NAME (Please Print): KEY

HONOR PLEDGE (Please Sign): ________________________________________________

Statistics 110

Homework 4

You are allowed to discuss problems with other students, but the final answers must be your own work.

For all problems that require calculation, YOU MUST ATTACH SEPARATE PAGES, NEATLY WRITTEN, THAT SHOW YOUR WORK. You may attach JMP-IN output.

Please mark your answer in the space provided. As a general rule, each blank counts for one point. If necessary work is not shown, or if that work is substantially wrong, then you will not get credit even if the answer is correct. (The obvious purpose of this seemingly draconian policy is to prevent people from mindlessly copying each other’s answers.)

Report all numerical answers to at least two correct decimal places.

DUE DATE: IN CLASS ON WEDNESDAY, OCTOBER 3.
1. Use the data that is available at:
   http://lib.stat.cmu.edu/DASL/Datafiles/agecondat.html
   Read the description. If your last name begins with a letter in A-G, do a linear regression that predicts the price of beef from the consumption of beef, the price of pork, the retail food price index adjusted by the CPI, and disposable income. If your last name begins with a letter in H-S, do a linear regression that predicts the price of pork from the price of beef, the consumption of pork, the adjusted retail food price index, and disposable income. If your last name begins with a letter in T-Z, do a linear regression that predicts the adjusted retail food price index from the consumption of beef, the consumption of pork, the price of pork, and disposable income.

   What is your regression equation?

   A-G: \( PBE = 122.51 - 1.16CBE + 0.55PPO - 0.43PFO + 0.008DINC \)
   H-S: \( PPO = 84.95 + 0.12PBE - 0.95CPO + 0.85PFO - 0.49DINC \)
   T-Z: \( PFO = -117.57 + 0.21CBE + 1.12CPO + 1.08PPO + 0.63DINC \)

   What is your prediction for a year in which the price of beef is 60 cents/lb, the price of pork is 50 cents/lb, the consumption of beef per capita is 55 pounds, the consumption of pork per capita is 58 pounds, the adjusted RFP is 850, and the disposable income per capita index is 35?
   A-G: 0.048; H-S: -4.64; T-Z, 9.51.

   What is the significance probability for deciding whether disposable income is a useful variable in your regression?

   \( t_{13} \)

   What distribution is used for determining the significance probability. (Include the df if appropriate.)

   What is the strongest correlation among two of the explanatory variables in your model? Which variables are these?

   A-G: the correlation between cbe and dinc are not high enough to cause concern.
   H-S: the correlation between pbe and dinc may be high enough to cause concern.
   T-Z: the correlation between cbe and dinc seem high enough to cause concern.

   Do you feel their correlation is sufficiently strong that there is a problem of multicollinearity?
Attach a residual plot for your analysis with respect to disposable income (1 point). Comment upon any issues that you see in it below.

There are lots of possibilities. I just want you to look at your plot and spot possible outliers, heteroscedasticity, and whatever.

Interpret the sign of your most significant coefficient. Identify the variable, indicate the significance probability associated with it, and try to explain whether it makes sense. Sometimes it is not reasonable to try to interpret the sign of the coefficient—if this is the case, please indicate that and don’t bother with an explanation.

Holding all other variables constant, we will increase (decrease) the y-variable if we increase explanatory variable for the positive (negative) slope.

2. In Finland, there was a study of monozygotic twin pairs in which one twin smoked cigarettes and the other did not. The study found that of 22 pairs in which one of the twins had died, in 17 pairs the smoker died first, and in 5 pairs the non-smoker died first. If one looked just at the 9 pairs in which the cause of death was coronary heart disease, in all 9 cases it was the smoker who died. And in the two cases in which the cause of death was lung cancer, in both cases it was the smoker who died. Please analyze and comment upon this experiment in a short attachment (probably about a typed page).

Monozygotic twins control for genetic and many environmental effects (but not necessarily lifestyle risks). If we assume that smoking is not associated with other risk factors, then we can build a simple binomial model for the data. Under the null hypothesis that smoking is not harmful, then whether or not the smoking twin dies first is like the toss of a fair coin. Under the alternative hypothesis that smoking is harmful, the outcome is like the toss of an unfair coin, with more than a 50% of the smoker dying first. For this data, the significance probability is \( P = \binom{22}{17} \cdot 0.5^{17} \cdot 0.5^5 = 0.0063 \). Thus there is strong evidence against the null hypothesis that smoking is not harmful.

3. Show that the ratio of the volume of a unit hypersphere to the volume of a unit hypercube goes to zero as the dimension of the space increases?

The volume of the unit hypercube in \( \mathbb{R}^p \) is \( 1^p = 1 \). The volume of the \( p- \)
dimensional sphere with radius \( r \) depends on whether \( p \) is even or odd. I’ll show the result for the even case, the simplest, and leave the odd case as bookwork. The volume formula is then \( \pi^{p/2}/2^{p/2} \). To prove the ratio goes to zero, we need to use Stirling’s approximation: \( x! \approx (x/e)^x \). Plugging this in, the limiting ratio is:

\[
\lim_{p \to \infty} \frac{\pi^{p/2}/2^{p/2}}{(p/2)^{p/2}} = \lim_{p \to \infty} \left(\frac{2\pi}{p}\right)^{p/2}.
\]

When \( p > 2e\pi \), the ratio is less than 1, and thus the limit goes to zero.

4. What is the probability that a particular datum does not get selected in a bootstrap sample?

\[
(1 - \frac{1}{n})^n
\]

\( e^{-1} \) What is the limit of this probability as \( n \to \infty \)?

Recall that \( e = \lim(1 + \frac{1}{n})^n \). So \( e^{-1} = \lim(1 + \frac{1}{n})^{-n} = \lim(1 - \frac{1}{n})^n \).

5. Use proof by contradiction to show that 3 is an irrational number.

Assume \( \sqrt{3} \) is rational. Then \( \sqrt{3} = a/b \) for integers \( a, b \). Then \( 3 = (a/b)^2 \) so \( 3b^2 = a^2 \). But \( a^2 \) and \( b^2 \) both have an even number of prime factors, but the lefthand side has an odd number of prime factors. Contradiction.

6. Use proof by induction to show that \( \sum_{i=1}^{n} i^2 = n(n + 1)(2n + 1)/6 \).

Trivial.

7. During World War II, the army tested inductees for syphilis. One approach is to give each man a test, but that is expensive. Alternatively, you can divide each man’s blood sample into two parts. Then one pools the half-samples from \( n \) men and tests the pool for syphilis. If the test is positive, one goes back and tests each of the remaining half-samples separately; if the test is negative, one need not. Suppose the rate of syphilis were 10%. What is the expected number of tests one would need if \( n = A + 3 \), where \( A \) is 1, 2 or 3, depending on whether your last name begins with A-G, H-S, or T-Z, respectively.

The probability that one test is needed is the probability that none of the \( A \) soldiers has syphilis, or \( \binom{n}{0} .1^0 .9^n = .9^n \). The probability that \( n + 1 \) tests are
needed is \(1 - .9^n\). Thus the expected number of tests is \(.9^n + (n + 1)(1 - .9^n)\).
For \(A=1, 2, 3\) the answers are 2.375, 3.048, and 3.811, respectively.