Review of multiple linear regression

Prof. Maria Tackett

01.10.22

Click for PDF of slides

Announcements

- Labs start Thursday
    - Install R and configure GitHub
- Office hours this week:
    - Thu 2 - 3pm & Fri 1 - 2pm online (links in Sakai)
    - Full office hours schedule starts Tue, Jan 19
- Fill out All About You Survey

Questions from last time?

Linear least squares regression (LLSR) vs.

Generalized linear models (GLM) vs.

Multilevel models

Assumptions for linear regression

Linearity: Linear relationship between mean response and predictor variable(s)

Independence: Residuals are independent. There is no connection between how far any two points lie above or below regression line.

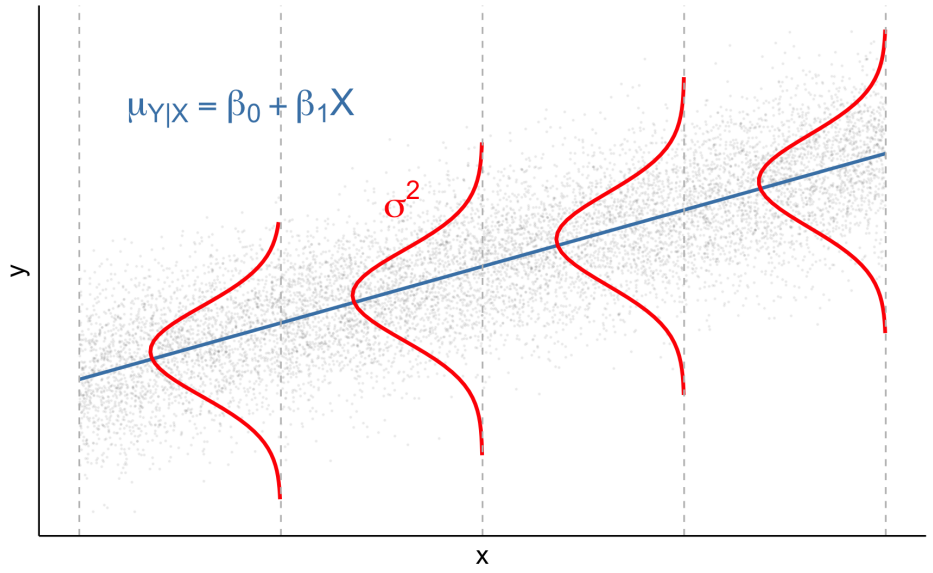
Normality: Response follows a normal distribution at each level of the predictor (or combination of predictors)

Equal variance: Variability (variance or standard deviation) of the response is equal for all levels of the predictor (or combination of predictors)

Use residual plots to check that the conditions hold before using the model for statistical inference.

Assumptions for linear regression

]



Modified from Figure 1.1. in BMLR

Linearity: Linear relationship between mean of the response **(Y)** and the predictor **(X)**

Independence: No connection between how any two points lie above or below the regression line

Normality: Response, **(Y)**, follows a normal distribution at each level of the predictor, **(X)** (indicated by red curves)

Equal variance: Variance (or standard deviation) of the response, **(Y)**, is equal for all levels of the predictor, **(X)**

Are the assumptions violated?

Click here for poll.

04:00

Beyond linear regression

- When we use linear least squares regression to draw conclusions, we do so under the assumption that L.I.N.E. are all met.
- **Generalized linear models** require different assumptions and can accommodate violations in L.I.N.E.
    - Relationship between response and predictor(s) can be nonlinear
    - Response variable can be non-normal
    - Variance in response can differ at each level of predictor(s)

But the independence assumption must hold!

- **Multilevel models** will be used for data with correlated observations

Review of multiple linear regression

Data: Kentucky Derby Winners

Today's data is from the Kentucky Derby, an annual 1.25-mile horse race held at the Churchill Downs race track in Louisville, KY. The data is in the file derbyplus.csv and contains information for races 1896 - 2017.

Response variable

- **`speed`**: Average speed of the winner in feet per second (ft/s)

Additional variable

- **`winner`**: Winning horse

Predictor variables

- **`year`**: Year of the race
- **`condition`**: Condition of the track (good, fast, slow)
- **`starters`**: Number of horses who raced

Data

```
derby <- read_csv("data/derbyplus.csv")
```

```
derby %>%
  head(5) %>% kable()
```

year	winner	condition	speed	starters
1896	Ben Brush	good	51.66	8
1897	Typhoon II	slow	49.81	6
1898	Plaudit	good	51.16	4
1899	Manuel	fast	50.00	5
1900	Lieut. Gibson	fast	52.28	7

Data analysis life cycle

Problem

✓Understanding and defining the problem

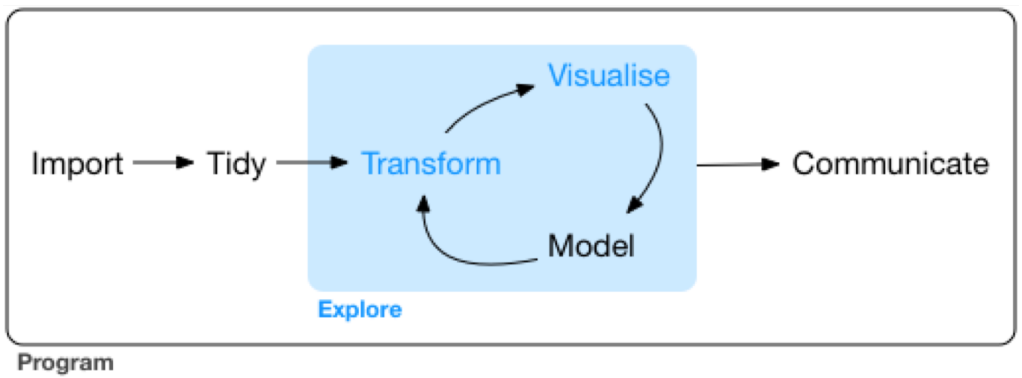
✓How do we go about answering the question?

Plan

✓What to measure and how?

✓How to collect the data: Study design? Recording? Collecting?

Collect data



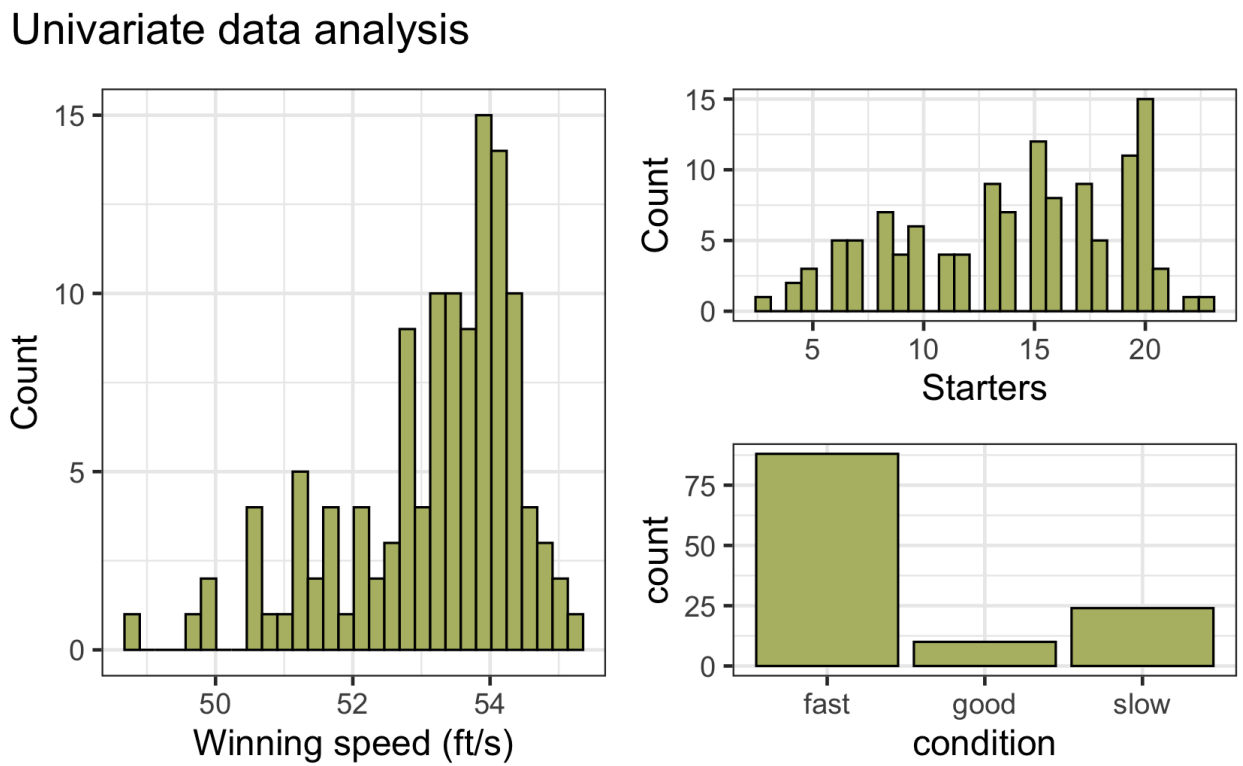
Exploratory data analysis (EDA)

- Once you're ready for the statistical analysis (explore), the first step should always be **exploratory data analysis**.
- The EDA will help you
    - begin to understand the variables and observations
    - identify outliers or potential data entry errors
    - begin to see relationships between variables
    - identify the appropriate model and identify a strategy
- The EDA is exploratory; formal modeling and statistical inference should be used to draw conclusions.

Plots for univariate EDA

Plot

Code



Plots for univariate EDA

Plot

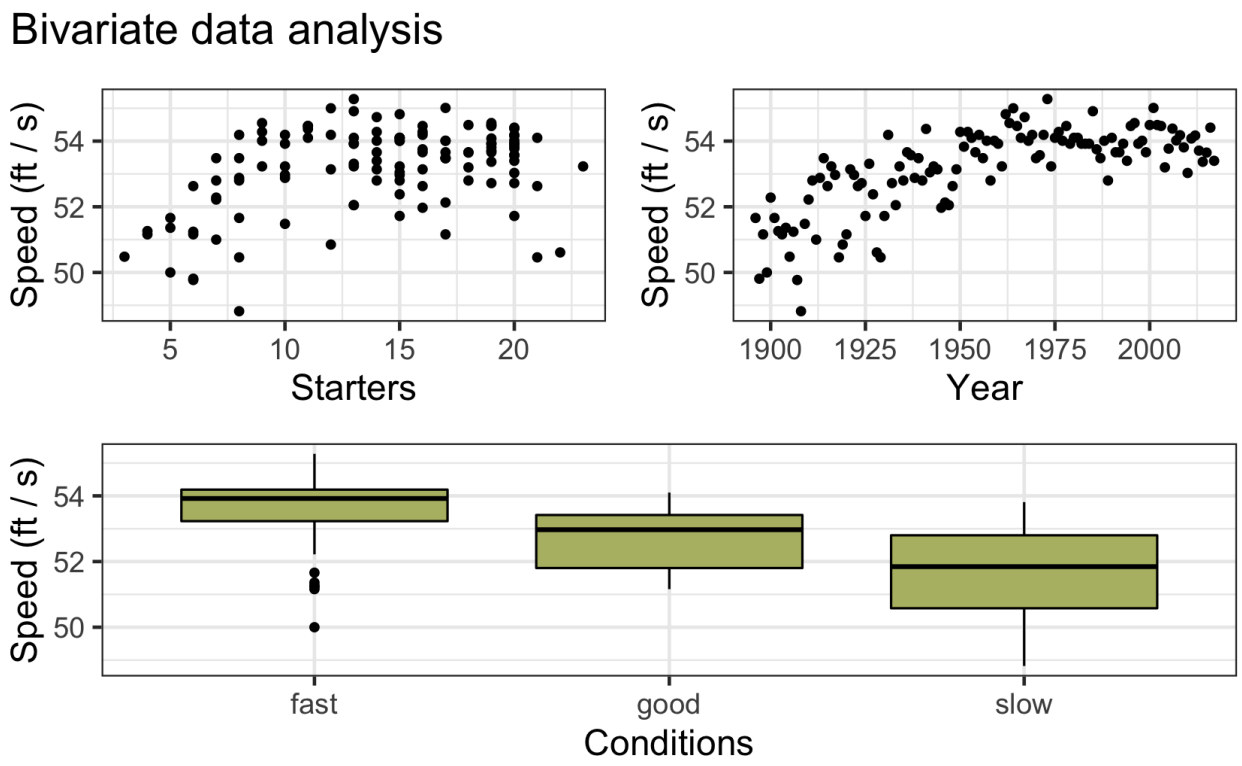
Code

```
p1 <- ggplot(data = derby, aes(x = speed)) +
  geom_histogram(fill = colors$green, color = "black") +
  labs(x = "Winning speed (ft/s)", y = "Count")

p2 <- ggplot(data = derby, aes(x = starters)) +
  geom_histogram(fill = colors$green, color = "black") +
  labs(x = "Starters", y = "Count")

p3 <- ggplot(data = derby, aes(x = condition)) +
   geom_bar(fill = colors$green, color = "black", aes(x = ))

p1 + (p2 / p3) +
  plot_annotation(title = "Univariate data analysis")
```

Plots for bivariate EDA

Plot

Code



Plots for bivariate EDA

Plot

Code

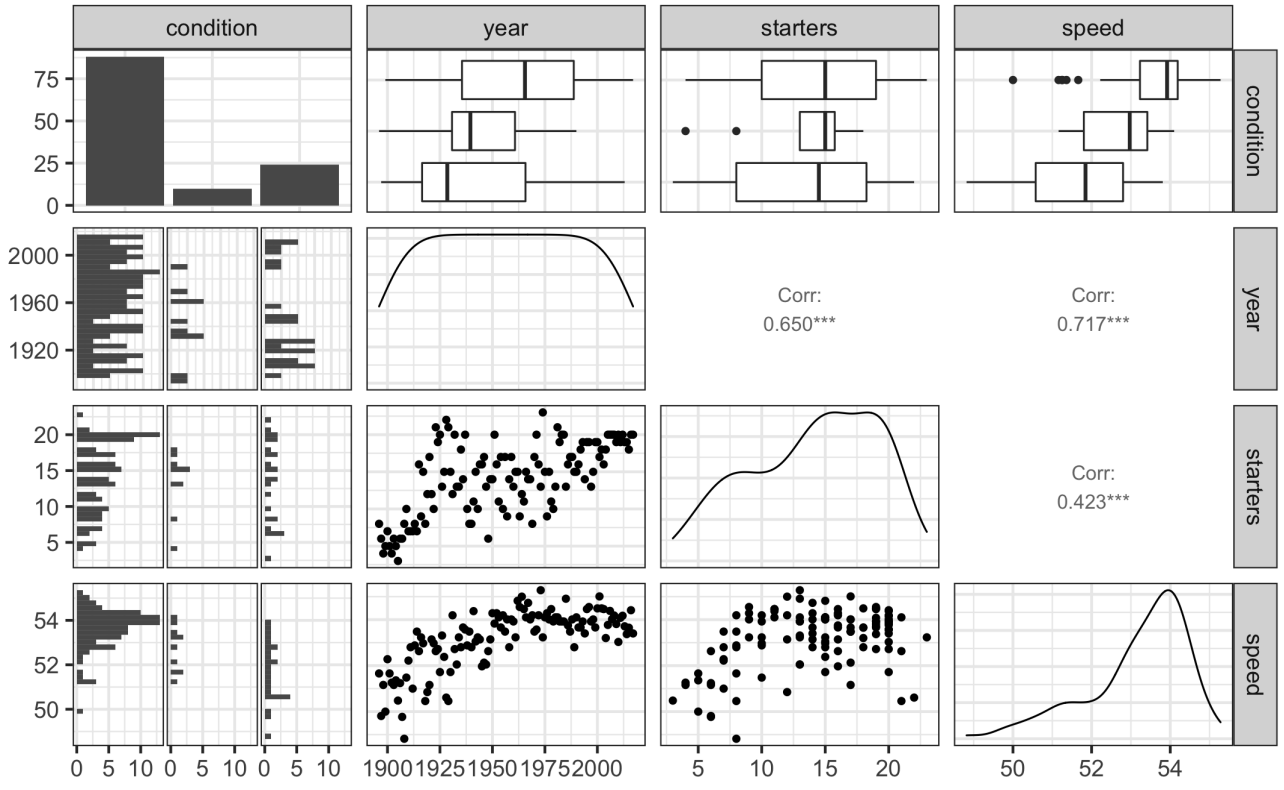
```
p4 <- ggplot(data = derby, aes(x = starters, y = speed)) +
  geom_point() +
  labs(x = "Starters", y = "Speed (ft / s)")

p5 <- ggplot(data = derby, aes(x = year, y = speed)) +
  geom_point() +
  labs(x = "Year", y = "Speed (ft / s)")

p6 <- ggplot(data = derby, aes(x = condition, y = speed)) +
  geom_boxplot(fill = colors$green, color = "black") +
  labs(x = "Conditions", y = "Speed (ft / s)")

(p4 + p5) + p6 +
  plot_annotation(title = "Bivariate data analysis")
```

Scatterplot matrix

A **scatterplot matrix** helps quickly visualize relationships between many variable pairs. They are particularly useful to identify potentially correlated predictors.

Plot

Code



Scatterplot matrix

A **scatterplot matrix** helps quickly visualize relationships between many variable pairs. They are particularly useful to identify potentially correlated predictors.

Plot

Code

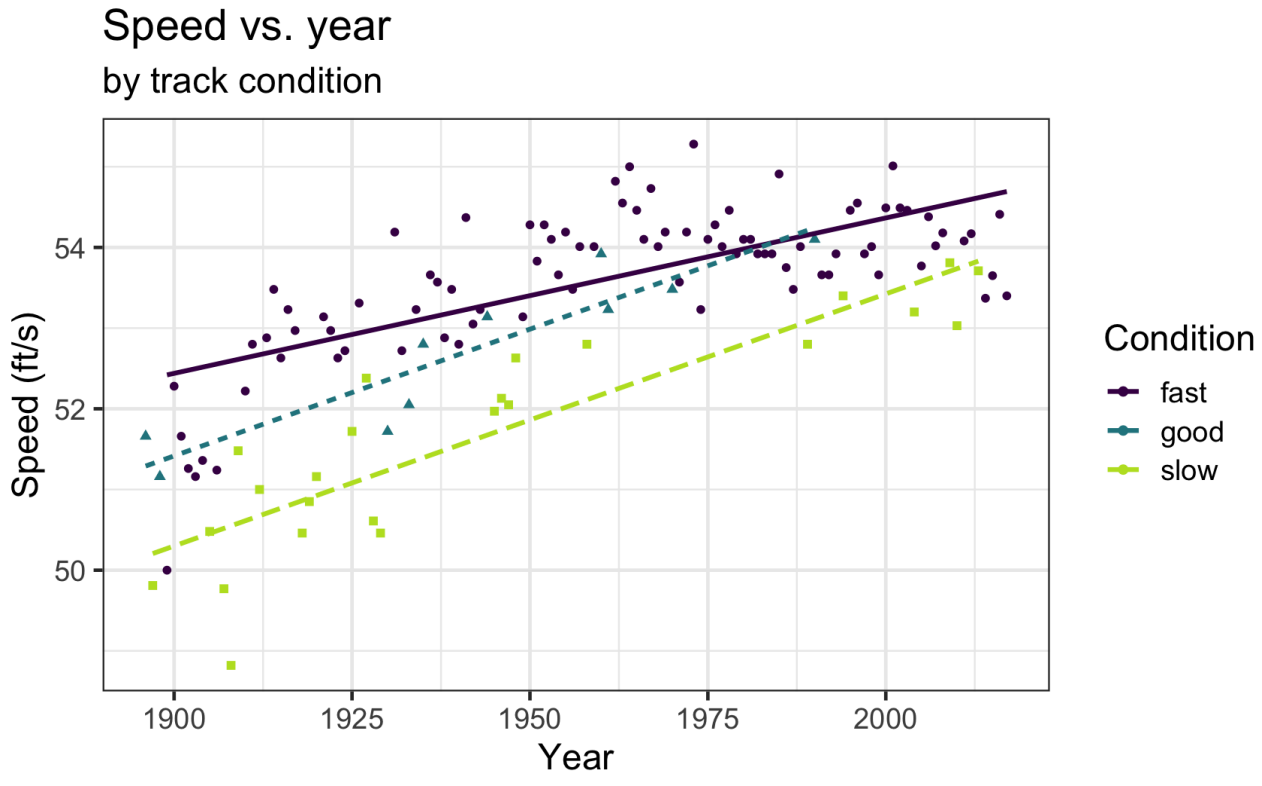
```
#library(GGally)
ggpairs(data = derby,
        columns = c("condition", "year", "starters", "speed"))
```

Plots for multivariate EDA

Plot the relationship between the response and a predictor based on levels of another predictor to assess potential interactions.

Plot

Code



Plots for multivariate EDA

Plot the relationship between the response and a predictor based on levels of another predictor to assess potential interactions.

Plot

Code

```
#library(viridis)
ggplot(data = derby, aes(x = year, y = speed, color = condition,
                         shape = condition, linetype = condition)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, aes(linetype = condition)) +
  labs(x = "Year", y = "Speed (ft/s)", color = "Condition",
       title = "Speed vs. year",
       subtitle = "by track condition") +
  guides(lty = FALSE, shape = FALSE) +
  scale_color_viridis_d(end = 0.9)
```

Model 1: Main effects model

Output

Code

term	estimate	std.error	statistic	p.value
(Intercept)	8.197	4.508	1.818	0.072
starters	-0.005	0.017	-0.299	0.766
year	0.023	0.002	9.766	0.000
conditiongood	-0.443	0.231	-1.921	0.057
conditionslow	-1.543	0.161	-9.616	0.000

Model 1: Main effects model

Output

Code

```
# Fit and display model
model1 <- lm(speed ~ starters + year + condition, data = derby)
tidy(model1) %>%
  kable(digits = 3)
```

Interpretation

$$\widehat{speed} = 8.197 - 0.005\ starters + 0.023\ year - 0.443\ good - 1.543\ slow$$

term	estimate	std.error	statistic	p.value
(Intercept)	8.197	4.508	1.818	0.072
starters	-0.005	0.017	-0.299	0.766
year	0.023	0.002	9.766	0.000
conditiongood	-0.443	0.231	-1.921	0.057
conditionslow	-1.543	0.161	-9.616	0.000

1. Write out the interpretations for **`starters`** and **`conditiongood`**.
2. Does the intercept have a meaningful interpretation?

Centering

Centering: Subtract a constant from each observation of a given variable

- Do this to make interpretation of model parameters more meaningful (particularly intercept)
- In STA 210, we used **mean-centering** where we subtracted the mean from each observation of given variable
- How does centering change the model?

Centering year

```
derby <- derby %>%
  mutate(yearnew = year - 1896) #1896 = starting year
```

term	estimate	std.error	statistic	p.value
(Intercept)	52.175	0.194	269.079	0.000
starters	-0.005	0.017	-0.299	0.766
yearnew	0.023	0.002	9.766	0.000
conditiongood	-0.443	0.231	-1.921	0.057
conditionslow	-1.543	0.161	-9.616	0.000

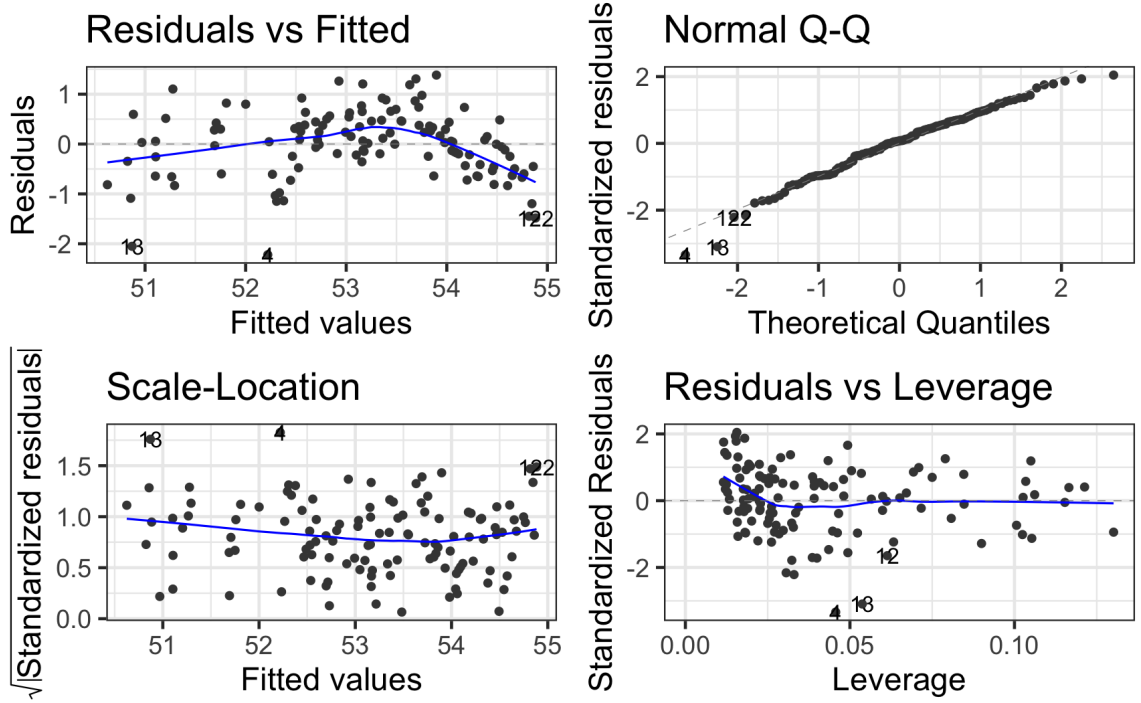
$$\widehat{speed} = 52.175 - 0.005\ starters + 0.023\ yearnew - 0.443\ good - 1.543\ slow$$

Model 1: Check model assumptions

Plots

Poll

```
#library(ggfortify)
autoplot(model1Cent)
```



Model 1: Check model assumptions

Plots

Poll

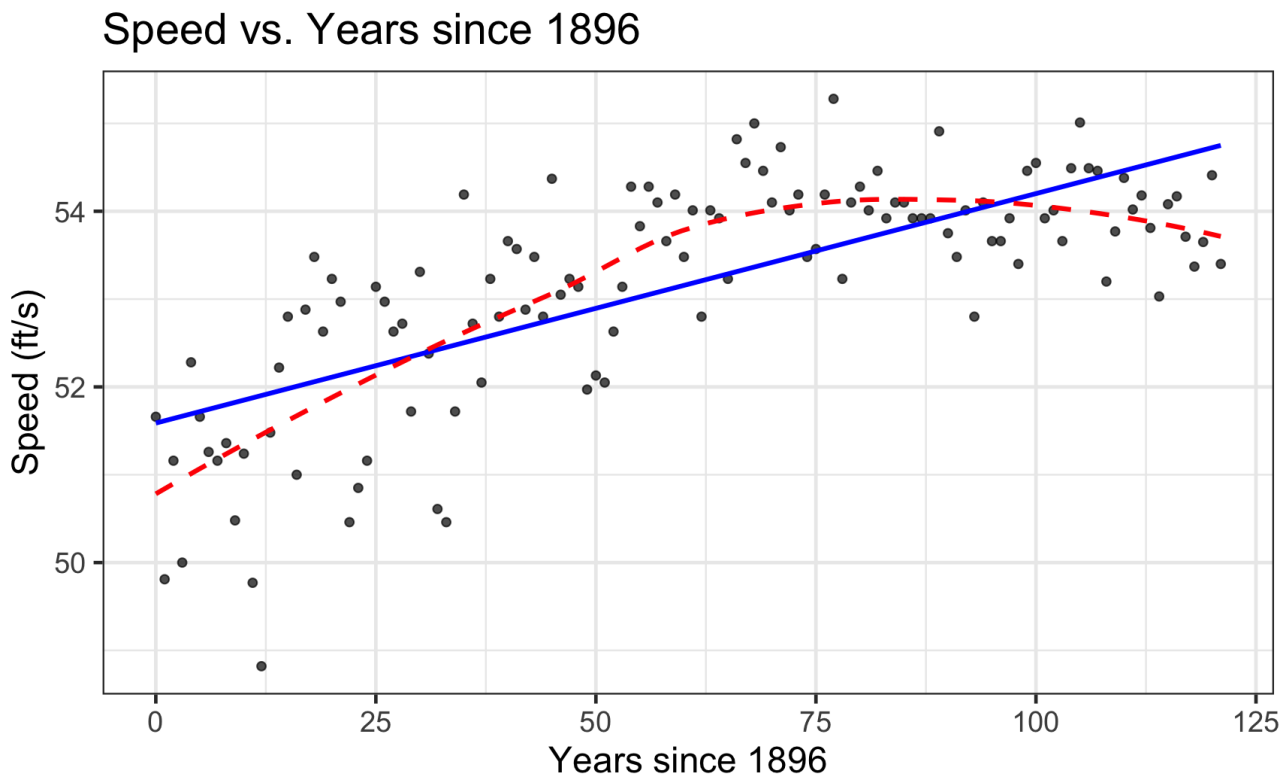
Click here for poll.

Model 2

Add quadratic effect for year?

Plot

Code



Add quadratic effect for year?

Plot

Code

```
ggplot(data = derby, aes(x = yearnew, y = speed)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  geom_smooth(se = FALSE, color = "red", linetype = 2) +
  labs(x = "Years since 1896", y = "Speed (ft/s)",
       title = "Speed vs. Years since 1896")
```

Model 2: Add $yearnew^2$

Output

Code

term	estimate	std.error	statistic	p.value
(Intercept)	51.4130	0.1826	281.5645	0.0000
starters	-0.0253	0.0136	-1.8588	0.0656
yearnew	0.0700	0.0061	11.4239	0.0000
I(yearnew^2)	-0.0004	0.0000	-8.0411	0.0000
conditiongood	-0.4770	0.1857	-2.5689	0.0115
conditionslow	-1.3927	0.1305	-10.6701	0.0000

Model 2: Add $yearnew^2$

Output

Code

```
model2 <- lm(speed ~ starters + yearnew + I(yearnew^2) + condition,
             data = derby)
tidy(model2) %>% kable(digits = 4)
```

Interpreting quadratic effects

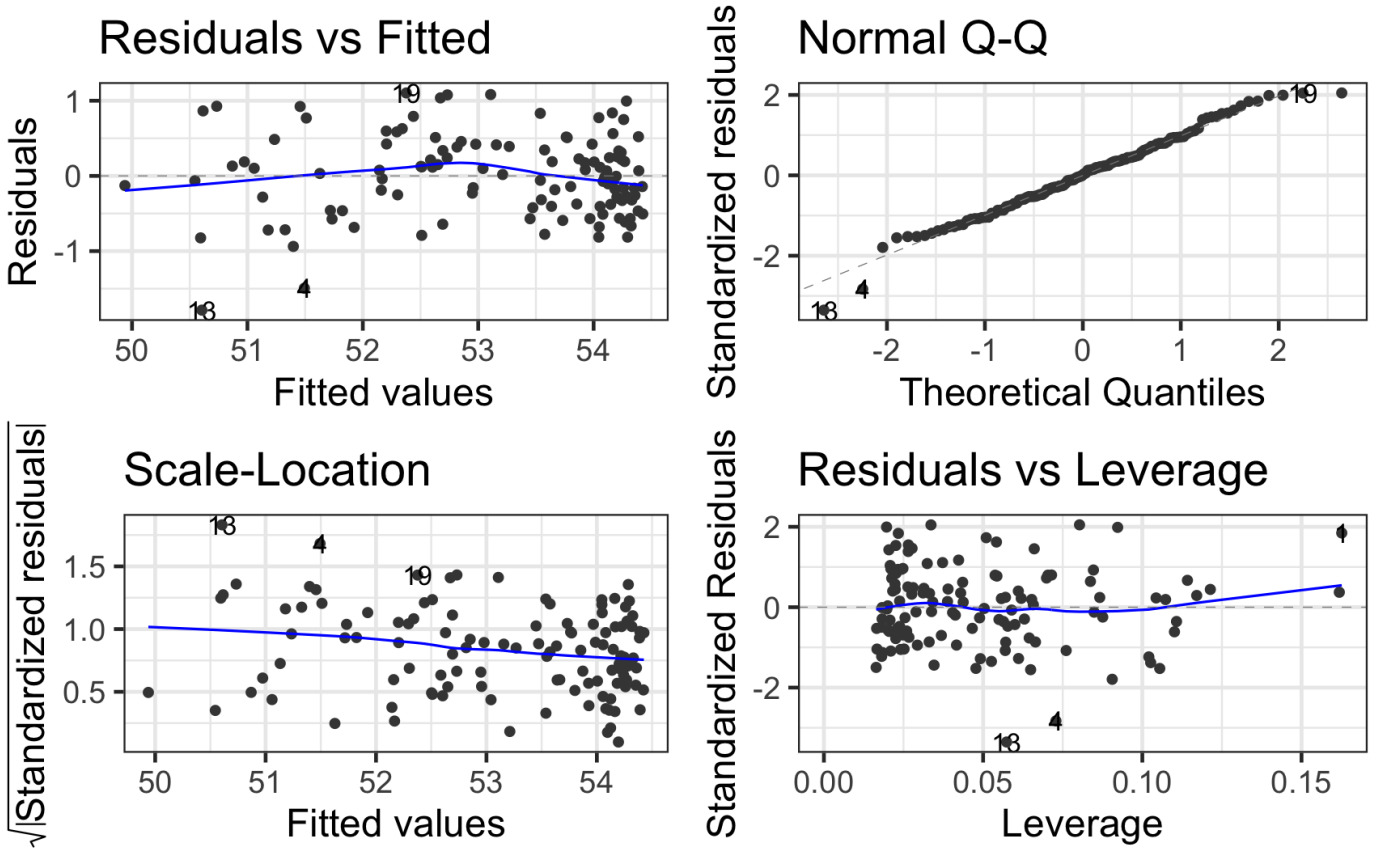
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \, x_1 + \hat{\beta}_2 \, x_2 + \hat{\beta}_3 \, x_2^2$$

General interpretation: When x_2 increases from a to b, y is expected to change by $\hat{\beta}_2(b - a) + \hat{\beta}_3(b^2 - a^2)$, holding x_1 constant.

$$\widehat{speed} = 51.413 - 0.025 \, starters + 0.070 \, yearnew - 0.0004 \, yearnew^2 - 0.477 \, good - 1.393 \, slow$$

Interpret the effect of year for the 5 most recent years (2013 - 2017).

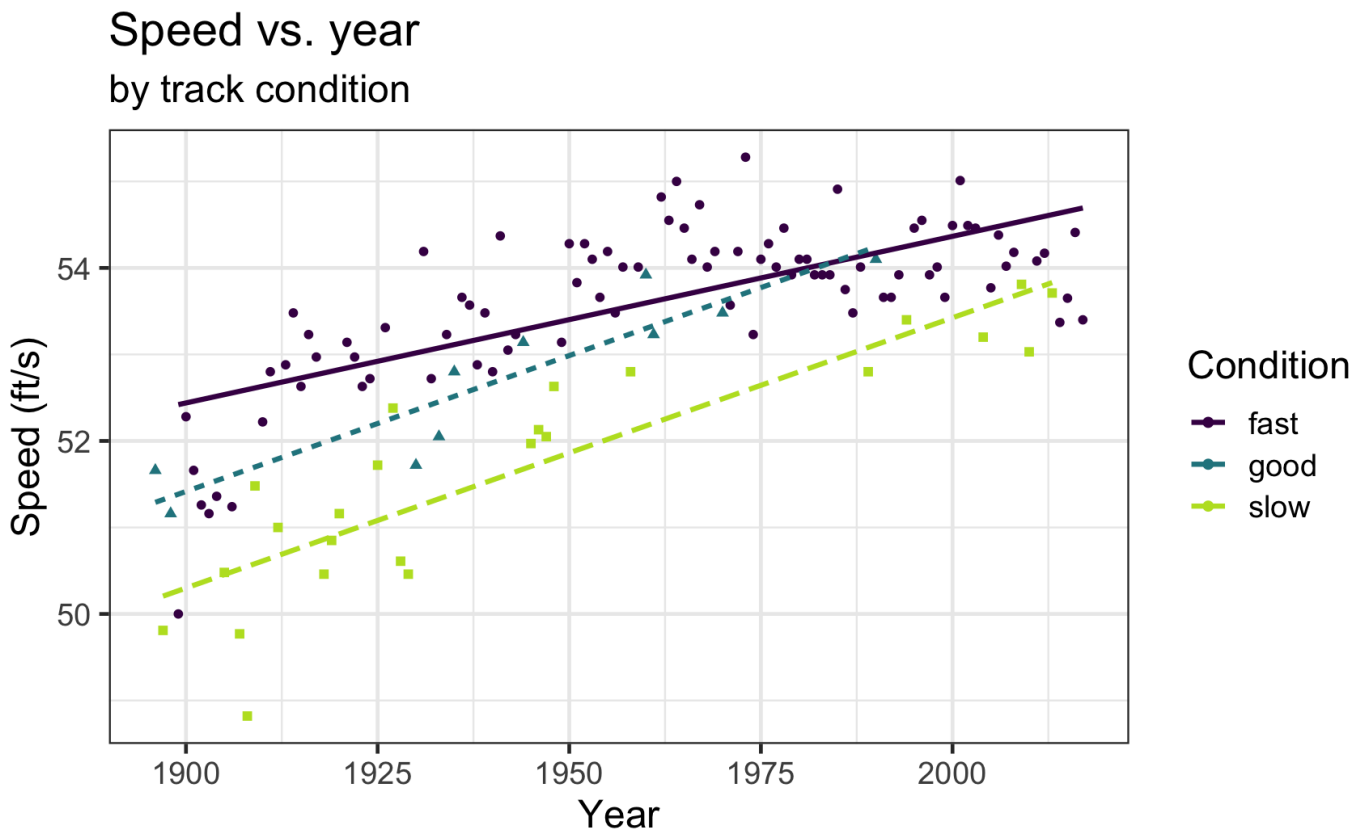
Model 2: Check model assumptions



Model 3

Include interaction term?

Recall from the EDA...



Model 3: Add interaction term

$$\widehat{speed} = 52.387 - 0.003\ starters + 0.020\ yearnew - 1.070\ good - 2.183\ slow + 0.012\ yearnew \times good + 0.012\ yearnew \times slow$$

Output

Code

Assumptions

term	estimate	std.error	statistic	p.value
(Intercept)	52.387	0.200	262.350	0.000
starters	-0.003	0.016	-0.189	0.850
yearnew	0.020	0.003	7.576	0.000
conditiongood	-1.070	0.423	-2.527	0.013
conditionslow	-2.183	0.270	-8.097	0.000
yearnew:conditiongood	0.012	0.008	1.598	0.113
yearnew:conditionslow	0.012	0.004	2.866	0.005

Model 3: Add interaction term

$$\widehat{speed} = 52.387 - 0.003\ starters + 0.020\ yearnew - 1.070\ good - 2.183\ slow + 0.012\ yearnew \times good + 0.012\ yearnew \times slow$$

Output

Code

Assumptions

```
model3 <- lm(speed ~ starters + yearnew + condition +
               yearnew * condition,
             data = derby)
tidy(model3) %>% kable(digits = 4)
```

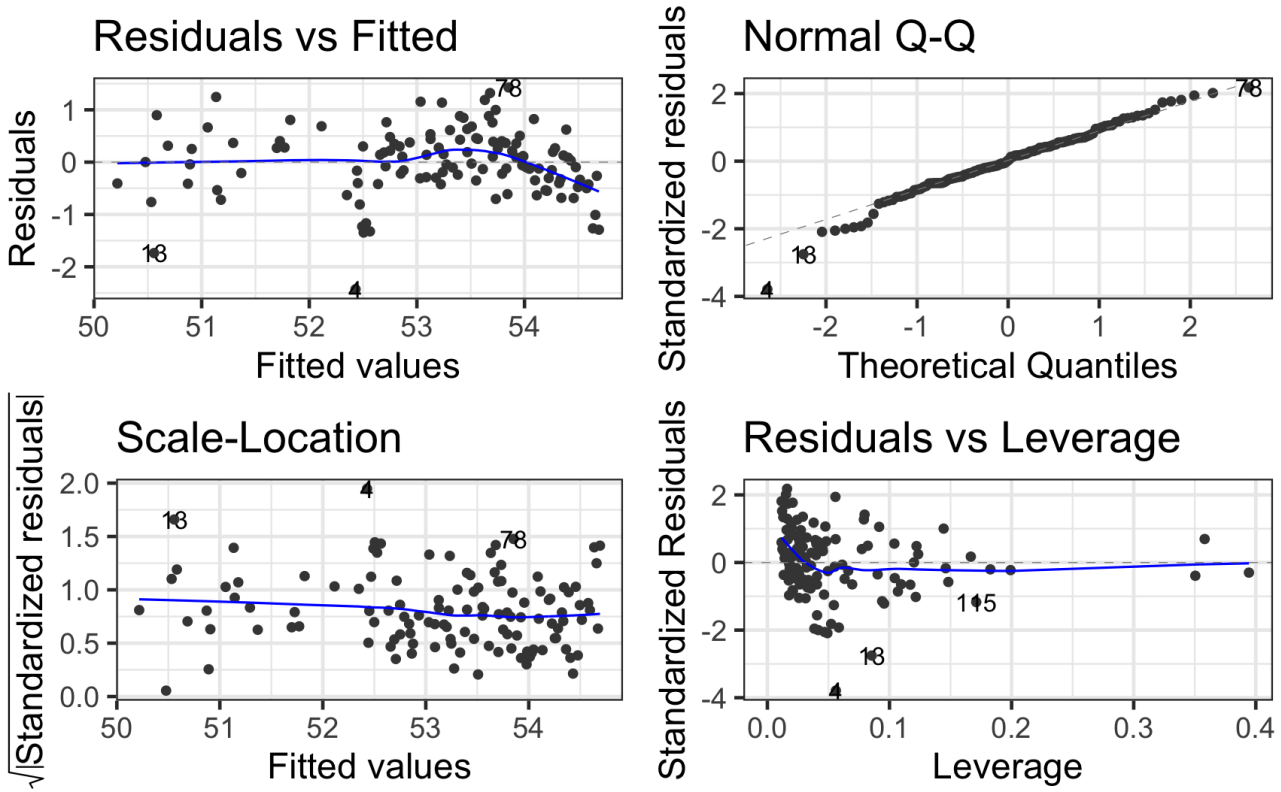
Model 3: Add interaction term

$$\widehat{speed} = 52.387 - 0.003\ starters + 0.020\ yearnew - 1.070\ good - 2.183\ slow + 0.012\ yearnew \times good + 0.012\ yearnew \times slow$$

Output

Code

Assumptions



Interpreting interaction effects

term	estimate	std.error	statistic	p.value
(Intercept)	52.387	0.200	262.350	0.000
starters	-0.003	0.016	-0.189	0.850
yearnew	0.020	0.003	7.576	0.000
conditiongood	-1.070	0.423	-2.527	0.013
conditionslow	-2.183	0.270	-8.097	0.000
yearnew:conditiongood	0.012	0.008	1.598	0.113
yearnew:conditionslow	0.012	0.004	2.866	0.005

Click here for poll

04:00

Which model would you choose?

Output

Code

Model 1: Main effects

r.squared	adj.r.squared	AIC	BIC
0.73	0.721	259.478	276.302

Model 2: Main effects + `(year^2)`

r.squared	adj.r.squared	AIC	BIC
0.827	0.819	207.429	227.057

Model 3: Main effects + interaction

r.squared	adj.r.squared	AIC	BIC

Which model would you choose?

Output

Code

```
# Model 1
glance(model1Cent) %>%
  select(r.squared, adj.r.squared, AIC, BIC) %>%
  kable(digits = 3)

# Model2
glance(model2) %>%
  select(r.squared, adj.r.squared, AIC, BIC) %>%
  kable(digits = 3)

# Model 3
glance(model3) %>%
  select(r.squared, adj.r.squared, AIC, BIC) %>%
  kable(digits = 3)
```

Characteristics of a "good" final model

- Model can be used to answer primary research questions
- Predictor variables control for important covariates
- Potential interactions have been investigated
- Variables are centered, as needed, for more meaningful interpretations
- unnecessary terms are removed
- Assumptions are met and influential points have been addressed
- model tells a "persuasive story parsimoniously"

List from Section 1.6.7 of BMLR

Acknowledgements

These slides are based on content in BMLR: Chapter 1 - Review of Multiple Linear Regression