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:

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

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

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

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

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

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 $\boldsymbol{z}_n$ are uncorrelated, so that $\mathbb{E}\left[\boldsymbol{x}_n \boldsymbol{z}_n^\intercal\right] = \boldsymbol{0}$.

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.)

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

Hint: by the Tower property,

$$\mathbb{E}\left[(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{Z}\right] = \mathbb{E}\left[(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \mathbb{E}\left[\boldsymbol{Z} | \boldsymbol{X}\right]\right].$$

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 $\mathbb{E}\left[\boldsymbol{z}_{\mathrm{new}} | \boldsymbol{x}_{\mathrm{new}}\right]$.

d

Using your result from (c), show that the predictions are biased at $\boldsymbol{x}_{\mathrm{new}}$ 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.

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

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

(For reference, I provide a proof in the notes for lecture 21.)

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

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

b

Using the Sherman-Woodbury formula, find an explicit expression for

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

in terms of $\boldsymbol{x}_n$, y_n, h_n, and $(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}$.

c

Combine (a) and (b) to derive an explicit expression for $\hat{\boldsymbol{\beta}}_{-n}$ in terms of $\boldsymbol{x}_n$, y_n, h_n, and $(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}$.

d

Using (c), derive an explicit expression for $\boldsymbol{x}_n^\intercal \hat{\boldsymbol{\beta}}_{-n}$ that depends only on h_n and $\hat{y}_n = \hat{\boldsymbol{\beta}}^\intercal \boldsymbol{x}_n$, where $\hat{y}_n$ is computed using the full dataset without leaving out row n.

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

f

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

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 under-estimates the leave-one-out cross-validation error.