

# Practice Questions for Test 1

(1) Consider fitting the usual simple linear regression model to these data:

$$\begin{array}{c|ccc} x & -1 & 0 & 1 \\ \hline y & 1 & -1 & 1 \end{array}$$

(a) Find  $\widehat{\beta}_1$ ,  $\widehat{\sigma}$ ,  $\widehat{\text{se}}$  and a 95 percent confidence interval for  $\beta_1$ .

(b) Construct the ANOVA table.

(c) Consider fitting the data to the model:  $Y_i = \beta_0 + \epsilon_i$ . Find  $\widehat{\beta}_0$  and  $\widehat{\sigma}$ .

---

(2) Suppose we fit the usual model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

assuming  $\mathbb{E}(\epsilon_i) = 0$  and  $\mathbb{V}(\epsilon_i) = \sigma^2$ . Suppose, however, that  $\mathbb{E}(\epsilon_i) = 5$  and  $\mathbb{V}(\epsilon_i) = x_i^2$ . How does this affect the mean and variance of  $\widehat{\beta}_1$ ?

---

(3) Let

$$\mathbb{X} = \begin{pmatrix} 1 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

(a) Describe the column space  $\mathcal{L}$  corresponding to  $\mathbb{X}$ ?

(b) Find  $H = \mathbb{X}(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T$  and verify that  $H$  is a projection matrix for  $\mathcal{L}$ .

---

(4) Suppose that

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

Further, assume that  $\beta_0$  and  $\beta_1$  are known. Given a value  $x_*$  of  $x$ , find a 95 percent prediction interval for  $Y = \beta_0 + \beta_1 x_* + \epsilon$ . Do you need to assume that  $\epsilon$  is Normal for your interval to be valid?

---

(5) In a multiple regression  $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_7 x_7 + \epsilon$ , explain how to test

$$H_0 : \beta_1 = \beta_3 = \beta_5 = 0.$$

What is the alternative hypothesis? What does it mean if we reject this test? What does it mean if we fail to reject this test?

---

(6) Suppose we have data  $(X_1, Y_1), \dots, (X_n, Y_n)$  and we want to fit the usual regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

By mistake, we fit the model  $Y_i = \beta_0 + \epsilon_i$ , that is, we omit  $\beta_1$ . Let  $\hat{\beta}_0$  be our estimated intercept. What is the mean of  $\hat{\beta}_0$ ? What is the variance of  $\hat{\beta}_0$ ?

---

(7) Draw a scatterplot that has:

- (a) A non-linear regression curve and  $\hat{\beta}_1 = 0$ .
  - (b) Two large outliers with high influence.
  - (c) Two influential points that are not outliers.
  - (d) Highly non-normal residuals.
- 

(8) Let  $R = \mathbb{E}(Y - \hat{Y})^2$  be the prediction error. Show that

$$R = \text{bias}^2 + \text{var}(\hat{Y}) + \sigma^2$$

where  $\text{bias} = \mathbb{E}(\hat{Y}) - \mathbb{E}(Y)$ .

---

(9) Consider the following data:

| X  | Y  |
|----|----|
| -1 | -1 |
| 1  | 1  |

Use the usual regression model. Find  $\hat{\beta}$ ,  $\text{se}(\hat{\beta})$ , the ANOVA table,  $R^2$  and the hat matrix.

---

(10) In the regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

suppose we replace each  $X_i$  with  $2X_i$ . What effect will this have on (i)  $\hat{\beta}_1$ , (ii)  $\text{se}(\hat{\beta}_1)$ , (iii)  $R^2$ , (iv) the test that  $\beta_1 = 0$ ?

---

(11) Explain the difference between residuals, standardized residuals and studentized residuals.

---

(12) Consider qq-plots for residuals. Draw a qq-plot for:

- (i) a distribution that is skewed,

(ii) a distribution with one large residual,

(iii) a distribution where the residuals only take 5 possible values.

---

**(13)** We want to fit the linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

However, we do not observe  $X_i$ . Rather we observe  $\tilde{X}_i = X_i + \delta_i$  where  $\delta_i \sim N(0, 1)$ . Suppose we perform the usual linear regression using the  $\tilde{X}_i$ 's in place of the  $X_i$ 's. How does this affect  $\hat{\beta}_1$ ?

---

**(14)** Show that, if the hat matrix is invertible, then it is the identity matrix.

---

**(15)** You are given data  $(X_1, Y_1), \dots, (X_n, Y_n)$  and you fit the following model:

$$Y_i = \beta X_i + \epsilon_i$$

using least squares. (Note that there is no intercept.) Suppose the data were actually generated from the model  $Y_i = X_i^2 + \epsilon_i$  where  $\epsilon_i \sim N(0, 1)$ . Find the mean and variance of  $\hat{\beta}$  (conditional on  $X_1, \dots, X_n$ ).

---

**(16)** Let

$$\mathbb{X} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

(a) Describe the column space  $\mathcal{L}$  corresponding to  $\mathbb{X}$ .

(b) Find  $H = \mathbb{X}(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T$ . Hint: if you look closely at your answer to (a) you can deduce the answer to (b) without doing any calculations.

---

**(17)** Suppose we have data  $(X_1, Y_1), \dots, (X_n, Y_n)$  and we want to fit the usual regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

We do not observe  $Y_i$ . Instead we observe

$$\tilde{Y}_i = Y_i + \delta_i$$

where  $\delta_i \sim N(0, 1)$ . Suppose we estimate  $\beta_1$  by regressing the  $\tilde{Y}_i$ 's on the  $X_i$ 's. Explain precisely how this affects the estimate  $\hat{\beta}_1$ .

(18) Suppose we have the following data:

$$\begin{array}{c|ccccc} X & -2 & -1 & 0 & 1 & 2 \\ Y & 0 & 0 & 0 & 0 & 0 \end{array}$$

Consider two regression models:

$$\text{Model 1} \quad Y_i = \beta_0 + \epsilon_i$$

$$\text{Model 2} \quad Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

(a) Use  $C_p$  to choose between these model. You may assume that  $\sigma^2 = 1$ .

(b) Fit the model  $Y_i = \beta X_i + \epsilon_i$  using the lasso. That is, find  $\hat{\beta}$  to minimize

$$\sum_i (Y_i - \beta X_i)^2 + \lambda |\beta|.$$

---

(19) Suppose we have data  $Y_1, \dots, Y_n$  and the model is

$$Y_i = \beta_0 + \epsilon_i$$

where  $\mathbb{E}(\epsilon_i) = 0$ .

(a) Find the least squares estimator  $\hat{\beta}_0$ .

(b) Find an explicit expression for the cross-validation score.

(b) Find an explicit expression for the generalized cross-validation score.