The General Linear Model

Correlation and Bivariate Regression

Martin Corley

Today

1. Correlation
   - Basics of Correlation
   - Covariance
   - Pearson's r & Spearman's ρ
2. Interpreting Correlation
   - Scatterplots
   - Statistical Significance
   - Caveats
3. Regression
   - Introduction
   - Basics of Regression
   - Example
   - Visualisation

Part I

Correlation

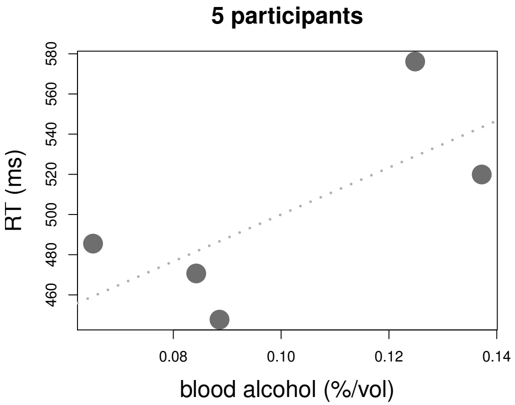
Correlation

- in **correlation**, both variables are ordinal or better
- aim of the game is to find out whether they're *related*
- no special status for 'IV' or 'DV', *other than by interpretation*

- is *blood alcohol* related to *reaction time*?

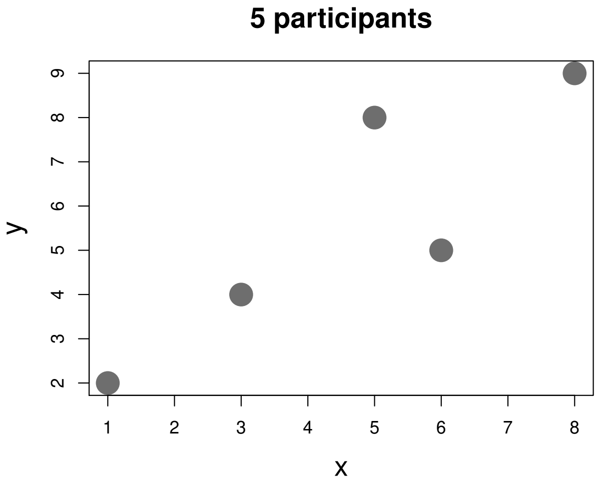
- as *blood alcohol* increases, does *reaction time* change systematically?

Scatterplot



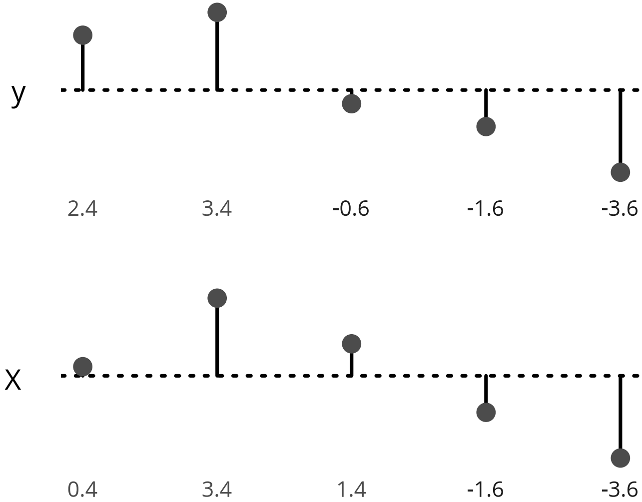
- each point represents pair of values for one participant

Simpler Data



- does y vary with x?
- equivalent to asking 'does y differ from its mean in the same way that x does?'

Covariance



- if observations of each variable differ *proportionately* from their means, it's likely the variables are related

Covariance

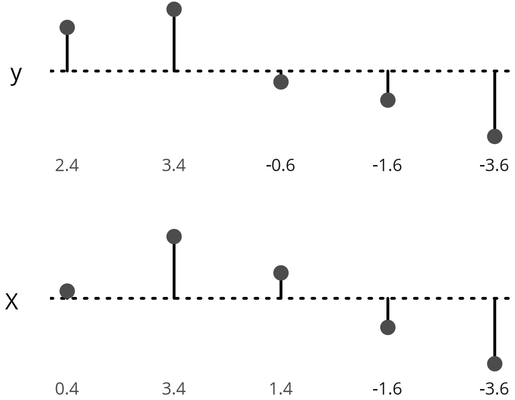
Variance

$$s^2 = \frac{\sum (x - \bar{x})^2}{N} = \frac{\sum (x - \bar{x})(x - \bar{x})}{N}$$

Covariance

$$\mathrm{cov}(x, y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{N}$$

Covariance



0.96 11.56 -0.84 2.56 12.96

$$\mathrm{cov}(x, y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{N} = \frac{27.2}{5} = 5.44$$

The Problem With Covariance

- covariance expresses the 'amount of shared variance'
- but it depends on the *units*

- imagine the last example was in *miles*...
- if we measured the same distances in km, the covariance would be 14.09 instead of 5.44

- we need some way to *standardise* covariance

Correlation Coefficient

- the standardised version of covariance is the **correlation coefficient**, r

$$r = \frac{\text{covariance}(x, y)}{\text{standard deviation}(x) \cdot \text{standard deviation}(y)}$$

Correlation Coefficient

Pearson's Correlation Coefficient

$$r = \frac{\frac{\sum(x-\bar{x})(y-\bar{y})}{N}}{\sqrt{\frac{\sum(x-\bar{x})^2}{N}}\sqrt{\frac{\sum(y-\bar{y})^2}{N}}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

$$= \frac{27.2}{\sqrt{33.2}\sqrt{29.2}} = \frac{27.2}{5.76 \cdot 5.40} = \frac{27.2}{31.14} = 0.87$$

Spearman's ρ

Spearman's Correlation Coefficient

Spearman's ρ is calculated in *exactly the same way* as Pearson's r, but uses the **ranks** of x and y (x_r and y_r) instead of their *values*

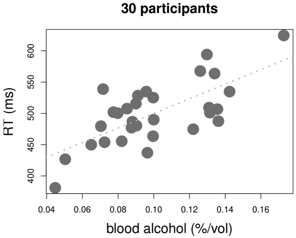
$$\rho = \frac{\sum (x_r - \bar{x}_r)(y_r - \bar{y}_r)}{\sqrt{\sum (x_r - \bar{x}_r)^2}\sqrt{\sum (y_r - \bar{y}_r)^2}}$$

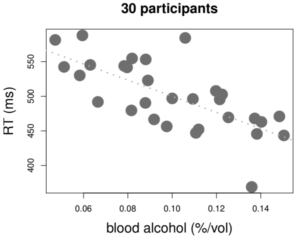
- for our toy data, $\rho = 0.9$

Correlation Coefficient

- measure of *how related* two variables are

- $-1 \leq r \leq 1$ (± 1 = perfect fit, 0 = no fit)
- *sign* tells you direction of slope

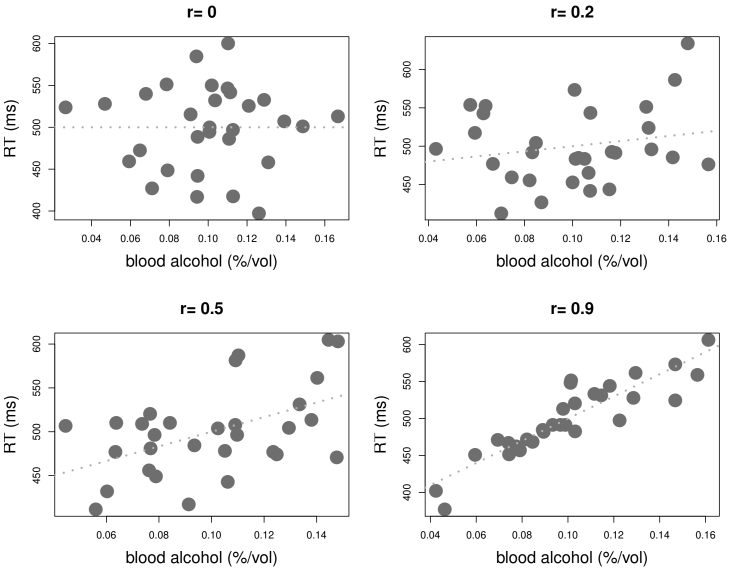




- $r = 0.7$

- $r = -0.7$

Scatterplots

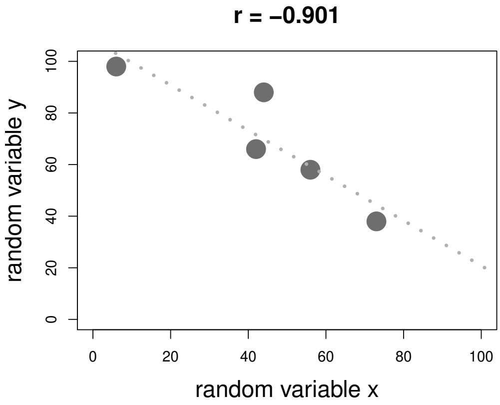


Significance of a Correlation

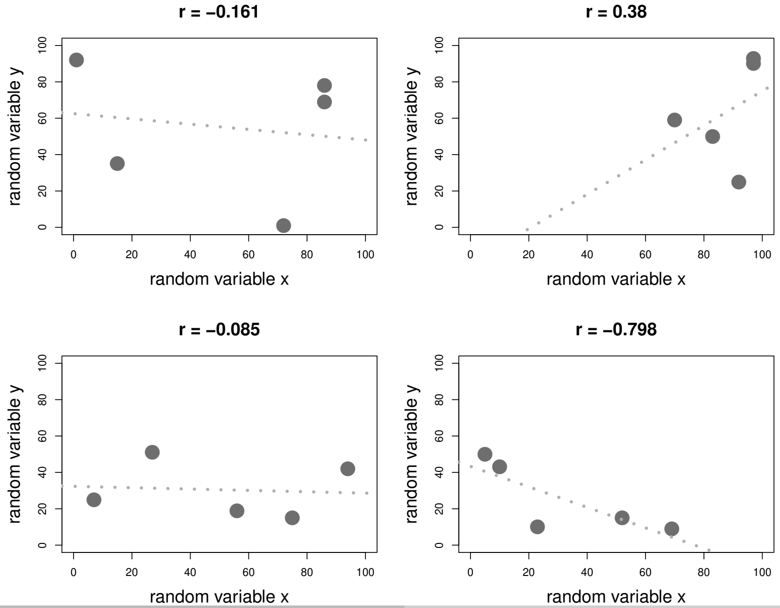
- we can measure a correlation using *r* or ρ as appropriate
- we want to know whether that correlation is *significant*
  - i.e., whether the probability of finding it *by chance* is low enough
- cardinal rule in NHST: compare everything to chance
- let's investigate...

Random Correlations

- pick 5 pairs of numbers at random...
y	66	58	98	88	38
x	42	56	6	44	73

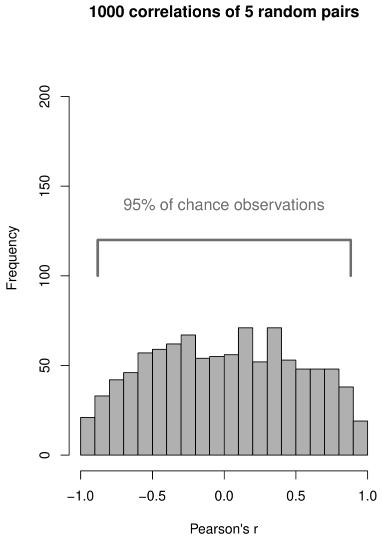


Random Correlations

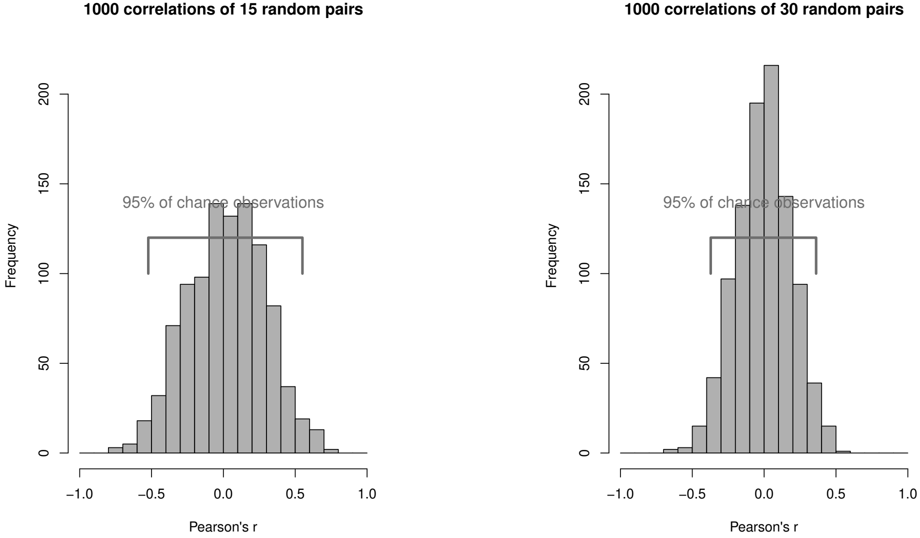


Lots of Random Correlations

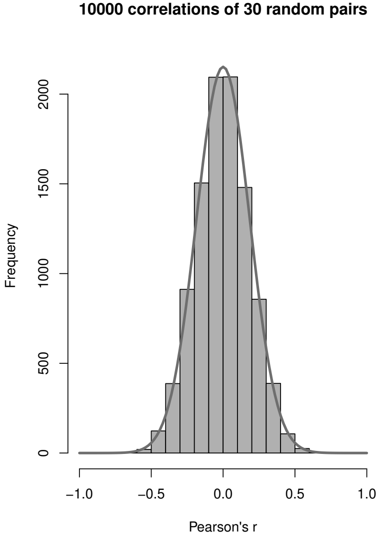
- histogram of random correlations
- (here, 1000 samples of 5 random pairs)



Lots of Random Correlations



As the Sample Tends to ∞



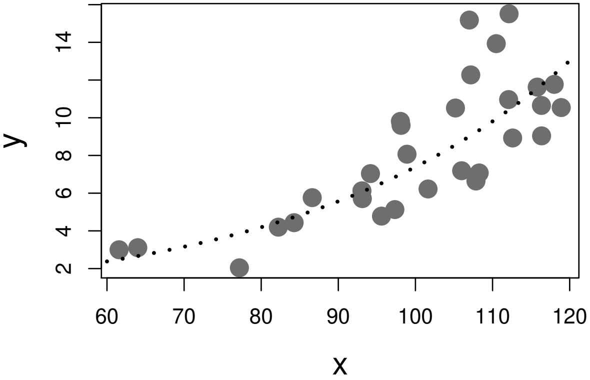
- distribution of random *r*s is *t* distribution

$$t = r\sqrt{\frac{N-2}{1-r^2}}$$

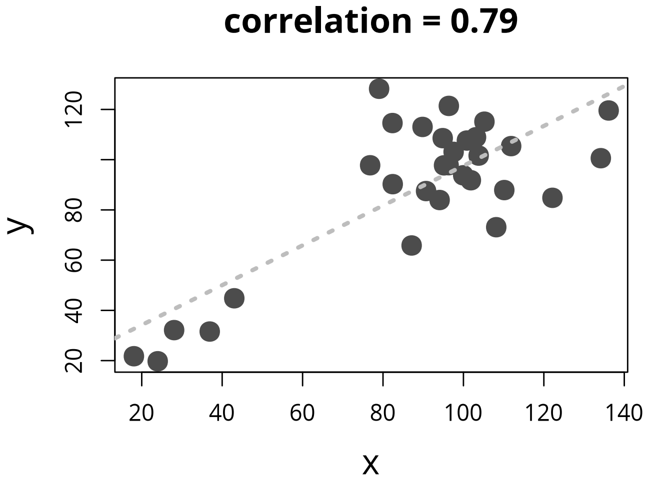
- makes it 'easy' to calculate probability of getting *r* for sample size *N* *by chance*
- in practice, use look-up tables

Beware False Positives

- correlations assume a *linear* relationship
- but the relationship might be something else. . .



Beware False Positives

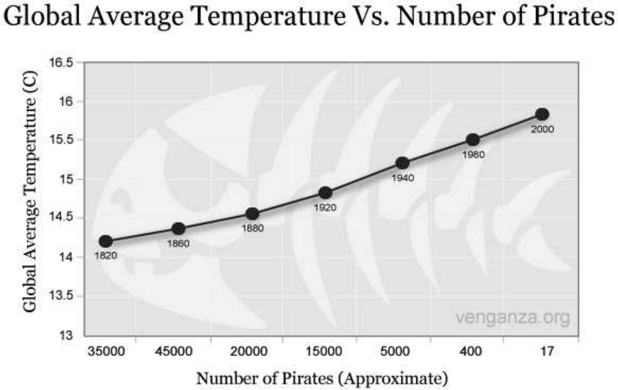


- correlation driven by a few unusual observations
- always look at scatterplots together with calculations

Interpreting Correlation

- correlation does not imply causation
- correlation merely suggests that two variables are related

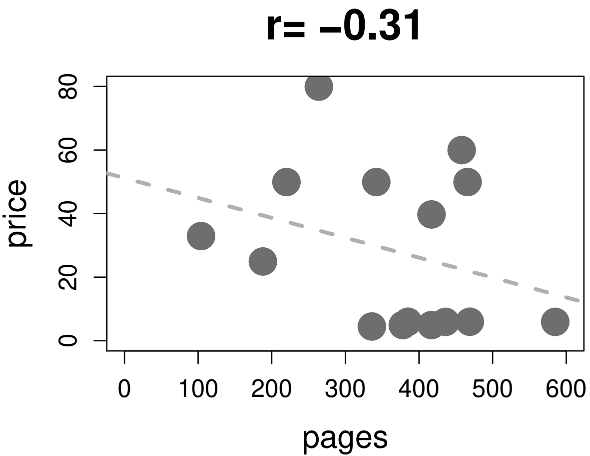
Pirates



- clear negative correlation between numbers of pirates and mean global temperature

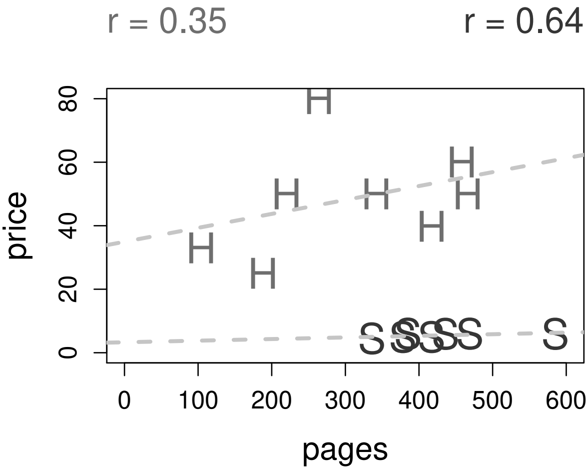
→ we need pirates to combat global warming

Books



- sample of books suggests that books with more pages cost less

Books



- hardbacks and softbacks mixed together
- an example of the **third variable** problem

(Utts, 1996)

Correlation

- correlation tests for the *relationship* between two variables
- *interpretation* of that relationship is key
- never rely on statistics such as r without looking at your data

Part II

Regression

A Word-Naming Experiment

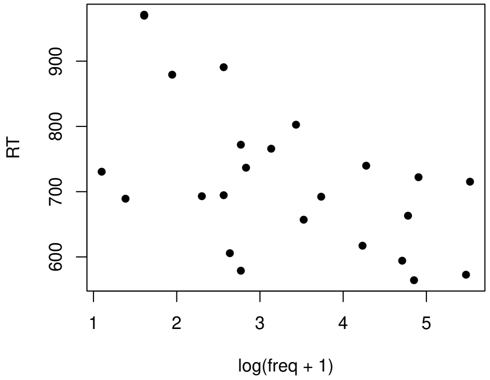
using entirely fictitious data

```
load(url("https://is.gd/refnet"))
ls()
## [1] "naming"
summary(naming)
##      length          freq        pos          RT
##  Min.   : 4   Min.   :   0   N:80   Min.   : 332
##  1st Qu.: 7   1st Qu.:   9   V:80   1st Qu.: 626
##  Median : 8   Median :  21   A:80   Median : 689
##  Mean   : 8   Mean   :  61          Mean   : 695
##  3rd Qu.: 9   3rd Qu.:  52          3rd Qu.: 770
##  Max.   :13   Max.   :1452          Max.   :1003
```

- `RT` = naming-aloud times (for 240 words)
- `length` in characters
- `freq` in wpm
- `pos`: `N`oun, `V`erb, or `A`djective

A Subset of the Data

```
with(n2, plot(RT ~ log(freq + 1), pch = 16))
```



(NB., add 1 to freq to avoid `log(0)`)

Correlation

- is word frequency related to time to name a word?

```
# could use cor.test(~RT+log(freq+1),data=n2)
with(n2, cor.test(RT, log(freq + 1)))
```

```
##
##  Pearson's product-moment correlation
##
## data:  RT and log(freq + 1)
## t = -3, df = 20, p-value = 0.006
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.78 -0.18
## sample estimates:
##   cor
## -0.55
```

- yes it is, negatively
- RT goes down as frequency goes up
- but is that really *all* we can say?

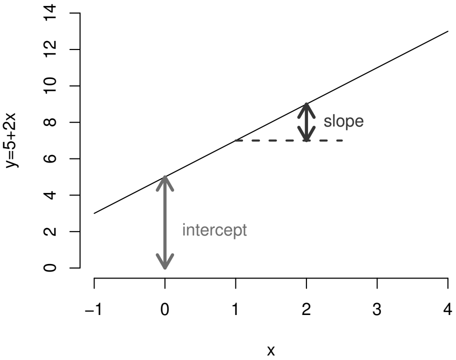
The Only Equation You Will Ever Need

A General Model of Observed Data

$$\text{outcome}_i = (\text{model}) + \text{error}_i$$

- to get further, we need to make *assumptions*
- nature of the **model** (linear)
- nature of the **errors** (normal)

Linear Models

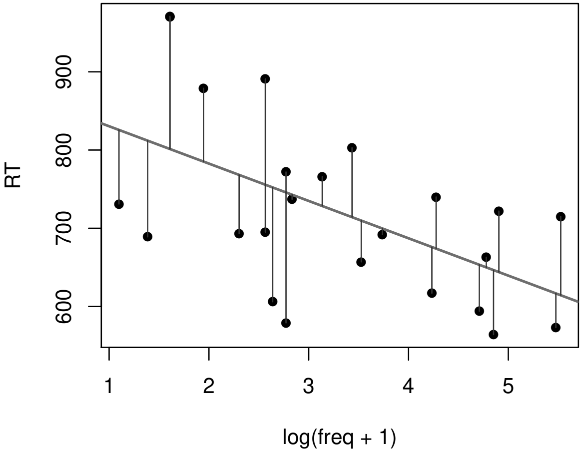


Linear Model

$$\hat{y}_i = b_0 \cdot 1 + b_1 \cdot x_i$$

`y ~ 1 + x`

A Linear Model



- a linear model describes the best line through the data
- the best-fit line minimizes the residuals

Residuals

- each $\hat{y}_i$ is an *estimate* according to the model
- the *real* observation for each x_i is y_i
- $y_i - \hat{y}_i$ is the **residual**, ϵ_i

$$\hat{y}_i = b_0 + b_1 x_i$$

the best-fit line

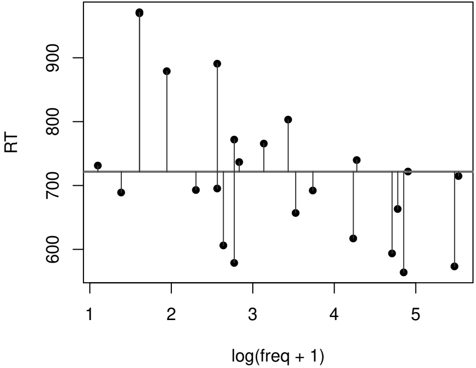
$$y_i = b_0 + b_1 x_i + \epsilon_i$$

the data

Total Sum of Squares (SS_{total})

$$SS_{total} = \sum (y - \bar{y})^2$$

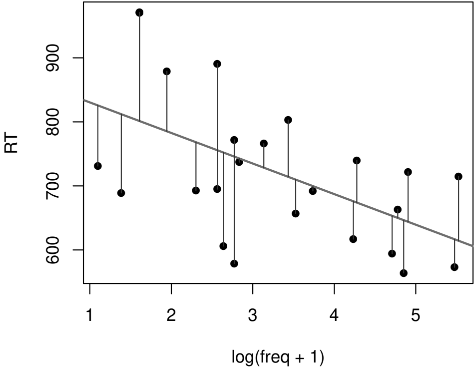
- sum of squared differences between observed y and mean $\bar{y}$
- how much does the observed data vary from a model which says 'there is no effect of x' (**null model**)?



Residual Sum of Squares ($SS_{residual}$)

$$SS_{residual} = \sum (y - \hat{y})^2$$

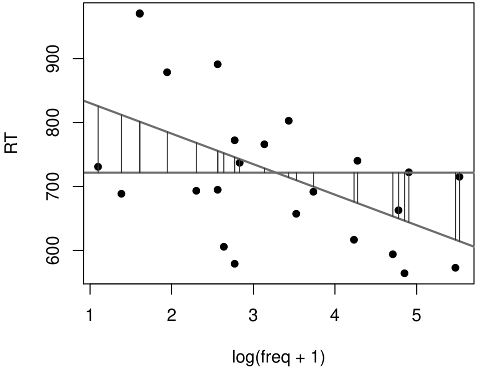
- sum of squared differences between observed y and predicted $\hat{y}$
- how much does the observed data vary from the existing model?



Model Sum of Squares (SS_{model})

$$\begin{aligned} SS_{model} &= \sum(\hat{y} - \bar{y})^2 \\ &= SS_{total} - SS_{residual} \end{aligned}$$

- sum of squared differences between predicted $\hat{y}$ and mean $\bar{y}$
- how much does the existing model vary from the null model?



Testing the Model: R^2

How much of the variance does the model account for?

$$R^2 = \frac{\text{SS}_{\text{model}}}{\text{SS}_{\text{total}}}$$

- indicates how much the model improves the prediction of $\hat{y}$ over the null model

- $O \leq R^2 \leq 1$
- we want R^2 to be *large*
- for a single predictor, $\sqrt{R^2} = |r|$ (where r is Pearson's correlation coefficient)

Testing the Model: F

- F-ratio depends on **mean squares**
- $\mathrm{MS}_x = \mathrm{SS}_x/\mathrm{df}_x$

How much does the model improve over chance?

$$F = \frac{\mathrm{MS}_{\mathrm{model}}}{\mathrm{MS}_{\mathrm{residual}}}$$

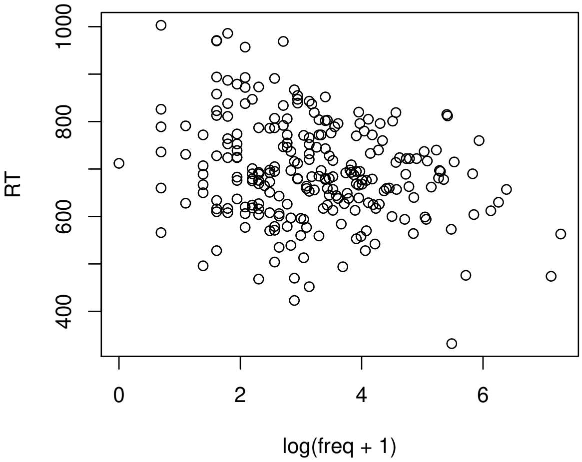
- indicates how much better the model predicts $\hat{y}$ compared to chance

- $0 < F$
- we want F to be *large*

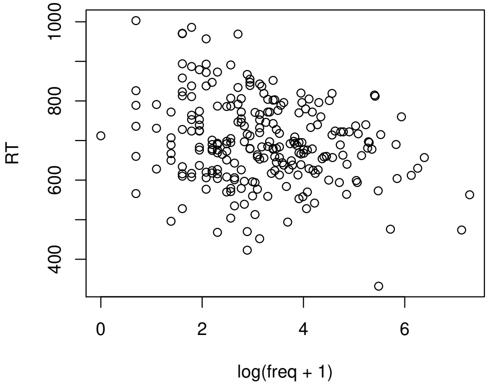
- significance of F does not always equate to a large (or theoretically sensible) effect

This Time, for Real

```
with(naming, plot(RT ~ log(freq + 1)))
```



Correlation

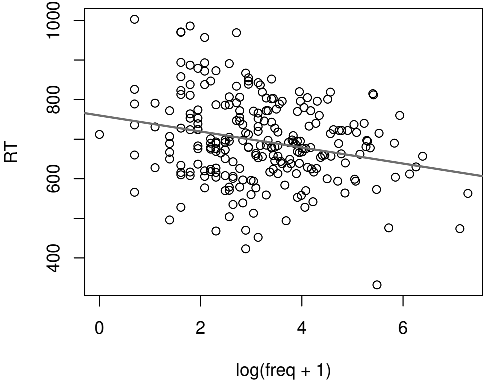


$$t = r\sqrt{\frac{N-2}{1-r^2}}$$

```
r <- with(naming, cor(RT, log(freq + 1)))
r
## [1] -0.24
pt(r * sqrt((length(naming[, 1] - 2)/(1 - r^2))), df = 22)
## [1] 0.00041
```

This Time, for Real

```
with(naming, plot(RT ~ log(freq + 1)))
```



- a linear model can tell us more about the data...

A Simple Linear Model

```
model <- lm(RT ~ log(freq + 1), data = naming)
summary(model)
```

```
## Call:
## lm(formula = RT ~ log(freq + 1), data = naming)
## ...
## Multiple R-squared:  0.0587,Adjusted R-squared:  0.0548
## F-statistic: 14.8 on 1 and 238 DF,  p-value: 0.00015
```

- R^2 and F are basic indicators of how ‘good’ a model is
- part of R’s output when summarising an `lm` object
- we’ll revisit adjusted R^2 later

A Simple Linear Model

```
summary(model)
```

```
## Call:
## lm(formula = RT ~ log(freq + 1), data = naming)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -316.9   -65.2    -6.1    70.4   263.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     759.87      18.04   42.13  < 2e-16 ***
## log(freq + 1)   -20.24       5.25   -3.85  0.00015 ***
## ...
```

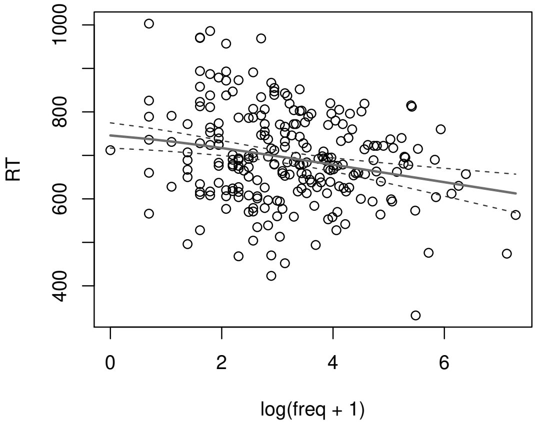
- glancing at `Residuals` gives an indication of whether they are roughly symmetrically distributed
- the `Coefficients` give you the model
- the `Estimate` for `(intercept)` is b_0
- the `Estimate` for `log(freq + 1)` is b_1, the **slope**

Coefficients

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       759.87      18.04   42.13  < 2e-16 ***
## log(freq + 1)     -20.24       5.25   -3.85  0.00015 ***
```

- *independently* of whether the model fit is 'good', coefficients can tell us about our data
- here, the `(Intercept)` b_0 isn't that useful
  - → it takes 760ms to name 'zero-frequency words'
- but the slope b_1 of `log(freq + 1)` is quite informative
  - → words are named 20ms faster per unit increase
  - this is a significant finding
  - calculated from the estimated coefficient and its `Std. Error`, using the t distribution

Visualisation (using `predict()`)



(confidence intervals for the *model*)

⏩ skip scaling

Scaling of Predictors

- 'words of zero frequency' may not be very meaningful
- can **rescale** predictor to make interpretation more useful
- can also be used to ameliorate collinearity

```
model.S <- lm(RT ~ I(log(freq + 1) - mean(log(freq + 1))), data = naming)
summary(model.S)
```

```
## ...
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   695.38       6.72  103.41  < 2e-16 ***
## I(lf)         -20.24       5.25   -3.85  0.00015 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 104 on 238 degrees of freedom
## Multiple R-squared:  0.0587,Adjusted R-squared:  0.0548
## F-statistic: 14.8 on 1 and 238 DF,  p-value: 0.00015
```

- slope unchanged
- 695ms corresponds to words of mean log frequency

Scaling of Predictors

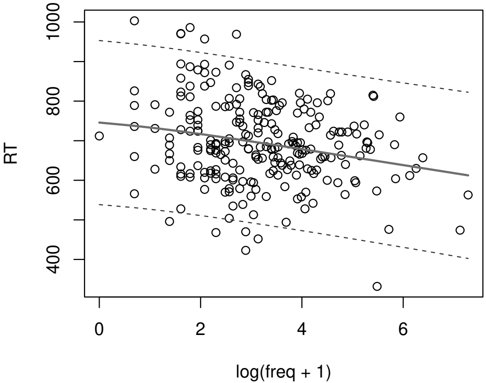
- *linear* scaling of predictors doesn't change model fit

```
summary(model)$r.squared
## [1] 0.059
summary(model.S)$r.squared
## [1] 0.059
summary(lm(RT ~ I(5 * log(freq + 1)), data = naming))$r.squared
## [1] 0.059
```

- *non-linear* scaling—like `log()` above—changes fit

```
summary(lm(RT ~ freq, data = naming))$r.squared
## [1] 0.044
```

Visualisation (using `predict()`)



(confidence intervals for *predicted observations*)