STAT 420: Homework 04

Spring 2020, Yu Wu (yuw5)

Due: Tuesday, February 25 by 11:30 PM CT

Contents

Assignment

Exercise 1 (Using `lm` for Inference)

For this exercise we will again use the `faithful` dataset. Remember, this is a default dataset in `R`, so there is no need to load it. You should use `?faithful` to refresh your memory about the background of this dataset about the duration and waiting times of eruptions of the Old Faithful geyser in Yellowstone National Park.

(a) Fit the following simple linear regression model in `R`. Use the eruption duration as the response and waiting time as the predictor.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Store the results in a variable called `faithful_model`. Use a t test to test the significance of the regression. Report the following:

- The null and alternative hypotheses
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.01$
- A conclusion in the context of the problem

When reporting these, you should explicitly state them in your document, not assume that a reader will find and interpret them from a large block of `R` output.

```
data("faithful")
fit <- lm(eruptions ~ waiting, data = faithful)

summary(fit)$coefficients[2, 3]
```

```
## [1] 34.08904
```

```
summary(fit)$coefficients[2, 4]
```

```
## [1] 8.129959e-100
```

- The null hypotheses is $\beta_1 = 0$ and alternative hypotheses is $\beta_1 \neq 0$.
- The value of the test statistic is 34.08904.
- The p-value of the test is 8.129959e-100.
- The statistical decision at $\alpha = 0.01$ is reject null hypothesis.

- The conclusion in the context of the problem is that the waiting time is a significant predictor, there is a linear relationship.

(b) Calculate a 99% confidence interval for β_1. Give an interpretation of the interval in the context of the problem.

```
confint(fit, "waiting", level = 0.99)
```

```
##             0.5 %    99.5 %
## waiting 0.0698727 0.0813832
```

When waiting time increases by 1 minute, we are 99% sure that the eruption time will increase between 0.0698727 to 0.0813832 minutes.

(c) Calculate a 90% confidence interval for β_0. Give an interpretation of the interval in the context of the problem.

```
confint(fit, "(Intercept)", level = 0.99)
```

```
##                 0.5 %    99.5 %
## (Intercept) -2.289453 -1.458579
```

When waiting time is at 0 minute, we are 99% sure that the eruption time will be between -2.289453 to -1.458579 minutes.

(d) Use a 95% confidence interval to estimate the mean eruption duration for waiting times of 75 and 80 minutes. Which of the two intervals is wider? Why?

```
x1 <- data.frame(waiting = 75)
x2 <- data.frame(waiting = 80)
predict(fit, x1, interval = c("confidence"), level = 0.95)
```

```
##       fit      lwr      upr
## 1 3.79808 3.736159 3.860002
```

```
predict(fit, x2, interval = c("confidence"), level = 0.95)
```

```
##       fit      lwr      upr
## 1 4.17622 4.104848 4.247592
```

```
mean(faithful$waiting)
```

```
## [1] 70.89706
```

The interval for 80 minutes is wider because it is further away from the center.

(e) Use a 95% prediction interval to predict the eruption duration for waiting times of 75 and 100 minutes.

```
x1 <- data.frame(waiting = 75)
x2 <- data.frame(waiting = 100)
predict(fit, x1, interval = c("prediction"), level = 0.95)
```

```
##       fit      lwr      upr
## 1 3.79808 2.818592 4.777569
```

```
predict(fit, x2, interval = c("prediction"), level = 0.95)
```

```
##        fit      lwr      upr
## 1 5.688779 4.701239 6.676319
```

(f) Create a scatterplot of the data. Add the regression line, 95% confidence bands, and 95% prediction bands.

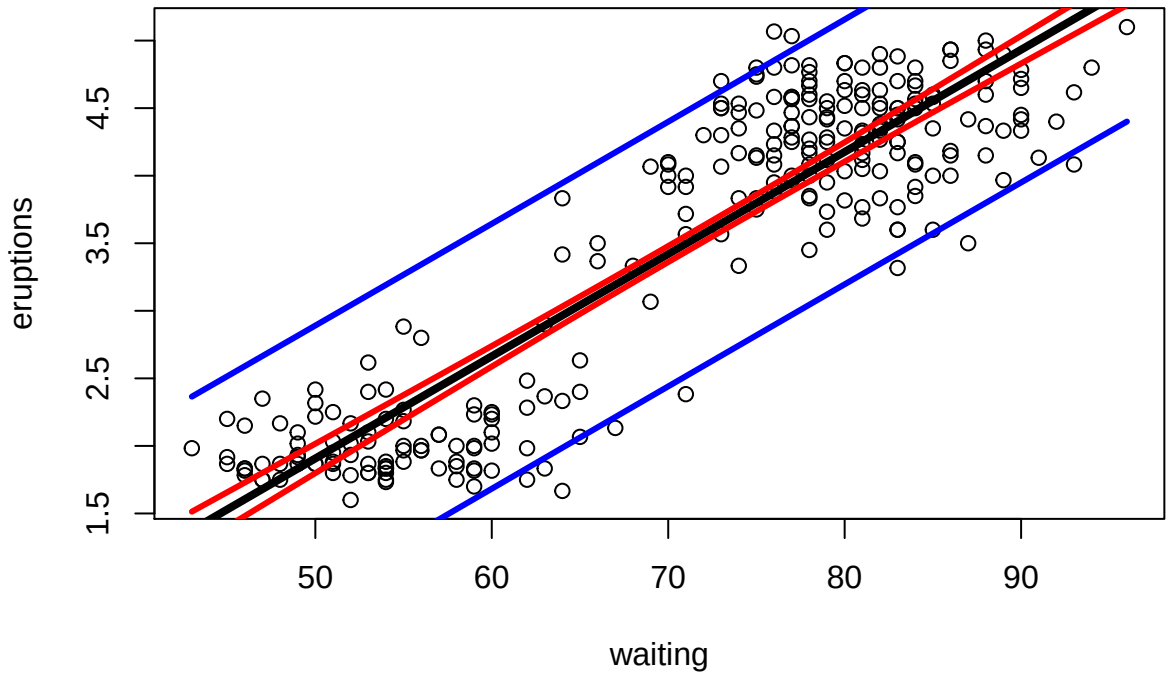
```r
x <- seq(min(faithful$waiting), max(faithful$waiting), by = 0.01)
ci <- predict(fit, newdata = data.frame(waiting = x), interval = "confidence")
pi <- predict(fit, newdata = data.frame(waiting = x), interval = "prediction")

plot(eruptions ~ waiting, data = faithful, xlab = "waiting", ylab = "eruptions")

# regression line
abline(fit, lwd = 4)

# confidence bands
lines(x, ci[,2], col = "red", lwd = 3)
lines(x, ci[,3], col = "red", lwd = 3)

# prediction bands
lines(x, pi[,2], col = "blue", lwd = 3)
lines(x, pi[,3], col = "blue", lwd = 3)
```



Exercise 2 (Using `lm` for Inference)

For this exercise we will again use the `diabetes` dataset, which can be found in the `faraway` package.

(a) Fit the following simple linear regression model in `R`. Use the total cholesterol as the response and weight as the predictor.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Store the results in a variable called `cholesterol_model`. Use an F test to test the significance of the regression. Report the following:

- The null and alternative hypotheses
- The ANOVA table (You may use `anova()` and omit the row for Total.)
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.05$
- A conclusion in the context of the problem

When reporting these, you should explicitly state them in your document, not assume that a reader will find and interpret them from a large block of `R` output.

```
library(faraway)
data('diabetes')
fit <- lm(chol ~ weight, data = diabetes)
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: chol
##            Df Sum Sq Mean Sq F value Pr(>F)
## weight      1   3505  3505.4  1.7932 0.1813
## Residuals 399 779991  1954.9
```

```
summary(fit)$fstatistic[1]
```

```
##    value
## 1.793178
```

```
summary(fit)$coefficients[2, 4]
```

```
## [1] 0.1813018
```

- The null hypotheses is $\beta_1 = 0$ and alternative hypotheses is $\beta_1 \neq 0$.
- The value of the test statistic is 1.7932.
- The anova table

```
knitr::kable(anova(fit))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
weight	1	3505.42	3505.420	1.793178	0.1813018
Residuals	399	779991.18	1954.865	NA	NA

- The p-value of the test is 0.1813018.
- The statistical decision at $\alpha = 0.05$ is fail to reject null hypothesis.
- The conclusion in the context of the problem is that the weight is not a significant predictor for cholesterol, there is no linear relationship bwtween cholesterol and weight.

(b) Fit the following simple linear regression model in `R`. Use HDL as the response and weight as the predictor.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Store the results in a variable called `hdl_model`. Use an F test to test the significance of the regression. Report the following:

- The null and alternative hypotheses
- The ANOVA table (You may use `anova()` and omit the row for Total.)

- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.05$
- A conclusion in the context of the problem

When reporting these, you should explicitly state them in your document, not assume that a reader will find and interpret them from a large block of `R` output.

```
fit <- lm(hdl ~ weight, data = diabetes)
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: hdl
##            Df Sum Sq Mean Sq F value    Pr(>F)
## weight      1  10100 10100.2  36.909 2.891e-09 ***
## Residuals 399 109188   273.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(fit)$fstatistic[1]
```

```
##    value
## 36.90875
```

```
summary(fit)$coefficients[2, 4]
```

```
## [1] 2.890526e-09
```

- The null hypotheses is $\beta_1 = 0$ and alternative hypotheses is $\beta_1 \neq 0$.
- The value of the test statistic is 36.90875.
- The anova table

```
knitr::kable(anova(fit))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
weight	1	10100.22	10100.2217	36.90875	0
Residuals	399	109187.89	273.6539	NA	NA

- The p-value of the test is 2.890526e-09.
- The statistical decision at $\alpha = 0.05$ is reject null hypothesis.
- The conclusion in the context of the problem is that the weight is a significant predictor for HDL, there is a linear relationship bwtween HDL and weight.

Exercise 3 (Inference "without" `lm`)

For this exercise we will once again use the data stored in `goalies.csv`. It contains career data for all 716 players in the history of the National Hockey League to play goaltender through the 2014-2015 season. The two variables we are interested in are:

- `W` - Wins
- `MIN` - Minutes

Fit a SLR model with `W` as the response and `MIN` as the predictor. Test $H_0 : \beta_1 = 0.008$ vs $H_1 : \beta_1 < 0.008$ at $\alpha = 0.01$. Report the following:

- $\hat{\beta}_1$
- $SE[\hat{\beta}_1]$

- The value of the t test statistic
- The degrees of freedom
- The p-value of the test
- A statistical decision at $\alpha = 0.01$

When reporting these, you should explicitly state them in your document, not assume that a reader will find and interpret them from a large block of `R` output.

You should use `lm()` to fit the model and obtain the estimate and standard error. But then you should directly calculate the remaining values. Hint: be careful with the degrees of freedom. Think about how many observations are being used.

```
goalies <- read.csv("goalies.csv")
fit <- lm(W ~ MIN, data = goalies)
summary(fit)
```

```
##
## Call:
## lm(formula = W ~ MIN, data = goalies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -81.814  -4.447   2.691   4.316 111.354
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.402e+00  7.698e-01  -5.719 1.58e-08 ***
## MIN          7.846e-03  5.071e-05 154.724  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.78 on 711 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.9712, Adjusted R-squared:  0.9711
## F-statistic: 2.394e+04 on 1 and 711 DF,  p-value: < 2.2e-16
```

```
df <- length(fitted(fit)) - 2
beta_hat <- summary(fit)$coef[2,1]
beta_0 <- 0.008
se <- summary(fit)$coef[2,2]
t <- (beta_hat-beta_0)/se
p <- pt(t, df)

beta_hat
```

```
## [1] 0.007845997
```

```
se
```

```
## [1] 5.070963e-05
```

```
t
```

```
## [1] -3.036956
```

```
df
```

```
## [1] 711
```

```
p
```

```
## [1] 0.0012386
```

- $-\ \hat{\beta}_1 = 0.007845997$.
- $SE[\hat{\beta}_1] = 5.070963e-05$.
- The value of the t test statistic is -3.036956.
- The degrees of freedom is 711.
- The p-value of the test is 0.0012386.
- The statistical decision at $\alpha = 0.01$ is reject the null hypothesis.

Exercise 4 (Simulating Sampling Distributions)

For this exercise we will simulate data from the following model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Where $\epsilon_i \sim N(0, \sigma^2)$. Also, the parameters are known to be:

- $\beta_0 = 4$
- $\beta_1 = 0.5$
- $\sigma^2 = 25$

We will use samples of size $n = 50$.

(a) Simulate this model 1500 times. Each time use `lm()` to fit a SLR model, then store the value of $\hat{\beta}_0$ and $\hat{\beta}_1$. Set a seed using **your** UIN before performing the simulation. Note, we are simulating the x values once, and then they remain fixed for the remainder of the exercise.

```
uin = 677405631
set.seed(uin)
n = 50
x = seq(0, 20, length = n)


beta_0 <- 4
beta_1 <- 0.5
sigma_square <- 25

m <- 1500
beta_0_hat <- numeric(m)
beta_1_hat <- numeric(m)

for (i in 1:m) {
  y <- beta_0 + beta_1 * x + rnorm(n, 0, sqrt(sigma_square))
  beta_0_hat[i] <- summary(lm(y ~ x))$coef[1, 1]
  beta_1_hat[i] <- summary(lm(y ~ x))$coef[2, 1]
}
```

(b) For the *known* values of x, what is the expected value of $\hat{\beta}_1$?

The expected value of $\hat{\beta}_1$ is 0.5.

(c) For the known values of x, what is the standard deviation of $\hat{\beta}_1$?

```r
Sxx <- sum((x - mean(x))^2)
sqrt(sigma_square/Sxx)
```

```
## [1] 0.120049
```

The standard deviation of $\hat{\beta}_1$ is 0.120049.

(d) What is the mean of your simulated values of $\hat{\beta}_1$? Does this make sense given your answer in **(b)**?

```r
mean(beta_1_hat)
```

```
## [1] 0.4982479
```

The mean of your simulated values of $\hat{\beta}_1$ is 0.4982479. It makes sense because these two values are very close.

(e) What is the standard deviation of your simulated values of $\hat{\beta}_1$? Does this make sense given your answer in **(c)**?

```r
sd(beta_1_hat)
```

```
## [1] 0.1218322
```

The standard deviation of your simulated values of $\hat{\beta}_1$ is 0.1218322. It makes sense because these two values are very close.

(f) For the known values of x, what is the expected value of $\hat{\beta}_0$?

The expected value of $\hat{\beta}_1$ is 4.

(g) For the known values of x, what is the standard deviation of $\hat{\beta}_0$?

```r
sqrt(sigma_square * (1/n + mean(x)^2 / Sxx))
```

```
## [1] 1.393261
```

The standard deviation of $\hat{\beta}_1$ is 1.393261.

(h) What is the mean of your simulated values of $\hat{\beta}_0$? Does this make sense given your answer in **(f)**?

```r
mean(beta_0_hat)
```

```
## [1] 4.019167
```

The mean of your simulated values of $\hat{\beta}_0$ is 4.019167. It makes sense because these two values are very close.

(i) What is the standard deviation of your simulated values of $\hat{\beta}_0$? Does this make sense given your answer in **(g)**?

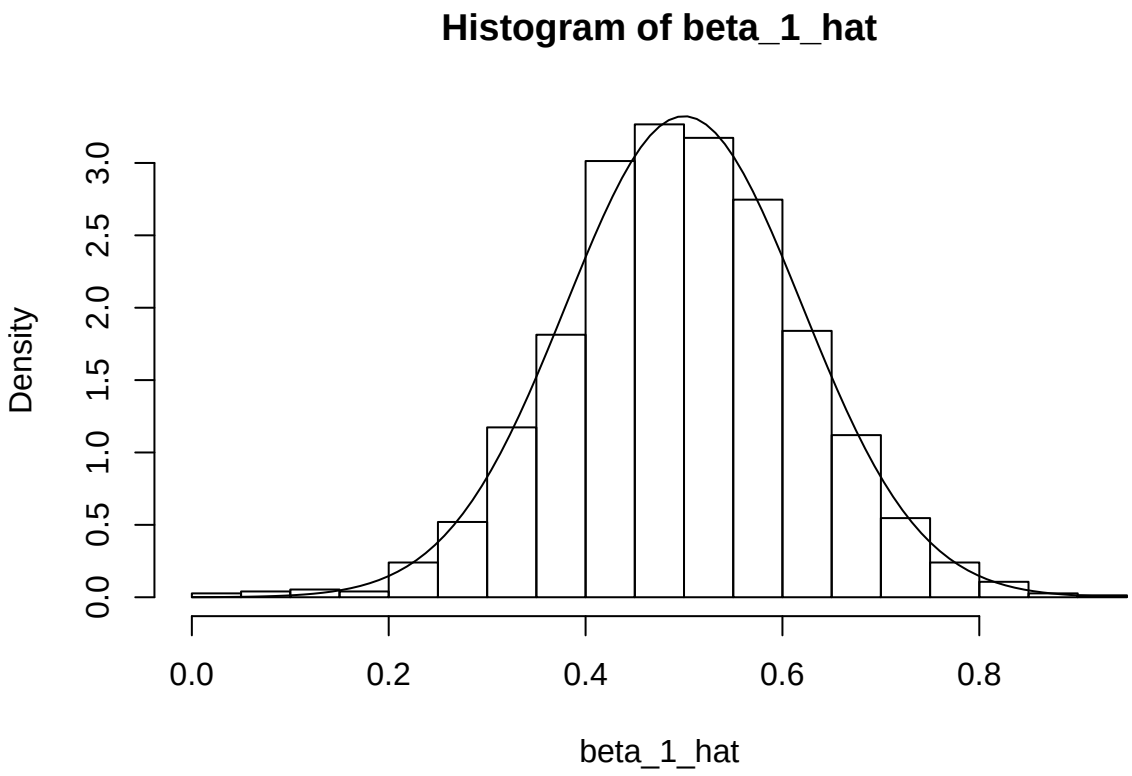
```r
sd(beta_0_hat)
```

```
## [1] 1.39312
```

The standard deviation of your simulated values of $\hat{\beta}_0$ is 1.39312. It makes sense because these two values are very close.

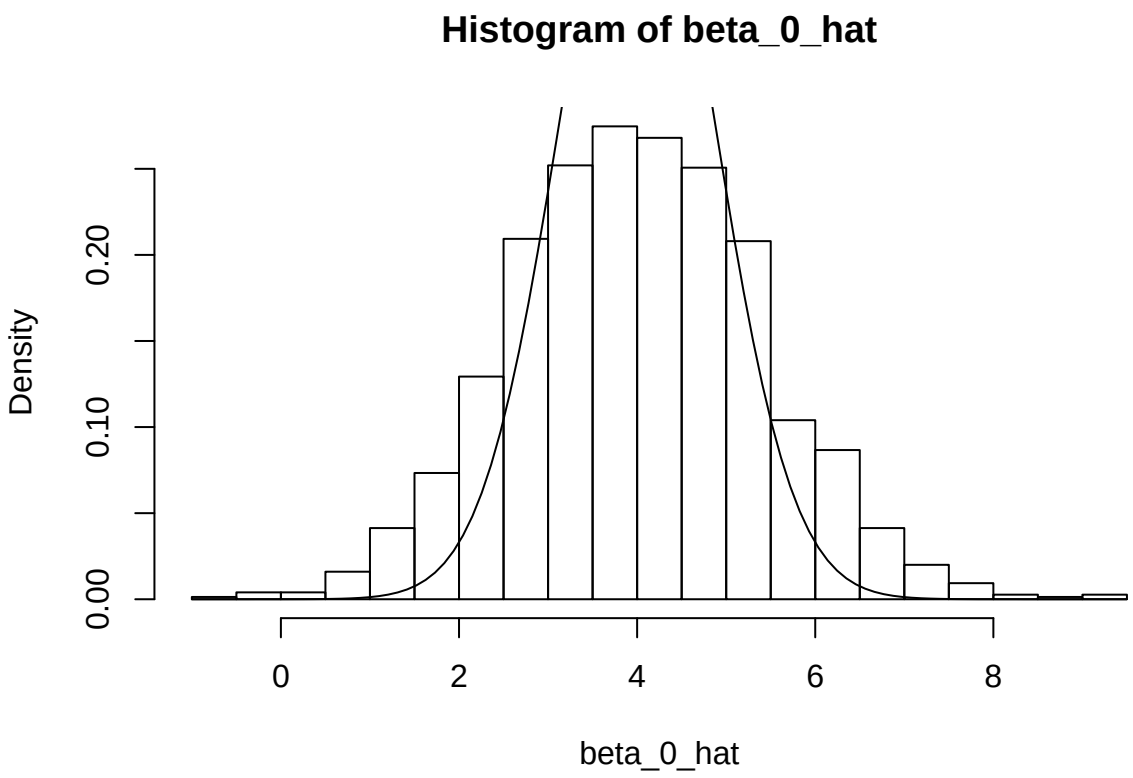
(j) Plot a histogram of your simulated values for $\hat{\beta}_1$. Add the normal curve for the true sampling distribution of $\hat{\beta}_1$.

```r
hist(beta_1_hat, breaks = 20, prob = TRUE)
curve(dnorm(x, beta_1, sqrt(sigma_square/Sxx)), add = TRUE)
```



(k) Plot a histogram of your simulated values for $\hat{\beta}_0$. Add the normal curve for the true sampling distribution of $\hat{\beta}_0$.

```
hist(beta_0_hat, breaks = 20, prob = TRUE)
curve(dnorm(x, beta_0, sqrt(sigma_square * (1/n + mean(x)^2 / Sxx))), add = TRUE)
```



Exercise 5 (Simulating Confidence Intervals)

For this exercise we will simulate data from the following model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Where $\epsilon_i \sim N(0, \sigma^2)$. Also, the parameters are known to be:

- $\beta_0 = 1$
- $\beta_1 = 3$
- $\sigma^2 = 16$

We will use samples of size $n = 20$.

Our goal here is to use simulation to verify that the confidence intervals really do have their stated confidence level.

(a) Simulate this model 2000 times. Each time use `lm()` to fit a SLR model, then store the value of $\hat{\beta}_0$ and s_e. Set a seed using **your** UIN before performing the simulation. Note, we are simulating the x values once, and then they remain fixed for the remainder of the exercise.

```
uin = 677405631
set.seed(uin)
n = 20
x = seq(-5, 5, length = n)

beta_0 <- 1
beta_1 <- 3
```

```
sigma_square <- 16

m <- 2000
beta_0_hat <- numeric(m)
se <- numeric(m)

for (i in 1:m) {
  y <- beta_0 + beta_1 * x + rnorm(n, 0, sqrt(sigma_square))
  beta_0_hat[i] <- summary(lm(y ~ x))$coef[1, 1]
  se[i] <- summary(lm(y ~ x))$sigma
}
```

(b) For each of the $\hat{\beta}_0$ that you simulated calculate a 90% confidence interval. Store the lower limits in a vector `lower_90` and the upper limits in a vector `upper_90`. Some hints:

- You will need to use `qt()` to calculate the critical value, which will be the same for each interval.
- Remember that `x` is fixed, so S_{xx} will be the same for each interval.
- You could, but do not need to write a `for` loop. Remember vectorized operations.

```
t <- abs(qt(0.1/2, n-2))
Sxx <- sum((x - mean(x))^2)

lower_90 <- beta_0_hat - t * se * sqrt(1/n + mean(x)^2 / Sxx)
upper_90 <- beta_0_hat + t * se * sqrt(1/n + mean(x)^2 / Sxx)
```

(c) What proportion of these intervals contain the true value of β_0?

```
mean(lower_90 < 1 & 1 < upper_90)
```

```
## [1] 0.9045
```

(d) Based on these intervals, what proportion of the simulations would reject the test $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ at $\alpha = 0.10$?

```
1 - mean(lower_90 < 0 & 0 < upper_90)
```

```
## [1] 0.2825
```

(e) For each of the $\hat{\beta}_0$ that you simulated calculate a 99% confidence interval. Store the lower limits in a vector `lower_99` and the upper limits in a vector `upper_99`.

```
t <- abs(qt(0.01/2, n-2))
Sxx <- sum((x - mean(x))^2)

lower_99 <- beta_0_hat - t * se * sqrt(1/n + mean(x)^2 / Sxx)
upper_99 <- beta_0_hat + t * se * sqrt(1/n + mean(x)^2 / Sxx)
```

(f) What proportion of these intervals contain the true value of β_0?

```
mean(lower_99 < 1 & 1 < upper_99)
```

```
## [1] 0.9945
```

(g) Based on these intervals, what proportion of the simulations would reject the test $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$ at $\alpha = 0.01$?

```
1 - mean(lower_99 < 0 & 0 < upper_99)
```

```
## [1] 0.0625
```