

# STAT151A Homework 6: Due April 19th

Your name here

## 1 Fit and regressors

Given a regression on  $\mathbf{X}$  with  $P$  regressors, and the corresponding  $\mathbf{Y}$ ,  $\hat{\mathbf{Y}}$ , and  $\hat{\varepsilon}$ , define the following quantities:

$$RSS := \hat{\varepsilon}^\top \hat{\varepsilon} \quad (\text{Residual sum of squares})$$

$$TSS := \mathbf{Y}^\top \mathbf{Y} \quad (\text{Total sum of squares})$$

$$ESS := \hat{\mathbf{Y}}^\top \hat{\mathbf{Y}} \quad (\text{Explained sum of squares})$$

$$R^2 := \frac{ESS}{TSS}.$$

**a**

1. Prove that  $RSS + ESS = TSS$ .
2. Express  $R^2$  in terms of  $TSS$  and  $RSS$ .
3. What is  $R^2$  when we include no regressors? ( $P = 0$ )
4. What is  $R^2$  when we include  $N$  linearly independent regressors? ( $P = N$ )
5. Can  $R^2$  ever decrease when we add a regressor? If so, how?
6. Can  $R^2$  ever stay the same when we add a regressor? If so, how?
7. Can  $R^2$  ever increase when we add a regressor? If so, how?
8. Does a high  $R^2$  mean the regression is correctly specified? Why or why not?
9. Does a low  $R^2$  mean the regression is incorrectly specified? Why or why not?

**Solutions:**

1. This follows from  $\hat{\mathbf{Y}}^\top \hat{\varepsilon} = \mathbf{0}$ .
2.  $R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$
3.  $R^2 = 0$  when we include no regressors
4.  $R^2 = 1$  when we include  $N$  linearly independent regressors?

5. No, it cannot, since you project onto the same or larger subspace.
6. Yes, if you add a regressor column that is colinear with the existing columns.
7. Yes, if you add a linearly independent regressor column.
8. No, you might overfit.
9. No, you might have low signal to noise ratio.

**b**

The next questions will be about the F-test statistic for the null  $H_0 : \beta = \mathbf{0}$ ,

$$\phi = \hat{\beta}^\top (\mathbf{X}^\top \mathbf{X}) \hat{\beta} / (P \hat{\sigma}^2)$$

1. Write the F-test statistic  $\phi$  in terms of  $TSS$  and  $RSS$ , and  $P$ .
2. Can  $\phi$  ever decrease when we add a regressor? If so, how?
3. Can  $\phi$  ever stay the same when we add a regressor? If so, how?
4. Can  $\phi$  ever increase when we add a regressor? If so, how?

**Solutions:**

1.  $\phi = \frac{\hat{\mathbf{Y}}^\top \hat{\mathbf{Y}}}{P \text{RSS}/(N-P)} = \frac{N-P}{P} \frac{TSS}{RSS}$
2. Yes,  $(N-P)/P$  decreases with  $P$ , and  $TSS$  and  $RSS$  can stay the same if the added regressor is colinear.
3. Yes,  $TSS/RSS$  might increase just as much as  $(N-P)/P$  decreases.
4. Yes,  $TSS/RSS$  might increase more than  $(N-P)/P$  decreases.

## 2 Omitted variable bias

For this problem, let  $(\mathbf{x}_n, \mathbf{z}_n, y_n)$  be IID random variables, where  $\mathbf{x}_n \in \mathbb{R}^{P_X}$  and  $\mathbf{z}_n \in \mathbb{R}^{P_Z}$ . Suppose that  $\mathbf{x}_n$  and  $\mathbf{z}_n$  satisfy  $\mathbb{E}[\mathbf{x}_n \mathbf{z}_n^\top] = \mathbf{0}$ .

Let  $y_n = \mathbf{x}_n^\top \beta + \mathbf{z}_n^\top \gamma + \varepsilon_n$ , where  $\varepsilon_n$  is mean zero, unit variance, and independent of  $\mathbf{x}_n$  and  $\mathbf{z}_n$ .

**a**

Take  $P_X = P_Z = 1$  (i.e. scalar regressors). Show that there exists  $x_n$  and  $z_n$  such that  $\mathbb{E}[x_n z_n] = 0$  but  $\mathbb{E}[z_n | x_n] \neq 0$  for some  $x_n$ . (A single counterexample will be enough.)

**Solution:**

An example is  $x_n \sim \mathcal{N}(0, 1)$ , and  $z_n = x_n^2$ . Then  $\mathbb{E}[x_n z_n^2] = \mathbb{E}[x_n^3] = 0$ .

A general construction is to start with a generic  $\mathbf{z}_n$  and define

$$\tilde{\mathbf{z}}_n = \mathbf{z}_n - \mathbb{E}[\mathbf{z}_n \mathbf{x}_n^\top] (\mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top])^{-1} \mathbf{x}_n.$$

Then

$$\mathbb{E}[\tilde{\mathbf{z}}_n \mathbf{x}_n^\top] = \mathbb{E}[\mathbf{z}_n \mathbf{x}_n^\top] - \mathbb{E}[\mathbf{z}_n \mathbf{x}_n^\top] (\mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top])^{-1} \mathbb{E}[\mathbf{x}_n \mathbf{x}_n^\top] = \mathbf{0},$$

but

$$\mathbb{E}[\tilde{\mathbf{z}}_n | \mathbf{x}_n^\top] \neq \mathbf{0}$$

if  $\mathbb{E}[\mathbf{z}_n | \mathbf{x}_n^\top]$  is not linear in  $\mathbf{x}_n$ . So if we start with  $\mathbf{z}_n = f(\mathbf{x}_n)$  for any nonlinear function  $f(\cdot)$ , we get a valid example.

**b**

Now return to the general case. Let  $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$  denote the OLS estimator from the regression on  $\mathbf{X}$  alone.

For simplicity, assume that  $\frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \mathbf{z}_n^\top = \mathbf{0}$ . (Note that, by the LLN,  $\frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \mathbf{z}_n^\top \rightarrow \mathbf{0}$  as  $N \rightarrow \infty$ , so this is a reasonable approximate assumption.)

Derive an expression for  $\mathbb{E}[\hat{\beta}]$ , where the expectation is taken over  $\mathbf{X}$ ,  $\mathbf{Y}$ , and  $\mathbf{Z}$ .

**Solution:**

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\beta + \mathbf{Z}\gamma + \boldsymbol{\varepsilon}) \\ &= \beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Z}\gamma + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon} \\ &= \beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon} \Rightarrow \\ \mathbb{E}[\hat{\beta}] &= \beta \end{aligned}$$

**c**

Using (b), derive an expression for the bias for a fixed  $\mathbf{x}_{\text{new}}$ , i.e.

$$\mathbb{E}[y_{\text{new}} - \mathbf{x}_{\text{new}}^\top \hat{\beta} | \mathbf{x}_{\text{new}}],$$

in terms of  $\beta$ ,  $\gamma$ , and the conditional expectation  $\mathbb{E}[\mathbf{z}_{\text{new}} | \mathbf{x}_{\text{new}}]$ .

**Solution:**

$$\begin{aligned}\mathbb{E} \left[ y_{\text{new}} - \mathbf{x}_{\text{new}}^T \hat{\beta} | \mathbf{x}_{\text{new}} \right] &= \mathbf{x}_{\text{new}}^T \beta + \mathbb{E} [\mathbf{z}_{\text{new}}^T | \mathbf{x}_{\text{new}}] \gamma + \mathbb{E} [\varepsilon_{\text{new}} | \mathbf{x}_{\text{new}}] - \mathbf{x}_{\text{new}}^T \mathbb{E} [\hat{\beta} | \mathbf{x}_{\text{new}}] \\ &= \mathbb{E} [\mathbf{z}_{\text{new}}^T | \mathbf{x}_{\text{new}}] \gamma.\end{aligned}$$

**d**

Using your result from (c), show that the predictions are biased at  $\mathbf{x}_{\text{new}}$  when omitting the variables  $\mathbf{z}_n$  from the regression precisely when  $\gamma^T \mathbb{E} [\mathbf{z}_n | \mathbf{x}_n] \neq 0$ . Using your result from (a), show that this bias can be expected to occur in general — that is, omitting variables can often induce biased predictions at a point.

**Solution:** This follows directly.

### 3 Estimating leave-one-out CV

This homework problem derives a closed-form estimate of the leave-one-out cross-validation error for regression. We will use the Sherman-Woodbury formula. Let  $A$  denote an invertible matrix, and  $\mathbf{u}$  and  $\mathbf{v}$  vectors the same length as  $A$ . Then

$$(A + \mathbf{u}\mathbf{v}^T)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^T A^{-1}}{1 + \mathbf{v}^T A^{-1}\mathbf{u}}.$$

We will also use the following definition of a “leverage score,”  $h_n := \mathbf{x}_n^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_n$ . We will discuss leverage scores more in the last lecture, but for now it’s enough that you know what it is. Note that  $h_n = (\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T)_{nn}$  is the  $n$ -th diagonal entry of the projection matrix  $\mathbf{P}_{\mathbf{X}}$ .

Let  $\hat{\beta}_{-n}$  denote the estimate of  $\hat{\beta}$  with the datapoint  $n$  left out. For leave-one-out CV, we want to estimate

$$MSE_{LOO} := \frac{1}{N} \sum_{n=1}^N (y_n - \mathbf{x}_n^T \hat{\beta}_{-n})^2.$$

Note that doing so naively requires computing  $N$  different regressions. We will derive a much more efficient formula.

Let  $\mathbf{X}_{-n}$  denote the  $\mathbf{X}$  matrix with row  $n$  left out, and  $\mathbf{Y}_{-n}$  denote the  $\mathbf{Y}$  matrix with row  $n$  left out.

**a**

Prove that

$$\hat{\beta}_{-n} = (\mathbf{X}_{-n}^\top \mathbf{X}_{-n})^{-1} \mathbf{X}_{-n}^\top \mathbf{Y}_{-n} = (\mathbf{X}^\top \mathbf{X} - \mathbf{x}_n \mathbf{x}_n^\top)^{-1} (\mathbf{X}^\top \mathbf{Y} - \mathbf{x}_n y_n)$$

**Solution**

This follows from  $\mathbf{X}^\top \mathbf{X} = \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^\top$  and  $\mathbf{X}^\top \mathbf{Y} = \sum_{n=1}^N \mathbf{x}_n y_n$ .

**b**

Using the Sherman-Woodbury formula, derive the following expression:

$$(\mathbf{X}^\top \mathbf{X} - \mathbf{x}_n \mathbf{x}_n^\top)^{-1} = (\mathbf{X}^\top \mathbf{X})^{-1} + \frac{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_n \mathbf{x}_n^\top (\mathbf{X}^\top \mathbf{X})^{-1}}{1 - h_n}$$

**Solution**

Direct application of the formula with  $\mathbf{u} = \mathbf{x}_n$  and  $\mathbf{v} = -\mathbf{x}_n$  gives

$$(\mathbf{X}^\top \mathbf{X} - \mathbf{x}_n \mathbf{x}_n^\top)^{-1} = (\mathbf{X}^\top \mathbf{X})^{-1} + \frac{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_n \mathbf{x}_n^\top (\mathbf{X}^\top \mathbf{X})^{-1}}{1 - \mathbf{x}_n^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_n}.$$

Then recognize the leverage score.

**c**

Combine (a) and (b) to derive the following explicit expression for  $\hat{\beta}_{-n}$ :

$$\hat{\beta}_{-n} = \hat{\beta} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_n \frac{1}{1 - h_n} \hat{\varepsilon}_n$$

**Solution**

We have

$$(\mathbf{X}^\top \mathbf{X} - \mathbf{x}_n \mathbf{x}_n^\top)^{-1} \mathbf{X}^\top \mathbf{Y} = \hat{\beta} + \frac{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_n \mathbf{x}_n^\top \hat{\beta}}{1 - h_n}.$$

and

$$(\mathbf{X}^\top \mathbf{X} - \mathbf{x}_n \mathbf{x}_n^\top)^{-1} \mathbf{x}_n y_n = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_n y_n + \frac{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_n h_n}{1 - h_n} y_n.$$

Combining,

$$\begin{aligned}
\hat{\beta}_{-n} &= \hat{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_n \left( \frac{\mathbf{x}_n^\top \hat{\beta}}{1 - h_n} - y_n - \frac{h_n}{1 - h_n} y_n \right) \\
&= \hat{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_n \left( \frac{1}{1 - h_n} \hat{y}_n - \left( 1 + \frac{h_n}{1 - h_n} \right) y_n \right) \\
&= \hat{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_n \left( \frac{1}{1 - h_n} \hat{y}_n - \frac{1}{1 - h_n} y_n \right) \\
&= \hat{\beta} - (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_n \frac{1}{1 - h_n} \hat{\varepsilon}_n
\end{aligned}$$

**d**

Using (c), derive the following explicit expression the leave-one-out error on the  $n$ -th observation:

$$y_n - \mathbf{x}_n^\top \hat{\beta}_{-n} = \frac{\hat{\varepsilon}_n}{1 - h_n}.$$

**Solution**

$$\begin{aligned}
y_n - \mathbf{x}_n^\top \hat{\beta}_{-n} &= y_n - \mathbf{x}_n^\top \hat{\beta} + \mathbf{x}_n^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_n \frac{1}{1 - h_n} \hat{\varepsilon}_n \\
&= y_n - \hat{y}_n + \frac{h_n}{1 - h_n} \hat{\varepsilon}_n \\
&= \left( 1 + \frac{h_n}{1 - h_n} \right) \hat{\varepsilon}_n \\
&= \frac{\hat{\varepsilon}_n}{1 - h_n}
\end{aligned}$$

**e**

Using (d), prove that

$$MSE_{LOO} := \frac{1}{N} \sum_{n=1}^N \frac{\hat{\varepsilon}_n^2}{(1 - h_n)^2},$$

where  $\hat{\varepsilon}_n = y_n - \hat{y}_n$  is the residual from the full regression without leaving any data out. Using this formula,  $MSE_{LOO}$  can be computed using only the original regression and  $(\mathbf{X}^\top \mathbf{X})^{-1}$ .

**Solution** Just plug it in.

**f**

Prove that  $\sum_{n=1}^N h_n = P$ , and  $0 \leq h_n \leq 1$ . Hint: if  $\mathbf{v}$  is a vector with a 1 in entry  $n$  and 0 otherwise, then  $h_n = \mathbf{v}^\top \mathbf{P}_X \mathbf{v}$ , and projection cannot increase a vector's norm. Recall also that  $\text{trace} \left( \mathbf{P}_X \right) = P$ .

**Solution**

$$0 \leq h_n = \mathbf{v}^\top \mathbf{P}_X \mathbf{v} \leq \|\mathbf{v}\|_2^2 = 1,$$

and

$$\sum_{n=1}^N h_n = \text{trace} \left( \mathbf{P}_X \right) = P.$$

**g**

Using (e) and (f), prove that  $MSE_{LOO} > RSS = \frac{1}{N} \sum_{n=1}^N \hat{\varepsilon}_n^2$ . That is, the  $RSS$  underestimates the leave-one-out cross-validation error.

**Solution**

We have  $1/(1 - h_n)^2 \geq 1$  because  $-1 < 0 \leq h_n \leq 1$ . At least some  $h_n > 0$ , since  $\sum_{n=1}^N h_n = P$ , so for at least one  $h_n$ ,  $1/(1 - h_n)^2 > 1$ . The result follows.