1. Growth rates of pigs

   (a) 10 pts. Here's my skeleton ANOVA / model. Others were accepted if they 1) included terms for the consistency of treatments across locations and 2) included a term for variability among pens. Location, diet, and monensin are crossed; pens are nested within location. If this were done at a single location, it would be a split plot with diets randomly assigned to pens (no blocks).

| Source | df | Comments |
|---|---|---|
| location | 2 | |
| diet | 4 | |
| diet*loc | 8 | random, because want broad sense inference (see 1b) |
| pen(diet*loc) | 30 | or pen(loc) - equivalent |
| monensin | 1 | |
| diet*mon | 4 | |
| location*diet*mon | 10 | could separate into loc*mon and loc*mon*diet<br>but both depend on consistency of monensin effects<br>so I would probably pool |
| mon*pen(diet*loc) | 30 | prob. pool with error |
| error = pig(trt) | 90 | truly is var between pigs receiving the same treatment<br>(pen, diet, mon) if you separate out mon*pen(diet*loc) |

   (b) 5 pts. Q tells you to do broad sense inference. Hence loc*diet at a minimum, which forces pen(diet*loc) to be random. Should also have loc*mon*diet. loc is optional.

   (c) 5 pts. diet*loc (because want broad sense inference)

2. Study time

   (a) 5 pts. Either:

$$\begin{aligned} Score_i &= S_{max} + \beta_2(H_{max} - Hours_i)^2\, I(Hours_i < H_{max}) + \varepsilon_i, \text{ or} \\ Score_i &= \begin{cases} S_{max} + \beta_2(H_{max} - Hours_i)^2 + \varepsilon_i & Hours_i \le H_{max} \\ S_{max} & Hours_i > H_{max} \end{cases} \end{aligned}$$

   This is a quadratic to a constant response model with unknown location of the maximum.

   (b) 5 pts. $\hat{S}_{max} = 79.6$, $\hat{H}_{max} = 63.1$, $\hat{\beta}_2 = -0.010$.
   Look at the se to decide how many digits to report. The se of $\hat{S}_{max}$ is 1.49. If you report 79.62119, you certainly do not know the 119 bit. I accepted up to 3 digits past the decimal point, but deducted points for reporting more digits.

   (c) 5 pts. I accepted either the Wald or T intervals:
   Wald: $63.1 \pm 1.960 \times 8.0 = (47.4,\ 78.8)$
   T: $63.1 \pm 2.074 \times 8.0 = (46.5,\ 79.7)$

   (d) 5 pts. The profile Sums-of-Squares trace must not be quadratic around $\hat{H}_{max}$.
   If you said the methods were different, you got a bit of credit (that is not sufficient because if the SS trace is quadratic, the methods are still different but the intervals are the same).

   (e) 5 pts. Profile. It makes fewer assumptions than the Wald.
   Some folks said Wald because it was shorter. That is a good property only if the interval

maintains the specified coverage (e.g. 95%). In this case, the Wald interval will not have 95% coverage.

3. Penalized splines

   (a) 5 pts. $20.66 = n - 2tr(S_{\lambda^2}) + tr(S'_{\lambda^2} S_{\lambda^2})$
   The error df **is not** $n -$ model df.

   (b) 5 pts. F = 46.5, central F with 2.783, 20.66 df. The calculations are:

| Model | df model | SS error |
|---|---|---|
| Intercept | 1 | 3718.1 |
| pen. spline | 3.783 | 512.3 |
| Difference | 2.783 | 3205.8 |

   so the model MS = 1,151.9. The error MS = 512.3 / 20.66 = 24.8 (Note use of model df in computing the model MS and error df in computing error MS). I deducted 1 point if you got F = 38.76 by using $n$ - error df to calculate model MS. I deducted 4 points if you thought the numerator df was 1.

   (c) 5 pts. Similar calculations get you F = -1.78 or -0.95, depending on your model df.

   (d) The quadratic model is not nested in the penalized spline model. Can not use a model comparison F test for non-nested models.
   Some folks said "different fixed effects". That's true but not sufficient. Any model comparison testing a fixed effect has models with different fixed effects.

4. 7 subject areas study.

   (a) 10 pts.

$$\begin{aligned}
Score_{ij} &= Smax_i + b_2(Hmax_i - hours_{ij})^2 I(hours_{ij} < Hmax_i) + \varepsilon_{ij} \\
i &\quad \text{identifies the subject area} \\
j &\quad \text{identifies the person within subject area} \\
Smax_i &\sim N(S_{max}, \sigma_s^2) \\
Hmax_i &\sim N(H_{max}, \sigma_h^2) \\
\varepsilon_{ij} &\sim N(0, \sigma_e^2)
\end{aligned}$$

   You could also separate the random effects into two pieces, a fixed effect and a random effect centered at 0

   (b) 5 pts. 14, 7 $Smax_i$ and 7 $Hmax_i$.
   My hint about columns of the $Z$ matrix was intended to steer you away from thinking about two random terms in the model.

   (c) 5 pts. $-2\Delta\text{lnL} = -2(-545.90 - (-545.25)) = 1.30$. $\chi^2_1$
   The output includes lnL values, so you can construct a LRT. You don't have the information for any other type of test. This null hypothesis **is not** on the boundary because a covariance can be positive or negative. Quite a few folks forgot the -2!

   (d) 5 pts. $-2\Delta\text{lnL} = -2(-602.09 - (-545.90)) = 112.38$. $0.5\chi^2_0 + 0.5\chi^2_1$
   This null hypothesis is on the boundary, so you have to use the adjusted distribution.

   (e) (No part e)

(f) 5 pts. $\widehat{score} = 96.075$
1) You need to use the blup's as the estimated parameters.
2) In this case, $hours_{ij} > Hmax_i$, so you don't want the quadratic piece.

(g) 5 pts. $\widehat{score} = 78.74 - 0.00816(72.196 - 30)^2 = 64.21$
You have no data for physics students, so you use the fixed effects (i.e., $\hat{S}max_i|data = S_{max} + 0$)

5. contagious bovine pleuropneumonia

(a) 5 pts. cbpp.m2: it has the smallest AIC
Some folks discussed what components they believed should be in the model. That's a very good exercise, but you need to consider what the data tell you. I accepted a verbal "what should be in the model" if you looked at estimated overdispersion and decided the data were overdispersed (which they are).

(b) 5 pts. Test statistic = 125.7 - 100.1 = 25.6, $\chi^2_3$
R reports both logLik and deviance. 125.7 is deviance for the model without period effects; 100.1 is that for the model with. This is a test of fixed effects, df = change in number of parameters, which is 4 - 1.

(c) 5 pts. 0.14 = -0.992 - (-1.129) (if using the R output) Same answer, different numbers in the calculation if using the SAS output
1) You are asked for a subject-specific estimate, so you need the GLMM output.
2) The R output includes parameter estimates for period 2 - period 1 = -0.992 and period 3 - period 1 = -1.129. The difference is the required quantity.
3) More than a few folks forgot that parameters in a logistic regresion are measured on the logit scale, so differences between parameters **is** the log odds ratio. The model form doesn't change when you add the random effect.

(d) 5 pts. $\hat{\pi} = \frac{1}{1+\exp(-(-1.269))} = 0.219$
1) You are asked for a population average, so you need GEE results.
2) Using GEE output, the estimated logit probability for period 1 = -1.269 (the estimated intercept + 0 in R)
3) Some folks forgot one of the minus signs, which gets you $\hat{\pi} = 0.78$. That would scare the minister of animal health! 4) I was thinking about the probability when I wrote the question. Incidence means "number of newly infected cattle" to some folks. That is easily obtained as $N\hat{\pi}$ if you know the total number of uninfected cattle. I was happy with just the probability.

There are 10 points for free.