# Homework 12 answers

S520

**Upload your answers as a PDF or HTML file through the Assignments tab on Canvas by 4 pm, Thursday 28th April. No late homeworks at all will be accepted this week.**

Trosset question numbers refer to the hardcover textbook. Draw all graphs in R and include all R code. You may work with others, but you must write up your homework independently — you should not have whole sentences in common with other students or other sources.

1. The psychologists Daniel Kahneman and Amos Tversky described the following situation:

   > The instructors in a flight school adopted a policy of consistent positive reinforcement recommended by psychologists. They verbally reinforced each successful execution of a flight maneuver. After some experience with this training approach, the instructors claimed that contrary to psychological doctrine, high praise for good execution of complex maneuvers typically results in a decrement of performance on the next try.[1]

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

   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. Trosset chapter 15.7 exercise 4

   (a) We answer this question using the normal distribution:

[1] Reprinted in *Judgement Under Uncertainty: Heuristics and Biases* (1982).

```
sister = c(69, 64, 65, 63, 65, 62, 65, 64, 66, 59, 62)
brother = c(71, 68, 66, 67, 70, 71, 70, 73, 72, 65, 66)
m = mean(brother)
s = sd(brother)
1 - pnorm(70, mean = m, sd = s)
```

gives an estimated proportion of brothers at least 5' 10" of 0.36 or 36%.

(b)
```
r = cor(sister, brother)
slope = r * sd(brother) / sd(sister)
intercept = mean(brother) - slope * mean(sister)
```

The regression line is $31.2 + 0.59\times$ sister's height. Plugging in 61 inches:

```
prediction = intercept + slope * 61
```

gives a predicted brother's height of 67.2 inches.

(c) The heights will have a normal distribution centered at the regression prediction (i.e. the answer to part (b). )There are a couple of ways to estimate the standard deviation. One is to do `s*sqrt(1-r^2)`, which gives 2.26 inches. A bit better is to do

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

and read off the residual standard error, which is 2.38 inches. Then the required probability is

```
1 - pnorm(70, mean = prediction, sd)
```

which gives 11% or 12%, depending on the SD used.

3. Trosset chapter 15.7 exercise 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.

4. *Apply the bivariate normal to the baseball wins data in* `baseball-wins.txt` *to estimate the following:*

(a) *The probability that a randomly selected team wins at least 84.5 games.*

```
baseball = read.table("baseball-wins.txt", header=TRUE)
y1wins = baseball$year1.wins
y2wins = baseball$year2.wins
1 - pnorm(84.5, mean(y2wins), sd(y2wins))
```

This gives a probability of 38%. (You can use year 1 wins instead of year 2 wins; it doesn't make a real difference.)

(b) *The probability that a team that won 95 games one season wins at least 84.5 games the next season.*

Fit the regression model:

```
r = cor(y1wins, y2wins)
slope = r * sd(y2wins) / sd(y1wins)
intercept = mean(y2wins) - slope * mean(y1wins)
```

(Or you can just use the `lm()` function.) Then get a regression prediction for a team that won 95 games:

```
pred95 = intercept + slope * 95
```

This is 88.2; this will be the mean of our normal distribution. To find the standard deviation, either get the residual standard error from the `lm()` function, or do

```
new.sd = sd(y2wins) * sqrt(1 - r^2)
```

It's 9.77 or 9.76 depending on method. The probability is then

```
1 - pnorm(84.5, pred95, new.sd)
```

giving 65%.

(c) *The probability that a team that won 75 games one season wins at least 84.5 games the next season.*

```
pred75 = intercept + slope * 75
1 - pnorm(84.5, pred75, new.sd)
```

The chance is about 25%.

5. *The file* ***examanxiety.txt*** *on Canvas contains information on a number of variables:*

- `Exam`: score on a math exam
- `Revise`: hours spend revising for the math exam
- `Anxiety`: "math anxiety" on a scale from 0 to 100 (100 is most anxious)

(a) *Find the regression line to predict exam score from anxiety. Write down your answer as an equation (do not just paste R output.)*

```
examanxiety = read.table("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) *Which of the following regression assumptions are met?*

i. Linearity
ii. Independence
iii. Equal variance (homoskedasticity)
iv. Normality of errors

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

(c) *Suppose we want to make probabilistic predictions of a student's exam score given their math anxiety. Should we use the bivariate normal? Why or why not?*

No. We need all of the assumptions to be at least approximately met to justify the bivariate normal for probabilistic prediction. That's not the case here.

6. *When 100 children at a certain school begin a grade, they're given an aptitude test. The top 20% go to class A, the next 20% to class B, and so on down to the bottom 20%, who go to class E. At the end of the year, they're given a similar aptitude test. The data is given in the file* `testscores.txt`.

*The Board of Education, none of whom have taken a statistics course, wants to measure teacher performance by looking at the average change (second test minus first test.) They get the following results:*

- Class A: Average change: $-2.15$
- Class B: Average change: 3.45
- Class C: Average change: 7.6
- Class D: Average change: 14.5
- Class E: Average change: 21.9

*Based on these results, the Board is deeply unhappy with the teacher of Class A. But did Class A really perform the worst compared to expectations? Do calculations and give a conclusion that the Board can understand.*

The decrease in scores of Class A could just be regression to the mean — that is, it's quite possible that some of the Class A students just got lucky on the first test. The teacher can hardly be blamed if they are merely failing to be lucky again on the second test.

A better strategy would be to fit a regression line that predict scores on the second test based on the first test, then compare the observed scores to these predicted scores.

```
testscores = read.table("testscores.txt", header=TRUE)
first.test = testscores$first.test
second.test = testscores$second.test
slope = cor(first.test, second.test) * sd(second.test) / sd(first.test)
int = mean(second.test) - slope * mean(first.test)
```

We find that

$$\text{Predicted second test score} = 0.193 \times \text{first test score} + 49.9$$

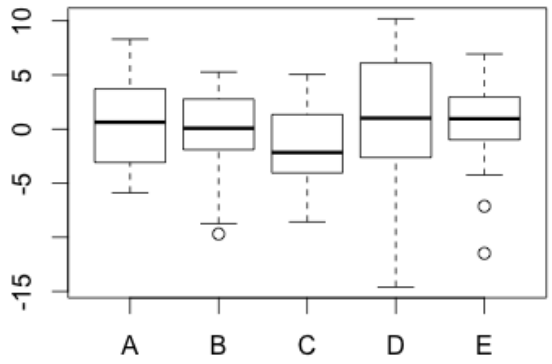
is the regression line — the best line for prediction.

Now find the residuals: the differences between actual scores and predicted scores.

```
predictions = slope * first.test + int
residuals = second.test - predictions
```

Graphically compare the residuals for the five classes. Does it look like there's any real difference between these five sets of residuals?

```
boxplot(residuals~testscores$class)
```



The boxplots look pretty similar — the one for class C is a bit lower but this might be luck. (We could test this formally with an ANOVA but (i) it's a bit tricky to set up, and (ii) it's superfluous.) There's no obvious sign that one class is doing better than any other in terms of residuals.