

ANOVA

Analysis of Variance

Stats and Genetics/Evolution

The Correlation between relatives on the supposition of Mendelian Inheritance

By R. A. FISHER, B.A.

Communicated by Professor J. ARTHUR THOMSON

With Four Figures in Text

(MS. received 15 June 1918. Read 8 July 1918. Issued separately 1 October 1918)

ON THE "PROBABLE ERROR" OF A COEFFICIENT OF CORRELATION DEDUCED FROM A SMALL SAMPLE

Fisher 1921

Author's Note (CMS 1.2a)

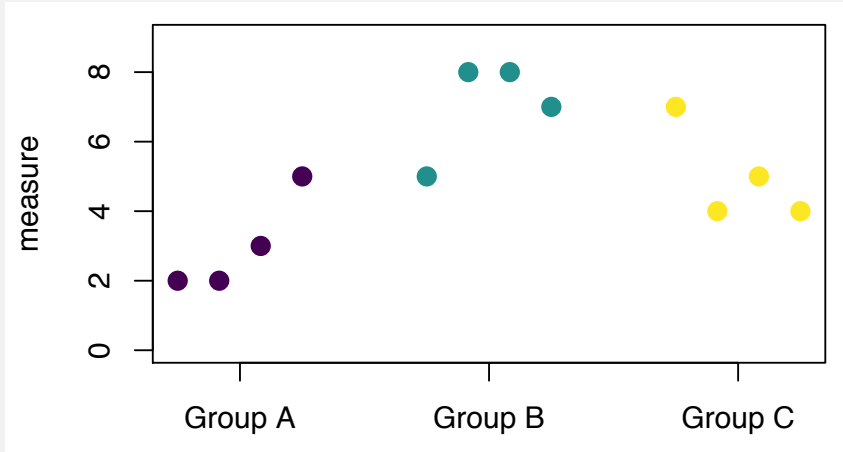
This is the second of three papers dealing with the sampling errors of correlation coefficients covering the cases (i) "The frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population," *Biometrika*, Vol. 10, pp. 507–521, 1915.

Analysis of Variance

- Used to compare the means among more than two groups
- If you are comparing three groups, for instance, you cannot just do three pair-wise t -tests – this approach would cause too many false positives
- ANOVA takes into account the fact that you are comparing multiple groups and controls the false positive rate.

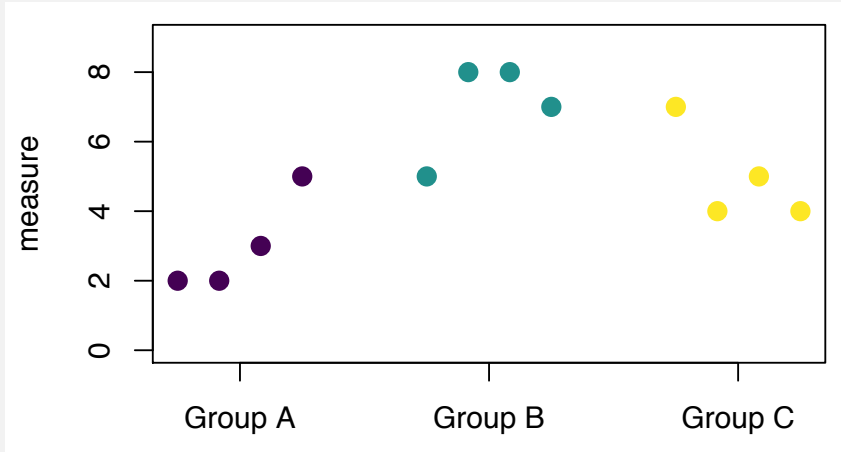
Analysis of Variance

A	B	C
2	5	7
2	8	4
3	8	5
5	7	4



Analysis of Variance

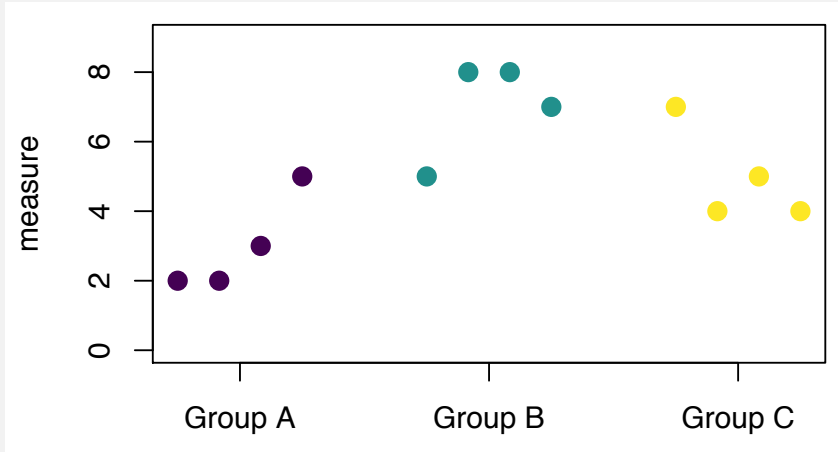
A	B	C
2	5	7
2	8	4
3	8	5
5	7	4



$$f \text{ statistic} = \frac{\frac{\sum_{j=1}^p n_j (\bar{x}_j - \bar{x})^2}{df_{ssb}}}{\frac{\sum_{j=1}^p \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{df_{ssw}}}$$

Analysis of Variance

A	B	C
2	5	7
2	8	4
3	8	5
5	7	4



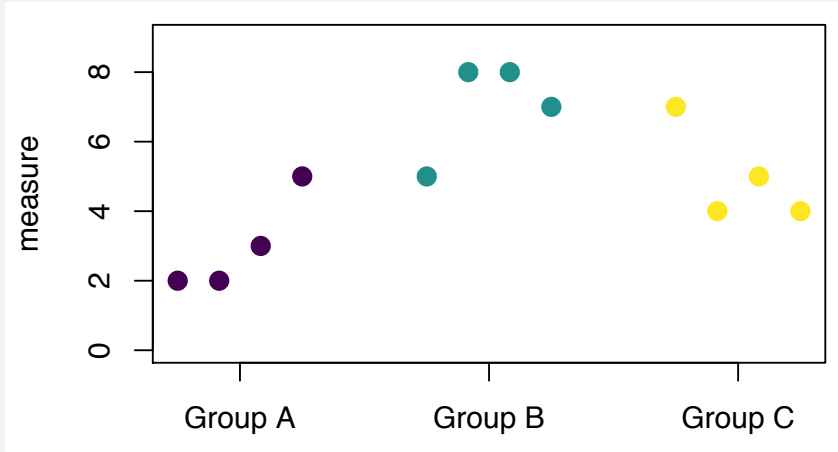
$$f \text{ statistic} = \frac{\frac{\sum_{j=1}^p n_j (\bar{x}_j - \bar{x})^2}{df_{ssb}}}{\frac{\sum_{j=1}^p \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{df_{ssw}}}$$

number of groups - 1

number of sample – number of groups

Analysis of Variance

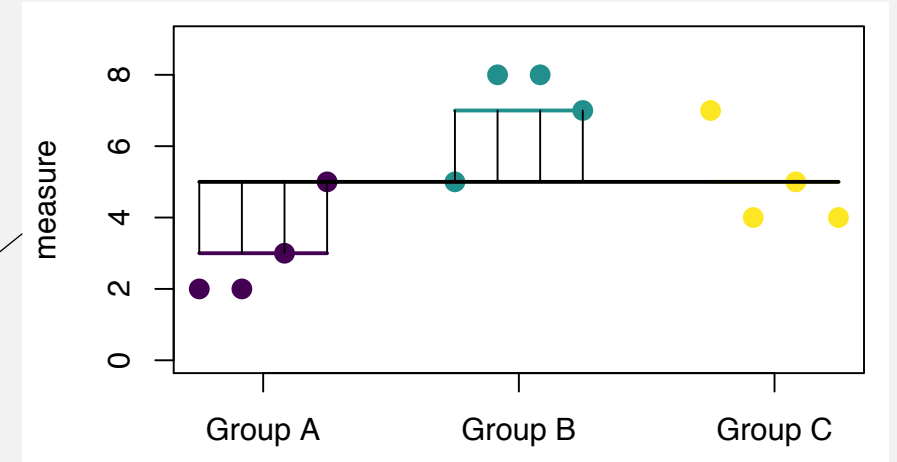
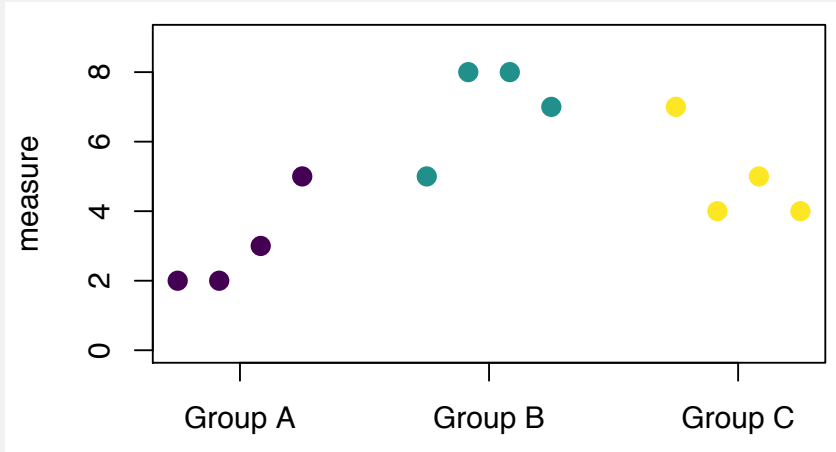
A	B	C
2	5	7
2	8	4
3	8	5
5	7	4



$$f \text{ statistic} = \frac{\frac{\sum_{j=1}^p n_j (\bar{x}_j - \bar{x})^2}{df_{ssb}}}{\frac{\sum_{j=1}^p \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{df_{ssw}}}$$

Analysis of Variance

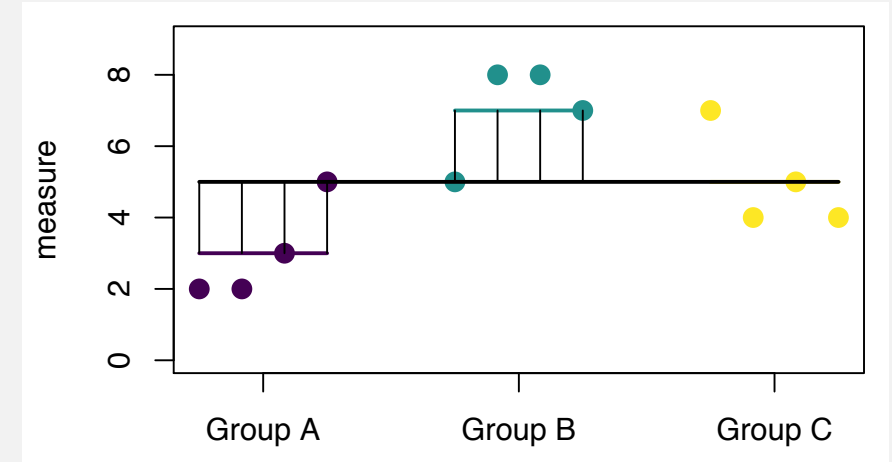
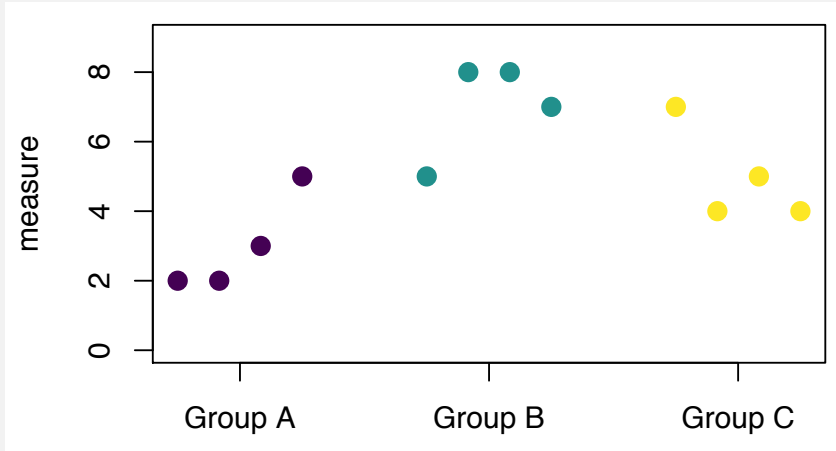
A	B	C
2	5	7
2	8	4
3	8	5
5	7	4



$$f \text{ statistic} = \frac{\frac{\sum_{j=1}^p n_j (\bar{x}_j - \bar{x})^2}{df_{ssb}}}{\frac{\sum_{j=1}^p \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{df_{ssw}}}$$

Analysis of Variance

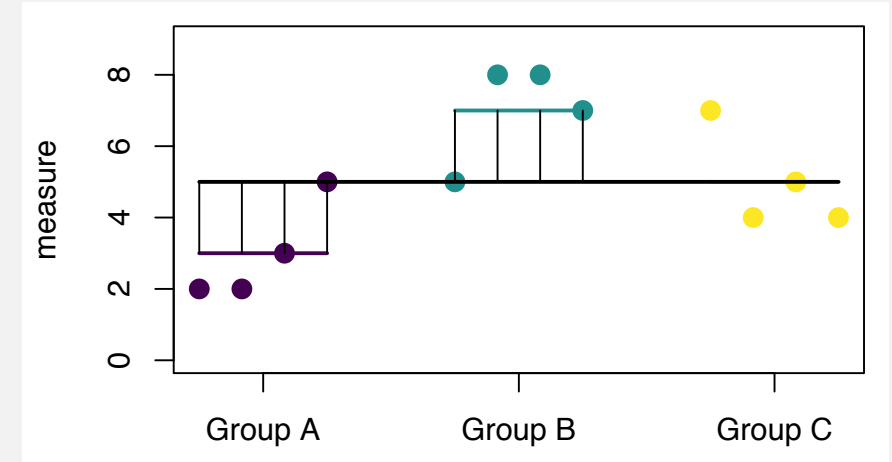
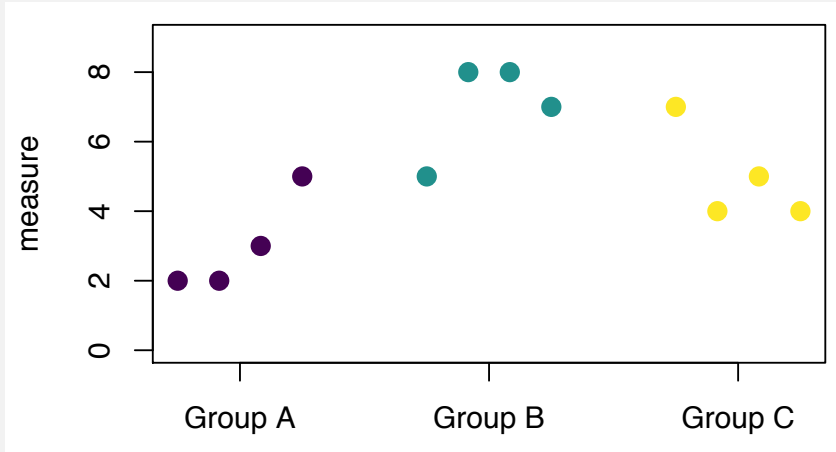
A	B	C
2	5	7
2	8	4
3	8	5
5	7	4



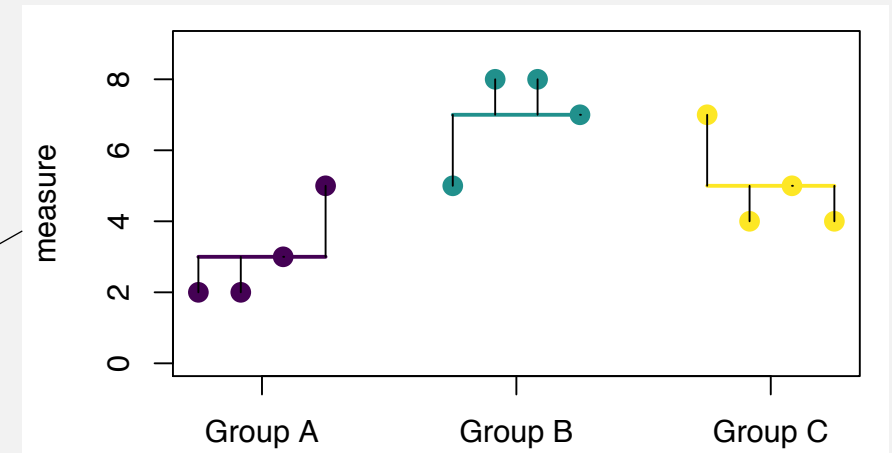
$$f \text{ statistic} = \frac{\frac{\sum_{j=1}^p n_j (\bar{x}_j - \bar{x})^2}{df_{ssb}}}{\frac{\sum_{j=1}^p \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{df_{ssw}}}$$

Analysis of Variance

A	B	C
2	5	7
2	8	4
3	8	5
5	7	4



$$f \text{ statistic} = \frac{\sum_{j=1}^p n_j (\bar{x}_j - \bar{x})^2}{df_{ssb}} \div \frac{\sum_{j=1}^p \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{df_{ssw}}$$



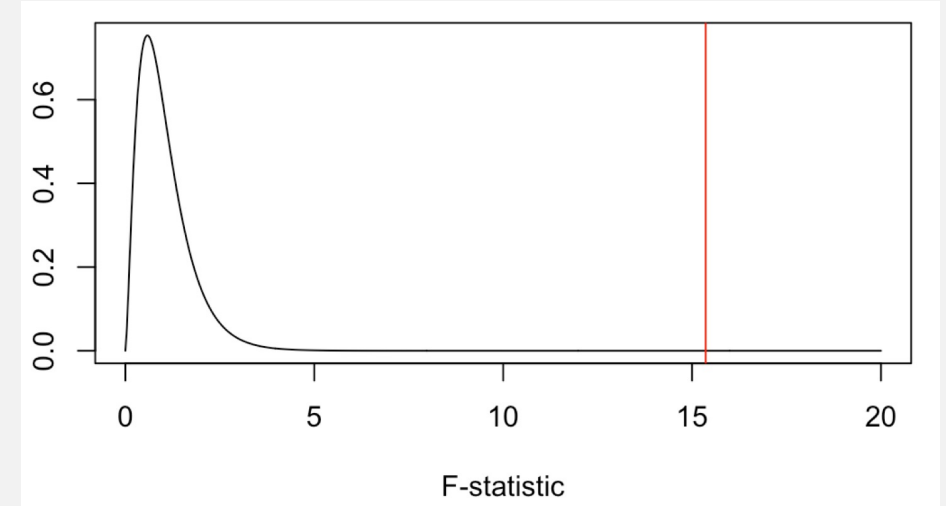
Running ANOVA in R

```
> data("chickwts")
> fit <- aov(weight~feed, data=chickwts)
> summary(fit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
feed	5	231129	46226	15.37	5.94e-10 ***
Residuals	65	195556	3009		

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

This significant result tells us that at least one of the groups of chickens have significantly different mean weights than at least one other groups. (significant ANOVA result allows us to reject the null that they are all the same)

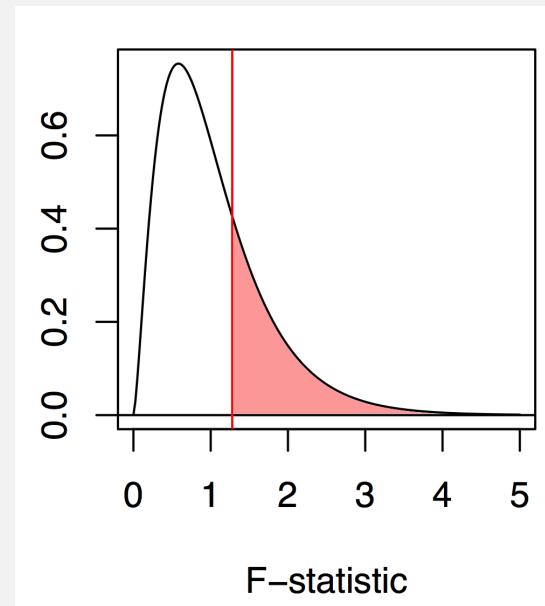
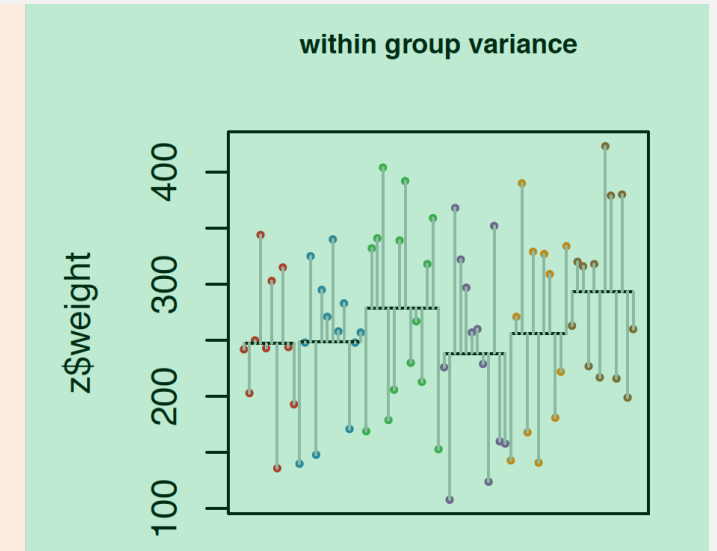
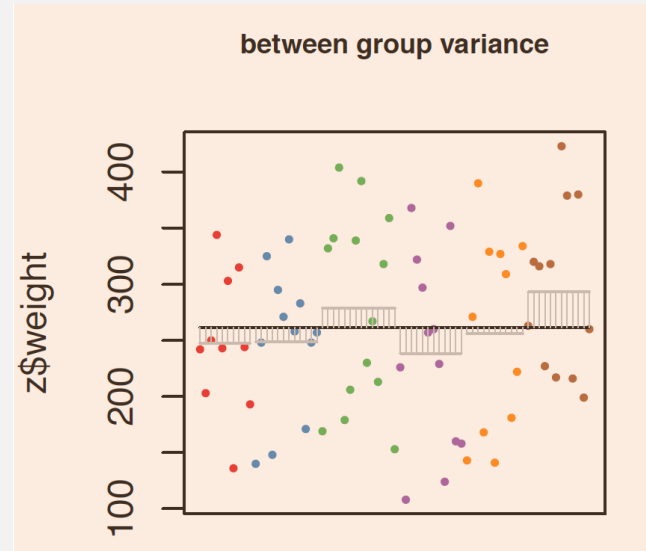


$$f \text{ statistic} = \frac{\sum_{j=1}^p n_j (\bar{x}_j - \bar{x})^2}{df_{ssb}} \div \frac{\sum_{j=1}^p \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}{df_{ssw}}$$

Running ANOVA in R

```
> data("chickwts")
> z <- chickwts
> z$weight <- sample(z$weight)
> fit <- aov(weight~feed, data=z)
> summary(fit)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
feed	5	32244	6449	1.063	0.389
Residuals	65	394441	6068		



Go To R

Post-hoc tests

If your ANOVA is significant, you may be interested in discovering which groups are different from one another

A variety of post-hoc comparisons of the means can be used

Fisher's LSD

- Least conservative test, basically uses t -tests to compare the means

Scheffe's method

- Performs all comparisons simultaneously, but has relatively low power

Tukey-Kramer method

- A pair-wise method, like a t -test, but corrected for multiple comparisons

Post-hoc tests

```
> data("chickwts")
> fit <- aov(weight~feed, data=chickwts)
> TukeyHSD(fit)
```

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = weight ~ feed, data = chickwts)

\$feed

	diff	lwr	upr	p adj
horsebean-casein	-163.383333	-232.346876	-94.41979	0.0000000
linseed-casein	-104.833333	-170.587491	-39.07918	0.0002100
meatmeal-casein	-46.674242	-113.906207	20.55772	0.3324584
soybean-casein	-77.154762	-140.517054	-13.79247	0.0083653
sunflower-casein	5.333333	-60.420825	71.08749	0.9998902
linseed-horsebean	58.550000	-10.413543	127.51354	0.1413329
meatmeal-horsebean	116.709091	46.335105	187.08308	0.0001062
soybean-horsebean	86.228571	19.541684	152.91546	0.0042167
sunflower-horsebean	168.716667	99.753124	237.68021	0.0000000
meatmeal-linseed	58.159091	-9.072873	125.39106	0.1276965
soybean-linseed	27.678571	-35.683721	91.04086	0.7932853
sunflower-linseed	110.166667	44.412509	175.92082	0.0000884
soybean-meatmeal	-30.480519	-95.375109	34.41407	0.7391356
sunflower-meatmeal	52.007576	-15.224388	119.23954	0.2206962
sunflower-soybean	82.488095	19.125803	145.85039	0.0038845

anov a and aov functions will both perform an ANOVA but the results are stored slightly differently. For this posthoc test we want the aov format

Interpreting post-hoc tests

```
> data("chickwts")
> fit <- aov(weight~feed, data=chickwts)
> TukeyHSD(fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level
```

```
Fit: aov(formula = weight ~ feed, data = chickwts)
```

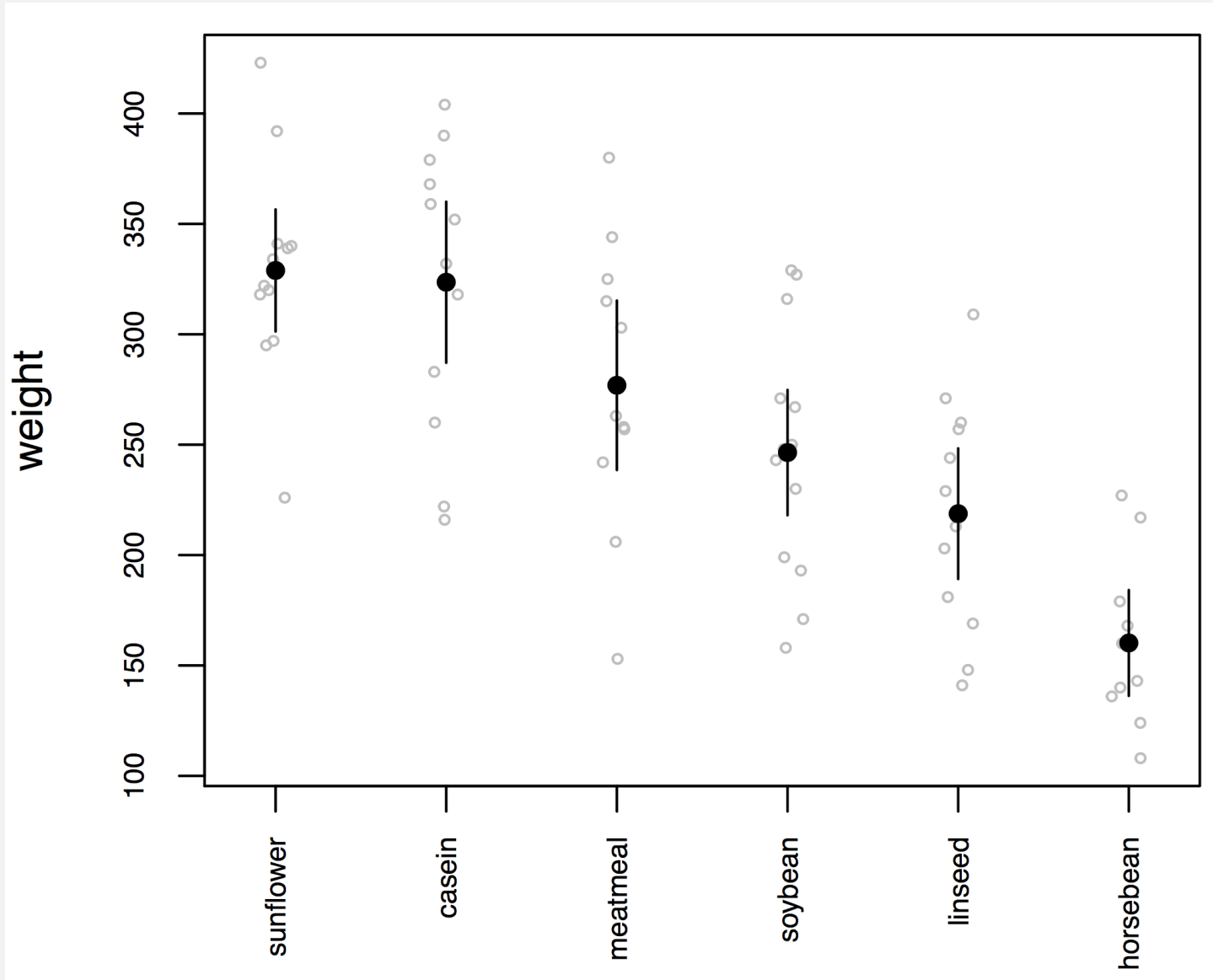
```
$feed
```

	diff	lwr	upr	p adj
horsebean-casein	-163.383333	-232.346876	-94.41979	0.0000000
linseed-casein	-104.833333	-170.587491	-39.07918	0.0002100
meatmeal-casein	-46.674242	-113.906207	20.55772	0.3324584
soybean-casein	-77.154762	-140.517054	-13.79247	0.0083653
sunflower-casein	5.333333	-60.420825	71.08749	0.9998902
linseed-horsebean	58.550000	-10.413543	127.51354	0.1413329
meatmeal-horsebean	116.709091	46.335105	187.08308	0.0001062
soybean-horsebean	86.228571	19.541684	152.91546	0.0042167
sunflower-horsebean	168.716667	99.753124	237.68021	0.0000000
meatmeal-linseed	58.159091	-9.072873	125.39106	0.1276965
soybean-linseed	27.678571	-35.683721	91.04086	0.7932853
sunflower-linseed	110.166667	44.412509	175.92082	0.0000884
soybean-meatmeal	-30.480519	-95.375109	34.41407	0.7391356
sunflower-meatmeal	52.007576	-15.224388	119.23954	0.2206962
sunflower-soybean	82.488095	19.125803	145.85039	0.0038845

When we examine all the significantly different ones we can draw several conclusions:

- 1) Chicks fed casein are significantly heavier than those fed horsebean, linseed, and soybean.
- 2) Chicks fed horsebean are significantly lighter than those fed meatmeal, soybean, and sunflower.
- 3) Chicks fed sunflower are significantly heavier than those fed linseed or soybean.

Plotting this kind of data



Our results from the ANOVA and Tukey match up pretty well with our rules of thumb about 95% CI overlaps

Example of code

```
#sets the order of the treatments
chickwts$feed <- factor(chickwts$feed,
                        levels=c("sunflower", "casein",
                                "meatmeal", "soybean",
                                "linseed", "horsebean"))

stripchart(weight ~ feed, data=chickwts,
            method = "jitter", vertical = TRUE, cex.axis = .7,
            col = "gray", pch = 1, cex = .5, las = 3)

#Add error bars:
#First calculate means and SDs
meanShift <- tapply(chickwts$weight, chickwts$feed, mean)
sdevShift <- tapply(chickwts$weight, chickwts$feed, sd)
n <- tapply(chickwts$weight, chickwts$feed, length)
feed_table <- data.frame(mean = meanShift,
                          std.dev = sdevShift, n = n)

#Now add the SEM for each group:
seShift <- 1.96 * sdevShift / sqrt(n)
segments(1:6, meanShift - seShift,
         1:6, meanShift + seShift)
points(meanShift ~ c(1:6), pch = 16)
```

Assumptions of the ANOVA

- The variable is normally distributed within each group
- The variance is the same in the different groups
- The design is balanced – you have the same sample size for each group
- But... ANOVA is fairly robust to violations of these assumptions

A Non-Parametric Alternative

- Kruskal-Wallis Test
- Based on ranks
- The multiple-group version of the Mann-Whitney U-test

R-implementation:

```
> kruskal.test(weight ~ feed, data = chickwts)
```

```
Kruskal-Wallis rank sum test
```

```
data: weight by feed
```

```
Kruskal-Wallis chi-squared = 37.343, df = 5, p-value = 5.113e-07
```

p-value suggests this
test has lower power
than ANOVA



A Non-Parametric post-hoc

- Dunn's test – is the non-parametric equivalent of the Tukey

Not in base R need to install:

```
install.packages("dunn.test", dependencies=TRUE)
```

```
> dunn.test(chickwts$weight, g=chickwts$feed,  
+           altp = T, method = "bonferroni")  
Kruskal-Wallis rank sum test
```

data: x and group

Kruskal-Wallis chi-squared = 37.3427, df = 5, p-value = 0

Comparison of x by group
(Bonferroni)

Col Mean-I					
Row Mean	casein	horsebea	linseed	meatmeal	soybean
horsebea	4.813069				
	0.0000*				
linseed	3.308292	-1.658736			
	0.0141*	1.0000			
meatmeal	1.415755	-3.364059	-1.819817		
	1.0000	0.0115*	1.0000		
soybean	2.499922	-2.602093	-0.933255	0.974144	
	0.1863	0.1390	1.0000	1.0000	
sunflowe	-0.182969	-4.987524	-3.491262	-1.594703	-2.689798
	1.0000	0.0000*	0.0072*	1.0000	0.1072

alpha = 0.05

Reject Ho if p <= alpha

ANOVA Summary

- ANOVA is the foundation of essentially all tests comparing multiple means
- Don't make it too complicated – the null hypothesis is simple: they are all the same.
- Post-hoc tests are important for determining which means are the source of a significant ANOVA.
- You can only justify a post-hoc test if the ANOVA is significant in the first place.
- Before applying ANOVA, check that your data fit the assumptions (consider transforming the data lots of times this will be based on your biological knowledge because you will have insufficient data to say much about the observed distribution)

ANOVA Practice Problems

- Use the chick weights dataset included in R data("chickwts"). Reduce the data down to just soybean and sunflower. Run an ANOVA and determine whether these foods lead to significant differences in weight.
- Use the offspring.csv file from the course website and determine whether XY and ZW systems have different numbers of male offspring.

Hypothesis testing has limits

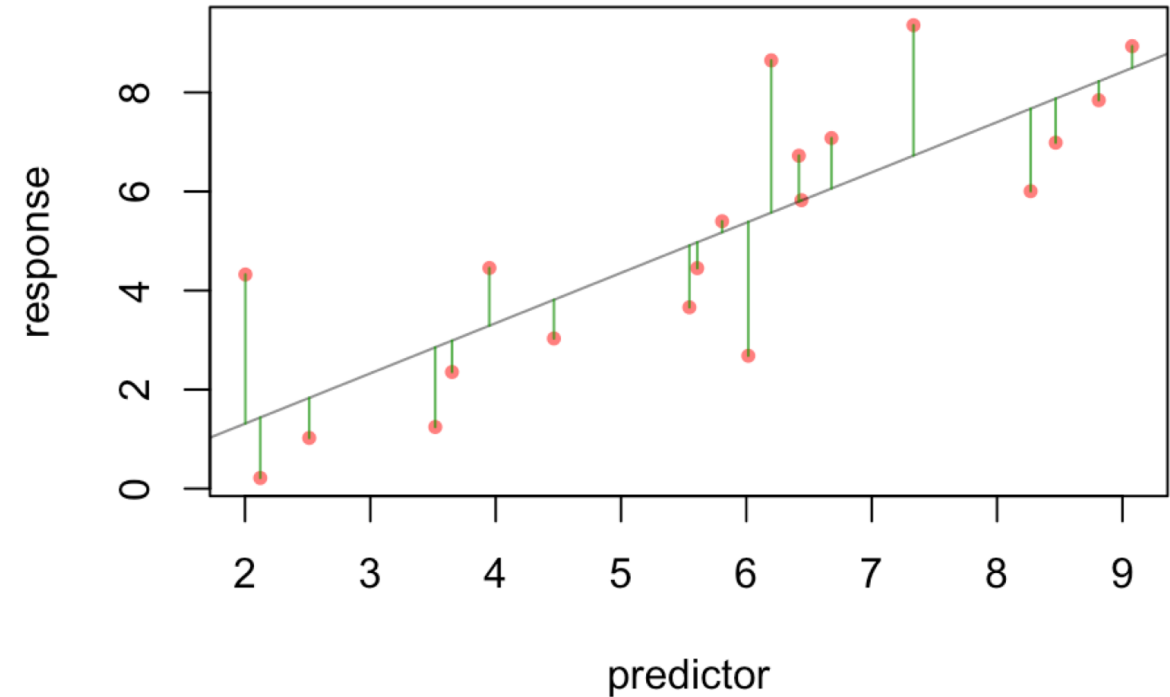
Often times we want to say more than something has an effect. We want to understand exactly how a predictor variable impacts a response variable.

Often times we have complex relationships where several variables (continuous and discrete) impact our response variable and we need to understand how all of these things work together to determine our observations.

The solution is to build a model that can describe the way that different variables make contributions to an observation.

Regression in R

- 1) With linear regression we find the linear equation that best predicts the values of Y based on the values of X.
- 2)
$$y = bx + a$$
- 3) Least-squares regression minimizes the squared deviations of the data points from that line.



Example of regression

```
set.seed(3)
x <- runif(min = 1, max = 10, 20)
y <- rnorm(20, mean = x, sd = 2)
fit.xy <- lm(y ~ x)
summary(fit.xy)
```

$$y = bx + a$$

$$t = \frac{\beta_0}{SE_b}$$

Call:
lm(formula = y ~ x)

Residuals:

Min	1Q	Median	3Q	Max
-2.7060	-0.9742	-0.4539	0.9479	3.0728

Coefficients:

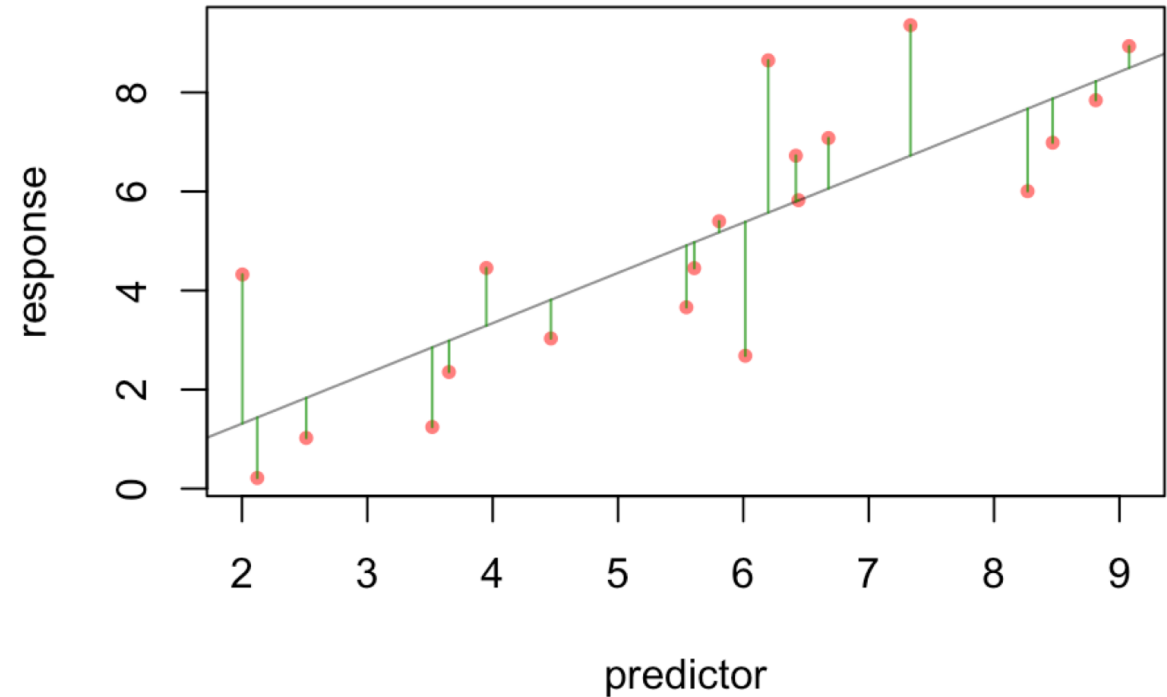
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.7173	1.0302	-0.696	0.495
x	1.0150	0.1708	5.943	1.27e-05 ***

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

Residual standard error: 1.625 on 18 degrees of freedom

Multiple R-squared: 0.6624, Adjusted R-squared: 0.6437

F-statistic: 35.32 on 1 and 18 DF, p-value: 1.267e-05



Example of regression

```
set.seed(3)
x <- runif(min = 1, max = 10, 20)
y <- rnorm(20, mean = x, sd = 2)
fit.xy <- lm(y ~ x)
summary(fit.xy)
```

$$y = bx + a$$

$$t = \frac{\beta_0}{SE_b}$$

Call:
lm(formula = y ~ x)

Residuals:

Min	1Q	Median	3Q	Max
-2.7060	-0.9742	-0.4539	0.9479	3.0728

Coefficients:

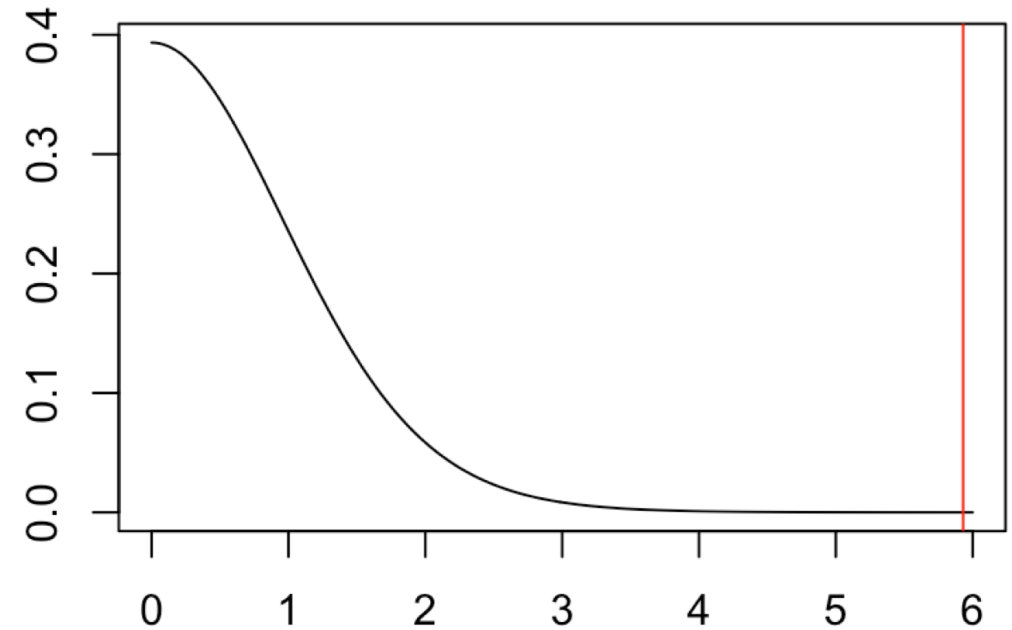
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.7173	1.0302	-0.696	0.495
x	1.0150	0.1708	5.943	1.27e-05 ***

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

Residual standard error: 1.625 on 18 degrees of freedom

Multiple R-squared: 0.6624, Adjusted R-squared: 0.6437

F-statistic: 35.32 on 1 and 18 DF, p-value: 1.267e-05



t-distribution

Example of regression

```
set.seed(3)
x <- runif(min = 1, max = 10, 20)
y <- rnorm(20, mean = x, sd = 2)
fit.xy <- lm(y ~ x)
summary(fit.xy)
```

Call:
lm(formula = y ~ x)

Residuals:

Min	1Q	Median	3Q	Max
-2.7060	-0.9742	-0.4539	0.9479	3.0728

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.7173	1.0302	-0.696	0.495
x	1.0150	0.1708	5.943	1.27e-05 ***

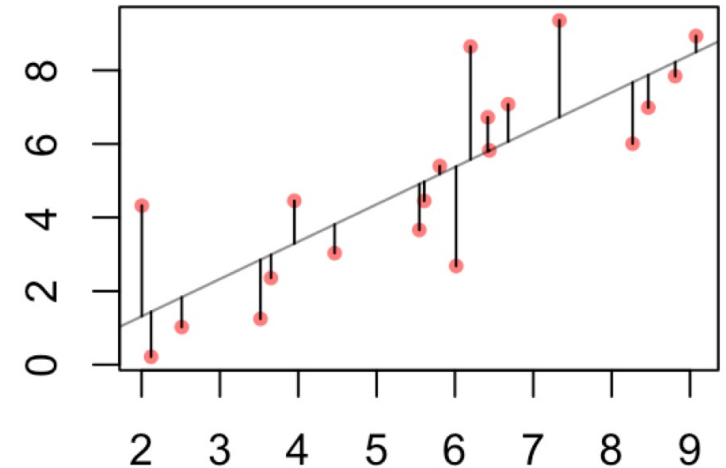
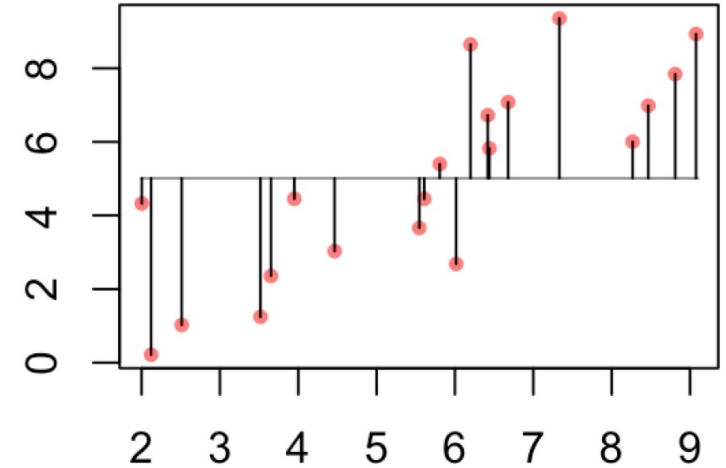
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.625 on 18 degrees of freedom

Multiple R-squared: 0.6624, Adjusted R-squared: 0.6437

F-statistic: 35.32 on 1 and 18 DF, p-value: 1.267e-05

This can help to justify the biological importance assuming you have a regression that is significant. It is the proportion of total variance explained by the regression.



Multiple vs Adjusted R-squared

Call:

```
lm(formula = y ~ x)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.7060	-0.9742	-0.4539	0.9479	3.0728

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.7173	1.0302	-0.696	0.495
x	1.0150	0.1708	5.943	1.27e-05 ***

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

Residual standard error: 1.625 on 18 degrees of freedom

Multiple R-squared: 0.6624, Adjusted R-squared: 0.6437

F-statistic: 35.32 on 1 and 18 DF, p-value: 1.267e-05

Adjusted R-squared penalizes for additional parameters

Linear regression uses

- Depict the relationship between two variables in an eye-catching fashion
- Test the null hypothesis of no association between two variables
 - The test is whether or not the slope is zero
- Predict the average value of variable Y for a group of individuals with a given value of variable X
 - variation around the line can make it very difficult to predict a value for a given individual with much confidence
 - Predictions outside of the range of observed data is generally discouraged
- Used both for experimental and observational studies

What are Residuals

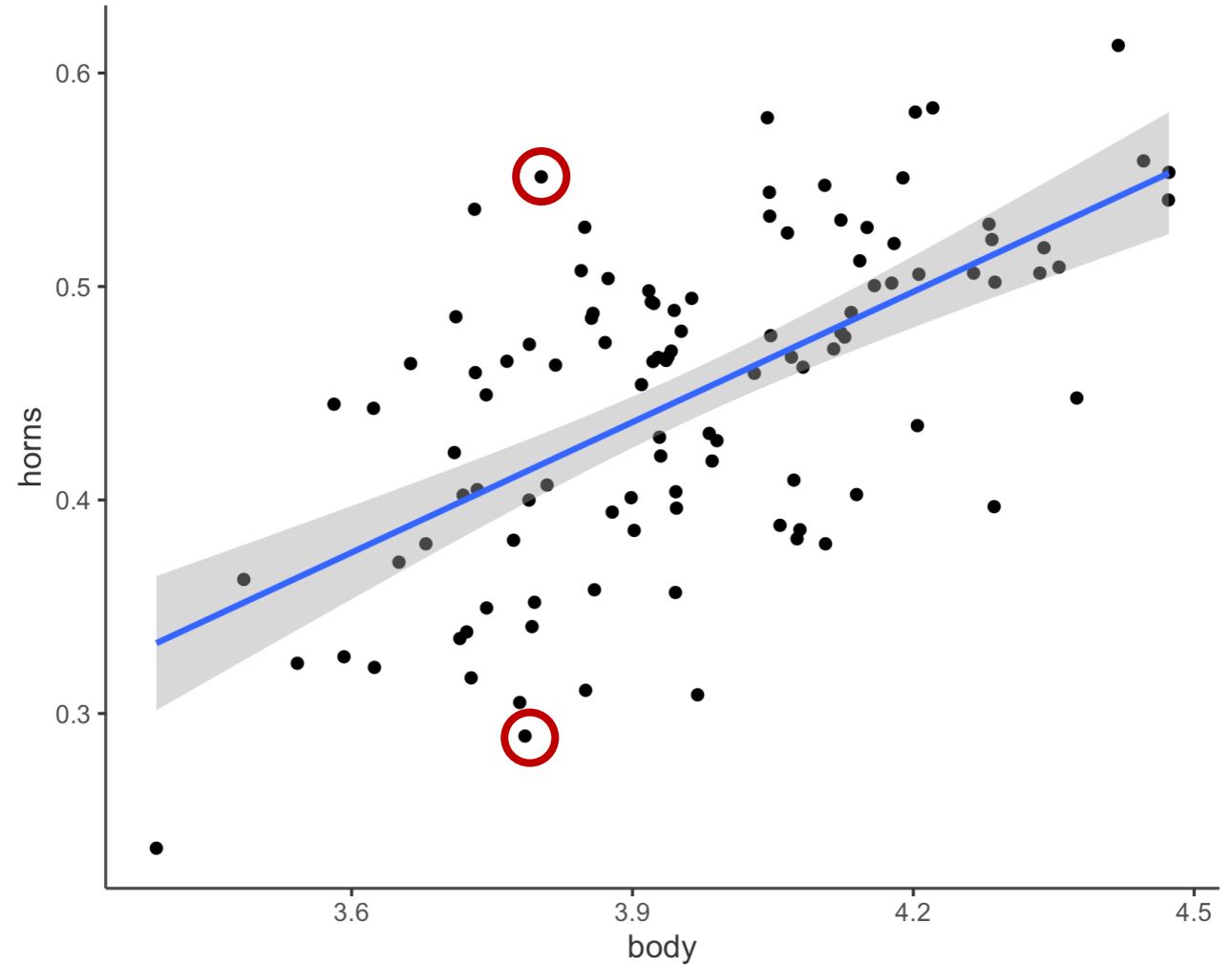
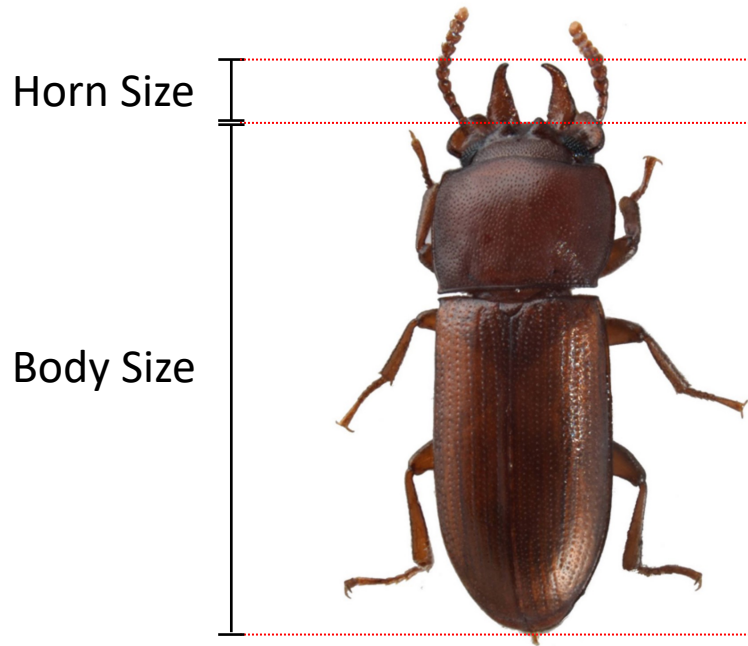
In general, the residual is the individual's departure from the value predicted by the model

In this case the model is simple – the linear regression – but residuals also exist for more complex models

For a model that fits better, the residuals will be smaller on average

Residuals can be of interest in their own right, because they represent values that have been ***corrected*** for relationships that might be obscuring a pattern.

What are Residuals



Making that plot

```
ggtheme <- theme_bw() + theme(panel.grid.major = element_blank(),  
                              panel.grid.minor = element_blank(),  
                              panel.background = element_blank(),  
                              panel.border=element_blank(),  
                              axis.line = element_line(colour="grey30"),  
                              axis.title = element_text(colour="grey20"),  
                              axis.text = (element_text(colour="grey30")),  
                              legend.title = element_text(colour="grey20"),  
                              legend.text = element_text(colour="grey30"))  
  
dat <- read.csv("gnatocerus.csv")  
ggplot(data = dat, aes(x=body, y=horn)) +  
  geom_point() + ggtheme +  
  geom_smooth(method='lm')
```


Assumptions of Linear Regression

- The true relationship must be linear
- At each value of X , the distribution of Y is normal (i.e., the residuals are normal)
- The variance in Y is independent of the value of X
- **Note that there are no assumptions about the distribution of X**

Common Problems

- Outliers
 - Regression is extremely sensitive to outliers
 - The line will be drawn to outliers, especially along the x-axis
 - Consider performing the regression with and without outliers
- Non-linearity
 - Best way to notice is by visually inspecting the plot and the line fit
 - Try a transformation to get linearity [often a log transformation]
- Non-normality of residuals
 - Can be detected from a residual plot
 - Possibly solved with a transformation
- Unequal variance
 - Usually visible from a scatterplot or from a residual plot

Moving past simple models

- The reason ANOVA is so widely used is that it provides a framework to simultaneously test the effects of multiple factors
- ANOVA also makes it possible to detect *interactions* among the factors
- ANOVA is a special case of a *general linear model*
- Linear regression is a special case of a *general linear model*

GLM and LM function in R

- The GLM and LM function in R takes equations that can be described with the following operators

+	+X include this variable
:	X:Z include the interaction between these variables
*	X*Y include these variables and the interactions between them
^	(X + Z + W)^3 include these variables and all interactions up to three way

R versus the math implied

`glm(y ~ X + W)`

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \epsilon_i$$

`glm(y ~ X * W)`

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + \beta_3 X_i W_i + \epsilon_i$$

Go to R

[Beta models](#)