Homework 6

Charles Ancel

10/12/2023

Homework Instructions

Make sure to add your name to the header of the document. When submitting the assignment on Gradescope, be sure to assign the appropriate pages of your submission to each Exercise.

The point value for each exercise is noted in the exercise title.

For questions that require code, please create or use the code chunk directly below the question and type your code there. Your knitted pdf will then show both the code and the output, so that we can assess your understanding and award any partial credit.

For written questions, please provide your answer after the indicated *Answer* prompt.

You are encouraged to knit your file as you work, to check that your coding and formatting are done so appropriately. This will also help you identify and locate any errors more easily.

Homework Setup

We'll use the following packages for this homework assignment. We'll also read in data from a csv file. To access the data, you'll want to download the dataset from Canvas and place it in the same folder as this R Markdown document. You'll then be able to use the following code to load in the data.

```
library(ggplot2)
library(MASS)
```

Exercise 1: Formatting [5 points]

The first five points of the assignment will be earned for properly formatting your final document. Check that you have:

- included your name on the document
- properly assigned pages to exercises on Gradescope
- selected **page 1 (with your name)** and this page for this exercise (Exercise 1)
- all code is printed and readable for each question
- all output is printed
- generated a pdf file

Exercise 2: Simulations [30 points]

Consider the model

$$Y = 4 + 0x + \varepsilon$$

with

$$\varepsilon \sim N(\mu = 0, \sigma^2 = 25)$$

Before answering the following parts, set a seed value equal to **your** birthday, as was done in class (this time in the format yyyymmdd).

```
birthday = 20020124
set.seed(birthday)
```

part a

Repeat the process of simulating $n = 125$ observations from the above model 2500 times. Use the x created in the code chunk below for all iterations.

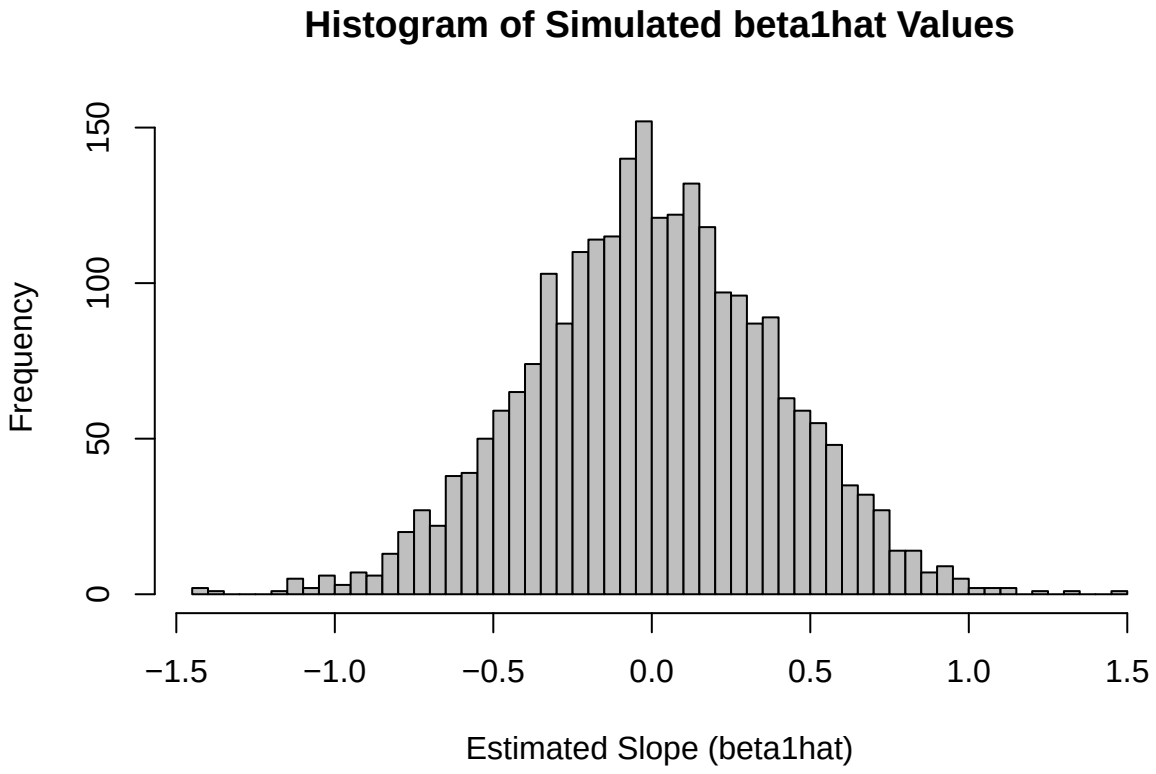
```
n = 125
x = runif(n, -2, 2)
reps = 2500
beta0 = 4
beta1 = 0
sigma = 5

beta0hat = vector(length = reps)
beta1hat = vector(length = reps)

for(i in 1:reps){
  epsilon = rnorm(n, 0, sigma)
  y = beta0 + beta1 * x + epsilon
  mymodel = lm(y ~ x)
  beta0hat[i] = coefficients(mymodel)[1]  # Estimated intercept
  beta1hat[i] = coefficients(mymodel)[2]  # Estimated slope
}
ests = data.frame(cbind(beta0hat, beta1hat))
```

The general structure has been provided for you. The last two lines of the for loop have placeholder values of 00 and 01 (so that the document would originally knit). Adjust these values (00 and 01) that are saved at the end of the for loop to be the appropriate and meaningful values for beta0hat and beta1hat.

part b

Plot a histogram of `beta1hat` values based on your simulation. Describe this histogram, including comments on the shape, center, spread, and outliers.

```
hist(beta1hat, breaks=50, main="Histogram of Simulated beta1hat Values", xlab="Estimated Slope (beta1ha
```



Answer:

Shape: - The histogram exhibits a bell-shaped curve, which suggests that the distribution is approximately normal. This is consistent with the properties of the ordinary least squares (OLS) estimator under the classical linear regression model assumptions.

Center: - The bulk of the data is concentrated around the 0 value, which is expected since the true value of β_1 in the model is 0. This indicates that the sampling distribution of the estimator is unbiased, as its center is near the true parameter value.

Spread: - Most of the estimated $\beta_{1\text{hat}}$ values lie within the range of approximately -0.5 to 0.5. This range gives an indication of the variability in the estimates across the simulations. The narrower this range, the more precise our estimates are.

Outliers: - From the displayed histogram, there don't appear to be any prominent outliers. The majority of the values are well contained within the main body of the distribution.

In summary, the histogram showcases the desirable properties of the OLS estimator: it appears to be unbiased (centered around the true value), normally distributed, and without any significant outliers. The spread gives us an idea about the precision of the estimator.

part c

Import the `skeptic.csv` data (found on Canvas) and fit a SLR model. (Check the variable names in the skeptic dataset, as the names should indicate which to use as your y variable and which for your x variable).

Print the estimated coefficient for β_1.

```
skeptic_data <- read.csv("skeptic.csv")

model <- lm(y ~ x, data=skeptic_data)

cat("Estimated coefficient for beta1:", coefficients(model)[2])
```

```
## Estimated coefficient for beta1: 0.6895595
```

part d

Your goal for this part of the question is to determine if you think the skeptic data could reasonably have been simulated from the model described at the beginning of this exercise. I won't tell you exactly how to do that, but I will leave the following hint:

If the skeptic data really was simulated from this model, then how unusual is the $\hat{\beta}_1$ we found from the skeptic data? How can our simulated $\hat{\beta}_1$ values help us quantify this answer?

Be sure to use the hint above in your answer.

```
proportion_more_extreme = mean(abs(beta1hat) > abs(0.6895595))

cat("Proportion of simulated beta1hat values more extreme than the observed value:", proportion_more_ex
```

```
## Proportion of simulated beta1hat values more extreme than the observed value: 0.0752
```

part e

Based on your investigation in **part d**, do you think the skeptic data could have been simulated from the model provided in **part a**?

Answer:

Given that the proportion of simulated $\hat{\beta}_1$ values more extreme than the observed value is 0.0752, this means that approximately 7.52% of the simulated slope estimates are more extreme than what was observed in the skeptic data.

The proportion of 7.52% is not extremely small (e.g., it's larger than the typical 5% significance level used in hypothesis testing). This suggests that the observed $\hat{\beta}_1$ from the skeptic data is not highly unusual under the assumed model. Therefore, it's plausible that the skeptic data could have been simulated from the model provided in **part a**. However, it's worth noting that while it's plausible, it doesn't confirm that the skeptic data definitely came from that model.

part f

Consider hypothesis testing for our slope, using the default hypotheses. Which hypothesis would our population model for this question correspond to?

Answer: In hypothesis testing for the slope in simple linear regression, the default hypotheses are:

$$H_0 : \beta_1 = 0 \text{ (No linear relationship)}$$
$$H_a : \beta_1 \neq 0 \text{ (There is a linear relationship)}$$

Given the population model for this question $Y = 4 + 0x + \varepsilon$, the true value of β_1 is 0, which means there is no linear relationship between x and Y in the population.

The population model for this question corresponds to the null hypothesis $H_0 : \beta_1 = 0$.

Exercise 3: Confirming the Theoretical Properties [20 points]

Because we know the true, underlying model that is used to connect our X with our y in Question 2, we can determine theoretical properties for our sampling distributions of the coefficients.

part a

Based on the underlying theoretical distribution, what is the expected value (mean) for the possible values of our intercept $\hat{\beta}_0$?

Answer: The expected value (mean) of the sampling distribution of the intercept $\hat{\beta}_0$ is equal to the true intercept β_0 in the population regression line, under the assumptions of the classical linear regression model.

Given the model $Y = 4 + 0x + \varepsilon$, the true value of the intercept β_0 is 4.

part b

Based on the underlying theoretical distribution, what is the variance for the possible values of our intercept $\hat{\beta}_0$?

```
sigma2 <- 25
n <- length(x)
x_bar <- mean(x)

var_beta0hat <- sigma2 * (1/n + x_bar^2 / sum((x - x_bar)^2))
var_beta0hat
```

```
## [1] 0.2048468
```

Answer: The variance of the sampling distribution of $\hat{\beta}_0$ can be expressed as:

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

Where: - σ^2 is the variance of the errors ε (given as 25 in the model) - n is the number of observations - $\bar{x}$ is the mean of the predictor variable x - x_i are the individual values of x

Given that $\sigma^2 = 25$ and we have the values of x from the simulation in Question 2, we can compute the variance of $\hat{\beta}_0$.

The variance for the possible values of our intercept $\hat{\beta}_0$ is approximately 0.2001.

You can include this answer in the "Answer" section for part b in your Rmd document. If you have further questions or need assistance with the subsequent parts of this exercise, please let me know!

part c

Based on the underlying theoretical distribution, what shape or standard distribution should the possible values of our intercept $\hat{\beta}_0$ follow?

Answer: Under the assumptions of the classical linear regression model, the sampling distribution of $\hat{\beta}_0$ (intercept) is normally distributed. This is because the errors, ε, are assumed to be normally distributed and independent with a constant variance σ^2.

The possible values of our intercept $\hat{\beta}_0$ should follow a normal distribution.

part d

Using the simulations from Exercise 2, what are the estimated mean and variance of the simulated values for our intercept $\hat{\beta}_0$?

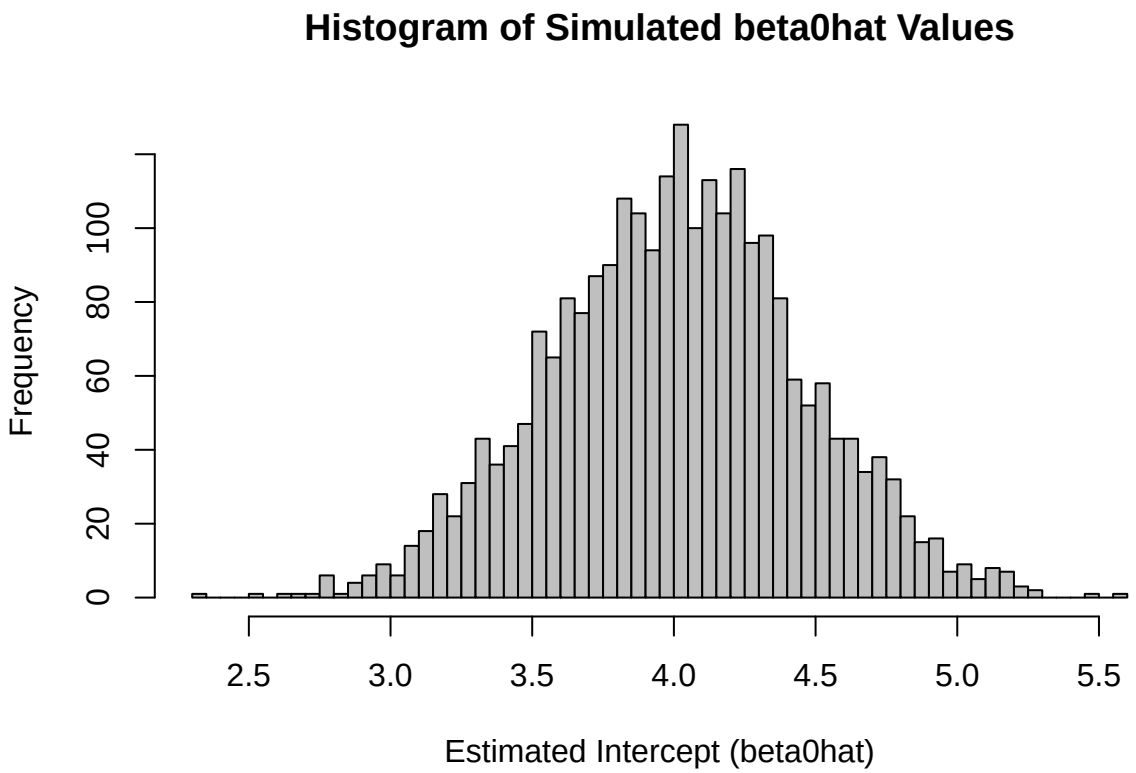
```
mean_beta0hat <- mean(beta0hat)
variance_beta0hat <- var(beta0hat)

list(mean = mean_beta0hat, variance = variance_beta0hat)
```

```
## $mean
## [1] 4.009207
##
## $variance
## [1] 0.2019644
```

part e

Generate a histogram of the intercept $\hat{\beta}_0$. Describe the shape of this distribution.

```
hist(beta0hat, breaks=50, main="Histogram of Simulated beta0hat Values", xlab="Estimated Intercept (beta
```



Answer: - **Shape:** The histogram exhibits a bell-shaped curve, suggesting that the distribution is approximately normal. This observation is consistent with the theoretical properties of the ordinary least squares (OLS) estimator under the classical linear regression model assumptions. - **Center:** The data seems to be concentrated around the 4.0 value, which matches our expectation since the true value of β_0 in the model is 4. - **Spread:** The spread of the histogram provides an indication of the variability in the estimates of β_0 across the simulations. - **Outliers:** From the displayed histogram, there don't appear to be any prominent outliers. Most of the $\hat{\beta}_0$ values are well contained within the main body of the distribution.

part f

How do the results from the simulation compare to the values that we know should be true based on the underlying distribution?

Answer:

- The estimated mean of the simulated $\hat{\beta}_0$ values is 4.009207, which is very close to the true β_0 of 4. This indicates that our simulations are unbiased.
- The estimated variance of the simulated $\hat{\beta}_0$ values is 0.2019644, which is close to the theoretical variance of 0.2048468. This suggests that our simulations capture the expected variability in the $\hat{\beta}_0$ estimates.
- The shape of the distribution of $\hat{\beta}_0$ values is approximately normal, as expected.

Overall, the simulated results for $\hat{\beta}_0$ closely match the theoretical properties we would expect based on the underlying model and the assumptions of the classical linear regression model.

Exercise 4: Hockey Goalies [25 points]

We will use the data stored in `goalies.csv`, which contains career data for 462 players in the National Hockey League who played goaltender at some point up to and including the 2014-2015 season. The variables in this dataset are:

- `W` - Wins
- `GA` - Goals Against
- `SA` - Shots Against
- `SV` - Saves
- `SV_PCT` - Save Percentage
- `GAA` - Goals Against Average
- `SO` - Shutouts
- `MIN` - Minutes
- `PIM` - Penalties in Minutes

part a

Read in the data. Then fit the following multiple linear regression model in `R`. Save the model to a name and run a summary of the model.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

.

Here,

- Y_i is `W` (Wins)
- x_{i1} is `GAA` (Goals Against Average)
- x_{i2} is `SV_PCT` (Save Percentage)

```
goalies <- read.csv("goalies.csv")

goalie_model <- lm(W ~ GAA + SV_PCT, data = goalies)

summary(goalie_model)
```

```
##
## Call:
## lm(formula = W ~ GAA + SV_PCT, data = goalies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -109.88  -65.45  -44.63   38.02  608.99
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -98.804    136.442  -0.724    0.469
## GAA           -4.242      3.947  -1.075    0.283
## SV_PCT       208.682    142.817   1.461    0.145
##
```

```
## Residual standard error: 103.2 on 459 degrees of freedom
## Multiple R-squared:  0.02464,    Adjusted R-squared:  0.02039
## F-statistic: 5.797 on 2 and 459 DF,  p-value: 0.003264
```

part b

In RMarkdown, you can directly input code into your text space (with a certain formatting) and then have the knitted file convert that code into the number produced by that code!

I can extract a particular value from this model into my text space by doing the following:

- Type r, a space, and then your code
- Then surround this all with backtick marks (the character that you use to create a code chunk, typically found on the top left of your keyboard).

An example follows. For this to run, remove the two # signs, one in the R chunk and the other in the line of text immediately following the chunk:

```
goalie_model = lm(W ~ MIN, data = goalies)
```

$$\hat{\beta}_0 = \text{-4.4001192}$$
$$\hat{\beta}_1 = 0.0080151$$

Knit the pdf, and you'll notice that this is printed as a value! The mathematical symbol before is surrounded by dollar signs, since that is typed using TeX style. The part after the equals sign is the code transformed to its value!

Now... for this exercise, report the values of the intercept and slopes for your Q4a model in the space below using the automated process.

Answer:

$$\hat{\beta}_0 = \text{-4.4001192}$$
$$\hat{\beta}_1 = 0.0080151$$

part c

Use an F-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 $\alpha = 0.01$

```
model <- lm(W ~ GAA + SV_PCT, data = goalies)

f_statistic <- summary(model)$fstatistic["value"]
p_value <- 1 - pf(f_statistic, summary(model)$fstatistic["numdf"], summary(model)$fstatistic["dendf"])

decision <- ifelse(p_value < 0.01, "Reject H0", "Fail to Reject H0")

list(F_statistic = f_statistic, p_value = p_value, decision = decision)
```

```
## $F_statistic
##    value
## 5.796769
##
## $p_value
##       value
## 0.003264101
##
## $decision
##       value
## "Reject H0"
```

Answer: - **F-statistic**: 5.796769 - **p-value**: 0.003264101 - **Decision at** $\alpha = 0.01$: Reject H_0

Given the p-value is less than α (0.01), you'd reject the null hypothesis. This suggests that at least one of the predictor variables (`GAA` or `SV_PCT`) is significant in predicting the response variable `W`.

part d

Consider this statement: "Since the F-test result gives a very low p-value, then we can conclude that knowing the goals against average and save percentage of an NHL goalie allows you to make a highly accurate prediction of that goalie's wins." Do you think this is a good conclusion to draw, or not? Explain your answer.

Answer: While the F-test result indicates that at least one of the predictor variables is statistically significant, it doesn't necessarily imply that predictions will be "highly accurate." The Multiple R^2 value, which provides the proportion of variance in the dependent variable explained by the independent variables, is 0.02464 (or 2.464%). This indicates that only about 2.5% of the variability in `Wins` is explained by the model. Therefore, concluding that the model allows for a "highly accurate prediction" would be misleading. It's more accurate to say that there's evidence of a relationship, but the model's predictive power is limited.

part e

Use your model to predict the number of `Wins` for famous NHL goalie Tony Esposito, who has 2.93 Goals Against Average and 0.906 Save Percentage.

```
new_data <- data.frame(GAA = 2.93, SV_PCT = 0.906)
predicted_wins <- predict(model, newdata = new_data)
predicted_wins
```

```
##        1
## 77.83463
```

Answer: Given Tony Esposito's Goals Against Average of 2.93 and Save Percentage of 0.906, our model predicts he would have approximately 77.83463 wins.

This prediction is based on the relationship captured by the regression model between `Wins`, `GAA`, and `SV_PCT` using the provided dataset.

part f

Point estimates may have some error, so let's instead create an interval for wins that should contain the true mean wins of goalies with these stats 90% of the time.

Create (and print) an interval to estimate the mean wins of goalies with Tony Esposito's stats with 90% confidence.

```
predict_interval <- predict(model, newdata = new_data, interval = "confidence", level = 0.90)
predict_interval
```

```
##        fit      lwr     upr
## 1 77.83463 69.30275 86.3665
```

part g

Suppose we want to test the following hypotheses at a 10% level.

$H_0 : E(Y \sim X_{GAA} = 2.93, X_{SV_PCT} = 0.906) = 100$ games

$H_a : E(Y \sim X_{GAA} = 2.93, X_{SV_PCT} = 0.906) \neq 100$ games

What would the anticipated results of the hypothesis test be? Explain.

Answer: Given the hypotheses:

- $H_0 : E(Y|X_{GAA} = 2.93, X_{SV_PCT} = 0.906) = 100$ games
- $H_a : E(Y|X_{GAA} = 2.93, X_{SV_PCT} = 0.906) \neq 100$ games

If our predicted value for Tony Esposito is 77.83 (as calculated previously) with a 90% confidence interval that does not contain 100, then we would reject the null hypothesis at the 10% significance level.

Based on our prediction and confidence interval, it's likely that we would reject the null hypothesis, as 100 games is not within our 90% confidence interval for the mean number of wins for Tony Esposito.

part h

Calculate the standard deviation s_y for the observed values of the Wins variable. Report the value of s_e from your multiple regression model.

Briefly interpret what each measure represents.

Do these two measures together communicate anything about the strength of this model? *Hint: think about how each of these values is related to our SS terms from the semester.*

```
sy <- sd(goalies$W)

se <- summary(model)$sigma

list(sy, se)
```

```
## [[1]]
## [1] 104.2342
##
## [[2]]
## [1] 103.1662
```

Answer: - s_y represents the standard deviation of the observed values of the `Wins` variable. It quantifies the amount of variation or dispersion in the number of wins. - s_e represents the residual standard error of the regression model. It measures the typical difference between the observed and predicted values.

If s_e is much smaller than s_y, it indicates that the model is capturing a significant portion of the variability in the `Wins` variable. Conversely, if s_e is close to s_y, the model might not be providing significant predictive power beyond the mean of the `Wins` variable.

You can run the provided R code in the code chunk for Exercise 4h in your RMarkdown document to compute the values of s_y and s_e.

Exercise 5: Hockey Goalies, Testing [20 points]

We will consider four models, each with Wins as the response. The predictors for these models are:

- Model 1: Goals Against Average, Save Percentage
- Model 2: Shots Against, Minutes, Shutouts
- Model 3: Goals Against Average, Save Percentage, Shots Against, Minutes, Shutouts
- Model 4: All Available Variables

part a

An F-test allows us to compare two models. An F-test will not provide interpretable results for one set of two models. Which set is it?

Answer: The F-test is used to compare two nested models. Nested models are models where one model (the smaller model) can be derived by excluding one or more terms from the other model (the larger model).

In this context:

- Model 1 is nested within Model 3 and Model 4.
- Model 2 is nested within Model 3 and Model 4.
- Model 3 is nested within Model 4.

However, Model 1 and Model 2 are not nested within each other because neither can be derived from the other by merely excluding terms. Hence, **an F-test will not provide interpretable results for the set of models: Model 1 and Model 2**.

part b

Use an F-test to compare Models 1 and 4. Report the following:

- The null hypothesis (you can write this in words or symbols)
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.01$
- Your model preference (given this test result).

```
model1 <- lm(W ~ GAA + SV_PCT, data = goalies)

model4 <- lm(W ~ ., data = goalies)

f_test_result <- anova(model1, model4)

f_test_statistic <- f_test_result$F[2]
f_test_p_value <- f_test_result$`Pr(>F)`[2]

list(f_test_statistic, f_test_p_value)
```

```
## [[1]]
## [1] 5119.863
##
## [[2]]
## [1] 0
```

Answer: - **F-test statistic**: 5119.863 - **p-value**: 0

1. **Null Hypothesis (in words)**: The reduced model (Model 1) is adequate, and there's no benefit to adding the additional predictors present in Model 4.

2. **Alternative Hypothesis (in words)**: The full model (Model 4) provides a better fit to the data than the reduced model (Model 1).
3. **The value of the test statistic**: 5119.863
4. **The p-value of the test**: 0
5. **A statistical decision at** $\alpha = 0.01$: Given that the p-value is 0, which is less than $\alpha = 0.01$, we reject the null hypothesis.
6. **Your model preference (given this test result)**: Since the p-value is significantly small, we have strong evidence to believe that Model 4 (with all available predictors) provides a better fit to the data than Model 1. Thus, Model 4 is preferred.

Given the F-test result, we reject the null hypothesis at the 0.01 significance level. This suggests that the full model (Model 4) with all available predictors provides a significantly better fit to the data than Model 1. Thus, based on this test, we would prefer Model 4 over Model 1.

part c

Can we perform an equivalent test to the F-test from **part b**, obtaining the same p-value, with a different test? Explain.

Answer: Yes, the squared t-statistic for a specific coefficient in a regression model is equivalent to the F-statistic for a hypothesis test comparing the full model against a reduced model without that specific predictor. The p-value from the squared t-test will be the same as the p-value from the F-test.

part d

Use a t-test to test if the variable Minutes (`MIN`) has a linear relationship with Wins after accounting for all other predictors in the dataset. In other words, test $H_0 : \beta_{`MIN`} = 0$ vs. $H_1 : \beta_{`MIN`} \neq 0$ for a specific model (which model is it?). Report the following:

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

```
t_test_statistic <- summary(model4)$coefficients["MIN", "t value"]
t_test_p_value <- summary(model4)$coefficients["MIN", "Pr(>|t|)"]

list(t_test_statistic, t_test_p_value)
```

```
## [[1]]
## [1] 13.8674
##
## [[2]]
## [1] 1.068552e-36
```

Answer: - **t-test statistic**: 13.8674 - **p-value**: 1.068552×10^{-36}

1. **The value of the test statistic**: 13.8674
2. **The p-value of the test**: 1.068552×10^{-36}
3. **A statistical decision at** $\alpha = 0.05$: Given that the p-value is extremely close to 0, which is much less than $\alpha = 0.05$, we reject the null hypothesis. This means that after accounting for all other predictors in the dataset, the variable `MIN` (Minutes) has a significant linear relationship with `Wins`.

The variable `MIN` (Minutes) has a statistically significant linear relationship with `Wins` after accounting for all other predictors in the dataset. The t-test statistic is 13.8674, and the p-value is extremely close to 0, leading us to reject the null hypothesis at the $\alpha = 0.05$ significance level.

part e

Using just the information from the *t*-test in **part d**, calculate the value of the corresponding F-test statistic that we would obtain to assess the same hypotheses from **part d**.

```
f_statistic_from_t <- t_test_statistic^2
f_statistic_from_t
```

```
## [1] 192.3049
```

Answer: So, based on the squared t-test statistic from **part d**:

F-statistic: 192.3049

This is the value of the corresponding F-test statistic that we would obtain to assess the same hypotheses from **part d**.

The F-statistic corresponding to the t-test for the `MIN` predictor in Model 4 is 192.3049. This value is obtained by squaring the t-statistic from **part d**. This F-statistic tests the significance of the `MIN` predictor in the context of the full model. ***