

## Biostat II: Lab 12, Fitting interaction and spline models in R

Date: 12 May 2008

We are interested modelling human weights as a function of age and gender. In this lab, we will imagine several different models, simulate some of our own data, and then fit the models using linear regression.

1. To start out, we may imagine the following model for weight (in kg) as a function of age (centered at 18 years) and gender:

$$\text{Weight}_i = \beta_0 + \beta_1 \cdot (\text{Age}_i - 18) + \beta_2 \cdot I(\text{Female}_i) + \epsilon_i$$

where  $\epsilon_i \sim N(0, \sigma^2)$ . Note that we use the subscript  $i$  to denote values for the  $i$ -th individual.

State the interpretation of the coefficients  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ , and  $\sigma^2$ .

2. It is natural to think that the relationship between weight and age varies by gender. Consider the interaction model:

$$\text{Weight}_i = \beta_0 + \beta_1 \cdot (\text{Age}_i - 18) + \beta_2 \cdot I(\text{Female}_i) + \beta_3 \cdot (\text{Age}_i - 18) \cdot I(\text{Female}_i) + \epsilon_i$$

- (a) Write down the specific case of this model for females by setting the indicator variable  $I(\text{Female})=1$ .
  - (b) Write down the specific case of this model for males by setting the indicator variable  $I(\text{Female})=0$ .
  - (c) State the interpretation of the coefficients  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$  and  $\sigma^2$ .
3. Perhaps weight is linearly related to age among children, but once people stop growing the rate of increase with age cannot possibly stay the same. To make our model flexible enough to handle children and adults, we can create a spline term for individuals over the age of 18:

$$\text{Weight}_i = \beta_0 + \beta_1 \cdot (\text{Age}_i - 18) + \beta_2 \cdot I(\text{Female}_i) + \beta_3 \cdot (\text{Age}_i - 18) \cdot I(\text{Female}_i) + \beta_4 \cdot (\text{Age}_i - 18)^+ + \epsilon_i$$

where  $(\text{Age}_i - 18)^+ = (\text{Age}_i - 18)$  when  $\text{Age} - 18$  is positive, and 0 otherwise.

- (a) Write down the specific case of this model for females under the age of 18.
- (b) Write down the specific case of this model for males under the age of 18.
- (c) Write down the specific case of this model for females over the age of 18.
- (d) Write down the specific case of this model for males over the age of 18.
- (e) State the interpretation of the coefficients  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$ ,  $\beta_4$  and  $\sigma^2$ .

4. Model 3 allowed the relationship between weight and age to differ for those above and below age 18, and forced the change to be the same across genders. We may want to make our model even more flexible, allowing different spline terms for the two genders. Consider the model:

$$\begin{aligned} \text{Weight}_i &= \beta_0 + \beta_1 \cdot (\text{Age}_i - 18) + \beta_2 \cdot I(\text{Female}_i) + \beta_3 \cdot (\text{Age}_i - 18) \cdot I(\text{Female}_i) \\ &+ \beta_4 \cdot (\text{Age}_i - 18)^+ + \beta_5 \cdot (\text{Age}_i - 18)^+ \cdot I(\text{Female}_i) + \epsilon_i \end{aligned}$$

- Write down the specific case of this model for females under the age of 18.
  - Write down the specific case of this model for males under the age of 18.
  - Write down the specific case of this model for females over the age of 18.
  - Write down the specific case of this model for males over the age of 18.
  - State the interpretation of the coefficients  $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$  and  $\sigma^2$ .
5. Instead of collecting actual data to measure the true values of all our  $\beta$  coefficients, we can simulate our own. We will create a data set of 500 observations, using model 4 as our underlying "true" model. We have to choose particular coefficients, which we specify as follows:

$$\begin{aligned} \text{Weight}_i &= 60 + 3.1 \cdot (\text{Age}_i - 18) - 10 \cdot I(\text{Female}_i) - 1.2 \cdot (\text{Age}_i - 18) \cdot I(\text{Female}_i) \\ &- 2.8 \cdot (\text{Age}_i - 18)^+ + 1.3 \cdot (\text{Age}_i - 18)^+ \cdot I(\text{Female}_i) + \epsilon_i \end{aligned}$$

where  $\epsilon_i \sim N(0, \sigma^2 = 15)$ . Note that we have no idea if these coefficients are close to the true values seen in nature, but we're just playing with this model for fun.

- Set your random seed so we all get the same set of simulated data:

```
set.seed(562342)
```

- Generate age values that are uniformly distributed between 5 and 40 years of age. To do so, we will first generate age values from a uniform distribution, and then round them to make it so the ages only take integer values.

```
Age = runif(500, min=5, max=40)
```

```
Age = round(Age)
```

- Generate the female indicators as Bernoulli( $p=0.5$ ) variables:

```
Female = rbinom(500, size=1, prob=0.5)
```

- Generate the error terms  $\epsilon_i$  as normal random variables with mean 0 and variance 10:

```
error = rnorm(500, mean=0, sd=sqrt(15))
```

- Create the spline variable  $(\text{Age} - 18)^+$ :

```
Age18.sp = ifelse(Age>18, Age-18, 0)
```

- Calculate the simulated weight variables according to our model:

```
Weight = 60 + 3.1 * (Age -18) - 10 * Female - 1.2 * (Age-18)* Female -  
2.8 * Age18.sp + 1.3 * Age18.sp * Female + error
```

6. Create a scatter plot of your simulated data with different point types and colors for males and females:

```
plot(Age, Weight, pch=(2:1)[Female+1], col=c("blue","red")[Female+1])
```

7. Now that we can simulated data from a known model, we can perform linear regression using a model of the same form to see how close our fitted model comes to the true model:

```
fit.full = lm(Weight ~ I(Age-18) + Female + I(Age-18)*Female
+ Age18.sp + Age18.sp*Female)
```

8. We could also fit the simpler model with no splines, specified as model 2.

```
fit.reduced = lm(Weight ~ I(Age-18) + Female + I(Age-18)*Female)
```

After performing this regression, calculate the p-value associated with the F-statistic to test whether or not the spline terms were helpful in the model. To calculate the F-statistic in R comparing nested models, we use the following code:

```
print(anova(fit.reduced, fit.full))
```

State the null and alternative hypotheses.

Report whether we reject or fail to reject the null hypothesis at level  $\alpha = 0.01$ .

9. How might you change model 4 to handle people over the age of 60?