Math Review Notes—Linear Regression

Gregory Faletto

Contents

Last updated August 12, 2020

Chapter 1

Linear Regression

These notes are based on my notes from *Time Series and Panel Data Econometrics* (1st edition) by M. Hashem Pesaran [Pesaran, 2015] and coursework for Economics 613: Economic and Financial Time Series I at USC taught by M. Hashem Pesaran, DSO 607 at USC taught by Jinchi Lv, Statistics 100B at UCLA taught by Nicolas Christou, GSBA 604: Regression and Generalized Linear Models for Business Applications at USC taught by Gourab Mukherjee, and the Coursera MOOC "Econometrics: Methods and Applications" from Erasmus University Rotterdam. I also borrowed from some other sources which I mention when I use them.

1.1 Chapter 1: Linear Regression

1.1.1 Preliminaries

Suppose the true model is $y_i = \alpha + \beta x_i + \epsilon_i$. Classical assumptions:

(i) $\mathbb{E}(\epsilon_i) = 0$

(ii) $\text{Var}(\epsilon_i \mid x_i = \sigma^2$ (constant)

(iii) $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ if $i \neq j$

(iv) ϵ_i is uncorrelated to x_i, or $\mathbb{E}(\epsilon_i \mid x_j) = 0$ for all i, j.

1.1.2 Estimation

$$\hat{\beta} = \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \frac{\sum_{i=1}^n x_i y_i - n\overline{xy}}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$

or

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{S_{XY}}{S_{XX}}$$

or

$$\hat{\beta} = r\frac{S_{YY}}{S_{XX}}$$

where r is the correlation coefficient.

Let

$$w_i = \frac{x_i - \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

so that

$$\hat{\beta} = \sum_{i=1}^n w_i(y_i - \overline{y}) = \sum_{i=1}^n w_i y_i - \overline{y}\frac{\sum_{i=1}^n x_i - \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} = \sum_{i=1}^n w_i y_i$$

since $\sum_{i=1}^n x_i - \overline{x} = 0$. Then a simple expression for $\text{Var}(\hat{\beta})$ is

$$\text{Var}(\hat{\beta}) = \sum_{i=1}^n w_i^2 \text{Var}(y_i \mid x_i) = \sum_{i=1}^n w_i^2 \text{Var}(\epsilon \mid x_i) = \sigma^2 \sum_{i=1}^n w_i^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sigma^2}{S_{XX}}$$

We can estimate these quantities as follows:

$$\hat{\sigma}^2 = \frac{1}{n-2} \cdot \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

Note that

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{t=1}^T (y_t - \hat{\alpha} - \hat{\beta} x_t)^2 = \frac{1}{n-2}\sum_{t=1}^T \left[(y_t - (\overline{y} - \hat{\beta}\overline{x}) - \hat{\beta} x_t)^2\right] = \frac{1}{n-2}\sum_{t=1}^T (y_t - \overline{y} - \hat{\beta}(x_t - \overline{x}))^2$$

$$= \frac{1}{n-2}\sum_{t=1}^T (y_t - \overline{y})^2 - 2\hat{\beta}(x_t - \overline{x})(y_t - \overline{y}) + \hat{\beta}^2 (x_t - \overline{x})^2$$

In the case where there is no intercept, we have

$$\hat{\sigma}^2 = \frac{1}{T-1}\sum_{t=1}^T (y_t - \hat{\beta} x_t)^2 = \frac{1}{T-1}\sum_{t=1}^T \left(y_t^2 - 2r\frac{S_{YY}}{S_{XX}} x_t y_t + r^2 \frac{S_{YY}^2}{S_{XX}^2} x_t^2\right)$$

Also,

$$\widehat{\text{Var}}(\hat{\beta}) = \frac{\hat{\sigma}^2}{S_{XX}} = \frac{1}{n-2} \cdot \frac{\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

Correlation coefficient:

$$r^2 = \frac{\left(\sum_{t=1}^T x_t y_t\right)^2}{\sum_{t=1}^T x_t^2 \sum_{t=1}^T y_t^2}$$

$$r = \frac{1}{T-1} \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$$

Remark. The formulas for the coefficients in univariate OLS can also be derived by considering (x, y) as a bivariate normal distribution and calculating the conditional expectation of y given x. (See Proposition (**??**).)

Proposition 1.1.2.1 (**Stats 100B homework problem**)**.** Consider the regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with x_i fixed and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, ϵ_i i.i.d. Let $e_i = y_i - \hat{y}_i$ be the residuals.

(a)

$$\sum_{i=1}^n e_i = 0$$

(b) $\text{Cov}(\overline{Y}, \hat{\beta}_1) = 0$ where $\overline{Y}$ is the sample mean of the y values.

(c)

$$\text{Cov}(e_i, e_j) = \sigma^2 \left(- \frac{1}{n} - \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^n (x_k - \bar{x})^2} \right)$$

(d) We can construct a confidence interval for σ^2 as

$$\Pr\left(\frac{\sum_{i=1}^n e_i^2}{\chi^2_{1-\frac{\alpha}{2}; n-2}} \leq \sigma^2 \leq \frac{\sum_{i=1}^n e_i^2}{\chi^2_{\frac{\alpha}{2}; n-2}} \right) = 1 - \alpha$$

Proof. (a)

$$\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n (y_i - [\bar{y} + \hat{\beta}_1 (x_i - \bar{x})])$$

$$= \sum_{i=1}^n \left(y_i - \bar{y} - \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} (x_i - \bar{x}) \right) = \sum_{i=1}^n y_i - n\bar{y} - \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} \sum_{i=1}^n (x_i - \bar{x})$$

$$= \sum_{i=1}^n y_i - n \frac{1}{n} \sum_{i=1}^n y_i - \left(\frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} \right) \left[\sum_{i=1}^n \left(x_i - \frac{1}{n} \sum_{i=1}^n x_i \right) \right]$$

$$= \sum_{i=1}^n (y_i - y_i) - \left(\frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} \right) \left[\sum_{i=1}^n x_i - \frac{1}{n} \cdot n \sum_{i=1}^n x_i \right] = 0 - 0 = \boxed{0}$$

Or:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n}(y_i - (\bar{y} - \hat{\beta}_1\bar{x}) - \hat{\beta}_1 x_i)$$

$$= \sum_{i=1}^{n}(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

(b)

$$\text{Cov}(\bar{Y}, \hat{\beta}_1) = \text{Cov}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i, \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) = \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\text{Cov}\left(\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n}(x_i - \bar{x})Y_i\right)$$

x_i is fixed, $\text{Cov}(Y_i, Y_j) = 0$ for $i \neq j$ by assumption of the model, $\text{Var}(Y_i) = \sigma^2$ by assumption of the model.

$$= \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\sum_{i=1}^{n}[(x_i - \bar{x})\text{Var}(Y_i)] = \frac{\sigma^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\sum_{i=1}^{n}(x_i - \bar{x}) = \boxed{0}$$

(c)

$$\text{Cov}(e_i, e_j) = \text{Cov}(y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x}), y_j - \bar{y} - \hat{\beta}_1(x_j - \bar{x}))$$

$$= \text{Cov}(y_i, y_j) - \text{Cov}(y_i, \bar{y}) - \text{Cov}(y_i, \hat{\beta}_1(x_j - \bar{x})) - \text{Cov}(\bar{y}, y_j) + \text{Cov}(\bar{y}, \bar{y}) + \text{Cov}(\bar{y}, \hat{\beta}_1(x_j - \bar{x})) - \text{Cov}(\hat{\beta}_1(x_i - \bar{x}), y_j)$$

$$+ \text{Cov}(\hat{\beta}_1(x_i - \bar{x}), \bar{y}) + \text{Cov}(\hat{\beta}_1(x_i - \bar{x}), \hat{\beta}_1(x_j - \bar{x}))$$

By assumption of the model, $\text{Cov}(y_i, y_j) = 0$.

$$= 0 - \text{Cov}(y_i, \bar{y}) - (x_j - \bar{x})\text{Cov}(y_i, \hat{\beta}_1) - \text{Cov}(\bar{y}, y_j) + \text{Var}(\bar{y}) + (x_j - \bar{x})\text{Cov}(\bar{y}, \hat{\beta}_1) - (x_i - \bar{x})\text{Cov}(\hat{\beta}_1, y_j)$$

$$+ (x_i - \bar{x})\text{Cov}(\hat{\beta}_1, \bar{y}) + (x_i - \bar{x})(x_j - \bar{x})\text{Cov}(\hat{\beta}_1, \hat{\beta}_1)$$

In part 7(b) we showed $\text{Cov}(\bar{y}, \hat{\beta}_1) = 0$. $\text{Var}(\bar{y}) = \sigma^2/n$. $\text{Cov}(\hat{\beta}_1, \hat{\beta}_1) = \text{Var}(\hat{\beta}_1) = \sigma^2/\sum(x_k - \bar{x})^2$. So this simplifies to

$$= -\text{Cov}(y_i, \bar{y}) - (x_j - \bar{x})\text{Cov}(y_i, \hat{\beta}_1) - \text{Cov}(y_j, \bar{y}) + \frac{\sigma^2}{n} + 0 - (x_i - \bar{x})\text{Cov}(y_j, \hat{\beta}_1) + 0 + (x_i - \bar{x})(x_j - \bar{x})\frac{\sigma^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

$$= -\text{Cov}(y_i, \bar{y}) - (x_j - \bar{x})\text{Cov}(y_i, \hat{\beta}_1) - \text{Cov}(y_j, \bar{y}) + \frac{\sigma^2}{n} - (x_i - \bar{x})\text{Cov}(y_j, \hat{\beta}_1) + (x_i - \bar{x})(x_j - \bar{x})\frac{\sigma^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2} \tag{1.1}$$

Find $\text{Cov}(y_i, \bar{y}), \text{Cov}(y_j, \bar{y}), \text{Cov}(y_i, \hat{\beta}_1)$, and $\text{Cov}(y_j, \hat{\beta}_1)$:

Using that x_i is fixed, $\text{Cov}(Y_i, Y_j) = 0$ for $i \neq j$ by assumption of the model, $\text{Var}(Y_i) = \sigma^2$ by assumption of the model:

$$\text{Cov}(y_i, \bar{y}) = \text{Cov}\left(y_i, \frac{1}{n}\sum_{k=1}^{n} y_k\right) = \frac{1}{n}\text{Cov}(y_i, y_i) = \frac{\sigma^2}{n}$$

Similarly,

$$\text{Cov}(y_j, \bar{y}) = \frac{\sigma^2}{n}$$

$$\text{Cov}(y_i, \hat{\beta}_1) = \text{Cov}\left(y_i, \frac{\sum_{k=1}^{n}(x_k - \bar{x})y_i}{\sum_{k=1}^{n}(x_k - \bar{x})^2}\right) = \frac{1}{\sum_{k=1}^{n}(x_k - \bar{x})^2}\text{Cov}\left(y_i, \sum_{k=1}^{n}(x_k - \bar{x})y_i\right)$$

$$= \frac{1}{\sum_{k=1}^{n}(x_k - \bar{x})^2}\text{Cov}(y_i, (x_i - \bar{x})y_i) = \frac{x_i - \bar{x}}{\sum_{k=1}^{n}(x_k - \bar{x})^2}\text{Var}(y_i) = \frac{x_i - \bar{x}}{\sum_{k=1}^{n}(x_k - \bar{x})^2}\sigma^2$$

Similarly,

$$\text{Cov}(y_j, \hat{\beta}_1) = \frac{x_j - \bar{x}}{\sum_{k=1}^{n}(x_k - \bar{x})^2}\sigma^2$$

Plugging these in to equation (1.1) yields

$$\text{Cov}(e_i, e_j) = -\frac{\sigma^2}{n} - (x_j - \bar{x})\frac{(x_i - \bar{x})\sigma^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2} - \frac{\sigma^2}{n} + \frac{\sigma^2}{n} - (x_i - \bar{x})\frac{(x_j - \bar{x})\sigma^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

$$+(x_i - \bar{x})(x_j - \bar{x})\frac{\sigma^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

$$= \frac{-\sigma^2}{n} - \sigma^2\frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

$$\text{Cov}(e_i, e_j) = \sigma^2\left(-\frac{1}{n} - \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2}\right)$$

(d) From class notes 08/29:

$$\frac{(n-2)S_e^2}{\sigma^2} \sim \chi^2_{n-2}$$

$$\implies \Pr\left(\chi^2_{\frac{\alpha}{2};n-2} \leq \frac{(n-2)S_e^2}{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2};n-2}\right) = 1 - \alpha$$

$$\implies \boxed{\Pr\left(\frac{(n-2)S_e^2}{\chi^2_{1-\frac{\alpha}{2};n-2}} \leq \sigma^2 \leq \frac{(n-2)S_e^2}{\chi^2_{\frac{\alpha}{2};n-2}}\right) = 1 - \alpha}$$

Since

$$S_e^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

this interval can be expressed as

$$\Pr\left(\frac{\sum_{i=1}^n e_i^2}{\chi^2_{1-\frac{\alpha}{2};n-2}} \leq \sigma^2 \leq \frac{\sum_{i=1}^n e_i^2}{\chi^2_{\frac{\alpha}{2};n-2}}\right) = 1-\alpha$$

□

Proposition 1.1.2.2 (Stats 100B homework problem). Suppose $Y_i = \beta_1 x_i + \epsilon_i$ (no intercept). Suppose x_i is fixed and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

(a) The maximum likelihood estimator of β_1 is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

which is unbiased. Its variance is $\frac{\sigma^2}{\sum_{i=1}^n x_i^2}$ and it is normally distributed.

(b) The maximum likelihood estimator of σ^2 is

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (y_i - \beta_1 x_i)^2.$$

Proof. (a) First we find the likelihood function to find the MLE. Assuming the n observations are independent,

$$L = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta_1 x_i)^2\right)$$

$$= (2\sigma^2\pi)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_1 x_i)^2\right)$$

Next,

$$\log(L) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_1 x_i)^2$$

$$\frac{\mathrm{d}\log(L)}{\mathrm{d}\beta_1} = \frac{\mathrm{d}}{\mathrm{d}\beta_1}\left(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_1 x_i)^2\right)$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^n x_i(y_i - \beta_1 x_i) = 0$$

$$\sum_{i=1}^n x_i y_i - \hat{\beta_1}\sum_{i=1}^n x_i^2 = 0$$

$$\implies \hat{\beta_1} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

Next we show that this estimator is unbiased.

$$\mathbb{E}(\hat{\beta}_1) = \mathbb{E}\Big(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\Big) = \frac{1}{\sum_{i=1}^n x_i^2}\mathbb{E}\Big(\sum_{i=1}^n x_i(\beta_1 x_i + \epsilon_i)\Big) = \frac{1}{\sum_{i=1}^n x_i^2}\Big[\mathbb{E}\Big(\sum_{i=1}^n x_i^2\beta_1\Big) + E\Big(\sum_{i=1}^n x_i\epsilon_i\Big)\Big]$$

Since x_i and β_1 are non-random and ϵ_i are independent, this can be written as

$$\frac{1}{\sum_{i=1}^n x_i^2}\Big[\sum_{i=1}^n x_i^2\beta_1 + \sum_{i=1}^n x_i\mathbb{E}(\epsilon_i)\Big] = \frac{1}{\sum_{i=1}^n x_i^2}\beta_1\sum_{i=1}^n x_i^2 = \beta_1$$

Next we find the variance.

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\Big(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\Big) = \frac{1}{(\sum_{i=1}^n x_i^2)^2}\mathrm{Var}\Big(\sum_{i=1}^n x_i(\beta_1 x_i + \epsilon_i)\Big)$$

$$= \frac{1}{(\sum_{i=1}^n x_i^2)^2}\Big[\mathrm{Var}\Big(\sum_{i=1}^n x_i^2\beta_1\Big) + \mathrm{Var}\Big(\sum_{i=1}^n x_i\epsilon_i\Big)\Big]$$

Since x_i and β_1 are non-random and ϵ_i are independent, this can be written as

$$\frac{1}{(\sum_{i=1}^n x_i^2)^2}\Big[0 + \sum_{i=1}^n x_i^2\mathrm{Var}(\epsilon_i)\Big] = \frac{1}{(\sum_{i=1}^n x_i^2)^2}\sigma^2\sum_{i=1}^n x_i^2 = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$$

β_1 is a linear combination of y_i which is normally distributed, therefore β_1 is normally distributed.

$$\implies \beta_1 \sim \mathcal{N}\Big(\beta_1, \frac{\sigma}{\sqrt{\sum_{i=1}^n x_i^2}}\Big)$$

(b)

$$\frac{\mathrm{d}\log(L)}{\mathrm{d}\sigma^2} = \frac{\mathrm{d}}{\mathrm{d}\sigma^2}\Big(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n(y_i - \beta_1 x_i)^2\Big)$$

$$= -\frac{n}{2}\frac{1}{2\pi\sigma^2}2\pi - \frac{1}{2}\Big(-\frac{1}{(\sigma^2)^2}\Big)\sum_{i=1}^n(y_i - \beta_1 x_i)^2 = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n(y_i - \beta_1 x_i)^2 = 0$$

$$\frac{1}{2(\hat{\sigma^2})^2}\sum_{i=1}^n(y_i - \beta_1 x_i)^2 = \frac{n}{2\hat{\sigma^2}}$$

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n(y_i - \beta_1 x_i)^2$$

□

Remark. More details on this problem available in Math 541A Homework 7.

1.2 Chapter 2: Multiple Regression

General OLS:

$$\hat{\beta} = (\boldsymbol{X}'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + u) = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u = \beta + (X'X)^{-1}X'u$$

$$\text{Var}(\hat{\beta}) = \text{Var}(\beta + (X'X)^{-1}X'u) = \text{Var}(\beta) + \text{Var}((X'X)^{-1}X'u) = 0 + \mathbb{E}[(X'X)^{-1}X'uu'X(X'X)^{-1}]$$

$$= \mathbb{E}\big[(X'X)^{-1}X'\mathbb{E}(uu' \mid X)X(X'X)^{-1}\big] = \sigma^2\mathbb{E}\big[(X'X)^{-1}X'I_TX(X'X)^{-1}\big] = \sigma^2\mathbb{E}\big[(X'X)^{-1}\big]$$

$$= \sigma^2(X'X)^{-1}$$

Or:

$$\text{Var}[(X^TX)^{-1}X^Ty] = (X^TX)^{-1}X^T\text{Var}[y]X(X^TX)^{-1} = (X^TX)^{-1}X^T[\sigma^2I_n]X(X^TX)^{-1}$$

$$= \sigma^2(X^TX)^{-1}$$

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{T-k}$$

Proposition 1.2.0.1 (**GSBA 604 Problem**)**.** A necessary and sufficient condition for $a'\beta$ to be estimable in a linear model is that a is in the column space of X.

Proof. Let $a \in \mathbb{R}^{(p+1)\times k}$. A linear combination $a^T\beta \in \mathbb{R}^k$ is estimable if and only if there exists a $\gamma \in \mathbb{R}^{n\times k}$ so that the linear combination γ^Ty satisfies $\mathbb{E}(\gamma^Ty) = a^T\beta$ for all β. That is,

$$\mathbb{E}(\gamma^Ty) = a^T\beta \iff \gamma^TX\beta = a^T\beta \iff X^T\gamma = a;$$

that is, the columns of a lie in the column space of X^T.

□

Proposition 1.2.0.2 (**GSBA 604 Problem**)**.** Suppose $a'\beta$ is estimable in a linear model. Then its best linear unbiased estimator (BLUE) is unique.

Proof. Let $a \in \mathbb{R}^{(p+1)\times k}$. The BLUE of $a'\beta \in \mathbb{R}^k$ is a random variable solving the equation

$$\underset{b\in\mathbb{R}^k, \mathbb{E}(b)=a'\beta}{\text{minimize}} \quad \text{Var}(b) = \underset{b\in\mathbb{R}^k, \mathbb{E}(b)=a'\beta}{\text{minimize}} \quad \mathbb{E}(b-\mathbb{E}(b))^2 = \underset{b\in\mathbb{R}^k, \mathbb{E}(b)=a'\beta}{\text{minimize}} \quad \mathbb{E}(b-a'\beta)^2$$

This objective is strictly convex in b. From Exercise 1 we know that the set of vectors $b \in \mathbb{R}^k$ satisfying $\mathbb{E}(b) = \alpha^T\beta$ are linear combinations of vectors in the column space of X^T; that is, this set of vectors is the column space of X^T, which is a subspace and therefore convex. Therefore the minimizer of this problem exists and is unique.

□

Remark. By Theorem 1.2.0.5, the best linear unbiased estimator is $A\hat{\beta}_{\text{OLS}}$. We have

$$A\hat{\beta}_{\text{OLS}} = A(A^TA)^{-1}A^Ty$$

which is unique if it exists (if A^TA is invertible).

Theorem 1.2.0.3 (Gauss-Markov Theorem, as stated in Pesaran [2015]). Suppose we have data generated by

$$y = X\beta + \epsilon$$

and we make the following assumptions.

1. $\mathbb{E}(\epsilon) = 0$.
2. Homoskedasticity: $\text{Var}(\epsilon_i \mid X) = \sigma^2 > 0 \qquad \forall i$
3. Uncorrelated errors: $\text{Cov}(\epsilon_i, \epsilon_j \mid X) = 0, \qquad \forall i \neq j$.
4. Orthogonality: $\mathbb{E}(\epsilon_i \mid X) = 0, \qquad \forall i$.

Then if $X\beta$ is estimable, then the best linear unbiased estimator (BLUE) of β is $\hat{\beta}_{OLS}$, the least squares estimate. That is, if $\tilde{\beta}$ is an alternative linear unbiased estimator, where $\tilde{\beta} = \hat{\beta}_{OLS} + C^Ty$ with $C \in \mathbb{R}^{n\times p}$ and $\mathbb{E}(\tilde{\beta}) = \beta$ for all β, then $\text{Var}(\tilde{\beta} \mid X) \geq \text{Var}(\hat{\beta}_{OLS} \mid X)$.

Proof (adapted from Pesaran [2015]). Note that $\text{Var}(\hat{\beta}_{OLS} \mid X) \leq \text{Var}(\tilde{\beta} \mid X) \iff \text{Var}(\tilde{\beta} \mid X) - \text{Var}(\hat{\beta}_{OLS} \mid X)$ is a positive semidefinite matrix. We have

$$\tilde{\beta} = [(X^TX)^{-1}X^T + C^T]y = [(X^TX)^{-1}X^T + C^T](X\beta + \epsilon) = \beta + (X^TX)^{-1}X^T\epsilon + C^TX\beta + C^T\epsilon$$

Since $\mathbb{E}(\tilde{\beta}) = \beta$ for all β, we have $C^TX\beta = 0$ for all β, so $C^TX = 0$. (Note that since $n \geq p$, $C^T \in \mathbb{R}^{p\times n}$ has at least an $n - p$-dimensional nullspace.)

$$\implies \tilde{\beta} = \beta + (X^TX)^{-1}X^T\epsilon + C^T\epsilon \iff \tilde{\beta} - \beta = [(X^TX)^{-1}X^T + C^T]\epsilon \tag{1.2}$$

Then

$$\text{Var}(\tilde{\beta}) = \mathbb{E}\left[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^T\right] = \mathbb{E}\left[\tilde{\beta}\tilde{\beta}^T - \tilde{\beta}\beta^T - \beta\tilde{\beta}^T + \beta\beta^T\right] \tag{1.3}$$

Using (1.2) we have

$$\mathbb{E}\left[\tilde{\beta}\tilde{\beta}^T\right] = \mathbb{E}\left[(\beta + (X^TX)^{-1}X^T\epsilon + C^T\epsilon)(\beta + (X^TX)^{-1}X^T\epsilon + C^T\epsilon)^T\right]$$

$$= \mathbb{E}\big[\beta\beta^T + \beta\epsilon^TX(X^TX)^{-1} + \beta\epsilon^TX + (X^TX)^{-1}X^T\epsilon\beta^T + (X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}$$

$$+(X^TX)^{-1}X^T\epsilon\epsilon^TC + C^T\epsilon\beta^T + C^T\epsilon\epsilon^TX(X^TX)^{-1} + C^T\epsilon\epsilon^TC\big]$$

$$= \beta\beta^T + (X^TX)^{-1}X^T\mathbb{E}\left[\epsilon\epsilon^T\right]X(X^TX)^{-1} + (X^TX)^{-1}X^T\mathbb{E}\left[\epsilon\epsilon^T\right]C + C^T\mathbb{E}\left[\epsilon\epsilon^T\right]X(X^TX)^{-1} + C^T\mathbb{E}\left[\epsilon\epsilon^T\right]C$$

$$= \beta\beta^T + \sigma^2(X^TX)^{-1} + \sigma^2(X^TX)^{-1}X^TC + \sigma^2C^TX(X^TX)^{-1} + \sigma^2C^TC$$

$$= \beta\beta^T + \sigma^2(X^TX)^{-1} + \sigma^2C^TC \tag{1.4}$$

Also,

$$\tilde{\beta}\beta^T = (\beta + (X^TX)^{-1}X^T\epsilon + C^T\epsilon)\beta^T = \beta\beta^T + (X^TX)^{-1}X^T\epsilon\beta^T + C^T\epsilon\beta^T$$

$$\implies \mathbb{E}(\tilde{\beta}\beta^T) = \beta\beta^T, \qquad \mathbb{E}\left(\beta\tilde{\beta}^T\right) = \mathbb{E}\left(\left[\tilde{\beta}\beta^T\right]^T\right) = \beta\beta^T. \tag{1.5}$$

Substituting (1.4) and (1.5) into (1.3), we have

$$\text{Var}(\tilde{\beta}) = \beta\beta^T + \sigma^2(X^TX)^{-1} + \sigma^2C^TC - 2\beta\beta^T + \beta\beta^T$$

$$= \sigma^2\left[(X^TX)^{-1} + C^TC\right]$$

Therefore (using $\text{Var}(\hat{\beta}_{OLS}) = \sigma^2(X^TX)^{-1}$)

$$\text{Var}(\tilde{\beta}) - \text{Var}(\hat{\beta}_{OLS}) = \sigma^2\left[(X^TX)^{-1} + C^TC - (X^TX)^{-1}\right] = \sigma^2C^TC$$

which is a positive semidefinite matrix since the inner product of a matrix with itself is positive semidefinite (see Proposition **??**).

□

Theorem 1.2.0.4 (Gauss-Markov Theorem, as stated in Faraway [2002]). Suppose

$$y = X\beta + \epsilon$$

with $X \in \mathbb{R}^{n \times p}$ a fixed full rank matrix and $n \geq p$, $\beta \in \mathbb{R}^p$, and $\epsilon \in \mathbb{R}^n$ with $\mathbb{E}(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2 I_n$. Suppose X is fixed ($\mathbb{E}(Y) = X\beta$). Let $\psi : \mathbb{R}^p \to \mathbb{R}$ be an estimable function: $\psi := c^T\beta$ for some $c \in \mathbb{R}^p$. Then in the class of all unbiased linear estimates of ψ, $\hat{\psi} = c^T\hat{\beta}$ has the minimum variance and is unique.

Proof (adapted from Faraway [2002]). Suppose for some $a \in \mathbb{R}^n$, $a^T y$ is another unbiased estimator of $c^T\beta$ so that

$$\mathbb{E}(a^T y) = a^T X\beta = c^T\beta, \qquad \forall \beta \in \mathbb{R}^p.$$

This implies

$$a^T X = c^T \iff X^T a = c, \tag{1.6}$$

Since X has full column rank, $X^T X$ is full rank and its column space is all of $\mathbb{R}^p$, so there exists a unique $\lambda \in \mathbb{R}^p$ such that

$$c = X^T X\lambda = X^T a \tag{1.7}$$

(Note that $a = X\lambda + k$ where $k \in \mathbb{R}^n$ is any vector in the $n - p$-dimensional nullspace of X^T.) Then

$$\text{Var}(a^T y) = \text{Var}(a^T y - c^T\hat{\beta} + c^T\hat{\beta}) = \text{Var}(a^T y - \lambda^T X^T\hat{y} + c^T\hat{\beta})$$

$$= \text{Var}(a^T y - \lambda^T X^T\hat{y}) + \text{Var}(c^T\hat{\beta}) + 2\text{Cov}(a^T y - \lambda^T X^T\hat{y}, c^T\hat{\beta}) \geq \text{Var}(c^T\hat{\beta}) + 2\text{Cov}(a^T y - \lambda^T X^T\hat{y}, c^T\hat{\beta}), \tag{1.8}$$

so we are done if the covariance in (1.8) is nonnegative.

$$\text{Cov}(a^T y - \lambda^T X^T\hat{y}, c^T\hat{\beta}) = \mathbb{E}\left[\left(a^T y - \lambda^T X^T\hat{y} - \mathbb{E}[a^T y - \lambda^T X^T\hat{y}]\right)\left(c^T\hat{\beta} - \mathbb{E}[c^T\hat{\beta}]\right)\right]$$

$$= \mathbb{E}\left[\left(a^T(X\beta + \epsilon) - \lambda^T X^T X\hat{\beta} - a^T X\beta + \lambda^T X^T X\beta\right)\left(c^T\hat{\beta} - c^T\beta\right)\right]$$

$$= \mathbb{E}\left[\left(a^T\epsilon - \lambda^T X^T X(\hat{\beta} - \beta)\right)c^T\left(\hat{\beta} - \beta\right)\right] = \mathbb{E}\left[\left(a^T\epsilon - \lambda^T X^T X(X^T X)^{-1}X^T\epsilon\right)c^T(X^T X)^{-1}X^T\epsilon\right]$$

$$= \mathbb{E}\left[a^T\epsilon c^T(X^T X)^{-1}X^T\epsilon\right] - \mathbb{E}\left[\lambda^T X^T\epsilon c^T(X^T X)^{-1}X^T\epsilon\right]$$

$$= (a^T - \lambda^T X^T)\mathbb{E}\left[\epsilon c^T (X^T X)^{-1} X^T \epsilon\right] = \left(a^T - \lambda^T X^T\right)\mathbb{E}\left[\epsilon(\lambda^T X^T \epsilon)\right]$$

Note that

$$\lambda^T X^T \epsilon = \sum_{j=1}^{n} (\lambda^T X^T)_j \epsilon_j \in \mathbb{R}^n$$

$$\implies \left(a^T - \lambda^T X^T\right)\mathbb{E}\left[\epsilon(\lambda^T X^T \epsilon)\right] = \left(a^T - \lambda^T X^T\right)\mathbb{E}\begin{bmatrix}\epsilon_1[\lambda^T X^T \epsilon] \\ \vdots \\ \epsilon_n[\lambda^T X^T \epsilon]\end{bmatrix}$$

$$= \left(a^T - \lambda^T X^T\right)\mathbb{E}\begin{bmatrix}\epsilon_1 \sum_{j=1}^n (\lambda^T X^T)_j \epsilon_j \\ \vdots \\ \epsilon_n \sum_{j=1}^n (\lambda^T X^T)_j \epsilon_j\end{bmatrix}$$

Using independence of ϵ_i and ϵ_j for $i \neq j$, we can write this as

$$= \left(a^T - \lambda^T X^T\right)\mathbb{E}\begin{bmatrix}\epsilon_1 (\lambda^T X^T)_1 \epsilon_1 \\ \vdots \\ \epsilon_n (\lambda^T X^T)_j n \epsilon_n\end{bmatrix} = \left(a^T - \lambda^T X^T\right)\sigma^2\begin{bmatrix}(\lambda^T X^T)_1 \\ \vdots \\ (\lambda^T X^T)_j n\end{bmatrix} = \sigma^2\left(a^T - \lambda^T X^T\right)X\lambda$$

$$= \sigma^2\left(a^T X - \lambda^T X^T X\right)\lambda = 0$$

since by (1.7) $c^T = a^T X = \lambda^T X^T X$.

□

Theorem 1.2.0.5 (Gauss-Markov Theorem, as stated in Faraway [2002], more general case). Suppose

$$y = X\beta + \epsilon$$

with $X \in \mathbb{R}^{n \times p}$ a fixed full rank matrix and $n \geq p$, $\beta \in \mathbb{R}^p$, and $\epsilon \in \mathbb{R}^n$ with $\mathbb{E}(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2 I_n$. Suppose X is fixed ($\mathbb{E}(Y) = X\beta$). Let $\psi : \mathbb{R}^p \to \mathbb{R}^k$ be an estimable function: $\psi := c^T\beta$ for some full rank $c \in \mathbb{R}^{p \times k}$. Then in the class of all unbiased linear estimates of ψ, $\hat{\psi} = c^T\hat{\beta}$ has the minimum variance and is unique. That is, for some $a \in \mathbb{R}^{n \times k}$ such that $a^T y$ is another unbiased estimator of $c^T\beta$,

$$\text{Var}(a^T y) \succeq \text{Var}(c^T\hat{\beta}).$$

Proof (adapted from Faraway [2002]). First, note that

$$\mathbb{E}(a^T y) = a^T X\beta = c^T\beta, \qquad \forall \beta \in \mathbb{R}^p.$$

This implies

$$a^T X = c^T \iff X^T a = c, \tag{1.9}$$

Since X has full column rank, $X^T X$ is full rank and its column space is all of $\mathbb{R}^p$, so there exists a unique $\lambda \in \mathbb{R}^{p \times k}$ such that

$$c = X^T X \lambda = X^T a \tag{1.10}$$

(Note that $a = X\lambda + b$ for some $b \in \mathbb{R}^{n \times k}$ whose columns lie in the $(n-p)$-dimensional nullspace of X^T.) Then

$$\text{Var}(a^T y) = \text{Var}(a^T y - c^T \hat{\beta} + c^T \hat{\beta}) = \text{Var}(a^T y - \lambda^T X^T \hat{y} + c^T \hat{\beta})$$

$$= \text{Var}(a^T y - \lambda^T X^T \hat{y}) + \text{Var}(c^T \hat{\beta}) + 2\text{Cov}(a^T y - \lambda^T X^T \hat{y}, c^T \hat{\beta}) \succeq \text{Var}(c^T \hat{\beta}) + 2\text{Cov}(a^T y - \lambda^T X^T \hat{y}, c^T \hat{\beta}), \tag{1.11}$$

so we are done if the covariance matrix in (1.11) is positive semidefinite.

$$\text{Cov}(a^T y - \lambda^T X^T \hat{y}, c^T \hat{\beta}) = \mathbb{E}\left[\left(a^T y - \lambda^T X^T \hat{y} - \mathbb{E}[a^T y - \lambda^T X^T \hat{y}]\right) \left(c^T \hat{\beta} - \mathbb{E}[c^T \hat{\beta}]\right)^T \right]$$

$$= \mathbb{E}\left[\left(a^T (X\beta + \epsilon) - \lambda^T X^T X \hat{\beta} - a^T X \beta + \lambda^T X^T X \beta\right) \left(c^T [\hat{\beta} - \beta]\right)^T \right]$$

$$= \mathbb{E}\left[\left(a^T \epsilon - \lambda^T X^T X (\hat{\beta} - \beta)\right) \left(\hat{\beta} - \beta\right)^T c \right] = \mathbb{E}\left[\left(a^T \epsilon - \lambda^T X^T X (X^T X)^{-1} X^T \epsilon\right) \epsilon^T X (X^T X)^{-1} c \right]$$

$$= \mathbb{E}\left[a^T \epsilon \epsilon^T X (X^T X)^{-1} c \right] - \mathbb{E}\left[\lambda^T X^T \epsilon \epsilon^T X (X^T X)^{-1} c \right]$$

$$= \sigma^2 a^T I_n X (X^T X)^{-1} c - \sigma^2 \lambda^T X^T I_n X (X^T X)^{-1} c = \sigma^2 \left(a^T X - \lambda^T X^T X\right) \lambda = 0$$

since by (1.10) $c^T = a^T X = \lambda^T X^T X$.

□

1.3 Chapter 3: Hypothesis testing in regression

In this section, I borrow from C. Flinn's notes "Asymptotic Results for the Linear Regression Model," available online at http://www.econ.nyu.edu/user/flinnc/notes1.pdf.

Proposition 1.3.0.1.

$$\hat{\beta} - \beta \sim \mathcal{N}_{p+1}\left(0, \sigma^2 (X^T X)^{-1}\right)$$

Proof.

$$\hat{\beta} = (X^T X)^{-1} X^T (X\beta + \epsilon) = \beta + (X^T X)^{-1} X^T \epsilon \sim \mathcal{N}_{p+1}\left(\beta, \sigma^2 (X^T X)^{-1} X^T I_n X (X^T X)^{-1}\right)$$

$$\iff \hat{\beta} - \beta \sim \mathcal{N}_{p+1}\left(0, \sigma^2 (X^T X)^{-1}\right) \implies \hat{\beta}_j - \beta_j \sim \mathcal{N}\left(0, \sigma^2 \left[(X^T X)^{-1}\right]_{jj}\right)$$

$$\iff \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 \left[(X^T X)^{-1}\right]_{jj}}} \sim \mathcal{N}(0,1). \tag{1.12}$$

□

Proposition 1.3.0.2.

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}.$$

Proof.

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\hat{\epsilon}^T\hat{\epsilon} = \frac{1}{n-p-1}\left[(I_n - P_X)(X\beta + \epsilon)\right]^T \left[(I_n - P_X)(X\beta + \epsilon)\right]$$

But $P_X X = X(X^T X)^{-1} X^T X = X \implies (I - P_X)X\beta = X\beta - X\beta = 0$, so we can write this as

$$= \frac{1}{n-p-1}\left[(I_n - P_X)\epsilon\right]^T \left[(I_n - P_X)\epsilon\right] = \frac{1}{n-p-1}\epsilon^T (I_n - P_X)^T (I_n - P_X)\epsilon = \frac{1}{n-p-1}\epsilon^T (I - P_X)\epsilon \tag{1.13}$$

where we used the fact that $I_n - P_X$ is symmetric and idempotent. We have

$$\mathrm{Tr}(I_n - P_X) = \mathrm{Tr}(I) - \mathrm{Tr}(P_x) = n - \mathrm{Tr}(X(X^T X)^{-1} X^T) = n - \mathrm{Tr}((X^T X)^{-1} X^T X) = n - \mathrm{Tr}(I_{p+1}) = n - p - 1.$$

Take the spectral decomposition of $I_n - P_X$:

$$I_n - P_X = QDQ^T$$

where $Q \in \mathbb{R}^{n \times n}$ is orthogonal and $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix containing the eigenvalues of $I_n - P_X$. Since $I_n - P_X$ is idempotent, all of its eigenvalues equal either 0 or 1, so it must have $n - p - 1$ eigenvalues equal to 1 and $p + 1$ eigenvalues equal to 0. That is,

$$D = \begin{bmatrix} I_{n-p-1} & 0 \\ 0 & 0 \end{bmatrix}$$

Therefore we can write (1.13) as

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\epsilon^T QDQ^T\epsilon = \frac{1}{n-p-1}\epsilon^T QDQ^T\epsilon = \frac{1}{n-p-1}(Q^T\epsilon)^T DQ^T\epsilon = \frac{1}{n-p-1}B^T B \tag{1.14}$$

where $B \in \mathbb{R}^n$ is defined by

$$B_i = \begin{cases} (Q^T\epsilon)_i & i \in \{1,\dots,n-p-1\} \\ 0 & i \in \{n-p,\dots,n\}. \end{cases}$$

Note that since $\epsilon \sim \mathcal{N}_n(0,\sigma^2 I_n)$,

$$Q^T\epsilon \sim \mathcal{N}_n\left(0, Q^T(\sigma^2 I_n)Q\right) = \mathcal{N}_n\left(0,\sigma^2 I_n\right)$$

(since $Q^TQ = I_n$), so

$$\frac{B_i}{\sigma} \sim \begin{cases} \mathcal{N}_{n-p-1}\left(0, I_{n-p-1}\right) & i \in \{1,\dots,n-p-1\} \\ 0 & i \in \{n-p,\dots,n\}. \end{cases} \tag{1.15}$$

Using (1.14) and (1.15) we have

$$(n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} = \frac{B^TB}{\sigma^2} \sim \chi^2_{n-p-1}. \tag{1.16}$$

□

Proposition 1.3.0.3 (GSBA 604 Problem).

$$\frac{(\hat{\beta}-\beta)^T X^T X(\hat{\beta}-\beta)}{(p+1)\hat{\sigma}^2} \sim F(p_1, n-p-1).$$

Proof. Recall that an F distribution can be defined as

$$X = \frac{U/d_1}{V/d_2} \implies X \sim F_{d_1,d_2}$$

where $U \sim \chi^2_{d_1}$, $V \sim \chi^2_{d_2}$, and $U \perp\!\!\!\perp V$. By Proposition 1.3.0.1,

$$\hat{\beta} - \beta \sim \mathcal{N}_{p+1}\left(0, \sigma^2(X^TX)^{-1}\right)$$

$$\implies \frac{\sqrt{X^TX}}{\sigma}(\hat{\beta}-\beta) \sim \mathcal{N}_{p+1}\left(0, I_{p+1}\right)$$

$$\implies \left[\sqrt{X^TX}/\sigma(\hat{\beta}-\beta)\right]^T\left[\sqrt{X^TX}/\sigma(\hat{\beta}-\beta)\right] = \frac{1}{\sigma^2}(\hat{\beta}-\beta)^T\left(\sqrt{X^TX}\right)^T\sqrt{X^TX}(\hat{\beta}-\beta)$$

$$= \frac{1}{\sigma^2}(\hat{\beta}-\beta)^TX^TX(\hat{\beta}-\beta) \sim \chi^2_{p+1}.$$

By Proposition 1.3.0.2,

$$(n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}.$$

Therefore

$$\frac{(\hat{\beta}-\beta)^TX^TX(\hat{\beta}-\beta)}{(p+1)\hat{\sigma}^2} = \frac{(\hat{\beta}-\beta)^TX^TX(\hat{\beta}-\beta)/[\sigma^2(p+1)]}{(n-p-1)\hat{\sigma}^2/[\sigma^2(n-p-1)]} \sim F_{p+1,n-p-1}.$$

□

Proposition 1.3.0.4. The elements of $\hat{\beta}$ are uncorrelated with (and independent of) the elements of $\hat{\epsilon}$.

Proof.

$$\text{Cov}(\hat{\beta},\hat{\epsilon}) = \mathbb{E}\left[(\hat{\beta}-\mathbb{E}(\hat{\beta}))(\hat{\epsilon}-\mathbb{E}(\hat{\epsilon}))^T\right]$$

$$= \mathbb{E}\left[((X^TX)^{-1}X^T(X\beta+\epsilon)-\beta)((I-P_x)(X\beta+\epsilon)-\mathbb{E}((I-P_x)(X\beta+\epsilon)))^T\right]$$

$$= \mathbb{E}\left[(\beta+(X^TX)^{-1}X^T\epsilon)-\beta)(X\beta+\epsilon-X\beta-P_x\epsilon-\mathbb{E}[X\beta+\epsilon-X\beta-P_x\epsilon])^T\right]$$

$$= \mathbb{E}\left[(X^TX)^{-1}X^T\epsilon(\epsilon-X(X^TX)^{-1}X^T\epsilon)^T\right]$$

$$= \mathbb{E}\left[(X^TX)^{-1}X^T\epsilon\epsilon^T-(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}X^T\right]$$

$$= \sigma^2(X^TX)^{-1}X^T-\sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1}X^T$$

$$= \sigma^2(X^TX)^{-1}X^T-\sigma^2(X^TX)^{-1}X^T = 0.$$

Since $\hat{\beta}$ and $\hat{\epsilon}$ are Gaussian and uncorrelated, they are independent.

□

Proposition 1.3.0.5.

$$\frac{\hat{\beta}_j-\beta_j}{\text{se}(\hat{\beta}_j)} \sim t_{n-p-1}.$$

Proof. By Proposition 1.3.0.1,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 \left[(X^T X)^{-1}\right]_{jj}}} \sim \mathcal{N}(0,1). \tag{1.17}$$

Recall that we define the *standard error* of $\hat{\beta}_j$ to be an estimate of the standard deviation of $\hat{\beta}_j$ that shows up in the denominator of (1.17) that uses the plug-in estimator $\hat{\sigma}^2$ for σ^2:

$$\text{se}(\hat{\beta}_j) := \sqrt{\hat{\sigma}^2 \left[(X^T X)^{-1}\right]_{jj}}. \tag{1.18}$$

By Proposition 1.3.0.4, $\hat{\beta} \perp\!\!\!\perp \hat{\epsilon}$, and since functions of independent random vectors are also independent and $\hat{\sigma}^2$ is a function of $\hat{\epsilon}$ and not $\hat{\beta}$,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 \left[(X^T X)^{-1}\right]_{jj}}} \perp\!\!\!\perp (n-p-1)\frac{\hat{\sigma}^2}{\sigma^2}. \tag{1.19}$$

Recall that a Student's t distribution can be defined as

$$T = \frac{Z}{\sqrt{V/v}} \implies T \sim t_v$$

where $Z \sim \mathcal{N}(0,1)$, $V \sim \chi^2_v$, and $Z \perp\!\!\!\perp V$. Using this characterization along with Proposition 1.3.0.2 (use (1.16)),(1.17), (1.18), and (1.19), we have

$$\frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 \left[(X^T X)^{-1}\right]_{jj}}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 \left[(X^T X)^{-1}\right]_{jj}}} \Bigg/ \sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}$$

$$= \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 \left[(X^T X)^{-1}\right]_{jj}}} \Bigg/ \sqrt{(n-p-1)\,\frac{\hat{\sigma}^2}{\sigma^2} \Big/ (n-p-1)} \sim t_{n-p-1}.$$

□

Lemma 1.3.0.6.

$$\frac{1}{n} \cdot X'\epsilon \xrightarrow{p} 0$$

Proof. Note that $\mathbb{E}\frac{1}{n} \cdot X'\epsilon = 0$ for any n. Then we have

$$\text{Var}\left(\frac{1}{n} \cdot X'\epsilon\right) = \mathbb{E}\left(\frac{1}{n} \cdot X'\epsilon\right)^2 = n^{-2}\mathbb{E}(X'\epsilon\epsilon' X) = n^{-2}\mathbb{E}(\epsilon\epsilon')X'X = \frac{\sigma^2}{n}\frac{X'X}{n}$$

implying that $\lim_{n\to\infty} \text{Var}\left(\frac{1}{n} \cdot X'\epsilon\right) = 0$. Therefore the result follows from Chebyshev's Inequality (Theorem **??**). □

Lemma 1.3.0.7. If ϵ is i.i.d. with $E(\epsilon_i) = 0$ and $\mathbb{E}(\epsilon_i^2) = \sigma^2$ for all i, the elements of the matrix X are uniformly bounded so that $|X_{ij}| < U$ for all i and j and for U finite, and $\lim_{n\to\infty} X'X/n = Q$ is finite and nonsingular, then

$$\frac{1}{\sqrt{n}} X'\epsilon \xrightarrow{d} \mathcal{N}(0, \sigma^2 Q)$$

Proof. If we have one regressor, then $n^{-1/2}\sum_{i=1}^n X_i\epsilon_i$ is a scalar. Let G_i be the cdf of $X_i\epsilon_i$. Let

$$S_n^2 = \sum_{i=1}^n \text{Var}(X_i\epsilon_i) = \sigma^2 \sum_{i=1}^n X_i^2$$

In this scalar case, $Q = \lim_{n\to\infty} n^{-1}\sum_i X_i^2$. By the Lindberg-Feller Theorem, a necessary and sufficient condition for $Z_n \to \mathcal{N}(0\sigma^2 Q)$ is

$$\lim_{n\to\infty} \frac{1}{S_n^2} \sum_{i=1}^n \int_{|\omega|>\nu S_n} \omega^2 dG_i(\omega) = 0$$

for all $\nu > 0$. Now $G_i(\omega) = F(\omega/|X_i|)$. Then rewrite the above equation as

$$\lim_{n\to\infty} \frac{n}{S_n^2} \sum_{i=1}^n \frac{X_i^2}{n} \int_{|\omega/X_i|>\nu S_n/|X_i|} \left(\frac{\omega}{X_i}\right)^2 dF(\omega/|X_i|) = 0$$

Since $\lim_{n\to\infty} S_n^2 = \lim_{n\to\infty} n\sigma^2 \sum_{i=1}^n X_i^2/n = n\sigma^2 Q$, we have $\lim_{n\to\infty} n/S_n^2 = (\sigma^2 Q)^{-1}$, which is a finite and nonzero scalar. Then we need to show

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n X_i^2 \delta_{i,n} = 0$$

where

$$\delta_{i,n} = \int_{|\omega/X_i|>\nu S_n/|X_i|} \left(\frac{\omega}{X_i}\right)^2 dF(\omega/|X_i|)$$

But $\lim_{n\to\infty} \delta_{i,n} = 0$ for all i and any fixed ν since $|X_i|$ is bounded while $\lim_{n\to\infty} X_n = \infty$, so the measure of the set $\{|\omega/X_i| > \nu S_n/|X_i|\}$ goes to 0 asymptotically. Since $\lim_{n\to\infty} n^{-1}\sum_i X_i^2$ is finite and $\lim_{n\to\infty} \delta_{i,n} = 0$ for all i, $\lim_{n\to\infty} n^{-1}\sum_i X_i^2\delta_{i,n} = 0$, so $\frac{1}{n}\cdot X'\epsilon \xrightarrow{p} 0$.

□

Theorem 1.3.0.8. Under the conditions of Lemma 1.3.0.7 (ϵ is i.i.d. with $E(\epsilon_i) = 0$ and $\mathbb{E}(\epsilon_i^2) = \sigma^2$ for all i, the elements of the matrix X are uniformly bounded so that $|X_{ij}| < U$ for all i and j and for U finite, and $\lim_{n\to\infty} X'X/n = Q$ is finite and nonsingular),

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma^2 Q^{-1})$$

Proof.

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{X'X}{n}\right)^{-1} \frac{1}{\sqrt{n}} X'\epsilon$$

Since $\lim_{n\to\infty}(X'X/n)^{-1} = Q^{-1}$ and by Lemma 1.3.0.7

$$\frac{1}{\sqrt{n}} X'\epsilon \xrightarrow{d} \mathcal{N}(0, \sigma^2 Q)$$

then

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma^2 Q^{-1} Q Q^{-1}) = \mathcal{N}(0, \sigma^2 Q^{-1})$$

□

t-test statistic:

$$t = \frac{\hat{\beta} - 0}{s.e.(\hat{\beta})}$$

F-test statistic:

$$F = \left(\frac{T - k - 1}{r}\right)\left(\frac{SSR_R - SSR_U}{SSR_U}\right)$$

Since

$$R^2 = \frac{\sum_t (y_t - \overline{y})^2 - \sum_t (y_t - \hat{y}_t)^2}{\sum_t (y_t - \overline{y})^2} = \frac{\sum_t (y_t - \overline{y})^2 - SSR_U}{\sum_t (y_t - \overline{y})^2}$$

we have

$$SSR_U = \sum_t (y_t - \overline{y})^2 - R^2 \sum_t (y_t - \overline{y})^2 = (1 - R^2) \sum_t (y_t - \overline{y})^2$$

yielding

$$F = \left(\frac{T - k - 1}{r}\right)\left(\frac{\sum_t (y_t - \overline{y})^2 - (1 - R^2)\sum_t (y_t - \overline{y})^2}{(1 - R^2)\sum_t (y_t - \overline{y})^2}\right) = \left(\frac{T - k - 1}{r}\right)\left(\frac{R^2}{1 - R^2}\right)$$

Confidence interval for sums of coefficients. (Two coefficient case.) Suppose we want to test H_0 : $\beta_1 + \beta_2 = k$. Let $\delta = \beta_1 + \beta_2 - k$, $\hat{\delta} = \hat{\beta}_1 + \hat{\beta}_2 - k$. Note that under the null hypothesis $\delta = 0$. We can construct a t-statistic

$$t_{\hat{\delta}} = \frac{\hat{\delta} - 0}{\sqrt{\hat{\text{Var}}(\hat{\delta})}} = \frac{\hat{\beta}_1 + \hat{\beta}_2 - k}{\sqrt{\hat{\text{Var}}(\hat{\delta})}}$$

where

$$\hat{\text{Var}}(\hat{\delta}) = \hat{\text{Var}}(\hat{\beta}_1) + \hat{\text{Var}}(\hat{\beta}_2) + 2\hat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2)$$

This means that a 95% confidence interval for δ can be constructed in the following way:

$$\hat{\delta} \pm t^* \sqrt{\hat{\text{Var}}(\hat{\delta})}$$

where t^* is the 95% critical value for the t-distribution.

1.3.1 ANOVA

$$g_1 = lm(R \sim x_1 + x_2 + x_3 + x_4), \qquad g_2 = lm(R \sim x_2 + x_3),$$

$$y = \beta_0 + \sum_{i=1}^{4} \beta_i x_i$$

$$H_0 : \beta_1 = \beta_4 = 0, \qquad H_A : \text{ at least one of } \beta_1, \beta_4 \text{ is not zero.}$$

1.4 Chapter 4: Heteroskedasticity

Under heteroskedasticity, the OLS estimator $\hat{\beta} = (X'X)^{-1}X'y$ is unbiased, but the true covariance matrix of $\hat{\beta}$ no longer matches the OLS formula. For instance, suppose we have

$$y_t = \sum_{i=1}^{K} \beta_i x_{ti} + u_t$$

where $\text{Var}(u_t) = \sigma^2 z_t^2$.

$$\hat{\beta} = (X'X)^{-1}X'y = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u = \beta + (X'X)^{-1}X'u$$

$$\implies \mathbb{E}(\hat{\beta}) = \mathbb{E}[\beta] + (X'X)^{-1}X'\mathbb{E}[u] = \beta$$

since $\mathbb{E}(u)$ is still 0. However,

$$\text{Var}(\hat{\beta}) = \mathbb{E}\big[\big(\hat{\beta} - \mathbb{E}(\hat{\beta})\big)\big(\hat{\beta} - \mathbb{E}(\hat{\beta})\big)'\big] = \mathbb{E}\big[\big(\beta + (X'X)^{-1}X'u - \beta\big)\big(\beta + (X'X)^{-1}X'u - \beta\big)'\big]$$

$$= \mathbb{E}\big[\big((X'X)^{-1}X'u\big)\big((X'X)^{-1}X'u\big)'\big] = \mathbb{E}\big[(X'X)^{-1}X'uu'X\big((X'X)^{-1}\big)'\big]$$

$$= (X'X)^{-1}X'\mathbb{E}[uu' \mid X]X(X'X)^{-1}$$

$$= (X'X)^{-1}X' \begin{bmatrix} \sigma^2 z_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 z_2^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2 z_3^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 z_T^2 \end{bmatrix} X(X'X)^{-1}$$

$$= \sigma^2 (X'X)^{-1}X' \begin{bmatrix} z_1^2 & 0 & 0 & \dots & 0 \\ 0 & z_2^2 & 0 & \dots & 0 \\ 0 & 0 & z_3^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & z_T^2 \end{bmatrix} X(X'X)^{-1}$$

which is different from the OLS estimator of the covariance matrix $\sigma^2(X'X)^{-1}$. Therefore the estimate of the variances of $\hat{\beta}$ will be biased if the OLS formulas are used, and the usual t and F tests for $\hat{\beta}$ will be invalid.

1.5 Chapter 5: Autocorrelated disturbances

1.5.1 Generalized Least Squares

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

where

$$\mathbb{E}(\boldsymbol{u} \mid \boldsymbol{X}) = 0 \ \ \forall \ t$$

$$\mathbb{E}(\boldsymbol{u}\boldsymbol{u}' \mid \boldsymbol{X}) = \boldsymbol{\Sigma}$$

where $\boldsymbol{\Sigma}$ is a positive definite matrix.

Suppose

$$y = X\beta + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \Sigma)$. Then

$$\Sigma^{-1/2}y = \Sigma^{-1/2}X\beta + \tilde{\epsilon}$$

where $\tilde{\epsilon} = \Sigma^{-1/2}\epsilon \sim \mathcal{N}(0, 1)$. Now we can do ordinary least squares since the errors of this transformed model are i.i.d.

$$\hat{\beta}_{GLS} = \left(\left[\Sigma^{-1/2} X \right]^T \Sigma^{-1/2} X \right)^{-1} \left[\Sigma^{-1/2} X \right]^T \Sigma^{-1/2} y = (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} y$$

$$\text{Var}(\hat{\beta}_{GLS}) = (X' \Sigma^{-1} X)^{-1}$$

Example application: spatial data.

1.5.2 Weighted Least Squares

A closely related idea to generalized least squares is weighted least squares. In this model, weights $a_i \in \mathbb{R}_{++}$ are assigned to each observation; higher weights correspond to observations that are more important for model fit, lower weights correspond to observations that are less important for model fit. If all weights equal 1, we recover ordinary least squares. Let $\boldsymbol{A} := \text{diag}(a_1, \dots, a_n) \in \mathbb{R}^{n \times n}$. Then weighted least squares minimizes the weighted average loss

$$\mathcal{L} := \frac{1}{n} \sum_{i=1}^{n} a_i \left(y_i - \beta^\top \boldsymbol{x}_i \right)^2 \tag{1.20}$$

where $\boldsymbol{x}_i \in \mathbb{R}^p$ and $\beta \in \mathbb{R}^p$.

Proposition 1.5.2.1. If $\boldsymbol{X}$ is invertible, the weighted least squares estimator (the unique minimizer of (1.20)) is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{A} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \boldsymbol{A} \boldsymbol{y}.$$

Remark. Note the relationship between generalized least squares and weighted least squares—generalized least squares is weighted least squares if the disturbances are independent and the weight for observation i is the inverse of the variance of the disturbance for observation i.

Proof. Since all the entries of $\boldsymbol{A}$ are on the diagonal and positive, $\boldsymbol{A}^{1/2}$ exists such that $\boldsymbol{A}^{1/2} \boldsymbol{A}^{1/2} = \boldsymbol{A}$. Also, since $\boldsymbol{A}^{1/2}$ only has diagonal entries, it is symmetric, so $\left(\boldsymbol{A}^{1/2} \right)^\top = \boldsymbol{A}^{1/2}$. We have

$$\begin{aligned}
\mathcal{L} &= \frac{1}{n} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{A} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= \frac{1}{n} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{A}^{1/2} \boldsymbol{A}^{1/2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\
&= \frac{1}{n} \left(\boldsymbol{y}^\top \boldsymbol{A}^{1/2} - \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{A}^{1/2} \right) \left(\boldsymbol{A}^{1/2} \boldsymbol{y} - \boldsymbol{A}^{1/2} \boldsymbol{X} \boldsymbol{\beta} \right) \\
&= \frac{1}{n} \left(\boldsymbol{A}^{1/2} \boldsymbol{y} - \boldsymbol{A}^{1/2} \boldsymbol{X} \boldsymbol{\beta} \right)^\top \left(\boldsymbol{A}^{1/2} \boldsymbol{y} - \boldsymbol{A}^{1/2} \boldsymbol{X} \boldsymbol{\beta} \right)
\end{aligned}$$

Let $\boldsymbol{y}^* := \boldsymbol{A}^{1/2} \boldsymbol{y}$ and let $\boldsymbol{X}^* := \boldsymbol{A}^{1/2} \boldsymbol{X}$. Note that since $\boldsymbol{A}^{1/2}$ is full rank (invertible), $\boldsymbol{X}^*$ has the same rank as $\boldsymbol{X}$. Then we can write this as

$$\begin{aligned}
\mathcal{L} &= \frac{1}{n}\left(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}\right)^\top \left(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}\right) \\
&= \frac{1}{n}\left(\left[\boldsymbol{y}^*\right]^\top - \boldsymbol{\beta}^\top \left[\boldsymbol{X}^*\right]^\top\right)\left(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}\right) \\
&= \frac{1}{n}\left(\left[\boldsymbol{y}^*\right]^\top \boldsymbol{y}^* - \left[\boldsymbol{y}^*\right]^\top \boldsymbol{X}^*\boldsymbol{\beta} - \boldsymbol{\beta}^\top \left[\boldsymbol{X}^*\right]^\top \boldsymbol{y}^* + \boldsymbol{\beta}^\top \left[\boldsymbol{X}^*\right]^\top \boldsymbol{X}^*\boldsymbol{\beta}\right) \\
&= \frac{1}{n}\left(\left[\boldsymbol{y}^*\right]^\top \boldsymbol{y}^* - 2\boldsymbol{\beta}^\top \left[\boldsymbol{X}^*\right]^\top \boldsymbol{y}^* + \boldsymbol{\beta}^\top \left[\boldsymbol{X}^*\right]^\top \boldsymbol{X}^*\boldsymbol{\beta}\right)
\end{aligned}$$

Now we take the gradient.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} &= \frac{1}{n}\mathbf{0} - \frac{2}{n}\left[\boldsymbol{X}^*\right]^\top \boldsymbol{y}^* + \frac{2}{n}\left[\boldsymbol{X}^*\right]^\top \boldsymbol{X}^*\boldsymbol{\beta} \\
&= -\frac{2}{n}\left[\boldsymbol{X}^*\right]^\top \boldsymbol{y}^* + \frac{2}{n}\left[\boldsymbol{X}^*\right]^\top \boldsymbol{X}^*\boldsymbol{\beta} \\
&= -\frac{2}{N}\left[\boldsymbol{X}^*\right]^\top \boldsymbol{y}^* + \frac{2}{N}\left[\boldsymbol{X}^*\right]^\top \boldsymbol{X}^*\boldsymbol{\beta} \\
&= -\frac{2}{N}\left[\boldsymbol{A}^{1/2}\boldsymbol{X}\right]^\top \boldsymbol{A}^{1/2}\boldsymbol{y} + \frac{2}{N}\left[\boldsymbol{A}^{1/2}\boldsymbol{X}\right]^\top \boldsymbol{A}^{1/2}\boldsymbol{X}\boldsymbol{\beta} \\
&= -\frac{2}{N}\boldsymbol{X}^\top \boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2}\boldsymbol{y} + \frac{2}{N}\boldsymbol{X}^\top \boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2}\boldsymbol{X}\boldsymbol{\beta} \\
&= -\frac{2}{N}\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{y} + \frac{2}{N}\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}\boldsymbol{\beta}
\end{aligned}$$

Set the gradient equal to 0 and solve for $\boldsymbol{\beta}$:

$$\begin{aligned}
\mathbf{0} &= -\frac{2}{n}\left[\boldsymbol{X}^*\right]^\top \boldsymbol{y}^* + \frac{2}{n}\left[\boldsymbol{X}^*\right]^\top \boldsymbol{X}^*\boldsymbol{\beta} \\
\left[\boldsymbol{X}^*\right]^\top \boldsymbol{y}^* &= \left[\boldsymbol{X}^*\right]^\top \boldsymbol{X}^*\boldsymbol{\beta} \\
\left(\left[\boldsymbol{X}^*\right]^\top \boldsymbol{X}^*\right)^{-1}\left[\boldsymbol{X}^*\right]^\top \boldsymbol{y}^* &= \left(\left[\boldsymbol{X}^*\right]^\top \boldsymbol{X}^*\right)^{-1}\left[\boldsymbol{X}^*\right]^\top \boldsymbol{X}^*\boldsymbol{\beta} \\
\left(\left[\boldsymbol{X}^*\right]^\top \boldsymbol{X}^*\right)^{-1}\left[\boldsymbol{X}^*\right]^\top \boldsymbol{y}^* &= \boldsymbol{\beta}
\end{aligned}$$

Lastly, we substitute back in $\boldsymbol{y}^* = \boldsymbol{A}^{1/2}\boldsymbol{y}$ and $\boldsymbol{X}^* = \boldsymbol{A}^{1/2}\boldsymbol{X}$ to get our final answer.

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\left[\boldsymbol{A}^{1/2}\boldsymbol{X}\right]^\top \boldsymbol{A}^{1/2}\boldsymbol{X}\right)^{-1}\left[\boldsymbol{A}^{1/2}\boldsymbol{X}\right]^\top \boldsymbol{A}^{1/2}\boldsymbol{y} \\
&= \left(\boldsymbol{X}^\top \boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top \boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2}\boldsymbol{y} \\
&= \left(\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{y} \\
&= \left(\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top \boldsymbol{A}\boldsymbol{y}.
\end{aligned}$$

□

1.6 Quantile Regression

Estimate the conditional *median* rather than the conditional mean (as in least squares). Least absolute deviation:

$$\arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} |y_i - x_i'\beta|$$

Problems: no closed form solution; have to solve by linear programming. Good: robust; less changed by fluctuations of outliers. Asymmetric loss:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (1-\tau) \underbrace{(y_i - x_i'\beta)_+}_{\text{under-estimation}} + \tau \underbrace{(x_i'\beta - y_i)_+}_{\text{over-estimation}}$$

In fixed dimensions:

$$\hat{\beta}(\tau) \xrightarrow{d} \mathcal{N}\left(\beta, (X^T X)^{-1} \frac{\tau(1-\tau)}{f_\epsilon^2(0)}\right)$$

Suppose $\epsilon \sim \mathcal{N}(0, \sigma^2)$; then $f_\epsilon(0) = 1/\sqrt{2\pi\sigma^2}$, so we have

$$\hat{\beta}(\tau) \xrightarrow{d} \mathcal{N}\left(\beta, 2\pi\sigma^2 (X^T X)^{-1} \tau(1-\tau)\right)$$

This is then maximized for $\tau = 1/2$, and the max value is $1/4 \cdot 2\pi = \pi/2$. For more information on other loss functions, see also Section **??**.

1.6.1 Detecting outliers in multiple dimensions

Hard because of curse of dimensionality. One thing: project onto lower-dimensional space and then compute distance there. multi-dimensional scaling or dimension reduction methods.

1. random projections
2. nonlinear methods: Iso-map, local linear embedding

Half-space depth (Tukey): make polygons of observations (outer one is the convex hull). Then create convex hulls of inside, keep going. See Figure 1.1.

1.7 Transformed Linear Models

1.7.1 Transformations of response

Consider the transformation

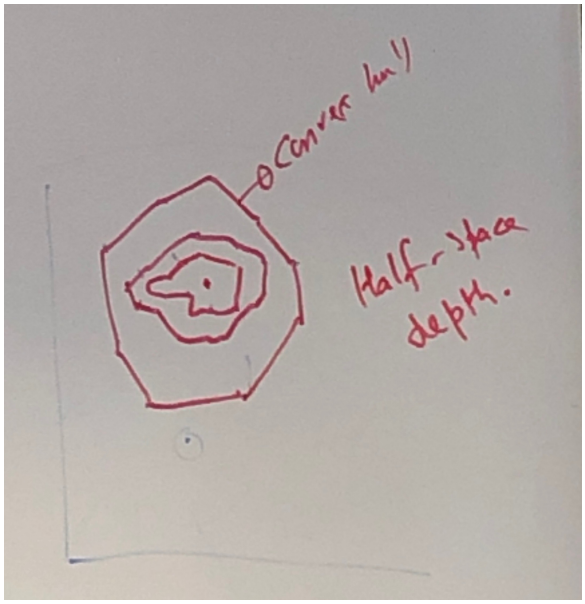

Figure 1.1: Illustration of half-space depth method for detecting outliers.

$$t(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y) = \lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda}, & \lambda = 0 \end{cases}$$

One example: **log linear model.**

$$\log y_i = \beta_0 + \beta_1 x_i + \epsilon.$$

Multiplicative model in the true response:

$$y_i = \exp(\beta_0) \cdot \exp(x_1 \beta_1) \cdot \exp(\epsilon_i)$$

If you increase y_i by one unit, then y_i is multiplied by $\exp(\beta_1)$. So in this case rather than talking about linear effects, you talk about percentage changes in the response: $(e^{\beta_1} - 1) \cdot 100\%$.

Square root transformations work well in Poisson models—variance stabilizes.

Consider

$$t_\lambda = x'\beta + \epsilon.$$

Then we have

$$SSE = (t_\lambda - x_i' \hat{\beta}_{OLS})^T (t_\lambda - x_i' \hat{\beta}_{OLS})$$

Try

$$\frac{n}{2}\log\left(\frac{SSE}{n}\right) + (\sum_{i=1}^{n}\log(y_i).$$

for $\lambda = -2, -2, -1/2, 0, 1/2, 1, 2$. We have $2L(\hat{\lambda}) - 2L(\lambda_{true}) \sim \chi^2_1$. If 1 is inside the 95% confidence interval, don't transform; if 1 isn't, do. In a lot of cases, changing the response can be a pain for interpretability.

1.7.2 Transforming predictor values

Add polynomial terms:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_d x_i^d + \epsilon.$$

Want to maintain hierarchy in selecting variables (backward elimination). Advantages: nonlinear curvature. Global: all data have influence on every predicted point.

Bibliography

J. J. Faraway. *Practical Regression and Anova using R*. 2002.

B. E. Hansen. *Econometrics*. August 2020.

M. H. Pesaran. *Time Series and Panel Data Econometrics*. Number 9780198759980 in OUP Catalogue. Oxford University Press, 2015. ISBN ARRAY(0x3bdaaf68). URL https://ideas.repec.org/b/oxp/obooks/9780198759980.html.