Statistical Modelling

Cheat Sheet · v1.0 · 2021

Mael Macuglia

D-MATH
ETH Zürich

1 Classical Linear Model

Model: Let $\mathcal{Y} := (Y_1, ..., Y_n)$ ind. copies of Y $\mathcal{Y} = f + \epsilon$ with $E\mathcal{Y} = f, f \in \mathcal{F} \subseteq \mathbb{R}^n$
Linear Model: $\mathcal{F} := \{f_i = a + \sum_{j=1}^{p} x_{ij}\beta_j, a \in \mathbb{R}, b \in \mathbb{R}^p\}$
Goal : find an estimate for the systematic component $\hat{E\mathcal{Y}} = f = X\hat{\beta}$

1.1 Modelling Effect of Covariates

(i) **Continous Covariates**: z has non-linear effect $\beta_1 f(z)$ with known f: Linear model $\rightarrow y_i = \beta_0 + \beta_1(xi)$ where $x_i := f(z_i) - \bar{f}$. z has approximatly polynomial effect $\beta_0 + \beta_1 z + ... + \beta_l z^l \rightarrow x_{ij} = z_i^j$ (apply centering) (use partial residuals to check the effect of modelling)
(ii) **Categorical Covariates** Let the covariate $x_i \in \{1, ..., c\}$ then apply Dummy Encoding: $x_{i1} = \mathbb{1}(x_i = 1), ..., x_{ic-1} = \mathbb{1}(x_{c-1} = 1)$ where the category omitted(needed for uniqueness of LS) is called the reference category
(iii) **Interactions** catcat/catcont: $y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \epsilon$ Think: x, z not need to be correlated has interaction has effect on y | Think: β_1 effect of having A β_2 = effect of having B β_3 = additional effect of having A and B | Think interaction when effect of cat. cov. do not translate.

1.2 Parameter Estimation β

(i) **Assumptions** Homoscedasticity and Uncorrelated errors: $cov(\epsilon) = \sigma^2 I_{nxn} \rightarrow$ same precision for all measurements| $rank(X) = p$.
(ii) **RSS** $:= \|Y - X\beta\|^2$
(iii) **minLS**: $\hat{\beta} = min\|Y - X\beta\|^2 = min_\beta \epsilon^T \epsilon$
(iv) **LSE**: $\hat{\beta} = (X^T X)^{-1} X^T Y$
(v) **Prediction** $\hat{EY} = \hat{Y} = X\hat{\beta} = HY = \mathcal{YPA}$
(vi) **Hat Matrix** $SVD(X) = P\Phi V^T \implies H = PP^T$ Trace: $tr(H) = p$ | Anti projection: $Q = (1 - H)$ Trace : $tr(Q) = n - p$ | both idempotent and symmetric
(vii) **Residuals** $\hat{\epsilon} = Y - X\hat{\beta} = QY = \mathcal{YA}\mathcal{X}$

1.3 Parameter Estimation σ

(i) **MLE**: $\hat{MLE(\sigma^2, \beta = \beta_{LS})} = \hat{\sigma^2_{MLE}} = \frac{1}{n}\|\hat{\epsilon}\|^2$ (biased)
(ii) **Unbiased Estimator**: $E\|\hat{\epsilon}\|^2 = (n-p)\sigma^2 \implies \hat{\sigma}^2 = \frac{1}{n-p}\|\hat{\epsilon}\|^2$ (unbiased)
(iii) **Residual Standard Error** $RSE := \hat{\sigma} = \sqrt{\frac{1}{n-p}\|\hat{\epsilon}\|^2}$

1.4 Geometric Properties

(i) **Normal Eq:** $\mu = EY = X\beta \in$ Span$(1, x_1, ..., x_{p-1})$ = p-dim Vector Space Minimize vector of residuals: $X^T(Y - X\hat{\beta}) = 0$
(ii) **Orthonormal Design:** In the regression of y onto $(x_0, x_1, ..., x_k)$ the coefficient of x_k is $\hat{\beta_k} = \frac{<z_k, y>}{<z_k, z_k>}$ where z_k is the residual from regressing x_k onto $(x_0, .., x_{k-1}) \implies z_k^T x_j = 0 \forall j \in \{0, 1, ..., k-1\}$
(iii) **Interpretation**: The multiple regression coeff. $\hat{\beta_k}$ represents the additional contribution of x_k on Y after x_k has been adjusted for $x_0, ..., x_{k-1}$

1.5 Analysis of Variance

(i) **Sample Variance**: $S^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
(ii) **Sample Pred. Var.**: $S^2_{\hat{Y}} = \frac{1}{n}\|\hat{y} - \bar{y}\|^2$
(iii) **Variance Decomp.** : $\|y - \bar{y}\|^2 = \|y - \hat{y} + \hat{y} - \bar{y}\|^2 = \|y - \hat{y}\|^2 + \|\hat{y} - \bar{y}\|^2 + 2 < y - \hat{y}, \hat{y} - \bar{y} >$ where we note that in case of LS the last scalar product is = 0
(iv) $SS_{tot} = SS_{reg} + SS_{res}$
(v) Residual variance is the variance we couldn't explain
(vi) **R Coeff**: $R^2 = \frac{SS_{reg}}{SS_{tot}} = \frac{\|\hat{y} - \bar{y}\|^2}{\|y - \bar{y}\|^2} = 1 - \frac{SS_{res}}{SS_{tot}}$
(vii) **Interp.**: R^2 proportion of variance explained by the model

Model Comparison using R^2 only meaningful if following 3 statements are fulfilled:

1. Every model has the same response y (not transform)
2. Every model has the same dimension of parameter space (nb of p param.)
3. Every model has an intercept

Note: By LLN the sample variance converges in probability to the variance as $n \rightarrow \infty$

1.6 Statistical Properties

(i) $E\hat{\beta} = \beta$ and $cov(\hat{\beta}) = \sigma^2(X^T X)^{-1}$
(ii) $Var(\hat{\beta}_j) = \frac{\sigma^2}{<z_j, z_j>} \rightarrow$ correlation between the covariates inflate the variance.
(iii) **Est. Var.** $\hat{cov}(\hat{\beta}) = \hat{\sigma}^2(X^T X)^{-1} | \hat{\sigma}^2 = \frac{1}{n-p}\epsilon^T \epsilon$
(iv) **Estimated Standard Error:** (Std.error in R output) $se_j = \hat{(cov}(\hat{\beta}))_{jj} = \hat{Var}(\hat{\beta}_j)^{\frac{1}{2}}$
(v) If $\epsilon \rightarrow N(0, \sigma^2 I)$ then $Y \rightarrow N(X\beta, \sigma^2) | \hat{\beta} \rightarrow N(\beta, \sigma^2(X^T X)^{-1})$

Theorem 1: Gauss Markov

$\mathbf{Y} = \mathbf{X}\beta + \epsilon \quad \mathbf{E}[\epsilon] = 0 \quad Cov(\epsilon) = \sigma^2 I \quad rank[X] = p$

Furthermore let $\mathbf{c} \in \mathcal{R}^p$ and $\hat{\beta}$ the LS estimator. Then $\mathbf{c}^T\hat{\beta}$ has minimal variance among all linear unbiased estimators of $\mathbf{c}^T\beta \rightarrow$ BLUE: Best Linear Unbiased Estimator.
Assume: ϵ has $N(0, \sigma^2 I)$. Then $\mathbf{c}^T\hat{\beta}$ has minimal variance among all unbiased estimator $\rightarrow$ UMVU

1.7 Asymptotic Properties of LSE

$\mathbf{Y_n} = \mathbf{X_n}\beta + \epsilon_\mathbf{n}$ $n :=$ nb of observations Assumptions:

(i) $(X_n^T X_n)^{-1} \rightarrow 0$ as $n \rightarrow \infty$
(ii) $max_j H_{jj} = max_j x_j (X^T X)^{-1} x_j^T \rightarrow 0$ as $n \rightarrow \infty$

Under Assumptions:

(i) $\hat{\beta_n}_{LS} \xrightarrow{p} \beta | \hat{\sigma_n}^2, \hat{\sigma_n}^2_{ML} \xrightarrow{p} \sigma^2$ (Consistency)
(ii) $(X_n^T X_n)^{\frac{1}{2}}(\hat{\beta} - \beta) \xrightarrow{D} N(0, \sigma^2 I)$

1.8 Statistical Properties of Residuals

(i) $\mathbf{E}\hat{\epsilon} = 0$
(ii) $cov(\hat{\epsilon}) = \sigma^2(I - H)$
(iii) $Var(\hat{\epsilon}_j) = \sigma^2(1 - H_{ii})$ (Heteroscedasticity)
(iv) $cov(\hat{\epsilon}_j, \hat{\epsilon}_i) \neq 0$ (Correlated)
(v) $\hat{epsilon} \rightarrow N(0, \sigma^2 Q)$
(vi) Res.Sum.Square : $\frac{\hat{\epsilon}^T\hat{\epsilon}}{\sigma^2} \rightarrow \mathcal{X}^2_{n-p}$
(vii) $\hat{\epsilon}^T\hat{\epsilon}$ Independant of $\hat{\beta}_{LS}$

2 Hypotheses Testing

2.1 Exact F test

Basics of testing: level α: $P(\phi(y, x, \beta_0) = 1) \leq \alpha$ Def P-value: Prob. to get a value as extrem as at least as extreme as result actually observed during the test, assuming that H_0 is correct

(i) **General linear Hypotheses:** $H_0 : \mathbf{C}\beta = \mathbf{d}, rank(\mathbf{C}) = r \leq k$
(ii) **Assumptions:** Gaussians Errors
(iii) **Pivot Statistic :** $F = \frac{n-p}{r}\frac{\Delta SSE}{SSE} \sim F_{r,n-p}$
(iv) Computation of Pivot: $F = \frac{1}{r}(C\hat{\beta} - d)'(\hat{\sigma}C(X'X)^{-1}C')^{-1}(C\hat{\beta} - d)$
(v) **Reject** H_0 **:** $F > F_{r,n-p}(1 - \alpha)$
(vi) $\Delta SSE = SSE_0 - SSE$ to compute $SSE_0 \rightarrow$ constrained Least square
(vii) ΔSSE get smaller the less the error the more likely to accept H_0

2.2 F-test for Specific Problems

(i) **Test of significance:** $H_0 : \hat{\beta}_j = 0$ (given all other predictors)
(ii) Pivot: $t_j = \frac{\hat{\beta}_j}{\hat{var}(\hat{\beta}_j)^{\frac{1}{2}}} = \frac{\hat{\beta}_j}{\hat{\sigma}}\|z_j\|^2 \sim t_{n-p}$
(iii) Reject H_0 (significance): $|t_j| > t_{n-p}(1 - \frac{\alpha}{2})$
(iv) **Composite test:** $\beta_{sub} = (\beta_1, ..., \beta_r)^T | H_0 : \beta_{sub} = 0 | H1 : \exists(\beta_{sub})_i \neq 0$
(v) Pivot: $F = \frac{1}{r}\hat{\beta}^T_{sub}\hat{cov}(\hat{\beta_{sub}})\hat{\beta}_{sub} \sim F_{r,n-p}$
(vi) Note: $\hat{cov}(\hat{\beta_{sub}})$ are sub element of full model $\hat{\sigma}^2(X^T X)^{-1}$
(vii) **Global F-test:** $H_0 : \beta_{full} = 0, \beta_{full} \in \mathcal{R}^k | H_1 \exists \beta_j \neq 0$
(viii) Pivot: $F = \frac{n-p}{k}\frac{R^2}{1-R^2}$
(ix) Under $H_0 : \beta_0 = \bar{y}$

2.3 Confidance Regions

(i) **CI** $H_0 : \beta_j = d_j$ with Pivot $t_j = \frac{\hat{\beta}_j - \beta_j}{se_j}$
(ii) $(1 - \alpha)CI = [\hat{\beta}_j + -se_j * t_{n-p}(1 - \frac{\alpha}{2})]$
(iii) **C El.** $\{\beta : \frac{1}{r}(\hat{\beta} - \beta)'\hat{cov}(\hat{\beta})(\hat{\beta} - \beta) \leq F_{r,n-p}\}$
(iv) **CI for** $Ey_0 = \mu_0$: $y_0 =$ future observation at location x_0
(v) $[x_0'\hat{\beta} + -t_{n-p}(1 - \frac{\alpha}{2})\hat{\sigma}(x_0'(X'X)^{-1}x_0)^{\frac{1}{2}}]$
(vi) **Prediction Interval:** we want an interval for future value y_0
(vii) $[x_0'\hat{\beta} + -t_{n-p}(1 - \frac{\alpha}{2})\hat{\sigma}(1 + x_0'(X'X)^{-1}x_0)^{\frac{1}{2}}]$
(viii) Note: Pred.interval bigger than Ci for mean

3 Multiple Testing

3.1 Problem Formulation

- **Hypotheses:** $H_0 = H_0^{(1)} \cap ... \cap H_0^{(m)}$
- **Discovery:** reject H_0 (at least accept one alternative)
- False Discovery $:= TypeI - Error$
- **P-Value**: $p := \mathbb{P}_{H_0}(T(X) > T(x)) \in [0, 1]$ for $\phi = \mathbb{1}_{(T(x) \notin [q_l, q_r])}$
- P-value $\sim Uni[0, 1]$
- **Proced.:** given $\{p^i\}_{i \in \{1,..,m\}} \xrightarrow{Procedure} \{H_0, H_A\}$
- **Level:** $\mathbb{P}_{H_0}(falsedisc.) = \mathbb{P}_{H_0}(p \leq \alpha) \leq \alpha$
- Naive approach: test separately $\mathbb{P}_{H_0}(\phi^i = 1; \exists i \in \{1, ..., m\}) = 1 - (1 - \alpha)^m$ as m grows last eq. goes to 1 $\implies$ get false discovery by chance.

3.2 Bonferroni Holms

- **Metrics** $FWER = \mathbb{P}(V \geq 1)$ with $V :=$nb false discovery
- **Procedure BF:** Reject $H_0^{(i)}$ if $p^{(i)} \leq \frac{\alpha}{m}$
- **Thm:** Bonferroni controls FWER with level α: $\mathbb{P}(V \geq 1) \leq \alpha$
- **Procedure BF Holms :** given ordered $\{p^(1) \leq ... \leq p^{(m)}\}$
- **Step1:** Reject if $p^{(1)} \leq \frac{\alpha}{m}$ else accept $\{H_0^(1)...H_0^(m)\}$
- **Step2:** Reject if $p^{(2)} \leq \frac{\alpha}{m+1}$ else accept $\{H_0^(2)...H_0^(m)\}$
- **Stepi:** Reject if $p^{(i)} \leq \frac{\alpha}{m-i+1}$ else accept $\{H_0^(i)...H_0^(m)\}$
- **Thm:** bf Holmes controls FWER at level α and is UMP for simple Bf procedure

3.3 Permutation Test

- **Idea:** non parametric test (not knowing about the dist.)
- (1)Sample $S_{xy} = \{x_1, ..., x_n, y_1, ..., y_m\}$
- (2)Randomly permuteS_{xy} and split into $\tilde{S}_x, \tilde{S}_y$
- (3)Compute statistic $T(\tilde{S}_x, \tilde{S}_y)$
- repeat B times (1,2,3): $\{T^{\tilde{(1)}}, ..., T^{\tilde{(B)}}\}$
- $\hat{p}_{val} = \frac{\sum_{l=1}^{B} \mathbb{1}(|T^{\tilde{(l)}}| \geq |T(S_x, S_y)|)+1}{B+1}$
- $T(S_x, S_y) :=$ observed statistic
- Decision: Reject if $\hat{p}_{val} \leq \alpha$

3.4 Benjamini Hochberg

- **Idea:** Controling FWER is very restrictive, hence low power. Gain in power by expecting $(1-FDR)$ of the discoveries to be true.
- $FDP = \frac{V}{|R|} = \frac{|H_0 \cap R|}{|R|}$
- $FDR := E[FDP]$
- **Procedure:** Given ordered pval $\{p^{(1)} \leq ... \leq p^{(m)}\}$
- $max_k(p^{(k)} \leq \frac{\alpha k}{m}) \rightarrow R := \{p^{(1)}, ..., p^{(k)}\}$
- **Thm:** if pval are independant then BH_α controls FDR at level α

4 Model Selection

4.1 Model Specification

- **Missing Varibles:** $\widehat{\beta}^M$ is biased | $Var(\widehat{\beta}^M) \leq Var(\beta_{true})$
- **Irrelevant Varibales:** $\widehat{\beta}^M$ is biased | $Var(\widehat{\beta}^M) \geq Var(\beta_{true})$

4.2 Metrics

- **SMSE:** sum of mean squared errors $\widehat{Y}_i^M$ from true value μ_i
- $SMSE(M) = E[\|\widehat{Y}^M - \mu\|^2] = |M|\sigma^2 + \sum_i^n (E[\widehat{Y}_i^M] - \mu_i)^2 = Var + bias^2$
- $\Gamma_p^M = \frac{SMSE}{\sigma^2}$
- **SPSE:** sum of prediction errors; $\widehat{Y}_i^M$ pred. of new variable $Y_{n+i} = \mu_i + \epsilon_{n+i}$
- $SPSE(M) = \sum_i^n E[(Y_{n+i} - \widehat{Y}_i^M)^2] = n\sigma^2 + SMSE(M)$
- **SSE:** sum of squared errors
- $SSE(M) = E[\|Y - \widehat{Y}^M\|^2] = SPSE(M) - 2|M|\sigma^2$
- Note: for all metrics μ_i, σ^2 are unkown $\rightarrow$ need estimates
- Note: Miniminzing one of the above metrics results in minimizing all of them

4.3 Model Choice Criteria

Overall Goal: select covariates that minimize the expected SPSE

- **Approach 1:** Split data into $\{train, validation\}$ Model on train and $\widehat{SPSE} = E[\|Y_{val} - \widehat{Y_{train}}^M\|^2]$ with validation
- **Approach 2:** not enough data for appr.1 compute $\widehat{\beta}^M$ with all data
- $\widehat{SPSE}(M) = \sum_i^n (y_i - \widehat{Y}_i^M)^2 + 2|M|\widehat{\sigma}^2$ ($\widehat{\sigma}^2$ based on full model)
- **Mallow's CP:** $\widehat{\Gamma}^M = C_p = \frac{\sum_i^n (y_i - \widehat{Y}_i^M)^2}{\widehat{\sigma}^2} - n + 2|M|$
- **AIC:** $AIC = -2l(\widehat{\beta}^M, \widehat{\sigma}^2_{ML}) + 2(|M|+1)$ wish to minimize AIC
- **BIC:** $BIC = -2l(\widehat{\beta}^M, \widehat{\sigma}^2_{ML}) + log(n)(|M|+1)$
- Note: BIC stronger penalization for complex model
- Better fit if small
- **Adj. R coef** $\bar{R}^2 = 1 - \frac{n-1}{n-p}(1-R^2)$
- **Cross Validation:** $CV_{score} = \sum_j^r \widehat{SPSE}_j^{-j}$ for data splitted in r chunk
- Special case Leave one out: $CV_{score} = \frac{1}{n}\sum_i^n (Y_i - \widehat{Y}_i^{-i,M})^2 = \frac{1}{n}\sum_i^n \frac{Y_i - \widehat{Y}_i^M}{1-H_{ii}}$
- big computational cost avoided as one only need to compute H^M once for each M and not one for all CV chunks

4.4 Model Selection Procedures

- Forward selection
- Backward selection
- Stepwise selection
- Best subset selection

4.5 Inference after Model Selection

- After model sel. the data has already been used $\Longrightarrow$ Statistical Inference rules are broken
- Solution: Randomly split data into 2 parts, apply model selection on first and use second for statistical inference (like CI,PI,etc...)

5 Model Diagnostic

5.1 Model Assumptions

- **Zero mean Errors**: ($\hat{\beta}$ biased | test,CI not valid)
- Transformations, included ommited covariates
- **Homo. Errors:** test and CI
- Transformation of response, general linear model
- **Uncorrolated Errors:** test and CI
- General linear model
- **Normal Ass.** test and CI only asymp. correct

5.2 Residual Plots

- **Stand. Residuals:** $r_i = \frac{\widehat{\epsilon}_i}{\widehat{\sigma}\sqrt{1-H_{ii}}}$
- **Stundentized Res:** $\tilde{r}_i = \frac{\widehat{\epsilon}_{(i)}}{\widehat{\sigma}^{(-i)}\sqrt{1-H_{ii}}} \sim t_{n-p-1}$
- design matrix wth i row $X^{-i} \rightarrow \widehat{\beta}^{(-i)} \rightarrow \widehat{\sigma}^{(-i)}$ and $\widehat{\epsilon_{(i)}} = y_i - x_i'\widehat{\beta}^{(-i)}$
- **Tuckey Anscombe:** Plot $\widehat{\epsilon}_i$ VS $\widehat{y}_i \rightarrow$ check zero means. Even if residuals not uncorr still good visual approx.
- fluctuations $\rightarrow$ non-linearity | omitted covariate
- **Stand Res VS Covariates:** if rdn spread around zero line model explained well effects of covariates on mean
- **Scale Location Plot :** $\sqrt{|r_i|}$ VS y_i to check Homosc. (if well spreaded around zero line)

5.3 Transformations

- **Goal:** Achieve mean function approx linear in the tranformed scale
- **Power Fam:**

$$\Psi(u, \lambda) = \begin{cases} u^\lambda & \lambda \neq 0 \\ log(u) & \lambda = 0 \end{cases}$$

- $\lambda \in \{-1, 0, 1/3, 1/3, 1\}$
- **Scaled Power Fam:** (transforming only covariates)

$$\Psi(x, \lambda) = \begin{cases} x^\lambda & \lambda \neq 0 \\ log(x) & \lambda = 0 \end{cases}$$

- Choose λ visually, ie model expression best linear relationship (ex: distr. not scewed) AND minimizing the SSE (res sum square)
- **Box Cox Procedure**

5.4 Collinearity Analysis

- **Goal:** Check wether covariates have lin dep. (correlation)
- **Variance Inflation Factor:** $VIF_j := \frac{1}{1-R_j}$
- Qtf. $var(\widehat{\beta}_j)|$ $R_j :=$ R coeff of $X_j \sim X^{(} - j)$
- Treshold: (> 10)

5.5 Outliers

- **Stud. Res:** $\tilde{r}_i \sim t_{n-p-1}|$ Test all $\tilde{r}_i$ simultaneously (BF correction)
- $\tilde{r}_i > q(\alpha/2)_{t_{n-p-1}} \rightarrow$ outlier
- **Leverage:** "Outlier in x-direction" $lev := H_{ii} \in [\frac{1}{n}, 1]$
- Influential data if : $var(\widehat{\epsilon}_i) = \sigma^2(1-H_{ii})$ if $H_{ii} \rightarrow 1$ (big) then Hyper plane passes through $(x_i, y_i) \rightarrow$ influential data .
- Rule of Thumb: $H_{ii} > \frac{2p}{n}$
- **Cook's Distance:** $D_i = \frac{\|\widehat{\mathbf{y}}^{(-i)} - \widehat{\mathbf{y}}\|^2}{p\widehat{\sigma}^2} = \frac{1}{p\widehat{\sigma}^2}\frac{\widehat{\epsilon}^{(-i)2}}{1-H_{ii}}\frac{H_{ii}}{1-H_{ii}}$
- Euclidean dist bwt Hyper plane Full$\widehat{y}$ and omitted i $\widehat{y}^{(} - i)$
- Check eq. $\rightarrow$ plot residuals vs leverage
- If $D_i > 0.5$

6 General Linear Model

6.1 Weighted Least Square

- **Gener. Model:** $Y = X\beta + \epsilon | E[\epsilon] = 0 | cov(\epsilon) = \sigma^2 W^{-1}$
- W assumed to be Pos. Definite.
- $W = P^T DP, \exists Bs.t W = BB^T$ and $W^{1/2}W^{1/2} = W$
- **OLS gener. Model:** $\widehat{\beta} = (X^T X)^{-1} X^T Y$
- $E[\widehat{\beta}] = 0$ but $cov(\widehat{\beta}) = \sigma^2 (X'X)^{-1} X'W^{-1} X (X'X)^{-1}$
- (CI and Test no longer valid) and GaussMarkov ass. brocken $cov(\epsilon) = \sigma^2 I$
- Hence OLS no longer UMVU
- **WLS :** $W^{1/2}Y = W^{1/2}X\beta + W^{1/2}\epsilon = \tilde{Y} = \tilde{X}\beta + \tilde{\epsilon}$
- $\widehat{\beta}_{WLS} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y} = (X'WX)^{-1}X'WY$
- $cov(\tilde{\epsilon}) = \sigma^2 I \rightarrow$ GaussMarkov $\widehat{\beta}_{WLS}$ is UMVU
- REML: $\widehat{\sigma}^2 = \frac{1}{n-p}\widehat{\epsilon}' W \widehat{\epsilon}$ and $\widehat{\epsilon} = Y - X\widehat{\beta}_{WLS}$
- if $W = diag(w_1, ..., w_n)$ then $WLS(\beta) = (Y - X\beta)'W(Y - X\beta) = \sum_i^n w_i (y_i - x_i'\beta)^2$
- $w_i \propto \frac{1}{var(\epsilon_i)}$ weights bigger impact on objective when var is small
- **Grouped Data (known Heter.):** $y = (\bar{y}_1, ..., \bar{y}_G)'$ hence $\epsilon = (\bar{\epsilon}_1, ..., \bar{\epsilon}_G)'$
- $cov(\bar{\epsilon}_i) = \frac{\sigma^2}{n_i}$
- $cov(\epsilon) = \sigma^2 W^{-1} \rightarrow W = diag\{\frac{1}{n_1}, ..., \frac{1}{n_G}\}$
- **Unknown Heter.** solution: transformations $\sqrt{y}, log(y)$
- **Solution 2 Stage LS:**
- $var(\epsilon_i) = E[\epsilon_i^2] = \sigma_i^2 \overset{lin.model}{\rightarrow} \epsilon_i^2 = \sigma_i^2 + \nu_i = z_i'\alpha + \nu_i$
- Note: $z_i :=$ cov. that affect the errors (usually X)
- errors unknown $\rightarrow \widehat{\epsilon}_i^2 \approx z_i'\alpha + \nu_i$
- $\widehat{E}[\widehat{\epsilon_i^2}] \approx \widehat{E}[\epsilon_i^2] = z_i'\widehat{\alpha}_{LS} \rightarrow \sigma_i^2 \approx z_i'\widehat{\alpha}_{LS}$
- WLS: $w_i := \frac{1}{z_i'\widehat{\alpha}_{LS}}$
- **White Esti.** estimation of $cov(\widehat{\beta})$ used for correction of stand. errors CI and tests for asymptotics
- $\widehat{cov}(\widehat{\beta}) = (X'X)^{-1}X'diag\{\widehat{\epsilon}_1^2, ..., \widehat{\epsilon}_n^2\}X(X'X)^{-1}$

7 Robust Regression

8 Generalized Linear Models

8.1 Binary Regression

- $y_i \in \{0,1\}$ and $y_i \sim Ber(\pi)$
- $E[y_i] = P(y_i = 1) = \pi_i$ and $var[y_i] = \pi_i(1-\pi_i)$
- **Response Func:** $\pi_i = h(\eta_i) \in [0,1]$
- **Link Func:** $h^{-1}(\pi_i) = g(\pi_i) = \eta_i = x_i'\beta$
- **Logit Model:** $log(\frac{\pi_i}{1-\pi_i}) = x_i'\beta \Leftrightarrow \pi_i = \frac{exp(\eta_i)}{1-exp(\eta_i)}$
- **Probit Model:** $\Phi^{-1}(\pi) = x'\beta \Leftrightarrow \pi = \Phi(\eta)$
- **Latent Variable:** $y_i = \mathbb{1}(z_i \geq 0)$ (observed) $z_i = x_i'\beta - \epsilon_i \rightarrow E[y_i] = P(y_i = 1) = P(z_i \geq 0) = P(\epsilon_i < x_i'\beta) = F(x_i'\beta)$
- if $\epsilon_i \sim N(0,\sigma) \rightarrow F = \Phi$ Probit model
- if $\epsilon_i \sim logistic \rightarrow F =$ logit model
- Note: Logistic variance $= \pi^2/3 \rightarrow \tilde{\beta} = \frac{\pi}{\sqrt{3}}\beta \rightarrow \pi(\eta) = h(x'\beta) = \Phi(x'\tilde{\beta})$
- resulting prob. are $\approx$ equal for adjusted model
- Note: to compare $\widehat{\beta}_{logit} vs \widehat{\beta}_{probit} \rightarrow$ scale: $\widehat{\beta}_{probit}\frac{\pi}{\sqrt{3}}$
- **Interpretation:** log ratio $\frac{P(y_i=1)}{P(y_i=0)} = exp(\beta_0)exp(x_1\beta_1)...$ (ex: if risk covariate categorical then see the effect of risk present $\frac{P(y_i=1)}{P(y_i=0)} = exp(1\widehat{\beta}_{risk}) * ...$ vs $= exp(0\widehat{\beta}_{risk}) * ...$
- **Grouped Data:** $\{n_i|\bar{y}_i = \frac{1}{n_i}\sum_i^{n_i} y_i|x_i\}, i \in \{1,..,G\}$ data grouped for identical covariates in G groups
- $y_i \sim Ber(\pi_i) \rightarrow \sum y_i \sim Bin(n_i, \pi_i)$
- $P_{\pi_i}(n_i\bar{y}_i) = P_{\pi_i}(\sum_{j=1}^{n_i} y_j) = (n_i, \sum y_j)\pi_i^{\sum_j y_j}(1-\pi_i)^{n_i - \sum_j y_j}$
- $E[\bar{y}_i] = \pi_i \rightarrow$ logit probit model
- $var(\bar{y}_i) = \frac{\pi(1-\pi_i)}{n_i}$
- **Overdispersion:** unobserved Heterogeneity or positive correlation (ex: when people sample comes from same cluster)
- Emprirical Var: $s = \frac{\bar{y}_i(1-\bar{y}_i)}{n_i} > \frac{\widehat{\pi}_i(1-\widehat{\pi}_i)}{n_i} \rightarrow$ Overdispersion
- Solution: adjust model with overdispersion parameter
- $var(\bar{y}_i) = \phi\frac{\pi(1-\pi_i)}{n_i}$

8.2 Max. Likelyhood Estimation

- **Score Func:** $s_\beta(y) = \frac{d}{d\beta}log(p_\beta(y))$
- **M Est:** $s_\beta(y)|_{\widehat{\beta}} = 0$
- **Observed Fisher Info:** $H(\beta) = \frac{d^2}{d^2\beta}l(\beta)$
- **Fisher Information:** $F(\beta) = E[s(\beta)s(\beta)'] = E[-\frac{d}{d\beta}s(\beta)]$
- **Z Estimator:** $\frac{d}{d\beta}\widehat{R}(\beta) = \frac{1}{n}\sum_{i=1}^n s_\beta(y) = s_\beta(\mathbb{Y})$
- **Alg. Newton :** $\beta^{t+1} = \beta^t + H(\beta^t)^{-1}s_{\beta^t}(\mathbb{Y})$
- **Alg. Fisher Scoring:** $\beta^{t+1} = \beta^t + F(\beta^t)^{-1}s_{\beta^t}(\mathbb{Y})$
- Note: $s_\beta(\mathbb{Y}) = \sum_i^n s_\beta(y_i)$
- Note: $\mathbb{F}(\beta) = nF(\beta)$
- **special case:** $y_i \sim Ber(\pi_i) \rightarrow s_\beta(\mathbb{Y}) = X'(Y-\pi) \rightarrow \mathbb{F}(\beta) = \sum_i^n x_i x_i'\pi_i(1-\pi_i) = X'VX$

8.3 Statistical Inference

- **Asymp. Theory :** $\widehat{\beta}_{MLE} \xrightarrow{D} N(\beta, F^{-1}(\beta))$
- $\widehat{cov}(\widehat{\beta}) = F^{-1}(\widehat{\beta})$
- $(F^(\widehat{\beta}))_{ii} = \widehat{var}(\widehat{\beta}_i) = se_i^2$ standard error for $\widehat{\beta}_i$
- **Testing:** Likelyhood ratio test $lr = 2(l(\widehat{\beta}) - l(\tilde{\beta})) \sim \chi_r^2$
- $H_0 : \beta_1 = 0 \rightarrow \widehat{\beta}$
- $H_A : \beta_1 \neq 0 \rightarrow \tilde{\beta}$
- **General Hyp:** Wald Statitic $w = (C\widehat{\beta} - d)'(CF^{-1}(\widehat{\beta})C')(C\widehat{\beta} - d) \sim_a \chi_r^2$ Note $C = r \times p$ with $rank(C) = r < p$
- **Sign test:** wald stat $w = z_j^2 = \frac{\widehat{\beta}_j^2}{F^{-1}(\widehat{\beta})_{ii}} \sim_a N(0,1)$
- **Model fit criteria :** Deviance $D(\widehat{\pi}) = -2l(\widehat{\pi})$ explicitly compare the fit with the perfect fit $l(\tilde{\pi} = 0)$ Like kullback leiber information. The smaller the better the fit
- Deviance for grouped data: Theoritical MLE known: $\tilde{\pi}_i = \bar{y}_i$ compare num to theoritical: $D = -2\sum_i^G(l_i(\widehat{\pi}_i) - l_i(\bar{y}_i))$
- Overdispersion estimation: $\widehat{\phi} = \frac{1}{n-p}D$
- Note: deviance good for comparing nested models

8.4 Count Data Regression

- $y_i \in \{0,1,2,...\}$ and $f_\beta(y_i) = \frac{\lambda_i^{y_i}}{y_i!}exp(-\lambda_i)$
- $E[y_i] = \lambda_i = var[y_i]$
- **Log Linear model:** $log(\lambda_i) = \eta_i = x_i'\beta$
- Overdispersion: variance higher in data than in model $\rightarrow$ adjust model $var[y_i] = \phi\lambda_i$
- $\widehat{\phi} = \frac{1}{n-p}D$

8.5 Unified Framework for GLM's

- **Exp Family:** $f_\theta(y) = exp[\frac{y\theta - b(\theta)}{\phi}w + C(\phi, w, y)]$
- $\theta :=$ natural parameter
- **Canonical Link Function:** $f_\gamma = exp[c(\gamma)T(y) - d(\gamma)]h(y) \rightarrow c(\gamma) := \theta$ c canonical link function
- Note: under natural paramter: $E[y] = \frac{d}{d\theta}b(\theta)$
- **Summary:** natural paramter modelled as linear $c(\gamma) = \theta = \eta = x'\beta$

8.6 Classification Metrics

- **Goal:** use predicted probabilities to classify (predict) new data: $y_{new} = \mathbb{1}(\widehat{pi} > t) \in \{0,1\}$ with t being the threshold
- $\widehat{\pi} = \frac{exp(x_{new}\widehat{\beta})}{1+exp(x_{new}\widehat{\beta})}$
- if $t = 0$ no mistake on oservation =1
- **Accuracy:** $\frac{TP+TN}{n}$
- **Sensitivity:** $TPR = \frac{TP}{TP+FN}$ (Recall)
- **Specificity:** $TNR = \frac{TN}{TN+FP}$
- $FPR = 1 - spec = \frac{FP}{TN+FP}$
- **ROC curve:** TPR vs FPR goal: reach point (0,1) and think threshold goes $1 \rightarrow 0$ left to right

9 Penalized Regression

- **Context:** If model is well specified the OLS estimator is unbiased but may have high variance $\rightarrow$ think subset of coeff that are close to zero bring the same variance as other coeff into the model. Overall idea is to trade some bias to reduce variance
- Case: if there is strong collinearity between covariates $\rightarrow X'X$ instable
- Case: High dimensional Regression $p > n$
- **Goal:** reduce SPSE by trading bias for some variance
- **Penalized Regr. :** $PLS(\beta) = \|Y - X\beta\|^2 + \lambda pen(\beta)$
- **Ridge:** $pen(\beta) = \|\beta\|^2 \rightarrow \widehat{\beta}_{ridge} = (X'X+I)^{-1}X'Y$
- **SVD:** $X = UDV'$ with $col(U) = Span(X)$ and $U'U = I$ with $U \in R^{n\times p}$
- $X\widehat{\beta}_{ridge} = \sum_{j=1}^p u_j\frac{d_j}{d_j+\lambda}u_j'Y$
- **Interpretation:** Ridge estimator shrinks the smaller principal component which correspond to small sample variances.
- $cov[\widehat{\beta}_{ridge}] \leq cov[\widehat{\beta}_{OLS}]$ BUT biased
- Ridge need covariates and response to be scaled and centered (centered because no penalty on intercept and scale because contrain on β so β_j must be of the same scale)
- Amount of shirnkage controlled by λ
- choosing $\lambda \rightarrow$ CV minimizing the SPSE and choose the simplest model
- **LASSO:** $pen(\beta) = |\beta|_1 \rightarrow \|Y - X\beta\|^2 + \lambda\sum_i^p|\beta_i|$
- Note: no close form
- **Graphical Overview :** $LS(\beta) = (\beta - \widehat{\beta})'X'X(\beta - \widehat{\beta}) + \widehat{\epsilon}'\widehat{\epsilon} \rightarrow$ ELLIPSOID (plot contour form and constraint on β)
- LASSO shrink coefficient to zero (more shrinkage for small coeff but less for big coeff)
- **Comparison of shrinkage:** Assume $X'X = I$ orhtogonal design and $\widehat{\beta}_j$ is OLS estimate
- Best subset selection: $\widehat{\beta}_j\mathbb{1}(|\widehat{\beta}_j| > \sqrt{\lambda})$
- Ridge: $\widehat{\beta}_j/(1+\lambda)$
- LASSO: $sgn(\widehat{\beta}_j)max\{|\widehat{\beta}_j| - \lambda/2\}$
- **Note on LASSO:** it performs soft thresholding and is like automatic variable selection as the coefficients are set to zero.

10 Maths Tricks Tutorials

Sub Differentials :

(i) **Convex Subset** : A subset $X \subseteq \mathbf{R}^k$ is called a convex subset iff $\forall x_1, x_2 \in X$, the segment that link them is contained in X

(ii) **Convex Function**: Let X be a convex subset and $f : X \rightarrow \mathbf{R}$ a function. Then f is convex iff $\forall x_1, x_2 \in X$ and $\forall t \in [0,1]$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

(iii) **Subdifferentials** : Let $f : I \rightarrow \mathbf{R}$ a real valued convex function on the open interval I. The subdifferential of f at $x_0 \in I$ is the set $[a,b]$ where

$$a = lim_{x\rightarrow x_0^-}\frac{f(x) - f(x_0)}{x - x_0}$$

$$b = lim_{x\rightarrow x_0^+}\frac{f(x) - f(x_0)}{x - x_0}$$

(iv) If f differential at x_0 then $a = b \implies [a,b] = \{f'(x_0)\}$

(v) **Proposition**: $x_0 \in I$ is a global minimum of f iff zero is contained in the subdifferential at x_0