1.

3. **(10%+10%+8%)** Suppose that there were 200 coupons for each of the discount percentages 5%, 10%, 15%, 20%, and 30% (i.e., the input values x_i, for $i = 1, \ldots, 1000$), and that the number of coupons redeemed for the above cases are 31, 52, 68, 101, and 144, respectively (i.e., the number of responses in each case that yields $y_i = 1$).

1. **(6%)** Solve Problem 3(c) of HW #1 using QDA.

$k=1 \Rightarrow y_i=1$
$k=2 \Rightarrow y_i=0$

$$\hat{\mu}_1 = \frac{1}{369}(0.05\times31 + 0.1\times52 + 0.15\times68 + 0.2\times101 + 0.3\times144) = 0.202$$

$$\hat{\mu}_2 = \frac{1}{604}(0.05\times169 + 0.1\times148 + 0.15\times132 + 0.2\times99 + 0.3\times56) = 0.132$$

$$\hat{\pi}_1 = \frac{396}{1000} \quad \hat{\pi}_2 = \frac{604}{1000}$$

$$\hat{\sigma}_1^2 = \frac{1}{396-1} \cdot \sum_{y_i=1}(x_i - 0.202)^2 = \frac{2.824}{395} = 0.0071$$

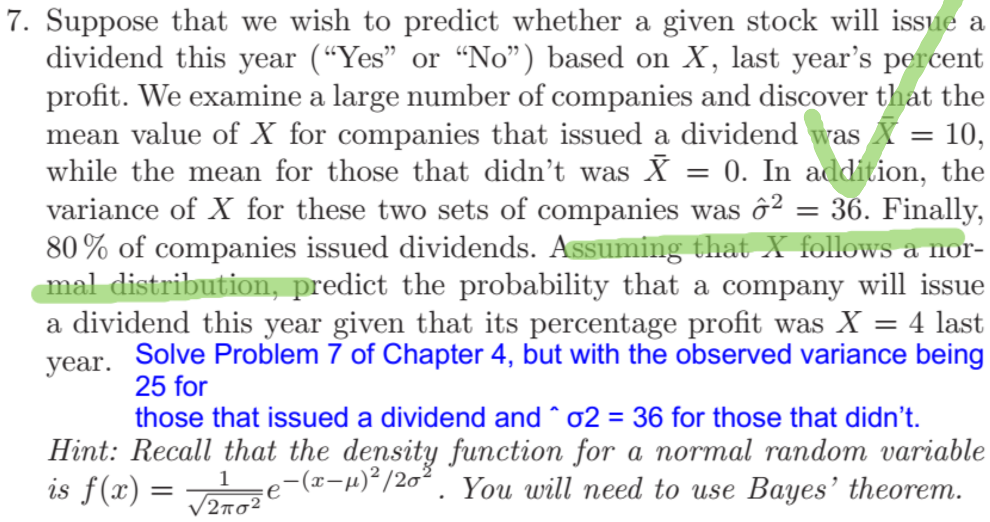
$$\hat{\sigma}_2^2 = \frac{1}{604-1} \sum_{y_i=0}(x_i - 0.132)^2 = \frac{3.368}{603} = 0.0055$$

$$\hat{\delta}_k(x) = \frac{-x^2}{2\sigma_k^2} + \frac{x\mu_k}{\sigma_k^2} - \frac{\mu_k^2}{2\sigma_k^2} + \log\pi_k.$$

$5\% = \delta_1(0.05) \Rightarrow 0.17$
$10\% : \delta_1(0.1) \Rightarrow 0.23$
$15\% = \delta_1(0.15) \Rightarrow 0.33$
$20\% = \delta_1(0.2) \Rightarrow 0.46$
$25\% = \delta_1(0.25) \Rightarrow 0.63$
$30\% = \delta_1(0.3) \Rightarrow 0.78$ #

← 各折扣所对应的兑换率.
(代入公式算得)

2.



7. Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X} = 10$, while the mean for those that didn't was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80 % of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.

Solve Problem 7 of Chapter 4, but with the observed variance being ˆ σ2 = 25 for those that issued a dividend and ˆ σ2 = 36 for those that didn't.

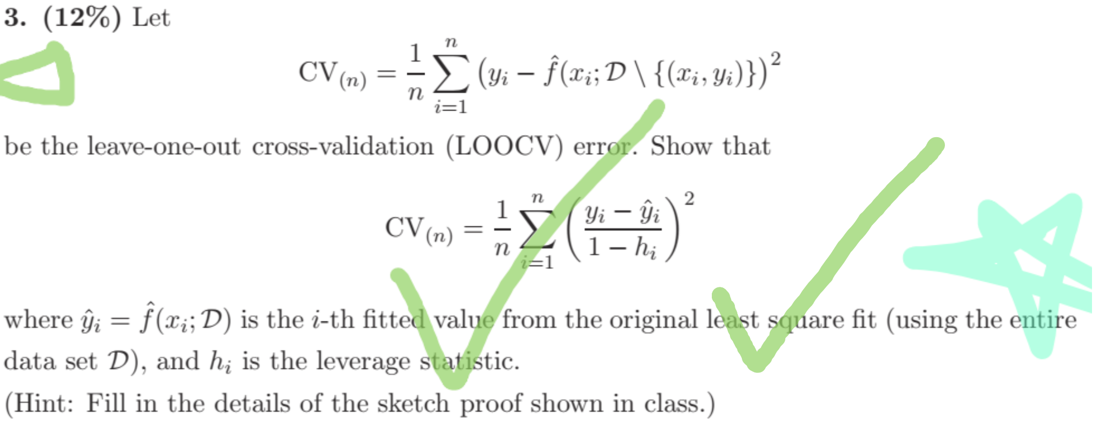
Hint: Recall that the density function for a normal random variable is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$. *You will need to use Bayes' theorem.*

2. $\hat{\sigma}^2 = 25$. $(\bar{X}_Y = 10., \ \bar{X}_N = 0)$

$$P_{Yes}(4) = \frac{0.8 \cdot \exp\left(-\frac{1}{2\times25} \cdot (4-10)^2\right) \cdot \frac{1}{5}}{\frac{1}{5} \cdot 0.8 \times \exp\left(-\frac{1}{2\times25} \cdot (4-10)^2\right) + (1-0.8) \cdot \exp\left(-\frac{1}{2\times36} \times (4-0)^2\right)\frac{1}{6}} = 0.74 \ \#$$

1/5 1/2*25

3.



3. (12%) Let

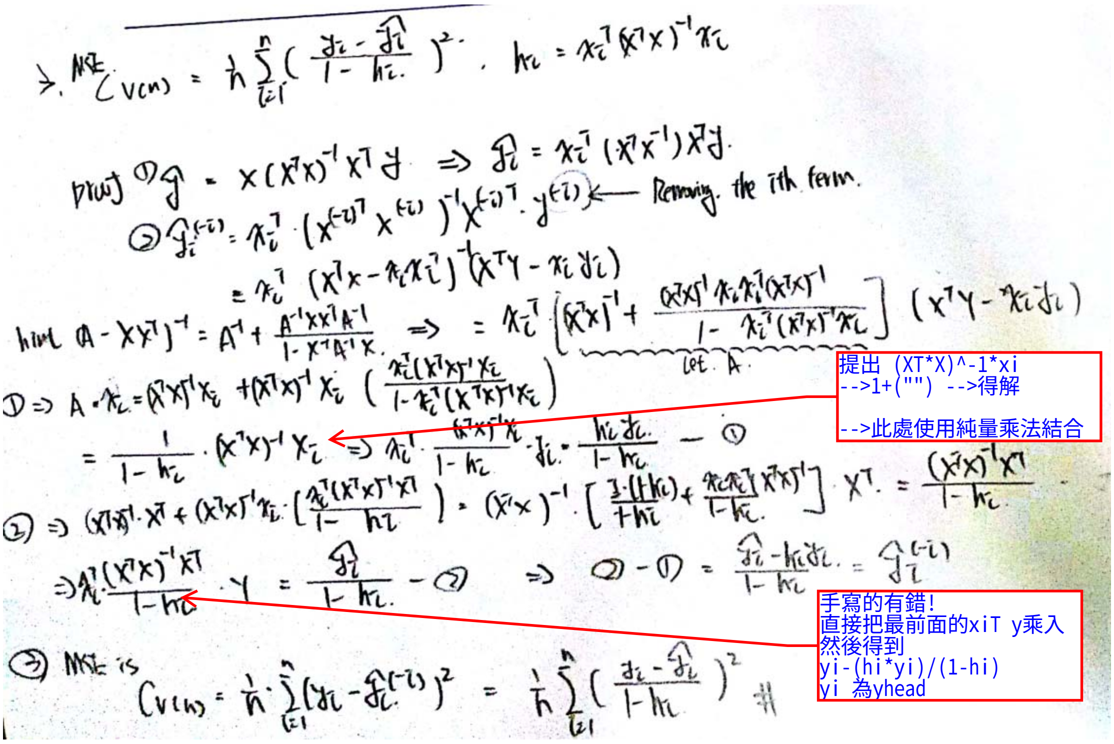
$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i; \mathcal{D}\setminus\{(x_i, y_i)\})\right)^2$$

be the leave-one-out cross-validation (LOOCV) error. Show that

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2$$

where $\hat{y}_i = \hat{f}(x_i;\mathcal{D})$ is the i-th fitted value from the original least square fit (using the entire data set $\mathcal{D}$), and h_i is the leverage statistic.

(Hint: Fill in the details of the sketch proof shown in class.)



3. NTS $CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2$. $h_i = x_i^T(X^TX)^{-1}x_i$

proof ① $\hat{y} = X(X^TX)^{-1}X^Ty \Rightarrow \hat{y}_i = x_i^T(X^TX^{-1})X^Ty.$

② $\hat{y}_i^{(-i)} = x_i^T\cdot(X^{(-i)T}X^{(-i)})^{-1}X^{(-i)T}\cdot y^{(-i)}$ ← Removing the ith term.

$= x_i^T(X^TX - x_ix_i^T)^{-1}(X^TY - x_iy_i)$

hint $(A - XX^T)^{-1} = A^{-1} + \frac{A^{-1}XX^TA^{-1}}{1-X^TA^{-1}X}$ $\Rightarrow$ $= x_i^T\underbrace{\left[(X^TX)^{-1} + \frac{(X^TX)^{-1}x_ix_i^T(X^TX)^{-1}}{1-x_i^T(X^TX)^{-1}x_i}\right]}_{\text{Let. } A}(X^TY - x_iy_i)$

① $\Rightarrow A\cdot x_i = (X^TX)^{-1}x_i + (X^TX)^{-1}x_i\left(\frac{x_i^T(X^TX)^{-1}x_i}{1-x_i^T(X^TX)^{-1}x_i}\right)$

$= \frac{1}{1-h_i}\cdot(X^TX)^{-1}x_i \Rightarrow x_i^T\cdot\frac{(X^TX)^{-1}x_i}{1-h_i}\cdot y_i = \frac{h_iy_i}{1-h_i}$ — ①

② $\Rightarrow (X^TX)^{-1}\cdot X^T + (X^TX)^{-1}x_i\cdot\left[\frac{x_i^T(X^TX)^{-1}X^T}{1-h_i}\right] = (X^TX)^{-1}\cdot\left[\frac{1\cdot(1-h_i)}{1-h_i} + \frac{x_ix_i^T(X^TX)^{-1}}{1-h_i}\right]\cdot X^T = \frac{(X^TX)^{-1}X^T}{1-h_i}$

$\Rightarrow x_i^T\frac{(X^TX)^{-1}X^T}{1-h_i}\cdot Y = \frac{\hat{y}_i}{1-h_i}$ — ② $\Rightarrow$ ② − ① $= \frac{\hat{y}_i - h_iy_i}{1-h_i} = \hat{y}_i^{(-i)}$

③ NTS is $CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^{(-i)})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2$ #

提出 (XT*X)^-1*xi
-->1+("") -->得解

-->此處使用純量乘法結合

手寫的有錯!
直接把最前面的xiT y乘入
然後得到
yi-(hi*yi)/(1-hi)
yi 為yhead

4.

4. **(10%+10%)** Suppose that the input and output variables of the n training data points can be expressed as $\mathbf{X} = (x_1, \ldots, x_n)^T$ and $\mathbf{y} = (y_1, \ldots, y_n)^T$, respectively. In ridge regression, the coefficient vector $\beta = (\beta_0, \ldots, \beta_p)^T$ is chosen such that $\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2$ is minimized for some $\lambda \geq 0$.

(a) Show that the resulting coefficient estimate is given by

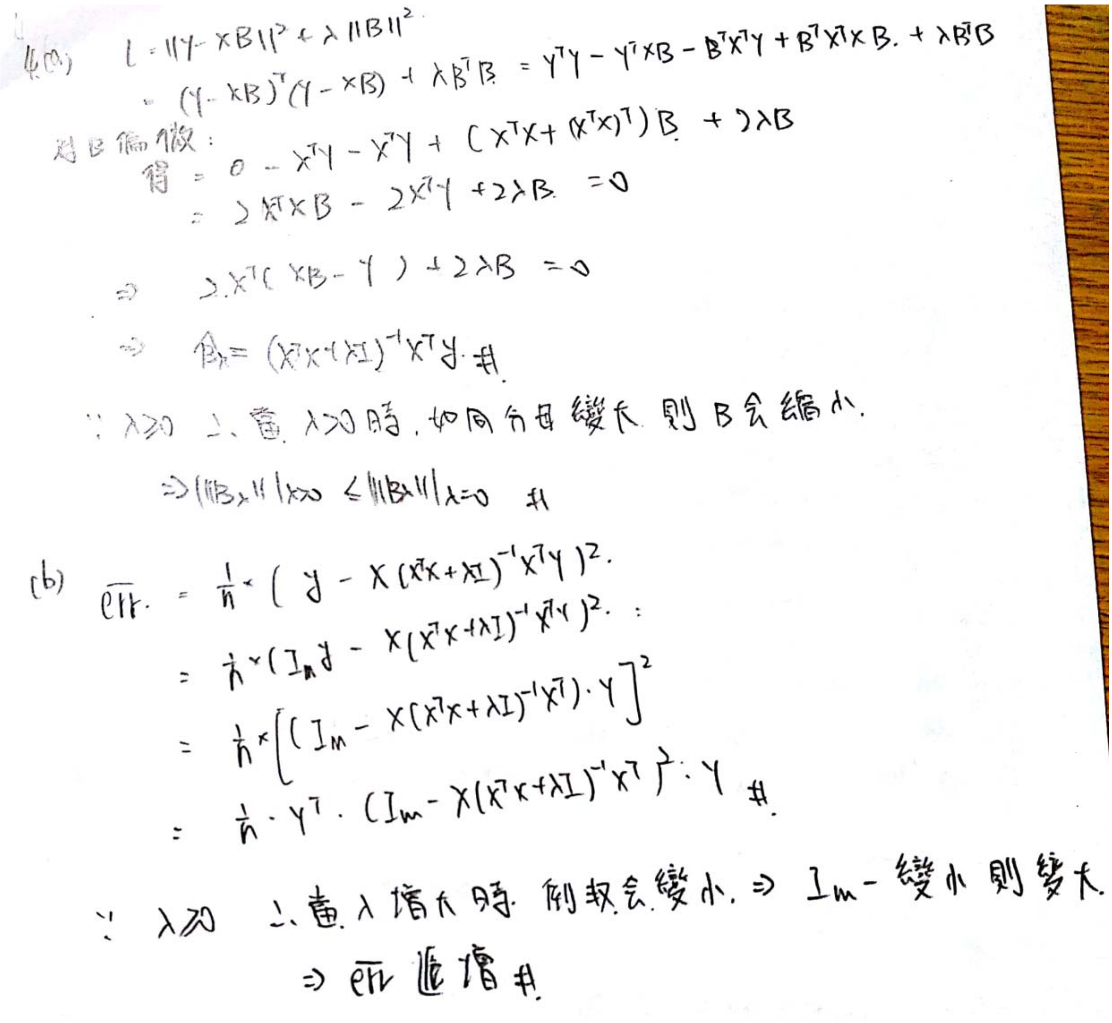
$$\hat{\beta}_\lambda = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

and that $\|\beta_\lambda\|\big|_{\lambda>0} \leq \|\beta_\lambda\|\big|_{\lambda=0}$

(b) Show that the training error is

$$\overline{\mathrm{err}} = \frac{1}{n}\mathbf{y}\left[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\right]^2\mathbf{y}$$

and that it is an increasing function of λ.



4(a) $L = \|Y - XB\|^2 + \lambda\|B\|^2$

$= (Y - XB)^T(Y - XB) + \lambda B^TB = Y^TY - Y^TXB - B^TX^TY + B^TX^TXB. + \lambda B^TB$

对B偏微：

得 $= 0 - X^TY - X^TY + (X^TX + (X^TX)^T)B + 2\lambda B$

$= 2X^TXB - 2X^TY + 2\lambda B = 0$

$\Rightarrow 2.X^T(XB - Y) + 2\lambda B = 0$

$\Rightarrow \hat{\beta}_\lambda = (X^TX + \lambda I)^{-1}X^Ty.$ #

∵ $\lambda \geq 0$ ∴ 當 $\lambda > 0$ 時，如同分母變大 則 B 会縮小.

$\Rightarrow \|\beta_\lambda\|\big|_{\lambda>0} \leq \|\beta_\lambda\|\big|_{\lambda=0}$ #

(b) $\overline{err} = \frac{1}{n} \times (y - X(X^TX + \lambda I)^{-1}X^TY)^2.$

$= \frac{1}{n} \times (I_n y - X(X^TX + \lambda I)^{-1}X^TY)^2.$

$= \frac{1}{n} \times \left[(I_m - X(X^TX + \lambda I)^{-1}X^T) \cdot Y\right]^2$

$= \frac{1}{n} \cdot Y^T \cdot (I_m - X(X^TX + \lambda I)^{-1}X^T)^2 \cdot Y$ #

∵ $\lambda \geq 0$ ∴ 當 λ 增大時 則扣会變小. $\Rightarrow$ $I_m -$ 變小 則變大.

$\Rightarrow$ $\overline{err}$ 遞增 #

5.

5. (10%+8%) Suppose that the available data set is $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), (x_3, y_3)\} = \{(1, 2), (3, 7), (5, 8)\}$, and that B bootstrap data sets $\mathcal{D}^{*1}, \dots, \mathcal{D}^{*B}$ (excluding those with only 1 distinct data point) are generated from $\mathcal{D}$, each with the same size as $\mathcal{D}$. Linear regression is performed on each data set to obtain coefficient estimates $\beta^{*1}, \dots, \beta^{*B}$, where

(a) For $B \to \infty$, find the standard errors of the coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

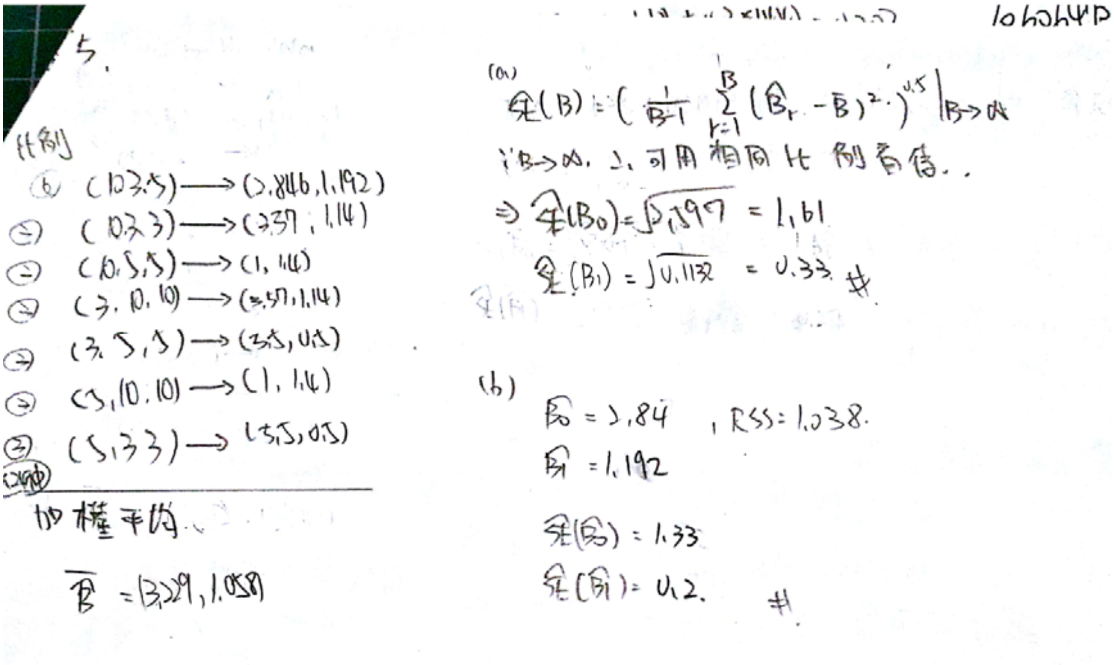
(b) Compare with the standard error estimates derived in Chapter 3 for original least squares (i.e., ridge regression with $\lambda = 0$), that is,

$$\widehat{\mathrm{SE}}(\hat{\beta}_0)^2 = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$$

and

$$\widehat{\mathrm{SE}}(\hat{\beta}_1)^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

where $n = 3$ in this case.



5.

比例

⑥ $(1,3,5) \longrightarrow (2.846, 1.192)$

③ $(1,3,3) \longrightarrow (2.37, 1.14)$

③ $(1,5,5) \longrightarrow (1, 1.4)$

③ $(3,1,1) \longrightarrow (2.57, 1.14)$

③ $(3,5,5) \longrightarrow (3.5, 0.5)$

③ $(5,1,1) \longrightarrow (1, 1.4)$

③ $(5,3,3) \longrightarrow (3.5, 0.5)$

(24)

加權平均

$\bar{B} = (3.229, 1.058)$

(a)

$\widehat{SE}(B) = \left(\frac{1}{B-1} \sum_{r=1}^{B} (\hat{B}_r - \bar{B})^2 \right)^{0.5} \Big|_{B \to \infty}$

∵$B \to \infty$, ∴ 可用相同比例看待..

$\Rightarrow \widehat{SE}(B_0) = \sqrt{2.597} = 1.61$

$\widehat{SE}(B_1) = \sqrt{0.1132} = 0.33$ #

(b)

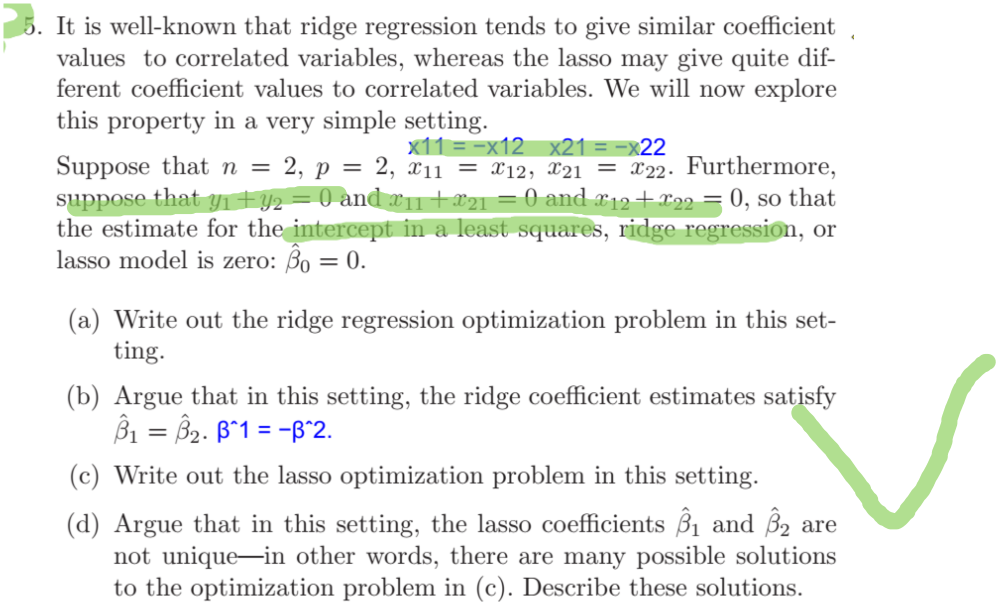
$\hat{B}_0 = 2.84$, RSS: 1.038.

$\hat{B}_1 = 1.192$

$\widehat{SE}(\hat{B}_0) = 1.33$

$\widehat{SE}(\hat{B}_1) = 0.2.$ #

6.



5. It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas the lasso may give quite different coefficient values to correlated variables. We will now explore this property in a very simple setting.

x11 = −x12   x21 = −x22

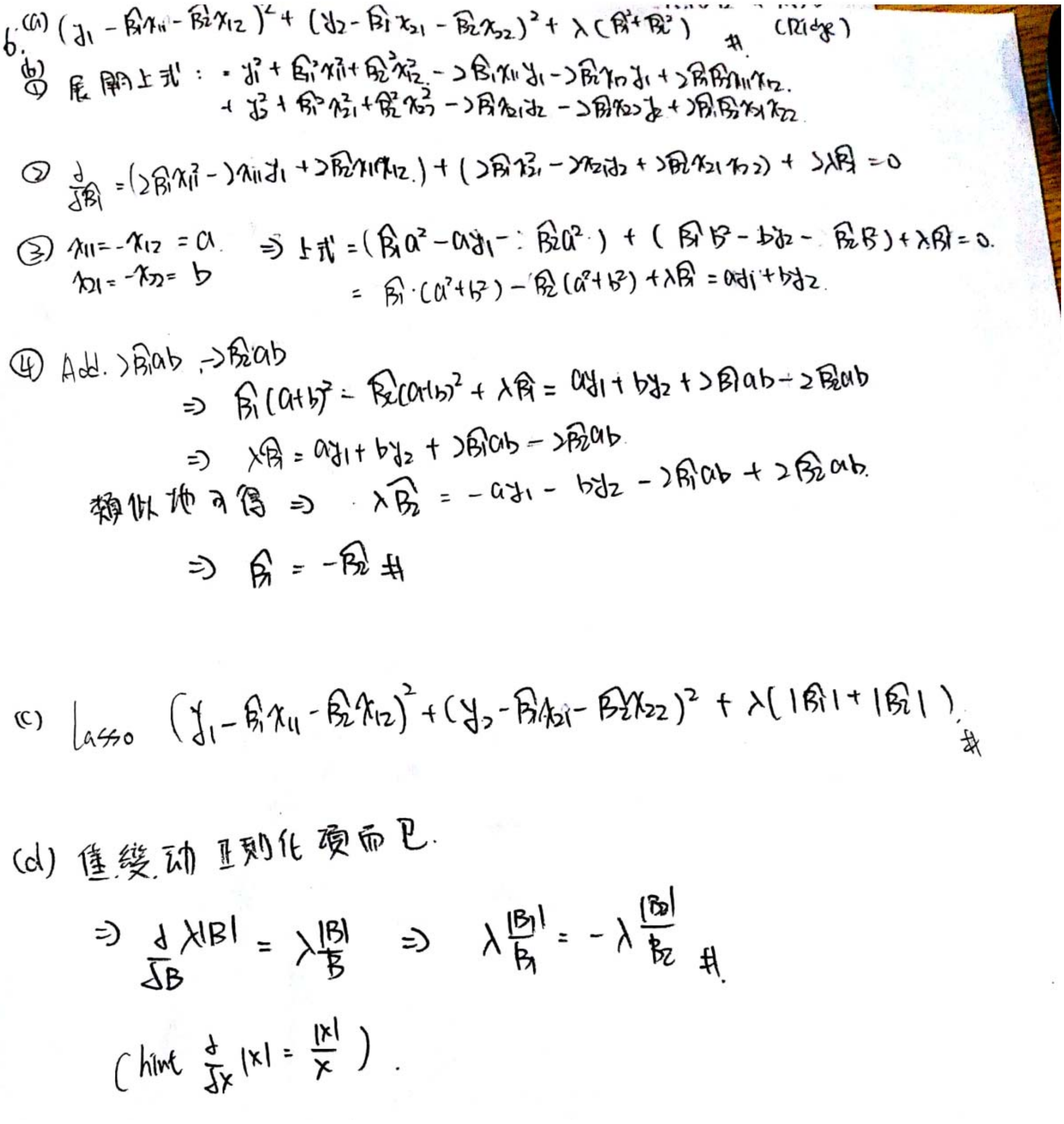
Suppose that $n = 2$, $p = 2$, $x_{11} = x_{12}$, $x_{21} = x_{22}$. Furthermore, suppose that $y_1 + y_2 = 0$ and $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$, so that the estimate for the intercept in a least squares, ridge regression, or lasso model is zero: $\hat{\beta}_0 = 0$.

(a) Write out the ridge regression optimization problem in this setting.

(b) Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$. β^1 = −β^2.

(c) Write out the lasso optimization problem in this setting.

(d) Argue that in this setting, the lasso coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are not unique—in other words, there are many possible solutions to the optimization problem in (c). Describe these solutions.



6. (a) $(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)$ # (Ridge)

(b) ① 展開上式：$= y_1^2 + \hat{\beta}_1^2 x_{11}^2 + \hat{\beta}_2^2 x_{12}^2 - 2\hat{\beta}_1 x_{11} y_1 - 2\hat{\beta}_2 x_{12} y_1 + 2\hat{\beta}_1\hat{\beta}_2 x_{11} x_{12}$
$+ y_2^2 + \hat{\beta}_1^2 x_{21}^2 + \hat{\beta}_2^2 x_{22}^2 - 2\hat{\beta}_1 x_{21} y_2 - 2\hat{\beta}_2 x_{22} y_2 + 2\hat{\beta}_1\hat{\beta}_2 x_{21} x_{22}$

② $\frac{d}{d\hat{\beta}_1} = (2\hat{\beta}_1 x_{11}^2 - 2x_{11}y_1 + 2\hat{\beta}_2 x_{11}x_{12}) + (2\hat{\beta}_1 x_{21}^2 - 2x_{21}y_2 + 2\hat{\beta}_2 x_{21}x_{22}) + 2\lambda\hat{\beta}_1 = 0$

③ $x_{11} = -x_{12} = a$ $\Rightarrow$ 上式 $= (\hat{\beta}_1 a^2 - ay_1 - \hat{\beta}_2 a^2) + (\hat{\beta}_1 b^2 - by_2 - \hat{\beta}_2 b^2) + \lambda\hat{\beta}_1 = 0.$
$x_{21} = -x_{22} = b$ $\quad = \hat{\beta}_1 \cdot (a^2 + b^2) - \hat{\beta}_2(a^2 + b^2) + \lambda\hat{\beta}_1 = ay_1 + by_2.$

④ Add. $2\hat{\beta}_1 ab$, $-2\hat{\beta}_2 ab$
$\Rightarrow \hat{\beta}_1(a+b)^2 - \hat{\beta}_2(a+b)^2 + \lambda\hat{\beta}_1 = ay_1 + by_2 + 2\hat{\beta}_1 ab - 2\hat{\beta}_2 ab$
$\Rightarrow \lambda\hat{\beta}_1 = ay_1 + by_2 + 2\hat{\beta}_1 ab - 2\hat{\beta}_2 ab.$
類似地可得 $\Rightarrow \lambda\hat{\beta}_2 = -ay_1 - by_2 - 2\hat{\beta}_1 ab + 2\hat{\beta}_2 ab.$
$\Rightarrow \hat{\beta}_1 = -\hat{\beta}_2$ #

(c) Lasso $(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda(|\hat{\beta}_1| + |\hat{\beta}_2|).$ #

(d) 僅變動正則化項而已.

$\Rightarrow \frac{d}{dB}\lambda|B| = \lambda\frac{|B|}{B} \Rightarrow \lambda\frac{|B_1|}{B_1} = -\lambda\frac{|B_2|}{B_2}$ #

(hint $\frac{d}{dx}|x| = \frac{|x|}{x}$).