Assessment: Least Squares Estimates, part 2

In Questions 7 and 8, you'll look again at female heights from **GaltonFamilies**.

Define **female_heights**, a set of mother and daughter heights sampled from **GaltonFamilies**, as follows:

```
set.seed(1989) #if you are using R 3.5 or earlier
set.seed(1989, sample.kind="Rounding") #if you are using R 3.6 or later
library(HistData)
data("GaltonFamilies")
options(digits = 3)    # report 3 significant digits
```

```
female_heights <- GaltonFamilies %>%
    filter(gender == "female") %>%
    group_by(family) %>%
    sample_n(1) %>%
    ungroup() %>%
    select(mother, childHeight) %>%
    rename(daughter = childHeight)
```

Question 7

2.0/2.0 points (graded)

Fit a linear regression model predicting the mothers' heights using daughters' heights.

What is the slope of the model?

0.31 ✔ **Answer:** 0.31

0.31

Explanation

The following code can be used to determine the slope:

```
fit <- lm(mother ~ daughter, data = female_heights)
fit$coef[2]
```

What the intercept of the model?

44.18 ✔ **Answer:** 44.2

44.18

Explanation
The following code can be used to determine the intercept:

```
fit$coef[1]
```

Submit You have used 2 of 10 attempts

Answers are displayed within the problem

Question 8

2.0/2.0 points (graded)
Predict mothers' heights using the model.

What is the predicted height of the first mother in the dataset?

65.6 ✔ **Answer:** 65.6

65.6

Explanation
The following code can be used to determine the height:

```
predict(fit)[1]
```

What is the actual height of the first mother in the dataset?

67 ✔ **Answer:** 67

67

Explanation
The following code can be used to determine the height:

```
female_heights$mother[1]
```

You have used 1 of 10 attempts

Submit

Answers are displayed within the problem

We have shown how BB and singles have similar predictive power for scoring runs. Another way to compare the usefulness of these baseball metrics is by assessing how stable they are across the years. Because we have to pick players based on their previous performances, we will prefer metrics that are more stable. In these exercises, we will compare the stability of singles and BBs.

Before we get started, we want to generate two tables: one for 2002 and another for the average of 1999-2001 seasons. We want to define per plate appearance statistics, keeping only players with more than 100 plate appearances. Here is how we create the 2002 table:

```
library(Lahman)
bat_02 <- Batting %>% filter(yearID == 2002) %>%
    mutate(pa = AB + BB, singles = (H - X2B - X3B - HR)/pa, bb = BB/pa) %>%
    filter(pa >= 100) %>%
    select(playerID, singles, bb)
```

Question 9

2.0/2.0 points (graded)

Now compute a similar table but with rates computed over 1999-2001. Keep only rows from 1999-2001 where players have 100 or more plate appearances, then calculate the average single rate (`mean_singles`) and average BB rate (`mean_bb`) per player over those three seasons.

How many players had a single rate `mean_singles` of greater than 0.2 per plate appearance over 1999-2001?

46 ✔ **Answer:** 46

46

Explanation

The following code can be used to determine the number of players:

```
bat_99_01 <- Batting %>% filter(yearID %in% 1999:2001) %>%
    mutate(pa = AB + BB, singles = (H - X2B - X3B - HR)/pa, bb = BB/pa) %>%
    filter(pa >= 100) %>%
    group_by(playerID) %>%
    summarize(mean_singles = mean(singles), mean_bb = mean(bb))
sum(bat_99_01$mean_singles > 0.2)
```

How many players had a BB rate `mean_bb` of greater than 0.2 per plate appearance over 1999-2001?

3 ✔ **Answer:** 3

3

Explanation
The following code can be used to determine the number of players:

```
sum(bat_99_01$mean_bb > 0.2)
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

Question 10

2.0/2.0 points (graded)
Use `inner_join` to combine the `bat_02` table with the table of 1999-2001 rate averages you created in the previous question.

What is the correlation between 2002 singles rates and 1999-2001 average singles rates?

0.5509222 ✔ **Answer:** 0.551

0.5509222

Explanation
The following code can be used to determine the correlation:

```
dat <- inner_join(bat_02, bat_99_01)
cor(dat$singles, dat$mean_singles)
```

What is the correlation between 2002 BB rates and 1999-2001 average BB rates?

0.7174787 ✔ **Answer:** 0.717

0.7174787

Explanation
The following code can be used to determine the correlation:

```
cor(dat$bb, dat$mean_bb)
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

Question 11

1.0/1.0 point (graded)
Make scatterplots of `mean_singles` versus `singles` and `mean_bb` versus `bb`.

Are either of these distributions bivariate normal?

- Neither distribution is bivariate normal.
- `singles` and `mean_singles` are bivariate normal, but `bb` and `mean_bb` are not.
- `bb` and `mean_bb` are bivariate normal, but `singles` and `mean_singles` are not.
- Both distributions are bivariate normal. ✔

Explanation
Both distributions are bivariate normal, as can be seen in the scatter plots made using the following code:

```
dat %>%
    ggplot(aes(singles, mean_singles)) +
    geom_point()
dat %>%
    ggplot(aes(bb, mean_bb)) +
    geom_point()
```

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Question 12

2.0/2.0 points (graded)
Fit a linear model to predict 2002 `singles` given 1999-2001 `mean_singles`.

What is the coefficient of `mean_singles`, the slope of the fit?

0.58813404 ✔ **Answer:** 0.588

0.58813404

Explanation

The linear model and slope can be generated using the following code:

```
fit_singles <- lm(singles ~ mean_singles, data = dat)
fit_singles$coef[2]
```

Fit a linear model to predict 2002 `bb` given 1999-2001 `mean_bb`.

What is the coefficient of `mean_bb`, the slope of the fit?

0.82904930 ✔ **Answer:** 0.829

0.82904930

Explanation

```
fit_bb <- lm(bb ~ mean_bb, data = dat)
fit_bb$coef[2]
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem