Lecture 2, Part I: Review of Linear Regression

Text references: 401 course material

Model Basics and Assumptions

- Recall our model building block from last time:

$$Y = r(X_1) + \varepsilon,$$

  where $\mathbb{E}(\varepsilon|X_1) = 0$. The regression function here is $r(X_1) = \mathbb{E}(Y|X_1)$, or $r(x) = \mathbb{E}(Y|X_1 = x)$. (We write X_1 here to reflect the fact that we just have one predictor, i.e., $X_1 \in \mathbb{R}$)

- In linear regression, we predict Y from a linear function of X_1, of the form $\beta_0 + \beta_1 X_1$. If we determine β_0, β_1 by minimizing mean squared error,

$$\mathrm{MSE}(\beta_0, \beta_1) = \mathbb{E}[(Y - \beta_0 - \beta_1 X_1)^2],$$

  then recall from last time that

$$\beta_1 = \frac{\mathrm{Cov}(X_1, Y)}{\mathrm{Var}(X_1)}, \quad \beta_0 = \mathbb{E}(Y) - \beta_1 \mathbb{E}(X_1)$$

Multiple Linear Regression

What happens now with p predictors, $X_1, \ldots X_p$?

We want to model Y as a linear function

Using MSE again as our criterion,

The optimal coefficients are given by

We will refer to these as the **population regression coefficients**.

Our multiple (linear regression) model is

$$Y = r(\mathbf{X}) + \varepsilon.$$

What are really our *assumptions* when using linear regression? Recall from your regression class,

[See catexample.R for a linear regression example]

Briefly, note that we can summarize these assumptions as

$$Y|\mathbf{X} \sim N(\beta_0 + \boldsymbol{\beta}^T \mathbf{X}, \sigma^2)$$

Remarks on these assumptions:

Linear Regression Estimates from Data

In practice, we don't have access to the distributions of $\mathbf{X}, Y$ so we cannot actually compute the population regression coefficients. Instead, we have, say, an iid sample of size n,

$$(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_n, Y_n)$$

from some distribution of $\mathbf{X}$ and Y. Note that each $\mathbf{X}_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$, but for simplicity, *we will from now on follow the notational convention of most statistical texts (including AOS, Witten et al., and Shalizi starting in Chapter*

3) and drop the bold-faced notation on the covariates (so a subscript i will denote ***observation*** *i).*

Hence, we write our linear model as

Now define the data matrix or the *design matrix* $\mathbb{X}$ as

Each subject corresponds to one row. **The number of columns of $\mathbb{X}$ corresponds to the number of features plus 1 for the intercept**, $q = p + 1$. Now define the vectors $\vec{Y}$, $\vec{\epsilon}$, β as

$$\vec{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad \vec{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}. \tag{1}$$

Then, we can write our linear model more concisely as

Under squared error loss, we minimize the residual sums of squares (RSS)

which gives as the **sample regression coefficients** (or just regression coefficients) $\widehat{\beta}$:

Think: What's the dimension of the matrix A? What's the distribution of $\widehat{\beta}_i$? Why, under what assumptions?

Properties of Least Squares Estimates

Theorem 1 *The estimators satisfy the following properties (conditional on $\mathbb{X}$)*

1. $\mathbb{E}(\widehat{\beta}) = \beta$.
2. $\mathbb{V}(\widehat{\beta}) = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$
3. $\widehat{\beta} \approx MVN(\beta, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1})$.
4. *An approximate $1-\alpha$ confidence interval for β_j is*
$$\widehat{\beta}_j \pm z_{\alpha/2}\, \widehat{\mathsf{se}}(\widehat{\beta}_j) \tag{2}$$
where $\widehat{\mathsf{se}}(\widehat{\beta}_j)$ is the square root of the appropriate diagonal element of the matrix $\widehat{\sigma}^2(\mathbb{X}^T\mathbb{X})^{-1}$.

Exercise: Prove the first two assertions[1]

[1]Recall: Let Y be a random vector. Denote the mean vector by μ and the covariance matrix by Σ. If a is a vector then
$$\mathbb{E}(a^TY) = a^T\mu, \quad \mathbb{V}(a^TY) = a^T\Sigma a.$$
If A is a matrix then
$$\mathbb{E}(AY) = A\mu, \quad \mathbb{V}(AY) = A\Sigma A^T.$$

Q: Why are the last two assertions useful? (Give examples)