KEY

Stat 511 Final Exam

May 5, 2004 **Prof. Vardeman**

22pts 1. In a study meant to determine the variability of diameters of widgets produced on a manufacturing line, an engineer measures $m = 10$ widgets produced on the line once each. Then the engineer measures the diameter of an 11^{th} widget $n = 8$ times. Suppose one models a measured widget diameter, y, as

$$y = x + \varepsilon$$

where x is the true diameter of the particular widget and ε is measurement error, for $x \sim \mathrm{N}(0, \sigma_x^2)$ independent of $\varepsilon \sim \mathrm{N}(0, \sigma^2)$. With s_y^2 the sample variance of the measurements on the 10 widgets, $\mathrm{E}s_y^2 = \sigma_x^2 + \sigma^2$, and with s^2 the sample variance of the repeat measurements on the 11^{th} widget, $\mathrm{E}s^2 = \sigma^2$. If the engineer observes $s_y = .05$ mm and $s = .01$ mm, find approximate 90% confidence limits for σ_x. (Hint: Cochran-Satterthwaite.) Consider the difference in "mean squares" $s_y^2 - s^2$. This has mean σ_x^2. Cochran-Satterthwaite "estimated d.f." are then

$$\hat{\nu} = \frac{\left((.05)^2 - (.01)^2\right)^2}{\frac{\left((.05)^2\right)^2}{9} + \frac{\left((.01)^2\right)^2}{7}} = 8.3$$

Rounding down to 8, since $\sqrt{s_y^2 - s^2} = .0490$, approximate confidence limits are then $\left(.0490\sqrt{\frac{8}{15.507}}, .0490\sqrt{\frac{8}{2.733}}\right)$ i.e. $(.0352, .0838)$

2. Printout #1 concerns the analysis of data taken from an article of Chowdhury and Mitra on a study intended to reduce defects produced by a wave soldering machine in an electronics plant. Numbers of "shorts" and numbers of "dry solder" defects were counted on a standard unit of product, as levels of 10 two-level factors (that we here simply call A through J) were changed. The data on the printout concern counts for 32 (out of $2^{10} = 1024$ possible) combinations of levels of these and a GLM analysis using a Poisson model and the canonical link function $h(\mu) = \ln(\mu)$.

6pts a) The first two models fit (to "shorts" and "dry solder") include main effects for the 10 two-level factors A through J. Why would it be hopeless to try to fit a model with main effects and all 2 factor interactions?

There are $\binom{10}{2} = 45$ pairs of factors, and (since we are dealing with 2-level factors) therefore 10 main effects plus 45 2 factor interactions to estimate based on only 32 observations ... all 55 of these (plus an overall "mean") simply can not be "estimable". (The model matrix could <u>not</u> be full rank.)

Based on the results for the first two models, the smaller 3^{rd} and 4^{th} models were fit. Use these smaller models to answer the rest of the questions about this scenario.

10pts b) Which factor seems to have the biggest effect on shorts? Which factor seems to have the biggest effect on dry solder defects?

biggest effect on shorts: D

biggest effect on dry solder defects: H

c) For the 3rd model (the smaller model fit to "shorts") the first fitted/predicted value is 7.968834. Show how this is obtained from the vector of estimated coefficients **β**. (Show calculation of this value.)

8pts

Since the link is the log link, we have

$$\widehat{\ln\mu} = 2.30714 + (-(-.10820)) + (-.23160) + (-.10820)$$
$$= 2.0755$$

and so $\hat{\mu} = e^{2.0755} = 7.9688$

8pts d) Some R code (and what is returned by the program) in this context is below. Use the result and show why a sensible `se.fit` for the first fitted value is 0.9388530. (Hints: "delta method" and "link function.")

```
> t(c(1,-1,1,-1))%*%vcov(glm.out3)%*%c(1,-1,1,-1)

        [,1]
[1,] 0.01388052
```

$\widehat{\mathrm{Var}\,\widehat{\ln\mu}} = .01388052$

Then $\hat{\mu} = \exp(\widehat{\ln\mu})$, so $\mathrm{Var}\,\hat{\mu} \approx \left(\frac{d}{dx}\exp(x)\Big|_{E\widehat{\ln\mu}}\right)^2 \mathrm{Var}\,\widehat{\ln\mu}$

which could be estimated as $(\exp(2.0755))^2(.01388052)$ and

$\sqrt{\text{above}} = .9388530$

8pts e) As it turns out, only factors A through I were factors whose levels could be set on the wave soldering machine. (Factor J was a factor whose levels fluctuated beyond the control of the process operators.) What levels of Factors C,D,E, and H do you recommend for future operation of the machine, if one wishes to minimize the total of shorts and dry solder defects at the worst level of J? What do you estimate to be the mean total under the conditions you recommend? only this affects both shorts and dry solder

C level 1 D level 2 E level 2 H level 2

comparing predictions for cases 23 and 26

estimated mean *shorts + dry solder defects* at the worst level of J 10.199 + 40.016 = 50.215

predictions for case #26

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 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 to be correlated. Printout #2 concerns the estimation of the correlation based on these 34 pairs.

7pts a) Give a point estimate of the "population" correlation between Bolt 3 torque and Bolt 4 torque, a bootstrap standard error for that estimate, and a "bias-corrected" version of the estimate.

bias-corrected estimate $= 2T_n - \overline{T}_n^*$

$= 2(.4702) - .4712$

estimate .4702 standard error .1315 bias-corrected estimate .4692

9pts b) What are 90% bootstrap percentile confidence limits for the population correlation? (Report two numbers.)

lower limit .2422 upper limit .6732

9pts c) As it turns out, the "acceleration factor" for the bootstrap samples represented on the printout is $\hat{a} = -.02290961$. The lower limit of a 90% BC_a confidence interval for the population correlation is at approximately which percentile of the bootstrapped correlations? (Show your work.)

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 - z_{\text{upper}}}{1-\hat{a}(\hat{z}_0 - z_{\text{upper}})}\right)$$ Now, about 47.3% of the bootstrapped T_n^*'s are less than $T_n = .4702$. So $\hat{z}_0 \approx -.067$ (as a matter of fact, a more detailed R report shows that $\hat{z}_0 \approx -.049$). Thus

$$\alpha_1 = \Phi\left(-.067 + \frac{-.067-1.645}{1-(-.023)(-.067-1.645)}\right) = \Phi(-1.849) = .032$$

3.2 percentile

4. Printout #3 concerns the analysis of some very old data of Rumford taken to study the cooling of a large cannon barrel that had been heated to a uniform (high) temperature by friction. Newton's Law of cooling predicted that when the source of heating was removed, the temperature would decline exponentially to the ambient temperature (that was apparently about 60°F on the day these data were collected). The model fit to the temperatures (in degrees F) and times (in minutes from the cessation of heating) is

$$y_i = \alpha + \beta \exp(-\theta t_i) + \varepsilon_i$$

12pts a) Give approximate 90% prediction limits for an unrecorded temperature at 45 minutes based on this model. (Plug numbers in everywhere, but you need not simplify.)

$$\hat{y} = 106.2 + 22.92\, e^{-(.04095)(45)} = 109.8$$

$$e^{-\hat{\theta}(45)} = .1584, \quad -\hat{\beta}(45)e^{-\hat{\theta}(45)} = -163.4$$

$$\frac{\partial y}{\partial \alpha} = 1, \quad \frac{\partial y}{\partial \beta} = e^{-\theta t}, \quad \frac{\partial y}{\partial \theta} = -\beta t e^{-\theta t}$$

So prediction limits are

$$109.8 \pm 1.812\sqrt{(.3418)^2 + (1, .158, -163)\begin{pmatrix} 1.03 & -.72 & .004 \\ -.72 & .59 & -.003 \\ .004 & -.003 & .00003 \end{pmatrix}\begin{pmatrix} 1 \\ .158 \\ -163 \end{pmatrix}}$$

6pts b) Notice from the plot on the printout, that over the first 45 minutes or so of cooling the fitted curve does a nice job of tracking the plotted points. There is, however, clear statistical evidence that this trend can not continue. Explain carefully/statistically. (The real temperature series had to eventually decline to about 60°F ... is that consistent with what has thus far been observed?)

An approximate t test of $H_0: \alpha = 60$ here has

$$T = \frac{\hat{\alpha} - 60}{s.e._{\hat{\alpha}}} = \frac{106.2 - 60}{1.015} = 45.5$$

which is huge. The p-value is tiny. This curve has an asymptote nowhere near 60 °F.

5. Printout #4 concerns the analysis of some data from a test of thermal battery lives. Batteries using 4 different Electrolytes were tested at 3 different Temperatures, in 9 different Batches/Tests, over a period of 3 Days (3 Batches per day). Each batch/test was run at a single temperature and included one battery of each electrolyte. Let

y_{ij} = the life of the battery with electrolyte i in test/batch j

$d(j)$ = the number of the day on which test/batch j is run

$t(j)$ = the level of temperature at which test/batch j is run

Consider fixed effects $\mu, \eta_i, \tau_t, \eta\tau_{it}$ (say, satisfying the sum restrictions) where the η's represent Electrolyte main effects, the τ's represent Temperature main effects, and the $\eta\tau$'s the Electrolyte by Temperature interactions. Suppose that $\beta_1, \beta_2, \ldots, \beta_9$ are iid $\mathrm{N}(0, \sigma_\beta^2)$, independent of $\delta_1, \delta_2, \delta_3$ that are iid $\mathrm{N}(0, \sigma_\delta^2)$, independent of ε_{ij}'s that are iid $\mathrm{N}(0, \sigma^2)$. We will work under the model

$$y_{ij} = \mu + \eta_i + \tau_{t(j)} + \eta\tau_{it(j)} + \beta_j + \delta_{d(j)} + \varepsilon_{ij}$$

(In addition to the fixed effects, there are random day and batch/test effects, along with the basic experimental error, ε.)

10pts a) Give approximate 95% confidence intervals for σ and σ_β.

for σ (.116, .224) for σ_β (.088, .350)

6pts b) How is it obvious from the printout that σ_δ is poorly determined? In retrospect, why should that be no surprise? The interval for σ_δ covers a ridiculous 50 orders of magnitude! This isn't terribly surprising because there are only 3 different days involved here. Estimating a variance / std deviation based on 3 observations is close to hopeless, even in 1-sample problems that don't involve all the other effects present in this complex situation.

10pts c) A bit of the printout is duplicated below. This gives 3 kinds of predictions for the first case in the data set. Show how to get each of these 3 predictions from other things on the printout. (Show sums that give these values.)

$\hat{\mu}$ $\hat{\eta}_1$ $\hat{\tau}_1$ $\widehat{\eta\tau}_{11}$

```
> predict(battlife,level=0:2)
  Day Test predict.fixed predict.Day predict.Test
1   1  1/1      1.890000    1.890073     2.128209
```

$1.89 = 1.785 + (.164) + (-.006) + (-.053)$

$1.890073 = 1.890000 + .000073$ $\hat{\delta}_1$ (approximate BLUP)

$2.128209 = 1.890073 + .23813565$ $\hat{\beta}_1$ (approximate BLUP)

d) A 10^{th} batch/test will be run tomorrow. Temperature 1 will be used. What do you predict for a battery life for the battery with electrolyte 1 and what is a standard error of prediction?

10pts The response will be $y_{new} = \mu + \eta_1 + \tau_1 + \eta\tau_{11} + \beta_{new} + \delta_{new} + \epsilon_{new}$

A predictor will be $\hat{\mu} + \hat{\eta}_1 + \hat{\tau}_1 + \widehat{\eta\tau}_{11} = 1.890000$

Note that $\hat{\mu}, \hat{\eta}_1, \hat{\tau}_1, \widehat{\eta\tau}_{11}$ are uncorrelated, so a std error will be

$$\sqrt{(.064)^2 + (.047)^2 + (.092)^2 + (.066)^2 + (.003)^2 + (.176)^2 + (.161)^2} = .28$$

$\hat{\sigma}_\delta$ $\hat{\sigma}_\beta$ $\hat{\sigma}$

prediction 1.890000 standard error of prediction $.28$

6. Miscellaneous Linear Models

12pts a) Write a matrix $\mathbf{C}$ so that with $\boldsymbol{\beta} = (\mu_{11}, \mu_{12}, \mu_{13}, \mu_{21}, \mu_{22}, \mu_{23}, \mu_{31}, \mu_{32}, \mu_{33})$ the testable hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ for (the cell means version of) the 3×3 factorial is the hypothesis that $\alpha_1 = \alpha_2$ and simultaneously $\beta_1 = \beta_3$.

$$C = \begin{pmatrix} 1 & 1 & 1 & -1 & -1 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & -1 \end{pmatrix}$$

7pts b) Carefully convince me using a simple "column space and projection" argument that simple linear regression for n points (x_i, y_i) gives the same the same predicted values as using the n points $(x_i - \bar{x}, y_i)$ (for $\bar{x}$ the sample mean of the x_i).

In "raw x" form, The X matrix here is $\begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$ while in "centered x" form it is $\begin{pmatrix} 1 & x_1 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix}$. These two have the same column spaces (note $\underset{\sim}{x} = \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix} + \bar{x}\mathbf{1}$). Since $\hat{Y}$ is the projection of Y onto $C(X)$, The two formulations give The same fitted/predicted values.