

# Notes on Linear Regression

Keng-Wit Lim<sup>\*1</sup>

<sup>1</sup>XXXX Los Angeles, CA, USA

March 19, 2018

## Abstract

Notes on linear regression.

---

Keywords: XXX

---

<sup>\*</sup>kengwit@gmail.com

# 1 Basics

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

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

## 2 Notation

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \quad (2)$$

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1p} \\ 1 & X_{21} & \dots & X_{2p} \\ & \vdots & & \\ 1 & X_{N1} & \dots & X_{Np} \end{bmatrix} \quad (3)$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad (4)$$

## 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

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

is linear in the inputs  $\mathbf{X}$ . Stated differently, we assume that  $\mathbf{X}\boldsymbol{\beta}$  is the correct model for the

mean; that is, the conditional expectation of  $E(\mathbf{y}|\mathbf{X})$  is linear in  $X_1, \dots, X_p$ .

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  $\sigma$ , i.e.,  $\epsilon_i \sim N(0, \sigma^2)$ . This means that

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

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

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

With the estimator vector, the predicted values are:

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

### 3.3 Properties of the Estimators

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

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

The variance (variance-covariance matrix) is

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

$$= \begin{bmatrix} var(\hat{\beta}_0) & cov(\hat{\beta}_0, \hat{\beta}_1) & \dots & cov(\hat{\beta}_0, \hat{\beta}_p) \\ cov(\hat{\beta}_1, \hat{\beta}_0) & var(\hat{\beta}_1) & \dots & cov(\hat{\beta}_1, \hat{\beta}_p) \\ \vdots & \vdots & \dots & \vdots \\ cov(\hat{\beta}_p, \hat{\beta}_0) & cov(\hat{\beta}_p, \hat{\beta}_1) & \dots & var(\hat{\beta}_p) \end{bmatrix} \tag{13}$$

so that

$$var(\hat{\beta}_j) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \tag{14}$$

Denote  $v_j$  as the  $j$ -th diagonal element of  $(\mathbf{X}^T \mathbf{X})^{-1}$  so that

$$\text{var}(\hat{\beta}_j) = \sigma^2 v_j = \sigma^2 (\mathbf{X}^T \mathbf{X})_{jj}^{-1} \quad (15)$$

$\mathbf{X}^T \mathbf{X}$  measures the spread so that as the spread increases, the variance of the estimator decreases (check this statement).

### 3.4 Hypothesis testing

For linear regression, the chief statements of the hypothesis test are expressed as

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_1 : \beta_1 &\neq 0 \end{aligned} \quad (16)$$

In this case, the hypothesis tests for the significance of regression. In other words, the test indicates if the fitted regression model is of value in explaining variations in the observations or if you are trying to impose a regression model when no true relationship exists between  $\mathbf{x}$  and  $\mathbf{Y}$ . Failure to reject  $H_0 : \beta_1 = 0$  implies that no linear relationship exists between  $\mathbf{x}$  and  $\mathbf{Y}$ . This result may be obtained when the scatter plots of against are as shown in (a) of the following figure and (b) of the following figure. (a) represents the case where no model exists for the observed data. In this case you would be trying to fit a regression model to noise or random variation. (b) represents the case where the true relationship between  $\mathbf{x}$  and  $\mathbf{Y}$  is not linear. (c) and (d) represent the case when  $H_0 : \beta_1 = 0$  is rejected, implying that a model does exist between  $\mathbf{x}$  and  $\mathbf{Y}$ . (c) represents the case where the linear model is sufficient. In the following figure, (d) represents the case where a higher order model may be needed.

First we need to say something about the variance of the estimators. Usually the actual variance  $\sigma$  is not known and it is estimated as

$$\hat{\sigma}^2 = \frac{1}{N - (p + 1)} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (17)$$

Then from (15), the estimate of the variance of the  $j$ -th estimator is

$$var(\hat{\beta}_j) \approx \hat{\sigma}^2 v_j = \hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})_{jj}^{-1} \quad (18)$$

For hypothesis testing, we first calculate the Z-score of the  $j$ -th estimator

$$\begin{aligned} z_j &= \frac{\hat{\beta}_j - \beta_j}{\sqrt{var(\hat{\beta}_j)}} \\ &= \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{v_j}} \end{aligned} \quad (19)$$

since  $\beta_j = 0$  (the null hypothesis).

Since we do not know the actual variance and we use an estimate of the variance, the appropriate distribution to be used would be the  $t$  distribution with  $N - (p + 1)$  degrees of freedom  $t_{N-(p+1)}$  (as opposed to a normal distribution); a large (absolute) value of  $z_j$  will lead to rejection of  $H_0 : \beta_1 = 0$ . Note: as the sample size increases, the difference between the tail quantiles of a  $t$ -distribution and a normal distribution becomes negligible.