STAT151A Homework 6: Due April 19th

Your name here

1 Fit and regressors

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

$$
\begin{aligned}
RSS &:= \hat{\varepsilon}^\intercal \hat{\varepsilon} && \text{(Residual sum of squares)} \\
TSS &:= \boldsymbol{Y}^\intercal \boldsymbol{Y} && \text{(Total sum of squares)} \\
ESS &:= \hat{\boldsymbol{Y}}^\intercal \hat{\boldsymbol{Y}} && \text{(Explained sum of squares)} \\
R^2 &:= \frac{ESS}{TSS}.
\end{aligned}
$$

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 $\hat{\boldsymbol{Y}}^\intercal \hat{\varepsilon} = \boldsymbol{0}$.
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 $H_0 : \boldsymbol{\beta} = \mathbf{0}$,

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

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

Solutions:

1. $\phi = \frac{\hat{\boldsymbol{Y}}^\intercal \hat{\boldsymbol{Y}}}{P \, RSS/(N-P)} = \frac{N-P}{P} \frac{TSS}{RSS}$
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 $(\boldsymbol{x}_n, \boldsymbol{z}_n, y_n)$ be IID random variables, where $\boldsymbol{x}_n \in \mathbb{R}^{P_X}$ and $\boldsymbol{z}_n \in \mathbb{R}^{P_Z}$. Suppose that $\boldsymbol{x}_n$ and $\boldsymbol{z}_n$ satisfy $\mathbb{E}\left[\boldsymbol{x}_n \boldsymbol{z}_n^\intercal\right] = \mathbf{0}$.

Let $y_n = \boldsymbol{x}_n^\intercal \beta + \boldsymbol{z}_n^\intercal \gamma + \varepsilon_n$, where ε_n is mean zero, unit variance, and indepdendent of $\boldsymbol{x}_n$ and $\boldsymbol{z}_n$.

a

Take $P_X = P_Z = 1$ (i.e. scalar regressors). Show that there exists x_n and z_n such that $\mathbb{E}\left[x_n z_n\right] = 0$ but $\mathbb{E}\left[z_n | x_n\right] \neq 0$ 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 $\mathbb{E}\left[x_n z_n^2\right] = \mathbb{E}\left[x_n^3\right] = 0$.

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

$$\tilde{\boldsymbol{z}}_n = \boldsymbol{z}_n - \mathbb{E}\left[\boldsymbol{z}_n \boldsymbol{x}_n^\intercal\right] \left(\mathbb{E}\left[\boldsymbol{x}_n \boldsymbol{x}_n^\intercal\right]\right)^{-1} \boldsymbol{x}_n.$$

Then

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

but

$$\mathbb{E}\left[\tilde{\boldsymbol{z}}_n | \boldsymbol{x}_n^\intercal\right] \neq 0$$

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

b

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

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

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

Solution:

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

c

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

$$\mathbb{E}\left[y_{\mathrm{new}} - \boldsymbol{x}_{\mathrm{new}}^\intercal \hat{\beta} | \boldsymbol{x}_{\mathrm{new}}\right],$$

in terms of β, γ, and the conditional expectation $\mathbb{E}\left[\boldsymbol{z}_{\mathrm{new}} | \boldsymbol{x}_{\mathrm{new}}\right]$.

Solution:

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

d

Using your result from (c), show that the predictions are biased at $\boldsymbol{x}_{\text{new}}$ when omitting the variables $\boldsymbol{z}_n$ from the regression precisely when $\gamma^{\intercal}\mathbb{E}\left[\boldsymbol{z}_n|\boldsymbol{x}_n\right] \neq 0$. 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 $\boldsymbol{u}$ and $\boldsymbol{v}$ vectors the same length as A. Then

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

We will also use the following definition of a "leverage score," $h_n := \boldsymbol{x}_n^{\intercal}(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}\boldsymbol{x}_n$. We will discuss leverage scores more in the last lecture, but for now it's enough that you know what it is. Note that $h_n = (\boldsymbol{X}(\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}\boldsymbol{X}^{\intercal})_{nn}$ is the n–th diagonal entry of the projection matrix $\underset{\boldsymbol{X}}{\boldsymbol{P}}$.

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

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

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

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

a

Prove that

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

Solution

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

b

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

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

Solution

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

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

Then recognize the leverage score.

c

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

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

Solution

We have

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

and

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

Combining,

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

d

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

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

Solution

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

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 $(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}$.

Solution Just plug it in.

f

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

Solution

$$0 \le h_n = \boldsymbol{v}^\intercal \underset{\boldsymbol{X}}{\boldsymbol{P}} \boldsymbol{v} \le \|\boldsymbol{v}\|_2^2 = 1,$$

and

$$\sum_{n=1}^{N} h_n = \mathrm{trace}\left(\underset{\boldsymbol{X}}{\boldsymbol{P}}\right) = P.$$

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 $1/(1-h_n)^2 \ge 1$ because $-1 < 0 \le h_n \le 1$ 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.