STA220H1: The Practice of Statistics I

Elena Tuzhilina

April 4, 2023

Please turn on your videos :)

Announcements

1. We have one more Quiz left.
2. The Midterm 2 regrade requests are due tonight.
3. We will have additional office hours:

Elena: Monday (April 10 and 17) at 11 am - 12 pm

Alice: Tuesday (April 18) at 1 - 2 pm

Vicky: Tuesday (April 18) at 7 - 8 pm

Ichiro: Wednesday (April 19) at 10 - 11 am

Agenda for today

- Recap: statistical testing for two samples
- Testing for two categorical variables
- Linear regression

Statistical testing

One sample: $x_1 \ldots x_n$

- z-test for proportion $H_0: p = p_0$
- t-test for population mean $H_0: \mu = \mu_0$

Two matching samples: $x_1 \ldots x_n \quad y_1 \ldots y_n$

- paired t-test $H_0: \mu_d = 0$
- signed test $H_0: p = 0.5$

Two non-matching samples: $x_1 \ldots x_n \quad y_1 \ldots y_m$

- z-test for proportions $H_0: p = q$
- t-test for two population means $H_0: \mu_x = \mu_y$

Statistical testing for two samples: mathching samples

T-test with matching samples compares two samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ with matching observations.

- Sample sizes are equal
- Samples are not independent

Statistical testing for two samples: mathching samples

Paired t-test: works for matching pairs.

- Create a sample that shows the difference in measurements

$$d_1, \ldots, d_n \text{ where } d_i = x_i - y_i$$

$\bar{d}$

- Perform statistical test on differences testing $H_0 : \mu_d = 0$ vs $H_a : \mu_d \neq 0$

Assumptions: requires the average difference $\bar{d}$ to come from Normal distribution

- d_i came from normal distribution
- n is large (CLT)

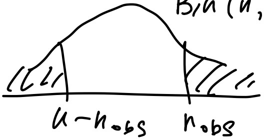
Statistical testing for two samples: mathching samples

Signed test: an alternative to paired t-test when assumptions are violated.

d_i

- Compute n_{obs} the number of positive differences
- Perform statistical test on the probability to get a positive difference H_0 : $p = 0.5$ vs H_a : $p \neq 0.5$ $\quad p > 0.5$
- Use null distribution $N \sim$ ~~*Bernoulli*~~ Binomial$(n, 0.5)$ to compute *p-value* $= P(N \geq n_{obs}) + P(N \leq n - n_{obs})$

$Bin(n, 0.5)$



Assumptions:

- No assumptions
- Works for small n

diff = after − before

$H_0 : p = 0.5 \quad H_a : p < 0.5 \quad | P(N \leq 3)$

Statistical testing for two samples: non-mathching samples

T-test with non-matching samples compares two samples $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$ with non-matching observations.

- Sample sizes can be different $n \neq m$
- Samples are independent

$n = m$

Statistical testing for two samples: non-mathching samples

Proportions: compares the probability of "successful" outcomes in $x_1, \dots, x_n$ and $y_1, \dots, y_m$.

0 1 . . . 1 0 1

- Perform statistical test on the probabilities $H_0 : p = q$ vs. $H_a : p \neq q$
- "Pool" two samples to approximate $p, q \approx \frac{n\bar{x}+m\bar{y}}{n+m}$
- Use test statistic

$$z_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{n\bar{x}+m\bar{y}}{n+m}\left(1 - \frac{n\bar{x}+m\bar{y}}{n+m}\right)\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

to find p-value

$p(1-p)$ $q(1-q)$

Assumptions: requires the difference in sample means $\bar{x} - \bar{y}$ to come from Normal distribution

- x_i and y_i came from normal distribution
- Both $n > 30$ and $m > 30$ (CLT)

$n + m > 30$

Statistical testing for two samples: non-mathching samples

Means: compares the population means of $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$.

- Perform statistical test on the probabilities $H_0 : \mu_x = \mu_y$ vs. $H_a : \mu_x \neq \mu_y$
- Use ugly formula to compute degrees-of-freedom

$$df = \frac{\left(s_x^2/n + s_y^2/m\right)^2}{\frac{(s_x^2/n)^2}{n-1} + \frac{\left(s_y^2/m\right)^2}{m-1}}$$

$S_x \to S_x^2$

$\mu_y - \mu_x \in [\ldots, \ldots]$

- Use test statistic

$$t_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \sim t_{df}$$

  and df to compute p-value

Assumptions: requires the difference in sample means $\bar{x} - \bar{y}$ to come from Normal distribution

- x_i and y_i came from normal distribution
- Both $n > 30$ and $m > 30$ (CLT)

Statistical testing for two samples: non-mathching samples

Means: compares the population means of $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$.

- Perform statistical test on the probabilities $H_0 : \mu_x = \mu_y$ vs. $H_a : \mu_x \neq \mu_y$
- Use "pooling" to approximate variance by

$$\sigma_x^2 = \sigma_y^2 \simeq s^2 \approx \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$

- Use test statistic

$$t_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{s^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

and $df = n + m - 2$ to compute p-value

Additional assumption: population variances are equal $\sigma_x = \sigma_y$

Two categorical variables

Titanic data set provides information on the fate of 891 passengers on the fatal maiden voyage of the ocean liner “Titanic”, summarized according to economic status (class), sex, age and survival.

PassengerId	Sex	Age	Class	Survived
1	male	22	3	no
2	female	38	1	yes
3	female	26	3	yes
4	female	35	1	yes
5	male	35	3	no
6	male	NA	3	no
7	male	54	1	no
8	male	2	3	no
9	female	27	3	yes
10	female	14	2	yes
11	female	4	3	yes
12	female	58	1	yes

Two categorical variables

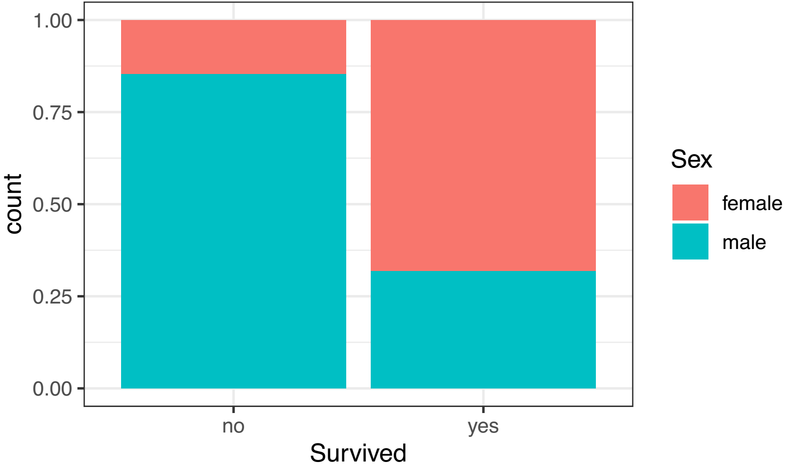
Is it true that women survived more often that men?

	no	yes	Sum
female	81	233	314
male	468	109	577
Sum	549	342	891

- **Marginal distribution** is the distribution of only one variable
- **Conditional distribution** is the distribution of one variable within a fixed value of a second value

Two categorical variables

Two variables are **independent** if conditional distribution of one variable is the same for all values of the other variable



Statistical testing for two categorical variables

Step 1: state your **null** hypothesis and the **alternative** hypothesis.

H_0 : sex and survived variables are independent

H_a : sex and survived variables are dependent

How would the table look like if null is true?

Statistical testing for two categorical variables

	no	yes	Sum
female	81	233	314
male	468	109	577
Sum	549	342	891

Multiply Both Sides by 891

If sex and survived variables are independent then (H_0)

$$\frac{81}{\cancel{891}} = P(no \cap female) = P(no) \cdot P(female) = \frac{549}{\cancel{891}} \cdot \frac{314}{891}$$

$$\frac{468}{891} = P(no \cap male) = P(no) \cdot P(male) = \frac{549}{891} \cdot \frac{577}{891}$$

$$= P(yes \cap female) = P(yes) \cdot P(female) =$$

$$= P(yes \cap male) = P(yes) \cdot P(male) =$$

Statistical testing for two categorical variables

	no	yes	Sum
female	81	233	314
male	468	109	577
Sum	549	342	891

Observed

expected

If sex and survived variables are independent then

observed counts = expected counts

$$81 = \#no\ and\ female = \frac{\#no \cdot \#female}{n} = \frac{549 \cdot 314}{891}$$

$$468 = \#no\ and\ male = \frac{\#no \cdot \#male}{n} = \frac{549 \cdot 577}{891}$$

$$223 = \#yes\ and\ female = \frac{\#yes \cdot \#female}{n} = \frac{342 \cdot 314}{891}$$

$$109 = \#yes\ and\ male = \frac{\#yes \cdot \#male}{n} = \frac{342 \cdot 577}{891} = 221$$

Statistical testing for two categorical variables

	no	yes	Sum
female	81	233	314
male	468	109	577
Sum	549	342	891

Step 2: summarize the data into a **test statistic**.

$$x^2_{obs} = \sum \frac{(observed - expected)^2}{expected} =$$

$$= \frac{(81 - *)^2}{*} + \frac{(468 - \dots)^2}{} +$$

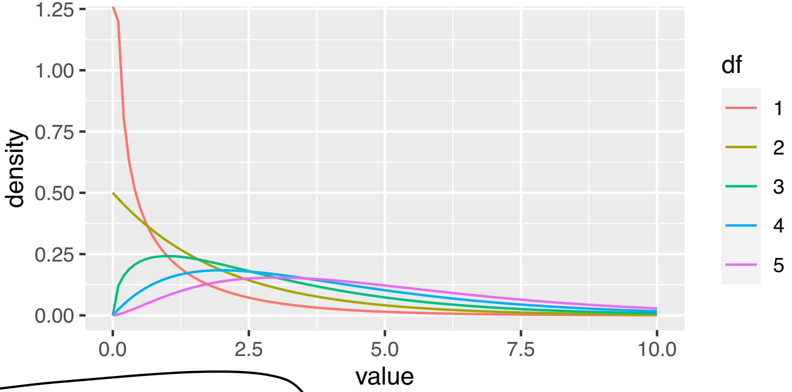
$$+ \frac{(233 - \dots)^2}{\dots} + \frac{(109 - 221)^2}{221} = \boxed{263}$$

Statistical testing for two categorical variables

t_{obs} , $T \sim t_{n-1}$

"Chi-square distribution with ... df"

Note that under the null, the test statistic $X^2 \sim \chi^2_{(r-1)(c-1)}$.



$\boxed{df = (2-1)\cdot(2-1) = 1}$ $\quad df = (r-1)\cdot(c-1)$

$df = (2-1)(3-1) = 2$ if Survived = Yes, No, Unkn.

Statistical testing for two categorical variables

263

Step 3: compute *p-value* $= P(X^2 > x^2_{obs})$ using the chi-square distribution table with $df = (2-1)(2-1)$.

Step 4: draw the conclusion. p-value < 0.05 $\Rightarrow$ reject H_0

```
chisq.test(x = sex, y = survived, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  sex and survived
## X-squared = 263.05, df = 1, p-value < 2.2e-16
```

Exercise

Perform statistical testing to check if there is association between Class and Survived variables.

	no	yes	Sum
1	80	136	216
2	97	87	184
3	372	119	491
Sum	549	342	891

$$\boxed{df = (3-1)(2-1) = 2}$$

Observed

80
136
97
87
372
119

Expected

$216 \cdot 549 / 891 \simeq 133$

$216 \cdot 342 / 891 \simeq 83$

$$\chi^2_{obs} = \frac{(80-133)^2}{133} + \frac{(136-83)^2}{83} + \dots$$

Exercise

```
chisq.test(x = class, y = survived, correct = FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  class and survived
## X-squared = 102.89, df = 2, p-value < 2.2e-16
```

Two quantitative variables

In Pearson's data set there are 1078 measurements of a father's height and his son's height.

fheight	sheight
65.04851	59.77827
63.25094	63.21404
64.95532	63.34242
65.75250	62.79238
61.13723	64.28113
63.02254	64.24221
65.37053	64.08231
64.72398	63.99574
66.06509	64.61338
66.96738	63.97944
59.00800	65.24451
62.93203	65.35102
63.67063	65.67992
64.07386	65.43664

Two quantitative variables

To quantify the relationship between quantitative $x_1, , ..., x_n$ and $y_1, ..., y_n$ we introduced **correlation coefficient.**

$$correlation = \frac{cov_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = r_{xy}$$

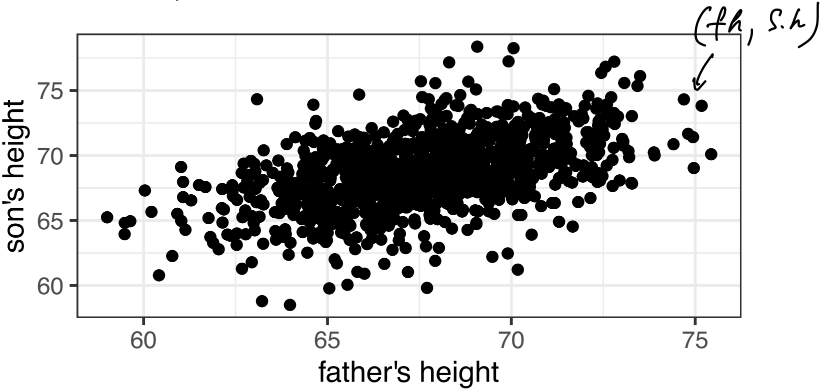
- Correlation value is between -1 and 1
- Positive correlation $\Rightarrow$ the variables tend to both increase together
- Negative correlation $\Rightarrow$ one variable tends to increase when the other decreases

```
cor(fheight, sheight)
```

```
## [1] 0.5013383
```

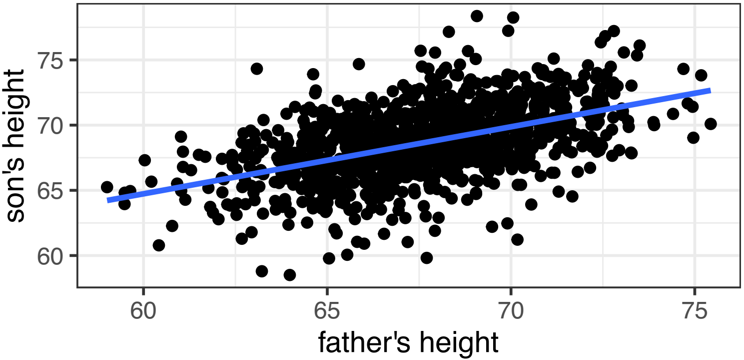
Two quantitative variables

The "trend" is positive!



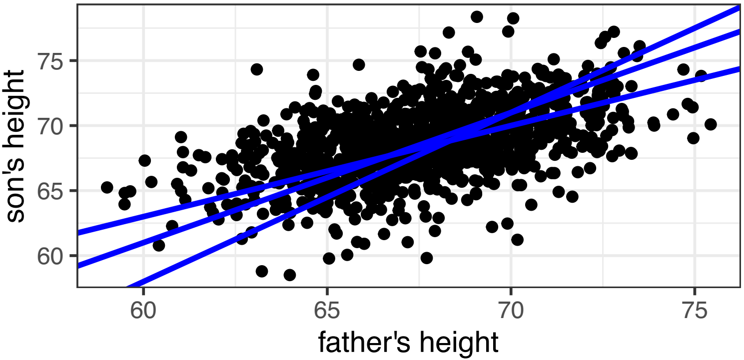
Linear regression

Goal: find the line that approximates the best the relationship between two variables.



Linear regression

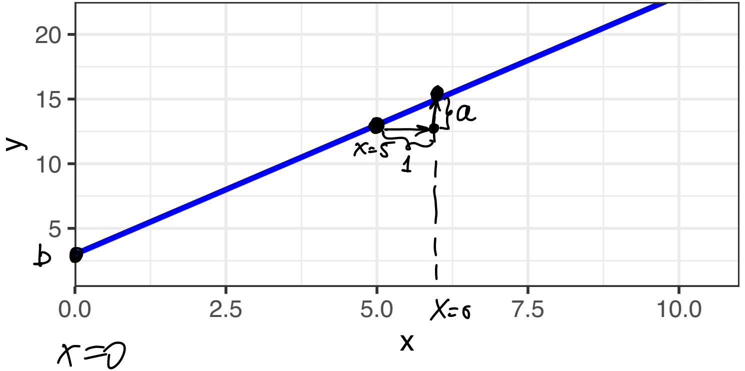
Which line is better?



Line equation

Any line can be written as $y = ax + b$.

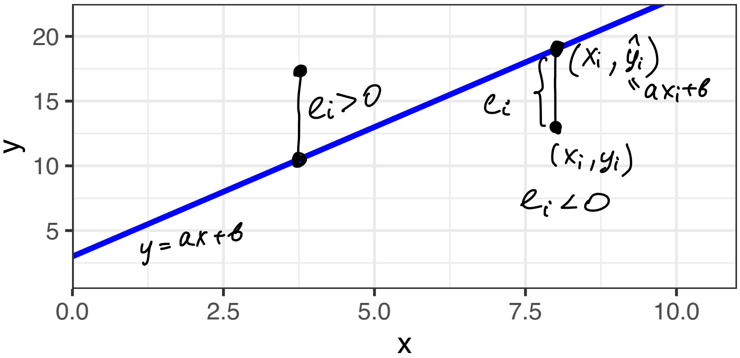
- a is **slope**, the value of y when $x = 0$ ← $y = a \cdot 0 + b = b$
- b is **intercept**, the change in y when x changes by 1 unit



Linear regression

Given a point (x_i, y_i)

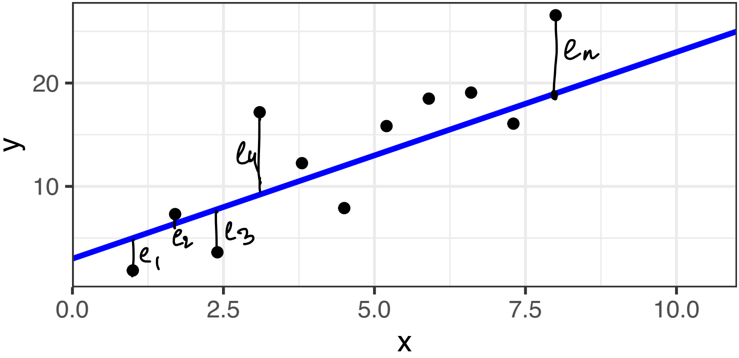
- vertical projection of the point on the line is $\hat{y}_i = ax_i + b$
- **residual** $e_i = y_i - \hat{y}_i$ measures how well the line approximates the point



Linear regression

Given a set of points $(x_1, y_1), \ldots, (x_n, y_n)$

- **residual sum of squares** $RSS = \sum_{i=1}^{n} e_i^2$, measures how well the line approximates the data



Linear regression

$$\sum_{i=1}^{n} y_i - a \cdot \sum_{i=1}^{n} x_i - n \cdot b = 0 \Rightarrow \underbrace{\frac{\sum y_i}{n}}_{\bar{y}} - a \underbrace{\frac{\sum x_i}{n}}_{\bar{x}} = b$$

Note that for different a and b we will get different RSS

$\hat{y}_i = ax_i + b$

unknown

$$RSS(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2 \longrightarrow \text{min w.r.t. } a, b$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

so we want to find a and b that minimize RSS.

cov

$$\boxed{\begin{aligned} a &= \frac{cov_{xy}}{s_x^2} = \frac{s_y}{s_x} r_{xy} \\ b &= \bar{y} - a\bar{x} \end{aligned}}$$

$$\begin{cases} \dfrac{d\, RSS(a,b)}{d a} = 0 \\ \dfrac{d\, RSS(a,b)}{d b} = 0 \end{cases}$$

$$\frac{d\, RSS(a,b)}{d b} = \sum_{i=1}^{n} \frac{d}{db} (y_i - ax_i - b)^2 = \sum_{i=1}^{n} (-2)(y_i - ax_i - b) =$$

$$= -2 \sum_{i=1}^{n} (y_i - ax_i - b) = 0$$

$$\sum_{i=1}^{n} y_i - a \cdot \sum_{i=1}^{n} x_i - n \cdot b = 0$$

Exercise

Find the regression line for Fisher's data set.

son = a · father + b
y x

```
c(mean(fheight), sd(fheight))
```

```
## [1] 67.687097  2.744868
```

$a = \boxed{\frac{2.8}{2.7} \cdot 0.5} \simeq 0.51$

$b = 68.7 - 0.51 \cdot 67.7$
$= 33$

```
c(mean(sheight), sd(sheight))
```

```
## [1] 68.684070  2.814702
```

```
cor(fheight, sheight)
```

```
## [1] 0.5013383
```

$$\boxed{\begin{array}{c} a = \dfrac{cov_{xy}}{s_x^2} = \dfrac{s_y}{s_x} r_{xy} \\ b = \bar{y} - a\bar{x} \end{array}}$$

cor

Exercise

y x

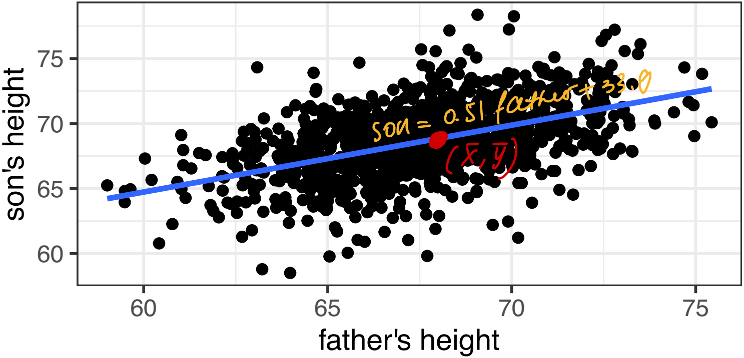
```
lm(sheight~fheight)
```

```
##
## Call:
## lm(formula = sheight ~ fheight)
##
## Coefficients:
## (Intercept)      fheight
##     33.8866       0.5141
```

$$\boxed{\text{Son} = 0.51 \cdot \text{father} + 33.9}$$

Linear regression: properties

- The line passes through $(\bar{x}, \bar{y})$ 67 68



Linear regression: properties

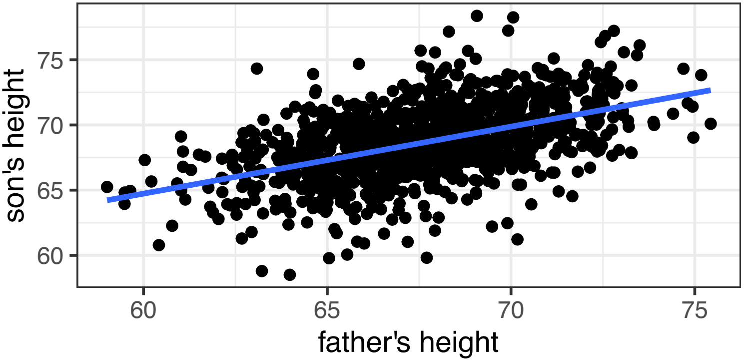
It is important which variable is x which is y.

In line equation $y = a \cdot x + b$

- y is called **response variable**
- x is called **explanatory variable**

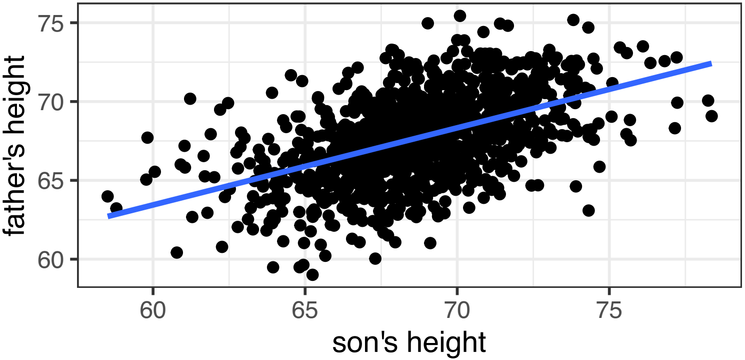
Linear regression: properties

$$son = 0.51 \cdot father + 33.89$$



Linear regression: properties

$$father = 0.49 \cdot son + 34.11$$



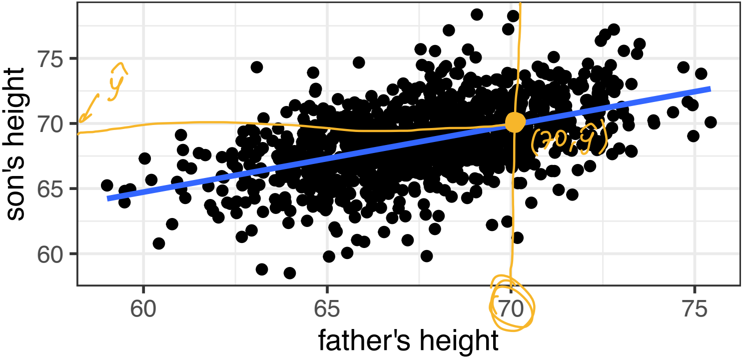
Linear regression: properties

Regression line is often used for prediction, $\hat{y}$ is often called predicted value.

What would be the son's height if father was 70 inch exactly?

70 $\qquad \hat{y} = a \cdot x_{new} + b$

$$son = 0.51 \cdot father + 33.89$$



Linear regression: properties

Regression coefficients have interpretation

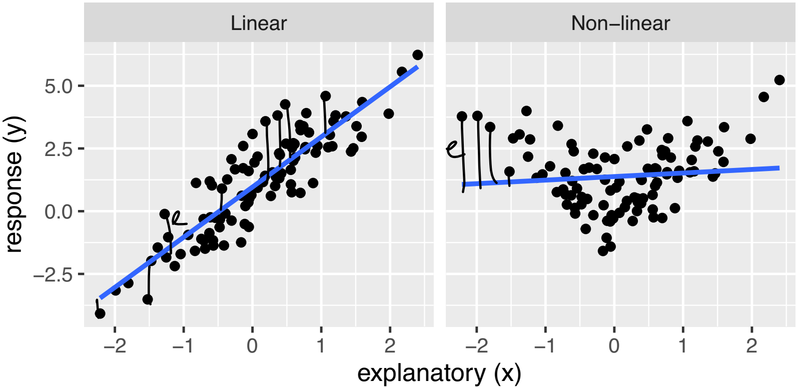
- b $\cancel{a}$ is the average value of y when $x = 0$ (if zero values make sense)
- a $\cancel{b}$ is the average change in y when x changes by 1 unit

$$son = \underset{a}{0.51} \cdot father + 33.89$$

If father's height increases by 1 inch
Son's height increases by 0.51 inch

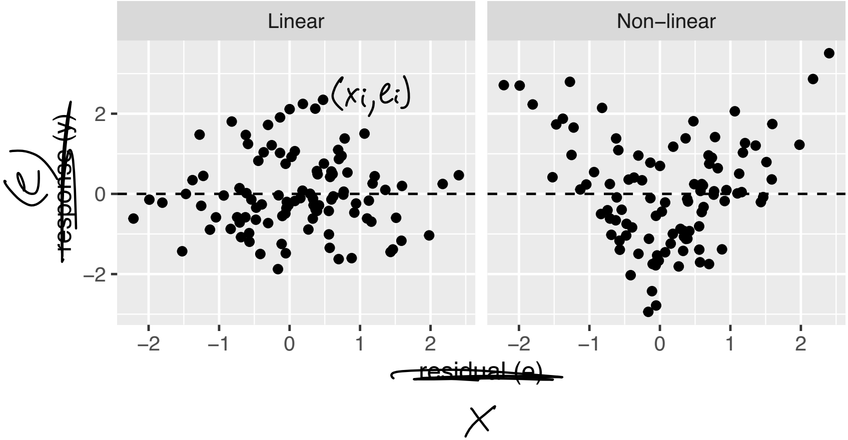
Linear regression: properties

Regression works great when the "trend" is linear.



Linear regression: properties

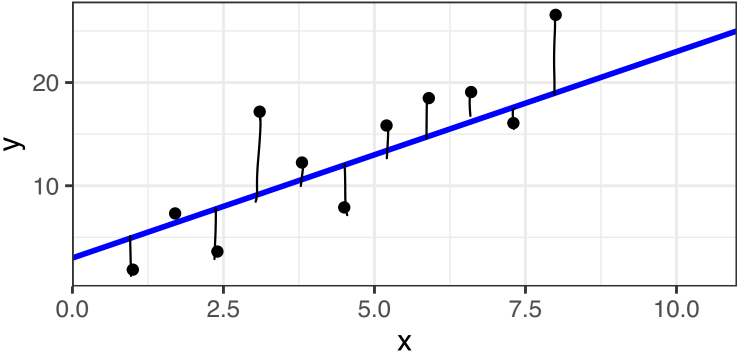
The **residual plot** will show whether a straight line is a good model for the data.



Linear regression: coefficient of determination

How to measure if line approximation is accurate?

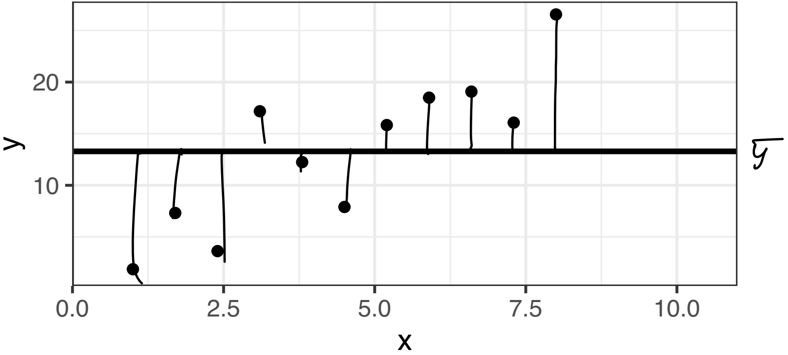
Residual sum of squares $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ measures how well the regression line approximates the data. BUT RSS dependents on the data scale.



Linear regression: coefficient of determination

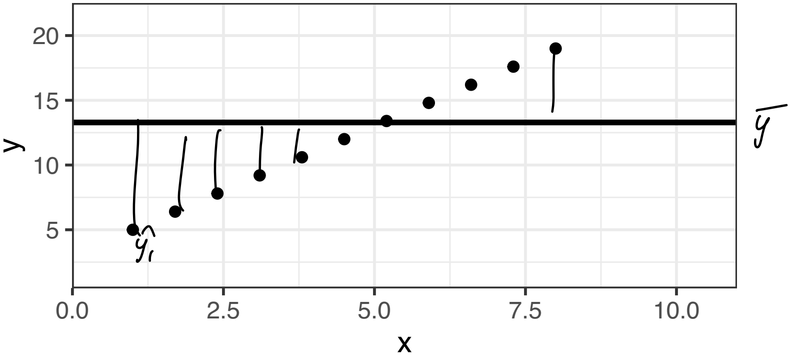
Total sum of squares $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ measures variation in the response variable.

$(n-1) \cdot S_y^2$



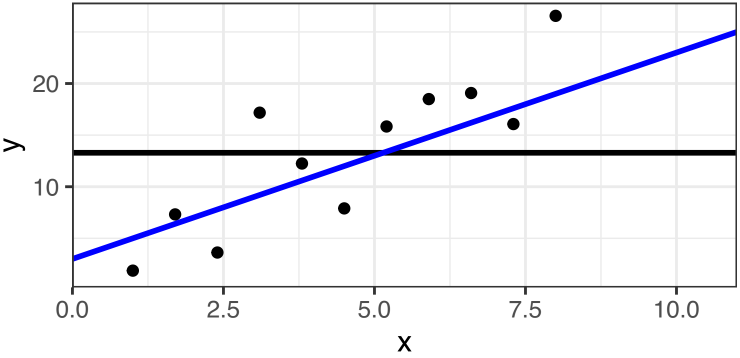
Linear regression: coefficient of determination

Explained sum of squares $ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ measures variation in the response variable that can be explained by the regression line.



Linear regression: coefficient of determination

$$TSS = ESS + RSS$$



Linear regression: coefficient of determination

Coefficient of determination measures the **proportion** of variation in response variable explained by the regression line.

$$R^2 = \boxed{\frac{ESS}{TSS}} = 1 - \frac{RSS}{TSS}$$

- $R^2 = 0$ none of the variation is explained (very bad fit)
- $R^2 = 1$ all of the variation is explained (perfect fit)

Cool fact: coefficient of determination is equal to the squared correlation between the response and explanatory variable (and prediction)

$$R^2 = cor^2(x, y) = cor^2(\hat{y}, y)$$

Exercise

Given RSS

```
sum((lm(sheight~fheight)$residuals)^2)
```

$TSS = \Sigma(y_i - \bar{y})^2$

$TSS = (n-1)S_y^2$

$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

```
## [1] 6388.001
```

standard deviation of sons height

$\boxed{R^2 = 0.25}$

```
sd(sheight)
```

```
## [1] 2.814702
```
$= S_y$

$TSS = 1077 \cdot (2.81)^2$

and $\boxed{n = 1078}$, find the coefficient of determination.

$R^2 = 1 - \frac{6388}{\ldots}$

TO DO

1. Module 11. Simple Linear Regression
2. Quiz 12 due Monday (April 10) @ 11:59 PM (EST)
3. Practice Problem Set 12