Assessment: Linear Models (Verified Learners only)

This assessment has 6 multi-part questions that will all use the setup below;

Game attendance in baseball varies partly as a function of how well a team is playing.

Load the **`Lahman`** library. The **`Teams`** data frame contains an **`attendance`** column. This is the total attendance for the season. To calculate average attendance, divide by the number of games played, as follows:

```
library(tidyverse)
library(broom)
library(Lahman)
Teams_small <- Teams %>%
    filter(yearID %in% 1961:2001) %>%
    mutate(avg_attendance = attendance/G)
```

Use linear models to answer the following 3-part question about **`Teams_small`**.

Question 1a

2.0/2.0 points (graded)

Use runs (`R`) per game to predict average attendance.

For every 1 run scored per game, attendance increases by how much?

4117 ✔ **Answer:** 4117

4117

Explanation

The increase in attendance can be found using the following code:

```
# find regression line predicting attendance from R and take slope
Teams_small %>%
  mutate(R_per_game = R/G) %>%
  lm(avg_attendance ~ R_per_game, data = .) %>%
  .$coef %>%
  .[2]
```

Use home runs (`HR`) per game to predict average attendance.

For every 1 home run hit per game, attendance increases by how much?

8113 ✔ **Answer:** 8113

8113

Explanation
The increase in attendance can be found using the following code:

```
Teams_small %>%
  mutate(HR_per_game = HR/G) %>%
  lm(avg_attendance ~ HR_per_game, data = .) %>%
  .$coef %>%
  .[2]
```

Submit You have used 3 of 10 attempts

Answers are displayed within the problem

Question 1b

2.0/2.0 points (graded)
Use number of wins to predict attendance; do not normalize for number of games.

For every game won in a season, how much does average attendance increase?

121.1 ✔ **Answer:** 121

121.1

Explanation
The increase in attendance can be found using the following code:

```
Teams_small %>%
  lm(avg_attendance ~ W, data = .) %>%
  .$coef %>%
  .[2]
```

Suppose a team won zero games in a season.

Predict the average attendance.

1129.22 ✔ **Answer:** 1129

1129.22

Explanation
The average attendance can be found using the following code:

```
Teams_small %>%
  lm(avg_attendance ~ W, data = .) %>%
  .$coef %>%
  .[1]
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

Question 1c

1.0/1.0 point (graded)
Use year to predict average attendance.

How much does average attendance increase each year?

244.5 ✔ **Answer:** 244

244.5

Explanation
The increase in attendance can be found using the following code:

```
Teams_small %>%
  lm(avg_attendance ~ yearID, data = .) %>%
  .$coef %>%
  .[2]
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

Question 2

2.0/2.0 points (graded)

Game wins, runs per game and home runs per game are positively correlated with attendance. We saw in the course material that runs per game and home runs per game are correlated with each other. Are wins and runs per game or wins and home runs per game correlated?

What is the correlation coefficient for wins and runs per game?

0.4116491 ✔ **Answer:** 0.412

0.4116491

Explanation

The following code will give the correlation coefficient:

```
cor(Teams_small$W, Teams_small$R/Teams_small$G)
```

What is the correlation coefficient for wins and home runs per game?

0.2744313 ✔ **Answer:** 0.274

0.2744313

Explanation

The following code will give the correlation coefficient:

```
cor(Teams_small$W, Teams_small$HR/Teams_small$G)
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

Stratify **`Teams_small`** by wins: divide number of wins by 10 and then round to the nearest integer. Keep only strata 5 through 10, which have 20 or more data points.

Use the stratified dataset to answer this three-part question.

Question 3a

1.0/1.0 point (graded)

How many observations are in the 8 win strata?
(Note that due to division and rounding, these teams have 75-84 wins.)

338 ✔ **Answer:** 338

338

Explanation

The number of observations can be found using the following code:

```
dat <- Teams_small %>%
    mutate(W_strata = round(W/10)) %>%
    filter(W_strata >= 5 & W_strata <= 10)

sum(dat$W_strata == 8)
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

Question 3b

2.0/2.0 points (graded)

Calculate the slope of the regression line predicting average attendance given runs per game for each of the win strata.

Which win stratum has the largest regression line slope?

- (●) 5
- ( ) 6
- ( ) 7
- ( ) 8
- ( ) 9

- 10

✔

Explanation
The slope can be found using the following code:

```
# calculate slope of regression line after stratifying by R per game
dat %>%
  group_by(W_strata) %>%
  summarize(slope = cor(R/G, avg_attendance)*sd(avg_attendance)/sd(R/G))
```

Calculate the slope of the regression line predicting average attendance given HR per game for each of the win strata.

Which win stratum has the largest regression line slope?

- (•) 5
- ( ) 6
- ( ) 7
- ( ) 8
- ( ) 9
- ( ) 10

✔

Explanation
The slope can be found using the following code:

```
# calculate slope of regression line after stratifying by HR per game
dat %>%
  group_by(W_strata) %>%
  summarize(slope = cor(HR/G, avg_attendance)*sd(avg_attendance)/sd(HR/G))
```

Submit You have used 1 of 2 attempts

ℹ Answers are displayed within the problem

Question 3c

1.0/1.0 point (graded)
Which of the followng are true about the effect of win strata on average attendance?
Select ALL that apply.

- [x] Across all win strata, runs per game are positively correlated with average attendance.
- [ ] Runs per game have the strongest effect on attendance when a team wins many games.
- [ ] After controlling for number of wins, home runs per game are not correlated with attendance.
- [x] Home runs per game have the strongest effect on attendance when a team does not win many games.
- [x] Among teams with similar numbers of wins, teams with more home runs per game have larger average attendance.

✔

Explanation
Looking at the data, we can see that runs per game are positively correlated with average attendance, that home runs per game have the strongest effect on attendance when teams don't win many games, and that teams with fewer wins have a larger average attendance with more home runs per game. We also see that runs per game have a stronger effect when teams win few, not many, games, and that home runs per game are in fact positively correlated with attendance in all win strata.

Submit You have used 2 of 2 attempts

ℹ Answers are displayed within the problem

Question 4

3.0/3.0 points (graded)
Fit a multivariate regression determining the effects of runs per game, home runs per game, wins, and year on average attendance. Use the original `Teams_small` wins column, not the win strata from question 3.

What is the estimate of the effect of runs per game on average attendance?

321.8 ✔ **Answer:** 322

321.8

Explanation
The estimate can be found using the following code:

```
fit <- Teams_small %>%
    mutate(R_per_game = R/G,
           HR_per_game = HR/G) %>%
    lm(avg_attendance ~ R_per_game + HR_per_game + W + yearID, data = .)
tidy(fit) %>%
    filter(term == "R_per_game") %>%
    pull(estimate)
```

What is the estimate of the effect of home runs per game on average attendance?

1798.4 ✔ **Answer:** 1798

1798.4

Explanation
The estimate can be found using the following code:

```
tidy(fit) %>%
    filter(term == "HR_per_game") %>%
    pull(estimate)
```

What is the estimate of the effect of number of wins in a season on average attendance?

116.7 ✔ **Answer:** 117

116.7

Explanation
The estimate can be found using the following code:

```
tidy(fit) %>%
    filter(term == "W") %>%
    pull(estimate)
```

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

Question 5

2.0/2.0 points (graded)

Use the multivariate regression model from Question 4. Suppose a team averaged 5 runs per game, 1.2 home runs per game, and won 80 games in a season.

What would this team's average attendance be in 2002?

16149.29 ✔ **Answer:** 16149

16149.29

Explanation

The average attendance can be found using the following code:

```
predict(fit, data.frame(R_per_game = 5, HR_per_game = 1.2, W = 80, yearID = 2002))
```

What would this team's average attendance be in 1960?

6504.751 ✔ **Answer:** 6505

6504.751

Explanation

The average attendance can be found using the following code:

```
predict(fit, data.frame(R_per_game = 5, HR_per_game = 1.2, W = 80, yearID = 1960))
```

Submit You have used 6 of 10 attempts

Answers are displayed within the problem

Question 6

1.0/1.0 point (graded)

Use your model from Question 4 to predict average attendance for teams in 2002 in the original `Teams` data frame.

What is the correlation between the predicted attendance and actual attendance?

0.5191942 ✔ **Answer:** 0.519

0.5191942

Explanation

The correlation can be found using the following code:

```
newdata <- Teams %>%
    filter(yearID == 2002) %>%
    mutate(avg_attendance = attendance/G,
           R_per_game = R/G,
           HR_per_game = HR/G)
preds <- predict(fit, newdata)
cor(preds, newdata$avg_attendance)
```

Submit

You have used 1 of 10 attempts

Answers are displayed within the problem