Math Review Notes—Linear Regression

Gregory Faletto

Contents

Last updated February 15, 2019

1 Linear Regression

These notes are based on my notes from *Time Series and Panel Data Econometrics* (1st edition) by M. Hashem Pesaran and coursework for Economics 613: Economic and Financial Time Series I at USC, DSO 607 at USC taught by Jinchi Lv, and Statistics 100B at UCLA taught by Nicolas Christou. 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 (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)
$$\text{Cov}(e_i, e_j) = \sigma^2 \bigg(-\frac{1}{n} - \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^n (x_k - \bar{x})^2} \bigg)$$

(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

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

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

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

Easy way (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,

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

Hard way: Find a matrix $\boldsymbol{A}$ such that $\boldsymbol{A}\vec{Y} = \begin{pmatrix}\vec{Y} & \bar{Y} & \hat{\beta}_1\end{pmatrix}'$. Then the covariance matrix $\text{Var}(\boldsymbol{A}\vec{Y}) = \boldsymbol{A}\text{Var}(\vec{Y})\boldsymbol{A}' = \boldsymbol{A}(\sigma^2\boldsymbol{I})\boldsymbol{A}' = \sigma^2\boldsymbol{A}\boldsymbol{A}'$ will contain $\text{Cov}(y_i, \bar{y}), \text{Cov}(y_j, \bar{y}), \text{Cov}(y_i, \hat{\beta}_1)$, and $\text{Cov}(y_j, \hat{\beta}_1)$.

$$\boxed{\bar{Y} = \frac{1}{n}\mathbf{1}'\vec{Y}}$$

$$\hat{\beta}_1 = \frac{\sum (x_k - \bar{x})y_k}{\sum (x_k - \bar{x})^2}$$

$$\boxed{\hat{\beta}_1 = \frac{\boldsymbol{q}'}{\boldsymbol{q}'\boldsymbol{q}}\vec{Y}}$$

(defining $\boldsymbol{q}$ as in question 7(b). Therefore

$$\begin{pmatrix}\boldsymbol{I} \\ \frac{1}{n}\mathbf{1}' \\ \frac{\boldsymbol{q}'}{\boldsymbol{q}'\boldsymbol{q}}\end{pmatrix}\vec{Y} = \begin{pmatrix}\vec{Y} \\ \bar{Y} \\ \hat{\beta}_1\end{pmatrix}$$

So

$$\boldsymbol{A} = \begin{pmatrix}\boldsymbol{I} \\ \frac{1}{n}\mathbf{1}' \\ \frac{\boldsymbol{q}'}{\boldsymbol{q}'\boldsymbol{q}}\end{pmatrix}$$

To find the variance matrix, we calculate

$$\sigma^2\boldsymbol{A}\boldsymbol{A}' = \sigma^2\begin{pmatrix}\boldsymbol{I} \\ \frac{1}{n}\mathbf{1}' \\ \frac{\boldsymbol{q}'}{\boldsymbol{q}'\boldsymbol{q}}\end{pmatrix}\begin{pmatrix}\boldsymbol{I} & \frac{1}{n}\mathbf{1} & \frac{\boldsymbol{q}}{\boldsymbol{q}'\boldsymbol{q}}\end{pmatrix} = \sigma^2\begin{pmatrix}\boldsymbol{I} & \frac{1}{n}\mathbf{1} & \frac{\boldsymbol{q}}{\boldsymbol{q}'\boldsymbol{q}} \\ \frac{1}{n}\mathbf{1}' & \frac{1}{n^2}\mathbf{1}'\mathbf{1} & \frac{1}{n\boldsymbol{q}'\boldsymbol{q}}\mathbf{1}'\boldsymbol{q} \\ \frac{\boldsymbol{q}'}{\boldsymbol{q}'\boldsymbol{q}} & \frac{1}{n\boldsymbol{q}'\boldsymbol{q}}\boldsymbol{q}'\mathbf{1} & \frac{1}{(\boldsymbol{q}'\boldsymbol{q})^2}\boldsymbol{q}'\boldsymbol{q}\end{pmatrix} = \sigma^2\begin{pmatrix}\boldsymbol{I} & \frac{1}{n}\mathbf{1} & \frac{\boldsymbol{q}}{\boldsymbol{q}'\boldsymbol{q}} \\ \frac{1}{n}\mathbf{1}' & \frac{n}{n^2} & \frac{1}{n\boldsymbol{q}'\boldsymbol{q}}\mathbf{1}'\boldsymbol{q} \\ \frac{\boldsymbol{q}'}{\boldsymbol{q}'\boldsymbol{q}} & \frac{1}{n\boldsymbol{q}'\boldsymbol{q}}\boldsymbol{q}'\mathbf{1} & \frac{1}{\boldsymbol{q}'\boldsymbol{q}}\end{pmatrix}$$

But

$$\mathbf{1}'\boldsymbol{q} = \boldsymbol{q}'\mathbf{1} = \sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n \left(x_i - \frac{1}{n}\sum_{j=1}^n x_j\right) = \sum_{i=1}^n x_i - \frac{1}{n}\cdot n\sum_{j=1}^n x_j = 0$$

So we have

$$\text{Var}(\boldsymbol{A}\vec{Y}) = \sigma^2\boldsymbol{A}\boldsymbol{A}' = \sigma^2\begin{pmatrix}\boldsymbol{I} & \frac{1}{n}\mathbf{1} & \frac{\boldsymbol{q}}{\boldsymbol{q}'\boldsymbol{q}} \\ \frac{1}{n}\mathbf{1}' & \frac{1}{n} & 0 \\ \frac{\boldsymbol{q}'}{\boldsymbol{q}'\boldsymbol{q}} & 0 & \frac{1}{\boldsymbol{q}'\boldsymbol{q}}\end{pmatrix}$$

This implies

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

and

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

Plugging these in to equation (1) yields

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

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

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

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

(d) From class notes 08/29:

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

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

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

Since

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

this interval can be expressed as

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

□

Proposition 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.

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

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

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

Next we find the variance.

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

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

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

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

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

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

(b)

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

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

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

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

□

1.2 Chapter 2: Multiple Regression

General OLS:

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

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

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

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

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

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.

Lemma 3.

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

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 4. 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 5. Under the conditions of Lemma 4 (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 4

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

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.

1.4 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

Generalized least squares model:

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

where

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

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

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

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

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

1.6 DSO 607

Generalized linear models:

$$f_n(z, \beta) = \prod_{i=1}^{n} \exp\left[\theta_i z_i - b(\theta_i) h(z_i)\right], \quad z = (z_1, \ldots, z_n)^T$$

Natural parameter θ_i: $\theta_i = x_i^T \beta$, $x_i = 9x_{ij} : j \in \mathscr{M}$

$h(z_i)$: normalization constant

linear regression: $b(\theta) = \frac{1}{2}\theta^2$

other: $b(\theta) = \log(1 + e^\theta)$

If $Y = (Y_1, \ldots, Y_n)^T \sim F_n(\cdot, \beta)$, then $\mathbb{E}(Y) = (b'(\theta_1), \ldots, b'(\theta_n))^T = \mu(\theta)$ and

$\text{Cov}(Y) = \text{diag}\{b''(\theta_1), \ldots, b''(\theta_n)\} = \Sigma(\theta)$ where $\theta = X\beta$ and $X = (x_1, \ldots, x_n)^T$ is the $n \times d$ design matrix.

Quasi-log-likelihood ("quasi" because error may be misspecified):

$$\ell_n(y, \beta) = y^T X\beta - \mathbf{1}^T b(X\beta) + \mathbf{1}^T h(y)$$

Like MLE, maximizing $\ell_n(y, \beta)$ with respect to β gives the quasi-MLE $\hat{\beta}_n$. Solution exists and is unique due to strict convexity of b, solves the score equation

$$\frac{\partial \ell_n(y, \beta)}{\partial \beta} = x^T[y - \mu(X\beta)] = \mathbf{0}$$

(Intuition of score equation: the columns of X are all orthogonal to the errors (uncorrelated if X is random).

KL Divergence:

$$I(g_n; f_n(\cdot, \beta)) = \sum_{i=1}^{n} \left[\int \right]$$

To minimize the KL divergence:

$$\frac{\partial I(g_n; f_n(\cdot, \beta))}{\partial \beta} = -X^T[\mathbb{E}(Y) = \mu(X\beta)] = 0$$

the inverse of the Fisher information matrix is the covariance of the MLE (?).

$$\vdots$$

For AIC, we minimize the KL divergence. For BIC, we maximize the Bayes factor (posterior probability for the model).

$$\vdots$$

$$B_n^{1/2} A_n(\hat{\beta}_n - \beta_{n,0}) = W_n \xrightarrow{D} \mathcal{N}(0, I_d)$$

$$\hat{\beta}_n - \beta_{n,0} = A_n^{-1} B_n^{1/2} W_n \implies \text{Cov}(\hat{\beta}_n) = \text{Cov}(\hat{\beta} - n - \beta_{n,0})$$

$$= \text{Cov}(A_n^{-1} B^{1/2} W_n) = A_n^{-1} B_n^{1/2} \text{Cov}(W_n) B_n^{1/2} A_n^{-1} = A_n^{-1} B_n^{1/2} I_d B_n^{1/2} A_n^{-1} = \boxed{A_n^{-1} B_n A_n^{-1}}$$

Note that if the model is correct, $A_n = B_n$ so this reduces to conventional asymptotic MLE theory ($\text{Cov}(\hat{\beta_n}) = A_n^{-1}$).

$$\vdots$$

A_n from working model, B_n from true model (unknown).

GBIC in misspecified models: $H_n = A_n^{-1} B_n$ (covariance contrast matrix). Note that when model is specified, $H_n = I_d$ so the log of its determinant is 0 so it vanishes. If not, then it is a misspecification penalty.

$$\vdots$$

Note: $\log(y, \hat{\beta}_n) > \log(y, \beta_{n,0})$ because $\hat{\beta}_n$ is by definition the MLE on the observed data. But $\mathbb{E}(\log(\tilde{y}, \beta_{n,0}) > \mathbb{E}(\log(\tilde{y}, \hat{\beta}_n)$ because $\beta_{n,0}$ is the true parameter. We have a systematic upward bias when we use the empirical estimate. (p.18 of week 2-2 slides)

1.7 Lasso

Consider the linear regression model $y = X\beta + \epsilon$. If we assume the errors ϵ have a multivariate Gaussian distribution, that is,

$$f_\epsilon(t) = \left(\frac{1}{2\sqrt{\pi\sigma^2}}\right)^n \exp\left(-\frac{t^T t}{2\sigma^2}\right), \quad t = (t_1, \ldots, t_n)^T$$

then the log likelihood is

$$\log(f(t)) = n \log[(2\pi\sigma^2)^{-1/2}] - t^T t/(2\sigma^2)$$

Suppose we want the MLE estimator. When we maximize the log likelihood, we can disregard the first term which does not include t (it is constant). So we seek

$$\arg\max_{\beta} -t^T t/(2\sigma^2) = \arg\max_{\beta} -\|y - X\beta\|_2^2/(2\sigma^2)$$

which is the same as

$$\arg\min_{\beta} \|y - X\beta\|_2^2/(2\sigma^2)$$

We commonly scale this with an n in the denominator to match the empirical risk; note that this does not affect the arguments which minimize the quantity. When the design matrix X multiplied by $n^{-1/2}$ is orthonormal ($X^T X = nI_p$), the penalized least squares reduces to the minimization of

$$\min_{\beta \in \mathbb{R}^p} \{ \quad \frac{1}{2n} \|y - X\hat{\beta}\|_2^2 + \frac{1}{2}\|\hat{\beta} - \beta\|_2^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \quad \}$$

where $\hat{\beta} = (X^T X)^{-1} X^T y = nX^T y$ is the OLS estimator. Disregarding the first term which does not contain β, we have a **separable** loss function (we can solve for one parameter at a time):

$$\min_{\beta \in \mathbb{R}^p} \{ \quad \frac{1}{2}\|\hat{\beta} - \beta\|_2^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \quad \}.$$

So we can consider the univariate penalized least squares function

$$\hat{\theta}(z) = \arg\min_{\theta \in \mathbb{R}} \{ \frac{1}{2}(z - \theta)^2 + p_\lambda(|\theta|).$$

Antoniadis and Fan (2001) showed that the PLS estimator $\hat{\theta}$ possesses the following properties:

- *sparsity* if $\min_{t \geq 0}\{t + p'_\lambda(t)\} > 0$;
- *approximate unbiasedness* if $p'_\lambda(t) = 0$ for large t;
- *continuity* if and only if $\arg\min_{t \geq 0}\{t + p'_\lambda(t)\} = 0$. Intuition: if you perturb data a little, the solution should remain similar.

In general, the singularity of the penalty function at the origin (i.e., $p'_\lambda(0+) < 0$) is needed for generating sparsity in variable selection and the concavity is needed to reduce the bias.

1.7.1 Soft Thresholding

Classical ideas of nonparametric models: kernels (locally constant/linear), splines (smooth basic functions). But wavelets are non-smooth. Why is this beneficial? Some real life functions are non-smooth. (example; image data with noise. There will be non-smooth edges to objects.) Also, the wavelet basis functions are orthonormal (which is closely related to the assumption we made above about the orthonormal design matrix). So when working with wavelets, we have a separable optimization problem. Soft thresholding is something like the lasso idea for wavelets (but before the lasso was developed).

Suppose we wish to recover an unknown function f on $[0, 1]$ from noisy data

$$d_i = f(t_i) + \sigma z_i, \quad i = 0, \dots, n-1$$

where $t_i = i/n$ and $z_i \sim \mathcal{N}(0, 1)$. The term de-noising is to optimize the mean squared error $n^{-1} \; E\|\hat{f} - f\|_2^2$. Donoho and Johnstone (1994) proposed a soft-thresholding estimator

$$\hat{\beta}_j = \operatorname{sgn}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$$

where γ is some small number. (So estimator gets shrunk by γ, and if γ is bigger than the original estimator, we set it equal to 0.) They applied this estimator to the coefficients of a wavelet transform of a function measured with noise, then back-transformed to obtain a smooth estimate of the function.

Example 1.1. Suppose we have an image in data in the form of $X \in \mathbb{R}^n$. We have a wavelet basis $W \in \mathbb{R}^{n \times n}$ where W is orthonormal. We transform the image into the frequency domain by

$$Wx \to \tilde{x}$$

where $\tilde{x}$ is the frequency domain representation. Then we apply soft-thresholding to $\tilde{x}$ to yield $\tilde{x}^*$, which we hope is de-noised. Finally, we bring the image back into the original domain according to

$$\hat{x} = W^{-1}\tilde{x}^* = W^T\tilde{x}^*.$$

The asymptotic risk of this estimator is

$$[2(\log p) + 1](\sigma^2 + R_{DP})$$

Note that the $2 \log p$ term is related to the result (described informally) below:

Proposition 6. if we have n i.i.d. $\mathcal{N}(0, 1)$ random variables, the maximum of them is near $\sqrt{2 \log n}$ if n is large. (The order is this large with high probability)

Remark. In the language of wavelets, sometimes ℓ_0 penalization is called "hard-thresholding."

1.7.2 Lasso theory

Drawbacks of previous techniques that lasso helps with: subset selection is interpretable but computationally intensive and not stable because it is a discrete process (small changes in the data can result in very different models being selected). Ridge regression is a continuous process and more stable, but it does not set any coefficients equal to 0 and hence does not give an easily interpretable model.

In the orthonormal design case $X^T X = nI_p$, the lasso solution can be shown to be the same as soft thresholding:

$$\hat{\beta}_j = \text{sgn}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$$

where $\gamma \geq 0$ is determined by the condition $\sum_{j=1}^p |\beta_j| = t$.

Geometry: the criterion $\sum_{i=1}^n (y_i - \sum_{j=1}^p \beta_j x_{ij})^2$ equals the quadratic function (plus a constant)

$$(\beta - \hat{\beta}^0 0^T X^T X(\beta - \hat{\beta}^0).$$

The contours (level sets) are elliptical and centered at the OLS estimates. If the constraint region does not have corners, zero solutions result with probability zero (see DSO 607 homework 2).

1.7.3 Non-Negative Garotte

This idea inspired the lasso. Proposed by Breiman (1995). It minimizes

$$\sum_{i=1}^n (y_i - \alpha - \sum_{j=1}^p c_j \hat{\beta}_j^o x_{ij})^2 \text{ subject to } c_j \geq 0, \sum_{j=1}^p c_j \leq t$$

It starts with OLS estimates and shrinks them by non-negative factors whose sum is constrained. It depends on both the sign and magnitude of OLS estimates. In contrast, lasso avoids the explicit use of OLS estimates.

1.7.4 LARS—Preliminaries and Intuition

Intuition: the algorithm takes steps from a model where all coefficients are 0 to the biggest model (the unpenalized OLS model). Covariates are considered from the highest correlation with y to the least. (The variable most highly correlated with y is the one at the "least angle" from y.) Recall the original definition of the lasso estimator:

$$\hat{\beta}_{lasso} = \arg\min_\beta \{\frac{1}{2n}\|y - X\beta\|_2^2\} \text{ subject to } \|\beta\|_1 \leq t \tag{2}$$

The more common version now:

$$\hat{\beta}_{lasso} = \arg\min_\beta \{\frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\} \tag{3}$$

One form can be changed to the other by applying Lagrangians[1]. Have to be careful because this is a convex program (quadratic with "linear" constraint—use a slack variable).

Taking the gradient of the loss function in (3) yields

$$\nabla\left(\frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\right) = \nabla\left(\frac{1}{2n}\|y - X\beta\|_2^2\right) + \lambda\nabla\big(\|\beta\|_1\big)$$

$$= -\frac{1}{n}X^T(y - X\beta) + \lambda\nabla\big(\|\beta\|_1\big) \tag{4}$$

We set this equal to zero. If the first term equals 0, the residual has to equal 0. For the second part to equal zero, we have to account for the fact that the gradient doesn't exist at 0. In the one-dimensional case $g(t) = |t|$, we have

$$g'(t) = \begin{cases} -1 & t < 0 \\ 1 & t > 0 \end{cases}$$

but it doesn't exist at 0. Instead of using the gradient, we will use ∂, the subdifferential, which is the set of all subgradients. We have a solution if 0 is in the subdifferential. We can rewrite (4) using the subdifferential instead of the gradient:

$$\partial\left(\frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\right) = \nabla\left(\frac{1}{2n}\|y - X\beta\|_2^2\right) + \lambda\partial\big(\|\beta\|_1\big) = -\frac{1}{n}X^T(y - X\beta) + \lambda\partial\big(\|\beta\|_1\big)$$

Then rather than setting the gradient equal to 0, our condition is

$$0 \in -\frac{1}{n}X^T(y - X\beta) + \lambda\partial\big(\|\beta\|_1\big)$$

Note that

$$\partial g(t) = \begin{cases} -1 & t < 0 \\ [-1, 1] & t = 0 \\ 1 & t > 0 \end{cases} = \begin{cases} \operatorname{sgn}(t) & t \neq 0 \\ [-1, 1] & t = 0 \end{cases}$$

so we have

$$0 \in -\frac{1}{n}X^T(y - X\beta) + \lambda \cdot \left[\begin{cases} \operatorname{sgn}(\beta_j) & t \neq 0 \\ [-1, 1] & \beta_j = 0 \end{cases}\right]$$

where

[1] However, the correspondence between t and λ is **not** one-to-one. Because with $t = \infty$, $\lambda = 0$. But a slightly smaller t would result in the same solution.

$$\begin{bmatrix} \begin{cases} \text{sgn}(\beta_j) & t \neq 0 \\ [-1, 1] & \beta_j = 0 \end{cases} \end{bmatrix} \in \mathbb{R}^p$$

(1) Examining the jth component of this separable equation, if $\beta_j \neq 0$, we have

$$0 = -\frac{1}{n} X_j^T (y - X\beta) + \lambda \cdot \text{sgn}(\beta_j) \iff \frac{1}{n} X_j^T (y - X\beta) = \lambda \cdot \text{sgn}(\beta_j)$$

Note that the left side contains the correlation between X_j and $e = y - X\beta$, the residual vector. **So if lasso chooses k variables, all k of them will have the same correlation with the residual (λ).**

(2) If $\beta_j \neq 0$, we have

$$0 \in -\frac{1}{n} X^T (y - X\beta) + \lambda \cdot [-1, 1] \iff \left| \frac{1}{n} X^T (y - X\beta) \right| \leq \lambda$$

So for unselected features, the (absolute) correlation should be bounded by λ.

These two conditions relate to the KKT conditions (first order conditions).

So if we start with λ very large and gradually decrease it, we will let in as the first feature the one that is most highly correlated with y—that is, the feature with the *least angle* between it and y.

1.7.5 LARS

In Figure 1, note that we choose feature X_1 first because it has the highest correlation with y. As the coefficient on X_1 increases, the correlation between X_1 and the residual with y decreases, while the correlation between X_2 and the residual remains constant (**increases?**). When the correlation between X_1 and the residual becomes equal to the correlation between X_2 and the residual, X_2 enters the lasso path.

Remark. Just like in lasso, in LARS the correlation between all included features and the residual are equal. However, LARS is a stepwise procedure—once we add a feature, it stays in the model. In the lasso, features can be dropped later in the path after they are selected—whenever β_j becomes 0, it is dropped from the current active set. A feature's sign cannot change in lasso—it is not possible. If we modify the LARS algorithm to have this property ("lasso modfiication"), then the result is the lasso estimator.

The LARS algorithm for lasso has order $\mathcal{O}(np \cdot \min\{n, p\})$. In particular, if $p > n$ it has order $\mathcal{O}(n^2 p)$.

1.8 Quadratic Loss

Theorem 7. Let $X : \Omega \to \mathbb{R}$ be a random variable with $\mathbb{E}X^2 < \infty$. Then $\mathbb{E}(X - t)^2$ is minimized for $t \in \mathbb{R}$ uniquely when $t = \mathbb{E}X$.

Proof. We seek

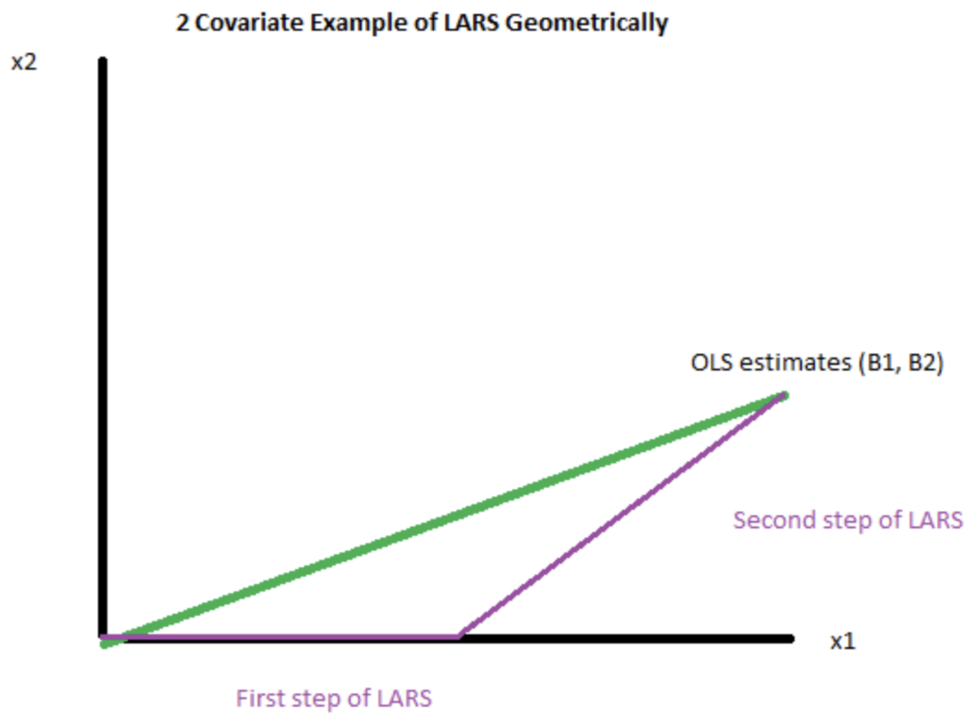

Figure 1: LARS figure in 2d case.

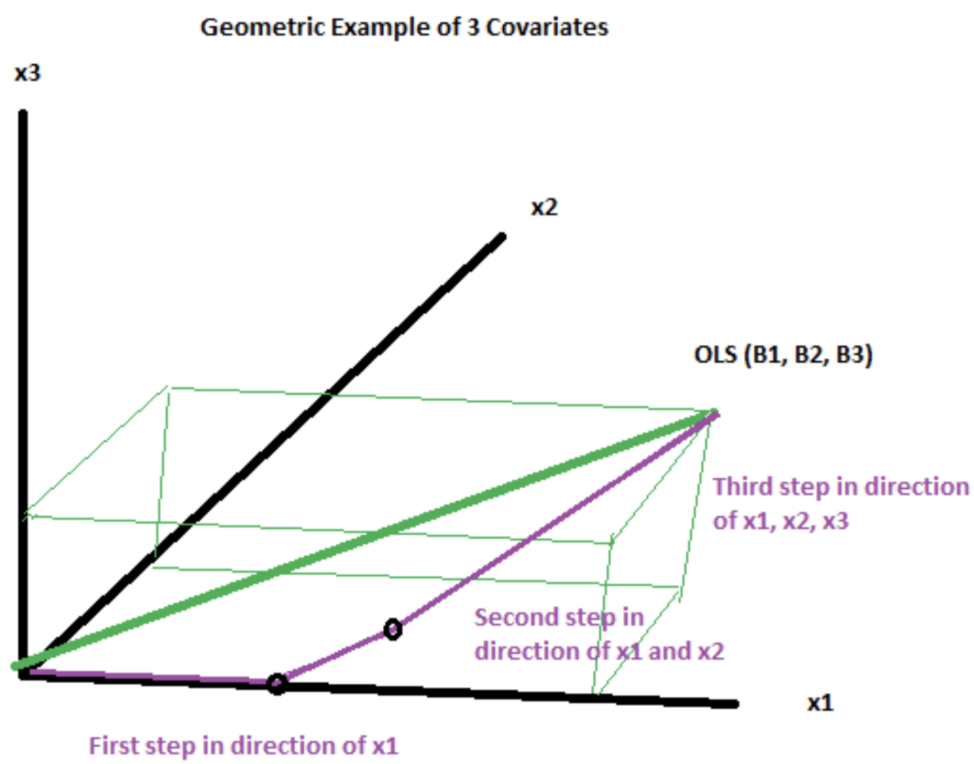

Figure 2: LARS figure in 3d case.

$$\arg\min_t \mathbb{E}(X-t)^2 = \arg\min_t \left[\mathbb{E}(X^2) - 2t\mathbb{E}(X) + t^2\right] = \arg\min_t \left[t^2 - 2t\mathbb{E}(X)\right]$$

where the last step follows because $\mathbb{E}(X^2)$ is independent of t. This expression is quadratic in t. Differentiating with respect to t and setting equal to 0, we have

$$2t - 2\mathbb{E}(X) = 0 \implies \boxed{\arg\min_t \mathbb{E}(X-t)^2 = \mathbb{E}(X)}$$

□