STAT151A Homework 3.

Your name here

This homework is due on Gradescope on **Friday October 11th at 9pm**.

1 Interpretation of transforms of the response

Suppose I have data of the where $n = 1, \dots N$ indexes households, y_n is the expenditure on food in a time period (so that $y_n > 0$ for all n). Suppose that households are randomly selected to either receive food stamps (for which $x_n = 1$) or to not receive food stamps (for which $x_n = 0$). Also, suppose we measure household income z_n.

(a)

Suppose we regress $y_n \sim \beta_0 + \beta_1 x_n + \beta_2 z_n$. Let $\hat{f}(x,z) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 z$ denote the fit.

Using this regression, we might estimate the effect of food stamps on food expenditure by $\hat{f}(1, z) - \hat{f}(0, z)$. How does this estimate depend on z?

(b)

Suppose we regress $\log y_n \sim \gamma_0 + \gamma_1 x_n + \gamma_2 z_n$. Let $\hat{g}(x,z) = \hat{\gamma}_0 + \hat{\gamma}_1 x + \hat{\gamma}_2 z$.

Using this regression, we might estimate the effect of food stamps on food expenditure by $\exp(\hat{g}(1, z)) - \exp(\hat{g}(0, z))$. How does this estimate depend on z?

(c)

The regressions in (a) and (b) make different implicit assumptions about how food stamps affect consumption for a particular household. State these assumptions in ordinary language.

2 Fit and regressors

Given a regression on $\boldsymbol{X}$ with P regressors and N data points, 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. Prove that $RSS + ESS = TSS$.
b. Express R^2 in terms of TSS and RSS.
c. What is R^2 when we include no regressors? ($P = 0$)
d. What is R^2 when we include N linearly independent regressors? ($P = N$)
e. Can R^2 ever decrease when we add a regressor? If so, how?
f. Can R^2 ever stay the same when we add a regressor? If so, how?
g. Can R^2 ever increase when we add a regressor? If so, how?
h. Does a high R^2 mean the regression is useful? (You may argue by example.)
i. Does a low R^2 mean the regression is not useful? (You may argue by example.)

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.

3 Prediction in the bodyfat example

This exercise will use the bodyfat example from the datasets. Suppose we're interested in predicting `bodyfat`, which is difficult to measure precisely, with other variables which are easier to measure: `Height`, `Weight`, and `Abdomen` circumference.

If we do so, we get the following sum of squared errors:

```
reg <- lm(bodyfat ~ Abdomen + Height + Weight, bodyfat_df)
print(reg$coefficients)
```

```
(Intercept)     Abdomen      Height      Weight
-36.6147193   0.9515631  -0.1270307  -0.1307606
```

```
print(sprintf("Error: %f", mean(reg$residuals^2)))
```

```
[1] "Error: 19.456161"
```

(a)

Noting that `Heigh`, `Weight`, and `Abdomen` are on different scales, your colleague suggests that you might get a better fit by normalizing them. But when you do, here's what happened:

```
bodyfat_df <- bodyfat_df %>%
    mutate(height_norm=(Height - mean(Height)) / sd(Height),
           weight_norm=(Weight - mean(Weight)) / sd(Weight),
           abdomen_norm=(Abdomen - mean(Abdomen)) / sd(Abdomen))
reg_norm <- lm(bodyfat ~ abdomen_norm + height_norm + weight_norm, bodyfat_df)
print(reg_norm$coefficients)
```

```
 (Intercept) abdomen_norm  height_norm  weight_norm
   19.150794    10.260778    -0.465295    -3.842945
```

```
print(sprintf("Error: %f", mean(reg_norm$residuals^2)))
```

```
[1] "Error: 19.456161"
```

Our coefficients changed, but our fitted error didn't change at all.

- Explain why the fitted error did not change.
- Explain why the coefficients did change.

(b)

Chastened, your colleague suggests that maybe it's the difference between normalized height and weight that would helps us predict. After all, it makes sense that height should only matter relative to weight, and vice versa. So they run the regresson on the difference:

```
bodyfat_df <- bodyfat_df %>%
    mutate(hw_diff = height_norm - weight_norm)
reg_diff <- lm(bodyfat ~ abdomen_norm + height_norm + weight_norm + hw_diff, bodyfat_df)
print(reg_diff$coefficients)
```

```
 (Intercept) abdomen_norm  height_norm  weight_norm      hw_diff
   19.150794    10.260778    -0.465295    -3.842945           NA
```

```
print(sprintf("Error: %f", mean(reg_diff$residuals^2)))
```

```
[1] "Error: 19.456161"
```

Now, our fitted error didn't change at all, but our difference coefficient wasn't even estimated.

- Explain why the fitted error did not change.
- Explain why the difference coefficient was not estimated by `R`.

(c)

Finally, your colleague suggests regressing instead on the *ratio* of weight to height. Here are the results:

```
bodyfat_df <- bodyfat_df %>%
    mutate(hw_ratio = height_norm / weight_norm)
reg_ratio <- lm(bodyfat ~ abdomen_norm + height_norm + weight_norm + hw_ratio, bodyfat_df)
print(reg_ratio$coefficients)
```

```
 (Intercept) abdomen_norm  height_norm  weight_norm     hw_ratio
19.149475485 10.271751693 -0.478683698 -3.848561820  0.009057317
```

```
print(sprintf("Error: %f", mean(reg_ratio$residuals^2)))
```

```
[1] "Error: 19.423560"
```

Our fitted error is different this time, and we could estimate this coefficient.

- Explain why we could we estimate the coefficient of the ratio of height to weight, but not the difference.
- Explain why the fitted error changed.

- It happened that, by including the regressor `hw_ratio`, the fitted error decreased. Your colleague tells you that had it been a bad regressor, the error would have increased. Are they correct?

(d)

Let $x_n = (\text{Abdomen}_n, \text{Height}_n, \text{Weight}_n)^\intercal$ denote our set of regressors.

Your colleague suggests a research project where you improve your fit by regressing $y_n \sim z_n$ for new regressors z_n of the form $z_n = \boldsymbol{A}x_n$, where the matrix $\boldsymbol{A}$ is chosen using a machine learning algorithm.

- Will this result produce a better fit to the data than simply regressing $y_n \sim \boldsymbol{x}_n$? Why or why not?

(e)

Finally, your colleague suggests a research project where you again regression $y_n \sim z_n$, but now you let $z_n = f(x_n)$ for any function f, where you use a neural network to find the best fit to the data over all possible functions $f(x_n)$.

- Will this result produce a better fit to the data than simply regressing $y_n \sim \boldsymbol{x}_n$? Why or why not?
- Do you think this result produce a useful prediction for new data? Why or why not?

4 Leaving a single datapoint out of regression

This homework problem derives a closed-form expression for the effect of leaving a datapoint out of the regression.

We will use the following result, known as the Woodbury formula (but also many other names, including the Sherman-Morrison-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 definition of a "leverage score" from lecture $h_n := \boldsymbol{x}_n^\intercal(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{x}_n$. 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 $\boldsymbol{P}_{\boldsymbol{X}}$.

Let $\hat{\boldsymbol{\beta}}_{-n}$ denote the estimate of $\hat{\beta}$ with the datapoint n left out. Similarly, 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 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

Letting $\hat{y}_{n,-n} = \boldsymbol{x}_n^\intercal \hat{\boldsymbol{\beta}}_{-n}$ denote the estimate of y_n after deleting the n–th observation. Using (c), derive the following explicit expression the change in $\hat{y}_n$ upon deleting the n–th observation:

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

This shows that the effect of deleting observation n on $\hat{y}_n$ is large only if both the residual and leverage score is large.