02 - Introduction to Linear Models Comprehension Check

Gabriele Mineo - Harvard Data Science Professional

Confounding: Are BBs More Predictive?

Question 1

Why is the number of home runs considered a confounder of the relationship between bases on balls and runs per game?

- Home runs is not a confounder of this relationship.
- Home runs are the primary cause of runs per game.
- The correlation between home runs and runs per game is stronger than the correlation between bases on balls and runs per game.
- Players who get more bases on balls also tend to have more home runs; in addition, home runs increase the points per game. [X]

Are these data considered “tidy” in R? Why or why not?

- Yes. These data are considered “tidy” because each row contains unique observations.
- Yes. These data are considered “tidy” because there are no missing data in the data frame.
- No. These data are not considered “tidy” because the variable “year” is stored in the header. [X]
- No. These data are not considered “tidy” because there are not an equal number of columns and rows.

Stratification and Multivariate Regression

Question 1

As described in the video, when we stratified our regression lines for runs per game vs. bases on balls by the number of home runs, what happened?

- The slope of runs per game vs. bases on balls within each stratum was reduced because we removed confounding by home runs. [X]
- The slope of runs per game vs. bases on balls within each stratum was reduced because there were fewer data points.
- The slope of runs per game vs. bases on balls within each stratum increased after we removed confounding by home runs.
- The slope of runs per game vs. bases on balls within each stratum stayed about the same as the original slope.

Linear Models

Question 1

We run a linear model for sons’ heights vs. fathers’ heights using the Galton height data, and get the following results:

```
> lm(son ~ father, data = galton_heights)

Call:
```

```
lm(formula = son ~ father, data = galton_heights)

Coefficients:
(Intercept)    father
    35.71       0.50
```

Interpret the numeric coefficient for “father.”

- For every inch we increase the son’s height, the predicted father’s height increases by 0.5 inches.
- For every inch we increase the father’s height, the predicted son’s height grows by 0.5 inches. [X]
- For every inch we increase the father’s height, the predicted son’s height is 0.5 times greater.

Question 2

We want the intercept term for our model to be more interpretable, so we run the same model as before but now we subtract the mean of fathers’ heights from each individual father’s height to create a new variable centered at zero.

```
galton_heights <- galton_heights %>%
    mutate(father_centered=father - mean(father))
```

We run a linear model using this centered fathers’ height variable.

```
> lm(son ~ father_centered, data = galton_heights)

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

Coefficients:
(Intercept)    father_centered
    70.45            0.50
```

Interpret the numeric coefficient for the intercept.

- The height of a son of a father of average height is 70.45 inches. [X]
- The height of a son when a father’s height is zero is 70.45 inches.
- The height of an average father is 70.45 inches.

Least Squares Estimates (LSE)

Question 1

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?

- 0.65
- 0.5 [X]
- 0.2
- 12

Question 2

The least squares estimates for the parameters $\beta_0, \beta_1 ... \beta_n$ **minimize** the residual sum of squares.

The lm Function

Question 1

Run a linear model in R predicting the number of runs per game based on the number of bases on balls and the number of home runs. Remember to first limit your data to 1961-2001.

What is the coefficient for bases on balls?

- 0.39 [X]
- 1.56
- 1.74
- 0.027

LSE are Random Variables

Question 1

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?

- They are approximately normally distributed.
- The expected value of each is the true value of $\beta_0$ 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 $\beta_0 = 1$.

Question 2

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

```
$\beta_0 $
> 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
$\beta_0 $
```

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.” [X]

Predicted Variables are Random Variables

Question 1

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

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

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

```
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)) [X]
```

~~~
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))
 ```
<br/>

## **Advanced dplyr: Tibbles**
### **Question 1**

What problem do we encounter when we try to run a linear model on our baseball data, grouping by home r

- There is not enough data in some levels to run the model.
~~~

```
- The ```lm``` function does not know how to handle grouped tibbles. [X]
- The results of the ```lm``` function cannot be put into a tidy format.

<br/>

### **Question 2**

Tibbles are similar to what other class in R?

- Vectors
- Matrices
- Data frames [X]
- Lists

<br/>

## **Tibbles: Differences from Data Frames**
### **Question 1**

What are some advantages of tibbles compared to data frames?

- Tibbles display better. [X]
- If you subset a tibble, you always get back a tibble. [X]
- Tibbles can have complex entries. [X]
- Tibbles can be grouped. [X]

<br/>

## **do**
### **Question 1**

What are two advantages of the do command, when applied to the tidyverse?

- It is faster than normal functions.
- It returns useful error messages.
- It understands grouped tibbles. [X]
- It always returns a data.frame. [X]

<br/>

### **Question 2**

You want to take the tibble ```dat```, which we’ve been using in this video, and run the linear model R

You’ve already written the function ```get_slope```, shown below.
```

get_slope <- function(data) { fit <- lm(R ~ BB, data = data) sum.fit <- summary(fit)

data.frame(slope = sum.fit*coefficients*[2, ”*Estimate*”], *se* = *sum.fit*coefficients[2, “Std. Error”], pvalue = sum.fit$coefficients[2, “Pr(>|t|)”]) }

```
What additional code could you write to accomplish your goal?
```

dat %>% group_by(HR) %>% do(get_slope)

dat %>% group_by(HR) %>% do(get_slope(.)) [X]

dat %>% group_by(HR) %>% do(slope = get_slope(.))

dat %>% do(get_slope(.))

```
<br/>

## **broom**
### **Question 1**

The output of a broom function is always what?

A data.frame [X]
A list
A vector

<br/>

### **Question 2**

You want to know whether the relationship between home runs and runs per game varies by baseball league
```

dat <- Teams %>% filter(yearID %in% 1961:2001) %>% mutate(HR = HR/G, R = R/G) %>% select(lgID, HR, BB, R)

```
What code would help you quickly answer this question?
```

dat %>% group_by(lgID) %>% do(tidy(lm(R ~ HR, data = .), conf.int = T)) %>% filter(term == "HR") [X]

dat %>% group_by(lgID) %>% do(glance(lm(R ~ HR, data = .)))

dat %>% do(tidy(lm(R ~ HR, data = .), conf.int = T)) %>% filter(term == "HR")

dat %>% group_by(lgID) %>% do(mod = lm(R ~ HR, data = .))

```
<br/>

## **Building a Better Offensive Metric for Baseball**
### **Question 1**

What is the final linear model we use to predict runs scored per game?

- ```lm(R ~ BB + HR)```
- ```lm(HR ~ BB + singles + doubles + triples)```
- ```lm(R ~ BB + singles + doubles + triples + HR) [X]```
```

```
- ```lm(R ~ singles + doubles + triples + HR)```

<br/>

### **Question 2**

We want to estimate runs per game scored by individual players, not just by teams. What summary metric c

Look at the code from the video for a hint:
```

pa_per_game <- Batting %>% filter(yearID == 2002) %>% group_by(teamID) %>% summarize(pa_per_game = sum(AB+BB)/max(G)) %>% .$pa_per_game %>% mean “‘

- `pa_per_game`: the mean number of plate appearances per team per game for each team
- `pa_per_game`: the mean number of plate appearances per game for each player
- `pa_per_game`: the number of plate appearances per team per game, averaged across all teams [X]

Question 3

Imagine you have two teams. Team A is comprised of batters who, on average, get two bases on balls, four singles, one double, and one home run. Team B is comprised of batters who, on average, get one base on balls, six singles, two doubles, and one triple.

Which team scores more runs, as predicted by our model?

- Team A
- Team B [X]
- Tie
- Impossible to know

On Base Plus Slugging (OPS)

Question 1

The on-base-percentage plus slugging percentage (OPS) metric gives the most weight to:

- Singles
- Doubles
- Triples
- Home Runs [X]

Regression Fallacy

Question 1

What statistical concept properly explains the “sophomore slump”?

- Regression to the mean [X]
- Law of averages
- Normal distribution

Measurement Error Models

Question 1

In our model of time vs. observed_distance, the randomness of our data was due to:

- sampling
- natural variability
- measurement error [X]

Question 2

Which of the following are important assumptions about the measurement errors in this experiment?

- The measurement error is random [X]
- The measurement error is independent [X]
- The measurement error has the same distribution for each time i [X]

Question 3

Which of the following scenarios would violate an assumption of our measurement error model?

- The experiment was conducted on the moon.
- There was one position where it was particularly difficult to see the dropped ball. [X]
- The experiment was only repeated 10 times, not 100 times.