3.7.0 → Overview

→ overview → recall that MLR relies on ordinary least squares to determine the coefficient estimates.

In the next subsection, we present alternatives that vary slightly from a MLR fit, namely:

→ shrinkage (regularization) methods → ridge & lasso

→ dimension reduction methods → PCA + partial least squares

→ These methods may assume that variables used will be centered or scaled

---

3.7.1 → Standardizing variables

→ overview → A centered variable is the result of subtracting the sample mean from a variable

→ A scaled variable is the result of dividing a variable by its sample st dev

→ A standardized variable is the result of first centering a variable, then scaling it.

→ Across various resources, it is common to see that scaling uses the biased sample st dev (denom of n rather than n−1) & others the unbiased version. In many cases, this choice does not have a large impact on the final results.

→ If an exam problem is not explicit, we anticipate scaling w/ the biased sample st dev, denote this s.

→ Example → original model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ ⇒ $\hat{\beta}_0 = 59.13$, $\hat{\beta}_1 = -2.29$, $\hat{\beta}_2 = -0.08$ ✓ (different)

→ Centered predictors: $y = \beta_0 + \beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \varepsilon$ ⇒ $\hat{\beta}_0 = 40$, $\hat{\beta}_1 = -2.29$, $\hat{\beta}_2 = -0.08$ (same)

⇒ only $\hat{\beta}_0$ is different, but it leads to the same fitted equation

$\hat{y} = 40 - 2.29(x_1 - 8) - 0.08(x_2 - 10)$ ⟨$\bar{y}$⟩ ⟨#s from example⟩

$= 40 - 2.29x_1 + 8(2.29) - 0.08x_2 - 0.08(10)$ ⟨rearrange⟩

$= 59.13 - 2.29x_1 - 0.08x_2$ ✓

→ Scaled model: $y = \beta_0 + \beta_1\left(\frac{x_1}{s_1}\right) + \beta_2\left(\frac{x_2}{s_2}\right) + \varepsilon$ ⇒ $\hat{\beta}_0 = 59.13$ (same as unscaled), $\hat{\beta}_1 = -4.57$, $\hat{\beta}_2 = -0.33$ (different — — —)

⇒ Again, same fitted equation ⇒ model is scale equivariant

$y = 59.13 - 4.57\left(\frac{x_1}{2}\right) - 0.33\frac{x_2}{4}$

$= 59.13 - 2.29x_1 - 0.08x_2$ ✓

→ when using OLS, note that:

→ Centering &/or scaling the predictors doesn't intrinsically change the fitted equation

→ when centered predictors are used, the intercept estimate will equal the sample mean of the response, while the rest of the estimated regression coefficients remain unchanged

→ when scaled predictors are used, the intercept estimate does not change. However, the other coefficient estimates will change by a factor equal to the sample st dev of the associated explanatory variable

⇒ Standardized model: $y = \beta_0 + \beta_1\left(\frac{x_1 - \bar{x}_1}{s_1}\right) + \beta_2\left(\frac{x_2 - \bar{x}_2}{s_2}\right) + \varepsilon$ ⇒ $\hat{\beta}_0 = 40$, $\hat{\beta}_1 = -4.57$, $\hat{\beta}_2 = -0.33$

⟨... same fitted equation again ✓ ...⟩

→ Training / testing

→ Suppose we pre-process training set w/ $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_p)$ & $s = (s_1, \ldots, s_p)$, when we pre-process the test set, we should reuse these SAME $\bar{x}$ & s to avoid making different standardizations

---

3.7.2 → Ridge regression

→ overview → Recall that ordinary least squares (OLS) aims to find regression coefficients that minimize SSE

$$SSE = \sum (y_i - \hat{y}_i)^2, \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$$

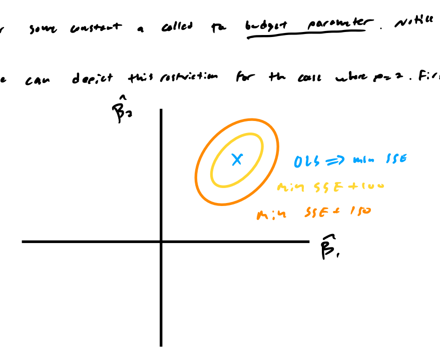
→ As p (ie flexibility) grows, OLS gets closer to having no bias. However, by the bias-variance tradeoff, this is at the expense of increasing the method's variance. One way to reduce variance (while allowing a bit more bias) is to restrict the possible values of the coefficient estimates. Specifically, we restrict by forcing the coefficient estimates to be closer to or shrunken towards 0.

→ The first shrinkage method is called ridge regression. It minimizes the same SSE expression w/ an added restriction of

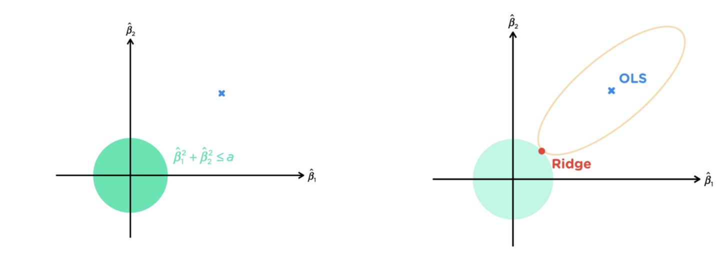
$$\sum_{j=1}^{p} \beta_j^2 \le a$$

for some constant a called the budget parameter. Notice that this restricts all the coefficient estimates except the intercept.

→ we can depict this restriction for the case where p = 2. First consider a plot of possible $\hat{\beta}_1$ & $\hat{\beta}_2$ values.



→ In the plot above, the cross marks the values $\hat{\beta}_1$ + $\hat{\beta}_2$ under OLS. Since OLS minimizes the SSE, any other point in this plot will result in a larger SSE relative to the cross. We illustrate SSE as a third dimension using contours. For each contour, every point along the contour has the same SSE value at those $\hat{\beta}_1$ + $\hat{\beta}_2$ coordinates. The further the contour is from the cross, the larger the SSE.



→ The plots above include the region $\hat{\beta}_1^2 + \hat{\beta}_2^2 \le a$ shaded in green. Ridge regression produces the coefficient estimates that minimize the SSE while restricted to the circle. Since the circle doesn't include the cross, the contour that first touches the edge of the circle satisfies the ridge regression criterion. Hence, this intersection is the closest to the cross (in terms of SSE) among all other points within the circle, & denotes the ridge estimates.

→ The plots demonstrate how the values of $\hat{\beta}_1$ + $\hat{\beta}_2$ would tend to shrink towards 0 for ridge regression. But it is possible that no shrinkage occurs, as when the green circle is large enough to include the OLS estimates.

→ In addition, the restriction can be incorporated directly into an optimization problem. Ridge regression seeks to minimize the expression:

$$\underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{SSE} + \underbrace{\lambda \sum_{j=1}^{p} \beta_j^2}_{penalty}, \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$$

where λ is the tuning parameter that controls the strength of the shrinkage. note that:

→ when λ = 0, there is no shrinkage penalty ⇒ $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ will equal the OLS estimates.

→ As λ approaches ∞, the entire expression is minimized when $\hat{\beta}_1, \ldots, \hat{\beta}_p = 0$.

→ This means λ is inversely related to flexibility. λ allows us to select a flexibility level between the inflexible null model (λ→∞) & the flexible MLR model (λ=0). The process of finding the best value of λ is called tuning, which can be achieved through cross-validation

→ Furthermore, some might find the notation $\sum_{j=1}^{p} \beta_j^2$ cumbersome. It can be simplified using the ℓ norm notation. First, let

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$$

→ Then the ℓ_2 norm of $\hat{\beta}$ is

$$\|\hat{\beta}\|_2 = \sqrt{\sum_{j=1}^{p} \hat{\beta}_j^2}$$

→ Consequently, the restriction inequality & the expression to be minimized for ridge regression can be simplified as follows:

$$\|\hat{\beta}\|_2^2 \le a$$

$$\sum (y_i - \hat{y}_i)^2 + \lambda \|\hat{\beta}\|_2^2$$

→ other important details regarding ridge regression include:

→ The x_j's should be scaled variables of the original. This puts $\hat{\beta}_1, \ldots, \hat{\beta}_p$ on the same scale, which is crucial b/c the shrinkage penalty puts equal weight on each one. This means the ridge estimates are not scale equivariant.

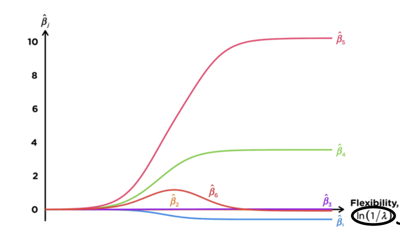
→ As λ increases, $\sum_{j=1}^{p} \beta_j^2$ must decrease (ie be shrunken towards 0) to minimize SSE + penalty. Yet, it is possible for an individual $\hat{\beta}_j$ to increase in absolute value (ie deviate away from 0) as λ increases.

→ Practically, none of the p ridge estimates will equal 0 as long as λ is finite. This means ridge regression does not drop variables from a model.

→ It is useful when dealing w/ high dimensions as only some of the predictors will have a meaningful estimated coefficient.

→ Example

→ 6 variables → After scaling the predictors, running the ridge procedure, & reversing the scale back to the original units, we see how the ridge coefficients change as a function of flexibility in the graph below.



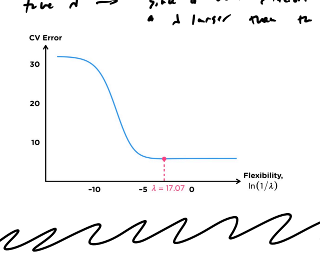
→ shrinkage appears from right to left (increasing penalty from right to left)

(λ→∞ ⇒ OLS estimate λ→0 ⇒ null model)

→ reciprocal b/c inversely related & ln() for plotting

→ note that a plot may use a penalty measure like λ in the horizontal axis instead of flexibility. Since flexibility + penalty are inversely related, both plots would look different despite having the same information

→ Can tune λ → since a lower flexibility (ie higher penalty) w/ no loss in model accuracy is preferred (ie parsimony), a λ larger than the λ_{CV} may be desired



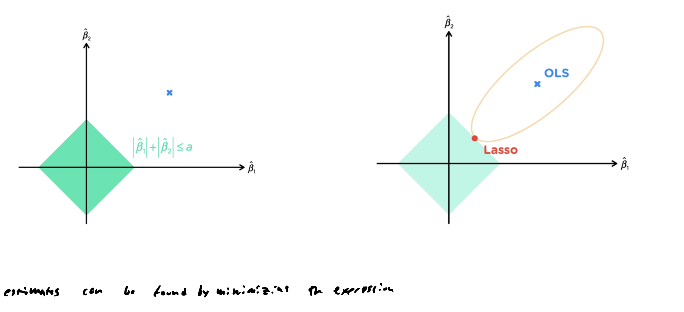
3.7.3 → Lasso regression

→ overview → Lasso regression is another shrinkage method; it bears a close resemblance to ridge regression. In minimizing the SSE, lasso imposes a restriction of

$$\sum_{j=1}^{p} |\beta_j| \le a$$

→ with budget parameter a. In short, where ridge takes the square of the $\hat{\beta}_j$'s, lasso takes their absolute value instead.

→ lets visualize the restriction for the case where p = 2 with a plot of $\hat{\beta}_1$ against $\hat{\beta}_2$



→ The lasso estimates can be found by minimizing the expression

$$\underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{SSE} + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}_{penalty}$$

where λ is the tuning parameter that controls the shrinkage strength & functions in the same way as in ridge regression

→ Moreover, the ℓ_1 norm of $\hat{\beta}$ is

$$\|\hat{\beta}\|_1 = \sum_{j=1}^{p} |\hat{\beta}_j|$$

→ thus, the restriction inequality & the expression to be minimized for lasso regression can be simplified as follows:

$$\|\hat{\beta}\|_1 \le a$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \|\hat{\beta}\|_1$$

→ other important details regarding lasso regression include:

→ x_j's should be scaled variables of the original for the same reason as mentioned in the previous section. Lasso estimates are not scale equivariant

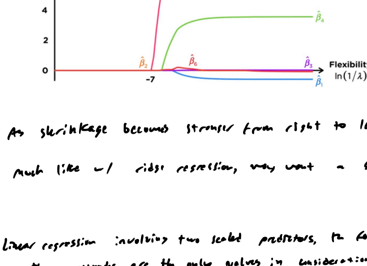
→ As λ increases, $\sum_{j=1}^{p} |\hat{\beta}_j|$ must decrease to minimize SSE plus penalty. Like ridge regression, it is possible for an individual $\hat{\beta}_j$ to increase in absolute value as λ increases.

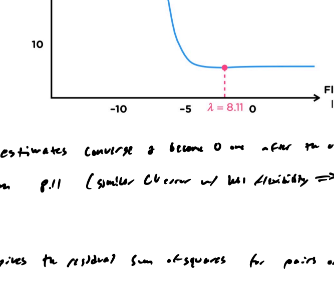
→ Unlike ridge regression, the lasso estimates can equal 0 for a large enough, finite λ. This means lasso regression drops variables from a model

→ It is useful when dealing w/ high dimensions

→ Examples

→ same one as ridge regression example





→ As shrinkage becomes stronger from right to left, lasso estimates converge & become 0 one after the other.

→ much like w/ ridge regression, may want a slightly less than 8.11 (similar CV error w/ less flexibility ⇒ more parsimony w/ little deterioration in performance

→ For linear regression involving two scaled predictors, the following table gives the residual sum of squares for pairs of β_1 & β_2 estimates, assuming these estimates are the only values in consideration

$\hat{\beta}_1$ \ $\hat{\beta}_2$	1	2	3
−1	99.4	96.0	79.3
0	98.4	97.8	82.1
1	105.5	114.7	90.8
2	~~164.2~~	~~130.3~~	~~186.4~~

→ with a budget parameter of 2, find $|\hat{\beta}_1 - \hat{\beta}_2|$ for lasso.

→ ① restriction: $|\hat{\beta}_1| + |\hat{\beta}_2| \le 2$

② minimize SSE + $\lambda \|\hat{\beta}\|_1$

$(\hat{\beta}_1, \hat{\beta}_2)$: (−1, 1) (0, 1) (1, 1) (0, 2)

SSE + $\lambda \|\hat{\beta}\|_1$: 99.4 + 2λ 98.4 + λ 105.5 + 2λ 97.8 + 2λ