Stat 511 Final Exam

May 4, 2009
Prof. Vardeman

(**This exam will scored on a 160 point basis.**)

I have neither given nor received unauthorized assistance on this exam.

KEY
____
Name

____
Name Printed

1. A marketing study used as an example in Neter et al. concerned counts of customers visiting a particular lumber store during a two-week period from each of $n = 110$ different census tracts (these are metropolitan areas with populations of about 4000 residents each). Various demographic characteristics of the tracts were also obtained. Available for each tract are

y = the number of customers visiting the store from the tract

x_1 = the number of housing units in the tract

x_2 = the average personal income in the tract (dollars/year)

x_3 = the average housing unit age in the tract (years)

x_4 = the distance from the tract to the nearest competing store (miles)

x_5 = the distance from the tract to the store (miles)

Here we will model customer counts as independent Poisson variables with $\mathrm{E}y_i = \lambda_i$ and

$$\ln \lambda_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i}$$

There is an V output for these data attached to this exam. Use it to help you answer the following.

8 pts

a) In the presence of all other predictors, **which** of the x's appears to be **least important** to the description of y? **Explain**.

All tests of $H_0: \beta_j = 0$ have fairly small p-values, but $H_0: \beta_3 = 0$ has the largest p-value. x_3 is the apparently least important predictor.

8 pts

b) **What** are approximate 95% confidence limits for the log-mean number of visits by tract #1 customers in a two week period? **What** are corresponding approximate 95% limits for the mean number of such visits?

Use $\widehat{\log\lambda_1} \pm z(\widehat{s.e.\log\lambda_1})$ i.e.

$2.512666 \pm 1.96(.05467314)$ i.e.

$2.512666 \pm .1072$ i.e. $(2.4055, 2.6198)$

Limits for λ are $(e^{2.4055}, e^{2.6198})$ i.e. $(11.08, 13.73)$

10 pts

c) **Find** an appropriate prediction standard error for the number of visits in a future two week period by tract #1 customers. (Hints: $\widehat{\ln\lambda} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5$. How is a standard error for $\hat{\lambda} = \exp\left(\widehat{\ln\lambda}\right)$ related to a standard error for $\widehat{\ln\lambda}$? How is $\mathrm{Var}Y_{new}$ related to λ?)

$$\mathrm{Var}\left(y_{new} - \hat{\lambda}\right) = \lambda + \mathrm{Var}\,\hat{\lambda} \quad \text{and} \quad \widehat{\mathrm{Var}\left(y_{new} - \hat{\lambda}\right)} = \hat{\lambda} + \widehat{\mathrm{Var}\,\hat{\lambda}}$$

$$\mathrm{Var}\,\hat{\lambda} \approx \left(\exp(\ln\lambda)\right)^2 \mathrm{Var}\left(\widehat{\ln\lambda}\right) \quad \text{and} \quad \widehat{\mathrm{Var}\,\hat{\lambda}} = \left(\exp\left(\widehat{\ln\lambda}\right)\right)^2 \left(\text{s.e.}\,\widehat{\ln\lambda}\right)^2$$

So a standard error is

$$\sqrt{12.337775 + (12.337775)^2(.05467314)^2} = 3.577$$

8 pts

d) Census tract #41 has one of the largest values of y in the data set (and corresponding large value of $\hat{\lambda}$) and thus is an important source of customers for the store. A competitor is about to open a new store only .1 mile away from this tract. **By what fraction** does the fitted model suggest that the mean number of visits from tract #41 customers will decrease? **Provide** 95% confidence limits for this fraction.

$\ln\lambda$ will change by $\beta_4(.1 - 4.90) = -\beta_4(4.80)$ - So 95% limits for $(-4.80)\beta_4$ are -4.80 times limits for β_4 i.e

$$\left((-4.80)(.2187),\ (-4.80)(.1177)\right)$$

$$(-1.0498, -.5650)$$

This means that the new λ will be between $e^{-1.0498}$ and $e^{-.5650}$ times the previous one. That is, it will be between

.35 and .57

times the previous one.

2. Consider the making of fitted values $\hat{y}_i$ from n pairs (x_i, y_i) under a model that says

$$y_i = \mu(x_i) + \varepsilon_i$$

for some unknown mean function $\mu(x)$ where the ε_i are iid with mean 0 and variance σ^2. A standard measure of the flexibility of the fitting method employed is

$$flex = \frac{1}{\sigma^2}\sum_{i=1}^{n} \mathrm{Cov}(\hat{y}_i, y_i)$$

For "linear" fitting methods (ones for which $\hat{\mathbf{Y}} = \mathbf{MY}$ for some fixed $n \times n$ matrix $\mathbf{M}$) this is fairly easily computed. (Consider the covariance matrix of the $2n \times 1$ vector $(\hat{\mathbf{Y}}', \mathbf{Y}')'$ computed beginning from $\mathrm{Var}\mathbf{Y} = \sigma^2 \mathbf{I}$.)

10 pts

a) **What** is numerical value of *flex* for simple linear regression fitting? (Note that here, $\mathbf{M} = \mathbf{P_X}$, the projection matrix onto the column space of the simple linear regression $\mathbf{X}$ matrix.) **Explain**.

$$\mathrm{Var}\begin{pmatrix}\hat{Y}\\ Y\end{pmatrix} = \mathrm{Var}\begin{pmatrix}P_X\\ I\end{pmatrix}\sigma^2 I \begin{pmatrix}P_X & I\end{pmatrix} = \sigma^2\begin{pmatrix}P_X P_X & P_X\\ P_X & I\end{pmatrix} = \sigma^2\begin{pmatrix}P_X & P_X\\ P_X & I\end{pmatrix}$$

i.e. the covariance matrix for $\hat{Y}$ is $\sigma^2 P_X$ and

$$flex = \frac{1}{\sigma^2}\,\mathrm{trace}\,(\sigma^2 P_X) = \mathrm{trace}\, P_X = \mathrm{rank}\, X = 2$$

10 pts

b) Many popular smoothing methods are linear methods. In particular, kernel smoothing is a linear method. For a small problem where $n = 4$, $x_1 = 1, x_2 = 2, x_3 = 3,$ and $x_4 = 4$, the kernel $K(u) = \exp(-u^2)$ used with bandwidth $b = 2$, produces the matrix $\mathbf{M}$ given below. **What** is *flex* in this context? **Explain**.

$$\mathbf{M} = \begin{pmatrix} .4440 & .3458 & .1634 & .0468\\ .2662 & .3418 & .2662 & .1258\\ .1258 & .2662 & .3418 & .2662\\ .0468 & .1634 & .3458 & .4440 \end{pmatrix}$$

As above, $\mathrm{Var}\,\hat{Y} = \sigma^2 M$

and $flex = \frac{1}{\sigma^2}\,\mathrm{trace}\,(\sigma^2 M)$

$= \mathrm{trace}\,(M)$

$= 1.57$

3. Several nominally identical bolts are used to hold face-plates on a model of transmission manufactured by an industrial concern. Some testing was done to determine the torque required to loosen bolts number 3 and 4 on $n = 34$ transmissions. Since the bolts are tightened simultaneously by two heads of a pneumatic wrench fed from a single compressed air line, it is natural to expect the torques on a single face-plate to be correlated. Printout #2 concerns several aspects of the analysis of 34 pairs of measured torques (in ft-lbs).

8 pts

a) **What** are approximate 90% confidence limits for the mean difference of bolt 4 and bolt 3 torques? **Explain**.

From page 12 of printout, use the .05 and .95 quantiles of the bootstrap dsn of bolt4 − bolt3, i.e.

$(-.1176, 1.0588)$

8 pts

b) **What** is a bootstrap standard error for the ratio of sample standard deviations $s_{\text{bolt3}} / s_{\text{bolt4}}$? **Explain**.

From the top of page 13, use the standard deviation of the bootstrap dsn of $s_{\text{bolt3}} / s_{\text{bolt4}}$, i.e.

.1597

8 pts

c) **Why** would you expect the bootstrap to fail in the estimation of the upper .01 point for the bolt 4 torques in this situation?

The probability that 34 observations are all less than the upper .01 point of a cont's dsn is $(.99)^{34} = .71$. It's likely that no value in the sample is larger than the θ of interest. Bootstrap samples will then never have values larger than θ. It's not even clear how one would try to define an upper .01 point of 34 observations. !!!

4. The book *Statistical Analysis of Designed Experiments* by Tamhane considers an experiment run to study the corrosion resistance of 4 types of coating for steel bars. Steel bars were coated, baked, and tested for corrosion resistance as follows. An oven was set to one of 3 different temperatures (360°F, 370°F, or 380°F), 4 bars (one coated with each different coating) were loaded into the oven, all were baked for a fixed time, the bars were then removed and cooled, and corrosion testing was done (no units of measurement are stated for the response variable). After running the oven at each temperature once, the whole protocol was repeated. Let

y_{ijk} = a measured corrosion resistance of coating i under baking temperature j
seen in the kth time the furnace is heated

and consider the model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \gamma_k + \varepsilon_{ijk} \qquad (*)$$

for fixed effects $\mu, \alpha_i, \beta_j, \alpha\beta_{ij}$ (Factor A being Coating Type and Factor B being Temperature), and random effects γ_k and ε_{ijk} that are independent mean 0 normal variables, the γ_k with variance σ_γ^2 and the ε_{ijk} with variance σ^2. The γ_k are "firing" effects potentially peculiar to each different time $k = 1, 2, \ldots, 6$ that the oven is heated.

If one defines $l(1) = l(2) = l(3) = 1$ and $l(4) = l(5) = l(6) = 2$ the variable $l(k)$ specifies the replication (first or second) of which firing k is a part. In the event that there was a substantial time period between replications 1 and 2 of the experiment, it might make sense to entertain a generalization of model (*)

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \delta_{l(k)} + \gamma_k + \varepsilon_{ijk} \qquad (**)$$

for δ_1 and δ_2 iid $\mathrm{N}(0, \sigma_\delta^2)$ independent of the γ_k and ε_{ijk}.

There is an V output based on Tamhane's data attached to this exam. Use it as needed to help you answer the following questions.

10 pts

a) **Exactly what** about the design of this study guarantees that the variance component σ_δ^2 of model (**) will be poorly determined? (Use 25 words or less, and do NOT to refer to any number from the data analysis on the output to answer this question.)

There is only a "sample" of 2 replications (2 δ's) to use in estimating σ_δ^2. That's a tiny sample to use in estimating a variance.

12 pts

b) The fundamental difference between models (*) and (**) is in their covariance structures. In terms of the variance components $\sigma_\gamma^2, \sigma_\delta^2$, and σ^2 **fill in the table** below comparing the models.

	Model (*)	**Model (**)**
Var y_{ijk}	$\sigma_\gamma^2 + \sigma^2$	$\sigma_\gamma^2 + \sigma_\delta^2 + \sigma^2$
covariance between two y's from different replications	0	0
covariance between two y's from the same replication but different firings	0	σ_δ^2
covariance between two y's from the same firing	σ_γ^2	$\sigma_\gamma^2 + \sigma_\delta^2$

12 pts

c) Based on the V output, **argue carefully** that REML can't really distinguish between models (*) and (**) here. (Your answer to b) might be useful.) The deviance values for the 2 fits are the same (189.2) to the number of digits displayed. Further, plugging estimates of variances from the printout into the formulas above produces essentially the same values left and right. (For all practical purposes $\widehat{\sigma_\delta^2} \approx 0$ and the values of $\widehat{\sigma_\gamma^2}$ and $\widehat{\sigma^2}$ are the same for the 2 fits.)

Henceforth use model (*).

12 pts

d) The two from observations from coating i under baking temperature j could possibly be used to compute a (sample size 2) sample variance s_{ij}^2. **Identify** a constant c and degrees of freedom ν so that $cs_{ij}^2 \sim \chi_\nu^2$. **Why** can one not conclude that $\sum_{i,j} cs_{ij}^2 \sim \chi_{6\nu}^2$?

12ν was intended, not 6ν

The 2 observations have the same mean but different γ's and ϵ's. So

$$\frac{(2-1)s^2}{\sigma_\gamma^2+\sigma^2} \sim \chi_1^2$$

$c = \left(\frac{2-1}{\sigma_\gamma^2+\sigma^2}\right)$ $\nu =$ 1

Why?:

These s_{ij}^2 are at least not obviously independent. Any pair with a given j are built on observations sharing common γ's. If they were, this sum would be χ_{12}^2. Some care would be required to conclude whether or not they are in fact independent.

endpoints 0 and 1.213842

12 pts

e) **Is there** definitive statistical evidence which of the two standard deviations σ_γ and σ is largest? **Explain**.

No. The interval for $\frac{\sigma_\gamma}{\sigma}$ on page 15 covers 1 (contains values both less than 1 and larger than 1).

NOTE: On HW6 I said that the MCMC sampling produces a credible interval for σ (which it does) and credible limits for the ratios of the other std deviations to σ. More correctly, it produces multipliers (left and right) that can be used with the interval for σ to get intervals for the other std deviations. (This is a slightly subtle difference from what I said earlier) The interval for σ_γ is thus $(0(12.68), 1.214(39.12))$.

12 pts

f) **Is there** definitive statistical evidence of a difference in Temperature 1 and Temperature 2 main effects? **Explain**.

$$\hat\beta_1-\hat\beta_2 = -44.50-(9.125) = -53.625$$

$$\text{s.e.}(\hat\beta_1-\hat\beta_2) = \sqrt{(1,-1)\begin{pmatrix}195.91 & -97.95\\ -97.95 & 195.91\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix}} = 20.03$$

An approximate 95% CI for $\beta_1-\beta_2$ is

$$-53.625 \pm \underbrace{1.96(20.03)}_{39.26}$$

By this standard, there is a statistically detectable difference in β_1 and β_2.

12 pts

g) If the object is to maximize mean corrosion resistance, **what combination of level**s of Coating and Temperature is indicated to be best by the study? Explain.

We look for the maximum value of $\hat\alpha_i+\hat\beta_j+\widehat{\alpha\beta}_{ij}$

$\hat\alpha_1=-6.46$ $\hat\alpha_2=-10.96$ $\hat\alpha_3=-5.46$ $\hat\alpha_4=\boxed{22.88}$

$\hat\beta_1=-44.5$ $\hat\beta_2=9.125$ $\hat\beta_3=\boxed{35.375}$

$\widehat{\alpha\beta}_{11}=-.167$ $\widehat{\alpha\beta}_{12}=-1.292$ $\widehat{\alpha\beta}_{13}=1.459$

$\widehat{\alpha\beta}_{21}=-5.167$ $\widehat{\alpha\beta}_{22}=17.208$ $\widehat{\alpha\beta}_{23}=-12.041$

$\widehat{\alpha\beta}_{31}=13.333$ $\widehat{\alpha\beta}_{32}=-.792$ $\widehat{\alpha\beta}_{33}=-12.541$

$\widehat{\alpha\beta}_{41}=-7.999$ $\widehat{\alpha\beta}_{42}=-15.124$ $\widehat{\alpha\beta}_{43}=\boxed{23.123}$

Clearly, Coating 4 and Temperature 3 maximize this sum