36-309/749 Math Review 2014

The math content of 36-309 is not high. We will use algebra, including logs. We will not use calculus or matrix algebra. This optional handout is intended to help those students who need to review their basic math skills. If you are at all unsure about your skills, you should *first* solve the problems, *then* look at the answers! See Howard or a TA if you need more help. You will not need all of this math at once, so you have time to re-learn it before you need to use it. In particular, problem 3 is for later in the course.

Problem 1: Distributions. A certain outcome is ***normally distributed*** and has mean 12.3 and standard deviation 1.5. The "Z score" for any particular value, X, is defined as $Z = \frac{X - mean}{sd}$. A. What are the Z scores for X=14.9 and X=9.1?

Answer: For X=14.9, Z=(14.9-12.3)/1.5=1.73. Caution: If you enter "14.9 – 12.3 / 1.5 ="on most calculators, you will get the wrong answer of 6.7. To get the right answer, either use the parentheses or enter "14.9 - 12.3 = / 1.5 =". A third method is to first solve 14.9-12.3=2.6 and then solve 2.6/1.5=1.73. Note that perhaps the easiest way to work is to type "(14.9-12.3)/1.5" in the Google search box to get the answer.

Try to verify the reasonableness of your answer by thinking about its *meaning*. A Z score tells us how far a value is from the mean in standard deviation units. So 1.73 indicates a value that is between 1 and 2 s.d. units (of 1.5 each) above the mean, while 6.7 indicates a value that is far above the mean, which 14.9 (the wrong answer above) is not.

Important note: Always round your answers to an appropriate number of decimal places, no more than one beyond what is in the data. **As a rule of thumb, final answers should have three significant figures.** *But*, if you need to use one set of results in calculation for a further set of results, you should include 2 or 3 extra significant figures in the intermediate results to avoid cumulative rounding errors.

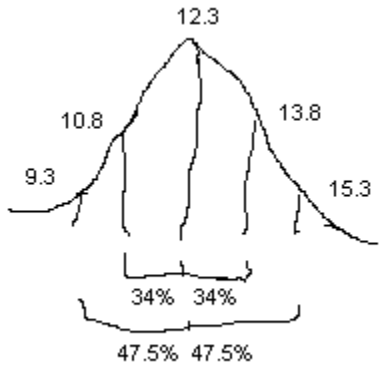
For X=9.1, Z = (9.1-12.3)/1.5 = -3.2/1.5 = -2.13. This is negative because 9.1 is below the mean.

B. We know that, for a normal distribution, half of the outcomes are above the mean, 68% are within 1 s.d. of the mean, and 95% are within 2 s.d. of the mean. Using the mean and standard deviation of part A, what interval holds 68% of the data? What interval holds 95%?

Answer: We want to calculate the intervals [mean-sd, mean+sd], and [mean-2sd, mean+2sd]. The first is [12.3-1.5, 12.3+1.5]=[10.8, 13.8]. The second is [12.3-2(1.5), 12.3+2(1.5)]. On approach is to first calculate 2*1.5=3.0. Then the answer is [12.3-3.0, 12.3+3.0]=[9.3, 15.3].

C. Using the mean and s.d of part A, what fraction of the values is above 10.8? below 15.3?

Answer: A picture helps:



Using the fact than 50% are above 12.3 and 50% below, 34+50=84% are above 10.8 and 50+47.5=97.5% are below 15.3.

Problem 2: Model equations. The predicted value of an outcome, Y, from a certain ***linear regression*** problem with explanatory variables X_1, X_2, X_3, and X_4 is $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$, where b_0=-2.0, b_1=3.5, b_2=4.9, b_3=0.8 and b_4=-10.0.

A. Fill in the chart:

X_1	X_2	X_3	X_4	Y
2.0	3.0	0	1	
2.5	0.5	0	0	
3.4	1.2	1	1	

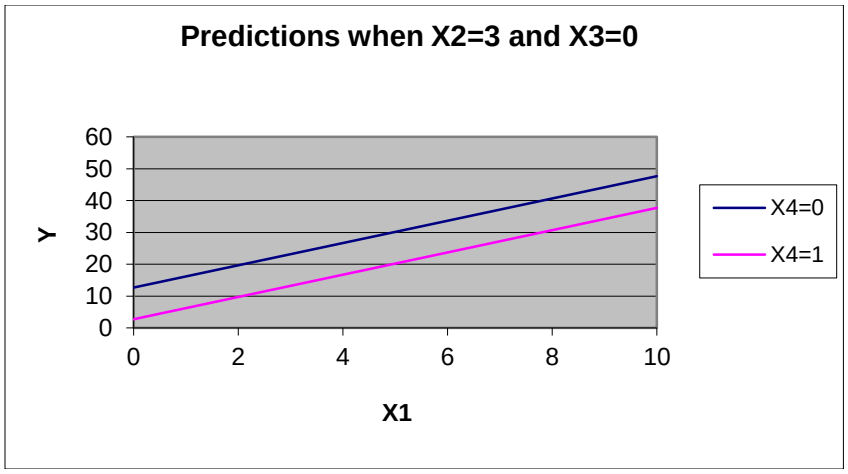
Answers:
Y=-2.0+3.5*2.0+4.9*3.0+0.8*0-10.0*1=9.70
Y=-2.0+3.5*2.5+4.9*0.5+0.8*0-10.0*0=9.20
Y=-2.0+3.5*3.4+4.9*1.2+0.8*1-10.0*1=6.58

B. What would Y be if the X's are like the second line, but with X_3=1?

The only difference is replacing +0.8*0 with +0.8*1, so the answer will be 0.8 higher, which is 10.0.

C. Draw a graph with X_1 on the horizontal axis and Y on the vertical axis with two lines representing the predicted values of the outcome for various values of X_1 between 0 and 10, when X_2=3.0 and X_3=0, for both X_4=0 and X_4=1. In other words, plug in X_2=3.0, X_3=0 and X_4=0, simplify the equation, and plot it. Then repeat for X_4=1.

Answer: Write $Y = -2.0 + 3.5X_1 + 4.9X_2 + 0.8X_3 – 10.0X_4$, or
$Y = -2.0 + 3.5X_1 + 4.9(3.0) + 0.8(0) – 10.0(0)$, or
$Y = -2.0 + 3.5X_1 + 14.7$ or $Y = 12.7 + 3.5\ X_1$. This represents a straight line with an intercept of 12.7 and a slope of 3.5, so graph it starting at Y=12.7 when X_1=0 and make the line rise 35 units in Y for each 10 unit increase in X_1. For X_4=1, the equation is $Y = -2.0 + 3.5X_1 + 14.7 –10.0$ or $Y = 2.7 + 3.5\ X_1$, which has the same slope, but an intercept of 2.7 (10.0 units lower).



We interpret this as follows: Y is directly proportional to X_1. (The slope is b_1 which is 3.5.) The intercept for X_4=1 is 2.5 lower than the intercept when X_4=0. (The above applies for any combinations of X_2 and X_3. As X_2 and X_3 change, the lines move up or down parallel to the lines shown.)

D. The model is now modified to include an **interaction** between X_1 and X_4 as follows: $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_{14}X_1X_4$ with b_{14}=2.0. Write the equations for Y as a function of X_1, X_2 and X_3 for X_4=0 and X_4=1. In other words, plug in 0 for X_4 and simplify the equation. Then repeat for X_4=1.

Answer: Write $Y = -2.0 + 3.5X_1 + 4.9X_2 + 0.8X_3 – 10.0X_4 + 2.0X_1X_4$. For X_4=0, this is $Y = -2.0 + 3.5X_1 + 4.9X_2 + 0.8X_3 – 10.0(0) + 2.0X_1(0)$, or
$Y = -2.0 + 3.5X_1 + 4.9X_2 + 0.8X_3$.
For X_4=1, we have $Y = -2.0 + 3.5X_1 + 4.9X_2 + 0.8X_3 – 10.0(1) + 2.0X_1(1)$, or
$Y = -2.0 + 3.5X_1 + 4.9X_2 + 0.8X_3 – 10.0 + 2.0X_1$ or
$Y = -12.0 + 5.5X_1 + 4.9X_2 + 0.8X_3$.
This shows that for an interaction, both the slope and the intercept change when the level of an explanatory variable changes.

Problem 3: **Logs**. (ln means natural log, log10 means base 10, log may mean either)

A. If ln(y) = 4.2 + 3.5x, what is y?

Answer: Exponentiate both sides of the equation. Since the logs are natural logs the answer is exp(ln(y)) = exp(4.2 + 3.5x) where exp(z) means e^z and e is approximately 2.718. We simplify exp(ln(y)) to y because exp() and ln() are inverse functions of each other, i.e. one undoes what the other does. We simplify exp(4.2 + 3.5x) to exp(4.2)*exp(3.5x) using the general rule:

$$a^{b+c} = a^b \cdot a^c$$

A calculator tells us that exp(4.2)=66.7. Many calculators require you to enter 4.2, press the "inverse function key", then press the ln key to get exp(). The final answer is y=66.7*exp(3.5x) or $y = 66.7e^{3.5x}$

B. If y = 3.7 + 3.0 ln(x), how much does x have to change for y to increase by 3.0?

Answer: y will increase by 3.0 if ln(x) increases by 1.0, and that happens when x is *multiplied* by 2.718.
Remember the way that $\log_{10}(x)$ works: $\log_{10}(10^3)=\log_{10}(1000)=3$, $\log10(10^2)=\log_{10}(100)=2$, $\log_{10}(10)=1$, $\log_{10}(1)=0$ (the latter is because 10 to the zero power is 1).
Natural log works the same way: $\ln(e^3)=\ln(20.08)=3$, $\ln(e^2)=\ln(7.39)=2$, $\ln(e)=\ln(2.718)=1$, $\ln(1)=0$.
In general, ln(x) will increase by 1.0 if x is multiplied by 2.718 (*e*). For instance, ln(10)=2.30 and ln(27.18)=3.30.

C. A ***logistic regression*** model says that: ln(odds(success))=3.5-(0.05)age. How much does the (natural) log of the odds of success change for a decade increase in age? How much do the odds change for a decade increase in age?

Answer: If age increases by 10, (0.05)age increases by 0.5 and 3.5-(0.05)age *de*creases by 0.5. E.g., log odds of success for newborns is 3.5, and log odds for 10 year olds is 3.0, and log odds for 20 year olds is 2.5, and log odds for 40 year olds is 3.0.

If ln(odds(success))=3.5-(0.05)age, then exponentiating on both sides of the equal sign, we find that odds(success) = $e^{3.5-(0.05)*age} = e^{3.5}\, e^{-(0.05)*age}$. This says that when the age increases by another 10 years, the odds get multiplied by another $e^{-0.5}=0.606$, which is the same as being divided by $e^{0.5}=1.65$. For the decades newborn to 40, the odds are $e^{3.5}=33.1$, $e^{3.0}=20.1$, $e^{2.5}=12.2$, and $e^{2.5}=7.39$ respectively. Note how the odds are not evenly spaced like the log odds; for each additional decade, the odds (of success, for whatever our outcome is) are 0.606 times the odds for the previous decade.

Here is an interpretation of the meaning of "odds" for this problem: If the outcome is "having two living parents", then for every one 10-year-old child who does not have two living parents there are 20.1 10-year-old children who do.