MAST30025
Linear Statistical Modelling

Shromann Majumder

2021

Contents

1 Full Rank Model **2**

1.1 Parameters 2

1.2 Variance 2

1.3 Diagnostics 2

1.4 Maximum Likelihood Estimation 3

1.5 Confidence Interval 3

1.6 Prediction Interval 4

1.7 Joint Confidence Interval 4

1.8 Generalized Least Squares 4

1.9 Nonlinearities 4

2 Inference for the full rank model **5**

2.1 Hypothesis Testing & Relevance 5

2.2 Sequential Testing 6

2.2.1 Forward selection 6

2.2.2 Backward elemination 6

2.2.3 Stepwise selection 6

2.3 Shrinkage 7

2.3.1 Ridge Regression 7

2.3.2 LASSO 7

3 Less than Full Rank **8**

3.1 Classification 8

3.2 Conditional Inverse 8

3.3 Normal Equations 9

3.4 Estimability 9

3.5 σ^2 9

3.6 Interval Estimation 9

1 Full Rank Model

1.1 Parameters

$$\boxed{\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}} \tag{1}$$

▷ **Implications**

1. Full Rank Property:
$$r(X) = k + 1 \implies (X^T X)^{-1}$$

2. Assumption $\boldsymbol{\epsilon}$:
$$\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma^2 I)$$

3. Estimating $\boldsymbol{\beta}$:
$$\frac{\partial}{\partial \mathbf{b}} \sum_{i=1}^{n} e_i^2 = \frac{\partial}{\partial \mathbf{b}} (\mathbf{y} - X\mathbf{b})^T (\mathbf{y} - X\mathbf{b}) = \mathbf{0}$$
$$\implies \mathbf{b} = (X^T X)^{-1} X^T \mathbf{y} \qquad \textbf{BLUE}$$

   $E[\mathbf{b}] = \boldsymbol{\beta}$

   $\text{Var } \mathbf{b} = (X^T X)^{-1} \sigma^2$

4. Predicting **response** $\mathbf{y}^*$:
$$\mathbf{t}^T = \begin{bmatrix} 1 & x_1^* & x_2^* & \dots & x_k^* \end{bmatrix} \to \mathbf{t}^T \mathbf{b} \to \mathbf{y}^*$$

1.2 Variance

$$\boxed{\hat{\sigma}^2 = \frac{SS_{Res}}{n - p} = \frac{(\mathbf{y} - X\boldsymbol{\beta})^T (\mathbf{y} - X\boldsymbol{\beta})}{n - (k + 1)}} \tag{2}$$

▷ **Properties**

1. unbiased for σ^2:
$$E[\hat{\sigma}^2] = \sigma^2$$

2. Regression through the origin:
$$p = k$$

1.3 Diagnostics

$$\boxed{\textit{To assess the fit of Linear Models}} \tag{3}$$

▷ **Tools**

1. Variance of residuals:
$$\text{Var } \mathbf{e} = \text{Var } (I - H)\mathbf{y} = (I - H)\sigma^2 I (I - H)^T = \sigma^2 (I - H)$$

   $(I - H)$ is symmetric and idempotent

   To make them comparable, we standardize $\mathbf{e}$
$$z_i = \frac{e_i}{\sqrt{s^2 (1 - H_{ii})}}$$

   leverage: diagonal of H, measure of how much influence H_i has on overall fit of the model

2. Cooks Distance: measures the change in estimated parameters **b** if a point is removed. Minimal change $\iff$ Minimal influence,

$$D_i = \frac{1}{k+1} z_i \left(\frac{H_{ii}}{1 - H_{ii}} \right)$$

behavior: $D_i \propto z_i,\ H_{ii}$

usual metric: $D > 1 \to$ large, $D < 0.5 \to$ small

1.4 Maximum Likelihood Estimation

$$\boxed{\frac{\partial}{\partial \boldsymbol{\beta}} \ln L(\boldsymbol{\beta}, \sigma^2) = 0 \implies (X^T X)\boldsymbol{\beta} = X^T \mathbf{y} \equiv \text{normal equations!}} \tag{4}$$

▷ **Method & Properties**

1. $\log(x)$ transformation to derive easily:

$$\ln L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta})^T (y - X\boldsymbol{\beta})$$

2. for $y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma^2 I)$:

$$\mathbf{b} = (X^T X)^{-1} X^T \mathbf{y}$$
$$\tilde{\sigma}^2 = \frac{SS_{Res}}{n} \to \textit{Biased!}$$

$\tilde{\sigma}^2$ is only **asymptotically** *unbiased*, and has the same property as $\hat{\sigma}^2$

3. Sufficiency: *Fisher-Neyman Factorization Theorem*

$$\mathbf{y} = u(\mathbf{x}) \text{ is sufficient for } \boldsymbol{\theta} \iff f(\mathbf{x}; \boldsymbol{\theta}) = g(\mathbf{y}; \boldsymbol{\theta}) h(\mathbf{x})$$

Use all 'relevant' information about the parameters that is contained in the observed response variables.

$$\mathbf{b} = (X^T X)^{-1} X^T \mathbf{y} \text{ sufficient for } \boldsymbol{\beta}$$
$$s^2 = \frac{SS_{Res}}{n-p} \text{ sufficient for } \sigma^2$$

UMVUE *from MLE's asymptotic analysis*

1.5 Confidence Interval

· 100(1−α)% Confidence Interval

$$\boxed{b_i \pm t_{\alpha/2} s \sqrt{c_i i}} \tag{5}$$

$$\boxed{\mathbf{t}^T \mathbf{b} \pm t_{\alpha/2} s \sqrt{\mathbf{t}^T (X^T X)^{-1} \mathbf{t}}} \tag{6}$$

▷ **Definations**

1. c is the covariance matrix of **b**

2. Given a set of inputs, a 95% confidence interval for the response gives an interval that contains the expected response 95% of the time.

3. confidence interval for the error variance σ^2:

$$\left(\frac{(n-p)s^2}{\chi^2_{n-p}\left(1 - \frac{\alpha}{2}\right)}, \frac{(n-p)s^2}{\chi^2_{n-p}\left(\frac{\alpha}{2}\right)} \right)$$

1.6 Prediction Interval

$$\boxed{(x^*)^T\mathbf{b} \pm t_{\alpha/2} s\sqrt{1 + (\mathbf{x}^*)^T(X^TX)^{-1}(\mathbf{x}^*)}} \tag{7}$$

▷ **Definations**

1. given a set of inputs, a 95% prediction interval produces an interval in which we are 95% sure that any given response with those inputs lies in.

1.7 Joint Confidence Interval

$$\boxed{(\mathbf{b} - \boldsymbol{\beta})^T X^T X(\mathbf{b} - \boldsymbol{\beta}) \leq ps^2 f_\alpha} \tag{8}$$

▷ **Definations**

- $f \sim F_{p,n-p}$, the critical value

1.8 Generalized Least Squares

$$\boxed{\mathbf{b} = (X^TV^{-1}X)^{-1}X^TV^{-1}\mathbf{y}} \tag{9}$$

▷ **Definations**

1. $\mathbf{y} \sim MVN(X\boldsymbol{\beta}, V)$
2. $\boldsymbol{\epsilon} \sim N$ and var $\boldsymbol{\epsilon} = V$ is positive definite:

$$\mathbf{e}^TV^{-1}\mathbf{e} = (y - X\mathbf{b})^TV^{-1}(y - X\mathbf{b}) \implies X^TV^{-1}X\mathbf{b} = X^TV^{-1}\mathbf{y}$$

3. Gauss-Markov model still alive, i.e the following are **BLUE**

   $E[\mathbf{b}] = \boldsymbol{\beta}$

   Var $\mathbf{b} = (X^TV^{-1}X)^{-1}$

4. Var $\boldsymbol{\epsilon} = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2) = V$:

$$(y - X\mathbf{b})^TV^{-1}(y - X\mathbf{b}) = \sum_{i=1}^{n}\left(\frac{e_i}{\sigma_i}\right)^2$$

1.9 Nonlinearities

$$\boxed{\text{Apply } t : f_{non-linear} \rightarrow g_{linear}} \tag{10}$$

▷ **Applications**

1. *exponential*:

$$y = e^x \rightarrow \ln y = \ln e^x = x$$

2. *products*:

$$y = x_1x_1 \rightarrow \ln y = \ln x_1x_2 = \ln x_1 + \ln x_2$$

3. *Log transformation is a standard starting point due to its nature as shown above*

2 Inference for the full rank model

2.1 Hypothesis Testing & Relevance

$$\boxed{H_0 : \boldsymbol{\beta} = \mathbf{0} \text{ vs } H_1 : \boldsymbol{\beta} \neq \mathbf{0}} \tag{1}$$

▷ **Purpose, Assumption & Methods**

- The first thing we want to test is model relevance: does our model contribute anything at all?
- If none of the x variables have any relevance for predicting y, then all the parameters $\boldsymbol{\beta}$ will be $\mathbf{0}$.
- *Assumption*: $\boldsymbol{\epsilon} \sim MVN$
- *ANOVA*:

$$\begin{aligned} \boldsymbol{\beta} = \mathbf{0} &\implies \mathbf{y} = \boldsymbol{\epsilon} && \implies \mathbf{y}^T\mathbf{y} = \mathbf{e}^T\mathbf{e}\ (SS_{Res}) \\ \boldsymbol{\beta} \neq \mathbf{0} &\implies \mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} && \implies \mathbf{y}^T\mathbf{y} = (X\boldsymbol{\beta} + \boldsymbol{\epsilon})^T(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ & && [SS_{Total}]\ \mathbf{y}^T\mathbf{y} = \mathbf{y}^T X(X^TX)^{-1}X^T\mathbf{y}\ [SS_{Reg}] + \mathbf{e}^T\mathbf{e}\ [SS_{Res}] \end{aligned}$$

$$\begin{aligned} \boldsymbol{\beta} = \mathbf{0} &\iff SS_{Total} \to SS_{Res} \\ \boldsymbol{\beta} \neq \mathbf{0} &\iff SS_{Total} \to SS_{Reg} \end{aligned}$$

- *General Linear Hypothesis*:

$$H_0 : C\boldsymbol{\beta} = \boldsymbol{\delta}^* \text{ vs } H_1 : C\boldsymbol{\beta} \neq \boldsymbol{\delta}^*$$

$$\textit{example.}\ \ (r \times p)\ C = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix},\ \ (r \times 1)\ \boldsymbol{\delta}^* = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \implies \begin{cases} \beta_1 - \beta_2 = 0 \\ \beta_2 - \beta_3 = 0 \end{cases} \implies \beta_1 = \beta_2 = \beta_3 = 0$$

$$E(C\mathbf{b} - \boldsymbol{\delta}^*) = C\boldsymbol{\beta}\boldsymbol{\delta}^*$$

$$Var(C\mathbf{b} - \boldsymbol{\delta}^*) = C(X^TX)^{-1}C^T\sigma^2$$

$$\frac{(C\mathbf{b} - \boldsymbol{\delta}^*)^T[C(X^TX)^{-1}C^T]^{-1}C\mathbf{b} - \boldsymbol{\delta}^*)}{SS_{Res}/(n-p)} \sim F_{r,n-p} \text{ (Under } H_0)$$

One-tailed test: null is **false** (reject null) if the the below would be greater than σ^2, otherwise lower.

$$\frac{(C\mathbf{b} - \boldsymbol{\delta}^*)^T[C(X^TX)^{-1}C^T]^{-1}C\mathbf{b} - \boldsymbol{\delta}^*)}{r}$$

- *Splitting $\boldsymbol{\beta}$*:

$$\boldsymbol{\beta} \neq \mathbf{0} \not\Longrightarrow \forall b \in \boldsymbol{\beta} : b = 0$$

If $\beta_i = 0$ then best to **remove** β_i, thus is a way to test other parts of $\boldsymbol{\beta}$ are 0 or not.

$$\boldsymbol{\beta} = \begin{bmatrix} \gamma_1 \\ \gamma_1 \end{bmatrix},\ H_0 : \gamma_1 = \mathbf{0} \text{ vs. } H_1 : \gamma_1 \neq \mathbf{0}$$

we're essentially comparing H_1 in full model and H_0 in the reduced model.

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \text{ vs } \mathbf{y} = X_2\gamma_2 + \boldsymbol{\epsilon}_2 : X = [X_1|X_2]$$

$$\frac{R(\gamma_1|\gamma_2)/r}{MS_{Res}} = \frac{R(\boldsymbol{\beta}) - R(\gamma_2)/r}{MS_{Res}} = \frac{SS_{Reg}(\boldsymbol{\beta}) - SS_{Reg}(\gamma_2)/r}{MS_{Res}} \sim F_{r,n-p},$$

**Reject null if this is too big*

- *Corrected SS*: $\boldsymbol{\beta} = \mathbf{0}$ instead of $\exists \beta_i : \beta_i = 0$

$$\begin{aligned} Corrected\ ss &: \mathbf{y}^T\mathbf{y} - R(\gamma_2) \\ Correction\ Factor &: R(\gamma_2) \end{aligned}$$

Same ANOVA as 'splitting $\boldsymbol{\beta}$' but $r = k$ (number of βs)

2.2 Sequential Testing

$$\boxed{\textit{pick variables to avoid overfitting and reduce noise in the model}} \tag{2}$$

▷ **Purpose**

1. To use minimal number of variables, the most useful one, to avoid **overfitting** and dealing with noise.
2. We do this by seeing which variables add more variation:

$$R(\beta_k|\beta_0, \ldots, \beta_{k-1}) = R(\beta_k) - R(\beta_0, \ldots, \beta_{k-1})$$

2.2.1 Forward selection

1. Start with an empty model.
2. Calculate the F-values for the null hypothesis $H_0 : \beta_i = 0$, for all parameters not in the model, in the presence of parameters already in the model
3. If none of the tests are significant (we do not reject any null hypothesis), then stop
4. Otherwise, add the most significant parameter (the parameter with the largest F- value)
5. Return to step 2

2.2.2 Backward elemination

1. Start with the full model.
2. Calculate the F-values for the null hypothesis $H_0 : \beta_i = 0$, for all parameters in the model, in the presence of the other parameters in the model.
3. If all of the tests are significant (we reject the null hypothesis), then stop.
4. Otherwise, we remove the least significant parameter (the parameter with the smallest F -value).
5. Return to stop 2.

2.2.3 Stepwise selection

1. Start with any model.
2. Compute the AIC of all models which either have one extra variable or one less variable than the current model.
3. If the AIC of all such models is more than the AIC of the current model, stop.
4. Otherwise, change to the model with the lowest AIC.
5. Return to step 2.

▷ **Definations**

- Akaike's Information Criterion (measure of 'goodness-of-fit'):

$$AIC = -2\ln(L(\beta, \sigma^2)) + 2p = n\ln\left(\frac{SS_{Res}}{n}\right) + 2p + c$$

- *Advantage* of Stepwise: allowed to add and remove, unlike forward and backward
- *Advantage* of Forward and Backward: only $number(\beta)$ iterations vs $number(\beta)^2$, gives global optimum unlike AIC, which only gives local optimum
- *t-test*: for a partial test of one parameter. That is, to test $H_0 : \beta_i = 0$ against $H_1 : \beta_i \neq 0$ in the presence of all other parameters.

this part isn't emphasized much; probably come back add stuff after working through problems

2.3 Shrinkage

$$\boxed{\textit{`shrink' all fitted parameters toward 0, so that irrelevant variables have little or no effect on the model.}} \tag{3}$$

2.3.1 Ridge Regression

$$\boxed{\sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=1}^{n} b_j^2} \tag{4}$$

▷ **Properties**

1. λ: controls the amount of 'shrinkage' of the parameters.
2. *penalized least squares estimators*:
$$\mathbf{b} = (X^T X + \lambda I)^{-1} X^T \mathbf{y}$$
3. never shrinks parameters to 0

2.3.2 LASSO

$$\boxed{\sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=1}^{n} |b_j|} \tag{5}$$

▷ **Properties**

1. The LASSO actually shrinks small parameters to 0, and can be used for variable selection by removing those variables.
2. *cross-validation*, choosing λ: estimates the predictive power of the model by removing parts of the dataset and using them as test sets.

3 Less than Full Rank

3.1 Classification

$$\boxed{\mathbf{b} = (X^T X)^c X^T \mathbf{y}} \tag{1}$$

▷ **Implications**

1. samples come from k distinct populations, and we wish to determine the differences between these populations.
2. *One-Way Classification Model*:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{21} \\ y_{22} \\ y_{k,n_k} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \dots \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_k \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{21} \\ \vdots \\ \epsilon_{k,n_k} \end{bmatrix}$$

   μ is the overall mean, and the first column of X is the sum of the remaining columns (not full rank).

3. *Reparametrization*: One way to address the issue stated above is to convert the less than full rank model to a full rank model.

   Theres *Differences in Populations* and *General Pooled Variance*, that is not included here

3.2 Conditional Inverse

$$\boxed{AA^cA = A} \tag{2}$$

▷ **Properties**

1. *Infinite* number of Conditional Inverses, they are just linear combinations of each other.
2. *Rank*:
$$r(A) = r(AA^c) = r(A^cA)$$
3. *Equalities*:
$$A = A(A^TA)^c(A^TA)$$
$$A^T = (A^TA)(A^TA)^cA^T$$
4. *Unique Expressions*: an expression involving a conditional inverse is unique if it is the same no matter what conditional inverse we use.
5. *Finding a Conditional Inverse*:
   1. Find a minor M of A which is non-singular and of dimension $r(A) \times r(A)$
   2. Replace M in A with $(M^{-1})^T$ and everything else with zero.
   3. Transpose the resulting matrix.

3.3 Normal Equations

$$\boxed{X^T X\mathbf{b} = X^T\mathbf{y} \implies \mathbf{b} = (X^T X)^c X^T\mathbf{y}} \tag{3}$$

▷ **Property**

- We can use the Theorems learnt in class to find a all solutions for b:

$$\mathbf{b} = (X^T X)^c X^T\mathbf{y} + [I - (X^T X)^c X^T X]\mathbf{z}, \quad \mathbf{z} \text{ is any vector}$$

- *Theorem 6.5*: Let $A\mathbf{x} = \mathbf{g}$ be a consistent system. Then $A^c g$ is a solution to the system, where A^c is any conditional inverse for A.

$$AA^c\mathbf{g} = AA^cA\mathbf{x}^* = A\mathbf{x}^* = \mathbf{g}.$$

3.4 Estimability

$$\boxed{\text{quantities that not change regardless of the solutions we use for the normal equations}} \tag{4}$$

▷ **Defination**

1. $\mathbf{t}^T\boldsymbol{\beta}$ is estimable if:

$$E[c^T\boldsymbol{\beta}] = \mathbf{t}^T\boldsymbol{\beta}$$

2. In the general linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\mathbf{t}^T\boldsymbol{\beta}$ is estimable $\iff$ $\mathbf{t}^T(X^T X)^c X^T X = \mathbf{t}^T$ (very **important**)
3. Take the *Transpose* of the above to get, $X^T X(X^T X)^c\mathbf{t} = \mathbf{t}$, so $\mathbf{z} = (X^T X)^c\mathbf{t}$ solves $X^T X\mathbf{x} = \mathbf{t}$
4. **BLUE** for $\mathbf{t}^T\boldsymbol{\beta}$ is:

$$\mathbf{z}^T X^T\mathbf{y}$$

where $\mathbf{z}$ is the solution to $X^T X\mathbf{z} = \mathbf{t}$. also the choice of $\mathbf{z}$ doesn't matter, can be any solution.

3.5 σ^2

$$\boxed{\hat{\sigma}^2 = \frac{SS_{Res}}{n-r}} \tag{5}$$

3.6 Interval Estimation

$$\boxed{\frac{(n-r)s^2}{\sigma^2} = \frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-r}} \tag{6}$$

▷ **Property**

- If $\mathbf{t}^T\boldsymbol{\beta}$ is estimable, then $\mathbf{t}^T\boldsymbol{\beta}$ is independent of s^2.
- The steps to derive a confidence interval are very similar to that for the full rank case, but with two small differences. Firstly, we can only find confidence intervals for quantities that are estimable!

$$\mathbf{t}^T\mathbf{b} \pm t_{\alpha/2}s\sqrt{\mathbf{t}^T(X^T X)^c\mathbf{t}}$$

- Secondly, we replace the inverse $(X^T X)^{-1}$ by the conditional inverse $(X^T X)^c$
- All other steps are the same