Formula sheet

Summary statistics

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}, \qquad r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Types of variables: categorical, quantitative, logical, date.

Descriptive words for univariate distributions:

- unimodal, bimodal, multimodal
- symmetric, right-skewed, left-skewed, uniform
- outliers

Descriptive words for bivariate distributions:

- shape: linear, non-linear, no relationship
- strength: weak, moderate, strong
- form: positive, negative

Tidy data

Verbs: gather, spread, nest/unnest, separate/unite

Wrangling data

Verbs: filter, arrange, select, mutate, summarise, group/ungroup

Grammar of graphics

There are seven components of the grammar that define a data plot: DATA, AESTHETICS/MAPPINGS, GEOM, STAT, POSITION, COORDINATE, FACET.

Colour palettes: sequential, diverging, qualitative

Optimization

One variable

For a single variable x and $f(x)$ a continuously differentiable function on $[a, b]$, recall that the conditions for a local optima are as follows:

$$\begin{aligned} f'(x) &= 0 \quad \text{First-order condition,} \\ f''(x) &< 0 \quad \text{Second-order condition: Max,} \\ f''(x) &> 0 \quad \text{Second-order condition: Min.} \end{aligned}$$

Two variables

For two variables x, y and $f(x, y)$ a continuously differentiable function on $[a, b] \times [a, b]$, recall that the conditions for a local optima are as follows:

$$\begin{pmatrix} \frac{\partial f(x,y)}{\partial x} \\ \frac{\partial f(x,y)}{\partial y} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{First-order condition,}$$

$$\frac{\partial^2 f(x,y)}{\partial x^2} < 0, \frac{\partial^2 f(x,y)}{\partial y^2} < 0, \left\{ \left(\frac{\partial^2 f(x,y)}{\partial x^2} \right) \left(\frac{\partial^2 f(x,y)}{\partial y} \right) - \frac{\partial^2 f(x,y)}{\partial x \partial y} \right\} > 0 \quad \text{Second-order condition: Max,}$$

$$\frac{\partial^2 f(x,y)}{\partial x^2} > 0, \frac{\partial^2 f(x,y)}{\partial y^2} > 0, \left\{ \left(\frac{\partial^2 f(x,y)}{\partial x^2} \right) \left(\frac{\partial^2 f(x,y)}{\partial y} \right) - \frac{\partial^2 f(x,y)}{\partial x \partial y} \right\} > 0 \quad \text{Second-order condition: Min.}$$

Models

Simple linear:

$Y = \beta_0 + \beta_1 X + \varepsilon$

- $\varepsilon \sim N(\mu, \sigma)$
- Fitted values: $\hat{Y} = b_0 + b_1 X$
- Residual: $e = Y - \hat{Y}$
- Estimates: $b_1 = r\frac{s_y}{s_x}$, $b_0 = \bar{Y} - b_1 \bar{X}$
- $R^2 = 1 - \frac{\sum e^2}{\sum Y^2}$
- $MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{(n-2)}$
- $RMSE = \sqrt{MSE}$
- $MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{(n-2)}$

Decision trees:

ANOVA criterion: $SS_T - (SS_L + SS_R)$, $SS_T = \sum (y_i - \bar{y})^2$, and SS_L, SS_R are the equivalent values for the two subsets created by partitioning.