Notes on Linear Regression

Keng-Wit Lim*[1]

[1]XXXX Los Angeles, CA, USA

March 16, 2018

Abstract

Notes on linear regression.

Keywords: XXX

*kengwit@gmail.com

1 Basics

Constant matrix $\mathbf{A}$ and vector $\mathbf{u}$. Note that variance of a vector results in a matrix:

$$\begin{aligned}
Var(\mathbf{Au}) &= E\left[\{\mathbf{Au} - E(\mathbf{Au})\}^2\right] \\
&= E\left[\{\mathbf{Au} - E(\mathbf{Au})\}\{\mathbf{Au} - E(\mathbf{Au})\}^T\right] \\
&= E\left[\{\mathbf{Au} - E(\mathbf{Au})\}\left\{\mathbf{u}^T\mathbf{A}^T - E(\mathbf{u}^T\mathbf{A}^T)\right\}\right] \\
&= E\left[\mathbf{Auu}^T\mathbf{A}^T - \mathbf{Au}E(\mathbf{u}^T\mathbf{A}^T) - E(\mathbf{Au})\mathbf{u}^T\mathbf{A}^T + E(\mathbf{Au})E(\mathbf{u}^T\mathbf{A}^T)\right] \\
&= E(\mathbf{Auu}^T\mathbf{A}^T) - E(\mathbf{Au})E(\mathbf{u}^T\mathbf{A}^T) - E(\mathbf{Au})E(\mathbf{u}^T\mathbf{A}^T) + E(\mathbf{Au})E(\mathbf{u}^T\mathbf{A}^T) \\
&= \mathbf{A}E(\mathbf{uu}^T)\mathbf{A}^T - \mathbf{A}E(\mathbf{u})E(\mathbf{u}^T)\mathbf{A}^T - \mathbf{A}E(\mathbf{u})E(\mathbf{u}^T)\mathbf{A}^T + \mathbf{A}E(\mathbf{u})E(\mathbf{u}^T)\mathbf{A}^T \\
&= \mathbf{A}\left[E(\mathbf{uu}^T) - E(\mathbf{u})E(\mathbf{u}^T)\right]\mathbf{A}^T \\
&= \mathbf{A}E\left[\{\mathbf{u} - E(\mathbf{u})\}\{\mathbf{u} - E(\mathbf{u})\}^T\right]\mathbf{A}^T \\
&= \mathbf{A}E\left[\{\mathbf{u} - E(\mathbf{u})\}^2\right]\mathbf{A}^T \\
&= \mathbf{A}Var(\mathbf{u})\mathbf{A}^T \qquad (1)
\end{aligned}$$

2 Notation

$$\mathbf{y} = \left\{ \begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_N \end{array} \right\} \tag{2}$$

$$\mathbf{X} = \left[\begin{array}{cccc} 1 & X_{11} & \dots & X_{1p} \\ 1 & X_{21} & \dots & X_{2p} \\ & \vdots & & \\ 1 & X_{N1} & \dots & X_{Np} \end{array} \right] \tag{3}$$

$$\boldsymbol{\beta} = \left\{ \begin{array}{c} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{array} \right\} \tag{4}$$

3 Linear Regression

3.1 Assumptions

The following assumptions will allow us to draw inferences about the estimators and linear regression model:

1. A linear regression model assumes that the regression function

$$\begin{aligned} f(\mathbf{X}) &= E(\mathbf{y}|\mathbf{X}) \\ &= \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j \\ &= \mathbf{X}\boldsymbol{\beta} \end{aligned} \tag{5}$$

is linear in the inputs $\mathbf{X}$. This means that (5) is assumed to be the correct model for the mean and that the conditional expectation of $E(\mathbf{y}|\mathbf{X})$ is linear in $X_1, \ldots, X_p$.

2. The *true* relation between a quantitative response $\mathbf{y}$ on the basis of predictors $\mathbf{X}$ is assumed to take the form

$$\begin{aligned} \mathbf{y} &= E(\mathbf{y}|\mathbf{X}) + \boldsymbol{\epsilon} \\ &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \end{aligned} \tag{6}$$

where $\boldsymbol{\epsilon}$ is the error or residual vector, and it is assumed that each element of $\boldsymbol{\epsilon}$ is normally distributed with zero mean and has (unobserved) variance of σ, i.e., $\epsilon_i \sim N(0, \sigma^2)$. This means that

$$\begin{aligned} Var(\boldsymbol{\epsilon}) &= E(\boldsymbol{\epsilon}^2) \\ &= E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) \\ &= \sigma^2\mathbf{I} \end{aligned} \tag{7}$$

where $\mathbf{I}$ is the $N \times N$ identity matrix. The assumed relation (6) means that the deviations of $\mathbf{y}$ around its expectation are additive and Gaussian.

3.2 Solution for the Estimators

In linear regression, we assume that there is approximately a linear relation between $\mathbf{y}$ and $\mathbf{X}$:

$$\mathbf{y} \approx \mathbf{X}\boldsymbol{\beta} \tag{8}$$

The least-squares solution for the estimator vector is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{9}$$

With the estimator vector, the predicted values are:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{10}$$

3.3 Properties of the Estimators

Now, we derive the mean and variance for the estimator. The mean is

$$
\begin{aligned}
E(\hat{\boldsymbol{\beta}}) &= E\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right] \\
&= E\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}\right)\right] \quad \text{using (6)} \\
&= E\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta}+\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\right] \quad \text{using (6)} \\
&= \underbrace{E(\boldsymbol{\beta})}_{\boldsymbol{\beta}}+\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\underbrace{E(\boldsymbol{\epsilon})}_{\mathbf{0}} \\
&= \boldsymbol{\beta}
\end{aligned} \tag{11}
$$

The variance is

$$
\begin{aligned}
Var(\hat{\boldsymbol{\beta}}) &= E\left[\left\{\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right\}\left\{\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right\}^T\right] \\
&= E\left[\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T - \hat{\boldsymbol{\beta}}E(\hat{\boldsymbol{\beta}})^T - E(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}^T + E(\hat{\boldsymbol{\beta}})E(\hat{\boldsymbol{\beta}})^T\right] \\
&= E(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T) - E(\hat{\boldsymbol{\beta}})E(\hat{\boldsymbol{\beta}})^T \\
&= E(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^T) - \boldsymbol{\beta}\boldsymbol{\beta}^T \quad \text{using (9)} \\
&= E\left(\left\{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right\}\left\{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right\}^T\right) - \boldsymbol{\beta}\boldsymbol{\beta}^T \quad \text{using (6)} \\
&= E\left(\left\{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}\right)\right\}\left\{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}\right)\right\}^T\right) - \boldsymbol{\beta}\boldsymbol{\beta}^T \\
&= E\left(\left\{\boldsymbol{\beta} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\right\}\left\{\boldsymbol{\beta} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\right\}^T\right) - \boldsymbol{\beta}\boldsymbol{\beta}^T \\
&= E\left(\left\{\boldsymbol{\beta} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\right\}\left\{\boldsymbol{\beta} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\right\}^T\right) - \boldsymbol{\beta}\boldsymbol{\beta}^T \quad \text{using (6)} \\
&= E\left(\left\{\boldsymbol{\beta} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\right\}\left\{\boldsymbol{\beta}^T + \boldsymbol{\epsilon}^T\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T\right\}\right) - \boldsymbol{\beta}\boldsymbol{\beta}^T \\
&= E\left(\boldsymbol{\beta}\boldsymbol{\beta}^T + \boldsymbol{\beta}\boldsymbol{\epsilon}^T\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\boldsymbol{\beta}^T + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T\right) \\
&\quad -\boldsymbol{\beta}\boldsymbol{\beta}^T \\
&= E(\boldsymbol{\beta}\boldsymbol{\beta}^T) + E\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T\right) - \boldsymbol{\beta}\boldsymbol{\beta}^T \\
&= E(\boldsymbol{\beta}\boldsymbol{\beta}^T) + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)\right)\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T - \boldsymbol{\beta}\boldsymbol{\beta}^T \\
&= \boldsymbol{\beta}\boldsymbol{\beta}^T + \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T - \boldsymbol{\beta}\boldsymbol{\beta}^T \quad \text{using (7) and (11)} \\
&= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-T} \\
&= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-T} \\
&= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \qquad (12)
\end{aligned}
$$