

## Homework #6

**Total Points: 26**

## Submission Instructions

You must submit this assignment to Gradescope by **Thursday, October 6th at 11:59 PM Pacific**.

You can work on this assignment in any way you like:

- One way is to download this PDF, print it out, and write directly on these pages (we've provided enough space for you to do so). Alternatively, if you have a tablet, you could save this PDF and write directly on it.
- You could also write your answers on a blank sheet of paper.
- You can also typeset your answers with  $\text{\LaTeX}$ . Overleaf is a great tool to get started!

Regardless of what method you choose, the end result needs to end up on Gradescope, as a PDF. If you wrote something on physical paper (like options 1 and 3 above), you will need to use a scanning application (e.g. CamScanner) in order to submit your work.

When submitting on Gradescope, you **must correctly assign pages to each question** (it prompts you to do this after submitting your work). This significantly streamlines the grading process for our readers. Failure to do this may result in a score of 0 for any questions that you didn't correctly assign pages to. If you have any questions about the submission process, please don't hesitate to ask on Piazza.

## Collaborators

Data science is a collaborative activity. While you may talk with others about the homework, we ask that you write your solutions individually. If you do discuss the assignments with others please include their names at the top of your submission.

## Normal equations

1. (12 points) In lecture, we discussed a geometric argument for the *normal equations* we solve to get the least squares estimator  $\hat{\theta}$ :

$$\mathbb{X}^\top(\mathbb{Y} - \mathbb{X}\theta) = 0.$$

Here, we are using  $\mathbb{X}$  to denote the design matrix:

$$\mathbb{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix} = \begin{bmatrix} | & | & | & \cdots & | \\ 1_n & X_1 & X_2 & \cdots & X_p \\ | & | & | & & | \end{bmatrix}$$

where  $1_n$  is the  $n$ -vector of all 1s and  $X_j$  is the  $n$ -vector  $[x_{1,j}, \dots, x_{n,j}]^\top$  holding all observations of the  $j$ th variable. We use  $\mathbb{Y}$  to denote the response vector as an  $n \times 1$  matrix (note that there would be nothing wrong with writing  $y - \mathbb{X}\theta$ , we would just interpret the vector  $y$  as a length- $n$  column vector, which is the same as an  $n \times 1$  matrix).

The normal equations are the estimating equations for OLS, i.e. (multiple) linear regression under the least squares loss. As we discussed in class, while there can be multiple values of  $\theta \in \mathbb{R}^{p+1}$  that satisfy these equations, there is always at least one solution.

To build intuition for these equations and relate them to the SLR estimating equations, we will derive them mathematically in several ways.

- (a) (2 points) Show that finding the optimal estimator  $\hat{\theta}$  by solving the normal equations is equivalent to requiring that the residual vector  $e = \mathbb{Y} - \mathbb{X}\hat{\theta}$  should average to zero, and should be orthogonal to  $X_j$  (i.e., should have zero dot product with  $X_j$ ) for every  $j$ , that is, show that

$$\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i = 0$$

and

$$X_j^\top e = \sum_i x_{i,j} e_i = 0$$

for all  $j = 1, \dots, p$ . That is, the equations above are the same as the normal equations, just translated to summation notation rather than matrix notation.

(b) (4 points) Remember that the (empirical) MSE for multiple linear regression is

$$\text{MSE}(\theta) = \frac{1}{n} \sum_{i=1}^n (y_i - \theta_0 - \theta_1 x_{i,1} - \cdots - \theta_p x_{i,p})^2$$

Use calculus to show that any  $\theta = [\theta_0, \theta_1, \dots, \theta_p]^\top$  that minimizes the MSE must solve the normal equations.

Hint: remember that, at a minimum of MSE, the partial derivatives of MSE with respect to every  $\theta_i$  must all be zero.

- (c) (3 points) Recall that we define the correlation  $r$  between two vectors  $u$  and  $v$  as

$$r_{u,v} = \frac{1}{n} \sum_i \left( \frac{u_i - \bar{u}}{\sigma_u} \right) \left( \frac{v_i - \bar{v}}{\sigma_v} \right).$$

- (i) Show that in general, for *any* two non-constant vectors  $u$  and  $v$ , if

$$u^\top v = 0 \text{ and } \bar{u} = 0,$$

then  $u$  and  $v$  are uncorrelated, i.e.  $r_{u,v} = 0$ .

By non-constant, we mean that the coordinates are not all equal to the same value.

- (ii) Use (i) to conclude that the residuals  $e = \mathbb{Y} - \mathbb{X}\hat{\theta}$  are uncorrelated with every non-constant predictor variable  $X_j$ , as long as the residuals are not constant (i.e. at least one prediction is not perfect), that is, show that

$$r_{e,X_j} = 0$$

for all  $X_j$ .

- (d) (3 points) Use the previous part to show that the residuals are also uncorrelated to the fitted values  $\hat{y} = \mathbb{X}\hat{\theta}$  as long as neither  $e$  nor  $\hat{y}$  is a constant vector, that is, show that

$$r_{e,\hat{y}} = 0.$$

## Multiple $R^2$

2. (14 points) In a simple linear regression, the  $r^2$  or squared correlation between  $x$  and  $y$ , is often used as a measure of how well the linear model fits the data. That is,  $r^2$  measures how strong is the linear relationship between the response  $y$  and the single predictor variable  $x$ . It is natural to ask how we can generalize this concept to the setting of a regression on multiple predictor variables  $X_1, \dots, X_p$ . That is, we would like to ask, how strong is the linear relationship between the response  $y$  and the full set of predictors  $X_1, \dots, X_p$ .

The usual way that people generalize this concept is the multiple  $R^2$ , which we defined in class as the variance of the fitted values divided by the variance of the response:

$$R^2 = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2}.$$

We mentioned in lecture that it is often referred to as the fraction or percentage of “**variance explained by the regression**,” but we did not explain why people call it that. In fact, the multiple  $R^2$  is a popular measure of how well the regression fits the data, because it has several interpretations:

1.  **$R^2$  is the square of the correlation of the prediction vector  $\hat{\mathbf{y}}$  with the response  $\mathbf{y}$** , so if it is very high, the response lines up almost exactly with the predictions, but if it is very low, the response has almost no association with the predictions. We will show this in part (e).
2.  **$1 - R^2$  is the variance of the residuals as a fraction of the original  $\sigma_y^2$** , so if it is large, the residuals are very small compared to the variation of  $y$ . We will show this in part (d).
3. In the special case of OLS regression with only a single predictor variable, we can still compute  $R^2$  and it is exactly the same as the squared correlation  $r^2$  between  $x$  and  $y$ . We will show this in part (f).

For all parts of this question, you can freely assume that any vectors like  $y, e, X_j$ , and so on are non-constant if we are calculating correlations (correlations with constant vectors are undefined).

- (a) (3 points) We have been using the notation  $\sigma_x^2$  and  $\sigma_y^2$  in class to denote the *variance* of vectors  $x$  and  $y$ , which are defined as the average squared deviation from the  $\bar{x}$  and  $\bar{y}$ , respectively:

$$\sigma_x^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2, \quad \sigma_y^2 = \frac{1}{n} \sum_i (y_i - \bar{y})^2.$$

This is closely related to the idea of the variance of a random variable, which we will cover later in the course, but for this problem you do not need to think about

random variables, we are only referring to this specific sum, which measures how “spread out” the coordinates of a vector are around the average.

To get practice working with the variance, show that if  $\hat{y}_i = a + bx_i$ , that the mean is

$$\bar{\hat{y}} = a + b\bar{x}$$

and the variance is

$$\sigma_{\hat{y}}^2 = b^2\sigma_x^2.$$

- (b) (3 points) For any two non-constant vectors  $u$  and  $v$ , let  $\sigma_u^2$  and  $\sigma_v^2$  denote their variances and let

$$r_{u,v} = \frac{1}{n} \sum_i \left( \frac{u_i - \bar{u}}{\sigma_u} \right) \left( \frac{v_i - \bar{v}}{\sigma_v} \right),$$

where the variances are defined as  $\sigma_u^2 = \frac{1}{n} \sum_i (u_i - \bar{u})^2$  and  $\sigma_v^2 = \frac{1}{n} \sum_i (v_i - \bar{v})^2$ .

Show that the variance of the sum of the two vectors is

$$\sigma_{u+v}^2 = \sigma_u^2 + \sigma_v^2 + 2\sigma_u\sigma_v r_{u,v}.$$

Hint: Write out  $\sigma_{u+v}^2$  using the definition of variance, using the fact that the mean of  $u + v$  is  $\bar{u} + \bar{v}$ . Remember that we can expand the square of any sum to get  $(a + b)^2 = a^2 + b^2 + 2ab$  (figure out what is the right  $a$  and  $b$  to use for this).

- (c) (2 points) Remember the conclusion of 1(d), that  $r_{e,\hat{y}} = 0$ . Use this, along with part (b) of this question, to show that the variance of the response  $y$  can be decomposed as a sum of the variance of the fitted values  $\hat{y}$  and the variance of the residuals  $e$ :

$$\sigma_y^2 = \sigma_{\hat{y}}^2 + \sigma_e^2,$$

where  $\sigma_e^2$  is the variance of the residuals:  $\sigma_e^2 = \frac{1}{n} \sum_i (e_i - \bar{e})^2$  and the variances  $\sigma_y^2$  and  $\sigma_{\hat{y}}^2$  are defined similarly. Conclude that  $0 \leq R^2 \leq 1$ .

Hint: remember that  $y = \hat{y} + e$ .

- (d) (2 points) Use the last part to show that  $1 - R^2 = \sigma_e^2 / \sigma_y^2$ .

Note on motivation: In other words, the residuals — i.e., the part of the response that is unexplained by the predictors — are less variable than the response, and the factor by which they shrink is exactly  $1 - R^2$ . So if, for example,  $R^2 = 0.7$ , the residuals from the regression are 70% smaller than the residuals for the constant model. This motivates the idea that the regression has therefore “explained away” 70% of the variance in the response.

- (e) (4 points) One of the interpretations of multiple  $R^2$  is that it measures the strength of the linear relationship between the predictions and the response. Show that  $R^2$  is the square of the correlation between the fitted values and the responses, that is, show that

$$R^2 = (r_{\hat{y},y})^2.$$

Hint: The problem will be easiest if you start by writing an expression for  $\sigma_y \sigma_{\hat{y}} r_{\hat{y},y}$ . Then, substitute  $y_i = \hat{y}_i + e_i$ , and use the distributive property to get an expression in terms of  $r_{e,\hat{y}}$  and  $\sigma_{\hat{y}}^2$ , and rearrange to solve for  $r_{\hat{y},y}$ .

- (f) (0 points) (Optional – no extra credit) At the beginning of this problem we motivated  $R^2$  as generalizing the concept of the correlation  $r_{x,y}$  from a simple linear regression with one predictor variable  $x$ . Show that for SLR,  $R^2$  coincides with the square of the correlation between  $x$  and  $y$ , i.e.  $R^2 = (r_{x,y})^2$ .

Hint: From the conclusion of part (e), we only need to show that  $(r_{\hat{y},y})^2 = (r_{x,y})^2$ . Remember you showed in part (a) that if  $\hat{y} = a + bx$ , then for any real numbers  $a, b$  and vector  $x$ , we have  $\bar{\hat{y}} = a + b\bar{x}$  and  $\sigma_{\hat{y}}^2 = b^2\sigma_x^2$ . As a result,  $\sigma_{\hat{y}} = |b|\sigma_x$ , because both  $\sigma_{\hat{y}}$  and  $\sigma_x$  are positive.