Homework 3 Solutions

by the honorable Asa Levi

p198 # 2 (5 pts)

There is definitely something wrong! High school and college GPAs are between 0 and 4. If for some HS GPA the regression line predicts a college GPA of 2 then the maximum error in this vertical strip is 2. This same idea shows that for a predicted GPA between 1 and 3 the maximum error is less than or equal to 3. We expect most of the predicted GPAs to be between 1 and 3, and the error is always bounded by 4, thus the RMS error should be less than 3.

p198 # 4 (10 pts)

Part (a)

The RMS error is $\text{SD(Y)}\sqrt{1-r^2} = 15\sqrt{1-0.6^2} = 12$. We expect 68% of the students to be within 12 points of the prediction, thus $32\% \approx \frac{1}{3}$ of the students, will be more than 12 points away from their prediction.

Part (b)

The equation for the regression line is

$$(y-55) = (0.6)\frac{15}{25}(x-50)$$

For $x = 80$ we get $y = 65.8$ which is our prediction.

Part (c)

The prediction is likely to be off by 1 RMS error from the actual value. So it is likely to be off by 12 points or so.

p198 # 5 (15 pts)

Part (a)

This doesn't have to do with regression, because we are not looking inside a vertical strip. We are looking over the whole population, so this is an old-fashioned z-table question. To do this we first convert 80 into SU and then use a z-table to figure out how much area is to the right of this value.

First 80 in SU is $\frac{80-55}{15} = 1.66$. The area to the right of $z = 1.66$ is about 4.945 %.

Part (b)

Here we are looking inside of a vertical strip, so we have to use the fancier method. When we are looking inside a vertical strip the average and SD is different. The average in the vertical strip is predicted by the regression line, while the SD in the vertical strip is the RMS error ONLY BECAUSE IT IS FOOTBALL SHAPED!!

We predicted the score for the a student whose midterm score is 80 in 4(b) to be 65.8. In 4(a) we computed the RMS error to be 12. Now we follow the same method as in Part (a) for a distribution for an average of 65.8 and SD of 12. So first we convert 80 to SU: $\frac{80-65.8}{12} = 1.18$. Now we want the area to the right of $z = 1.18$, which is approximately 11.5%.

p215 # 9 (10 pts)

Grouping the data in this way makes it more correlated, because the points plotted are closer to the averages in the vertical strips. Its regression line will be close to the regression line of all the data, thus we have that the slope of the line is around the slope of the regression line which is $r\frac{\mathrm{SD(Y)}}{\mathrm{SD(X)}} = (0.5)\frac{15}{45000} = \frac{1}{6000}$.

p215 # 10 (10 pts)

We can predict the average income of a child with IQ 110 using the regression line, with IQ as the independent variable and income as the dependent variable. The equation for the regression line is:

$$(y - 90000) = (0.5)\frac{45000}{15}(x - 100)$$

In our case we have a child with ID 110, so we plug that into the regression line and get a predicted income of \$105,000. So the estimate of \$150,000 is likely to be too high.

p235 # 6 (10 pts)

Part (a)

True.

Part (b)

True.

Part (c)

False. The two events are dependent, so the chance is not simply the product of the chances. It is $\frac{1}{52} \times \frac{1}{51}$.

p236 # 12 (10 pts)

You need one of your three balls to be drawn on the first time, one of the two remaining to be drawn on the second draw and the final must be your final ball. Since there is no replacement the odds of this happening are:

$$\frac{3}{100}\frac{2}{99}\frac{1}{98} = \frac{6}{970200} \approx 6.184 \times 10^{-6}$$

Worksheet 1 (10 pts)

This is a question about the whole population, so it is just a straight up z-table question. A person is actually obese if their BF % is over 25 %, so we just need to convert this to SU and then find the percent that are to the right of it. 25% in SU is $\frac{25-24}{6} = 0.16$. The area to the right of $z = 0.16$ is about 44 %, so 44% of the population is actually obese.

Worksheet 2 (10 pts)

This is another whole population question, so we follow the same method as in the previous question. A person is BMI-obese if their BMI is above 30, so we convert this to SU and find the area to the right of it using a z-table. In SU 30 is $\frac{30-26}{5} = 0.8$. The area to the right of $z = 0.8$ is about 21.185%, so 21.185% of the population is BMI-obese.

Worksheet 3 (40 pts)

Part (a)

This is a question about vertical strips, so we need to follow the more complex approach. First the vertical strip is above a BMI, so BMI is the independent variable, while BF % is dependent. The average in the vertical strip is given by the regression line:

$$y - 24 = (0.66)\frac{6}{5}(x - 26)$$

Plugging in $x = 30$ we get the average in this vertical strip, which is 27.168 % BF.

Next we also need the SD for the vertical strip. Since our scatter diagram is football shaped this is the RMS error which is $\mathrm{SD(Y)}\sqrt{1 - r^2} = 6\sqrt{1 - 0.66^2} \approx 4.5$. Now, we have a normal distribution with average 27.168 % and SD 4.5 % and we want to figure out what % is to the right of 25%. So we convert 25 to SU: $\frac{25-27.168}{4.5} \approx -0.48$. The area to the right of $z = -0.48$ is about 69.145%. So in conclusion 69.145% of people with BMI 30 are actually obese.

Part (b)

This is just a z-table question. We need first to convert 30 and 33 to SU: 30 in SU is $\frac{30-26}{5} = 0.8$, while 33 in SU is $\frac{33-26}{5} = 1.4$. So we are looking for the area in under the normal curve between 0.8 and 1.4. This is 13.11%.

Part (c)

The simplification lets us say that for a vertical strip above any BMI between 30 and 33 the % of people in that strip that are actually obese is 69.145%. Now the (b) tells us that 13.11% of the total population has a BMI between 30 and 33, and the percent of these people that are actually obese is 69.145%. Thus the percent of the population with BMI between 30 and 33 that is also actually obese is $69.145\% \times 13.11\% \approx 9.06\%$.

Part (d)

First we need to figure out what % of people with BMI 33 are actually obese. By plugging $x = 33$ into the regression line, we get that the average in this vertical strip is 29.544% BF. Now, in this vertical strip 25% in SU is $\frac{25-29.544}{4.5} \approx -1$. So the percent of people with BMI 33 who are actually obese is about 84.135.

Now we need to figure out the percent of people with BMI between 33 and 36. This is just a z-table question. 33 in SU is $\frac{33-26}{5} = 1.4$ and 36 in SU is $\frac{36-26}{5} = 2$. Using a z-table we find that the area we want is 5.8%.

So the final step is just to multiply these, which tells us that about 4.88% of the population has BMI between 33 and 36 and is actually obese.

Part (e)

The total % of the population that is both BMI-obese and actually obese is just the sum of the actually obese people with BMI between 30 and 33, between 33 and 36 and above 36. We computed the first two in the last two parts. In this part is says to just assume that everybody with BMI over 36 is actually obese, so we just assume that the % of the population with BMI over 36 that is actually obese is just the % of the population with BMI over 36. Figuring out the % of the population with BMI over 36 is another z-table question. By doing it you find that 2.175% of the population has BMI over 36.

Thus the total percent of the population that is BMI-obese and actually obese is $9.06\% + 4.88\% + 2.175\% = 16.115\%$.

Worksheet 4 (5 pts)

This is just your answer from 3(e) divided by your answer from 2. So that becomes $100 \times \frac{16.115}{21.185} \approx 76\%$

Worksheet 5 (5 pts)

This is just your answer from 3(e) divided by your answer from 1. In this case it is $100 \times \frac{16.115}{44} \approx 36.625\%$.

Worksheet 6 (10 pts)

The specificity is pretty high, which means that BMI-obesity is pretty accurate. That is, if you are BMI-obese then there is a good chance that you are actually obese. However, the sensitivity is low. This means that just because you're not BMI-obese doesn't mean that you're not actually obese :(