Lecture - 7

March 24, 2012

Linear Regression

Linear Regression is one of the most widely used statistical methods. We shall start with Multiple linear regression analysis. Our discussion of the multiple linear regression uses matrix algebra discussed earlier.

The classical linear Regression Model

The multiple linear regression model consist of a single response variable with multiple predicators and assumes, the form

$$Y_i = \beta_0 + \beta_1 Z_{i_1} + \beta_2 Z_{i_2} + \cdots + \beta_r Z_{r_i} + \epsilon_i \ , \quad i = 1(1)n \tag{1}$$

Where the error terms ϵ_i satisfy

1. $E(\epsilon_i) = 0$ for all i
2. $Var(\epsilon_i) = \sigma^2$, a constant; and
3. $Cov(\epsilon_i, \epsilon_j) = 0$ if $i \neq j$.

In matrix notation, Eq(1) becomes

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & Z_{11} & Z_{12} & \cdots & Z_{1r} \\ 1 & Z_{21} & Z_{22} & \cdots & Z_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & Z_{n1} & Z_{n2} & \cdots & Z_{nr} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

or,

$$Y_{n\times 1} = Z_{n\times(r+1)}\beta_{(r+1)\times 1} + \epsilon_{n\times 1}$$

and the assumptions become

1. $E(\epsilon) = 0$ and
2. $Cov(\epsilon) = E(\epsilon\epsilon^{'}) = \sigma^2 I_{n\times n}$

The matrix Z is referred to as the design matrix.

Least Square Estimation

The least square estimation (LSE) of β is obtained by minimizing the sum of squares of errors

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 Z_{i_1} - \beta_2 Z_{i_2} - \cdots - \beta_r Z_{r_i})^2 \\ &= \big(Y - Z\beta\big)^{'}\big(Y - Z\beta\big) \end{aligned}$$

Before deriving the LSE, we first define $\tilde{\beta} = \big(Z^{'}Z\big)^{-1}Z^{'}Y$ and $H = Z\big(Z^{'}Z\big)^{-1}Z^{'}$, assuming that the inverse exists. Then, it is easy to see that H is symmetric and $H^2 = H$ i.e. H is idempotent matrix.

Define,

$$\begin{aligned} \tilde{\epsilon} &= Y - Z\tilde{\beta} \\ &= Y - Z\big(Z^{'}Z\big)^{-1}Z^{'}\underline{Y} \\ &= \big[I - H\big]\underline{Y} \end{aligned}$$

Then using $Z^{'}H = Z^{'}$, we have

$$Z^{'}\tilde{\epsilon} = Z^{'}\big[I - H\big]Y = \big[Z^{'} - Z^{'}H\big]Y = \big[Z^{'} - Z^{'}\big]Y = 0$$

In other words, $\tilde{\epsilon}$ is orthogonal to the column space of the design matrix Z.

<u>Result :</u> Assume that Z is of full rank $(r+1)$; where $r+1 \leq n$. The LSE of β is

$$\hat{\beta} = \big(Z^{'}Z\big)^{-1}Z^{'}Y$$

Proof.

$$\begin{aligned}
\big(Y - Z\underline{\beta}\big) &= \big(Y - Z\tilde{\beta} + Z\tilde{\beta} - Z\beta\big) \\
&= \big(Y - Z\tilde{\beta}\big) + Z(\tilde{\beta} - \beta)
\end{aligned}$$

We have,

$$\begin{aligned}
S(\beta) &= \big(Y - Z\beta\big)^{'}\big(Y - Z\beta\big) \\
&= \big[\big(Y - Z\tilde{\beta}\big) + Z(\tilde{\beta} - \beta)\big]^{'}\big[\big(Y - Z\tilde{\beta}\big) + Z(\tilde{\beta} - \beta)\big] \\
&= \big(Y - Z\tilde{\beta}\big)^{'}\big(Y - Z\tilde{\beta}\big) + (\tilde{\beta} - \beta)^{'}(ZZ^{'})(\tilde{\beta} - \beta)
\end{aligned}$$

because $\big(Y - Z\tilde{\beta}\big)^{'}Z = \tilde{\epsilon}^{'}Z = 0$. The first term of $S(\beta)$ doesnot depend on β and the second term is the squared length of $Z(\tilde{\beta} - \beta)$. Because Z has full rank, $Z(\tilde{\beta} - \beta) \neq 0$ if $\tilde{beta} \neq \beta$, so the minimum of $S(\beta)$ is unique and occurs at $\beta = \tilde{\beta}$. Consequently, the LSE of β is $\hat{\beta} = \tilde{\beta} = \big(Z^{'}Z\big)^{-1}Z^{'}Y$. *Q.E.D.*

∎

The derivations

$$\hat{\epsilon} = Y_i - \beta_0 - \beta_1 Z_{i_1} - \beta_2 Z_{i_2} - \cdots - \beta_r Z_{r_i}\ , \quad i = 1(1)n$$

are the residuals of the MLR model.

Let $\hat{Y} = Z\hat{\beta} = HY$ denote the filled values of Y. The H matrix is called the hat matrix. The LS residuals then become

$$\begin{aligned}\hat{\epsilon} &= Y - \hat{Y} \\ &= \big[I - Z\big(Z^{'}Z\big)^{-1}Z^{'}\big]Y \\ &= \big[I - H\big]Y\end{aligned}$$

Note that $(I - H)$ is also a symmetric and idempotent matrix.

Our prior discussion shows that the residuals satisfy (a) $Z^{'}\hat{\epsilon} = 0$ and (b) $\hat{Y}^{'}\hat{\epsilon} = 0$. Also the residual sum of squares (RSS) is $RSS = \sum_{i=1}^{n}\hat{\epsilon_i}^2 = \hat{\epsilon}^{'}\hat{\epsilon} = Y^{'}[I - H]Y$.

Sum of Square decomposition :

Since, $\hat{Y}^{'}\hat{\epsilon} = 0$, we have

$$Y^{'}Y = \big(\hat{Y} + \hat{\epsilon}\big)^{'}\big(\hat{Y} + \hat{\epsilon}\big) = \hat{Y}^{'}\hat{Y} + \hat{\epsilon}^{'}\hat{\epsilon}$$

Since, the first column of Z is 1, the result $Z^{'}\hat{\epsilon} = 0$ includes $0 = 1^{'}\hat{\epsilon} = 1^{'}(Y - \hat{Y})$. Consequently, we have $\overline{Y} = \overline{\hat{Y}}$, subtracting $n\overline{Y}^2 = n(\overline{\hat{Y}})^2$ from the prior decomposition, we have

$$Y^{'}Y - n\overline{Y}^2 = \hat{Y}^{'}\hat{Y} - n(\overline{\hat{Y}})^2 + \hat{\epsilon}^{'}\hat{\epsilon}$$

or,

$$\sum_{i=1}^{n}\big(Y_i - \overline{Y}\big)^2 = \sum_{i=1}^{n}\big(\hat{Y}_i - \overline{Y}\big)^2 + \sum_{i=1}^{n}\hat{\epsilon_i}^2$$

The quantity

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\hat{\epsilon_i}^2}{\sum_{i=1}^{n}\big(Y_i - \overline{Y}\big)^2}$$

is the co-efficient of determination. It is the proportion of total variation in $Y_i^{'}s$ explained by the predictors $Z_1, \cdots, Z_p$.

Sampling Properties :

1. The LSE $\hat{\beta}$ satisfies $E(\hat{\beta}) = \beta$, $\big[E\big((Z'Z)^{-1}Z'Y\big) \;=\; (Z'Z)^{-1}(Z'Z)\beta\,;\; E(Y) = Z\beta \;=\; \beta\big]$

$$\begin{aligned} Var(\hat{\beta}) &= V\big((Z'Z)^{-1}Z'Y\big) \\ &= (Z'Z)^{-1}Z'V(Y)Z(Z'Z)^{-1} \\ &= (Z'Z)^{-1}(Z'Z)(Z'Z)^{-1}\sigma^2 \\ &= (Z'Z)^{-1}\sigma^2 \end{aligned}$$

2. The residuals satisfy $E(\hat{\epsilon}) = 0$ and $V(\hat{\epsilon}) = V((I-H)Y) = (I-H)\sigma^2$

3. $E(\hat{\epsilon}'\hat{\epsilon}) = (n-r-1)\sigma^2$ so that, defining

$$S^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-r-1} = \frac{Y'[I-H]Y}{n-r-1}$$

.

Proof. Using

$$tr(H) = tr(Z(Z'Z)^{-1}Z') = tr((Z'Z)^{-1}Z'Z) = tr(I_{r+1}) = r+1$$

$$\begin{aligned} E(\hat{\epsilon}'\hat{\epsilon}) &= E(tr(\hat{\epsilon}'\hat{\epsilon})) \\ &= E[tr(\hat{\epsilon}\hat{\epsilon}')] \\ &= tr[Cov(\hat{\epsilon}\hat{\epsilon})] \\ &= tr[\sigma^2(I-H)] \\ &= \sigma^2[tr(I) - tr(H)] \\ &= \sigma^2[n-r-1] \end{aligned}$$

∎

Gauss's Least Squares Theorem :

Theorem 0.1. *Let* $Y = Z\underline{\beta} + \underline{\epsilon}$*, where* $E(\epsilon) = 0$ *,* $Cov(\epsilon) = \sigma^2 I$ *, and* $Rank(Z) = r + 1$*. For any c, the estimator* $c^{'}\hat{\beta}$ *of* $c^{'}\beta$ *has the smallest possible variance among all linear estimators of the form* $a^{'}Y$ *that are unbiased for* $c^{'}\beta$*.*

Proof. For any fixed c, let $a^{'}Y$ be an unbiased estimator of $c^{'}\beta$. Then $E(a^{'}Y) = c^{'}\beta$, whatever the value of β. But $E(a^{'}Y) = E(a^{'}(Z\beta + \epsilon)) = a^{'}Z\beta$. Consequently, $c^{'}\beta = a^{'}Z\beta$ or equivalently, $(c^{'} - a^{'}Z)\beta = 0$ for all β.

In particular , choosing $\beta = (c^{'} - a^{'}Z)$, we obtained $c^{'} = a^{'}Z$ for any unbiased estimator.

Next $c^{'}\hat{\beta} = c^{'}(Z^{'}Z)^{-1}Z^{'}Y = a^{*}Y$, where $a^{*} = Z(Z^{'}Z)^{-1}c$

Since $E(\hat{\beta}) = \beta$, as $c^{'}\hat{\beta} = a^{*}Y$ is an unbiased estimator of $c^{'}\beta$. the result of the last paragraph says that $c^{'} = (a^{*})^{'}Z$.

Finally for any $'a'$ satisfying the unbiased requirement $c^{'} = a^{'}Z$, we have

$$
\begin{aligned}
V(a^{'}Y) &= V(a^{'}Z\beta + a^{'}\epsilon) \\
&= V(a^{'}\epsilon) \\
&= \sigma^2 \underline{a^{'}a} \\
&= \sigma^2[(a - a^{*} + a^{*})^{'}(a - a^{*} + a^{*})] \\
&= \sigma^2[(a - a^{*})^{'}(a - a^{*}) + a^{*'}a^{*}]
\end{aligned}
$$

where,

$$\begin{aligned}
(a-a^*)^{'}a^* &= (a-a^*)^{'}Z(Z^{'}Z)^{-1}c \\
&= (a^{'}Z - a^{*'}Z)(Z^{'}Z)^{-1}c \\
&= (c^{'} - c^{'})(Z^{'}Z)^{-1}c \\
&= 0.
\end{aligned}$$

Because a^* is fixed and $(a-a^*)^{'}(a-a^*)$ is positive unless $a = a^*$, $Var(a^{'}Y)$ is minimized by the choice of $a = a^*$ and we have,

$$a^{*'}Y = c^{'}(Z^{'}Z)^{-1}Z^{'}Y = c^{'}\hat{\beta}$$

The result says that the LSE $c^{'}\beta$ is the best linear unbiased estimator $(BLUE)$ for $c^{'}\beta$. ∎

Inference

For the multiple linear regression model we further assume that $\underline{\epsilon} \sim N_n(0, \sigma^2 I)$.

Result : $\hat{\beta}$ is also maximum likelihood estimate of β. In addition, $\hat{\beta} \sim N_{\overline{r+1}}\left[\beta, \sigma^2(Z^{'}Z)^{-1}\right]$ and is independent of the residuals $\hat{\epsilon} = \left(Y - Z\hat{\beta}\right)$. Let $\tilde{\sigma}^2$ be the maximum likelihood estimate of σ^2. Then

$$n\tilde{\sigma}^2 = \hat{\epsilon}^{'}\hat{\epsilon} \sim \sigma^2 \chi^2_{\overline{n-r-1}}$$

Note that the MLE of $\tilde{\sigma}^2$ of σ^2 is $\frac{\hat{\epsilon}^{'}\hat{\epsilon}}{n}$, which is different from the LSE of $\hat{\sigma}^2$.

Proof. The model is

$$\underline{Y} = X\underline{\beta} + \underline{\epsilon} \quad \text{where} \quad \underline{\epsilon} \sim N_n(0, \sigma^2 I),\ \underline{Y} \sim N_n(X\underline{\beta}, \sigma^2 I)$$

$$f(\underline{y}; \underline{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma^n} e^{-\frac{1}{2\sigma^2}(Y-X\beta)^{'}(Y-X\beta)}$$

Maximizing $f(\underline{y};\underline{\beta},\sigma^2)$ is equivalent to minimizing $(Y-X\beta)^{'}(Y-X\beta)$ i.e. the LSE when σ^2 is given.

Therefore, $\hat{\beta}$ is also maximum likelihood estimate of β.

Now, incorporating $\hat{\beta}$ in the likelihood we get log-likelihood of σ^2 as

$$L(\sigma^2) = -n\ln\sigma - \frac{1}{2\sigma^2}(Y-X\beta)^{'}(Y-X\beta)$$

Now, taking derivative w.r.t. σ^2 we get $\sigma^2 = \frac{\hat{\epsilon}^{'}\hat{\epsilon}}{n}$

Therefore maximum likelihood estimate of σ^2 is $\hat{\sigma}^2 = \frac{\hat{\epsilon}^{'}\hat{\epsilon}}{n}$ which is different from least square estimate of $\sigma^2 = \left(\frac{\hat{\epsilon}^{'}\hat{\epsilon}}{n-r-1}\right)$.

$$\hat{\epsilon}^{'}\hat{\epsilon} = (Y-X\hat{\beta})^{'}(Y-X\hat{\beta}) = (Y-X\beta)^{'}(Y-X\beta) - (\hat{\beta}-\beta)X^{'}X(\hat{\beta}-\beta)$$

Using moment generating function we can prove $\hat{\epsilon}^{'}\hat{\epsilon} \sim \chi^2_{n-r-1}$ as $(Y-X\beta)^{'}(Y-X\beta) \sim \chi^2_n$ and $(\hat{\beta}-\beta)^{'}X^{'}X(\hat{\beta}-\beta) \sim \chi^2_{r+1}$

∎

<u>Result :</u> For Gaussian MLR model, a $100(1-\alpha)\%$ confidence region for β is given by

$$(\beta-\hat{\beta})^{'}Z^{'}Z(\beta-\hat{\beta}) \leq (r+1)s^2F_{r+1,n-r-11}(\alpha)$$

Where, $s^2 = \left(\frac{\hat{\epsilon}^{'}\hat{\epsilon}}{n-r-1}\right)$ is the LSE of σ^2.

Also' simultaneous $100(1-\alpha)\%$ confidence intervals for the β_j are

$$\hat{\beta}_i \pm \sqrt{Var(\hat{\beta}_i)}\sqrt{(r+1)F_{r+1,\overline{n-r-1}}(\alpha)}\ ; \quad i = 0,1,\cdots,r$$

Where, $Var(\hat{\beta}_i)$ is the diagonal element of $s^2(Z^{'}Z)^{-1}$ corresponding to β_i.

Proof. Consider the vector $V = (Z^{'}Z)^{1/2}(\hat{\beta}-\beta)$ which is normally distributed with mean zero and variance matrix $\sigma^2 I_{\overline{r+1}}$.

Consequently $\qquad V'V \sim \chi^2_{r+1}.$

In addition,

$$(n-r-1)s^2 = \hat{\epsilon}'\hat{\epsilon} \sim \sigma^2 \chi^2_{n-r+1}$$

and is independent of V. The result then follows. ∎

Likelihood ratio tests for the regression parameters :

Consider

$$H_0 : \beta_{q+1} = \beta_{q+2} = \cdots = \beta_r = 0 \quad vs \quad H_a : \beta_i \neq 0 \quad \text{for some} \;\; q+1 \leq i \leq r \tag{2}$$

under H_0, the model is

$$Y = Z_1\beta_1 + \underline{\epsilon} \tag{3}$$

under H_a, the model is

$$Y = Z_1\beta_1 + Z_2\beta_2 + \underline{\epsilon} \tag{4}$$

where $Z = [Z_1 \;\; Z_2]$, $\qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$

<u>Result :</u> Let Z have full rank $(r+1)$ and $\epsilon \sim N_n(0, \sigma^2 I)$. The likelihood ratio test the null hypothesis $H_0 : \beta_2 = 0$ is

$$\frac{\frac{\left[SS_{res}(Z_1) - SS_{res}(Z_2)\right]}{(r-q)}}{s^2} \sim F_{r-q,n-r-1}$$

Where, $SS_{res}(Z_1)$ and $SS_{res}(Z_2)$ are the sum of squares of the models in (3) and (4), respectively.

Proof. Under the model in (4), the maximized likelihood function is

$$L(\hat{\beta}, \hat{\sigma}^2) = \frac{1}{(2\pi)^{n/2}\hat{\sigma}^n} e^{-(n/2)}$$

. where, $\hat{\beta} = (Z'Z)^{-1}Z'Y$ and $\hat{\sigma^2} = \frac{(Y-Z\hat{\beta})'(Y-Z\hat{\beta})}{n}$ on the other hand under the sub-model in (3), the maximize likelihood function is

$$L(\hat{\beta}_1, \hat{\sigma_1}^2) = \frac{1}{(2\pi)^{n/2}\hat{\sigma_1}^n} e^{-(n/2)}$$

Where, $\hat{\beta}_1 = (Z_1'Z_1)^{-1}Z_1'Y$ and $\hat{\sigma_1^2} = \frac{(Y-Z_1\hat{\beta}_1)'(Y-Z_1\hat{\beta}_1)}{n}$

Thus, the likelihood ratio is

$$\frac{L(\hat{\beta}_1, \hat{\sigma_1}^2)}{L(\hat{\beta}, \hat{\sigma}^2)} = \left(\frac{\hat{\sigma^2}}{\hat{\sigma_1^2}}\right)$$

which gives rise to the test statistic

$$\frac{\frac{n(\hat{\sigma_1^2} - \hat{\sigma^2})}{(r-q)}}{\frac{n\hat{\sigma_1^2}}{(n-r-1)}} = \frac{\frac{\left[SS_{res}(Z_1) - SS_{res}(Z_2)\right]}{(r-q)}}{s^2} \sim F_{r-q, n-r-1}$$

Alternatively, one can construct a matrix C such that the null hypothesis becomes $H_0 : C\beta = 0$.

In this way, $C\hat{\beta} \sim N_{r-q}(C\beta, \sigma^2 C(Z'Z)^{-1}C')$ which can be used to perform the test. ■

Inferences From the Fitted Model

Consider a specific point of interest, say $z_0 = (1, z_{01}, \cdots, z_{0r})'$, in the design-matrix space. Then model says $E\big(Y_0|z_0\big) = z_0'\beta$, and LSE of this expection is $z_0'\hat{\beta}$.

In addition, $Var(z_0'\hat{\beta}) = \sigma^2 z_0'(Z'Z)^{-1}z_0$. Consequently, under the normality assumption, a $100(1-\alpha)\%$ confidence interval for $z_0'\beta$ is

$$z_0'\hat{\beta} \pm t_{n-r-1}(\alpha/2)\sqrt{z_0'(Z'Z)^{-1}z_0 s^2}$$

Forecasting : The point prediction of Y at z_0 is $z_0'\hat{\beta}$, which is an unbiased estimator, Since $Y_0 = z_0'\beta + \epsilon_0$, the variance of the forecast is $\sigma^2(1 + Z'(Z'Z)^{-1}z_0)$, where we use the property $\hat{\beta}$ and ϵ_0 are uncorrelated. Therefore, a $100(1-\alpha)\%$ prediction interval for Y_0 is

$$z_0'\hat{\beta} \pm t_{n-r-1}(\alpha/2)\sqrt{s^2(1+z_0'(Z'Z)^{-1}z_0)}$$

.