Stat 500 — 2009 Midterm II — Solution

Notes about my answers are marked by •
General things about my grading:

- If I wrote math error, you had the right idea/equation, but made a mistake in the calculation. I deducted one point for these.
- The exam had many places where results from early parts were used in subsequent parts. I deducted points for the mistake when (and only when) they occurred. I did not deduct subsequent points because your answer(s) in later parts didn't match mine. If you did the right things in subsequent parts, I gave you full credit.
- These solutions indicate some common mistakes. If you don't understand a comment on your exam or something in these solutions, please come and see me (or call/e-mail if you're offcampus).
- If I misadded points, or you don't understand why I deducted points, please see / call / e-mail me.

1. Health of factory workers

   (a) correlation coefficient, r=0.838. You are interested in association between two observed quantities, not prediction.
   • If you wanted to use a regression slope, which slope is the correct one, i.e. should X be the H or the L variable?

   (b) se $= s_e/\sqrt{\Sigma(x_i - \bar{x})^2} = 4.95/\sqrt{1428} = 0.131$

   (c) T $= b_1/se = 1.91/0.131 = 15.1$. df=101, p < 0.0001.

   (d)

Source	d.f.	SS	MS	F
Model	1	5576	5576	227.6
Error	101	2474.2	24.50	
c.total	102	8050.2		

   • Common mistakes were:
   2 df for model: remember model SS are the comparison of intercept model to intercept + slope. That's a difference of 1 d.f.
   Misplaced SS: c.total SS is the error SS for the intercept only model
   error is the error SS for the regression. Model SS is the difference

   (e) the prediction for $L_i = 32$.
   • No need to calculate prediction sd's. For a simple linear regression, the most precise prediction is at the mean X. $L_i = 32$ is closest to the mean.

   (f)

Assumption	diagnosis	explanation
lack of fit	no evid.	resid vs. pred plot is flat
equal var.	no evid.	resid vs. pred plot has equal spread
non-normality	no evid.	QQ plot is nearly linear
independence	no evid.	design suggests o.u. = e.u.

   "no evid." above means "no evidence of a problem"
   • You can't tell independence from the plots. Need to look at the design.

(g) Test lack of fit by comparing straight line to a more complicated model. This leads to a full (more complicated) /reduced (straight line) model F test. Details depend on the choice of full model; conclusion does not. My details are for the loess alternative.

Model	error d.f.	error SS
linear	101	2474.2
loess	97.5	2391.4

Change in df = 3.5; change in error SS = 82.8

F = (82.8/3.5) / (2391.4/97.5) = 0.96. Compare to 3.5, 97.5 F distribution. (or 3,100 F dn in tables). p-value is large. No evidence of lack of fit.

(h) $\hat{L} = 30.8$

- Obtained by solving 5.4 = -55.6 + 1.98 L.

(i) This requires Fisher Z transformations.

$Z_r = \frac{1}{2} \log\left(\frac{1+0.838}{1-0.838}\right) = 1.21$

$Z_\rho = \frac{1}{2} \log\left(\frac{1+0.8}{1-0.8}\right) = 1.099$

$Z = (1.21 - 1.099)\sqrt{103-3} = 1.16$

Test is a one-sided alternative. p-value = P[Z > 1.16] = 1-0.877 = 0.12. No evidence that $\rho > 0.8$.

- Some folks calculated a 2-sided p-value = ca 0.23. Problem called for a 1-sided p-value.
- A few folks calculated a T statistic from r. That T statistic tests Ho:$\rho = 0$. Problem called for comparing r to 0.8.

2. gasoline demand

(a) Holding all other variables in the model constant, an increase of 1 dollar in the gas price is associated with a decline of 23.65 million barrels gas consumption.

- Saying xx CAUSED yy is not appropriate because these are obserational data.
- Omitting "holding all other variables constant" omits a crucial part of the definition of a partial regression coefficient.

(b) F = $\frac{(800.96-207.64)/7}{207.64/(27-11)} = 6.53$

p < 0.0001

(c) Various conclusions are acceptable, including:

at least one β for .... is not zero

at least one .... contributes to predicting gas consumption

the simpler model can be improved by adding at least one of ...

In all cases ... is the list of variables in model 2 but not in model 3

- claiming all variables contribute (or are needed, or have $\beta > 0$) is not appropriate.

(d) This test is not possible because the two models are not nested.

- There is no "full" or "reduced" model here. Each model has variables not in the other.

(e) $\hat{Y} = 157.0$ (or ca. 161) depending on rounding

- common mistakes included omitting the intercept

or matching the wrong slope with the wrong value (e.g. -8.26*7000)

(f) $s_p = \sqrt{MSE + MSEx(X'X)^{-1}x'} = \sqrt{34.82(1 + 0.083)} = 6.14$

- Many answers were $s_{\hat{y}} = \sqrt{MSEx(X'X)^{-1}x'} = \sqrt{34.82(0.083)} = 1.7$. This is the s.e. of the line, i.e. of the mean prediction at that X. I asked for the sd of an **observation**.

(g) Lack of fit is big concern. For this prediction, the prediction is too high; se is too high.

- Why is this prediction too high? $\hat{Y} \approx 160$. Looking at the resid vs predicted value plot, you see that the residuals for predictions ca 160 tend to be ca -10. $e_i = Y_i - \hat{Y}_i \approx -10$, so $\hat{Y}_i \approx Y_i + 10$. Too high, at least for $\hat{Y} \approx 160$.
- If the residuals are near zero for a specific predicted value, that prediction is ok, even if other predictions are not. • The s.d. is likely too high because lack of fit increases MSE.

(h) No problem with multicollinearity: all VIF values are < 10.

- some folks looked at the scatterplot matrix and commented that the X variables seemed correlated. Definitely true. However, correlated X's are common, especially with observational data. The concern is whether the correlation is sufficiently large to be a concern. That's VIF.

(i) none. All DFBETA values are < 1 in absolute value.

- Answering using DFFITS or Cook's D doesn't address the question, which was about the effect of outliers on $\hat{\beta}$ for gasprice.
- Obs 21 has DFBETA = -0.97. It's really close to being a concern.
- 27 obs is a small-medium data set; hence the < 1 criterion. If you used the large data set cut point, I didn't argue.

(j) No problem - the plot doesn't argue for anything other than a straight line.

- Lots of different answers here. Some misinterpreted the added variable plot as 'good = no pattern'. No. The slope in the added variable plot is the partial regression slope. You want to see a pattern, that is if you want to find a non-zero partial regression coefficient. You want to see a straight line pattern. (Remember the wiggle in the HW made-up example. That sort of pattern would be a problem.)
- Quite a few argued this is not a nice line of points. You're right; there is a blob in the middle and a few big and a few small points. If the slope is driven by an outlier, you would see this in the DFBETA values. No problem there. Even though this is not a nice line, it isn't anything worse. There isn't a clear curve or wiggle, which would lead to a concern over lack of fit.

(k) DW = 0.738 (or autocorrelation coefficient = 0.631). p < 0.0001. Very strong evidence of autocorrelation.

3. models of vaccine effectiveness

(a) $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i$
X_i is the vaccine dose
Z_i is 1 with the adjuvant and 0 without

- Other equivalent equations were accepted for full credit

Writing an ANOVA model (different means for each dose) lost a couple of points because that doesn't honor “relationship between DOSE and RESPONSE is linear”.

(b) In my model, β_2.

(c) $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + \varepsilon_i$
No new variables needed.

- Some folks proposed $Y_i = \beta_0 + \beta_1 X_i + \beta_3 X_i Z_i + \varepsilon_i$. That model does allow unequal slopes so it was accepted for full credit here.

(d) Test whether $\beta_3 = 0$, either using a t-test or a model comparison.

- If you proposed $Y_i = \beta_0 + \beta_1 X_i + \beta_3 X_i Z_i + \varepsilon_i$ for part c and a model comparison here, you lost points here. This is because the models are not nested. Your "unequal slopes" model forces intercepts to be the same. To use model comparison to test equal slopes, both models have to allow unequal intercepts.

4. Computer processors

(a) $X_{\text{min}} = \frac{b_1}{2b_2} = 2.54$.

- most folks made the transition from finding the maximum (in class) to finding the minimum (the - goes away).

(b) $\hat{N} = \exp(\hat{X}_{\text{min}} + \log 32) = 408$.

- Some folks used log base 10 which gives an answer over 10,000. I intended to accept that answer because one form of the model had log base 10. See me if I deducted points that I shouldn't have.
- Some folks answered with the minimum Time, i.e. Y at X_{min}. That's answering the wrong question.

(c) The best approach is to test whether $N_{\text{min}} = 350$, i.e. $X_{\text{min}} = 2.392$ by testing whether $b_1 + 2\,2.392\,b_2 = 0$. You have what you need to compute the se because $\text{Cov}(b_1, b_2) = 0$ (in the problem text). I got T = 31.49 / 16.9 = 1.85. p> 0.05. The min is consistent with 350.

- I got lots of answers here, some were close and some were very close. I planned this to be the hardest Q on the exam, which is why it was last. When I read the first 10 tests and saw nobody getting close, I revised my grading scheme and was more generous in allocating points. When I subsequently saw very close answers, I gave bonus points (+something).
- The most frequent error was to compute a T statistic as $(\hat{N} - 350)/seX_{\text{min}}$. The numerator is on the N scale but the denominator is on the X scale.
- Calculating a T statistic as $(\hat{X}_{\text{min}} - 2.392)/seX_{\text{min}}$ assumes $\hat{X}_{\text{min}}$ is normally distributed. X_{min} is a ratio, so it is unlikely to be normally distributed. There is no guarantee that $(\hat{X}_{\text{min}} - 2.392)/se\hat{X}_{\text{min}}$ has a T distribution. That could be a problem in small samples.
- The best approach is to test $C = b_1 + 2 * 2.392\,b_2 = 0$ because the b's are normally distributed (under usual assumptions), so C is normally distributed and t-based inference is reasonable.