Homework 12 answers

S320/520

Due at the beginning of class, Thursday 10th December.

Please write "S320" or "S520" at the top of your homework.
Show working and include all graphs you are asked to draw (in R.)

1. From Kahneman and Tversky:

   > Regression is inevitable in flight maneuvers because performance is not perfectly reliable and progress between successive maneuvers is slow. Hence, pilots who did exceptionally well on one trial are likely to deteriorate on the next, regardless of the instructors' reaction to the initial success. The experience flight instructors actually discovered the regression but attributed it to the detrimental effect of positive reinforcement. This true story illustrates a saddening aspect of the human condition. We normally reinforce others when their behavior is good and punish them when their behavior is bad. By regression alone, therefore, they are most likely to improve after being punished and most likely to deteriorate after being rewarded. Consequently, we are exposed to a lifetime schedule in which we are most often rewarded for punishing others, and punished for rewarding.

2. (a)
```
x = c(-0.2, -0.9, -0.4, 0.6, 0.4)
y = c(0.4, -0.3, -0.3, 0.5, 1.1)
cor(x,y)
```
   The correlation is 0.82.

   (b)
```
slope = cor(x,y) * sd(y) / sd(x)
intercept = mean(y) - slope * mean(x)
```
   The regression line is
   $$\hat{y} = 0.36 + 0.80x$$

   (c)
```
slope2 = cor(x,y) * sd(x) / sd(y)
intercept2 = mean(x) - slope2 * mean(y)
```
   The regression line is
   $$\hat{x} = -0.34 + 0.85y$$

   Or, if we want y to be the subject,

```
intercept3 = -intercept2/slope2
slope3 = 1/slope2
```
   $$y = 0.40 + 1.18\hat{x}$$

(d)
```
plot(x,y)
abline(intercept, slope)
abline(intercept3, slope3, col="red")
```

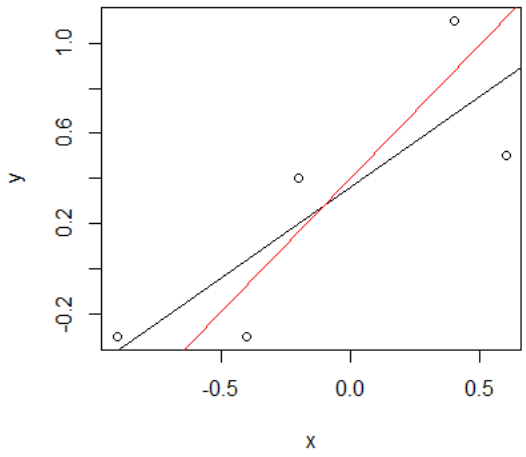

Figure 1: Regression lines for Q1(d).

The regression lines are not the same. Unless there is perfect correlation, the line to predict y from x is not the same as the line to predict x from y.

3. (Trosset 15.7.5.)

(a) This is just r^2, which is 0.31.

(b) We test the null hypothesis that the slope is 0:

```
std.error = sd(brother)/sd(sister) * sqrt((1-r^2)/(n-2))
t.stat = slope / std.error
# Two-sided P-value
2 * (1 - pt(abs(t.stat), df = n-2))
```

The P-value is 0.07, which is some evidence, but not quite convincing at the 0.05 level (of course, it's a small sample).

(c)
```
slope - qt(0.95, df=n-2) * std.error
slope + qt(0.95, df=n-2) * std.error
```

A 90% confidence interval runs from 0.05 to 1.13.

(d) The width of the confidence interval is

$$2 \times q_t \times \frac{s_y}{s_x}\sqrt{\frac{1-r^2}{n-2}}.$$

Set this equal to 0.1 and make n the subject:

$$\begin{aligned} 0.1 &= 2 \times q_t \times \frac{s_y}{s_x}\sqrt{\frac{1-r^2}{n-2}} \\ \frac{0.05 s_x}{q_t s_y} &= \sqrt{\frac{1-r^2}{n-2}} \\ \frac{0.0025 s_x^2}{q_t^2 s_y^2} &= \frac{1-r^2}{n-2} \\ n &= \frac{q_t^2 s_y^2 (1-r^2)}{0.0025 s_x^2} + 2. \end{aligned}$$

Using the observed values of s_x, s_y, and r, and making the approximation $q_t \approx 1.96$, we get $n \approx 1189$.

4. (Trosset 15.7.8)

   (a) This is a bad suggestion: for a start, Test 2 has lower scores than Test 1. We can assign a score using regression. The slope of the regression line is $0.5 \times 12/10 = 0.6$, and the intercept is $64 - 0.6 \times 75 = 19$. The prediction is $19 + 0.6 \times 80 = 67$.

   (b) This is a bad suggestion because of the regression effect. The regression effect states that individuals that do well on one test (in terms of standard units) will tend to do somewhat less well on another moderately correlated test (again, in terms of standard units). So somebody one standard deviation above the mean on one test will, on average, do somewhat less well on another test. Instead, we fit another regression line: the slope is $0.5 \times 10/12 = 5/12$, and the intercept is $75 - (5/12)64 = 48 + 1/3$. The prediction is $48 + 1/3 + (5/12)76 = 80$: in other words, half a standard deviation above the mean.

5. (a)
```
examanxiety = read.table("Teaching/examanxiety.txt", header=TRUE)
anxiety.lm = lm(Exam ~ Anxiety, data=examanxiety)
summary(anxiety.lm)
```
   The regression line is

$$\text{Predicted exam score} = 111.2 - 0.73 \times \text{anxiety score}$$

   (b) Firstly, independence is (approximately) satisfied, since students' should have negligible direct effect on each other's scores. Draw some plots:

```
Anxiety = examanxiety$Anxiety
Exam = examanxiety$Exam
ExamResiduals = anxiety.lm$residuals
plot(Anxiety, ExamResiduals)
qqnorm(ExamResiduals)
```

   There's no clear trend in the residuals, so the linear fit is reasonable. However, there's more spread on the right hand side of the plot, so error variance isn't constant — the data is heteroskedastic. The QQ plot bends a little at each end (because there's a maximum and minimum possible scores), so the normal distribution isn't a great fit for the errors. In summary:

i. Linearity: quite possibly
ii. Independence: quite possibly
iii. Equal variance (homoskedasticity): nope
iv. Normality of errors: probably not

6. You're free to use whatever model you see fit. For example, you might be interested in anxiety and amount of time spent revising matter but not gender. Then you'd do

```
lm(Exam ~ Anxiety + Revise, data=examanxiety)
```

This gives the model

$$\text{Predicted exam score} = 90.3 - 0.523 \times \text{anxiety score} + 0.272 \times \text{hours spent revising}$$

This is a perfectly acceptable model. Is it the best? That depends — the definition of "best" depends on exactly what the goal of your regression is.