Introduction to Research Data Analysis

Myo Minn Oo

PNGIMR

2022-04-09

Outline

Section 1

Research Questions

A Good Research Question

FINER

- Feasible
- Interesting
- Novel
- Ethical
- Relevant

Our Study Outline - SARS-COV-2 Prevalence Study

Research Question: Among samples tested for COVID-19 infection at PNGIMR lab, what are the positivity rates of COVID-19 infection and predictors for being positive?

Study design: cross-sectional study

Subjects: samples that were tested RT-PCR for COVID-19 infection between September 2021 and March 2022

Outcome/Dependent variable: being positive on RT-PCR

Predictors/Independent variable: age, sex, residence, being symptomatic, previous contact, travel history, COVID-19 vaccination status, vaccination dose, number of symptoms

Potential confounders: reasons for testing

Type of Study Designs

- Cross-sectional
- Cohort
- Case-control
- Randomized control trials (RCT)

Section 2

Introduction to Statistics

What is statistics?

- Statistics - the practice and study of collecting and analyzing data
- Summary statistics - understanding or summary of some data

What can statistics do?

- How likely is a sample to be positive on a PCR test? Are samples more likely to be positive if people were tested at different clinic or purpose?
- What is the risk of dying if a person tested positive for COVID-19 infection? What interventions can we do to reduce the risk of death?
- A/B tests: which ads or adovacy method is more effective in getting people to test for COVID-19?

What **can't** statistics do?

- *Why is antivax working?*

Instead

- Are people who were tested for COVID-19 favorable toward COVID-19 vaccination?

But . . .

- Even so this won't tell us if more testing would lead to more vaccination.

Types of Statistics

Descriptive

- describe and summarize data

Examples:

- 50% of samples tested positive
- 25% of people tested for COVID-19 were symptomatic
- 5% died during hospitalization

Inferential

- Use a sample of data to make inferences about a larger population

Examples:

- What percent of test samples were tested positive?
- etc

Types of data

Numeric (quantitative)

- Continuous (measured)
  - age, weight, height
  - time from sample collection to date of test result reported
- Discrete (counted)
  - number of contact persons
  - number of symptoms

Categorical (Qualitative)

- Nominal (Unordered)
  - sex (male / female)
  - province where patients live, country of origin
- Ordinal (Ordered)
  - Dose of vaccination: 0, 1, 2+ doses
  - likert scale: strong diagree, disagree, neutral, agree, strongly agree

Categorical data represented as numbers

Nominal (unordered)

- sex: male / female -> 1 / 2
- country of origin (1, 2, 3, . . .)

Ordinal (ordered)

- Dose of vaccination: 0, 1, 2
- likert scale: 1, 2, 3, 4, 5

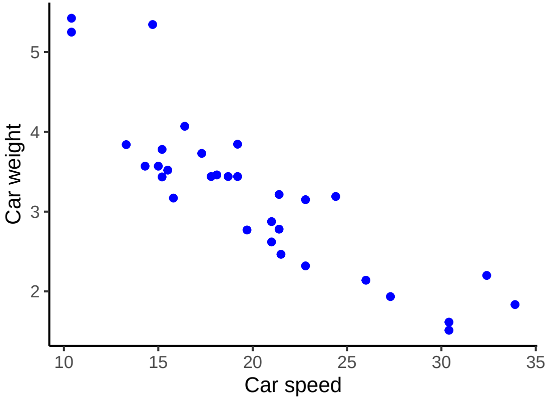
Why does data type matter?

Summary Statistics

```
mtcars %>%
    summarize(avg_speed = mean(mpg))
```

```
##   avg_speed
## 1  20.09062
```

Plot



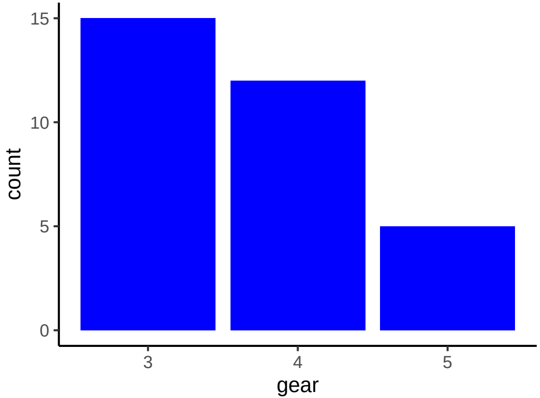
Why does data type matter?

Summary Statistics

```
mtcars %>%
    tabyl(gear) %>%
    adorn_pct_formatting()
```

```
##  gear  n percent
##     3 15   46.9%
##     4 12   37.5%
##     5  5   15.6%
```

Plot



Section 3

Introduction to R and RStudio

See R Handout!

Let's practice R!

Section 4

Descriptive Statistics: Number Summary

mtcars

```
mtcars
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C         17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE        16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
```

How is mpg distributed in this dataset?

- What is a typical value?
- Where is the center of the data?
    - mean
    - median
    - mode

```r
mtcars %>%
    select(mpg)
```

```
##                      mpg
## Mazda RX4           21.0
## Mazda RX4 Wag       21.0
## Datsun 710          22.8
## Hornet 4 Drive      21.4
## Hornet Sportabout   18.7
## Valiant             18.1
## Duster 360          14.3
## Merc 240D           24.4
## Merc 230            22.8
## Merc 280            19.2
```

Measure of center: mean

$$mean\_mpg = \frac{21.0 + 21.0 + 22.8 + \ldots}{32} = 20.09062$$

```
mtcars %>%
    summarize(avg_speed = mean(mpg))
```

```
##   avg_speed
## 1  20.09062
```

Measure of center: median

1. sort the numbers in ascending order
2. if odd number, take the middle one
3. if even number, add the middle twos and divide the summation by 2.

```
mtcars %>%
    arrange(mpg) %>%
    unlist(mpg, use.names = FALSE) %>%
    .[16:17]
```

```
## [1] 19.2 19.2
```

$$median\_mpg = \frac{19.2 + 19.2}{2} = 19.2$$

```
mtcars %>%
    summarize(median_mpg = median(mpg))
```

```
##   median_mpg
## 1       19.2
```

Measure of center: mode

```r
mtcars %>%
    count(mpg, sort = TRUE)
```

```
##     mpg n
## 1  10.4 2
## 2  15.2 2
## 3  19.2 2
## 4  21.0 2
## 5  21.4 2
## 6  22.8 2
## 7  30.4 2
## 8  13.3 1
## 9  14.3 1
## 10 14.7 1
## 11 15.0 1
```

Measure of Spread: standard deviation

$$sd\_mpg = \sqrt{\frac{\sum (x_i - \mu)^2}{n-1}}$$

```
mtcars %>%
    summarize(sd_mpg = sd(mpg))
```

```
##     sd_mpg
## 1 6.026948
```

Measure of Spread: Interquartile Range

- points that equally divide your data into four quartiles
- first quartile, Q1 = 25%
- second quartile, Q2 = 50% ~ median
- thrid quartile, Q3 = 75%
- interquartile Range (Q1, Q3)

```
mtcars %>%
    summarise(median = median(mpg),
              q1 = quantile(mpg, probs = 0.25),
              q3 = quantile(mpg, probs = 0.75))
```

```
##   median     q1   q3
## 1   19.2 15.425 22.8
```

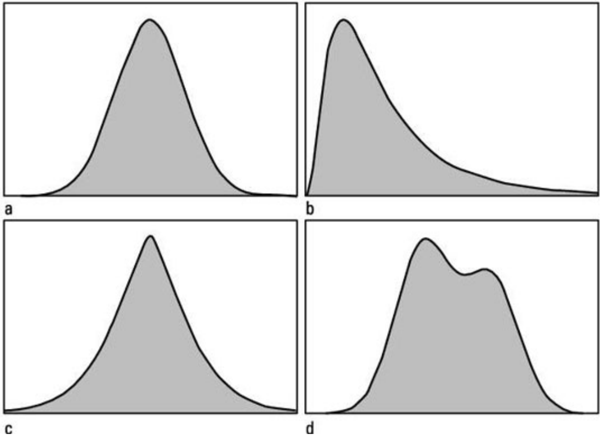
Other measures of spread: range

- minimum and maximum values
- useful for checking invalid values

```
mtcars %>%
    summarise(mean = mean(mpg),
              minimum = min(mpg),
              maximum = max(mpg))
```

```
##        mean minimum maximum
## 1 20.09062    10.4    33.9
```

Distributions

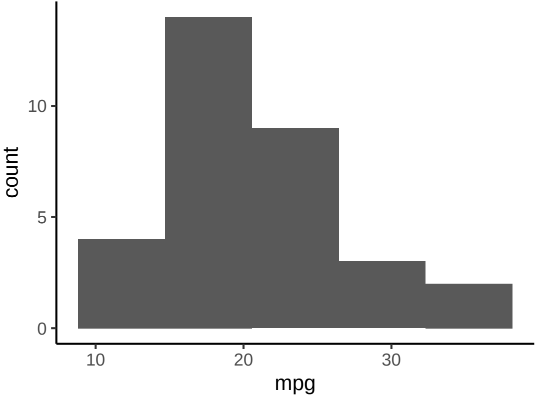


See R Handout!

Let's practice in R!

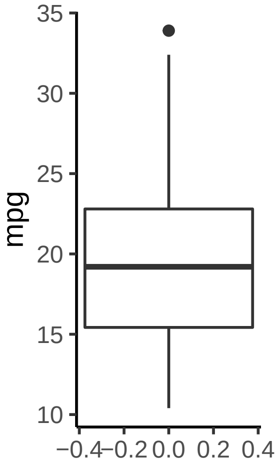
Visualization: histogram

```r
mtcars %>%
    ggplot(aes(mpg)) +
    geom_histogram(bins = 5) +
    theme_classic()
```



Visualization: boxplot

```
mtcars %>%
    ggplot(aes(mpg)) +
    geom_boxplot() +
    coord_flip() +
    theme_classic()
```



See R Handout!

Let's practice in R!

Section 5

Descriptive Statistics: categorical data

Displaying frequency and proportions

Gear ratio

```
mtcars %>%
    tabyl(gear) %>%
    adorn_totals("row") %>%
    adorn_pct_formatting()
```

```
##    gear  n percent
##       3 15   46.9%
##       4 12   37.5%
##       5  5   15.6%
##   Total 32  100.0%
```

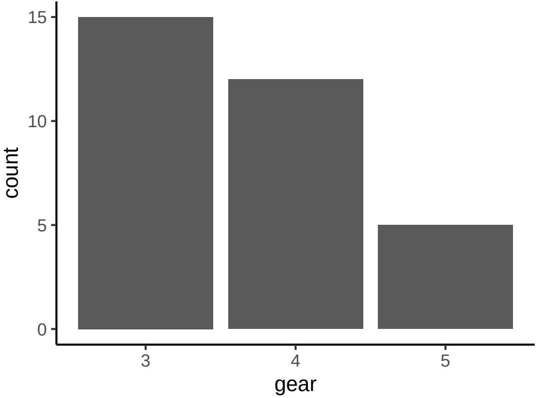
Automatic transmission

```
mtcars %>%
    tabyl(am) %>%
    adorn_totals("row") %>%
    adorn_pct_formatting()
```

```
##      am  n percent
##       0 19   59.4%
##       1 13   40.6%
##   Total 32  100.0%
```

Barplots

```
mtcars %>%
    ggplot(aes(gear)) +
    geom_bar() +
    theme_classic()
```



See R Handout!

Let's practice in R!

Section 6

Relationship between two variables

Relationship between two variables

- categorical $\sim$ categorical » cross-tabulation (contigency table)
- categorical $\sim$ numerical » grouped (stratified) summary measures
- numerical $\sim$ numerical » pearson's correlation (**r**)

categorical ~ categorical

```r
mtcars %>%
    tabyl(gear, am) %>%
    adorn_totals(c("row", "col")) %>%
    adorn_percentages("row") %>%
    adorn_pct_formatting(digits = 1, affix_sign = FALSE) %>%
    adorn_ns("front")
```

```
##    gear             0           1       Total
##       3 15 (100.0)  0     (0.0) 15 (100.0)
##       4  4  (33.3)  8    (66.7) 12 (100.0)
##       5  0   (0.0)  5   (100.0)  5 (100.0)
##   Total 19  (59.4) 13    (40.6) 32 (100.0)
```

categorical ∼ numerical

```
mtcars %>%
    group_by(gear) %>%
    summarise(mean = mean(mpg),
              sd = sd(mpg))
```

```
## # A tibble: 3 x 3
##    gear  mean    sd
##   <dbl> <dbl> <dbl>
## 1     3  16.1  3.37
## 2     4  24.5  5.28
## 3     5  21.4  6.66
```

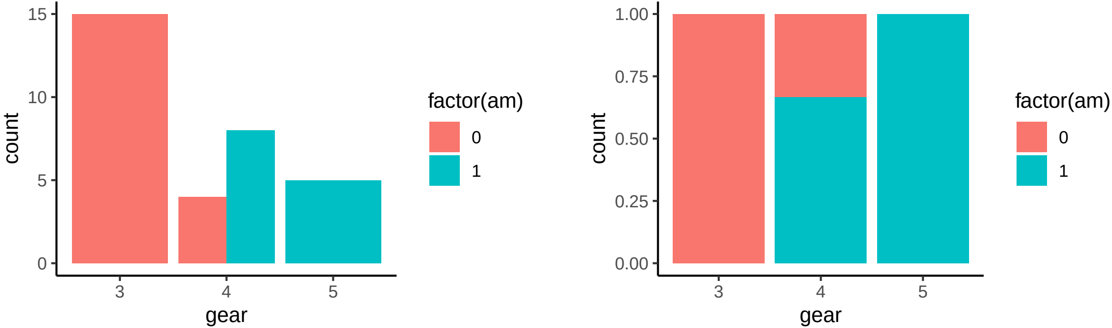
numerical ~ numerical

- correlation value ranges from −1 to +1
- value 0 = no correlation
- -1 = absolute negative association
- +1 = absoluate positive association
- around 0.4 = weak association
- around 0.8 = strong association
- **Assumption of linear association**

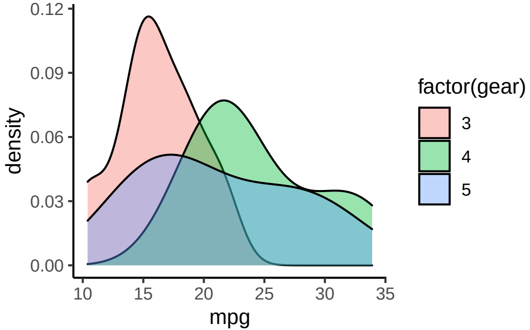
```
mtcars %>%
    summarise(correlation = cor(mpg, wt))
```

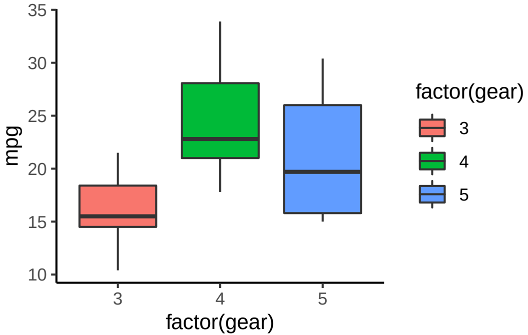
```
##   correlation
## 1  -0.8676594
```

Visualization: categorical ~ categorical

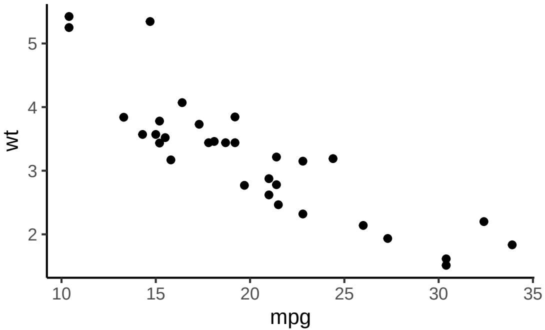


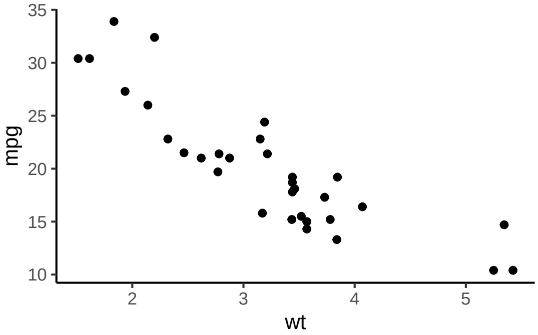
Visualization: categorical $\sim$ numerical





Visualization: numerical $\sim$ numerical





See R Handout!

Let's practice in R!

Section 7

Inferential Statistics

Inferential Statistics

- Confidence Intervals (CI)
- p-value

Confidence Intervals

- If 100 similar studies were carried out, the true unknown value of the population will lie in the range of confidence intervals for 95 times.
- mostly commonly used cut-off $= 95\%$
- If no true association exists, there is 5% chance that we will find a false association.
- more robust than a single estimate like means or proportion

p-value

- how likely our data would have occured by random change
- decision on whether to reject null hypothesis; does not mean it's true.
- arbitarily set value = 0.05 or 5% » .95 or 95% Confidence Interval
- largely depend on sample size: increasing sample size will likely lower p-value.
- **No p-value hacking!!**

statistical tests and data type

- categorical ~ categorical » cross-tabulation (contigency table)
  - chi-squared test of independence (each cell must be greater than 5!)
  - If not, use Fisher's exact test.
  - if you have ordered data, use different tests.
- categorical ~ numerical » grouped (stratified) summary measures
  - t-test (normal distribution, equal variance) to compare two means
  - if not, use Wilcoxon tests.
- numerical ~ numerical » pearson's correlation (**r**)
  - to compare more than two groups, use ANOVA (same assumptions as t-test)
  - if not, use Kruskal Wallis test.

Due to time constraint, we won't cover them in details. Interpretation of p-value is the same across all tests.

In addition, there are many other statistical tests that are out of scope for this workshop.

See R Handout!

Let's practice in R!

Section 8

Confounding versus interaction

Confounding variable

- a variable that is associated with both your outcome (dependent variable) and predictors (independent variables)
- Example:
    - coffee drinking is strongly associated with lung cancer.
    - seems like smokers are also heavey coffee drinker.
    - So smoking confounds the non-association between coffee drinking and lung cancer.

Interaction

- the effect of a variable on outcome depends on a second variable.
- Example:
    - type of food (icecream versus hotdogs) + condiments (chocolate sauce versus mustard)
    - Do you prefer ketchup or chocolate sauce on your food?
    - It actually depends on the type of food!

Confounding versus interaction

- if there are confounders, leave confoundees (variables affected by confounders) out of your analysis
- if there are interactions, account the interaction effect in your model.

How do you know which is which? - you don't know in most cases - check during variable selection aka model building - literature or previous knowledge - biological pathways or plausibility - pathway analysis like DAG or SEM??

Section 9

Regression

Regression

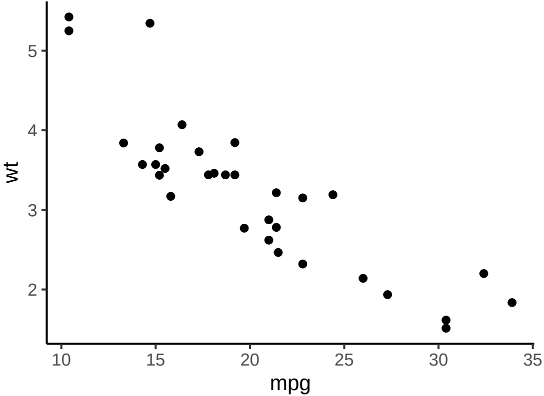
- statistical model
  - to explore relationship between an outcome and predictors
  - to predict outcome based on the values of predictors
- Jargons
  - outcome a.k.a dependent variable or `y`
  - predictors a.k.a independent variables or `x`

Linear versus logistic

- linear regression: outcome is continus
- logistic regression: outcome is logical

Visualizing linear relationship

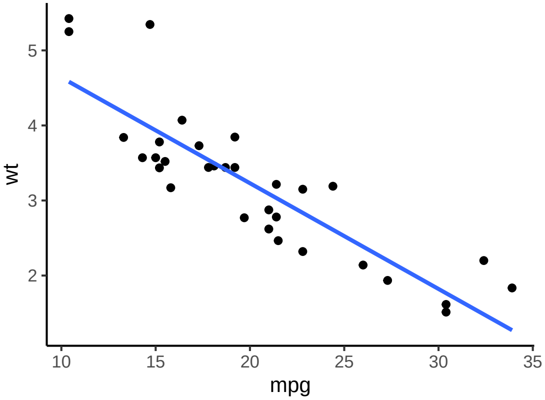
```r
mtcars %>%
    ggplot(aes(mpg, wt)) +
    geom_point() +
    theme_classic()
```



Adding a linear trend

```
mtcars %>%
    ggplot(aes(mpg, wt)) +
    geom_point() +
    geom_smooth(method = "lm",
                se = FALSE) +
    theme_classic()
```

```
## 'geom_smooth()' using formula 'y ~ x
```



Straight lines defined by two things

Intercept The `y` value at the point when `x` is zero

Slope The amount `y` value increase if `x` value increases by one unit

Equation

$$y = intercept + slope * x$$

Running a linear model

```r
lm(mpg ~ wt, data = mtcars)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Coefficients:
## (Intercept)           wt
##      37.285       -5.344
```

Interpreting the model

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Coefficients:
## (Intercept)           wt
##      37.285       -5.344
```

$$mpg = 37.285 + (-5.344) * wt$$

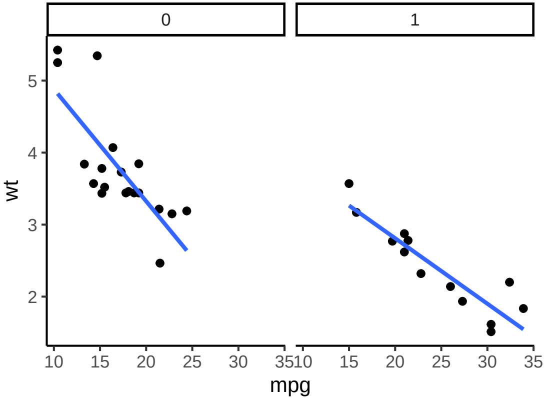
In every 1000 lbs increase, 5.344 mile per gallon will reduce.

Adding a categorical variable

Let's add am indicating automatic transmission by 0 and manual by 1

```
mtcars %>%
    ggplot(aes(mpg, wt)) +
    geom_point() +
    geom_smooth(method = "lm",
                se = FALSE) +
    facet_grid(cols = vars(am)) +
    theme_classic()
```

```
## `geom_smooth()` using formula 'y ~ x
```



Adding a categorical variable to the linear model

We need to convert am to a factor type to indicate as categorical data. Converting it as character also works.

```
lm(mpg ~ wt + factor(am), data = mtcars)
```

```
##
## Call:
## lm(formula = mpg ~ wt + factor(am), data = mtcars)
##
## Coefficients:
## (Intercept)           wt  factor(am)1
##    37.32155     -5.35281     -0.02362
```

Interpreting a linear model with two predictors

```
##
## Call:
## lm(formula = mpg ~ wt + factor(am), data = mtcars)
##
## Coefficients:
## (Intercept)           wt  factor(am)1
##    37.32155     -5.35281     -0.02362
```

- while keeping the same weight, cars with manual transmission will reduce additional 0.02362 miles per gallon, compared to cars with automatic transmission.

Assessing model fit

- coefficient of determination - R-squared » just squaring Pearson's correlation `r`
    - always increase when number of predictors increases - not good for multivariable model
- adjusted R-squared
    - penalized for number of predictors
- RSE = residual standard error
    - typical difference between prediction and observed values
- AIC
- BIC

Checking model fit

```
mpg_model <- lm(mpg ~ wt + factor(am), data = mtcars)
summary(mpg_model)
```

```
##
## Call:
## lm(formula = mpg ~ wt + factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5295 -2.3619 -0.1317  1.4025  6.8782
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.32155    3.05464  12.218 5.84e-13 ***
## wt          -5.35281    0.78824  -6.791 1.87e-07 ***
```

Checking model fit using glance()

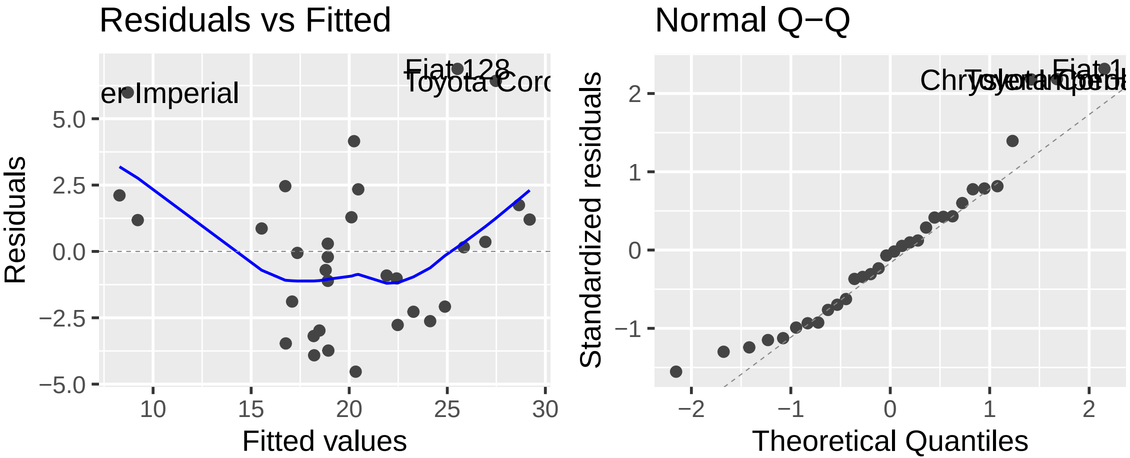
```
library(broom)
mpg_model %>%
    glance()
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic       p.value    df logLik   AIC
##       <dbl>         <dbl> <dbl>     <dbl>         <dbl> <dbl>  <dbl> <dbl>
## 1     0.753         0.736  3.10      44.2 0.00000000158     2  -80.0  168.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

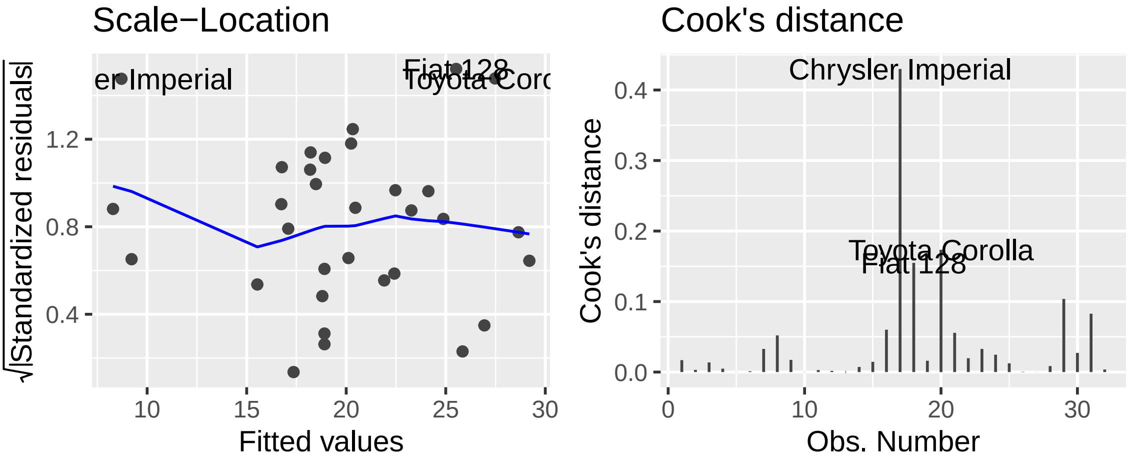
We look at sigma for RSE. The value of RSE is 3.10.

typical difference between predicted mpg and observed mpg is 3.10 miles per gallon.

Visualizing model fit



Visualizing model fit 2



What is next!

- model diagnostics in details
- quadratic or cubed model
- modelling on transformed data
- prediction
- outliers, leverage or influential points
- model building or variable selection

See R Handout!

Let's practice in R!

Section 10

Logistic Regression

Logistic Regression

- another type of linear regression
- The outcome variable is logical » meaning 1 and 0.

We can still run a linear model on the binary outcome.

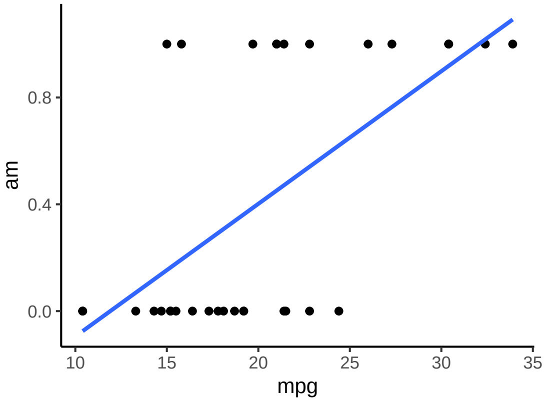
```
lm(am ~ mpg, data = mtcars)
```

```
##
## Call:
## lm(formula = am ~ mpg, data = mtcars)
##
## Coefficients:
## (Intercept)          mpg
##    -0.59149      0.04966
```

Visualizing linear model on binary outcome

Let's add am indicating automatic transmission by 0 and manual by 1

```
mtcars %>%
    ggplot(aes(mpg, am)) +
    geom_point() +
    geom_smooth(method = "lm",
                se = FALSE) +
    theme_classic()
```



Odds Ratio

$$OddsRatio = \frac{Probability\_of\_something\_happening}{probability\_of\_something\_not\_happening}$$

$$OddsRatio = \frac{probability}{(1 - probability)}$$

$$OddsRatio = \frac{0.25}{(1 - 0.25)} = \frac{1}{3}$$

Running a logistic regression

```
logm1 <- glm(am ~ mpg, data = mtcars, family = binomial)
summary(logm1)
```

```
##
## Call:
## glm(formula = am ~ mpg, family = binomial, data = mtcars)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5701  -0.7531  -0.4245   0.5866   2.0617
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.6035     2.3514  -2.808  0.00498 **
## mpg           0.3070     0.1148   2.673  0.00751 **
```

Displaying coefficient estimates

```
logm1 %>%
    tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   -6.60      2.35      -2.81 0.00498
## 2 mpg            0.307     0.115      2.67 0.00751
```

With 1 mpg increase, there is 0.307 log odds chance of being manual.

it is hard to understand log odds scale without visualization.

Calculating odds ratios

```
cbind(
    exp(coef(logm1)),
    exp(confint(logm1))
)
```

```
## Waiting for profiling to be done...

##                                 2.5 %     97.5 %
## (Intercept) 0.001355579 4.425443e-06 0.06255158
## mpg         1.359379288 1.129764e+00 1.79946863
```

See R Handout!

Let's practice in R!