CORRELATION AND REGRESSION

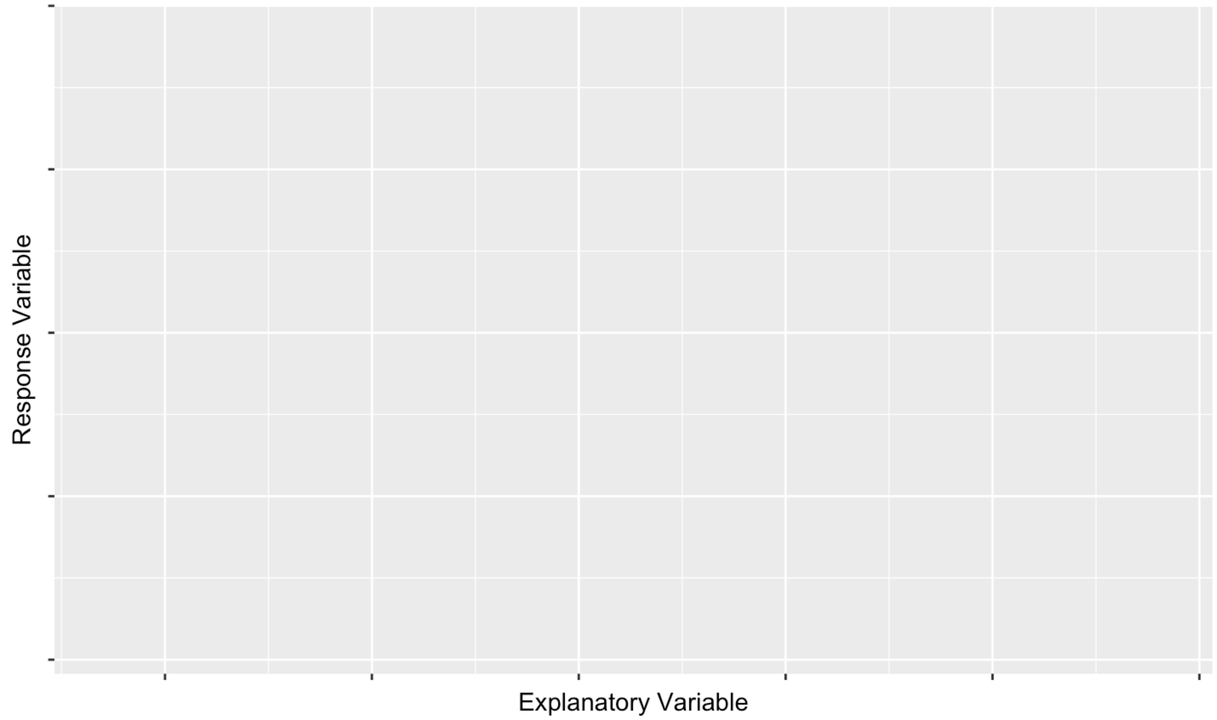
# Modeling bivariate relationships

# Bivariate relationships

- Both variables are numerical
- Response variable
  - a.k.a. *y*, dependent
- Explanatory variable
  - Something you think might be related to the response
  - a.k.a. *x*, independent, predictor

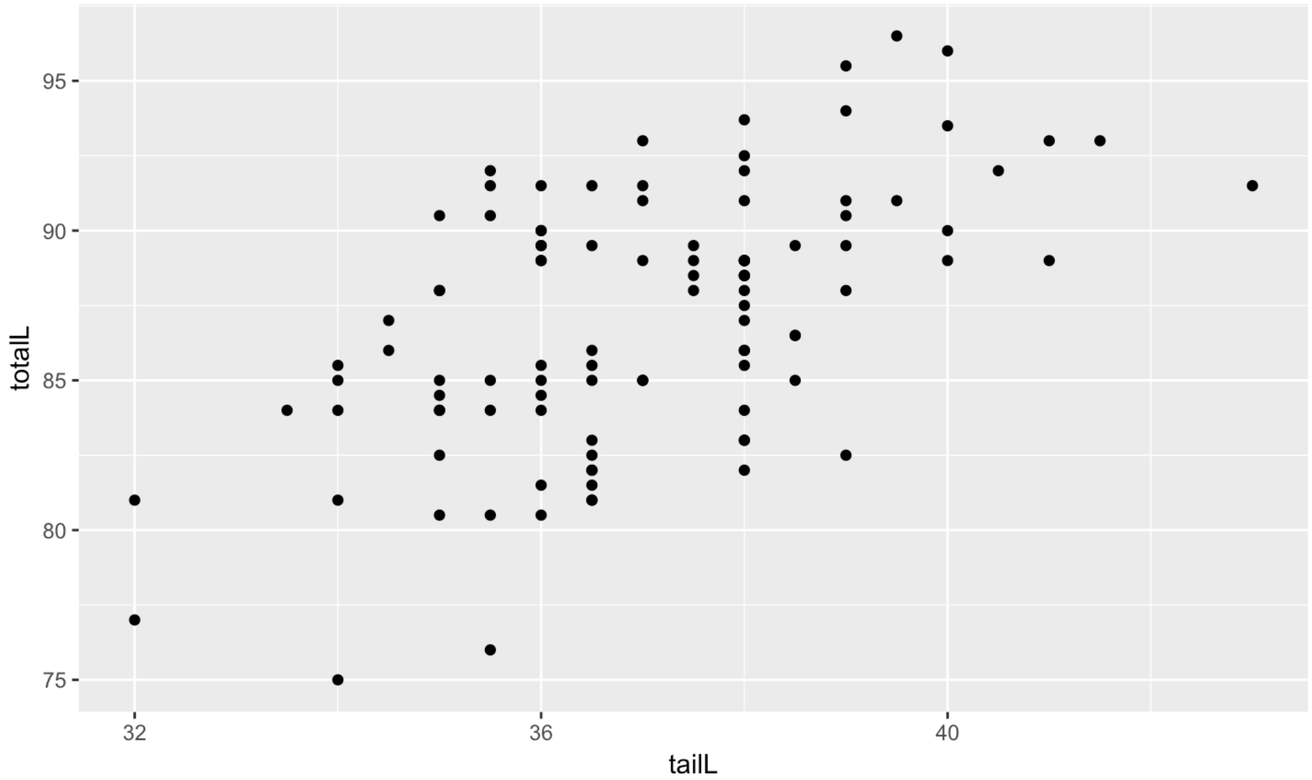
# Graphical representations

- Put response on vertical axis
- Put explanatory on horizontal axis



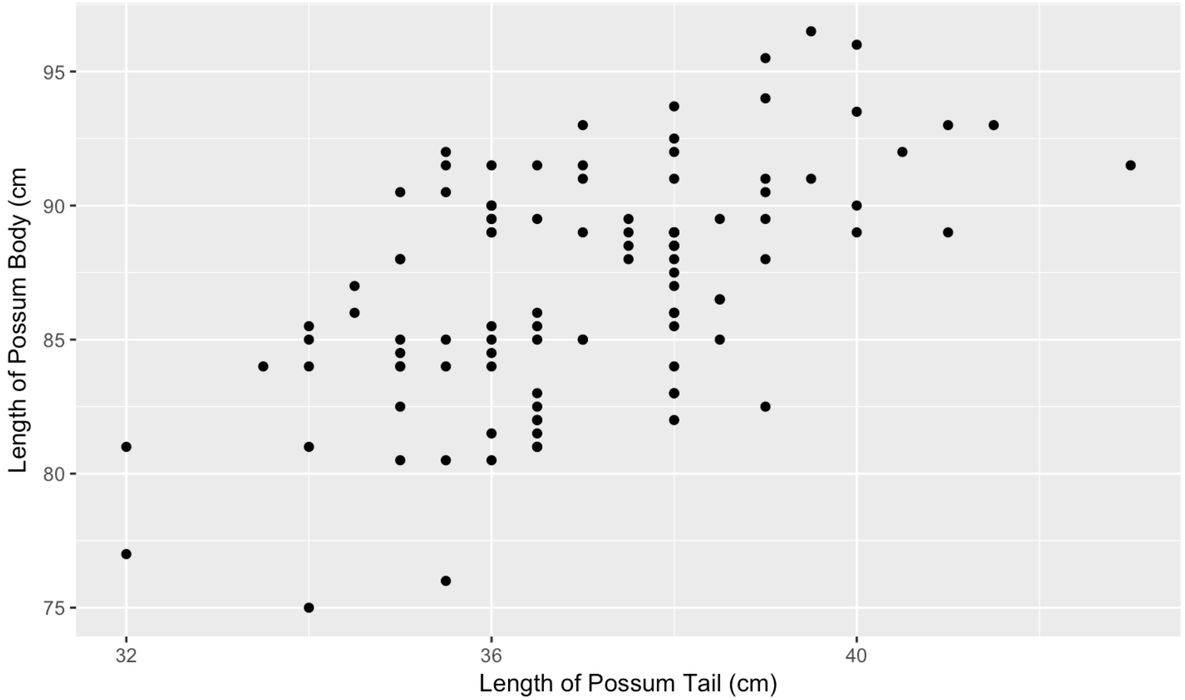
# Scatterplot

```
> ggplot(data = possum, aes(y = totalL, x = tailL)) +
    geom_point()
```



# Scatterplot

```
> ggplot(data = possum, aes(y = totalL, x = tailL)) +
    geom_point() +
    scale_x_continuous("Length of Possum Tail (cm)") +
    scale_y_continuous("Length of Possum Body (cm)")
```

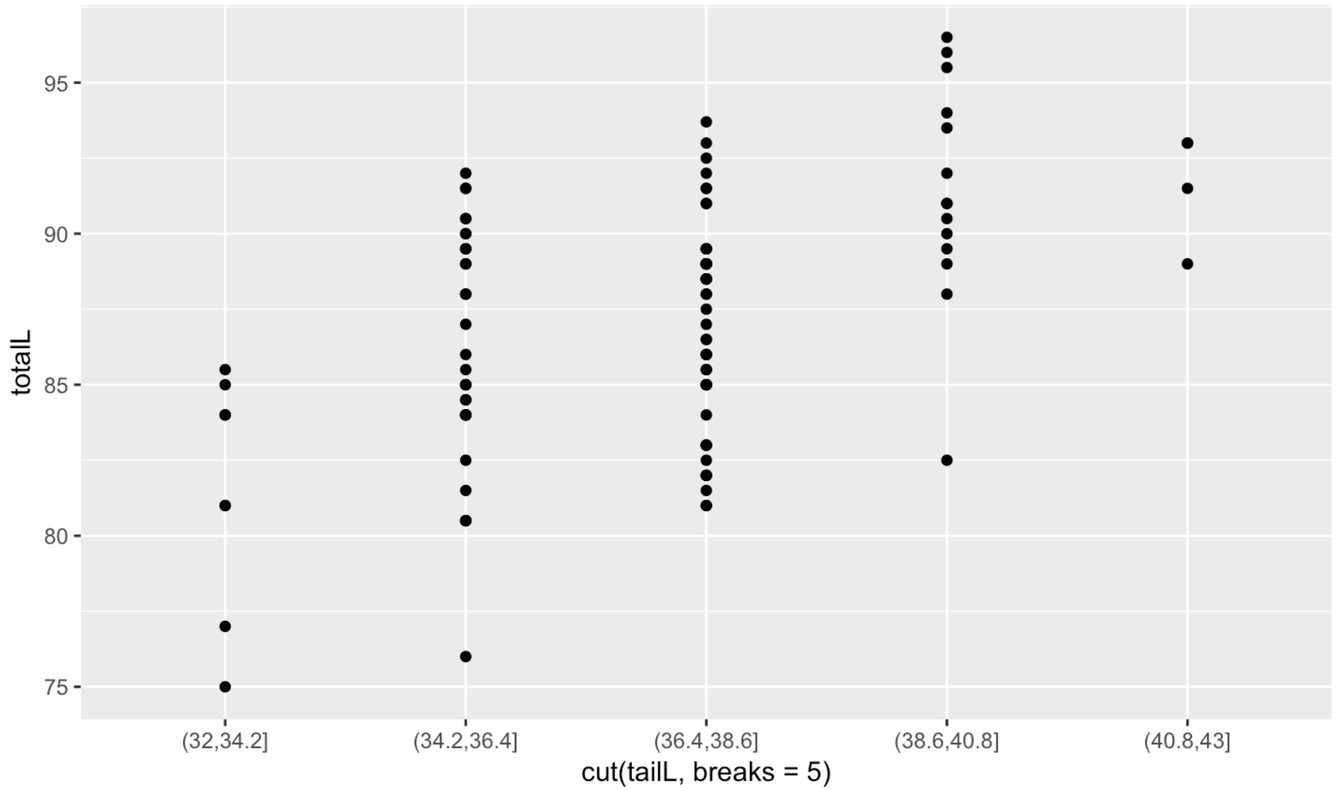


# Bivariate relationships

- Can think of boxplots as scatterplots...
  - ...but with discretized explanatory variable
- `cut()` function discretizes
  - Choose appropriate number of "boxes"

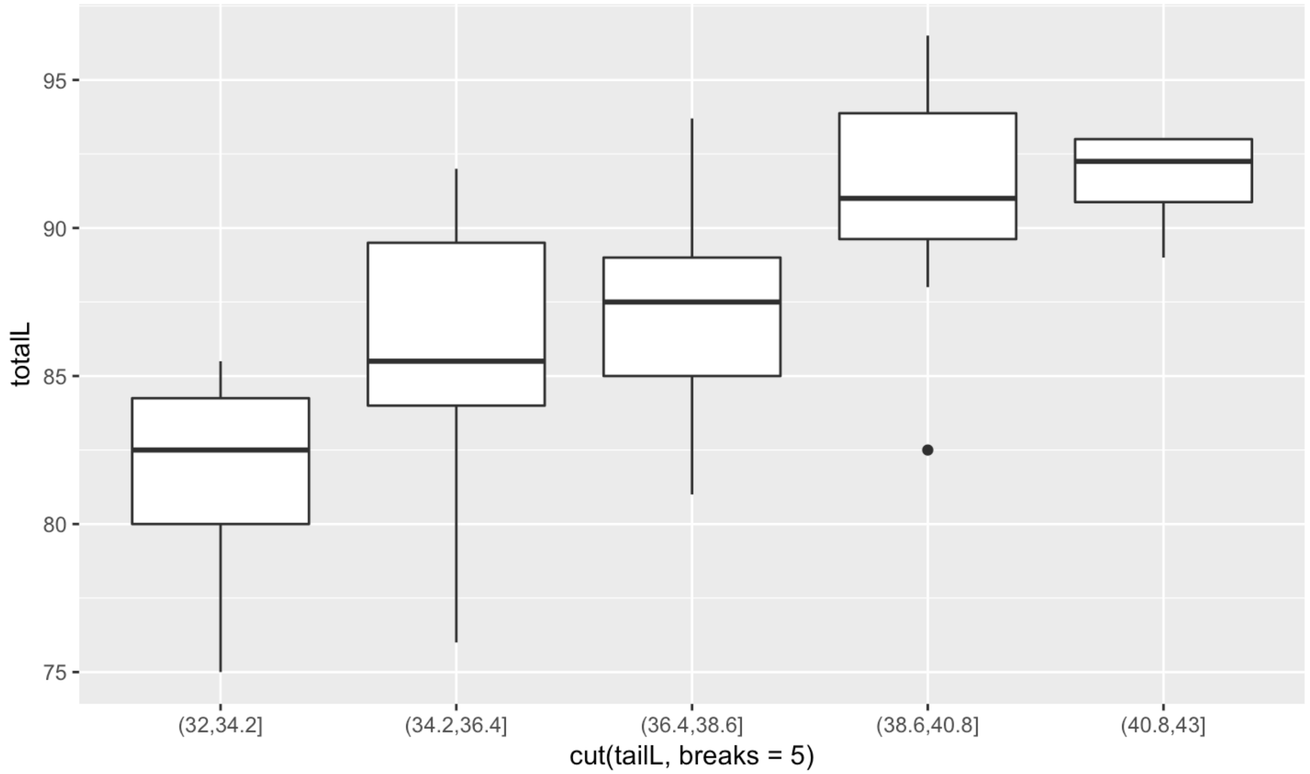
# Scatterplot

```
> ggplot(data = possum, aes(y = totalL, x = cut(tailL, breaks = 5))) +
    geom_point()
```



# Scatterplot

```
> ggplot(data = possum, aes(y = totalL, x = cut(tailL, breaks = 5))) +
    geom_boxplot()
```



CORRELATION AND REGRESSION

# Let’s practice!

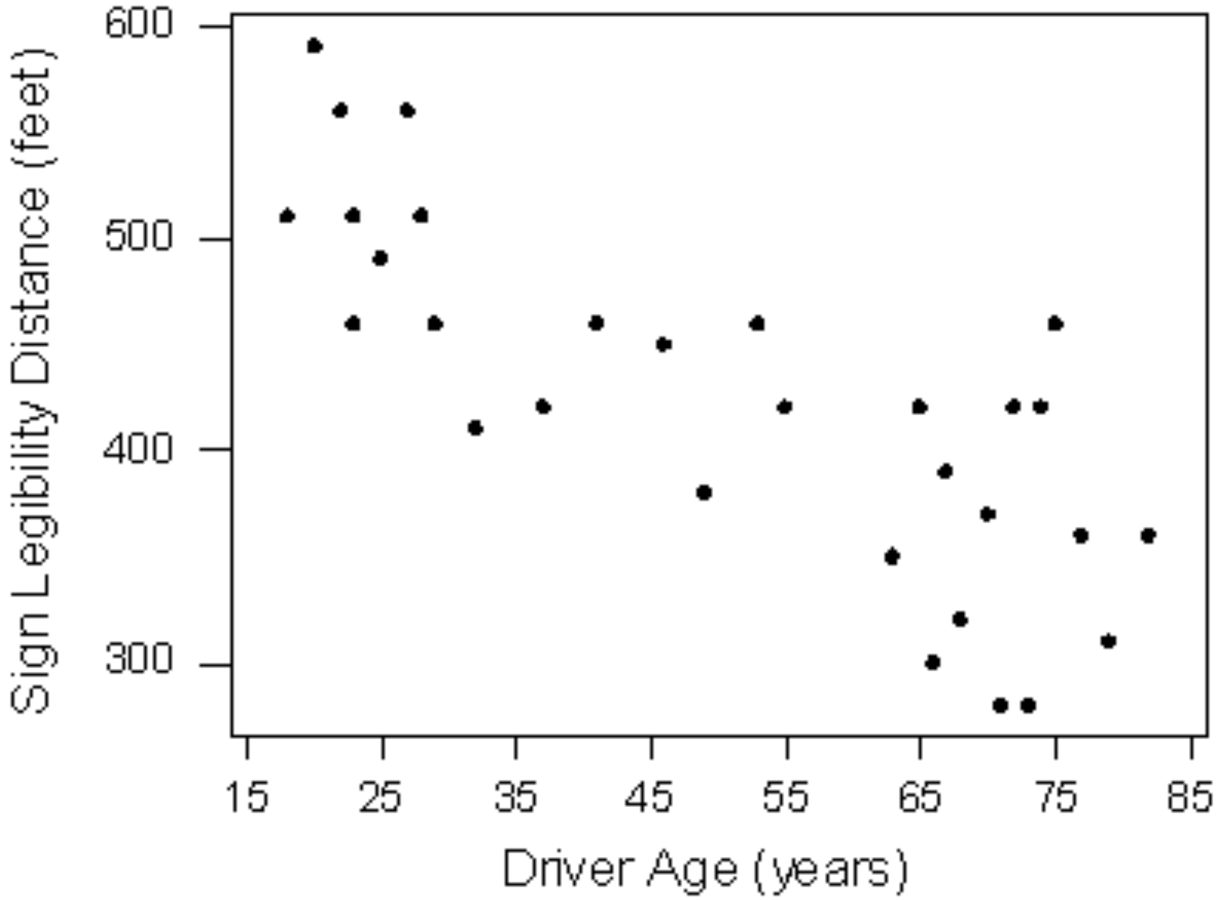
CORRELATION AND REGRESSION

# Characterizing bivariate relationships

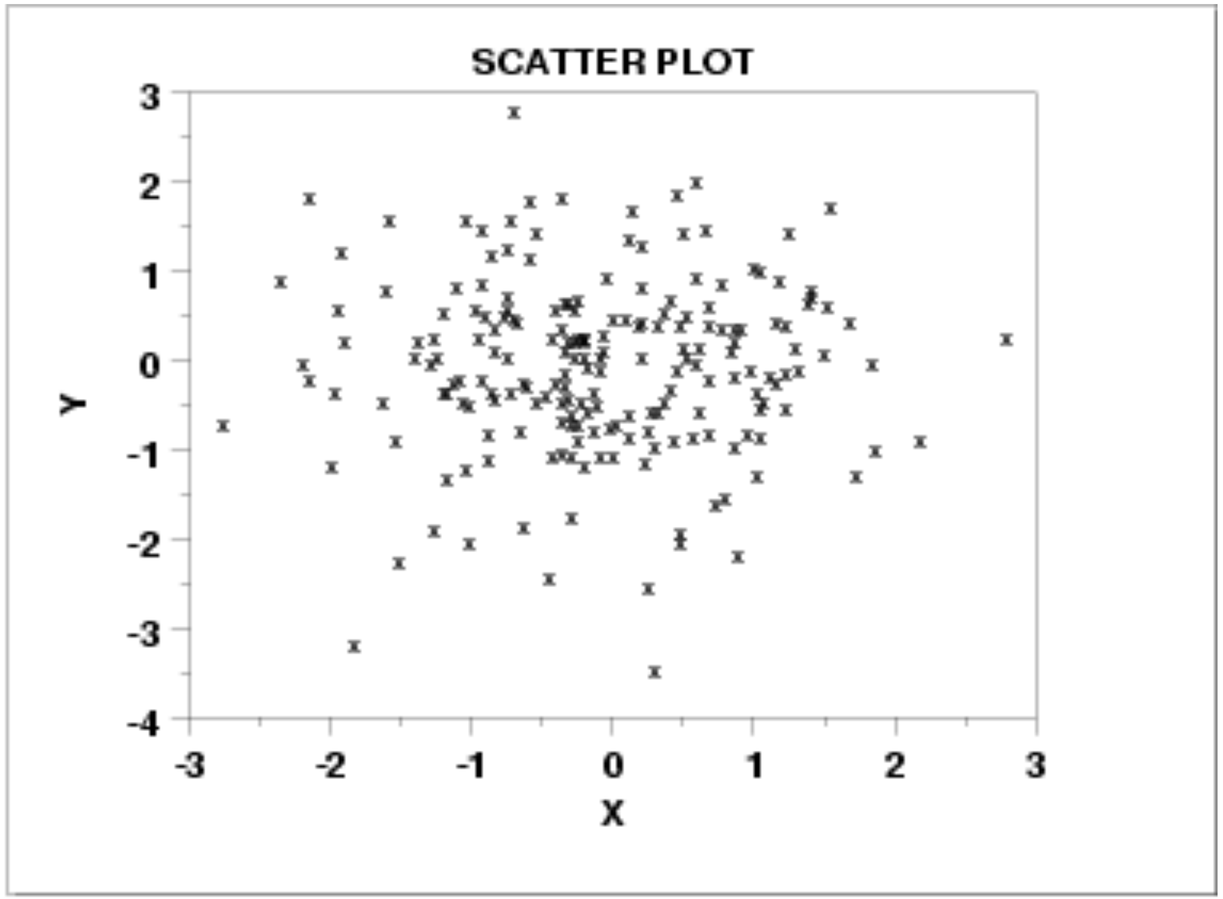
# Characterizing bivariate relationships

- Form (e.g. linear, quadratic, non-linear)
- Direction (e.g. positive, negative)
- Strength (how much scatter/noise?)
- Outliers

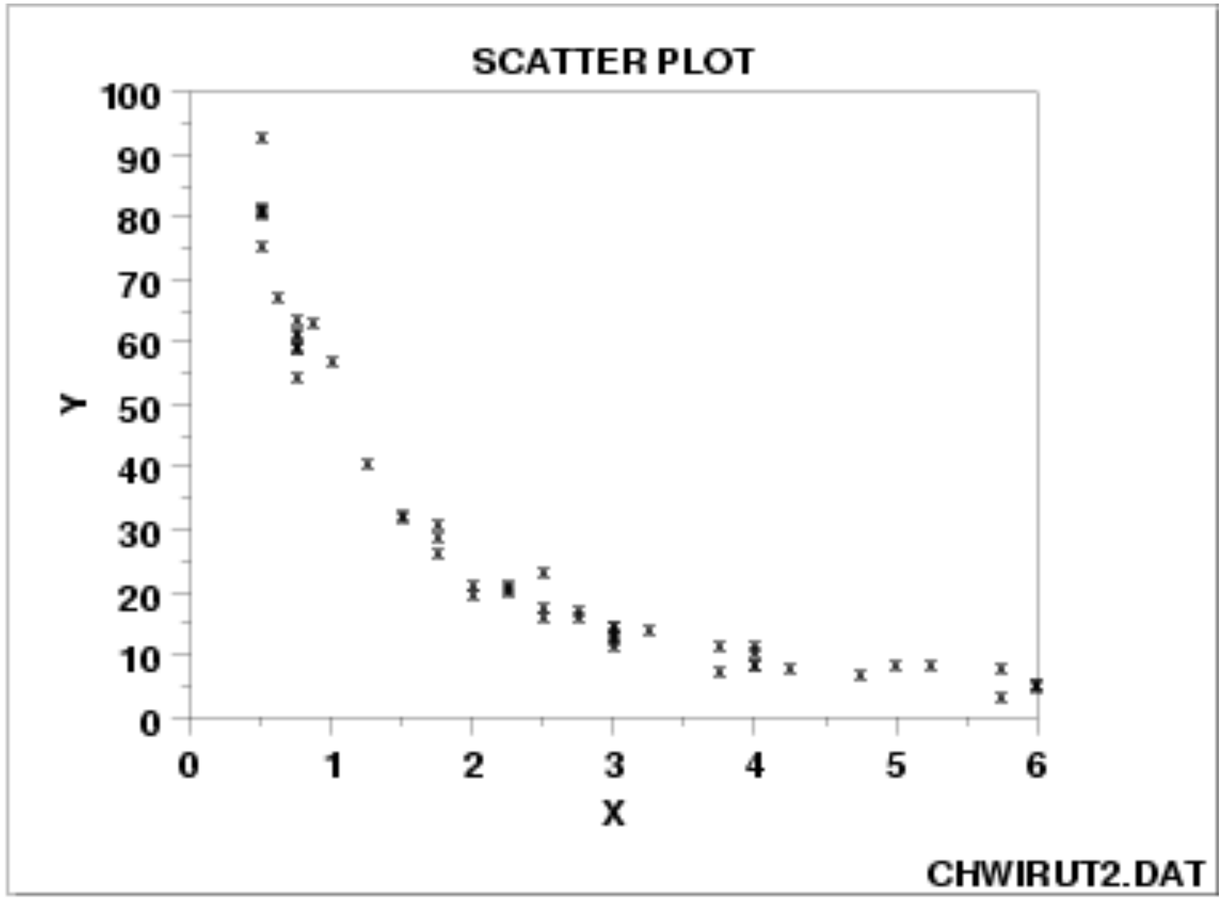
# Sign legibility



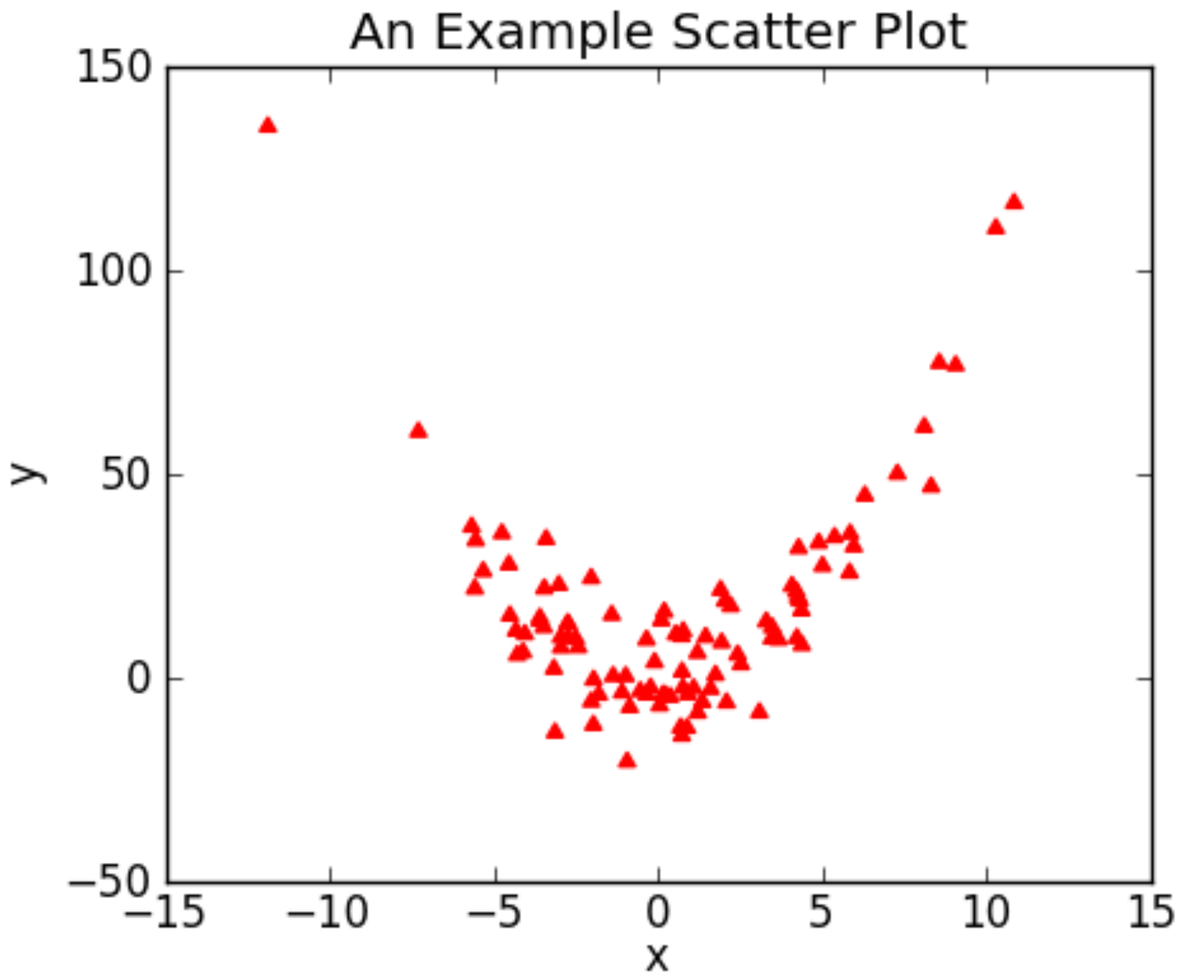
# NIST



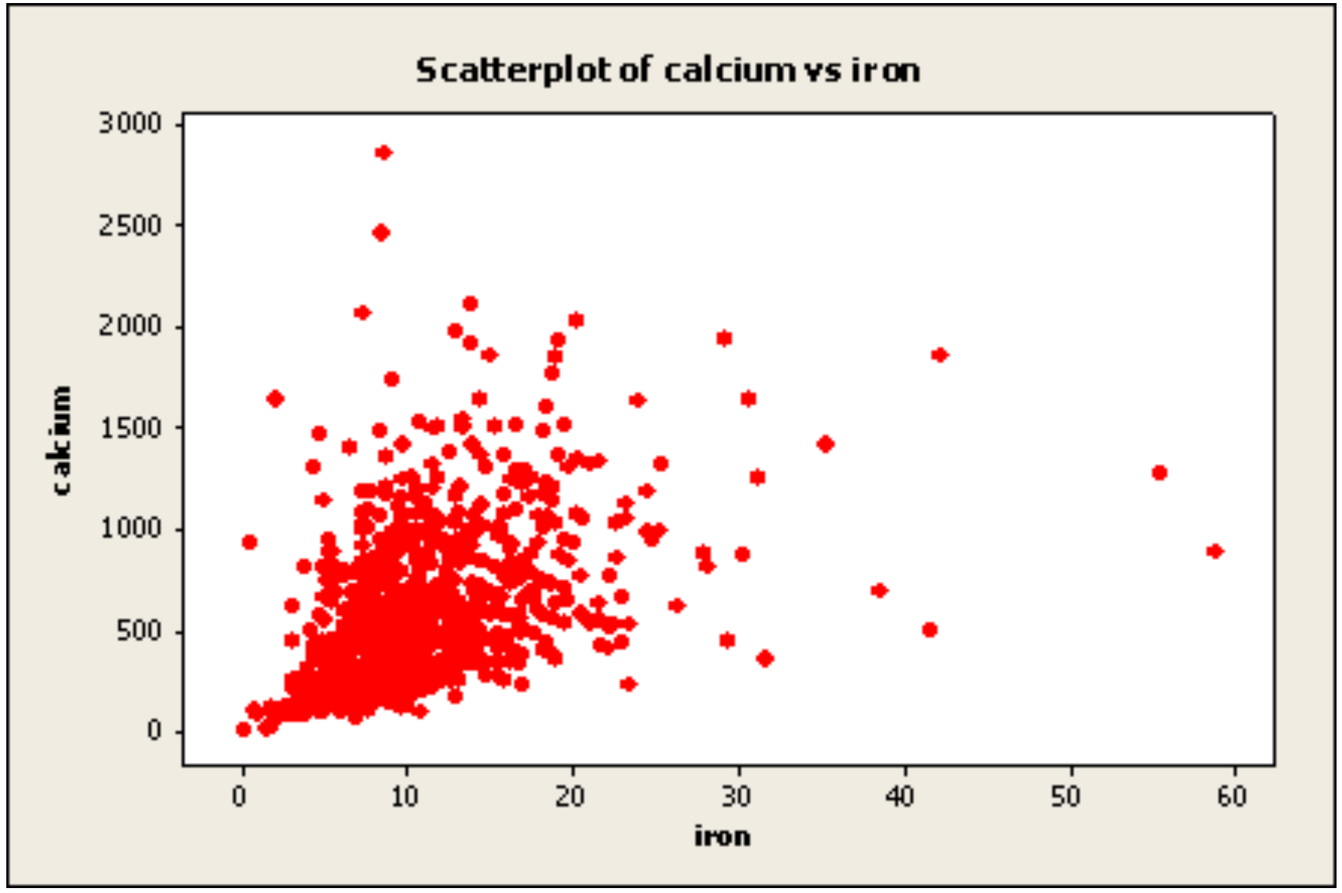
# NIST 2


SCATTER PLOT
100
90
80
70
60
50
40
30
20
10
0
Y
0
1
2
3
4
5
6
X
CHWIRUT2.DAT

# Non-linear



# Fan shape


Scatterplot of calcium vs iron
calcium
iron
3000
2500
2000
1500
1000
500
0
0
10
20
30
40
50
60

CORRELATION AND REGRESSION

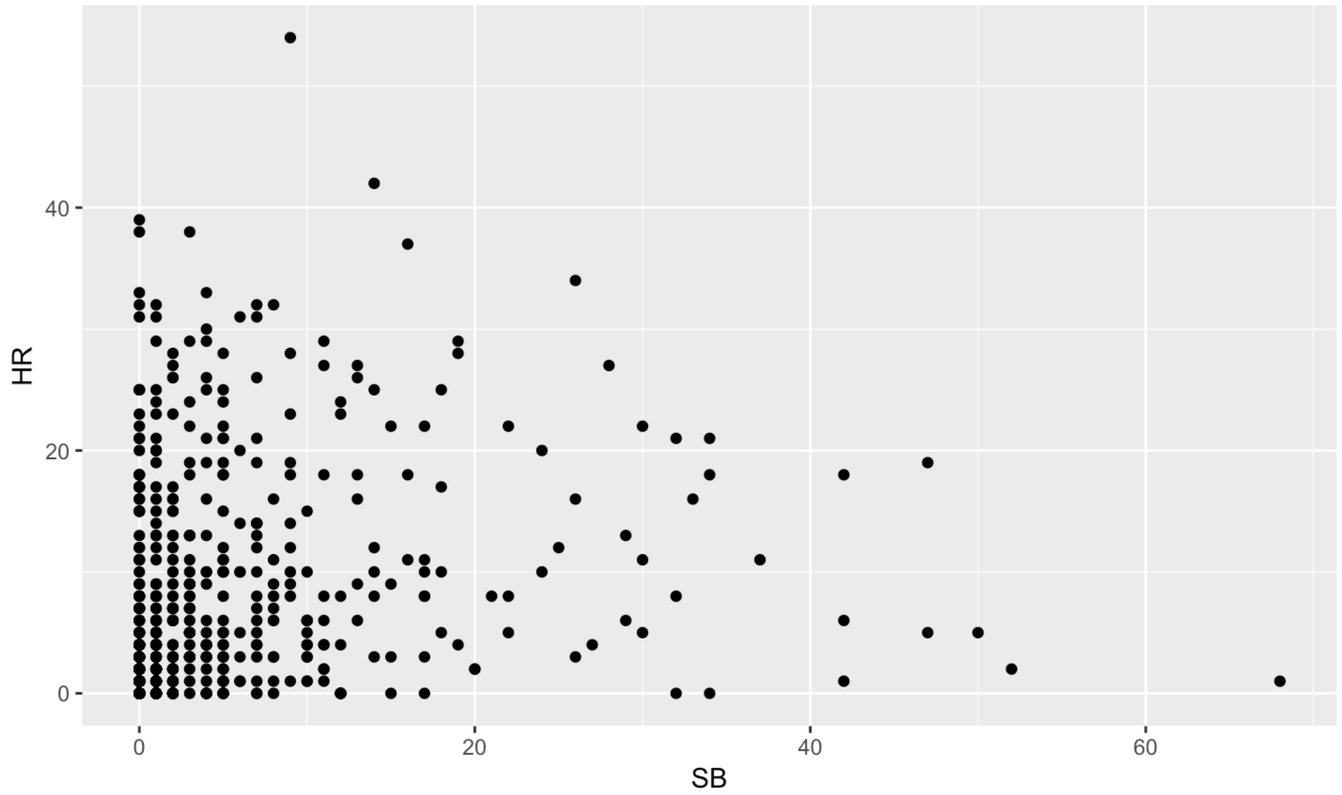
# Let’s practice!

CORRELATION AND REGRESSION

# Outliers

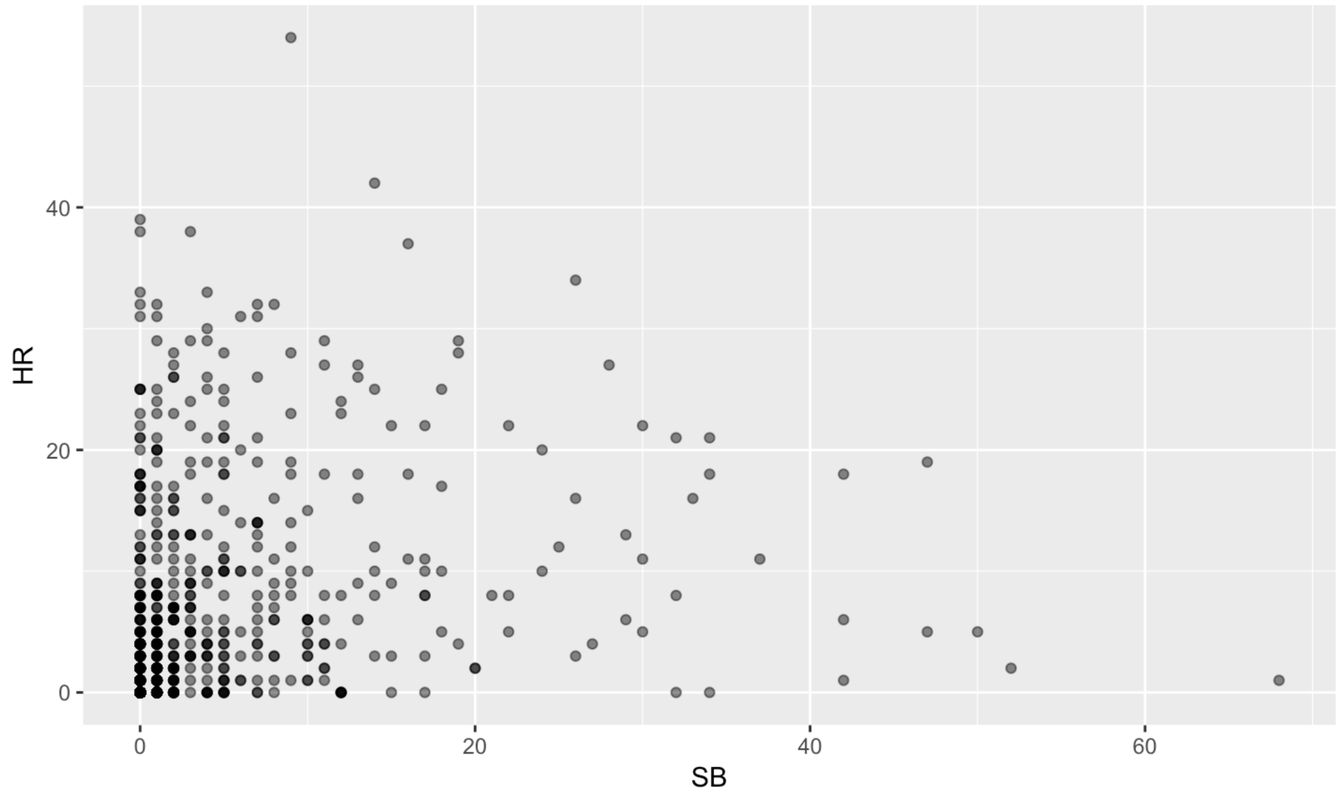
# Outliers

```
> ggplot(data = mlbBat10, aes(x = SB, y = HR)) +
    geom_point()
```



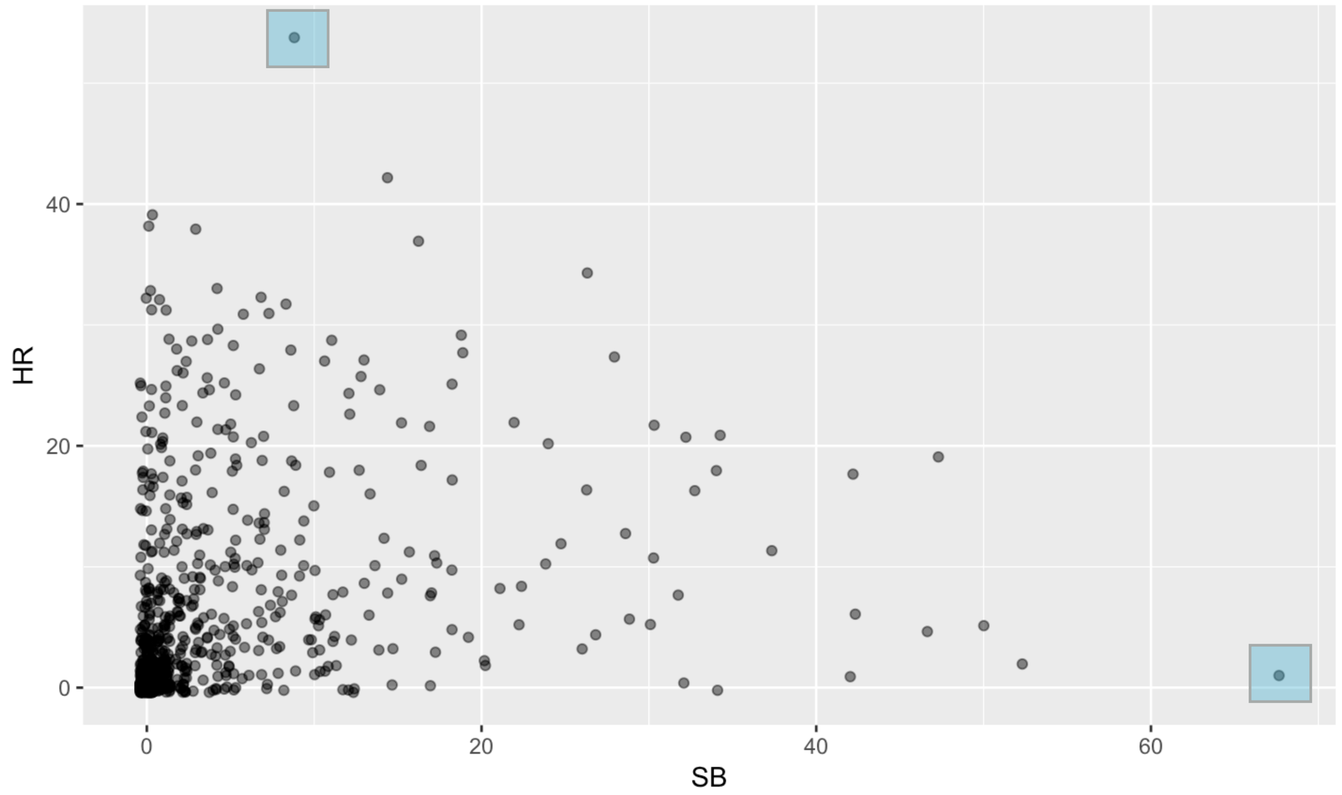
# Add transparency

```
> ggplot(data = mlbBat10, aes(x = SB, y = HR)) +
    geom_point(alpha = 0.5)
```



# Add some jitter

```
> ggplot(data = mlbBat10, aes(x = SB, y = HR)) +
    geom_point(alpha = 0.5, position = "jitter")
```



# Identify the outliers

```
> mlbBat10 %>%
    filter(SB > 60 | HR > 50) %>%
    select(name, team, position, SB, HR)

  ##          name team position SB HR
  ## 1   J Pierre  CWS       OF 68  1
  ## 2 J Bautista  TOR       OF  9 54
```

CORRELATION AND REGRESSION

# Let’s practice!

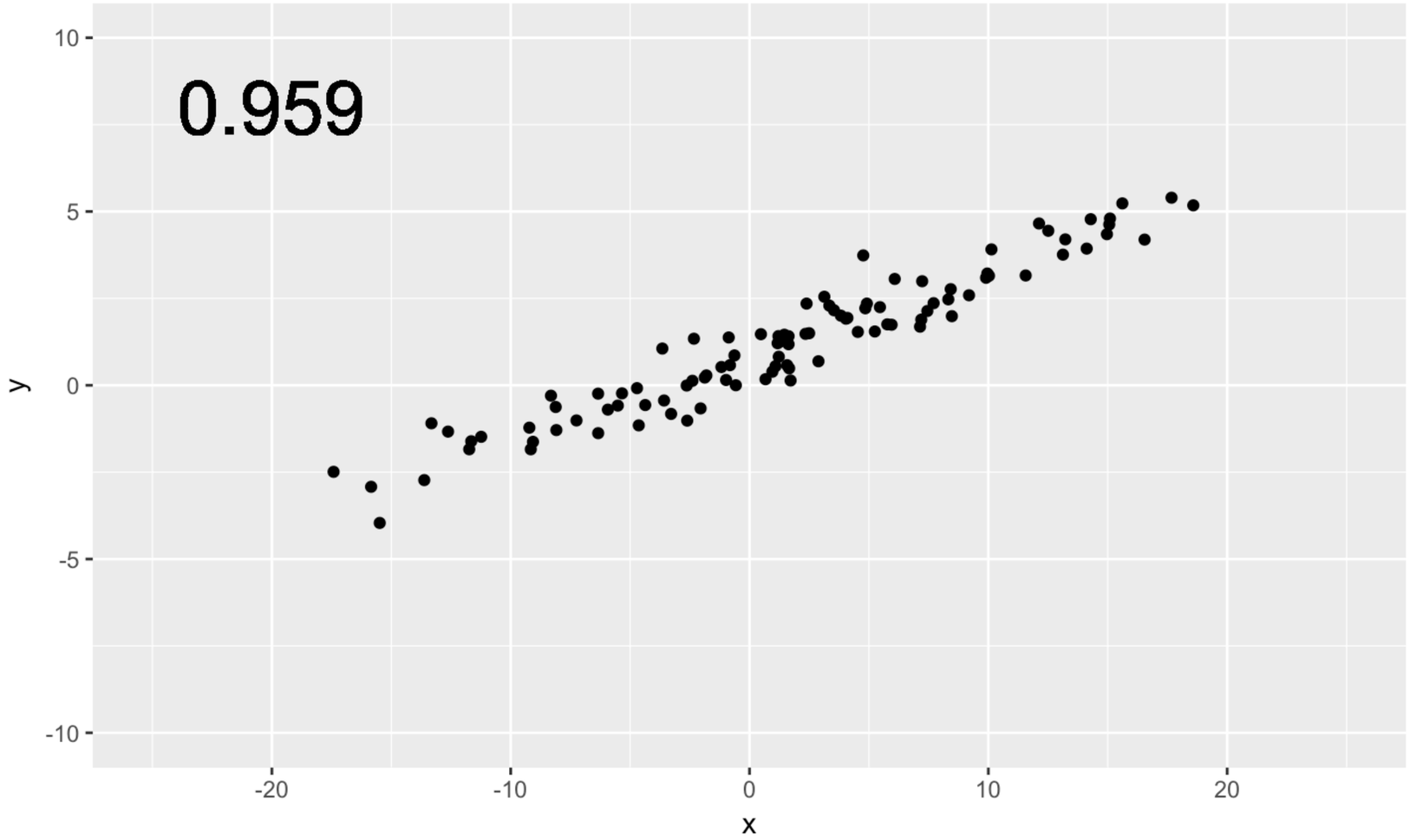
CORRELATION AND REGRESSION

# Correlation

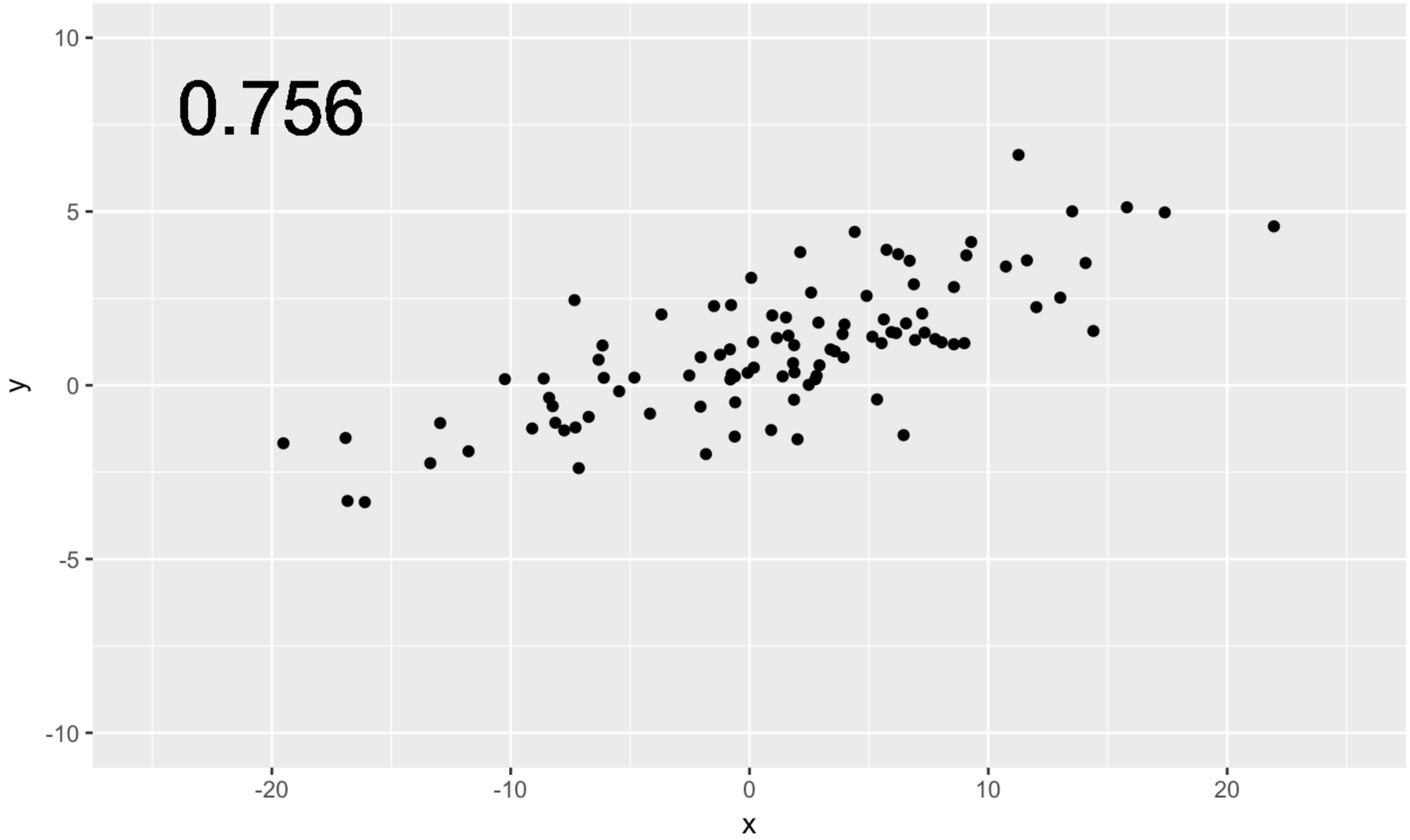
# Correlation

- Correlation coefficient between -1 and 1
- Sign —> direction
- Magnitude —> strength

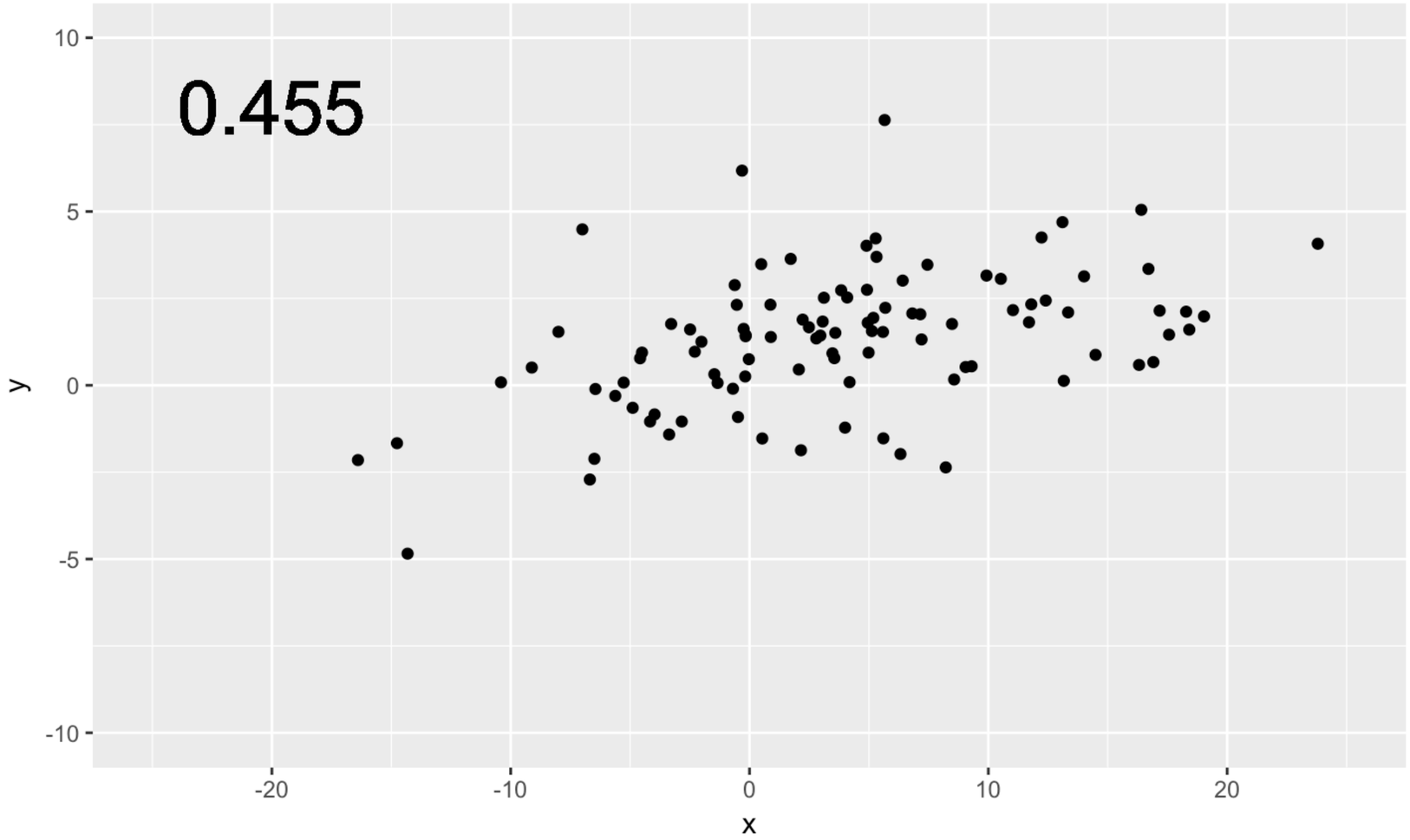
# Near perfect correlation



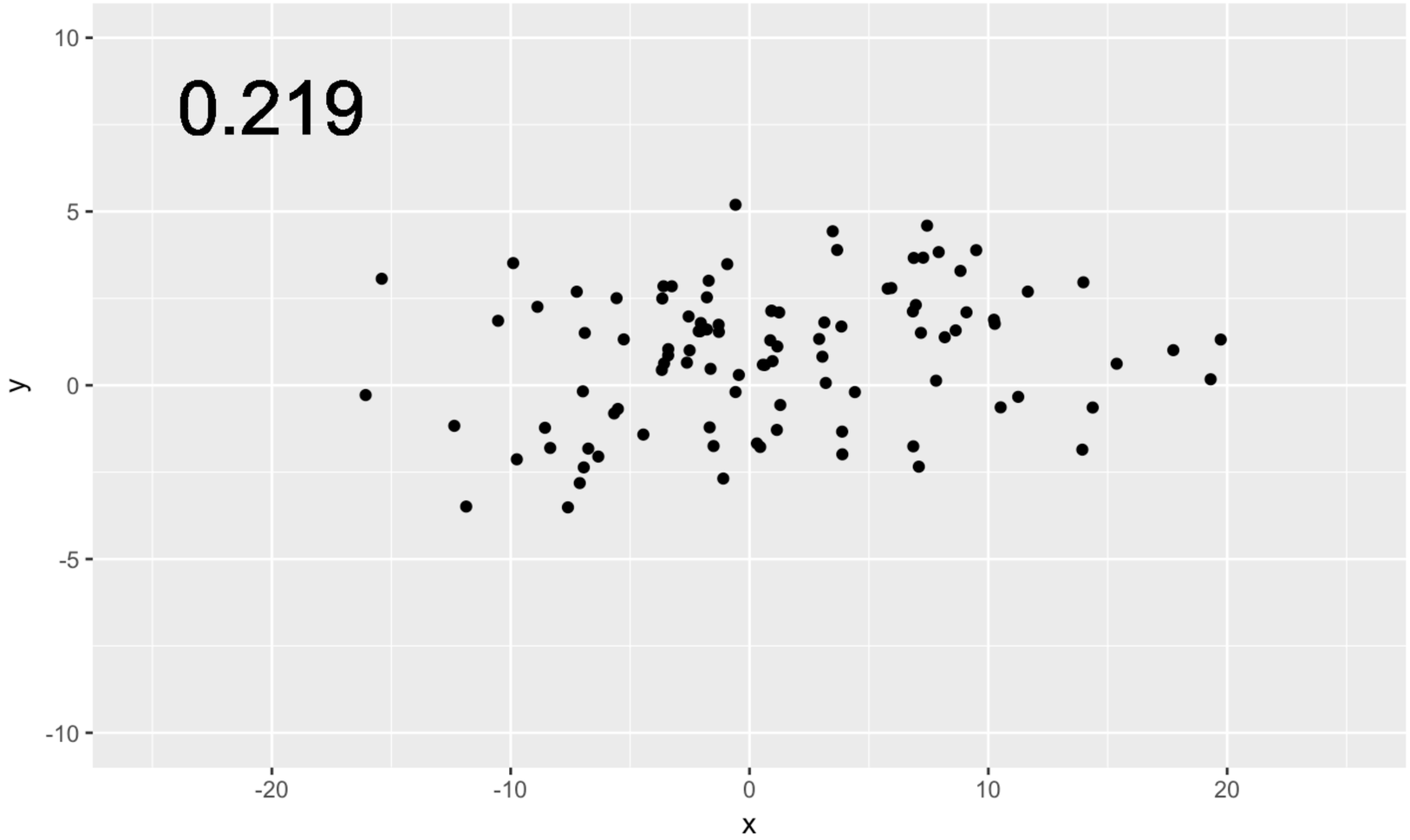
# Strong



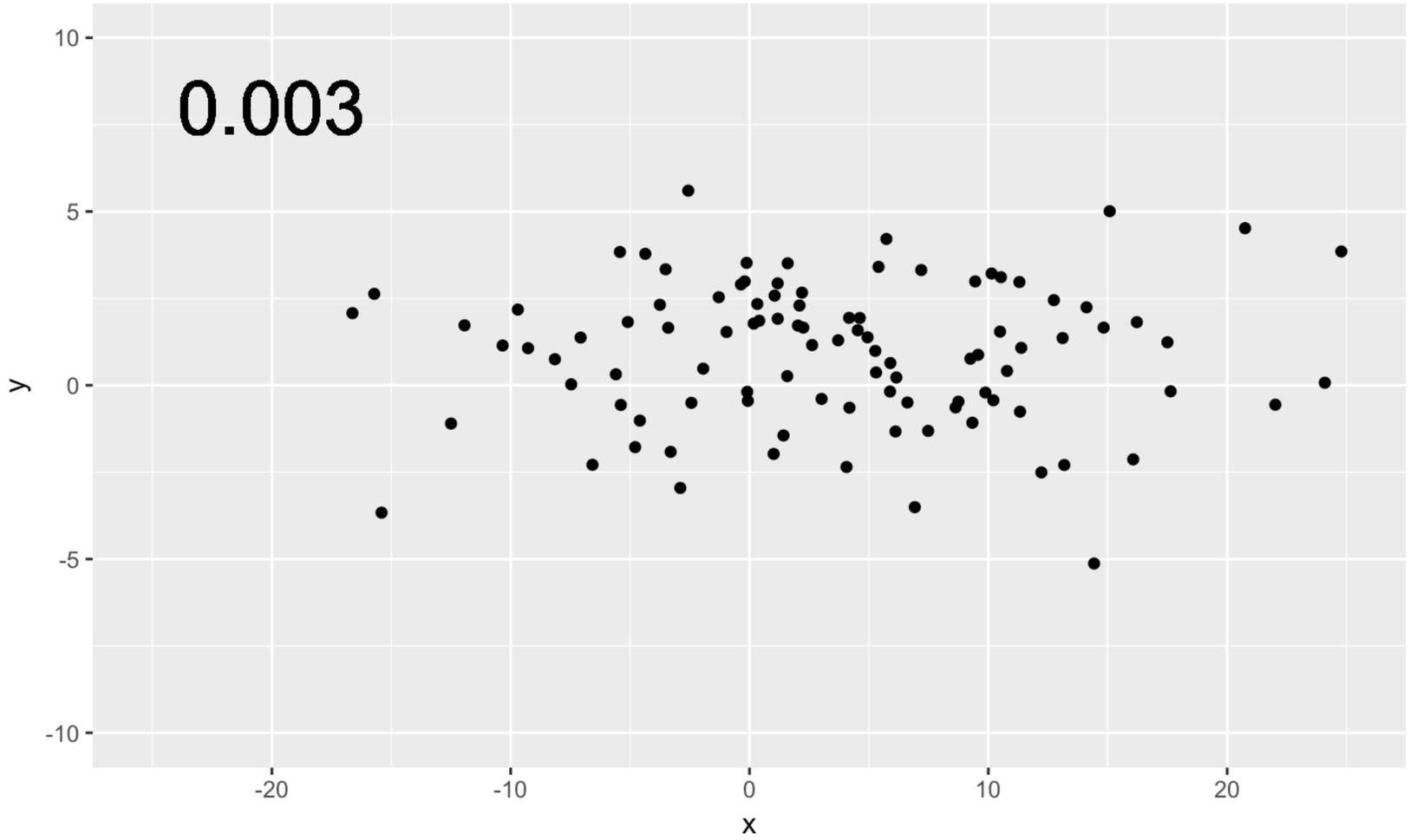
# Moderate



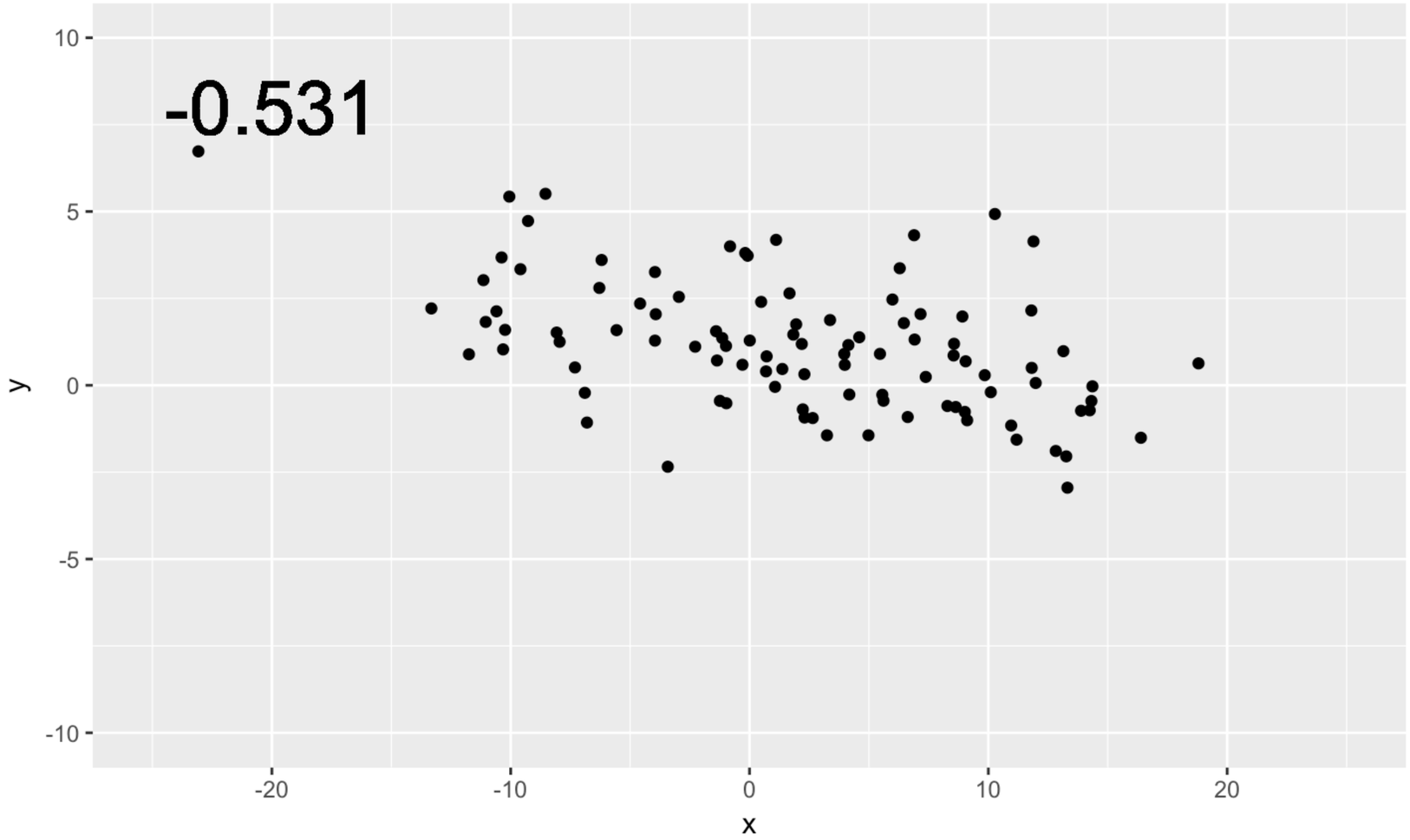
# Weak



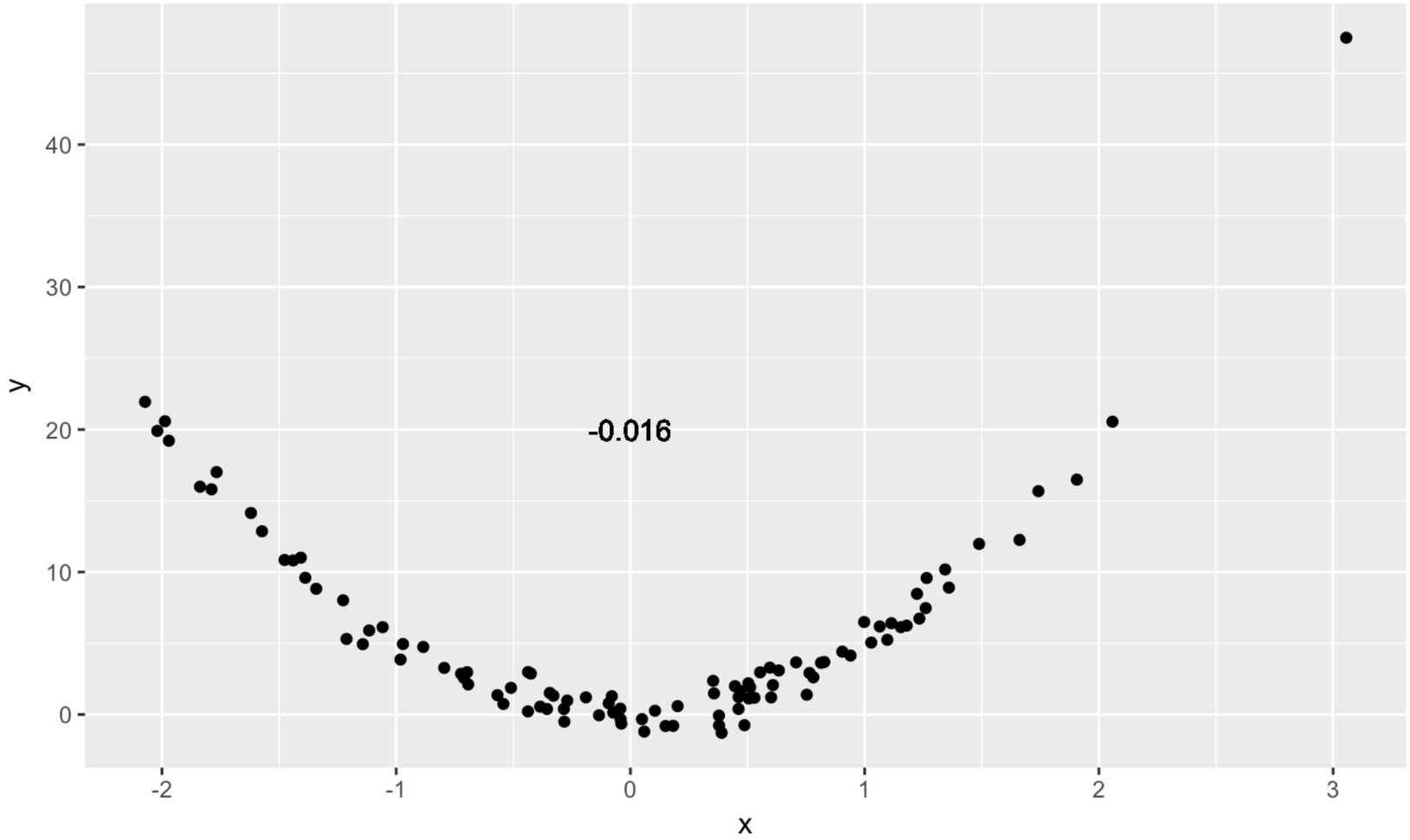
# Zero



# Negative

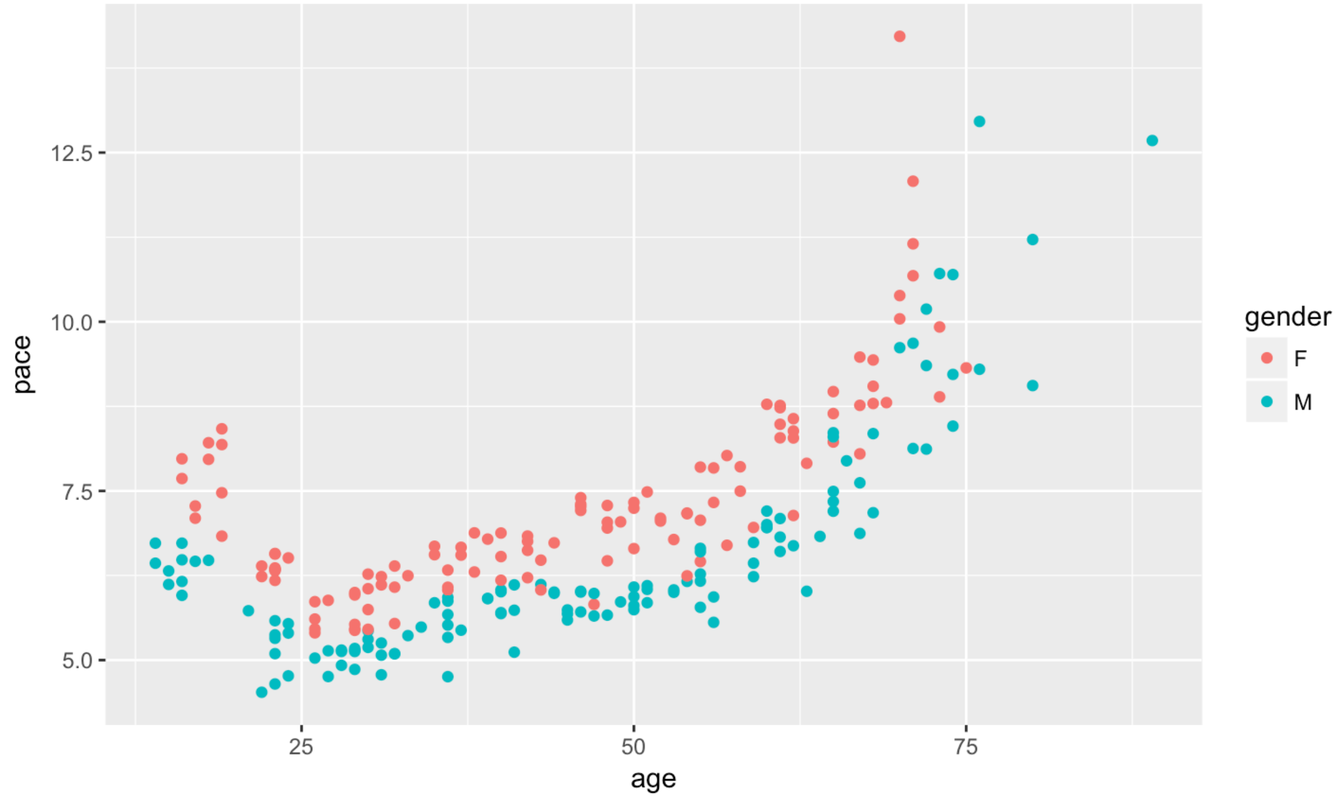


# Non-linear



# Non-linear correlation

```
> run10 %>%
    filter(divPlace <= 10) %>%
    ggplot(aes(x = age, y = pace, color = gender)) +
    geom_point()
```



# Pearson product-moment correlation

$$r(x, y) = \frac{Cov(x, y)}{\sqrt{SXX \cdot SYY}}$$

# Pearson product-moment correlation

$$r(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}$$

CORRELATION AND REGRESSION

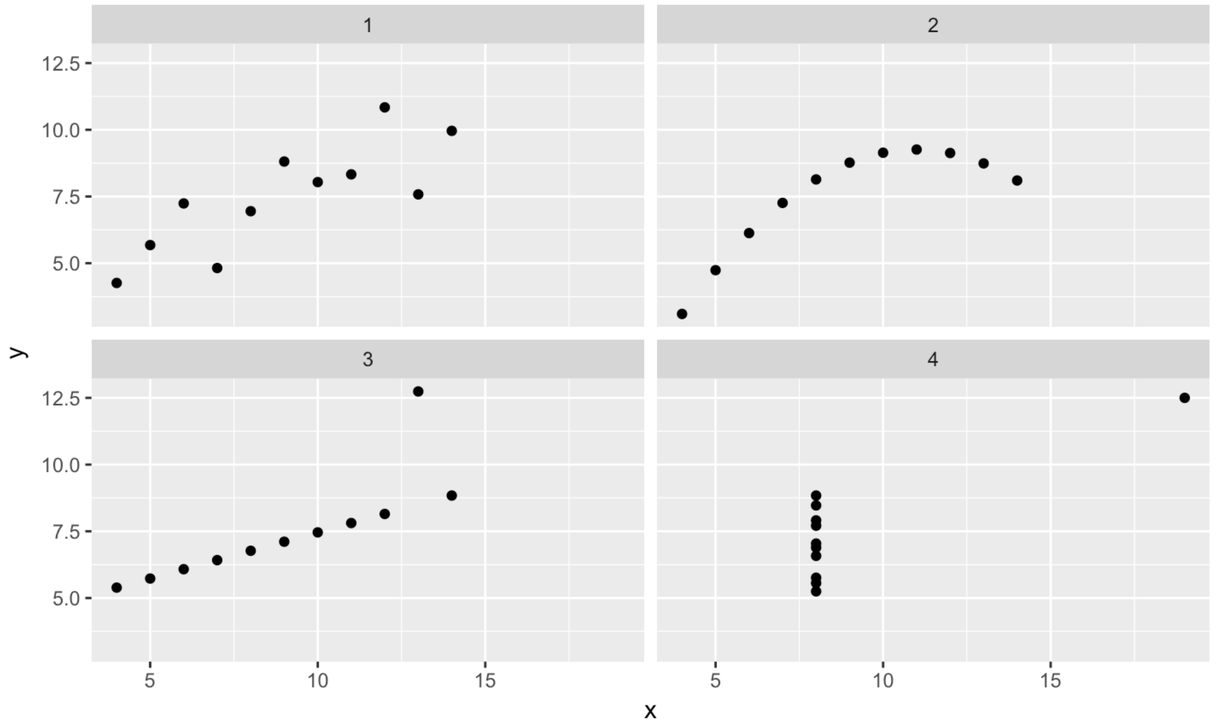
# Let’s practice!

CORRELATION AND REGRESSION

# Correlation

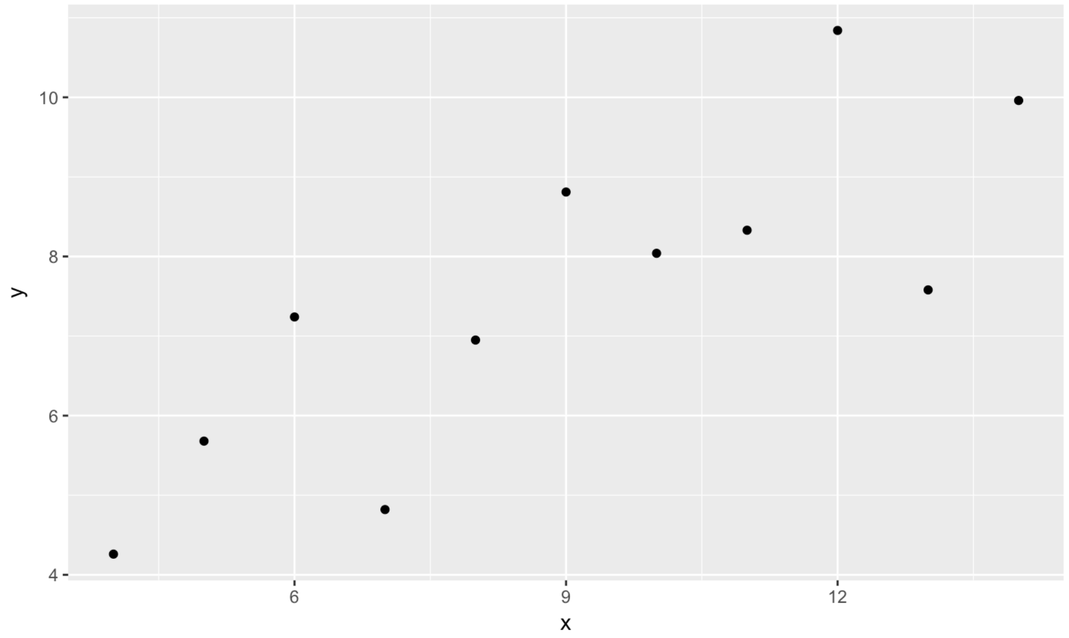
# Anscombe

```
> ggplot(data = Anscombe, aes(x = x, y = y)) +
    geom_point() +
    facet_wrap(~ set)
```



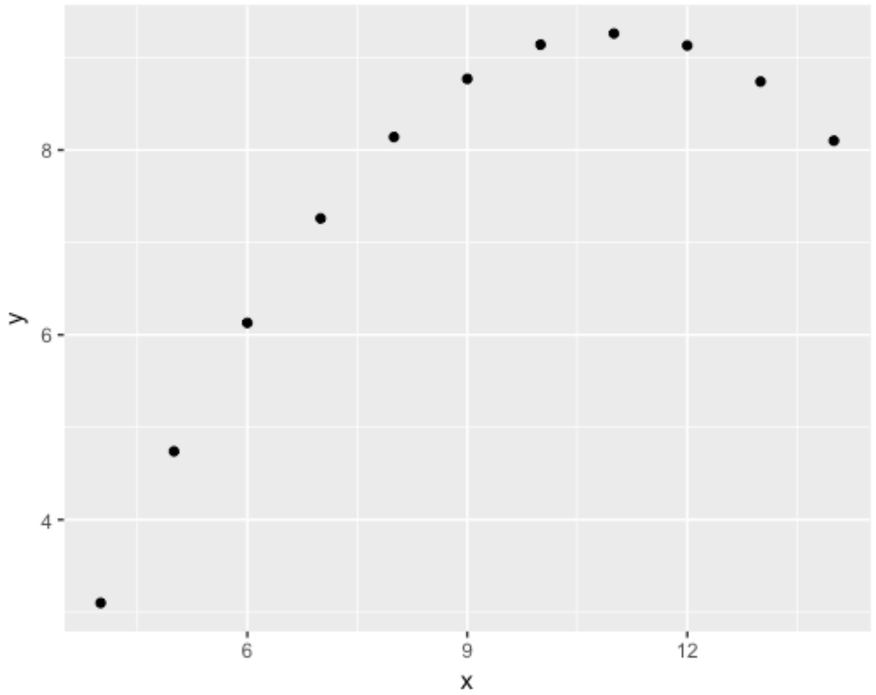
# Anscombe 1

```
> Anscombe %>%
    filter(set == 1) %>%
    ggplot(aes(x = x, y = y)) +
    geom_point()
```



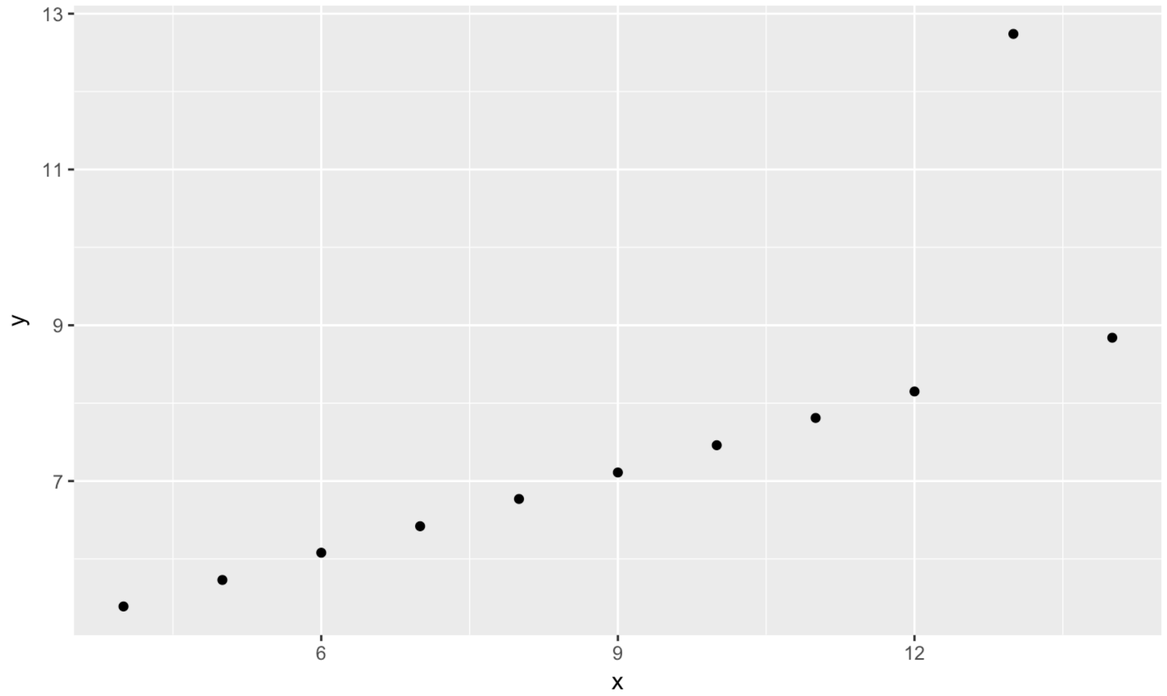
# Anscombe 2

```
> Anscombe %>%
    filter(set == 2) %>%
    ggplot(aes(x = x, y = y)) +
    geom_point()
```



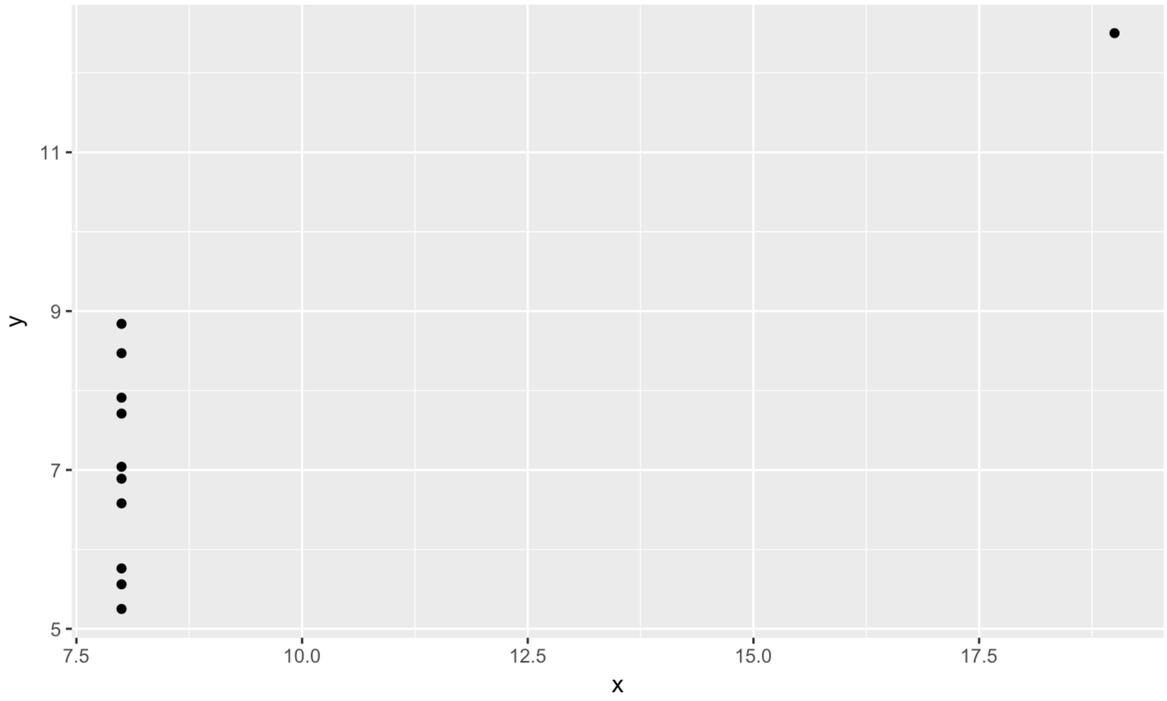
# Anscombe 3

```
> Anscombe %>%
    filter(set == 3) %>%
    ggplot(aes(x = x, y = y)) +
    geom_point()
```



# Anscombe 4

```
> Anscombe %>%
    filter(set == 4) %>%
    ggplot(aes(x = x, y = y)) +
    geom_point()
```



CORRELATION AND REGRESSION

# Let’s practice!

CORRELATION AND REGRESSION

# Interpretation of Correlation

# Exercise and beer

Getty Images

Email
Share
Tweet
Save
More

For many people, working out and alcohol are closely linked. Sports teams and training partners celebrate victories, bemoan defeats or mark the end of training sessions with a beer or three. Beer, in fact, provides a substantial portion of some exercisers' fluid intake after workouts.

But whether exercise encourages people to drink and, likewise, whether drinking encourages people to exercise has been in dispute.

Now two new studies suggest that exercise may well influence when and how much people drink. Drinking may even affect whether people exercise, and, the findings suggest, the interplay between

*Source: http://well.blogs.nytimes.com/2015/12/02/the-close-ties-between-exercise-and-beer/*

# Exercise and beer

Past epidemiological studies have shown that people who exercise tend numerically also to be people who drink, and vice versa. In a typical study from 2001, for example, researchers found that men and women who qualified as moderate drinkers, meaning they downed about a drink a day, were twice as likely to exercise regularly as teetotalers.

**PHYS ED**
Gretchen Reynolds on the science of fitness.

But most of these previous studies had limitations. They relied, for instance, on people's ability to recall their exercise and drinking habits over the course of, say, the past year, which can be notoriously unreliable. They also rarely took into account participants' ages and gender, which affect how much people exercise and drink.

And perhaps most problematic, these past studies rarely

# Exercise and beer

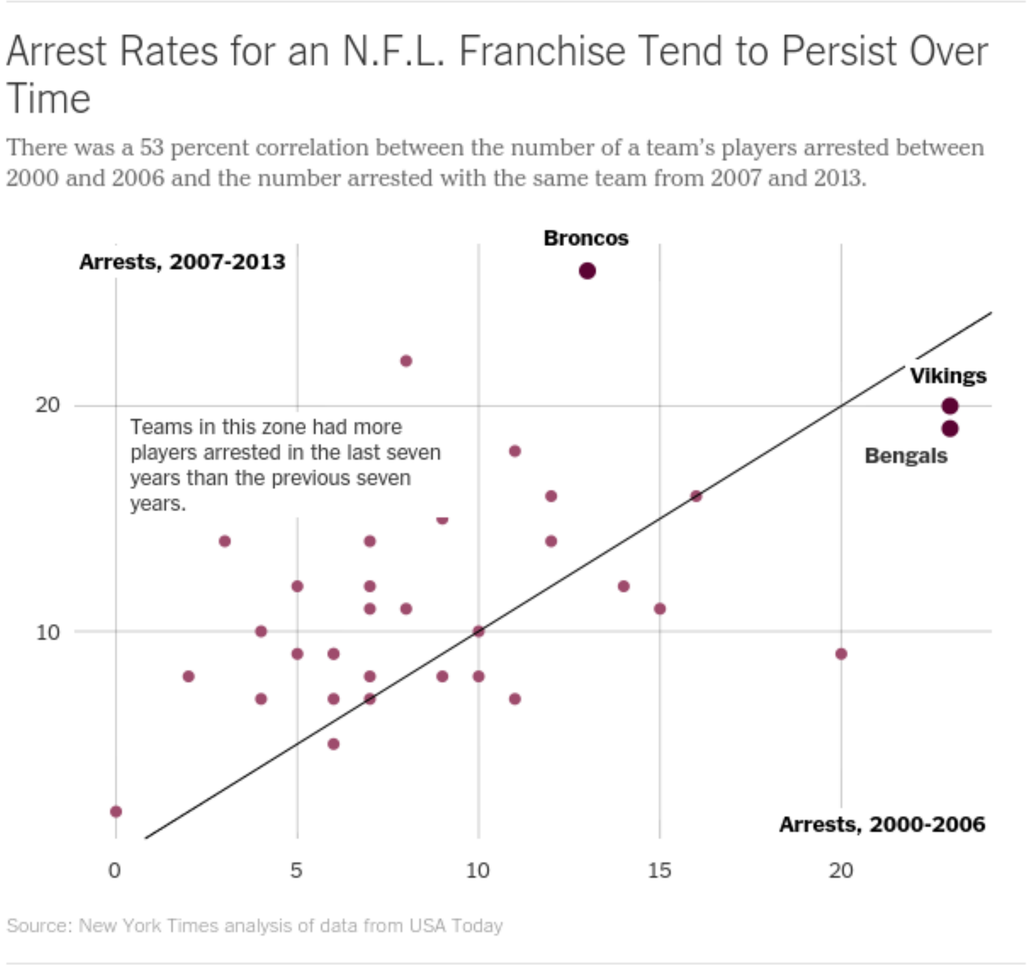
write in their study, which was published recently in Health Psychology, "people drank more than usual on the same days that they engaged in more physical activity than usual."

This relationship held true throughout all seasons of the year and whether someone was a man or a woman, a collegian or a retiree. Age and gender did not affect the results.

Thankfully, the data did not show that exercise incited or exacerbated problem drinking. Only very rarely during the study did anyone report drinking heavily, which the researchers defined as downing more than four drinks in succession for a woman and five for a man.

But of course this kind of epidemiological study cannot determine why working out and drinking should be associated at all, which makes the second study, a newly published review of past, related experiments, especially those involving animals, so compelling.

# NFL arrests




*Source: https://www.nytimes.com/2014/09/13/upshot/what-the-numbers-show-about-nfl-player-arrests.html*

# NFL arrests

But there's a simple way to test that. If the results were random, you would expect there to be no correlation between the number of player arrests in one time period with a subsequent time period. You could even imagine a negative correlation, if teams that had a run of players getting in trouble took extra care not to sign players reputed to have character issues.

But that is not what happened over the last 14 years. If you chart the number of arrests of players from each franchise in the first seven years of the data, 2000 to 2006, versus the number of arrests that franchise experienced from 2007 to 2013, the correlation is a pretty solid 53 percent. A scatter plot shows a clear pattern in which those franchises with high numbers of arrests in the early years also tended to have high number arrests in later years and vice versa.

The data don't tell us anything about why these patterns are so persistent,

# Correlation vs. regression

By JONATHAN WEISMAN NOV. 1, 2012

WASHINGTON — The Congressional Research Service has withdrawn an economic report that found no correlation between top tax rates and economic growth, a central tenet of conservative economic theory, after Senate Republicans raised concerns about the paper's findings and wording.

The decision, made in late September against the advice of the agency's economic team leadership, drew almost no notice at the time. Senator Charles E. Schumer, Democrat of New York, cited the study a week and a half after it was withdrawn in a speech on tax policy at the National Press Club.

*Source: http://www.nytimes.com/2012/11/02/business/questions-raised-on-withdrawal-of-congressional-research-services-report-on-tax-rates.html*

# Can you plot a correlation?

They plotted a correlation between various types of abuse – including racism, misogyny and homophobia – and political positions including voting to Leave:

**Correlation Overview (UK)**

| | RACISM | TRANSPHOBIA | HOMOPHOBIA | MASCULINITY | MISOGYNY | VOTED YES TO LEAVE EU | CON 2015 | UKIP | TURNOUT 2015 | LONG TERM UNEMPLOYMENT |
|---|---|---|---|---|---|---|---|---|---|---|
| RACISM | | -0.295 | 0.478 | 0.279 | 0.632 | 0.0139 | -0.0658 | -0.047 | 0.086 | 0.145 |
| TRANSPHOBIA | -0.295 | | -0.315 | -0.443 | -0.334 | 0.334 | -0.0813 | 0.095 | -0.197 | 0.152 |
| HOMOPHOBIA | 0.478 | -0.315 | | 0.157 | 0.752 | 0.138 | -0.0691 | 0.200 | -0.142 | 0.163 |
| MASCULINITY | 0.279 | -0.443 | 0.157 | | 0.376 | -0.077 | -0.251 | -0.054 | 0.128 | 0.101 |
| MISOGYNY | 0.632 | -0.334 | 0.752 | 0.376 | | 0.114 | -0.111 | 0.273 | -0.307 | 0.405 |

*Source: http://heatst.com/world/no-correlation-between-voting-for-brexit-and-racism-study-finds/*

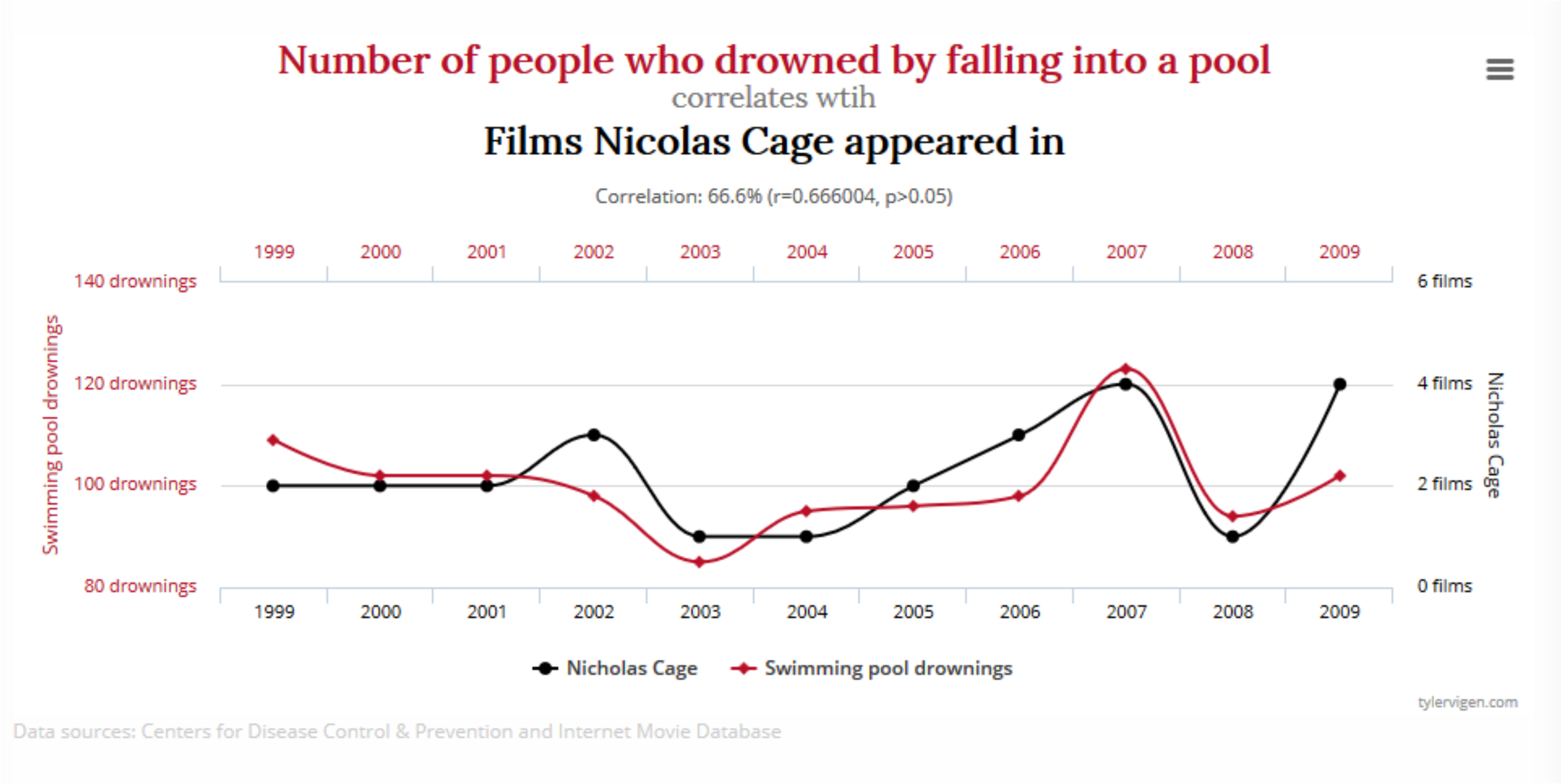
CORRELATION AND REGRESSION

# Let’s practice!

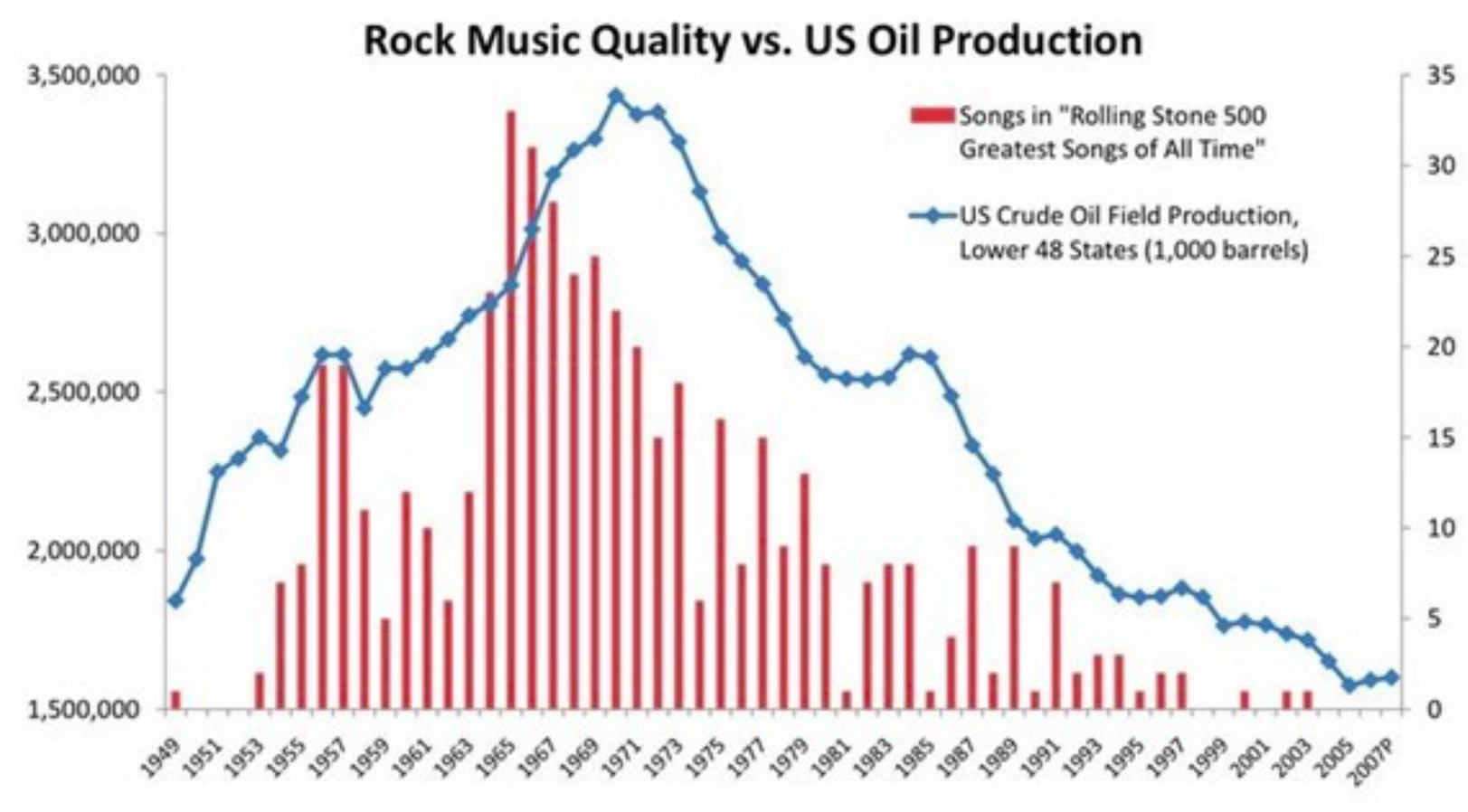


# Spurious correlations

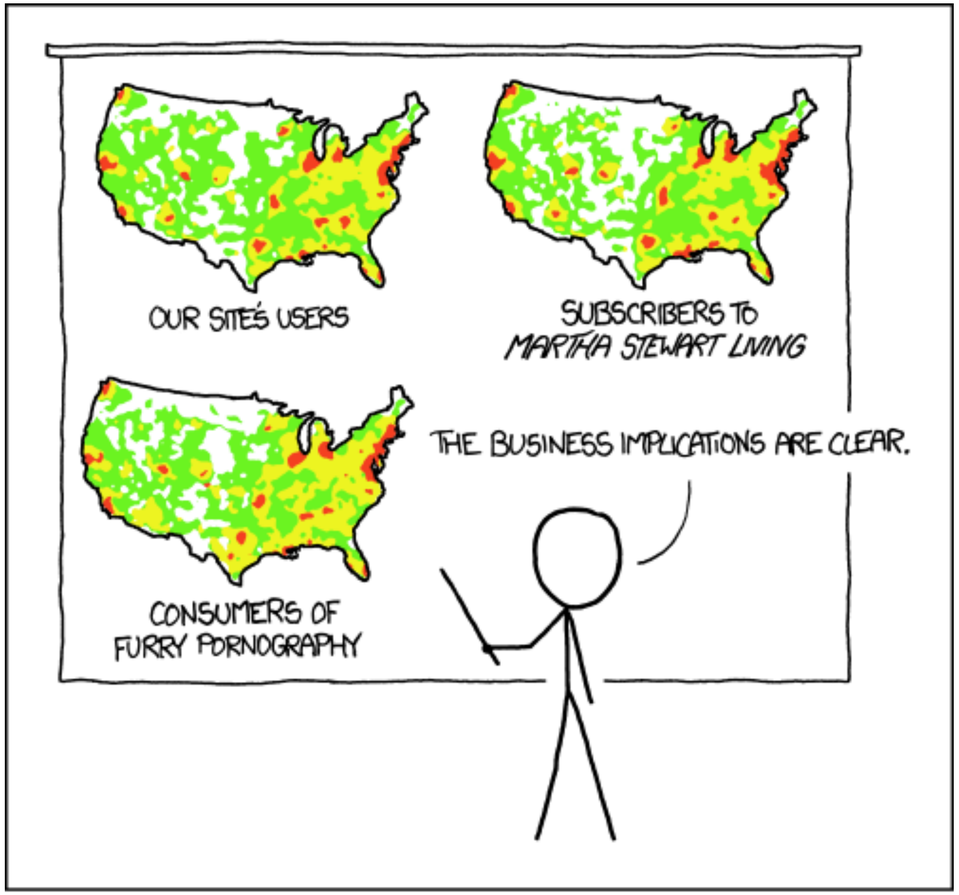
# Spurious over time



# Spurious over time

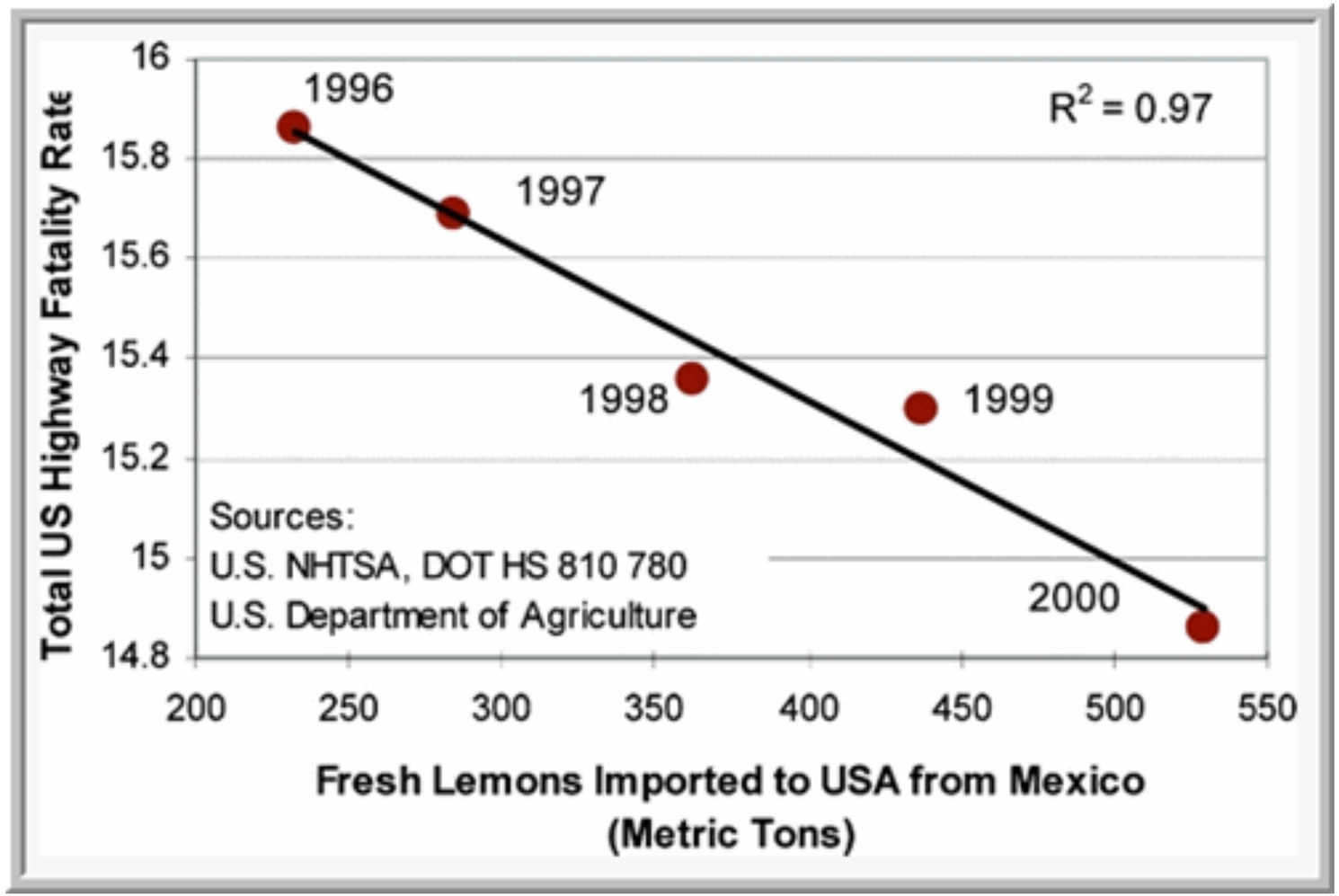


# Spurious over space




PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

# Spurious for whatever reason



CORRELATION AND REGRESSION

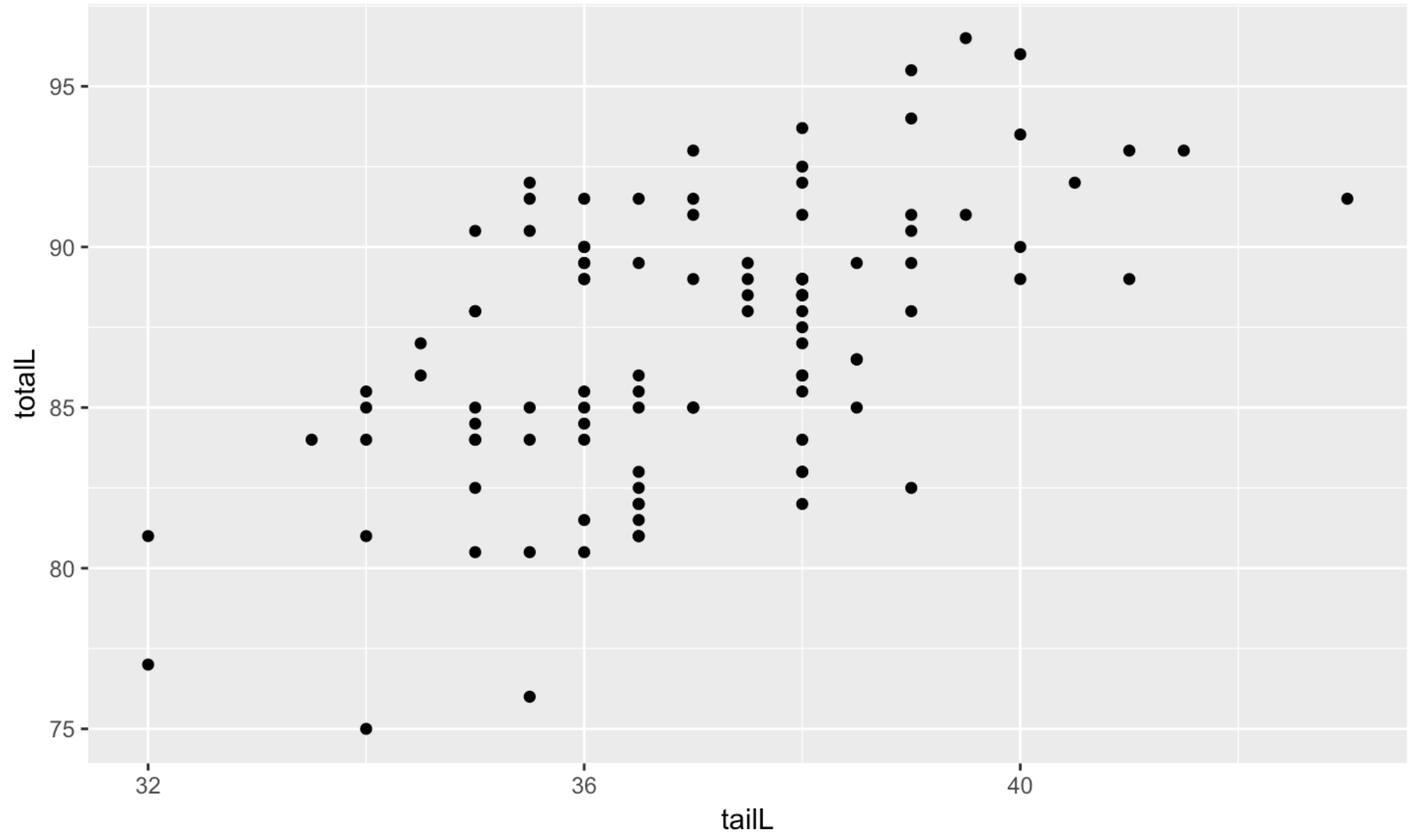
# Let’s practice!

CORRELATION AND REGRESSION

# Visualization of Linear Models

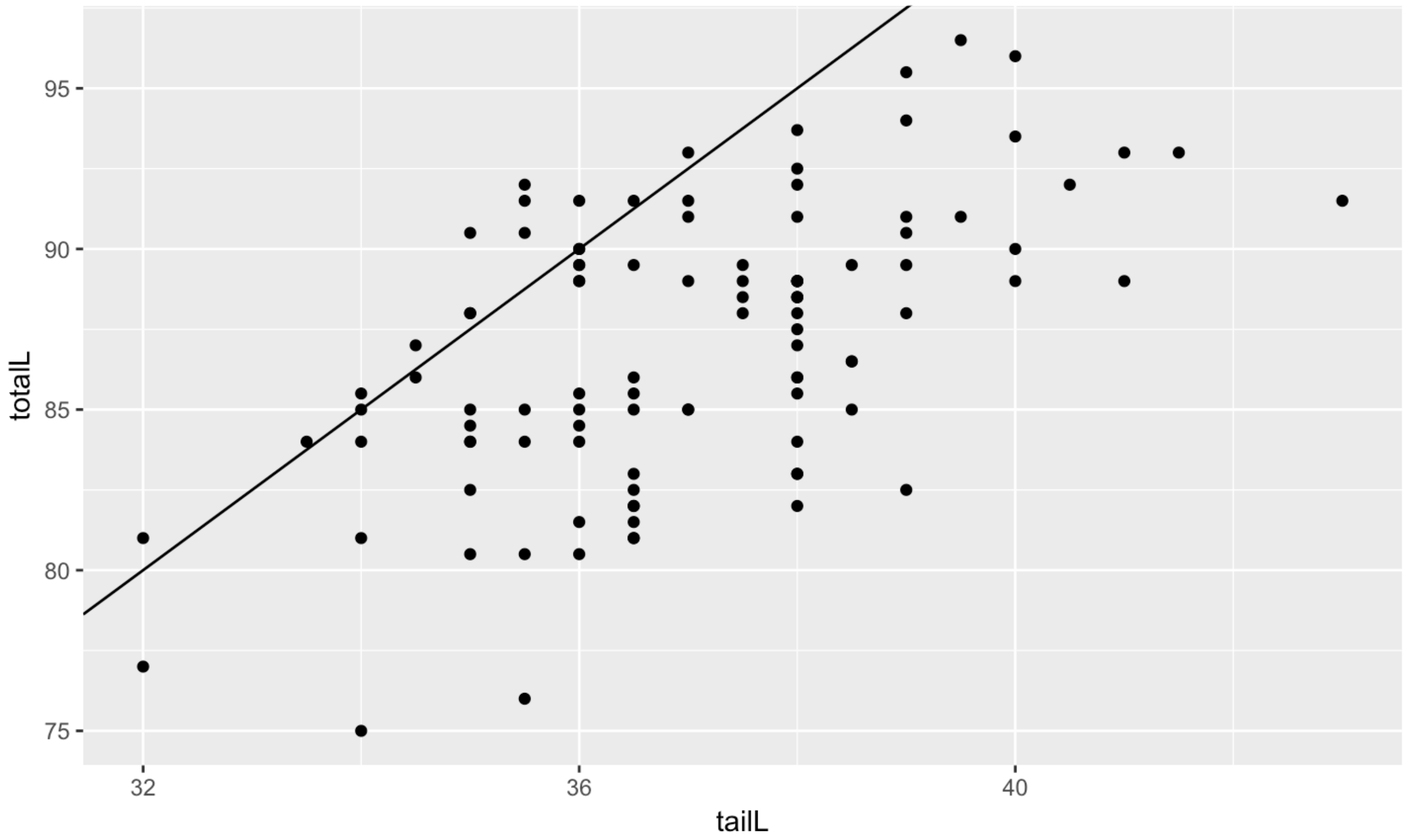
# Possums

```
> ggplot(data = possum, aes(y = totalL, x = tailL)) +
    geom_point()
```



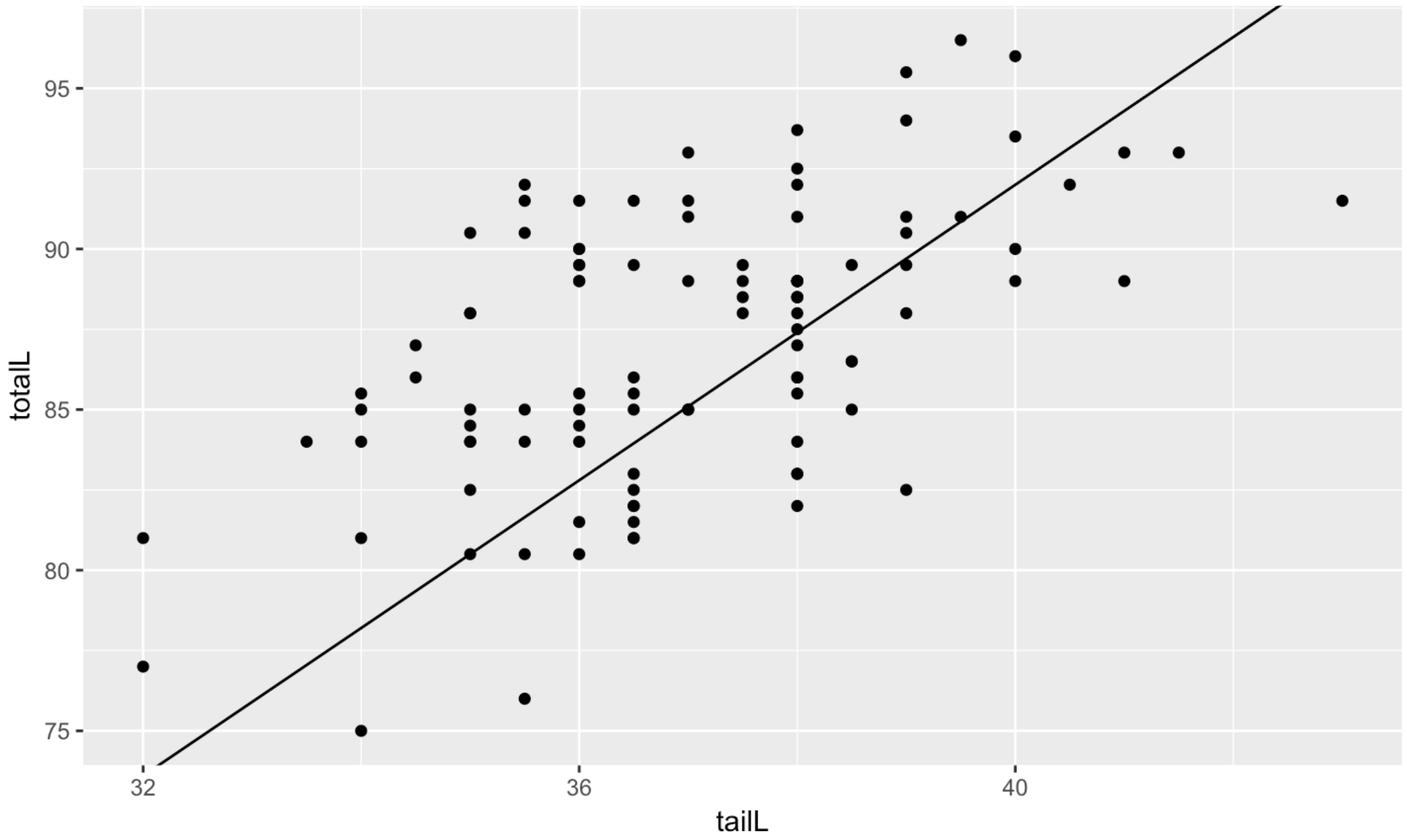
# Through the origin

```
> ggplot(data = possum, aes(y = totalL, x = tailL)) +
    geom_point() + geom_abline(intercept = 0, slope = 2.5)
```



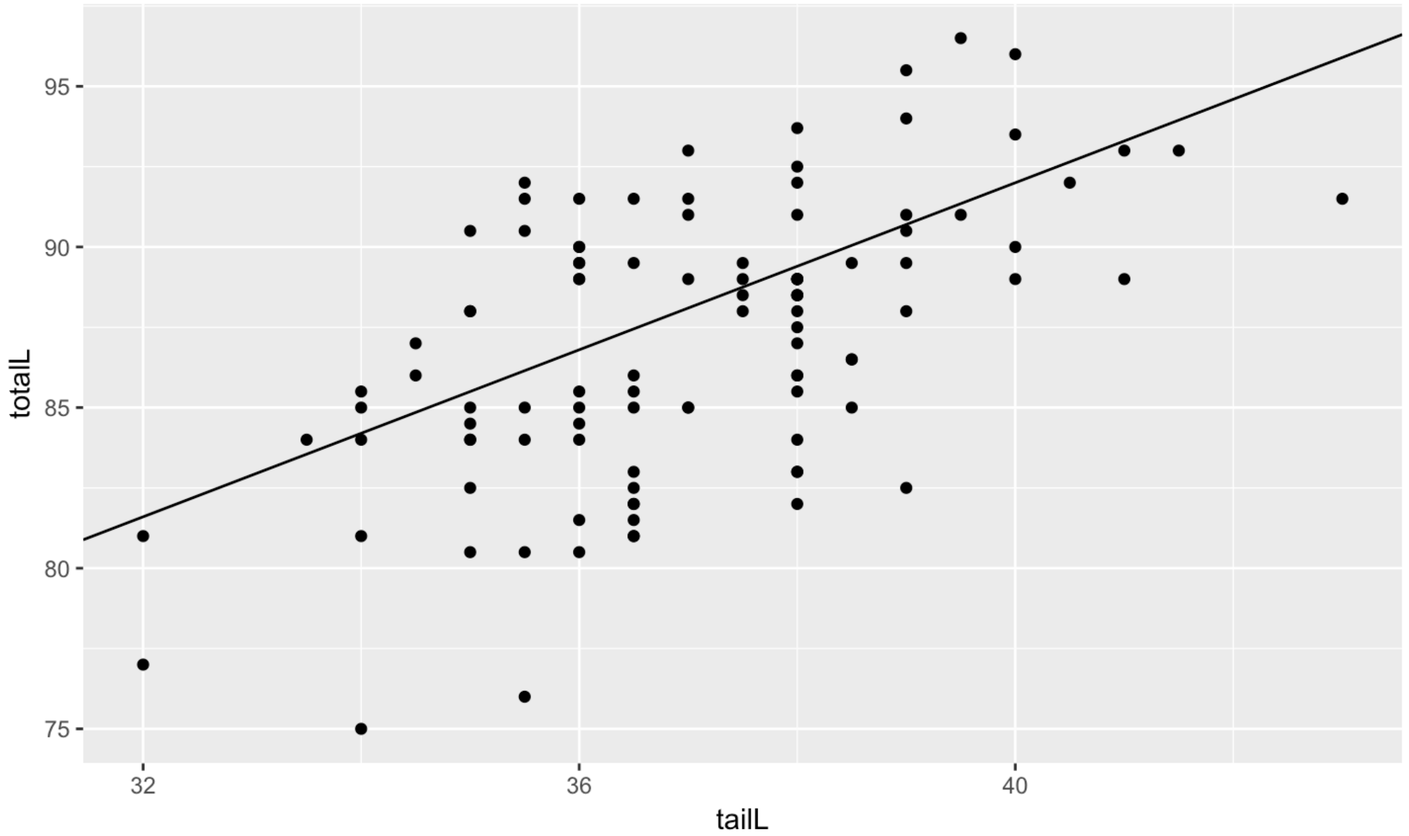
# Through the origin, better fit

```
> ggplot(data = possum, aes(y = totalL, x = tailL)) +
    geom_point() + geom_abline(intercept = 0, slope = 1.7)
```



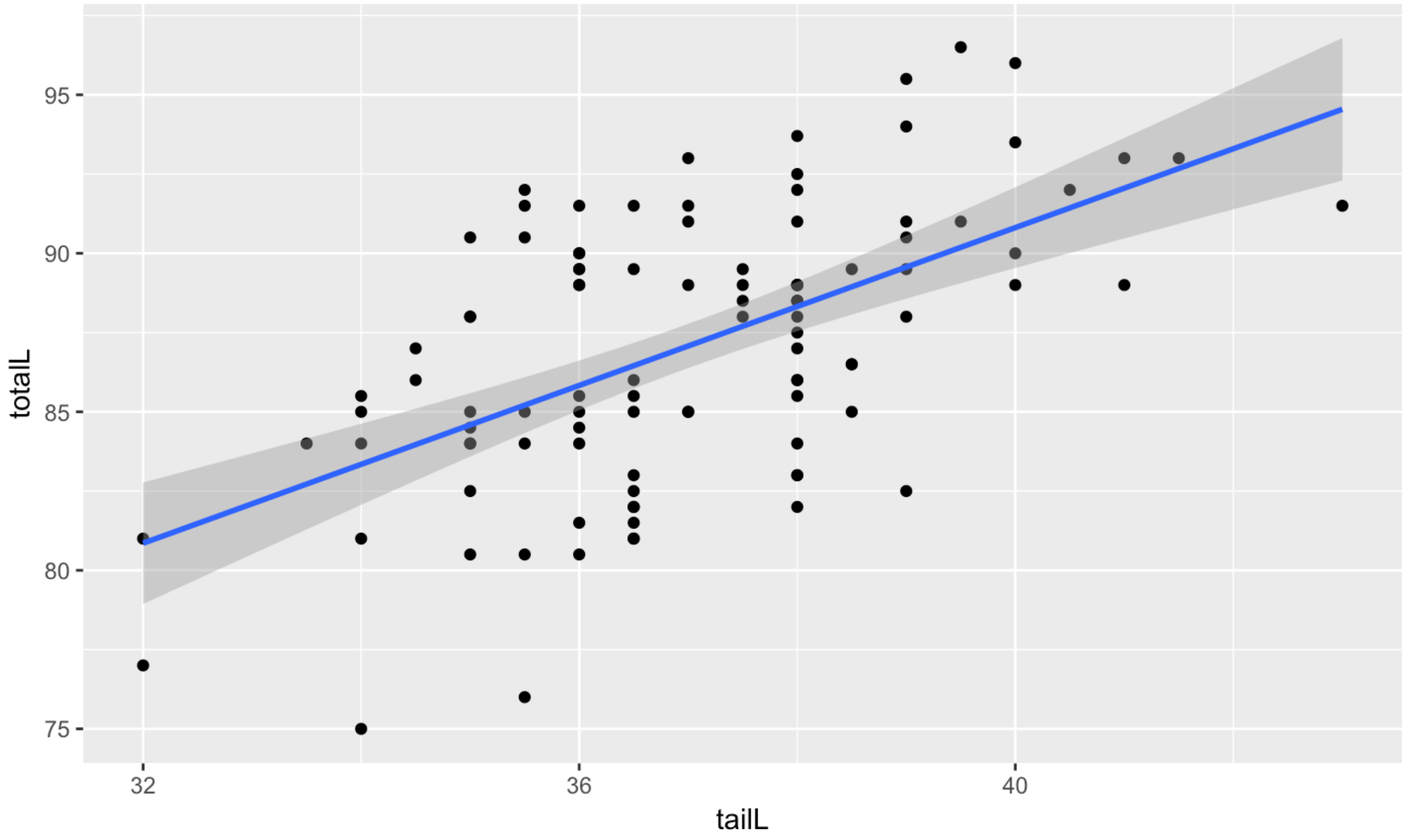
# Not through the origin

```
> ggplot(data = possum, aes(y = totalL, x = tailL)) +
    geom_point() + geom_abline(intercept = 40, slope = 1.3)
```



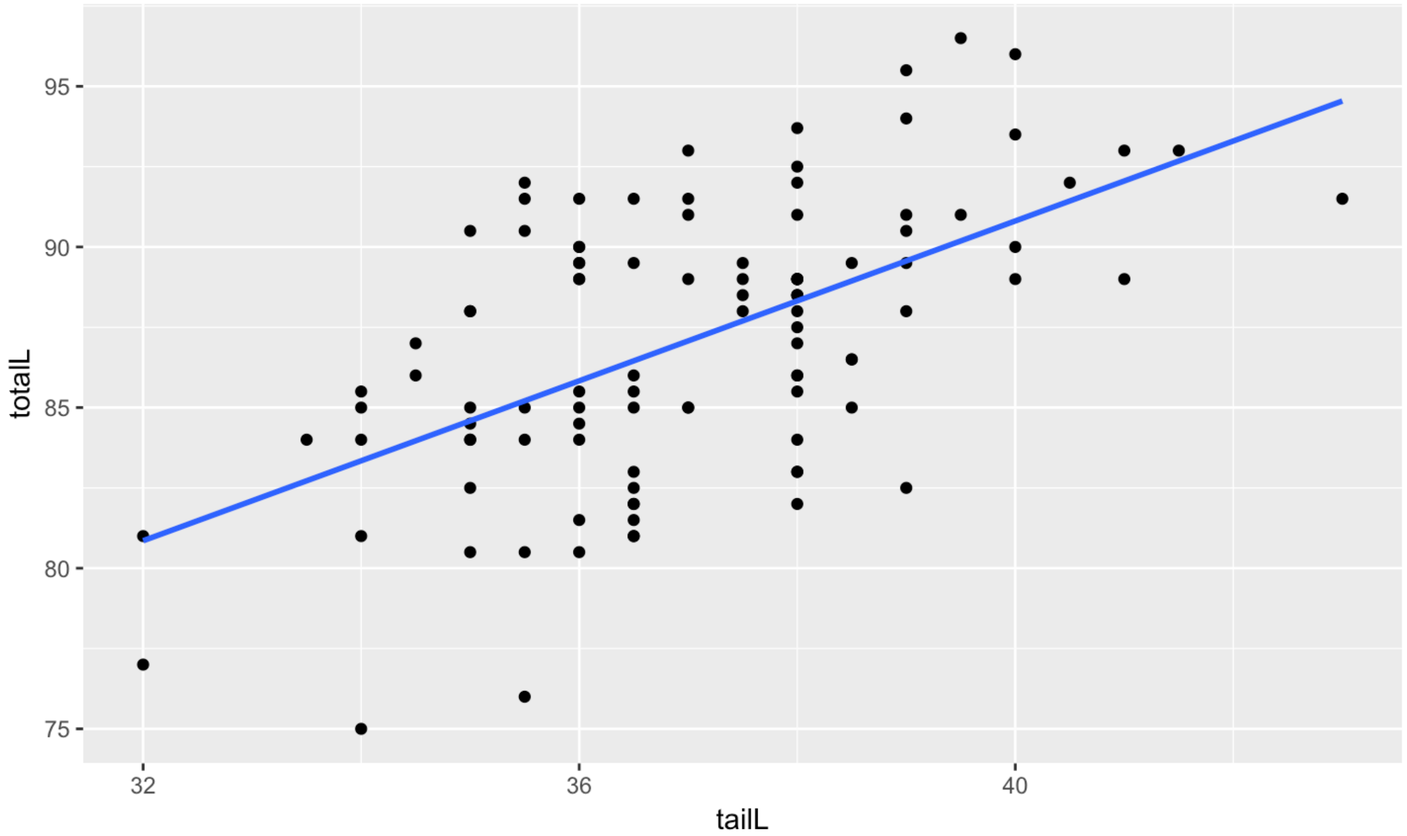
# The "best" fit line

```
> ggplot(data = possum, aes(y = totalL, x = tailL)) +
    geom_point() + geom_smooth(method = "lm")
```



# Ignore standard errors

```
> ggplot(data = possum, aes(y = totalL, x = tailL)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE)
```



CORRELATION AND REGRESSION

# Let’s practice!

CORRELATION AND REGRESSION

# Understanding the linear model

# Generic statistical model

$$response = f(explanatory) + noise$$

# Generic linear model

$$response = intercept + (slope * explanatory) + noise$$

# Regression model

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon\,, \qquad \epsilon \sim N(0, \sigma_\epsilon)$$

# Fitted values

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$$

# Residuals

$$e = Y - \hat{Y}$$

# Fitting procedure

- Given $n$ observations of pairs $(x_i, y_i)$...
- Find $\hat{\beta}_0, \hat{\beta}_1$ that minimize $\sum_{i=1}^{n} e_i^2$

# Least squares

- Easy, deterministic, unique solution
- Residuals sum to zero
- Line must pass through $(\bar{x}, \bar{y})$
- Other criteria exist—just not in this course

# Key concepts

- Y-hat is expected value given corresponding X
- Beta-hats are estimates of true, unknown betas
- Residuals (e's) are estimates of true, unknown epsilons
- "Error" may be misleading term—better: noise

CORRELATION AND REGRESSION

# Let’s practice!

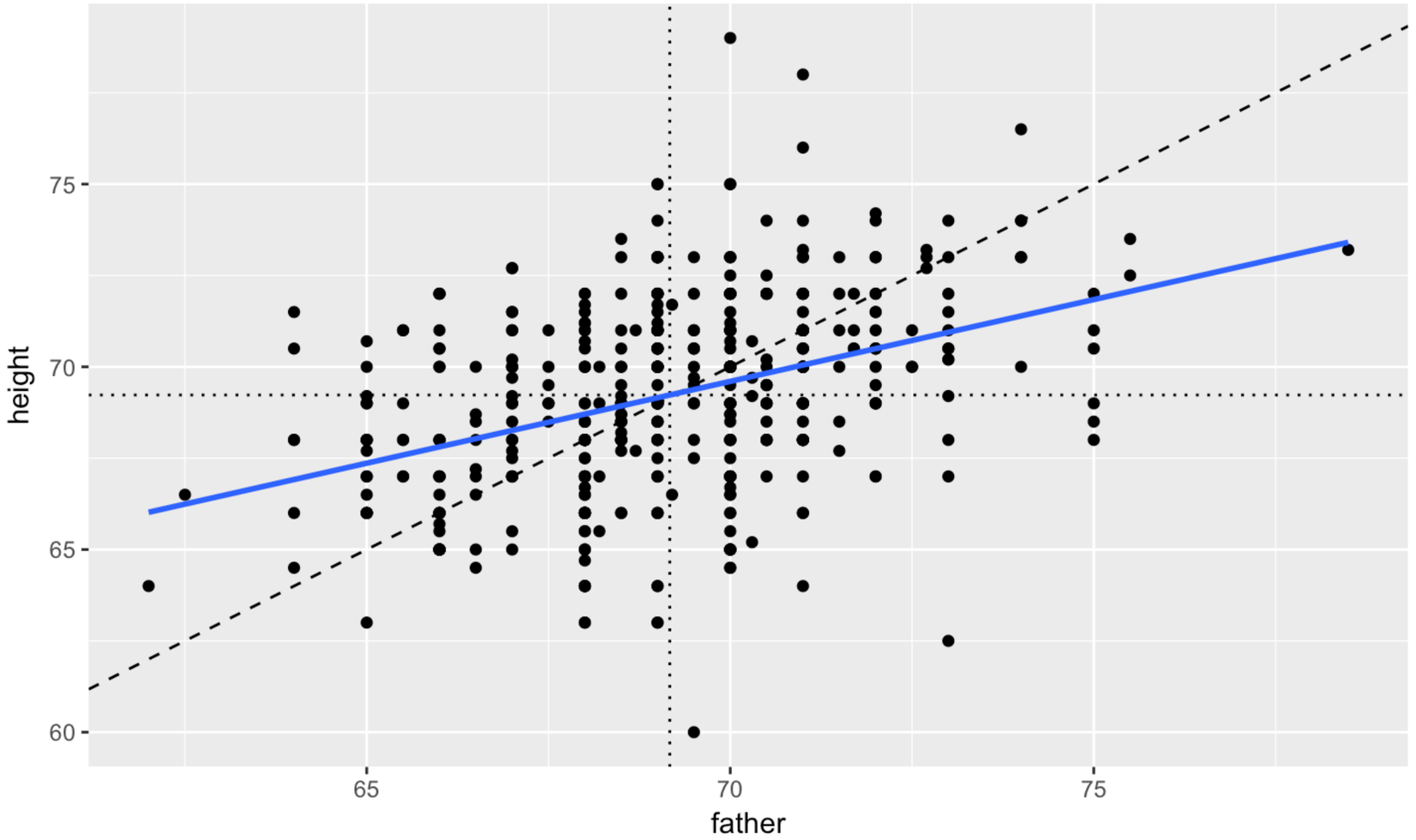


# Regression vs. regression to the mean

# Heredity

- Galton's "regression to the mean"
- Thought experiment: consider the heights of the children of NBA players

# Galton's data



# Regression modeling

- "Regression": techniques for modeling a quantitative response
- Types of regression models:
  - Least squares
  - Weighted
  - Generalized
  - Nonparametric
  - Ridge
  - Bayesian
  - ...

CORRELATION AND REGRESSION

# Let’s practice!

CORRELATION AND REGRESSION

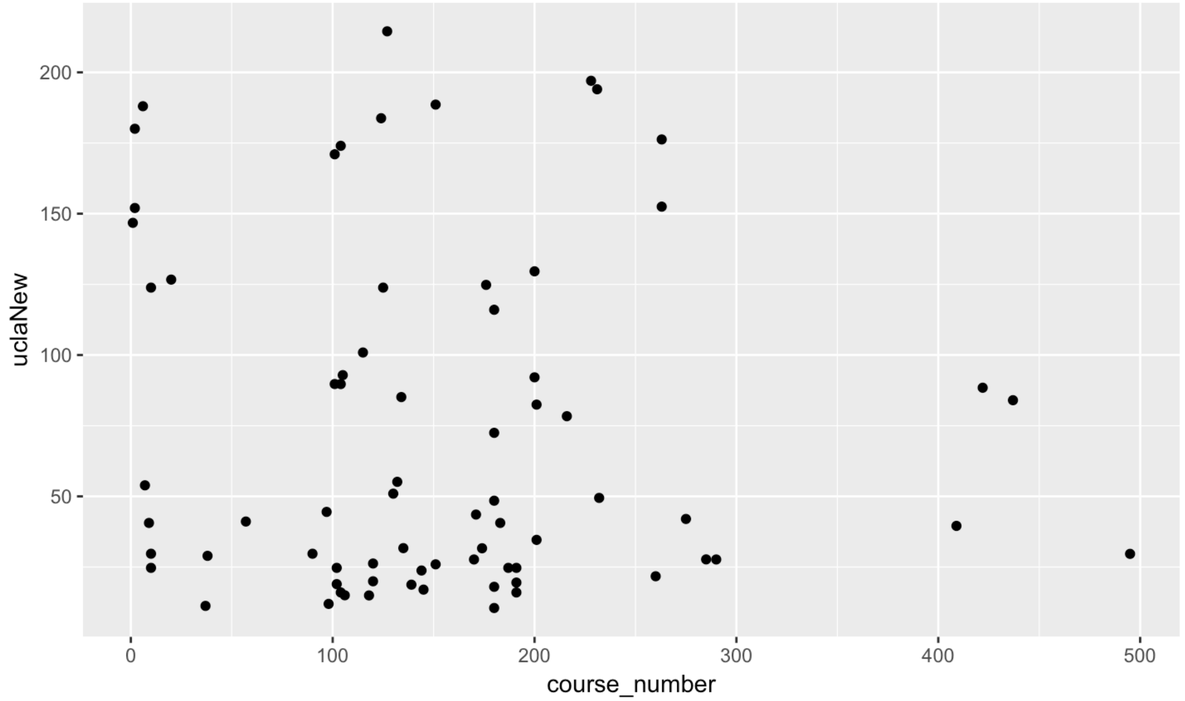
# Interpretation of regression coefficients

# Is that textbook overpriced?

```
> head(textbooks)
  deptAbbr course           isbn uclaNew amazNew more  diff
1   Am Ind   C170 978-0803272620   27.67   27.95    Y -0.28
2   Anthro      9 978-0030119194   40.59   31.14    Y  9.45
3   Anthro   135T 978-0300080643   31.68   32.00    Y -0.32
4   Anthro  191HB 978-0226206813   16.00   11.52    Y  4.48
5  Art His  M102K 978-0892365999   18.95   14.21    Y  4.74
6  Art His   118E 978-0394723693   14.95   10.17    Y  4.78
```

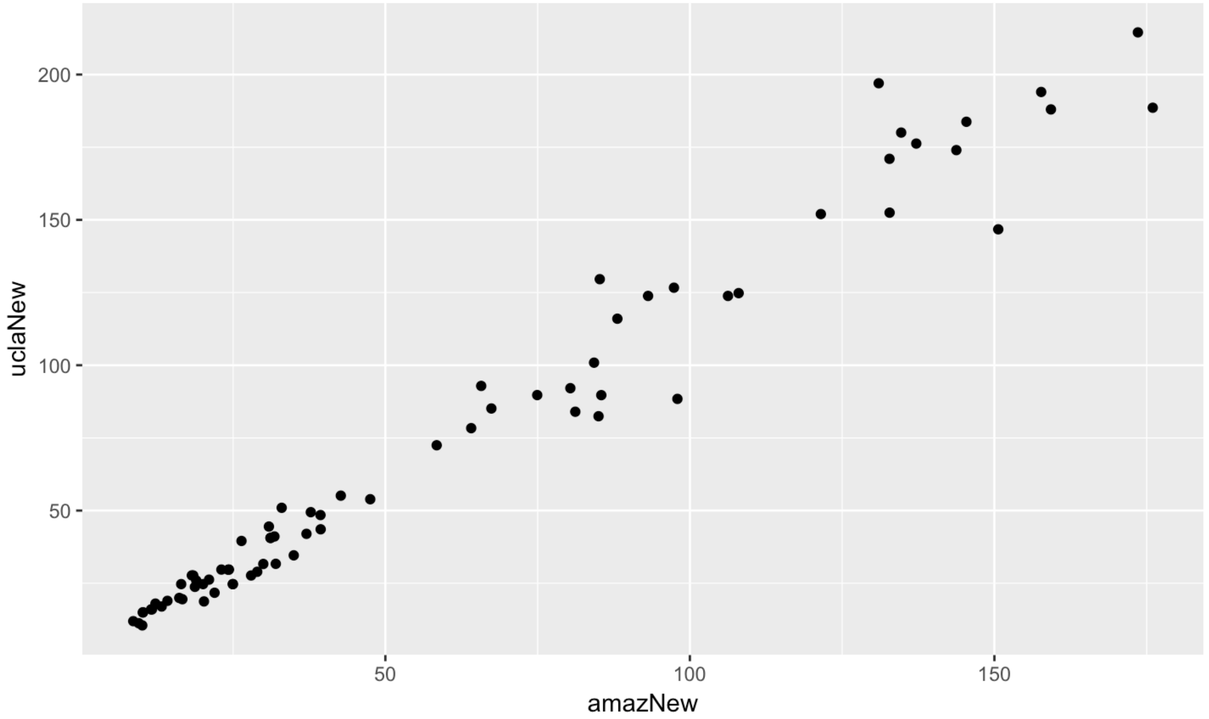
# Compared to the course number?

```
> textbooks %>%
    mutate(course_number = readr::parse_number(course)) %>%
    ggplot(aes(x = course_number, y = uclaNew)) +
    geom_point()
```



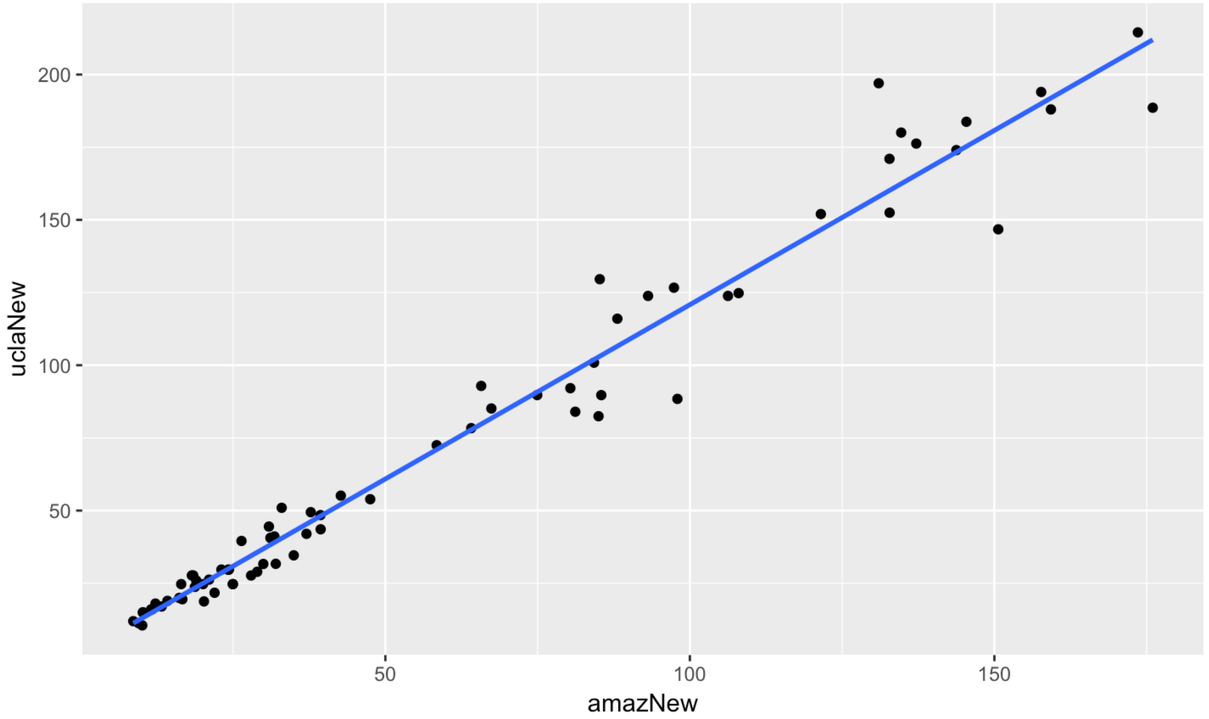
# Compared to Amazon?

```
> ggplot(data = textbooks, aes(x = amazNew, y = uclaNew)) +
    geom_point()
```



# Compared to Amazon?

```
> ggplot(data = textbooks, aes(x = amazNew, y = uclaNew)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE)
```



# Slope and intercept

```
> lm(uclaNew ~ amazNew, data = textbooks)

Call:
lm(formula = uclaNew ~ amazNew, data = textbooks)

Coefficients:
(Intercept)       amazNew
      0.929         1.199
```

$$\widehat{uclaNew} = 0.929 + 1.199 \cdot amazNew$$

# Units and scale

```
> textbooks %>%
    mutate(amazNew_cents = amazNew * 100) %>%
    lm(uclaNew ~ amazNew_cents, data = .)

Call:
lm(formula = uclaNew ~ amazNew_cents, data = .)

Coefficients:
(Intercept)  amazNew_cents
      0.929          0.01199
```

CORRELATION AND REGRESSION

# Let’s practice!

CORRELATION AND REGRESSION

# Your linear model object

# Is that textbook overpriced?

```
> mod <- lm(uclaNew ~ amazNew, data = textbooks)
> class(mod)
[1] "lm"
```

# Print

```
> mod

Call:
lm(formula = uclaNew ~ amazNew, data = textbooks)

Coefficients:
(Intercept)       amazNew
      0.929         1.199
```

# Fitted coefficients

```
> coef(mod)
(Intercept)      amazNew
      0.929        1.199
```

# Summary

```
> summary(mod)

Call:
lm(formula = uclaNew ~ amazNew, data = textbooks)

Residuals:
   Min     1Q Median     3Q    Max
-34.78  -4.57   0.58   4.01  39.00

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9290     1.9354    0.48     0.63
amazNew       1.1990     0.0252   47.60   <2e-16

Residual standard error: 10.5 on 71 degrees of freedom
Multiple R-squared:  0.97,   Adjusted R-squared:  0.969
F-statistic: 2.27e+03 on 1 and 71 DF,  p-value: <2e-16
```

# Fitted values

```
> fitted.values(mod)
    1      2      3      4      5      6      7      8      9     10
34.44  38.27  39.30  14.74  17.97  13.12  24.98  20.90 128.32  16.83
   11     12     13     14     15     16     17     18     19     20
36.84 106.55  23.05  20.68 117.69  57.89  90.77 160.12 146.61 130.42
   21     22     23     24     25     26     27     28     29     30
14.92  23.64  15.60  27.25  38.27  35.64  20.29  46.19  39.03  40.46
   31     32     33     34     35     36     37     38     39     40
37.94 102.84  42.83 118.37  98.26  12.32  13.16 162.42 173.29 211.95
   41     42     43     44     45     46     47     48     49     50
181.53 175.26 209.03 158.00 189.99 165.40  30.84 191.91  28.59  26.16
   51     52     53     54     55     56     57     58     59     60
52.10  48.13 103.08 112.59  81.74 160.14  30.08  30.84 103.38  13.01
   61     62     63     64     65     66     67     68     69     70
79.74 101.96  11.24  70.97  97.29  77.77  45.34  25.16  48.10  32.55
   71     72     73
29.93  23.37  22.77
```

# Residuals

```
> residuals(mod)
       1        2        3        4        5        6        7
-6.77105  2.32413 -7.61701  1.25854  0.98322  1.82719 -0.28093
       8        9       10       11       12       13       14
-1.40433 -4.48287  0.17228 -5.20906  9.45100  4.61946  4.02348
      15       16       17       18       19       20       21
 8.98228 -3.99352 -1.04014 10.87962  5.39236 -5.62112  1.07869
      22       23       24       25       26       27       28
 2.31195  2.39526 -5.51705  2.32413 -6.69006 -0.34284  3.25873
      29       30       31       32       33       34       35
 2.05677 10.48996  6.55786 -20.39409 -8.23406 -29.95115 -14.26390
      36       37       38       39       40       41       42
-1.06948  1.84122 17.60753  0.71458 -23.37321 -34.78455  8.48623
      43       44       45       46       47       48       49
 5.47235 39.00185  4.01249 10.85401 -6.14405 -3.90591  1.11007
      50       51       52       53       54       55       56
 0.08405  3.02765 -4.57365 26.51611 11.24803  3.37834 -7.66436
...
```

# broom

```
> library(broom)
> augment(mod)
   uclaNew amazNew .fitted .se.fit    .resid    .hat .sigma   .cooksd
1    27.67   27.95   34.44   1.460  -6.77105 0.01944 10.515 4.227e-03
2    40.59   31.14   38.27   1.418   2.32413 0.01834 10.543 4.687e-04
3    31.68   32.00   39.30   1.407  -7.61701 0.01806 10.507 4.955e-03
4    16.00   11.52   14.74   1.721   1.25854 0.02700 10.546 2.059e-04
5    18.95   14.21   17.97   1.674   0.98322 0.02555 10.546 1.186e-04
6    14.95   10.17   13.12   1.745   1.82719 0.02776 10.545 4.469e-04
7    24.70   20.06   24.98   1.577  -0.28093 0.02268 10.547 8.544e-06
8    19.50   16.66   20.90   1.632  -1.40433 0.02430 10.546 2.295e-04
9   123.84  106.25  128.32   1.700  -4.48287 0.02637 10.533 2.548e-03
10   17.00   13.26   16.83   1.690   0.17228 0.02605 10.547 3.716e-06
11   31.63   29.95   36.84   1.433  -5.20906 0.01874 10.528 2.407e-03
12  116.00   88.09  106.55   1.422   9.45100 0.01844 10.485 7.794e-03
13   27.67   18.45   23.05   1.603   4.61946 0.02343 10.532 2.390e-03
14   24.70   16.47   20.68   1.636   4.02348 0.02439 10.536 1.891e-03
15  126.67   97.38  117.69   1.554   8.98228 0.02202 10.491 8.468e-03
```

CORRELATION AND REGRESSION

# Let’s practice!

CORRELATION AND REGRESSION

# Using the linear model

# Is that textbook overpriced?

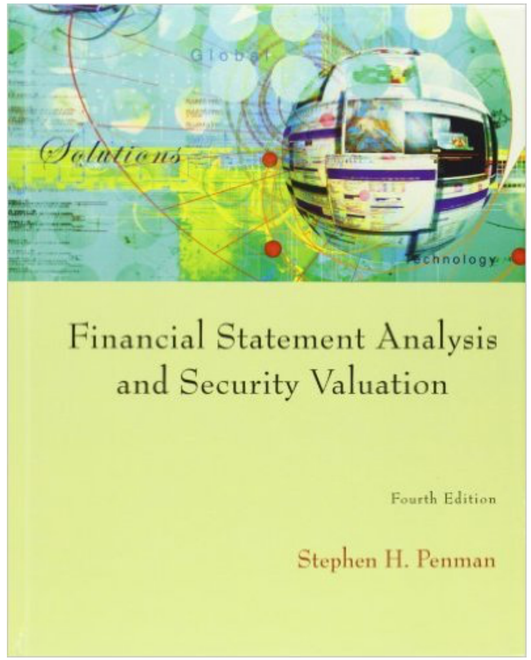
```
> mod <- lm(uclaNew ~ amazNew, data = textbooks)
```

# Examining residuals

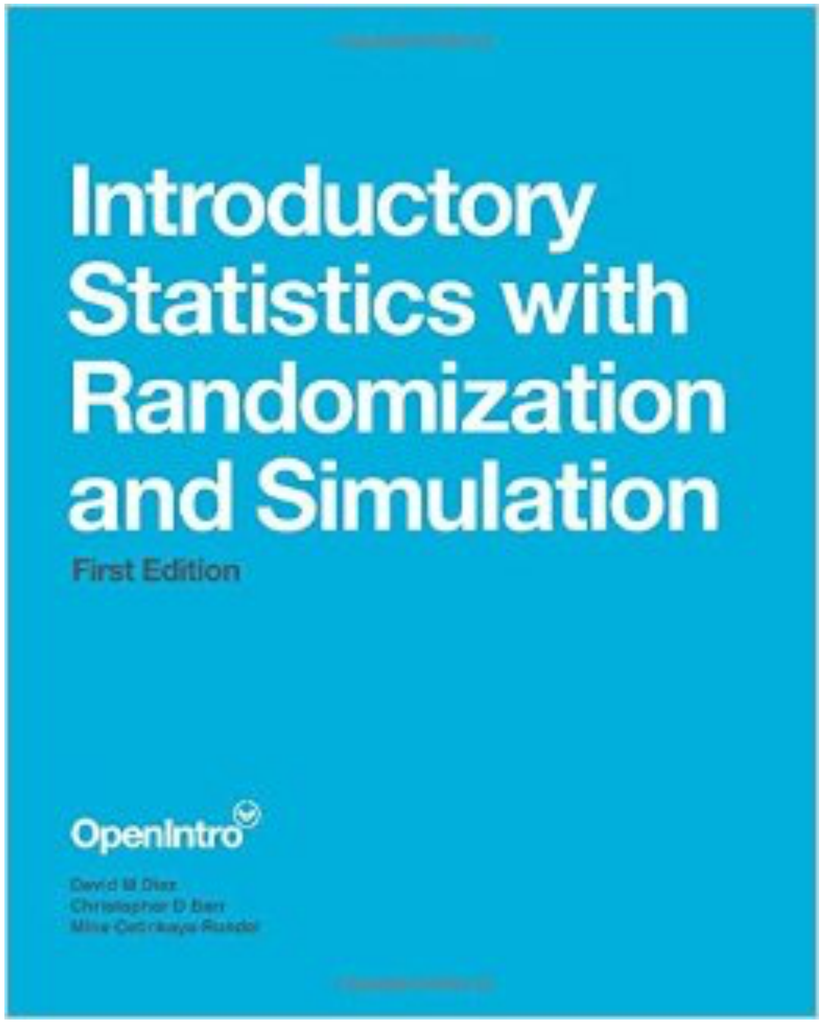
```
> library(broom)
> augment(mod) %>%
    arrange(desc(.resid)) %>%
    head()
 uclaNew amazNew .fitted .se.fit .resid    .hat .sigma .cooksd .std.resid
1  197.00  131.00  158.00   2.179  39.00 0.04331  9.409 0.32816      3.808
2  129.60   85.20  103.08   1.387  26.52 0.01753 10.051 0.05822      2.554
3  180.03  134.69  162.42   2.257  17.61 0.04644 10.324 0.07219      1.722
4   92.88   65.73   79.74   1.236  13.14 0.01393 10.428 0.01128      1.264
5  123.84   93.13  112.59   1.491  11.25 0.02026 10.459 0.01217      1.085
6  171.00  132.77  160.12   2.216  10.88 0.04479 10.463 0.02649      1.063
```

# Markup

```
> textbooks %>%
    filter(uclaNew == 197)
  deptAbbr course           ibsn uclaNew amazNew more diff
1     Mgmt    228 978-0073379661     197      131    Y   66
```



# Making predictions



# Making predictions

```
predict(lm)
```

fitted values for existing data

# Making predictions

```
predict(lm, newdata)
```

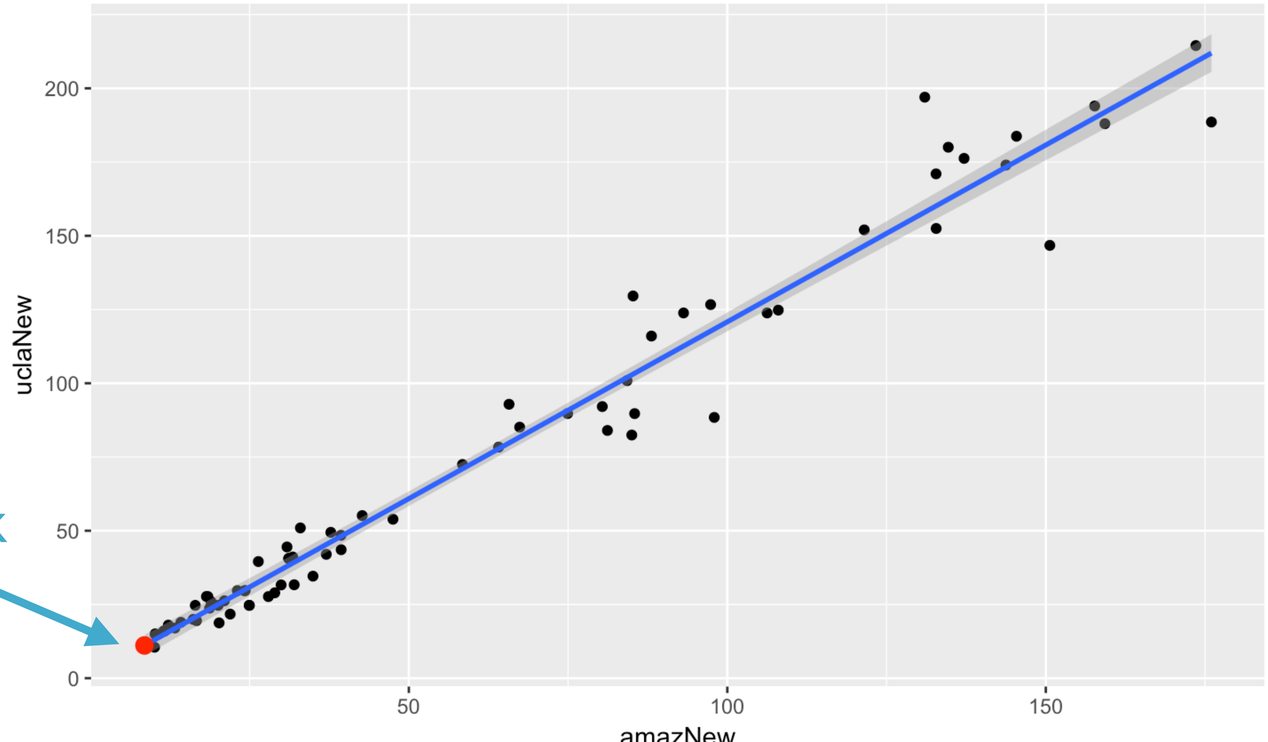
fitted values for any new data

# New data

```
> new_data <- data.frame(amazNew = 8.49)
> predict(mod, newdata = new_data)
    1
11.11
```

# Visualize new observations

```
> isrs <- broom::augment(mod, newdata = new_data)
> ggplot(data = textbooks, aes(x = amazNew, y = uclaNew)) +
    geom_point() + geom_smooth(method = "lm") +
    geom_point(data = isrs, aes(y = .fitted), size = 3, color = "red")
```



CORRELATION AND REGRESSION

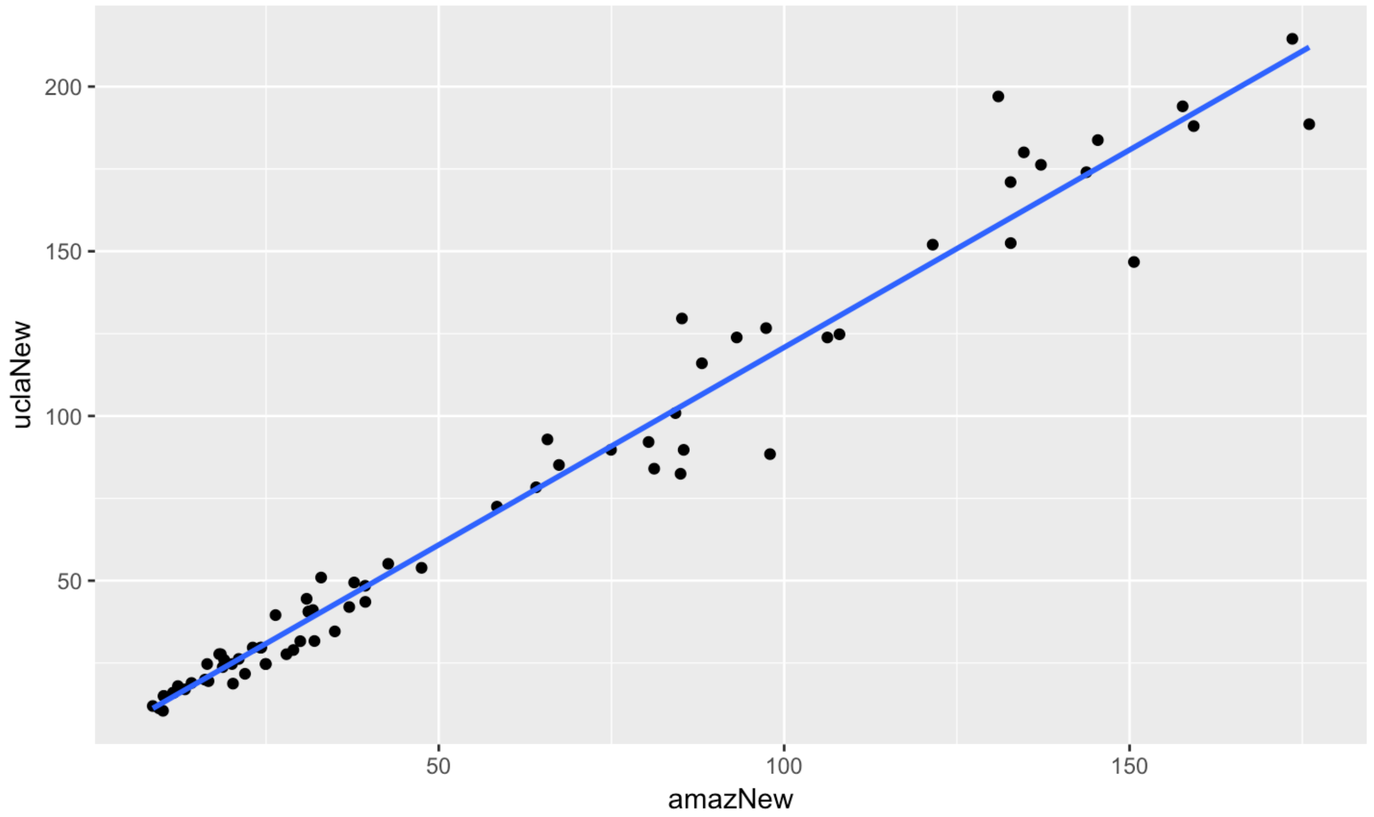
# Let’s practice!

CORRELATION AND REGRESSION

# Assessing model fit

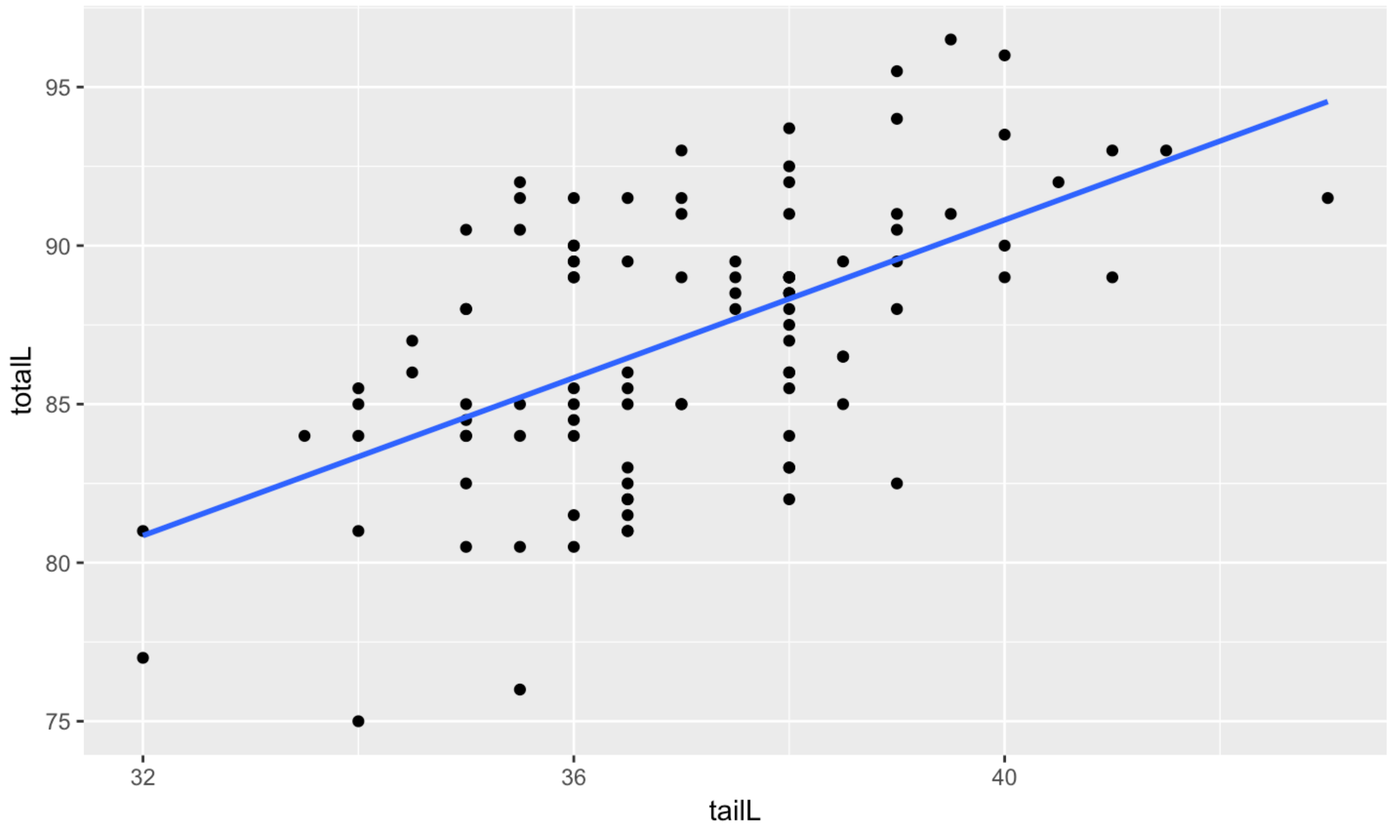
# How well does our textbook model fit?

```
> ggplot(data = textbooks, aes(x = amazNew, y = uclaNew)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE)
```

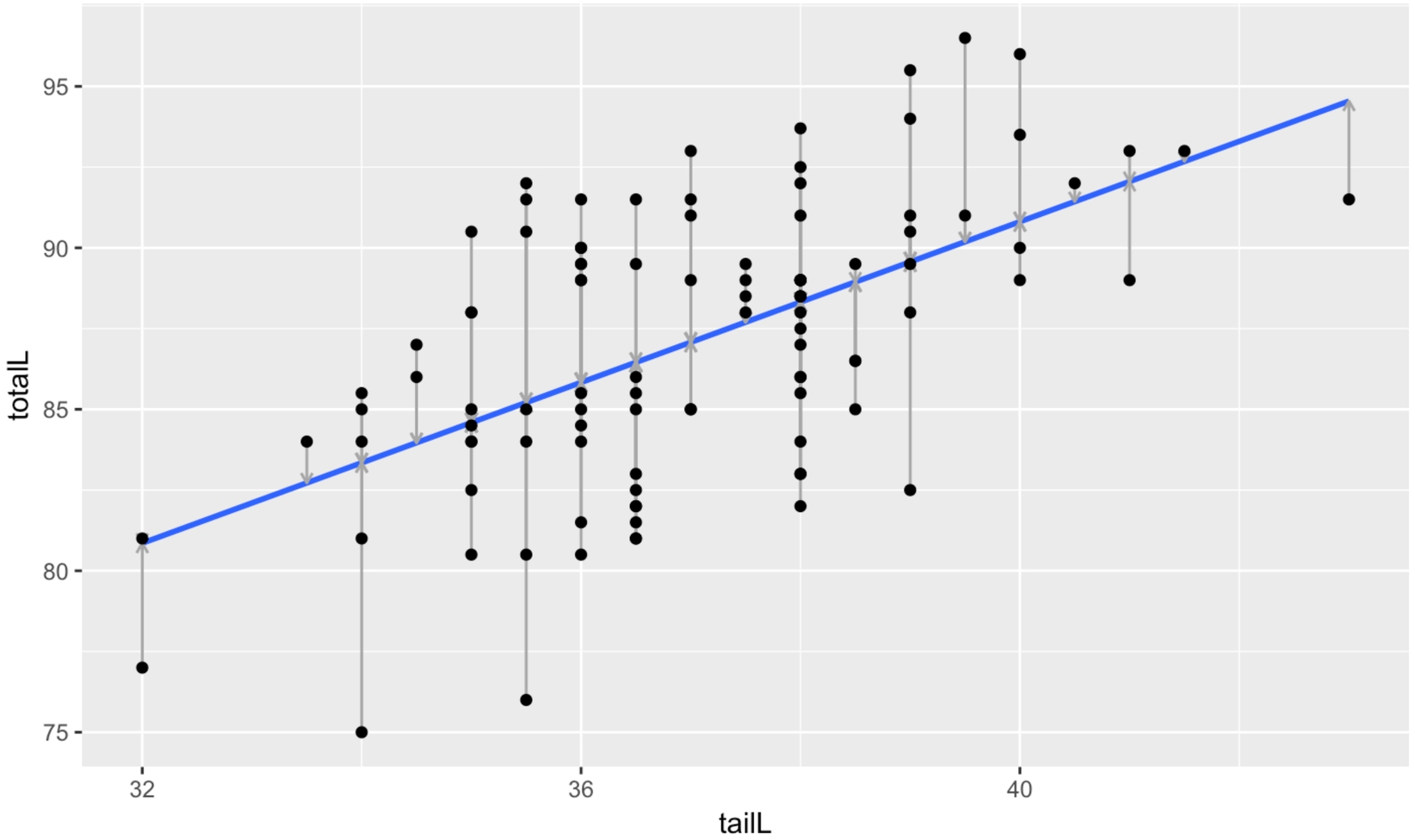


# How well does our possum model fit?

```
> ggplot(data = possum, aes(y = totalL, x = tailL)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE)
```



# Sums of squared deviations



# SSE

```
> library(broom)
> mod_possum <- lm(totalL ~ tailL, data = possum)
> mod_possum %>%
    augment() %>%
    summarize(SSE = sum(.resid^2),
              SSE_also = (n() - 1) * var(.resid))
   SSE SSE_also
1 1301     1301
```

# RMSE

$$RMSE = \sqrt{\frac{\sum_i e_i^2}{d.f}} = \sqrt{\frac{SSE}{n-2}}$$

# Residual standard error (possums)

```
> summary(mod_possum)

Call:
lm(formula = totalL ~ tailL, data = possum)

Residuals:
   Min     1Q Median     3Q    Max
-9.210 -2.326  0.179  2.777  6.790

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    41.04       6.66    6.16  1.4e-08
tailL           1.24       0.18    6.93  3.9e-10

Residual standard error: 3.57 on 102 degrees of freedom
Multiple R-squared:  0.32,   Adjusted R-squared:  0.313
F-statistic:   48 on 1 and 102 DF,  p-value: 3.94e-10
```

# Residual standard error (textbooks)

```
> lm(uclaNew ~ amazNew, data = textbooks) %>%
    summary()

Call:
lm(formula = uclaNew ~ amazNew, data = textbooks)

Residuals:
   Min     1Q Median     3Q    Max
-34.78  -4.57   0.58   4.01  39.00

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9290     1.9354    0.48     0.63
amazNew       1.1990     0.0252   47.60   <2e-16

Residual standard error: 10.5 on 71 degrees of freedom
Multiple R-squared:  0.97,   Adjusted R-squared:  0.969
F-statistic: 2.27e+03 on 1 and 71 DF,  p-value: <2e-16
```

CORRELATION AND REGRESSION

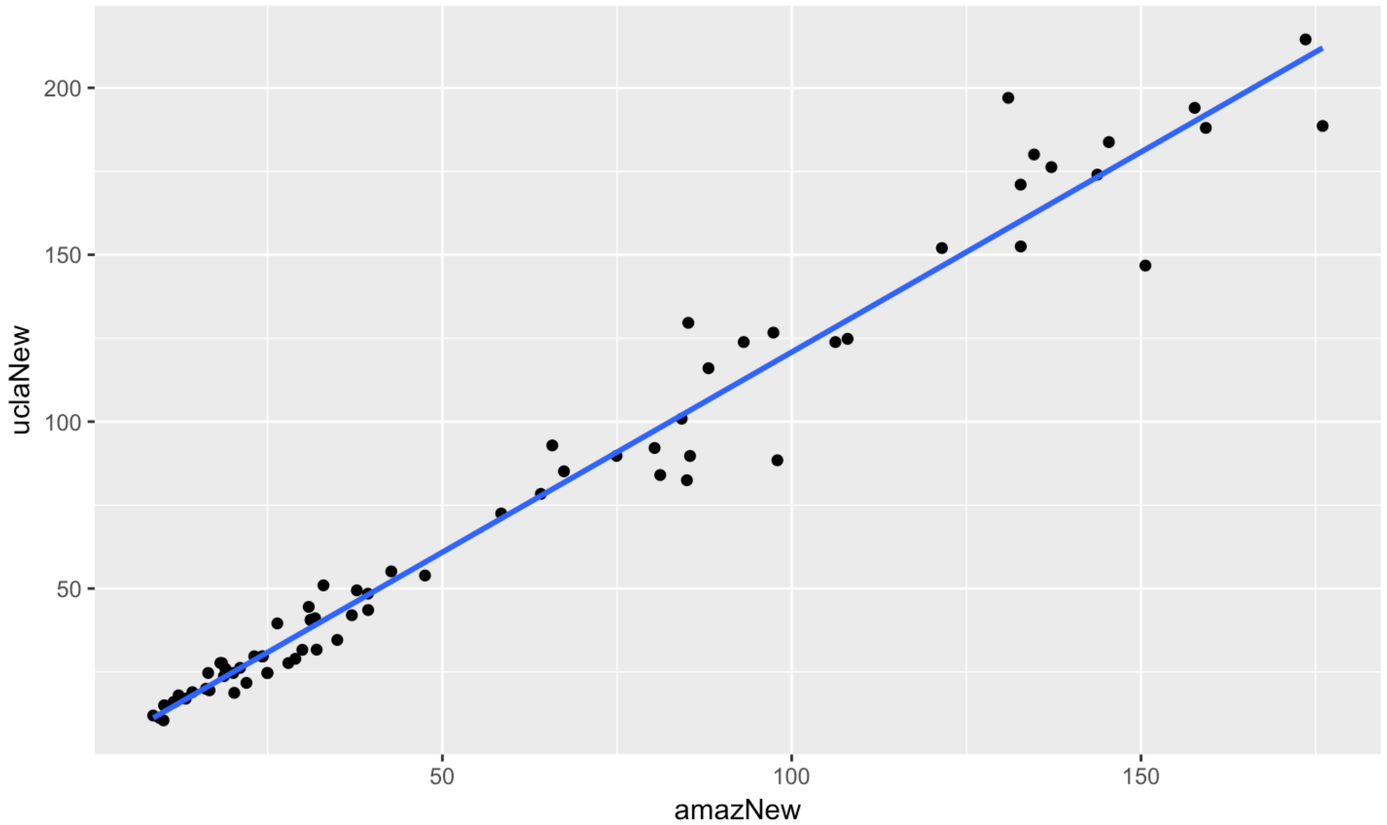
# Let’s practice!

CORRELATION AND REGRESSION

# Comparing model fits

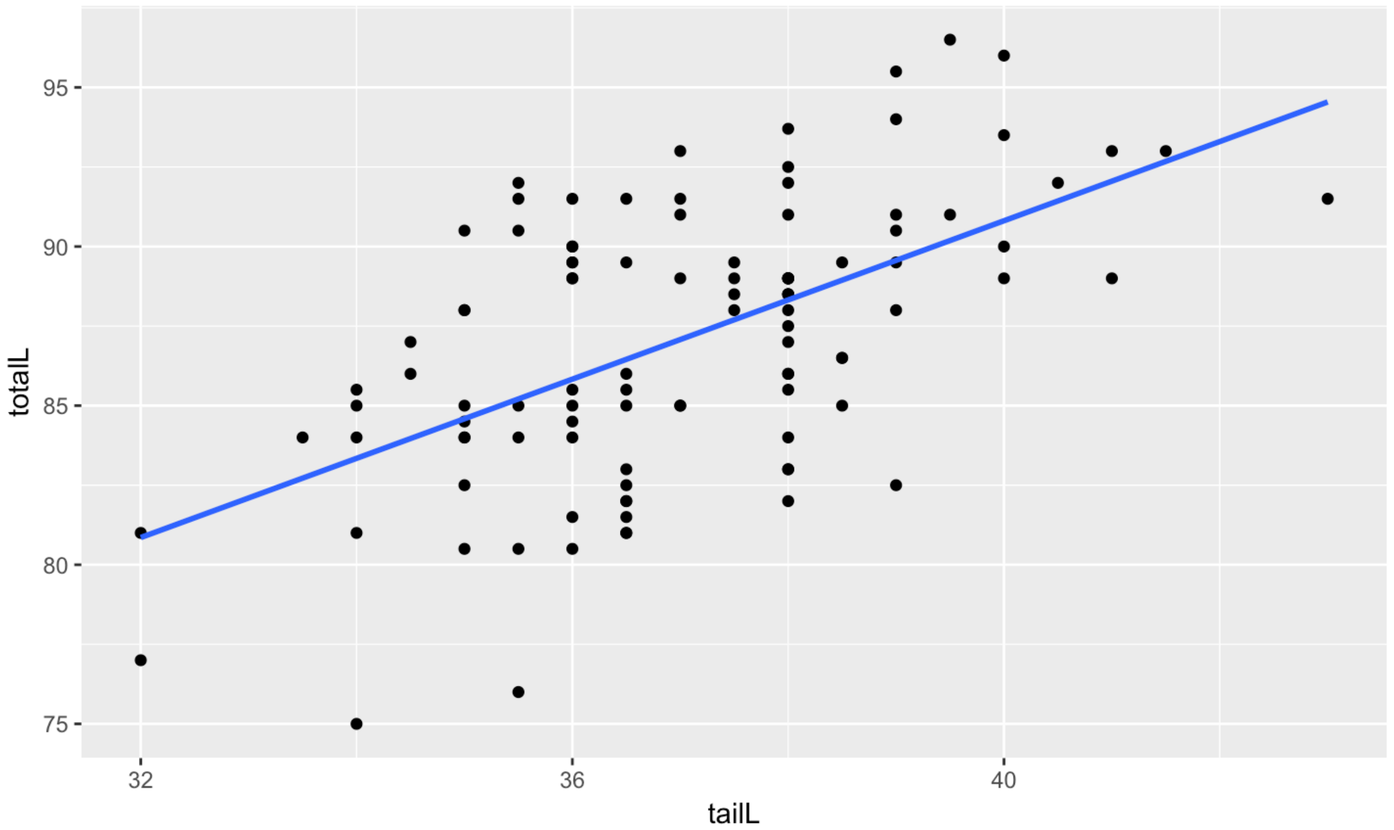
# How well does our textbook model fit?

```
> ggplot(data = textbooks, aes(x = amazNew, y = uclaNew)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE)
```



# How well does our possum model fit?

```
> ggplot(data = possum, aes(y = totalL, x = tailL)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE)
```

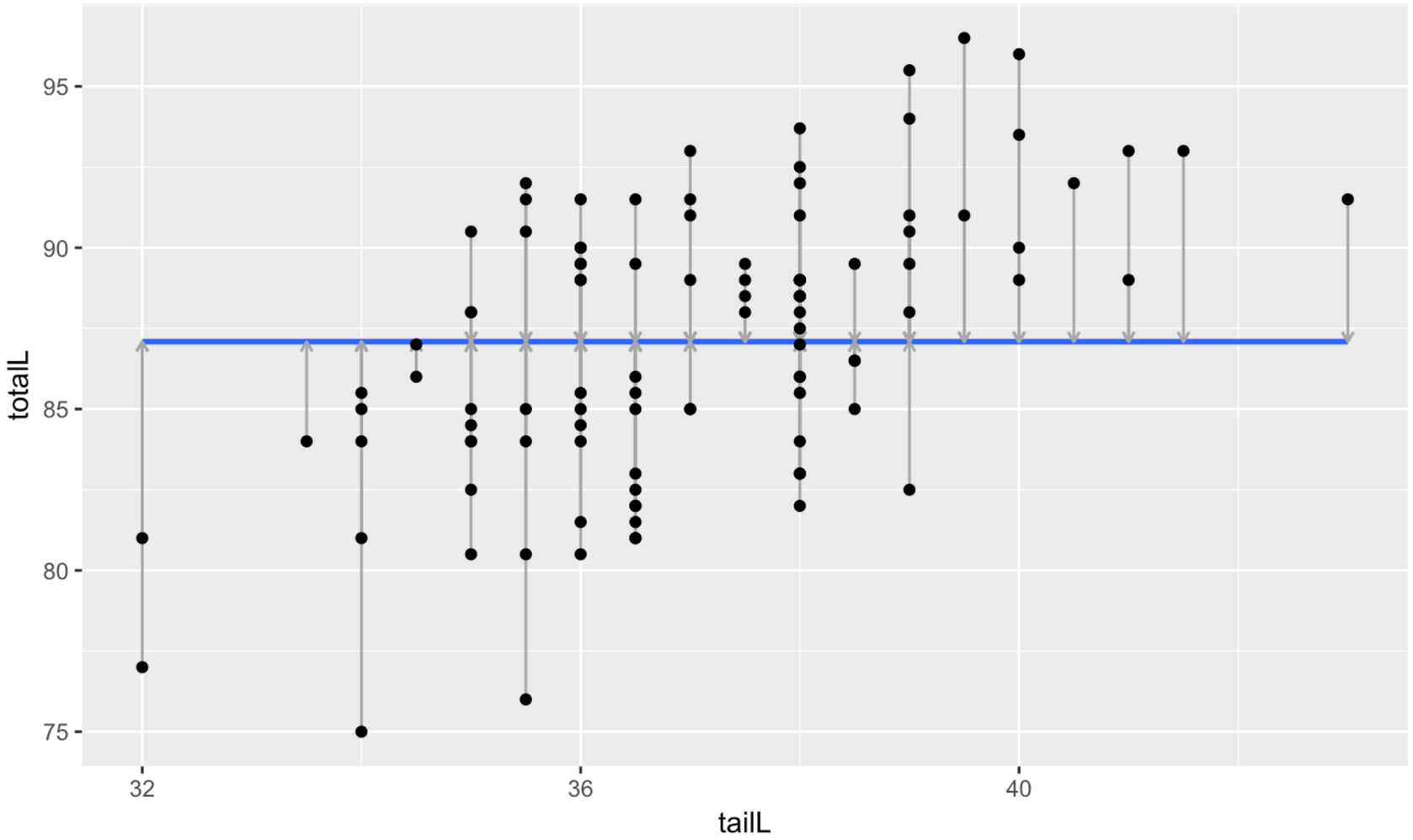


# Null (average) model

- For all observations...

$$\hat{y} = \bar{y}$$

# Visualization of null model



# SSE, null model

```
> mod_null <- lm(totalL ~ 1, data = possum)
> mod_null %>%
    augment(possum) %>%
    summarize(SST = sum(.resid^2))
   SST
1 1914
```

# SSE, our model

```
> mod_possum <- lm(totalL ~ tailL, data = possum)
> mod_possum %>%
    augment() %>%
    summarize(SSE = sum(.resid^2))
   SSE
1 1301
```

# Coefficient of determination

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{Var(e)}{Var(y)}$$

# Connection to correlation

- For simple linear regression...

$$r^2_{x,y} = R^2$$

# Summary

```
> summary(mod_possum)

Call:
lm(formula = totalL ~ tailL, data = possum)

Residuals:
   Min     1Q Median     3Q    Max
-9.210 -2.326  0.179  2.777  6.790

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    41.04       6.66    6.16  1.4e-08
tailL           1.24       0.18    6.93  3.9e-10

Residual standard error: 3.57 on 102 degrees of freedom
Multiple R-squared:  0.32,   Adjusted R-squared:  0.313
F-statistic:   48 on 1 and 102 DF,  p-value: 3.94e-10
```

# Over-reliance on R-squared

*"Essentially, all models are wrong, but some are useful."*

**- George Box**

CORRELATION AND REGRESSION

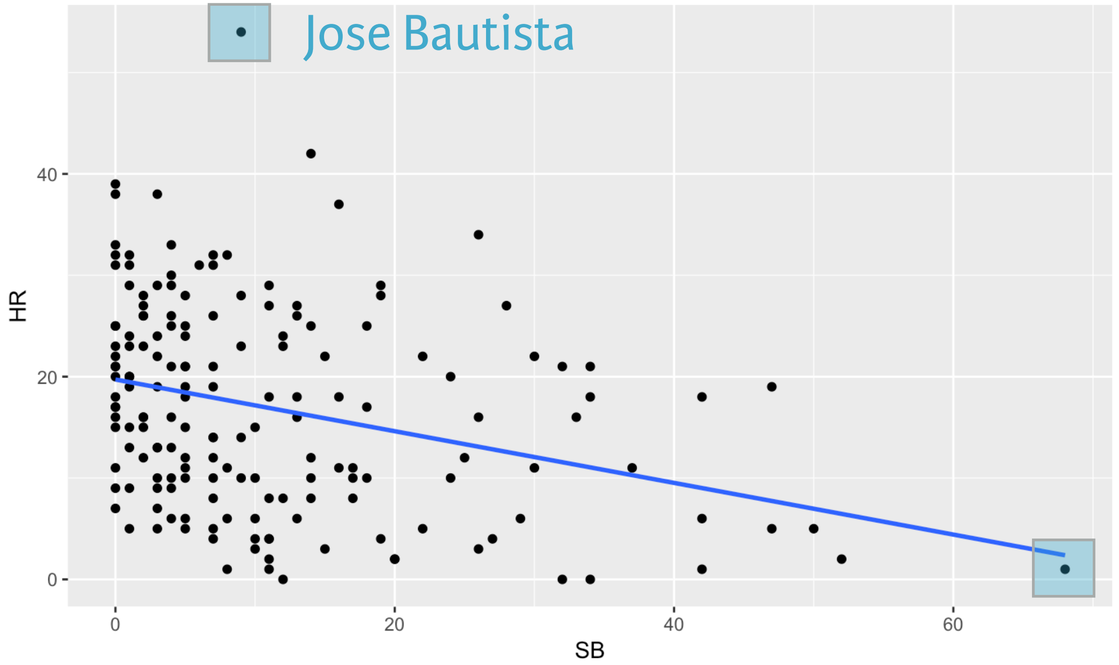
# Let’s practice!

CORRELATION AND REGRESSION

# Unusual points

# Unusual points

```
> regulars <- mlbBat10 %>%
    filter(AB > 400)
> ggplot(data = regulars, aes(x = SB, y = HR)) +
    geom_point() +
    geom_smooth(method = "lm", se = 0)
```



# Leverage

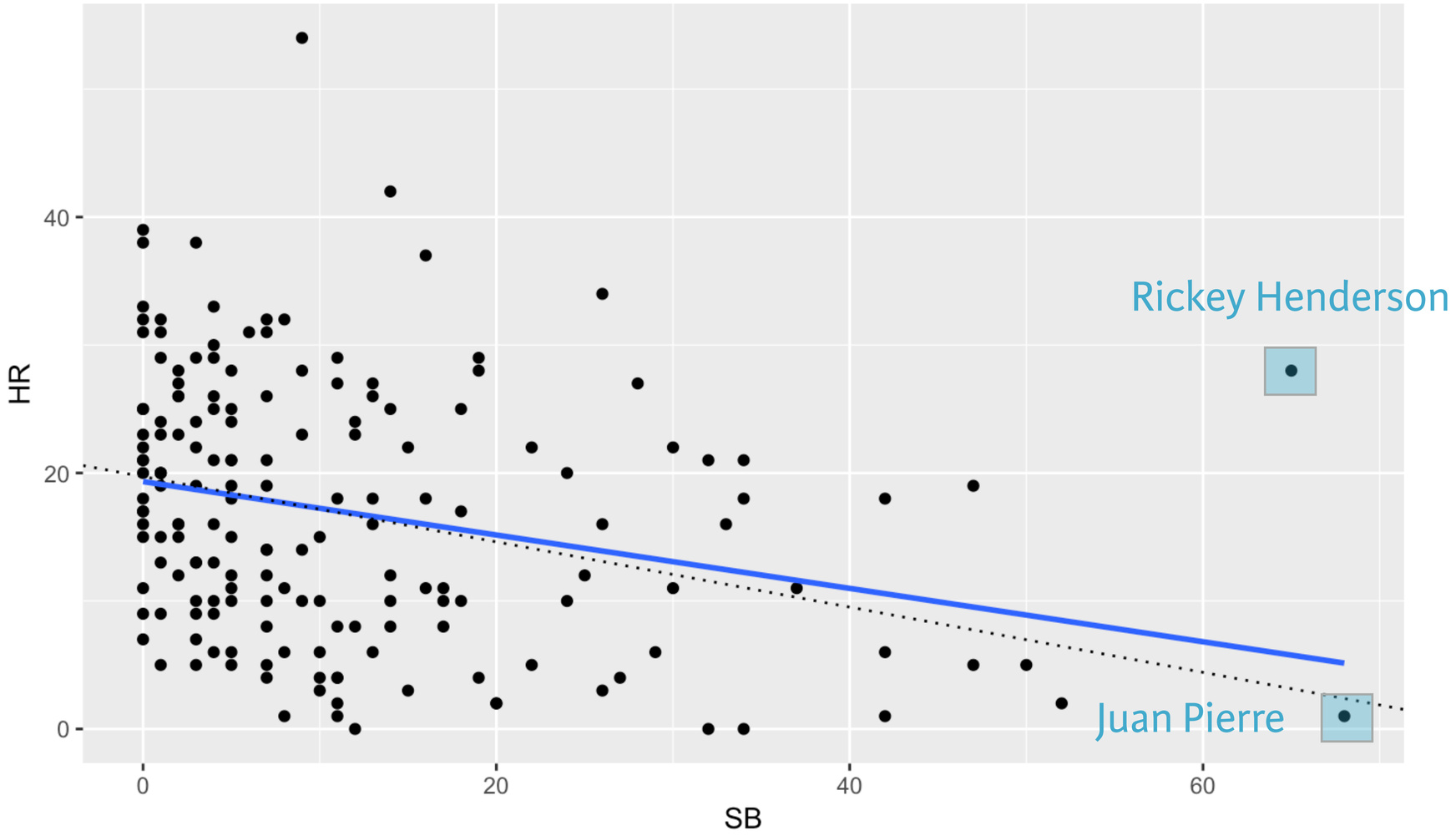
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

# Leverage computations

```
> library(broom)
> mod <- lm(HR ~ SB, data = regulars)
> mod %>%
    augment() %>%
    arrange(desc(.hat)) %>%
    select(HR, SB, .fitted, .resid, .hat) %>%
    head()
  HR SB .fitted .resid    .hat
1  1 68   2.383 -1.383 0.13082
2  2 52   6.461 -4.461 0.07034
3  5 50   6.971 -1.971 0.06417
4 19 47   7.736 11.264 0.05550
5  5 47   7.736 -2.736 0.05550
6  1 42   9.010 -8.010 0.04261
```

Juan Pierre

# Consider Rickey Henderson...



# Influence via Cook's distance

```
> mod <- lm(HR ~ SB, data = regulars_plus)
> mod %>%
    augment() %>%
    arrange(desc(.cooksd)) %>%
    select(HR, SB, .fitted, .resid, .hat, .cooksd) %>%
    head()
  HR SB .fitted .resid     .hat .cooksd
1 28 65   5.770 22.230 0.105519 0.33430
2 54  9  17.451 36.549 0.006070 0.04210
3 34 26  13.905 20.095 0.013150 0.02797
4 19 47   9.525  9.475 0.049711 0.02535
5 39  0  19.328 19.672 0.010479 0.02124
6 42 14  16.408 25.592 0.006061 0.02061
```

Henderson

CORRELATION AND REGRESSION

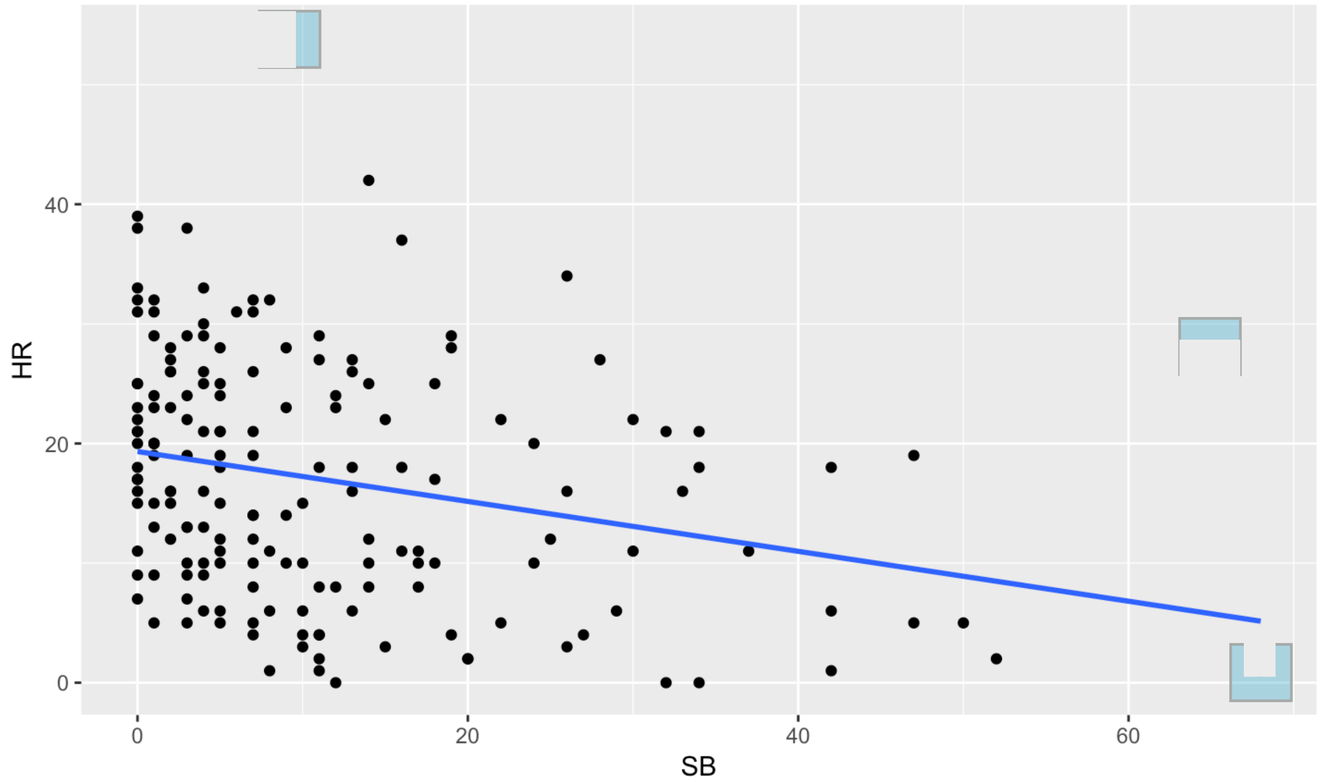
# Let’s practice!

CORRELATION AND REGRESSION

# Dealing with outliers

# Dealing with outliers

```
> ggplot(data = regulars_plus, aes(x = SB, y = HR)) +
    geom_point() +
    geom_smooth(method = "lm", se = 0)
```



# The full model

```
> coef(lm(HR ~ SB, data = regulars_plus))
(Intercept)          SB
    19.3282     -0.2086
```

# Removing outliers that don't fit

```
> regulars <- regulars_plus %>%
    filter(!(SB > 60 & HR > 20)) # remove Henderson
> coef(lm(HR ~ SB, data = regulars))
(Intercept)          SB
    19.7169     -0.2549
```

- What is the justification?
- How does the scope of inference change?

# Removing outliers that do fit

```
> regulars_new <- regulars %>%
    filter(SB < 60) # remove Pierre
> coef(lm(HR ~ SB, data = regulars_new))
(Intercept)          SB
    19.6870     -0.2514
```

- What is the justification?
- How does the scope of inference change?

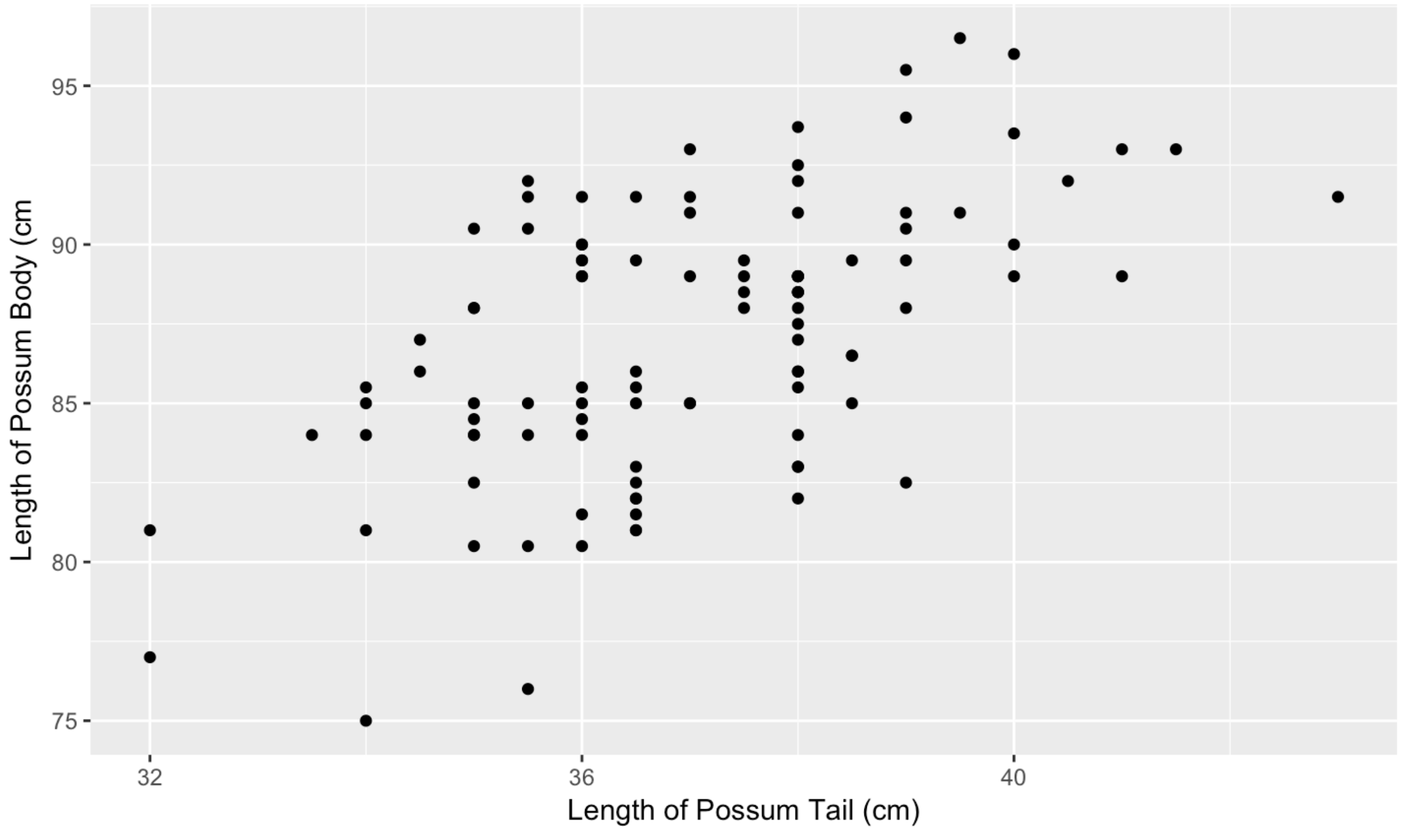
CORRELATION AND REGRESSION

# Let’s practice!

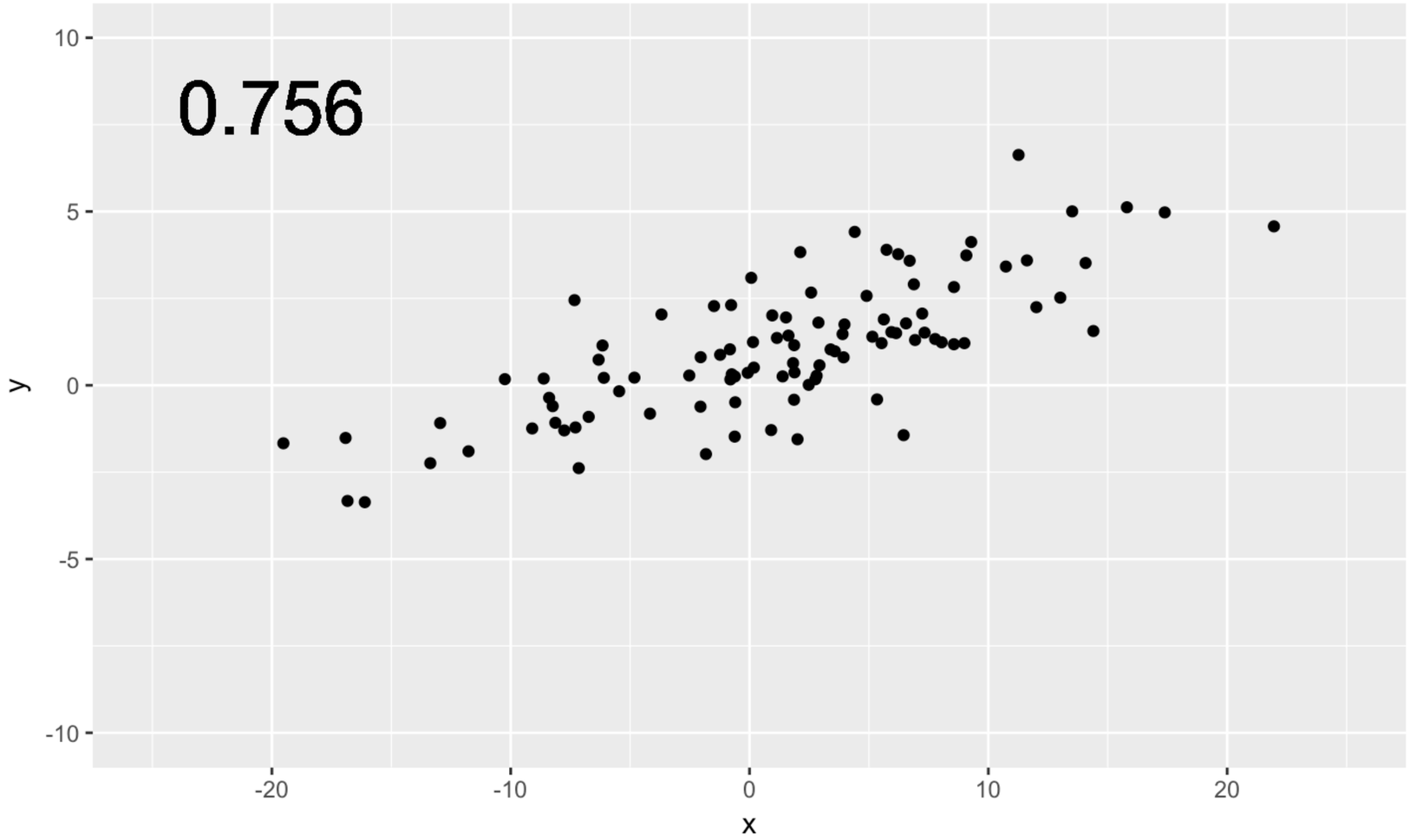
CORRELATION AND REGRESSION

# Conclusion

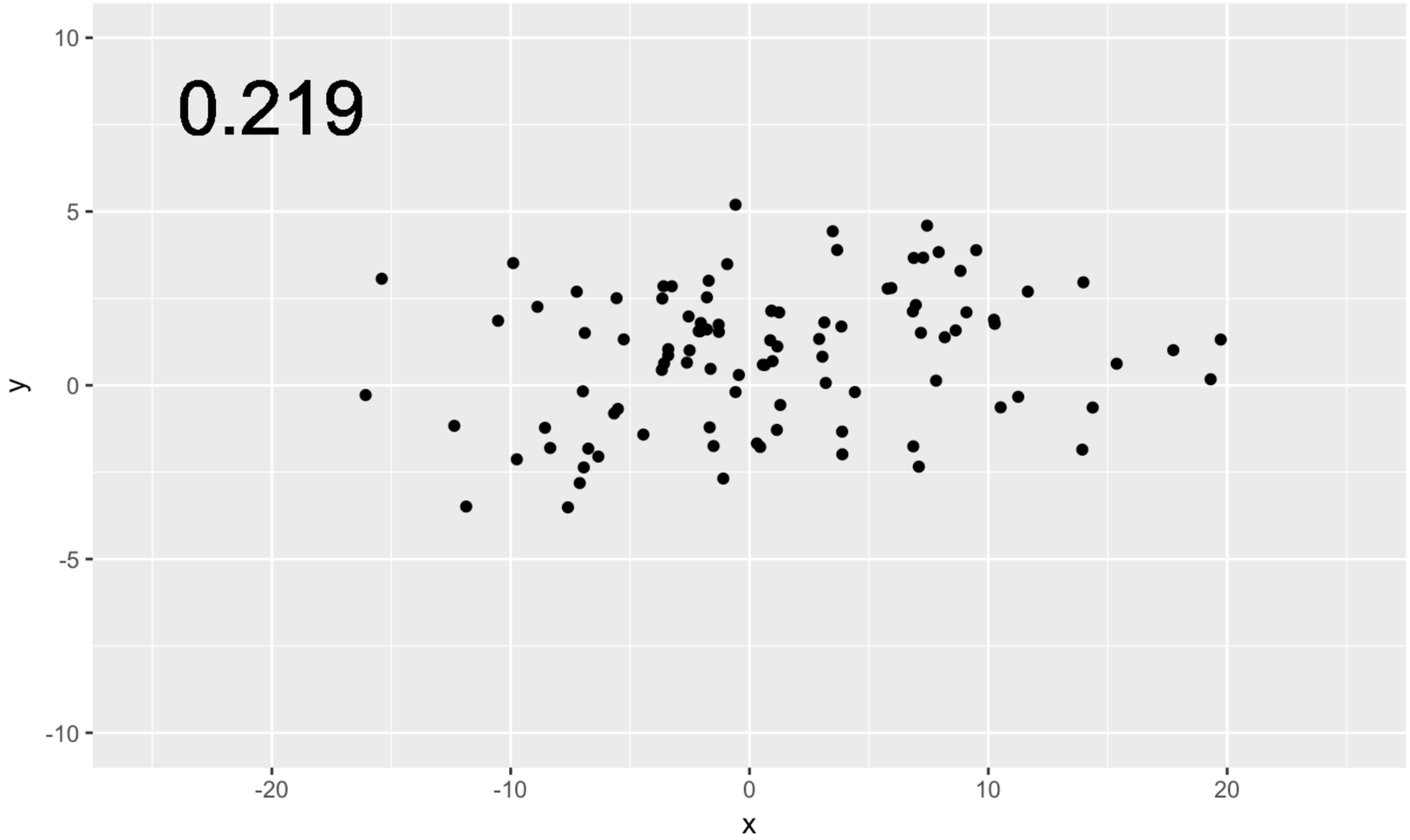
# Graphical: scatterplots



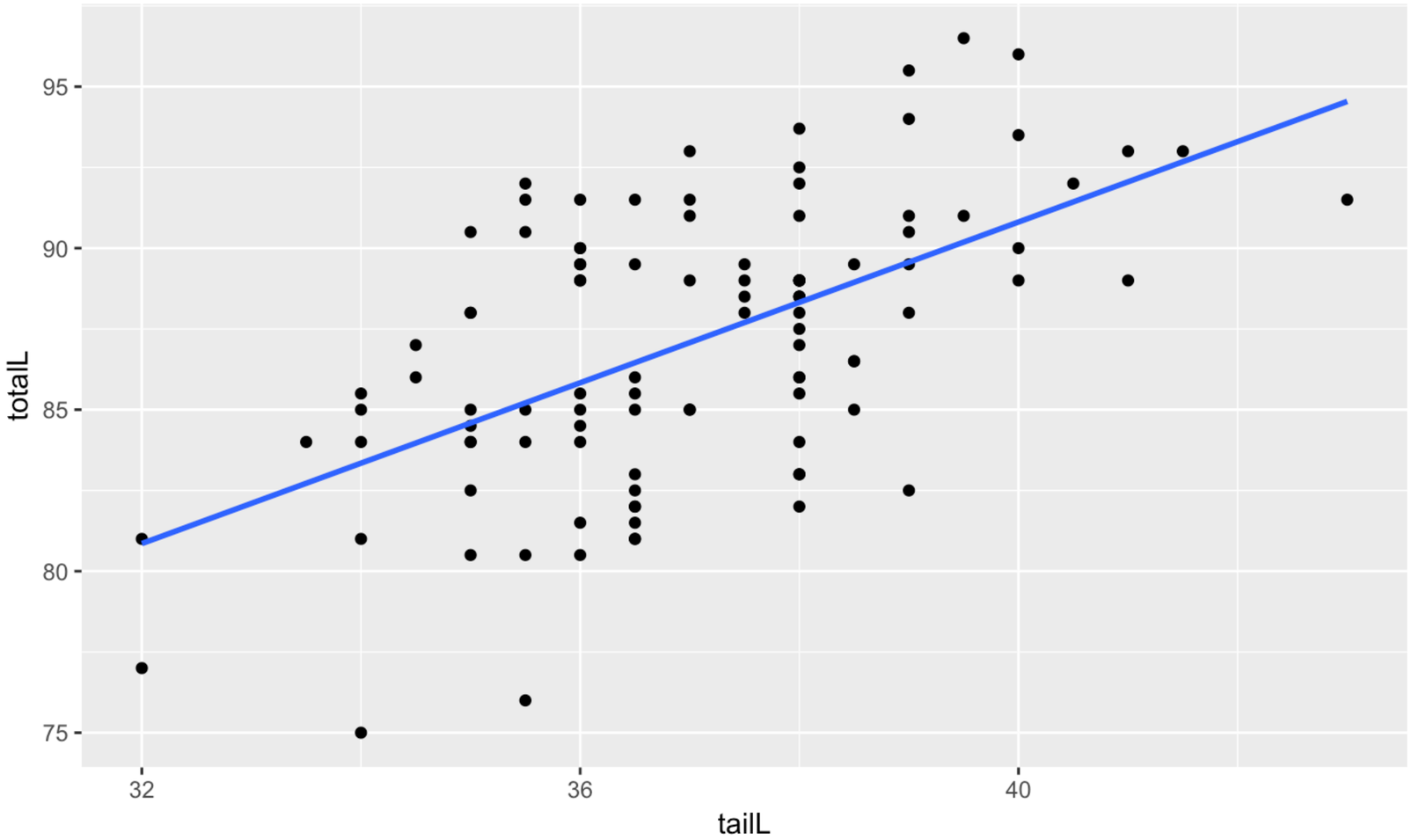
# Numerical: correlation



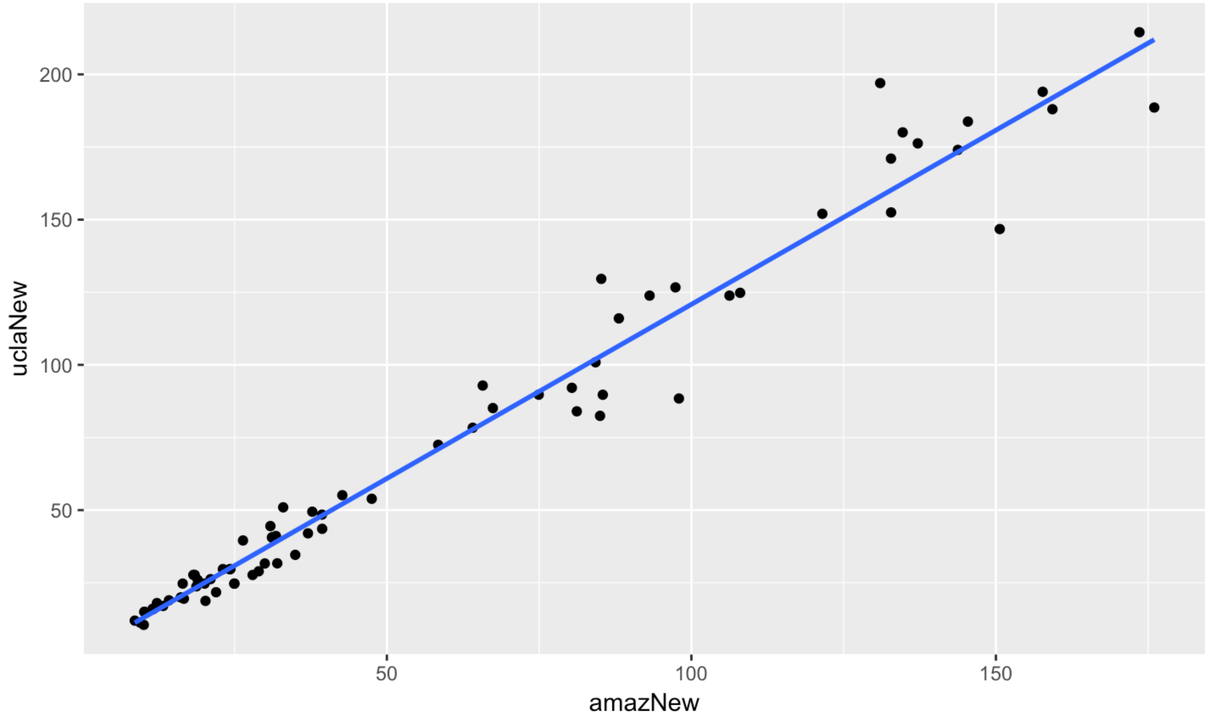
# Numerical: correlation



# Modular: linear regression



# Focus on interpretation

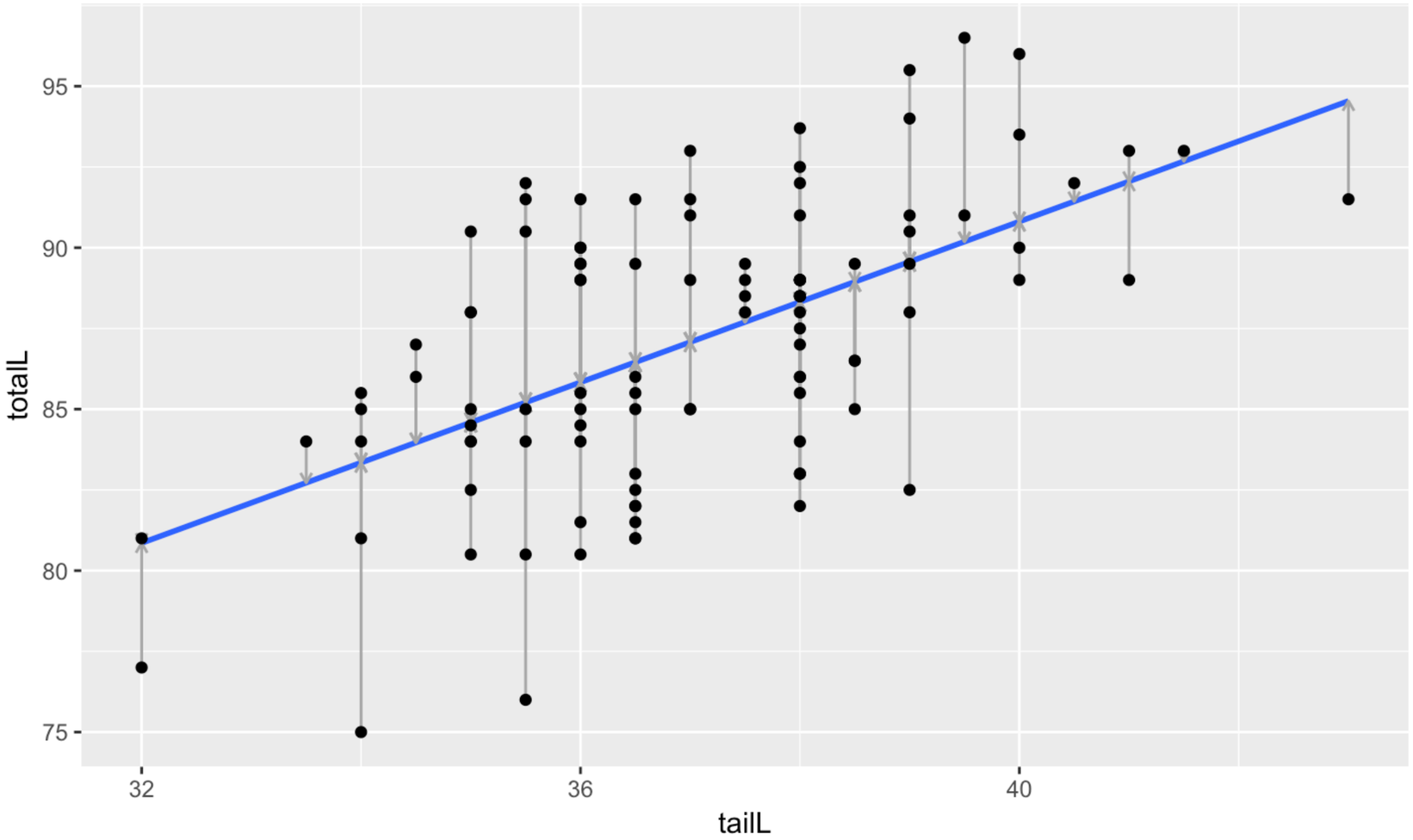



$$\widehat{uclaNew} = 0.929 + 1.199 \cdot amazNew$$

# Objects and formulas

```
> summary(mod)

Call:
lm(formula = uclaNew ~ amazNew, data = textbooks)

Residuals:
   Min     1Q Median     3Q    Max
-34.78  -4.57   0.58   4.01  39.00

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9290     1.9354    0.48     0.63
amazNew       1.1990     0.0252   47.60   <2e-16

Residual standard error: 10.5 on 71 degrees of freedom
Multiple R-squared:  0.97,   Adjusted R-squared:  0.969
F-statistic: 2.27e+03 on 1 and 71 DF,  p-value: <2e-16
```

# Model fit





# Thanks!