

Linear modeling I

STEPHANIE J. SPIELMAN, PHD

BIO5312, FALL 2017

A solid orange horizontal bar at the bottom of the slide.

General notes on Homework

Do not use Rstudio-specific commands in R chunks

- View() is not an R command
- May prevent from knitting

Only set working directory once

- Add to top chunk with "root.dir"

Load all libraries in the top chunk

Chi-squared assumptions refer to **expected** counts

Don't give yourself more work than you have to

Don't make your hypotheses fancy

When to use median vs. mean for directional conclusions?

Announcing Final Project (25%)

Identify a dataset and ask **four scientific questions** about the data

- Use any of the statistical approaches we have learned to answer each question
- Make a descriptive figure for each scientific question

Scientific questions

Recall HW7:

- Do PHA levels tend to differ between the birds that received supplemental carotenoids and those that did not? Based on your results, can you infer anything about immune differences between birds that did and did not receive carotenoids?
- **Use a Mann Whitney U test to answer the scientific question**
- "Run a Mann Whitney U test on PHA levels between bird treatments" is not a scientific question
- What figure might be good to make here?
- What figures would not be good here?

Final project proposal

Homework due 11/28 will be a proposal

- Identify your dataset and give 1-2 paragraphs of background
 - **IN YOUR OWN WORDS**
 - Pose four scientific questions
 - Explain how you will solve each question
 - What statistical procedure and why
 - Indicate how you will visualize your data
- } 3-4 sentences total per question

Around 1 written page, single-spaced.

Updated schedule

Date	Topic
10/24	Linear modeling I
10/31	Linear modeling II and logistic regression
11/7	Model selection and evaluation
11/14	Principal Components Analysis (PCA) and clustering
11/21	Thanksgiving break
11/28	Advanced R grabbag and/or overflow
12/5	Advanced R grabbag and/or in-class office hours for final project Email me for special topic requests.
12/12	Final project due (by 11:59 pm on 12/12)

Linear Modeling

ANOVA and friends

Correlation

Regression

Multiple regression

ANOVA: Analysis of Variance

Used to compare more than 2 means (among k groups)

Ho: All means are the same, i.e. $\mu_1 = \mu_2 = \dots = \mu_k$

Ha: At least one mean is different, i.e. at least one $\mu_{<1-k>}$ differs

Why "can't" we use a t -test?

- We can do all the comparisons and use a P-value correction
- ANOVA is preferred

ANOVA Example

A clinical trial asks if there is a difference in mean daily calcium intake in adults with normal bone density, adults with osteopenia, and adults with osteoporosis. Each participant's daily calcium intake is measured based on reported food intake.

Normal Bone Density	Osteopenia	Osteoporosis
1200	1000	490
1000	1100	650
980	700	200
900	800	300
750	500	400
800	700	350

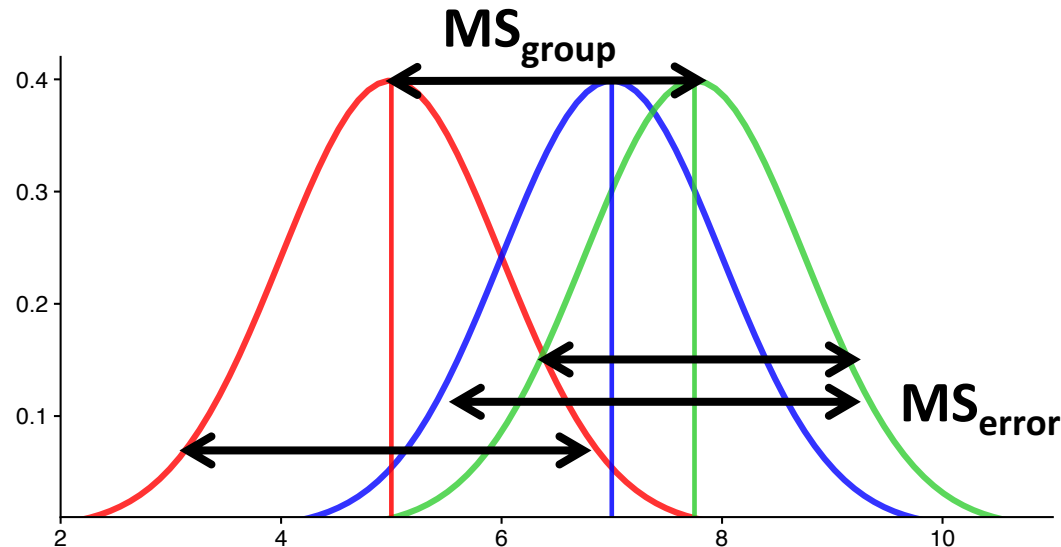
Is there a difference in mean calcium intake across groups?

ANOVA compares sources of *variance* using the F statistic

Variance *among* groups is the **group mean square** (MS_{group})

Variance *within* each group is the **error mean square** (MS_{error})

- Pooled sample variance



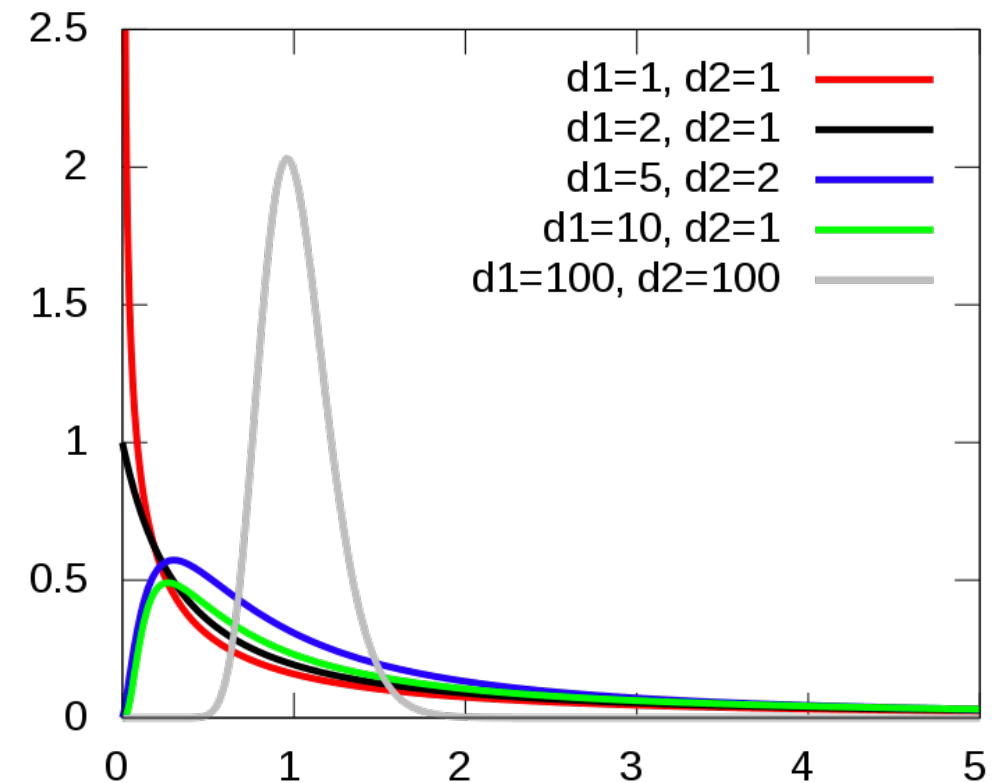
$$F = \frac{MS_{\text{group}}}{MS_{\text{error}}}$$

The F statistic

$$F = \frac{MS_{group}}{MS_{error}}$$

$$\text{df group} = k - 1$$

$$\text{df error} = N - k = \sum n - 1$$



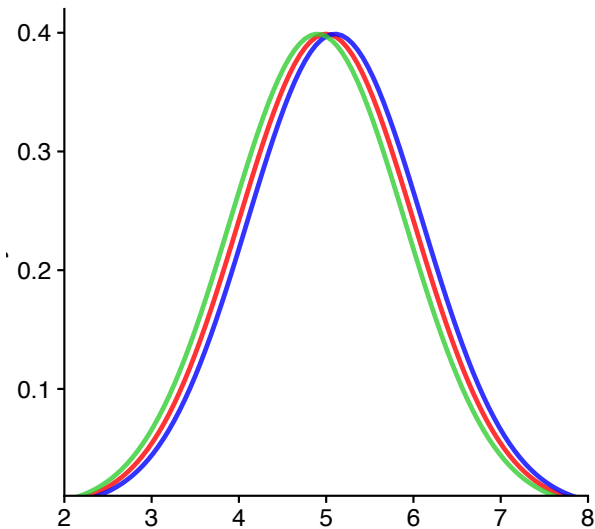
We use only the upper tail for P-value

https://en.wikipedia.org/wiki/F-distribution#/media/File:F-distribution_pdf.svg

Interpreting the F statistic

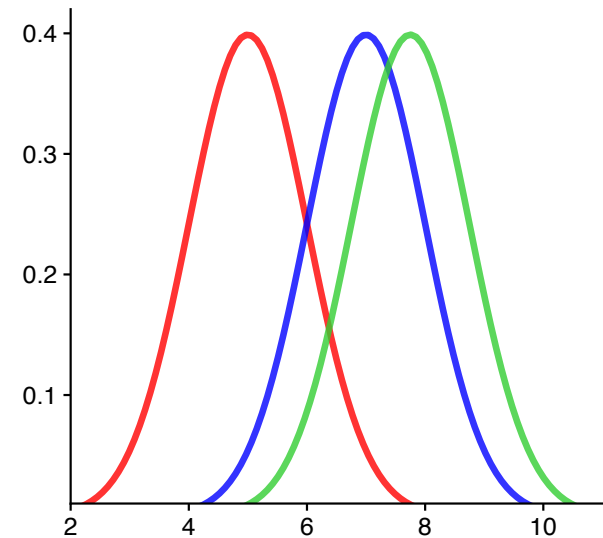
If H_0 is true, $F \cong 1$

$$MS_{\text{group}} \cong MS_{\text{error}}$$



If H_0 is false, $F > 1$

$$MS_{\text{group}} > MS_{\text{error}}$$



Computing the F statistic

$$F = \frac{MS_{group}}{MS_{error}} \quad MS = \frac{SS}{df}$$

Sum of squares

s_i^2 = Standard deviation of group i

$df_i = n_i - 1$, where n_i = sample size of group i

\bar{X}_i = mean of group i

\bar{X} = grand mean (mean of *all* numbers)

$$MS_{error} = \frac{\sum s_i^2 (n_i - 1)}{N - k}$$

$$MS_{group} = \frac{\sum n_i (\bar{X}_i - \bar{X})^2}{k - 1}$$

Error degrees of freedom =

$$df_{error} = \sum df_i = \sum (n_i - 1) = N - k$$

Computing the F statistic

$$F = \frac{MS_{group}}{MS_{error}} \quad MS = \frac{SS}{df}$$

$$MS_{error} = \frac{\sum s_i^2 (n_i - 1)}{N - k}$$

$$MS_{group} = \frac{\sum n_i (\bar{X}_i - \bar{X})^2}{k - 1}$$

Normal Bone Density	Osteopenia	Osteoporosis
1200	1000	490
1000	1100	650
980	700	200
900	800	300
750	500	400
800	700	350

Running the ANOVA

Ho: Groups have the same mean calcium intake.

Ha: At least one group has a different calcium intake.

```
data <- tibble("normal"      = c(1200, 1000, 980, 900, 750, 800),
               "osteopenia"  = c(1000, 1100, 700, 800, 500, 700),
               "osteoporosis" = c(490, 650, 200, 300, 400, 350))

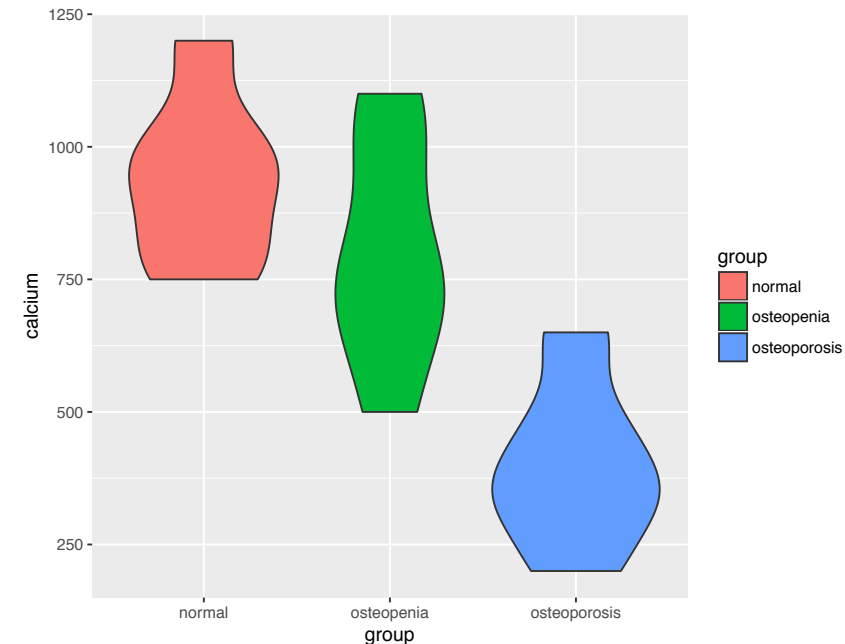
data %>% gather(group, calcium, normal:osteoporosis) -> tidy.data
```

	group	calcium
1	normal	1200
2	normal	1000
3	normal	980
4	normal	900
5	normal	750
6	normal	800
7	osteopenia	1000
8	osteopenia	1100
9	osteopenia	700
10	osteopenia	800
11	osteopenia	500
12	osteopenia	700
...		

Visualize the data

It is **always** the right idea to view your data before modeling it

```
ggplot(tidy.data, aes(x = group, y = calcium, fill= group)) + geom_violin()
```



Running the ANOVA

```
> aov(calcium ~ group, data = tidy.data)
```

Call:

```
  aov(formula = calcium ~ group, data = tidy.data)
```

Terms:

		error
	group	Residuals
Sum of Squares	944144.4	493166.7
Deg. of Freedom	2	15

Residual standard error: 181.3223

Estimated effects may be unbalanced

Obtaining the ANOVA table

```
> summary( aov(calcium ~ group, data = tidy.data) )
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
group	2	944144	472072	14.36	0.000328	***
error Residuals	15	493167	32878			

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

```
> 1-pf(14.36, 2, 15)
[1] 0.0003277806
```

$$MS_{group} = \frac{\sum n_i (\bar{X}_i - \bar{X})^2}{k - 1}$$

$$MS_{error} = \frac{\sum s_i^2 df_i}{N - k}$$

$$F = \frac{MS_{error}}{MS_{group}}$$

Reports and conclusions

Our $P = 0.000328$, which is less than α . We reject the null hypothesis and we have evidence that at least one mean (normal bone density, osteopenia, or osteoporosis calcium intake) differs from the other.

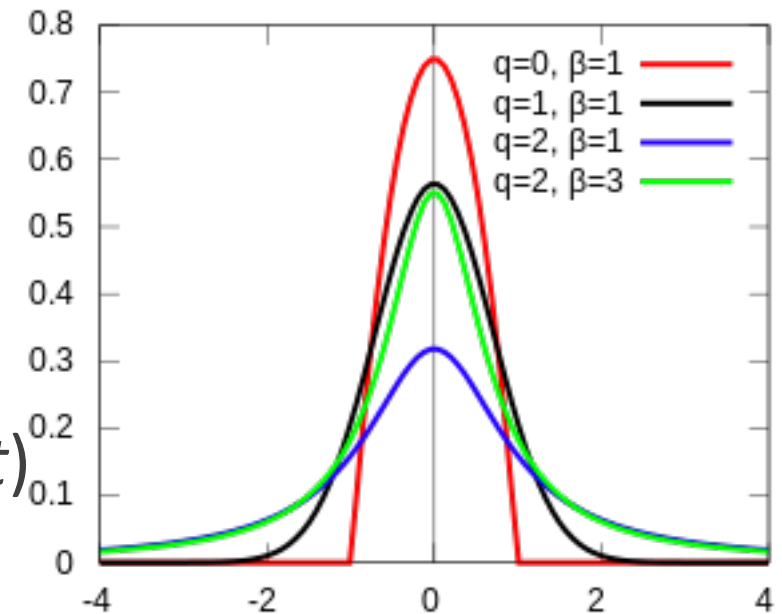
Unplanned comparisons with the Tukey-Kramer Method

AKA Tukey's test, Tukey's method, Tukey's HSD (honest significant difference) test

Tests all pairs of means

- Normal vs. osteopenia
- Normal vs. osteoporosis
- Osteopenia vs. osteoporosis

Roughly, multiple t-tests where FWER is controlled but using the q -statistic (similar to t)



Running Tukey's test on ANOVA

```
> TukeyHSD( aov(calcium ~ group, data = tidy.data) )
```

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

```
Fit: aov(formula = calcium ~ group, data = tidy.data)
```

```
$group
```

	diff	lwr	upr	p adj
osteopenia-normal	-138.3333	-410.2534	133.5867	0.4054988
osteoporosis-normal	-540.0000	-811.9200	-268.0800	0.0003238
osteoporosis-osteopenia	-401.6667	-673.5867	-129.7466	0.0043335

Reports and conclusions, round 2

Our $P = 0.000328$, which is less than α . We reject the null hypothesis and we have evidence that at least one mean (normal bone density, osteopenia, or osteoporosis calcium intake) differs from the other.

After running the *post-hoc* Tukey's test, we find that osteoporosis groups have a significantly higher average calcium intake than normal groups ($P=0.0003$), and that osteoporosis groups have a significantly higher average calcium intake than osteopenia groups ($P=0.004$). However, we do not find a significant difference in calcium intake between normal and osteopenia groups.

ANOVA assumptions

Samples are random

Samples are normally distributed

- Robust to violations when study is **large**

Samples have the same variance

- Robust to violations when study is **balanced**

Kruskal-Wallis is the non-parametric alternative

```
> kruskal.test(calcium ~ as.factor(group), data = tidy.data)
```

Kruskal-Wallis rank sum test

```
data:  calcium by as.factor(group)  
Kruskal-Wallis chi-squared = 11.439, df = 2, p-value = 0.003281
```

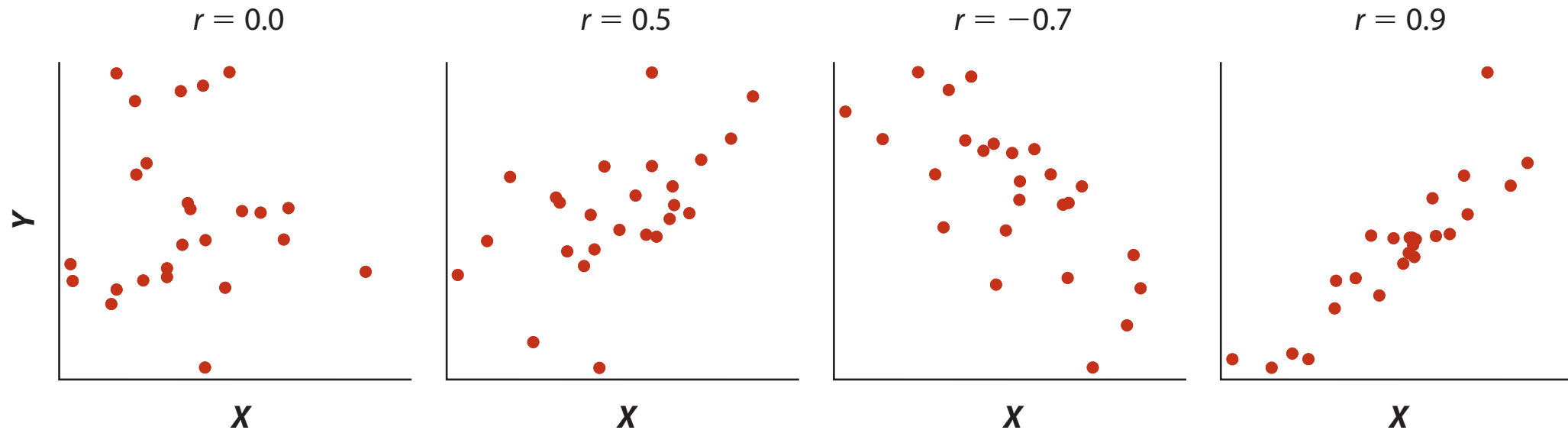
Unless something is **really really weird**, you "can" use ANOVA

Exercise break

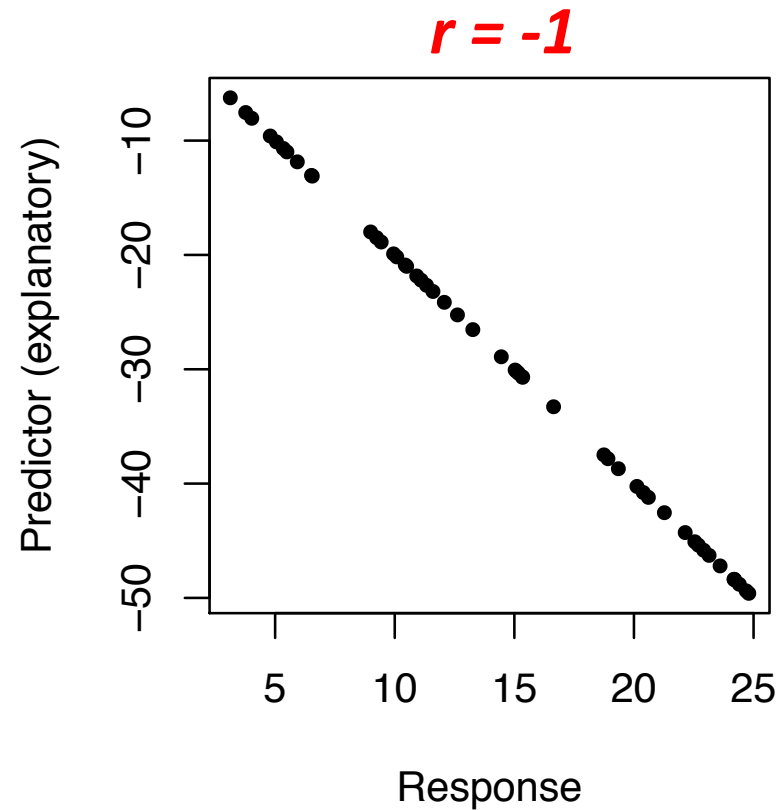
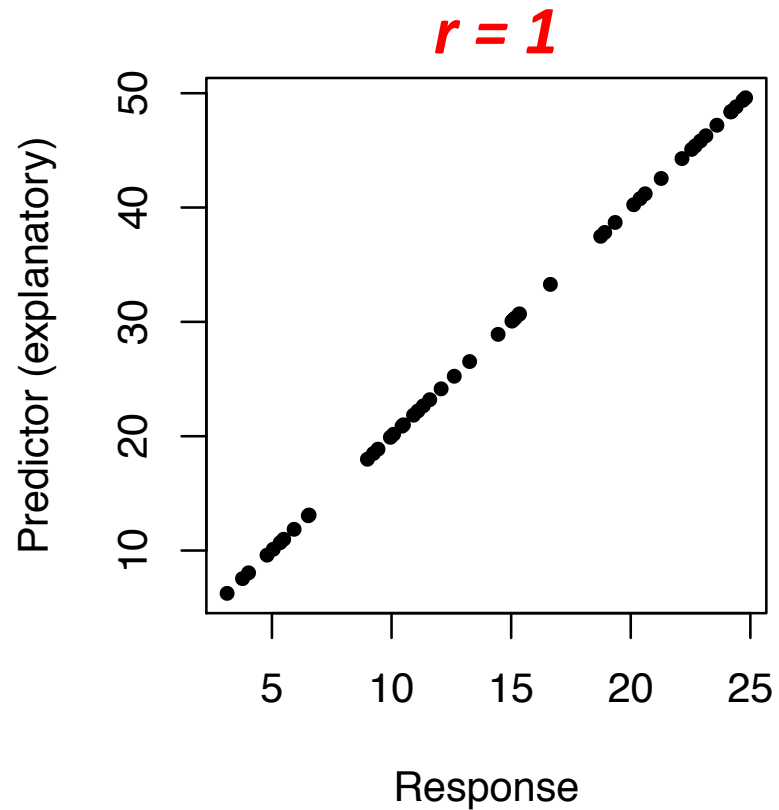
Correlation

Measures the strength and direction of the **linear association** between two numeric variables

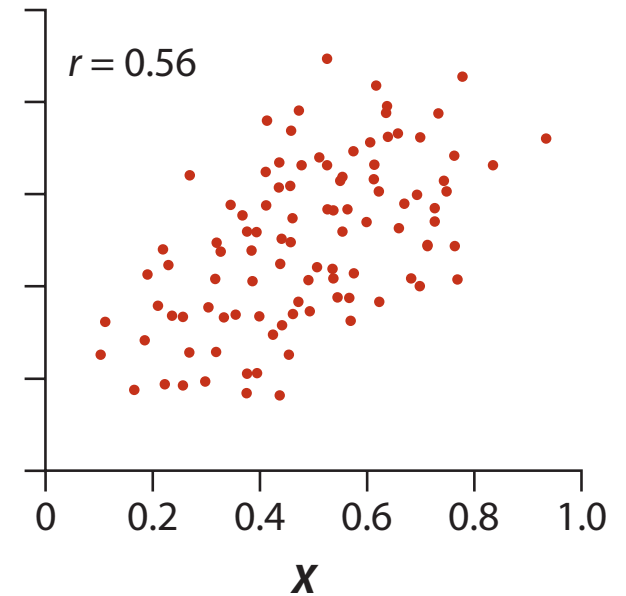
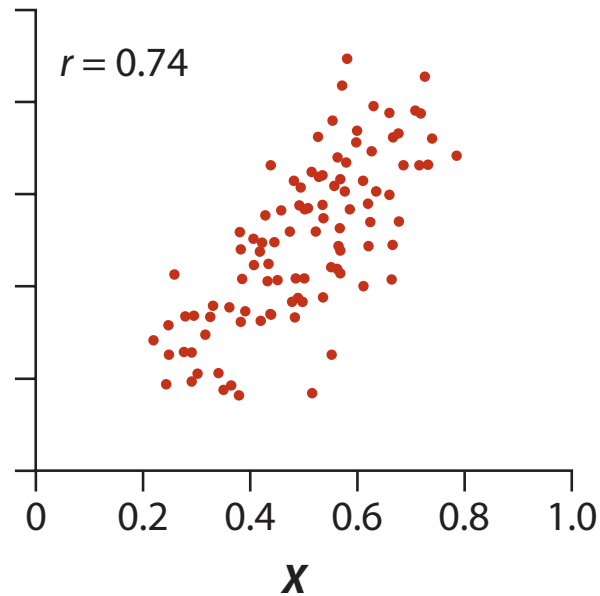
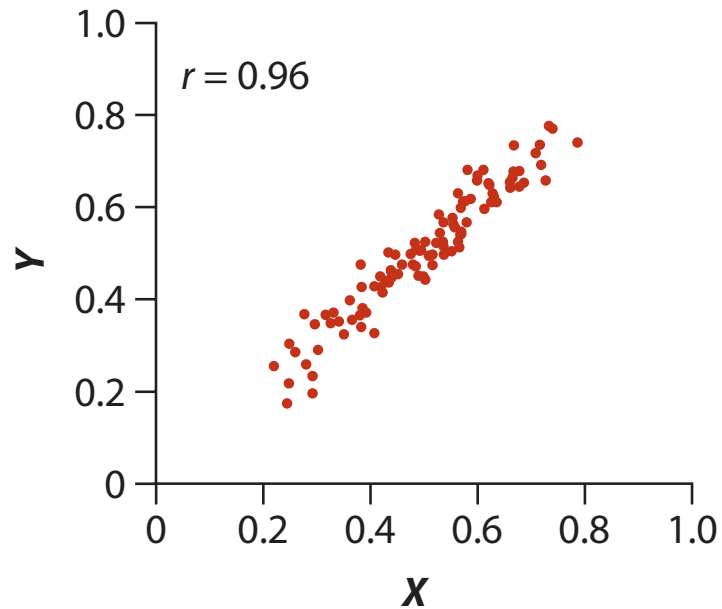
$$-1 \leq r \leq 1$$



Perfect correlations



Variability (error) has a substantial influence



Calculating the correlation coefficient

$$r = \frac{\textit{cov}(X, Y)}{s_X s_Y}$$

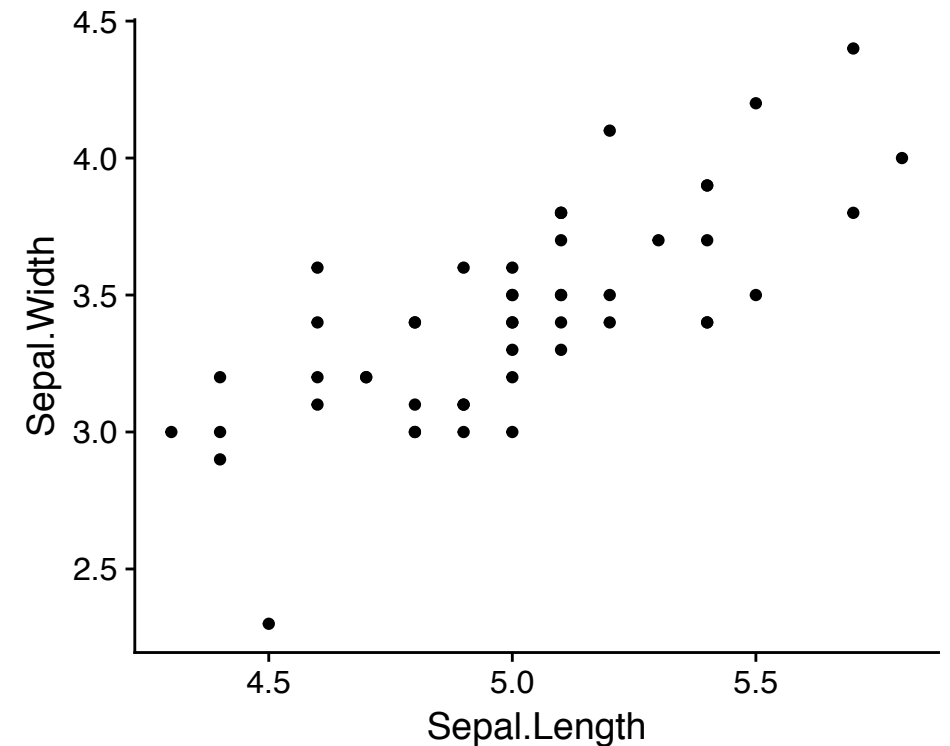
$$= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{\sum (X_i - \bar{X})^2}{n-1}} \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{n-1}}}$$

$$= \frac{1}{n-1} \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}}$$

Example: correlation between irises

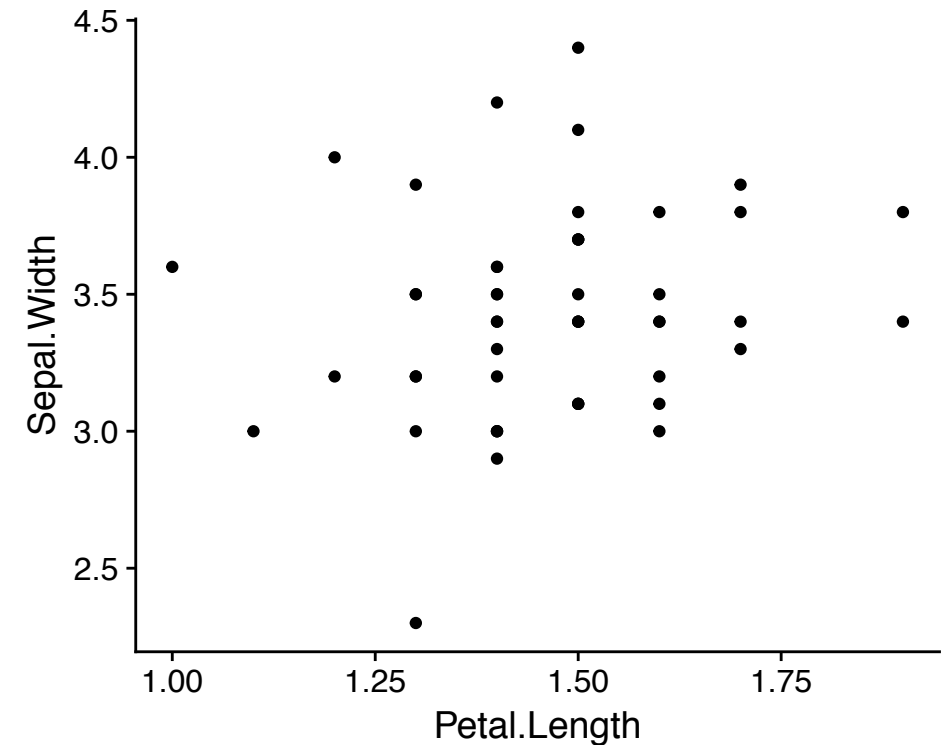
```
> setosa <- iris %>% filter(Species == "setosa")  
> cor(setosa$Sepal.Length, setosa$Sepal.Width)  
[1] 0.7425467
```

```
> cor(setosa$Sepal.Width, setosa$Sepal.Length)  
[1] 0.7425467
```



Example: correlation between irises

```
> cor(setosa$Petal.Length, setosa$Sepal.Width)  
[1] 0.1777
```



Hypothesis testing with correlations

H_0 : Petal length and sepal width are not correlated ($r=0$)

H_A : Petal length and sepal width are correlated ($r \neq 0$)

```
> cor.test(setosa$Sepal.Width, setosa$Sepal.Length)
```

Pearson's product-moment correlation

data: setosa\$Sepal.Width and setosa\$Sepal.Length

$t = 7.6807$, $df = 48$, $p\text{-value} = 6.71e-10$

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.5851391 0.8460314

sample estimates:

cor

0.7425467

With $P=6.7e-10$, which is much less than 0.05, we reject the null hypothesis. We find evidence that setosa sepal widths and sepal lengths are correlated. The correlation coefficient $r=0.74$, with a 95% CI of [0.58, 0.85] This value is above 0, indicating a positive relationship, and it is fairly large, indicating a strong correlation.

Hypothesis testing

```
> cor.test(setosa$Petal.Length, setosa$Sepal.Width)
```

Pearson's product-moment correlation

data: setosa\$Petal.Length and setosa\$Sepal.Width

t = 1.2511, df = 48, p-value = 0.217

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.1058851 0.4345536

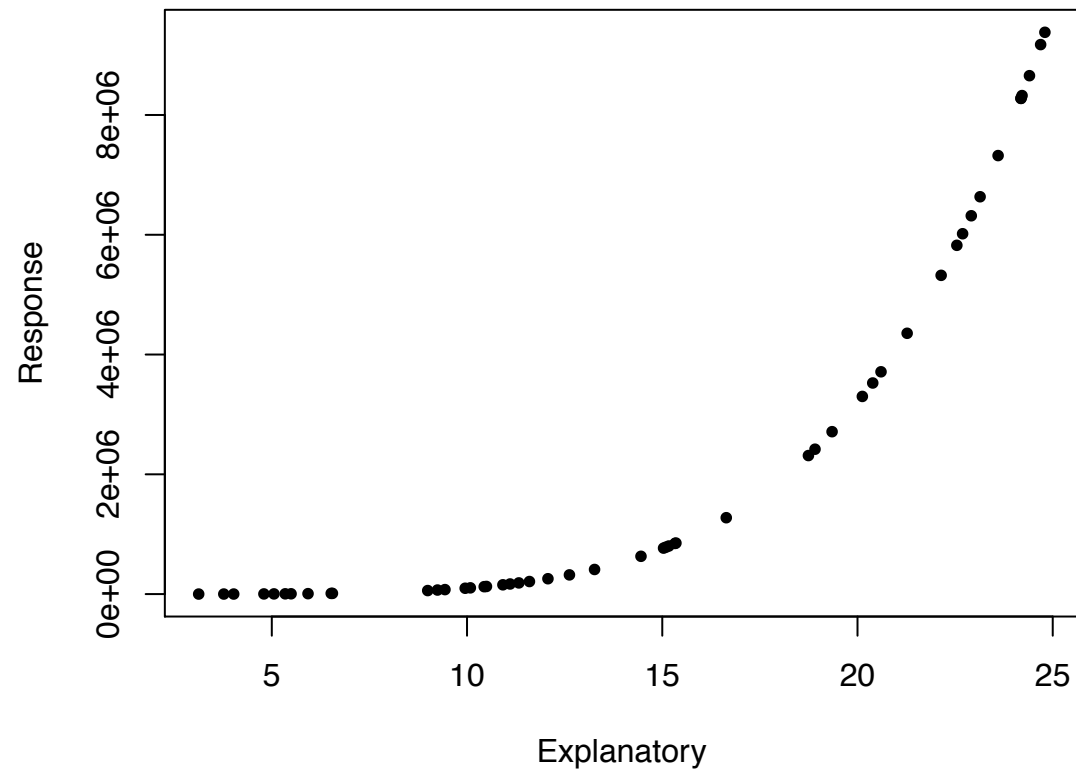
sample estimates:

cor

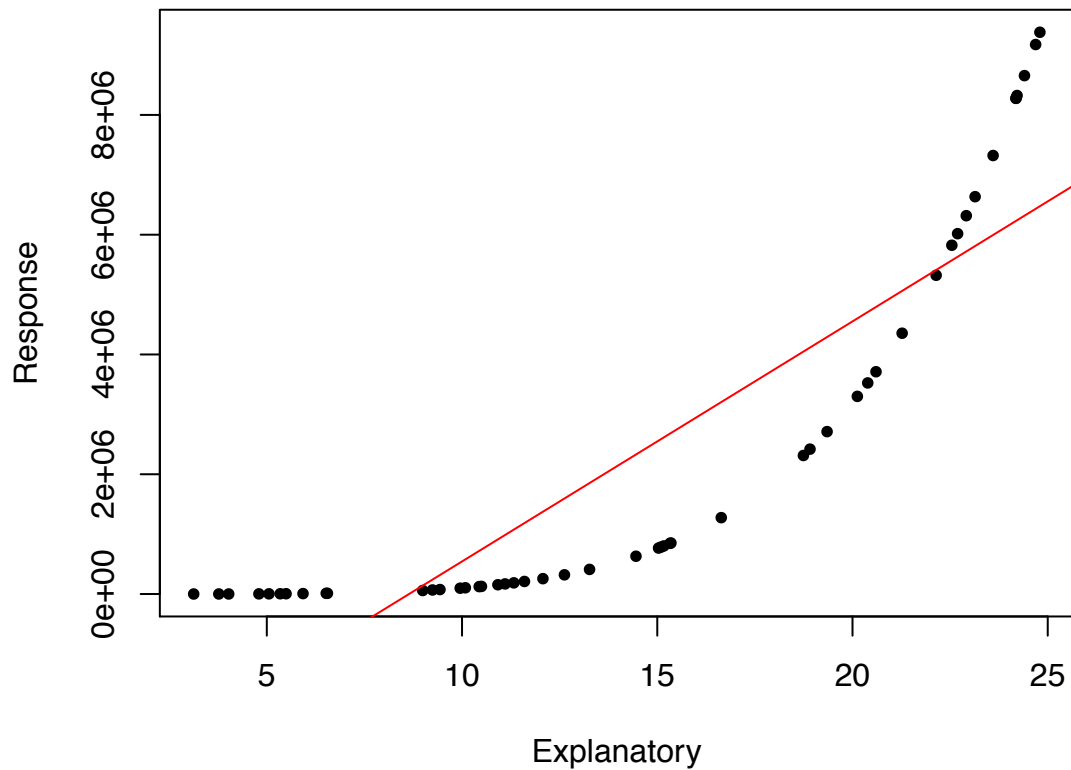
0.1777

We fail to reject the null hypothesis. There is no evidence that the correlation between petal lengths and sepal width in setosa irises differs from 0.

Nonlinear data



Nonlinear data



```
> cor.test(x, y)
```

Pearson's product-moment correlation

```
data: x and y
t = 13.089, df = 48, p-value < 2.2e-16
alternative hypothesis: true correlation is
      not equal to 0
95 percent confidence interval:
 0.8030401 0.9327180
sample estimates:
      cor
0.8838302
```

Spearman rank nonparametric correlation

Assumes data is *monotonic* (ordinal)

```
> cor.test(x, y, method = "spearman" )
```

Spearman's rank correlation rho

data: x and y

$S = 0$, p-value < $2.2e-16$

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

1

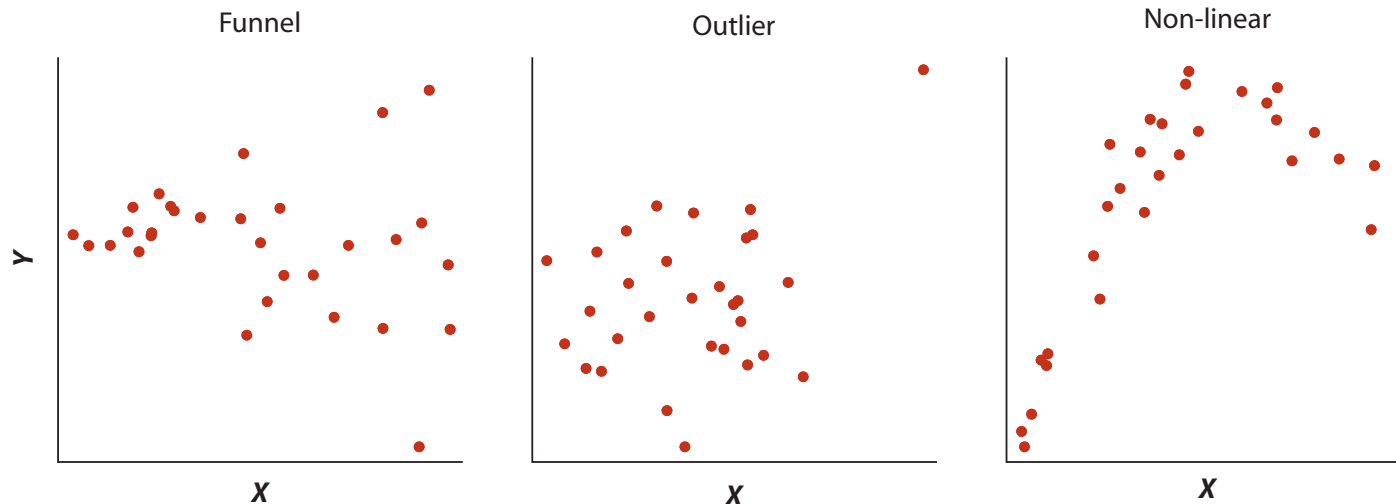
Assumptions: Check by plotting

Data are linearly related without any severe outliers

Both X and Y are normally distributed

- Robust to large N

Cloud of points is not "funnel-shaped" (fans out at end(s))



Exercise break

Regression

The simplest type of *linear model*

Predicts the value of one numeric variable from another via "line of best fit"

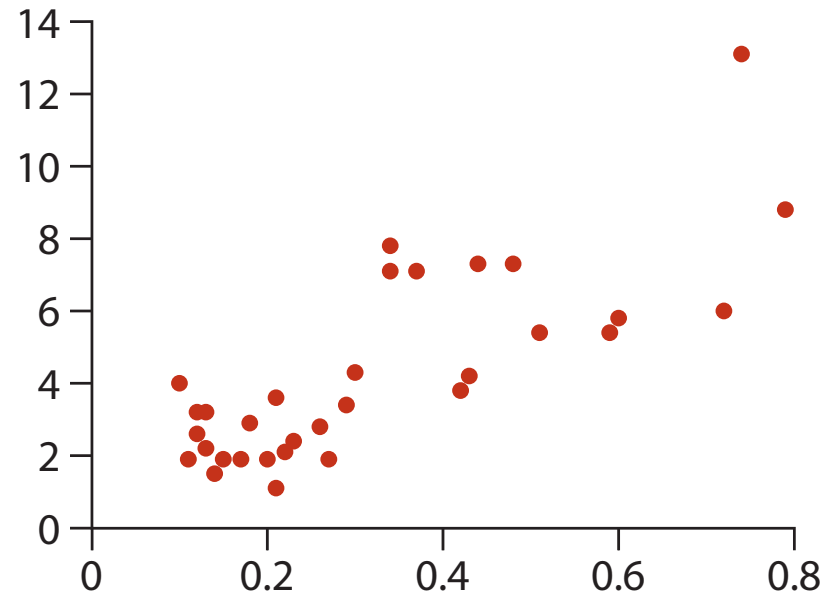
$$Y = a + bX$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

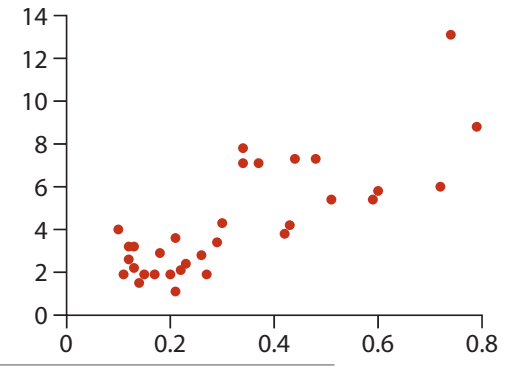
Residuals: ε is a random error component that measures how far above/below the line the **actual** value of Y for a given X lies. Mean is 0.

Least squares approach

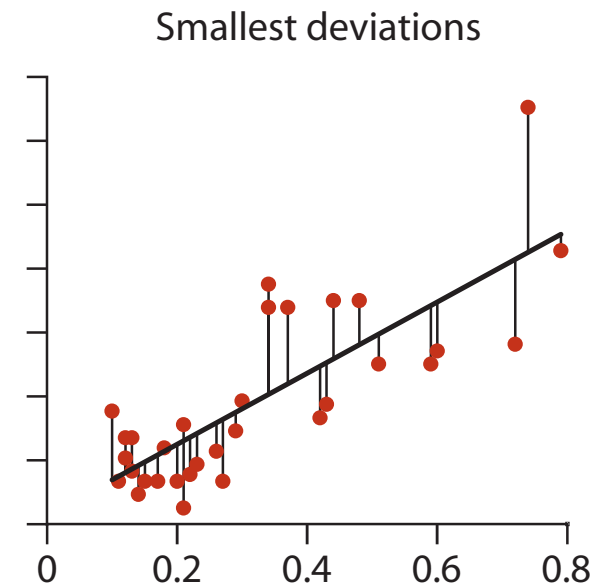
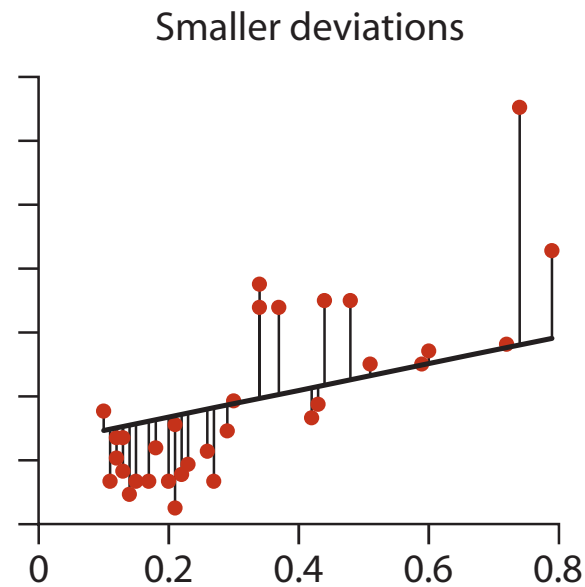
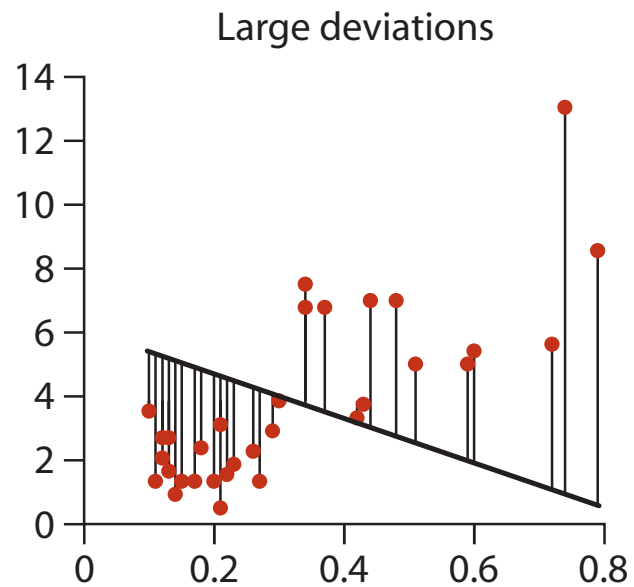
Find the line which **minimizes** the sum of squared residuals



Least squares approach



Find the line which **minimizes** the sum of squared residuals



Calculating slope and intercept

$$Y = a + bX$$

$$b = \frac{\text{cov}(X, Y)}{s_X^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n-1} \sum (X_i - \bar{X})^2}$$

$$a = \bar{Y} - b\bar{X}$$

The point (\bar{X}, \bar{Y}) always goes through the regression line

Executing a linear model

```
### lm(Y ~ X, data = <data frame>) ###
```

```
> lm(Sepal.Length ~ Sepal.Width, data = setosa)
```

Call:

```
lm(formula = Sepal.Length ~ Sepal.Width, data = setosa)
```

Coefficients:

(Intercept)	Petal.Width
2.6390	0.6905

$$Y = 2.64 + 0.69X$$

Testing a linear model

```
> summary(lm(Sepal.Length ~ Sepal.Width, data = setosa))
```

Call:

```
lm(formula = Sepal.Length ~ Sepal.Width, data = setosa)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.52476	-0.16286	0.02166	0.13833	0.44428

Five number summary of the distribution of residuals

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	2.6390	0.3100	8.513	3.74e-11	***
Sepal.Width	0.6905	0.0899	7.681	6.71e-10	***

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

Residual standard error: 0.2385 on 48 degrees of freedom

SE of ϵ

Multiple R-squared: 0.5514, Adjusted R-squared: 0.542

F-statistic: 58.99 on 1 and 48 DF, p-value: 6.71e-10

Test for model improvement over slope=0

Coefficients

Y X
lm(Sepal.Length ~ Sepal.Width)

Test for null hypothesis
that coefficient != 0

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	2.6390	0.3100	8.513	3.74e-11 ***	Expected value of Y when X=0
Sepal.Width	0.6905	0.0899	7.681	6.71e-10 ***	Expected unit increase in Y for every 1 unit increase in X

What can we conclude?

On average, setosa Sepal Length (**Y**) increases by 0.6905 cm (+/-0.0899 SE) for every 1 cm of Sepal Width (**X**).

[If $P < \alpha$, don't conclude this..]

R^2

R^2 is the percent of variation in Y than can be explained by X

- $0 \leq R^2 \leq 1$

Multiple R-squared: 0.5514, Adjusted R-squared: 0.542

$$R^2 = \left[\frac{cov(X,Y)}{s_X s_Y} \right]^2 = 1 - \frac{SS_{res}}{SS_{total}} = \frac{\textit{Explained variation}}{\textit{Total variation}}$$

What can we conclude?

~55% of the variation in Setosa sepal lengths (**Y**) can be explained by Setosa sepals widths (**X**).

[If $P < \alpha$, don't conclude this..]

Broom to the rescue

```
> summary( lm(Sepal.Length ~ Sepal.Width, data = setosa) )
```

```
Call:
```

```
lm(formula = Sepal.Length ~ Sepal.Width, data = setosa)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-0.52476	-0.16286	0.02166	0.13833	0.44428

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	2.6390	0.3100	8.513	3.74e-11	***
Sepal.Width	0.6905	0.0899	7.681	6.71e-10	***

```
---
```

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

```
Residual standard error: 0.2385 on 48 degrees of freedom
```

```
Multiple R-squared:  0.5514, Adjusted R-squared:  0.542
```

```
F-statistic: 58.99 on 1 and 48 DF,  p-value: 6.71e-10
```

Broom to the rescue

```
> library(broom)
> model <- lm(Sepal.Length ~ Sepal.Width, data = setosa)
```

```
##### Coefficients and Pvalues #####
```

```
> tidy(model)
```

	term	estimate	std.error	statistic	p.value
1	(Intercept)	2.6390012	0.31001431	8.512514	3.742438e-11
2	Sepal.Width	0.6904897	0.08989888	7.680738	6.709843e-10

```
##### Concise *one row* summary #####
```

```
> glance(model)
```

	r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC
1	0.5513756	0.5420292	0.2385422	58.99373	6.709843e-10	2	1.734067	2.531865

	BIC	deviance	df.residual
1	8.267934	2.731315	48

Broom to the rescue

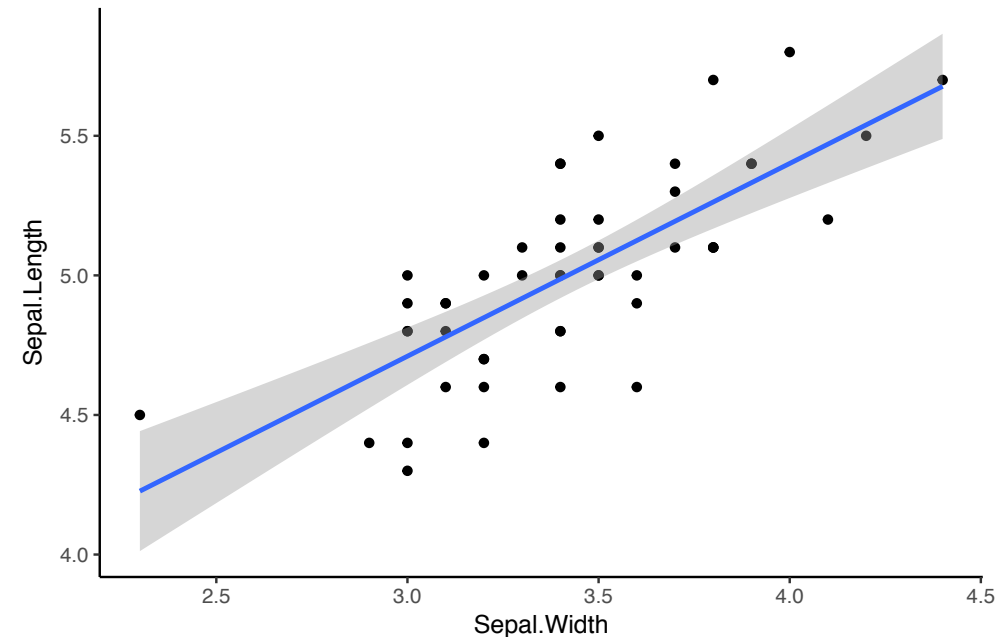
```
##### Add columns from fit to the original data that was modeled #####
```

```
> augment(model)
```

	Sepal.Length	Sepal.Width	.fitted	.se.fit	.resid	.hat	.sigma	.cooksd	.std.resid
1	5.1	3.5	5.055715	0.03435031	0.04428474	0.02073628	0.2409782	3.726311e-04	0.18760265
2	4.9	3.0	4.710470	0.05117134	0.18952960	0.04601750	0.2393991	1.596010e-02	0.81347001
3	4.7	3.2	4.848568	0.03947370	-0.14856834	0.02738325	0.2400630	5.614273e-03	-0.63152438
4	4.6	3.1	4.779519	0.04480537	-0.17951937	0.03528008	0.2395878	1.073468e-02	-0.76620575
5	5.0	3.6	5.124764	0.03710984	-0.12476423	0.02420180	0.2403616	3.476539e-03	-0.52947419
6	5.4	3.9	5.331911	0.05420835	0.06808885	0.05164186	0.2408507	2.339099e-03	0.29310589

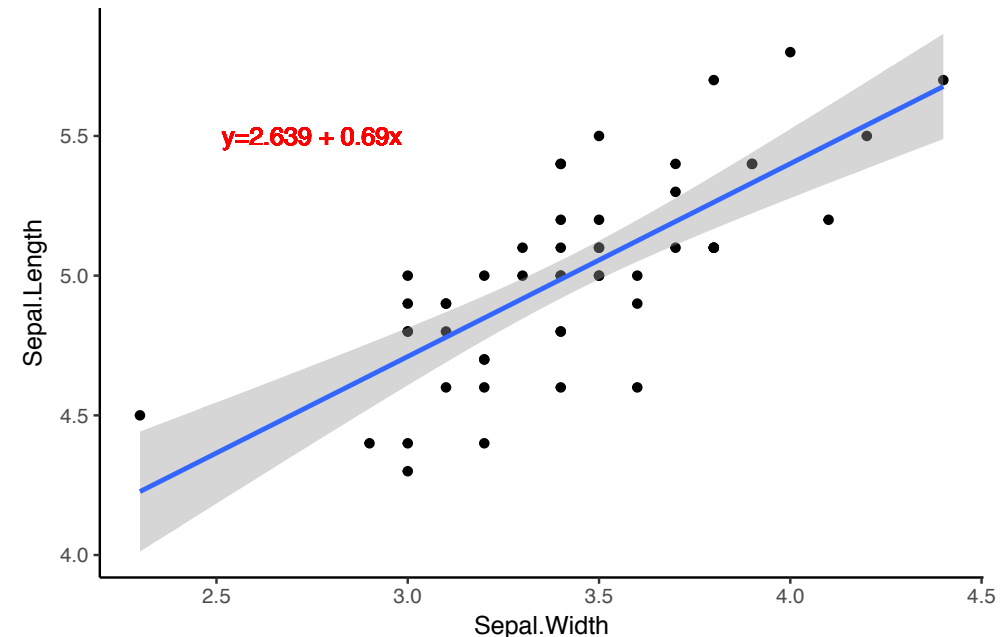
Visualizing the regression

```
ggplot(setosa, aes(x = Sepal.Width, y = Sepal.Length)) +  
  geom_point() + geom_smooth(method = "lm")
```



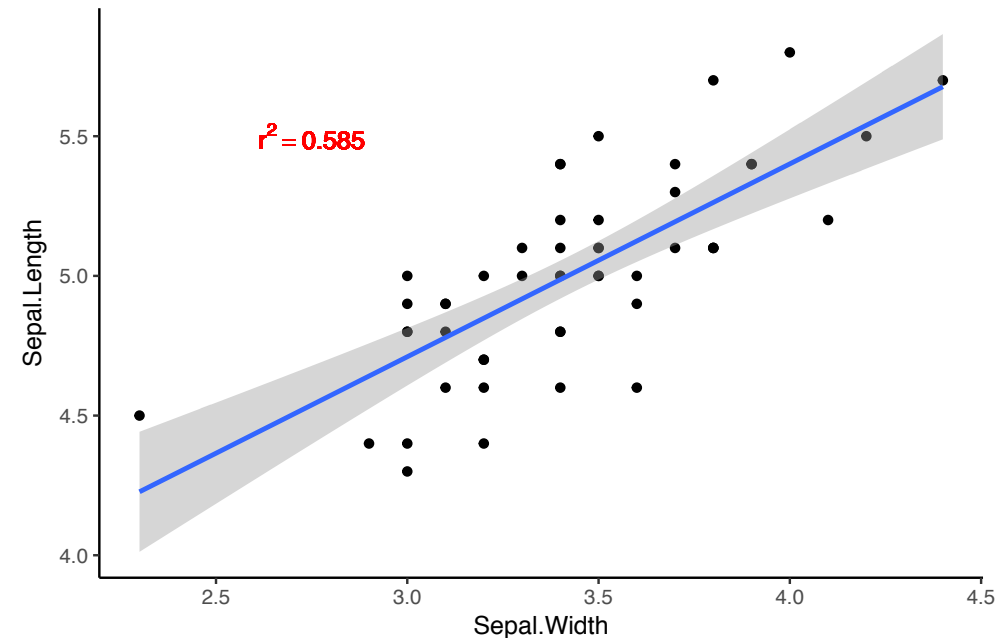
Visualizing the regression

```
ggplot(setosa, aes(x = Sepal.Width, y = Sepal.Length)) +  
  geom_point() + geom_smooth(method = "lm") +  
  geom_text(x = 2.75, y = 5.5, label = "y=2.639 + 0.69x", color="red")
```



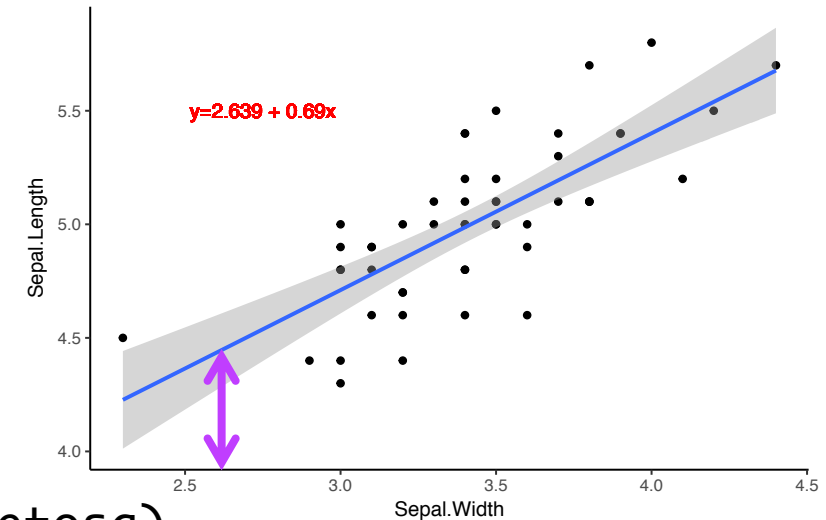
Visualizing the regression

```
ggplot(setosa, aes(x = Sepal.Width, y = Sepal.Length)) +  
  geom_point() + geom_smooth(method = "lm") +  
  geom_text(x=2.75, y=5.5, label="r^2 == 0.585", parse=TRUE, color="red")
```



Using the model: Predicting responses

What is sepal length for a sepal width of 2.6?



```
> model <- lm(Sepal.Length ~ Sepal.Width, data = setosa)
> new.data <- tibble(Sepal.Width = 2.6) ## Same column name as model's predictor
> predict(model, new.data)
      1
4.434275
```

Predicting with intervals

Confidence interval

- Range that is likely to contain **the mean response**

$$\hat{y}_h \pm t_{(\alpha/2, n-2)} \times \sqrt{MSE \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}$$

Prediction interval

- Range that is likely to contain **the response value of a single new observation**
- Wider than CI due to added uncertainty for predicting a single point

$$\hat{y}_h \pm t_{(\alpha/2, n-2)} \times \sqrt{MSE \times \left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}$$

Predicting with intervals

```
> predict(model, new.data)
```

```
      1  
4.434275
```

```
> predict(model, new.data, interval = "confidence")
```

```
      fit      lwr      upr  
1 4.434275 4.269957 4.598592
```

```
> predict(model, new.data, interval = "predict")
```

```
      fit      lwr      upr  
1 4.434275 3.927287 4.941262
```

```
> predict(model, new.data, interval = "confidence", level = 0.9)
```

```
      fit      lwr      upr  
1 4.434275 4.297205 4.571344
```

Assumptions of linear models

Residuals are normally distributed

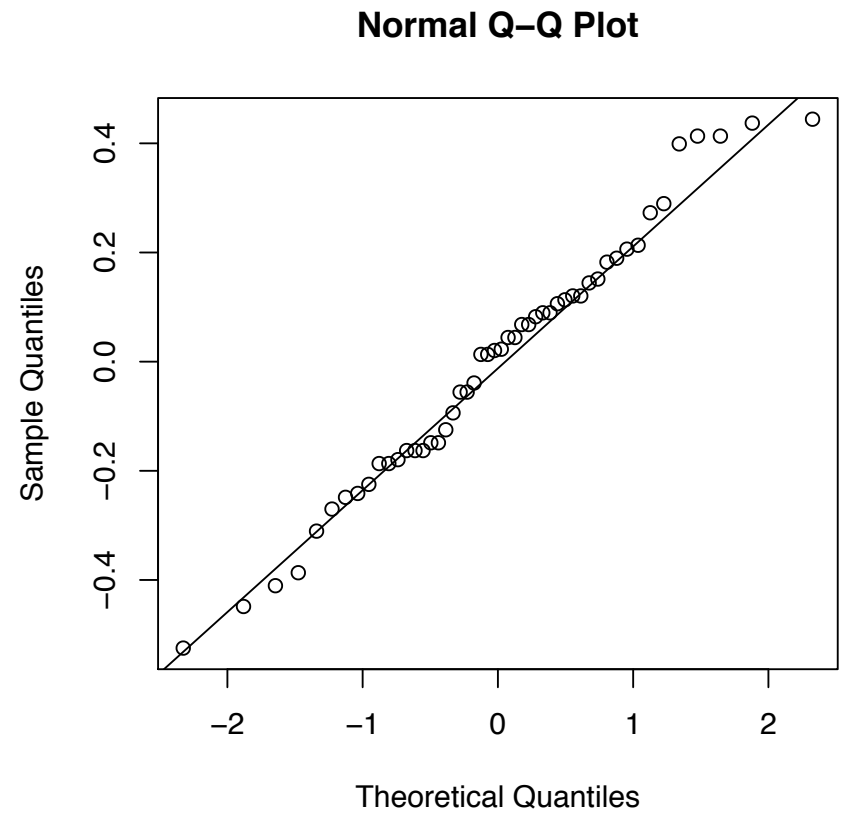
The variance is the same for all predictors*

Predictors are independent of each other*

The relationship between response and any numeric predictors is linear*

Normality of residuals

```
> augmented <- augment(model)
> qqnorm(augmented$.resid)
> qqline(augmented$.resid)
```



How to check regression assumptions

- 1. Plot response and predictor against each other to ensure linearity**
 - Critically important
- 2. Plot the residuals to ensure normality**
 - Important, usually overlooked
 - Most times we are robust to departures

Exercise break

Linear Models

`lm(Numeric response ~ <predictors>)`

Linear Models

`lm(Numeric response ~ <predictors>)`

Single numeric predictor: Regression

Single categorical predictor: ANOVA

Multiple numeric predictors: multiple regression

Multiple categorical predictors: n -way ANOVA

Single categorical and n numeric predictors: ANCOVA

Multiple categorical and n numeric predictors: linear model

ANOVA meets linear modeling

	group	calcium
1	normal	1200
2	normal	1000
3	normal	980
4	normal	900
5	normal	750
6	normal	800
7	osteopenia	1000
8	osteopenia	1100
9	osteopenia	700
10	osteopenia	800
11	osteopenia	500
12	osteopenia	700
...		

```
> summary(aov(calcium ~ group, data = tidy.data))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2	944144	472072	14.36	0.000328 ***
Residuals	15	493167	32878		

$$\frac{944144}{944144 + 493167} = 0.656$$

```
> summary(lm(calcium ~ group, data = tidy.data))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	938.33	74.02	12.676	2.04e-09 ***
grouposteopenia	-138.33	104.69	-1.321	0.206168
grouposteoporosis	-540.00	104.69	-5.158	0.000117 ***

Residual standard error: 181.3 on 15 degrees of freedom

Multiple R-squared: 0.6569, Adjusted R-squared: 0.6111

F-statistic: 14.36 on 2 and 15 DF, p-value: 0.000328

On average, the normal group consumes 938.33 calcium

~~Compared to the normal group, the osteopenia group consumes on average -138.33 less calcium.~~

Compared to the normal group, the osteoporosis group consumes on average -540 less calcium.

Which group you belong to explains ~66% of the variation in calcium intake

Briefly, bootstrapping the regression

```
> library(slipper)
> setosa %>%
  slipper_lm(Sepal.Length ~ Sepal.Width, B=1e3)%>% head()
      term      value      type
1 (Intercept) 2.6390012 observed
2 Sepal.Width 0.6904897 observed
3 (Intercept) 2.0900929 bootstrap
4 Sepal.Width 0.8467474 bootstrap
5 (Intercept) 2.7629316 bootstrap
6 Sepal.Width 0.6527575 bootstrap

setosa %>%
  slipper_lm(Sepal.Length ~ Sepal.Width, B=1e3) %>%
  filter(type == "bootstrap", term == "Sepal.Width") %>%
  summarize(mean = mean(value),
            ci_low = quantile(value,0.025),
            ci_high = quantile(value,0.975))
      mean      ci_low      ci_high
1 0.6945918 0.5302098 0.8961058
```