STA286 Lecture 33 - NOT ON EXAM

Neil Montgomery

Last edited: 2017-04-14 16:15

case study - regression

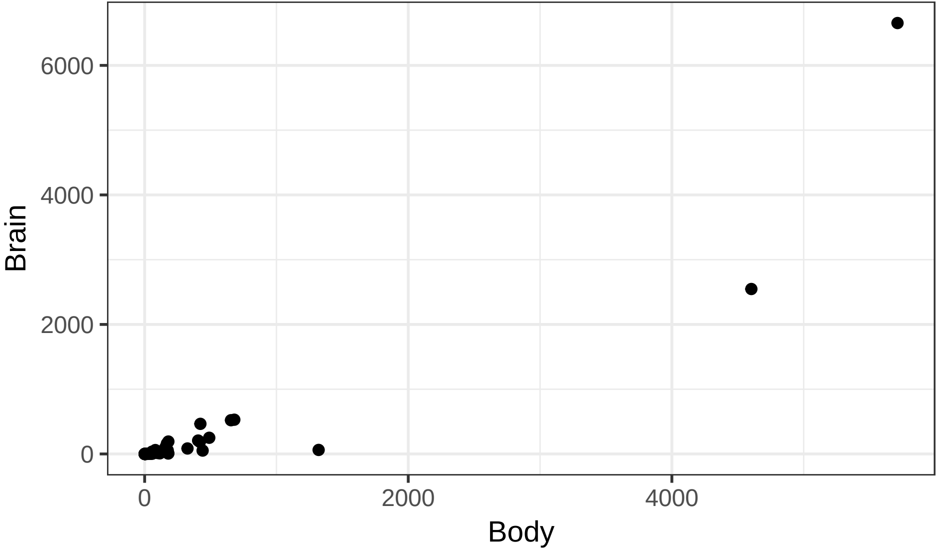
some data

There's a "classic" dataset with the weights of the bodies (kg) and brains (g) of 62 animals.

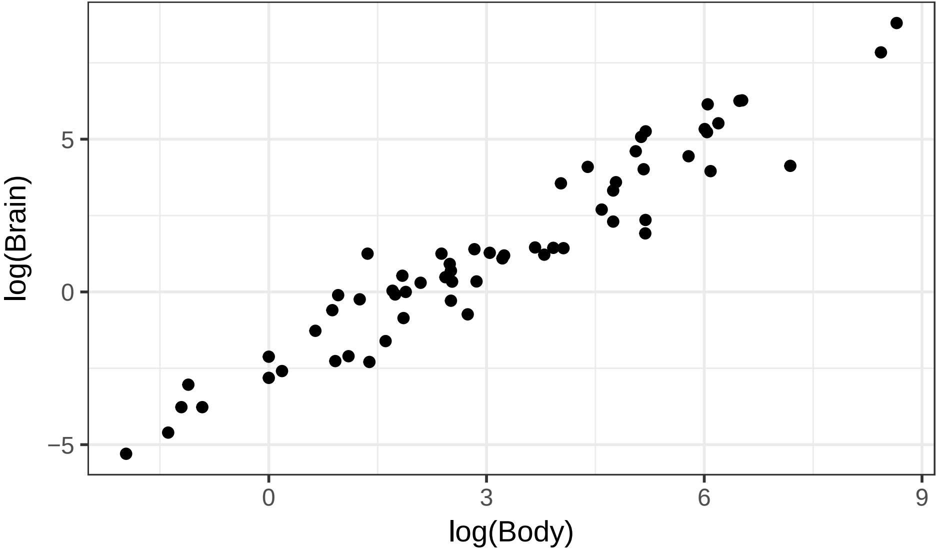
Here's a glance at the data:

Index	Brain	Body
1	3.38	44.50
2	0.48	15.50
3	1.35	8.10
4	465.00	423.00
5	36.33	119.50
6	27.66	115.00
7	14.83	98.20
8	1.04	5.50
9	4.19	58.00
10	0.42	6.40
11	0.10	4.00
12	0.92	5.70

two numerical variables - scatterplot



log-log scales



the simple regression model

A simple model in the form of:

$$\text{Output} = \text{Input} + \text{Noise}$$

is the simple linear regression model with normal noise:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

with ε i.i.d. $N(0, \sigma)$ for some unknown σ (which is constant - not a function of x.)

The data has n rows, so $i \in \{1, \ldots, n\}$.

the variables

the variables

y gets variously called the “output variable”, the “dependent variable”, the “outcome variable”, the “target variable”, the “y” variable, and probably others.

the variables

y gets variously called the "output variable", the "dependent variable", the "outcome variable", the "target variable", the "y" variable, and probably others.

x gets variously called the "input variable", the "independent variable", the "predictor", "explanatory", "risk factor", "feature", "x", and probably others.

the variables

y gets variously called the “output variable”, the “dependent variable”, the “outcome variable”, the “target variable”, the “y” variable, and probably others.

x gets variously called the “input variable”, the “independent variable”, the “predictor”, “explanatory”, “risk factor”, “feature”, “x”, and probably others.

The input variable is not treated as though it came from a random process (even if it did.) It is treated as a vector of constants.

The output variable is therefore a sum of constants plus a vector of random noise.

the parameters, and the model re-expressed

So this is another way of writing the model:

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma)$$

There are three parameters in this model: β_0, β_1, and σ^2. These will be estimated using data.

the parameters, and the model re-expressed

So this is another way of writing the model:

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma)$$

There are three parameters in this model: β_0, β_1, and σ^2. These will be estimated using data.

For convenience call the data: $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

the parameters, and the model re-expressed

So this is another way of writing the model:

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma)$$

There are three parameters in this model: β_0, β_1, and σ^2. These will be estimated using data.

For convenience call the data: $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

A likelihood function is:

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2; D) &= \prod_{i=1}^{n} f(y_i, x_i, \beta_0, \beta_1, \sigma^2) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^{n} \frac{1}{2\sigma^2} (y_i - (\beta_0 + \beta_1 x_i))^2\right) \\ \ell(\beta_0, \beta_1, \sigma^2) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2 \end{aligned}$$

the MLEs

Set the three partial deriavitives equal to zero, and solve away! Tedious, but do-able.

The estimators are called $\hat{\beta}_0$, $\hat{\beta}_1$ and $\widehat{\sigma^2}$. We end up with the *fitted regression line*:

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

the MLEs

Set the three partial deriavitives equal to zero, and solve away! Tedious, but do-able.

The estimators are called $\hat{\beta}_0$, $\hat{\beta}_1$ and $\widehat{\sigma^2}$. We end up with the *fitted regression line*:

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

When you evaluate this line at the input values from the data, you end up with the *fitted values*:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

the MLEs

Set the three partial deriavitives equal to zero, and solve away! Tedious, but do-able.

The estimators are called $\hat{\beta}_0$, $\hat{\beta}_1$ and $\widehat{\sigma^2}$. We end up with the *fitted regression line*:

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

When you evaluate this line at the input values from the data, you end up with the *fitted values*:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

The distances between the fitted values and the true values are called the *residuals*:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

MLE for σ^2

MLE for σ^2

Of particular interest is $\widehat{\sigma^2}$, which is the solution of the following (with $\hat{\beta}_0$ and $\hat{\beta}_1$ plugged in):

$$
\begin{aligned}
0 &= \frac{\partial \ell}{\partial \sigma^2} \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)^2 \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2
\end{aligned}
$$

MLE for σ^2

Of particular interest is $\widehat{\sigma^2}$, which is the solution of the following (with $\hat{\beta}_0$ and $\hat{\beta}_1$ plugged in):

$$
\begin{aligned}
0 &= \frac{\partial \ell}{\partial \sigma^2} \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2 \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2
\end{aligned}
$$

The solution is:

$$
\widehat{\sigma^2} = \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}
$$

the unbiased estimators for the parameters

After all the work is done, we end up with:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$MSE = \widehat{\sigma^2} \frac{n}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$$

the unbiased estimators for the parameters

After all the work is done, we end up with:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i (y_i - \overline{y})$$

$$MSE = \widehat{\sigma^2} \frac{n}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$$

the distribution of $\hat{\beta}_1$

Since the y_i have normal distributions, it means $\hat{\beta}_1$ has a normal distribution. In particular:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma}{\sqrt{S_{xx}}}\right)$$

So who wants to guess that distribution this thing has?

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{xx}}}$$

R regression output

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.50907    0.18408  -13.63   <2e-16
## log(Body)    1.22496    0.04638   26.41   <2e-16
##
## Residual standard error: 0.8863 on 60 degrees of freedom
## Multiple R-squared:  0.9208, Adjusted R-squared:  0.9195
## F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```