STA286 Lecture 34 Probably a Waste of Time to Have Prepared

Neil Montgomery

Last edited: 2017-04-13 13:10

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

main results

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{xx}}} \sim t_{n-2}$$

The "canonical" regression hypothesis test is $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.

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
```

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
```

So the fitted line is: $y = -2.509 + 1.225x$

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
```

So the fitted line is: $y = -2.509 + 1.225x$

The standard deviation of $\hat{\beta}_1$ is $\sqrt{MSE/S_{xx}} = 0.046$

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
```

So the fitted line is: $y = -2.509 + 1.225x$

The standard deviation of $\hat{\beta}_1$ is $\sqrt{MSE/S_{xx}} = 0.046$

The p-value for the canonical hypothesis test is 9.836×10^{-35}.

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
```

So the fitted line is: $y = -2.509 + 1.225x$

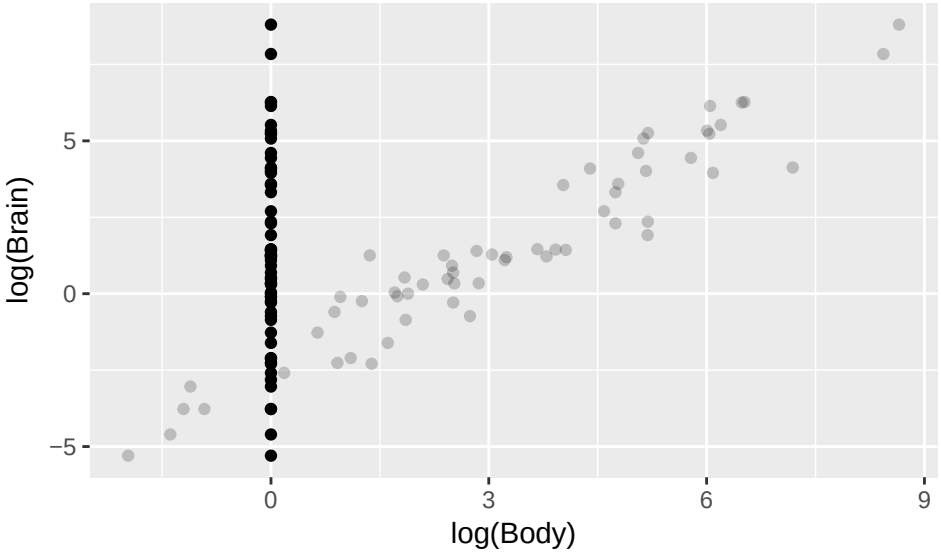
The standard deviation of $\hat{\beta}_1$ is $\sqrt{MSE/S_{xx}} = 0.046$

The p-value for the canonical hypothesis test is 9.836×10^{-35}.

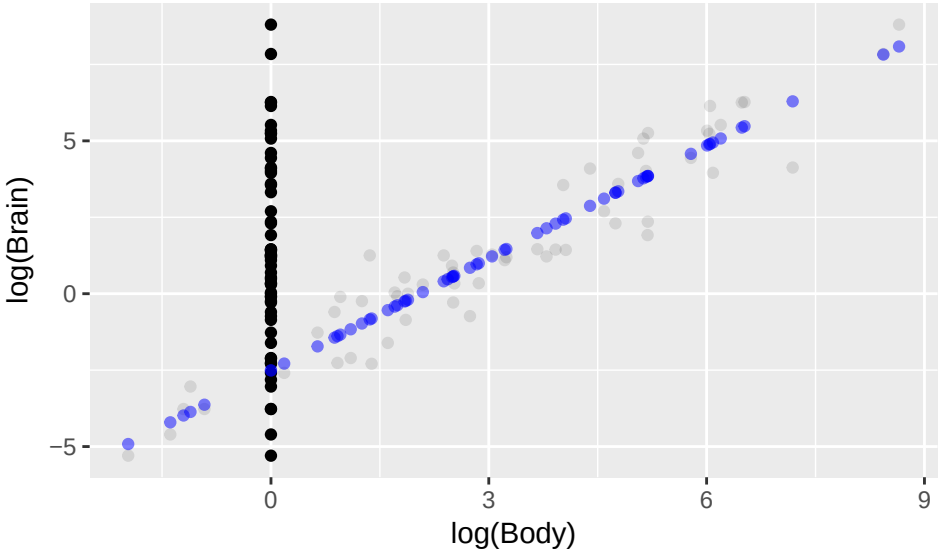
Also on the output is $\sqrt{MSE}$ itself: 0.886

the rest of the information depends on a “sum of squares” result

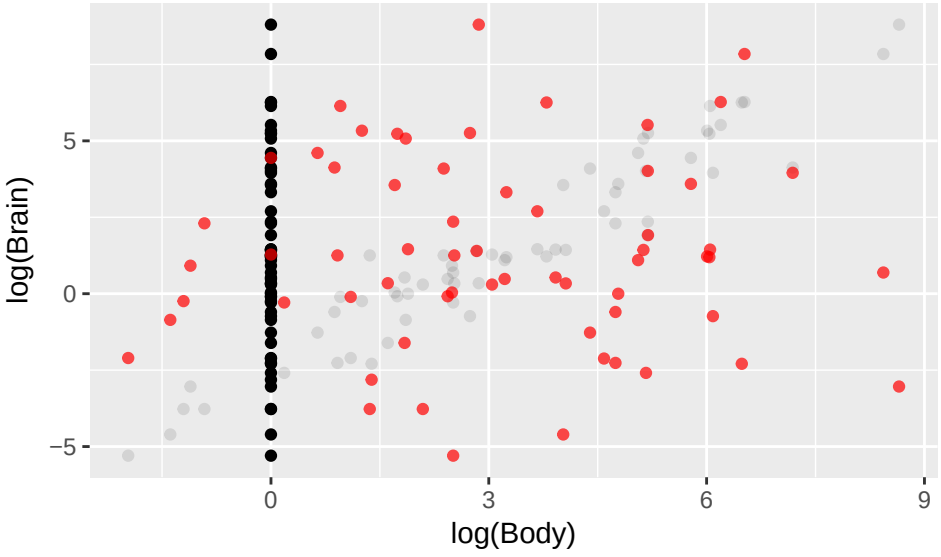
consider the y values - why are they different?



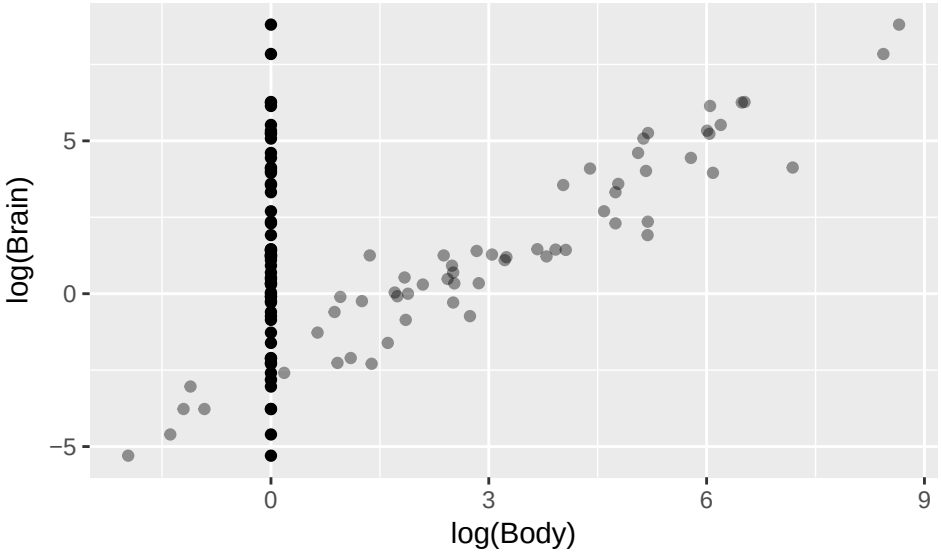
consider the y values - different only because of the line (no noise)



consider the y values - different only because of the noise (no line)



consider the y values - different because of line, and noise



towards an "index" of strength of fit

We would usually measure the model-free variation of y by, say, its sample variance:

$$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}$$

towards an "index" of strength of fit

We would usually measure the model-free variation of y by, say, its sample variance:

$$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1}$$

Consider the numerator only. We can split it into three pieces:

$$\begin{aligned}\sum_{i=1}^{n} (y_i - \overline{y})^2 &= \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2 \\ &= \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n} (\hat{y}_i - \overline{y})(y_i - \hat{y}_i)\end{aligned}$$

the last term is always 0

The sum of the residuals $y_i - \hat{y}_i$ is always 0.

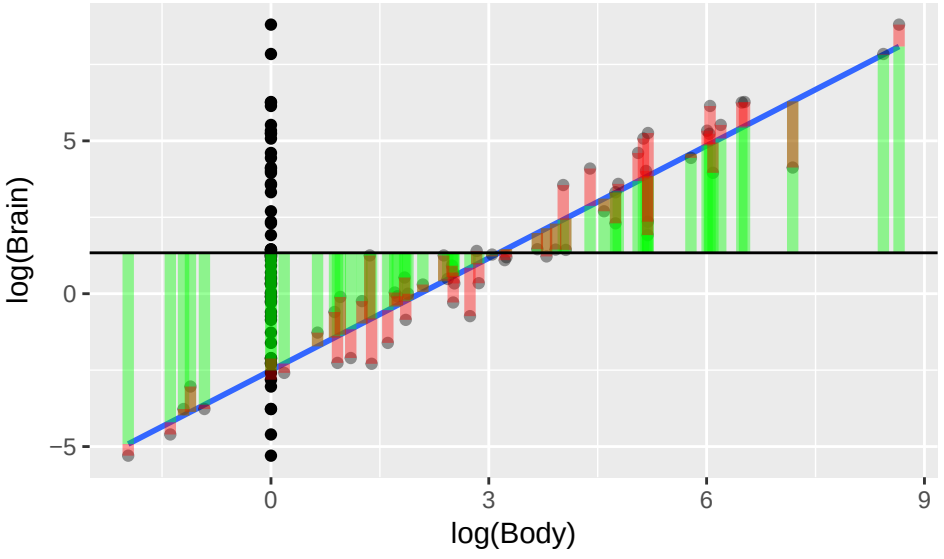
$$\begin{aligned}
\sum_{i=1}^{n} (\hat{y}_i - \overline{y})(y_i - \hat{y}_i) &= \sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 x_i - \overline{y}\right)(y_i - \hat{y}_i) \\
&= \left(\hat{\beta}_0 - \overline{y}\right) \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)}_{\text{always 0}} + \hat{\beta}_1 \underbrace{\sum_{i=1}^{n} x_i \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)}_{\text{Solution of } \frac{\partial \ell}{\partial \beta_1} = 0} \\
&= 0
\end{aligned}$$

Note: this was a nice slide.

the sum of squares decomposition

$$\begin{array}{ccccc} \sum_{i=1}^{n}(y_i-\overline{y})^2 & = & \sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2 & + & \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \\ \text{Total} & = & \text{Regression} & + & \text{Error} \\ SST & = & SSR & + & SSE \end{array}$$

the sum of squares decomposition - a plot nobody will give a shit about



an "index" of strength of fit

The first thing we can do with the SS decomposition is divide through by SST:

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

We define $R^2 = \frac{SSR}{SST}$. It is "the proportion of variation explained by the regression".

It is bounded between 0 and 1, where 1 is a perfect fit and 0 is no relationship at all.*

*Only linear relationships under consideration.

the regression F test

The other thing we can do with the SS decomposition is to ponder the distributions of the three parts:

$$\begin{array}{ccccc} \sum_{i=1}^{n} (y_i - \overline{y})^2 & = & \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 & + & \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ \text{Total} & = & \text{Regression} & + & \text{Error} \\ SST & = & SSR & + & SSE \end{array}$$

the regression F test

The other thing we can do with the SS decomposition is to ponder the distributions of the three parts:

$$\begin{array}{ccccc} \sum_{i=1}^{n}(y_i-\bar{y})^2 & = & \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 & + & \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \\ \text{Total} & = & \text{Regression} & + & \text{Error} \\ SST & = & SSR & + & SSE \\ \chi^2_{n-1} & = & \chi^2_1 & + & \chi^2_{n-2} \end{array}$$

Another way to test $H_0 : \beta_1 = 0$ is by using an F distribution, which is the ratio of χ^2 distrubutions. In this case:

$$\frac{SSR/1}{SSE/(n-2)} \sim F_{1,n-2}$$

In the case of simple regression, this is equivalent to the t distribution version of the test.

the output again

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.5091     0.1841   -13.6   <2e-16
## log(Body)     1.2250     0.0464    26.4   <2e-16
##
## Residual standard error: 0.886 on 60 degrees of freedom
## Multiple R-squared:  0.921,  Adjusted R-squared:  0.919
## F-statistic:  697 on 1 and 60 DF,  p-value: <2e-16
```

how to screw up a regression analysis

Regression modeling is only valid if there is actually a linear relationship between input and output, and if the variance is constant.

You should check this by looking at a plot of residuals versus fitted values.

how to screw up a regression analysis

Regression modeling is only valid if there is actually a linear relationship between input and output, and if the variance is constant.

You should check this by looking at a plot of residuals versus fitted values.

The most common bad idea done by engineers, lawyers, doctors, and various hucksters and wannabes goes by the odd term "ecological fallacy".

This is what happens when you have more than one y value at each x value, and you do the following reasonable-sounding thing: 1. at each x value, calculation the average of the y values. 2. fit the regression with the average y values versus the x values.

log-log body/brain residuals versus fitted values