

# Topics in Applied Regression, LTH, 2007

## Take-home exam exercise

John Maindonald

June 4, 2007

Marks (~15%) will be given for insightful comments!

For marking before I leave on June 17, the due date is June 11, with the possibility of extension to June 13. If this deadline is too tight, please contact me and organize to send me your submission by June 22 at the latest.

**A. Linear regression exercise** Consider the two models

1.  $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1$
2.  $\mathbf{y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2$

where it is assumed that  $\boldsymbol{\epsilon}_1 \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$ , and similarly for  $\boldsymbol{\epsilon}_2$ .

M1 is nested within M2 if the columns of  $\mathbf{X}_1$  are a subset of those of  $\mathbf{X}_2$ . More generally, it is enough that the columns of  $\mathbf{X}_2$  are a basis for the column space of  $\mathbf{X}_1$ . Suppose that  $\mathbf{X}_2$  is  $n \times k_2$  and that  $\mathbf{X}_1$  is  $n \times k_1$ . Under these conditions, assuming that M1 is the true model, the difference in the residual sum of squares is distributed as  $\sigma^2\chi_\nu^2$ , where  $\nu = k_2 - k_1$ .

What however happens if M2 is not nested within M1? We can investigate this, in specific cases, using a bootstrap or permutation distribution approach.

1. For the hillraces data (data set `racess2000`<sup>1</sup> in the *DAAG* package, but limiting attention to hill races), fit the three models:
  - (a) P1 is the sum of polynomial of degree 2 in  $\log(\text{dist})$  and a linear term in  $\log(\text{climb})$
  - (b) M1 is the sum of natural spline terms of degree 2 in  $\log(\text{dist})$  and of degree 3 in  $\log(\text{climb})$ .
  - (c) M2 is the sum of natural spline terms of degree 3 in  $\log(\text{dist})$  and of degree 4 in  $\log(\text{climb})$ .

Summarize the results in graphical form, e.g., use `termplot()`.

[10 marks]

2. These three models are not nested. A reasonable way to assess how close P1 is to a model that is nested in M1 may be to express each of the columns of the model matrix of P1 as a linear combination of the columns of the

---

<sup>1</sup>In version 0.93 of *DAAG* or later, these are the data in `hills2000`.

model matrix for M1 and calculate the statistic  $1 - R^2$  for each column. (This is a rare occasion when something akin to  $R^2$  may be useful!). For example:

```
P1 <- with(lhills2k, model.matrix(~ poly(ldist,2) + lclimb))
M1 <- with(lhills2k, model.matrix(~ ns(ldist,3) + ns(lclimb,2)))
qrM1 <- qr(M1)
qdashP1 <- qr.qty(qrM1, P1[, -1])
p <- dim(M1)[2]
r2P1M1 <- 1 - apply(qdashP1, 2, function(x)sum(x[-(1:p)]^2)/sum(x[-1]^2))
print(r2P1M1)
```

For more flexible use, create a function that does this calculation for any pair of models. Use it to compare M1 with P1 and to compare M2 with M1. Annotate the code to explain each step.

[10 marks]

- Using (a) bootstrap sampling, and (b) the permutation distribution, obtain  $p$ -values i) for the comparison between M1 and P1, and ii) for the comparison between M2 and M1. Compare the  $p$ -values with the theoretical values for the case when the models are nested. Comment on your results.

[10 marks]

- Use the function `gam()` in *mgcv* to fit a natural spline model. Use the permutation test approach to compare this GAM model with M2.

[10 marks]

- Choose one of the models and examine model diagnostics.

[10 marks]

**[Part A total is 50 marks]**

**B. Multi-way table & Logistic Regression Exercise** This will use the dataset `nassCDS`, in the *DAAGxtras* package.

- Use the function `excessRisk()` to calculate, for the various categories, the excess risk from airbags when account is taken of both of the remaining factors `seatbelt` and `dvcat` (a speed of impact measure). Use bootstrap sampling to estimate confidence intervals for the excess risk in each of the multi-way categories.

[15 marks]

- Compare the estimates and standard errors with the theoretical estimates when a model is fitted using `glm()`, assuming that errors are binomial.

[10 marks]

**[Part B total is 25 marks]**

**C. Viva Voce** Please organize a time (20-30 minutes) on June 11, or on one of the days June 13-15. June 12, early in the day, may also be possible.

**[25 marks]**