Notes on Linear Regression

Keng-Wit Lim*[1]

[1]XXXX Los Angeles, CA, USA

January 14, 2019

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:

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

2 Notation

$$\mathbf{y} = \begin{Bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{Bmatrix} \tag{2}$$

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

$$\boldsymbol{\beta} = \begin{Bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{Bmatrix} \tag{4}$$

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 $\mathbf{X}$. 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 $X_1, \ldots, X_p$.

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

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

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

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

The variance (variance-covariance matrix) is

$$
\begin{aligned}
\mathrm{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) \\
&= \begin{bmatrix}
\mathrm{Var}(\hat{\beta}_0) & \mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_1) & \dots & \mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_p) \\
\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_0) & \mathrm{Var}(\hat{\beta}_1) & \dots & \mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_p) \\
\vdots & \vdots & \dots & \vdots \\
\mathrm{Cov}(\hat{\beta}_p,\hat{\beta}_0) & \mathrm{Cov}(\hat{\beta}_p,\hat{\beta}_1) & \dots & \mathrm{Var}(\hat{\beta}_p)
\end{bmatrix} \qquad (13)
\end{aligned}
$$

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

$$\text{Var}(\hat{\beta}_j) \quad = \quad \sigma^2 v_j = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{jj} \tag{14}$$

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

Typically, the population variance σ^2 is unknown and estimated as:

$$\hat{\sigma}^2 \quad = \quad \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{15}$$

The $N-p-1$ rather than N in the denominator makes $\hat{\sigma}^2$ an unbiased estimate of σ^2, i.e., $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$.

Based on the above, we have

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}) \tag{16}$$

(How do you know that it is normally distributed? Based on properties of sum and multiplication by a constant for normal distributions?). Furthermore, since the expectation (mean) of a chi-squared distribution of $N-p-1$ is:

$$\mathbb{E}(\chi^2_{N-p-1}) = N-p-1 \tag{17}$$

we have

$$\mathbb{E}\left(\frac{1}{N-p-1}\sum(y_i - \hat{y}_i)^2\right) = \mathbb{E}(\hat{\sigma}^2) = \sigma^2 \tag{18}$$

Therefore,

$$(N-p-1)(\hat{\sigma}^2) \sim \sigma^2 \chi^2_{N-p-1} \tag{19}$$

3.4 Case of single explanatory variable

When there is a single explanatory variable, the model (6) reduces to:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n \tag{20}$$

and

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ & \vdots \\ 1 & x_n \end{bmatrix} \tag{21}$$

$$\boldsymbol{\beta} = \begin{Bmatrix} \beta_0 \\ \beta_1 \end{Bmatrix} \tag{22}$$

so that

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \tag{23}$$

and formulas become more transparent. For example, the *standard error* of the estimated slope β_1 is

$$\widehat{\mathrm{SE}}(\hat{\beta}_1) \equiv \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_1)} = \sqrt{[\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}]_{22}} = \sqrt{\frac{n\hat{\sigma}^2}{n\sum x_i^2 - (\sum x_i)^2}}. \tag{24}$$

The denominator can be written as

$$n\sum_i (x_i - \bar{x})^2 \tag{25}$$

Therefore,

$$\widehat{\mathrm{SE}}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2}} \tag{26}$$

with

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_i \hat{\epsilon}_i^2 \tag{27}$$

The $n - 2$ term accounts for the loss of 2 degrees of freedom in the estimation of the intercept and the slope. This explanation of degrees of freedom is standard but is in fact technically *not correct*; see interesting discussion on degrees of freedom on https://stats.stackexchange.com/questions/16921/how-to-understand-degrees-of-freedom

3.5 The Chi-Squared Distribution

A chi-square random variable is the result of squaring Normal random variables. Let $Y \sim N(0,1)$. We call this a standard normal random variable. What happens when we square Y? Let $Z = Y^2$. Then Z is a chi-square random variable with 1 degree of freedom (df). We can write this as:

$$Z = Y^2 \sim \chi^2_1 \tag{28}$$

where the subscript 1 means that there is 1 df.

The more general chi-squared distribution, with different degrees of freedom, is obtained by summing up the squares of independent, normally distributed random variables. Let $Y_1, \ldots, Y_d \overset{\text{iid}}{\sim} N(0,1)$ be d independent standard normal random variables. If we square them all and then sum them up we get a Chi-square random variable with d degrees of freedom.

$$\begin{aligned} W &= Y_1^2 + Y_2^2 + \ldots + Y_d^2 \\ W &\sim \chi^2_d \end{aligned} \tag{29}$$

We would say this as "W comes from a chi-squared distribution with d degrees of freedom." Note that the chi-square distribution has a pdf, just like the Normal distribution; those interested can find out much more about the chi-square distribution by looking at the relevant Wikipedia page. It will be helpful to know the mean of a chi-square random variable, which is equal to the degrees of freedom:

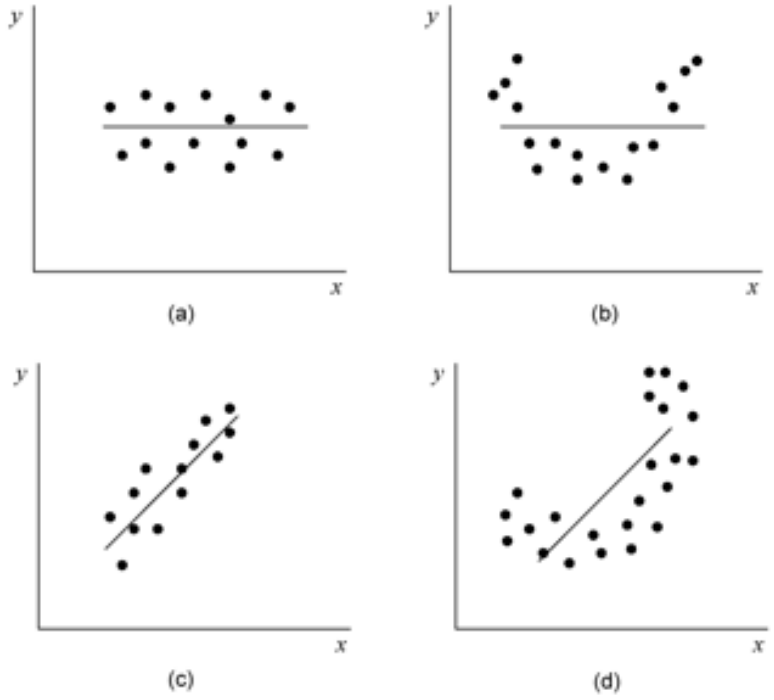
$$\text{If } W \sim \chi^2_d, \text{ then } E(W) = d \tag{30}$$

3.6 Hypothesis testing

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

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0 \tag{31}$$

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 exits for the observed data (note: the slope or β_1 is zero). 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 σ is not known and it is estimated as

$$\hat{\sigma}^2 = \frac{1}{N-(p+1)} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{32}$$

Then from (14), the estimate of the variance of the j-th estimator is

$$\text{Var}(\hat{\beta}_j) \approx \hat{\sigma}^2 v_j = \hat{\sigma}^2 \left(\mathbf{X}^T \mathbf{X}\right)^{-1}_{jj} \tag{33}$$

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

$$\begin{aligned} z_j &= \frac{\hat{\beta}_j - \beta_j}{\sqrt{\text{Var}(\hat{\beta}_j)}} \\ &= \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}} \end{aligned} \tag{34}$$

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.

3.7 Confidence Interval or Set

Recall from (16) that

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^T \mathbf{X}\right)^{-1}) \tag{35}$$

If one isolates β_j, we have

$$\begin{aligned} & \hat{\beta}_j - \beta_j \sim N(0, \sigma^2 v_j) \quad \text{using (14)} \\ \rightarrow \quad & \frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{v_j}} \sim \frac{1}{\sigma\sqrt{v_j}} N(0, \sigma^2 v_j) \\ \rightarrow \quad & \frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{v_j}} \sim N(0, 1) \end{aligned} \tag{36}$$

Based on the discussion in Section 3.5, we have

$$\begin{aligned} & \left(\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{v_j}} \right)^2 \sim \chi_1^2 \\ \rightarrow \quad & \left(\frac{\hat{\beta}_j - \beta_j}{\sqrt{v_j}} \right)^2 \sim \sigma^2 \chi_1^2 \end{aligned} \tag{37}$$