Week 11: Correlation Analysis

EMSE 4197 | John Paul Helveston | March 25, 2020

Visualizing coronavirus

Choropleth maps

- World, Robinson projection [NY Times]
- US, Albers projection [Newsweek]
- US animation, Albers projection [NY Times]

Bubble maps

- US, Albers projection [NY Times]
- US animation, Albers projection [NY Times]

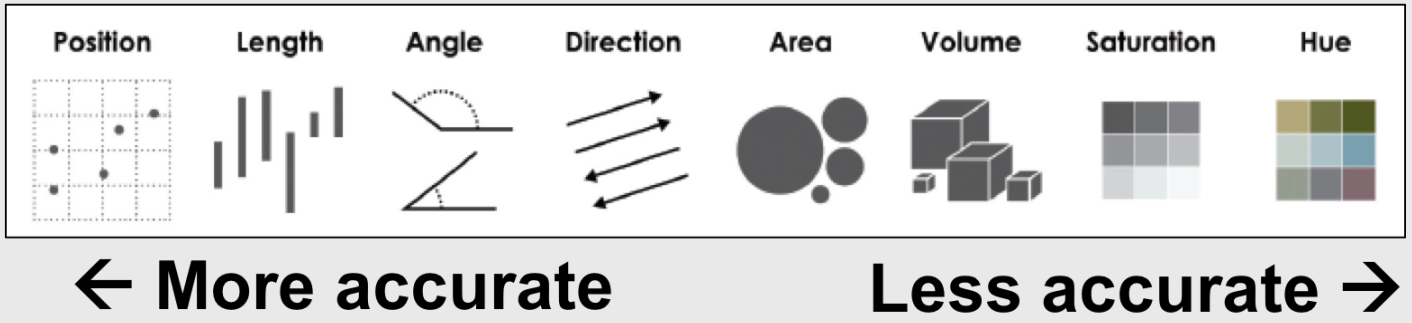
Make good design choices

Static charts

- The Hammer and the Dance, by Tomas Pueyo
- Heatmap, [NPR]
- Log line charts, [Financial Times]

Dashboards

- ArcGIS dashboard, Center for Systems Science and Engineering (CSSE), Johns Hopkins U.
- R Shiny dashboard, Christoph Schönenberger



General proposal feedback

- Your research question should be a question (e.g. "what...?", "why...?", "how...?").
- Your projects are *exploratory*: your goal isn't to "prove" something true or false, but rather to identify evidence that *generates* hypotheses.
- Testing hypotheses is "confirmatory" analysis, which we won't cover in this class.

Today's data

```
wildlife_impacts <- read_csv(here::here('data', 'wildlife_impacts.csv'))
msleep           <- read_csv(here::here('data', 'msleep.csv'))
```

New package:

```
install.packages('HistData')
install.packages('GGally')
```

Correlation Analysis

1. What is correlation?
2. Visualizing correlation
3. Linear models
4. Visualizing linear models

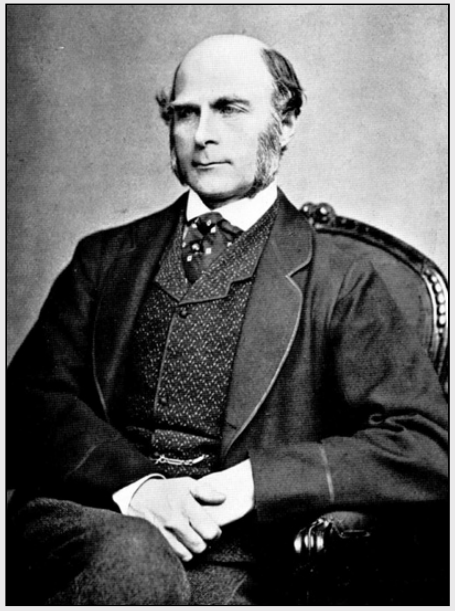
Correlation Analysis

1. What is correlation?
2. Visualizing correlation
3. Linear models
4. Visualizing linear models

Origins in Eugenics ("Well Born")

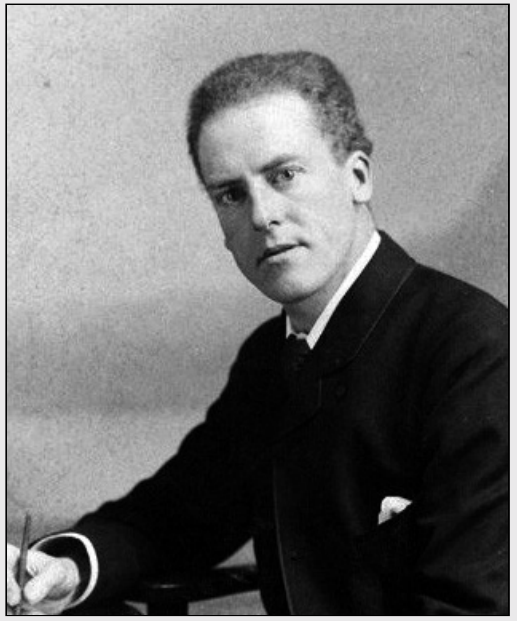
Sir Francis Galton (1822 - 1911)

- Charles Darwin's cousin.
- "Father" of Eugenics.
- Studied correlations between parents & children.



Karl Pearson (1857 - 1936)

- Galton's protégé (to the verge of hero worship).
- Developed equation for correlation coefficient.
- "Father" of mathematical statistics.



Galton's family data

Galton, F. (1886). "Regression towards mediocrity in hereditary stature". *The Journal of the Anthropological Institute of Great Britain and Ireland* 15: 246-263.

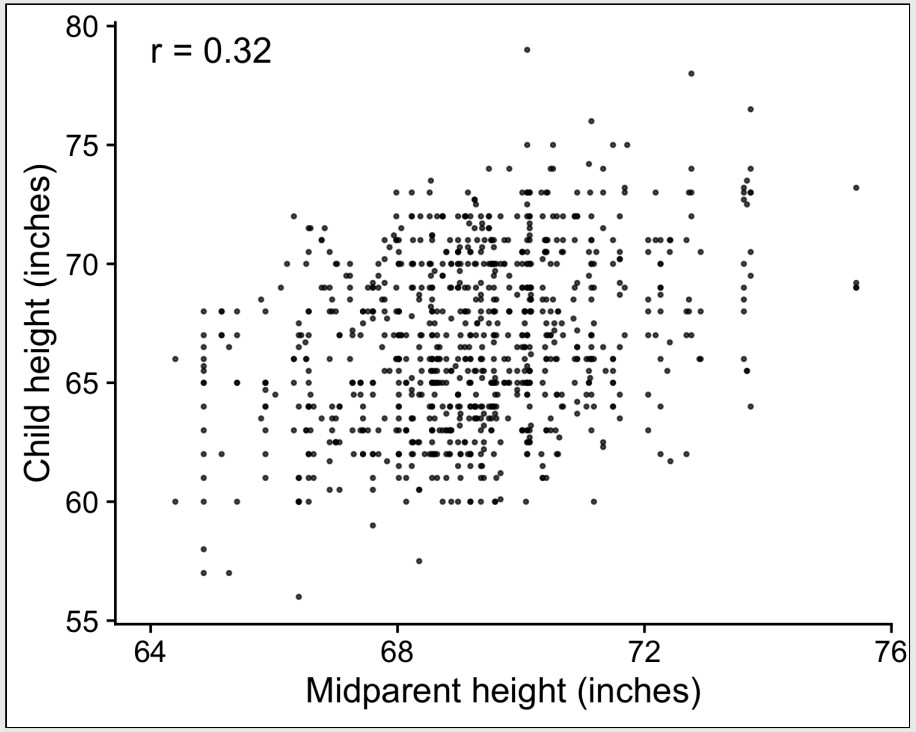
Galton's research question: Does marriage selection indicate a relationship between the heights of husbands and wives?
(He called this "assortative mating")

Btw, "midparent height" is just a scaled average:

```
midparentHeight =  (father + 1.08*mother)/2
```

```
library(HistData)

galtonScatterplot <- ggplot(GaltonFamilies) +
    geom_point(aes(x = midparentHeight,
                   y = childHeight),
               size = 0.5, alpha = 0.7) +
    theme_half_open() +
    labs(x = 'Midparent height (inches)',
         y = 'Child height (inches)')
```

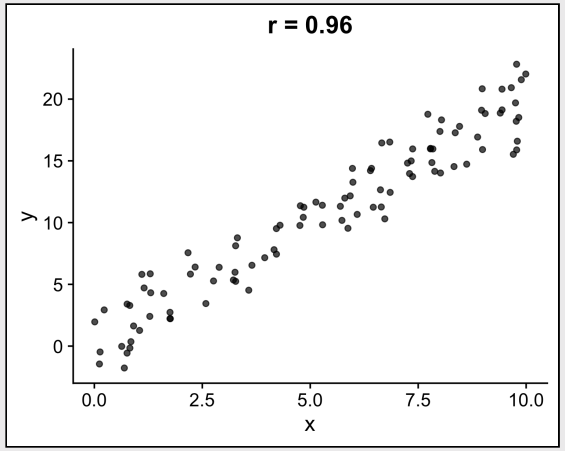


How do you measure correlation?

$$\text{Pearson: } r = \frac{\text{Cov}(x,y)}{\text{sd}(x) * \text{sd}(y)}$$

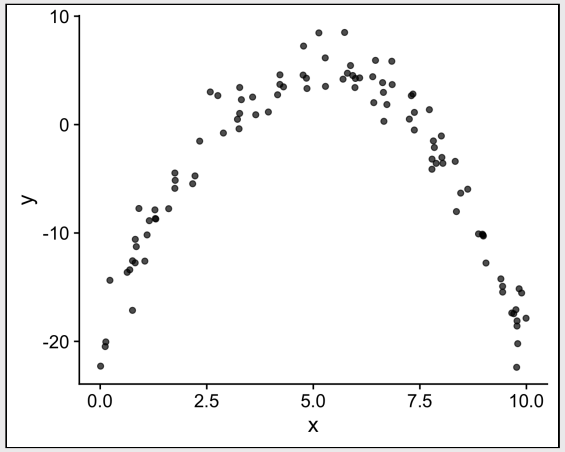
How do you measure correlation?

$$r = \frac{\mathrm{Cov}(x,y)}{\mathrm{sd}(x)*\mathrm{sd}(y)}$$



Assumptions:

1. Variables must be interval or ratio
2. Linear relationship

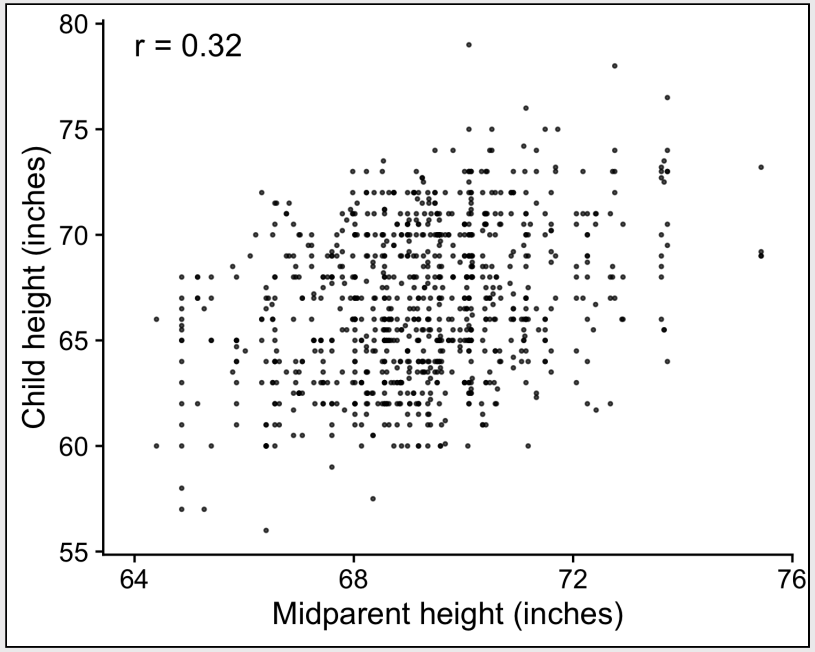


How do you *interpret* r?

$$r = \frac{\text{Cov}(x,y)}{\text{sd}(x)*\text{sd}(y)}$$

Interpretation:

- $-1 \leq r \leq 1$
- Closer to 1 is stronger correlation
- Closer to 0 is weaker correlation

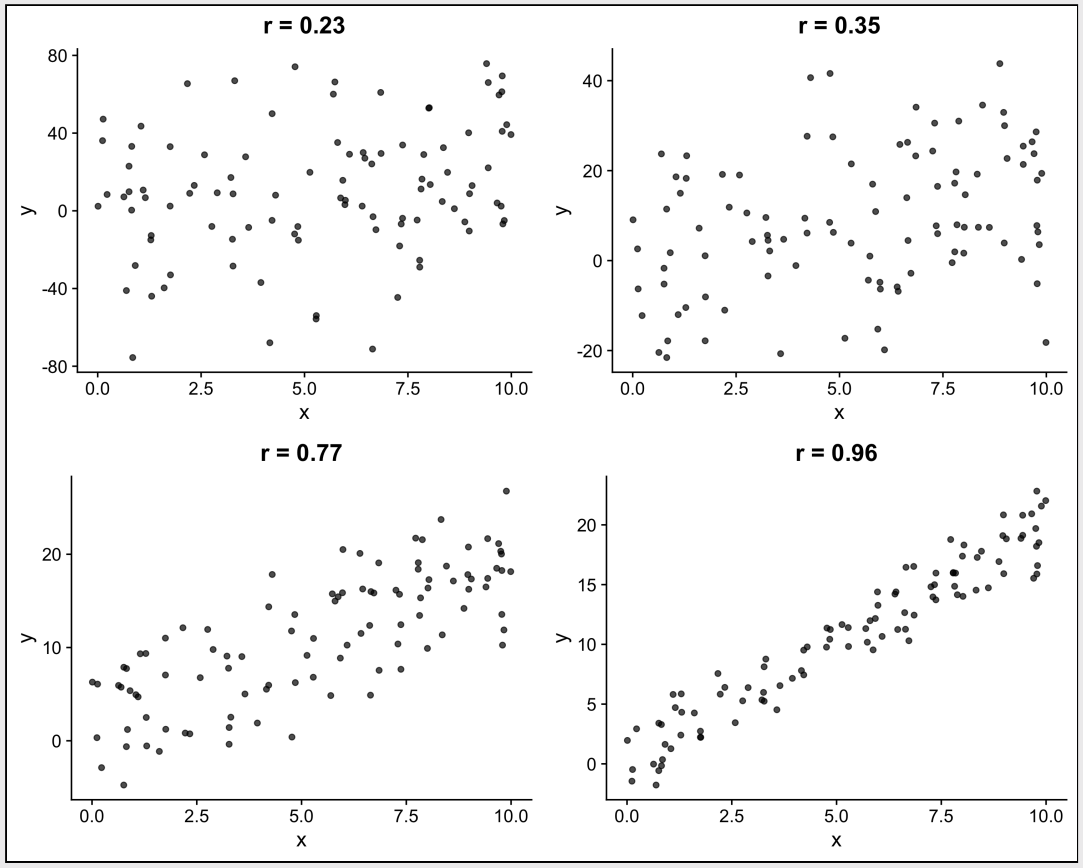



```
round(cor(
    x = GaltonFamilies$midparentHeight,
    y = GaltonFamilies$childHeight,
    method = 'pearson'), 2)
```

```
## [1] 0.32
```

What does r mean?

- $\pm 0.1 - 0.3$: Weak
- $\pm 0.3 - 0.5$: Moderate
- $\pm 0.5 - 0.8$: Strong
- $\pm 0.8 - 1.0$: Very strong

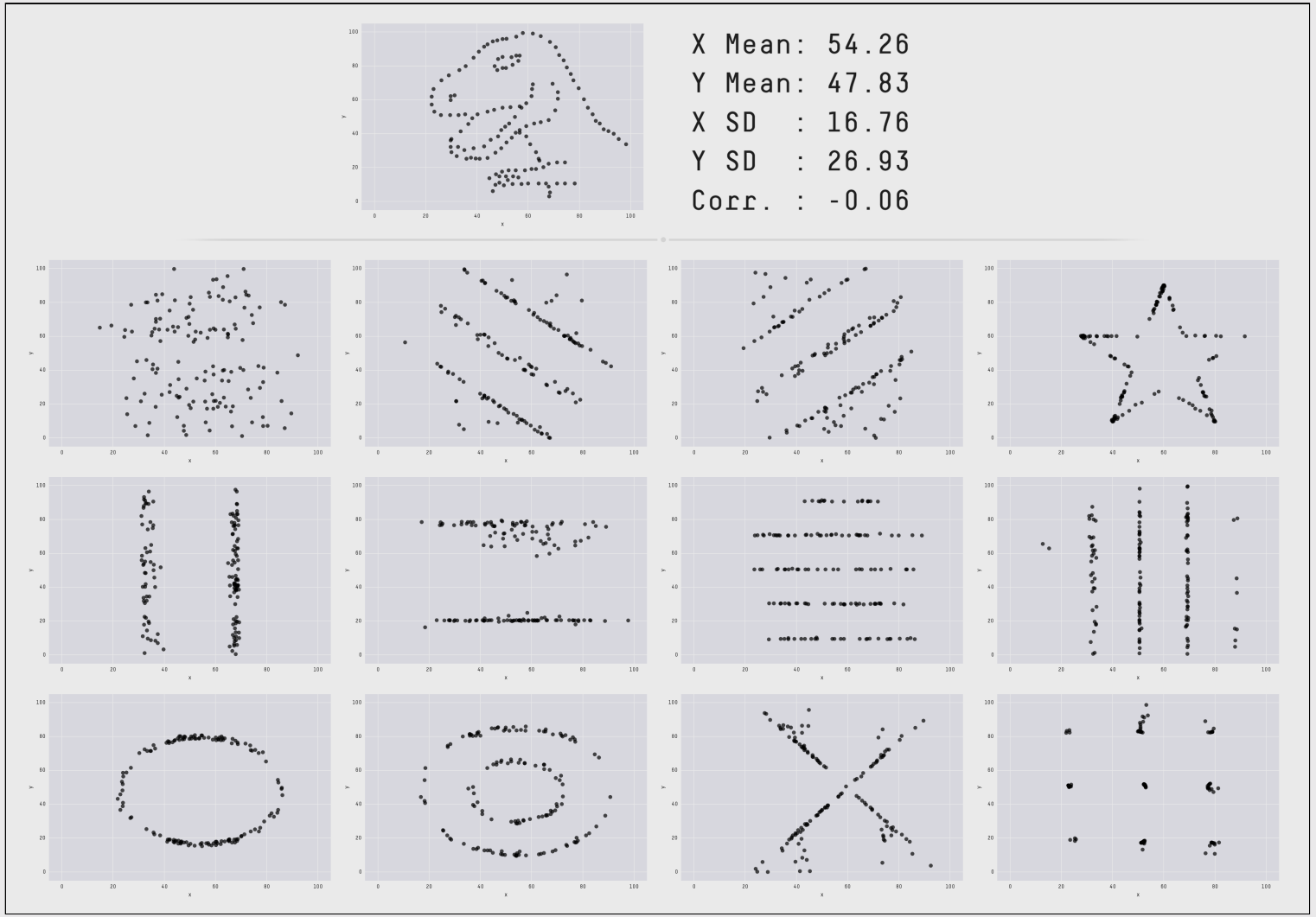


Visualizing correlation is...um...easy!

guessthecorrelation.com

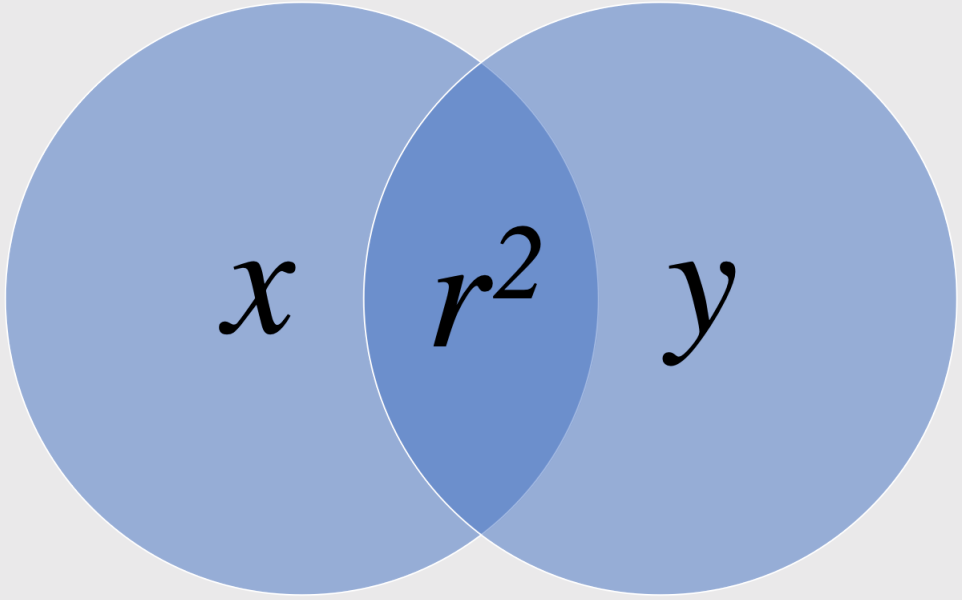
The datasaurus

See more here



Coefficient of determination: r^2

Percent of variance in one variable
that is explained by the other variable



r	r^2
0.1	0.01
0.2	0.04
0.3	0.09
0.4	0.16
0.5	0.25
0.6	0.36
0.7	0.49
0.8	0.64
0.9	0.81
1.0	1.00

You should report both r and r^2

Correlation between parent and child height is 0.32, therefore 10% of the variance in the child height is explained by the parent height.

Correlation != Causation

X causes Y

- Training causes improved race performance

Y causes X

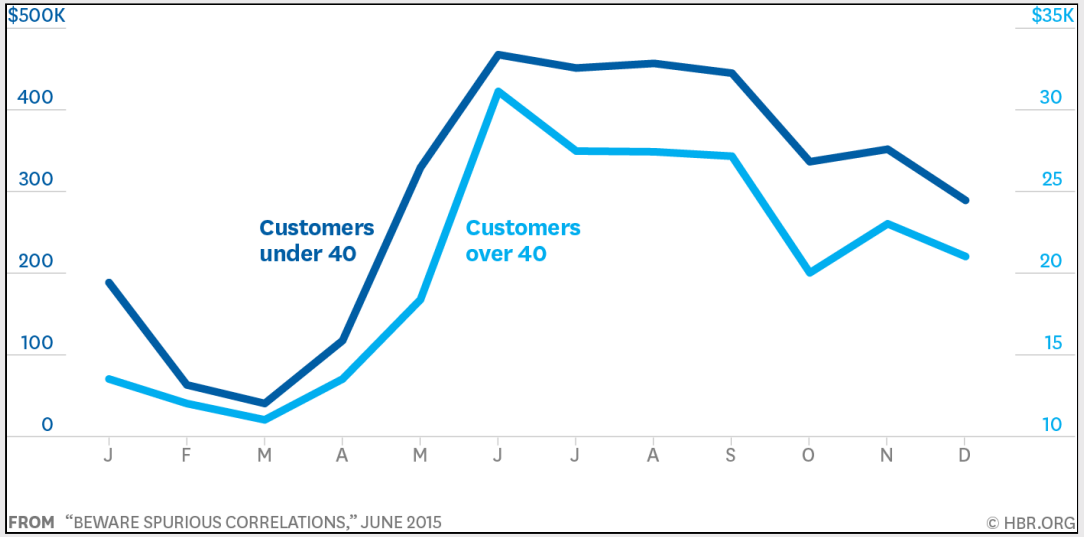
- Race performance causes people to train harder.

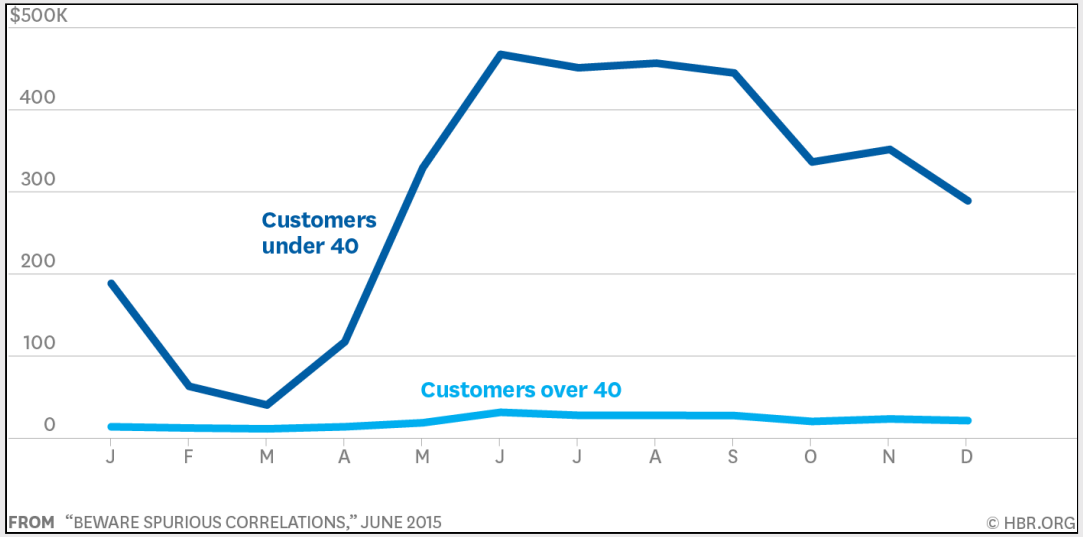
Z causes both X & Y

- Commitment and motivation cause increased training and better race performance.

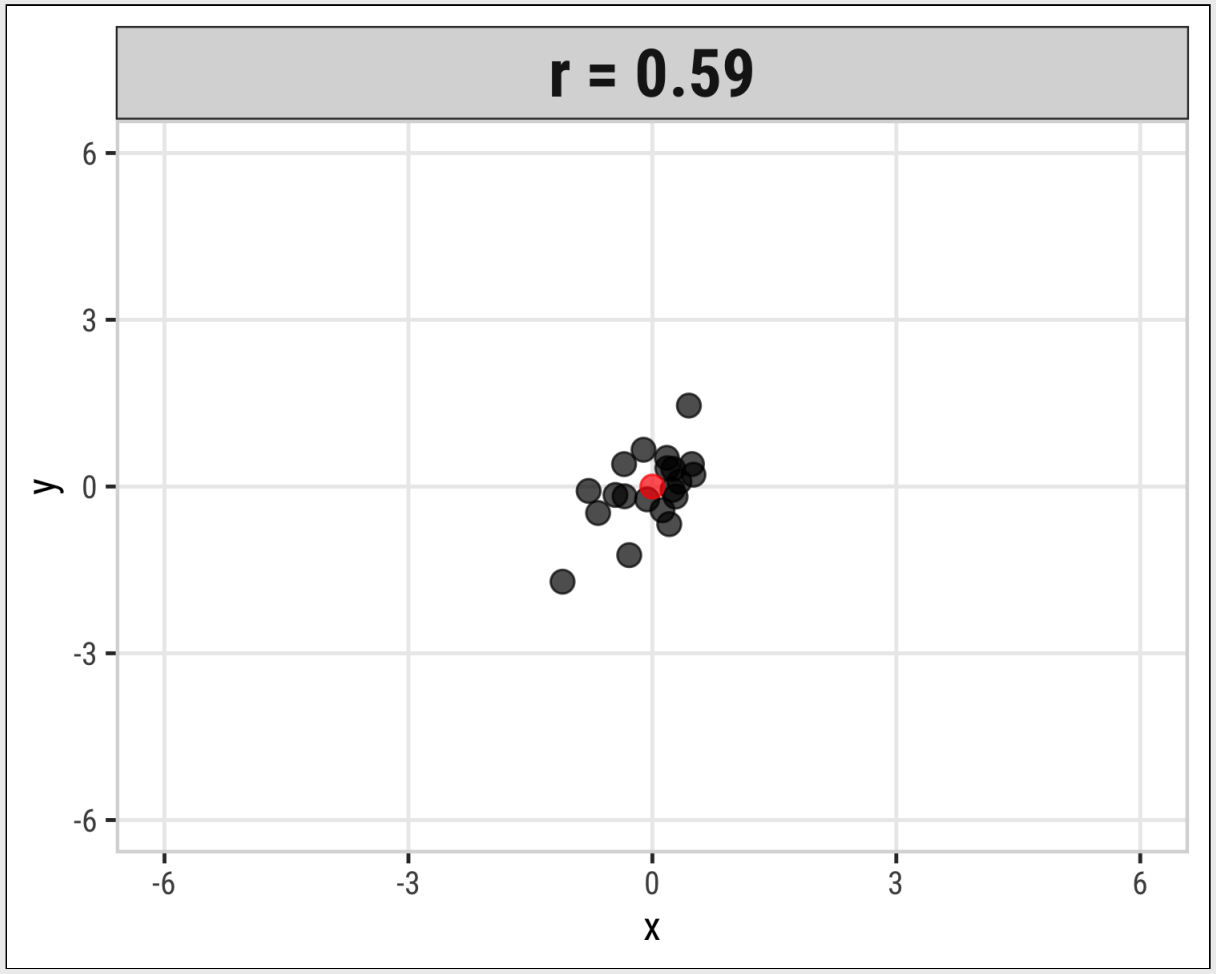
Be weary of dual axes!

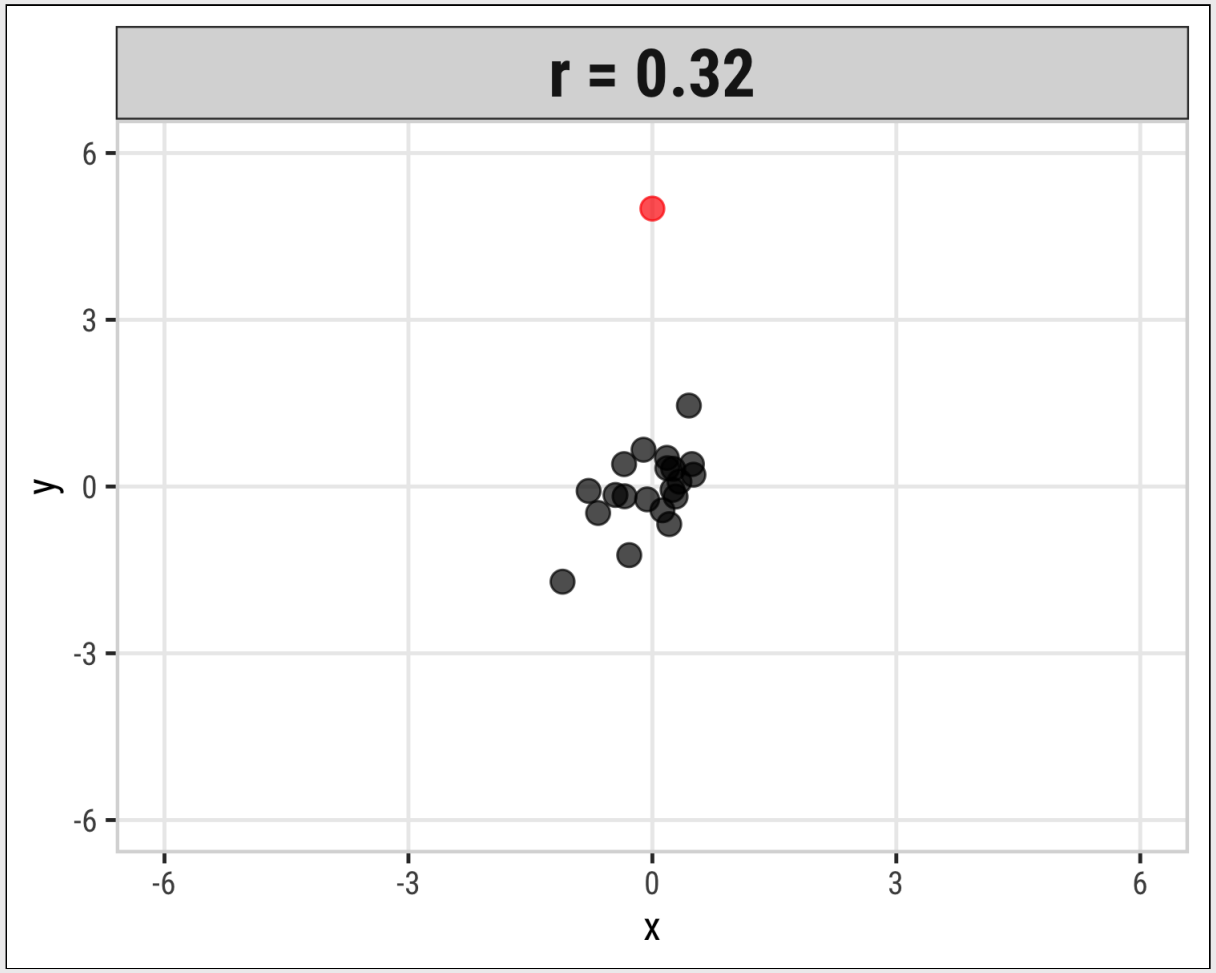
They can cause spurious correlations

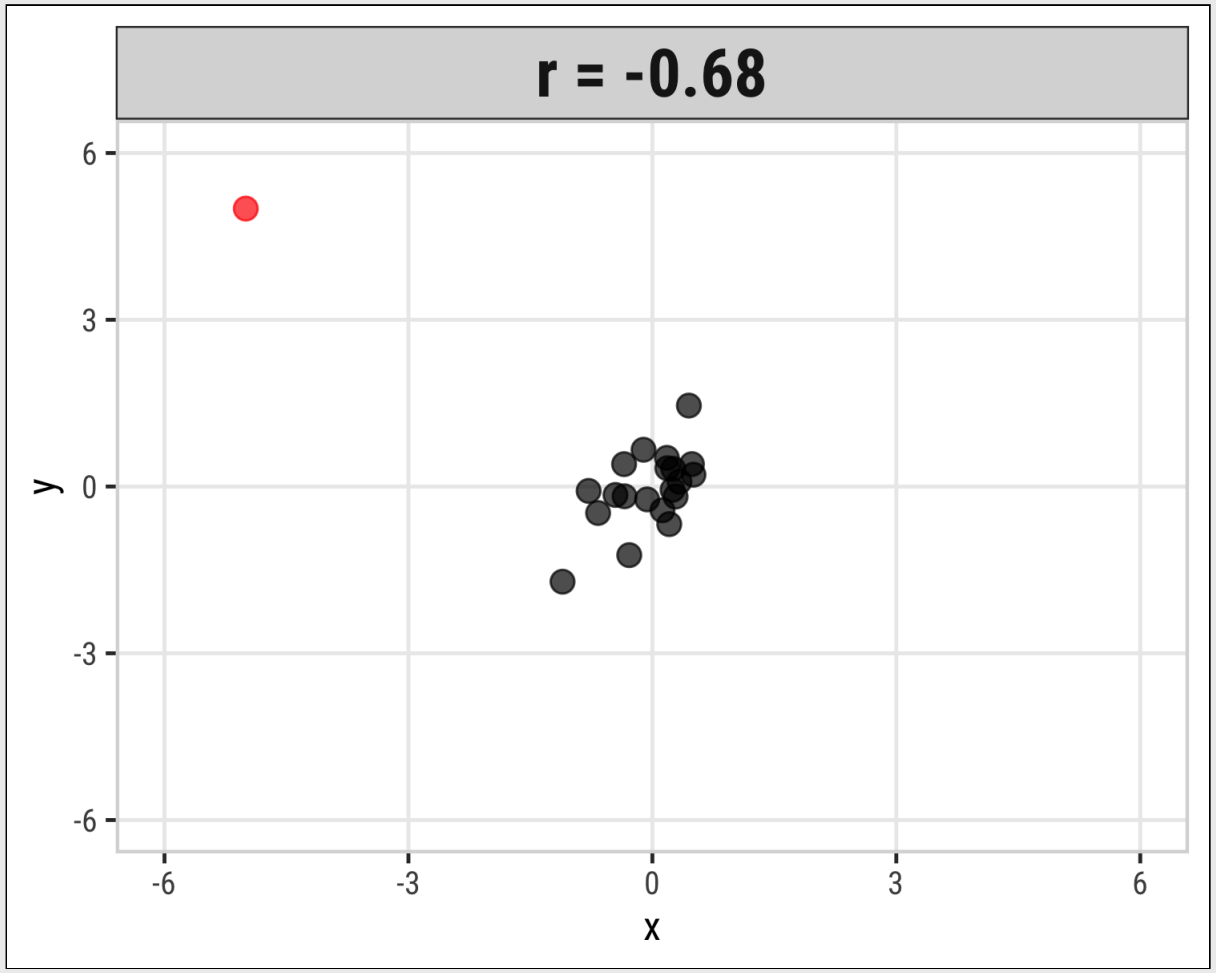




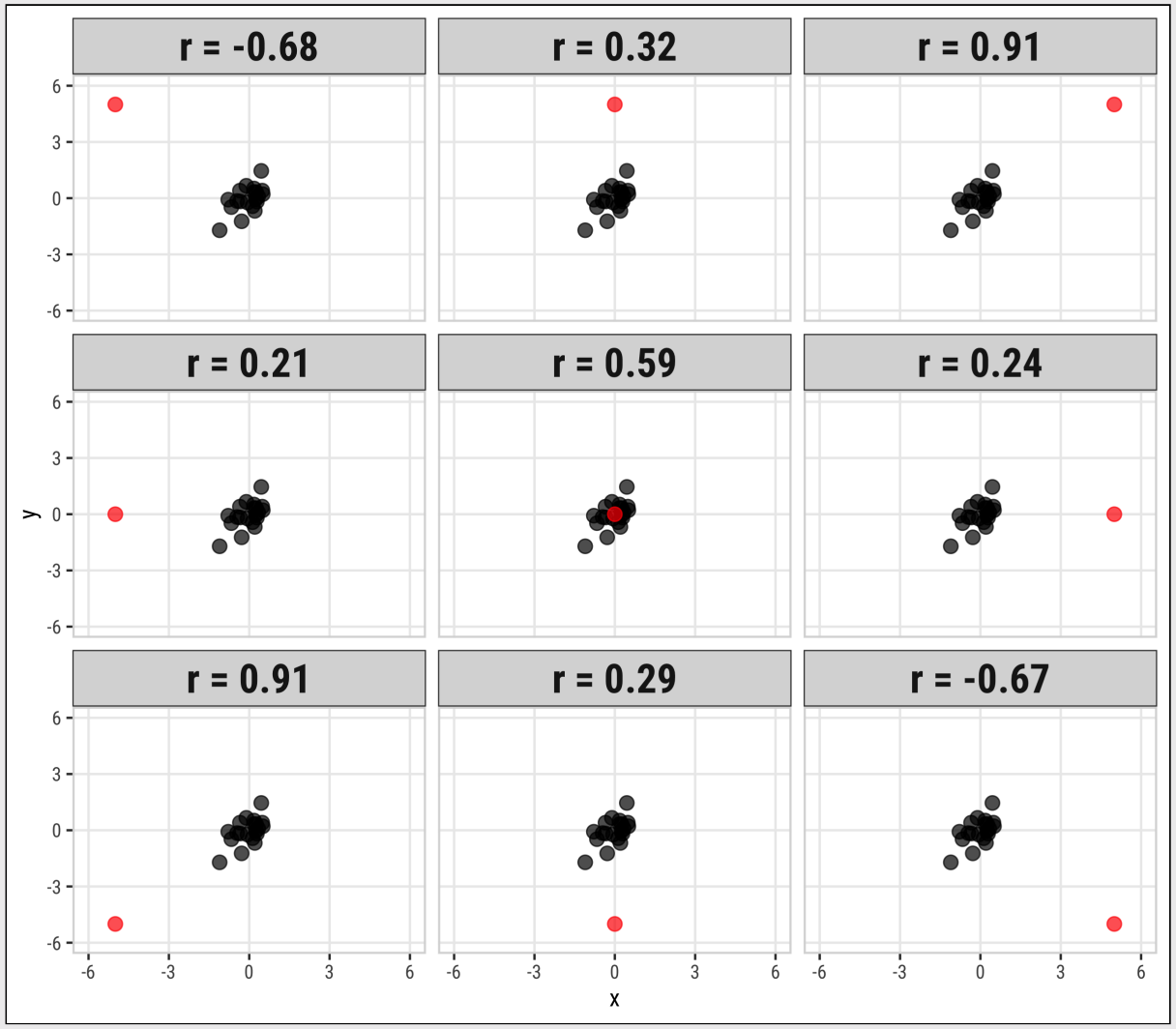
Outliers


r = 0.59
6
3
0
-3
-6
y
-6
-3
0
3
6
x


r = 0.32
6
3
0
-3
-6
y
-6
-3
0
3
6
x


r = -0.68
6
3
0
-3
-6
y
-6
-3
0
3
6
x

Pearson correlation is highly sensitive to outliers



Spearman's rank-order correlation

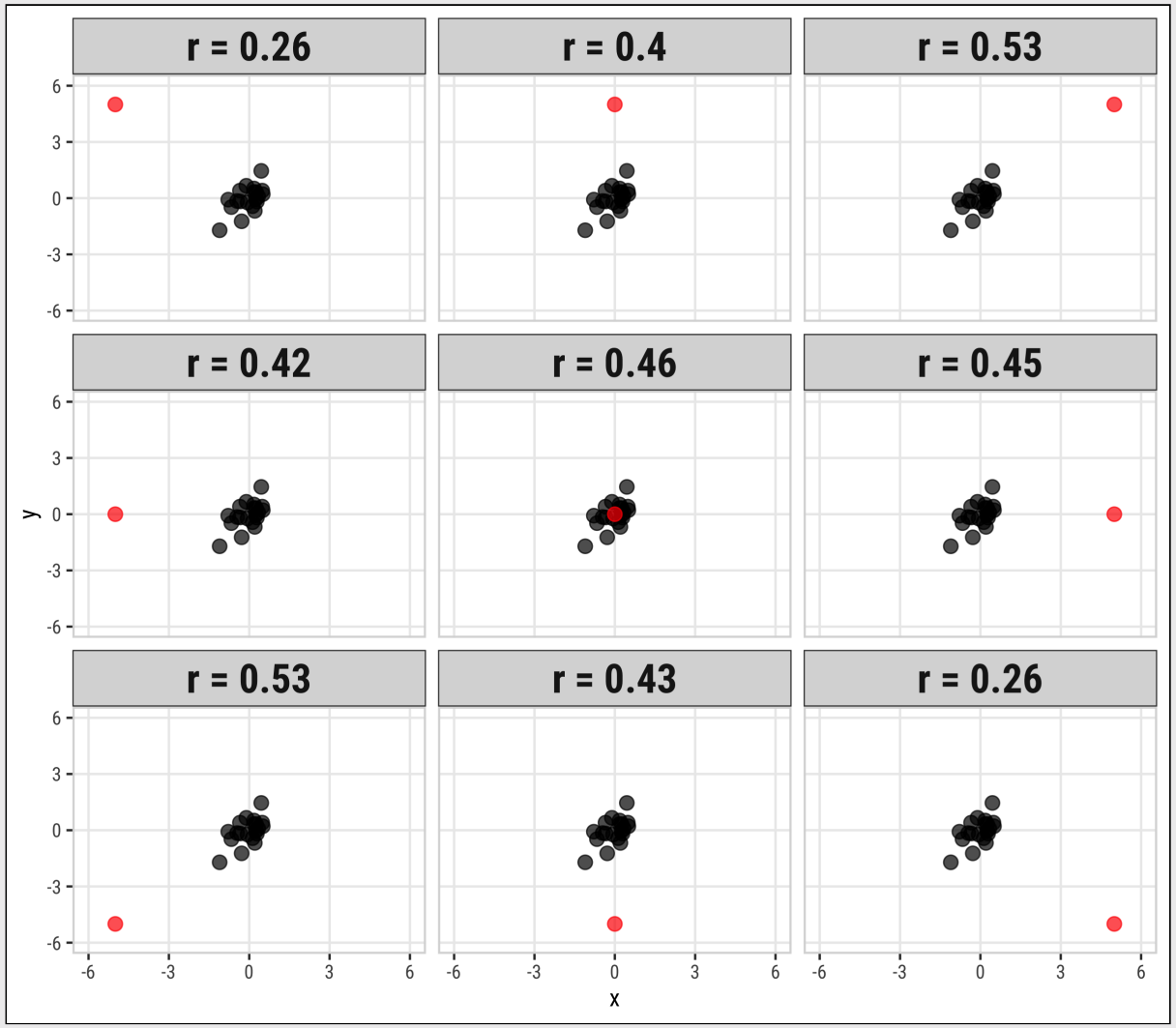
$$r = \frac{\mathrm{Cov}(x,y)}{\mathrm{sd}(x)*\mathrm{sd}(y)}$$

- Separately rank the values of X & Y.
- Use Pearson's correlation on the *ranks* instead of the x & y values.

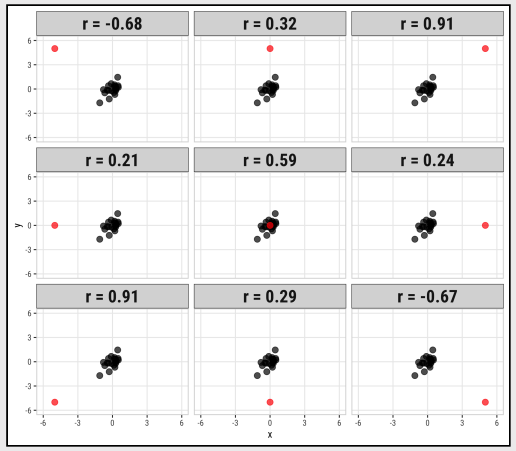
Assumptions:

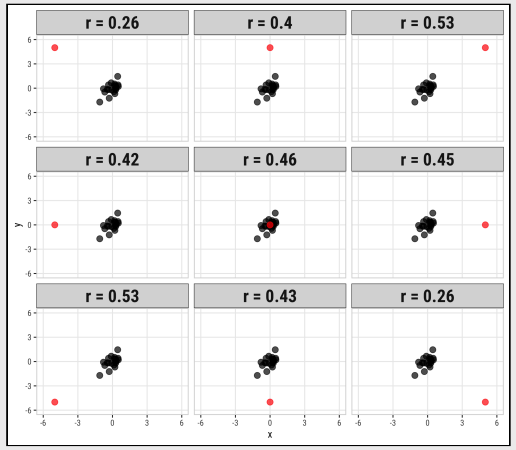
- Variables can be ordinal, interval or ratio
- Relationship must be monotonic (i.e. does not require linearity)

Spearman correlation more robust to outliers

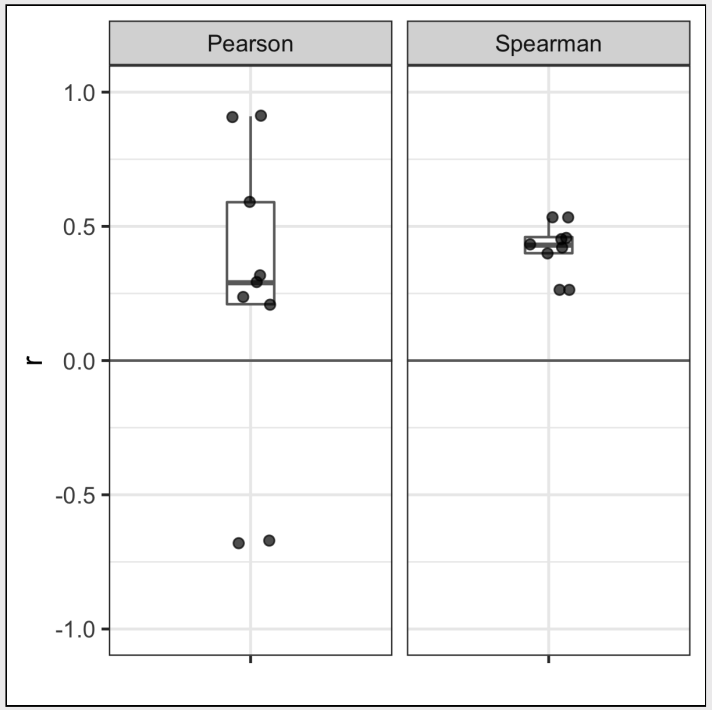


Spearman correlation more robust to outliers





Pearson	Spearman
-0.56	0.53
0.39	0.69
0.94	0.81
0.38	0.76
0.81	0.79
0.31	0.70
0.95	0.81
0.51	0.75
-0.56	0.53



Summary of correlation

- **Pearon's correlation**: Described the strength of a **linear** relationship between two variables that are interval or ratio in nature.
- **Spearman's rank-order correlation**: Describes the strength of a **monotonic** relationship between two variables that are ordinal, interval, or ratio. **It is more robust to outliers**.
- The **coefficient of determination** describes the amount of variance in one variable that is explained by the other variable.
- Correlation != Causation

R command (hint: add `use = "complete.obs"` to drop NA values)

```
pearson  <- cor(x, y, method = "pearson", use = "complete.obs")
spearman <- cor(x, y, method = "spearman", use = "complete.obs")
```

Correlation Analysis

1. What is correlation?
2. Visualizing correlation
3. Linear models
4. Visualizing linear models

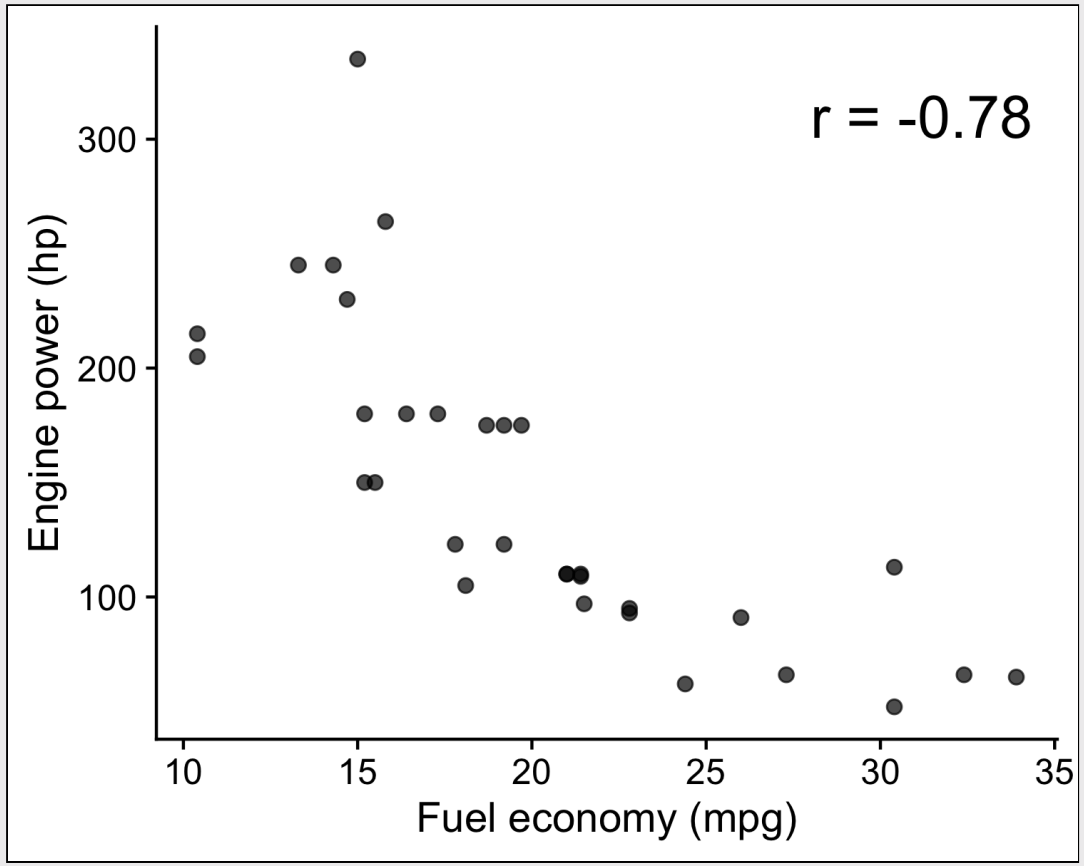
Visualizing correlations: the scatterplot

Compute the correlation

```
mtcarsCorr <- round(cor(
  mtcars$mpg, mtcars$hp,
  method = 'pearson'), 2)
```

Make the plot

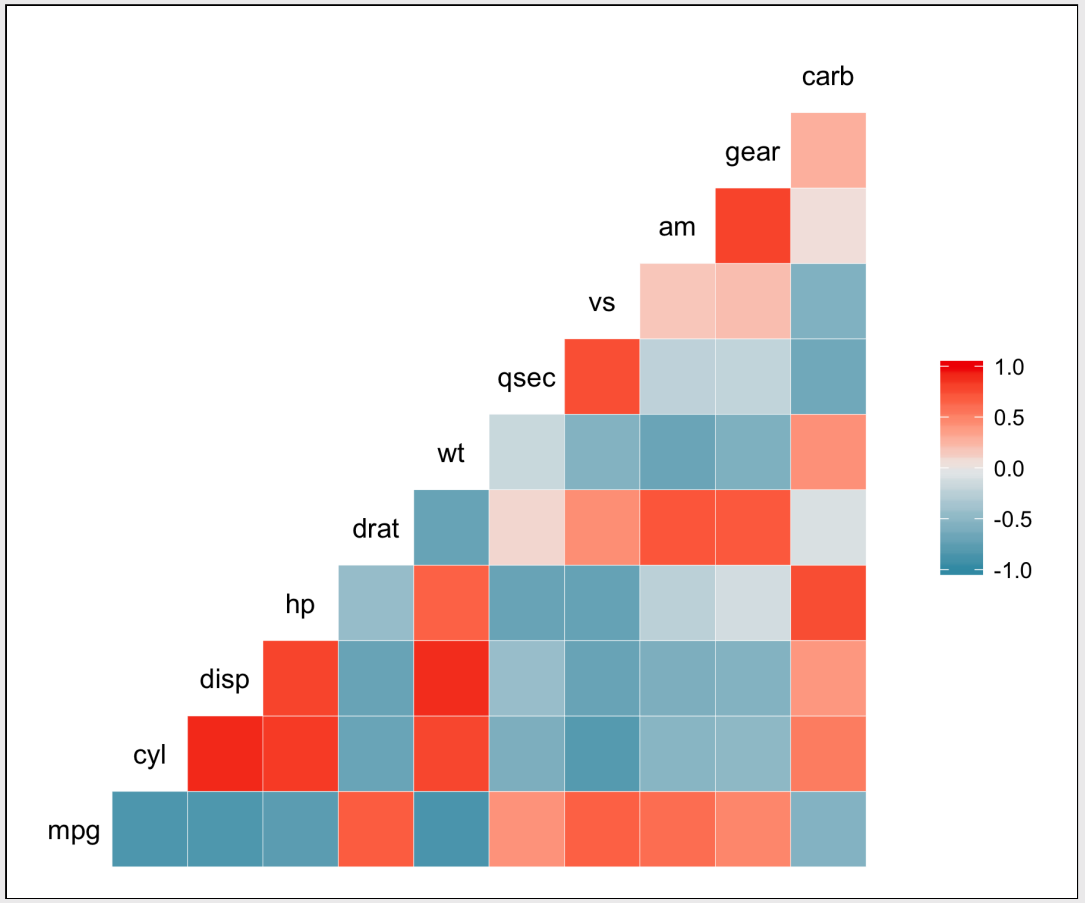
```
ggplot(mtcars) +
    geom_point(aes(x = mpg, y = hp),
               size = 2, alpha = 0.7) +
    annotate(geom = 'text', x = 28, y = 310,
             label = str_c('r = ', mtcarsCorr),
             hjust = 0, size = 7) +
    theme_half_open() +
    labs(x = 'Fuel economy (mpg)',
         y = 'Engine power (hp)')
```



Visualizing correlations: ggcorr()

```
library('GGally')
```

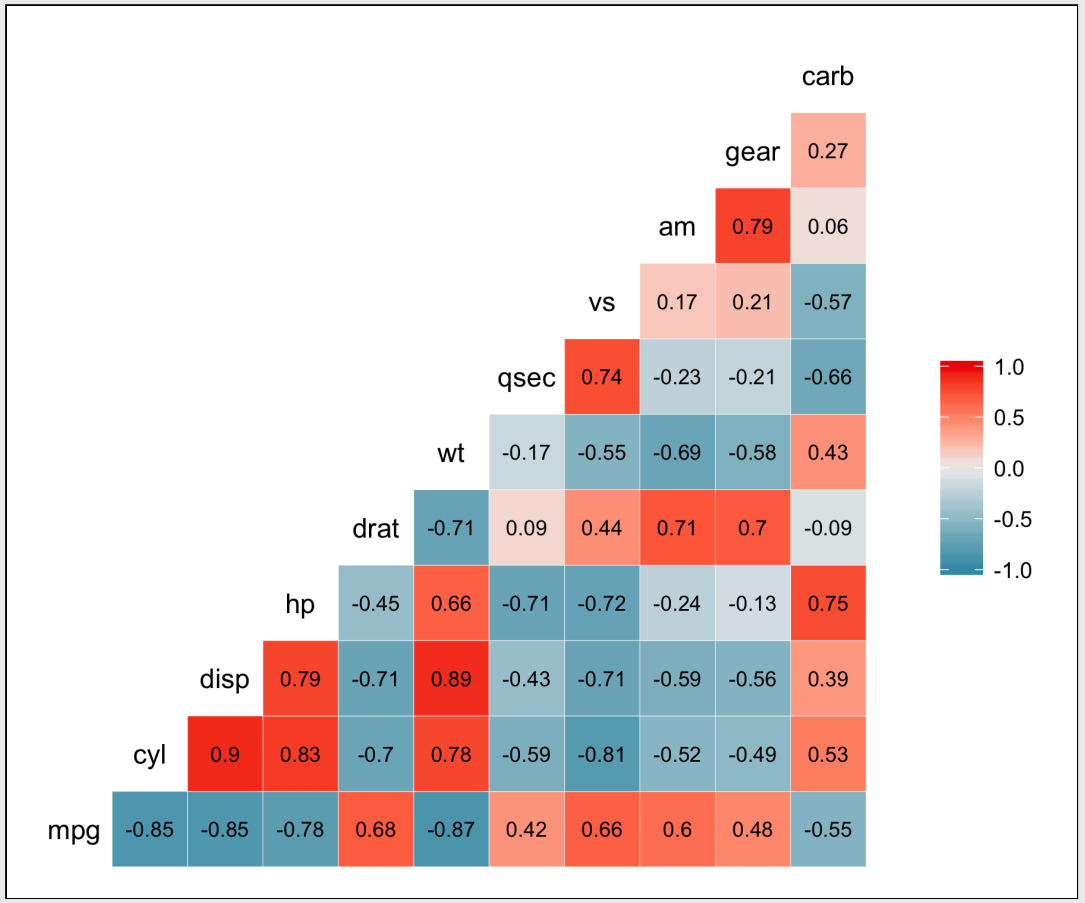
```
mtcars %>%
    ggcorr()
```



Visualizing correlations: ggcorr()

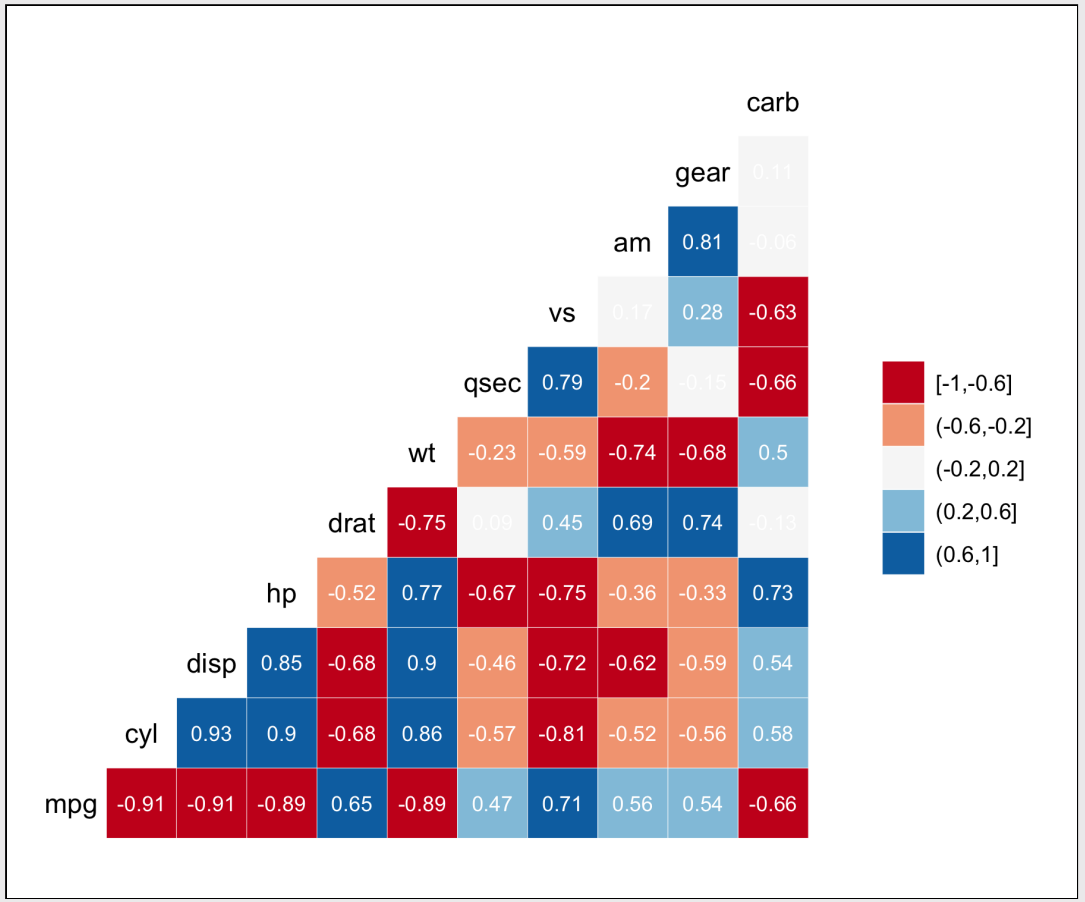
```
library('GGally')
```

```
mtcars %>%
    ggcorr(label = TRUE,
           label_size = 3,
           label_round = 2)
```



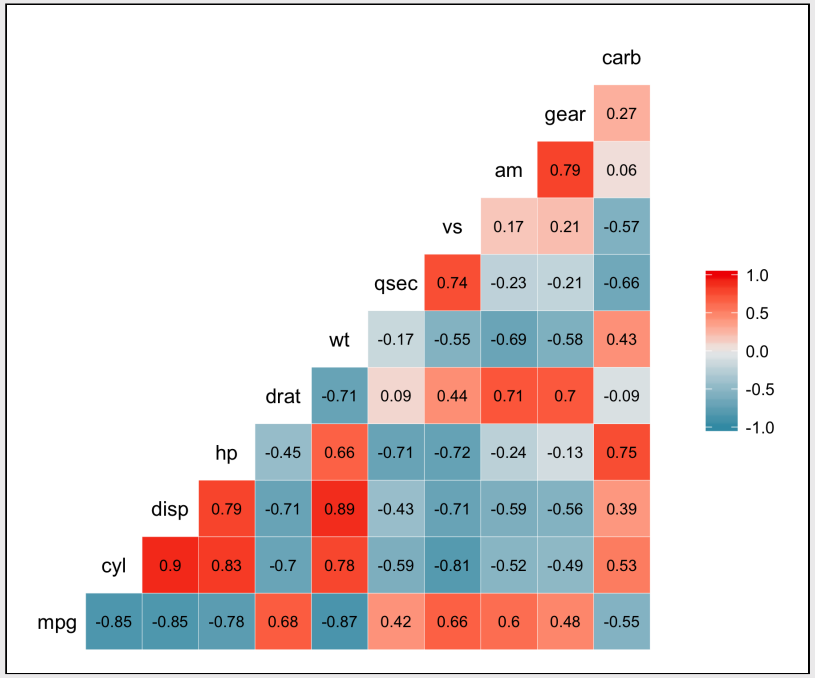
Visualizing correlations: ggcorr()

```
ggcor_mtcars_final <- mtcars %>%
    ggcorr(label = TRUE,
           label_size = 3,
           label_round = 2,
           label_color = 'white',
           nbreaks = 5,
           palette = "RdBu")
```



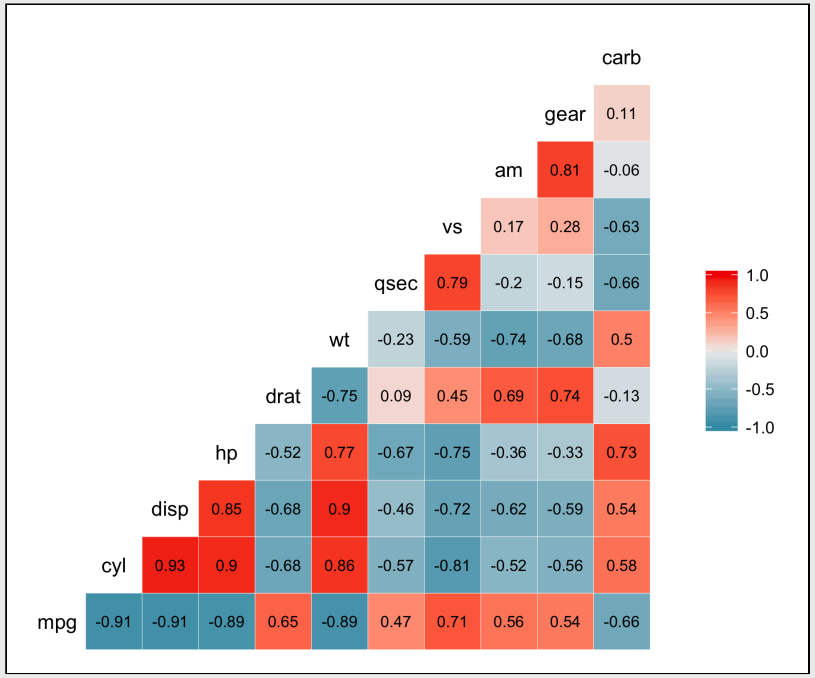
Pearson

```
mtcars %>%
    ggcorr(label = TRUE,
           label_size = 3,
           label_round = 2,
           method = c("pairwise", "pearson"))
```



Spearman

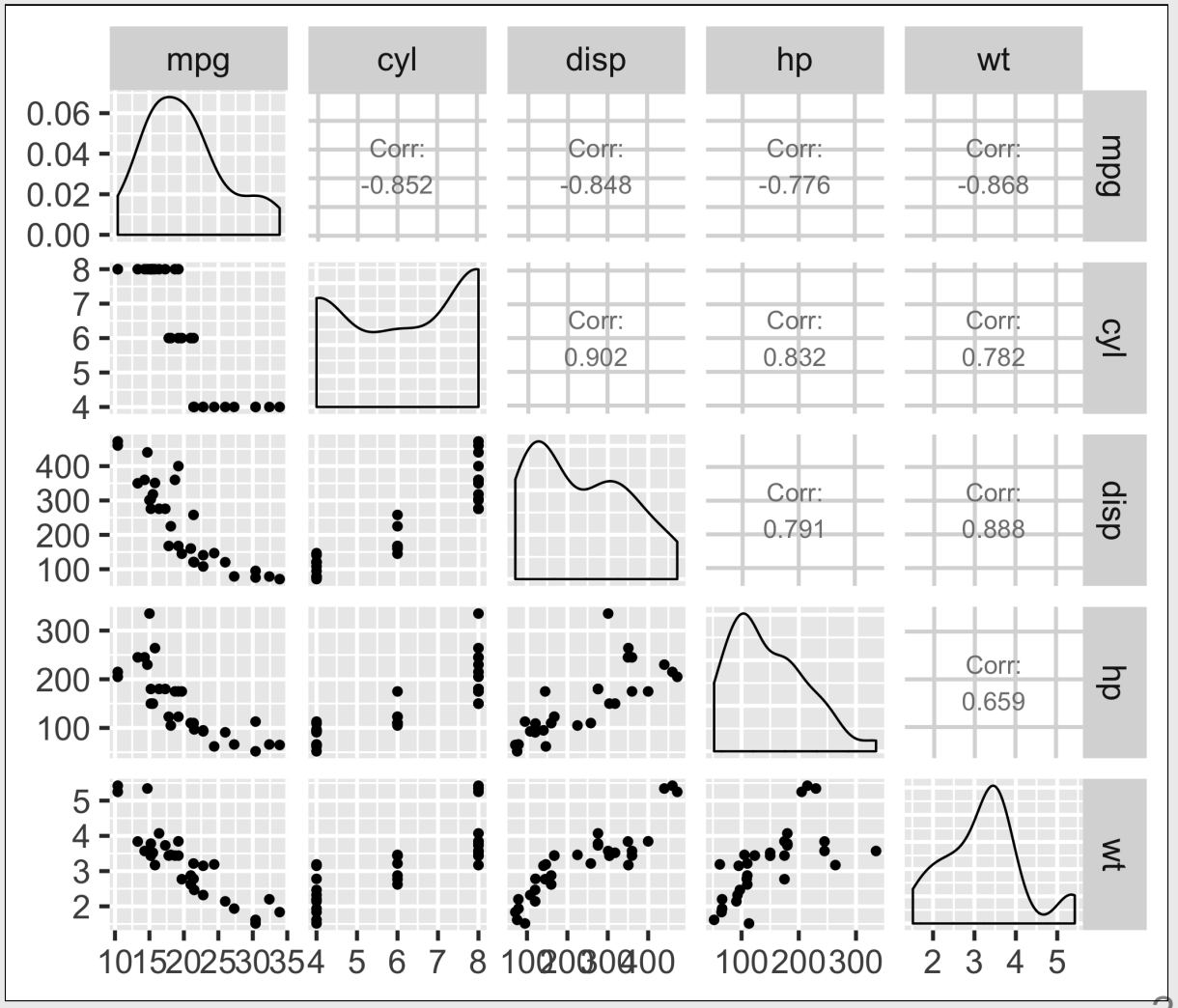
```
mtcars %>%
    ggcorr(label = TRUE,
           label_size = 3,
           label_round = 2,
           method = c("pairwise", "spearman"))
```



Correlograms: `ggpairs()`

```
library('GGally')
```

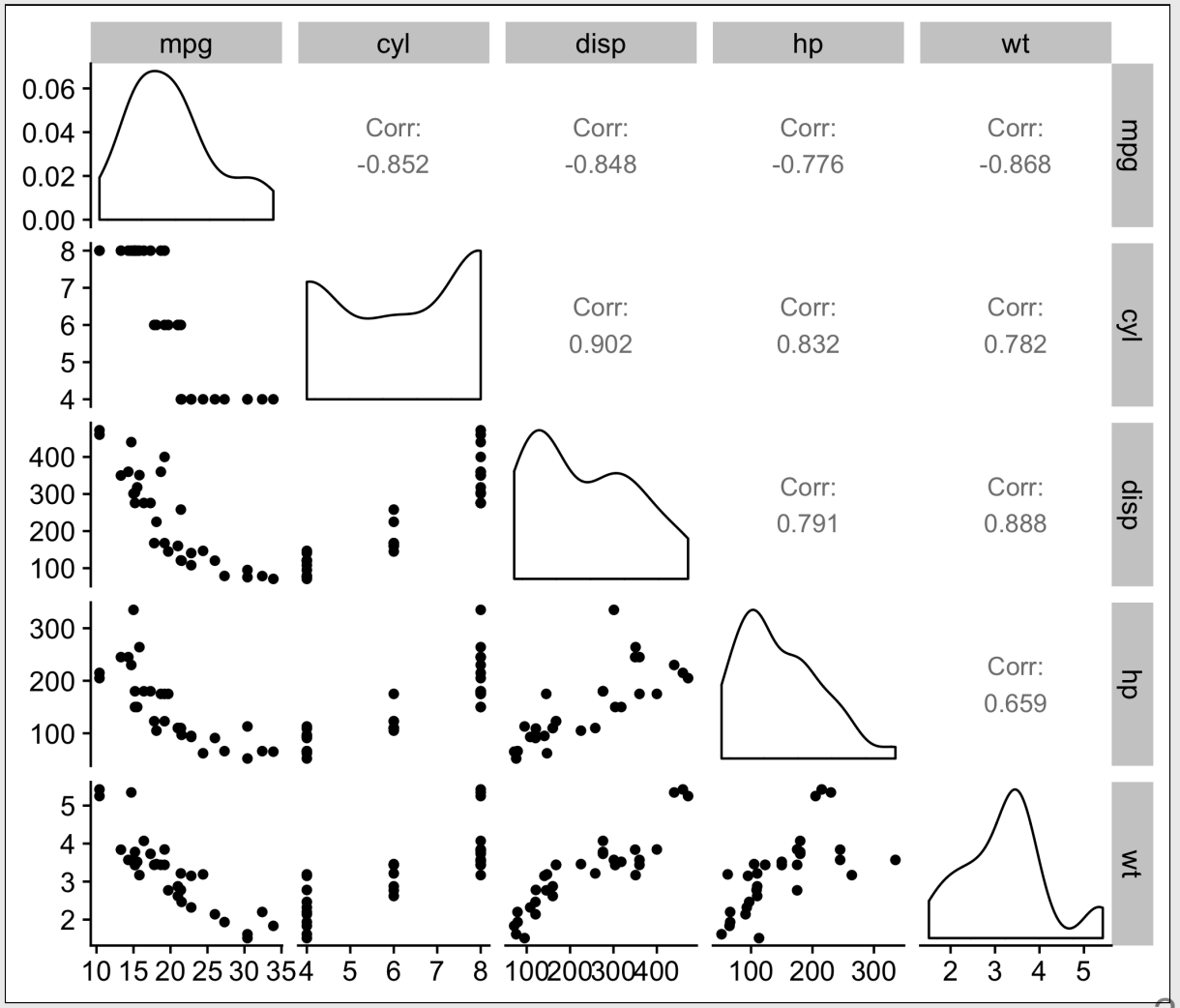
```
mtcars %>%
    select(mpg, cyl, disp, hp, wt) %>%
    ggpairs()
```



Correlograms: `ggpairs()`

```
library('GGally')
```

```
mtcars %>%
    select(mpg, cyl, disp, hp, wt) %>%
    ggpairs() +
    theme_half_open()
```



Your turn

15:00

Using the `wildlife_impacts` data frame:

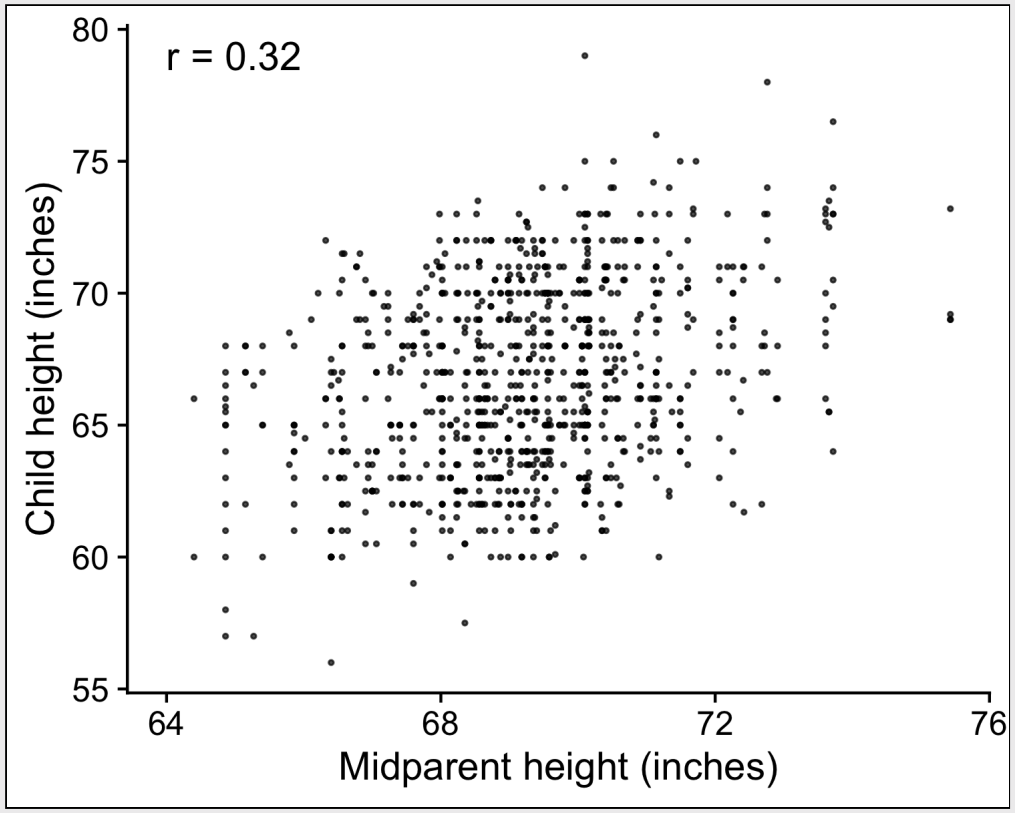
1. Find the two variables with the largest correlation in absolute value (i.e. closest to -1 or 1).
2. Create a scatter plot of those two variables. Include an annotation for the Pearson correlation coefficient.

Correlation Analysis

1. What is correlation?
2. Visualizing correlation
3. Linear models
4. Visualizing linear models

Galton's Height Data

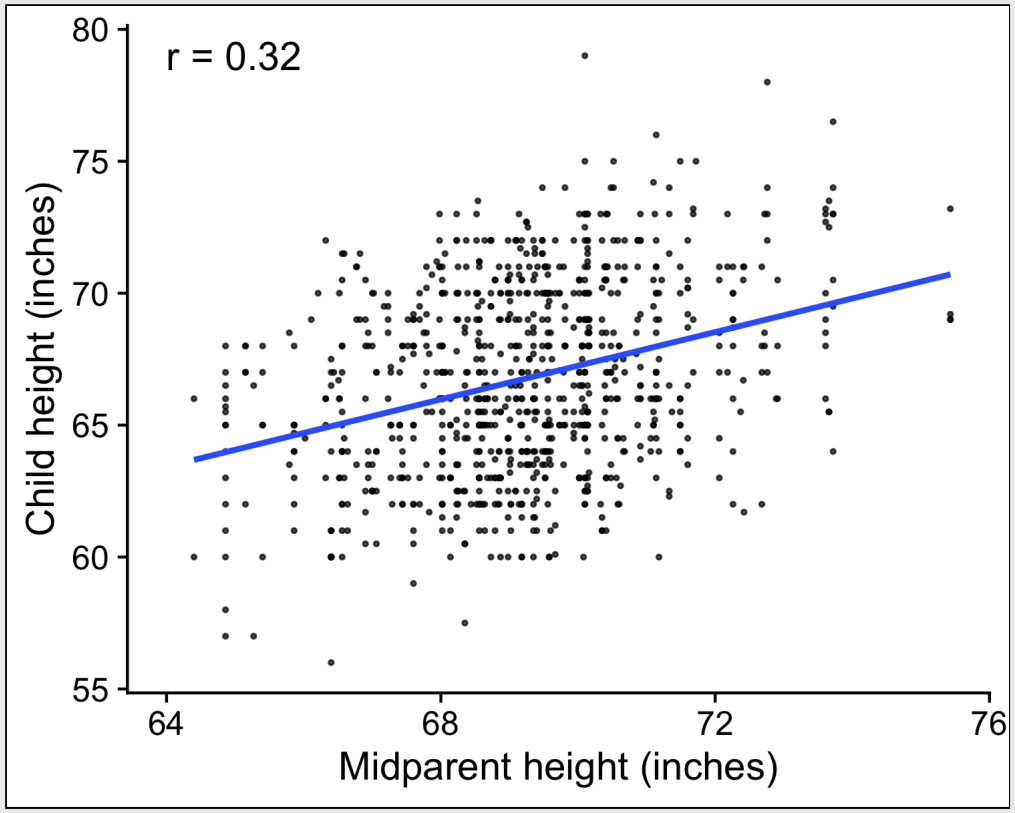
We already know that the correlation is 0.32, which means that the midparent height explains about 10% of the variation in the child height.



Galton's Height Data

We already know that the correlation is 0.32, which means that the midparent height explains about 10% of the variation in the child height.

To make predictions, we need to fit a model to these points.

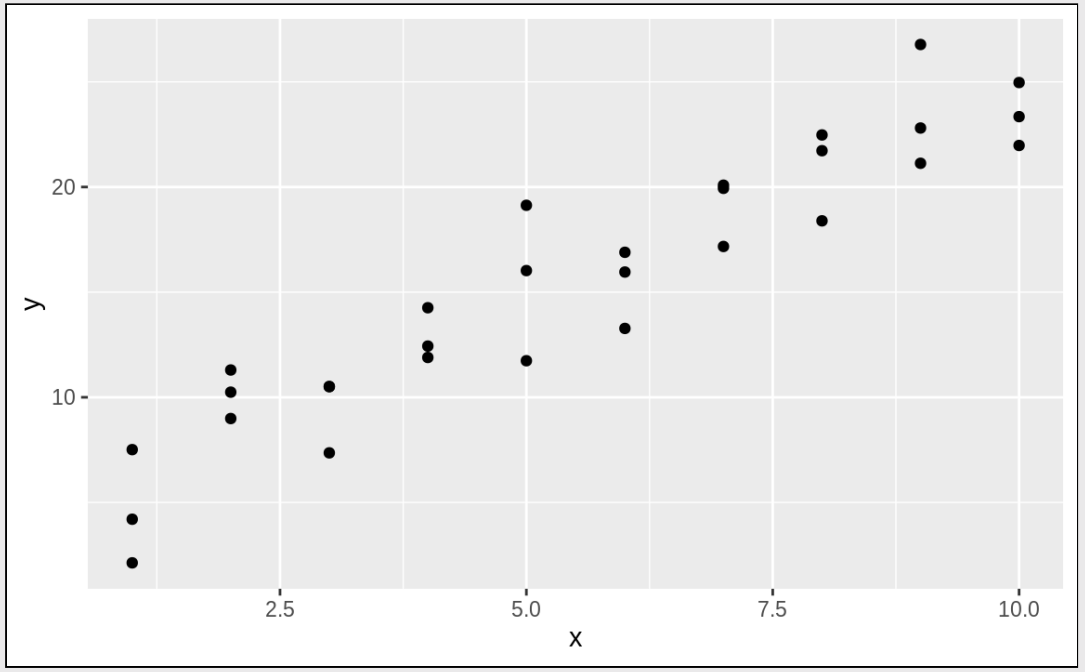


Modeling basics

Two parts to a model:

1. **Model family**: for example, $y = ax + b$
2. **Fitted model**: for example, $y = 3x + 7$

Here is some simulated data

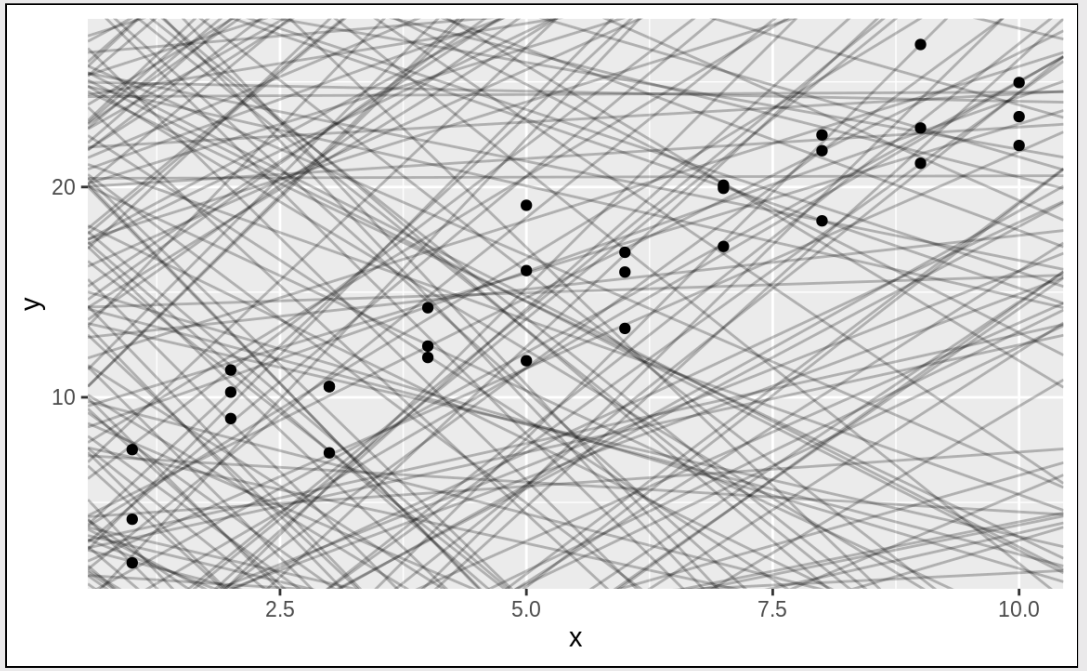


Modeling basics

Two parts to a model:

1. **Model family**: linear model: $y = ax + b$

There are an infinite number of possible models

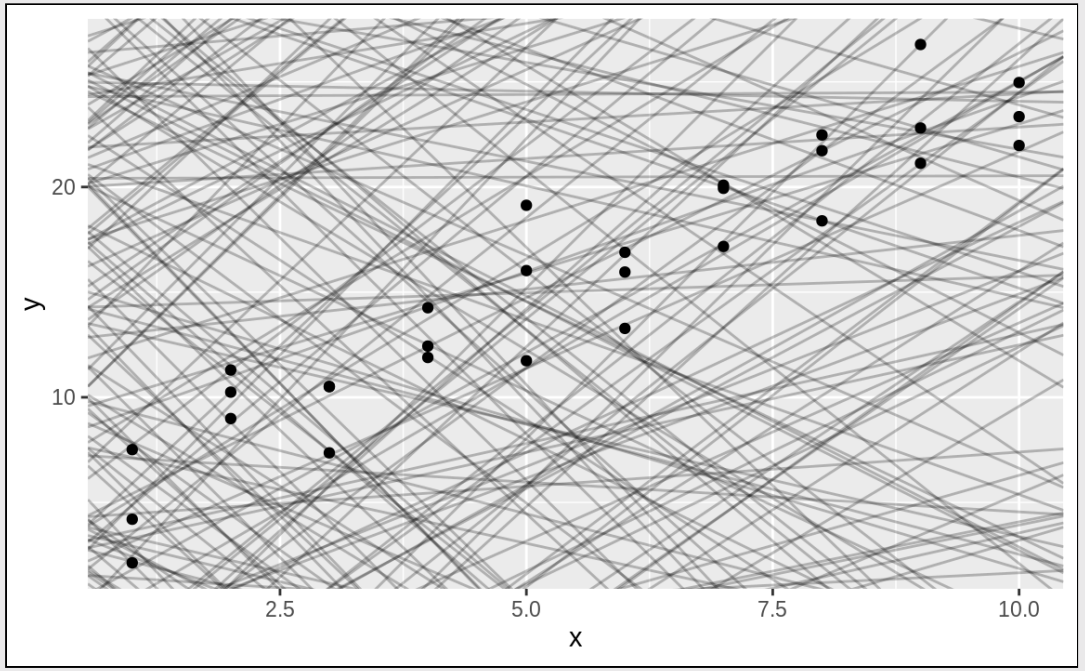


Modeling basics

Two parts to a model:

1. **Model family**: linear model: $y = ax + b$
2. **Fitted model**: How to choose the "best" a and b?

There are an infinite number of possible models



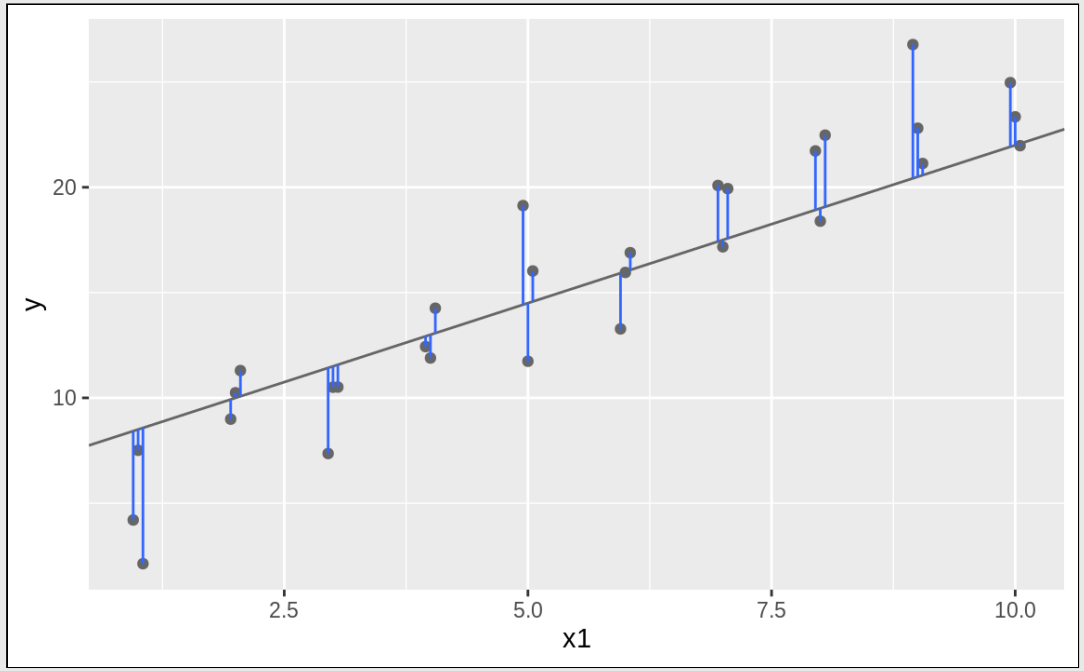
Modeling basics

Two parts to a model:

1. **Model family**: linear model: $y = ax + b$
2. **Fitted model**: How to choose the "best" a and b?

We need to come up with some measure of "distance" from the model to the data

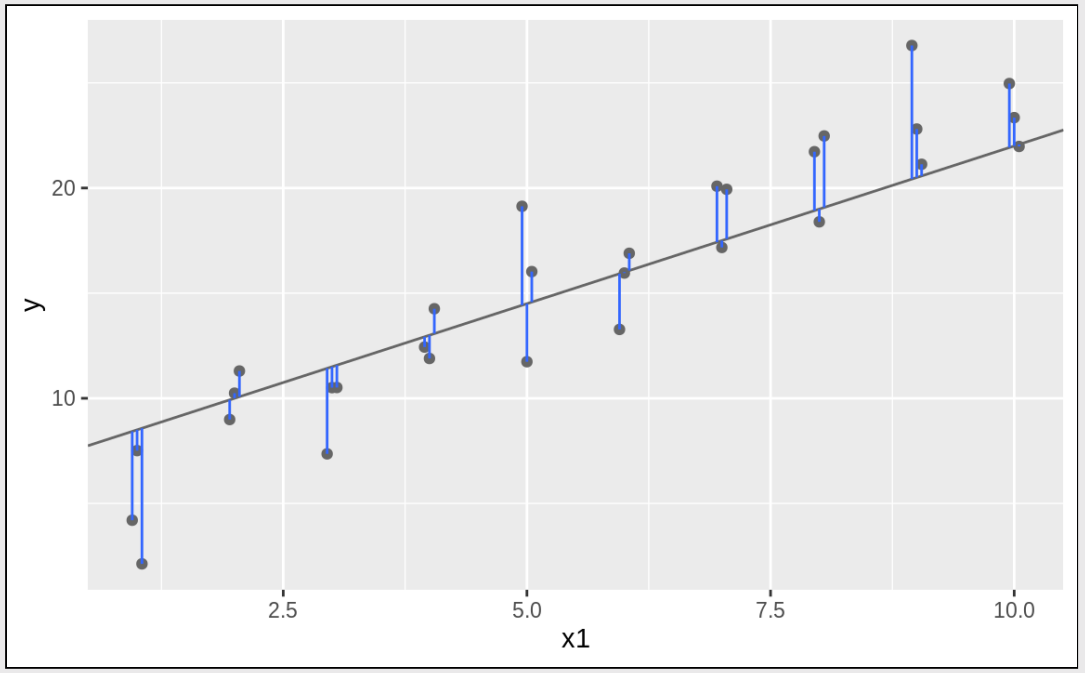
Compute the **"residuals"**:
The distance between the model line and the data



Residuals

"residual": The distance between the model line and the data

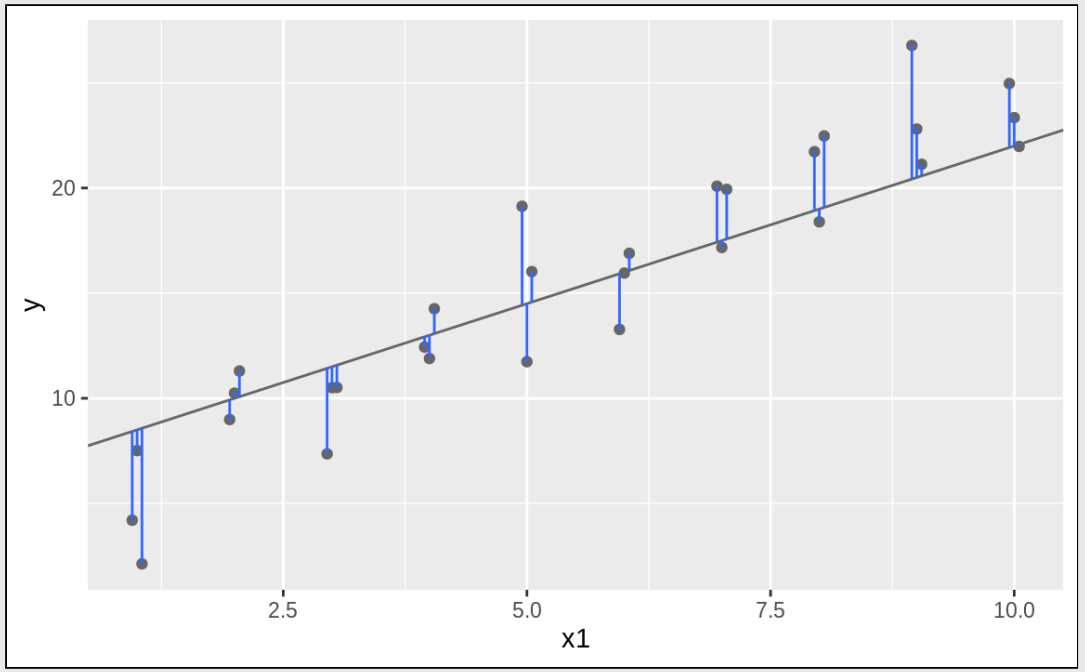
residual = $y_i - y_i'$



All the residuals

"residual": The distance between the model line and the data

residual = $y_i - y_i'$



Sum of squared residuals:

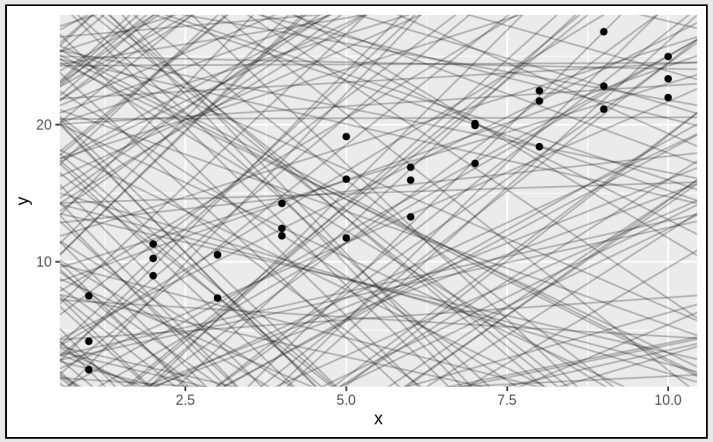
$$\text{SSR} = \sum_{i=1}^{n} (y_i - y_i')^2$$

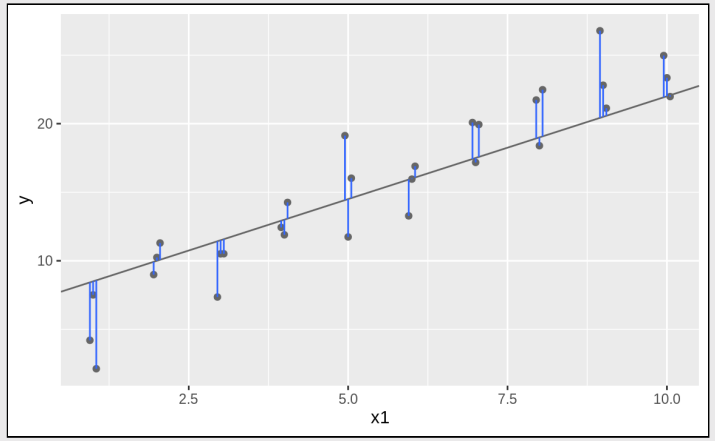
Search algorithm

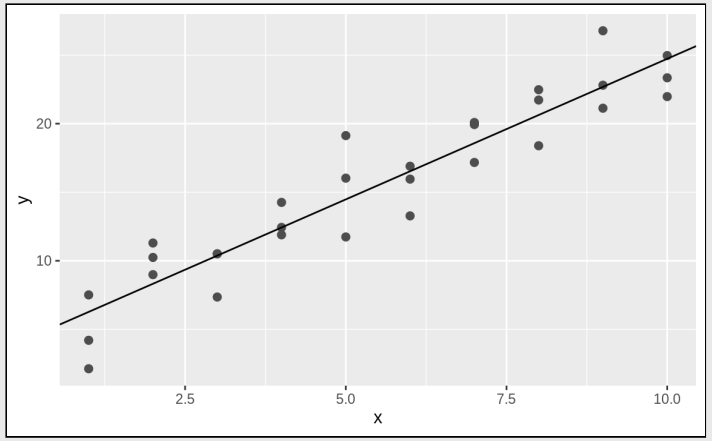
1): Choose a model:
$y = ax + b$

2): Compute the SSR:
$\text{SSR} = \sum_{i=1}^{n}(y_i - y_i')^2$

3): Repeat steps 1 & 2 until the smallest SSR is found







Fitting a linear model in R

```
model <- lm(formula = y ~ x,
            data = data)
```

Example: Galton's height data

```
model <- lm(
  formula = childHeight ~ midparentHeight,
  data    = GaltonFamilies)
```

Get coefficients

```
coef(model)
```

```
##     (Intercept) midparentHeight
##      22.6362405       0.6373609
```

Fitting a linear model in R

```
model <- lm(formula = y ~ x,
            data = data)
```

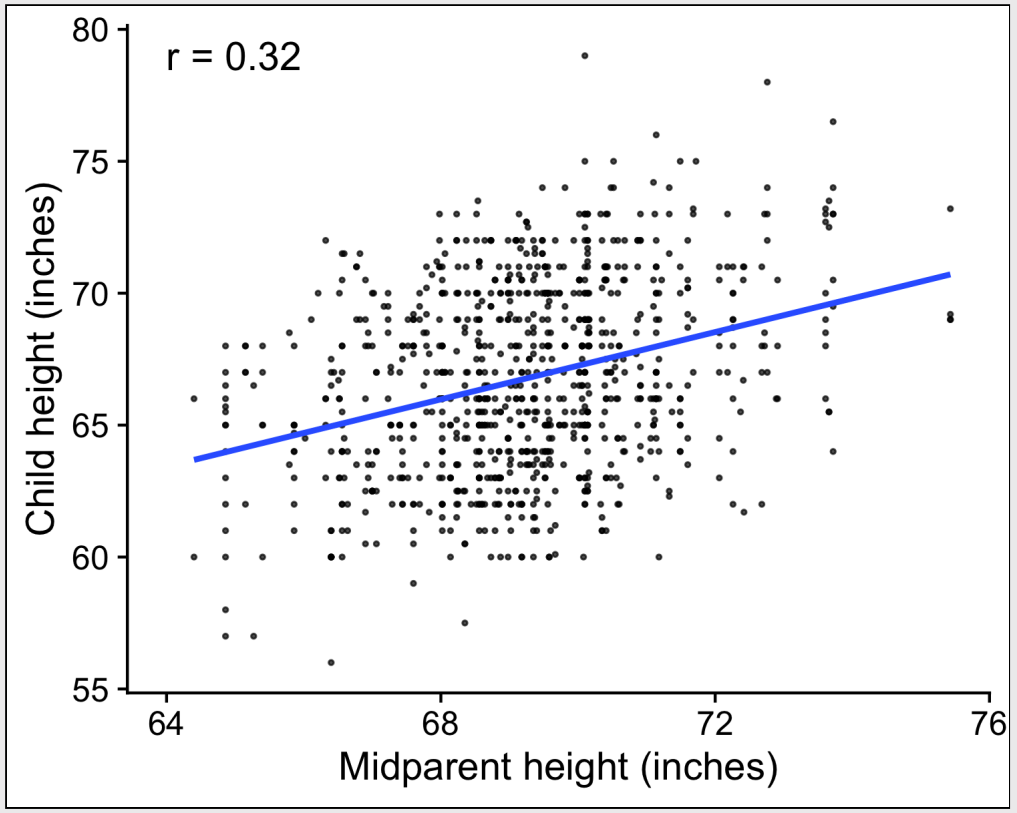
Example: Galton's height data

```
model <- lm(
  formula = childHeight ~ midparentHeight,
  data    = GaltonFamilies)
```

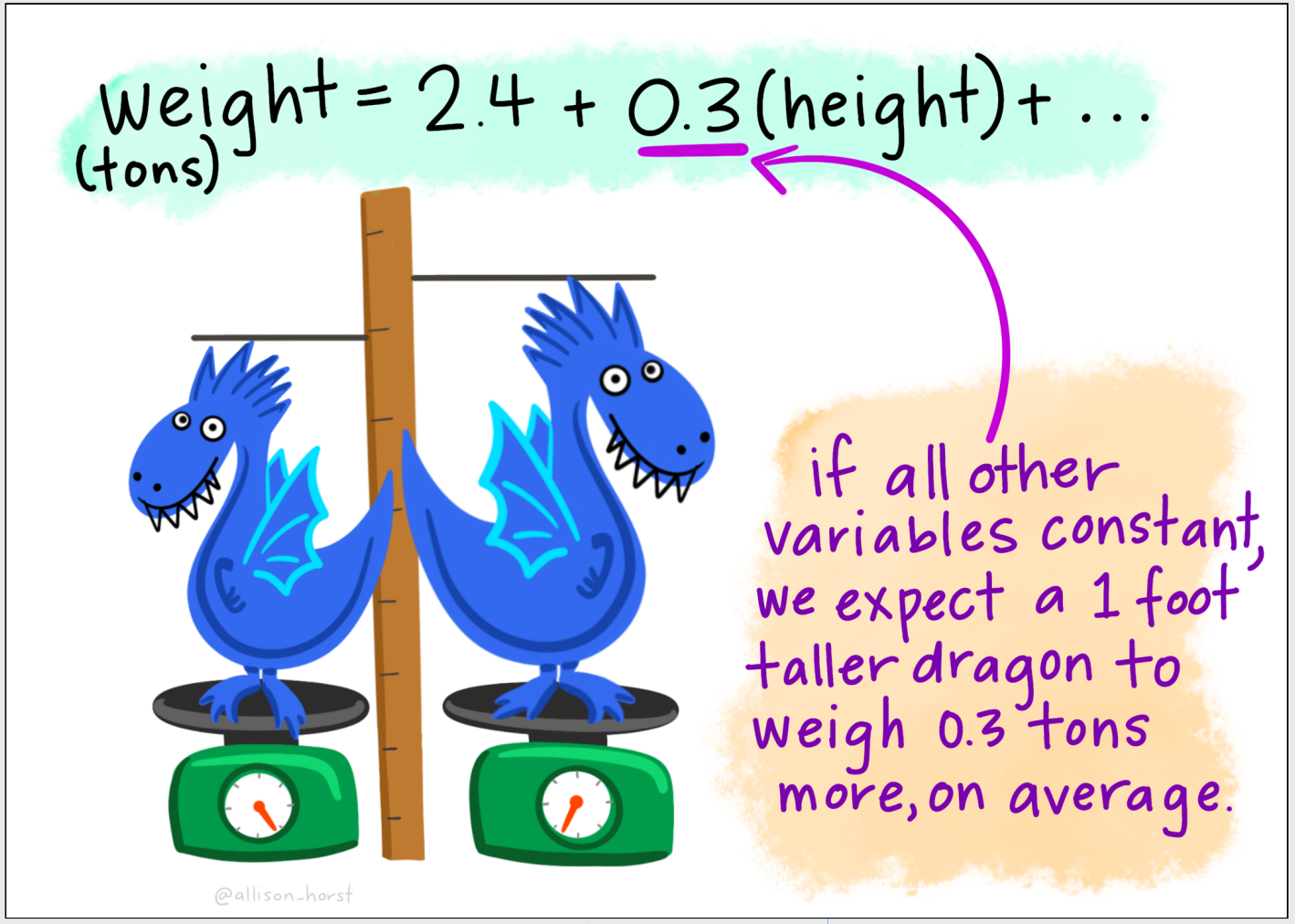
Get coefficients

```
coef(model)
```

```
##     (Intercept) midparentHeight
##      22.6362405       0.6373609
```



Interpreting results

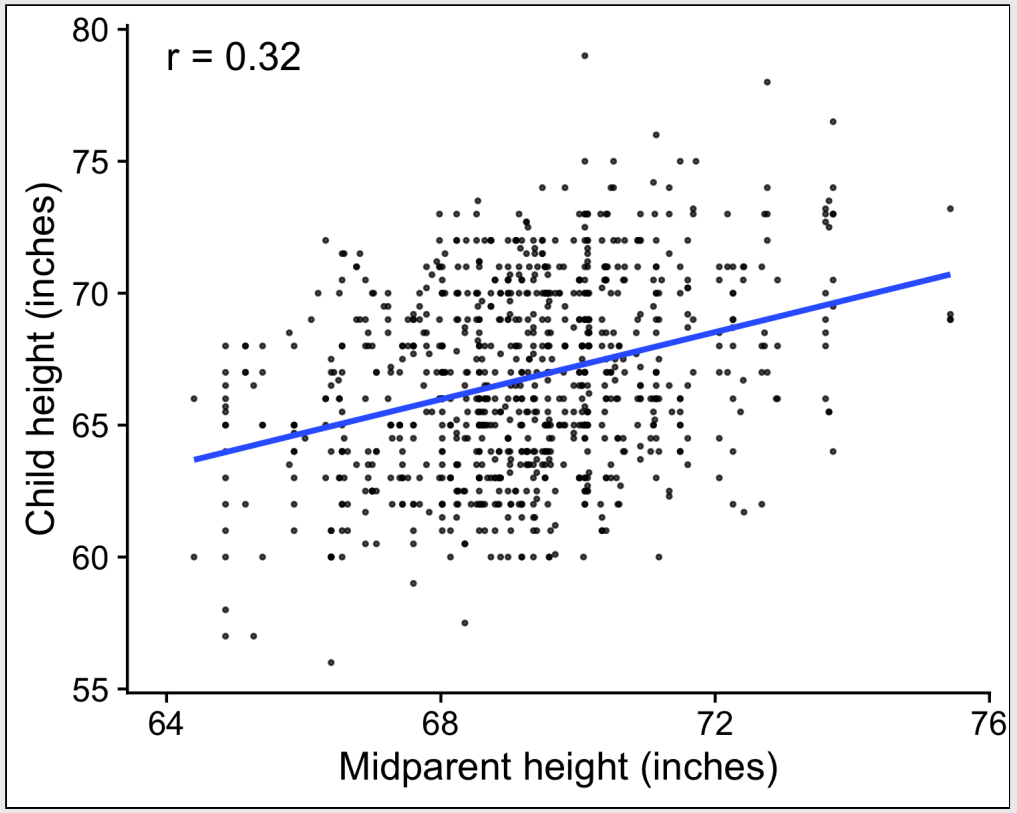


Art by Allison Horst

Example write up for Galton's height data

The correlation between midparent height and child height is **0.32**. Therefore, **10%** of the variance in child height is explained by midparent height.

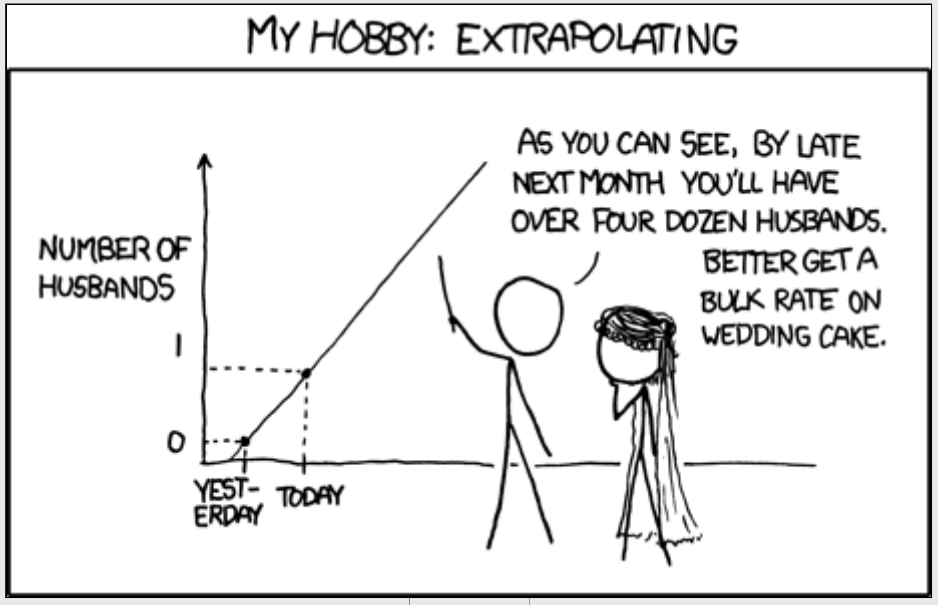
The slope of the best fitting regression line indicates that child height increased by **0.64** inches as midparent height increased by one inch.



Making predictions

Interpolation is OK: You may predict values of y for values of x that were not observed but are within the range of the observed values of x.

Extrapolation is BAD: You should NOT predict values of y using values of x that are outside the observed range of x.

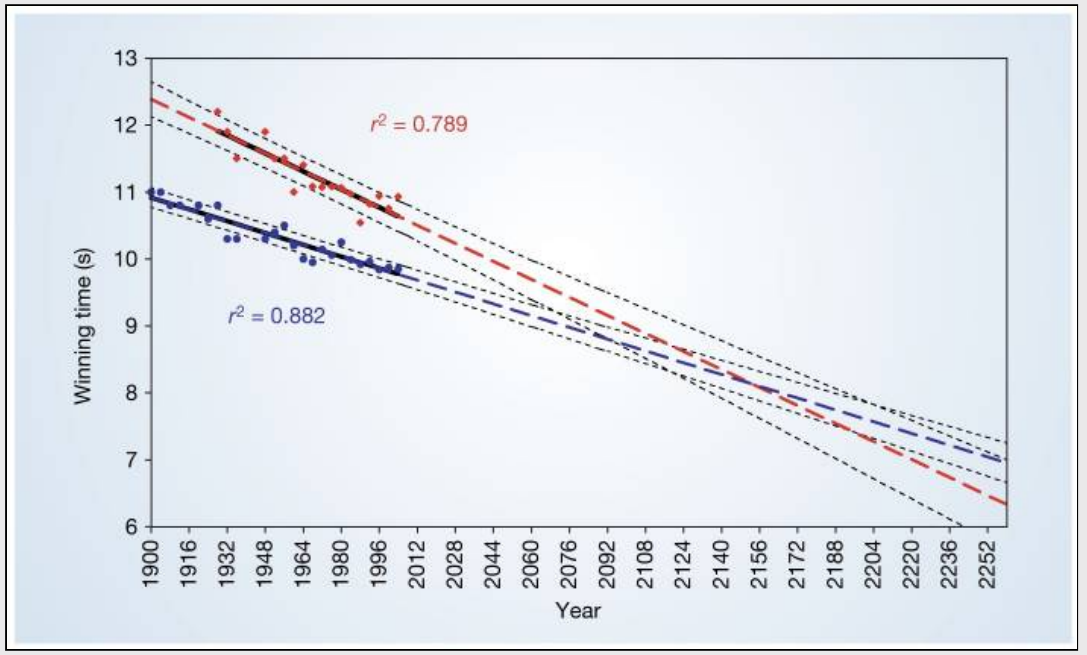


xkcd

Repeat: Extrapolation is BAD

Tatem, A. J., Guerra, C. A., Atkinson, P. M., & Hay, S. I. (2004). Momentous sprint at the 2156 Olympics? *Nature*, 431(7008), 525-525. View online

"Extrapolation of these trends to the 2008 Olympiad indicates that the women's 100-metre race could be won in a time of 10.57±0.232 seconds and the men's event in 9.73±0.144 seconds. **Should these trends continue, the projections will intersect at the 2156 Olympics, when — for the first time ever — the winning women's 100-metre sprint time of 8.079 seconds will be lower than that of the men's winning time of 8.098 seconds (Fig. 1).**"

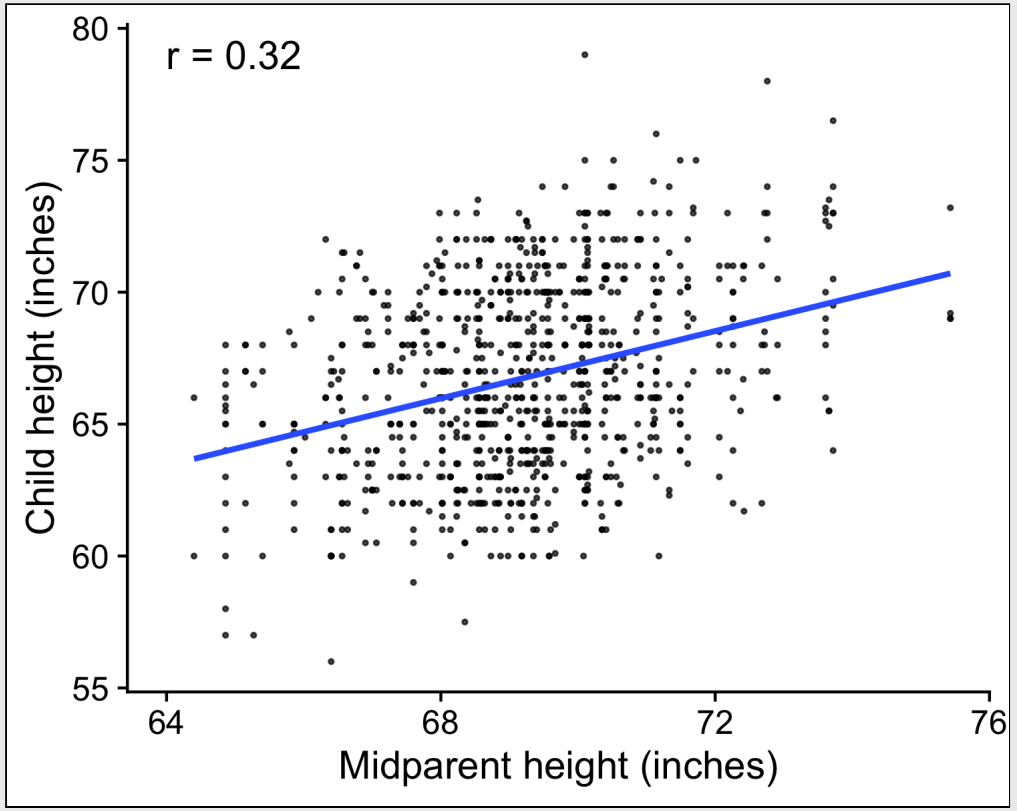


Symantics

The all mean the same thing:

- "Use X to predict Y"
- "Regress Y on X"
- "Regression of Y on X"

```
model <- lm(formula = y ~ x,
            data = data)
```



Symantics

Y: Dependent variable

- Outcome variable
- Response variable
- Regressand
- Left-hand variable

X: Independent variable

- Predictor variable
- Explanatory variable
- Regressor
- Right-hand variable

```
model <- lm(formula = y ~ x,
            data = data)
```

Correlation Analysis

1. What is correlation?
2. Visualizing correlation
3. Linear models
4. Visualizing linear models

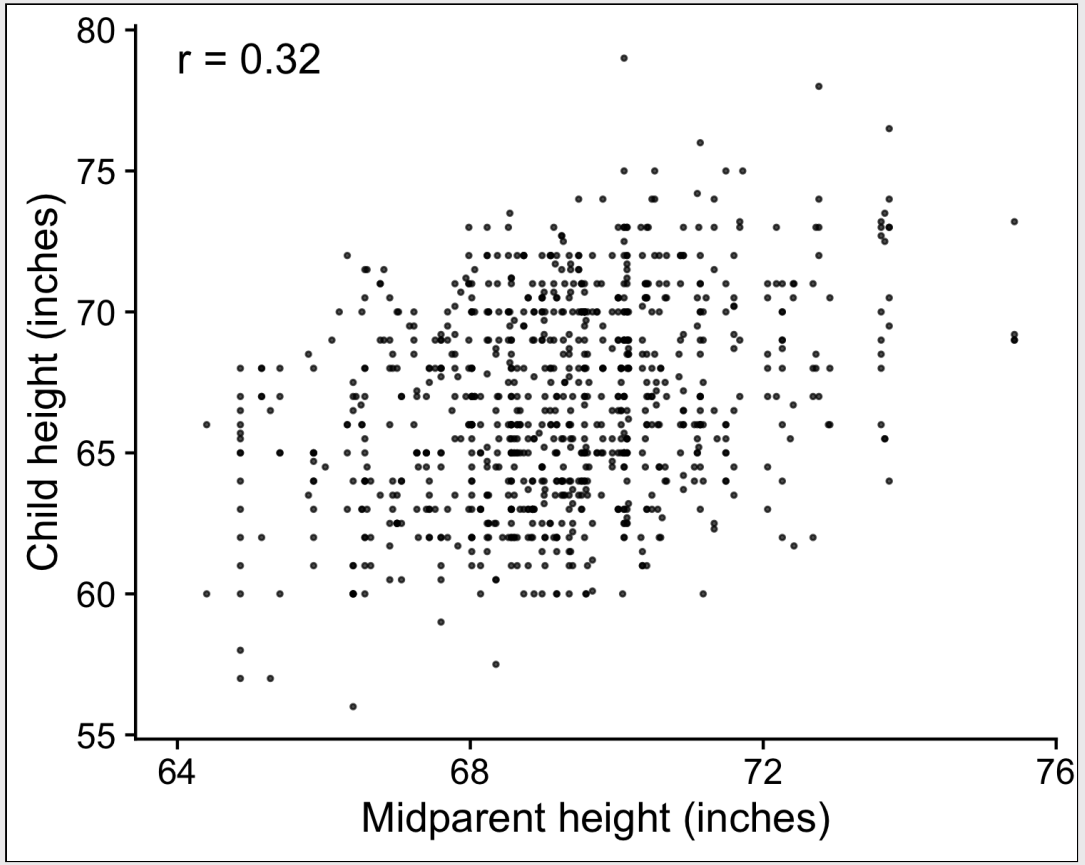
Visualizing models

Compute the correlation

```
galtonCorr <- round(cor(
    GaltonFamilies$mpg, GaltonFamilies$hp,
    method = 'pearson'), 2)
```

Make the plot

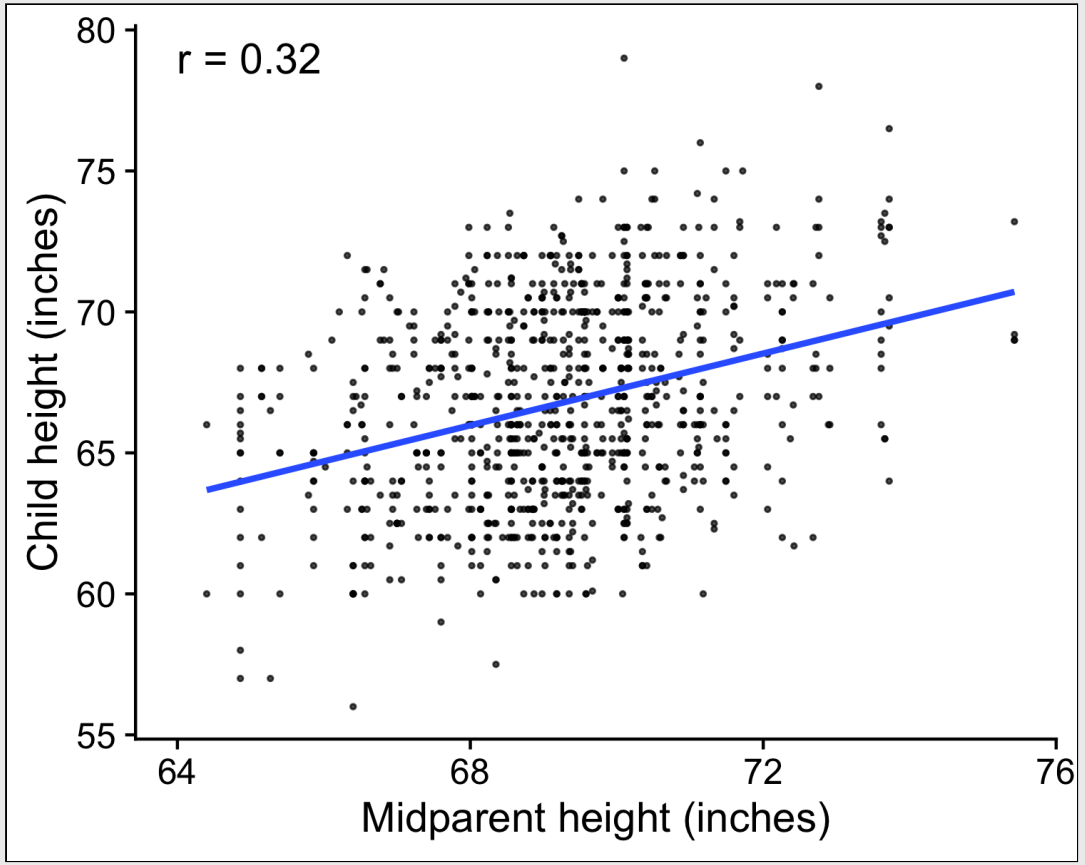
```
plot <- ggplot(GaltonFamilies) +
    geom_point(aes(x = midparentHeight,
                   y = childHeight),
               size = 0.5, alpha = 0.7) +
    annotate(geom = 'text', x = 64, y = 79,
             label = str_c('r = ', galtonCorr),
             hjust = 0, size = 5) +
    theme_half_open() +
    labs(x = 'Midparent height (inches)',
         y = 'Child height (inches)')
```



Visualizing models

Add linear model with `geom_smooth()`

```
plot +
    geom_smooth(aes(x = midparentHeight,
                    y = childHeight),
                method = 'lm',
                se = FALSE)
```



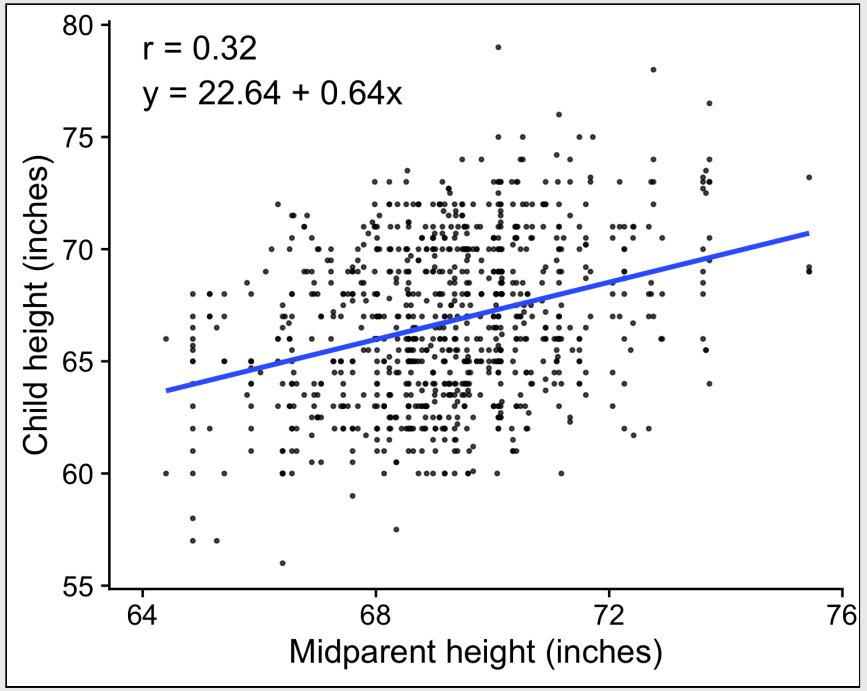
Visualizing models

Make equation label

```
model <- lm(
  formula = childHeight ~ midparentHeight,
  data = GaltonFamilies)
coefs <- round(coef(model), 2)
eqLabel <- str_c('y = ', coefs[1], ' + ', coefs[2], 'x')
```

Add linear model equation with `annotate()`

```
plot +
    geom_smooth(aes(x = midparentHeight,
                    y = childHeight),
                method = 'lm',
                se = FALSE) +
    annotate(geom = 'text', x = 64, y = 77,
             label = eqLabel, hjust = 0,
             size = 5)
```



Your turn

15:00

Using the `msleep` data frame:

1. Create a scatter plot of `brainwt` versus `bodywt`.
2. Include an annotation for the Pearson correlation coefficient.
3. Include an annotation for the best fit line.

Bonus: Compare your results to a log-linear relationship by converting the x and y variables to the log of x and y.