MATH3091 Computer Lab 3

14 Feb 2022

The dataset in `birth.csv` contains data on the weight of 24 newborn babies. Read the data into a variable `birth`.

There are two explanatory variables; sex (`Sex`) and gestational age in weeks (`Age`) together with the response variable, birthweight in grams (`Weight`).

`Sex` is a categorical variable, taking two values, 1 for male and 2 for female. Because it is a categorical variable, it should be be declared as a factor in `R`, using

```
birth$Sex <- as.factor(birth$Sex)
```

The data can be plotted, with males and females distinguished, using

```
plot(birth$Age, birth$Weight, xlab = "Age (weeks)", ylab = "Birthweight (grams)",
     col = birth$Sex)
legend("bottomright", legend = c("Male", "Female"), pch = 1, col = 1:2)
```

> **Question:** Do you see what the ‘legend‘ command has done? Read the help file on legend by issuing the command ‘?legend‘. Can you move the legend to the top left of the plot?

We could plot the data in several other ways. For each command below, guess at what type of plots the command will produce. Then run them with `R`: is the result what you expected?

```
plot(Weight ~ Age + Sex, data = birth)
```

```
pairs(birth)
```

> **Question:** Which of the plots you have produced for this data do you think is the most informative?

We can now fit models which include both continuous and categorical explanatory variables. Such models are sometimes called *Analysis of Covariance* models. For example

```
birth_lm1 <- lm(Weight ~ Sex + Age, data = birth)
```

fits the model

$$Y_i = \mu + \alpha_{s_i} + \beta x_i + \epsilon_i, \quad i = 1, \dots, n \tag{1}$$

where Y_i is the `Weight`, s_i is the `Sex`, x_i is the `Age` of the ith baby. Remember to use the `summary` command to see the results of the fitting.

> **Question:** Do you think you should remove any variables from the model?

Note that α_{s_i} can only take two possible values, α_1 and α_2, because s_i can only take two possible values: 1 (male) or 2 (female). By default, `R` forces the first of the two possible values of α_{s_i} to be zero. In this case, this means $\alpha_1 = 0$ since 1 is the first level of the factor `Sex`: check `levels(birth$Sex)`. If you have text names for the possible levels of a categorical variable, by default the levels are ordered alphabetically. The coefficient corresponding to the first level of the factor (the first one alphabetically) will be set to zero.

The parameter α_2 may be interpreted as the average difference in weight between the female and male babies. The model will be

$$Y_i = \mu + \beta x_i + \epsilon_i$$

when the ith baby is male and

$$Y_i = \mu + \alpha_2 + \beta x_i + \epsilon_i$$

when the ith baby is female. In the model for female babies $\mu + \alpha_2$ is the intercept while in the model for male babies μ will be the intercept. The two models will define a pair of parallel straight lines for $E(Y)$.

Model (1) states that `Sex` and `Age` affect `Weight` but that they do so independently. The difference between the expected `Weight` for two different values of `Age` is the same for either `Sex`. Similarly, the difference between the expected `Weight` between the two levels of `Sex` is the same for every value of `Age`. However, we can also incorporate an interaction between a factor and a continuous explanatory variable.

```
birth_lm2 <- lm(Weight ~ Sex + Age + Sex:Age, data = birth)
```

fits the model

$$Y_i = \mu + \alpha_{s_i} + \beta x_i + \gamma_{s_i} x_i + \epsilon_i, \quad i = 1, \ldots, n \tag{2}$$

where α_s and γ_s are constrained to be equal to 0 at the first level of `Sex`, so $\alpha_1 = \gamma_1 = 0$. The interpretation of this model is that the expected `Weight` depends linearly on `Age`, and that the linear relationship has a different intercept *and a different slope* for the different levels of `Sex`. The parameter γ_s describes the differences in the slopes.

1. > **Question:** Which model do you feel best describes the relationship between birthweight, sex and gestational age?

2. Let us do prediction for a female baby with 39 weeks of gestational age and a male baby with 42 weeks of gestation.

   ```
   newdata <- data.frame(Age = c(39, 42), Sex = c("2", "1"))
   predict(birth_lm1, newdata)
   predict(birth_lm1, newdata, interval = "confidence")
   predict(birth_lm1, newdata, interval = "prediction")
   ```

Question: What is the difference between ‘interval = "confidence"‘ and ‘interval = "prediction"‘ here?

3. We check the above results by hand, by using the parameter estimates from `birth_lm1`:

   ```
   coef(birth_lm1)
   ```

   ```
   # For female aged 39 weeks
   120.8943 * 39 - 163.0393 - 1610.2825
   # For male aged 42 weeks
   120.8943 * 42 - 1610.2825
   ```

   Question: How would you predict the birthweight for a female child aged 41 weeks?

4. We would like to add the fitted regression lines to our plot of the data. First, we re-do our original plot, using colours to distinguish between the sexes:

   ```
   plot(birth$Age, birth$Weight, xlab = "Age (weeks)", ylab = "Birthweight (grams)",
        col = birth$Sex)
   legend("bottomright", legend = c("Male", "Female"), pch = 1, col = 1:2)
   ```

   We can add a fitted regression line for the males with:

   ```
   abline(-1610, 120.9, lty = 2, col = 1)
   ```

   Question: How would you add fitted regression line for the females? Can you make this line red, to match the female data points?