Chapter 5

Least Squares

Inconsistent Systems

In regression (and many other applications) we have a system of equations we'd like to solve:

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$$

However, this system does not have an exact solution. (i.e. all of our data points don't lie exactly on a *flat* surface)

- The best we can do is consider an equation with error and try to minimize that error:

$$\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{y} + \boldsymbol{\epsilon}$$

$$\mathbf{X}\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{y}}$$

- $\widehat{\mathbf{y}}$ is the vector of predicted values.
- $\widehat{\boldsymbol{\beta}}$ is the vector of parameter estimates.
- $\mathbf{X}$ is the design matrix.
- $\boldsymbol{\epsilon} = \widehat{\mathbf{y}} - \mathbf{y}$ is a vector of residuals

The Normal Equations

Since we can't solve $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$, we want to solve $\mathbf{X}\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{y}}$, where

$$\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\widehat{\mathbf{y}} - \mathbf{y})^T(\widehat{\mathbf{y}} - \mathbf{y}) \text{ is minimized.}$$

- (Remember, $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$ is just the sum of squared error.)
- Then $\widehat{\boldsymbol{\beta}}$ is called a **least-squares solution.**

The Normal Equations

Since we can't solve $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$, we want to solve $\mathbf{X}\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{y}}$, where

$$\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\widehat{\mathbf{y}} - \mathbf{y})^T(\widehat{\mathbf{y}} - \mathbf{y}) \ \text{ is minimized.}$$

The Normal Equations

Since we can't solve $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$, we want to solve $\mathbf{X}\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{y}}$, where

$$\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\widehat{\mathbf{y}} - \mathbf{y})^T(\widehat{\mathbf{y}} - \mathbf{y}) \text{ is minimized.}$$

- The set of least-squares solutions is precisely the set of solutions to the **Normal Equations**,

$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}.$$

The Normal Equations

Since we can't solve $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$, we want to solve $\mathbf{X}\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{y}}$, where

$$\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\widehat{\mathbf{y}} - \mathbf{y})^T(\widehat{\mathbf{y}} - \mathbf{y}) \text{ is minimized.}$$

- The set of least-squares solutions is precisely the set of solutions to the **Normal Equations**,

$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}.$$

- There is a unique solution if and only if $\mathbf{X}$ has full rank.

The Normal Equations

Since we can't solve $\mathbf{X}\beta = \mathbf{y}$, we want to solve $\mathbf{X}\widehat{\beta} = \widehat{\mathbf{y}}$, where

$$\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\widehat{\mathbf{y}} - \mathbf{y})^T(\widehat{\mathbf{y}} - \mathbf{y}) \text{ is minimized.}$$

- The set of least-squares solutions is precisely the set of solutions to the **Normal Equations**,

$$\mathbf{X}^T\mathbf{X}\widehat{\beta} = \mathbf{X}^T\mathbf{y}.$$

- There is a unique solution if and only if $\mathbf{X}$ has full rank.
    - Linear independence of variables.

The Normal Equations

Since we can't solve $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$, we want to solve $\mathbf{X}\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{y}}$, where

$$\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\widehat{\mathbf{y}} - \mathbf{y})^T(\widehat{\mathbf{y}} - \mathbf{y}) \text{ is minimized.}$$

- The set of least-squares solutions is precisely the set of solutions to the **Normal Equations**,

$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}.$$

- There is a unique solution if and only if $\mathbf{X}$ has full rank.
  - Linear independence of variables.
  - #NoPerfectMulticollinearity

The Normal Equations

$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$$

When $\mathbf{X}$ has full rank, $\mathbf{X}^T\mathbf{X}$ is invertible. So we can multiply both sides by the inverse matrix:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

And then by definition, our predicted values are

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

The Normal Equations

$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$$

When $\mathbf{X}$ has full rank, $\mathbf{X}^T\mathbf{X}$ is invertible. So we can multiply both sides by the inverse matrix:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

And then by definition, our predicted values are

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

The Intercept

Remember that we generally have an intercept built into our model:

$$\beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_p \mathbf{x}_p = \mathbf{y}$$

This means our *design* matrix, **X**, has a built-in column of ones:

$$\underbrace{\begin{array}{c} \\ obs_1 \\ obs_2 \\ \vdots \\ obs_n \end{array} \begin{array}{c} \begin{array}{ccccc} & \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_p \end{array} \\ \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \end{array}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}}_{\boldsymbol{\beta}} = \underbrace{\begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{y}}$$