Notes for 751 and 752

2013 751 and 752 class

November 25, 2013

1 The multivariate normal

Let X be an $n \times p$ matrix, with $X \sim N(\mu, \Sigma)$ where μ is $p \times 1$ and Σ is $p \times p$. Then:

$$f(X; \mu, \Sigma) = (2\pi)^{-1/2} |\Sigma|^{-1/2} \exp\{\frac{-1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)\}$$

- $\Sigma A = E[(X-\mu)(X-\mu')] = E[XX'] - \mu\mu'$
- $X \sim N(A\mu, A\Sigma A')$ provided $A\Sigma A' > 0$
- $\widehat{\mu} = \bar{X} = X(J_n'J_n)^{-1}J_n$, where J_n is an $n \times 1$ column vector of 1's
- $\widehat{\Sigma} = \frac{1}{n-p} X(I_n - H)X'$, where $H = J_n(J_n'J_n)^{-1}J_n'$ (unbiased)
- $E[X'AX] = \mathrm{tr}(A\Sigma) + E[X']AE[X]$

  Proof.

$$\begin{aligned} E[X'AX] &= E[\mathrm{tr}(X'AX)] \\ &= E[\mathrm{tr}(AXX')] \\ &= \mathrm{tr}(E[AXX']) \\ &= \mathrm{tr}(AE[XX']) \\ &= \mathrm{tr}(A(\Sigma + E[X]E[X'])) \\ &= \mathrm{tr}(A\Sigma) + \mathrm{tr}(AE[X]E[X']) \\ &= \mathrm{tr}(A\Sigma) + \mathrm{tr}(E[X]AE[X']) \\ &= \mathrm{tr}(A\Sigma) + E[X]AE[X'] \end{aligned}$$

□

1.1 Gaussian Copula

Let $X \sim N(\mu, \Sigma)$ be $p \times 1$ and let F be the distribution function for X.

$$F_i := N(\mu_i, \theta_{ii}) = \Theta\big(\frac{X - \mu_i}{\sqrt{\theta_{ii}}}\big).$$

Given a function G, define the random variables:

$$U := (F_1(X_1), ..., F_p(X_p))$$

$$Y := (G_1^{-1}(U_1), ..., G_p^{-1}(U_p))$$

- $F \sim U[0, 1]$.
- $P(Y_i \leq y) = P(G_i^{-1}(F(X_i)) \leq y) = G_i(y)$
- We can use the multivariate normal distribution to create more complex multivariate distributions with desired marginals (while maintaining the overall structure of the multivariate normal)

1.2 Multivariate Delta Method

Define a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, where $\nabla f = \begin{pmatrix} \frac{\delta f}{\delta X_1} \\ \vdots \\ \frac{\delta f}{\delta X_n} \end{pmatrix}$ Then:

$$n^{1/2}(f(\bar{X}) - f(\mu)) \sim N(0, \nabla f(\mu)' \Sigma \nabla f(\mu)),$$

and a $(1 - \alpha)100\%$ confidence interval for $f(\mu)$ is:

$$f(\bar{X}) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n} \nabla f(\bar{X})' \Sigma \nabla f(\bar{X})}$$

2 Least squares and the geometry of linear models

Let Y be an $n \times 1$ matrix, X be an $n \times p$ matrix, and β be a $p \times 1$ matrix, and assume $\operatorname{rank}(X) = p$

- $Y \sim X\beta + \epsilon$
- $\widehat{\beta} = (X'X)^{-1}X'Y$ minimizes $\|Y - X\beta\|^2$ (Least Squares)

Proof.

$$\begin{aligned}\|Y - X\beta\|^2 &= \sum_{i=1}^{n}(Y_i - X_i'\beta)^2 \\ &= (Y - X\beta)'(Y - X\beta) \\ &= Y'Y - \beta'X'Y + \beta'X'X\beta - Y'X\beta \\ &= Y'Y - 2\beta'X'Y + \beta'X'X\beta\end{aligned}$$

$$\frac{\delta}{\delta\beta} = 0 - 2X'Y + 2X'X\beta$$

$$\begin{aligned}X'X\widehat{\beta} &= X'Y \\ \widehat{\beta} &= (X'X)^{-1}X'Y\end{aligned}$$

□

- $H := X(X'X)^{-1}X'$ ("Hat Matrix")
  - H is the orthogonal projection of Y onto the columnspace of X
  - $\widehat{Y} = HY$ (i.e. $H : \mathbb{R}^n \to \mathcal{P}$ where $\mathcal{P} := \{\widehat{Y} | \widehat{Y} = X\beta\}$ is a p dimensional subspace of $\mathbb{R}^n$)
  - $\text{rank}(H) = p$, $\text{rank}(I_n - H) = n - p$
  - H and $I_n - H$ are both symmetric and idempotent
- We approximate Y by projecting onto the linear basis associated with X
  - How much of Y have we explained by the columnspace if X?
  - Our least-squares estimation of $\widehat{\beta}$ gives the $\widehat{Y}$ that minimize the distance from Y to its projection onto $\mathcal{P}$
  - Any linear reorganization of the columns of X yields the same approximation (i.e. if $Y = WX\widetilde{\beta} + \epsilon$, then $\widehat{\beta} = W^{-1}\widetilde{\widehat{\beta}}$)

2.1 Residuals

- Residuals measure the distance between the outcome values and the the fitted value
- Least squares minimizes the sum of the residuals
- $e = (I_n - X(X'X)^{-1}X')Y = (I_n - H)Y$

Proof.

$$\begin{aligned} Y - \widehat{Y} &= Y - X\beta \\ &= Y - (X(X'X)^{-1}X'Y) \\ &= (I_n - X(X'X)^{-1}X')Y \\ &= (I_n - H')Y \end{aligned}$$

□

- The residuals are independent from anything in $\mathcal{C}(X)$

2.2 Statistical Properties

Suppose $Y|X \overset{iid}{\sim} N(X\beta, I\sigma^2)$

- $\widehat{\beta}$ is unbiased
- $\mathrm{Var}(\widehat{\beta}) = (X'X)^{-1}\sigma^2$
- $S^2 = \frac{e'e}{n-p}$ is an unbiased estimator for σ^2

  Proof. In general: if $Y \sim N(\mu, \Sigma)$ then $E[Y'AY] = \mathrm{tr}(A\Sigma) + E[Y']AE[Y]$

  $$\begin{aligned} E[e'e] &= E[Y'(I_n - X(X'X)^{-1}X')Y] \\ &= \mathrm{tr}[(I_n - X(X'X)^{-1}X')I_n\sigma^2] + E[Y'](I - X(X'X)^{-1}X')E[Y] \text{ since } E[Y] \in \mathcal{C}(X) \\ &= \mathrm{tr}[(I_n - X(X'X)^{-1}X')I_n\sigma^2] \\ &= \mathrm{tr}(I_n - X(X'X)^{-1}X')\sigma^2 \\ &= [\mathrm{tr}(I_n) - \mathrm{tr}(X(X'X)^{-1}X')]\sigma^2 \\ &= [\mathrm{tr}(I_n) - \mathrm{tr}(X'(X'X)^{-1}X)]\sigma^2 \\ &= [\mathrm{tr}(I_n) - \mathrm{tr}(I_p)\sigma^2] \text{ (for symmetric idempotent matrices, trace=rank)} \\ &= (n-p)\sigma^2 \end{aligned}$$

  In contrast, the MLE of σ^2 is $\frac{e'e}{n}$, and is therefore biased. □

2.3 Quadratic Forms

Let $X \sim N(\mu, \Sigma)$

- $Z = \Sigma^{-1/2}(X - \mu) \sim N(0, I_n)$
- $(X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi^2_n$
- $Z'AZ \sim \chi^2_{\mathrm{rank}(A)}$ where A is an $n \times n$, symmetric, idempotent matrix

- $X'AX \sim \chi^2_{\text{rank}(A)}$ with a non-centrality parameter $\frac{1}{2}\mu' A\mu$

Let $X \sim N(\mu, \sigma^2 I_n)$

- $\frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$ where $\widehat{\sigma}^2 = \frac{e'e}{n-p}$

  Proof.

$$\begin{aligned}
\frac{(n-p)\widehat{\sigma}^2}{\sigma^2} &= \frac{e'e}{\sigma^2} \\
&= \frac{Y'(I - X(X'X)^{-1}X')Y}{\sigma^2} \\
&= Y'\frac{I - X(X'X)^{-1}X'}{\sigma^2}Y \\
&= Y'AY \quad \text{where} \quad A = \frac{I - X(X'X)^{-1}X'}{\sigma^2} \\
A(\sigma^2 I_n) &= I - X(X'X)^{-1}X' \quad \text{is symmetric and idempotent}
\end{aligned}$$

□

2.4 Decomposition of Variation

Let $Y = X\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I_n)$, and X contains an intercept column.

- $J_n \in \mathcal{C}(X)$, where J_n is a column vector of 1's.
- $(I_n - X(X'X)^{-1}X')J_n = 0$, (i.e. $(I_n - H) \perp J_n$), so $J_n = X(X'X)^{-1}X')J_n$

We can decompose the variation in Y into the variation explained by the model (sums of squares regression) and the variation that remains unexplained (sums of squares residuals/ error):

$$\begin{aligned}
\|Y - \bar{Y}\|^2 &= (Y - \bar{Y})'(Y - \bar{Y}) \\
&= (Y - J(J'J)^{-1}J'Y)'(Y - J(J'J)^{-1}J'Y) \\
&= Y'Y - Y'J(J'J)^{-1}J'Y - Y'J(J'J)^{-1}J'Y + Y'J(J'J)^{-1}J'J(J'J)^{-1}J'Y \\
&= Y'Y - Y'J(J'J)^{-1}J'Y - Y'J(J'J)^{-1}J'Y + Y'J(J'J)^{-1}J'Y \\
&= Y'Y - Y'X(X'X)^{-1}X'Y + Y'X(X'X)^{-1}X'Y - Y'J(J'J)^{-1}J'Y \\
&= Y'(I - X(X'X)^{-1}X')Y + Y'(X(X'X)^{-1}X' - J(J'J)^{-1}J)Y \\
&= (Y - \widehat{Y})'(Y - \widehat{Y}) + (\widehat{Y} - \bar{Y})'(\widehat{Y} - \bar{Y}) \\
&= \|Y - \widehat{Y}\|^2 + \|\widehat{Y} - \bar{Y}\|^2 \\
\sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \widehat{Y})^2 \\
SS_{tot} &= SS_{reg} + SS_{res} \\
R^2 &= \frac{SS_{reg}}{SS_{tot}}
\end{aligned}$$

2.5 Analysis of Variance

3 Likelihood and estimation for linear models

Suppose $Y|X \overset{iid}{\sim} N(X\beta, \sigma^2 I_n)$ Then,

$$\mathcal{L}(Y; X\beta, \sigma^2) = \big(\frac{1}{2\pi}\big)^{\frac{n}{2}} \big(\frac{1}{\sigma^2}\big)^{\frac{n}{2}} \exp\{\frac{-n}{2}(Y - X\beta)' \frac{1}{\sigma^2}(Y - X\beta)\}$$

$$\ell(Y; X\beta, \sigma^2) = \frac{-n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{n}{2}(Y - X\beta)' \frac{1}{\sigma^2}(Y - X\beta)$$

$$\begin{aligned} \frac{\delta}{\delta\sigma^2} &= \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}(Y - X\beta)'(Y - X\beta) \\ \widehat{\sigma}^2 &= \frac{(Y - X\beta)'(Y - X\beta)}{n} \\ &= \frac{e'e}{n} \end{aligned}$$

Suppose $Y|X \overset{iid}{\sim} N(X\beta, \Sigma^2)$ Then,

$$\mathcal{L}(Y; X\beta, \sigma^2) = \big(\frac{1}{2\pi}\big)^{\frac{n}{2}} \big(\frac{1}{|\Sigma|}\big)^{\frac{n}{2}} \exp\{\frac{-n}{2}(Y - X\beta)'\Sigma^{-1}(Y - X\beta)\}$$

$$\ell(Y; X\beta, \Sigma) = \frac{-n}{2}\ln(2\pi) - \frac{n}{2}\ln|\Sigma| - \frac{n}{2}(Y - X\beta)'\Sigma^{-1}(Y - X\beta)$$

$$\begin{aligned} \frac{\delta}{\delta\beta} &= -2X'\Sigma^{-1}Y + 2X'\Sigma^{-1}X\beta \\ \widehat{\beta}(\Sigma) &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y \end{aligned}$$

4 Inference for linear models

Let $Y = X\beta + \epsilon$ where Y, X, β are $n \times 1, n \times p, p \times 1$ and $\epsilon \sim N(0, \sigma^2 I_n)$. Suppose K is a $D \times p$ matrix, so that $K(X'X)^{-1}K'$ is $D \times D$ full rank, and let m be a $D \times 1$ vector. We'll test:

$$H_0 : K\beta = m$$

- Under the null hypothesis:

$$K\widehat{\beta} - m \sim N(0, K(X'X)^{-1}X'K'\sigma^2)$$

$$(K\widehat{\beta} - m)'(K(X'X)^{-1}X'K'\sigma^2)^{-1}(K\widehat{\beta} - m) \sim \chi^2_D$$

- If σ^2 is unknown: F-distribution

$$e \perp \widehat{\beta} \Rightarrow e'e \perp (K\widehat{\beta} - m)'(K(X'X)^{-1}X'K'\sigma^2)^{-1}(K\widehat{\beta} - m)$$

$$\frac{(n-p)s^2}{\sigma^2} = \frac{e'e}{\sigma^2} \sim \chi^2_{n-p}$$

$$\frac{1/D}{1/(n-p)} \frac{(K\widehat{\beta} - m)'(K(X'X)^{-1}X'K'\sigma^2)^{-1}(K\widehat{\beta} - m)}{(n-p)s^2/\sigma^2} = \frac{\chi^2_D/D}{\chi^2_{n-p}/(n-p)} \sim F_{D,n-p}$$

4.1 Residual Maximum Likelihood Estimation (REML)

Let $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$

- Let $\widetilde{Y} = (I - H)Y$, where $H = X(X'X)^{-1}X'$. Then $\widetilde{Y} \sim SN(0, (I - H)\sigma^2)$
- Since $\text{rank}(I-H) = n-p$, taking the first $n-p$ entries of $\widetilde{Y}$ will yield a normal distribution.
- We estimate σ^2 in the projection space of X.

$$\widehat{\sigma}^2 = \frac{e'e}{n-p}$$

4.2 Estimability

$Y_{ij} = \mu + \beta_i + \epsilon_{ij}$, where $i = 1, ..., I$ and $j = 1, ..., J$

- The design matrix X is $J \times (I + 1)$, where the first column is all 1's $\Rightarrow$ not full rank
- Define $\theta := \begin{pmatrix} \mu & \beta_1 & ... & \beta_I \end{pmatrix}'$
- $\widehat{\theta} = (X'X)^{\theta}X'Y$ is not unique and depends upon a generalized inverse
- We need a linear constraint on the parameters

Estimability

- A linear function of parameters $q'\beta$ is *estimable* if it is a linear combination of the $E[Y]$ (i.e. it has an unbiased estimator that is a linear function of Y)
- Estimable functions are constant

Best Linear Unbiased Estimator Let q' be an estimable function of the parameters

$$q'\beta = t'X\beta = t'E[Y]$$

Its estimator $q'\widehat{\beta}$ is BLUE

1. Linear in Y
2. Unbiased

3. Minimum variance among collection of linear unbiased estimators

$$\begin{aligned}\mathrm{Var}(q'\widehat{\beta}) &= q'\mathrm{Var}(\widehat{\beta})q \\ &= q'(X'X)^{-1}\sigma^2 q \\ &= t'X(X'X)^{-1}\sigma^2 X't\end{aligned}$$

5 Diagnostics for linear models

5.1 Leverage

Potential influence of a data point

- How far outside the data cloud X is (exert influence depending upon Y)
- $h_{ii} := X_i(X'X)^{-1}X_i', 0 \leq h_{ii} \leq 1$
- $\mathrm{Var}(e_i) = (1 - h_{ii})\sigma^2$
- Studentized residuals: comparability across studies

$$\frac{e_i}{\widehat{\sigma}\sqrt{1 - h_{ii}}} \text{ internally studentized}$$

$$\frac{e_i}{\widehat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} \text{ externally studentized}$$

5.2 Influence

How much the inclusion of a data point influences the fit of your model

- Predicted Residual Sums of Squares: leave one out measure of cross-validation
- PRESS Residual =residual you would obtain if you take observed data point and subtract off fitted value had you removed that point, fit the model, and tried to predict that point. Overfitting is penalized because the error will be high when you leave out the point.

$$e_{i,-i} := Y_i - \widehat{Y}_{i_1-i} = \frac{e_i}{1 - h_{ii}}$$

$$PRESS := \sum_{i=1}^{n}(Y_i - \widehat{Y}_{i,-i})^2$$

DFFITS

- Measures point wise realized influence on fitted values

$$\begin{aligned} DFFITS &:= \frac{\widehat{Y}_i - \widehat{Y}_{i,-i}}{S_{-i}\sqrt{h_{ii}}} \\ &= \frac{Y_i - \widehat{Y}_i}{S_{-i}\sqrt{1-h_{ii}}}\Big(\frac{h_{ii}}{1-h_{ii}}\Big)^{1/2} \\ &= \text{RStudent}\Big(\frac{h_{ii}}{1-h_{ii}}\Big)^{1/2} \end{aligned}$$

- How much influence does the i^{th} point have (divided by the SE of the residual with the i^{th} point removed)

DFBETAS

$$DFBETAS := \frac{\widehat{\beta}_j - \widehat{\beta}_{j,-i}}{S_{-i}\sqrt{c_{ii}}}, \quad \text{where} \quad c_{jj} = [(X'X)^{-1}]_{jj}$$

- Measures point wise realized influence on coefficients
- Compare the j^{th} component of $\widehat{\beta}$ when point i is included and excluded

Cook's Distance

$$D_i := \frac{(\widehat{\beta} - \widehat{\beta}_{-i})(X'X)(\widehat{\beta} - \widehat{\beta}_{-i})}{pS^2}$$

- Measures the effect of deleting a given observation
- Jointly analyze the fitted β with or without the i^{th} observation

6 Confounding, causal inference and the propensity score

Confounding occurs when other variables mediate the relationship you are studying between your predictor and response variables.

Let $Y = X_1\beta_1 + X_2\beta_2 + \epsilon = X\beta + \epsilon$ where X_1 is $n \times p_1$ and X_2 is $n \times p_2$, so that β_1 and β_2 are $p_1 \times 1$ and $p_2 \times 1$, respectively.

- Define $MSE(\widehat{\beta}) := E[(\widehat{\beta}-\beta)'(\widehat{\beta}-\beta)]$ and $B(\widehat{\beta}) := E[\widehat{\beta}]-\beta$

$$\begin{aligned}
MSE(\widehat{\beta}) &= E[(\widehat{\beta}-\beta)'(\widehat{\beta}-\beta)] \\
&= E[\widehat{\beta}'\widehat{\beta}] - E[\widehat{\beta}'\beta] - E[\beta'\widehat{\beta}] + E[\beta'\beta] \\
&= E[\widehat{\beta}'\widehat{\beta}] - E[\widehat{\beta}]'E[\widehat{\beta}] + E[\widehat{\beta}]'E[\widehat{\beta}] - E[\widehat{\beta}'\beta] - E[\beta'\widehat{\beta}] + E[\beta'\beta] \\
&= E[\widehat{\beta}'\widehat{\beta}] - E[\widehat{\beta}]'E[\widehat{\beta}] + E[\widehat{\beta}]'E[\widehat{\beta}] - E[\widehat{\beta}']\beta - \beta'\widehat{E}[\beta] + \beta'\beta \\
&= E[\widehat{\beta}'\widehat{\beta}] - E[\widehat{\beta}]'E[\widehat{\beta}] + (E[\widehat{\beta}]-\beta)'(E[\widehat{\beta}]-\beta) \\
&= E[\widehat{\beta}'\widehat{\beta}] - E[\widehat{\beta}]'E[\widehat{\beta}] + B(\widehat{\beta})'B(\widehat{\beta}) \\
&= \mathrm{tr}\{\mathrm{Var}(\widehat{\beta})\} + B(\widehat{\beta})'B(\widehat{\beta})
\end{aligned}$$

- If we include β_2 in the model, but $\beta_2 = 0$ then:

$$\begin{aligned}
B(\widehat{\beta}) &= B(X'X)^{-1}X'Y) \\
&= E[(X'X)^{-1}X'Y] - \beta \\
&= (X'X)^{-1}X'E[Y] - \beta \\
&= (X'X)^{-1}X'X\beta - \beta \\
&= 0
\end{aligned}$$

If we include an unnecessary covariate, we still get an unbiased estimate.

$$\begin{aligned}
\mathrm{Var}(\widehat{\beta}) &= \mathrm{Var}((X'X)^{-1}X'Y) \\
&= (X'X)^{-1}X'\mathrm{Var}(Y)X(X'X)^{-1} \\
&= (X'X)^{-1}X'\Sigma X(X'X)^{-1} \\
&= (X'X)^{-1}\sigma^2 \quad \text{if} \quad \Sigma = \sigma^2 I_n \\
&= \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1} \sigma^2
\end{aligned}$$

$$\begin{aligned}
MSE(\widehat{\beta}) &= \mathrm{tr}\{\mathrm{Var}(\widehat{\beta})\} + B(\widehat{\beta})'B(\widehat{\beta}) \\
&= \mathrm{tr}\{\mathrm{Var}(\widehat{\beta})\} \\
&= \mathrm{tr}\{\begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1} \sigma^2\} \\
&= \mathrm{tr}\{\begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1}\}\sigma^2
\end{aligned}$$

If we include an unnecessary covariate, the MSE will be increased.

- If we exclude β_2 from the model, but $\beta_2 \neq 0$ then, $\widehat{\beta} = \widehat{\beta}_1$ and:

$$\begin{aligned} B(\widehat{\beta}_1) &= B(X_1'X_1)^{-1}X_1'Y) \\ &= E[(X_1'X_1)^{-1}X_1'Y] - \beta_1 \\ &= (X_1'X_1)^{-1}X_1'E[Y] - \beta_1 \\ &= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2) - \beta_1 \\ &= (X_1'X_1)^{-1}X_1'X_1\beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 - \beta_1 \\ &= (X_1'X_1)^{-1}X_1'X_2\beta_2 \end{aligned}$$

  If we exclude a necessary covariate, we still will get a biased estimate of β_1, unless X_2 is orthogonal to X_1

$$\begin{aligned} \mathrm{Var}(\widehat{\beta}_1) &= \mathrm{Var}((X_1'X_1)^{-1}X_1'Y) \\ &= (X_1'X_1)^{-1}X'\mathrm{Var}(Y)X_1(X_1'X_1)^{-1} \\ &= (X_1'X_1)^{-1}\sigma^2 \end{aligned}$$

  If we exclude a necessary covariate, the variability of our coefficients will decrease??

$$\begin{aligned} MSE(\widehat{\beta}_1) &= \mathrm{tr}\{\mathrm{Var}(\widehat{\beta}_1)\} + B(\widehat{\beta}_1)'B(\widehat{\beta}_1) \\ &= \mathrm{tr}\{(X_1'X_1)^{-1}\sigma^2\} + \beta_2'X_2'X_1(X_1'X_1)^{-1}(X_1'X_1)^{-1}X_1'X_2\beta_2 \\ &= \mathrm{tr}\{(X_1'X_1)^{-1}\}\sigma^2 + \beta_2'X_2'X_1(X_1'X_1)^{-1}(X_1'X_1)^{-1}X_1'X_2\beta_2 \end{aligned}$$

  If we exclude a necessary covariate, the MSE will.....

- $\widehat{\beta}_1 = (e'_{X_1|X_2}e_{X_1|X_2})^{-1}e'_{X_1|X_2}e_{Y|X_2}$, where

$$e_{X_1|X_2} = (I - X_2(X_2'X_2)^{-1}X_2')X_1$$

$$e_{Y|X_2} = (I - X_2(X_2'X_2)^{-1}X_2')Y$$

- If $X_1 \perp X_2$ then:

$$\widehat{\beta} = \begin{pmatrix} (X_1'X_1)^{-1}X_1'Y \\ (X_2'X_2)^{-1}X_2'Y \end{pmatrix}$$

To make inferences on the effects of treatments, we need to estimate the response Y of observation X had he received a different treatment.

- Need potential outcomes given treatment to be independent from treatment assignment (Z)
- Assume potential outcomes are independent from treatment assignment

- Balancing score: $f(X)$ such that treatment assignment is effectively random within levels.
- Propensity Score: "best" (coarsest) function e so that at fixed value, treatment assignment and potential outcomes are independent

$$X \perp Z|b(X)$$

Let $Z = 0$ for controls and $Z = 1$ for treated, and define $Y_\ell(Z)$ as the potential outcome of person ℓ, so that $Y_\ell(1) - Y_\ell(0)$ estimates the person specific effect of treatment.

$$e(X) := P(Z = 1|X),$$

where we assume

$$P(Z_1, ..., Z_n|X_1, ..., X_n) = \Pi_{\ell=1}^n e(X_\ell)^{Z_\ell}\{1 - e(X_\ell)\}^{1-Z_\ell}$$

To estimate $\widehat{e}(X)$, we reduce the dimensions of X.

- Pair matching

7 Asymptotics for linear models

8 Model selection

Bias-Variance Tradeoff

1. $Y = X_1\beta_1 + \epsilon$
2. $Y = X_1\beta_1 + X_2\beta_2 + \epsilon$

Underfitting: we fit model 1 when model 2 is actually true

$$\begin{aligned} E[\widehat{\beta}_1^{(1)}] &= E[(X_1'X_1)^{-1}X_1'Y] \\ &= (X_1'X_1)^{-1}X_1'E[Y] \\ &= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2) \\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 \\ B[\widehat{\beta}_1^{(1)}] &= (X_1'X_1)^{-1}X_1'X_2\beta_2 \end{aligned}$$

Our coefficient estimate is biased

$$\begin{aligned} \mathrm{Var}(\widehat{\beta}_1^{(1)}) &= \mathrm{Var}((X_1'X_1)^{-1}X_1'Y) \\ &= (X_1'X_1)^{-1}X_1'\mathrm{Var}(Y)X_1(X_1'X_1)^{-1} \\ &= (X_1'X_1)^{-1}X_1'\sigma^2X_1(X_1'X_1)^{-1} \\ &= (X_1'X_1)^{-1}\sigma^2 \end{aligned}$$

$$
\begin{aligned}
E[(Y-\widehat{Y})'(Y-\widehat{Y})] &= E[Y'(I-X_1(X_1'X_1)^{-1}X_1'Y] \\
&= \mathrm{tr}(I-X_1(X_1'X_1)^{-1}X_1)\sigma^2 + E[Y'](I-X_1(X_1'X_1)^{-1}X_1')E[Y] \\
&= \mathrm{tr}(I-X_1(X_1'X_1)^{-1}X_1)\sigma^2 + \beta'X'(I-X_1(X_1'X_1)^{-1}X_1')X\beta \\
&= (n-p_1)\sigma^2 + (X_1\beta_1+X_2\beta_2)'(I-X_1(X_1'X_1)^{-1}X_1')(X_1\beta_1+X_2\beta_2) \\
&= (n-p_1)\sigma^2 + (X_2\beta_2)'(I-X_1(X_1'X_1)^{-1}X_1')(X_2\beta_2) \\
E[S^2_{(1)}] &= \frac{E[(Y-\widehat{Y})'(Y-\widehat{Y})]}{n-p_1} \\
&= \sigma^2 + \frac{(X_2\beta_2)'(I-X_1(X_1'X_1)^{-1}X_1')(X_2\beta_2)}{n-p_1}
\end{aligned}
$$

Our variance estimate is biased

Overfitting: we fit model 2 when model 1 is actually true

$$
\begin{aligned}
E[\widehat{\beta}_1^{(2)}] &= \beta_1 \\
B[\widehat{\beta}_1^{(2)}] &= 0
\end{aligned}
$$

Our coefficient estimate is unbiased

$$
\begin{aligned}
\mathrm{Var}(\widehat{\beta}^{(2)}) &= (X'X)^{-1}\sigma^2 \\
\mathrm{Var}(\widehat{\beta}_1^{(2)}) &= (X'X)^{-1}_{11}\sigma^2 \\
&\geq \mathrm{Var}(\widehat{\beta}_1^{(1)}) \\
VIF &= \frac{\mathrm{Var}(\widehat{\beta}_1^{(2)})}{\mathrm{Var}(\widehat{\beta}_1^{(1)})} \\
&= \frac{(X'X)^{-1}}{(X_1'X_1)^{-1}}
\end{aligned}
$$

$$
\begin{aligned}
E[(Y-\widehat{Y})'(Y-\widehat{Y})] &= E[Y'(I-X(X'X)^{-1}X'Y] \\
&= \mathrm{tr}(I-X(X'X)^{-1}X)\sigma^2 + E[Y'](I-X(X'X)^{-1}X')E[Y] \\
&= \mathrm{tr}(I-X(X'X)^{-1}X)\sigma^2 + \beta'X'(I-X(X'X)^{-1}X')X\beta \\
&= (n-p_1-p_2)\sigma^2 + \beta'X'(I-X(X'X)^{-1}X')X\beta \\
&= (n-p_1+p_2)\sigma^2 \\
E[S^2_{(2)}] &= \frac{E[(Y-\widehat{Y})'(Y-\widehat{Y})]}{n-p_1-p_2} \\
&= \sigma^2
\end{aligned}
$$

Our variance estimate is unbiased

$$
\begin{aligned}
\frac{(n-p_1-p_2)S^2_{(2)}}{\sigma^2} &\sim \chi^2_{n-p_1-p_2} \\
\mathrm{Var}\big(\frac{(n-p_1-p_2)S^2_{(2)}}{\sigma^2}\big) &= 2(n-p_1-p_2) \\
\mathrm{Var}(S^2_{(2)}) &= \frac{2\sigma^4}{n-p_1-p_2}
\end{aligned}
$$

In contrast, the variance estimate for model 1 has:

$$
\begin{aligned}
\frac{(n-p_1)S^2_{(1)}}{\sigma^2} &\sim \chi^2_{n-p_1} \\
\mathrm{Var}\big(\frac{(n-p_1)S^2_{(1)}}{\sigma^2}\big) &= 2(n-p_1) \\
\mathrm{Var}(S^2_{(1)}) &= \frac{2\sigma^4}{n-p_1}
\end{aligned}
$$

However the degrees of freedom is now reduced so that the variance estimate in the over fitted model has higher variance, and a confidence interval for the variance estimate will therefore be wider.

9 Bayesian linear models