ST 517: Data Analytics I

Multiple Linear Regression

Outline

The Multiple Linear Regression (MLR) Model

- Example
- R Output
- Interpretations

Multiple Linear Regression

In simple linear regression we model the mean of *Y* as a function of *X*:

$$\mu(Y\ |X) = \beta_0 + \beta_1 X_1$$

In multiple linear regression with two explanatory variables, X_1 and X_2, we model the mean of *Y* as a function of both X_1 and X_2.

For example:

$$\mu(Y\ |X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

Other Examples

Each of the following is an example of a multiple linear regression model:

- $\mu(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- $\mu(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
- $\mu(Y \mid X_2) = \beta_0 + \beta_1 X_2 + \beta_2 X_2^2$

Notice that these models are linear in the β's, but not necessarily in the X's.

Example: Brain Size

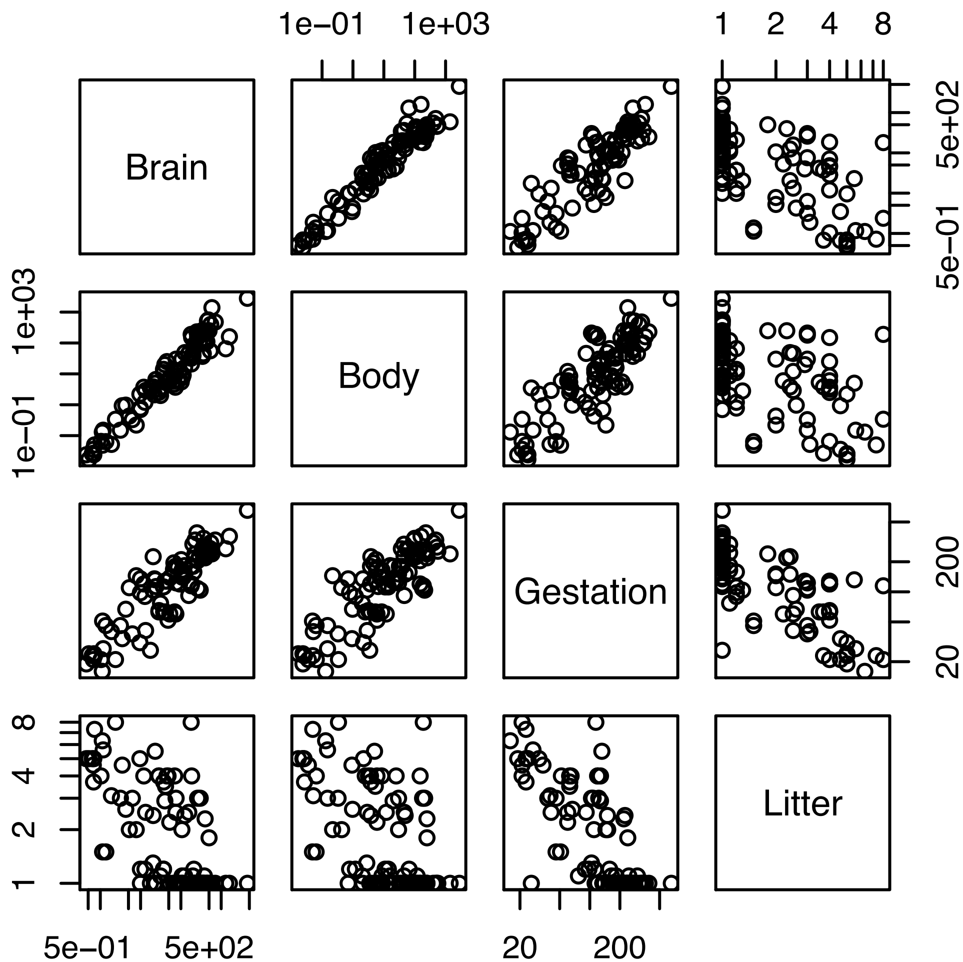
Why do some mammals have large brains for their body size? We will consider a dataset which contains:

- Ninety-six species of mammal
- Average brain weight (g), average body weight (kg), average litter size, average gestation period (days)

Question: After accounting for differences in body weight, is brain weight associated with litter size and/or gestation period?

Why should we be worried about body weights?

Scatter Plots (log axes)



Some Questions

Before we start constructing a MLR for these data, let's consider the following questions:

- Does brain weight appear to be associated with body weight?
- Does brain weight appear to be associated with gestation period?
- Does brain weight appear to be associated with litter size?
- Do any of the explanatory variables appear to be associated with each other?

A MLR for Brain Size

We'll work with all 4 variables on the log scale. The multiple linear regression (MLR) model we will consider is:

$$\mu(lbrain\,|lbody, lgest, llit) = \beta_0 + \beta_1 lbody + \beta_2 lgest + \beta_3 llit$$

- β_0 is still called an intercept—it is the value of $\mu(lbrain\,|lbody, lgest, llit)$ when $lbody = lgest = llit = 0$ (i.e. the mean response when all explanatory variables are zero)
- $\beta_1, \beta_2,$ and β_3 are called slope terms, although we'll have to be a little careful with our interpretations of them.

Some R Output

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.85482    0.66167   1.292  0.19962
lbody         0.57507    0.03259  17.647  < 2e-16 ***
llitter      -0.31007    0.11593  -2.675  0.00885 **
lgest         0.41794    0.14078   2.969  0.00381 **

Residual standard error: 0.4748 on 92 degrees of freedom
Multiple R-squared:  0.9537, Adjusted R-squared:  0.9522
F-statistic: 631.6 on 3 and 92 DF,  p-value: < 2.2e-16
```

Interpreting the R Output

Each line in the Coefficients table of the MLR output corresponds to a parameter in the model.

```
Coefficients:
              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    0.85482     0.66167    1.292   0.19962
lbody          0.57507     0.03259   17.647   < 2e-16 ***
llitter       -0.31007     0.11593   -2.675   0.00885 **
lgest          0.41794     0.14078    2.969   0.00381 **
```

The *p*-values in the last column correspond, respectively, to the null hypotheses:

$$H_{00}\colon\ \beta_0 = 0; H_{01}\colon\ \beta_1 = 0; H_{02}\colon\ \beta_2 = 0; H_{03}\colon\ \beta_3 = 0$$

Interpreting R Output

In this case, each of the *p*-values, except for the one corresponding to the intercept term, are quite small (< 0.01)

- The small *p*-values indicate that we have strong evidence that each of the regression coefficients, β_1, β_2, and β_3 are different from zero.
- Even though there is no evidence that the *y*-intercept is different from zero, we typically leave it in the regression model.
  - In part, this allows us to interpret R^2 as a proportion of variation in the response explained by the model.
  - It will also introduce bias into the estimates of the other regression parameters if we remove the intercept term.

Interpreting R Output

The bottom part of the MLR output looks a lot like that from SLR.

```
Residual standard error: 0.4748 on 92 degrees of freedom
Multiple R-squared:  0.9537, Adjusted R-squared:  0.9522
F-statistic: 631.6 on 3 and 92 DF,  p-value: < 2.2e-16
```

- Residual standard error is the MLR estimate of σ.
- R^2 is the proportion of variation in *lbrain* explained by the MLR model.
- The *F*-statistic is for an *F*-test comparing the MLR to the model with just an intercept term.

Interpreting MLR Coefficients

The estimated regression model for the Brain Size data is:

$$\mu(lbrain\ |lbody, lgest, llit) = 0.85 + 0.58lbody + 0.42lgest - 0.31llit$$

- 0.58 is the estimated amount by which mean log brain size changes for a unit change in log body size *when log gestation and log litter are held fixed.*
- 0.42 is the estimated amount by which mean log brain size changes for a unit change in log gestation *when log body size and log litter are held fixed.*
- -0.31 is the estimated amount by when mean log brain size changes for a unit change in log litter *when log body size and log gestation are held fixed.*

Interpreting MLR Coefficients

Consider the two models:

$$\mu(Y \mid \mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

and

$$\mu(Y \mid \mathbf{X}) = \beta_0 + \beta_1 X_1$$

The coefficient β_1 has a different interpretation in these two models:

- In the first model: β_1 is the amount by which the mean of *Y* changes for a unit increase in X_1 when X_2 is held fixed.
- In contrast, in the second model, β_1 is the amount by which the mean of *Y* changes for a unit increase in X_1.