Linear regression an ANOVA

February 11, 2014

Aims

- First describe linear regression and multiple linear regression

Aims

- First describe linear regression and multiple linear regression
- Look at analysis of variance ANOVA

Aims

- First describe linear regression and multiple linear regression
- Look at analysis of variance ANOVA
- Apply these methods using R

Aims

- First describe linear regression and multiple linear regression
- Look at analysis of variance ANOVA
- Apply these methods using R
- Look at how to apply these methods to genetic data

Linear regression

The basic idea of a linear regression is to describe the relationship between two variables

Linear regression

The basic idea of a linear regression is to describe the relationship between two variables

- In the first instance we will look at two continuous variables.

Linear regression

The basic idea of a linear regression is to describe the relationship between two variables

- In the first instance we will look at two continuous variables.
- The best way to explain this method is through an example

Linear regression

We have some data detailing tick counts on some cattle. For each cattle, we have a weight measurement, their breed and also whether they were housed inside or outside.

Linear regression

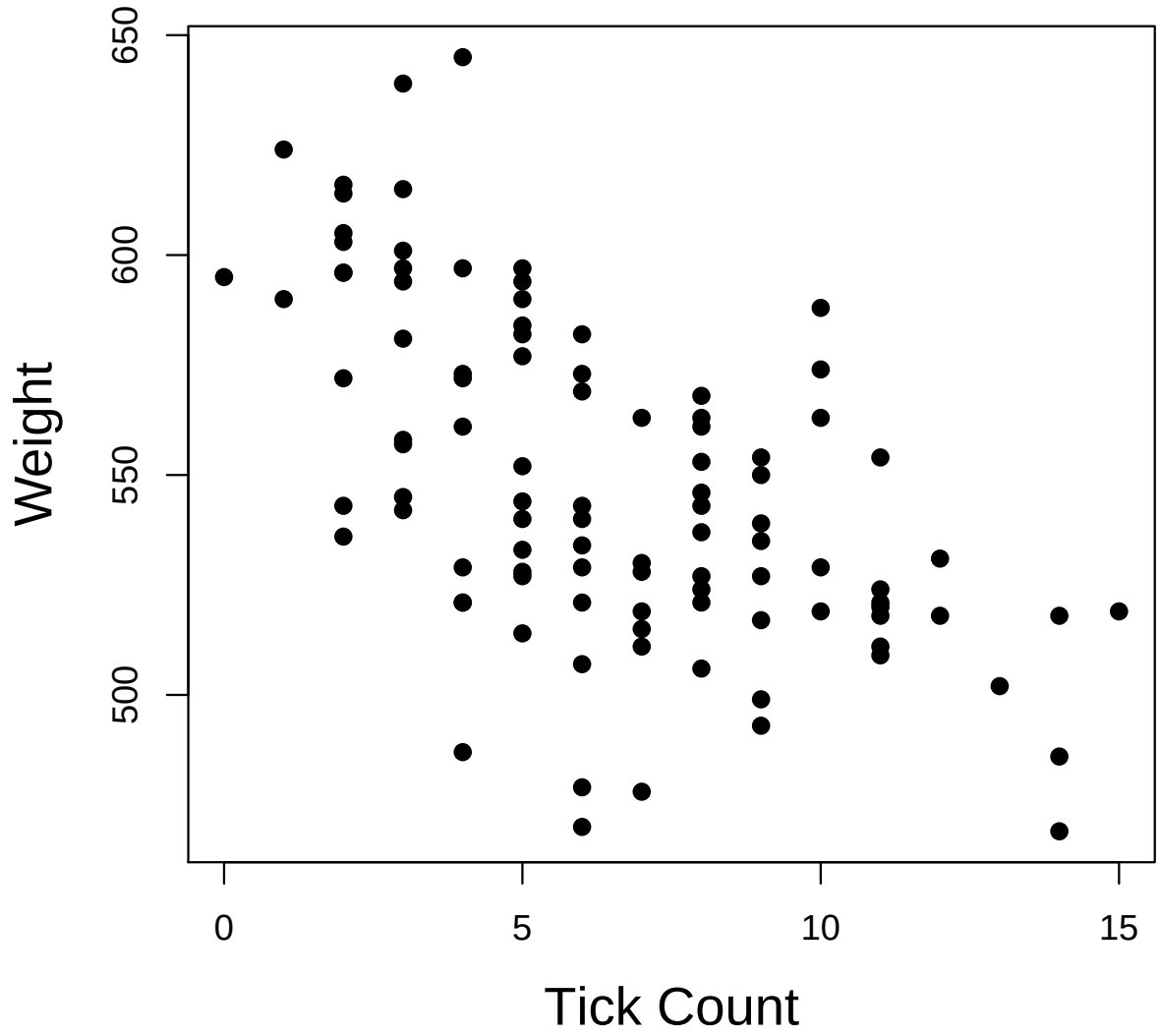
We have some data detailing tick counts on some cattle. For each cattle, we have a weight measurement, their breed and also whether they were housed inside or outside.

Linear regression

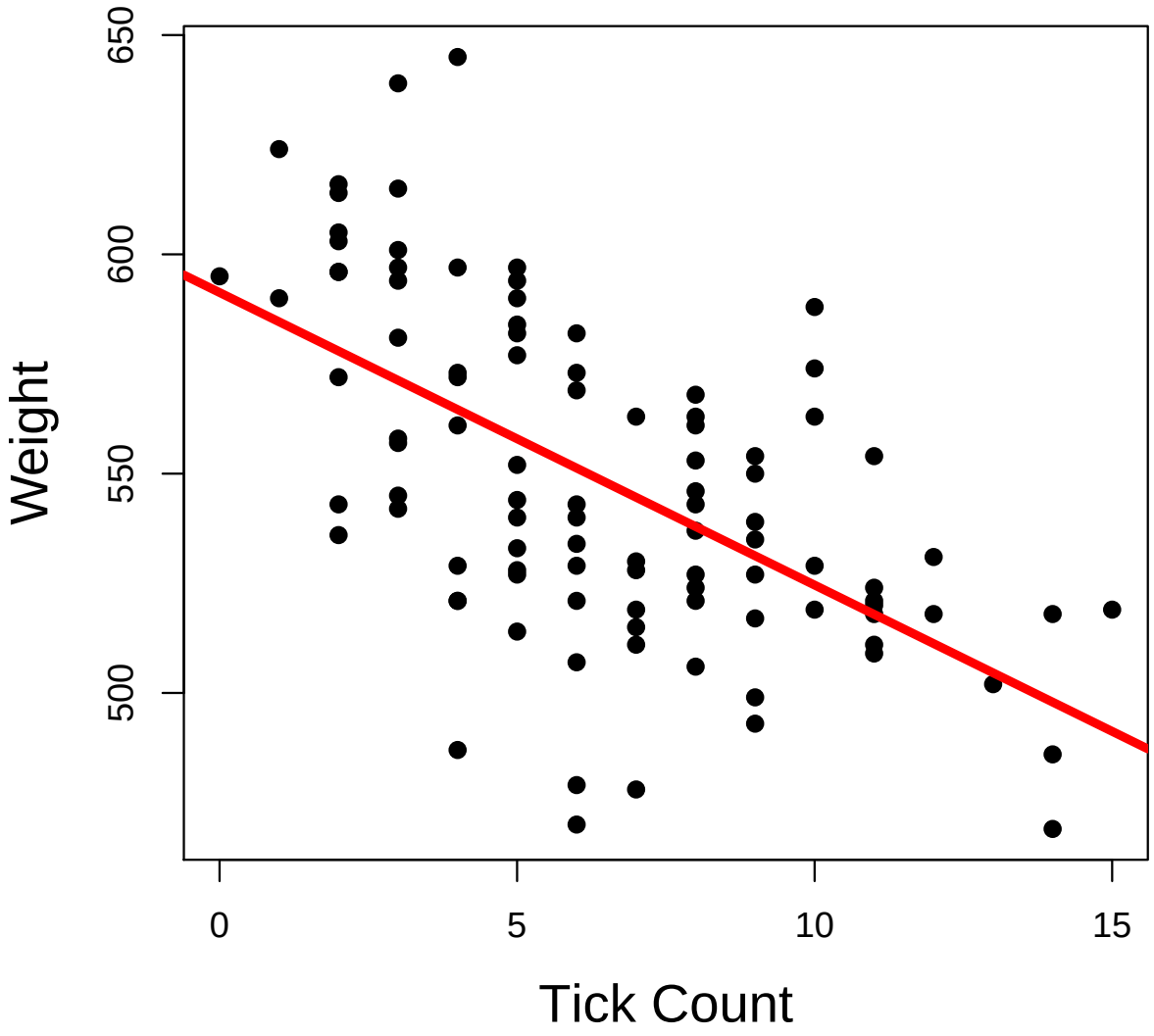
We have some data detailing tick counts on some cattle. For each cattle, we have a weight measurement, their breed and also whether they were housed inside or outside.

- We can display this data in a scatterplot

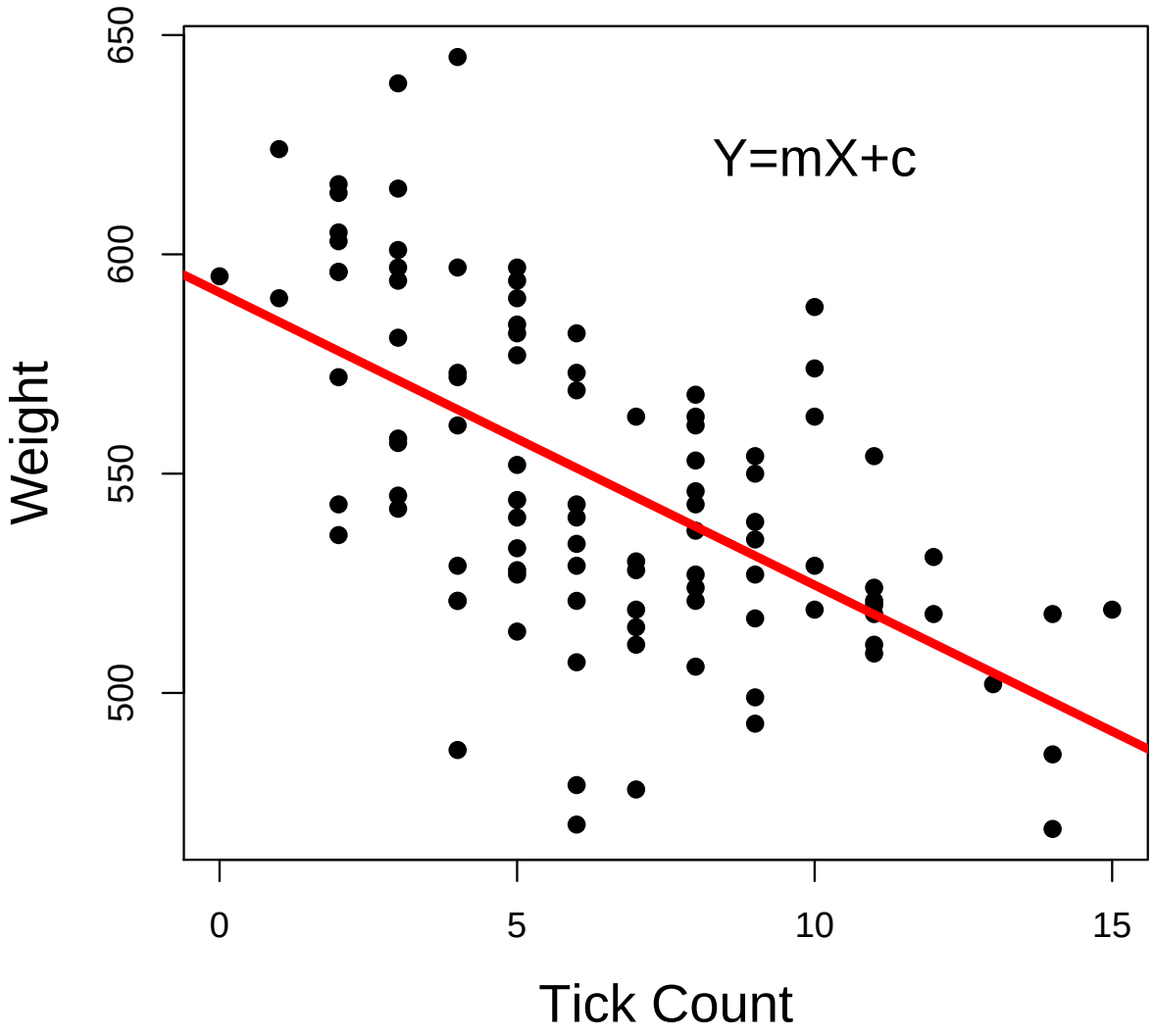
Linear regression



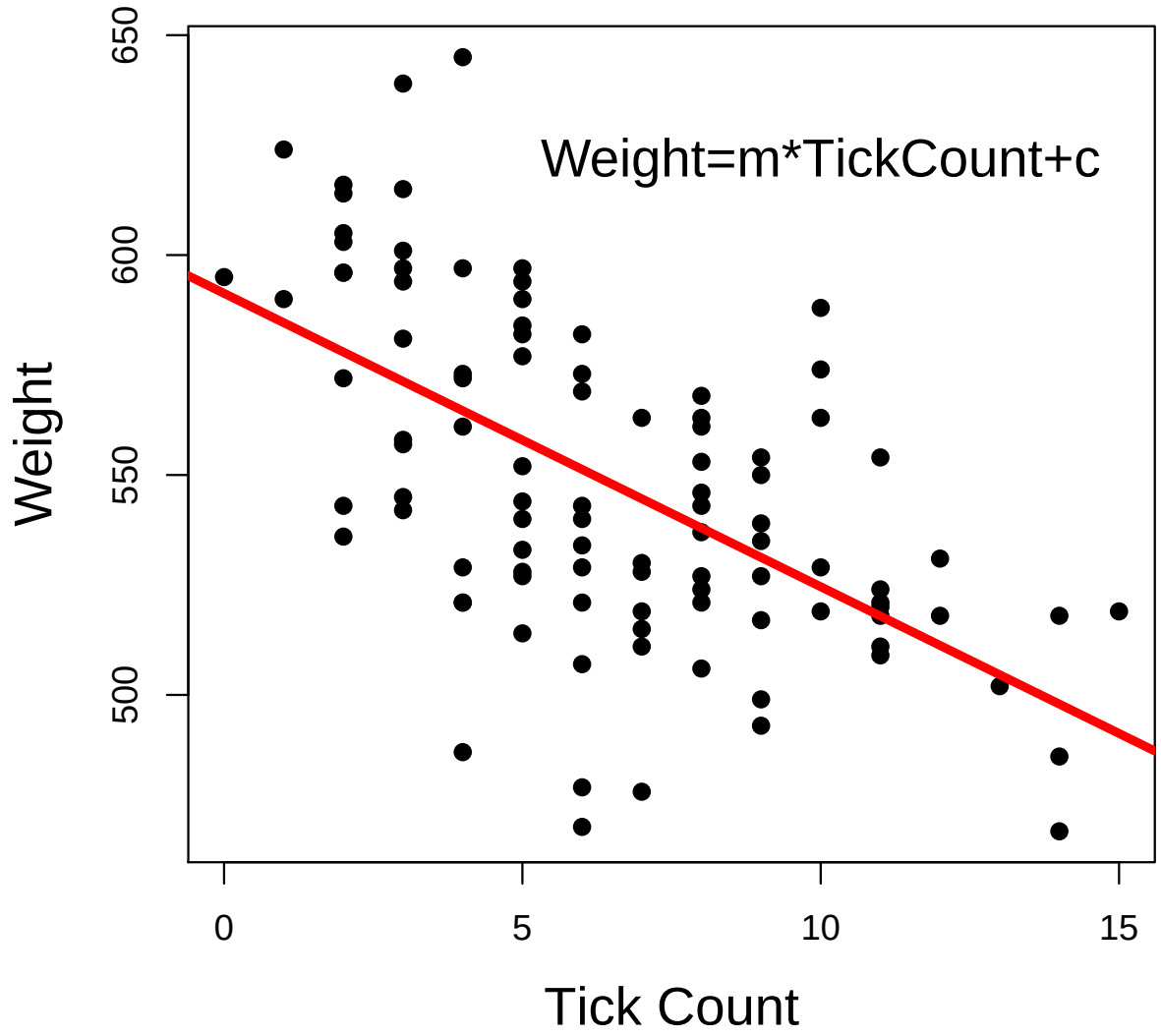
Linear regression



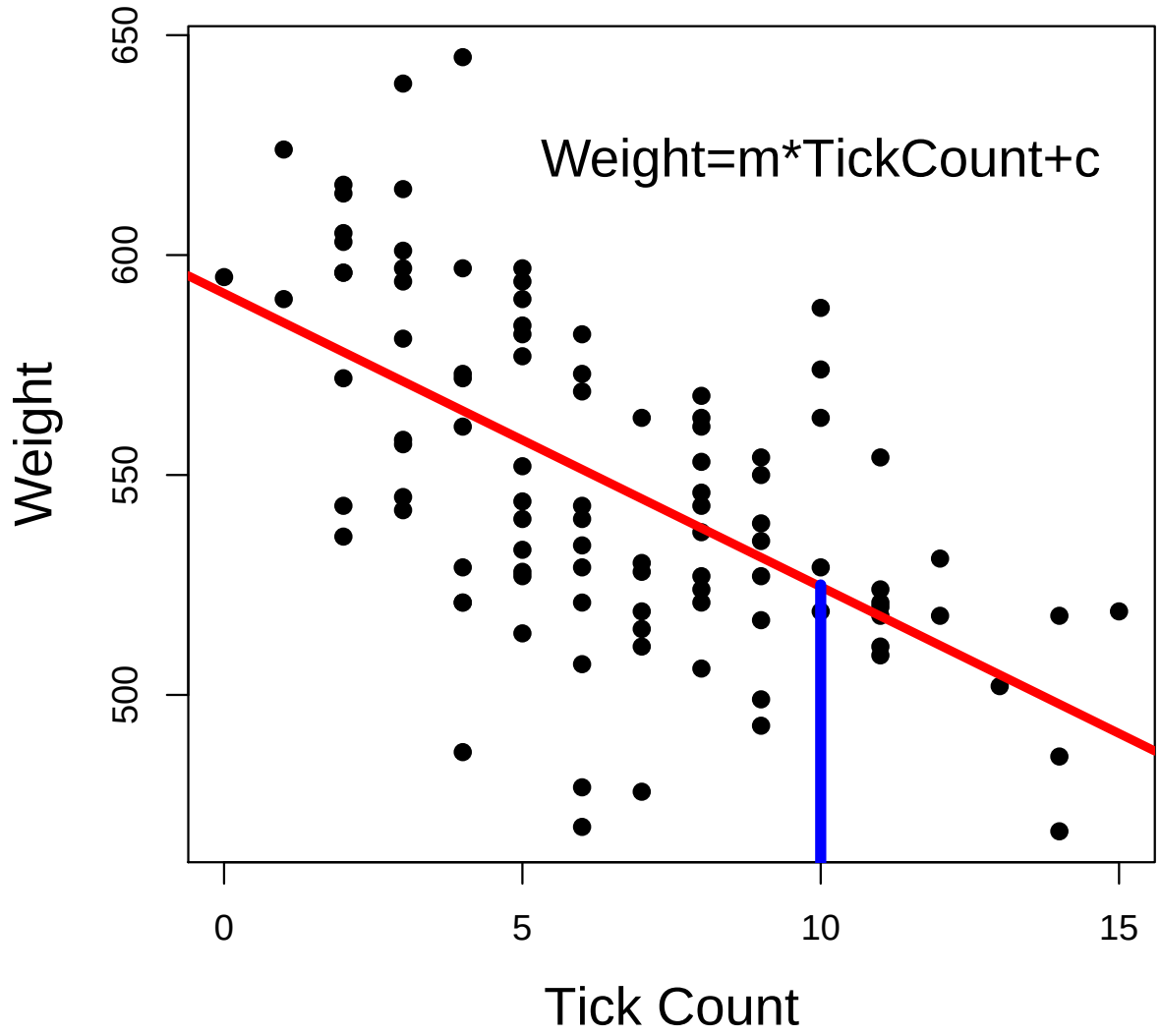
Linear regression



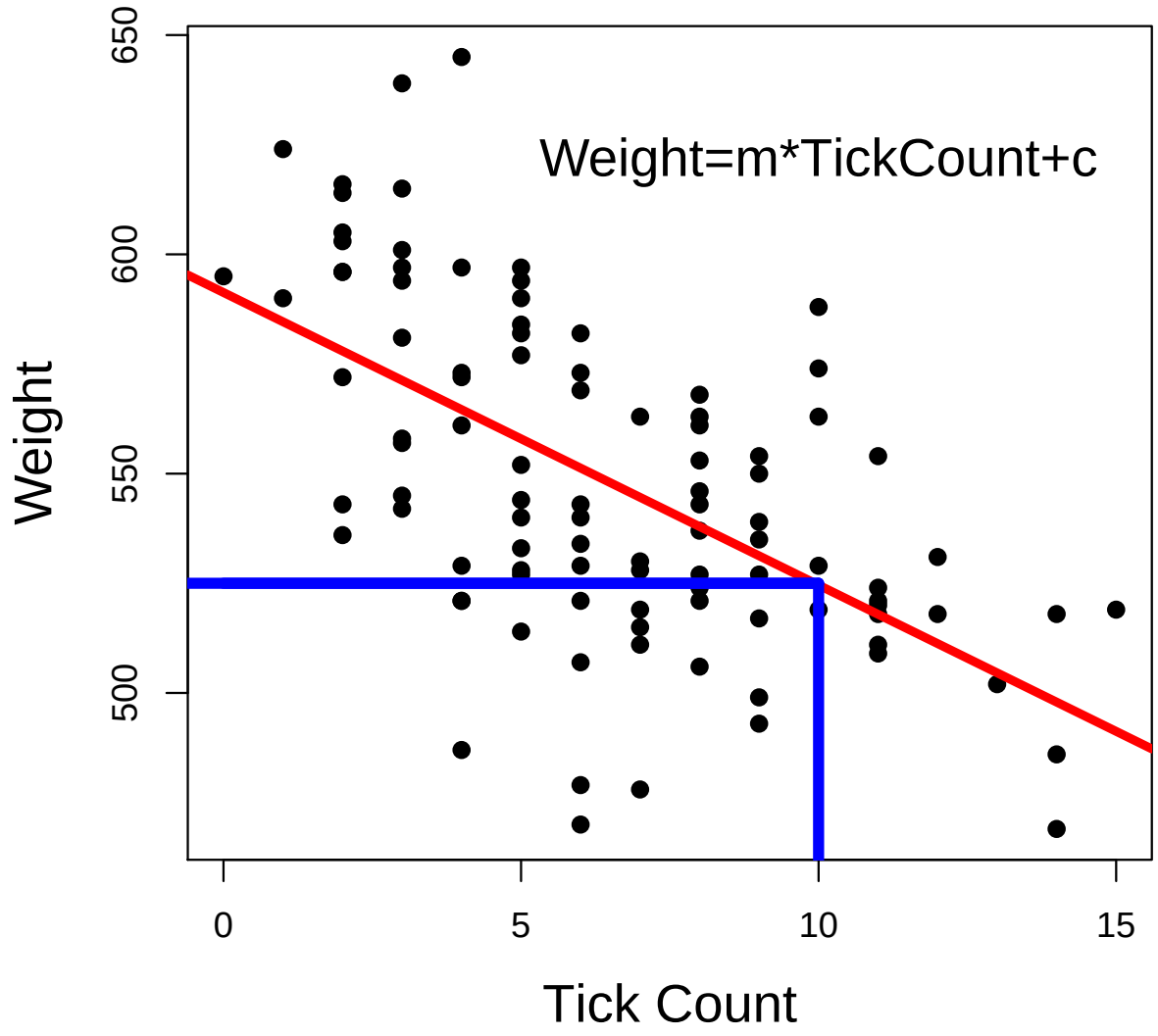
Linear regression



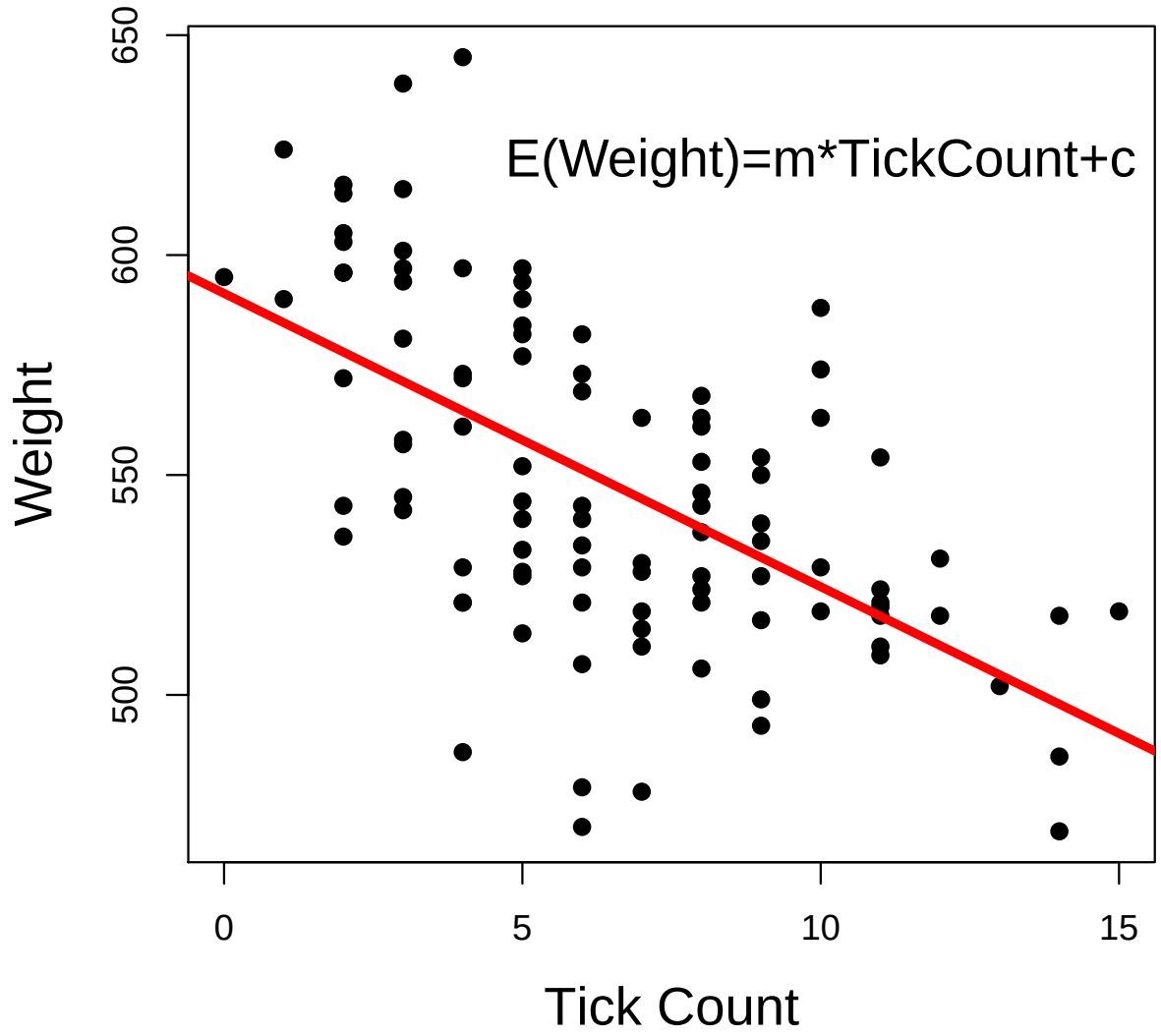
Linear regression



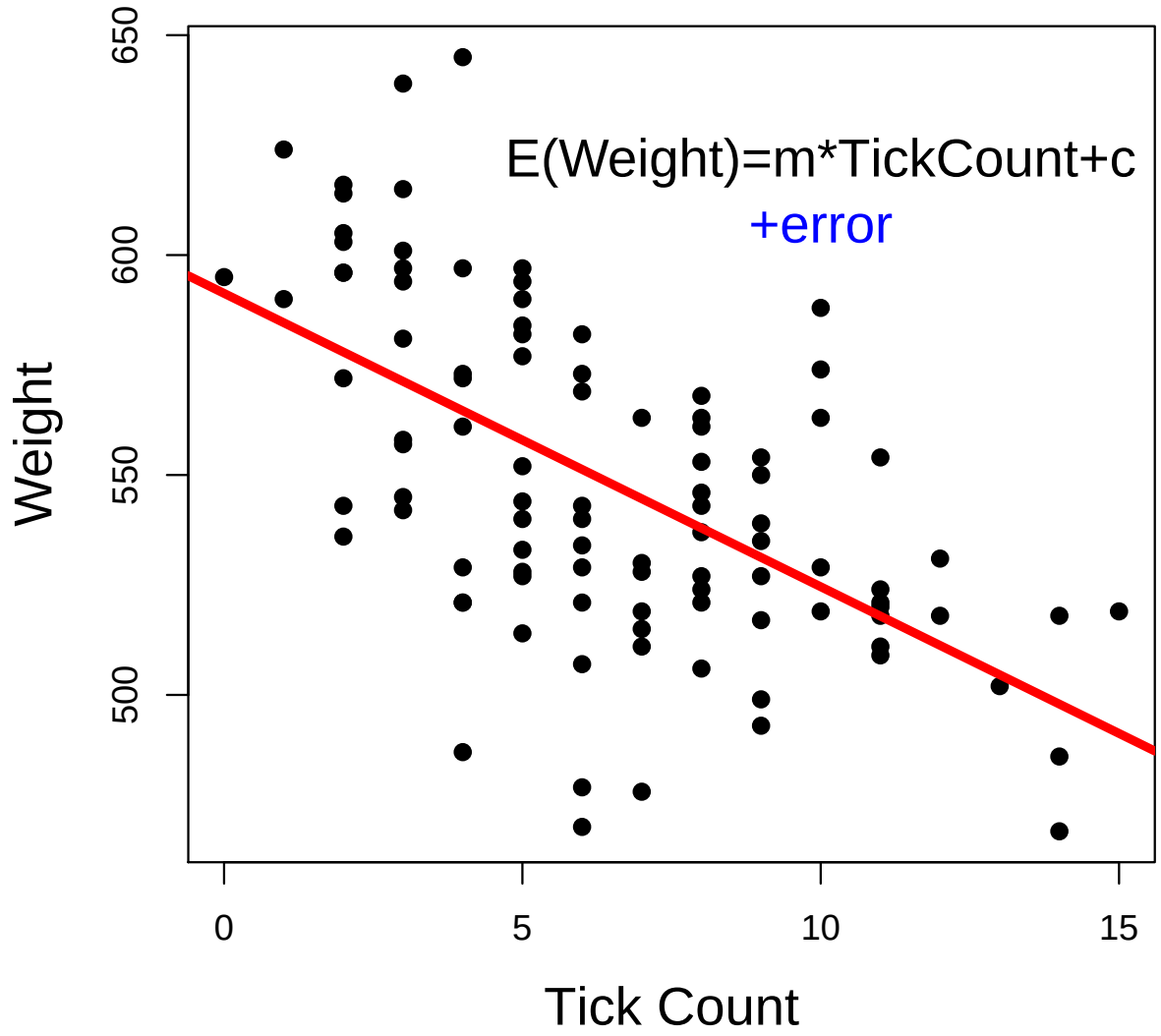
Linear regression



Linear regression



Linear regression



Linear regression

- Linear regression methods just aim to estimate the parameters m and c and say something about the additional error term.

Linear regression

- Linear regression methods just aim to estimate the parameters m and c and say something about the additional error term.
- There is a function in R which can do this.

Linear regression

- Linear regression methods just aim to estimate the parameters m and c and say something about the additional error term.
- There is a function in R which can do this.
- But most statistical test make some assumptions about data

Linear regression

- Linear regression methods just aim to estimate the parameters m and c and say something about the additional error term.
- There is a function in R which can do this.
- But most statistical test make some assumptions about data
- If we want to use a function in R, then we always need to make sure the data meet the assumptions

Linear regression

What are the assumptions of a simple linear regression?

Linear regression

What are the assumptions of a simple linear regression?

1. We have independent samples

2. The response data are normally distributed

3. The errors are normally distributed with constant variance and zero mean

Linear regression

What are the assumptions of a simple linear regression?

1. We have independent samples
   - We assume this to be true here.
   - This assumption does not hold in larger genetic studies.
2. The response data are normally distributed

3. The errors are normally distributed with constant variance and zero mean

Linear regression

What are the assumptions of a simple linear regression?

1. We have independent samples
   - We assume this to be true here.
   - This assumption does not hold in larger genetic studies.
2. The response data are normally distributed
   - In this case, this is the weight data
   - We can look at a histogram to see if this is OK
3. The errors are normally distributed with constant variance and zero mean

Linear regression

What are the assumptions of a simple linear regression?

1. We have independent samples
   - We assume this to be true here.
   - This assumption does not hold in larger genetic studies.
2. The response data are normally distributed
   - In this case, this is the weight data
   - We can look at a histogram to see if this is OK
3. The errors are normally distributed with constant variance and zero mean
   - This is something we can check from the R output

Linear regression

What are the assumptions of a simple linear regression?

1. We have independent samples
   - We assume this to be true here.
   - This assumption does not hold in larger genetic studies.
2. The response data are normally distributed
   - In this case, this is the weight data
   - We can look at a histogram to see if this is OK
3. The errors are normally distributed with constant variance and zero mean
   - This is something we can check from the R output
4. Other types of regressions make slightly different assumptions.

Linear regression

Statistical distributions are useful for a variety of reasons

Linear regression

Statistical distributions are useful for a variety of reasons

- They describe a set of observations.

Linear regression

Statistical distributions are useful for a variety of reasons

- They describe a set of observations.
- They can be used for prediction

Linear regression

Statistical distributions are useful for a variety of reasons

- They describe a set of observations.
- They can be used for prediction
- The more common and powerful statistical tools make distributional assumptions

Linear regression

Statistical distributions are useful for a variety of reasons

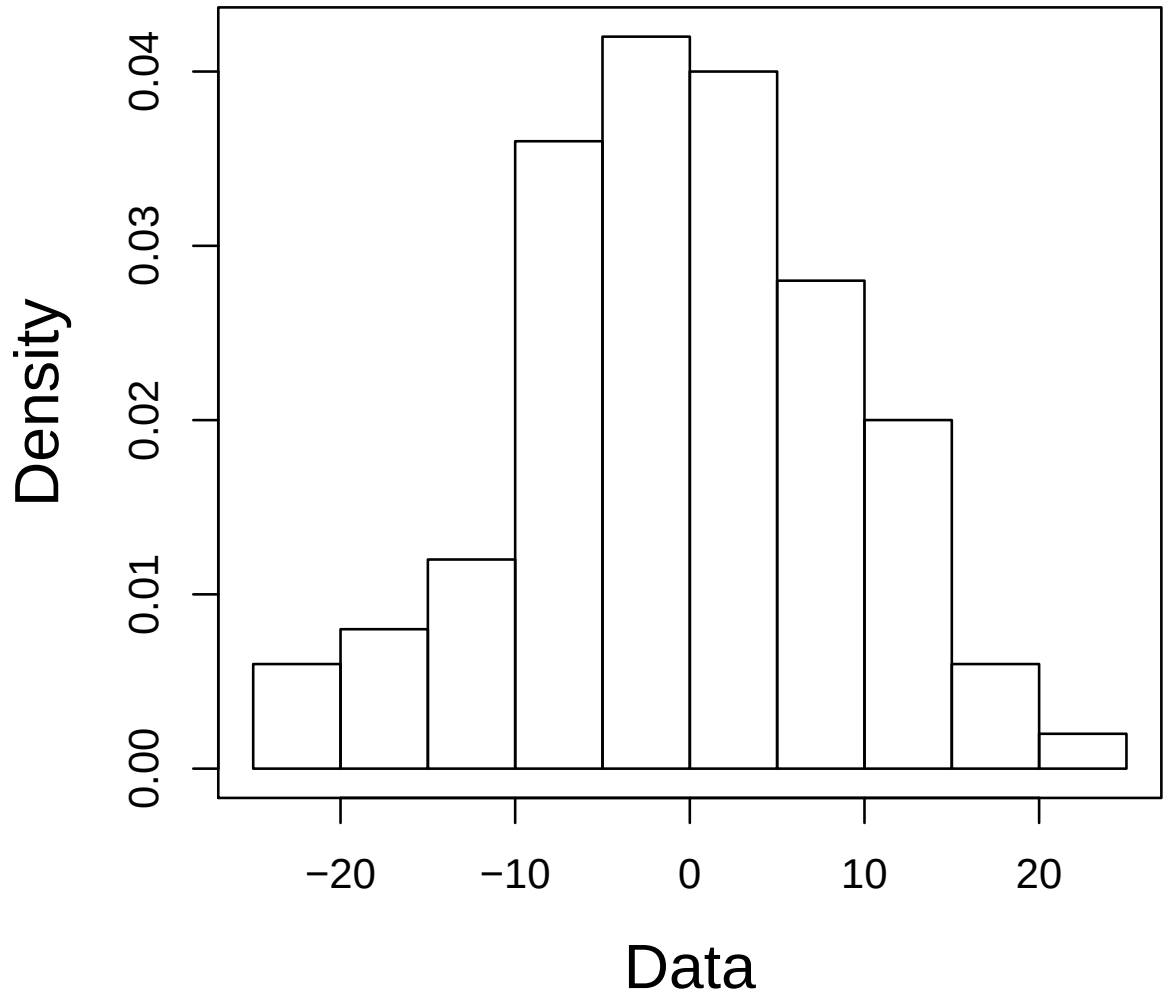
- They describe a set of observations.
- They can be used for prediction
- The more common and powerful statistical tools make distributional assumptions
- The most common distribution is the normal distribution

Linear regression

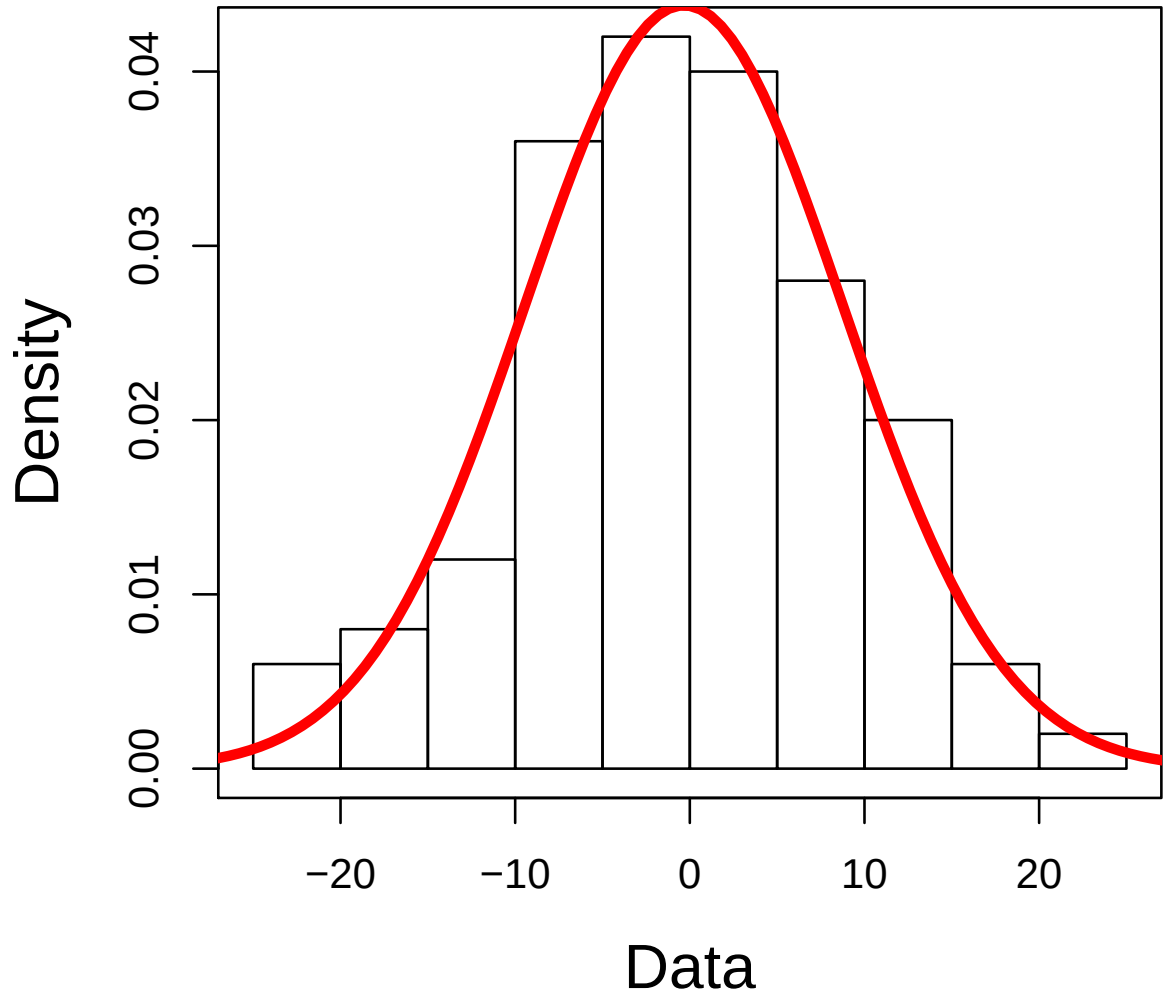
Statistical distributions are useful for a variety of reasons

- They describe a set of observations.
- They can be used for prediction
- The more common and powerful statistical tools make distributional assumptions
- The most common distribution is the normal distribution
- It assumes the data are centered around some mean value with some level or variation

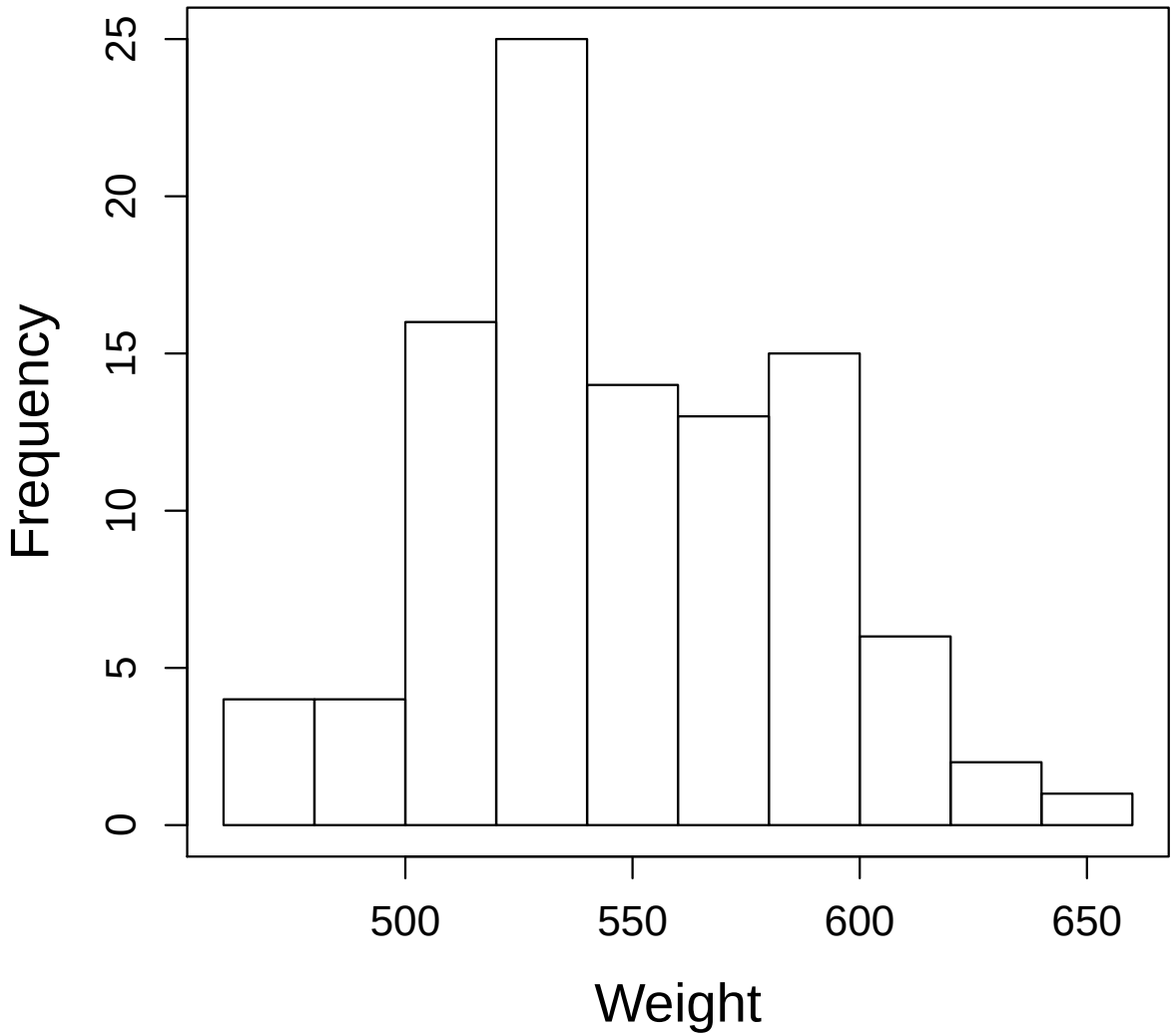
Linear regression



Linear regression



Linear regression



Linear regression

1. To do this test in R, we use the *lm* function

Linear regression

1. To do this test in R, we use the *lm* function
2. lm(weight $\sim$ tick, data=Cattle)

Call:
lm(formula = weight $\sim$ tick, data = Cattle)

Residuals:

Min	1*Q*	*Median*	3*Q*	*Max*
-81.275	-19.028	2.577	22.819	80.384

Coefficients:

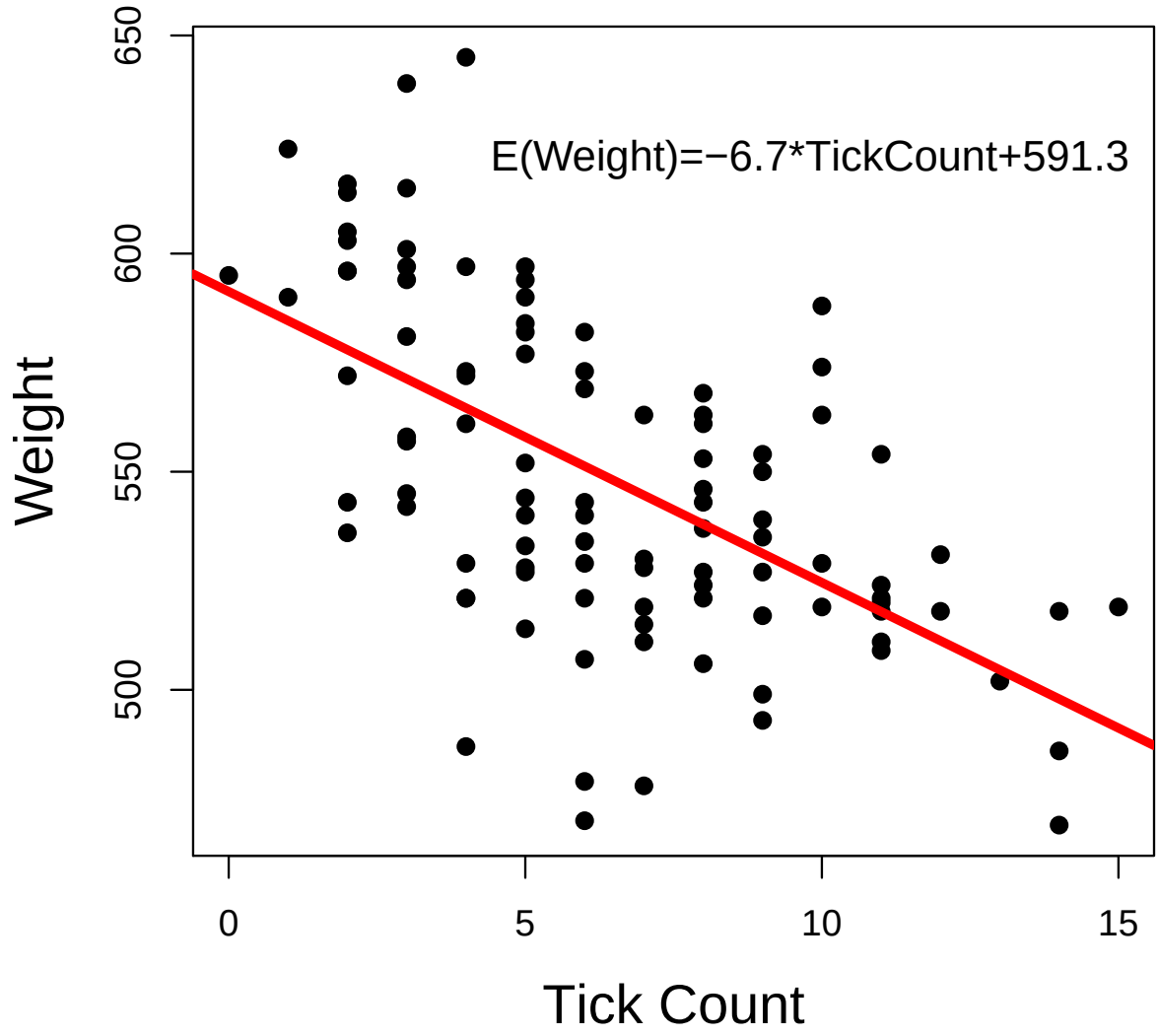
	EstimateStd.	*Error*	*tvalue*	*Pr*$(>\|t\|)$	
(*Intercept*)	**591.2975**	6.7512	87.585	$< 2e-16$	$***$
tick	**-6.6704**	0.9269	-7.196	$1.26e-10$	$***$

*Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1*

Residual standard error: 31.01 on 98 degrees of freedom Multiple R-squared: 0.3457, Adjusted R-squared:

0.3391 F-statistic: 51.79 on 1 and 98 DF, p-value: 1.258e-10

Linear regression

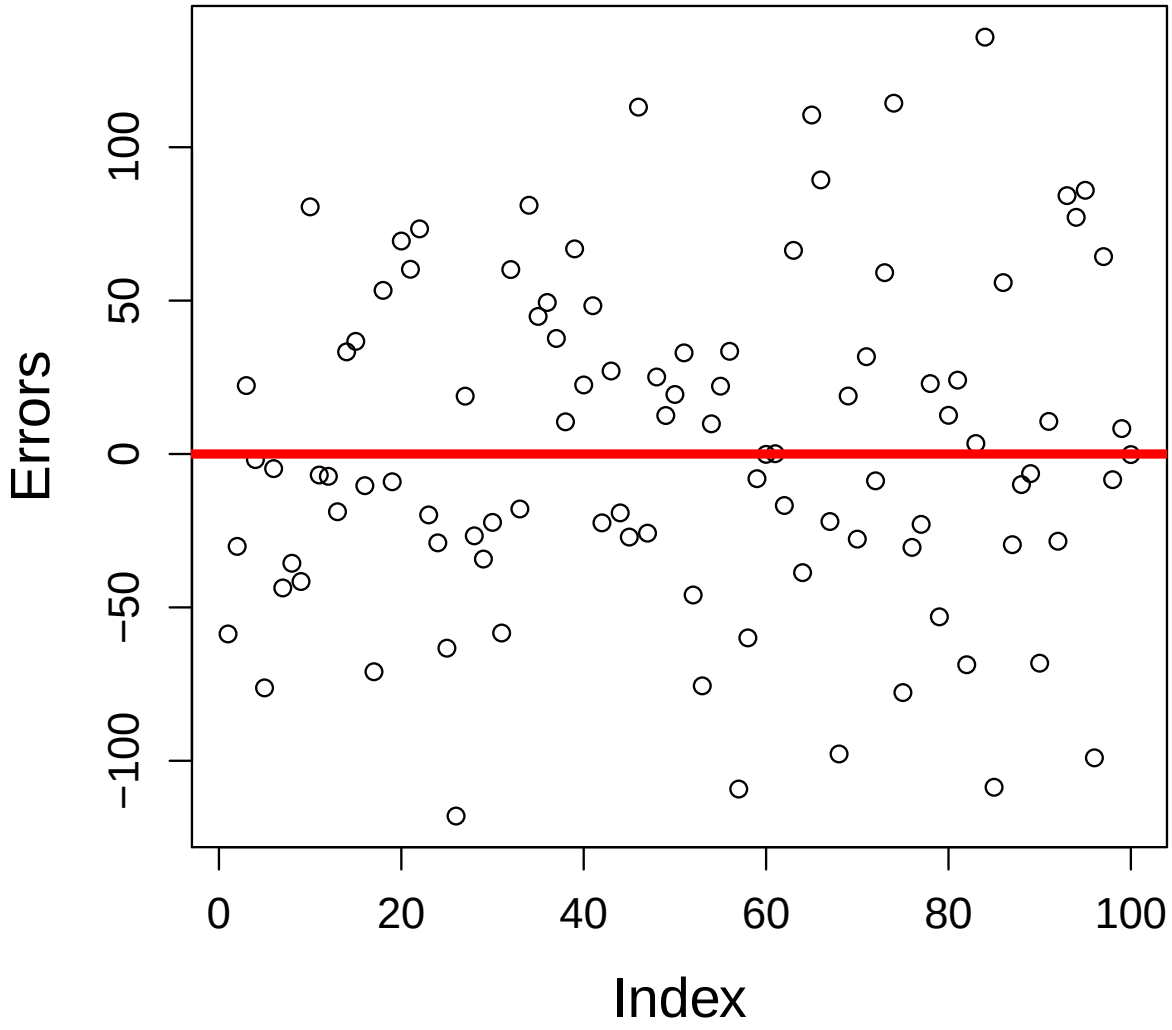


Linear regression

What are the assumptions of a simple linear regression?

1. We have independent samples
   - We assume this to be true here.
   - This assumption does not hold in larger genetic studies.
2. The response data are normally distributed
   - In this case, this is the weight data
   - We can look at a histogram to see if this is OK
3. The errors are normally distributed with constant variance and zero mean
   - This is something we can check from the R output
4. Other types of regressions make slightly different assumptions.

Question 2



Linear regression

Suppose we repeated this experiment. Then we would have different data and different estimates of the two parameters

Linear regression

1. To do this test in R, we use the *lm* function
2. lm(weight $\sim$ tick, data=Cattle)

Call:
lm(formula = weight $\sim$ tick, data = Cattle)

Residuals:

Min	1*Q*	*Median*	3*Q*	*Max*
-81.275	-19.028	2.577	22.819	80.384

Coefficients:

	EstimateStd.	*Error*	*tvalue*	*Pr*$(>\|t\|)$	
(*Intercept*)	**591.2975**	6.7512	87.585	$< 2e-16$	$***$
tick	**-6.6704**	0.9269	-7.196	$1.26e-10$	$***$

*Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1*

Residual standard error: 31.01 on 98 degrees of freedom Multiple R-squared: 0.3457, Adjusted R-squared:

0.3391 F-statistic: 51.79 on 1 and 98 DF, p-value: 1.258e-10

Linear regression

Suppose we repeated this experiment. Then we would have different data and different estimates of the two parameters

Linear regression

Suppose we repeated this experiment. Then we would have different data and different estimates of the two parameters

- Is there a range of plausible values for each of these values.

Linear regression

Suppose we repeated this experiment. Then we would have different data and different estimates of the two parameters

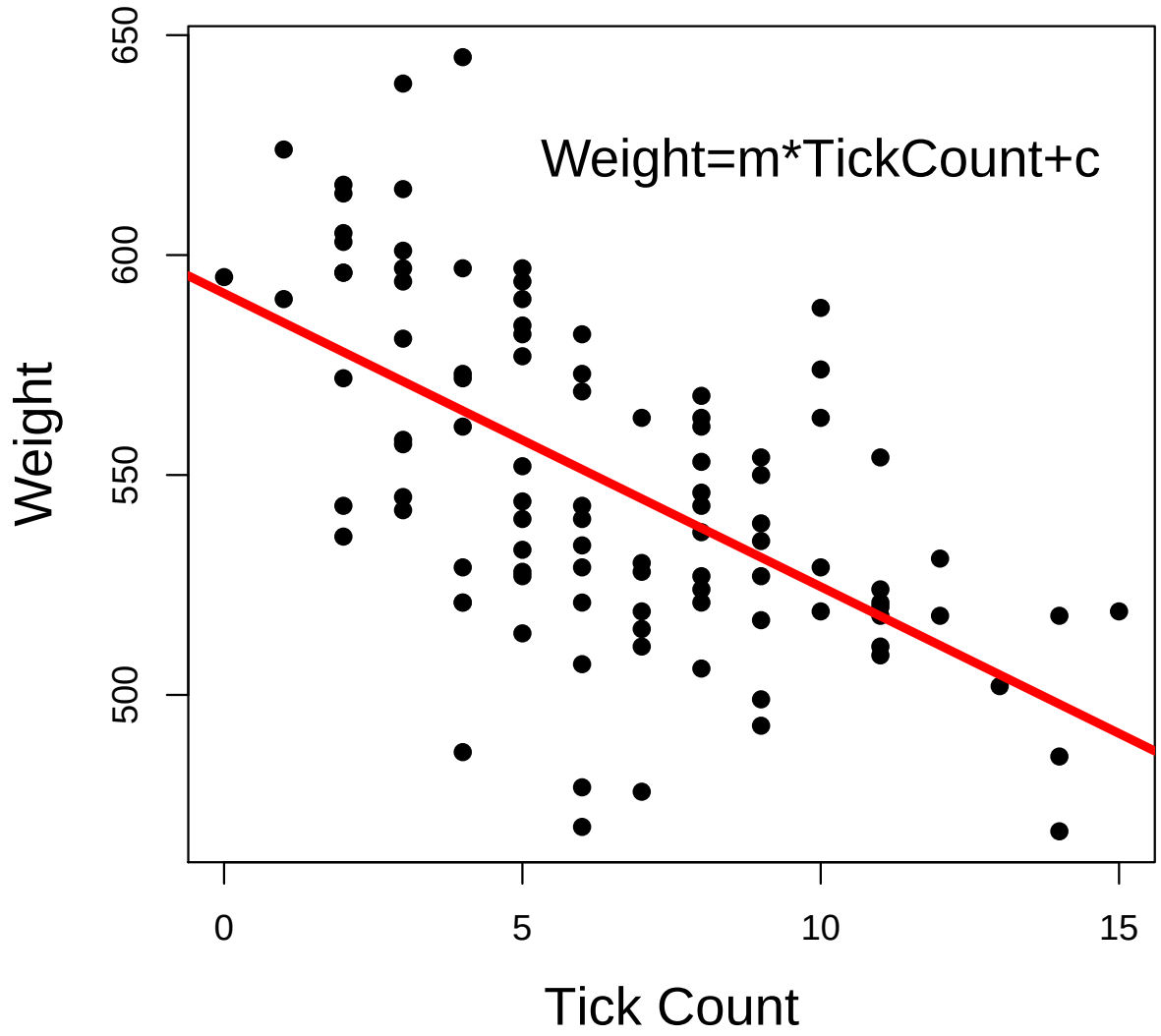
- Is there a range of plausible values for each of these values.
- We are particularly interested in whether these parameters are equal to zero

Linear regression

Suppose we repeated this experiment. Then we would have different data and different estimates of the two parameters

- Is there a range of plausible values for each of these values.
- We are particularly interested in whether these parameters are equal to zero
- If the parameter m was equal to zero, then there would be no relationship between tick count and weight

Linear regression



Linear regression

1. To do this test in R, we use the *lm* function
2. lm(weight $\sim$ tick, data=Cattle)

Call:
lm(formula = weight $\sim$ tick, data = Cattle)

Residuals:

Min	1*Q*	*Median*	3*Q*	*Max*
−81.275	−19.028	2.577	22.819	80.384

Coefficients:

	EstimateStd.	*Error*	*tvalue*	*Pr*(> |*t*|)	
(*Intercept*)	**591.2975**	6.7512	87.585	**< 2e − 16**	* * *
tick	**−6.6704**	0.9269	−7.196	**1.26e − 10**	* * *

*Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1*

Residual standard error: 31.01 on 98 degrees of freedom Multiple R-squared: 0.3457, Adjusted R-squared:

0.3391 F-statistic: 51.79 on 1 and 98 DF, p-value: 1.258e-10

Hypothesis tests

- All p-values relate to a hypothesis test.

Hypothesis tests

- All p-values relate to a hypothesis test.
- They are probabilities and so lie between zero and one.

Hypothesis tests

- All p-values relate to a hypothesis test.
- They are probabilities and so lie between zero and one.
- In this case, we are interested in whether these parameters are equal to zero.

Hypothesis tests

- All p-values relate to a hypothesis test.
- They are probabilities and so lie between zero and one.
- In this case, we are interested in whether these parameters are equal to zero.
- Therefore, these p-values are testing the hypothesis that these parameters are equal to zero.
  - If the null hypothesis is true then the p-value will be "large".
  - If the null hypothesis is not true, then the p-value will be "small".

Hypothesis tests

- All p-values relate to a hypothesis test.
- They are probabilities and so lie between zero and one.
- In this case, we are interested in whether these parameters are equal to zero.
- Therefore, these p-values are testing the hypothesis that these parameters are equal to zero.
  - If the null hypothesis is true then the p-value will be "large".
  - If the null hypothesis is not true, then the p-value will be "small".
- A standard cut off value for "small" is 0.05

Hypothesis tests

- All p-values relate to a hypothesis test.
- They are probabilities and so lie between zero and one.
- In this case, we are interested in whether these parameters are equal to zero.
- Therefore, these p-values are testing the hypothesis that these parameters are equal to zero.
  - If the null hypothesis is true then the p-value will be "large".
  - If the null hypothesis is not true, then the p-value will be "small".
- A standard cut off value for "small" is 0.05
- This is normally called the significance level of the test.

Hypothesis tests

In summary

Hypothesis tests

In summary

- If the p-value is less than 0.05 then we can reject the null hypothesis.

- If the p-value is greater than 0.05 then we can accept the null hypothesis.

Hypothesis tests

In summary

- If the p-value is less than 0.05 then we can reject the null hypothesis.
  - This means that there is a significant effect. So tick count has a significant effect on weight.
- If the p-value is greater than 0.05 then we can accept the null hypothesis.

Hypothesis tests

In summary

- If the p-value is less than 0.05 then we can reject the null hypothesis.
  - This means that there is a significant effect. So tick count has a significant effect on weight.
- If the p-value is greater than 0.05 then we can accept the null hypothesis.
  - This means that there is no significant effect. So tick count does not have a significant effect on weight.

Linear regression

1. To do this test in R, we use the *lm* function

Linear regression

1. To do this test in R, we use the *lm* function
2. lm(weight $\sim$ tick, data=Cattle)

Call:
lm(formula = weight $\sim$ tick, data = Cattle)

Residuals:

Min	1*Q*	*Median*	3*Q*	*Max*
−81.275	−19.028	2.577	22.819	80.384

Coefficients:

	EstimateStd.	*Error*	*tvalue*	*Pr*(> |*t*|)	
(*Intercept*)	**591.2975**	6.7512	87.585	**< 2e − 16**	* * *
tick	**−6.6704**	0.9269	−7.196	**1.26e − 10**	* * *

*Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1*

Residual standard error: 31.01 on 98 degrees of freedom Multiple R-squared: 0.3457, Adjusted R-squared:

0.3391 F-statistic: 51.79 on 1 and 98 DF, p-value: 1.258e-10

Linear regression

1. There are many extension to a simple linear regression

Linear regression

1. There are many extension to a simple linear regression
   - A simple linear regression has one explanatory variable

Linear regression

1. There are many extension to a simple linear regression
   - A simple linear regression has one explanatory variable
   - A multiple linear regression has multiple explanatory variables.

Linear regression

1. There are many extension to a simple linear regression
   - A simple linear regression has one explanatory variable
   - A multiple linear regression has multiple explanatory variables.
2. This model assumes that the multiple explanatory variables act additively

$$Weight = c + m_1 TickCount$$

Linear regression

1. There are many extension to a simple linear regression
   - A simple linear regression has one explanatory variable
   - A multiple linear regression has multiple explanatory variables.
2. This model assumes that the multiple explanatory variables act additively

$$Weight = c + m_1 TickCount + m_2 Breed + m_3 Housing + \ldots + error$$

ANOVA

1. The idea behind a regression model is to explain variation in a response variable given an explanatory variable
2. This is a similar idea to an ANOVA

ANOVA

1. The idea behind a regression model is to explain variation in a response variable given an explanatory variable
2. This is a similar idea to an ANOVA
3. The ANOVA is used to see if there is a difference in some response variable between different groups

ANOVA

1. The idea behind a regression model is to explain variation in a response variable given an explanatory variable
2. This is a similar idea to an ANOVA
3. The ANOVA is used to see if there is a difference in some response variable between different groups
4. It does this by assessing the amount of variation within groups compared to the amount of variation between groups

ANOVA

1. The idea behind a regression model is to explain variation in a response variable given an explanatory variable
2. This is a similar idea to an ANOVA
3. The ANOVA is used to see if there is a difference in some response variable between different groups
4. It does this by assessing the amount of variation within groups compared to the amount of variation between groups
5. There is a function in R which does anova.

ANOVA

1. ANOVA can be illustrated using the same data set.

ANOVA

1. ANOVA can be illustrated using the same data set.
2. There are three breeds of cattle: Angus, Jersey and Holstein. We want to see if breed has a significant effect on weight.

ANOVA

1. ANOVA can be illustrated using the same data set.
2. There are three breeds of cattle: Angus, Jersey and Holstein. We want to see if breed has a significant effect on weight.
3. This can be done using an ANOVA.

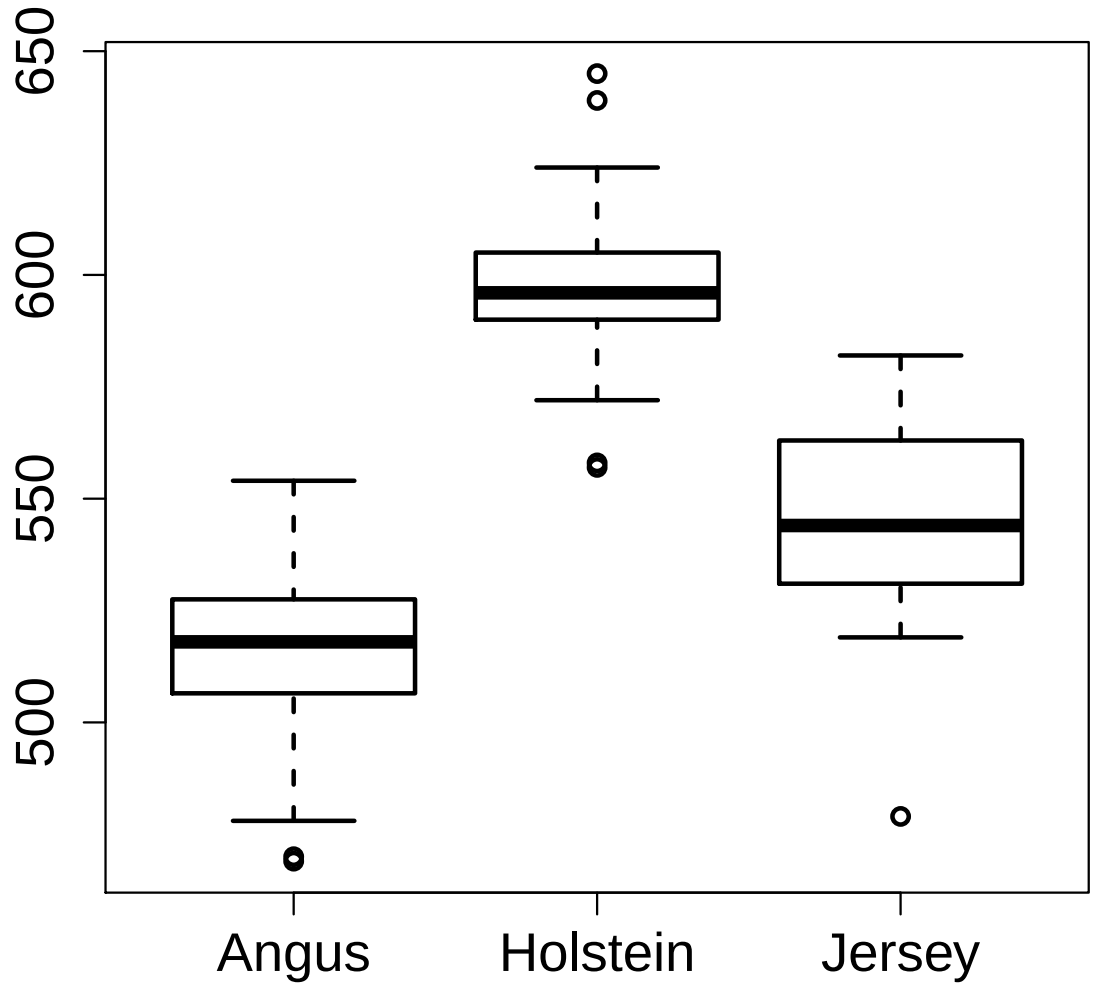
ANOVA

1. ANOVA can be illustrated using the same data set.
2. There are three breeds of cattle: Angus, Jersey and Holstein. We want to see if breed has a significant effect on weight.
3. This can be done using an ANOVA.
4. We want to see if breed can explain a significant amount of the variation in weight.

ANOVA

1. ANOVA can be illustrated using the same data set.
2. There are three breeds of cattle: Angus, Jersey and Holstein. We want to see if breed has a significant effect on weight.
3. This can be done using an ANOVA.
4. We want to see if breed can explain a significant amount of the variation in weight.
5. Again, we can look at the data graphically.

ANOVA



ANOVA

As with the linear regression, the ANOVA makes some assumptions.

- Sample are independent
- Each group is normally distributed
- The three groups have similar variances

ANOVA

As with the linear regression, the ANOVA makes some assumptions.

- Sample are independent
  - We will assume this to be true.
- Each group is normally distributed

- The three groups have similar variances

ANOVA

As with the linear regression, the ANOVA makes some assumptions.

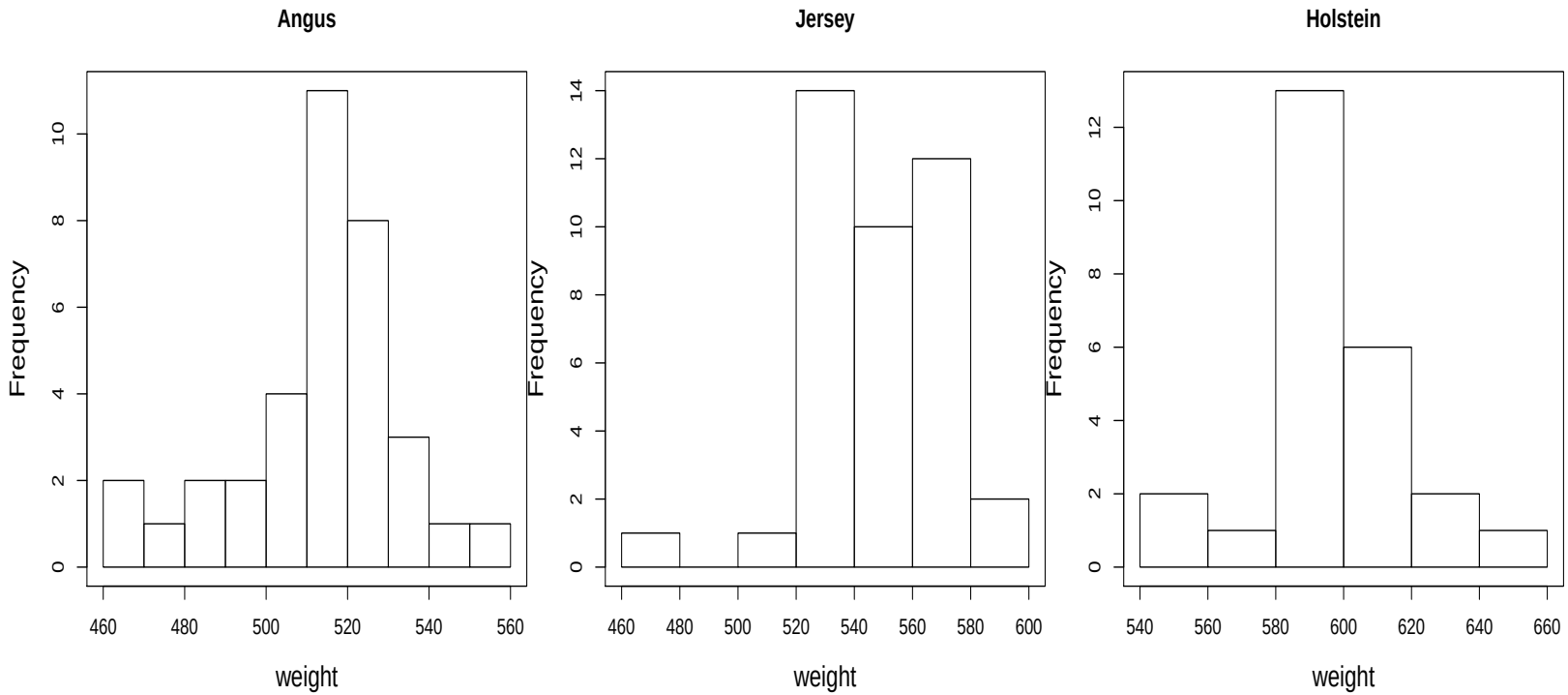
- Sample are independent
  - We will assume this to be true.
- Each group is normally distributed
  - Look at histograms of each group separately to see if they look normal
- The three groups have similar variances

ANOVA

As with the linear regression, the ANOVA makes some assumptions.

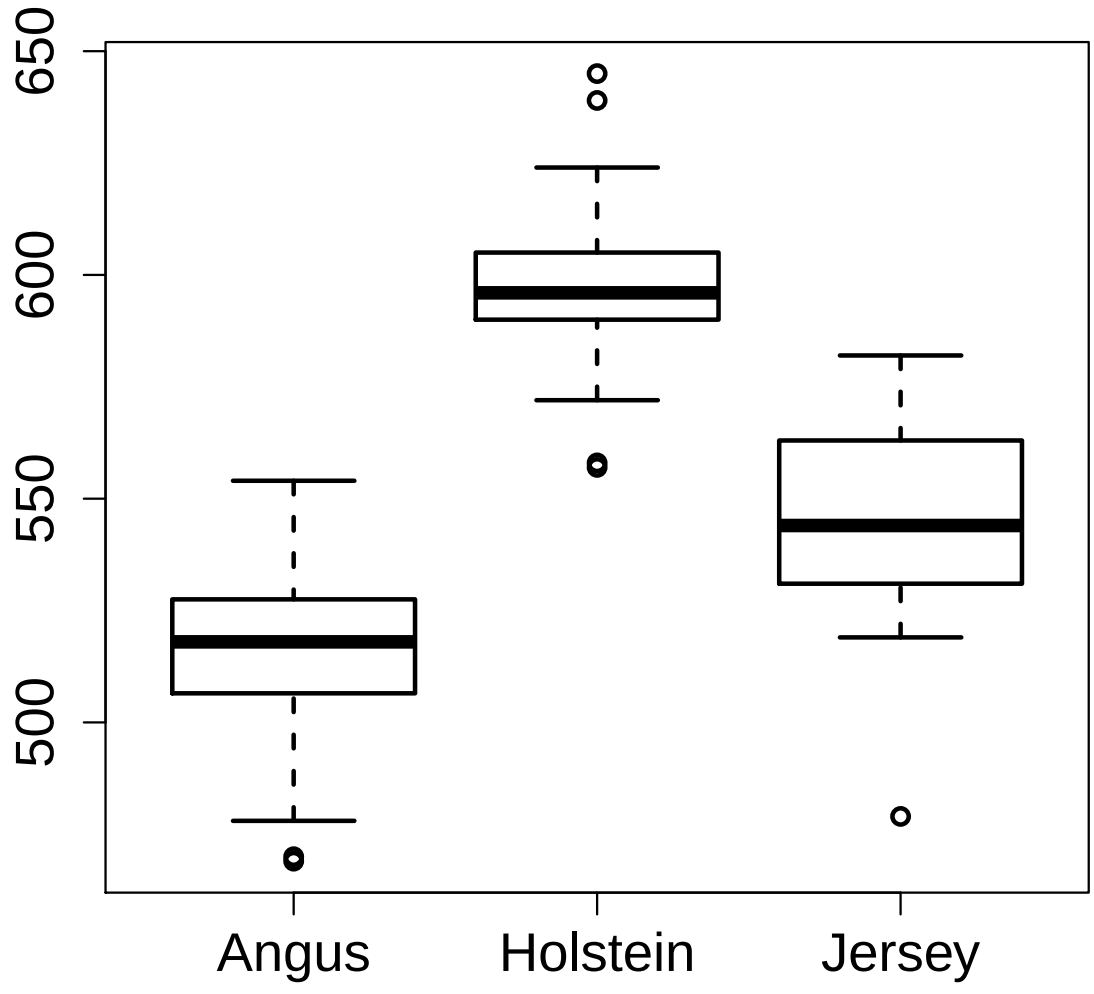
- Sample are independent
  - We will assume this to be true.
- Each group is normally distributed
  - Look at histograms of each group separately to see if they look normal
- The three groups have similar variances
  - Look at the spreads of the histograms and boxplots to see if they look the same
  - Use the *var()* function in R for each of the three groups and see if they are around the same.

ANOVA


Angus
Frequency
0
2
4
6
8
10
460
480
500
520
540
560
weight
Jersey
Frequency
0
2
4
6
8
10
12
14
460
480
500
520
540
560
580
600
weight
Holstein
Frequency
0
2
4
6
8
10
12
540
560
580
600
620
640
660
weight

(a) (b) (c)

ANOVA



ANOVA

- The ANOVA aims to test the hypothesis is

The mean weight from Angus = the mean weight from Holstein = mean weight from Jersey

ANOVA

- The ANOVA aims to test the hypothesis is

The mean weight from Angus = the mean weight from Holstein = mean weight from Jersey

- If we reject this hypothesis, then we only know that there is at least one difference between the three breeds and not where the differences lie.

ANOVA

- The ANOVA aims to test the hypothesis is

The mean weight from Angus = the mean weight from Holstein = mean weight from Jersey

- If we reject this hypothesis, then we only know that there is at least one difference between the three breeds and not where the differences lie.
- We can do this in R using the *aov* function

ANOVA

$aov(weight \sim breed)$

Terms:

	Df	*SumSq*	*MeanSq*	$F - value$	$Pr(>F)$	
breed	2	101768	50884	116.9	$< 2e-16$	$***$
Residuals	97	42227	435			

ANOVA

$aov(weight \sim breed)$

Terms:

	Df	*SumSq*	*MeanSq*	$F - value$	$Pr(>F)$	
breed	2	101768	50884	116.9	$< \mathbf{2e-16}$	$***$
Residuals	97	42227	435			

Example

Suppose with have height measurements from 1000 people and we have data from one locus with two alleles A and B. We want to test if this locus is associated with height.

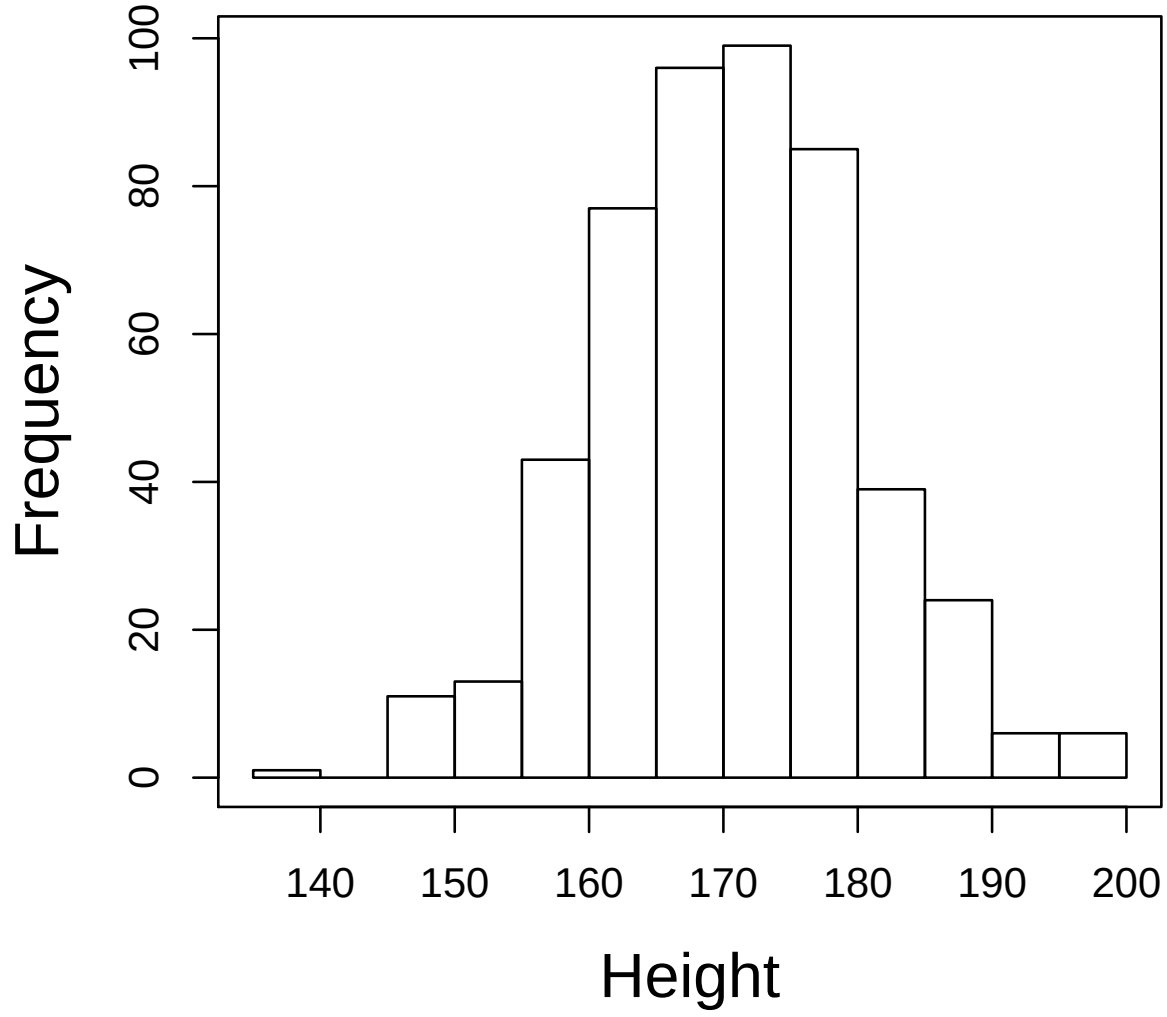
ID	Height	Genotype
1	150	AB
2	161	AA
3	145	BB
4	189	AA
5	141	BA
⋮	⋮	⋮

Example

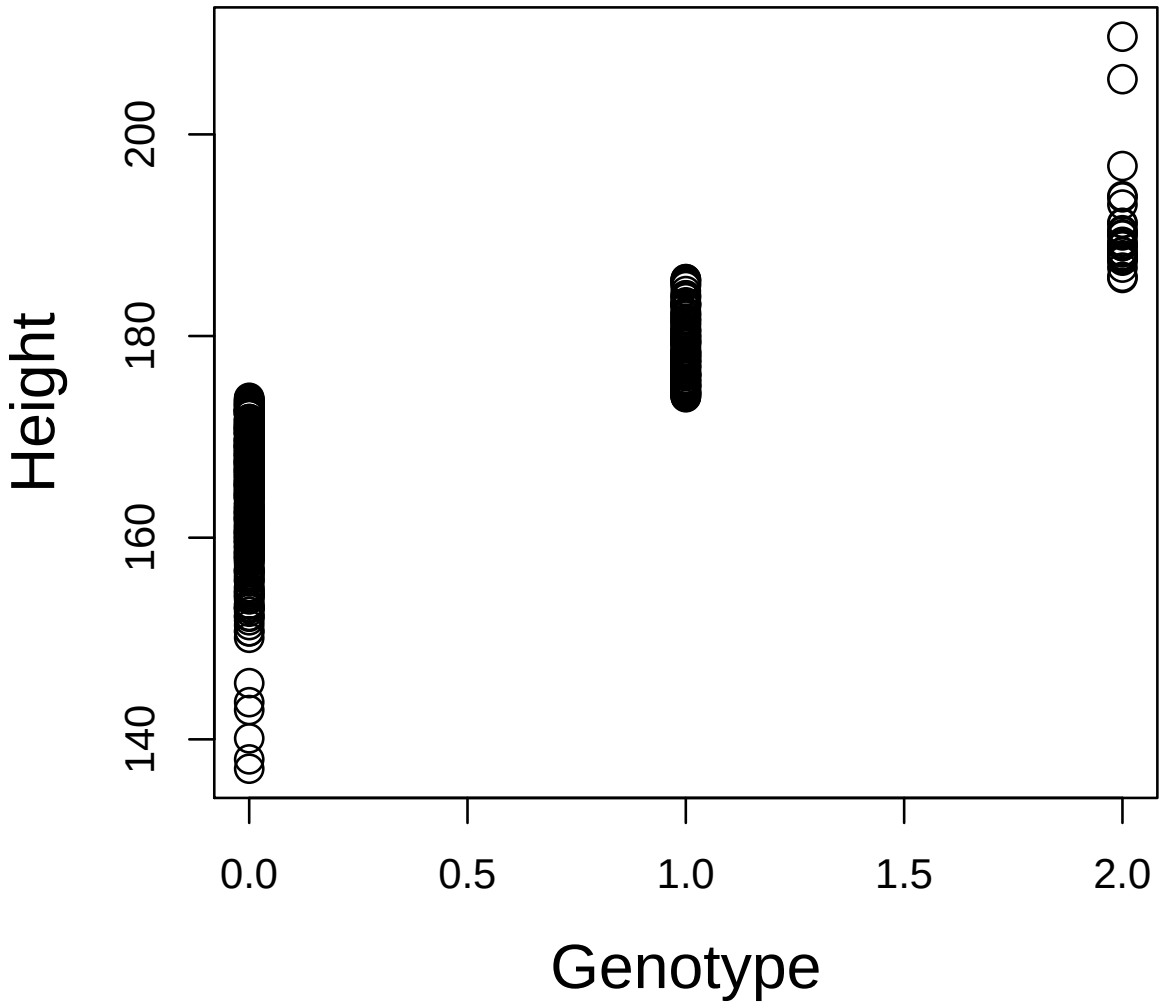
Suppose with have height measurements from 1000 people and we have data from one locus with two alleles A and B. We want to test if this locus is associated with height.

ID	Height	Genotype
1	150	1
2	161	0
3	145	2
4	189	0
5	141	1
⋮	⋮	⋮

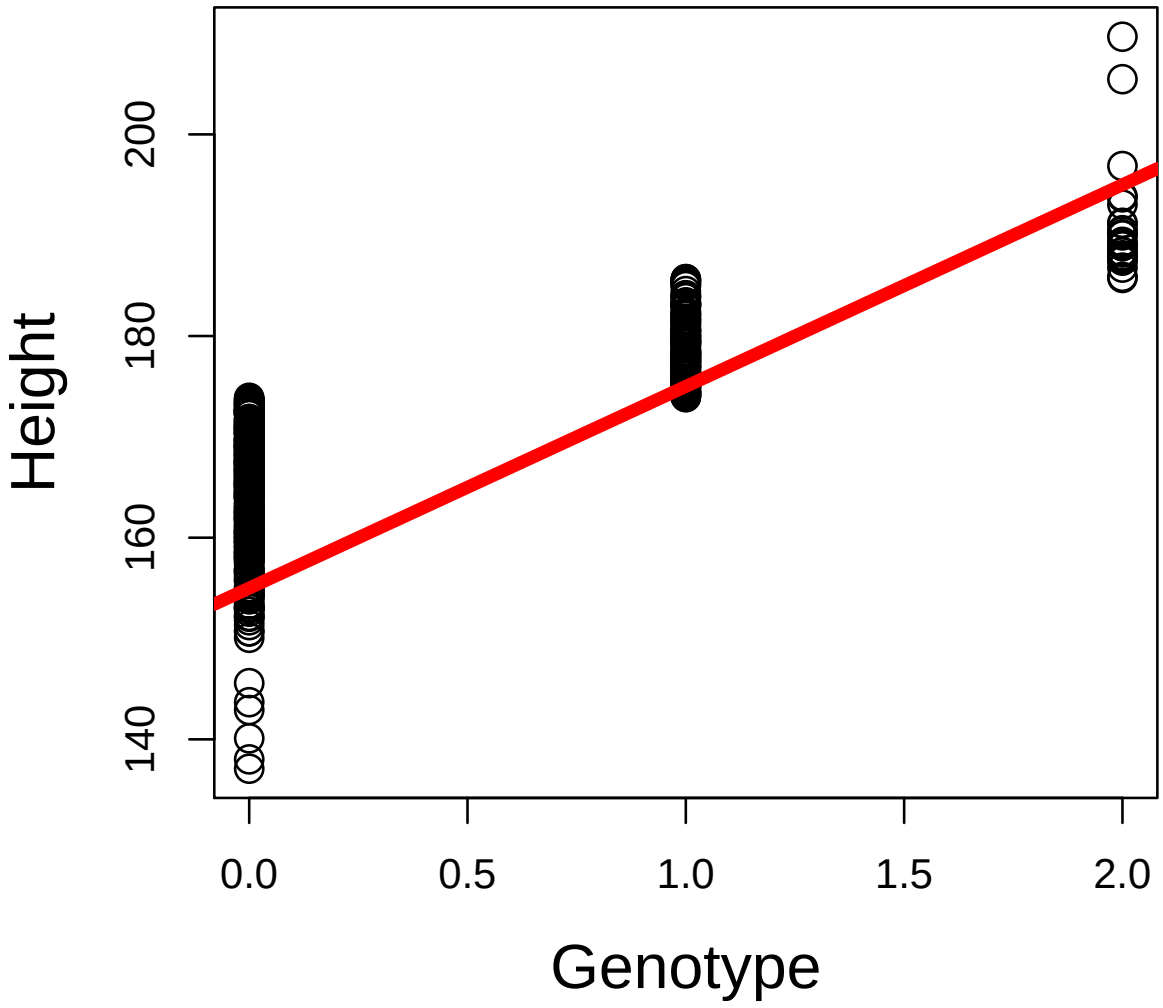
Example



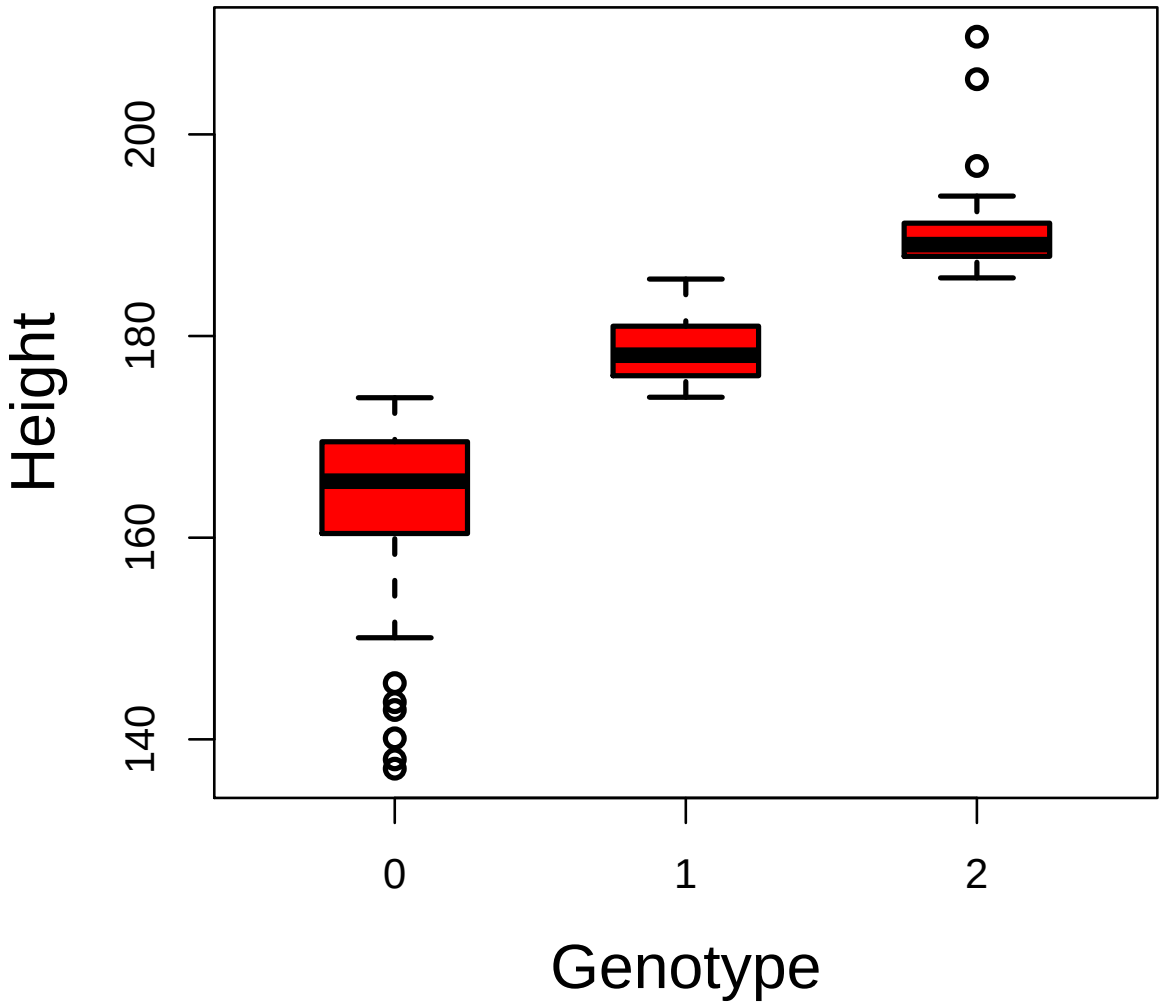
Example



Example - Linear regression



Example - ANOVA



Example - Linear regression

Call:
lm(formula = height ~ genotype)

Residuals:

Min	*1Q*	*Median*	*3Q*	*Max*
−27.4684	−3.3333	0.3393	4.4179	17.8228

Coefficients:

	Estimate	*Std.Error*	*tvalue*	*Pr(>|t|)*	
(*Intercept*)	164.5621	0.3090	532.5	$< 2e-16$	***
genotype	13.6526	0.4433	30.8	$< 2e-16$	***

*Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1*

Residual standard error: 5.831 on 498 degrees of freedom Multiple R-squared: 0.6558, Adjusted R-squared: 0.6551 F-statistic: 948.6 on 1 and 498 DF, p-value: ¡ 2.2e-16

Example - Linear regression

Call:
lm(formula = height ~ genotype)

Residuals:

Min	*1Q*	*Median*	*3Q*	*Max*
−27.4684	−3.3333	0.3393	4.4179	17.8228

Coefficients:

	Estimate	*Std.Error*	*tvalue*	*Pr(>|t|)*	
(*Intercept*)	164.5621	0.3090	532.5	$< 2e-16$	***
genotype	13.6526	0.4433	30.8	$< 2e-16$	***

*Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1*

Residual standard error: 5.831 on 498 degrees of freedom Multiple R-squared: 0.6558, Adjusted R-squared: 0.6551 F-statistic: 948.6 on 1 and 498 DF, p-value: ¡ 2.2e-16

Example - ANOVA

$aov(height \sim genotype)$

Terms:

	Df	*SumSq*	*MeanSq*	*Fvalue*	$Pr(>F)$	
genotype	1	32257	32257	948.6	$< 2e-16$	$***$
Residuals	498	16934	34			

*Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1*

Example - ANOVA

$aov(height \sim genotype)$

Terms:

	Df	*SumSq*	*MeanSq*	*Fvalue*	$Pr(>F)$	
genotype	1	32257	32257	948.6	$< 2e-16$	$***$
Residuals	498	16934	34			

*Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1*