Statistics 305/605: Introduction to Biostatistical Methods for Health Sciences

Chapter 19, part 1: Multiple Linear Regression Models

Jinko Graham

2018-11-11

Multiple Regression

- Have been looking at *simple linear regression*, to understand the relationship between the response and a single explanatory variable.
- Now, transition to *multiple linear regression*, in which we study the relationship between the response and multiple explanatory variables.
- Why multiple regression?
  - A model with more than one explanatory variable may better explain the the mean of the response variable.
  - Allows for possible synergy between explanatory variables.
  - Allows adjustment for confounding variables
    - (Recall: A confounding variable is an extraneous variable that is associated with both the outcome and the exposure of interest).
  - Can improve the precision of our predictions.

Example and Notation

Consider the `lbwt` dataset

- The response variable, Y, is head circumference, with observed values denoted by y.
- One explanatory variable is gestational age, X_1, with observed values denoted by x_1.
- Let's consider a second explanatory variable, birth weight, X_2, with observed values denoted by x_2.

```r
head(lbwt,n=3)
```

```
##   headcirc length gestage birthwt momage toxemia
## 1       27     41      29    1360     37       0
## 2       29     40      31    1490     34       0
## 3       30     38      33    1490     32       0
```

Multiple Linear Regression (MLR) Model

- Let's fit a model with **two** explanatory variables:
  - X_1, the gestational age, as before, and a new variable
  - X_2, the birth weight.

```
fit2 <- lm(headcirc ~ gestage + birthwt, data=lbwt)
summary(fit2)$coefficients
```

```
##                Estimate  Std. Error  t value     Pr(>|t|)
## (Intercept) 8.308015388 1.578942936 5.261758 8.535816e-07
## gestage     0.448732848 0.067245982 6.673006 1.555501e-09
## birthwt     0.004712283 0.000631179 7.465843 3.596527e-11
```

- Consider the following interpretation of the fitted MLR model:

  Among infants **of the same birth weight**, *a one week increase in gestational age is associated with a 0.45cm increase in head circumference.*

- We will motivate this intepretation in what follows.

Multiple Regression Overview

- We have one response variable, Y.
- We have q explanatory variables denoted by $X_1, X_2, \ldots, X_q$, with observed values $x_1, x_2, \ldots, x_q$, respectively.
- The regression model describes how the average value of Y changes as $x_1, \ldots, x_q$ change.
- We use the method of least squares to fit the model to our data.
- Under modelling assumptions, we can
  - infer the slopes of the regression model in the population from the slopes fitted in our sample, and
  - make predictions from the model we have fitted to our data.
- Model assumptions are checked *after* the model is fit to our sample of data.

Model Overview

- Model components remain the same as in SLR:
    1. linear predictor,
    2. normal error terms, and
    3. constant SD, σ_y.
- The linear predictor component is generalized to include more explanatory variables, but the normal errors and constant SD assumptions are as-before.
- Also, as in SLR, we assume independent observations.

Linear Predictor

- In SLR, the linear predictor for the population mean response is

$$\mu_{y|x} = \alpha + \beta x,$$

  where x is a value of the single explanatory variable X.
- In MLR, the linear predictor is generalized to

$$\mu_{y|x_1,\ldots,x_q} = \alpha + \beta_1 x_1 + \ldots + \beta_q x_q$$

- $\mu_{y|x_1,\ldots,x_q}$ is the population mean value of Y for all data points with $X_1 = x_1, \ldots, X_q = x_q$.
- Individual regression coefficients β_k are the change in $\mu_{y|x_1,\ldots,x_q}$ for a one-unit increase in x_k *holding all other x's fixed*.
  - In the above example with gestational age (X_1) and birth weight (X_2), the interpretation of β_1 is the increase in $\mu_{y|x_1,x_2}$ for a one week increase in gestational age, holding birth weight fixed.

Interpreting Fitted Regression Coefficients

- Fit MLR model to `lbwt` data.

```
fit2 <- lm(headcirc ~ gestage + birthwt, data=lbwt)
summary(fit2)$coefficients
```

```
##                Estimate  Std. Error  t value     Pr(>|t|)
## (Intercept) 8.308015388 1.578942936 5.261758 8.535816e-07
## gestage     0.448732848 0.067245982 6.673006 1.555501e-09
## birthwt     0.004712283 0.000631179 7.465843 3.596527e-11
```

- Interpretation of fitted coefficient $\hat{\beta}_1$ for gestational age:
    - "For a given birth weight, a one week increase in gestational age is associated with a 0.45cm increase in head circumference."
- Interpretation of fitted coefficient $\hat{\beta}_2$ for birth weight:
    - "For a given gestational age, a one gram increase in birth weight is associated with a 0.0047cm increase in head circumference."
    - Or (since a 1g increase in weight is too fine-grained) "For a given gestational age, a **100** gram increase in birth weight is associated with a 0.47cm increase in head circumference."

Software Notes

- `lm()` will fit simple *and* multiple regression models.
- To use `lm()` to fit multiple regression models, we include multiple explanatory variables on the right-hand side of the formula, separated by the + sign.
  - In the example, the model formula
    `headcirc ~ gestage + birthwt`
    includes both `gestage` and `birthwt` as explanatory variables.