19. Multiple Regression

If one explanatory variable helped to predict the value of the response variable, then, perhaps, more than one would be give better predictions. This is the idea behind multiple (linear) regression.

19.1. The Model

For simple linear regression, we denoted the mean of our response variable, y, as

$$\mu_{y|x}$$

to indicate the dependence of the mean of y on the single explanatory variable, x. We extend this notation to include q response variables, $x_1, x_2, \ldots, x_q$:

$$\mu_{y|x_1, x_2, \ldots, x_q}$$

Here, too, we will restrict ourselves to modeling this mean as a linear function:

$$\mu_{y|x_1, x_2, \ldots, x_q} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$$

The interpretation of the parameters is similar to the simple linear regression case:

- α is the intercept—the value of the mean of y when each of the explanatory variables is zero
- the β_k's indicate the rate of change of the mean of y with the x_k explanatory variable, k=1,…,q or the change in the

mean of y with a one unit change in x_k, all else being equal.

Again, we model the "error" or spread of y values about the mean as a (normal) mean zero random variable:

$$y = \mu_{y|x_1,x_2,\ldots,x_q} + \varepsilon = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \varepsilon$$

The assumptions of the above model are essentially the same as the simple linear model. See POB, page 450.

19.1.1. The Least-Squares Regression Equation

Estimation of the regression model parameters again follows the least squares criterion

$$\min_{\alpha,\beta_1,\ldots,\beta_q}\left(\sum_{i=1}^{n}\left[y_i - (\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_q x_{qi})\right]^2\right)$$

The solution of which yields the estimated regression model

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}1x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_q x_{qi}$$

or

$$y_i = \hat{\alpha} + \hat{\beta}1x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_q x_{qi} + e_i$$

We use the residuals, e_i, in the same way as we do for simple linear regression.

19.1.2. Inference for Regression Parameters

We proceed with regression analysis as we did before. In particular, the estimated regression coefficients, under the model assumptions, will be normally distributed with some standard error:

$$\hat{\alpha}_k \sim N\left(\alpha_k, se(\hat{\alpha}_k)\right)$$

$$\hat{\beta}_k \sim N\left(\beta_k, se(\hat{\beta}_k)\right)$$

where we'll estimate the se's by estimating $\sigma_{y|x1,x2,\ldots xq}$ with

$$s_{y|x_1,\ldots x_q} = \sqrt{\frac{1}{(n-q-1)}\sum_{i=1}^{n} e_i^{\,2}} = \sqrt{\frac{1}{(n-q-1)}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$

($\sigma_{y|x1,x2,\ldots xq}$ is part of the equations for each se (not shown)).

Thus, we can perform hypothesis tests using the t-statistics as before, e.g.,

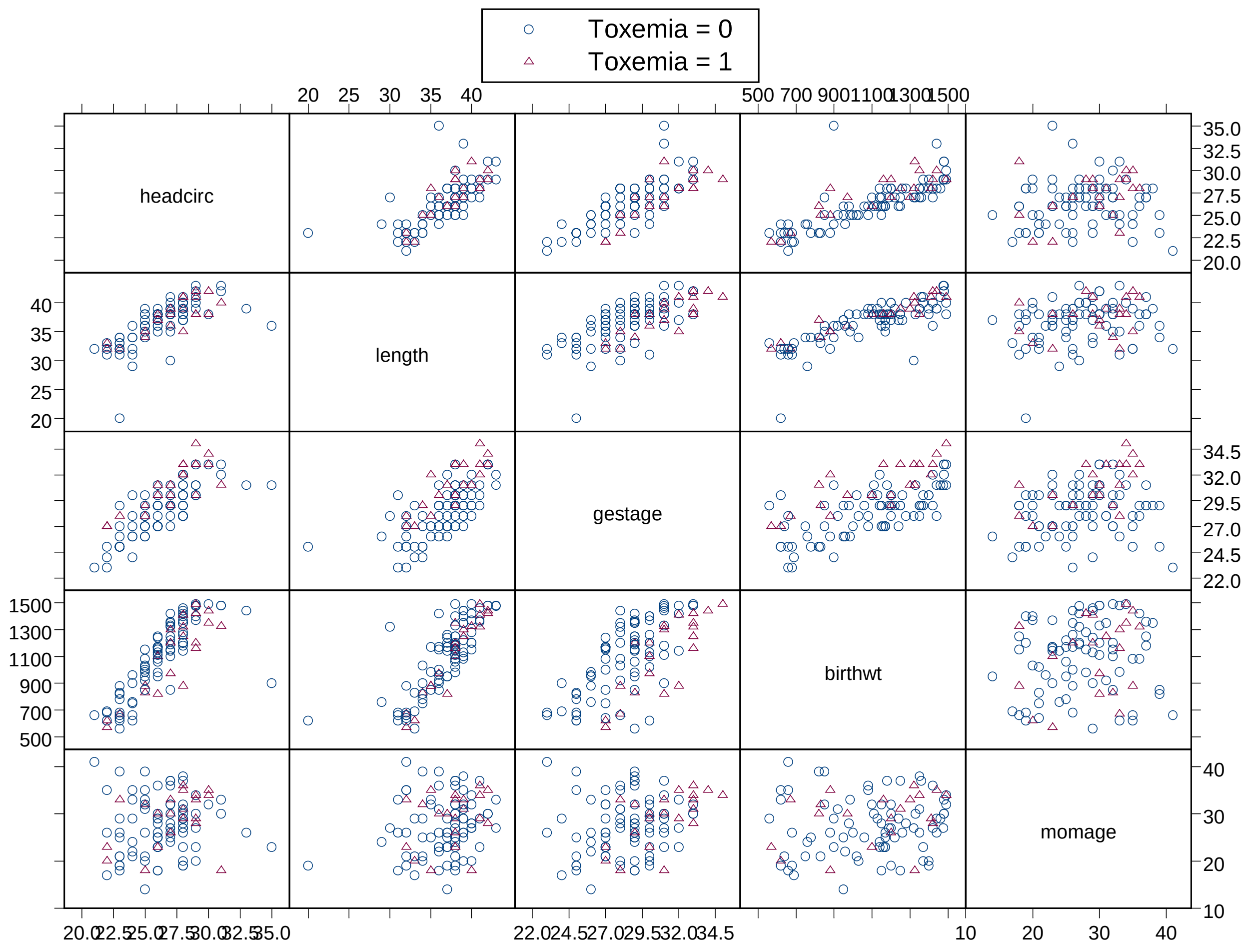
H_0:β_k=β_{k0}
H_A:$\beta_k \neq \beta_{k0}$

$$t = \frac{\hat{\beta}_k - \beta_{k0}}{se(\hat{\beta}_k)}$$

We can also construct confidence intervals for the regression coefficients in the obvious way, and make "predictions" with appropriate intervals. We do not give the details of the calculations, but, instead, proceed to illustrate multiple regression with an example. This will also serve as coverage of sections 19.1.3, 19.1.4, and 19.1.5.

E.g., Modeling infant head circumference as a (linear) function of various explanatory variables. We'll follow the analysis in chapter 19 so you can see the relationship between the text results and those in S-Plus.

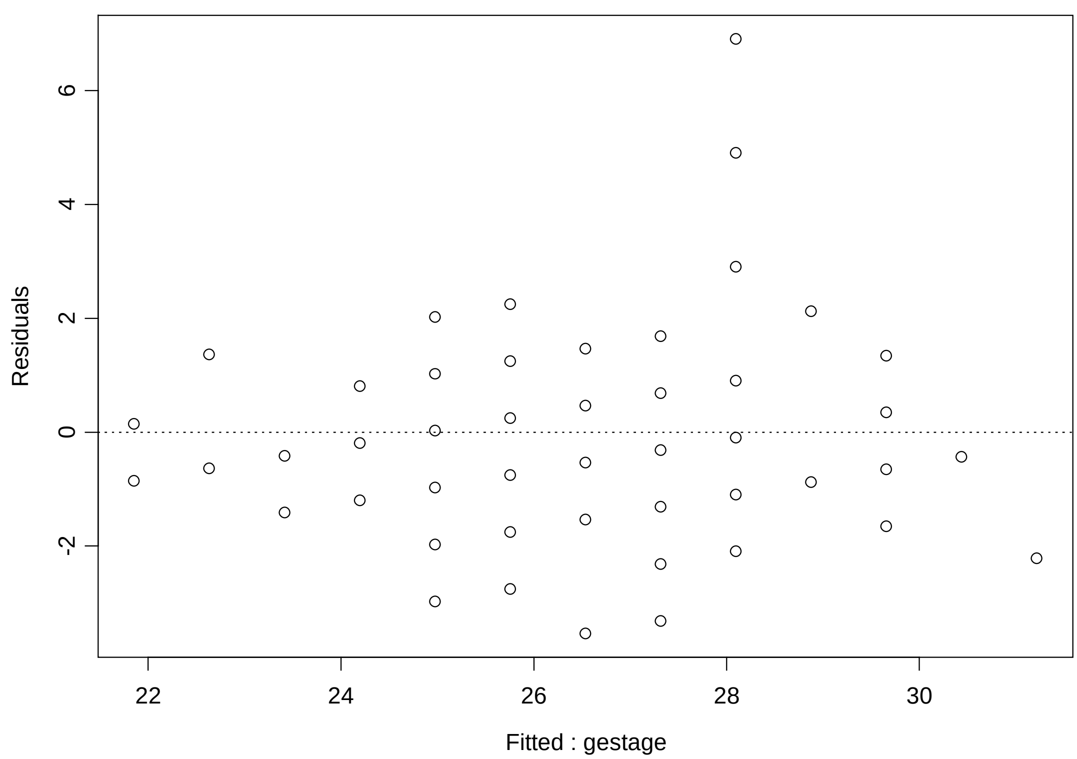
First, some exploratory graphical analysis

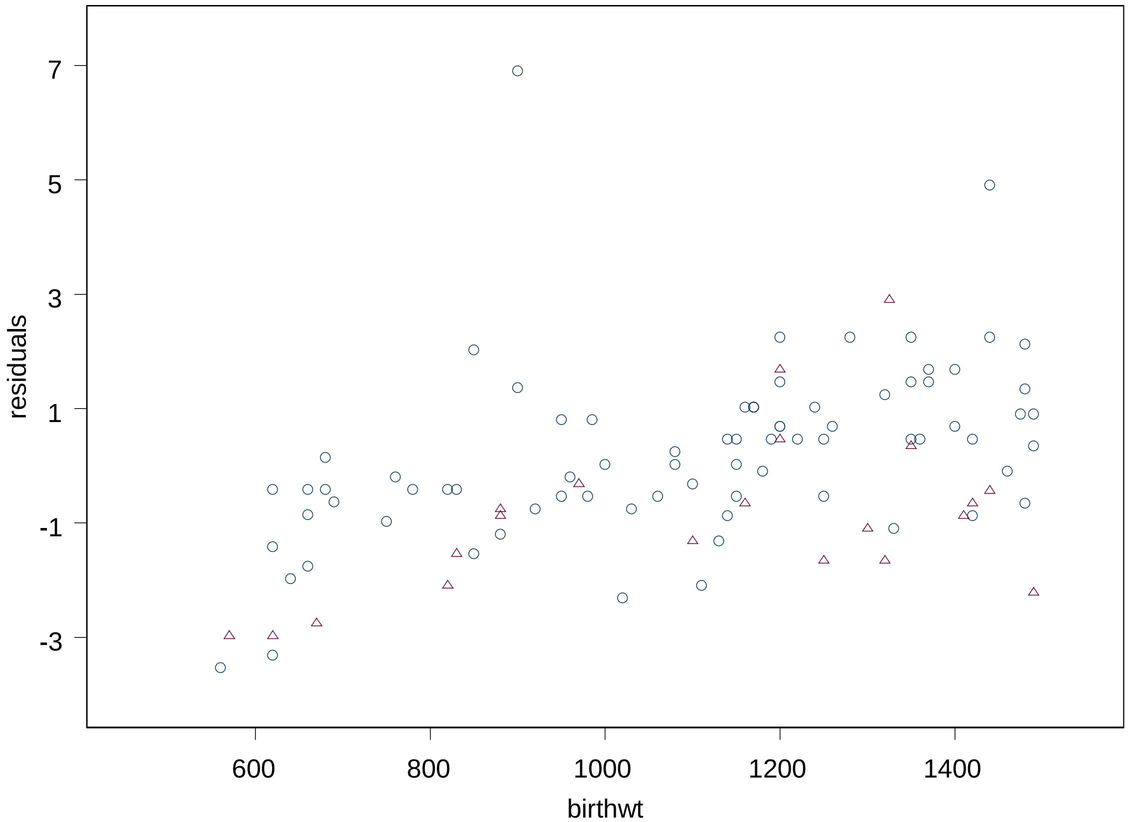


Looks like several variables are related to infant head circumference.

Start simple:
headcirc vs. gestage (as in book p450, see chapter 18, too)

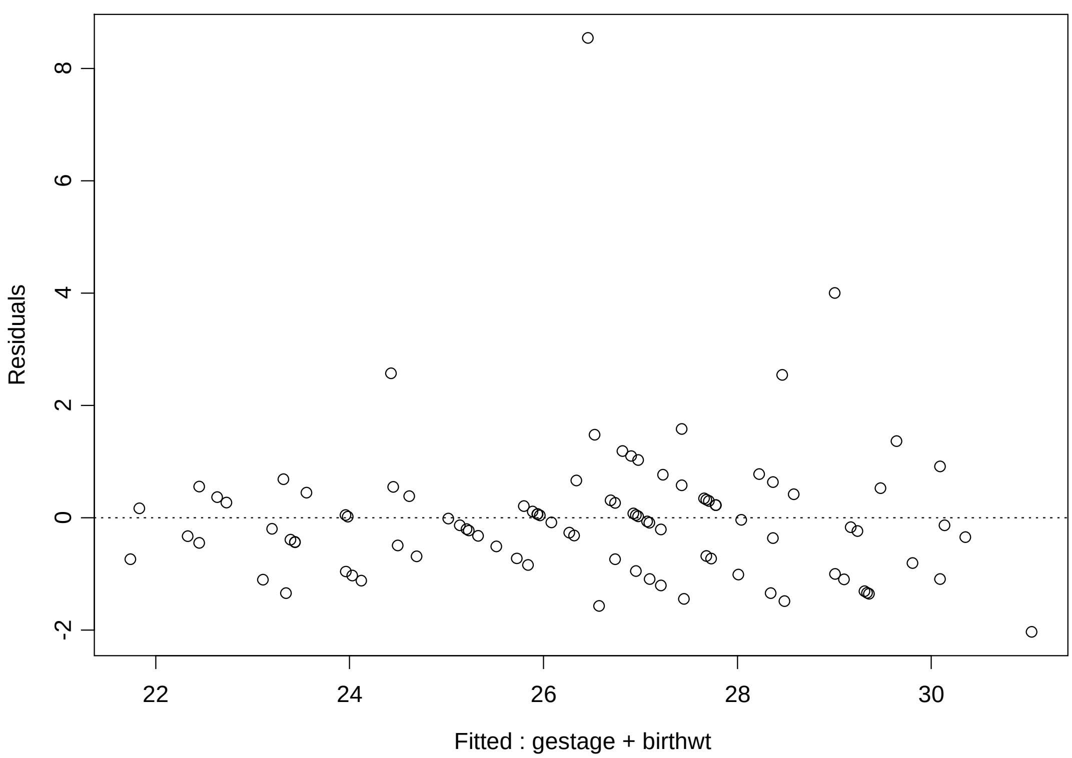
```
       *** Linear Model ***

Call: lm(formula = headcirc ~ gestage, data = low.birth.weight.infants, na.action =
na.exclude)
Residuals:
    Min     1Q  Median     3Q   Max
 -3.536 -0.876 -0.1458 0.9041 6.904

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  3.9143  1.8291     2.1399  0.0348
    gestage  0.7801  0.0631    12.3672  0.0000

Residual standard error: 1.59 on 98 degrees of freedom
Multiple R-Squared: 0.6095
F-statistic: 152.9 on 1 and 98 degrees of freedom, the p-value is 0

Analysis of Variance Table

Response: headcirc

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value Pr(F)
  gestage  1  386.8674 386.8674 152.9474     0
Residuals 98  247.8826   2.5294
```

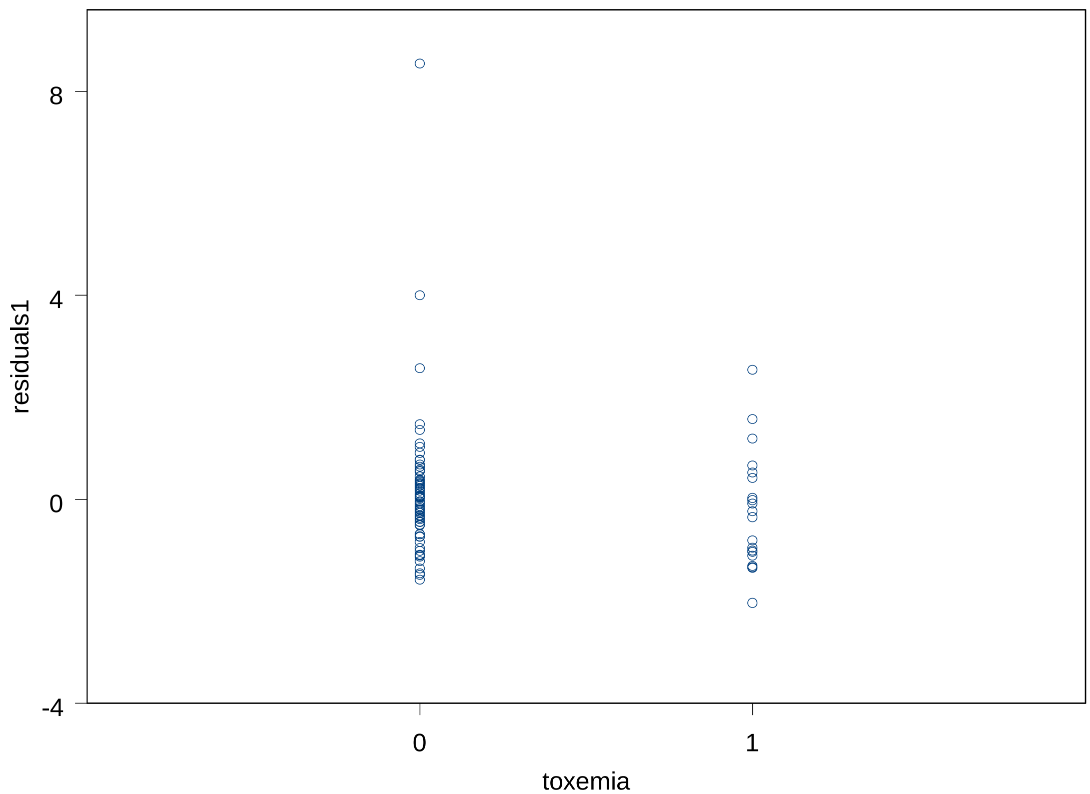



7
5
3
1
-1
-3
residuals
600
800
1000
1200
1400
birthwt

headcirc vs gestage and birthwt (as in book, p 451)

```
       *** Linear Model ***

Call: lm(formula = headcirc ~ gestage + birthwt, data = low.birth.weight.infants, na.action =
na.exclude)
Residuals:
    Min      1Q   Median     3Q  Max
 -2.035 -0.7271 -0.07653 0.3472 8.54

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 8.3080 1.5789     5.2618  0.0000
    gestage 0.4487 0.0672     6.6730  0.0000
    birthwt 0.0047 0.0006     7.4658  0.0000

Residual standard error: 1.274 on 97 degrees of freedom
Multiple R-Squared: 0.752
F-statistic: 147.1 on 2 and 97 degrees of freedom, the p-value is 0

Analysis of Variance Table

Response: headcirc

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value         Pr(F)
  gestage  1  386.8674 386.8674 238.3776 0.000000e+000
  birthwt  1   90.4595  90.4595  55.7388 3.596523e-011
Residuals 97  157.4231   1.6229
```




8
4
0
-4
residuals1
0
1
toxemia

headcirc vs. gestage birthwt and toxemia (not in book)

```
       *** Linear Model ***

Call: lm(formula = headcirc ~ gestage + birthwt + toxemia, data = low.birth.weight.infants,
na.action = na.exclude)
Residuals:
    Min      1Q  Median    3Q   Max
 -1.852 -0.6677 -0.1265 0.391 8.236

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  7.0958  1.7976     3.9472  0.0002
    gestage  0.5080  0.0794     6.3989  0.0000
    birthwt  0.0044  0.0007     6.4116  0.0000
    toxemia -0.5128  0.3693    -1.3886  0.1682

Residual standard error: 1.268 on 96 degrees of freedom
Multiple R-Squared: 0.7569
F-statistic: 99.62 on 3 and 96 degrees of freedom, the p-value is 0

Analysis of Variance Table

Response: headcirc

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq  F Value     Pr(F)
  gestage  1  386.8674 386.8674 240.6588 0.0000000
  birthwt  1   90.4595  90.4595  56.2722 0.0000000
  toxemia  1    3.0998   3.0998   1.9283 0.1681605
Residuals 96  154.3233   1.6075
```

headcirc vs. gestage and toxemia (as in book, p455-6)

```
       *** Linear Model ***

Call: lm(formula = headcirc ~ gestage + toxemia, data = low.birth.weight.infants, na.action =
na.exclude)
Residuals:
    Min      1Q   Median     3Q   Max
 -3.843 -0.8427 -0.05252 0.8109 6.409

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   1.4956   1.8680     0.8006   0.4253
    gestage   0.8740   0.0656    13.3222   0.0000
    toxemia  -1.4123   0.4062    -3.4773   0.0008

Residual standard error: 1.507 on 97 degrees of freedom
Multiple R-Squared: 0.6528
F-statistic: 91.18 on 2 and 97 degrees of freedom, the p-value is 0
```

headcirc vs. gestage and toxemia (as in book, p457-8)

```
       *** Linear Model ***

Call: lm(formula = headcirc ~ gestage + toxemia + gestage:toxemia, data =
low.birth.weight.infants, na.action =
       na.exclude)
Residuals:
    Min      1Q   Median    3Q   Max
 -3.837 -0.8366 -0.09276 0.791 6.434

Coefficients:
                   Value Std. Error  t value Pr(>|t|)
    (Intercept)   1.7629   2.1023     0.8386   0.4038
        gestage   0.8646   0.0739    11.7001   0.0000
        toxemia  -2.8150   4.9851    -0.5647   0.5736
gestage:toxemia   0.0462   0.1635     0.2823   0.7783

Residual standard error: 1.515 on 96 degrees of freedom
Multiple R-Squared: 0.6531
F-statistic: 60.23 on 3 and 96 degrees of freedom, the p-value is 0
```