Multivariate Statistical Analysis, Exercise 9, Fall 2011, Prof. S.K. Jeng

December 18, 2011 TA: H.C. Cheng

(Statistical Properties of Linear Regression with Least Squares)

When we use the linear regression model with least squares to fit the underlying data as

$$y_{n\times 1} = X_{n\times p+1}\beta_{p+1\times 1} + \epsilon_{n\times 1} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1,p} \\ 1 & x_{21} & x_{22} & \dots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{n,p} \end{pmatrix} \beta_{p+1\times 1} + \epsilon_{n\times 1},$$

and

$$\hat{\beta} = (X^T X)^{-1} X^T y,$$

we also concern with the statistical properties of $\hat{\beta}$. There are several questions:

(1) Is the least square estimate $\hat{\beta}$ unbiased?

(2) What is the covariance matrix of $\hat{\beta}$?

(3) How to estimate the the covariance matrix?

(4) How to infer about β?

Before answering this questions, we must have a statistical model for the deviation (or called noise) ϵ. The simplest model, which we will refer to as the standard statistical model:

$$\begin{aligned} \mathbb{E}(\epsilon) &= 0 \\ \mathrm{Cov}(\epsilon) &= \sigma^2 I_{(p+1)\times(p+1)}. \end{aligned}$$

Note that if we make no assumption, we will hardly get anything. There is no free lunch! However, how to make the statistical model and how goodness of it fits to the practical data is the core problem called "Model Selection" in Statistical Learning Theory. One my use Vapnik-Chervonenkis (VC) dimension techniques or Minimum Description Length method to address this. However, this is beyond the scope of this class. We only assume the standard statistical model.

Now we derive

$$\begin{aligned} \hat{\beta} &= \mathbb{E}[(X^T X)^{-1} X^T Y] \\ &= \mathbb{E}[(X^T X)^{-1} X^T (X\beta + \epsilon)] \\ &= \beta + (X^T X)^{-1} X^T \mathbb{E}(\epsilon) = \beta. \end{aligned}$$

Hence the estimate $\hat{\beta}$ is unbiased.

Before derive the covariance matrix of $\hat{\beta}$, recall the identity that with a fixed matrix A, the covariance matrix of $Z = AY$ is $\Sigma_{ZZ} = A\Sigma_{YY}A^T$. Hence the covariance matrix of the estimate $\hat{\beta}$ is

$$\begin{aligned}\Sigma_{\hat{\beta}\hat{\beta}} &= (X^TX)^{-1}X^{\Sigma}_{\epsilon\epsilon}X(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}.\end{aligned}$$

However, in order the use the above formula for further development, σ^2 is sometimes known and needed to be estimated. First, we calculus the sum of squared residuals:

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 &= \|Y - PY\|^2 \\ &= \|(I-P)Y\|^2 \\ &= Y^T(I-P)^T(I-P)Y \\ &= Y^T(I-P)Y,\end{aligned}$$

where $P = X(X^TX)^{-1}X^T$. Second, we use the lemma:

Lemma Let Z be a random vector with mean μ and covariance Σ, and let A be a fixed matrix. Then

$$\mathbb{E}(Z^TAZ) = \text{trace}(A\Sigma) + \mu^TA\mu.$$

The proof can be easily derived by comparing both side element-wise. Hence

$$\mathbb{E}(Y^T(I-P)Y) = [\mathbb{E}(Y)]^T(I-P)[\mathbb{E}(Y)] + \sigma^2\text{trace}(I-P).$$

Now $\mathbb{E}(Y) = X\beta$, so

$$(I-P)\mathbb{E}(Y) = [I - X(X^TX)^{-1}X^T]X\beta = 0.$$

Furthermore, use the identity that

$$\text{trace}(I-P) = \text{trace}(I) - \text{trace}(P),$$

and the community of trace (*i.e.* $\text{trace}(AB) = \text{trace}(BA)$, we have

$$\begin{aligned}\text{trace}(P) &= \text{trace}[X(X^TX)^{-1}X^T] \\ &= \text{trace}[X^TX(X^TX)^{-1}] \\ &= \text{trace}(I_{(p+1)\times(p+1)}) = p+1.\end{aligned}$$

We have

$$\mathbb{E}(\|Y - \hat{Y}\|^2) = (n - (p+1))\sigma^2.$$

Hence the unbiased estimate of σ^2 is

$$s^2 = \frac{\|Y - \hat{Y}\|^2}{n - (p+1)}.$$

This result will be used to construct confidence intervals and the hypothesis testing that will be exact under the assumption of normality (because $\hat{\beta}$ may be expressed as a linear combination of the independent random variables ϵ_i, a version of the central limit theorem with certain assumptions on X implies the approximate result).

Under the normality assumption, it can be shown as the t-test:

$$t = \frac{\hat{\beta}_i - \beta_i}{\mathrm{SE}(\hat{\beta}_i)} \sim t_{n-(p+1)},$$

where $\mathrm{SE}(\hat{\beta}_i)$ denotes the standard deviation of $\hat{\beta}_i$. It follows that a $100(1-\alpha)\%$ confidence interval for β_i is

$$\hat{\beta}_i \pm t_{1-\alpha/2;n-(p+1)}\mathrm{SE}(\hat{\beta}).$$

Similarly, in testing $\beta_i = 0$, we reject the hypothesis at the significance level α if $|t| \geq t_{1-\alpha/2;n-(p+1)}$. (Note that $t_{1-\alpha/2;n}$ denotes $1-\alpha/2$ quantile of the t-distribution with n degrees of freedom.)

Now with the data from Exercise 8:

	Sales	Price	Advert.	Ass. Hours
1	230	125	200	109
2	181	99	55	107
3	165	97	105	98
4	150	115	85	71
5	97	120	0	82
6	192	100	150	103
7	181	80	85	111
8	189	90	120	93
9	172	95	110	86
10	170	125	130	78

we know that $\hat{\beta} = [65.6696, -0.2158, 0.4852, 0.8437]^T$. With significance level $\alpha = 0.1$, in testing $\beta_i = 0 \quad i = 2, 3, 4$, do we reject the hypothesis?