Linear Regression, part 2

March 23

- Today
  - Finish up 1 topic relating to simple regression
  - Review some previously-discussed concepts
  - Lecture on multiple regression
- Tomorrow
  - Practical
- Homework
  - due in 2 weeks (April 10)
- Next week
  - no lecture
  - Practical (more practice with regression)
- Next lecture
  - April 10
  - no practical that week

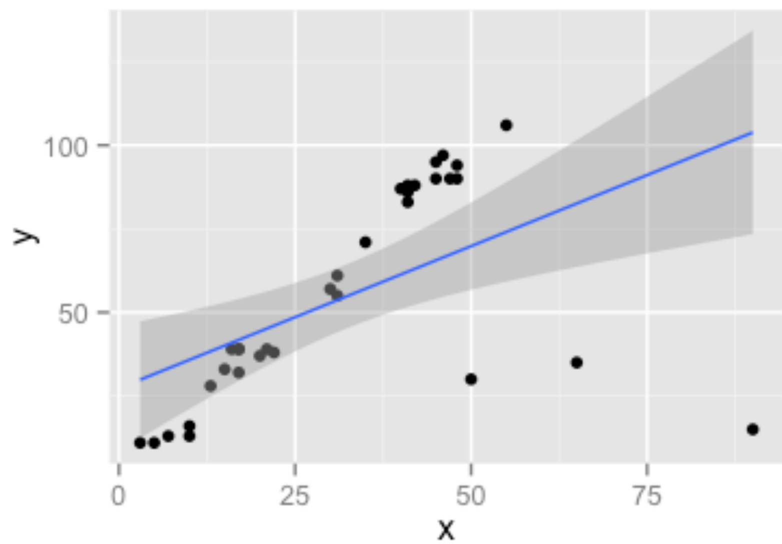
March 23

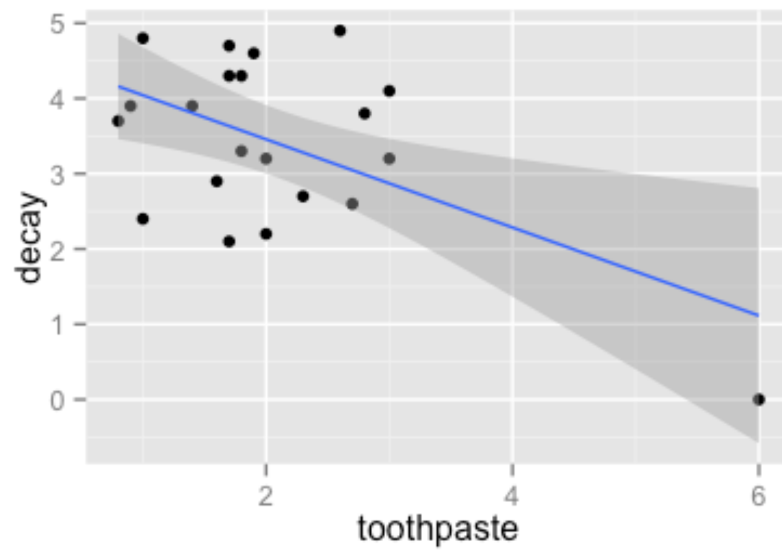
- End-of-quarter dates
  - Final Practical: May 12 (used for review)
  - Exam: May 15
  - Final assignment available: May 15
  - Deadline final assignment: May 29
  - Final grades: June 12
  - Resit final assignment available: June 12
    - ideally completed in pairs
  - Resit deadline for final assignment/Resit exam day: June 26
  - Resit final grades: July 6
- Resit of whole course is also available next year

What happens when you violate assumptions?

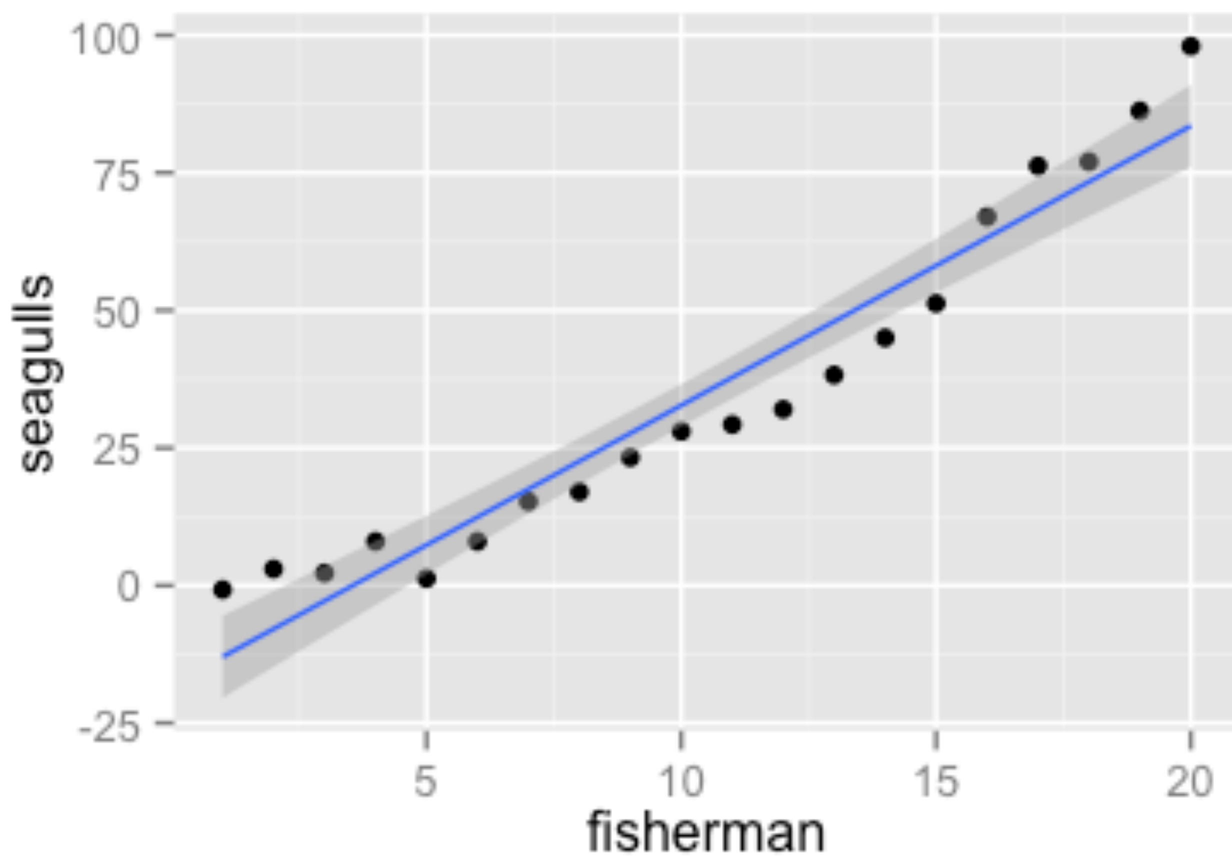
- Model doesn't generalize
- What can you do?
  - residuals
    - transformation of data? (e.g., take the log values of one [or more] of your variables)
      - must transform ALL values of the variable
    - choose a different method
  - highly influential points
    - if good theoretical reason, remove
    - run model with and without the outlier

To exclude or not to exclude





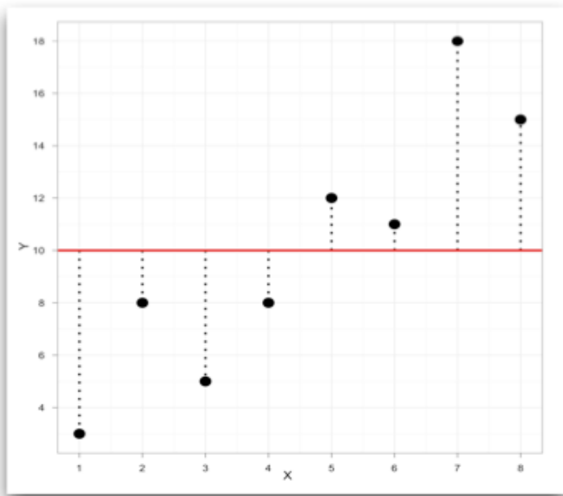
Nonlinear Relationships

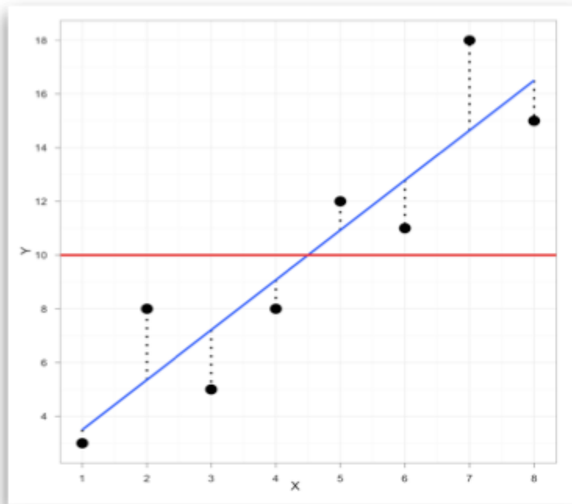


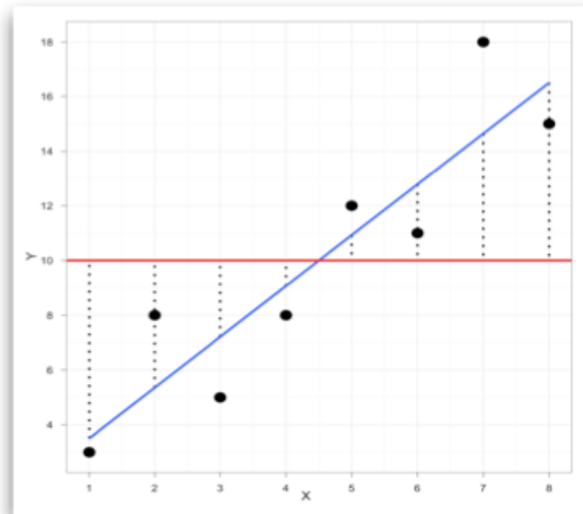
Fitting a Polynomial Curve In RStudio

Review of R^2

Total Sum of Squares (SSt) - Residual Sum of Squares (SSr) = Model Sum of Squares (SSm)





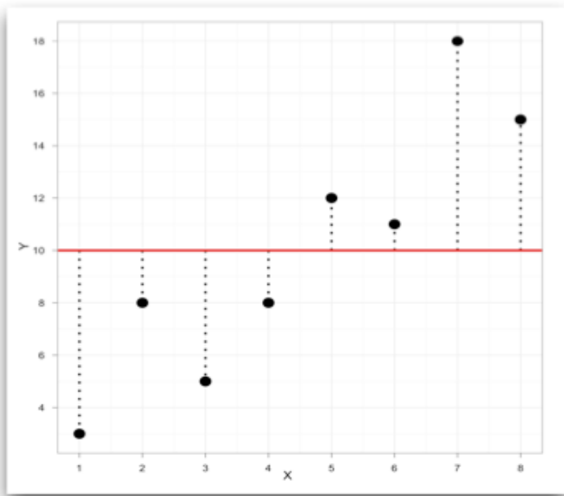


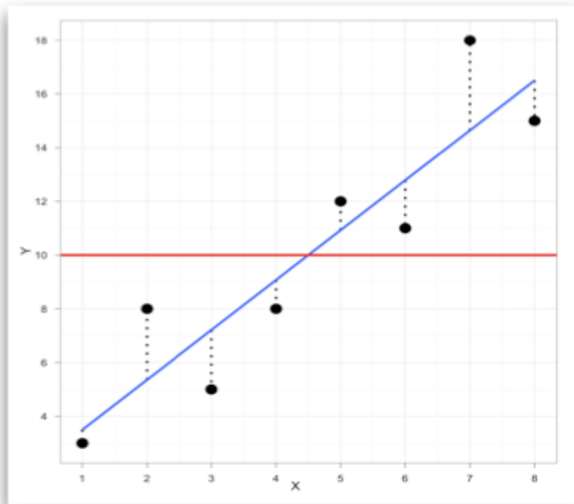
$$\frac{\text{Model Sum of Squares}}{\text{Total Sum of Squares}} = \frac{\text{Explained Variance}}{\text{Total Variance}} = R^2$$

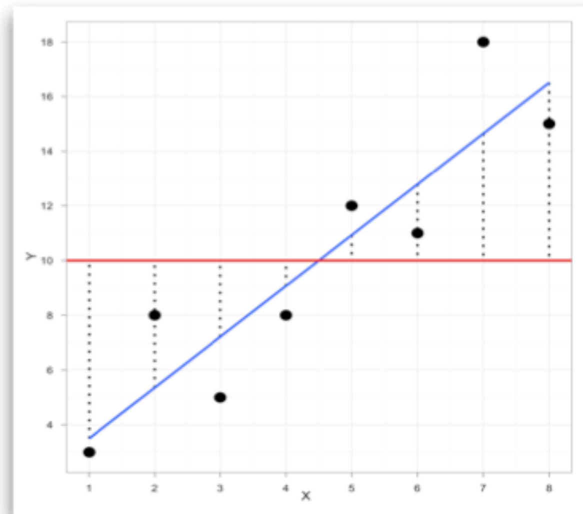
R^2 x 100 = Percentage Explained Variance (so R^2 of 1 would represent perfect explanation)

Review of F

Total Sum of Squares (SSt) - Residual Sum of Squares (SSr) = Model Sum of Squares (SSm)







Residual Mean Squares (MSr)

Mean Squares for the Model (MSm)

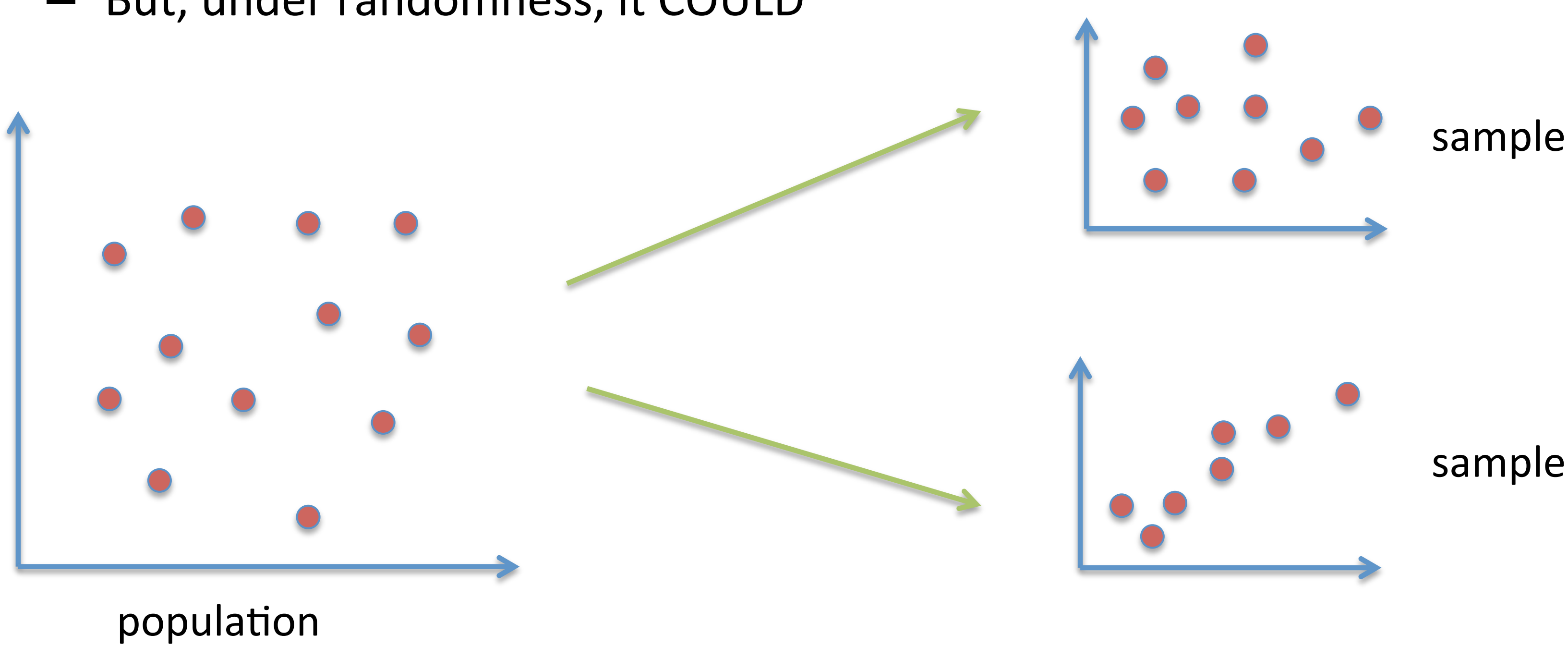
$$\frac{\text{Mean Squares for the Model}}{\text{Residual Mean Squares}} = \frac{\text{Explained Variance}}{\text{Unexplained Variance}} = F$$

R^2 versus F

- R^2
  - explained variance / total variance
  - measure of effect size
  - measure of fit of regression line to data
- F
  - explained variance / unexplained variance
  - test statistic
  - used to determine whether the fit of regression line to data is significant

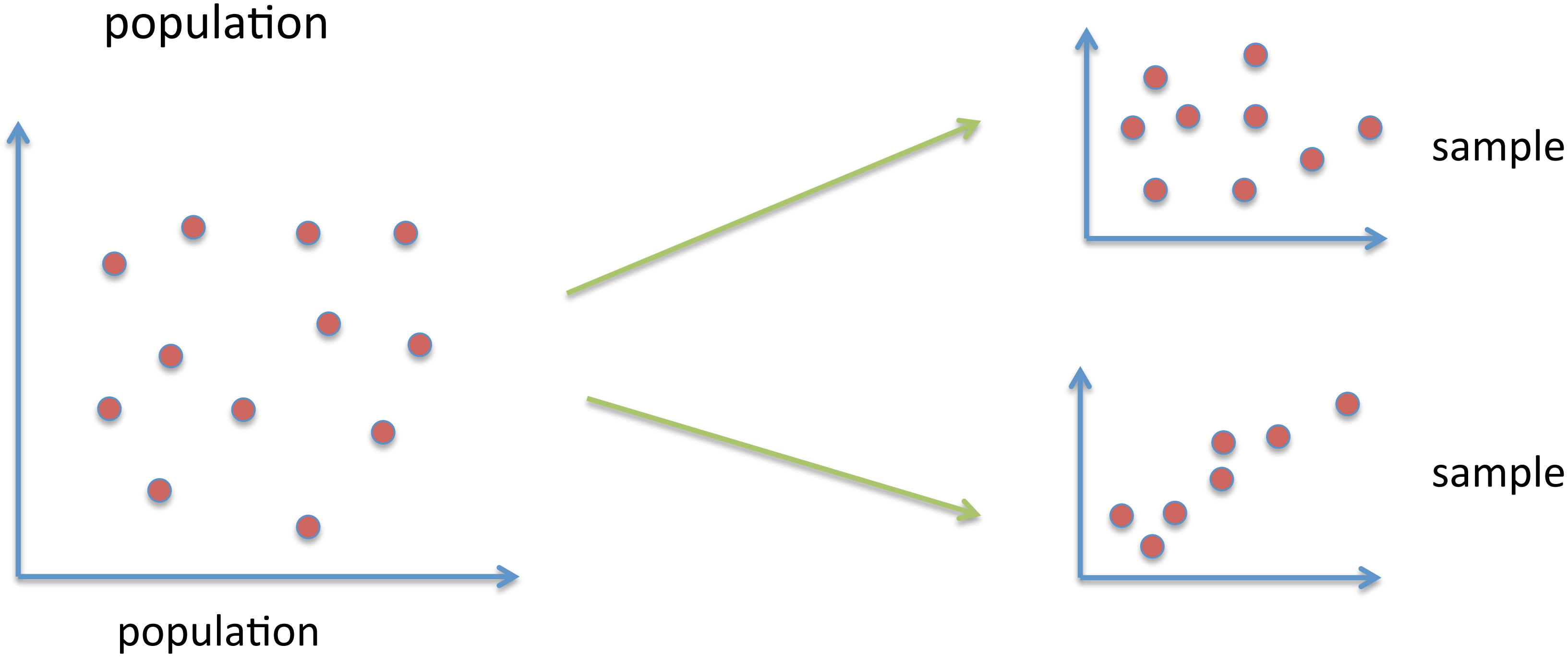
Review of test statistics & statistical significance

- Test statistics (e.g., F, t) compare explained to unexplained variance
- The probabilities of obtaining particular values of test statistics under the null hypothesis are known
  - Assume there is no pattern in the population (=null hypothesis)
  - We take a random sample from the population
  - Most likely, the random sample will have no pattern in it either
  - But, under randomness, it COULD



Review of test statistics & statistical significance

- We compute the ratio of explained to unexplained variance (F)
  - If there is a strong effect in the sample, F will be high
  - You are unlikely to get that high of an F (that is, that high of a ratio of explained to unexplained variance) if there's no pattern in the population
  - Mathematicians have figured out the probabilities for every possible value of F (see the F distribution) assuming no pattern in the population



Review of test statistics & statistical significance

- Statistical significance
  - When the probability (p-value) of a test statistic falls below a threshold (conventionally 0.05)
  - Less than 5% chance you would get a random sample with this high of a ratio of explained to unexplained variance in it if the population you sampled from has no such pattern in it

Review of t Statistic

- Like F, t also compares explained to unexplained variance
- But we use t to examine whether the betas are significantly different from 0
  - explained variance: difference between a beta and 0
  - unexplained variance: how variable the value of beta would see across different samples
    - standard deviation of the sampling distribution
    - estimated via standard error

$$t = \frac{\text{beta}}{\text{standard error}_{\text{beta}}}$$

Role of R^2, F, and t in lm() output

summary(albumSales.1)

>Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.341e+02	7.537e+00	17.799	<2e-16 ***
adverts	9.612e-02	9.632e-03	9.979	<2e-16 ***

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

Residual standard error: 65.99 on 198 degrees of freedom

Multiple R-squared: 0.3346, Adjusted R-squared: 0.3313

F-statistic: 99.59 on 1 and 198 DF, p-value: < 2.2e-16

What do we actually use linear regression for

- Prediction
- Stating that there is a relationship between two variables
  - but isn't this what correlation does?
  - true power of linear regression is when there are <u>more than one predictor variable</u>
    - multiple regression

Multiple Regression

- What if we wanted to predict album sales based on
  - amount of money spent on advertising
  - how often the song was played on the radio
- Our regression equation changes

$$Outcome_i = b_0 + b_1 * Predictor_{1i} + b_2 * Predictor_{2i} \ldots + b_n Predictor_{ni}$$

- We are no longer dealing with a line!
  - if n=2, then we have a regression <u>surface</u>
  - if n>2, difficult to visualize

Additional Assumption of Multiple Regression: Problem of Multicollinearity

- multicollinearity: when your predictors are correlated with each other
  - increases the standard error of the betas
    - your sample's betas less likely to be representative of the population's betas
  - difficult to know which predictors are important

Assessing Multicollinearity

vif(name_of_model)

- returns VIF values for each predictor
- problems:
  - largest VIF > 10
  - average VIF is substantially > 1
- if problems: use cor() to check which pairwise combinations of predictors are collinear
  - Pearson's r > 0.8 indicates highly correlated

R^2 with Multiple Regression

- SSt, SSr, SSm calculated similarly for multiple regression
- Multiple R^2 goes up with more predictors
  – adding predictors, even meaningless ones, will eat up unexplained variance randomly by chance
- We we also must pay attention to Adjusted R^2
  – Penalizes you for having many predictors
  – Tells us how much explained variance we would expect in the population

Significance in Multiple Regression

- Use overall p-value the same way as in simple regression
- Now the p-values of the betas matter!
  - indicate whether each predictor is significant
  - …we still don't care much about the intercept's p-value though ☹

Multiple Regression in RStudio

The problem of multiple possible models

- another possible predictor for album sales
  - physical attractiveness of the band
- possible models
  - sales ~ adverts + airplay + attract
  - sales ~ adverts + airplay
  - sales ~ airplay + attract
  - sales ~ adverts + attract
  - sales ~ adverts
  - …

The problem of multiple possible models

- Bigger is not always better
  - Remember R^2 penalty
- We need a process for finding a model containing the "best" combination of predictors
  - method of progressing through various models
  - method of comparing models to know which one is better

Methods of Model Selection

- Disagreement across authors/statisticians
- Hierarchical (the one advocated by the book)
    1. model 1 includes predictors shown meaningful by previous research
    2. model 2 includes additional predictors you hypothesize to be important
    3. models 3+ remove "statistically redundant" predictors
- Backward Step-wise
    1. model 1 contains all predictors
    2. models 2+ remove predictors 1-by-1 until you arrive at a "best" model
- Put all predictors in and leave them there!

Methods of Model Selection

- Our approach
  - Combination of hierarchical and backward step-wise
  - model 1: we'll start with all predictors that we consider to be (potentially) theoretically important
  - models 2+: we'll remove 1 predictor at a time, considering whether the new model is an improvement over the previous one

Akaike Information Criterion (AIC)

- measure of fit that penalizes the model for having more predictors
  - similar to multiple R^2
- bigger AIC values indicate worse fit
- We use AIC to compare different models
  - these models must have the same data

Comparing models

- drop1(name_of_model, test="F")
  - returns information about AIC of current model and different models if you were to drop particular predictors
- What to consider
  - does AIC drop (people sometimes say by more than 2)?
  - is the predictor non-significant?
  - does the predictor make theoretical sense?

Model Selection in RStudio

Reporting a Linear Regression

- Reproduce the information from the summary() function regarding the betas in a new table, and include this table as an appendix
  - include coefficients, SEs, t scores, and p-values
- In the text itself:
  - The final model's formula was sales ~ adverts + airplay + attract. All main effects were very significant: $p<0.001$, and the model was highly significant overall ($F_{3,\ 196}=129.5$, $p<0.001$) and achieved a high variance explanation (mult. $R^2=0.6647$, adj. $R^2=0.6595$). All regression coefficients, as well as their standard errors, t scores, and p-values, are provided in the appendix, and checking of model assumptions revealed no problems.