Statistical Inference and Data Analysis

Prof: Stefan Van Aelst

January 18th 2023

1. Consider a random sample $X_1, \ldots, X_n$ from $X \sim \text{Geom}(p)$. For X it holds that

$$P(X = k) = p(1-p)^k \quad \text{for } k = 0, 1, 2, \ldots$$

so $F_X(k) = 1 - (1-p)^{k+1}$, $E[X] = \frac{1-p}{p}$, $\text{Var}[X] = \frac{1-p}{p^2}$ and $\phi_X(t) = \frac{p}{1-(1-p)e^{it}}$.

   (a) Calculate the MLE for p.

   (b) Use the properties of the MLE to determine its asymptotic distribution.

   (c) Construct an estimator for the asymptotic variance of the MLE and find its asymtotic distribution.

   (d) Consider the hypothesis test

$$H_0 : p = 1/2$$
$$H_1 : p \neq 1/2$$

   Construct the Wald test and score tests for this problem. Are they different or not?

2. Consider $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma})$ and split $\mathbf{X}, \mu, \mathbf{\Sigma}$ as follows

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}$$

Here $\mathbf{X}_1 \in \mathbb{R}^{q \times 1}$ and $\mathbf{X}_2 \in \mathbb{R}^{p-q \times 1}$.

   (a) Show that

$$\text{Cor}(\mathbf{a}^T\mathbf{X}_1, \mathbf{b}^T\mathbf{X}_2) = \frac{\mathbf{a}^T\mathbf{\Sigma}_{12}\mathbf{b}}{(\mathbf{a}^T\mathbf{\Sigma}_{11}\mathbf{a})^{1/2}(\mathbf{b}^T\mathbf{\Sigma}_{22}\mathbf{b})^{1/2}}$$

   (b) Use appropriate transformations to show that

$$\max_{a\neq 0, b\neq 0}\text{Cor}(\mathbf{a}^T\mathbf{X}_1, \mathbf{b}^T\mathbf{X}_2) = \max_{u\neq 0, v\neq 0}\frac{(\mathbf{u}^T\mathbf{M}\mathbf{v})^2}{(\mathbf{u}^T\mathbf{u})(\mathbf{v}^T\mathbf{v})}$$

   with $\mathbf{M} = \mathbf{\Sigma}_{11}^{-1/2}\mathbf{\Sigma}_{12}\Sigma_{22}^{-1/2}$.

(c) Use the Cauchy-Schwarz inequality to prove that

$$\max_{u\neq 0, v\neq 0} \frac{(\mathbf{u}^T\mathbf{M}\mathbf{v})^2}{(\mathbf{u}^T\mathbf{u})(\mathbf{v}^T\mathbf{v})} \leq \max_{v\neq 0} \frac{\mathbf{v}^T\mathbf{M}^T\mathbf{M}\mathbf{v}}{\mathbf{v}^T\mathbf{v}}$$

with equality if $\mathbf{u} = c\mathbf{M}\mathbf{v}$ for some constant $c \neq 0$.

(d) Show that

$$\max_{a\neq 0, b\neq 0} \text{Cor}(\mathbf{a}^T\mathbf{X}_1, \mathbf{b}^T\mathbf{X}_2) = \sqrt{\rho_1}$$

with ρ_1 the largest eigenvalue of the matrix $\mathbf{\Sigma}_{22}^{-1/2}\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\Sigma_{22}^{-1/2}$ and is reached for $\mathbf{b} = \mathbf{\Sigma}_{22}^{-1/2}\mathbf{e}_1$ and $\mathbf{a} = \mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{b}$ with $\mathbf{e}_1$ the eigenvector corresponding to ρ_1.

(e) Now, consider a random sample $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ distributed as $\mathbf{X} \sim N_p(\mu, \mathbf{\Sigma})$. Construct a likelihood ratio test for the null hypothesis of indepence between $\mathbf{X}_1$ and $\mathbf{X}_2$, i.e. $\mathbf{\Sigma}_{12} = 0$ based on the random sample.

3. We consider a dataset with a representative sample of 30 General Motor (GM) cars from the year 2015. The goal of this study is to estimate retail prices of future cars. The data provides information on the variables 'Price, Mileage, Make (Buick, Cadillac or Pontiac)'. *(Explanations of variables are given.) Next, there was given R-output from a linear regression model*[1]*, including the coefficients and significance codes. The fitted model was also shown in a figure.*

(a) Give the expression for the three regression lines shown in Figure 1.

(b) Given that the variance of the response (log(Price)) equals 46.08, calculate the coefficient of determination (R^2) and the adjusted coefficient of determination (R_a^2) for this model.

(c) Explain in words what the coefficient of determination is.

(d) Consider the following anova table. *(Next, the anova table was given as R-output)* Explain which hypothesis test is performed here and calculate tje value of the F-statistic that is missing in the table. What is your conclusion based on this result?

(e) Let $\mathbf{X}$ denote the design matrix of the linear model. Given the matrix

$$(\mathbf{X}^t\mathbf{X})^{-1} = \textit{6x6- matrix given}$$

construct simultaneous confidence intervals for β_2 and β_3. Motivate your method of choice and give the expressions for the confidence intervals.

(f) Which assumptions need to hold for the above inference results to be valid? Explain for each of your assuptions how they can be verified based on the plots in Figure 2 and formulate your conclusion. *Figure 2 consisted of 5 plots: normal QQ-plot, standardized residuals (for both Fitted values and Mileage), boxplots of standardized residuals and Cook's distances.*

[1]The variable 'Make' was a dummy variable, consisting of three classes, so one class is used as a reference class and didn't appear in the output.