Regression analysis (in R)

Catalina Vallejos (The Alan Turing Institute and UCL)
Aaron Lun (Cancer Research UK - Cambridge Institute)

December 9th, 2016

Outline for the course

- **Introduction to statistics**
- **Hypothesis testing (in R)**
- **Regression analysis (in R)**

Regression analysis

In statistics, **regression analysis** is a tool to quantify relationships between 2 or more variables

Regression analysis

In statistics, **regression analysis** is a tool to quantify relationships between 2 or more variables

In regression, variables are clasified as

- **Response**: outcome variable of interest (dependent)
- **Explanatory**: covariate(s) to explain the response (independent)

Examples?

Regression analysis

There are different types of regression models, which one to use depends on the properties of the analysed variables. For example,

- For a binary response (e.g. failure/success), a common choice is **logistic regression**

Regression analysis

There are different types of regression models, which one to use depends on the properties of the analysed variables. For example,

- For a binary response (e.g. failure/success), a common choice is **logistic regression**
- For a continuous response (e.g. blood sugar level), a typical choice is **linear regression**

Regression analysis

There are different types of regression models, which one to use depends on the properties of the analysed variables. For example,

- For a binary response (e.g. failure/success), a common choice is **logistic regression**
- For a continuous response (e.g. blood sugar level), a typical choice is **linear regression**

Today, we will focus on linear regression

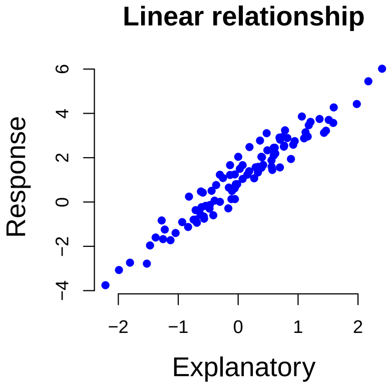
Linear regression

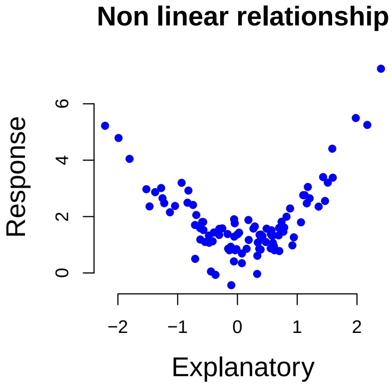
Linear regression

Linear regression is a statistical tool to capture **linear** relationships between a single **continuous** response variable and one or more explanatory variables (may be discrete or continuous)

Linear regression

Linear regression is a statistical tool to capture **linear** relationships between a single **continuous** response variable and one or more explanatory variables (may be discrete or continuous)


Linear relationship

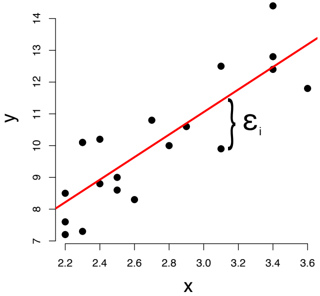

Non linear relationship

Simple linear regression

In **simple linear regression**, the aim is to capture a linear relationship between a response (y) and a **single** covariate (x)

The simple linear regression model is writen as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, \ldots, n$$

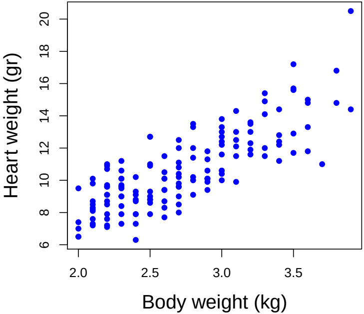


- β_0 is the intercept
- β_1 is the slope (gradient)
- ϵ_i is the error term

(*) Plot the response on the vertical axis and the covariate on the horizontal axis

Simple linear regression: an example

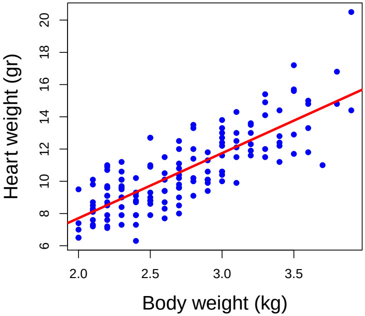
The data in Fisher (1947) shows heart and body weights for 144 cats



Is there a relationship between heart and body weight?

Simple linear regression: an example

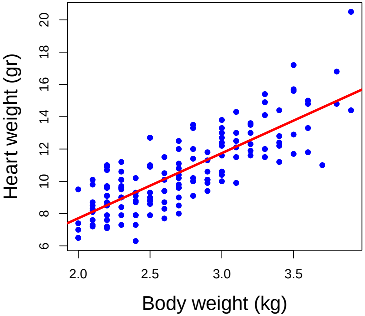
The data in Fisher (1947) shows heart and body weights for 144 cats



Is there a linear relationship between heart and body weight?

Simple linear regression: an example

The data in Fisher (1947) shows heart and body weights for 144 cats



We can fit a simple linear regression

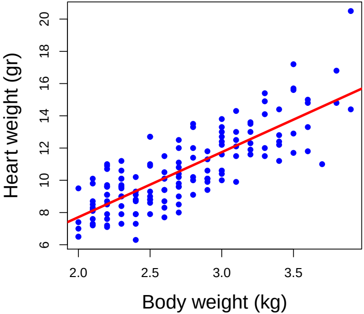
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where

- y_i = heart weight of cat i
- x_i = body weight of cat i

Simple linear regression: an example

The data in Fisher (1947) shows heart and body weights for 144 cats



We can fit a simple linear regression

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

where

- y_i = heart weight of cat i
- x_i = body weight of cat i

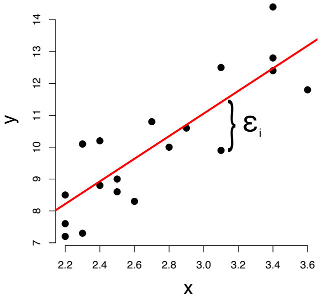
How to estimate β_0 and β_1?

Simple linear regression

Idea: find β_0 and β_1 such that the line is a good fit for the data

Simple linear regression

Idea: find β_0 and β_1 such that the line is a good fit for the data

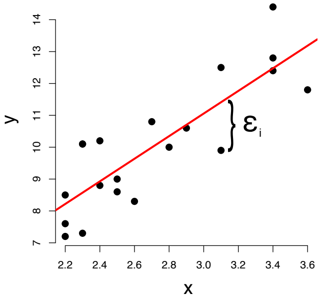


For example, by minimizing the **residuals** of the regression

$$\epsilon_i = Y_i - (\beta_0 + \beta_1 x_i) \equiv Y_i - \hat{Y}_i$$

Simple linear regression

Idea: find β_0 and β_1 such that the line is a good fit for the data



For example, by minimizing the **residuals** of the regression

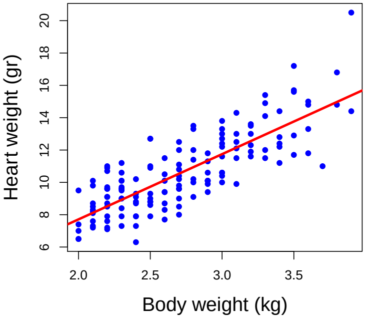
$$\epsilon_i = Y_i - (\beta_0 + \beta_1 x_i) \equiv Y_i - \hat{Y}_i$$

Typically, this is done by minimizing the **Sum of Square Errors (SSE)**

$$\text{SSE} = \sum_{i=1}^{n} \epsilon_i^2$$

Simple linear regression: an example

The data in Fisher (1947) shows heart and body weights for 144 cats



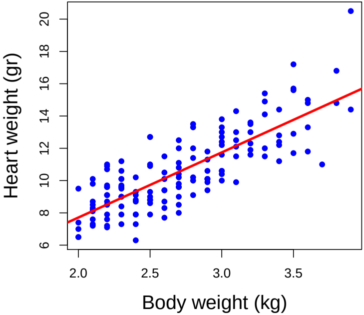
In this example we have

- β_0 = -0.36
- β_1 = 4.03

How to interpret this?

Simple linear regression: an example

The data in Fisher (1947) shows heart and body weights for 144 cats



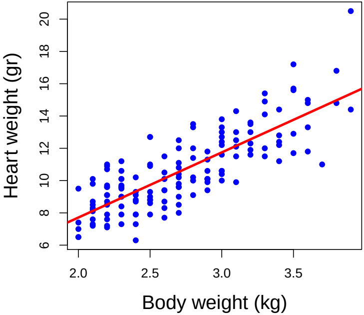
In this example we have

- $\beta_0 = -0.36$
- $\beta_1 = 4.03$

How to interpret this?
For a 1kg increase in body weight, we expect heart weight to increase by 4.03gr

Simple linear regression: an example

The data in Fisher (1947) shows heart and body weights for 144 cats



In this example we have

- $\beta_0 = -0.36$
- $\beta_1 = 4.03$

How to interpret this?

For a 1kg increase in body weight, we expect heart weight to increase by 4.03gr

Is this increase statistically significant?

Simple linear regression

We can answer this using hypothesis testing:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

Recall: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

→ if $\beta_1 = 0$, y_i and x_i are not (linearly) dependent

Simple linear regression

We can answer this using hypothesis testing:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

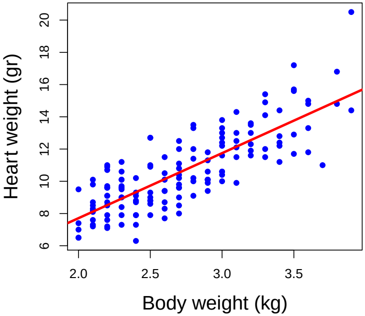
Recall: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

→ if $\beta_1 = 0$, y_i and x_i are not (linearly) dependent

Assumming $\epsilon_i \sim \text{Normal}(0, \sigma^2)$, we can derive a *t*-test for β_1

Simple linear regression: an example

The data in Fisher (1947) shows heart and body weights for 144 cats



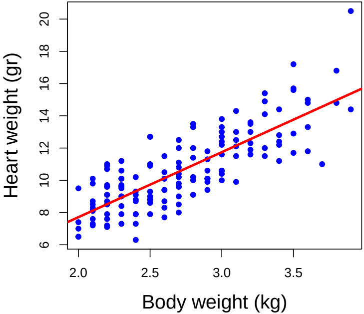
In this example we have

- $\beta_1 = 4.03$

The *p*-value is $< 2 \times 10^{-16}$

Simple linear regression: an example

The data in Fisher (1947) shows heart and body weights for 144 cats



In this example we have

- $\beta_1 = 4.03$

The *p*-value is $< 2 \times 10^{-16}$

We reject H_0, i.e. we conclude that the linear relationship between heart and body weight is statistically significant ($\alpha = 0.05$)

Simple linear regression: Analysis of Variance (ANOVA)

ANOVA aims to decompose the total variance of y

Recall: $\mathrm{Var}(y) = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$, with $\bar{y} = (\sum_{i=1}^{n} y_i)/n$

Simple linear regression: Analysis of Variance (ANOVA)

ANOVA aims to decompose the total variance of y

Recall: $\text{Var}(y) = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$, with $\bar{y} = (\sum_{i=1}^{n} y_i)/n$

$$\underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{SSR (Regression)}} = \underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SST (Total)}} - \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{SSE (Error)}}$$

- SST quantifies the total variability of y
- SSE quantifies residual variability of y (unexplained by x)
- SSR quantifies how much of the variability of y is explained by x

Simple linear regression: Analysis of Variance (ANOVA)

The **coefficient of determination** R^2 is defined as the proportion of total variability of y that is explained by x

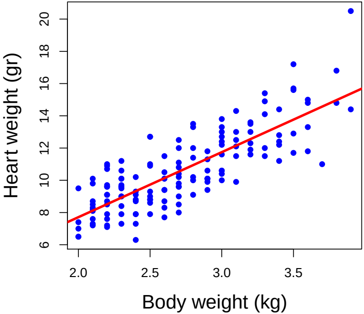
$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

- $0 \leq R^2 \leq 1$
- If $R^2 = 0$, x explains none of the variability of y
- If $R^2 = 1$, x explains all of the variability of y (perfect fit!)

Warning: this assumes the relationship between x and y is linear

Simple linear regression: an example

The data in Fisher (1947) shows heart and body weights for 144 cats



In this example we have

- $R^2 = 0.65$

Body weight explains 65% of the variability of a cat's heart weight

Simple linear regression: Analysis of Variance (ANOVA)

This variance decomposition is summarized by the **ANOVA table**

Source	**Degrees of freedom**	**Sum of Squares**	**Mean Square**	**F ratio**
Regression	1	SSR	SSR/1	$\frac{\text{SSR}}{1} \ / \ \frac{\text{SSE}}{n-2}$
Error	$n-2$	SSE	SSE/$(n-2)$	
Total	$n-1$	SST	SST/$(n-1)$	

- Assuming $\epsilon_i \sim \text{Normal}(0, \sigma^2)$, the F ratio can be used to test

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0$$

How does this compare to the t-test introduced earlier?

Simple linear regression versus correlation

Pearson's correlation ($\rho_{x,y}$) quantifies linear dependency between variables x and y

Simple linear regression versus correlation

Pearson's correlation ($\rho_{x,y}$) quantifies linear dependency between variables x and y

How does this relate to simple linear regression?

Simple linear regression versus correlation

Pearson's correlation ($\rho_{x,y}$) quantifies linear dependency between variables x and y

How does this relate to simple linear regression?

Both methods are closely related:

$$\rho_{xy} = \sqrt{R^2}$$

Simple linear regression versus correlation

Pearson's correlation ($\rho_{x,y}$) quantifies linear dependency between variables x and y

How does this relate to simple linear regression?

Both methods are closely related:

$$\rho_{xy} = \sqrt{R^2}$$

However,

- in linear regression, x and y receive an **asymmetric** treatment
- in correlation, x and y receive a **symmetric** treatment

Simple linear regression versus correlation

Pearson's correlation ($\rho_{x,y}$) quantifies linear dependency between variables x and y

How does this relate to simple linear regression?

Both methods are closely related:

$$\rho_{xy} = \sqrt{R^2}$$

However,

- in linear regression, x and y receive an **asymmetric** treatment
- in correlation, x and y receive a **symmetric** treatment

Examples?

Simple linear regression versus correlation

Pearson's correlation ($\rho_{x,y}$) quantifies linear dependency between variables x and y

How does this relate to simple linear regression?

Both methods are closely related:

$$\rho_{xy} = \sqrt{R^2}$$

However,

- in linear regression, x and y receive an **asymmetric** treatment
- in correlation, x and y receive a **symmetric** treatment

Examples?

Let's play: `http://guessthecorrelation.com`

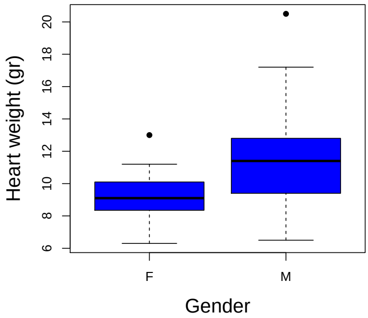
Questions + practical

http://bioinformatics.cruk.cam.ac.uk/apps/winter-school/linear/

Simple linear regression: categorical covariates

Simple linear regression: an example

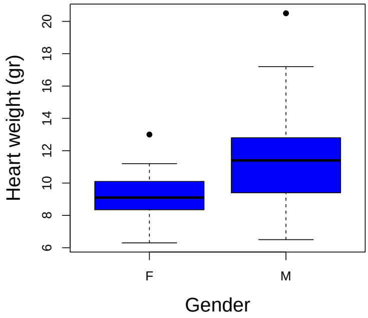
The data in Fisher (1947) also contains gender information



Is there a relationship between gender and a cat's heart weight?

Simple linear regression: an example

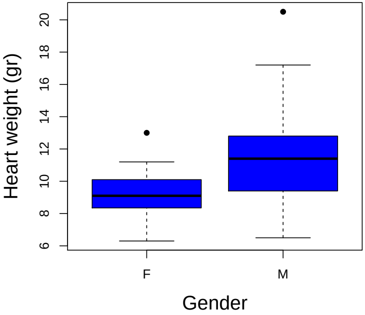
The data in Fisher (1947) also contains gender information



In principle, we can answer this question using a *t*-test

Simple linear regression: an example

The data in Fisher (1947) also contains gender information



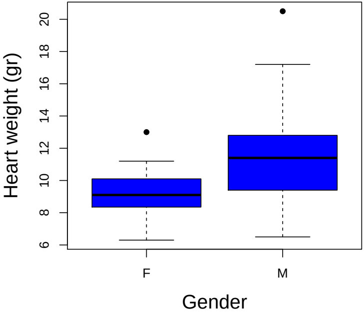
In principle, we can answer this question using a *t*-test

In this example we have:

$$t = \text{-}5.35$$
$$p\text{-value} = 3.38 \times 10^{-7}$$

Simple linear regression: an example

The data in Fisher (1947) also contains gender information



In principle, we can answer this question using a *t*-test

In this example we have:

$$t = \text{-}5.35$$
$$p\text{-value} = 3.38 \times 10^{-7}$$

Can we answer the same question using linear regression?

Simple linear regression: categorical covariates

Linear regression also allows us to answer this question by treating the grouping variable (e.g. gender) as a **categorical covariate**

Simple linear regression: categorical covariates

Linear regression also allows us to answer this question by treating the grouping variable (e.g. gender) as a **categorical covariate**

However, **categorical covariates** require an special treatment ...

Recall: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

Instead of defining a single regression effect β_1, we need to estimate an effect for **each level** of the categorical covariate

Simple linear regression: categorical covariates

Linear regression also allows us to answer this question by treating the grouping variable (e.g. gender) as a **categorical covariate**

However, **categorical covariates** require an special treatment . . .

Recall: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

Instead of defining a single regression effect β_1, we need to estimate an effect for **each level** of the categorical covariate

This can be done using **dummy variables**

Simple linear regression: categorical covariates

In this context, **dummy variables** are used as **binary** indicators associated to particular levels of a categorical covariate

Simple linear regression: categorical covariates

In this context, **dummy variables** are used as **binary** indicators associated to particular levels of a categorical covariate

For example,

$$D_i = \begin{cases} 1, \text{if cat } i \text{ is male;} \\ 0, \text{if cat } i \text{ is female.} \end{cases}$$

Simple linear regression: categorical covariates

In this context, **dummy variables** are used as **binary** indicators associated to particular levels of a categorical covariate

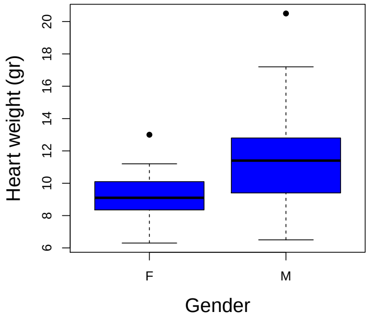
For example,

$$D_i = \begin{cases} 1, \text{if cat } i \text{ is male;} \\ 0, \text{if cat } i \text{ is female.} \end{cases}$$

Note: gender has **two** levels, but we only define **one** dummy variable (here we left *female* as a **reference category**)

Simple linear regression: an example

The data in Fisher (1947) also contains gender information

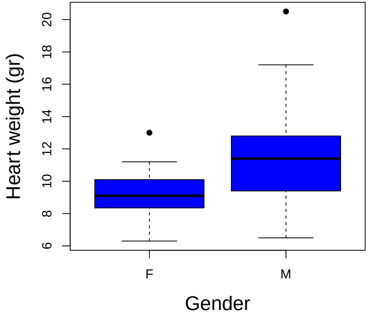


In this example we have

$$y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

Simple linear regression: an example

The data in Fisher (1947) also contains gender information



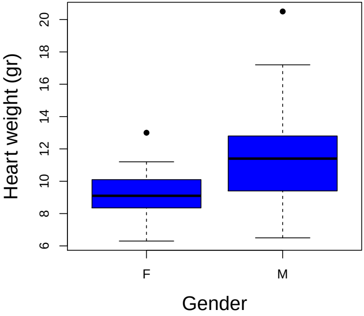
In this example we have

$$y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

- For a male cat: $\hat{y}_i = \beta_0 + \beta_1$
- For a female cat: $\hat{y}_i = \beta_0$

Simple linear regression: an example

The data in Fisher (1947) also contains gender information



In this example we have

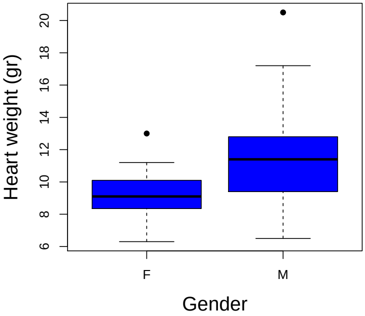
$$y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

- For a male cat: $\hat{y}_i = \beta_0 + \beta_1$
- For a female cat: $\hat{y}_i = \beta_0$

$\Rightarrow$ β_1 quantifies the difference between female and male cats

Simple linear regression: an example

Is there a relationship between gender and a cat's heart weight?

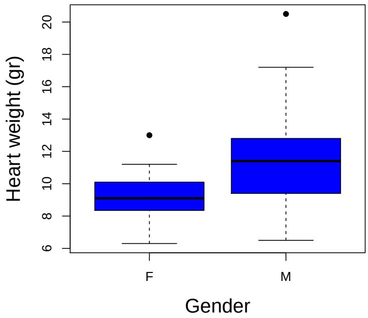


We can answer this using the test

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

Simple linear regression: an example

Is there a relationship between gender and a cat's heart weight?



We can answer this using the test

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

Assuming $\epsilon_i \sim \text{Normal}(0, \sigma^2)$, we derive a t-test:

$$\beta_1 = 2.12$$
$$t = 5.35$$
$$p\text{-value} = 3.38 \times 10^{-7}$$

Multiple linear regression

Multiple linear regression

Multiple linear regression is an extension of simple linear regression that allows more than 1 covariate

Multiple linear regression

Multiple linear regression is an extension of simple linear regression that allows more than 1 covariate

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip} + \epsilon_i$$

Multiple linear regression: an example

For example, with 2 continuous covariates:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

where

- y_i: heart weight for cat i
- x_{i1}: body weight for cat i
- x_{i2}: age for cat i

Multiple linear regression

In multiple linear regression, there are 2 types of **hypothesis**:

Multiple linear regression

In multiple linear regression, there are 2 types of **hypothesis**:

- **Marginal**: to assess the significance of a single covariate

$$H_0 : \beta_j = 0 \text{ vs } H_1 : \beta_j \neq 0$$

Multiple linear regression

In multiple linear regression, there are 2 types of **hypothesis**:

- **Marginal**: to assess the significance of a single covariate

$$H_0 : \beta_j = 0 \text{ vs } H_1 : \beta_j \neq 0$$

- **Global**: to assess the joint effect of all covariates

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \text{ vs } H_1 : \text{at least one } \beta_j \neq 0$$

Multiple linear regression

In multiple linear regression, there are 2 types of **hypothesis**:

- **Marginal**: to assess the significance of a single covariate

  $$H_0 : \beta_j = 0 \text{ vs } H_1 : \beta_j \neq 0$$

  Assuming $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, we can use a *t*-test

- **Global**: to assess the joint effect of all covariates

  $$H_0 : \beta_1 = \cdots = \beta_p = 0 \text{ vs } H_1 : \text{at least one } \beta_j \neq 0$$

Multiple linear regression

In multiple linear regression, there are 2 types of **hypothesis**:

- **Marginal**: to assess the significance of a single covariate

  $$H_0 : \beta_j = 0 \text{ vs } H_1 : \beta_j \neq 0$$

  Assuming $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, we can use a *t*-test

- **Global**: to assess the joint effect of all covariates

  $$H_0 : \beta_1 = \cdots = \beta_p = 0 \text{ vs } H_1 : \text{at least one } \beta_j \neq 0$$

Assuming $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, we can use an *F*-test (from ANOVA table)

Multiple linear regression: an example

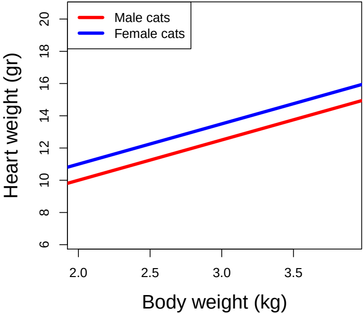
. . . or mixing a continuous and a categorical covariate

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

where

- y_i: heart weight for cat i
- x_{i1}: body weight for cat i
- $x_{i2} = \begin{cases} 1, \text{if cat } i \text{ is male;} \\ 0, \text{if cat } i \text{ is female.} \end{cases}$

Multiple linear regression: an example



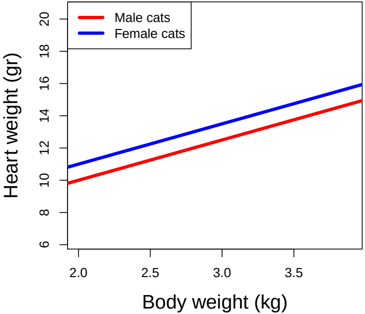
For a male cat:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2$$

For a female cat:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1}$$

Multiple linear regression: an example



For a male cat:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2$$

For a female cat:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1}$$

⇒ parallel lines, effect of body weight is independent of gender

⇒ β_2 quantifies a global difference between female and male cats

Multiple linear regression: an example

What if we think the effect of body weight depends on gender?

Multiple linear regression: an example

What if we think the effect of body weight depends on gender?

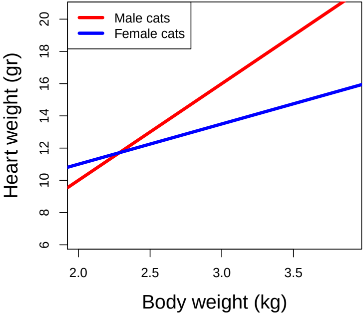
We can define an **interaction effect**

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i12} + \epsilon_i,$$

Recall:

- y_i: heart weight for cat i
- x_{i1}: body weight for cat i
- $x_{i2} = \begin{cases} 1, \text{if cat } i \text{ is male;} \\ 0, \text{if cat } i \text{ is female.} \end{cases}$

Multiple linear regression: an example



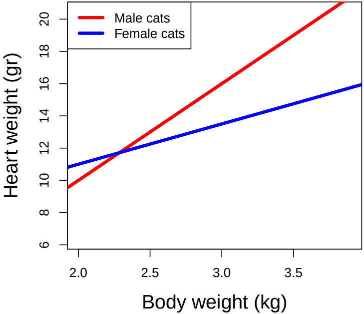
For a male cat:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 + \beta_3 x_{i1}$$

For a female cat:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1}$$

Multiple linear regression: an example



For a male cat:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 + \beta_3 x_{i1}$$

For a female cat:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1}$$

$\Rightarrow$ crossing lines, effect of body weight depends on gender

$\Rightarrow$ β_2 quantifies a global difference between female and male cats

$\Rightarrow$ β_3 quantifies the differential effect of body weight

Questions + practical

Assumptions in linear regression

Assumptions in linear regression

Recall: the multiple linear regression model is written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip} + \epsilon_i$$

What assumptions underlie this model?

Assumptions in linear regression

- Firstly, we need more observations than variables ($n > p$)

Assumptions in linear regression

- Firstly, we need more observations than variables ($n > p$)

This might not hold in some genomics applications ...

... but that's outside the scope of this course!

Assumptions in linear regression

- Firstly, we need more observations than variables ($n > p$)

This might not hold in some genomics applications ...

... but that's outside the scope of this course!

- Secondly, we assume that residuals ϵ_i are **independent** and **identically distributed** with

$$\epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

Assumptions in linear regression

- Firstly, we need more observations than variables ($n > p$)

This might not hold in some genomics applications ...

... but that's outside the scope of this course!

- Secondly, we assume that residuals ϵ_i are **independent** and **identically distributed** with

$$\epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

How can we assess this?

Assumptions in linear regression

Idea: after estimating the regression coefficients, define residuals as

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

Use these estimated residuals to diagnose model quality

Assumptions in linear regression

Idea: after estimating the regression coefficients, define residuals as

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

Use these estimated residuals to diagnose model quality

For example, we can prepare a _ _ _ _ _ _ _ _ _ to see if residuals are normally distributed

Assumptions in linear regression

Idea: after estimating the regression coefficients, define residuals as

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

Use these estimated residuals to diagnose model quality

For example, we can prepare a _ _ _ _ _ _ _ _ _ to see if residuals are normally distributed

More about this in the practical ...

Assumptions in linear regression

What if things go wrong?

Assumptions in linear regression

What if things go wrong?

If the residuals are not good enough to pass the diagnostic criteria, we need to revisit the model. For example

- We can transform some of the covariates or
- We can transform the response variable

Assumptions in linear regression: covariate transformation

Suppose a simple linear regression model

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

Covariate transformation can be useful in situations where the relationship between y and x is **not linear**

Assumptions in linear regression: covariate transformation

Suppose a simple linear regression model

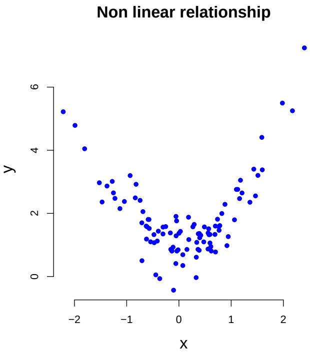
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Covariate transformation can be useful in situations where the relationship between y and x is **not linear**

Idea: replace x_i by a transformed version of x_i (e.g. $x_i^* = x_i^2$)

Assumptions in linear regression: covariate transformation

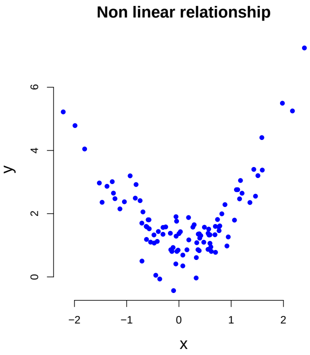
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

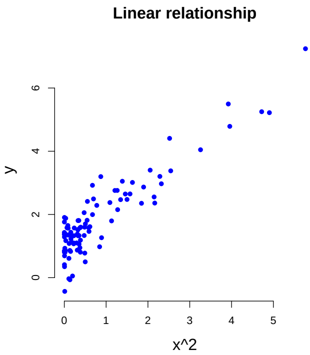


Assumptions in linear regression: covariate transformation

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

$$y_i = \beta_0 + \beta_1 x_i^2 + \epsilon_i$$





Assumptions in linear regression: response transformation

Suppose a simple linear regression model

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

Response transformation can be useful in situations where the variance of y is not constant as a function of x

We refer to this as **heteroskedastic** errors

Assumptions in linear regression: response transformation

Suppose a simple linear regression model

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

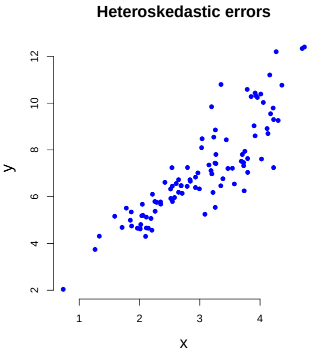
Response transformation can be useful in situations where the variance of y is not constant as a function of x

We refer to this as **heteroskedastic** errors

Idea: replace y_i by a transformed version of y_i (e.g. $y_i^* = \log(y_i)$)

Assumptions in linear regression: response transformation

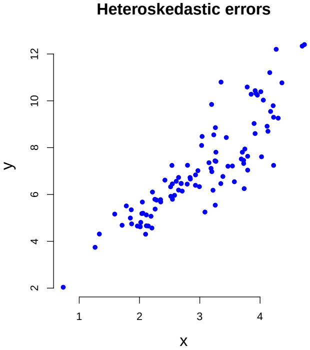
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

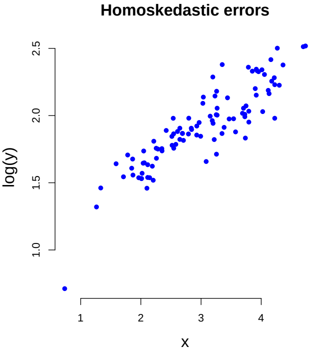


Assumptions in linear regression: response transformation

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

$$\log(y)_i = \beta_0 + \beta_1 x_i + \epsilon_i$$





Assumptions in linear regression

Finally, an important assumption underlying all what we discussed today is that the response variable y is **continuous**

Assumptions in linear regression

Finally, an important assumption underlying all what we discussed today is that the response variable *y* is **continuous**

What if this is not true?

Assumptions in linear regression

Finally, an important assumption underlying all what we discussed today is that the response variable y is **continuous**

What if this is not true?

- **Count-based response**

e.g. y_i: no. of relapses after treatment

Assumptions in linear regression

Finally, an important assumption underlying all what we discussed today is that the response variable y is **continuous**

What if this is not true?

- **Count-based response**

  e.g. y_i: no. of relapses after treatment

- **Binary response**

  e.g. $y_i = 1$: if treatment is effective; $y_i = 0$ if not

Assumptions in linear regression

Finally, an important assumption underlying all what we discussed today is that the response variable y is **continuous**

What if this is not true?

- **Count-based response**

  e.g. y_i: no. of relapses after treatment

- **Binary response**

  e.g. $y_i = 1$: if treatment is effective; $y_i = 0$ if not

We can use **Generalized Linear Models ...**

Questions + practical