# Math Review Notes—Linear Regression

Gregory Faletto

# Contents

Last updated June 14, 2019

# 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, 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) $\epsilon_i$ 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 $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, $\epsilon_i$ 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 $\sigma^2$ 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 $\text{Cov}(y_j, \hat{\beta}_1)$:

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) 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 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 $\beta_1$ 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 $\sigma^2$ 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,

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

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

Next,

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

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

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

$$\sum_{i=1}^{n} x_i y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

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

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 $\beta_1$ are non-random and $\epsilon_i$ 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 $\beta_1$ are non-random and $\epsilon_i$ are independent, this can be written as

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

$\beta_1$ is a linear combination of $y_i$ which is normally distributed, therefore $\beta_1$ is normally distributed.

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

(b)

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

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

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

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

□

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

## 1.2 Chapter 2: Multiple Regression

General OLS:

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

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

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

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

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

## 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 $\epsilon$ 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 $\nu$ 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 ($\epsilon$ 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

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

□

$t$-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 $\delta$ 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

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

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

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:

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

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

$h(z_i)$: normalization constant

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

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

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

$\text{Cov}(Y) = \text{diag}\{b''(\theta_1),\dots,b''(\theta_n)\} = \Sigma(\theta)$ where $\theta = X\beta$ and $X = (x_1,\dots,x_n)^T$ 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 $\beta$ 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).

### 1.6.1 Akaike Information Criterion (AIC)

AIC: proposed by Akaike [1973] to choose a model by minimizing the Kullback-Leibler (KL) divergence of the fitted model from the true model (or equivalently, maximize the expected log-likelihood). Recall the KL Divergence

$$I(\theta;\theta_0) := 2\mathbb{E}_{\theta_0}\left[\log\left(f(X \mid \theta_0)\right)\right] - 2\mathbb{E}_{\theta_0}\left[\log\left(f(X \mid \theta)\right)\right].$$

We will try to maximizing the KL Divergence by estimating $\theta_0$ as best as we can by maximizing the **probabilistic negentropy**

$$\mathbb{E}_Z I(\theta;\hat{\theta}_0(Z)) := 2\mathbb{E}_{\theta_0}\left[\log\left(f(X \mid \theta_0)\right] - 2\mathbb{E}_{\theta_0,Z}\left[\log\left(f\left(X \mid \hat{\theta}_0(Z)\right)\right)\right].$$

Because the true model $\theta_0$ is unknown we cannot carry out this maximization directly. Note that as the number of independent observations increases, the **mean log-likelihood ratio**

$$\hat{I}(\theta;\theta_0) := \frac{2}{n}\sum_{i=1}^{n}\log\frac{f(x_i \mid \theta_0)}{f(x_i \mid \theta)} \xrightarrow{p} I(\theta;\theta_0).$$

Because of this, Akaike reasons that maximizing the mean log-likelihood ratio over $\theta_0$ (i.e. computing the maximum likelihood estimate) tend to maximize the entropy. So the maximum likelihood estimate $\hat{\theta}_0(Z)$ is substituted for the unknown $\theta_0$.

Way we wrote KL Divergence in DSO 607: density $f$ from density $g$:

$$I(g_n; f_n(\cdot,\beta)) = \int [\log g(z)]g(z)dz - \int [\log f(z)]g(z)dz$$

Akaike [1973] found that up to an additive constant, the KL divergence of the fitted model from the true model can be asymptotically expanded as

$$-\ell_n(\hat{\theta}) + \lambda \dim(\hat{\theta}) = -\ell_n(\hat{\theta}) + \lambda\sum_{j=1}^{p}\mathbf{1}_{\{\hat{\theta}_j \neq 0\}}$$

where $\ell_n(\theta)$ is the log-likelihood function and $\lambda = 1$. This leads to the Akaike information criterion (AIC) for comparing models:

$$AIC\left(\hat{\theta}_k(Z)\right) := n\hat{I}\left(\hat{\theta}_k(Z;\hat{\theta}_0(Z)\right) + 2\|\hat{\theta}_k(Z)\|_0 = 2\sum_{i=1}^{n}\log\frac{f(x_i \mid \hat{\theta}_0(Z))}{f(x_i \mid \hat{\theta}_k(Z))} + 2\|\hat{\theta}_k(Z)\|_0$$

Way we wrote this is DSO 607:

$$AIC(\hat{\theta}) := -2\ell_n(\hat{\theta}) + 2\|\hat{\theta}\|_0$$

Intuition: $\log g(x)$ is the log likelihood. Penalty term can be interpreted as penalty, or as a bias correction since you are doing training and feature selection simultaneously on the same data.

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

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 more information on KL Divergence, see Section **??**). For AIC, we minimize the KL divergence. For BIC, we maximize the Bayes factor (posterior probability for the model).

### 1.6.2 Bayesian Information Criterion (BIC)

A typical Bayesian model selection procedure is to first give nonzero prior probability $\alpha_M$ om each model $M$ and then prescribe a prior distribution $\mu_M$ for the parameter vector in the corresponding model. The Bayesian principle of model selection is to choose the most probable model *a posteriori*; that is, to choose a model that maximizes the log-marginal likelihood (or the Bayes factor)

$$\log \int \alpha_M \exp[\ell_n(\theta)] d\mu_m(\theta).$$

Schwarz [1978] took a Bayesian approach with prior distributions that have nonzero prior probabilities on some lower dimensional subspaces of $\mathbb{R}^p$ and showed that the negative log-marginal likelihood can be asymptotically expanded as

$$-\ell_n(\hat{\theta}) + \lambda \|\hat{\theta}\|_0$$

where $\lambda = (\log n)/2$. This asymptotic expansion leads to the Bayesian information criterion (BIC) for comparing models:

$$BIC(\hat{\theta}) := -2 \log \Big( f(x \mid \hat{\theta}; \hat{\theta}_{MLE}) \Big) + (\log n) \|\hat{\theta}\|_0.$$

where $f$ is the density function parameterized by $\hat{\theta}_{MLE}$, the maximum likelihood estimate for the density given the data $x$.

Way we wrote this in DSO 607:

$$BIC(\hat{\theta}) := -2\ell_n(\hat{\theta}) + (\log n) \|\hat{\theta}\|_0.$$

$$\vdots$$

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

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

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

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)

**Proposition 6** (**Result from *"Econometrics: Methods and Applications"* homework**)**.** Consider the usual linear model, where $y = X\beta + \epsilon$. Suppose we compare two regressions, which differ in how many variables are included in the matrix $X$. In the full (unrestricted) model $p_1$ regressors are included. In the restricted model only a subset of $p_0 < p_1$ regressors are included. Then for large $n$, selection based on AIC corresponds to an $F$-test with a critical value of approximately 2.

*Proof.* Let $e_R$ be the vector of residuals for the restricted model with $p_0$ parameters and $e_U$ the vector of residuals for the full unrestricted model with $p_1$ parameters. Then we have the sample standard deviations

$$s_0^2 = \frac{1}{n-p_0}e_R{}'e_R, s_1^2 = \frac{1}{n-p_1}e_U{}'e_U \tag{2}$$

Recall the AIC:

$$\log(s^2) + \frac{2k}{n}$$

where $k$ is the number of regressors included in the model.
For the small model, we have

$$AIC_0 = \log(s_0^2) + \frac{2p_0}{n}.$$

For the big model, we have

$$AIC_1 = \log(s_1^2) + \frac{2p_1}{n}.$$

Therefore the smallest model is preferred according to the AIC if

$$AIC_0 < AIC_1$$

$$\iff \log(s_0^2) + \frac{2p_0}{n} < \log(s_1^2) + \frac{2p_1}{n} \iff \log(s_0^2) - \log(s_1^2) < \frac{2p_1}{n} - \frac{2p_0}{n} \iff \log(\frac{s_0^2}{s_1^2}) < \frac{2}{n}(p_1 - p_0)$$

$$\iff \frac{s_0^2}{s_1^2} < e^{\frac{2}{n}(p_1 - p_0)} \tag{3}$$

If $n$ is very large, $\frac{2}{n}(p_1 - p_0)$ is small. Therefore, using the first order Taylor approximation $e^x \approx 1 + x$ we can approximate that

$$e^{\frac{2}{n}(p_1 - p_0)} \approx 1 + \frac{2}{n}(p_1 - p_0)$$

(if $n$ is very large.) Substituting this expression into the right side of (3) yields

$$\frac{s_0^2}{s_1^2} < 1 + \frac{2}{n}(p_1 - p_0) \iff \frac{s_0^2}{s_1^2} - 1 < \frac{2}{n}(p_1 - p_0) \iff \frac{s_0^2}{s_1^2} - \frac{s_1^2}{s_1^2} < \frac{2}{n}(p_1 - p_0)$$

$$\iff \frac{s_0^2 - s_1^2}{s_1^2} < \frac{2}{n}(p_1 - p_0)$$

for $n$ very large. Plugging in the expressions from (2), we have

$$\frac{\frac{1}{n-p_0}{e_R}'e_R - \frac{1}{n-p_1}{e_U}'e_U}{\frac{1}{n-p_1}{e_U}'e_U} < \frac{2}{n}(p_1 - p_0).$$

For large values of $n$, $n - p_0 \approx n - p_1 \approx n$. This yields

$$\frac{\frac{1}{n}{e_R}'e_R - \frac{1}{n}{e_U}'e_U}{\frac{1}{n}{e_U}'e_U} < \frac{2}{n}(p_1 - p_0)$$

$$= \frac{{e_R}'e_R - {e_U}'e_U}{{e_U}'e_U} < \frac{2}{n}(p_1 - p_0) \tag{4}$$

Now recall the $F$ statistic:

$$F = \frac{({e_R}'e_R - {e_U}'e_U)/g}{{e_U}'e_U/(n-k)} \tag{5}$$

where $k$ is the number of explanatory factors in the unrestricted model, and $g$ is the number of explanatory factors removed from the unrestricted model to create the restricted model. Under this test, we believe there is significant evidence to suggest that $\beta \neq 0$ (so the unrestricted model is preferred) if $F > F_{critical}$. Therefore a larger model is preferred if $F > F_{critical}$, and we stay with (prefer) a smaller model if $F < F_{critical}$.

Let $F_{critical} = 2$. Then a smaller model is preferred if $F < 2$:

$$\frac{(e_R{}'e_R - e_U{}'e_U)/g}{e_U{}'e_U/(n-k)} < 2$$

In this case, with $p_1$ factors in the unrestricted model and $p_0$ in the restricted model, we get

$$\frac{(e_R{}'e_R - e_U{}'e_U)/(p_1-p_0)}{e_U{}'e_U/(n-p_1)} < 2$$

$$\frac{(e_R{}'e_R - e_U{}'e_U)}{e_U{}'e_U} < \frac{2(p_1-p_0)}{n-p_1}$$

If $n$ is very large, $n - p_1 \approx n$. Substituting this in yields

$$\frac{(e_R{}'e_R - e_U{}'e_U)}{e_U{}'e_U} < \frac{2(p_1-p_0)}{n} \tag{6}$$

which equals (4). Our condition for preferring a restricted model when doing an F-test with $F_{critical} = 2$ (and when $n$ is very large) is approximately the same as our condition for preferring a restricted model when using the AIC (when $n$ is very large).

□

## 1.7 Ridge Regression

Suppose $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown vector, and for all $1 \leq i \leq n$, there are known vectors $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^p$. Our observed data are $\boldsymbol{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$. Let $\boldsymbol{X}$ be the $n \times p$ matrix so that the $i^{th}$ row of $\boldsymbol{X}$ is the row vector $x^{(i)}$. Assume that $p \leq n$ and the matrix $\boldsymbol{X}$ has full rank. Let $\lambda > 0$ and consider the quantity

$$\sum_{i=1}^{n} \left(y_i - {x^{(i)}}^T \boldsymbol{\beta}\right)^2 + \lambda \|\boldsymbol{\beta}\|_2^2 = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \tag{7}$$

The term $\|\boldsymbol{\beta}\|_2^2$ penalizes $\boldsymbol{\beta}$ from having large entries. By Lagrange Multipliers, a critical point $\boldsymbol{\beta}$ of the constrained minimization problem

$$\text{minimize } \sum_{i=1}^{n} (\boldsymbol{y}_i - \langle x^{(i)}, \boldsymbol{\beta} \rangle)^2 \quad \text{subject to } \|\boldsymbol{\beta}\|_2^2 \leq 1$$

is equivalent to the existence of a $\lambda \in \mathbb{R}$ such that $\boldsymbol{\beta}$ is a critical point of (7). We call the $\hat{\beta}$ that minimizes (7) the **ridge regression** estimator for $\beta$.

**Proposition 7 (Math 541A Homework Problem).** The value of $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ that minimizes (7) is $\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

*Proof.*

$$\sum_{i=1}^{n} \left(y_i - {x^{(i)}}^T \boldsymbol{\beta}\right)^2 + \lambda \|\boldsymbol{\beta}\|^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T\boldsymbol{\beta}$$

$$= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta} = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}$$

where $\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}$ because a scalar equals its transpose. Differentiating with respect to $\boldsymbol{\beta}$ yields

$$-2\boldsymbol{y}^T\boldsymbol{X} + 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{\beta}^T = 0 \iff \boldsymbol{\beta}^T(2\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{I}_p) = 2\boldsymbol{y}^T\boldsymbol{X}$$

$$\iff (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y} \iff \hat{\boldsymbol{\beta}}_{\text{ridge}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

where $\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p$ is invertible by the following argument. $\boldsymbol{X}^T\boldsymbol{X}$ must be positive semidefinite. In fact, it is positive definite because $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ has full rank; that is, $\text{rank}(\boldsymbol{X}) = p$, so $\boldsymbol{X}^T\boldsymbol{X} \in \mathbb{R}^{p\times p}$ has rank $p$ (full rank) and is invertible. So $\boldsymbol{X}^T\boldsymbol{X}$ is positive definite (all positive eigenvalues). Then since $\text{Tr}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p) > \text{Tr}(\boldsymbol{X}^T\boldsymbol{X})$, the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p$ are also all positive, which means the determinant of $\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p$ is nonzero, which means it is invertible.

□

**Proposition 8** (**DSO 607 Homework Problem**)**.** Suppose $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$.

(a) The asymptotic behavior of the ridge estimator is as follows: as $\lambda \to \infty$, $\hat{\boldsymbol{\beta}}_{\text{ridge}} \to \boldsymbol{0}$, and as $\lambda \to 0$, $\hat{\boldsymbol{\beta}}_{\text{ridge}} \to X^\dagger(X\beta + \epsilon)$.

(b) For any fixed $\lambda > 0$,

*Proof.* (a) Since $X$ is fixed, as $\lambda \to \infty$ we have

$$\hat{\beta}_{\text{ridge}} = (X^TX + \lambda I_p)^{-1}X^Ty \to (\lambda I_p)^{-1}X^Ty = \frac{1}{\lambda}I_pX^T(X\beta + \epsilon) = \frac{1}{\lambda}(X^TX\beta + X^T\epsilon) \to \boldsymbol{0}$$

where $\boldsymbol{0}$ is a $p$-dimensional vector of zeroes. As $\lambda \to 0^+$ we have

$$\hat{\beta}_{\text{ridge}} = (X^TX + \lambda I_p)^{-1}X^Ty \to X^\dagger y = X^\dagger(X\beta + \epsilon)$$

where we substitute the pseudoinverse instead of the inverse because since $X^TX$ is rank deficient, $(X^TX)^{-1}$ does not exist.

(b) We have

$$\hat{\beta}_{\text{ridge}} = (X^TX+\lambda I_p)^{-1}X^Ty = (X^TX+\lambda I_p)^{-1}X^T(X\beta+\epsilon) = (X^TX+\lambda I_p)^{-1}X^TX\beta+(X^TX+\lambda I_p)^{-1}X^T\epsilon$$

Let $e_i$ be a selection vector, with the $i$th entry equal to 1 and all other entries equal to 0. Let the $i$th entry of $\hat{\beta}_{\text{ridge}}$ be $\hat{\beta}^{(i)}_{\text{ridge}} = e_i^T\hat{\beta}_{\text{ridge}}$. We have

$$\Pr(\hat{\beta}^{(i)}_{\text{ridge}} = 0) = \Pr\left(e_i^T[(X^TX + \lambda I_p)^{-1}X^TX\beta + (X^TX + \lambda I_p)^{-1}X^T\epsilon] = 0\right)$$

$$= \Pr\left(e_i^T(X^TX + \lambda I_p)^{-1}X^T\epsilon = -e_i^T(X^TX + \lambda I_p)^{-1}X^TX\beta\right)$$

Since every entry of $\epsilon$ is distributed continuously, the probability of it equaling a particular value is 0. Therefore the probability that each component of the ridge estimator equals 0 is 0. (For an intuitive argument as to why this is, see Figure 1.)

□

## 1.8 Lasso

From KKT theory, the correlation between all selected features and residual will be $\lambda$ (see the remark in Section lars.prelims for an explanation why).

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

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

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

$$\arg\max_{\beta\in\mathbb{R}^p} \left\{-t^T t/(2\sigma^2)\right\} = \arg\max_{\beta\in\mathbb{R}^p} \left\{-\|y - X\beta\|_2^2/(2\sigma^2)\right\}$$

which is the same as

$$\arg\min_{\beta\in\mathbb{R}^p} \left\{\|y - X\beta\|_2^2/(2\sigma^2)\right\}$$

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

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

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 $\beta$, we have a **separable** loss function (we can solve for one parameter at a time):

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

So we can consider the univariate penalized least squares function

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

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.

To recap: constrained version:

$$\begin{aligned} \hat{\beta}_{\text{lasso}} = &\arg\min_{\beta \in \mathbb{R}^p} && \frac{1}{2n}\|y - X\beta\|_2^2 \\ &\text{subject to} && \|\beta\|_1 \leq t \end{aligned}$$

Unconstrained version:

$$\hat{\beta}_{\text{lasso}} = \arg\min \left\{ \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \right\}$$

Use $1/n$ to rescale RSS due to $\|1\| - 2 = \sqrt{n}$.

**Proposition 9** (**Math 541A Homework Problem**)**.** Suppose $\boldsymbol{\beta} \in \mathbb{R}^p$ is an unknown vector, and for all $1 \leq i \leq n$, there are known vectors $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$. Our observed data are $\boldsymbol{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$. Let $\boldsymbol{X}$ be the $n \times p$ matrix so that the $i^{th}$ row of $\boldsymbol{X}$ is the row vector $x^{(i)}$. Assume that $p \leq n$ and the matrix $\boldsymbol{X}$ has full rank. Let $\lambda > 0$ and consider the quantity

$$\sum_{i=1}^{n} \left( \boldsymbol{y}_i - {x^{(i)}}^T \boldsymbol{\beta} \right)^2 + \lambda \sum_{i=1}^{p} |\boldsymbol{\beta}_i| \tag{8}$$

Then there exists a $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ that minimizes this quantity (this $\hat{\boldsymbol{\beta}}$ is known as the LASSO, or least absolute shrinkage and selection operator).

*Proof.* We can write (8) as

$$\sum_{i=1}^{n} \left(\boldsymbol{y}_i - {x^{(i)}}^T \boldsymbol{\beta}\right)^2 + \lambda \sum_{i=1}^{p} |\boldsymbol{\beta}_i| = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$$

$$= \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \tag{9}$$

By Proposition **??**, $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ is convex, and by Proposition **??**, $\lambda\|\boldsymbol{\beta}\|_1$ is convex. Therefore by Proposition **??**, (9) is convex. Differentiating and setting equal to 0 yields

$$-2\boldsymbol{y}^T\boldsymbol{X} + 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X} + \lambda\left[\text{sgn}(\boldsymbol{\beta}_i)\right] = 0 \tag{10}$$

where $\left[\text{sgn}(\boldsymbol{\beta}_i)\right]$ is vector resulting from the sgn function being applied elementwise to $\boldsymbol{\beta}$. Since (10) is linear in $\boldsymbol{\beta}$, it has one solution. Since (9) is convex, the unique solution to (10) minimizes (8).

□

**Remark.** The $L_1$ penalization term in (8) is better at penalizing large entries of $\boldsymbol{\beta}$ (a similar observation applies in the compressed sensing literature). Unfortunately, there is no closed form solution to (8) in general. The constrained minimization problem

$$\text{minimize } \sum_{i=1}^{n} (\boldsymbol{y}_i - \langle x^{(i)}, \boldsymbol{\beta}\rangle)^2 \quad \text{subject to } \sum_{i=1}^{n} |\boldsymbol{\beta}_i| \leq 1$$

is morally equivalent to (8), but technically Lagrange Multipliers does not apply since the constraint is not differentiable everywhere.

### 1.8.1 Soft Thresholding

Classical ideas of nonparametric models: kernels (locally constant/linear), splines (smooth basis 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

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

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

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

where $\gamma$ is some small number. (So estimator gets shrunk by $\gamma$, and if $\gamma$ 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 10.** 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 $\ell_0$ penalization is called "hard-thresholding."

### 1.8.2 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 $X^TX = nI_p$, 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)

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

*Proof.*

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 = \sum_{i=1}^{n}\left(y_i - X_i\hat{\beta}\right)^2 = (\boldsymbol{y} - \boldsymbol{X}\hat{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}) = [\boldsymbol{X}(\beta^0 - \hat{\beta})]^T[\boldsymbol{X}(\beta^0 - \hat{\beta})]$$

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

□

The contours (level sets) are therefore elliptical and centered at the OLS estimates. If the constraint region does not have corners, as in ridge regression, zero solutions result with probability zero (see Proposition 8 and Figure 1).

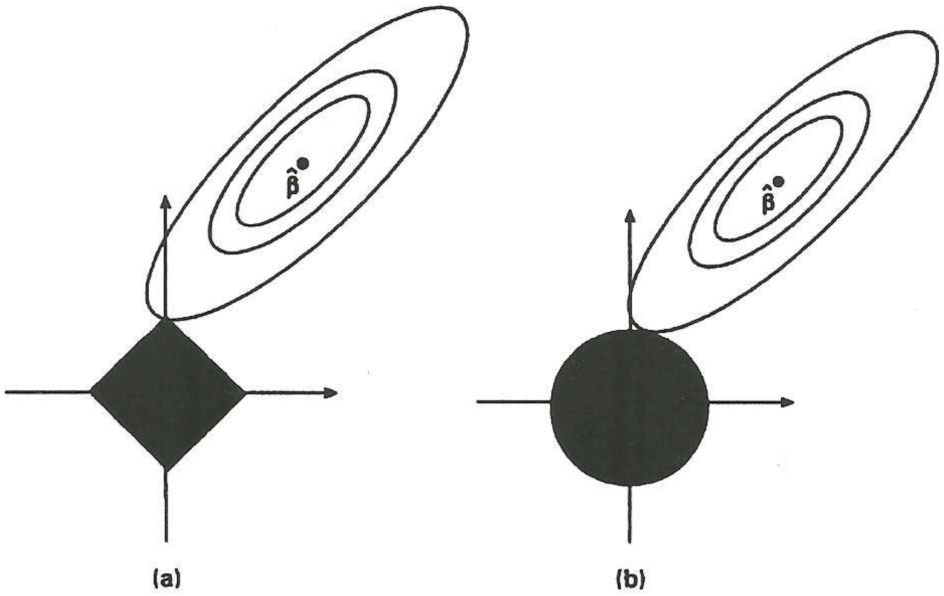


Figure 1: Level sets of least squares loss function with feasible sets for (a) lasso and (b) ridge regression in the case of $\beta \in \mathbb{R}^2$.

### 1.8.3 Non-Negative Garotte

This idea inspired the lasso. Proposed by Breiman [1995]. It minimizes

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

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.

### 1.8.4 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:

$$\hat{\beta}_{lasso} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 \right\} \text{ subject to } \|\beta\|_1 \leq t \tag{11}$$

The more common version now:

$$\hat{\beta}_{lasso} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \tag{12}$$

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 (12) yields

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

$$= -\frac{1}{n} X^T (y - X\beta) + \lambda \nabla\big(\|\beta\|_1\big) \tag{13}$$

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 $\partial$, the subdifferential, which is the set of all subgradients. We have a solution if 0 is in the subdifferential. We can rewrite (13) 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

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

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

Note that

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

so we have

$$0 \in -\frac{1}{n} X^T (y - X\beta) + \lambda \cdot \left[ \begin{cases} \text{sgn}(\beta_j) & t \neq 0 \\ [-1,1] & \beta_j = 0 \end{cases} \right] \tag{14}$$

where

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

is a vector with each entry as specified.

**Remark.** (1) Examining the $j$th component of the separable equation (14), 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 ($\lambda$).**

(2) If $\beta_j = 0$, 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 $\lambda$.**

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

So if we start with $\lambda$ 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 $y$.

### 1.8.5 LARS

In Figure 2, 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.

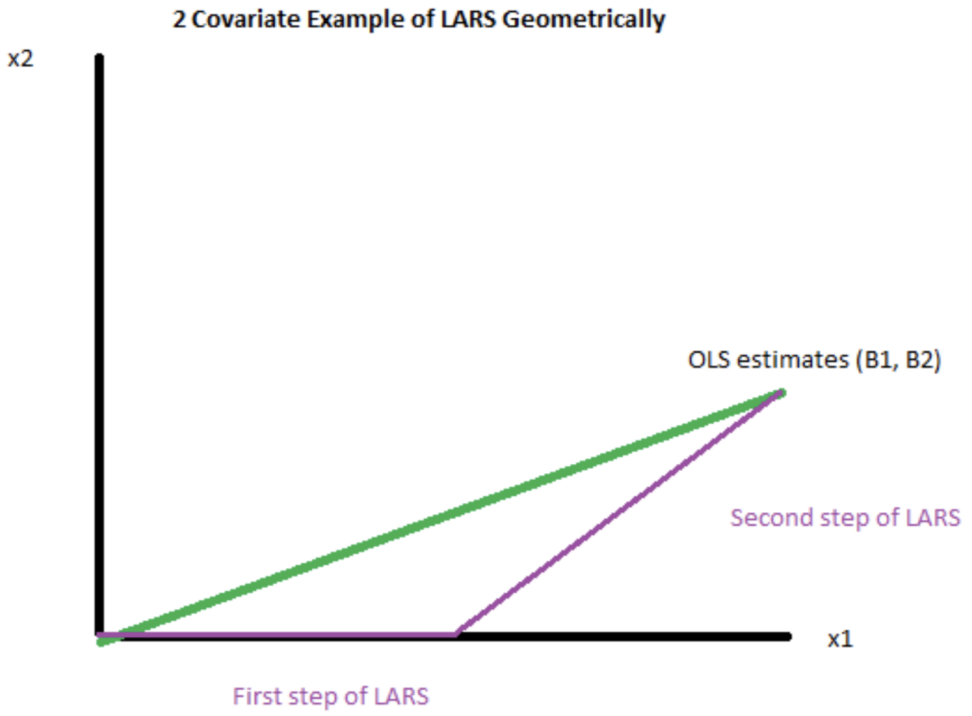


Figure 2: LARS figure in 2d case.

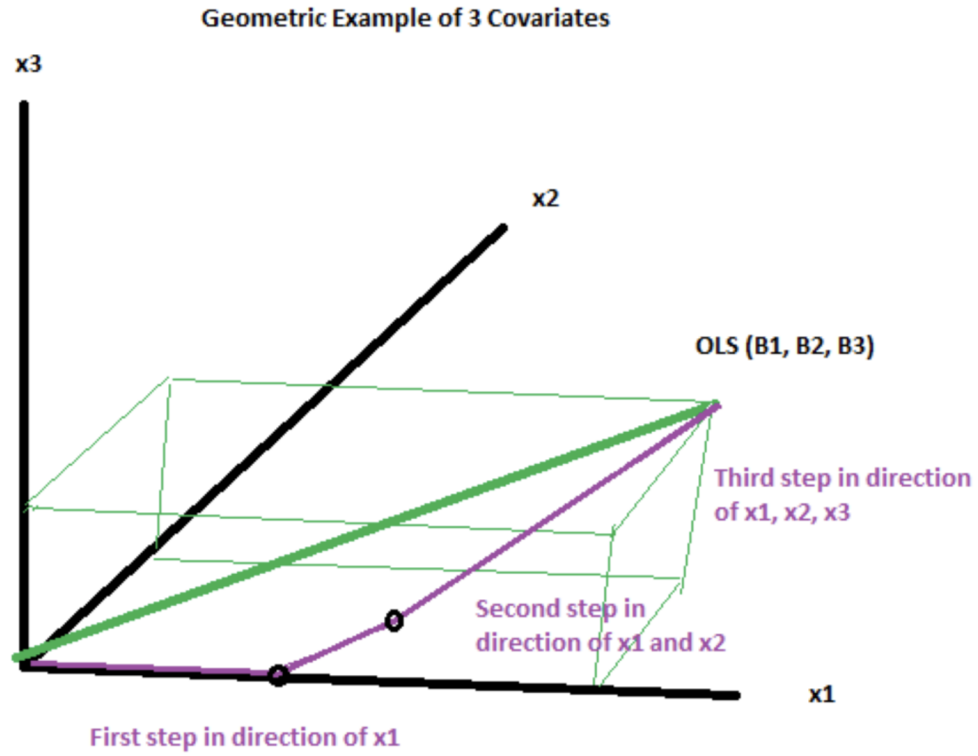


Figure 3: LARS figure in 3d case.

**Remark.** Just like in lasso, in LARS the correlation between all included features and the residual are equal (see the remark in Section 1.8.4). 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 $\beta_j$ 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 modification"), 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 $\mathcal{O}(n^2 p)$.

## 1.9 Quadratic Loss

**Theorem 11.** 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

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

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

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

□

### 1.9.1 Feature Selection properties

Model selection consistency: $\Pr(\text{supp}(\hat{\beta}) = \text{supp}(\beta_0)) \to 1$.

Oracle property: model selection consistency, asymptotic efficiency as efficient as if true model were known ("efficiency" having to do with the variance given $n$).

**Definition 1.1** (**Oracle property**)**.** Let $\beta^0$ denote the true parameter vector for data generated from a linear model. Let $S_0$ be the true support; that is, $S_0 = \{j : \beta_j^0 \neq 0, j = 1, \ldots, p\}$. Denote $\hat{\beta}(\delta)$ the coefficient estimator for fitting procedure $\delta$. We call $\delta$ an **oracle procedure** if $\hat{\beta}(\delta)$ asymptotically has the following properties:

- Identifies the right subset model (consistency): $\{j : \hat{\beta}_j \neq 0\} = S_0$.
- Has the optimal estimation rate: $\sqrt{n}(\hat{\beta}(\delta)_{S_0} - \beta^0_{S_0}) \xrightarrow{d} \mathcal{N}(0, \Sigma_0)$ where $\Sigma_0$ is the covariance matrix knowing the true subset model.

The lasso problem is convex but not necessarily strictly convex if $p > n$. That is, there is some flat region, so the minimizer may not be unique. Consider the KKT conditions from convex optimization:

$$g(\beta) = \arg\min\left\{\frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\right\} = \arg\min\left\{f_1(\beta) + f_2(\beta)\right\}$$

Then $\hat{\beta}$ is a lasso solution if and only if 0 is in the subdifferential of $g(\hat{\beta})$. Note that

$$\partial g(\hat{\beta}) = \nabla f_1 + \partial f_2 = \frac{1}{n}X^T(X\beta - y) + \lambda\begin{bmatrix}\vdots\\ \partial|\beta_j|\\ \vdots\end{bmatrix} = \frac{1}{n}X^T(X\beta - y) + \lambda\begin{bmatrix}\vdots\\ \begin{cases}\text{sgn}(\beta_j) & \beta_j \neq 0\\ [-1,1] & \beta_j = 0\end{cases}\\ \vdots\end{bmatrix}$$

Now assume $\text{supp}(\hat{\beta}) = \text{supp}(\beta_0)$ (that is, assume lasso recovers the correct support). Suppose the first $s$ features are nonzero and consider one of them (so we know that we should have $\hat{\beta}_j \neq 0$):

$$0 \in \partial g(\hat{\beta}) \implies 0 \in \partial_j g(\hat{\beta}) = \left[\frac{1}{n}X^T(X\beta - y)\right]_j + \lambda\,\text{sgn}(\hat{\beta}_j)$$

Therefore

$$\frac{1}{n}X_A^T(X\hat{\beta} - y) + \lambda\,\text{sgn}(\hat{\beta}_j) = 0 \tag{15}$$

where $X_A$ is a submatrix of $X$ containing the columns corresponding to the features in the true support, is our first condition. Next, consider what happens for $j > s$ (features not in the true support). We have

$$0 \in \partial g(\hat{\beta}) \implies 0 \in \partial_j g(\hat{\beta}) = \left[\frac{1}{n}X^T(X\beta - y)\right] + \lambda[-1,1]$$

$$\implies \left\|\frac{1}{n}X_{A^C}^T(X\hat{\beta} - y)\right\|_\infty \leq \lambda \tag{16}$$

where $X_{A^C}$ is a submatrix of $X$ containing the columns corresponding to the features not in the true support, is our boundary condition. Recall the true model

$$y = X\beta_0 + \epsilon$$

and consider the case $X = \begin{bmatrix}X_1 & X_2\end{bmatrix}$ where $X_1$ are the features in the true model and $X_2$ are noise features; that is, $\beta_0 = \begin{bmatrix}\beta_1\\ 0\end{bmatrix}$. Then we are assuming

$$\hat{\beta}_{\text{lasso}} = \begin{bmatrix}\hat{\beta}_1\\ 0\end{bmatrix}.$$

We have from (15)

$$0 = \frac{1}{n} X_1^T (X\hat{\beta} - y) + \lambda \operatorname{sgn}(\hat{\beta}_1) = \frac{1}{n} X_1^T (X_1\hat{\beta}_1 - X_1\beta_1 - \epsilon) + \lambda \operatorname{sgn}(\hat{\beta}_1)$$
$$\iff \frac{1}{n} X_1^T X_1 (\hat{\beta}_1 - \beta_1) = \frac{1}{n} X_1^T \epsilon - \lambda \operatorname{sgn}(\hat{\beta}_1)$$

Let's assume that $\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(\beta_0)$ (sign consistency).

$$\iff \frac{1}{n} X_1^T X_1 (\hat{\beta}_1 - \beta_1) = \frac{1}{n} X_1^T \epsilon - \lambda \operatorname{sgn}(\beta_1)$$

which is linear in $\hat{\beta}$. Solving, we have

$$\iff \hat{\beta}_1 - \beta_1 = (X_1^T X_1)^{-1}\big(X_1^T \epsilon - n\lambda \operatorname{sgn}(\beta_1)\big) \iff \hat{\beta}_1 = \beta_1 + (n^{-1} X_1^T X_1)^{-1}\big(n^{-1} X_1^T \epsilon - \lambda \operatorname{sgn}(\beta_1)\big) \quad (17)$$

Looking at the second (boundary) condition (16), we have

$$\left\| \frac{1}{n} X_2^T (X\hat{\beta} - y) \right\|_\infty \leq \lambda. \quad (18)$$

Consider that

$$X\hat{\beta} - y = X_1\hat{\beta}_1 - X_1\beta_1 - \epsilon = X_1(\hat{\beta}_1 - \beta_1) - \epsilon$$

Substituting in the result from (17) yields

$$X\hat{\beta} - y = X_1\big[(n^{-1} X_1^T X_1)^{-1}\big(n^{-1} X_1^T \epsilon - \lambda \operatorname{sgn}(\beta_1)\big)\big] - \epsilon$$

which when we plug into (18) yields

$$\left\| \frac{1}{n} X_2^T \left[ X_1 (n^{-1} X_1^T X_1)^{-1}\big(n^{-1} X_1^T \epsilon - \lambda \operatorname{sgn}(\beta_1)\big) - \epsilon \right] \right\|_\infty \leq \lambda.$$

$$\iff \left\| \frac{1}{n} X_2^T X_1 (n^{-1} X_1^T X_1)^{-1}\big(n^{-1} X_1^T \epsilon - \lambda \operatorname{sgn}(\beta_1)\big) - \frac{1}{n} X_2^T \epsilon \right\|_\infty \leq \lambda.$$

Using the Triangle Inequality, we have

$$\left\| \frac{1}{n} X_2^T X_1 (n^{-1} X_1^T X_1)^{-1}\big(n^{-1} X_1^T \epsilon - \lambda \operatorname{sgn}(\beta_1)\big) - \frac{1}{n} X_2^T \epsilon \right\|_\infty$$

$$\leq \left\| \frac{1}{n} X_2^T X_1 (n^{-1} X_1^T X_1)^{-1}\big(n^{-1} X_1^T \epsilon - \lambda \operatorname{sgn}(\beta_1)\big) \right\|_\infty + \left\| \frac{1}{n} X_2^T \epsilon \right\|_\infty$$

$$\leq \left\| \frac{1}{n} X_2^T X_1 (n^{-1} X_1^T X_1)^{-1} \right\|_\infty \cdot \left\| n^{-1} X_1^T \epsilon - \lambda \operatorname{sgn}(\beta_1) \right\|_\infty + \left\| \frac{1}{n} X_2^T \epsilon \right\|_\infty \tag{19}$$

Assume that the $j$th column of $X$ has $L_2$ norm $n^{1/2}$ (as it would if all entries equaled 1). We have

$$\left\| n^{-1} X_1^T \epsilon \right\|_\infty \leq \lambda/2, \qquad \left\| n^{-1} X_2^T \epsilon \right\|_\infty \leq \lambda/2$$

$$\left\| n^{-1} X^T \epsilon \right\|_\infty \leq \lambda/2 \text{ with large probability}$$

Recall that $\lambda = \sigma \sqrt{\frac{c \log p}{n}}$ for some $c > 2$. Then we have (continuing from (19)), and using $\left\| n^{-1} X_2^T \epsilon \right\| \leq \lambda/2$,

$$\leq \left\| n^{-1} X_1^T \epsilon \right\|_\infty + \left\| \lambda \operatorname{sgn}(\beta_1) \right\|_\infty$$

$$\left\| n^{-1} X_2^T X_1 (n^{-1} X_1^T X_1)^{-1} \right\|_\infty \cdot \underbrace{\|\cdot\|_\infty}_{3/2\lambda} + \underbrace{\|\cdot\|_\infty}_{\lambda/2} \leq \lambda$$

$$\left\| \underbrace{n^{-1} X_2^T X_1}_{\text{corr. between noise and true}} \left( \underbrace{n^{-1} X_1^T X_1}_{\text{sample covariance matrix}} \right)^{-1} \right\|_\infty \leq 1/3 \tag{20}$$

It turns out we're fine as long as it's less than or equal to 1. This is known as the **irrepresentable condition**. Note that the sample covariance matrix is the same as the sample correlation since the columns are standardized. So this is the correlation between the true variables. Note that this matrix has dimension $(p - s) \times s$ where $s$ is the dimension of the true support. Note that

$$n^{-1} X_2^T X_1 (n^{-1} X_1^T X_1)^{-1} = (X_2^T X_1 (X_1^T X_1)^{-1}]^T = (X_1^T X_1)^T X_1^T X_2$$

which is ordinary least squares for regressing $X_2$ on $X_1$. In the end, the irrepresentable condition says the correlation between the noise and true variables can't be too high.

## 1.10 Dantzig Selector

Dantzig selector:

$$\begin{aligned} \hat{\beta}_{\text{Dantzig}} = & \underset{\beta \in \mathbb{R}^p}{\arg\min} \quad \|\beta\|_1 \\ & \text{subject to} \quad \|n^{-1} X^T (y - X\beta)\|_\infty \leq \lambda \end{aligned}$$

Can be recast as a linear program:

$$\begin{aligned} \hat{\beta}_{\text{Dantzig}} = &\arg\min_{u \in \mathbb{R}^p} && \sum_{i=1}^{p} u_i \\ &\text{subject to} && -u \leq \beta \leq u \\ & && -\lambda_p \sigma \mathbf{1} \leq n^{-1} X^T (y - X\beta) \leq \lambda_p \sigma \mathbf{1} \end{aligned} \tag{21}$$

where $|u|$ denotes the absolute value of $u$ componentwise. (This is a benefit because linear programming is easy to use and very popular in industry and other applications.) Note that $n^{-1}X^t(y-X\beta)$ corresponds to the correlations between the residuals and the design matrix. Recall that in OLS this correlation is 0—the design matrix is orthogonal to the residuals. In the Dantzig selector we relax this, bounding the $L_\infty$ norm by $\lambda$. Recall that the gradient of the log-likelihood is the **score function**, in this case $n^{-1}X6t(y - X\beta)$. For example, the score equation in linear regression is $n^{-1}X^Ty = n^{-1}X^TX\beta$. Note:

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

Note for Theorem 1: in original paper, assumed columns had $L_2$ norm 1, resulting in $\lambda_p = \sqrt{2\log p}$. We are instead assuming each column has $L_2$ norm $\sqrt{n}$, which results in $\lambda = \sigma \cdot \sqrt{\frac{c\log p}{n}}$. Intuition of $\log p$ term:

By a theorem in James et al. [2009], the lasso and Dantzig selector estimates equal each other under certain conditions:

**Theorem 12.** Let $I_l$ be the support of the lasso estimate $\hat{\beta}_{\text{lasso}}$. Let $\boldsymbol{X}_L$ be the $n \times |I_L|$ matrix constructed by taking $\boldsymbol{X}_{I_L}$ and multiplying its columns by the signs of the corresponding coefficients in $\hat{\beta}_{\text{lasso}}$. Suppose that $\lambda_{\text{lasso}} = \lambda_{\text{Dantzig}}$. Then $\hat{\beta}_{\text{lasso}} = \hat{\beta}_{\text{Dantzig}}$ if $\boldsymbol{X}_L$ has full rank and

$$\boldsymbol{u} = \left(\boldsymbol{X}_L^T\boldsymbol{X}_L\right)^{-1}\mathbf{1} \succeq 0 \text{ and } \|\boldsymbol{X}^T\boldsymbol{X}_L\boldsymbol{u}\|_\infty \leq 1$$

where $\mathbf{1}$ is an $|I_L|$-vector of ones and the vector inequality is understood componentwise.

**Corollary 12.1.** If $\boldsymbol{X}$ is orthonormal ($\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_p$), then the entire lasso and Dantzig selector coefficient paths are identical.

*Proof.* For each index set $\boldsymbol{I}$, $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_{|\boldsymbol{I}|}$, so clearly both of the conditions of Theorem 12 are satisfied.

□

The entire paths can be identical under another condition presented in the same paper.

**Theorem 13.** Suppose that all pairwise correlations between the columns of $\boldsymbol{X}$ are equal to the same value $\rho$ where $0 \leq \rho < 1$. Then the entire Lasso and Dantzig selector coefficient paths are identical. In addition, when $p = 2$, the same holds for every $\rho \in (-1, 1)$.

## 1.11 Coordinate Descent

Start with $\beta_1$ varying and all other $\beta$s fixed. Optimize $\beta_1$. Then cycle through each $\beta_j$, run until convergence.

### 1.12 Nonconvex Learning

### 1.13 Yinfei Kong—Innovated Interaction Screening for HIgh-Dimensional Nonlinear Classification

### 1.14 Total Variational Distance

## References

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csáki, editors, *2nd International Symposium on Information Theory*, pages 267–281, Tsahkadsor, Armenia, USSR, 1973.

A. Antoniadis and J. Fan. Regularization of Wavelet Approximations. *Journal of the American Statistical Association*, 96:939–967, 2001. ISSN 0162-1459. doi: 10.1198/016214501753208942. URL https://www.tandfonline.com/action/journalInformation?journalCode=uasa20.

L. Breiman. Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, 37(4):373–384, 1995. URL https://www-jstor-org.libproxy2.usc.edu/stable/pdf/1269730.pdf?refreqid=excelsior{%}3A76eea9bd08301e990d7d6edd86067262.

D. L. Donoho and I. M. Johnstone. Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika*, 81 (3):425–455, 1994. URL https://www-jstor-org.libproxy2.usc.edu/stable/pdf/2337118.pdf?refreqid=excelsior{%}3Afb36dc2b9ad4d3d57225eebb75e05df2.

G. M. James, P. Radchenko, and J. Lv. DASSO: connections between the Dantzig selector and lasso. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 71(1):127–142, 2009. URL https://rss-onlinelibrary-wiley-com.libproxy1.usc.edu/doi/pdf/10.1111/j.1467-9868.2008.00668.x.

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.

G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978. URL https://www.andrew.cmu.edu/user/kk3n/simplicity/schwarzbic.pdf.