do so,

i. (3pts) Create a QQ-plot to check that the residuals are normally distributed. Interpret your plot and state whether you believe the errors are approximately normal.

```
rhr.fitted <- rhr.X%*%rhr.beta.cap
rhr.eps <- rhr.fitted - rhr.y
probs <- ppoints(length(rhr.fitted))
rhr.quants <- quantile(rhr.eps,probs)
rhr.normal <- qnorm(probs,mean(rhr.eps),sd(rhr.eps))
plot(rhr.quants,rhr.normal,xlab="In practice",ylab = "In theory")
abline(0,1)
```



Because most of the residuals' quantiles fall onto the abline, that implies that the residuals' quantiles are the same as a normal distribution's quantiles', thereby implying that epsilon/the residuals ~Norm(0,σ^2).

ii. (3pts) Assuming the errors are approximately normal, state the observed value of the test statistic T , state the observed p-value, and state your conclusion of the test in the context of the problem.

```
sE <- sqrt((sum(rhr.eps^2))/(length(rhr.eps)-3))
XtXinv <- qr.solve((t(rhr.X))%*%rhr.X)
t.val <- (rhr.beta.cap[3,1]-0)/(sE*sqrt(XtXinv[3,3]))
p.val <- 2*(pt(-abs(t.val),length(rhr.eps)-3))
c(t.val,p.val)
```

```
## [1] 2.901434076 0.006669324
```

Observed test statistic T: 2.901434076

Observed p-value: 0.006669324

Conclusion: We reject the null hypothesis, being that age has no effect on the relative decrease of in RHR, under any practical level of significance.

(e) (4pts) While researchers expected that exercise would reduce resting heart rate, they want to formally test whether it reduces it more than 1% per extra hour of exercise. To do, test

$$H_0 : \beta_2 = -.01 \text{ vs } H_a : \beta_2 < -.01$$

Show all of your work: state the test statistic you use, the distribution of this test statistic under the null, provide the p-value, and state your conclusion of the test in the context of the problem.

Hint. Note that the standard hypothesis testing procedure from class will not work. Ask yourself, under H0, what modified version of your T test statistic follows a t-distribution with n – p degrees of freedom? To compute the p-value, what values of the test statistic are more in favor of Ha?

(So we need to perform a one-tailed t-test, where we consider the rejection region, that which is in the lowest section of the null hypothesis' t-distribution, rather than on both extreme's of the t-distribution, since T being large supports the null hyopthesis more than the alternative hypothesis)

```
t.val <- (rhr.beta.cap[2,1]-(-.01))/(sE*sqrt(XtXinv[2,2]))
p.val <- pt(t.val,length(rhr.eps)-3)
c(t.val,p.val)
```

```
## [1] -2.630455013  0.006501768
```

The test statistic I used was the T = (β_2-(-0.1))/(sE*sqrt(XtXinv_(2,2))

This test statistic follows a degree of freedom of n-p when the null hypothesis is true

The p-value is 0.006501768

My conclusion is that we reject the null hypothesis, being that exercise decreases rhr by 1% per every hour of exercise, in favor of the alternative hypothesis, being that exercise decreases rhr by significantly more than 1% per every hour of exercise, under any practical level of significance.

Testing:

```
rhr.summary <- lm(RHRdec ~ exercise + age,rhrdat)
show(rhr.summary)
```

```
## 
## Call:
## lm(formula = RHRdec ~ exercise + age, data = rhrdat)
## 
## Coefficients:
## (Intercept)     exercise          age  
##    -0.20816     -0.03878      0.00436
```