Linear models lecture

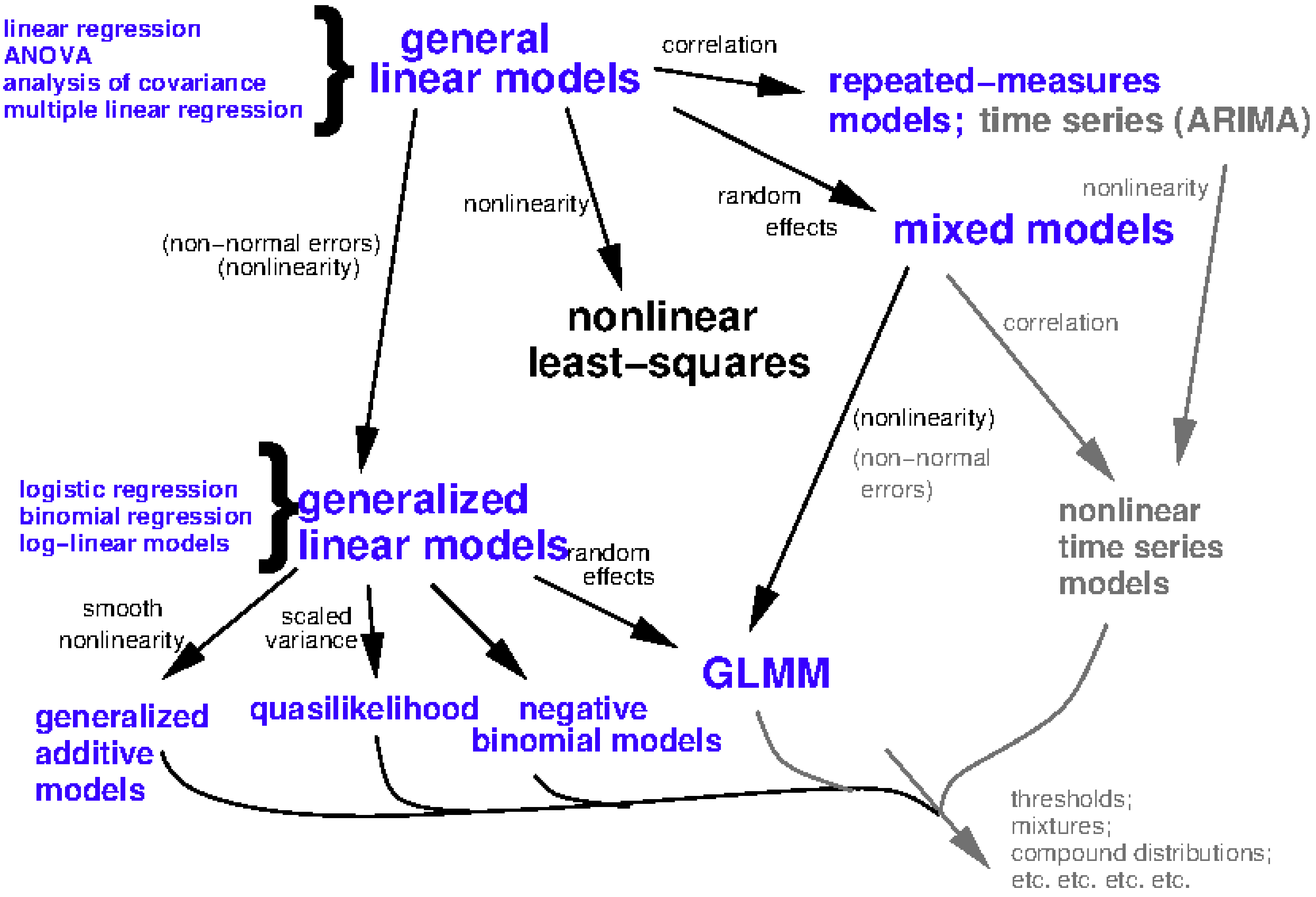
Kate Langwig (input from BB and JD)

Introduction

History

- ANOVA, ANCOVA, regression, t test are all variations of the same animal, the *general* linear model
- Many people (including the R project) call it a linear model (`lm`) to distinguish it from the *generalized* linear model (`glm`)
- these models are typical fit by ordinary least squares

(part of) the statistical universe



4/34

Extended linear models

- *Generalized* linear models can incorporate:
  - (Some) non-linear relationships
  - Non-normal response *families* (binomial, Poisson, …)
- *Mixed* models incorporate *random effects*
  - Categories in data represent samples from a population
  - e.g. species, sites, genes …
  - Traditionally used to account for experimental blocks

Basic theory

Assumptions

- *Response variables* are linear functions of *input variables*, in turn based on *predictor variables*
    - Can have one or more input variables per predictor variable
    - Each input variable is associated with an estimated parameter (more about this later)
- *Errors* or *residuals* are Normally distributed
    - In other words, the difference between our model *predictions* and our observations is Normal
    - *not* assuming the marginal distribution is Normal (e.g. a histogram of an input variable)
    - (Go to board for residual drawing)
- Predictor variables are independent

Machinery

- *least squares* fit - we get parameters that minimize the squared differences between predictions and observations
- Least squares fits have a lot of nice properties
- Sensitive to some departures from the assumptions
- anomalous events tend to have a larger effect than they should

One-parameter variables

- Continuous predictor variable: estimate a straight line with one parameter
  - Also implies one *input variable*
- $Y = a + bX$: Y is the response, X is the input variable
  (b is the *slope* - expected change in Y per unit change in X)
- Categorical predictor variable with two categories: only one parameter
  - difference in predicted value between levels
- Parameters are (usually) easy to interpret
- Can think in terms of *confidence intervals* for the parameter
- see code

Multi-parameter variables

- With more than two categories, there is more than one input variable (parameter) associated with a single predictor variable
- A note: Non-linear response to a predictor variable
  - Might be able to use a linear model!
  - $Y = a + bX + cX^2$ is linear in a and b (the unknowns)
- see code

Interpreting multi-parameter variables is hard

- We can get a p value for the variable as a whole (e.g. is place important?)
- But we can only get CIs on the parameters (e.g. field / forest)
- there are also different ways to parameterize (*contrasts*)
- Think clearly about the *scientific* questions you have for this variable
- If you're just trying to control for it, just put it in and then ignore it!
- If you do have a clear scientific question, you should be able to construct *contrasts* in such a way that you can test it.
- You can do pairwise comparisons to test each pair of variables for differences and make an `aabbc` list

Interactions

- Interactions allow the value of one predictor to affect the relationship between another predictor and the response variable
- Interpreting *main effects* in the presence of interactions is tricky
- Your estimate of the effect of variable B is no longer constant
- You need to pick a fixed point, or average in some way
- Example: $Y = a + b_1X_1 + b_2X_2 + b_{12}X_1 * X_2$
- The response to X_1 is $Y = (a + b_2X_2) + (b_1 + b_{12}X_2)X_1$
  the response to X_1 *depends on* the value of X_2.

An experimental example

- You want to know whether a drug treatment changes the metabolism of some rabbits
- You're using adult, prime-aged rabbits and keeping them under controlled conditions, so you don't expect their metabolism to change *without* the drug.
    - What's wrong here?

An experimental example

- You want to know whether a drug treatment changes the metabolism of some rabbits
- You're using adult, prime-aged rabbits and keeping them under controlled conditions, so you don't expect their metabolism to change *without* the drug.
  - What's wrong here?
- You also introduce some control rabbits and treat them exactly the same, including giving them fake pills. You find no significant change in the metabolism of the control rabbits through time
  - There is still a problem…

Testing interactions

- Time is an important factor
- Use an *interaction*:

$$M = a + B_x X + B_t t + B_{xt} X t$$

- The interaction term B_{xt} represents the *difference in the response* between the two groups.
- It asks: **did the treatment group change differently than the control group**?
- Could also write:

$$M = a + B_1 trmt + B_2 time + B_3 trmt * time$$

- [Draw plot here]

Interactions and parameters

- We can use CIs, and coefficient plots, and get a pretty good idea what's going on
- In more complicated cases, interaction terms may have many parameters
- These have all the interpretation problems of other multi-parameter variables
- Think about "differences in differences"

Interactions: example

- Bear road-crossing
- Predictor variables: sex (categorical), road type (categorical: major/minor), road length (continuous)
- Two-way interactions
    - sex × road length: "are females more sensitive to amount of road than males?"
    - sex × road type: "do females vary behaviour between road type more than males?"
    - road type × road length: "does amount of road affect crossings differently for different road types?"

Statistical philosophy

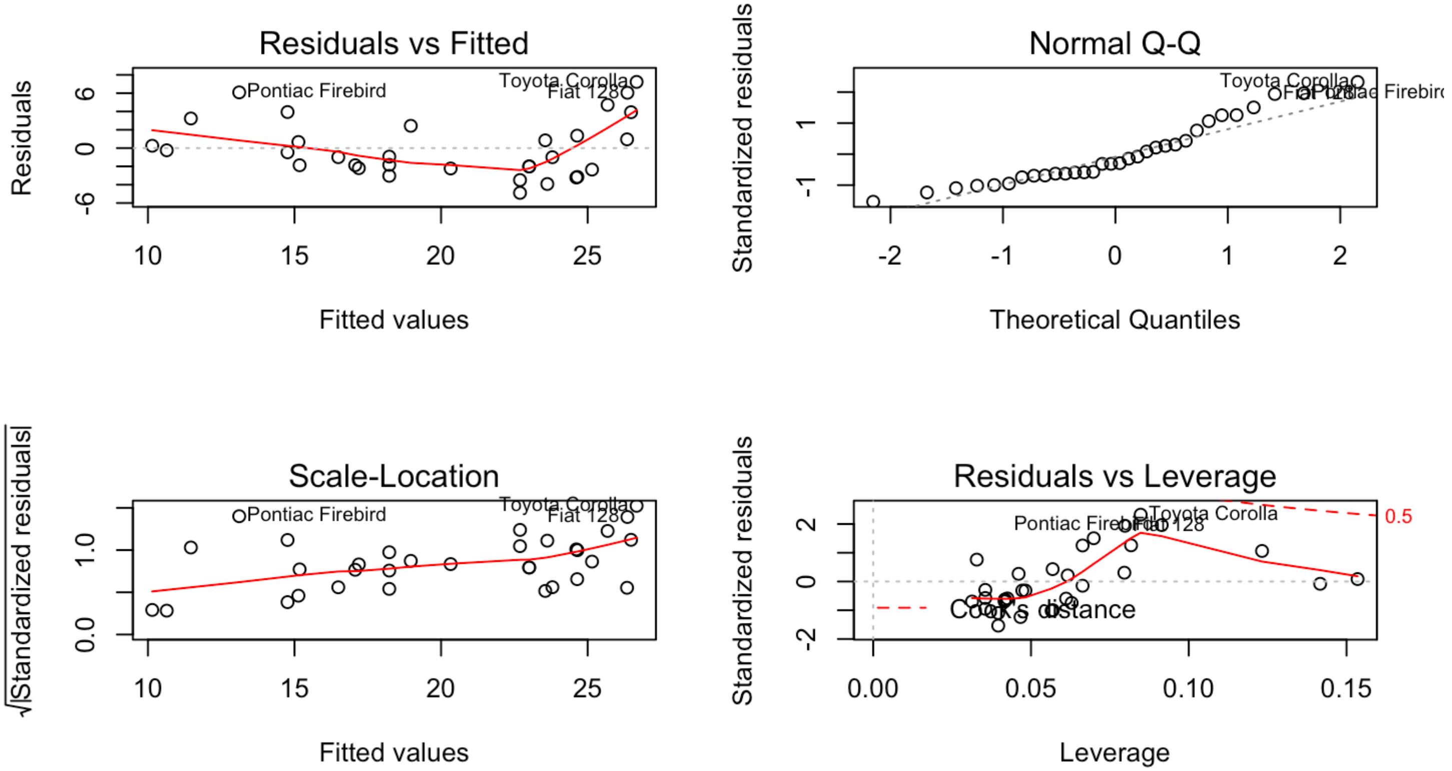
- Don't just accept the null hypothesis
  - Don't throw out predictors you wanted to test because they're not significant
  - Don't throw out interactions you wanted to test because they're not significant
  - Think carefully about your hypotheses and construct interactions if appropriate
  - There are techniques to deal with multiple predictors e.g. lasso regression, AIC, step-wise regression (although this violates above)
  - There are ways to estimate main effects in the presence of interactions (google "sum-to-zero contrasts" or "orthogonal interactions")

Diagnostics

- Because the linear model is sensitive (sometimes!) to assumptions, it is good to evaluate them
- Concerns:
  - *Heteroscedasticity* (does variance change across the data set, check diagnostic plots)
  - Linearity (does your model fit well?)
  - Normality (assuming no overall problems, do your **residuals** look Normal?)
    - What test would I use for this?
- Independence (no autocorrelation)
  - current value is independent of the previous (historic) values
  - we will discuss this more later in the semester - important for time series data
- Normality is the **least important** of these assumptions

Default plots in R

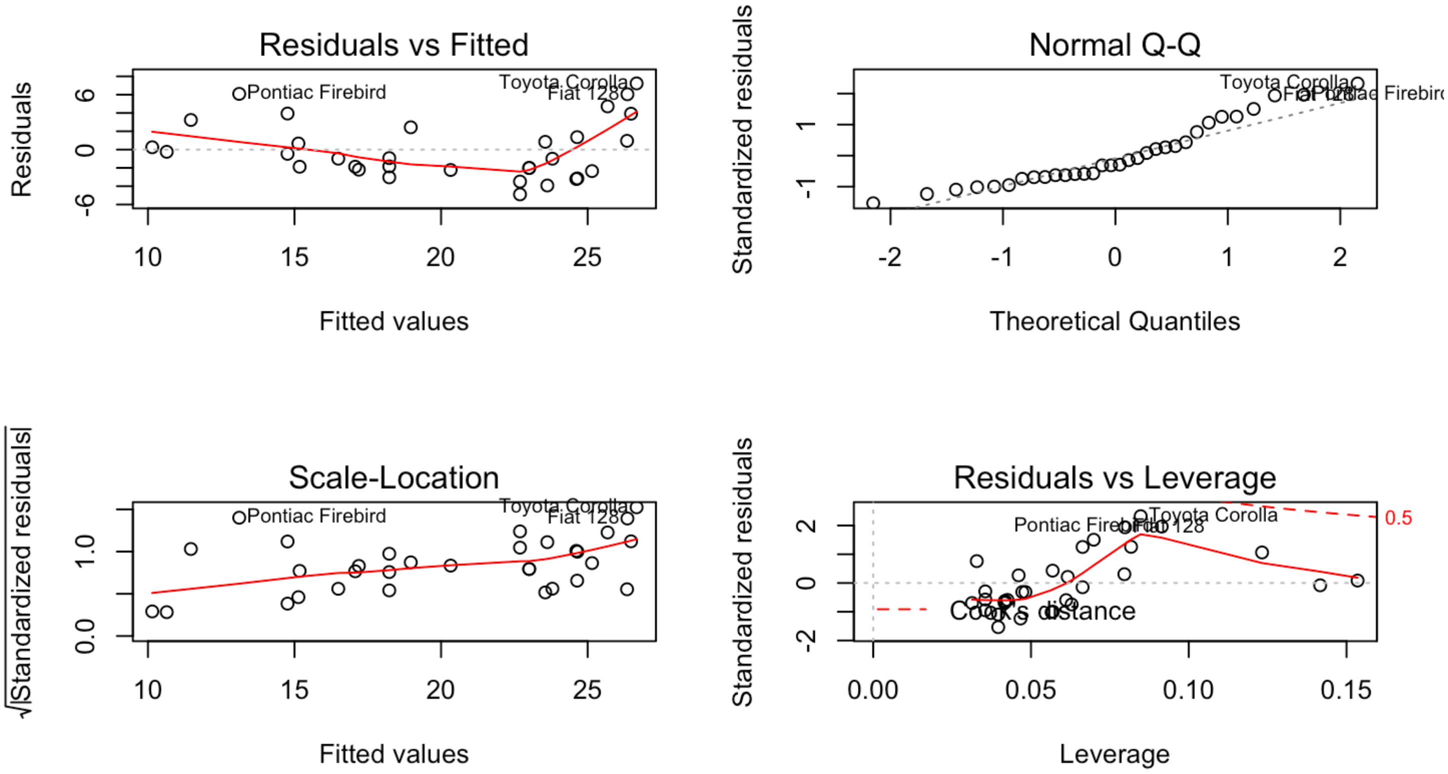
```
par(mfrow=c(2,2))  # set 2 rows and 2 column plot layout
mod_1 <- lm(mpg ~ disp, data=mtcars)  # linear model
plot(mod_1)
```



20/34

Default plots in R - Heteroscedasticity

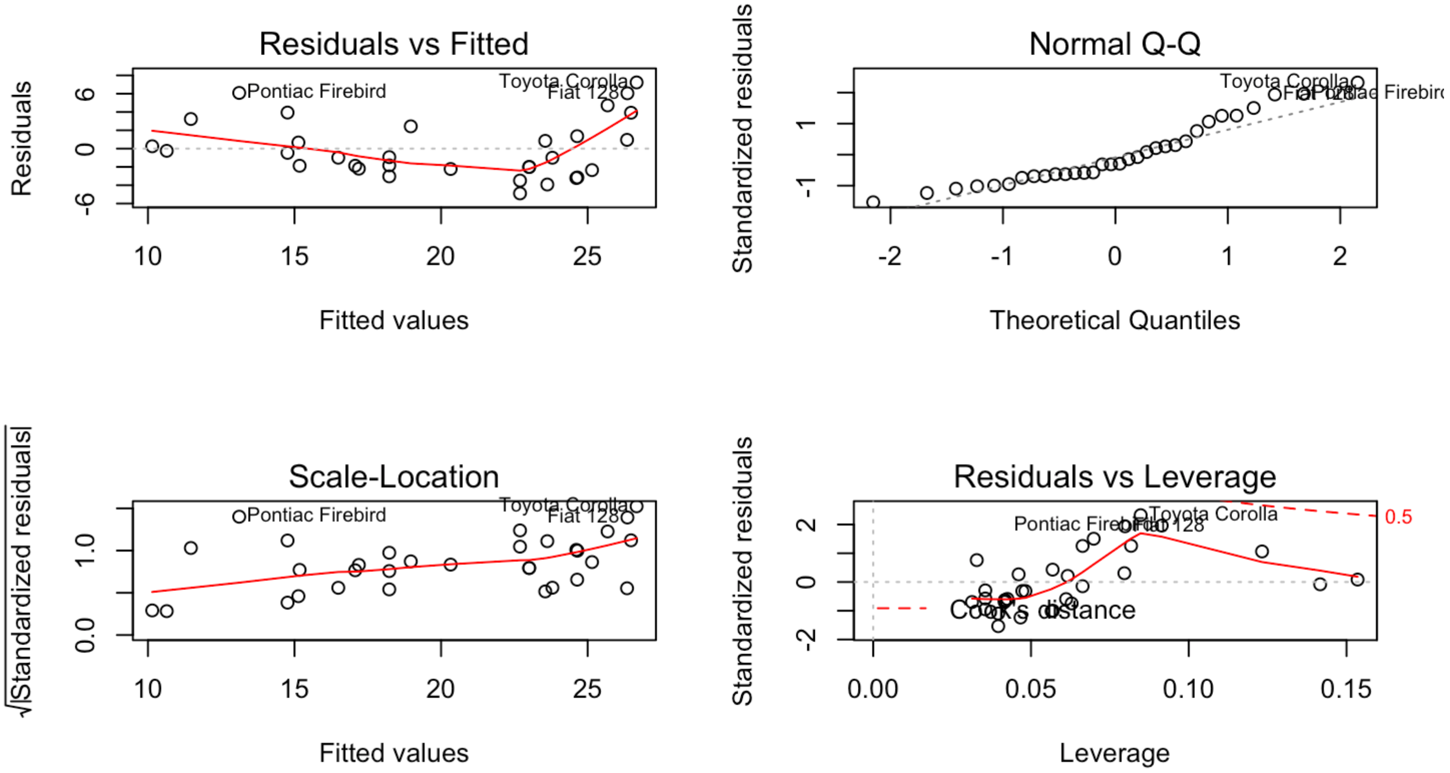
- The plots on the left look at variance across the range of fitted values
- They ask - as variable one increases, is it's explanatory ability similar across all values? These plots should be flat.



21/34

Default plots in R - Normality of Residuals

- The plot on the top right examines normality of residuals - this line should be 1:1. It may be easier to examine with a shapiro-wilk test

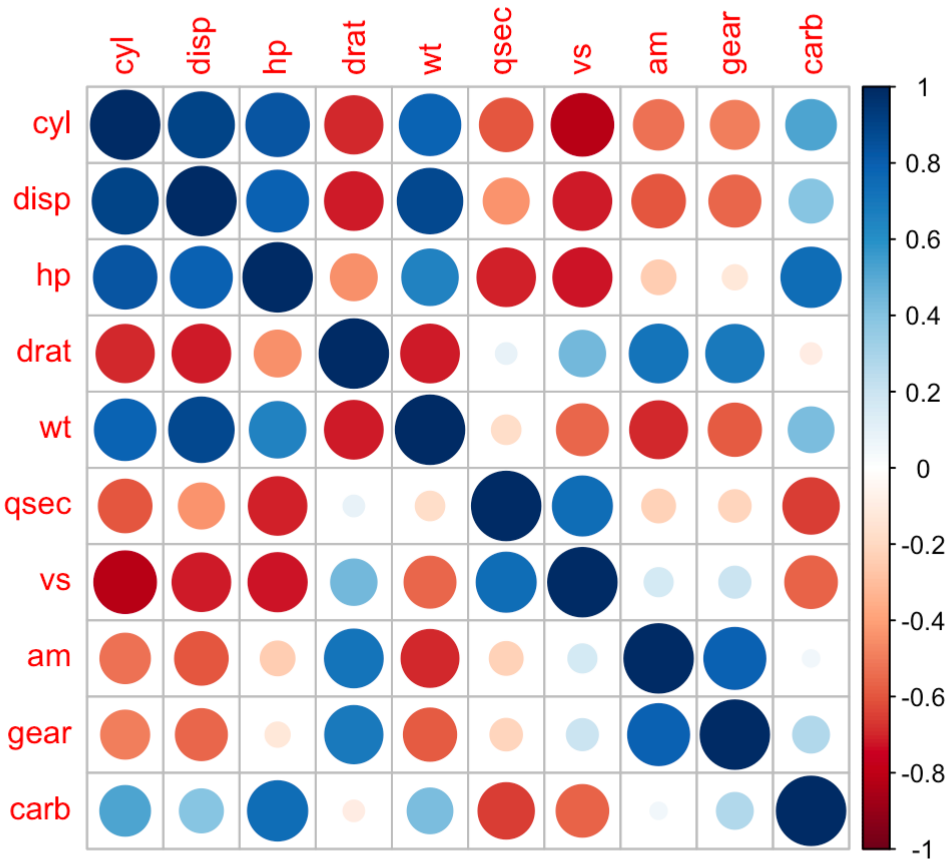


Diagnostic plots in code…

Some other assumptions

- the number of datapoints is greater than the number of predictors
- there is some variability in the values of your predictor (e.g. x is not the same)
- your predictors aren't perfectly correlated (e.g. multicollinearity)
- you can check this with corrplot

Correlation plots - which are correlated?



Resources

http://r-statistics.co/Assumptions-of-Linear-Regression.html

Transformations

- One way to deal with problems in model assumptions is by transforming one or more of your variables
- Transformations are not cheating: a transformed scale may be as natural (or more natural) a way to think about your data as your original scale
- The linear scale (no transformation) often has direct meaning, if you are adding things up or scaling them (as in our ant example)
- The log scale is often the best scale for thinking about physical quantities: 1:10 as 10:?

Transformation tradeoffs

- A transformation may help you meet model assumptions
  - Homoscedasticity
  - Linearity
  - Normality
- But there is no guarantee that you can fix them all
- Piles of zeros are hard too (consider GLMs)

Transformations to consider

- log-lin, lin-log and log-log for various sorts of exponential and power relationships
- Box-Cox and Yeo-Johnson (takes negative values)
  - Note: *Box-Cox transformation* tries out transformations of the form $(y^{\lambda} - 1)/\lambda$ ($\lambda = 0$ corresponds to log-transformation)
- Avoid classical 'transform then linear model' recommendations for
  - probability data or count data
  - Generally better to respect the structure of these data with a GLM

Deciding whether to transform

- It's **not OK** to pick transformations based on trying different ones and looking at P values
- It's probably OK to decide based on a measure of Normality of residuals

Tools for fitting and inference

Basic tools

- `lm` fits a linear model
- `summary` prints statistics associated with the *parameters* that were fitted
- [see code]

Multiple comparisons

- One standard of practice is to take a variable-level P value and then evaluate patterns in the response to significant variables
- Straightforward, but maybe not conservative
- `TukeyHSD` does multiple comparison tests on objects produced by `aov`. Can also use `glht` in the `multcomp` package more generally.
- Note: `aov` is just another way of calling `lm`, whereas `anova` *compares* different model fits.
- [see code]

Plotting

- `plot` can be applied to an `lm` object to give you a nice set of diagnostic tests.
- `predict` can give predicted values, and standard errors.
- In `ggplot`, `geom_smooth(method="lm")` fits a linear model to each group of data (i.e. each group that you have identified by plotting it in a different colour, within a different facet, etc.
    - [see code]