Math Review Notes—Linear Regression

Gregory Faletto

Contents

Last updated October 2, 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) 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 $\text{Cov}(y_j, \hat{\beta}_1)$:

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 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,

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

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

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

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

Next we find the variance.

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

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

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

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

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

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

(b)

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

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

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

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

□

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

1.2 Chapter 2: Multiple Regression

General OLS:

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

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

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

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

Or:

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

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

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

Theorem 3 (**Gauss-Markov Theorem, as stated in Pesaran [2015]**). Suppose we have data generated by

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

and we make the following assumptions.

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

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

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

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

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

$$\implies \tilde{\beta} = \beta + (X^T X)^{-1} X^T \epsilon + C^T \epsilon \iff \tilde{\beta} - \beta = [(X^T X)^{-1} X^T + C^T]\epsilon \tag{2}$$

Then

$$\text{Var}(\tilde{\beta}) = \mathbb{E}\left[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^T\right] = \mathbb{E}\left[\tilde{\beta}\tilde{\beta}^T - \tilde{\beta}\beta^T - \beta\tilde{\beta}^T + \beta\beta^T\right] \tag{3}$$

Using (2) we have

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

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

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

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

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

$$= \beta\beta^T + \sigma^2(X^TX)^{-1} + \sigma^2C^TC \tag{4}$$

Also,

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

$$\implies \mathbb{E}(\tilde{\beta}\beta^T) = \beta\beta^T, \qquad \mathbb{E}\left(\beta\tilde{\beta}^T\right) = \mathbb{E}\left(\left[\tilde{\beta}\beta^T\right]^T\right) = \beta\beta^T. \tag{5}$$

Substituting (4) and (5) into (3), we have

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

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

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

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

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

□

Theorem 4 (**Gauss-Markov Theorem, as stated in Faraway [2002]**). Suppose

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

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

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

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

This implies

$$a^T X = c^T \iff X^T a = c, \tag{6}$$

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

$$c = X^T X\lambda = X^T a \tag{7}$$

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

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

$$= \text{Var}(a^T y - \lambda^T X^T\hat{y}) + \text{Var}(c^T\hat{\beta}) + 2\text{Cov}(a^T y - \lambda^T X^T\hat{y}, c^T\hat{\beta}) \geq \text{Var}(c^T\hat{\beta}) + 2\text{Cov}(a^T y - \lambda^T X^T\hat{y}, c^T\hat{\beta}), \tag{8}$$

so we are done if the covariance in (8) is nonnegative.

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

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

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

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

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

Note that

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

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

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

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

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

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

since by (7) $c^T = a^T X = \lambda^T X^T X$.

□

Theorem 5 (**Gauss-Markov Theorem, as stated in Faraway [2002], more general case**). Suppose

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

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

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

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

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

This implies

$$a^T X = c^T \iff X^T a = c, \tag{9}$$

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

$$c = X^T X\lambda = X^T a \tag{10}$$

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

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

$$= \text{Var}(a^T y - \lambda^T X^T \hat{y}) + \text{Var}(c^T\hat{\beta}) + 2\text{Cov}(a^T y - \lambda^T X^T\hat{y}, c^T\hat{\beta}) \succeq \text{Var}(c^T\hat{\beta}) + 2\text{Cov}(a^T y - \lambda^T X^T\hat{y}, c^T\hat{\beta}), \quad (11)$$

so we are done if the covariance matrix in (11) is positive semidefinite.

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

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

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

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

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

since by (10) $c^T = a^T X = \lambda^T X^T X$.

□

1.3 Chapter 3: Hypothesis testing in regression

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

Lemma 6.

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

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

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

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

Lemma 7. If ϵ is i.i.d. with $E(\epsilon_i) = 0$ and $\mathbb{E}(\epsilon_i^2) = \sigma^2$ for all i, the elements of the matrix X are uniformly bounded so that $|X_{ij}| < U$ for all i and j and for U finite, and $\lim_{n\to\infty} X'X/n = Q$ is finite and nonsingular, then

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

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

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

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

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

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

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

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

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

where

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

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

□

Theorem 8. Under the conditions of Lemma 7 (ϵ is i.i.d. with $E(\epsilon_i) = 0$ and $\mathbb{E}(\epsilon_i^2) = \sigma^2$ for all i, the elements of the matrix X are uniformly bounded so that $|X_{ij}| < U$ for all i and j and for U finite, and $\lim_{n\to\infty} X'X/n = Q$ is finite and nonsingular),

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

Proof.

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

Since $\lim_{n\to\infty}(X'X/n)^{-1} = Q^{-1}$ and by Lemma 7

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

then

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

□

t-test statistic:

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

F-test statistic:

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

Since

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

we have

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

yielding

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

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

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

where

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

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

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

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

1.3.1 ANOVA

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

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

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

1.4 Chapter 4: Heteroskedasticity

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

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

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

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

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

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

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

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

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

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

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

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

1.5 Chapter 5: Autocorrelated disturbances

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.

Suppose

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

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

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

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

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

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

Example application: spatial data.

1.6 Quantile Regression

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

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

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

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

In fixed dimensions:

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

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

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

This is then maximized for $\tau = 1/2$, and the max value is $1/4 \cdot 2\pi = \pi/2$. For more information on other loss functions, see also Section 1.14.

1.6.1 Detecting outliers in multiple dimensions

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

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

Half-space depth (Tukey): make polygons of observations (outer one is the convex hull). Then create convex hulls of inside, keep going. See Figure 1.

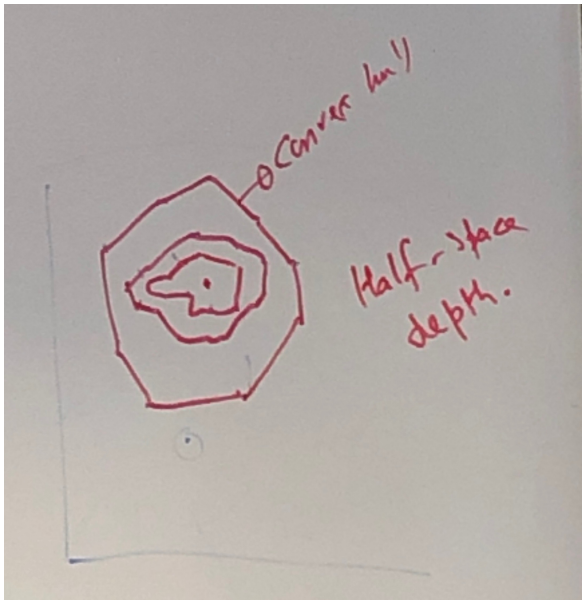

Figure 1: Illustration of half-space depth method for detecting outliers.

1.7 Transformed Linear Models

1.7.1 Transformations of response

Consider the transformation

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

One example: **log linear model.**

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

Multiplicative model in the true response:

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

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

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

Consider

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

Then we have

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

Try

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

for $\lambda = -2, -2, -1/2, 0, 1/2, 1, 2$. We have $2L(\hat{\lambda}) - 2L(\lambda_{true}) \sim \chi_1^2$. If 1 is inside the 95% confidence interval, don't transform; if 1 isn't, do. In a lot of cases, changing the response can be a pain for interpretability.

1.7.2 Transforming predictor values

Add polynomial terms:

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

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

Broken stick regression/segmented regression: useful if data has two different groups.

$$y = B_\ell(x) + B_r(x) = \beta_0 + \beta_1(c - x)_+ + \beta_2(x - c)_+$$

where c is the breaking point and is fixed (if you use data to choose c, becomes a nonlinear problem). Advantages: more localized

Splines: like combination of broken stick regression and polynomial regression. fit a polynomial in each segment.

1.8 Segmented regression, local regression, splines

for splines: knot selection is important (bias/variance tradeoff)

1.8.1 Local Regression

Local regression: weight chosen by kernel:

$$w_i = \frac{K((x_i - x_0)/\sqrt{v})}{\sum_{i=1}^{M(s)} K((x_i - x_0)/\sqrt{v})}$$

$$\text{dnorm}(x_i, \text{mean} = x_0; \text{sd} = \sqrt{v})$$

$$K((x_i - x_0)/\sqrt{v}) = \phi((x_i - x_0)/\sqrt{v}) \cdot v^{-1/2} = \frac{1}{\sqrt{2\pi v}} \exp - \left(\frac{(x_i - x_0)^2}{2v} \right)$$

so if the distance between two points increases than the weight decreases. This is called kernel density with a Gaussian kernel.

$$x = x_0; \hat{\beta}(x_0); \hat{\sigma}(x_0)$$

Two tuning parameters: v (bandwidth) and s. Disadvantages: hard to interpret, starts to do poorly in high dimensions.

1.8.2 Curse of Dimensionality (brief discussion)

Suppose $x_i : i \in [n] \sim F$ i.i.d. with fixed dimension, and F has bounded support $[a, b]^d$. Let $x \in \text{supp}(F)$. Then

$$\min_x d_2(x, x_i) \sim \frac{1}{n^{1/d}}$$

so goes to 0 as $n \to \infty$ for fixed d. But bad in d; for example, if $d = 5$, need $n = 10^5$ for $d = 0.1$. As this minimum distance increases, regularity becomes more difficult.

1.9 Dimension Reduction methods

1.9.1 Principal components regression

(See Section **??** for more details on principal components.) Given data (y, X), don't look at y for now. Look for maximum variability direction of X:

$$P_1 = \underset{\|a\|_2=1}{\arg\max} \text{Var}(a^T x) = a^T (X^T X) a.$$

then (letting $\mathcal{P}_1$ be the projection matrix for projecting onto P_1)

$$P_2 = \underset{\|a\|_2=1, \mathcal{P}_1 a=0}{\arg\max} \operatorname{Var}(a^T x) = a^T (X^T X) a.$$

and so on. We will regress y against the first M principal components of X for some $X \leq p$, where the principal components are the columns of U:

$$\hat{y}^{\text{pcr}}_{(M)} = \overline{y} + \sum_{m=1}^{M} \hat{\theta}_m z_m,$$

where $\hat{\theta}_m = (z_m^T y)/(z_m^T z_m)$ since all the z_m are orthogonal. Also, since the z_m are linear combinations of the original x_j, the solution can be expressed in terms of coefficients in the original feature space:

$$\hat{\beta}^{\text{pcr}}(M) = \sum_{m=1}^{M} \hat{\theta}_m v_m,$$

Works well when high variability directions in the explanatory variables are also interesting attributes to study.

Sometimes if data are in a mainfold that isn't linearly parsed, good to use an ISOMAP or LLE (local linear embedding).

(read: starting on p.63 of ESL)

1.9.2 Partial least squares

Does look at y, unlike PCR. $X_{n\times p} \to W_{n\times 3}$. Start by standardizing all x_j to have mean 0 and unit variance. Then compute $\hat{\phi}_{1j} = \langle x_j, y \rangle$ for each j. Then the derived input $z_1 = \sum_j \hat{\phi}_{1j} x_j$ is the first partial least squares direction. Then y is regressed on z_1 giving coefficient $\hat{\theta}_1$, and then $x_1, \ldots, x_j$ are orthogonalized with respect to z_1. Continue until $M \leq p$ directions have been obtained. (see p. 80 of ESL)

$$\vdots$$

notes from GSBA 604: First reduced dimension:

$$w^{(1)}_{n\times 1} = X^{(1)}_{n\times 1} + X^{(2)}_{n\times 1} + \ldots + X^{(p)}_{n\times 1}$$

proportional correlation $(y, x^{(1)})$ to

1.9.3 Dimension reduction by random matrix

Let $R \in \mathbb{R}^{p\times d}$, $d < p$, with $R_{ij} \sim \mathcal{N}(0,1)$. Let $\tilde{X} = XR$. By the Johnson-Lindenstrauss lemma, for any $\epsilon > 0$ there exists an R such that

$$(1-\epsilon)\|\tilde{X}_i - \tilde{X}_j\|_2^2 \leq \|X_i - X_j\|_2^2 \leq (1+\epsilon)\|\tilde{X}_i - \tilde{X}_j\|_2^2.$$

(isometric transformation—distances are maintained)

1.10 Goodness of fit, residuals, residual diagnostics, leverage

Goodness of fit: F test. Assume $\text{Var}(\epsilon) = \sigma^2 I$ and recall $\hat{\epsilon} = (I - H)\epsilon$. So $\text{Var}(\hat{\epsilon}) = \sigma^2(I - H)$ and $\text{Var}(\hat{\epsilon}_i) = \sigma^2(1 - H_{ii})$. So $h_i := H_{ii}$ is called the **leverage** for the ith case. Some properties: if X is **ill-conditioned** (condition number—ratio of largest to smallest eigenvalue of Gram matrix $X^T X$—is high), then H_{ii} will vary a lot.

Properties: $\sum_i h_i = p + 1$ (because H_{ii} is idempotent, its trace is equal to $p + 1$—all its eigenvalues are 0 or 1, and it is nonsingular, so all of its eigenvalues are 1). And, all $h_i \geq 1/n$.

Larger h_i result in smaller $\text{Var}(\hat{\epsilon}_i)$, which forces the fit to be close to y_i. Average leverage: $(p+1)/n$. Rule of thumb: leverages of more than $2(p+1)/n$ should be looked at closely (they have a large influence on the slope, so if they are incorrect then it's a big problem for the fit).

Standardized or Studentized results:

$$r_i = \frac{\hat{\epsilon}_i}{\text{se}(\hat{\epsilon}_i)} = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

Let $\hat{\beta}_i$ and $\hat{\sigma}_i^2$ be the estimates from the regression with the ith case excluded. Let x_i^T denote the ith row of X, write $X_{(i)}$ for X without the ith row.

$$\vdots$$

Outlier test: can test for outliers using the fact that each $t_i \sim t(n - p - 2)$ if no outliers are present. Need to do a multiplicity correction for multiple testing (say Bonferroni correction). have to do residual diagnostics.

1.10.1 Residual diagnostics

Consider leverage points (outliers, influential points: points that change the slope a lot).

1. **Partial residual plot:** kth variable. $\hat{\epsilon}^{(k)} = y - \sum_{j \neq k} \beta_j x_j$: residuals without the kth explanatory variable. Plot this against x_k and regress. If the resulting slope $\hat{\beta}_k$ is large, this is an influential point.
2. **Added variable plot:** Regress x_k by other explanatory variables $\hat{\delta}^{(k)}$. Plot $\hat{\epsilon}^{(k)}$ against $\hat{\delta}^{(k)}$.

1.11 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 θ_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 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).

1.11.1 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 θ_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 θ_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 θ_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 θ_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 Sections **??** and **??**). For AIC, we minimize the KL divergence. For BIC, we maximize the Bayes factor (posterior probability for the model).

1.11.2 Bayesian Information Criterion (BIC)

A typical Bayesian model selection procedure is to first give nonzero prior probability α_M om each model M and then prescribe a prior distribution μ_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 9 (**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

$$s_0^2 = \frac{1}{n-p_0}e_R{}'e_R, s_1^2 = \frac{1}{n-p_1}e_U{}'e_U \tag{12}$$

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

$$AIC_0 < AIC_1$$

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

$$\iff \frac{s_0^2}{s_1^2} < e^{\frac{2}{n}(p_1 - p_0)} \tag{13}$$

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 (13) 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 (12), 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

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

$$= \frac{{e_R}'e_R - {e_U}'e_U}{{e_U}'e_U} < \frac{2}{n}(p_1 - p_0) \tag{14}$$

Now recall the F statistic:

$$F = \frac{({e_R}'e_R - {e_U}'e_U)/g}{{e_U}'e_U/(n-k)} \tag{15}$$

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

$$\frac{(e_R{}'e_R - e_U{}'e_U)}{e_U{}'e_U} < \frac{2(p_1-p_0)}{n} \tag{16}$$

which equals (14). 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).

□

1.12 Ridge Regression

If p is large, it tends to be better to shrink the last squares estimator. (Even though this introduces bias, it will likely reduce variance, and the tradeoff will often help for some amount of shrinkage.) This is related to the Stein estimator.

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

$$\sum_{i=1}^{n} \left(y_i - {x^{(i)}}^T \boldsymbol{\beta}\right)^2 + \lambda \|\boldsymbol{\beta}\|_2^2 = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \tag{17}$$

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 (17). We call the $\hat{\beta}$ that minimizes (17) the **ridge regression** estimator for β.

Proposition 10 (Math 541A Homework Problem). The value of $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ that minimizes (17) is $\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

Proof.

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

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

where $\boldsymbol{y}^T\boldsymbol{X\beta} = \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}$ 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 11 (DSO 607 Homework Problem). Suppose $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$ 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 $\lambda > 0$, the probability that each component of the ridge estimator $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ equals 0 is 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 (and because the Moore-Penrose pseudoinverse minimizes the ℓ_2 norm, exactly what the ridge solution will do).

(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 ith entry equal to 1 and all other entries equal to 0. Let the ith 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 ϵ 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 2.)

□

GSBA 604: using SVD $X = UDV^T$, we have

$$\hat{y}_{ridge}(\lambda) = X\hat{\beta}_{ridge}(\lambda) = X(X^TX + \lambda I)^{-1}X^Ty = UDV^T(VD^2V^T + \lambda I)^{-1}VDU^Ty$$

$$= UDV^T(V(D^2 + \lambda I)V^T)^{-1}VDU^Ty = UDV^T[(D^2 + \lambda I)V^T]^{-1}V^TVDU^Ty = UDV^TV(D^2 + \lambda I)^{-1}DU^Ty$$

$$= UD(D^2 + \lambda I)^{-1}DU^Ty = \sum_{j=1}^{n} \frac{d_j^2}{d_j^2 + \lambda} u_j u_j^T y \tag{18}$$

also,

$$\hat{\beta}_{LS} = \sum_{j=1}^{n} v_j(u_j^T y)$$

so by comparison, what ridge regression is doing is changing the weights of the columns (by weights that are between 0 and 1 for any $\lambda > 0$). Higher d_j's means higher weights. So the directions that have higher variability are shrunk less.

So in the limit of (18) as $\lambda \to 0^+$, we have

$$\hat{\beta}_{\text{ridge}} = (X^TX + \lambda I_p)^{-1}X^Ty \to X^\dagger y = X^\dagger(X\beta + \epsilon) \tag{19}$$

where we substitute the pseudoinverse instead of the inverse because since X^TX is rank deficient, $(X^TX)^{-1}$ does not exist (and because the Moore-Penrose pseudoinverse minimizes the ℓ_2 norm, exactly what the ridge solution will do).

$$\hat{y}_{ridge}(\lambda) = X\hat{\beta}_{ridge}(\lambda) \to \sum_{j=1}^{n} u_j u_j^T y = UU^T y,$$

which is the least squares solution in the case that $p \leq n$ and this exists, or (19) in the general case.

1.13 Lasso

From KKT theory, the correlation between all selected features and residual will be λ (see the remark in Section 1.13.4 for an explanation why).

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

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

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 β, 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:

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

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 12 (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 $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^p$. Our observed data are $\boldsymbol{y} = (y_1, \dots, y_n)^T \in \mathbb{R}^n$. 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

$$\sum_{i=1}^{n} \left(\boldsymbol{y}_i - {x^{(i)}}^T \boldsymbol{\beta} \right)^2 + \lambda \sum_{i=1}^{p} |\boldsymbol{\beta}_i| \tag{20}$$

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 (20) as

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

$$= \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{21}$$

By Proposition **??**, $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ is convex, and by Proposition **??**, $\lambda\|\boldsymbol{\beta}\|_1$ is convex. Therefore by Proposition **??**, (21) is convex. Differentiating and setting equal to 0 yields

$$-2\boldsymbol{y}^T\boldsymbol{X} + 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X} + \lambda\left[\text{sgn}(\boldsymbol{\beta}_i)\right] = 0 \tag{22}$$

where $\left[\text{sgn}(\boldsymbol{\beta}_i)\right]$ is vector resulting from the sgn function being applied elementwise to $\boldsymbol{\beta}$. Since (22) is linear in $\boldsymbol{\beta}$, it has one solution. Since (21) is convex, any solution to (22) minimizes (20).

□

Remark. The L_1 penalization term in (20) 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 (20) 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 (20), but technically Lagrange Multipliers does not apply since the constraint is not differentiable everywhere.

1.13.1 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 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 13. 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."

1.13.2 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 11 and Figure 2).

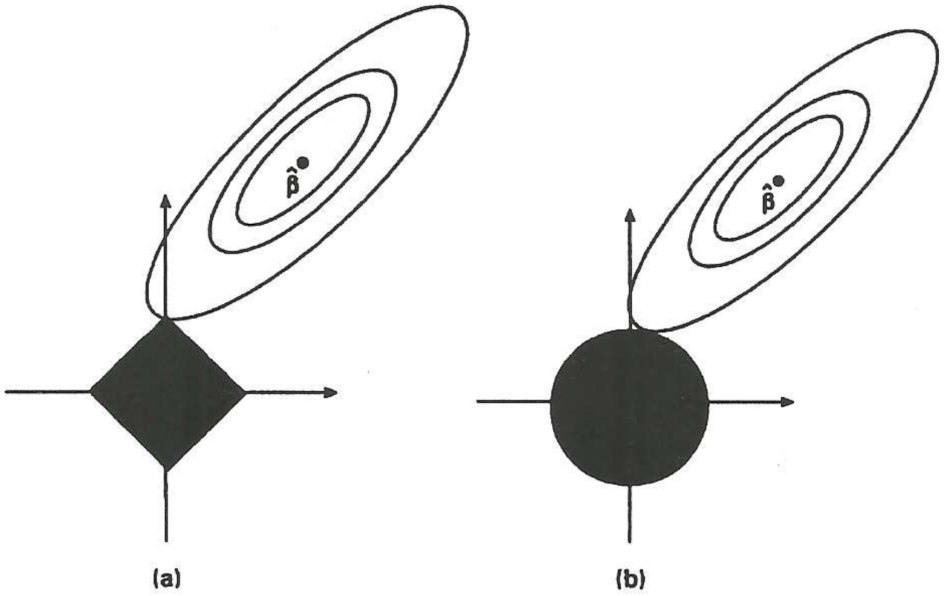

Figure 2: Level sets of least squares loss function with feasible sets for (a) lasso and (b) ridge regression in the case of $\beta \in \mathbb{R}^2$.

Proposition 14 (2018 DSO Statistics Group In-Class Screening Exam, Question 5). Consider the optimization problem

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{23}$$

where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, and $\lambda > 0$.

(a) The following problem is a dual of (23):

$$\underset{u \in \mathbb{R}^n}{\text{maximize}} \quad \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - u\|_2^2 \quad \text{subject to} \quad \|X^T u\|_\infty \leq \lambda.$$

Also, $\hat{u} = y - X\hat{\beta}$, where $\hat{\beta}$ is a solution of (23) and $\hat{u}$ is a solution of the dual.

(b) $\hat{\beta}$ is not necessarily unique, but $\hat{u}$, $\|y - X\hat{\beta}\|_2^2$, and $\|\hat{\beta}\|_1$ are.

(c) Suppose $y = X\beta^* + \epsilon$, and suppose the tuning parameter λ is chosen to satisfy $\lambda \geq \|X^T\epsilon\|_\infty$. Then

(i)
$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2}\|\epsilon\|_2^2 + \lambda\|\beta^*\|_1.$$

(ii)
$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \geq \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|X\beta^*\|_2^2.$$

(iii)
$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \geq \frac{1}{2}\|\epsilon\|_2^2 - \lambda\|\beta^*\|_1.$$

Remark. We can express the original optimization problem (23) as

$$\begin{aligned} \underset{\beta\in\mathbb{R}^p, z\in\mathbb{R}^n}{\text{minimize}} \quad & \frac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 \\ \text{subject to} \quad & z = X\beta. \end{aligned} \tag{24}$$

We will also refer to another expression of the lasso optimization problem,

$$\begin{aligned} \underset{\beta\in\mathbb{R}^p}{\text{minimize}} \quad & \frac{1}{2}\|y - X\beta\|_2^2 \\ \text{subject to} \quad & \|\beta\|_1 \leq t \end{aligned} \tag{25}$$

for some $t > 0$.

Before proving the main results, we will show a few simpler results. Whenever $\lambda > 0$, the lasso objective function (23) is the Lagrangian of (25). We will prove a useful lemma about the relationship between these functions.

Lemma 15. For a given $\lambda > 0$, let $\hat{\beta}$ minimize (23). Then there is exactly one $t = \|\hat{\beta}\|_1$ such that any $\hat{\beta}$ minimizing (23) also minimizes (25).

Proof. This must be true by contradiction. First of all, since the objective function of (25) is continuous and the feasible region $\|\beta\|_1 \leq t$ is compact, a minimum of (25) is guaranteed to exist. Now suppose $\hat{\beta}$ minimizes (23) for a fixed λ, with $\|\hat{\beta}\|_1 = t$, but there is a different solution $\hat{\beta}^*$ that is feasible for (25) and achieves a lower value. That is,

$$\frac{1}{2}\|y - X\hat{\beta}^*\|_2^2 < \frac{1}{2}\|y - X\hat{\beta}\|_2^2.$$

and $\|\hat{\beta}^*\|_1 \leq \|\hat{\beta}\|_1 = t$. Since $\lambda > 0$, $\|\hat{\beta}\|_1 < \|\hat{\beta}_{global}\|_1$, where $\hat{\beta}_{global}$ is a global minimum for $\frac{1}{2}\|y - X\hat{\beta}\|_2^2$. Since (25) is convex and all global minima lie outside the feasible region, $\hat{\beta}^*$ lies on the boundary; that is, $\|\hat{\beta}^*\|_1 = \|\hat{\beta}\|_1 = t$. But then

$$\frac{1}{2}\|y - X\hat{\beta}^*\|_2^2 < \frac{1}{2}\|y - X\hat{\beta}\|_2^2 \iff \frac{1}{2}\|y - X\hat{\beta}^*\|_2^2 + \lambda\|\hat{\beta}^*\|_1 < \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1$$

which contradicts the fact that $\hat{\beta}$ minimizes (23).

□

Another useful result follows in a simple way from Lemma 15.

Proposition 16. Let $\mathcal{B}$ be the set of all $\hat{\beta}$ that minimize (23) for some fixed $\lambda > 0$. Then for any two $\hat{\beta}_1, \hat{\beta}_2 \in \mathcal{B}$, $\|\hat{\beta}_1\|_1 = \|\hat{\beta}_2\|_1$. That is, $\|\hat{\beta}\|_1$ is unique.

Proof. Suppose $\hat{\beta}_1$ and $\hat{\beta}_2$ both minimize (23), and (without loss of generality) $\|\hat{\beta}_1\|_1 < \|\hat{\beta}_2\|_1$. By Lemma 15, these values both minimize (25) with $t = \|\hat{\beta}_2\|_1$ (we cannot choose $t = \|\hat{\beta}_1\|_1$ because $\hat{\beta}_1$ is not feasible for that problem). Because the global minimum of (25) lies outside the feasible region and (25) is convex, all solutions to (25) lie on the boundary of the feasible region. But $\|\hat{\beta}_1\|_1 < \|\hat{\beta}_2\|_1$, so $\hat{\beta}_1$ is not on the boundary of the feasible region, contradiction. Therefore $\|\hat{\beta}_1\|_1 = \|\hat{\beta}_2\|_1$ for all solutions $\hat{\beta}_1$, $\hat{\beta}_2$ to (23); that is, $\|\hat{\beta}\|_1$ is unique. (See Osborne et al. [2000] for more details.)

□

Now we are ready to prove Proposition 14.

Proof of Proposition 14. (a) The Lagrangian of (24) is

$$\mathcal{L}(\beta, z, u) = \frac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 + u^T(z - X\beta),$$

so the Lagrange dual function is

$$\inf_{\beta,z}\{\mathcal{L}(x,u)\} = \inf_{\beta,z}\left\{\frac{1}{2}\|y - z\|_2^2 + \lambda\|\beta\|_1 + u^T(z - X\beta)\right\}$$

$$= \inf_{\beta,z}\left\{\frac{1}{2}(y-z)^T(y-z) + u^Tz + \lambda\|\beta\|_1 - u^TX\beta\right\}$$

This minimization is separable:

$$= \inf_z\left\{\frac{1}{2}\left(y^Ty - 2y^Tz + z^Tz\right) + u^Tz\right\} + \inf_\beta\left\{\lambda\|\beta\|_1 - u^TX\beta\right\} \tag{26}$$

We will handle each part of (26) separately. First, the left side:

$$\inf_z\left\{\frac{1}{2}\left(y^Ty - 2y^Tz + z^Tz\right) + u^Tz\right\} = \inf_z\left\{\frac{1}{2}z^Tz + (u-y)^Tz + \frac{1}{2}y^Ty\right\}$$

Since this is a convex quadratic form, differentiate with respect to z and set equal to zero:

$$z + (u - y) = 0 \implies z = y - u \tag{27}$$

$$\implies \inf_z\left\{\frac{1}{2}z^Tz + (u-y)^Tz + \frac{1}{2}y^Ty\right\} = \frac{1}{2}(y-u)^T(y-u) + (u-y)^T(y-u) + \frac{1}{2}y^Ty$$

$$= \frac{1}{2}\left(y^Ty - 2u^Ty + u^Tu\right) + 2u^Ty - y^Ty - u^Tu + \frac{1}{2}y^Ty = -\frac{1}{2}u^Tu + u^Ty = \frac{1}{2}y^Ty - \frac{1}{2}y^Ty + u^Ty - \frac{1}{2}u^Tu$$

$$= \frac{1}{2}y^Ty - \frac{1}{2}(y^Ty - 2u^Ty + u^Tu) = \frac{1}{2}y^Ty - \frac{1}{2}(y-u)^T(y-u) = \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y-u\|_2^2$$

Next we will minimize the right side of (26):

$$\inf_\beta\left\{\lambda\|\beta\|_1 - u^TX\beta\right\} = \inf_\beta\left\{\lambda\sum_{i=1}^p|\beta_i| - \sum_{i=1}^p\left[u^TX\right]_i\beta_i\right\} = \inf_\beta\left\{\sum_{i=1}^p\left(\lambda|\beta_i| - \left[u^TX\right]_i\beta_i\right)\right\}$$

$$= \inf_{\beta} \left\{ \sum_{i=1}^{p} \left(\text{sgn}(\beta_i)\lambda - \left[u^T X\right]_i \right) \beta_i \right\} = \sum_{i=1}^{p} \inf_{\beta_i} \left\{ \left(\text{sgn}(\beta_i)\lambda - \left[u^T X\right]_i \right) \beta_i \right\}.$$

Notice that when β_i is negative, if $\left(\text{sgn}(\beta_i)\lambda - \left[u^T X\right]_i \right) = -\left(\lambda + \left[u^T X\right]_i \right)$ is positive there is no lower bound on the quantity we are minimizing; otherwise, when β_i is negative the infimum is 0. When β_i is positive, if $\left(\text{sgn}(\beta_i)\lambda - \left[u^T X\right]_i \right) = \left(\lambda - \left[u^T X\right]_i \right)$ is negative there is no lower bound on the quantity we are minimizing; otherwise, when β_i is negative the infimum is 0. That is, the only dual feasible points satisfy for all i

$$-\left(\lambda + \left[u^T X\right]_i \right) \leq 0, \qquad \lambda - \left[u^T X\right]_i \geq 0 \iff \left[u^T X\right]_i \geq -\lambda, \qquad \left[u^T X\right]_i \leq \lambda$$

which is equivalent to the condition

$$\|u^T X\|_\infty \leq \lambda.$$

Therefore the Lagrange dual function is

$$\inf_{\beta, z} \{\mathcal{L}(x, u)\} = \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - u\|_2^2 \tag{28}$$

subject to the constraint $\|u^T X\|_\infty \leq \lambda$. This quantity represents a lower bound on the minimum value of the original optimization problem for all $u \in \mathbb{R}^p$. The dual problem is to find the best lower bound by maximizing over u; that is, the dual problem is

$$\begin{aligned} \underset{u \in \mathbb{R}^p}{\text{maximize}} \quad & \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - u\|_2^2 \\ \text{subject to} \quad & \|u^T X\|_\infty \leq \lambda. \end{aligned} \tag{29}$$

Lastly, suppose $\hat{\beta}$ and $\hat{u}$ satisfy

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \quad \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \quad ,$$

$$\hat{u} = \begin{array}{ll} \underset{u \in \mathbb{R}^p}{\arg\max} & \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - u\|_2^2 \\ \text{subject to} & \|u^T X\|_\infty \leq \lambda \end{array} = \begin{array}{ll} \underset{u \in \mathbb{R}^p}{\arg\min} & -\frac{1}{2}\|y\|_2^2 + \frac{1}{2}\|y - u\|_2^2 \\ \text{subject to} & \|u^T X\|_\infty \leq \lambda \end{array}$$

Then since (27) is a requirement for dual feasibility of u and strong duality applies, we have $\hat{u} = y - X\hat{\beta}$.

Remark. Note that we could also find the arguments maximizing (29) by

$$\begin{array}{ll} \underset{u \in \mathbb{R}^p}{\arg\min} & -\frac{1}{2}\|y\|_2^2 + \frac{1}{2}\|y - u\|_2^2 \\ \text{subject to} & \|u^T X\|_\infty \leq \lambda. \end{array} = \begin{array}{ll} \underset{u \in \mathbb{R}^p}{\arg\min} & \frac{1}{2}\|y - u\|_2^2 \\ \text{subject to} & \|u^T X\|_\infty \leq \lambda. \end{array}$$

where the first step follows from the fact that arguments that maximize a function are the same as the arguments that minimize the negative of a function, and the second step follows from the fact that the $-\frac{1}{2}\|y\|_2^2$ term does not include u. Therefore we see that the residual vector u from a lasso fit can be thought of as the projection of y onto the convex polyhedron $C \subset \mathbb{R}^n$ defined by $C := \{u : \|X^T u\|_\infty \leq \lambda\}$.

Another way of saying this is that the lasso estimate $\hat{y} = X\hat{\beta}_{\text{lasso}}$ itself is the residual from projecting y onto C; that is,

$$X\hat{\beta}_{\text{lasso}} = (I - P_C)y,$$

where P_C is the operator projecting y onto C.

(b) (i) **Not necessarily unique.** Per Tibshirani [2013], if $\text{rank}(X) < p$, the lasso solution is not necessarily unique. Intuitively, this is because the columns of X are linearly dependent, so there may exist more than one linear combination of the columns that minimizes (23). **Jacob's suggestion: counterexample. X is two columns that are equal; then convex combinations of two solutions are equal as long as same sign (can't be opposite sign because then ℓ_1 could be smaller by setting one equal to 0.**

(ii) **Necessarily unique.** The dual problem (29) is strictly concave, so the value $\hat{u}$ that maximizes it is unique.

(iii) **Necessarily unique** (except in the trivial case $\lambda = 0$). Per part 5(b)(iv), $\|\hat{\beta}\|_1$ is unique. (23) is convex, so the minimum $\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1$ is unique. Therefore $\|y - X\hat{\beta}\|_2^2$ must be unique.
Jacob's solution: Since $\hat{u}$ is unique and by (27) $\hat{u} = y - X\hat{\beta}$, we must have that $\|\hat{u}\| = \|y - X\hat{\beta}\|$ is unique.

(iv) **Necessarily unique** (except in the trivial case $\lambda = 0$). This is immediate from Proposition 16.

(c) (i) Since β^* is clearly feasible for (23) and $\hat{\beta}$ achieves the minimum, we have

$$\frac{1}{2}\|y - X\beta^*\|_2^2 + \lambda\|\beta^*\|_1 \geq \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \iff \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2}\|\epsilon\|_2^2 + \lambda\|\beta^*\|_1$$

(ii) We know that the expression in the dual problem (29) is a lower bound for the solution of the primal problem (23) for any u feasible for (29) (that is, any u satisfying $\|u^T X\|_\infty \leq \lambda$). Therefore we have

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\beta\|_1 \geq \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - u\|_2^2.$$

Since by assumption $\lambda \geq \|X^T\epsilon\|_\infty$, ϵ is feasible for (29). Therefore we have

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\beta\|_1 \geq \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|y - \epsilon\|_2^2 = \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|X\beta^*\|_2^2 \tag{30}$$

as desired.

(iii) We can rewrite the right side of (30) as

$$\frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|X\beta^*\|_2^2 = \frac{1}{2}\|X\beta^*\|_2^2 + \frac{1}{2}\|\epsilon\|_2^2 + \epsilon^T X\beta^* - \frac{1}{2}\|X\beta^*\|_2^2 = \frac{1}{2}\|\epsilon\|_2^2 + \epsilon^T X\beta^*. \tag{31}$$

By assumption, we have

$$\lambda \geq \|X^T\epsilon\|_\infty \iff \lambda\mathbf{1} - X^T\epsilon \succeq 0 \implies \lambda\mathbf{1}\beta^* - X^T\epsilon\beta^* \succeq 0$$

$$\iff -\lambda\|\beta^*\|_1 \leq \epsilon^T X\beta^* \leq \lambda\|\beta^*\|_1.$$

By Hölder's Inequality, we have for any two vectors $u, v \in \mathbb{R}^n$, $|u^T v| \leq \|u\|_\infty\|v\|_1$. Therefore

$$|\epsilon^T X\beta^*| = |(X^T\epsilon)^T\beta^*| \leq \|X^T\epsilon\|_\infty \|\beta^*\|_1 \leq \lambda\|\beta^*\|_1$$

where the last step used the assumption $\|X^T\epsilon\|_\infty \leq \lambda$. So we have

$$\frac{1}{2}\|\epsilon\|_2^2 + \lambda\|\beta^*\|_1 \leq \frac{1}{2}\|\epsilon\|_2^2 + \epsilon^T X\beta^*.$$

Substituting in to (30), using the identity in (31), and using the result from part 5(c)(iii) yields

$$\frac{1}{2}\|\epsilon\|_2^2 + \lambda\|\beta^*\|_1 \leq \frac{1}{2}\|\epsilon\|_2^2 + \epsilon^T X\beta^* = \frac{1}{2}\|y\|_2^2 - \frac{1}{2}\|X\beta^*\|_2^2 \leq \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\beta\|_1$$

as desired.

(iv) We see from parts (i) and (iii) that

$$\frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\beta\|_1 - \lambda\|\beta^*\|_1 \leq \frac{1}{2}\|\epsilon\|_2^2 \leq \frac{1}{2}\|y - X\hat{\beta}\|_2^2 + \lambda\|\beta\|_1 + \lambda\|\beta^*\|_1$$

$$\iff \frac{1}{n}\|y - X\hat{\beta}\|_2^2 + \frac{2}{n}\lambda\|\beta\|_1 - \frac{2}{n}\lambda\|\beta^*\|_1 \leq \frac{1}{n}\|\epsilon\|_2^2 \leq \frac{1}{n}\|y - X\hat{\beta}\|_2^2 + \frac{2}{n}\lambda\|\beta\|_1 + \frac{2}{n}\lambda\|\beta^*\|_1$$

that is, we can lower bound and upper bound $\frac{1}{n}\|\epsilon\|_2^2$ by taking the quantity $\frac{1}{n}\|y - X\hat{\beta}\|_2^2 + \frac{2}{n}\lambda\|\beta\|_1$ and adding or subtracting $\frac{2}{n}\lambda\|\beta^*\|_1$. Therefore it seems that the quantity in the middle of this interval, $\frac{1}{n}\|y - X\hat{\beta}\|_2^2 + \frac{2}{n}\lambda\|\beta\|_1$, is a reasonable estimator for $\sigma^2 = \mathbb{E}\left[n^{-1}\|\epsilon\|_2^2\right]$.

□

1.13.3 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.

1.13.4 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:

$$\hat{\beta}_{lasso} = \underset{\beta\in\mathbb{R}^p}{\arg\min}\left\{\frac{1}{2n}\|y - X\beta\|_2^2\right\} \text{ subject to } \|\beta\|_1 \leq t \tag{32}$$

The more common version now:

$$\hat{\beta}_{lasso} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \tag{33}$$

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 (33) yields

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

$$= -\frac{1}{n} X^T (y - X\beta) + \lambda \nabla\big(\|\beta\|_1\big) \tag{34}$$

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 (34) 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

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

so we have

[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.

$$0 \in -\frac{1}{n}X^T(y - X\beta) + \lambda \cdot \begin{bmatrix} \begin{cases} \text{sgn}(\beta_j) & t \neq 0 \\ [-1, 1] & \beta_j = 0 \end{cases} \end{bmatrix} \tag{35}$$

where

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

is a vector with each entry as specified.

Remark. (1) Examining the jth component of the separable equation (35), 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 $\beta_j = 0$, 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.

1.13.5 LARS

In Figure 3, 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 (see the remark in Section 1.13.4). 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 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 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 $\mathcal{O}(n^2 p)$.

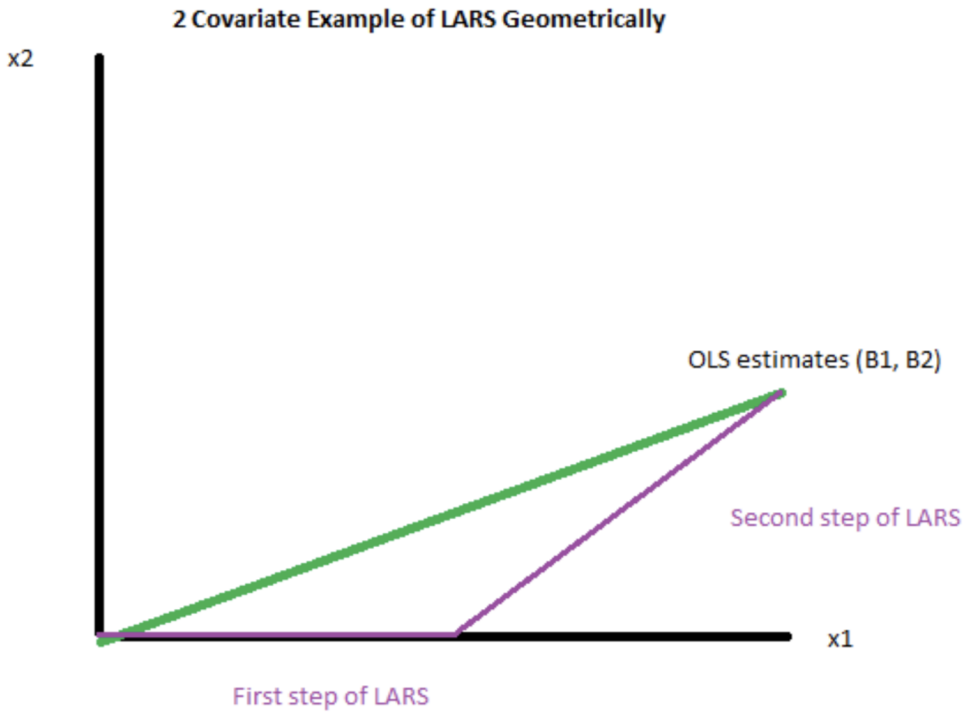

Figure 3: LARS figure in 2d case.

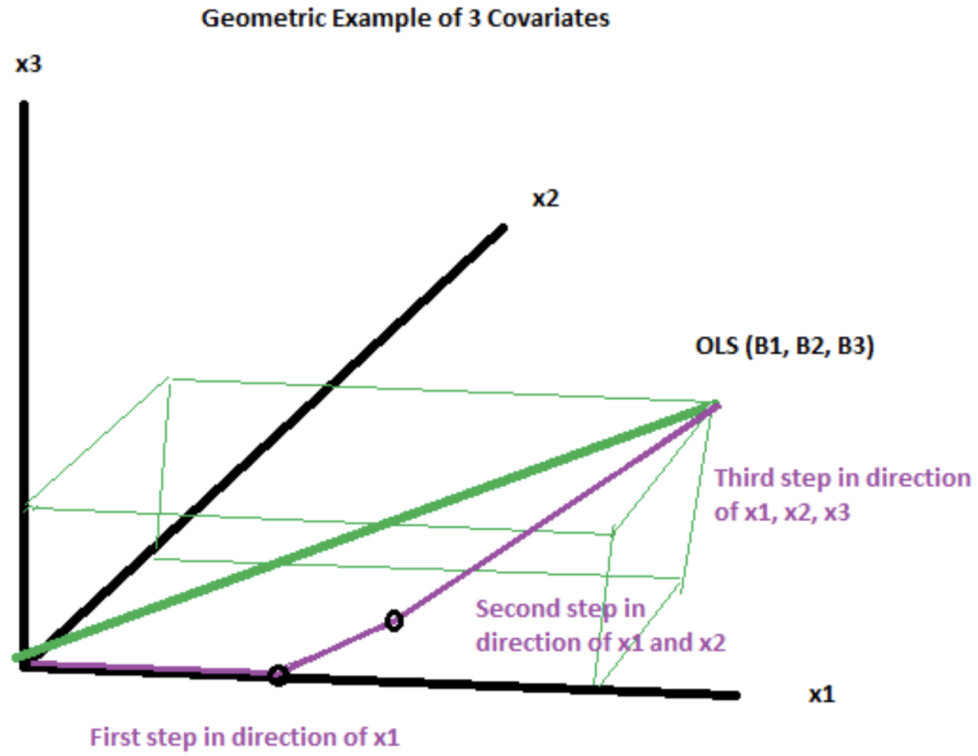

Figure 4: LARS figure in 3d case.

1.14 Loss Functions

Asymmetric loss: may have asymmetry between overestimation and underestimation. For example: in a supply chain, estimate inventory level $y_1, \dots, y_n$. Let $y_i = x_i'\beta + \epsilon_i$. Underestimating is really bad because then you don't have enough product for customers and they might not come back; overestimating not so bad. Then our loss function:

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

you could choose $\tau = 0.1$ for a 9 to 1 ratio of loss (i.e., underestimation is 9 times as costly as overestimation).

Theorem 17 (Loss: quadratic). 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

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

□

Huber loss: combines benefits of squared loss (unbiasedness) with MAD loss (robust).

$$L_\delta^H(y, \hat{y}) = (y-\hat{y})^2 \cdot I\left\{|y-\hat{y}| \le \delta\right\} + |y-\hat{y}| \cdot I\left\{|y-\hat{y}| > \delta\right\}.$$

Only downside: not that easy to estimate (loss function is not differentiable). Instance of **M-estimation** (more generalized than regression):

$$\hat{\beta}_M = \arg\min_{\beta,\sigma} \sum_{i=1}^{n} \rho_M\left(\frac{y_i - x_i'\beta}{\sigma}\right)$$

where ρ_M is a loss function. Suppose $y_i = \sum_{j=1}^p x_{ij}\beta_j + \epsilon_i$. Then differentiating the loss function with respect to β yields

$$\frac{\partial}{\partial \beta_k}\left[\sum_{i=1}^{n} \rho_M\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right] = 0 \qquad \text{for } k \in \{1, \dots, p\}.$$

$$= \sum_{i=1}^{n} \rho'(y_i - x_i'\beta) \cdot \frac{(-x_{ij})}{\sigma}$$

Let $u_i := (y_i - x_i'\beta)/\hat{\sigma}$.

$$\implies \sum_{i=1}^{n} \underbrace{\frac{\rho'(u_i)}{u_i}}_{\text{weight}} \cdot x_{ij}(y_i - x_i'\beta) = 0$$

Compare to least squares: normal equations $X^T(y - X^T\beta_{OLS}) = 0$, or

$$\sum_{i=1}^{n} x_{ij}(y_i - x_i'\beta) = 0$$

Now we have this extra weighting term. Algorithm for computing:

1. Use $\hat{\beta}_{OLS}$ as the initial solution; then compute

$$u_i^{(0)} = \frac{y_i' - x_i'\hat{\beta}_{OLS}}{\hat{\sigma}_{OLS}}$$

2. With $w_i^{(0)} = \rho'(u_i)/u_i$, compute

$$\hat{\beta}_{LS}^{(1)} = (X^T W X)^{-1} X^T W y.$$

3. Update weights with $\hat{\beta}_{LS}^{(1)}$.
4. Repeat until convergence. (Convergence is a little tricky but should work.)

For absolute deviation loss, see Section 1.6 on quantile regression.

1.14.1 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 β^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 δ. We call δ 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 Σ_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:

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

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

$$\frac{1}{n} X_A^T (X\hat{\beta} - y) + \lambda \, \text{sgn}(\hat{\beta}_j) = 0 \tag{36}$$

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

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

$$\implies \left\| \frac{1}{n} X_{A^C}^T (X\hat{\beta} - y) \right\|_\infty \leq \lambda \tag{37}$$

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 (36)

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

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

$$\iff \hat{\beta}_1 - \beta_1 = (X_1^T X_1)^{-1}\big(X_1^T \epsilon - n\lambda \operatorname{sgn}(\beta_1)\big) \iff \hat{\beta}_1 = \beta_1 + (n^{-1} X_1^T X_1)^{-1}\big(n^{-1} X_1^T \epsilon - \lambda \operatorname{sgn}(\beta_1)\big) \quad (38)$$

Looking at the second (boundary) condition (37), we have

$$\left\| \frac{1}{n} X_2^T (X\hat{\beta} - y) \right\|_\infty \leq \lambda. \quad (39)$$

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 (38) 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 (39) 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

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

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

$$\leq \left\| \frac{1}{n} X_2^T X_1 (n^{-1} X_1^T X_1)^{-1} \right\|_\infty \cdot \left\| n^{-1} X_1^T \epsilon - \lambda \operatorname{sgn}(\beta_1) \right\|_\infty + \left\| \frac{1}{n} X_2^T \epsilon \right\|_\infty \tag{40}$$

Assume that the jth 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 (40)), and using $\left\| n^{-1} X_2^T \epsilon \right\| \leq \lambda/2$,

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

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

$$\left\| \underbrace{n^{-1} X_2^T X_1}_{\text{corr. between noise and true}} \; (\; \underbrace{n^{-1} X_1^T X_1}_{\text{sample covariance matrix}} \;)^{-1} \right\|_\infty \leq 1/3 \tag{41}$$

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.

1.15 Dantzig Selector

Dantzig selector:

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

Can be recast as a linear program:

$$\begin{aligned}\hat{\beta}_{\text{Dantzig}} = &\arg\min_{u\in\mathbb{R}^p} \quad \sum_{i=1}^p u_i \\ &\text{subject to} \quad -u \le \beta \le u \\ &\qquad\qquad -\lambda_p\sigma\mathbf{1} \le n^{-1}X^T(y-X\beta) \le \lambda_p\sigma\mathbf{1}\end{aligned} \tag{42}$$

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_∞ norm by λ. 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 18. 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

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

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

Corollary 18.1. 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 18 are satisfied.

□

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

Theorem 19. Suppose that all pairwise correlations between the columns of $\boldsymbol{X}$ are equal to the same value ρ 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 $\rho \in (-1, 1)$.

1.16 Coordinate Descent

Start with β_1 varying and all other βs fixed. Optimize β_1. Then cycle through each β_j, run until convergence.

1.17 Total Variational Distance

1.18 Non-parametric regression

One example: LOESS

Another: GAM

1.18.1 Generalized additive models

Suppose $y_i = f(x_i) + \epsilon_i$. Suppose $0 \leq x_1 \leq \ldots \leq x_n \leq 1$. Then express

$$f(x) = \sum_{i=1}^{\infty} \theta_i \phi_i(x)$$

where $\int_{\mathbb{R}} f^2(x)\ dx < \infty$ and the ϕ_i form an orthonormal basis. Assume sparsity:

$$f(x) \approx \sum_{i=1}^{p} \theta_i \phi_i(x).$$

So

$$y_i = \sum_{i=1}^{p} \theta_i \phi_i(x) + \epsilon_i, \qquad i \in \{1, \ldots, n\}$$

and if f is smooth then the coefficient basis is sparse, so θ is sparse, so

$$\|\theta\|_p = \left[\sum_{i=1}^{n} |\theta_i|^p \right]^{1/p}$$

is small.

Can choose a wavelet basis.

1.19 Mixture regression

Suppose

$$y_i \sim \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}\left(X_i'\beta_k, \sigma_k^2 I\right).$$

Mixture modeling is very useful. Typically estimated by expectation maximization (see Section **??**). One example: spatial analysis. See Figure 5. In the black region, slopes and standard deviations will be much higher; outside, will be much lower.

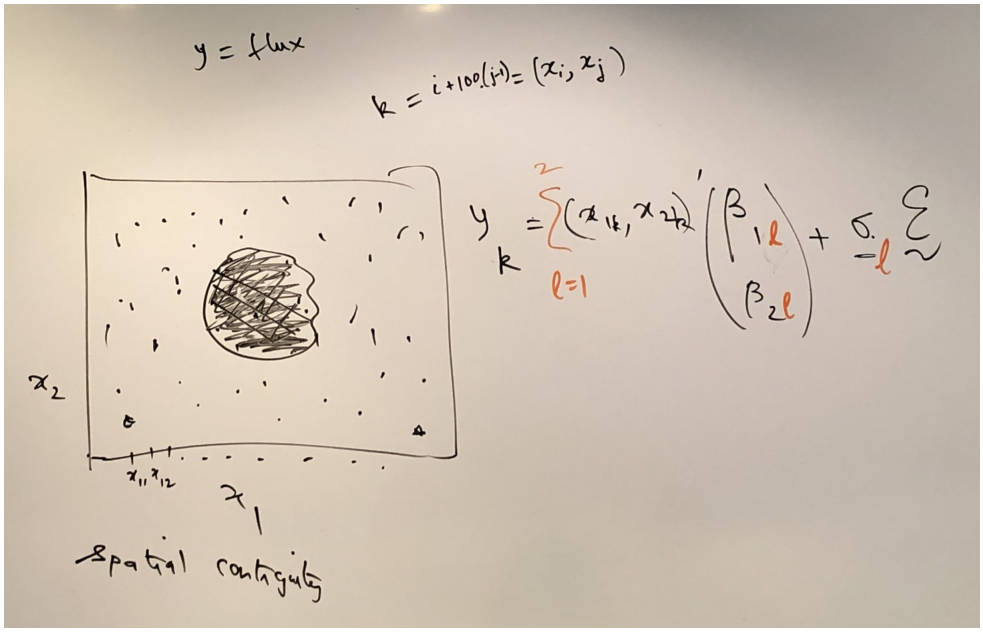

Figure 5: Mixture model example.

1.20 Missing observations

1.21 Generalized linear models

Exponential family: binomial, Poisson, Beta, negative binomial, etc. See Section **??** for more information on exponential families. The density is

$$g_\eta(y) = c(y)\exp\{\eta y - \psi(\eta)\}$$

where $c(y)$ is the **carrier density**, η is the **natural parameter**, y is a **sufficient statistic**, and $\psi(\cdot)$ is

the **normalizing function** (or **cumulant generating function**), chosen so that it is a proper density: that is, choose $\psi(\eta)$ such that

$$\int_{\mathbb{R}} g_\eta(y)\ dy = 1,$$

so

$$e^{-\psi(y)} = \left(\int_{\mathbb{R}} c(y)e^{\eta y}\ dy\right)^{-1}$$

The density is simplest to express in the natural parameter.

Skewness: third centered moment.

$$\frac{\mathbb{E}(Y - \mathbb{E}(Y))^3}{\text{Var}(Y)^{3/2}}$$

kurtosis: 4th centered moment

$$\frac{\mathbb{E}(Y - \mathbb{E}(Y))^4}{\text{Var}(Y)}$$

get excess kurtosis by subtracting 3 (which is Gaussian kurtosis).

In general: cumulant, either centered or raw.

$$K_\gamma = \mathbb{E}(Y - \mathbb{E}(Y))^\gamma, \qquad K_\gamma = \mathbb{E}(Y^\gamma)$$

Cumulant generating function:

$$\psi(\eta) - \psi(\eta_0) = \sum_{\gamma=1}^{\infty} K_\gamma \frac{(\eta - \eta_0)^r}{r!}$$

if you know all the cumulants, you know the distribution exactly.

1. **Gaussian variables:** Suppose $y_1, y_2, \dots y_n$ are i.i.d. $\mathcal{N}(\mu, 1)$.
2. **Gaussian response:** Suppose (y_i, x_i) are i.i.d. pairs in $\mathbb{R}^{p+1}$ with $y_i \sim \mathcal{N}(x_i'\beta, 1)$.
3. **Exponential family variables:** Suppose $Y_1, \dots, Y_n$ are i.i.d. in an exponential family. Then the density of each is $g_\eta(y) = e^{\eta y 0 \psi(\eta)} c(y)$ and the joint density is

$$\prod_{i=1}^n g_\eta(y_i) = \exp\left(\eta \sum_{i=1}^n y_i - n\psi(\eta)\right) \cdot \prod_{i=1}^n c(y_i). \tag{43}$$

   Then the density of $\overline{y}$ is simply

$$\exp\{n\eta\overline{y} - n\psi(\eta)\}$$

now the natural parameter is $n\eta$ and the normalizing function is $n\psi(\eta)$. So

$$\overline{Y} \sim \left[n\frac{\mathrm{d}\psi(\eta)}{\mathrm{d}\eta}, n\frac{\mathrm{d}\psi^2(\eta)}{\mathrm{d}\eta^2}\right].$$

The log likelihood is the log of (43):

$$\log L(\eta) = \ell(\eta) = \sum_{i=1}^{n} \eta y_i - n\psi(\eta) + \sum_{i=1}^{n} \log c(y_i).$$

The maximum likelihood estimate solves

$$\hat{\eta}_{MLE} := \arg\max_{\eta} \ell(\eta)$$

we have

$$\frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta) = \sum_{i=1}^{\eta} y_i - n\frac{\mathrm{d}}{\mathrm{d}\eta}\psi(\eta), \qquad \frac{\mathrm{d}^2}{\mathrm{d}\eta^2}\ell(\eta) = -n\frac{\mathrm{d}^2}{\mathrm{d}\eta^2}\psi(\eta) = -n\mathrm{Var}(\eta) < 0$$

so since this function is concave down it has a unique maximum at

$$\sum_{i=1}^{n} y_i - n\frac{\mathrm{d}}{\mathrm{d}\eta}\psi(\hat{\eta}_{MLE}) \implies \psi(\hat{\eta}_{MLE}) = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y} \implies \hat{\mu}_{MLE} = \overline{y}$$

That is, the MLE of the mean parameter in an exponential family is the sample mean. Due to the equivariance property of the maximum likelihood estimator, the MLE for any function of the mean parameter is simply the function of the MLE for the mean parameter (see Proposition **??**).

We call the derivative $\frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta)$ the **score function**. If $\xi = h(\eta)$, then by the chain rule

$$\frac{\mathrm{d}}{\mathrm{d}\xi}\ell(\eta) = \frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta) \Big/ \frac{\mathrm{d}\xi}{\mathrm{d}\eta}$$

For the mean parameter in particular, we have

$$\frac{\mathrm{d}}{\mathrm{d}\mu}\ell(\eta) = \frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta) \Big/ \frac{\mathrm{d}\mu}{\mathrm{d}\eta} = n(\overline{y} - \mu)/\mathrm{Var}_{\eta}(Y) \tag{44}$$

(since $\frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta) = n\overline{y} - n\frac{\mathrm{d}}{\mathrm{d}\eta}\psi(\eta)$ and $\mathrm{Var}_{\eta}(Y) = \frac{\mathrm{d}\mu}{\mathrm{d}\eta}$). Also, since the expectation of the score function $\mathbb{E}(\frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta)) = 0$, we hace

$$\mathrm{Var}\left(\frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta)\right) = \mathbb{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta)\right)^2\right] - \left[\mathbb{E}\left(\frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta)\right)\right]^2 = \mathbb{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta)\right)^2\right]$$

We call this the **Fisher information:**

$$i_{\eta}^{(n)}(\xi) := \mathbb{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\eta}\ell(\eta)\right)^2\right]. \tag{45}$$

(See also Definiton **??** and Section **??** for more on this.) In the case of the mean, we can write this as

$$i_\eta^{(n)}(\eta) = \mathbb{E}\left[\left(n\overline{y} - n\frac{\mathrm{d}}{\mathrm{d}\eta}\psi(\eta)\right)^2\right] = \mathrm{Var}\left(n\overline{y} - n\frac{\mathrm{d}}{\mathrm{d}\eta}\psi(\eta)\right) = \mathrm{Var}(n\overline{y}) = n^2\mathrm{Var}(\overline{y}) = n^2\mathrm{Var}_\eta(y)/n = n\mathrm{Var}_\eta(y)$$

Then substituting (44) into (45) we have

$$i_\eta^{(n)}(\mu) = \mathbb{E}\left[\frac{\left(n\overline{y} - n\frac{\mathrm{d}}{\mathrm{d}\eta}\psi(\eta)\right)^2}{\mathrm{Var}_\eta(y)^2}\right] = \frac{\mathrm{Var}\left[n\overline{y} - n\frac{\mathrm{d}}{\mathrm{d}\eta}\psi(\eta)\right]}{\mathrm{Var}_\eta(y)^2} = \frac{n\mathrm{Var}_\eta(y)}{\mathrm{Var}_\eta(y)^2} = \frac{n}{\mathrm{Var}_\eta(y)}. \tag{46}$$

Note that as n increases, the Fisher information increases linearly. The Fisher information is inversely proportional to the underlying variance. Again, by Proposition **??**, the plug-in estimator for the Fisher Information using the maximum likelihood estimator for the variance is the MLE for the Fisher information:

$$i_\eta^{(n)}(\mu)\Big|_{\eta=\hat{\eta}_{MLE}} = \frac{n}{\widehat{\mathrm{Var}}_\eta(y)} = \mathbb{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\mu}\ell(\eta)\right)^2\right] = -\mathbb{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\ell(\eta)\right]\Big|_{\eta=\hat{\eta}_{MLE}}$$

In the case of the mean parameter, we can write this as

$$= -\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\ell(\eta)\Big|_{\eta=\hat{\eta}_{MLE}}$$

Cramer-Rao lower bound (see Theorem **??**):

$$\mathrm{Var}_\eta(\hat{\xi}) \geq \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\eta}\left[\mathbb{E}_\eta(\hat{\eta})\right]\right)^2}{i_\eta^{(n)}(\xi)}$$

Any estimator that reaches (or approaches asymptotically) this lower bound is **(asymptotically) efficient**. Consider an unbiased estimator $\hat{\xi}$ ($\mathbb{E}(\hat{\xi} = \xi)$: then the bound reduces to

$$\mathrm{Var}_\eta(\hat{\xi}) \geq \frac{1}{i_\eta^{(n)}(\xi)}$$

Further, suppose we have an unbiased estimator of the mean parameter $\hat{\mu}$. Then using (46) we have

$$\mathrm{Var}_\eta(\hat{\mu}) \geq \frac{1}{i_\eta^{(n)}(\mu)} = \frac{\mathrm{Var}_\eta(y)}{n}$$

Since we saw earlier that the MLE has exactly this variance ($\mathrm{Var}(\overline{y}) = \mathrm{Var}_\eta(y)/n$), it is the efficient unbiased estimator for μ. (However, it turns out that the plug-in MLE estimator for a function of this mean parameter is not in general the efficient unbiased estimator.)

4. **Exponential family model (GLM):**

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.

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

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. R. Osborne, B. Presnell, and B. A. Turlach. On the LASSO and its Dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000. ISSN 1537-2715. doi: 10.1080/10618600.2000.10474883. URL https://www.tandfonline.com/action/journalInformation?journalCode=ucgs20.

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.

R. J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7(1):1456–1490, 2013. ISSN 19357524. doi: 10.1214/13-EJS815.