Formula sheet

Hypothesis testing

Statistic	Population	Sample
Mean	μ	$\bar{X}$
Proportion	π	p
Correlation	ρ	r

- one sample test: $H_0 : \mu = \mu_0$ and $H_a : \mu > (or <)\mu_0$, $H_a : \mu \neq \mu_0$
- two sample: $H_0 : \mu_1 = \mu_2$ and $H_a : \mu_1 > (or <)\mu_2$, $H_a : \mu_1 \neq \mu_2$
- α: Probability of Type I Error
- β: Probability of Type II Error
- One sample test statistic: $t = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$

Statistical distributions

- Uniform, e.g. $P(X = x) = f(x) = 1/10, \quad x \in \{0, ..., 9\}$
- Normal: $N(\mu, \sigma)$, $f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$
- Exponential: $Exp(\lambda)$, $f(x \mid \lambda) = e^{-\lambda x} \quad x \geq 0$
- Poisson: $P(X = x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \quad x \in \{0, 1, 2, ...\}$
- Binomial: $P(X = x \mid n, p) = \begin{pmatrix} n \\ p \end{pmatrix} p^x (1-p)^{n-x} \quad x \in \{0, 1, 2, ..., n\}$
- Pareto: $f(x \mid \alpha, \lambda) = \frac{\alpha\lambda^\alpha}{(\lambda+x)^{\alpha+1}} \quad x > 0, \alpha > 0, \lambda > 0$
- Weibull: $f(x \mid \lambda, k) = \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{(-x/\lambda)^k}, \quad x \geq 0$
- Gamma: $f(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-x\beta}, \quad x \geq 0 \;\; \alpha, \beta > 0$

Likelihood function:

$L(X_1, \ldots, X_n|\theta) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n \mid \theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i; \theta)$

Regression 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: $b_1 = r\frac{s_y}{s_x}$, $b_0 = \bar{Y} - b_1\bar{X}$
- $R^2 = 1 - \frac{\sum e^2}{\sum Y^2}$

Diagnostics:

- Leverage: cutoff $2p/n$
- Influence, CooksD: cutoff $4/n$
- Collinearity, VIF: cutoff 10

Model fit and significance:

- $MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-p)}$
- Confidence interval for β_k: $b_k \pm t_{\alpha/2,n-2} SE(b_k)$
- Confidence interval for predicted value: $\hat{y} \pm t_{\alpha/2,n-2}\sqrt{MSE(\frac{1}{n} + \frac{n(x-\bar{X})^2}{n\sum_{i=1}^{n}(X_i-\bar{X})^2})}$
- Prediction interval: $\hat{y} \pm t_{\alpha/2,n-2}\sqrt{MSE(1 + \frac{1}{n} + \frac{n(x-\bar{X})^2}{n\sum_{i=1}^{n}(X_i-\bar{X})^2})}$

Generalised linear model:

- Poisson: $y = exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2) + \varepsilon$
- Binomial: $y = \frac{exp(\beta_0+\beta_1 x_1+\beta_2 x_2)}{1+exp(\beta_0+\beta_1 x_1+\beta_2 x_2)} + \varepsilon$

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.

$RMSE = \sqrt{MSE}$

Bayesian thinking

$P(A \text{ and } B) = P(A)P(B|A)$

Bayes Theorem: $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$