Multiple Linear Regression Review

Rebecca C. Steorts (slide adaption from Maria Tacket) and material from Chapter 1 of Roback and Legler text.

Computing set up

```
library(tidyverse)
library(tidymodels)
library(GGally)
library(knitr)
library(patchwork)
library(viridis)

ggplot2::theme_set(ggplot2::theme_bw(base_size = 16))

colors <- tibble::tibble(green = "#B5BA72")
```

Topics

- Define statistical models
- Motivate generalized linear models and multilevel models
- Review multiple linear regression

Notes based on Chapter 1 of @roback2021beyond unless noted otherwise.

Statistical models

Models and statistical models

Suppose we have observations $y_1, \ldots, y_n$

- **Model**: Mathematical description of the process we think generates the observations
- **Statistical model**: Model that includes an equation describing the impact of explanatory variables (**deterministic part)** and probability distributions for parts of the process that are assumed to be random variation (**random part)**

Definitions adapted from @stroup2012generalized

Statistical models

A statistical model must include

1. The observations
2. The deterministic (systematic) part of the process
3. The random part of the process with a statement of the presumed probability distribution

Definitions adapted from @stroup2012generalized

Example

Suppose we have the model for comparing two means:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

where

- $i = 1, 2$: group
- $j = 1, \dots, n$: observation number
- n_i: number of observations in group i
- μ_i: mean of group i
- y_{ij}: j^{th} observation in the i^{th} group
- ϵ_{ij} : random error (variation) associated with ij^{th} observation

Adapted from @stroup2012generalized

Example

$$y_{ij} = \mu_i + \epsilon_{ij}$$

- y_{ij}: the observations
- μ_i: deterministic part of the model, no random variability
- e_{ij} : random part of the model, indicating observations vary about their mean
- Typically assume ϵ_{ij} are independent and identically distributed (i.i.d.) $N(0, \sigma^2)$

Adapted from @stroup2012generalized

Practice

Suppose y_{ij}'s are observed outcome data and x_i's are values of the explanatory variable. Assume a linear regression model can used to describe the process of generating y_{ij} based on the x_i.

1. Write the specification of the statistical model.
2. Label the 3 components of the model equation (observation, deterministic part, random part)
3. What is $E(y_{ij})$, the expected value of the observations?

Practice Solution

Consider

$$y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_{ij}.$$

- y_{ij} are the observed data.
- ϵ_{ij} is random.
- β_0, β_1, and x_i are deterministic, fixed.

$E[y_{ij}] = \beta_0 + \beta_1 x_i$

Motivating generalized linear models (GLMs) and 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)

Assumptions for linear regression

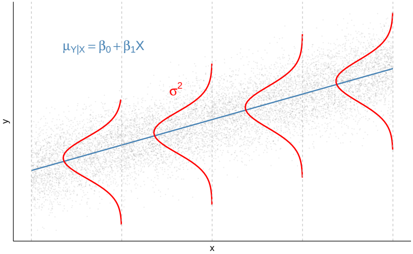


Figure 1: Modified from Figure 1.1. in BMLR

- **Linearity**: Linear relationship between mean of the response Y and the predictor X
- **Independence**: No connection between how far any two points lie from regression line
- **Normality**: Response Y follows a normal distribution at each level of the predictor X (red curves)
- **Equal variance**: Variance of the response Y is equal for all levels of the predictor X

Violations in assumptions

Do wealthy households tend to have fewer children compared to households with lower income? Annual income and family size are recorded for a random sample of households.

- The response variable is number of children in the household.
- The predictor variable is annual income in US dollars.

Which assumption(s) are obviously violated, if any?

Violations in assumptions

Medical researchers investigated the outcome of a particular surgery for patients with comparable stages of disease but different ages. The 10 hospitals in the study had at least two surgeons performing the surgery of interest. Patients were randomly selected for each surgeon at each hospital. The surgery outcome was recorded on a scale of 1 - 10.

- The response variable is surgery outcome, 1 - 10.
- The predictor variable is patient age in years.

Which assumption(s) are obviously violated, if any?

Beyond linear regression

- When drawing conclusions from linear regression models, we do so assuming LINE are all met

Beyond linear regression

- When drawing conclusions from linear regression models, we do so assuming LINE are all met
- **Generalized linear models** require different assumptions and can accommodate violations in LINE

Beyond linear regression

- When drawing conclusions from linear regression models, we do so assuming LINE are all met
- **Generalized linear models** require different assumptions and can accommodate violations in LINE
  - Relationship between response and predictor(s) can be nonlinear

Beyond linear regression

- When drawing conclusions from linear regression models, we do so assuming LINE are all met
- **Generalized linear models** require different assumptions and can accommodate violations in LINE
  - Relationship between response and predictor(s) can be nonlinear
  - Response variable can be non-normal

Beyond linear regression

- When drawing conclusions from linear regression models, we do so assuming LINE are all met
- **Generalized linear models** require different assumptions and can accommodate violations in LINE
  - 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)

Beyond linear regression

- When drawing conclusions from linear regression models, we do so assuming LINE are all met
- **Generalized linear models** require different assumptions and can accommodate violations in LINE
    - 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)
    - **The independence assumption still must hold!**

Beyond linear regression

- When drawing conclusions from linear regression models, we do so assuming LINE are all met
- **Generalized linear models** require different assumptions and can accommodate violations in LINE
  - 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)
  - **The independence assumption still must hold!**
- **Multilevel models** are used to model data that violate the independence assumption, i.e. correlated observations

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 data/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

Goal: Understand variability in average winner speed based on characteristics of the race.

Data

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

Data

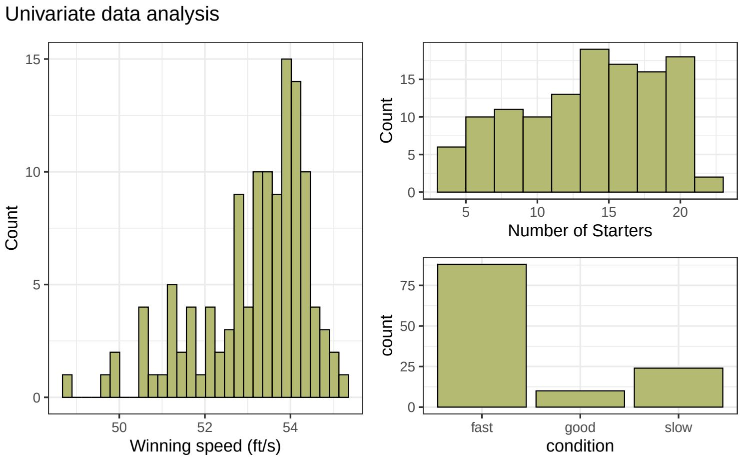
```r
derby |>
  head(5) |> kable()
```

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

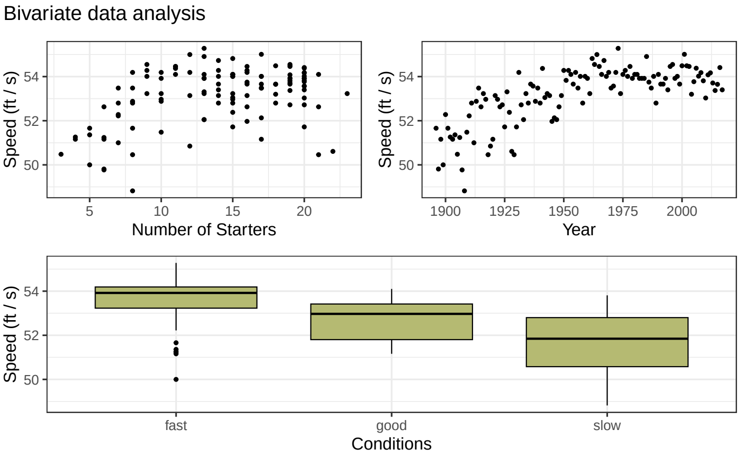
Exploratory data analysis (EDA)

- Once you're ready for the statistical analysis, 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 are used to draw conclusions.

Univariate EDA



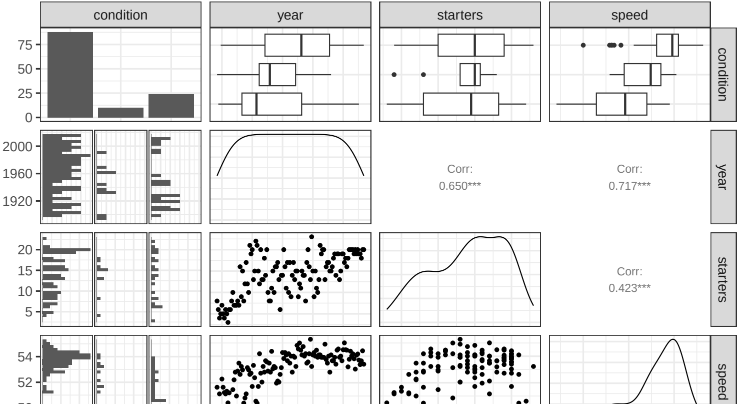
Bivariate EDA



Scatterplot matrix

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

```
## `stat_bin()` using `bins = 30`. Pick better value with `
## `stat_bin()` using `bins = 30`. Pick better value with `
## `stat_bin()` using `bins = 30`. Pick better value with `
```



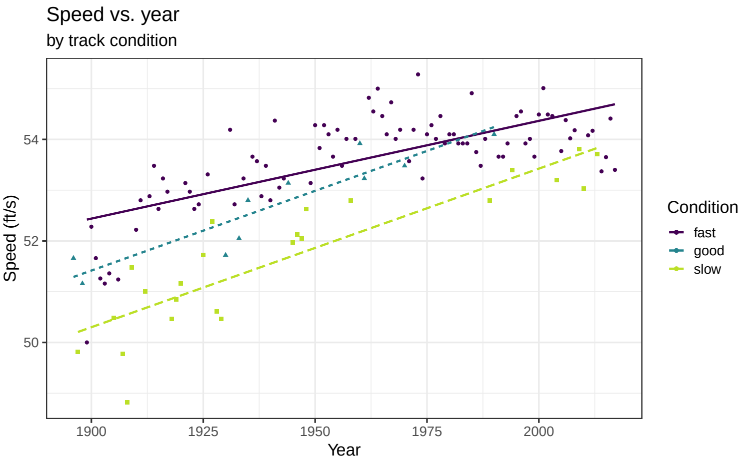
Scatterplot matrix code

Create using the `ggpairs()` function in the GGally package.

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

Multivariate EDA

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



Candidate models

Model 1: Main effects model (year, condition, starters)

```
model1 <- lm(speed ~ starters + year +
                 condition, data = derby)
```

Model 2: Main effects + *year*^{2}, the quadratic effect of year

```
model2 <- lm(speed ~ starters + year +
                 I(year^2) + condition,
               data = derby)
```

Model 3: Main effects + interaction between year and condition

```
model3 <- lm(speed ~ starters + year +
                 condition + year * condition,
               data = derby)
```

Inference for regression

Use statistical inference to

- Evaluate if predictors are statistically significant (not necessarily practically significant!)
- Quantify uncertainty in coefficient estimates
- Quantify uncertainty in model predictions

If LINE assumptions are met, we can use inferential methods based on mathematical models. If at least linearity and independence are met, we can use simulation-based inference methods.

Inference for regression

When LINE assumptions are met... . . .

- Use least squares regression to obtain the estimates for the model coefficients $\beta_0, \beta_1, \ldots, \beta_j$ and for σ^2
- $\hat{\sigma}$ is the **regression standard error**

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p-1}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-p-1}}$$

where p is the number of non-intercept terms in the model (e.g., $p = 1$ in simple linear regression)

- Goal is to use estimated values to draw conclusions about β_j
  - Use $\hat{\sigma}$ to calculate $SE_{\hat{\beta}_j}$. Click here for more detail.

Hypothesis testing for β_j

1. **State the hypotheses**. $H_0 : \beta_j = 0$ vs. $H_a : \beta_j \neq 0$, given the other variables in the model.

2. **Calculate the test statistic.**

$$t = \frac{\hat{\beta}_j - 0}{SE_{\hat{\beta}_j}}$$

3. **Calculate the p-value.** The p-value is calculated from a t distribution with $n - p - 1$ degrees of freedom.

$$\text{p-value} = 2P(T > |t|) \quad T \sim t_{n-p-1}$$

4. **State the conclusion in context of the data.**

   - Reject H_0 if p-value is sufficiently small.

Confidence interval for β_j

The $C\%$ confidence confidence interval for β_j is

$$\hat{\beta}_j \pm t^* \times SE_{\hat{\beta}_j}$$

where the critical value $t^* \sim t_{n-p-1}$

General interpretation for the confidence interval (LB, UB):

We are $C\%$ confident that for every one unit increase in x_j, the response is expected to change by LB to UB units, holding all else constant.

Inference for regression

When LINE assumptions are met… . . .

- Use least squares regression to obtain the estimates for the model coefficients $\beta_0, \beta_1, \ldots, \beta_j$ and for σ^2
- $\hat{\sigma}$ is the **regression standard error**

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p-1}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-p-1}}$$

  where p is the number of non-intercept terms in the model (e.g., $p = 1$ in simple linear regression)

- Goal is to use estimated values to draw conclusions about β_j
  - Use $\hat{\sigma}$ to calculate $SE_{\hat{\beta}_j}$. Click here for more detail.

Hypothesis testing for β_j

1. **State the hypotheses**. $H_0 : \beta_j = 0$ vs. $H_a : \beta_j \neq 0$, given the other variables in the model.

2. **Calculate the test statistic.**

$$t = \frac{\hat{\beta}_j - 0}{SE_{\hat{\beta}_j}}$$

3. **Calculate the p-value.** The p-value is calculated from a t distribution with $n - p - 1$ degrees of freedom.

$$\text{p-value} = 2P(T > |t|) \quad T \sim t_{n-p-1}$$

4. **State the conclusion in context of the data.**
   - Reject H_0 if p-value is sufficiently small.

Confidence interval for β_j

The $C\%$ confidence confidence interval for β_j is

$$\hat{\beta}_j \pm t^* \times SE_{\hat{\beta}_j}$$

where the critical value $t^* \sim t_{n-p-1}$

General interpretation for the confidence interval (LB, UB):

We are $C\%$ confident that for every one unit increase in x_j, the response is expected to change by LB to UB units, holding all else constant.