SDS 383C: Statistical Modeling I **Fall 2016**

Lecture 6: September 20

Lecturer: Purnamrita Sarkar *Scribes: Ciara Nugent*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These scribe notes have been slightly proofread and may have typos etc.*

6.1 Review of Previous Lecture

6.1.1 Linear Regression

Last lecture we learned about linear regression. Recall the model for linear regression:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon} \quad \text{with } \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

where:

$$\boldsymbol{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & x_{12} & \dots & x_{1p} \\ 1 & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \dots & x_{np} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \text{and } \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Recall that the MLE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

Note that to solve this equation, $\boldsymbol{X}^T\boldsymbol{X}$ must be invertible. $\boldsymbol{X}^T\boldsymbol{X}$ is not invertible if it is singular. A square matrix, such as $\boldsymbol{X}^T\boldsymbol{X}$, will be singular if any row of the matrix is a linear combination of any of the other rows, this is called collinearity. In the case of linear regression, colinearity means that one of your covariates can be entirely explained by some combination of the other covariates. Later in this course we will discuss ways to address colinearity.

6.2 Gauss-Markov theorem

6.2.1 Theorem

Theorem 6.1 ***(Gauss-Markov Theorem)*** *In statistics, the Gauss-Markov theorem states that in a linear regression model in which the errors are distributed iid from a $N(0, \sigma^2)$ distribution, for all linear estimators of the form $\boldsymbol{Ay}$, the least squares estimate has the smallest variance. Consider an alternate estimator $\tilde{\boldsymbol{\beta}}$ such that*

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{Ay} \text{ and } E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}.$$

Let the covariance matrix of $\tilde{\boldsymbol{\beta}}$ be given by $\tilde{\boldsymbol{\Sigma}}$ and the covariance of $\hat{\boldsymbol{\beta}}$ be given by $\boldsymbol{\Sigma}$. Then we have $\tilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} \succeq 0$, where $\boldsymbol{A} \succeq 0$ denotes that $\boldsymbol{A}$ is positive semidefinite.

[0]These notes are partially based on those of Li Kang (Kelly) and Su Chen.

Note: Does this mean that the marginal variances of each element of $\hat{\boldsymbol{\beta}}$ are also smaller? Yes! Recall that the definition of positive semi-definiteness arises from the fact that for a positive semi-defnitie matrix $\boldsymbol{A}$, $\forall \boldsymbol{x} \neq \boldsymbol{0}$, $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq 0$. Take $\boldsymbol{x}$ to be a zero vector with a one at the i^{th} place. Now, $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = A_{ii}$, so $\text{var}(\hat{\beta}_i) \leq \text{var}(\tilde{\beta}_i)$ for every i.

6.2.2 Proof

Consider the model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{with } \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

Let us consider two unbiased estimators, $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$, for $\boldsymbol{\beta}$.

$\hat{\boldsymbol{\beta}}$ is the MLE, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. Under the model:

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \quad Cov(\hat{\boldsymbol{\beta}}) = \sigma^2((\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

$$E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}, \quad Cov(\tilde{\boldsymbol{\beta}}) = \sigma^2((\boldsymbol{B}^T\boldsymbol{B})^{-1})$$

$$E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}, \quad Cov(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$$

Let $\boldsymbol{B} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + \boldsymbol{G}$, then:

$$\begin{aligned} Cov(\tilde{\boldsymbol{\beta}}) &= \sigma^2(\boldsymbol{B}^T\boldsymbol{B})^{-1} \\ &= \sigma^2(((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + \boldsymbol{G})((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + \boldsymbol{G})^T) \\ &= \sigma^2((\boldsymbol{X}^T\boldsymbol{X})^{-1} + \boldsymbol{G}\boldsymbol{G}^T + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{G}^T + \boldsymbol{G}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \\ &= \sigma^2((\boldsymbol{X}^T\boldsymbol{X})^{-1} + \boldsymbol{G}\boldsymbol{G}^T). \end{aligned} \tag{6.1}$$

Proving that $Cov(\tilde{\boldsymbol{\beta}}) - Cov(\hat{\boldsymbol{\beta}}) \succeq 0$, and hence the Gauss-Markov Theorem.

6.3 Hypothesis Tests

6.3.1 Hypothesis Tests

Define a null space, Θ_0, and an alternate space, Θ_a. Define the null hypothesis as $H_0 : \theta \in \Theta_0$, and the alternate hypothesis as $H_a : \theta \in \Theta_a$. We are interested in testing H_0 vs. H_a.

There are two types of error, Type I Error and Type II Error. Type I Error occurs when you reject the null hypothesis but should not have. Type II Error occurs when you fail to reject the null hypothesis when you should have. Hypothesis tests deal with Type I Error, while Type II Error is addressed by the Power of a test. We will focus on hypothesis tests. Define a hypothesis test as the probability of a Type I error, P(Type I Error). For a hypothesis test at level α, you are calculating P(Type I Error)$= \alpha$.

6.3.2 Hypothesis Tests for Linear Regression

For a linear model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{\sigma^2}\boldsymbol{I}).$$

For each β_j we wish to test:

$$H_0 : \beta_j = 0 \text{ vs. } H_a : \beta_j \neq 0.$$

Since, $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$, for each β_j, $\hat{\beta}_j \sim N(\beta_j, \sigma^2(X^TX)^{-1}_{jj})$. This can also be expressed as:

$$\frac{\hat{\beta}_j}{\sigma\sqrt{(X^TX)^{-1}_{jj}}} \sim N(0,1).$$

If σ^2 is known, this is just a Z-test with $z_j = \frac{\hat{\beta}_j}{\sigma\sqrt{(X^TX)^{-1}_{jj}}} \sim N(0,1)$. This is a case of a two-sided test, so for $\alpha = 0.05$ you would reject H_0 if $|z_j| \geq 1.96$. For a one sided test, and $\alpha = 0.05$, you would look at $z_j \geq 1.645$ or $z_j \leq -1.645$ depending on which tail you were interested in.

For σ^2 unknown, you would use a t-test. First you would plug-in $\hat{\sigma}$ for σ, where $\hat{\sigma}^2 = \frac{RSS}{n-p-1} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p-1}$. Where n is the number of observations and p is the number of covariates. Then, $\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2(X^TX)^{-1}_{jj}}} \sim t_{n-p-1}$, where t_{n-p-1} is a t-distribution with n-p-1 degrees of freedom. As n increases the t-distribution approaches the normal distribution, given p is not increasing with n.