

**TA** Meng Song  
**Office** 1270 MSC  
**Phone** 262-3230  
**Email** songm@stat.wisc.edu  
**Office hour** 11:00-12:00 AM and 1:00-2:00 PM Tue.  
**My Website** www.stat.wisc.edu/~songm

## 1 Question 5 of HW6

Model the velocity with these five increasingly complicated nested models.

- (a)  $\sim 1$
- (b)  $\sim \text{sqrt}(\text{loading})$
- (c)  $\sim \text{sqrt}(\text{loading}) + \text{symmetry}$
- (d)  $\sim \text{sqrt}(\text{loading}) + \text{genus}$
- (e)  $\sim \text{sqrt}(\text{loading}) + \text{species}$

The most complicated model is:

$$E[\text{velocity}] = \beta_1 + \beta_2 * \text{sqrt}(\text{loading}) + \beta_3 * 1_{\text{greenash}} + \beta_4 * 1_{\text{redmaple}} + \beta_5 * 1_{\text{silvermaple}} + \beta_6 * 1_{\text{sugarmaple}} + \beta_7 * 1_{\text{tuliptree}} + \beta_8 * 1_{\text{whiteash}}$$

Use the anova function in R to compare the five models. For each p-value in the anova table, specify the null and alternative hypotheses in terms of the parameters  $\beta_1, \dots, \beta_8$  and the reference distribution on which the p-value is based.

R output:

```

> anova(lm1,lm2,lm3,lm4,lm5)
Analysis of Variance Table

Model 1: velocity ~ 1
Model 2: velocity ~ sqrt(loading)
Model 3: velocity ~ sqrt(loading) + symmetry
Model 4: velocity ~ sqrt(loading) + genus
Model 5: velocity ~ sqrt(loading) + species

   Res.Df    RSS   Df Sum of Sq    F Pr(>F)
1       201 290900
2       200 254270   1     36630 356.15 <2e-16 ***
3       199  20085   1    234185 2276.98 <2e-16 ***
4       198  20082   1         3   0.03  0.87
5       194 19953   4        130   0.32  0.87
  
```

## 2 Simulation for a generalized linear model

Data from an experiment to see how an educational program on the importance of using gloves affected the rate of glove use by a group of nurses in an inner-city pediatric hospital emergency department. Without their knowledge, the nurses were observed during vascular access procedures before and one, two, and five months after an educational program to see how often they wore gloves. Each procedure by a nurse was counted as a separate observation.

Missing values are indicated by large dots.

1. Period: Observation period (1 = before intervention, 2 = one month after intervention, 3 = two months after )
2. Observed: Number of times the nurse was observed
3. Gloves: Number of times the nurse used gloves
4. Experience: Years of experience of nurse

See <http://lib.stat.cmu.edu/DASL/Datafiles/Nurses.html> for details.

Here we want to see the effect of educational program and nurse's experience, therefore we will ignore the effect of individuals and use Period and Experience as explanatory variables.

(a) Fit the data by a generalized linear model. Simulate 1000 reasonable values of  $\beta$  and  $\sigma$  from the model you fit. Compare the coefficient estimates and the corresponding standard deviation.

(b) What's the probability that a five-year experienced nurse wears gloves before intervention? What's the probability this nurse wears gloves one month after intervention? Compare the probabilities with those you get from the simulation.

## 3 Checking models

Suppose we have two groups of individuals with sample sizes 150 and 100. The model is  $y = \mu + \beta_1 * x_1 + \beta_2 * x_2 + \epsilon$ , where  $x_1 \sim \text{exp}(5)$ ,  $x_2 \sim N(3.5, 2.7)$ , and the intercept is 12 for F group and 7 for M group. The error standard deviation is 11. Now use fake data to check the coefficient estimates.