Linear Algebra

Definitions

- [Pseudoinverse] Let $A \in \mathbb{R}^{n\times p}$ and $A = U\Sigma V^T$ be its singular value decomposition, then $A^\dagger = V\Sigma^\dagger U^T = \sum_{i=1}^{\text{rank}(A)} \sigma_i^{-1} v_i u_i^T$
  - $AA^\dagger A = A$
  - $A^\dagger A A^\dagger = A^\dagger$
- [Gamma Function] $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\mathrm{d}z,\ z > 0$
  - $\Gamma(n) = (n-1)!$
- [Beta Function] $\mathrm{Beta}(z_1, z_2) = \int_0^1 x^{z_1-1}(1-x)^{z_2-1}\mathrm{d}x$
  - $\mathrm{Beta}(z_1, z_2) = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1+z_2)}$
- [Chi-Squared Distribution] $X \sim \chi_m^2$, $f(x) = \frac{1}{\Gamma\left(\frac{m}{2}\right)2^{\frac{m}{2}}} x^{\frac{m}{2}-1}e^{-\frac{x}{2}}$
- [Gamma Distribution] $X \sim \Gamma(\alpha, \beta)$, $\alpha, \beta > 0$, $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}e^{-\beta x}$ for $x > 0$
  - $f(x) \propto x^{\alpha-1}e^{-\beta x}$
  - $\mathbb{E}[X] = \frac{\alpha}{\beta}, \mathrm{Var}[X] = \frac{\alpha}{\beta^2}$
  - $\mathbb{E}[\log X] = \psi(\alpha) - \log\beta\,, \mathrm{Var}[\log X] = \psi'(\alpha)$
- [Beta Distribution] $X \sim \mathrm{B}(\alpha, \beta)$, $\alpha, \beta > 0$, $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ for $x \in (0,1)$
  - $f(x) \propto x^{\alpha-1}(1-x)^{\beta-1}$
  - $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}, \mathrm{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}$
  - $\mathbb{E}[\log X] = \psi(\alpha) - \psi(\alpha+\beta), \mathrm{Var}[\log X] = \psi'(\alpha) - \psi'(\alpha+\beta)$
- [Gram Schmidt]
  - $x_1 = u_1$
  - $x_2 = \hat{\beta}_{x_2|u_1}u_1 + u_2$ (OLS guarantees $u_1 \perp u_2$)
  - $x_3 = \hat{\beta}_{x_3|u_1}u_1 + \hat{\beta}_{x_3|u_2}u_2 + u_3$ ($u_1 \perp u_2 \Rightarrow$ reduces to univariate regression)
  - $x_k = \sum_{i=1}^{k-1}\hat{\beta}_{x_k|u_i}u_i + u_k$
- [QR Decomposition] $X \in \mathbb{R}^{n\times p}$, $Q \in \mathbb{R}^{n\times p}$ orthogonal columns, $R \in \mathbb{R}^{p\times p}$ upper triangular.
  - $X = Q\begin{bmatrix} \|u_1\| & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \|u_p\| \end{bmatrix}\begin{bmatrix} 1 & \hat{\beta}_{x_2|u_1} & \cdots & \hat{\beta}_{x_p|u_1} \\ 0 & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \hat{\beta}_{x_p|u_{p-1}} \\ 0 & 0 & \cdots & 1 \end{bmatrix} = QR$
  - $Q = [q_1 \quad \cdots \quad q_p]$ where $q_i = \frac{u_i}{\|u_i\|}$
  - $R\hat{\beta} = Q^TY$
- [Jacobian] $\mathrm{d}s\ \mathrm{d}t = \begin{vmatrix} \frac{\partial s}{\partial u} & \frac{\partial s}{\partial v} \\ \frac{\partial t}{\partial u} & \frac{\partial t}{\partial v} \end{vmatrix}\mathrm{d}u\ \mathrm{d}v$
  - $Dg = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} \end{bmatrix}$
- [Change of Measure] Let $g: \mathbb{R}^n \to \mathbb{R}^n$ be an invertible map and $Y = g(X)$. Then:
  - $f_Y\big(g(x)\big) = |Dg^{-1}|f_X(x)$
  - $f_Y\big(g(x)\big)dy = \mathbb{P}[Y \in (g(x), g(x)+\mathrm{d}y)] = \mathbb{P}[X \in (x, x+|Dg^{-1}|\mathrm{d}y)] = f_X(x)|Dg^{-1}|\mathrm{d}y$

Block Matrices

- $\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D-CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D-CA^{-1}B)^{-1} \\ -(D-CA^{-1}B)^{-1}CA^{-1} & (D-CA^{-1}B)^{-1} \end{bmatrix}$

- $\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}$
- [7.2] Let $X = [X_1 \quad X_2]$. Then $(X^TX)^{-1}_{11} = (X_1^TX_1 - X_1^TX_2(X_2^TX_2)^{-1}X_2^TX_1)^{-1} = \left(\tilde{X}_1^T\tilde{X}_1\right)^{-1}$ where $\tilde{X}_1 = (\mathbb{I} - H_2)X_1$

Sherman Morrison Woodbury

- $(\mathbb{I} + wv^T)^{-1} = \mathbb{I} - \frac{wv^T}{1+v^Tw}$
- $(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}$
- $(A + UV)^{-1} = A^{-1} - A^{-1}U(\mathbb{I} + VA^{-1}U)^{-1}VA^{-1}$
- $(X^TX - x_nx_n^T)^{-1} = (X^TX)^{-1} + \frac{(X^TX)^{-1}x_nx_n^T(X^TX)^{-1}}{1-x_n^T(X^TX)^{-1}x_n}$

Schur's Complement

- Let $\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(0, \begin{bmatrix} \Sigma_{XX} & \Sigma_{\mathrm{XY}} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}\right)$, then:
  - $\mathbb{E}[X|Y] = \mathbb{E}[X] + \Sigma_{XY}\Sigma_{YY}^{-1}(Y - \mathbb{E}[Y])$
  - $\mathrm{Cov}[X|Y] = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}$

Statistical Distributions

- Let $X \sim \chi_m^2$ and $Y \sim \chi_n^2$ independent. Then $\frac{X}{X+Y} \sim \mathrm{Beta}\left(\frac{m}{2}, \frac{n}{2}\right)$
- [B.1] Let $X \sim \Gamma(\alpha, \theta), Y \sim \Gamma(\beta, \theta)$ and $X \perp Y$. Then:
  - $X + Y \sim \Gamma(\alpha + \beta, \theta)$
  - $\frac{X}{X+Y} \sim \mathrm{Beta}(\alpha, \beta)$
  - $X + Y \perp \frac{X}{X+Y}$
- [B.1] $\chi_n^2 \sim \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$
- [B.2] Let $X \sim \Gamma(\alpha, \beta)$, then $\mathbb{E}[X] = \frac{\alpha}{\beta}$, $\mathrm{Var}[X] = \frac{\alpha}{\beta^2}$
- [B.4] Let $X \sim \mathrm{Beta}(\alpha, \beta)$, then $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$, $\mathrm{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Results

- $H \in \mathbb{R}^{n\times n}$ projects onto $C(X)$
- $\mathbb{I}_n - H \in \mathbb{R}^{n\times n}$ projects onto $C(X)^\perp$
- $H(\mathbb{I}_n - H) = 0$

Problem Solving

Problem-Specific Computations

- [Averaging Matrix] $A_n = \frac{1}{n}\mathbb{1}\mathbb{1}^T$
    - $A_n Y = \bar{y}\mathbb{1}_n$
    - A_n is a projection matrix
- [Centering Matrix] $C_n = \mathbb{I}_n - A_n = \mathbb{I}_n - \frac{1}{n}\mathbb{1}\mathbb{1}^T$
    - $C_n Y = Y - \bar{y}\mathbb{1}_n = \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}$
    - C_n is a projection matrix
    - $y^T C_n y = \sum_{i=1}^n (y_i - \bar{y})^2 = (n-1)\hat{\sigma}_y^2$
    - Let $X \in \mathbb{R}^{n\times d}$, then $X^T C_n X = (n-1)\widehat{\text{Cov}}[X] \in \mathbb{R}^{d\times d}$
        - $\widehat{\text{Cov}}[X]_{ij} = \hat{\sigma}_{ij}$ is the sample covariance of covariate i and covariate j
        - $(X^T C_n X)_{ij} = \sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) = (n-1)\hat{\sigma}_{ij}$
- [Stratum Indicator] $S = \begin{bmatrix} \mathbb{1}_{n_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbb{1}_{n_k} \end{bmatrix} \in \mathbb{R}^{n\times k}$, where k is number of stratums
    - $S(S^T S)^{-1} S^T = \begin{bmatrix} \frac{1}{n_1}\mathbb{1}_{n_1}\mathbb{1}_{n_1}^T & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{1}{n_k}\mathbb{1}_{n_k}\mathbb{1}_{n_k}^T \end{bmatrix}$ averages groupwise
    - $\mathbb{I}_n - S(S^T S)^{-1} S^T$ centers groupwise

Single-Variate Regression

- $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- $\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \hat{\rho}_{xy}\frac{\hat{\sigma}_y}{\hat{\sigma}_x}$; $\hat{\beta}_1 = \frac{\text{Cov}[x,y]}{\text{Var}[x]}$
- Under homoskedasticity:
    - [5.8] $\text{RSS} = \sum_{i=1}^n \hat{\epsilon}_i^2 = \left(1 - \hat{\rho}_{xy}^2\right)\sum_{i=1}^n (y_i - \bar{y})^2$
    - $\text{Var}\left[\hat{\beta}_1\right] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$
    - [5.8] t-statistic associated with $\hat{\beta}_1$ is: $\frac{\hat{\rho}_{xy}}{\sqrt{\frac{1-\hat{\rho}_{xy}}{n-2}}} \sim t_{n-2}$ (i.e. testing $H_0\colon \beta_1 = 0$)
- $t_{y\sim x} = t_{x\sim y}$
- $R^2 = \hat{\rho}_{xy}^2 = \frac{\left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}$

Multivariate Regression

- $y_i = \hat{\alpha} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2^T x_{i2} + \hat{\epsilon}_i$
- $\hat{\beta}_1 = \frac{\sum_{i=1}^n \tilde{x}_i \tilde{y}_i}{\sum_{i=1}^n \tilde{x}_i^2} = \frac{\sum_{i=1}^n \tilde{x}_i y_i}{\sum_{i=1}^n \tilde{x}_i^2}$ where $\tilde{x}_i$ is the residual from regressing x_1 on x_2
- [8.4] Under homoskedasticity:
    - t-statistic associated with $\hat{\beta}_1$ is: $\frac{\hat{\rho}_{yx_1|x_2}}{\sqrt{\frac{\left(1-\hat{\rho}_{yx_1|x_2}^2\right)}{n-p}}}$ where p is total number of regressors
    - $\text{Var}\left[\hat{\beta}_1\right] = \sigma^2 (X^T X)_{11}^{-1} = \frac{\sigma^2}{\tilde{x}_1{}^T \tilde{x}_1}$; $X_1 \sim \mathbb{1} + X_{[-1]}$ gives the residual $\tilde{X}_1$
- $R_{yx_1|x_2}^2 = \hat{\rho}_{yx_1|x_2}^2$
- [8.1] Let $X, Y, W \in \mathbb{R}^n$, then $\hat{\rho}_{X,Y|W} = \frac{\hat{\rho}_{X,Y} - \hat{\rho}_{Y,W}\hat{\rho}_{X,W}}{\sqrt{1-\hat{\rho}_{Y,W}^2}\sqrt{1-\hat{\rho}_{X,W}^2}}$

Two Sample t-Test

- $z_1, \ldots, z_m \sim N(\mu_1, \sigma^2)$ i.i.d., $w_1, \ldots, w_n \sim N(\mu_2, \sigma^2)$ i.i.d.
- Under $H_0: \mu_1 = \mu_2$, $t_{\text{equal}} = \frac{\bar{z} - \bar{w}}{\hat{\sigma}\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$ where $\hat{\sigma}^2 = \frac{(m-1)S_z^2 + (n-1)S_w^2}{m+n-2}$
- Equivalently, it is the same as the t-statistic of $H_0: \beta_1 = 0$ in $Y = X\beta + \epsilon$ with $Y = [z_1, \ldots, z_m, w_1, \ldots, w_n]^T$, $X_i = [1,1]$ for z_i and $X_i = [1,0]$ for w_i, $\beta = [\beta_0, \beta_1]$
- $z_1, \ldots, z_m \sim \mu_1, \sigma_1^2$ i.i.d., $w_1, \ldots, w_n \sim \mu_2, \sigma_2^2$ i.i.d.
- $t_{\text{unequal}} = \frac{\bar{z} - \bar{w}}{\sqrt{\frac{S_z^2}{m} + \frac{S_w^2}{n}}} \to N(0,1)$ as $(m, n) \to \infty$
- Same as $H_0: \beta_1 = 0$ in heteroskedastic linear regression with HC2 correction

ANOVA

- [ANOVA] $Y = X_1\beta_1 + X_2\beta_2 + \epsilon$, $\epsilon \sim N(0, \sigma^2 \mathbb{I}_n)$, $\beta_1 \in \mathbb{R}^{p_1}, \beta_2 \in \mathbb{R}^{p_2}$
  - $H_0: \beta_2 = 0$ i.e. under null, $Y = X_1\beta_1 + \epsilon$
  - $\text{RSS}_{\text{long}} = Y^T(\mathbb{I}_n - H)Y$
  - $\text{RSS}_{\text{short}} = Y^T(\mathbb{I}_n - H_1)Y$
  - $F_{\text{ANOVA}} = \frac{\frac{\text{RSS}_{\text{short}} - \text{RSS}_{\text{long}}}{p_2}}{\frac{\text{RSS}_{\text{long}}}{n-p}} = \frac{\text{RSS}_{\text{short}} - \text{RSS}_{\text{long}}}{p_2 \hat{\sigma}^2}$
- [8.2] $F_{\text{ANOVA}} = F_{\text{Wald}}$

Ordinary Least Squares

Gauss-Markov Model

- [Set-Up] The true model is $Y = X\beta + \epsilon$ s.t.:
  - $X \in \mathbb{R}^{n\times d}$ is a fixed design matrix with linearly independent columns
  - ϵ is s.t. $\mathbb{E}[\epsilon] = 0$, $\text{Cov}[\epsilon] = \sigma^2 \mathbb{I}_n$ (i.e. homoskedasticity)
  - (β, σ^2) fixed but unknown
- [Estimators]
  - [OLS] $\hat{\beta} = (X^T X)^{-1} X^T Y$; then $\hat{Y} = X\hat{\beta} = HY$, $\hat{\epsilon} = Y - \hat{Y} = (\mathbb{I} - H)Y$
  - [Residual Sum of Squares] $\text{RSS} = \sum_{i=1}^n \hat{\epsilon}_i^2$
  - [Variance Estimator] $\hat{\sigma}^2 = \frac{RSS}{n-p} = \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{n - \sum_{i=1}^n h_{ii}}$ is unbiased for σ^2
- [Results]
  - $\mathbb{E}[\hat{\beta}] = \beta$, $\text{Cov}[\hat{\beta}] = \sigma^2 (X^T X)^{-1}$
  - $\mathbb{E}\left[\begin{bmatrix}\hat{Y}\\ \hat{\epsilon}\end{bmatrix}\right] = \begin{bmatrix}X\beta\\ 0\end{bmatrix}$; $\text{Cov}\left[\begin{bmatrix}\hat{Y}\\ \hat{\epsilon}\end{bmatrix}\right] = \sigma^2 \begin{bmatrix}H & 0\\ 0 & \mathbb{I}_n - H\end{bmatrix}$
- [Gauss-Markov Theorem] Under the Gauss-Markov model, for any other $\tilde{\beta}$ s.t.
  - $\tilde{\beta}$ is unbiased i.e. $\mathbb{E}[\tilde{\beta}] = \beta$
  - $\tilde{\beta}$ is linear estimator in Y i.e. $\tilde{\beta} = AY$ for some $A \in \mathbb{R}^{p\times n}$

  Then $\text{Cov}[\tilde{\beta}] \succcurlyeq \text{Cov}[\hat{\beta}]$ i.e. $\hat{\beta}$ is the best linear unbiased estimator (i.e. with least variance)
- [t-Statistic] $t_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{(X^T X)^{-1}_{jj}}}$
  - $H_0: \beta_j = 0$
- [F-Statistic] $F = \frac{\hat{\beta}_{1:l}{}^T \left((X^T X)^{-1}_{1:l,1:l}\right)^{-1} \hat{\beta}_{1:l}}{l\hat{\sigma}^2} = \frac{\frac{\text{RSS}(Y\sim\mathbb{1}+X_2) - \text{RSS}(Y\sim\mathbb{1})}{l}}{\frac{\text{RSS}(Y\sim\mathbb{1})}{n-p}}$
  - $H_0: \beta_{1:l} = 0$ where $\beta \in \mathbb{R}^p$

Normal Linear Model

- [Set-Up] The true model is $Y = X\beta + \epsilon$ s.t.:
  - $X \in \mathbb{R}^{n\times p}$ is a fixed design matrix, linearly independent columns
  - $\epsilon \sim N(0, \mathbb{I}_n)$ independent
  - (β, σ^2) fixed but unknown
- [Estimators]
  - [OLS] $\hat{\beta} = (X^T X)^{-1} X^T Y$; then $\hat{Y} = X\hat{\beta} = HY$, $\hat{\epsilon} = Y - \hat{Y} = (\mathbb{I} - H)Y$
  - [Residual Sum of Squares] $\text{RSS} = \sum_{i=1}^n \hat{\epsilon}_i^2$
  - [Variance Estimator] $\hat{\sigma}^2 = \frac{RSS}{n-p} = \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{n - \sum_{i=1}^n h_{ii}}$ is unbiased for σ^2
    - $\hat{\sigma}^2 \sim \frac{\sigma^2}{n-p}\chi^2_{n-p}$
- [5.1] $\begin{bmatrix}\hat{\beta}\\ \hat{\epsilon}\end{bmatrix} \sim N\left(\begin{bmatrix}\beta\\ 0\end{bmatrix}, \sigma^2 \begin{bmatrix}(X^T X)^{-1} & 0\\ 0 & \mathbb{I}_n - H\end{bmatrix}\right)$
  - $\hat{\beta} \perp\!\!\!\perp \hat{\epsilon}$, thus $\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$
- [5.2] $\begin{bmatrix}\hat{Y}\\ \hat{\epsilon}\end{bmatrix} \sim N\left(\begin{bmatrix}X\beta\\ 0\end{bmatrix}, \sigma^2 \begin{bmatrix}H & 0\\ 0 & \mathbb{I}_n - H\end{bmatrix}\right)$
  - $\hat{Y} \perp\!\!\!\perp \hat{\epsilon}$
- [5.3] Let $c \in \mathbb{R}^p$.
  - $c^T(\hat{\beta} - \beta) \sim N(0, \sigma^2 c^T (X^T X)^{-1} c)$
  - $\frac{c^T(\hat{\beta} - \beta)}{\hat{\sigma}\sqrt{c^T (X^T X)^{-1} c}} \sim t_{n-p}$
  - $C_{1-\alpha} = \left[c^T\hat{\beta} - t_{n-p}\left(1 - \frac{\alpha}{2}\right)\hat{\sigma}\sqrt{c^T (X^T X)^{-1} c}, c^T\hat{\beta} + t_{n-p}\left(1 - \frac{\alpha}{2}\right)\hat{\sigma}\sqrt{c^T (X^T X)^{-1} c}\right]$
  - [Hypothesis Testing]

- $H_0: c^T\beta = d$, $H_1: c^T\beta \neq d$; Reject H_0 if $d \notin C_{1-\alpha}$

- [5.4] Let $C \in \mathbb{R}^{k\times p}$. Assume $k \le p$, C is row independent i.e. $C^T\beta = 0 \Rightarrow \beta = 0$
  - $C(\hat\beta - \beta) \sim N(0, \sigma^2 C(X^TX)^{-1}C^T)$
  - $\frac{(C\hat\beta - C\beta)^T \left(C(X^TX)^{-1}C^T\right)^{-1}(C\hat\beta - C\beta)}{k\hat\sigma^2} \sim F_{k,n-p}$
  - $C_{1-\alpha} = \left\{v: \frac{(C\hat\beta - v)^T\left(C(X^TX)^{-1}C^T\right)^{-1}(C\hat\beta - v)}{k\hat\sigma^2} \le F_{k,n-p}(1-\alpha)\right\}$
  - [Hypothesis Testing]
    - $H_0: C\beta = v$, $H_1: C\beta \neq v$; Reject H_0 if $v \notin C_{1-\alpha}$
- [Prediction Interval]
  - $\frac{y_{n+1} - x_{n+1}^T\hat\beta}{\hat\sigma\sqrt{1 + x_{n+1}^T(X^TX)^{-1}x_{n+1}}} \sim t_{n-p}$ (Warning: notice the extra 1 in denominator)
  - $P_{1-\alpha} = \left[x_{n+1}^T\hat\beta - t_{n-p}\left(1-\frac{\alpha}{2}\right)\hat\sigma\sqrt{1 + x_{n+1}^T(X^TX)^{-1}x_{n+1}}, x_{n+1}^T\hat\beta + t_{n-p}\left(1-\frac{\alpha}{2}\right)\hat\sigma\sqrt{1 + x_{n+1}^T(X^TX)^{-1}x_{n+1}}\right]$

Heteroskedastic Linear Model

- Key idea: Heteroskedasticity affects the standard error of β
- [Heteroskedastic Linear Model] The true model is: $y_i = x_i^T\beta + \epsilon_i$
  - ϵ_i independent, $\mathbb{E}[\epsilon_i] = 0$, $\mathrm{Var}[\epsilon_i] = \sigma_i^2$
  - X fixed, linearly independent
  - $(\beta, \sigma_1^2, \ldots, \sigma_n^2)$ unknown parameters
  - Assume $\lim_{n\to\infty} B_n = B$ and $\lim_{n\to\infty} M_n = M$ where B, M are finite.
- [EHW] $\hat V_{EHW} = (X^TX)^{-1}\left(X^T\hat\Omega X\right)(X^TX)^{-1}$
  - $\hat\Omega = \begin{bmatrix} \hat\epsilon_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat\epsilon_n^2 \end{bmatrix}$

Heteroskedastic Linear Model (Results)

- [6.1] Under heteroskedastic linear model, $\hat\beta \to \beta$ in probability
- $B_n = \frac{1}{n}X^TX = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$
- $M_n = \frac{1}{n}X^T\Omega X = \frac{1}{n}\sum_{i=1}^n \sigma_i^2 x_i x_i^T$
- $V := \mathrm{Cov}[\hat\beta] = \frac{1}{n}B_n^{-1}M_nB_n^{-1}$
  - Note that it consists of $\{\sigma_i^2\}_{i=1}^n$ which are unknowns
- $\hat V_{EHW} := (X^TX)^{-1}\left(X^T\hat\Omega X\right)(X^TX)^{-1} = \frac{1}{n}B_n^{-1}\hat M_n B_n^{-1}$
  - $\hat\Omega = \begin{bmatrix} \hat\epsilon_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat\epsilon_n^2 \end{bmatrix}$ is the natural estimator for Ω
- $\hat\beta \sim N\left(\beta, \hat V_{EHW}\right)$ asymptotically
- $\hat V_{EHW,k} = \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \hat\epsilon_{i,k}^2 x_i x_i^T\right)\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^T\right)^{-1}$
- $\hat\epsilon_{i,k} = \begin{cases} \hat\epsilon_i, & k = 0, \mathrm{HC0} \\ \hat\epsilon_i\sqrt{\frac{n}{n-p}}, & k = 1, \mathrm{HC1} \\ \frac{\hat\epsilon_i}{\sqrt{1-h_{ii}}}, & k = 2, \mathrm{HC2} \\ \frac{\hat\epsilon_i}{1-h_{ii}}, & k = 3, \mathrm{HC3} \\ \frac{\hat\epsilon_i}{(1-h_{ii})^{\min\left\{2, \frac{nh_{ii}}{2p}\right\}}}, & k = 4, \mathrm{HC4} \end{cases}$

Partial Regression

Definitions

- [Long Regression] $Y = [X_1 \quad X_2]\begin{bmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{bmatrix} + \hat{\epsilon}$
- [Short Regression] $Y = X_2\tilde{\beta}_2 + \tilde{\epsilon}$
- [Correlation] Given $(x_i, y_i)_{i=1}^n$, the <u>sample correlation</u> is $\hat{\rho}_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$
- [Partial Correlation] Given $(w_i, x_i, y_i)_{i=1}^n$,
  - Perform $Y \sim \mathbb{1} + W$ to get residuals ξ_Y, RSS_y
  - Perform $X \sim \mathbb{1} + W$ to get residuals ξ_X, RSS_x

  The <u>sample partial correlation</u> between x, y given w is $\hat{\rho}_{xy|w} = \frac{\sum_{i=1}^n \hat{\xi}_{x_i}\hat{\xi}_{y_i}}{\sqrt{\sum_{i=1}^n \hat{\xi}_{x_i}{}^2 \sum_{i=1}^n \hat{\xi}_{y_i}{}^2}} = \frac{\sum_{i=1}^n \hat{\xi}_{x_i}\hat{\xi}_{y_i}}{\sqrt{\mathrm{RSS}_x \mathrm{RSS}_y}}$
- [Omitted Variable Bias] Refers to the bias in the estimates of the parameters, due to model leaving out one or more relevant covariates.
  - Model attributes effect of missing covariates to those included in the model
- [Set-Up for Omitted Variable Bias]
  - [Observed Regression] $Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 Z_i + \tilde{\beta}_2^T X_i + \tilde{\epsilon}_i$
  - [True Regression] $Y_i = \hat{\beta}_0 + \hat{\beta}_1 Z_i + \hat{\beta}_2 X_i + \hat{\beta}_3 U_i + \hat{\epsilon}_i$
  - Z_i: parameter of interest e.g. treatment
  - X_i: observed covariates e.g. known confounders
  - U_i: unobserved covariates e.g. unobserved confounders
  - $\hat{\beta}_1$: true effect
  - $\tilde{\beta}_1$: observed effect
- [Confounding Bias] Bias in treatment effect due to presence of unobserved confounders
  - $\mathrm{Bias} = \tilde{\beta}_1 - \hat{\beta}_1 = \hat{\beta}_3\hat{\delta}_1$
  - Scale dependent on Z_i

Theorems

- [7.1 FWL] Let $Y = [X_1 \quad X_2]\begin{bmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{bmatrix} + \hat{\epsilon}$ where $X_1 \in \mathbb{R}^{n\times p_1}$ and $X_2 \in \mathbb{R}^{n \times p_2}$ and $Y = X_2\tilde{\beta}_2 + \tilde{\epsilon}$.

  Let $H_1 = X_1(X_1^TX_1)^{-1}X_1^T$.
  - $\hat{\beta}_2 = [(X^TX)^{-1}X^TY]_{\text{last } p_2 \text{ elements}}$
  - $\hat{\beta}_2 = (X_2^T(\mathbb{I}_n - H_1)X_2)^{-1}X_2^T(\mathbb{I}_n - H_1)Y$
  - $\hat{\beta}_2 = \left(\tilde{X}_2^T\tilde{X}_2\right)^{-1}\tilde{X}_2^TY$ where $\tilde{X}_2 = (\mathbb{I}_n - H_1)X_2$
    - $\hat{\beta}_2$ equals OLS coefficient from regressing Y on $\tilde{X}_2$, the residual matrix from regressing X_2 on X_1
    - $\hat{\beta}_2$ measures the residual "impact" of X_2 on Y after accounting for X_1
    - $\hat{\beta}_2$ as the "impact" of X_2 on Y holding X_1 constant
  - $\hat{\beta}_2 = \left(\tilde{X}_2^T\tilde{X}_2\right)^{-1}\tilde{X}_2^T\tilde{Y}$ where $\tilde{Y} = (\mathbb{I}_n - H_1)Y$
    - You must as well just take out the proportion of Y explained by X_1
    - OLS coefficient as the partial regression coefficient
  - $\tilde{\beta}_2 = (X_2^TX_2)^{-1}X_2^TY$
- [7.2] Let $V \coloneqq \mathrm{Cov}\left[\hat{\beta}_2\right]$. Under homoskedasticity assumption, obtain $\hat{V} = \hat{\sigma}^2(X^TX)^{-1}_{p_2\times p_2}$ from long regression and $\tilde{V} = \tilde{\sigma}^2\left(\tilde{X}_2^T\tilde{X}_2\right)^{-1}$ from short regression.
  - $(n - p_1 - p_2)\hat{V} = (n - p_2)\tilde{V}$
- [7.2] Under heteroskedasticity assumption:
  - $\hat{V}_{EHW} = \left((X^TX)^{-1}X^T\hat{\Omega}X(X^TX)^{-1}\right)_{p_2\times p_2} = \left(\tilde{X}_2{}^T\tilde{X}_2\right)^{-1}\tilde{X}_2{}^T\tilde{\Omega}\tilde{X}_2\left(\tilde{X}_2{}^T\tilde{X}_2\right)^{-1} = \tilde{V}_{EHW}$
- [7.3] Suppose $X_1^TX_2 = 0$, then $\tilde{X}_2 = X_2$ and $\hat{\beta}_2 = \tilde{\beta}_2$.

- [Partial Coefficient via FWL]
  - $\hat{\beta}_{Y\sim X|W} = \hat{\rho}_{XY|W}\sqrt{\frac{RSS_{Y\sim W}}{RSS_{X\sim W}}} = \hat{\rho}_{XY|W}\frac{\hat{\sigma}_{Y\sim W}}{\hat{\sigma}_{X\sim W}}$
    - $\tilde{Y} = (\mathbb{I}_n - H_{\mathbb{1},W})Y$
    - $\tilde{X} = (\mathbb{I}_n - H_{\mathbb{1},W})X$
    - $\hat{\beta}_{Y\sim X|W}$ from $\tilde{Y} \sim \tilde{X}$
  - $\hat{\beta}_{Y\sim X|W}$ from OLS coefficient of X in $Y \sim \mathbb{1} + W + X$
- [8.1] Let $w, x, y \in \mathbb{R}^n$. Then: $\hat{\rho}_{xy|w} = \frac{\hat{\rho}_{xy} - \hat{\rho}_{xw}\hat{\rho}_{yw}}{\sqrt{1-\hat{\rho}_{xw}^2}\sqrt{1-\hat{\rho}_{yw}^2}}$
- [9.1 Cochran] Let $Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\epsilon}$ and $Y = X_2\tilde{\beta}_2 + \tilde{\epsilon}$ and $X_1 = X_2\hat{\delta} + \hat{U}$
  - $\tilde{\beta}_2 = \hat{\beta}_2 + \hat{\delta}\hat{\beta}_1$
- [Cinelli-Hazlett] $|\tilde{\beta}_1 - \hat{\beta}_1|^2 = R^2_{Y\sim U|Z,X}\frac{R^2_{Z\sim U|X}}{1-R^2_{Z\sim U|X}}\frac{RSS(Y\sim\mathbb{1}+Z+X)}{RSS(Z\sim\mathbb{1}+X)}$

Model Fitting, Checking and Misspecification

Definition

- Key ideas:
  - [Fitting] How good do multiple covariates linearly represent the response? (R^2, CC)
  - [Checking] How sensitive / robust is the model to the data? (h_{ii})
  - [Misspecification] If the linear model is wrong, what does β represent?
- [$\hat{\rho}_{xy}$] Given $(x_i, y_i)_{i=1}^n$, $\hat{\rho}_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$
- [$\hat{\rho}_{xy|w}$] Given $(x_i, y_i, w_i)_{i=1}^n$, $\hat{\rho}_{xy|w} = \frac{\sum_{i=1}^n (x_i - \hat{x}_i)(y_i - \hat{y}_i)}{\sqrt{\sum_{i=1}^n (x_i - \hat{x}_i)^2}\sqrt{\sum_{i=1}^n (y_i - \hat{y}_i)^2}} = \hat{\rho}_{\xi_x, \xi_y} = \frac{\sum_{i=1}^n \xi_{x,i}\xi_{y,i}}{\sqrt{\xi_{x,i}^2}\sqrt{\xi_{y,i}^2}}$
  - Perform $Y \sim \mathbb{1} + W$ to get residuals ξ_y, RSS_y
  - Perform $X \sim \mathbb{1} + W$ to get residuals ξ_x, RSS_x
- [R^2] Let Y be a vector and $X \in \mathbb{R}^{n\times(p-1)}$ i.e. excluding $\mathbb{1}_n$. Let $\hat{Y}$ be obtained from $Y \sim \mathbb{1}_n + X$ i.e. p total covariates.
  - $R^2 = \frac{\mathrm{RegSS}}{\mathrm{TSS}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$
    - Proportion of variance explained by the regression
  - $R^2 = \hat{\rho}_{Y\hat{Y}}^2 = \frac{\left(\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y})\right)^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}$
    - Correlation between predicted $\hat{Y}$ and Y
  - $R^2 = \frac{\mathrm{RSS}_{\mathrm{short}} - \mathrm{RSS}_{\mathrm{long}}}{\mathrm{RSS}_{\mathrm{short}}} = \frac{\mathrm{RSS}(Y \sim \mathbb{1}_n) - \mathrm{RSS}(Y \sim \mathbb{1}_n + X)}{\mathrm{RSS}(Y \sim \mathbb{1}_n)}$
- [Partial R^2]
  - $R_{Y,X|W}^2 = R_{\tilde{\epsilon}_Y, \tilde{\epsilon}_X}^2$ where $Y = \mathbb{1}_n \tilde{\beta}_0 + W\tilde{\beta}_1 + \tilde{\epsilon}_Y$ and $X = \mathbb{1}_n \tilde{\delta}_0 + W\tilde{\delta}_1 + \tilde{\epsilon}_X$
  - $R_{Y,X|W}^2 = \frac{\mathrm{RSS}(Y \sim \mathbb{1}_n + W) - \mathrm{RSS}(Y \sim \mathbb{1}_n + X + W)}{\mathrm{RSS}(Y \sim \mathbb{1}_n + W)}$
  - $R_{Y,X|W}^2 = \frac{R_{Y,XW}^2 - R_{Y,W}^2}{1 - R_{Y,W}^2}$
- [Canonical Correlation] Let $x \in \mathbb{R}^p, y \in \mathbb{R}^k$ have joint covariance matrix $\begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$. Then, $CC(x,y) = \max_{a \in \mathbb{R}^p, b \in \mathbb{R}^k} \rho(y^T a, x^T b)$.
  - $\alpha, \beta = \operatorname*{arg\,max}_{a \in \mathbb{R}^p, b \in \mathbb{R}^k} \rho(y^T a, x^T b) = \Sigma_{yy}^{-\frac{1}{2}} v_{\max}\left(\Sigma_{yy}^{-\frac{1}{2}}\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-\frac{1}{2}}\right), \Sigma_{xx}^{-\frac{1}{2}} v_{\max}\left(\Sigma_{xx}^{-\frac{1}{2}}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\Sigma_{xx}^{-\frac{1}{2}}\right)$
  - $CC(x,y) = \left\|\Sigma_{yy}^{-\frac{1}{2}}\Sigma_{yx}\Sigma_{xx}^{-\frac{1}{2}}\right\|_{\mathrm{op}}$
- [Leverage Scores] $h_{ii} = (X(X^TX)^{-1}X^T)_{ii} = x_i^T(X^TX)^{-1}x_i \in \left[\frac{1}{n}, 1\right]$
  - Measure of how much of an outlier x_i is compared to the center of data $\bar{x}$
  - $\sum_{i=1}^n h_{ii} = \mathrm{rank}(H) = p$
  - $\frac{\partial \hat{y}_i}{\partial y_i} = h_{ii}$ i.e. h_{ii} measures contribution of y_i to its own fitted value $\hat{y}_i$
  - $\mathrm{Var}[\hat{y}_i] = \sigma^2 h_{ii}$
- [Leave One Out Setup] Let $X_{[-i]}$ denote the design matrix with row i left out. Then:
  - $\hat{\beta}_{[-i]} = \left(X_{[-i]}^T X_{[-i]}\right)^{-1} X_{[-i]}^T Y_{[-i]}$: OLS estimator when row i is left out
  - $\hat{\epsilon}_{[-i]} = y_i - x_i^T \hat{\beta}_{[-i]}$: residual when y_i is predicted with the leave-ith-row-out estimator

Theorems

- [Fact] R^2 is symmetric w.r.t. Y and X i.e. $R_{Y,X}^2 = R_{X,Y}^2$

- [10.1] $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
    - $\text{TSS} = \text{RegSS} + \text{RSS}$
    - $\text{RSS} = (1 - R^2)\text{TSS} = (1 - R^2)\sum_{i=1}^{n}(y_i - \bar{y})^2$
    - $\text{RegSS} = R^2\text{TSS} = R^2\sum_{i=1}^{n}(y_i - \bar{y})^2$
- [10.1] $R^2 = \hat{\rho}_{y\hat{y}}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}}$
- [10.5] Under the normal linear model i.e. $Y = \mathbb{1}\beta_0 + X\beta_1 + \epsilon$ where $\dim\beta_1 = p$ and $\epsilon_i \sim N(0, \sigma^2)$ independent, then: $\beta_1 = 0 \Rightarrow R^2 \sim \text{Beta}\left(\frac{p-1}{2}, \frac{n-p}{2}\right)$
- $\hat{y}_i = \sum_{j=1}^{n} h_{ij}y_j = h_{ii}y_i + \sum_{j\neq i}^{n} h_{ij}y_j$
- [11.1] Let $X = [\mathbb{1}_n \quad X_2]$, $H = \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T$, $S = \frac{1}{n-1}X_2^T(\mathbb{I} - H)X_2$, $D_i^2 = (x_{i2} - \bar{x}_2)^T S^{-1}(x_{i2} - \bar{x}_2)$. Then: $h_{ii} = \frac{1}{n} + \frac{D_i^2}{n-1}$
    - h_{ii} is a monotone function of D_i i.e. a measure of how far x_i is from $\bar{x}$
- [11.2] $\hat{\beta}_{[-i]} = \hat{\beta} - (1 - h_{ii})^{-1}(X^TX)^{-1}x_i\hat{\epsilon}_i$ provided that $h_{ii} \neq 1$ (Sherman-Morrison)
- [11.3] $\hat{\epsilon}_{[-i]} = \frac{\hat{\epsilon}_i}{1-h_{ii}}$
    - Under Gauss-Markov model:
        - $\text{Var}[\hat{\epsilon}_i] = \sigma^2(1 - h_{ii})$
        - $\text{Var}\left[\hat{\epsilon}_{[-i]}\right] = \frac{\sigma^2}{1-h_{ii}} = \frac{\sigma^2}{1-x_i^T(X^TX)^{-1}x_i} = \sigma^2\left(1 + x_i^T\left(X_{[-i]}^T X_{[-i]}\right)^{-1}x_i\right)$

Manipulations

- [R^2 and RSS]
    - $R_{Y,X}^2 = \frac{\text{RSS}(Y\sim\mathbb{1}) - \text{RSS}(Y\sim\mathbb{1}+X)}{\text{RSS}(Y\sim\mathbb{1})}$
    - $1 - R_{Y,X}^2 = \frac{\text{RSS}(Y\sim\mathbb{1}+X)}{\text{RSS}(Y\sim\mathbb{1})}$
    - $R_{Y,XZ}^2 = \frac{\text{RSS}(Y\sim\mathbb{1}) - \text{RSS}(Y\sim\mathbb{1}+X+Z)}{\text{RSS}(Y\sim\mathbb{1})}$
    - $R_{Y,X|Z}^2 = \frac{\text{RSS}(Y\sim\mathbb{1}+Z) - \text{RSS}(Y\sim\mathbb{1}+Z+X)}{\text{RSS}(Y\sim\mathbb{1}+Z)}$
- [Variance and RSS]
    - $\text{Var}[Y] = \text{RSS}(Y \sim \mathbb{1})$
    - $\text{Var}[Y|X] = \text{RSS}(Y \sim \mathbb{1} + X)$
    - $\text{Var}[Y|X, U] = \text{RSS}(Y \sim \mathbb{1} + X + U)$
- [Correlation and RSS]
    - Let $\hat{Y} = (Y \sim \mathbb{1} + X)$, then $\rho_{Y,\hat{Y}}^2 = R_{Y,X}^2$
    - $\rho_{Y,Z|X,U}^2 = R_{Y,Z|X,U}^2$
- [Coefficient and RSS] $Y \sim \mathbb{1} + X$
    - $\hat{\beta}_1 = \sqrt{\frac{\text{RSS}(Y\sim\mathbb{1}) - \text{RSS}(Y\sim\mathbb{1}+X)}{\text{RSS}(X\sim\mathbb{1})}}$
- [R^2 and F] $F = \frac{n-p}{p-1}\frac{R^2}{1-R^2}$ (i.e. always true) where F and R^2 are for the model $Y = \mathbb{1}_n\hat{\beta}_0 + X\hat{\beta} + \hat{\epsilon}$ and $Y = \mathbb{1}_n\tilde{\beta}_0 + \tilde{\epsilon}$
    - Under normal linear model, if $\beta = 0$, then $R^2 \sim \text{Beta}\left(\frac{p-1}{2}, \frac{n-p}{2}\right)$

Extra

- [Huber] Let $Y = X\beta + \epsilon$ be the true model, where X fixed, ϵ i.i.d. mean 0, variance $\sigma^2 < \infty$ not necessarily normal. Then, any linear combination of $\hat{\beta} = (X^TX)^{-1}X^TY$ is asymptotically normal if and only if $\lim_{n\to\infty}\max_{1\leq i\leq n} h_{ii} = 0$.

Population OLS

Definitions

- [Set-Up] Let $(x_i, y_i) \sim (x, y)$ i.i.d. $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$
  - In particular, X is no longer fixed
  - $\mathbb{E}[Y|X]$ is the best estimator for Y given X, but we restrict to linear estimators
- [Population OLS Coefficient] $\beta = \arg\min_{b \in \mathbb{R}^p} \mathbb{E}_{x,y}[(y - x^T b)^2]$, $\hat{y} = x^T b$
  - $\beta = \mathbb{E}[XX^T]^{-1}\mathbb{E}[XY] = \mathbb{E}[XX^T]^{-1}\mathbb{E}\big[X\mathbb{E}[Y|X]\big]$
  - $\mathrm{Cov}[y - \hat{y}, \hat{y}] = 0$, $\mathrm{Cov}[y, \hat{y}] = \mathrm{Var}[\hat{y}]$
- [Population Residual] $\epsilon \coloneqq y - x^T\beta$
  - [Uncorrelatedness] $\mathbb{E}[x\epsilon] = 0$
- [Population R^2] $R^2 = \frac{\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}}{\sigma_y^2}$
  - [12.5] $R^2 = \frac{\mathrm{Var}[\hat{y}]}{\mathrm{Var}[y]}$
  - [12.6] $R^2 = \max_{b \in \mathbb{R}^{p-1}} \rho^2(y, x^T b) = \rho^2(y, \hat{y})$
- [Population Partial R^2] $R^2_{yx|w} = R^2_{\tilde{y}\tilde{x}}$
  - [12.7] $\rho_{XY|W} = \frac{\rho_{XY} - \rho_{XW}\rho_{YW}}{\sqrt{1-\rho_{XW}^2}\sqrt{1-\rho_{YW}^2}}$
- [Restricted Mean Model] The true model is: $\mathbb{E}[y|x] = x^T\beta$
  - β is parameter of interest
- [Regression Model] Generate (x, ϵ) under constraints e.g. $\mathbb{E}[\epsilon|x] = 0$, then generate $y = x^T\beta + \epsilon$
  - Stronger assumption than correlation model
- [Correlation Model] Start with (x, y), decompose $y = x^T\beta + \epsilon$ where $\mathrm{Cov}[x^T\beta, \epsilon] = 0$

Theorems

- [12.1] Let m be any function. Then: $\mathbb{E}\left[\left(y - m(x)\right)^2\right] = \mathbb{E}\big[\mathrm{Var}[y|x]\big] + \mathbb{E}\left[\left(\mathbb{E}[y|x] - m(x)\right)^2\right]$
  - $\mathbb{E}[y|x] = \arg\min_m \mathbb{E}\left[\left(y - m(x)\right)^2\right]$
- [LLSE] For scalar x, y, the best linear predictor is $\hat{y} = \hat{\alpha} + \hat{\beta}x$
  - $\hat{\beta} = \frac{\mathrm{Cov}[x,y]}{\mathrm{Var}[x]} = \rho_{xy}\sqrt{\frac{\mathrm{Var}[y]}{\mathrm{Var}[x]}}$
  - $\hat{\alpha} = \mathbb{E}[y] - \mathbb{E}[x]\hat{\beta}$
- [Population FWL] Let $y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_{p-1}x_{p-1} + \hat{\epsilon}$ be the population OLS decomposition and $\tilde{y} = \tilde{\beta}_k \tilde{x}_k + \tilde{\epsilon}$.
  - $\hat{\beta}_k = \frac{\mathrm{Cov}[\tilde{x}_k, y]}{\mathrm{Var}[\tilde{x}_k]} = \frac{\mathrm{Cov}[\tilde{x}_k, \tilde{y}]}{\mathrm{Var}[\tilde{x}_k]} = \tilde{\beta}_k$ (Apply $\mathrm{Cov}[\tilde{x}_k, \cdot]$ to the partial regressions)
  - $\hat{\epsilon} = \tilde{\epsilon}$
- [Population Cochran] Let $y = \beta_1^T x_1 + \beta_2^T x_2 + \epsilon$ where x_1, x_2 are random vectors. Let $y = \tilde{\beta}_2^T x_2 + \tilde{\epsilon}$ and $x_1 = \delta^T x_2 + u$, then: $\tilde{\beta}_2 = \beta_2 + \delta\beta_1$

Inference

- $\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i y_i\right)$
- $\sqrt{n}\big(\hat{\beta} - \beta\big) \to N(0, B^{-1}MB^{-1})$ where $B = \mathbb{E}[xx^T]$ and $M = \mathbb{E}[\epsilon^2 xx^T]$
- $\hat{V}_{EHW} = \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2 x_i x_i^T\right)\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^T\right)^{-1}$
- [12.8] Let $(x_i, y_i)_{i=1}^n \sim (x, y)$ i.i.d. with $\mathbb{E}[\|x\|^4] < \infty$ and $\mathbb{E}[y^4] < \infty$, then $\sqrt{n}\big(\hat{\beta} - \beta\big) \to N(0, B^{-1}MB^{-1})$ and $n\hat{V}_{EHW} \to B^{-1}MB^{-1}$ in probability.
- EHW standard error is robust to heteroskedasticity of errors and to misspecification of linear model

Algorithms

Outlier Detection & Model Checking the Normal Linear Model

- [Standardised Residual] $\text{standr}_i = \frac{\hat{\epsilon}_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}$
  - (-) Exact distribution unknown
- [Studentised Residual] $\text{studr}_i = \frac{\hat{\epsilon}_{[-i]}}{\sqrt{\frac{\hat{\sigma}^2{}_{[-i]}}{(1-h_{ii})}}} = \frac{y_i - x_i^T\hat{\beta}_{[-i]}}{\sqrt{\frac{\hat{\sigma}^2{}_{[-i]}}{(1-h_{ii})}}} \sim t_{n-p-1}$
  - $y_i, \hat{\beta}_{[-i]}, \hat{\sigma}^2{}_{[-i]}$ mutually independent
- [Cook Distance] $\text{cook}_i = \frac{(X\hat{\beta}_{[-i]} - X\hat{\beta})^T(X\hat{\beta}_{[-i]} - X\hat{\beta})}{p\hat{\sigma}^2}$
  - cook_i measures change in OLS fitted value after leaving (x_i, y_i) out
  - $\text{cook}_i = \text{standr}_i^2 \frac{h_{ii}}{p(1-h_{ii})}$

Jackknife

- Crude but versatile strategy for bias and variance estimation (and thus bias reduction)
  - Utilises leave-one-out idea; work with pseudo-values
  - Can be used for cross-validation
- $\hat{\theta}_{[-i]}$: estimator of θ without observation i
- [Pseudo-value] $\tilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{[-i]}$
- [Jackknife Point Estimator] $\hat{\theta}_J = \frac{1}{n}\sum_{i=1}^n \tilde{\theta}_i$
- [Jackknife Variance Estimator] $\hat{V}_J = \frac{1}{n(n-1)}\sum_{i=1}^n (\tilde{\theta}_i - \hat{\theta}_J)(\tilde{\theta}_i - \hat{\theta}_J)^T$
- In the context of linear models:
  - [Pseudo-value] $\tilde{\beta}_i = \hat{\beta} + (n-1)\frac{1}{1-h_{ii}}(X^TX)^{-1}x_i\hat{\epsilon}_i$
  - [Jackknife Point Estimator] $\hat{\beta}_J = \hat{\beta} + \frac{n-1}{n}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i \frac{\hat{\epsilon}_i}{1-h_{ii}}\right)$
  - [Jackknife Variance Estimator] $\hat{V}_J = \frac{n-1}{n}(X^TX)^{-1}\left(\sum_{i=1}^n \left(\frac{\hat{\epsilon}^2}{1-h_{ii}}\right)^2 x_i x_i^T\right)(X^TX)^{-1}$

Gauss Updating Algorithm

- Idea: data (x_t, y_t) comes in a stream; want to compute $\hat{\beta}_{(t)}$ online
- [11.4] $\hat{\beta}_{(n+1)} = \hat{\beta}_{(n)} + \gamma_{(n+1)}\hat{\epsilon}_{[n+1]}$
  - $\gamma_{(n+1)} = \left(X_{(n+1)}^T X_{(n+1)}\right)^{-1} x_{n+1}$
  - $\hat{\epsilon}_{[n+1]} = y_{n+1} - x_{n+1}^T\hat{\beta}_{(n)}$: predicted residual of the $(n+1)$th outcome
- [Gauss Updating Algorithm]
  - [Initialise] $V_{(n)} = \left(X_{(n)}^T X_{(n)}\right)^{-1}$, $\hat{\beta}_{(n)}$
  - $V_{(n+1)} = V_{(n)} - \left(1 + x_{n+1}^T V_{(n)} x_{n+1}\right)^{-1} V_{(n)} x_{n+1} x_{n+1}^T V_{(n)}$ `// new inverse via Sherman-Morrison`
  - $\gamma_{(n+1)} = V_{(n+1)}x_{n+1}$, $\hat{\epsilon}_{(n+1)} = y_{n+1} - x_{n+1}^T\hat{\beta}_{(n)}$ `// 11.4`
  - $\hat{\beta}_{(n+1)} = \hat{\beta}_{(n)} + \gamma_{(n+1)}\hat{\epsilon}_{(n+1)}$ `// 11.4`

Conformal Predictions

- Key idea: leverage on i.i.d. distribution and exchangeability to conduct prediction
- Under $H_0: y_{n+1} = y^*$:
  - Obtain residuals $\hat{\epsilon}_i(y^*) = y_i - x_i^T\hat{\beta}(y^*)$ for $i \in \{1, \dots, n+1\}$
  - $\{|\epsilon_i^*(y^*)|\}_{i=1}^{n+1}$ are exchangeable
  - Define the rank of $|\epsilon_j^*(y^*)|$ as $\hat{R}_j(y^*) = 1 + \sum_{i\neq j}^{n+1} \mathbb{1}\{|\hat{\epsilon}_i(y^*)| \leq |\hat{\epsilon}_j(y^*)|\}$
  - $\hat{R}_{n+1}(y^*) \sim \text{Uniform}(\{1, \dots, n+1\})$
  - $\mathbb{P}[\hat{R}_{n+1}(y^*) \leq \lceil(1-\alpha)(n+1)\rceil] \geq 1 - \alpha$

Model Selection

Multicollinearity

- [Variance Inflation Factor] A measure of amount of multicollinearity
  - [Set-Up] $y_i = f(x_i) + \epsilon_i$ is the true model, $\mathbb{E}[\epsilon_i] = 0$, $\mathrm{Var}[\epsilon_i] = \sigma^2$, ϵ_i uncorrelated
  - [Long Regression] $Y \sim \mathbb{1} + X_1 + \cdots + X_p$, giving $Y = \hat{\beta}_0 + \cdots + \hat{\beta}_p X_p + \hat{\epsilon}$
  - [Short Regression] $Y \sim \mathbb{1} + X_j$, giving $Y = \tilde{\beta}_0 + \tilde{\beta}_j X_j + \tilde{\epsilon}$
  - [Variance Inflation Factor] $\frac{1}{1-R_j^2}$ where R_j^2 is the R^2 value from $X_j \sim \mathbb{1} + X_{[-j]}$
  - $\mathrm{Var}\left[\tilde{\beta}_j\right] = \frac{\sigma^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} = \frac{\sigma^2}{\mathrm{RSS}(X_j \sim \mathbb{1} + X_{[-j]})}$
  - $\mathrm{Var}\left[\hat{\beta}_j\right] = \frac{\mathrm{Var}[\tilde{\beta}_j]}{1-R_j^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \frac{1}{1-R_j^2}$

Model Selection Criterions

- [RSS, R^2] Strictly favours large models
- [Adjusted R^2] $\bar{R}^2 = 1 - \frac{n-1}{n-p}(1-R^2) = 1 - \frac{\frac{\mathrm{RSS}(Y\sim\mathbb{1}+X)}{n-p}}{\frac{\mathrm{RSS}(Y\sim\mathbb{1})}{n-1}} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2}$
  - Chooses the model with the smallest estimated variance $\hat{\sigma}^2$ as the best
  - Still favours unnecessarily large models due to upper quantile of F statistic
- [Akaike Information Criterion] $\mathrm{AIC} = n \log\left(\frac{\mathrm{RSS}}{n}\right) + 2p$
  - Selects model that minimises prediction error if the linear model is misspecified
  - Recommended, since linear model assumption cannot be justified in practice
- [Bayes Information Criterion] $\mathrm{BIC} = n \log\left(\frac{\mathrm{RSS}}{n}\right) + p \log n$
  - Consistently selects true model if the linear model is correct
- [Predicted Residual Error Sum of Squares] $\mathrm{PRESS} = \sum_{i=1}^n \hat{\epsilon}_{[-i]}^2 = \sum_{i=1}^n \frac{\hat{\epsilon}_i^2}{(1-h_{ii})^2}$
  - Leave-one-out cross validation; sums up the predicted residuals
  - Analog of RSS (in-sample): "leave-one-out" RSS
- [Generalised Cross Validation] $\mathrm{GCV} = \sum_{i=1}^n \frac{\hat{\epsilon}_i^2}{\left(1-\frac{p}{n}\right)^2} = \frac{\mathrm{RSS}}{\left(1-\frac{p}{n}\right)^2}$
  - Approximation to PRESS
  - As $\frac{p}{n} \to 0$, $\log \mathrm{GCV} \approx \frac{\mathrm{AIC}}{n} + \log n$
- [K-Fold Cross Validation] Computationally attractive
  - Randomly shuffle the observations
  - Split the data into K folds
  - For each fold, use all other folds as the training data; compute the predicted errors on fold $k \in \{1, \dots, K\}$
  - Aggregate prediction errors across the K folds, denoted as K-CV

Model Selection Algorithms

- [Best Subset Selection] Enumerate all 2^p models
- [Forward Selection] Start with $\mathbb{1}$ and greedily include the best covariate; select the best model out of the sequence of models
  - Generally prefer this; works for $p > n$
- [Backward Selection] Start with all covariates and greedily exclude the worst covariate; select the best model out of the sequence of models

Propositions

- [13.2] Consider testing two nested models: $Y = X_1\beta_1 + \epsilon$ and $Y = X_1\beta_1 + X_2\beta_2 + \epsilon$. Then $F > 1 \iff \bar{R}_1^2 < \bar{R}_2^2$.
  - This is equivalent to testing if $\beta_2 = 0$

Ridge and LASSO

Definitions

- [Ridge Regression] $\hat{\beta}_\lambda = (X^TX + \lambda\mathbb{I})^{-1}X^TY = X^T(XX^T + \lambda\mathbb{I})^{-1}Y$
  - $\hat{\beta}_\lambda = \arg\min_\beta\{\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2\} = \arg\min_\beta \left\| \begin{bmatrix} Y \\ 0 \end{bmatrix} - \begin{bmatrix} X \\ \sqrt{\lambda}\mathbb{I} \end{bmatrix}\beta \right\|_2^2$
  - Not invariant to transformations: $X^T\mathbb{1}_n = 0, = 1, Y^T\mathbb{1}_n = 0$
- [Principal Component Analysis] Let X be centered, $X = U\Sigma V^T$
  - The kth principal component of X is $u_k = \frac{1}{\sigma_k}Xv_k$
  - $v_1 = \arg\max_{v:\|v\|=1} v^TX^TXv$
  - $v_2 = \arg\max_{v:\|v\|=1, v\perp v_1} v^TX^TXv$
- [LASSO] $\hat{\beta} = \arg\min_\beta\{\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1\}$

Results

- [Properties of Ridge]
  - $\mathbb{E}[\hat{\beta}_\lambda] \neq \beta$ in general i.e. ridge estimator is biased
  - $\mathrm{Var}[\hat{\beta}_\lambda] = \sigma^2 V \mathrm{diag}\left(\frac{\sigma_1^2}{(\sigma_1^2+\lambda)^2}, \dots, \frac{\sigma_n^2}{(\sigma_n^2+\lambda)^2}\right)V^T$
  - $\mathrm{MSE}(\lambda) = \lambda^2 \sum_{i=1}^p \frac{(v_i^T\beta)^2}{(\sigma_i^2+\lambda)^2} + \sigma^2 \sum_{i=1}^p \frac{\sigma_i^2}{(\sigma_i^2+\lambda)^2}$ where $\{v_i\}_{i=1}^p$ were vectors in V
  - $\lim_{\lambda\to 0} \hat{\beta}_\lambda = X^\dagger Y$
- [Choice of λ]
  - $\lambda_{HKB} = \frac{p\hat{\sigma}^2}{\|\hat{\beta}\|^2}$
  - $\lambda_{LW} = \frac{p\hat{\sigma}^2}{\hat{\beta}^T\Sigma^2\hat{\beta}}$ where Σ is from the SVD of X
- [14.2] Let $\hat{\beta}(\lambda)$ be the ridge estimator as a function of λ and $\hat{\epsilon}(\lambda) = Y - X\hat{\beta}(\lambda)$. Let $H(\lambda) = X(X^TX + \lambda\mathbb{I})^{-1}X^T$ and $h_{ii}(\lambda) = x_i^T(X^TX + \lambda\mathbb{I})^{-1}x_i$.
  - $\hat{\beta}_{[-i]}(\lambda) = \hat{\beta}(\lambda) - \frac{1}{1-h_{ii}(\lambda)}(X^TX + \lambda\mathbb{I})^{-1}x_i\hat{\epsilon}_i(\lambda)$
  - $\hat{\epsilon}_{[-i]}(\lambda) = \frac{\hat{\epsilon}_i(\lambda)}{1-h_{ii}(\lambda)}$
  - $\mathrm{PRESS}(\lambda) = \sum_{i=1}^n \left(\hat{\epsilon}_{[-i]}(\lambda)\right)^2 = \sum_{i=1}^n \frac{(\hat{\epsilon}_i(\lambda))^2}{(1-h_{ii}(\lambda))^2}$
  - $\mathrm{GCV}(\lambda) = \sum_{i=1}^n \frac{(\hat{\epsilon}_i(\lambda))^2}{\left(1-\frac{\mathrm{tr}(H(\lambda))}{n}\right)^2}$
- [SVD Form]
  - [OLS] $\hat{y} = \sum_{i=1}^p \langle u_j, y\rangle u_j$: considers all principal components
  - [Ridge] $\hat{y} = \sum_{i=1}^p \frac{\sigma_j^2}{\sigma_j^2+\lambda}\langle u_j, y\rangle u_j$
    - Deprioritise the less important principal components (too much noise)
  - [Principal Component Regression] $\hat{y} = \sum_{i=1}^{p'} \langle u_j, y\rangle u_j$, $p' \leq p$
    - $Y \sim u_1, \dots, u_{p'}$ i.e. drop the less important principle components

Variants of Least Squares

Transformations

• Key idea: Transform data to hope that residuals become approximately normal
• [Log Transform] $\log y_i = x_i^T \beta + \epsilon_i$
    ○ β interpreted as proportional increase in average outcome
    ○ $\log y_i = x_i^T \beta + \epsilon_i$ and x_j-elasticity of y
• [Box-Cox Transformation] $g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, \ \lambda \neq 0 \\ \log y, \ \lambda = 0 \end{cases}$
• [Basis] $y_i = \sum_{j=1}^{J_1} \beta_{1j} S_j(x_{i1}) + \cdots + \sum_{j=1}^{J_p} \beta_{pj} S_j(x_{ip}) + \epsilon_i$ where S_j are basis functions
• [Polynomial] $S_j(x) = x^j$
• [Discontinuity] $\mathbb{1}\{x > c\}$
• [Kinks] $\mathbb{1}\{x > c\}(x - c) = \max(0, x - c)$

Interactions
• Key idea: interplay of two or more variables acting simultaneously on an outcome • Just add $x_1 x_2$ terms
Restricted OLS

• $\hat{\beta}_r = \underset{b \in \mathbb{R}^p : Cb = r}{\arg\min} \|Y - Xb\|^2$, C full row rank i.e. $\operatorname{rank}(C) = l < p$
• [18.1] If $X^T X$ is invertible, then $\hat{\beta}_r = \hat{\beta} - (X^T X)^{-1} C^T (C (X^T X)^{-1} C^T)^{-1} (C\hat{\beta} - r)$
    ○ Prove by Lagrangian
    ○ $\hat{\beta}_r - \beta = M_r(\hat{\beta} - \beta)$
• [18.2] If $r = 0$, then $\hat{\beta}_r = (\mathbb{I} - (X^T X)^{-1} C^T (C (X^T X)^{-1} C^T)^{-1} C)\hat{\beta} = M_r \hat{\beta}$
    ○ $M_r (X^T X)^{-1} C^T = 0$, $C M_r = 0$, $(\mathbb{I} - C^T (C C^T)^{-1} C) M_r = M_r$
• [18.3] Under Gauss-Markov model, $\mathbb{E}[\hat{\beta}_r] = \beta$, $\operatorname{Cov}[\hat{\beta}_r] = \sigma^2 M_r (X^T X)^{-1} M_r^T$
• [18.2] Under normal linear model, $\hat{\beta}_r \sim N(\beta, \sigma^2 M_r (X^T X)^{-1} M_r^T)$
    ○ $\hat{\sigma}_r^2 = \frac{\|\hat{\epsilon}_r\|^2}{n - p + l}$ is unbiased for σ, where $\hat{\epsilon}_r = Y - X\hat{\beta}_r$
    ○ $\hat{\beta}_r \perp \hat{\sigma}_r^2$

Mechanics

Definitions

- [Sample Correlation Coefficient] $\hat{\rho}_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$
- [Efficiency] Let $\hat{\theta}_1, \hat{\theta}_2 \in \mathbb{R}^n$ be estimators. Then $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if $\mathrm{Cov}[\hat{\theta}_2] \succcurlyeq \mathrm{Cov}[\hat{\theta}_1]$ i.e. $\mathrm{Var}[l^T \hat{\theta}_2] \geq \mathrm{Var}[l^T \hat{\theta}_1]\ \forall l \in \mathbb{R}^n$
- [Rayleigh Quotient] $r(x) = \frac{x^T A x}{x^T x}$, $x \in \mathbb{R}^n$
  - $\lambda_{\max}(A) = \max_{x \neq 0} r(x)$
  - $\lambda_{\min}(A) = \min_{x \neq 0} r(x)$
  - $\lambda_{\min}(A) \leq A_{ii} \leq \lambda_{\max}(A)$
- [Projection Matrix] A matrix $H \in \mathbb{R}^{n \times n}$ is a projection matrix if it is symmetric and $H^2 = H$.
  - Eigenvalues of H must be 0 or 1
  - $\mathrm{tr}(H) = \mathrm{rank}(H)$
- [Pseudoinverse] Let $A = U\Sigma V^T$. Then $A^\dagger = V\Sigma^\dagger U^T$.
  - $AA^\dagger A = A$
  - $A^\dagger A A^\dagger = A^\dagger$
- [Gamma Function] $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \mathrm{d}z$, $z > 0$
  - $\Gamma(n) = (n-1)!$
- [Digamma] $\psi(z) = \frac{\mathrm{d}\log\Gamma}{\mathrm{d}z}$
- [Trigamma] $\psi'(z)$
- [Chi-squared Random Variable] Let $Q_\nu \sim \chi_\nu^2$ be a chi-squared random variable with ν degrees of freedom.
  - If $\nu \in \mathbb{N}$, $Q_\nu = \sum_{i=1}^\nu Z_i^2$ where $Z_i \sim N(0,1)$ i.i.d.
  - $f_\nu(q) = \frac{q^{\frac{\nu}{2}-1} e^{-\frac{q}{2}}}{2^{\frac{\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)}$, $q > 0$
  - $\chi_\nu^2 \sim \Gamma\left(\frac{\nu}{2}, \frac{1}{2}\right)$
- [t Random Variable] A t random variable with degrees of freedom ν is represented as $t_\nu = \frac{Z}{\sqrt{\frac{Q_\nu}{\nu}}}$, where $Z \sim N(0,1)$, $Q_\nu \sim \chi_\nu^2$ and $Z \perp Q_\nu$
- [F Random Variable] A F random variable with degrees of freedom (r, s) is represented as $F = \frac{\frac{Q_r}{r}}{\frac{Q_s}{s}}$ where $Q_r \sim \chi_r^2$, $Q_s \sim \chi_s^2$ and $Q_r \perp Q_s$
- [Gamma Distribution] $X \sim \Gamma(\alpha, \beta)$, $\alpha, \beta > 0$, $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ for $x > 0$
  - $\mathbb{E}[X] = \frac{\alpha}{\beta}, \mathrm{Var}[X] = \frac{\alpha}{\beta^2}$
  - $\mathbb{E}[\log X] = \psi(\alpha) - \log\beta\,, \mathrm{Var}[\log X] = \psi'(\alpha)$
- [Beta Distribution] $X \sim \mathrm{B}(\alpha, \beta)$, $\alpha, \beta > 0$, $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ for $x \in (0,1)$
  - $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}, \mathrm{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)}$
  - $\mathbb{E}[\log X] = \psi(\alpha) - \psi(\alpha+\beta), \mathrm{Var}[\log X] = \psi'(\alpha) - \psi'(\alpha+\beta)$
- [Gumbel Distribution] Let $X_0 \sim \mathrm{Expo}(1)$. Then $Y = \mu - \beta\log X \sim \mathrm{Gumbel}(\mu, \beta)$
  - Let $Y \sim \mathrm{Gumbel}(0,1)$, then $F(y) = e^{-e^{-y}}$, $y \in \mathbb{R}$ and $f(y) = e^{-e^{-y}} e^{-y}$, $y \in \mathbb{R}$
- [Characteristic Function] Let $X \in \mathbb{R}^n$ be a random vector. Then the characteristic function is: $\phi_X(t) = \mathbb{E}\left[e^{it^T X}\right]$ for $t \in \mathbb{R}^n$.
- [Convergence of Random Vectors] Let $X_n, X \in \mathbb{R}^k$ be random vectors. Then $(X_n)_n \to X$ in probability if $\lim_{n\to\infty} \mathbb{P}[\|X_n - X\| > \epsilon] = 0\ \forall \epsilon > 0$

  - If $(X_n)_n \to X$ and $(Y_n)_n \to Y$ in probability, then $(X_n, Y_n)_n \to (X, Y)$ in probability
  - Let $X_1, X_2, \ldots$ be i.i.d. with mean $\mu \in \mathbb{R}^k$, then $\frac{1}{n}\sum_{i=1}^n X_i \to \mu$ in probability
- [Convergence in Distribution] Let $(X_n)_n, X \in \mathbb{R}^k$. Then $(X_n)_n \to X$ in distribution if $\forall$ continuous point z of $t \longmapsto \mathbb{P}[X \le t]$, $\lim_{n\to\infty} \mathbb{P}[X_n \le t] = \mathbb{P}[X \le t]$
- [M-Estimator] Let θ be a parameter and $\hat{\theta}$ be an estimator for θ. Then $\hat{\theta}$ is an $\underline{M}$-estimator if it is a solution to a set of equations of the form $\sum_{i=1}^n U(Y_i; \hat{\theta}) = 0$ where Y_i are i.i.d. observed data.
  - U has same dimensions as θ i.e. $U: \mathbb{R}^n, \mathbb{R}^p \to \mathbb{R}^p$ and must satisfy some regularity conditions
- [Sandwich Covariance Estimator] $\left(\sum_{i=1}^n \frac{\partial}{\partial b}\mathbb{E}[m(Y_i, \hat{\beta})]\right)^{-1}\left(\sum_{i=1}^n m(Y_i, \hat{\beta}) m(Y_i, \hat{\beta})^T\right)\left(\sum_{i=1}^n \frac{\partial}{\partial b}\mathbb{E}[m(Y_i, \hat{\beta})]\right)^{-T}$
  - It is the plug-in estimator of the covariance of $\hat{\beta}$.
  - It is a covariance matrix estimator

Propositions

- Let $A \in \mathbb{R}^{n\times m}$ be of rank k. Then $A = BC$ for some $B \in \mathbb{R}^{n\times k}$ and $C \in \mathbb{R}^{k\times m}$.
- Let $A \in \mathbb{R}^{n\times n}$ be symmetric. Then $A = \sum_{i=1}^n \lambda_i \gamma_i \gamma_i^T$ for orthonormal γ_i.
- [Polar Decomposition] Let $A \in \mathbb{R}^{n\times n}$ with $A = U\Sigma V^T$, then $A = (AA^T)^{\frac{1}{2}}\Gamma$ where $\Gamma = UV^T$ is an orthogonal matrix.
- [B.8] Let $Y_1, Y_2 \sim \text{Expo}(\lambda)$. Then $Y = Y_1 - Y_2 \sim \text{Laplace}\left(0, \frac{1}{\lambda}\right)$
- [B.9] Let $Y_i \sim \text{Gumbel}(\mu, \beta)$ i.i.d. Then $\max_{1\le i\le n} Y_i \sim \text{Gumbel}(\log \sum_{i=1}^n e^{\mu_i}, 1)$
- Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$, then $\text{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^T] \in \mathbb{R}^{n\times m}$
  - $\text{Cov}[X, Y]_{ij} = \mathbb{E}[(X_i - \mathbb{E}[X_i])(Y_j - \mathbb{E}[Y_j])] = \text{Cov}[X_i, Y_j]$
  - $\text{Cov}[X] = \mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}[X]^T$
  - $\text{Cov}[AX + B, CY + D] = A\text{Cov}[X, Y]C^T$
  - $\text{Cov}[AX + BY] = A\text{Cov}[X, Y]B^T + B\text{Cov}[Y, X]A^T$
- [Multivariate Normal] Let $Y \sim N(\mu, \Sigma) \in \mathbb{R}^n$. Then $Y = \mu + AZ$ where $AA^T = \Sigma$ for some k s.t. $A \in \mathbb{R}^{n\times k}$ and $Z \sim N(0, \mathbb{I}_k)$
  - The distribution $\mu + AZ$ is unique regardless of the decomposition $\Sigma = AA^T$, particularly for singular Σ
  - Generally, use $Y = \mu + \Sigma^{\frac{1}{2}}Z$
- [B.14] Let $Z \sim N(0, \mathbb{I}_n)$ and Γ be an orthogonal matrix. Then $\Gamma Z \sim N(0, \mathbb{I}_n)$
- [Properties of Characteristic Function]
  - $\phi_X(t) = \phi_Y(t)$ if and only if $X = Y$ in law
  - If X, Y independent, then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$
  - $X_n \to X$ in distribution if and only if $\phi_{X_n}(t) \to \phi_X(t)\ \forall t \in \mathbb{R}^n$
- [Properties of Multivariate Normal] Let $X \sim N_p(\mu, \Sigma)$ with $\Sigma > 0$
  - $Y = \Sigma^{-\frac{1}{2}}(X - \mu) \sim N_p(0, \mathbb{I}_p)$
  - $X = \Sigma^{\frac{1}{2}}Y + \mu$ where $Y \sim N_p(0, \mathbb{I}_p)$
  - $\mathbb{E}[X] = \mu$, $\text{Var}[X] = \Sigma$
  - Let $v \in \mathbb{R}^p$, then $v^T X$ is univariate normal $\sim N(v^T\mu, v^T\Sigma v)$
  - $U = (X - \mu)^T \Sigma^{-1}(X - \mu) \sim \chi_p^2$
- [B.16] Let $X \sim N(\mu, \sigma^2\mathbb{I}_n)$. If $AB^T = 0$, then $AX \perp BX$.
- [C.4] Let $(X_n)_n \in \mathbb{R}^k$ be zero-mean and with $\text{Cov}[X_n] = a_n C_n$ where $(a_n)_n \to 0$ and $(C_n)_n \to C < \infty$, then $(X_n)_n \to 0$ in probability.
- [C.5] Let $(X_n)_n \to X$ in probability and $\|X_n\| \le \|X\|$ with $\mathbb{E}[\|X\|] < \infty$, then $\mathbb{E}[X_n] \to \mathbb{E}[X]$
  - Prove by subsequence converges a.s., then dominated convergence theorem

- [C.6] Let $(X_n)_n$ be random vectors. Then, $(X_n)_n \to c$ in probability is equivalent to $(X_n)_n \to c$ in distribution.

Theorems

- [A.5 Projection Matrix] Let $X \in \mathbb{R}^{n\times p}$ be of rank p, then $H = X(X^TX)^{-1}X^T \in \mathbb{R}^{n\times n}$ is a projection matrix.
- [A.5 Projection Matrix] Let $H \in \mathbb{R}^{n\times n}$. If H is of rank p, then $H = X(X^TX)^{-1}X^T$ for some $X \in \mathbb{R}^{n\times p}$.
- [B.1] Let $X \sim \Gamma(\alpha,\theta), Y \sim \Gamma(\beta,\theta)$ and $X \perp Y$. Then:
  - $X + Y \sim \Gamma(\alpha+\beta,\theta)$
  - $\frac{X}{X+Y} \sim \beta(\alpha,\beta)$
  - $X + Y \perp \frac{X}{X+Y}$
- [B.4] $\mathrm{Cov}[X,Y] = \mathbb{E}\big[\mathrm{Cov}[X,Y|Z]\big] + \mathrm{Cov}\big[\mathbb{E}[X|Z],\mathbb{E}[Y|Z]\big]$
- [B.5] Let $\begin{bmatrix}X_1\\X_2\end{bmatrix} \sim N\left(\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix},\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\\Sigma_{21} & \Sigma_{22}\end{bmatrix}\right)$, then $X_1 \perp X_2$ if and only if $\Sigma_{12} = \Sigma_{21} = 0$
- [B.6 Lévy-Cramér] Let $X_1 \perp X_2$ and $X_1 + X_2$ be normal. Then both X_1 and X_2 must be normal.
- [B.7] Let $\begin{bmatrix}X_1\\X_2\end{bmatrix} \sim N\left(\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix},\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\\Sigma_{21} & \Sigma_{22}\end{bmatrix}\right)$
  - $X_1 \sim N(\mu_1,\Sigma_{11})$
  - $X_2 \sim N(\mu_2,\Sigma_{22})$
  - If $\Sigma_{22} \succ 0$, then $X_1|X_2 = x_2 \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$
    - Variance of X_1 can only decrease after knowing X_2
    - Σ_{22}^{-1} is rescaling the information gained from X_2
  - $X_1 - \Sigma_{12}\Sigma_{22}^{-1}(X_2-\mu_2) \sim N(\mu_1, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$
  - $X_2 \perp X_1 - \Sigma_{12}\Sigma_{22}^{-1}(X_2-\mu_2)$
- [B.8] Let Y be s.t. $\mathbb{E}[Y] = \mu$, $\mathrm{Cov}[Y] = \Sigma$ and A be a symmetric matrix. Then $\mathbb{E}[Y^TAY] = \mathrm{tr}(A\Sigma) + \mu^TA\mu$
  - $\mathbb{E}[Y^TY] = \Sigma + \mu\mu^T$
- [B.9] Let $Y \sim N(\mu,\Sigma)$ and A be a symmetric matrix, then $\mathrm{Var}[Y^TAY] = 2\mathrm{tr}(A\Sigma A\Sigma) + 4\mu^TA\Sigma A\mu$
- [B.10]
  - Let $Y \sim N(\mu,\Sigma)$ with $\Sigma \succ 0$, then $(Y-\mu)^T\Sigma(Y-\mu) \sim \chi_n^2$. If $\mathrm{rank}(\Sigma) = k < n$, then $(Y-\mu)^T\Sigma^\dagger(Y-\mu) \sim \chi_k^2$
  - Let $Y \sim N(0,\mathbb{I}_n)$ and H be projection matrix of rank k, then $Y^THY \sim \chi_k^2$
  - Let $Y \sim N(0,H)$ where H is a projection matrix of rank k, then $Y^TY \sim \chi_k^2$
- [Lindeberg-Feller CLT] Let $n \in \mathbb{N}$ and $X_{n,1},\dots,X_{n,k_n}$ be independent random vectors s.t. $\mathrm{Cov}\big[X_{n,i}\big] < \infty$. Assuming the following conditions hold:
  - (LF1) $\lim\limits_{n\to\infty}\sum_{i=1}^{k_n}\mathbb{E}\left[\|X_{n,i}\|^2\mathbb{1}\{\|X_{n,i} > c\|\}\right] = 0\ \forall c > 0$
  - (LF1') $\lim\limits_{n\to\infty}\sum_{i=1}^{k_n}\mathbb{E}\left[\|X_{n,i}\|^{2+\delta}\right] = 0$ for some $\delta > 0$
    - (LF1') ⇒ (LF1)
  - (LF2) $\lim\limits_{n\to\infty}\sum_{i=1}^{k_n}\mathrm{Cov}\big[X_{n,i}\big] = \Sigma$

  Then $\sum_{i=1}^{k_n}\big(X_{n,i} - \mathbb{E}\big[X_{n,i}\big]\big) \to N(0,\Sigma)$ in distribution
- [Huber; Asymptotic Normality under Arbitrary Errors] Let $Y = X\beta + \epsilon$ where X is fixed (but n, p are allowed to scale) and ϵ i.i.d., not necessarily normal, with mean 0 and finite variance σ^2. Let $\hat\beta = (X^TX)^{-1}X^TY$ and $H = X(X^TX)^{-1}X^T$. Any linear combination of $\hat\beta$ is asymptotically normal if and only if $\lim\limits_{n\to\infty}\max\limits_{1\le i\le n} H_{ii} = 0$ (referred to as the leverage score condition)
  - [Leverage Score] H_{ii} is the <u>leverage score</u> of unit i

  - [Maximum Leverage Score] $\kappa = \max_{1 \le i \le n} H_{ii}$
- [Continuous Mapping Theorem] Let $f: \mathbb{R}^n \to \mathbb{R}^m$ be continuous except on a measure 0 set. Then $(X_n)_n \to X$ in probability implies $\left(f(X_n)\right)_n \to f(X)$ in probability.
  - $(X_n)_n \to X$ in distribution implies $\left(f(X_n)\right)_n \to f(X)$ in distribution
- [Slutsky's Theorem] Let $(X_n)_n$ and $(Y_n)_n$ be random vectors. Let $(X_n)_n \to X$ in distribution and $(Y_n)_n \to c$ in probability (and equivalently in distribution). Then:
  - $(X_n + Y_n)_n \to X + c$ in distribution
  - $(X_n Y_n)_n \to cX$ in distribution
  - $\left(\frac{X_n}{Y_n}\right)_n \to \frac{X}{c}$ in distribution provided $c \neq 0$
- [Delta Method] Let $f: \mathbb{R}^n \to \mathbb{R}^m$ and $Df \in \mathbb{R}^{n \times m}$. Then $\sqrt{n}(X_n - \theta) \to N(\mu, \Sigma)$ in distribution implies $\sqrt{n}\left(f(X_n) - f(\theta)\right) \to N\left(\left(Df(\theta)\right)^T \mu, \left(Df(\theta)\right)^T \Sigma \left(Df(\theta)\right)\right)$
- [Properties of M-Estimator] Let $\mathbb{E}[U(Y_i; \theta_0)] = 0$ i.e. estimating equation is unbiased:
  - $\hat{\theta}_n$ is asymptotically consistent for θ_0
  - $\hat{\theta}_n$ has an asymptotic distribution of $N(\theta_0, A(\theta_0)^{-1} B(\theta_0) A(\theta_0)^{-1})$
    - $A(\theta_0) = \mathbb{E}\left[-\frac{\partial}{\partial \theta} U(Y_i; \theta)|_{\theta = \theta_0}\right]$
    - $B(\theta_0) = \mathbb{E}[U(Y_i; \theta_0) U(Y_i; \theta_0)^T]$
- [D.1] Let $(Y_i)_{i=1}^n$ be i.i.d. Suppose the true parameter $\beta \in \mathbb{R}^p$ is the unique solution of $\mathbb{E}[m(Y, \beta)] = 0$ and the estimator $\sum_{i=1}^n m\left(Y_i, \hat{\beta}\right) = 0$. Under regularity conditions, $\sqrt{n}\left(\hat{\beta} - \beta\right) \to N(0, B^{-1} M B^{-T})$ in distribution, where $B = -\frac{\partial}{\partial b} \mathbb{E}[m(Y, \beta)]$ and $M = \mathbb{E}[m(Y, \beta) m(Y, \beta)^T]$
  - $\hat{\beta}$ is asymptotically consistent for β
  - $\hat{\beta}$ has an asymptotic distribution of $N(\beta, B^{-1} M B^{-T})$
- [D.2] Let $(Y_i)_{i=1}^n$ be independent. Suppose the true parameter $\beta \in \mathbb{R}^p$ is the unique solution of $\mathbb{E}[m(Y, \beta)] = 0$ and the estimator $\sum_{i=1}^n m\left(Y_i, \hat{\beta}\right) = 0$. Under regularity conditions, $\sqrt{n}\left(\hat{\beta} - \beta\right) \to N(0, B^{-1} M B^{-T})$ in distribution, where $B = -\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial b} \mathbb{E}[m(Y_i, \beta)]$ and $M = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n \mathrm{Cov}[m(Y_i, \beta)]$
  - $\hat{\beta}$ is asymptotically consistent for β
  - $\hat{\beta}$ has an asymptotic distribution of $N(\beta, B^{-1} M B^{-T})$