Miniproject: Four small projects

Ulrich Halekoh and Søren Højsgaard Created: November 7, 2011

1 Diets of Rats

TOPICS: Linear model with quadratic terms in x; calculate p–value by hand.

The growth rate, y, of rats was studied for different amounts of supplement to their diet, x. Data for this exercise can be found in the data frame 'dietsup' of the course package `LiSciData`.

1. Read in the data and plot the rate of growth as a function of the amount of supplement.

```
data(dietsup, package="LiSciData")
head(dietsup)
```

```
   x  y
1 10 73
2 10 78
3 15 85
4 20 90
5 20 91
6 25 87
```

2. Based on the plot, a model where the growth rate is related to the amount of supplement through a quadratic equation could be reasonable. Formulate and fit such a model with the `glm`-function.

   What are the estimates of the parameters in this model?

3. Add the fitted values to your plot of data.

4. Consider the regression model without the quadratic term. Write down this model. Fit the model and plot the residuals against the predicted values for model checking.

5. Test whether there is a need for a quadratic term in the model using the `anova`-function.

6. The test statistic you just found is calculated in the following way. In the column `Deviance` of the output of `anova` one finds actually the **difference** in the residual deviances $D_0 - D_1$ = 686.4-45.2=641.2 Hence the F-test statistic is

$$\frac{(D_0 - D_1)/(7-8)}{D_1/7} = \frac{641.2/(8-7)}{45.2/7} = 99.3$$

You can calculate the p-value 'by hand'. You calculate the probability that an F-distributed random variable with degrees of freedom 1 and 7 is larger than 99.3:

```
pf(99.3,1,7,lower.tail=FALSE)
```

```
[1] 2.189525e-05
```

A histogram of 1000 observations from the F-distribution with degrees of freedom 1 and 7 showing the 95 percent quantile of this F-distribution can be created as follows. Generating random number from the F-distribution with 1 and 7 degrees of freedom

```
set.seed(998)
y  <- rf(n=1000,1,7)
```

The 95% quantile of the distribution

```
q95 <- qf(0.95,1,7)
```

and a histogram:

```
hist(y,probability=TRUE)
lines(density(y))
abline(v=q95)
```

(a) How large is the area to the right of the vertical line?

(b) On which side from the vertical line in the histogram must the F-test statistic lie, such that a F-test is significant?

2 Gestational Age

TOPICS: Linear normal model; two regression lines

The data are from a study, where for boys and girls their gestational(prenatal) weights were recorded. In the data set 'gestation' of the course package `LiSciData` the data are given in 'long format', e.g. each row consists of measurements of a boy and a girl. We first reshape the data, such that each row contains just the measurements for one boy or girl. We add a variable `sex`, and make it a factor, to indicate the sex of the individual.

1. Reading the data

```
data(gestation, package="LiSciData")
gestation<- reshape(gestation,direction='long',
             varying=list(age=c('boys.age','girls.age'),
             weight=c('boys.weight','girls.weight')),
             v.names=c('age','weight'),
             timevar='sex',times=c('boy','girl'))
gestation$sex<-factor(gestation$sex)
head(gestation)
```

```
      sex age weight id
1.boy boy  40   2968  1
2.boy boy  38   2795  2
3.boy boy  40   3163  3
4.boy boy  35   2925  4
5.boy boy  36   2625  5
6.boy boy  37   2847  6
```

2. First, plot the weight against gestational age, differentiating between boys and girls. Then add a smooth line for both sexes.(You can use the `lowess`-function.)

3. Formulate two models.

   M1: A model with different intercepts but common slope for regressions of weight on the gestational age.

   M2: A model that additionally differentiates the two slopes for the sexes.

   Fit these models using the `glm`-function.

4. In model M1 find and interpret the estimates of the model parameters. Use the `esticon`-function of the package `doBy` to find the estimate of the intercept for girls.

   Here we note that a simpler way to obtain this intercept is to declare the model M1 by

   ```
   m1.alter <- glm(weight ˜ sex + age - 1, data = gestation)
   coef(summary(m1.alter))[,c(1,2)]
   ```

   Also find and interpret the estimates of the model parameters in model M2.

5. Use the function `predict` to calculate the fitted values of M1. Then plot these fitted lines together with the observations.

6. In the model M1 find confidence intervals for the parameters.

7. We want to test whether it is necessary to have a slope for each sex. Use the function `anova` to compare the two models M1 and M2.

8. Use the function `drop1` on `m1`. What does the result tell you?

9. Plots for model checking can be obtained.

```
m1 <- glm(weight~sex+age-1,data=gestation)
par(mfrow=c(2,2))
plot(m1,pch=c(16,1)[gestation$sex],cex=2)
```

Does the model fit data?

3 Sexism Score

TOPICS: Linear normal model; two–way anova; LSMEANS

A study was conducted to compare the sexist attitudes of students at various types of colleges in the US. The colleges-types are: mixed (gender) college with at least 75% male students, mixed college with less than 75% male students, and single sex college. For each gender, random samples of each 10 undergraduate students were selected from each of the three types of colleges. Each student filled in a questionnaire, from which a score for 'degree of sexism'-defined as the extent to which a student considered males and females to have different life roles-was determined

1. Reading the data

   ```
   data(sexism,package='LiSciData')
   sexism <- transform(sexism, score = sexism)
   sexism <- transform(sexism, type = factor(type))
   sexism <- transform(sexism, gender.type = interaction(gender,type))
   head(sexism)
   ```

   ```
     sexism type gender score gender.type
   1     50    1   male    50      male.1
   2     35    1   male    35      male.1
   3     37    1   male    37      male.1
   4     32    1   male    32      male.1
   5     46    1   male    46      male.1
   6     38    1   male    38      male.1
   ```

2. Plot the data to visualize the effect of gender and type on the sexism score. (The function `stripchart` can be used.)

3. Formulate a model with a separate mean for each combination of gender and type. Call the model M1. Fit the model M1 using the `glm`-function and find the estimates of the model parameters.

4. Use the function `predict` to calculate the fitted values and plot these fitted values. The function `coplot` can be used here. If `m1` is the object holding your model fit of M1 the following will work

```
sexism <- transform(sexism,fit=predict(m1))
coplot(fit ~ type | gender, data = sexism ,panel = panel.smooth, pch = c(18))
```

Does the plot indicate an interaction between sex and type?

Another possibility for such a conditional plot is via trellis-graphics implemented in the `lattice` package

```
library(lattice)
print(
 xyplot(fit~type|gender,data=sexism,type='l')
 )
```

or

```
print(
 xyplot(fit~type,groups=gender,data=sexism,type='l',auto.key=TRUE)
 )
```

5. Now consider also the model with no interaction between gender and type

$$M_2: \quad E(y_{jk}) = \mu_{jk} = \mu + \alpha_j + \beta_k, \quad k = 1, 2, 3, \quad j = \text{boy, girl}$$

Fit this model using the `glm`-function and test the hypothesis of no interaction using the function `anova`.

6. Use the model M1 to obtain an estimate for the average sexism score for males across the types. Weight the three levels of type equally and use the `esticon function`. The resulting mean is called the LSMEANS (or population average) for males across type.

7. Similarly one can calculate the LSMEAN for females across type. One should consider these LSMEANS as possibly interesting predictions of the general level of sexism for males and females. But because there is an interaction between gender and type, it is not an interesting scientific hypothesis, that these LSMEANS should be the same. Nevertheless, it is mathematically possible to test the hypothesis.

In R you can do this with the `drop1` function, but for the contrast of the factors in the model you must use the `contrast.sum` contrast, not the `contr.treatment` we normally use.

```
old<-options()$contrasts
options(contrasts=c('contr.sum','contr.poly'))
m1L<- glm(score ~ gender + type+ gender:type,data=sexism)
drop1(m1L,.~.,test='F')
```

```
Single term deletions

Model:
score ~ gender + type + gender:type
```

```
            Df Deviance    AIC F value     Pr(F)    
<none>           1365.3 371.76                      
gender       1   1593.5 379.03  9.0237  0.004034 ** 
type         2   2022.7 391.34 13.0013 2.459e-05 ***
gender:type  2   1624.6 378.19  5.1279  0.009141 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The row with `gender` gives the test that the averages for males and females (equally weighted across type-levels) are the same. This hypothesis should by no means be considered as testing the main-effects of gender accounted for type, as it is sometimes expressed. Because there is interaction between gender and type there is no point in speaking of a main effect.

Finally we reset the contrasts

```
options(old)
```

4 Bacteria

TOPICS: Linear normal model; log-transformation.

In an experiment on nutrients on bacterial growth, bacteria were grown in different media with the nutrients sucrose and leucine added. After fours days the numbers of bacteria, `density`, were counted.

```
data(bactsucrose, package="LiSciData")
head(bactsucrose)
```

```
  day sucrose leucine  density
1   1       1       1 2.47e+07
2   1       1       2 3.81e+07
3   1       1       3 3.05e+08
4   1       2       1 1.22e+07
5   1       2       2 8.93e+07
6   1       2       3 1.54e+09
```

1. Assume the responses to be normally distributed and use an identity link. Assume the predictors `day`, `sucrose` and `leucine` as factors. Fit a model which is additive in `day` and with an interaction between the other two factors.
2. Look at the residuals. What do you observe?
3. Fit the model, where you assume that the transformed observations `log(density)` are normally distributed. Look at the residuals.