Assessment: Least Squares Estimates, part 1

Question 1

1/1 point (graded)

The following code was used in the video to plot RSS with $\beta_0 = 25$.

```
beta1 = seq(0, 1, len=nrow(galton_heights))
results <- data.frame(beta1 = beta1,
                      rss = sapply(beta1, rss, beta0 = 25))
results %>% ggplot(aes(beta1, rss)) + geom_line() +
  geom_line(aes(beta1, rss), col=2)
```

In a model for sons' heights vs fathers' heights, what is the least squares estimate (LSE) for β_1 if we assume $\hat{\beta}_0$ is 36?

Hint: modify the code above to do your analysis.

○ 0.65

● 0.5

○ 0.2

○ 12

✓

Answer

Correct: Correct. You can tell from a plot of RSS vs β_1 that the minimum estimate is 0.5

Explanation

Using the code from the video, you can plot RSS vs β_1 to find the value for β_1 that minimizes the RSS. In this case, that value is 0.5 when we assume that $\hat{\beta}_0$ is 36.

When we assumed that $\hat{\beta}_0$ was 25, as in the sample code, the LSE for β_1 was 0.65.

Submit

You have used 1 of 2 attempts

ℹ Answers are displayed within the problem

Question 2

1/1 point (graded)

The least squares estimates for the parameters $\beta_0, \beta_1, \ldots, \beta_n$ minimize ✓

Answer: minimize the residual sum of squares.

Explanation

The least squares estimates minimize, not maximize, the residual sum of squares.

Submit You have used 1 of 1 attempt

ℹ Answers are displayed within the problem

Question 3

1/1 point (graded)

Load the `Lahman` library and filter the `Teams` data frame to the years 1961-2001. Run a linear model in R predicting the number of runs per game based on *both* the number of bases on balls *and* the number of home runs.

What is the coefficient for bases on balls?

● 0.39

○ 1.56

○ 1.74

○ 0.027

✓

Answer

Correct: Correct.

Explanation

The coefficient for bases on balls is 0.39; the coefficient for home runs is 1.56; the intercept is 1.74; the standard error for the BB coefficient is 0.027.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Question 4

1/1 point (graded)

We run a Monte Carlo simulation where we repeatedly take samples of N = 100 from the Galton heights data and compute the regression slope coefficients for each sample:

```
B <- 1000
N <- 100
lse <- replicate(B, {
    sample_n(galton_heights, N, replace = TRUE) %>%
    lm(son ~ father, data = .) %>% .$coef
})

lse <- data.frame(beta_0 = lse[1,], beta_1 = lse[2,])
```

What does the central limit theorem tell us about the variables beta_0 and beta_1?

Select ALL that apply.

- ☑ They are approximately normally distributed.
- ☑ The expected value of each is the true value of β_0 and β_1(assuming the Galton heights data is a complete population).
- ☐ The central limit theorem does not apply in this situation.
- ☐ It allows us to test the hypothesis that $\beta_0 = 0$ and $\beta_1 = 0$.

✓

Answer

Correct:

Correct. With a large enough N, the distributions of both beta_0 and beta_1 are approximately normal.

Explanation

With a large enough N, the central limit theorem applies and tells us that the distributions of both beta_0 and beta_1 are approximately normal. The expected values of beta_0 and beta_1 are the true values of β_0 and β_1, assuming that the Galton heights data are a complete population.

For hypothesis testing, we assume that the errors in the model are normally distributed.

Submit

You have used 1 of 2 attempts

Answers are displayed within the problem

Question 5

1/1 point (graded)

In an earlier video, we ran the following linear model and looked at a summary of the results.

```
> mod <- lm(son ~ father, data = galton_heights)
> summary(mod)

Call:
lm(formula = son ~ father, data = galton_heights)

Residuals:
   Min     1Q  Median     3Q    Max
-5.902 -1.405  0.092   1.342  8.092

Coefficients:
                 Estimate  Std. Error  t value     Pr(>|t|)
(Intercept)     35.7125     4.5174       7.91    2.8e-13 ***
father           0.5028     0.0653       7.70    9.5e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What null hypothesis is the second p-value (the one in the father row) testing?

○ $\beta_1 = 1$, where β_1 is the coefficient for the variable "father."

○ $\beta_1 = 0.503$, where β_1 is the coefficient for the variable "father."

● $\beta_1 = 0$, where β_1 is the coefficient for the variable "father."

✓

Explanation

The p-value for "father" tests the null hypothesis that $\beta_1 = 0$, i.e., the fathers' heights are not associated with the sons' heights, where β_1 is the coefficient for the variable father.

Submit

You have used 1 of 1 attempt

Answers are displayed within the problem

Question 6

1/1 point (graded)

Which R code(s) below would properly plot the predictions and confidence intervals for our linear model of sons' heights?

Select ALL that apply.

☐

```
galton_heights %>% ggplot(aes(father, son)) +
    geom_point() +
    geom_smooth()
```

☑

```
galton_heights %>% ggplot(aes(father, son)) +
    geom_point() +
    geom_smooth(method = "lm")
```

☑

```
model <- lm(son ~ father, data = galton_heights)
predictions <- predict(model, interval = c("confidence"), level = 0.95)
data <- as.tibble(predictions) %>% bind_cols(father = galton_heights$father)

ggplot(data, aes(x = father, y = fit)) +
    geom_line(color = "blue", size = 1) +
    geom_ribbon(aes(ymin=lwr, ymax=upr), alpha=0.2) +
    geom_point(data = galton_heights, aes(x = father, y = son))
```

☐

```
model <- lm(son ~ father, data = galton_heights)
predictions <- predict(model)
data <- as.tibble(predictions) %>% bind_cols(father = galton_heights$father)

ggplot(data, aes(x = father, y = fit)) +
    geom_line(color = "blue", size = 1) +
    geom_point(data = galton_heights, aes(x = father, y = son))
```

✓

Answer

Correct:

Correct. This is one way to plot predictions and confidence intervals for a linear model of sons' heights vs. fathers' heights. This is one of two correct answers.

Correct. This code uses the `predict` command to generate predictions and 95% confidence intervals for the linear model of sons' heights vs. fathers' heights. This is one of two correct answers.

Explanation

If using the `geom_smooth` command, you need to specify that `method = “lm”` in your `geom_smooth` command, otherwise the smooth line is a loess smooth and not a linear model.

If using the `predict` command, you need to include the confidence intervals on your figure by first specifying that you want confidence intervals in the `predict` command, and then adding them to your figure as a `geom_ribbon`.

Submit You have used 2 of 2 attempts

ℹ Answers are displayed within the problem