STAT553 Summary

Wenjie Li

March 2019

1 Materials Before Exam1

1.1 Linear Models without the Normal Assumption

1. $\hat{\beta} = (X^TX)^{-1}X^TY$ minimizes $\|Y - X\beta\|_2$, can be proved by using the fact that any β can be written as $(X^TX)^{-1}z$ for some z, and $(X^TX)^{-1}$ is positive semi-definite.

2. $\text{Var}(l^TY) = l^T\Sigma l$, $\text{Cov}(AY) = A\Sigma A^T$

3. $\mathbb{E}(Y^TAY) = \mu^TA\mu + \text{tr}(A\Sigma)$

4. $\hat{\beta} = (X^TX)^{-1}X^TY$,$\text{Var}(l^T\beta) = \sigma^2 l^T(X^TX)^{-1}l$

5. If the information matrix (X^TX) is non-singular iff $c^T\beta$ is linearly unbiasedly estimable for any c

6. $\sum e_i^2 = Y^T(I-P)Y$, $P = X(X^TX)^{-1}X^T$, $\text{rank}(P) = \text{tr}(P) = \text{tr}(I_p) = p$, so p eigenvalues of P are non-zero(1's) and $n-p$ are zero. **Because** $PX_i = X_i$

7. $PX = X$, therefor $P\mathbf{1} = \mathbf{1}$, since X has a column of 1's. We can infer $\sum e_i = 0$

8. $\mathbb{E}(e^Te) = \mathbb{E}[Y^T(I-P)Y] = \sigma^2\text{tr}(I-P) = (n-p)\sigma^2$

9. The residuals are uncorrelated with the fitted values, *i.e.* $\sum e_i\hat{y}_i = 0$

10. Sample correlation of y and $\hat{y}$ is $r = (\sum y_i\hat{y}_i - \bar{y}^2)/(\sum(y_i - \bar{y}_i)^2\sum(\hat{y}_i - \bar{y})^2) = \frac{SSR}{\sqrt{SSR*SST}} = \sqrt{SSR/SST} = R$

1.2 Linear Models with the Normal Assumption

1. Confidence Interval for $c^T\beta$. Note that $c^T\hat{\beta} = c^T(X^TX)^{-1}X^TY \sim \mathcal{N}(c^T\beta, \sigma^2c^T(X^TX)^{-1}c)$. Hence the $100(1-\alpha)\%$ confidence interval for $c^T\hat{\beta}$ is $c^T\beta \pm t_{\alpha,n-p}\hat{\sigma}\sqrt{c^T(X^TX)^{-1}c}$.

2. However, for the above conclusion, we need the independence of $\hat{\beta}$ and $\hat{\sigma}^2$. Since $\hat{\beta} = (X^TX)^{-1}X^TY$, and $SSE = Y^T(I-P)Y$, they are independent (Y^TAY and BY are independent when $A\Sigma B^T = 0$).

3. Y^TAY and l^TY are independent iff $Al = 0$.

4. Let $Y \sim \mathcal{N}(\mu, \Sigma)$, $\psi = (Y-\mu)^T \Sigma^{-1}(Y-\mu) \sim \chi^2_p$ by standardizing Y

5. Since $\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2(X^TX)^{-1}, (\hat{\beta}-\beta)^T\sigma^{-2}(X^TX)(\hat{\beta}-\beta)^T) \sim \chi^2_p$, $P((\hat{\beta}-\beta)^T\sigma^{-2}(X^TX)(\hat{\beta}-\beta)^T \leq \chi^2_{p,\alpha}) = 1-\alpha$, which is an ellipsoid on β

6. If we replace σ^2 by $\hat{\sigma}^2$, under the normality assumption, $(\hat{\beta}-\beta)^T(X^TX)(\hat{\beta}-\beta)^T$ and $\hat{\sigma^2}$ are independent χ^2. Hence $\frac{(\hat{\beta}-\beta)^T(X^TX)(\hat{\beta}-\beta)^T/p}{\hat{\sigma^2}/n-p}$ their division follows $F_{p,n-p}$

1.3 Distributional Results on quadratic forms

1. Let $Y \sim N(0, \sigma^2 I)$, $Q = (Y^TAY)/\sigma^2$ is distributed as χ^2 iff A is idempotent, and the degree of freedom of Q is rank(A)

2. Suppose $Y \sim N(0, \sigma^2 I)$,$Q_1 = (Y^TAY)/\sigma^2$, $Q_2 = (Y^TBY)/\sigma^2$. Then Q_1, Q_2 are independent iff $AB = 0$

3. If $w \sim \chi^2(b), u \sim \chi^2(a), a < b$. Suppose $v = w - u \geq 0$, Then $v \sim \chi^2(b-a)$ and v is independent of u

4. $Y^TY \sim \chi^2$, Y^TPY is also χ^2 when P is idempotent and symmetric. Then $Y^T(I-P)Y$ is also χ^2 and it is independent of Y^TPY. (Error is independent of estimation).

1.4 Hypothesis Testing

1. **Gauss Markov Thm I.** $c^T\hat{\beta}$ is the best linear unbiased estimate (BLUE) of $c^T\beta$. i.e. Var$(l^TY) \geq$ Var$(c^T\hat{\beta})$ (By $c = X^Tl$)

2. **Gauss Markov Thm II.** . Let AY be any LUE of β, then Cov$(AY) \geq$ Cov$(\hat{\beta})$ (linear ordering, by $A = (X^TX)^{-1} + B$ and $BX = 0$)

3. Test $c^T\beta = \gamma$, $t_c = (c^T\hat{\beta} - \gamma)/\sqrt{\sigma^2 c^T(X^TX)^{-1}c} \sim t_{n-p}$

4. Suppose we want to simultaneously test $c_1^T\beta = \gamma_1, c_2^T\beta = \gamma_2, ..., c_k^T\beta = \gamma_k$, then the null is $H\beta = \gamma$, which is called a general linear hypothesis.

5. When testing $c^T\beta = \gamma$ Fit the full model and restricted model respectively, then get the RSS. Difference in RSS is $\frac{(c^T\beta-\gamma)^2}{c^T(X^TX)^{-1}c}$

6. Similarly, the difference when testing $H\beta = \gamma$ is $M = (H\hat{\beta}-\gamma)^T[H^T(X^TX)^{-1}H)]^{-1}(H\hat{\beta}-\gamma)$, so $M/\sigma^2 \sim \chi^2(k)$ and $\frac{M/k}{\hat{\sigma^2}/(n-p)} \sim F_{k,n-p}$

7. Prediction of new X_0 is $\hat{Y_0} = X_0^T\beta$, and $\hat{Y_0} - Y_0 \sim N(0, \sigma^2(1 + X_0^T(X^TX)^{-1}X_0))$

2 Materials Before Exam 2

2.1 Effect of Model Mis-specification on β

1. If the mean is mis-specified, $X = (X_1, X_2), \beta = (\beta_1, \beta_2)$, then

   - True Model: $Y = X\beta + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon$
   - Reduced Model: $Y = X_1\beta_1 + \epsilon$
   - $\mathbb{E}(\hat{\beta}_1) = \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2$, bias is 0 if $X_1^T X_2 = 0$
   - $\text{Cov}_{Full}(\hat{\beta}_1) \geq \text{Cov}_{Reduced}(\hat{\beta}_1)$ (Matrix sense), because $(X^T X)^{-1}_{11} \geq (X_1^T X_1)^{-1}$

2. If the variance is mis-specified, $\epsilon \sim N(0, \sigma^2 \Sigma)$, then

   - True Model: $Y = X\beta + \epsilon$
   - Generalized Least Square Estimates. $\Sigma^{-1/2} Y = \Sigma^{-1/2} X\beta + \Sigma^{-1/2}\epsilon$
   - $\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$, $\mathbb{E}(\hat{\beta}) = \beta$
   - If we use I instead of Σ, the covariance would be larger. $\text{Cov}((X^T X)^{-1} X^T Y) = \sigma^2 (X^T X)^{-1} X^T \Sigma X (X^T X)^{-1} \geq \sigma^2 (X^T \Sigma X)^{-1}$ (Matrix sense)

2.2 Effect of Model Mis-specification on σ^2

1. If the mean is mis-specified, $X = (X_1, X_2), \beta = (\beta_1, \beta_2)$, then

   - $Y = X_1\beta_1 + \epsilon$. By mis-specifying the mean, we will overestimate σ^2.

2.3 Asymptotics of $\hat{\beta}$

1. If the variance is mis-specified, $\Sigma \neq I$, then

   - The wrong estimate is $\hat{\beta}_R = (X^T X)^{-1} X^T Y$ $\text{Cov}(\hat{\beta}_R) = \sigma^2 (X^T X)^{-1} (X^T \Sigma X)(X^T X)^{-1}$
   - Take $\Sigma = Q^T \Delta Q$, and λ_{max} = largest eigenvalue of Σ, $\text{Cov}(\hat{\beta}_R) \leq \sigma^2 \lambda_{max} (X^T X)^{-1}$
   - Conclusion: If we use the wrong $\hat{\beta}_R = (X^T X)^{-1} X^T Y$, then $\hat{\beta}_R$ is unbiased and consistent if $\lambda_{max}(\Sigma) = O(1)$ and $\lambda_{min}(X^T X) \to \infty$.

2.4 Ridge Estimate

1. $\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y = (I + \lambda (X^T X)^{-1})^{-1} \hat{\beta}$ is called the ridge estimate of β.

2. $\text{Cov}(\hat{\beta}_\lambda) = \sigma^2 (I + \lambda (X^T X)^{-1})(X^T X)^{-1} (I + \lambda (X^T X)^{-1})$

3. There is a bias-variance tradeoff between the ridge estimate and OLS estimate $\hat{\beta}$

4. Bayes Estimate of β. $Y \sim N(X\beta, \sigma^2 I), \beta \sim N(0, \frac{1}{\lambda}I)$, then $\beta|Y \sim N((X^TX + \lambda I)^{-1}X^TY, \frac{\sigma^2}{\lambda\sigma^2+1}I)$

5. $\mathbb{E}_\beta(Y^TY) = \beta^TX^TX\beta + n\sigma^2, \mathbb{E}(Y^TY) = \frac{1}{\lambda}\text{tr}(X^TX) + n\sigma^2$

6. The empirical Bayes estimate for $\frac{1}{\lambda}$ is $\frac{Y^TY - n\sigma^2}{\text{tr}(X^TX)} \approx \frac{Y^TY - RSS}{\text{tr}(X^TX)} = \frac{Y^TPY}{\text{tr}(X^TX)}$

7. With this choice of λ, $\beta_\lambda = (X^TX + \frac{\text{tr}(X^TX)}{Y^TPY}I)^{-1}X^TY$.

8. Suppose now $X^TX = nI, \hat{\beta}_\lambda = (1 + \frac{p}{n\hat{\beta}^T\hat{\beta}})^{-1}\hat{\beta}$. This is the Stein estimate of β in the linear regression model for the orthogonal case.

9. $\hat{\beta}_\lambda = (X^TX + \lambda I)^{-1}X^TY = P^T(\Lambda + \lambda I)^{-1}PX^TY$

10. $\mathbb{E}(\hat{\beta}_\lambda) = (\sum_{i=1}^p \frac{s_i^2}{s_i^2+\lambda}u_iu_i^T)\beta \approx (\sum_{i=1}^p(1 - \frac{\lambda}{s_i^2}u_iu_i^T)\beta + o(|\lambda|)$, $\mathbb{E}(\hat{\beta}_\lambda) - \beta = -\lambda(X^TX)^{-1}\beta + o(\lambda)$.

11. The case of unknown Σ,we use the model $Y \sim N(X\beta, \Sigma)$, $Y = X\beta + z_1\alpha_1 + z_2\alpha_2 + \cdots z_r\alpha_r + \epsilon$, $\text{Cov}(\alpha_i, \alpha_j) = 0, \text{Cov}(\alpha_i) = \sigma^2 I$. There exists some matrix A such that $AY \sim N(0, A\Sigma A^T)$

2.5 Approximation Theory

1. For any normed linear space, $A \subset B$ is a finite dimensional subspace, then given $f \in B$, there exists a best approximation to f from A

2. To ensure the uniqueness of best approximation, make A a strictly convex set or the norm a strictly convex norm.

3. A is called strictly convex if for $x, y \in A, 0 < \lambda < 1$, $\lambda x + (1-\lambda)y \in A^o$ (interior of A)

4. Example: find a best linear polynomial approximation to a function f. The best approximation has an interesting property that $\|f - \hat{f}\|_\infty = |f(0) - \hat{f}(0)|, |f(1) - \hat{f}(1)|, |f(x_0) - \hat{f}(x_0)|$ for some x_0 and they alternate in sign

5. If the norm is strictly convex and A is a finite-dimensional subspace, then any $f \in B$ has a unique best approximation from A.

6. Hilbert space is a vector space with the property that every Cauchy sequence in B has a limit that belongs to B.

7. Let $f \in B$, then an $\hat{f}$ on A is a best approximation to f from A iff $\langle f - \hat{f}, g\rangle = 0, \forall g \in A$.

8. How to find the best linear approximate $\hat{f}$: $\hat{f} = \sum_j \frac{\langle f, \phi_j\rangle}{\|\phi_j\|_2^2}\phi_j$.

9. Let $\hat{f}$ be the best approximate to f, X be a linear operator. $\|f - X(f)\| \le \|f - \hat{f}\|(1 + \|X\|)$