Statistical Learning

Homework 3

Part 1: Conceptual Questions

Problem 1. Consider a regression of Y_i against a constant and X_i. Let $\hat{\beta}_0$, $\hat{\beta}_1$, and s^2 denote the estimated intercept, estimated slope parameter, and estimator of the variance of errors from that regression. Let T denote the t-statistic for testing H_0 that the slope parameter is zero in that regression. Let *pval* be the corresponding p-value. Now, let c_1 and c_2 be two constants ($c_2 \neq 0$). Define a new dependent variable and a new regressor as

$$Y_i^* = c_1 Y_i,$$
$$X_i^* = c_2 X_i.$$

Let $\hat{\beta}_0^*$, $\hat{\beta}_1^*$, and s_*^2 denote the estimated intercept, estimated slope parameter, and estimator of the variance of errors from the regression of Y_i^* against a constant and X_i^*. Let T^* denote the t-statistic for testing H_0 that the slope parameter in the regression of Y_i^* against a constant and X_i^* is zero. Let $pval^*$ be the corresponding p-value.

1. Find an expression for $\hat{\beta}_1^*$ in terms of $\hat{\beta}_1, c_1$, and c_2.
2. Find an expression for $\hat{\beta}_0^*$ in terms of $\hat{\beta}_0$ and c_1.
3. Find an expression for s_*^2 in terms of s^2 and c_1.
4. What is the relationship between T and T^*?
5. What is the relationship between *pval* and $pval^*$?

Solution.

(a) $\hat{\beta}_1^* = \frac{\sum_i (X_i^* - \bar{X}^*) Y_i^*}{\sum_i (X_i^* - \bar{X}^*)^2} = \frac{\sum_i (c_2 X_i - c_2 \bar{X}) c_1 Y_i}{\sum_i (c_2 X_i - c_2 \bar{X})^2} = \frac{c_1 c_2 \sum_i (X_i - \bar{X}) Y_i}{c_2^2 \sum_i (X - \bar{X})^2} = \frac{c_1}{c_2} \hat{\beta}_1.$

(b) $\hat{\beta}_0^* = \bar{Y}^* - \hat{\beta}_1^* \bar{X}^* = c_1 \bar{Y} - \frac{c_1}{c_2} \hat{\beta}_1 c_2 \bar{X} = c_1 \bar{Y} - c_1 \hat{\beta}_1 \bar{X} = c_1 \hat{\beta}_0.$

(c) First, $\hat{U}_i^* = Y_i^* - \hat{\beta}_0^* - \hat{\beta}_1^* X_i^* = c_1 Y_i - c_1 \hat{\beta}_0 - \frac{c_1}{c_2} \hat{\beta}_1 c_2 X_i = c_1 Y_i - c_1 \hat{\beta}_0 - c_1 \hat{\beta}_1 X_i = c_1 \hat{U}_i.$
Next, $s_*^2 = \frac{1}{n-2} \sum_i \left(\hat{U}_i^* \right)^2 = \frac{1}{n-2} \sum_i \left(c_1 \hat{U}_i \right)^2 = c_1^2 s^2.$

(d) For $H_0 : \beta_1^* = 0$, we have

$$
\begin{aligned}
T^* &= \hat{\beta}_1^* / \sqrt{s_*^2 / \sum_i (X_i^* - \bar{X}^*)^2} \\
&= \frac{c_1}{c_2}\hat{\beta}_1 / \sqrt{c_1^2 s^2 / \sum_i (c_2 X_i - c_2 \bar{X})^2} \\
&= \frac{c_1}{c_2}\hat{\beta}_1 / \sqrt{(c_1/c_2)^2 s^2 / \sum_i (X_i - \bar{X})^2} \\
&= \hat{\beta}_1 / \sqrt{s^2 / \sum_i (X_i - \bar{X})^2} \\
&= T.
\end{aligned}
$$

Note that T is the test statistic for testing $H_0 : \beta_1 = 0$.

(e) Since $T = T^*$ and df's are the same in both cases, $pval = pval*$. Thus, rescaling the dependent variable and regressor has no effect on testing for significance of the slope parameter.

Problem 2. ISL (2nd edition) Page 219, Question 1.

Solution. Compute

$$
\begin{aligned}
\text{Var}\left[\alpha X + (1-\alpha) Y\right] &= \text{Var}\left[\alpha X\right] + \text{Var}\left[(1-\alpha) Y\right] + 2\text{Cov}\left[\alpha X, (1-\alpha) Y\right] \\
&= \alpha^2 \text{Var}\left[X\right] + (1-\alpha)\text{Var}\left[Y\right] + 2\alpha(1-\alpha)\text{Cov}\left[X, Y\right] \\
&= \sigma_X^2 \alpha^2 + \sigma_Y^2 (1-\alpha)^2 + 2\sigma_{XY}\left(-\alpha^2 + \alpha\right).
\end{aligned}
$$

Take derivative:

$$
\frac{d}{d\alpha}\text{Var}\left[\alpha X + (1-\alpha) Y\right] = 2\alpha\sigma_X^2 + 2\sigma_Y^2 (1-\alpha)(-1) + 2\sigma_{XY}(-2\alpha + 1).
$$

The solution to

$$
0 = \frac{d}{d\alpha}\text{Var}\left[\alpha X + (1-\alpha) Y\right]
$$

is

$$
\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}.
$$

Problem 3. ISL (2nd edition) Page 284, Question 5.

Solution. (a) According to this setting ($x_{11} = x_{12} = x_1$ and $x_{21} = x_{22} = x_2$), the ridge regression seeks to minimize

$$
(y_1 - b_1 x_1 - b_2 x_1)^2 + (y_2 - b_1 x_2 - b_2 x_2)^2 + \lambda\left(b_1^2 + b_2^2\right).
$$

(b) By taking the derivative with respect to (b_1, b_2):

$$
b_1\left(x_1^2 + x_2^2 + \lambda\right) + b_2\left(x_1^2 + x_2^2\right) = y_1 x_1 + y_2 x_2
$$

and

$$b_1 \left(x_1^2 + x_2^2\right) + b_2 \left(x_1^2 + x_2^2 + \lambda\right) = y_1 x_1 + y_2 x_2.$$

The solution $\left(\hat{\beta}_1, \hat{\beta}_2\right)$ to the above equations satisfy $\hat{\beta}_1 = \hat{\beta}_2$.

(c) The LASSO optimization problem seeks to minimize

$$(y_1 - b_1 x_1 - b_2 x_1)^2 + (y_2 - b_1 x_2 - b_2 x_2)^2 + \lambda \left(|b_1| + |b_2|\right).$$

(d) Use the alternate form of the LASSO optimization problem: minimize

$$(y_1 - b_1 x_1 - b_2 x_1)^2 + (y_2 - b_1 x_2 - b_2 x_2)^2 \text{ subject to } |b_1| + |b_2| \leq s.$$

Substitute $x_1 + x_2 = 0$ and $y_1 + y_2 = 0$ into the objective function to get

$$2 \left(y_1 - (b_1 + b_2) x_1\right)^2 \geq 0.$$

The unconstrained solution $\left(\hat{\beta}_1, \hat{\beta}_2\right)$ must satisfy $\hat{\beta}_1 + \hat{\beta}_2 = y_1/x_1$. The constrained solution of

$$\min_{b_1, b_2} 2 \left(y_1 - (b_1 + b_2) x_1\right)^2 \text{ subject to } |b_1| + |b_2| \leq s$$

must be on the edges of the diamond of the constraints. The set of solutions must be either of the two entire edges:

$$\{(b_1, b_2) : b_1 \geq 0, b_2 \geq 0, b_1 + b_2 = s\} \tag{1}$$

and

$$\{(b_1, b_2) : b_1 \leq 0, b_2 \leq 0, b_1 + b_2 = -s\}. \tag{2}$$

Finding the solutions boils down to comparing $(y_1 - s \cdot x_1)^2$ and $(y_1 + s \cdot x_1)^2$. In case of $(y_1 - s \cdot x_1)^2 \geq (y_1 + s \cdot x_1)^2$, (2) is the set of solutions. In case of $(y_1 - s \cdot x_1)^2 \leq (y_1 + s \cdot x_1)^2$, (1) is the set of solutions. The constrained minimizer cannot occur at the interior of the other two edges

$$\{(b_1, b_2) : b_1 \geq 0, b_2 \leq 0, b_1 - b_2 = s\}$$

and

$$\{(b_1, b_2) : b_1 \leq 0, b_2 \geq 0, -b_1 + b_2 = s\}.$$

Suppose that $b_1 \geq 0, b_2 \leq 0, b_1 - b_2 = s$. Then, substitute $b_1 - b_2 = s$ into $\left(y_1 - (b_1 + b_2) x_1\right)^2$ to get $\left(y_1 - (s + 2b_2) x_1\right)^2$. Now choose $b_2 \in [-s, 0]$ to minimize it. It is clear that the minimizer must be on the boundary, since the objective $\left(y_1 - (s + 2b_2) x_1\right)^2$ is monotone in b_2.

Problem 4. ISL (2nd edition) Page 285, Question 7. Read “Bayesian Interpretation for Ridge Regression and the Lasso” on Page 248.

Solution.

(a) The likelihood:

$$f(Y \mid X, \beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2}{2\sigma^2} \right)$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right).$$

(b) The posterior distribution:

$$p(\beta \mid X, Y) \propto f(Y \mid X, \beta)\, p(\beta)$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right) \left[\frac{1}{2b} \exp\left(-\frac{|\beta|}{b} \right) \right],$$

where $|\beta| = \sum_{j=1}^{p} |\beta_j|$.
(c) Rearrange:

$$f(Y \mid X, \beta)\, p(\beta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \left(\frac{1}{2b}\right) \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 - \frac{|\beta|}{b} \right).$$

Take log:

$$\log\left(f(Y \mid X, \beta)\, p(\beta)\right)$$

$$= \log\left(\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \left(\frac{1}{2b}\right) \right) - \left(\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \frac{|\beta|}{b} \right).$$

The posterior mode is

$$\underset{\beta}{\operatorname{argmax}} \log\left(f(Y \mid X, \beta)\, p(\beta)\right) = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \frac{|\beta|}{b} \right)$$

$$= \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \frac{2\sigma^2 |\beta|}{b} \right)$$

$$= \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right),$$

where $\lambda = 2\sigma^2/b$. The posterior mode is equal to the LASSO estimator with penalty $\lambda = 2\sigma^2/b$.

(c) The posterior distribution:

$$
\begin{aligned}
p(\beta \mid X, Y) &\propto f(Y \mid X, \beta)\, p(\beta) \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right)\left(\frac{1}{\sqrt{2\pi c}}\right)^p \exp\left(-\frac{1}{2c}\sum_{j=1}^{p}\beta_j^2\right) \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \left(\frac{1}{\sqrt{2\pi c}}\right)^p \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 - \frac{1}{2c}\sum_{j=1}^{p}\beta_j^2\right).
\end{aligned}
$$

(d) The posterior mode is

$$
\begin{aligned}
\underset{\beta}{\operatorname{argmax}} \log\left(f(Y \mid X, \beta)\, p(\beta)\right) &= \underset{\beta}{\operatorname{argmin}} \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \frac{1}{2c}\sum_{j=1}^{p}\beta_j^2 \\
&= \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2,
\end{aligned}
$$

where $\lambda = \sigma^2/c$. The posterior mode is equal to the ridge estimator with penalty $\lambda = \sigma^2/b$. The posterior distribution is normal. Therefore, the mode is equal to the mean.

Problem 5. Another resampling method is called jackknife, which is similar to LOOCV. Suppose that $\hat{\theta} = \varphi_n(Z_1, Z_2, ..., Z_n)$ is the estimator of an parameter θ. Denote $\hat{\theta}_{-j} = \varphi_{n-1}(Z_1, ..., Z_{j-1}, Z_{j+1}, ..., Z_n)$. $\hat{\theta}_{-j}$ is an estimator obtained by removing the j-th observation from the entire sample. The variation in $\left\{\hat{\theta}_{-j} : j = 1, ..., n\right\}$ should be informative about the population variance of $\hat{\theta}_n$. Denote $\overline{\hat{\theta}} = n^{-1}\sum_{j=1}^{n}\hat{\theta}_{-j}$. The Jackknife standard error is

$$
\widehat{se}_{jk} = \sqrt{\frac{n-1}{n}\sum_{j=1}^{n}\left(\hat{\theta}_{-j} - \overline{\hat{\theta}}\right)^2}.
$$

An approximate 95% confidence interval is $\left[\hat{\theta}_n - 2\cdot\widehat{se}_{jk}, \hat{\theta}_n + 2\cdot\widehat{se}_{jk}\right]$. Consider the following simple example: for i.i.d. random variables $X_1, X_2, ..., X_n$, where $X_i \sim \mathrm{N}(\theta, \sigma^2)$, $\hat{\theta}_n = n^{-1}\sum_{i=1}^{n} X_i$ is an estimator of θ. Argue that when n is large, $\Pr\left[\hat{\theta}_n - 2\cdot\widehat{se}_{jk} \le \theta \le \hat{\theta}_n + 2\cdot\widehat{se}_{jk}\right]$ is approximately 95% by showing that $(n-1)\sum_{j=1}^{n}\left(\hat{\theta}_{-j} - \overline{\hat{\theta}}\right)^2$ is equal to the sample variance.

Solution. Easy to compute

$$
\begin{aligned}
\hat{\theta}_{-j} &= \frac{1}{n-1}\left(n\overline{X} - X_j\right) \\
\frac{1}{n}\sum_{j=1}^{n}\hat{\theta}_{-j} &= \frac{1}{n(n-1)}\sum_{j=1}^{n}\left(n\overline{X} - X_j\right) = \overline{X}.
\end{aligned}
$$

For this simple case,

$$\hat{\theta}_{-j} - \overline{\hat{\theta}} = \frac{1}{n-1}\left(n\overline{X} - X_j\right) - \overline{X} = \frac{1}{n-1}\left(\overline{X} - X_j\right).$$

We have

$$(n-1)\sum_{j=1}^{n}\left(\hat{\theta}_{-j} - \overline{\hat{\theta}}\right)^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2,$$

which is the sample variance that is a consistent and unbiased estimator for σ^2. Therefore,

$$\widehat{se}_{jk}^2 = \frac{1}{n}\cdot\left(\frac{1}{n-1}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2\right)$$

and

$$\frac{\hat{\theta}_n - \theta}{\widehat{se}_{jk}} \sim t_{n-1}$$

and it is approximately normally distributed when n is large.

Part 2: Applied Questions

Problem 6. ISL (2nd edition) Page 220, Question 5.

Problem 7. ISL (2nd edition) Page 221, Question 6.

Problem 8. ISL (2nd edition) Page 285, Question 8.

Problem 9. ISL (2nd edition) Page 286, Question 9 (a,b,c,d).