Multiple Regression

David Gerard

2018-12-07

Objectives

- Introduce Multiple Linear Regression
- Chapters 9 and 10 in the book.

Brain Size

- What variables are associated with brain weight?
- Collected information on 96 different species.
- We know that body weight is already associated with brain weight,
    - So what variables are associated with brain weight **after controlling for body weight**.
- Possible variables: Body weight (kg), gestation period (days), litter size

Brain Size

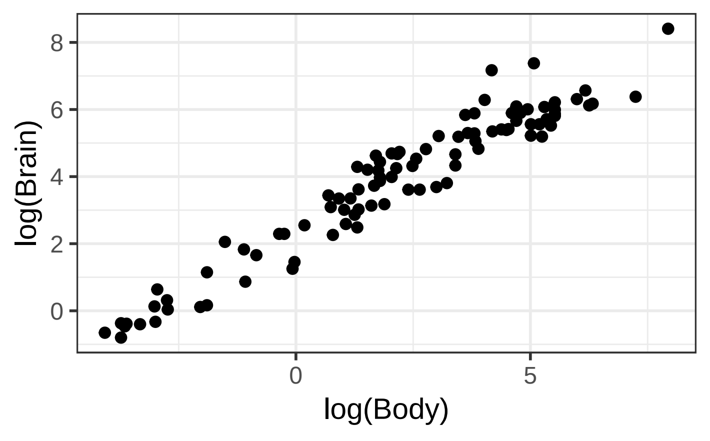
```r
library(Sleuth3)
data("case0902")
head(case0902)
```

```
##             Species  Brain    Body Gestation Litter
## 1          Aardvark    9.6    2.20        31    5.0
## 2          Acouchis    9.9    0.78        98    1.2
## 3 African elephant 4480.0 2800.00       655    1.0
## 4           Agoutis   20.3    2.80       104    1.3
## 5         Axis deer  219.0   89.00       218    1.0
## 6            Badger   53.0    6.00        60    2.2
```

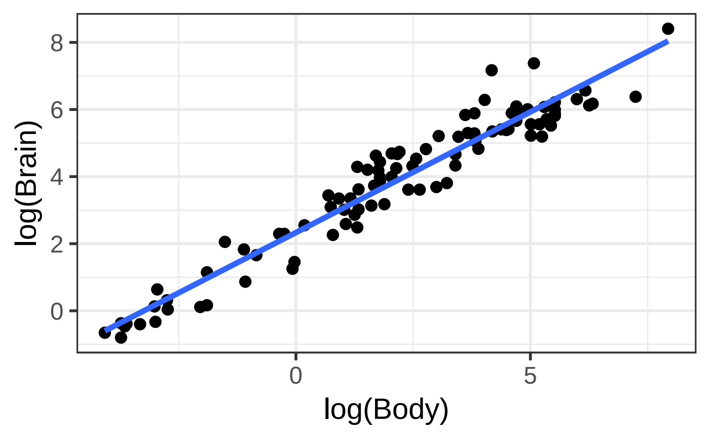
Simple Linear Regression

- One quantitative response variable (Y).
- One quantitative explanatory variable (X).
- The mean of Y is a linear function of X.
- Model the conditional distribution of Y given X.
- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

Simple Linear Regression



Simple Linear Regression



Multiple Linear Regression Model

- One quantitative response (Y).
- **Multiple** quantitative explanatory variables ($X_1, X_2, \ldots, X_p$).
- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \epsilon_i$

Multiple Linear Regression Model

- One quantitative response (Y).
- **Multiple** quantitative explanatory variables ($X_1, X_2, \dots, X_p$).
- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \epsilon_i$
- E.g. X_{2i} is the value of the second explanatory variable for observational unit i.
- E.g., when we have two explanatory variables, this equation is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

- ϵ_i is still ideally normally distributed with mean 0 and constant variance σ^2.

Multiple Linear Regression: Interpreting Coefficients

- Consider Model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$
- To interpret β_1 in a **randomized experiment**:
  - Conceptually fix X_{2i} at a value.
  - Add one to X_{1i}
  - Y_i changes by β_1
- β_1 is how much Y_i increases when we add one to X_{1i} but keep X_{2i} fixed.
- "A one-unit increase in light intensity causes the mean number of flowers to increase by β_1."

Multiple Linear Regression: Interpreting Coefficients

- Consider Model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$
- To interpret β_1 in an **observational study**:
  - The X_i's cannot be fixed independently of one another.
  - Consider a subpopulation that has the same values of the X_j's, where $j \neq 1$. Then the expected difference in means between species that differ in X_1 only by one is β_1.
- β_1 is the expected difference in Y's when we compare species with X_1 and $X_1 + 1$.
- "For any subpopulation of mammal species with the same body weight, species with a one-day longer gestation length tend to have a mean brain-weight β_1 larger."

A useful notation

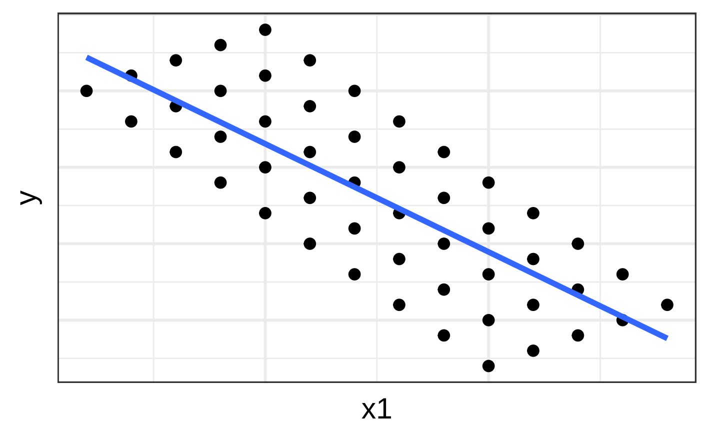
- $\mu(brain|gestation) = \beta_0 + \beta_1 gestation$
    - The mean of brain is equal to β_0 plust β_1 times the gestation time.
- $\mu(brain|gestation, body) = \beta_0 + \beta_1 gestation + \beta_2 body$
    - The mean of brain is equal to β_0 plust β_1 times the gestation time plus β_2 times the body weight.

Interpretation of coefficients changes when the model differs.

- $\mu(brain|gestation) = \beta_0 + \beta_1 gestation$
  - β_1 is the mean difference in brain weight as we compare different gestation periords 1 day apart **in the population of all mammal species**.
- $\mu(brain|gestation, body) = \beta_0 + \beta_1 gestation + \beta_2 body$
  - β_1 is the mean difference in brain weight as we compare different gestation periords 1 day apart **in subpopulations that have the same body weight**.

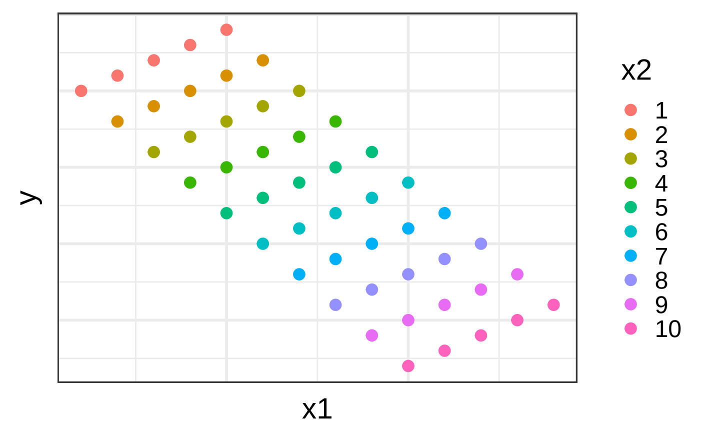
Interpretation of β_1 without X_2 in model

- $\mu(brain|gestation) = \beta_0 + \beta_1 gestation$
- Slope looks negative



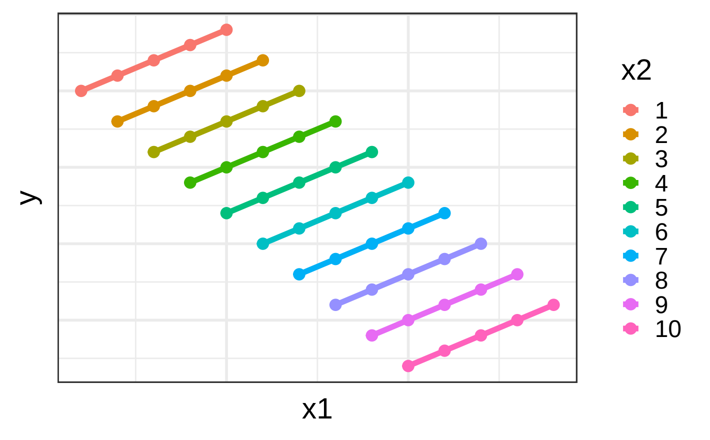
Interpretation of β_1 with X_2 in model

- $\mu(brain|gestation, body) = \beta_0 + \beta_1 gestation + \beta_2 body$
- Slope looks positive at each level of X_2

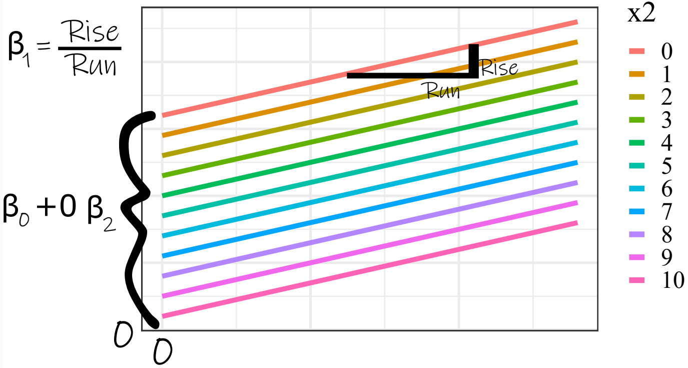


Interpretation of β_1 with X_2 in model

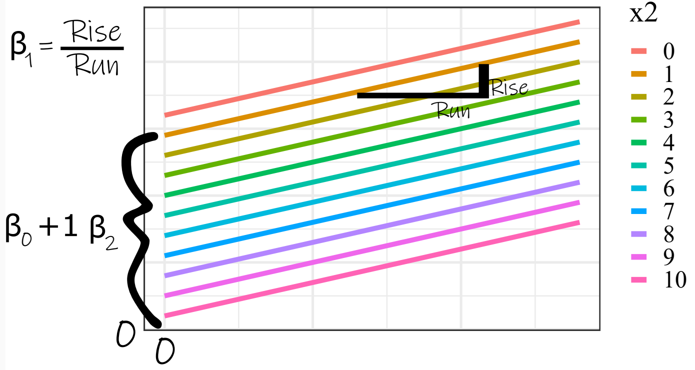
- $\mu(brain|gestation) = \beta_0 + \beta_1 gestation + \beta_2 body$
- Slope looks positive at each level of X_2



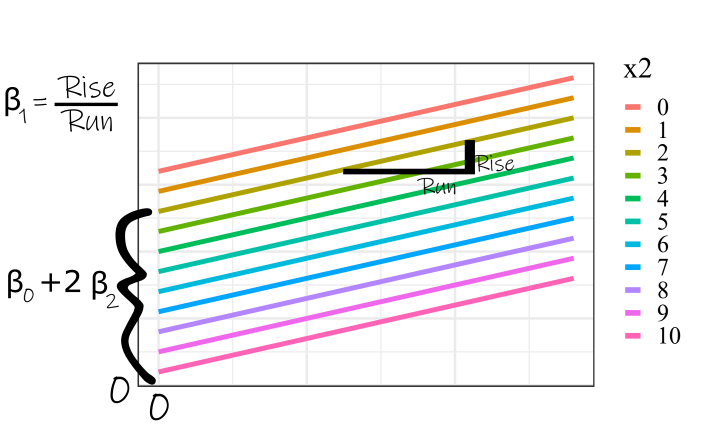
Interpretation of β_1 with X_2 in model



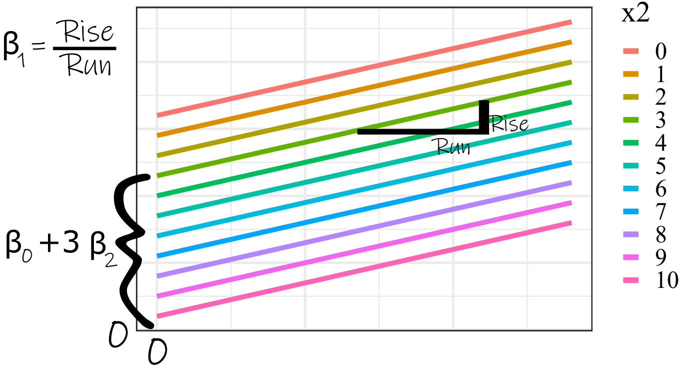
Interpretation of β_1 with X_2 in model



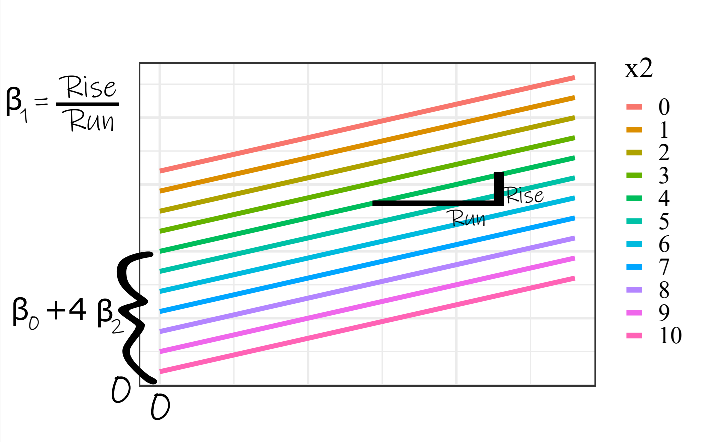
Interpretation of β_1 with X_2 in model



Interpretation of β_1 with X_2 in model



Interpretation of β_1 with X_2 in model



Fitting Multiple Linear Regression in R

How do we estimate the regression coefficients?

- Want to fit:
  $\mu(\textit{brain}|\textit{gestation}, \textit{body}) = \beta_0 + \beta_1 \textit{gestation} + \beta_2 \textit{body}$
- β_0, β_1, β_2 are parameters (we don't know them)
- We can **estimate** them by minimizing the sum of the square residuals.
- Residuals: $Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})$
- The resulting estimates are the OLS (ordinary least squares) estimates: $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$

Syntax

- Want to fit:
  $\mu(brain|gestation, body) = \beta_0 + \beta_1 gestation + \beta_2 body$
- Use `lm()` and **always save the output**.

```
lmout <- lm(Brain ~ Gestation + Body, data = case0902)
lmout
```

```
##
## Call:
## lm(formula = Brain ~ Gestation + Body, data = case0902)
##
## Coefficients:
## (Intercept)    Gestation         Body
##     -112.19         1.45         1.03
```

Table

```
##
## Call:
## lm(formula = Brain ~ Gestation + Body, data = case0902)
##
## Coefficients:
## (Intercept)    Gestation         Body
##     -112.19         1.45         1.03
```

- Interpreting output:

(Intercept)	Gestation	Body
$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$

Another Example

- Want to fit:
  $\mu(brain|gestation, body) = \beta_0 + \beta_1 gestation + \beta_2 body + \beta_3 litter$

```
lmout <- lm(Brain ~ Gestation + Body + Litter,
            data = case0902)
lmout
```

```
##
## Call:
## lm(formula = Brain ~ Gestation + Body + Litter, data = c
##
## Coefficients:
## (Intercept)    Gestation          Body        Litter
##    -225.292        1.809         0.986        27.649
```

Table

```
##
## Call:
## lm(formula = Brain ~ Gestation + Body + Litter, data = d
##
## Coefficients:
## (Intercept)    Gestation         Body       Litter
##    -225.292        1.809        0.986       27.649
```

- Interpreting output:

(Intercept)	Gestation	Body	Litter
$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$

Specific Language we Use

- We estimate that species with 1 day longer gestation time tend to have a brain weight 1.8 grams heavier, after adjusting for body weight and litter size.
- We estimate that species with an average body weight 1 kg heavier tend have a brain weight 0.99 grams heavier, after adjusting for gestation time and litter size.
- We estimate that species with a litter size of one offspring larger tend to have a brain weight 27.6 grams heavier, after adjusting for body weight and gestation time.

Inference

- We are usually interested in testing if $\beta_i = 0$.
- We are usually interested in getting confidence intervals on the β_i's.
- We can use the usual t-tools to get these.

Inference in R

- $\mu(brain|gestation, body) = \beta_0 + \beta_1 gestation + \beta_2 body$

```
lmout <- lm(Brain ~ Gestation + Body, data = case0902)
summary(lmout)
```

```
##
## Call:
## lm(formula = Brain ~ Gestation + Body, data = case0902)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1091.5   -63.2     8.2    67.1  1025.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -112.1920    43.0595   -2.61    0.011
## Gestation      1.4499     0.2752    5.27  8.9e-07
## Body           1.0326     0.0903   11.44  < 2e-16
##
```

Inference in R

- $\mu(brain|gestation, body) = \beta_0 + \beta_1 gestation + \beta_2 body$

```
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -112.192    43.0595  -2.606 1.068e-02
## Gestation      1.450     0.2752   5.268 8.889e-07
## Body           1.033     0.0903  11.436 1.984e-19
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	$\hat{\beta}_0$	$SE(\hat{\beta}_0)$	$\hat{\beta}_0/SE(\hat{\beta}_0)$	p-value for $H_0 : \beta_0 = 0$
Gestation	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$\hat{\beta}_1/SE(\hat{\beta}_1)$	p-value for $H_0 : \beta_1 = 0$
Body	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$\hat{\beta}_2/SE(\hat{\beta}_2)$	p-value for $H_0 : \beta_2 = 0$

Confidence Intervals in R

```r
confint(lmout)
```

```
##                  2.5 %  97.5 %
## (Intercept) -197.6997 -26.684
## Gestation      0.9033   1.996
## Body           0.8533   1.212
```

Interpreting p-values

- The interpretations of significance depends on what other variables are in the model.
- $\mu(brain|gestation, body) = \beta_0 + \beta_1 gestation$
- $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$
- Model under Null: $\mu(brain|gestation, body) = \beta_0$
- Model under Alternative: $\mu(brain|gestation, body) = \beta_0 + \beta_1 gestation$

Language we use

- If we reject H_0, then we say "we have strong evidence that gestation is related to brain weight."
- If we fail to reject H_0, then we say "we do not have strong evidence that gestation is related to brain weight."

Interpreting p-values

- The interpretations of significance depends on what other variables are in the model.
- $\mu(brain|gestation, body) = \beta_0 + \beta_1 gestation + \beta_2 body$
- $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$
- Model under Null: $\mu(brain|gestation, body) = \beta_0 + \beta_2 body$
- Model under Alternative: $\mu(brain|gestation, body) = \beta_0 + \beta_1 gestation + \beta_2 body$

Language we use

- If we reject H_0, then we say "we have strong evidence that gestation is related to brain weight after adjusting for body size."
- If we fail to reject H_0, then we say "we do not have strong evidence that gestation is related to brain weight after adjusting for body size."