Hickman PS 12

Keith Hickman

December 2, 2017

Problem 1:

Guess the correlation. *Please see attachment in Canvas*

Problem 2:

Is there a simpler explanation for the decreased performance following praise? What does this have to do with Chapter 15?

Answer

The simplest explanation is a that the two variables are causally unrelated. I assume that complex manuevers are difficult, and that positive reinforcement wouldn't cause or predict further successful attempts. Further, I would argue that the number of attempts would probably be a better predictor - the higher the number of attempts, the greater the likelihood that the next attempt will be successful.

In chapter 15, we are concerned with explaining the relationship between two variables - do they move in the same way, and is there potentially a causative effect? In this example, we see that there might be a correlative relationship, which it seems was confused with a causal relationship. Simply because two variables are correlated does not mean that one causes the other.

Problem 3:

Trosset chapter 15.7 exercise 4 Assume that the population of all sister-brother heights has a bivariate normal distribution and that the data in Table 14.4 were sampled from this population.

Answer

Let's set up our environment:

```
sibs <- read.csv("sisbro.csv", header=TRUE)
## sibs
sister <- sibs$Sister
brother <- sibs$Brother
# qqnorm(sister)
# qqnorm(brother)
# mean(sister)
# mean(brother)
# sd(sister)
# sd(brother)
# cor(sister, brother)
# lm(sister ~ brother)
summary(lm(sister ~ brother))
```

```
##
## Call:
```

```
## lm(formula = sister ~ brother)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0541 -1.2635  0.4189  0.5000  3.9459
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.6351    18.0371   1.532   0.1599
## brother       0.5270     0.2612   2.018   0.0744 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.247 on 9 degrees of freedom
## Multiple R-squared:  0.3114, Adjusted R-squared:  0.2349
## F-statistic:  4.07 on 1 and 9 DF,  p-value: 0.07442
```

(a) Consider the population of all sister-brother heights. Estimate the proportion of all brothers who are at least 5' 10".

First, let's solve for the probability that any given brother is taller than 70" $P(70 \leq Y)$. We'll find Z-score for 70 in the variable `brother`, then plug the standardized value into the `pnorm` function.

```
(70 - mean(brother)) / sd(brother)
```

```
## [1] 0.3676073
```

```
1 - pnorm(.36)
```

```
## [1] 0.3594236
```

Roughly 36% of the our population is at least as tall as 5'10".

(b) Suppose that Carol is 5' 1". Predict her brother's height.

```
# Prediction for Carol's brother's height:

slope = cor(sister, brother) * sd(brother) / sd(sister)
intercept = mean(brother) - slope * mean(sister)
predict.61 = intercept + 61 * slope
print(predict.61)
```

```
## [1] 67.22727
```

Our prediction is that Carol's brother is `67.2"` tall.

(c) Consider the population of all sister-brother heights for which the sister is 5' 1" *I assumed that the question was asking for sister heights of **at least** 61 inches - there are no sisters that are 61" and only two that are within +/- 1"*. Estimate the proportion of these brothers who are at least 5' 10".

First, we find the $E(Y|X \geq 61)$:

```
r <- cor(brother, sister)
slope <- r * (sd(brother)/sd(sister))
pop <- mean(sister) + (slope * (61 - mean(sister)))
pop
```

```
## [1] 62.22727
```

Next, we find $Var(Y|X \geq 61)$

```
var.set <- (1 - r^2) * sd(brother)
var.set
```

```
## [1] 1.873126
```

Now, we can find the proportion of `61"` sisters with brothers who are at least `5'10"` by using the pnorm function, plugging in the `pop` and `var.set` variables:

```
pnorm(70, 62.22, 1.87)
```

```
## [1] 0.9999841
```

Alternatively -

1. Get the regression prediction:

```
predict.61
```

```
## [1] 67.22727
```

2. Estimate the prediction error. We have a small sample size, so this may not work well:

```
r
```

```
## [1] 0.5580547
```

```
pred.error = sd(brother) * sqrt(1 - r^2)
pred.error
```

```
## [1] 2.257311
```

3. Use the normal distribution:

```
pnorm(70, mean=predict.61, sd=pred.error)
```

```
## [1] 0.8903388
```

This answer is a bit closer than my original answer of 99%.
Based on empirical observation of the data, the number should have been closer to 60%.

Using

```
mean61 = 31.1818 + 0.5909 * 61
rse = 2.379
pnorm(70, mean61, rse)
```

```
## [1] 0.8781406
```

Again, using the mean and the residual square error gets us to 87%, which is the closest answer so far.

Problem 4:

Trosset chapter 15.7 exercise 5

(a) Compute the sample coefficient of determination, the proportion of variation "explained" by simple linear regression.

```
summary(lm(brother ~ sister))
```

```
##
## Call:
## lm(formula = brother ~ sister)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5909 -1.2273 -0.9545  1.1136  4.0000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31.1818    18.7584   1.662   0.1308
## sister        0.5909     0.2929   2.018   0.0744 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.379 on 9 degrees of freedom
## Multiple R-squared:  0.3114, Adjusted R-squared:  0.2349
## F-statistic:  4.07 on 1 and 9 DF,  p-value: 0.07442
```

The coefficient of determination is .3114, or 31% of the variance explained. (Dr. Luen's objections noted)

(b) Let $\alpha = 0.05$. Do these data provide convincing evidence that knowing a sister's height x helps one predict her brother's height y?

No, there is not enough evidence from our dataset to show that knowing a sister's height will help predict her brother's height at $\alpha = .05$. With a small sample size and an α relatively close to our significance level, we might have a different result with more data.

(c) Construct a 0.90-level confidence interval for the slope of the population regression line for predicting y from x.

```
b <- ((1 - slope) / 9) * (sd(brother) / sd(sister))
upper.bound <- slope + qt(.95, df=9) * sqrt(b)
lower.bound <- slope - qt(.95, df=9) * sqrt(b)

print(c(upper.bound, lower.bound))
```

```
## [1] 0.9930700 0.1887482
```

Thus, our 90% confidence interval for the slope of the population regression line is from .18 to .99

(d) Suppose that you are planning to conduct a more comprehensive study of sibling heights. Your goal is to better estimate the slope of the population regression line for predicting y from x. If you want to construct a 0.95-level confidence interval of length 0.1, then how many sister-brother pairs should you plan to observe?

```
q <- qnorm(1 - .05/2)
q
```

```
## [1] 1.959964
```

```
L <- 2 * q * sd(brother)/sqrt(length(brother))
##this gives us 1.95, so I know I need to go significantly higher

L10 <- 2 * q * sd(brother)/sqrt(length(110))
L10
```

```
## [1] 10.66336
```

I'd need about 100 pairs of siblings.

Problem 5:

Trosset chapter 15.7 exercise 8 A class of 35 students took two midterm tests. Jack missed the first test and Jill missed the second test. The 33 students who took both tests scored an average of 75 points on the first test, with a standard deviation of 10 points, and an average of 64 points on the second test, with a standard deviation of 12 points. The scatter diagram of their scores is roughly ellipsoidal, with a correlation coefficient of r = 0.5. Because Jack and Jill each missed one of the tests, their professor needs to guess how each would have performed on the missing test in order to compute their semester grades.

(a) Jill scored 80 points on Test 1. She suggests that her missing score on Test 2 be replaced with her score on Test 1, 80 points. What do you think of this suggestion? What score would you advise the professor to assign?

Start with creating the variables required for our linear model:

```
avg.a <- 75
avg.b <- 64

sd.a <- 10
sd.b <- 12

r <- .5
```

The scatter plot is roughly ellipsoidal, we can proceed under the assumption of bivariate normality (also because that's explicit in the instructions).

To make a prediction of $\overline{y}$ given a value of $\overline{x}$, We know that $\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 * \overline{x}$ where β_0 = the Y-intercept and β_1 = the slope of regression line.

We find β_1 (slope) by multiplying r by the ratio of standard deviations of y and x:

```
beta1 <- r * (sd.b/sd.a)
beta1
```

```
## [1] 0.6
```

Next We find the intercept β_0 by subtracting from $\overline{y}$ (avg of test b) the product of β_1 and $\overline{x}$ (average of test a)

```
beta0 = avg.b - (beta1 * avg.a)
beta0
```

```
## [1] 19
```

Now we can use our beta coefficients to make a linear point prediction for Jill's second test after scoring 80 on the first test:

```
# Prediction for Jill:
19 + 0.6 * 80
```

```
## [1] 67
```

This seems about right - Jill's scores moved closer to the mean on the second attempt.

(b) Jack scored 76 points on Test 2, precisely one standard deviation above the Test 2 mean. He suggests that his missing score on Test 1 be replaced with a score of 85 points, precisely one standard deviation above the Test 1 mean. What do you think of this suggestion? What score would you advise the professor to assign?

If we know Y and want to predict X, we need to change the nomenclature:

```r
#Renaming Test1 as our Y variable
avg.y <- 75

#Test 2 is now X, as we know Jack's score on Test 2.
avg.x <- 64

#SD and correlations stay the same:
sd.y <- 10
sd.x <- 12

r <- .5

beta.one <- r * (sd.y/sd.x)
beta.one
```

```
## [1] 0.4166667
```

```r
beta.nt = avg.y - (beta.one * avg.x)
beta.nt
```

```
## [1] 48.33333
```

```r
# Prediction for Jack:
48.3 + .41 * 76
```

```
## [1] 79.46
```

Again, this result makes sense as Jack's scores have regressed to the mean of Y (Test 1)