Lecture 2: More on Linear Models

(Text Sections 2.4, 6.2.2, 6.4, 6.5)

Types of Predictor Variables

Predictor variables are either *numeric* or *categorical*. Numeric variables take on meaningful numeric quantities. They are further classified as *continuous* or *discrete*. Continuous variables can take on any value in a specified interval (e.g. lot size of house, price of house can be anything from 0 to infinity). Discrete variables can take on only specified values, possibly from an infinite set (e.g. suite = yes or no, number of bedrooms = 0, 1, 2,...).

Categorical variables (sometimes called *factors*) are not numeric, but rather take on various specified *levels* (e.g. house type = condo, townhouse, or detached). If the levels are ordered, we refer to the variable as an *ordered categorical variable* (e.g. house size = small, medium, or large).

Example 3: Numeric predictor variables

Numeric predictor variables are easy to incorporate in the model. To represent a continuous variable, we require only one column in the design matrix (say, column k). We simply let x_{ik} be the value of the predictor variable associated with observation Y_i. We say that there is 1 *degree of freedom* (df) associated with this variable (corresponding to the one column used in the design matrix).

Consider the following linear model predicting selling price, Y_i, as a function of size of house, x_i:

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i,$$

where the ϵ_i's are iid $N(0, \sigma^2)$ random variables. The expected value of Y_i is then

$$\mathrm{E}[Y_i] = \beta_1 + \beta_2 x_i.$$

Any observed deviations from this relationship are assumed to be random fluctuations which have no systematic pattern.

Example 4: Categorical predictor variables (ANOVA)

Categorical predictor variables require a coding system (called *contrasts*) in order to be represented in the design matrix. We need $q - 1$ columns of the design matrix to represent a factor with q levels (and hence say that there are $q - 1$ df associated with this factor).

Consider the `hotdog` data set, which consists of observations on the number of calories in beef, "meat", and poultry hotdogs. The factor `MeatType` can take on 3 possible values. Let

$$x_{i2} = \begin{cases} 1, & i^{th} \text{ hotdog is meat} \\ 0, & \text{otherwise} \end{cases}$$

$$\text{and } x_{i3} = \begin{cases} 1, & i^{th} \text{ hotdog is poultry} \\ 0, & \text{otherwise} \end{cases} .$$

These are called the *treatment* contrasts.

The following linear regression model relates calories to meat type:

$$Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i.$$

We can then compute

$$\mathrm{E}[Y_i] = \begin{cases} \beta_1, & i^{th} \text{ hotdog is beef} \\ \beta_1 + \beta_2, & i^{th} \text{ hotdog is meat} \\ \beta_1 + \beta_3, & i^{th} \text{ hotdog is poultry} \end{cases} .$$

In other words, this model allows a different mean number of calories for each level of meat type. This model is called a *1-way ANOVA model*.

Question: How do we interpret β_2?

<u>More About Contrasts</u>

S-PLUS uses *Helmert* contrasts as its default coding of categorical predictor variables. For example, in the `hotdog` data set, the factor `MeatType` has 3 levels (beef, meat, and poultry). We therefore require $3 - 1 = 2$ predictor variables to code for this factor. The S-PLUS command `contrasts(hotdog$MeatType)` yields

	[,1]	[,2]
Beef	−1	−1
Meat	1	−1
Poultry	0	2

In other words, S-PLUS defines the 2 predictor variables as

$$x_{i1} = \begin{cases} -1, & i^{th} \text{ hotdog is beef} \\ 1, & i^{th} \text{ hotdog is meat} \\ 0, & i^{th} \text{ hotdog is poultry} \end{cases}$$

and

$$x_{i2} = \begin{cases} -1, & i^{th} \text{ hotdog is beef} \\ -1, & i^{th} \text{ hotdog is meat} \\ 2, & i^{th} \text{ hotdog is poultry} \end{cases} .$$

The default ordering of the levels of each factor is *alphabetical*. To change the order, we use the command `ordered`. For example, for the order Poultry, Meat, Beef, we'd type

```
hotdog$MeatType_ordered(hotdog$MeatType,c("Poultry","Meat","Beef"))
```

followed by a statement to redefine the contrasts based on this ordering, e.g.

```
contrasts(hotdog$MeatType)_contr.helmert(3)
```

Upon doing a linear regression of Calories (Y_i) on MeatType (using the original ordering), S-PLUS gives estimates of coefficients named `MeatType1` and `MeatType2`. These coefficients are equal to β_1 and β_2 in the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i.$$

To interpret the S-PLUS coefficients, we compute

$$\mathrm{E}[Y_i] = \begin{cases} \beta_0 - \beta_1 - \beta_2, & i^{th} \text{ hotdog is beef} \\ \beta_0 + \beta_1 - \beta_2, & i^{th} \text{ hotdog is meat} \\ \beta_0 + 2\beta_2, & i^{th} \text{ hotdog is poultry} \end{cases}.$$

The coefficients don't have a really easy interpretation. However, we can think of β_0 as the intercept, and of $-\beta_1 - \beta_2$, $\beta_1 - \beta_2$, and $2\beta_2$ as the deviations from this intercept for beef, meat, and poultry hotdogs, respectively.

More importantly, we can see that

$$\begin{aligned} 2\beta_1 &= \text{the mean difference in calories between a meat and beef hotdog} \\ \beta_1 + 3\beta_2 &= \text{the mean difference in calories between a poultry and beef hotdog} \\ -\beta_1 + 3\beta_2 &= \text{the mean difference in calories between a poultry and meat hotdog.} \end{aligned}$$

Therefore, in hypothesis tests about differences in the mean number of calories of different types of hotdogs,

$$\begin{array}{rll} \beta_1 = 0 & \text{implies} & \text{Meat and beef hotdogs have the same mean no. of calories} \\ \beta_1 = -3\beta_2 & \text{implies} & \text{Poultry and beef hotdogs have the same mean no. of calories} \\ \beta_1 = 3\beta_2 & \text{implies} & \text{Meat and poultry hotdogs have the same mean no. of calories.} \end{array}$$

And, if $\beta_1 = \beta_2 = 0$, then all three types of hotdogs have the same mean number of calories.

Estimation of Linear Models: Least-Squares

One way of estimating the unknown parameters $\boldsymbol{\beta}$ is find $\hat{\boldsymbol{\beta}}$ which minimizes the sum of squares of the residuals, S. Let $\mathbf{x}_i'$ be the i^{th} row of $\mathbf{X}$. Then

$$\begin{aligned} S &\equiv \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} (Y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 \\ &= (\mathbf{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\mathbf{Y} - \boldsymbol{X}\boldsymbol{\beta}). \end{aligned}$$

The minimum value of S occurs at $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, where $\mathbf{X}'$ is the transpose and $\mathbf{X}^{-1}$ is the inverse of $\mathbf{X}$.

The distribution of $\hat{\boldsymbol{\beta}}$ is multivariate normal with mean $\boldsymbol{\beta}$ and variance-covariance matrix $\boldsymbol{\Sigma} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. In other words, the diagonal entries of $\boldsymbol{\Sigma}$ give the variances of the estimates $\hat{\beta}_j$, and the off-diagonal entries give the covariances between $\hat{\beta}_j$ and $\hat{\beta}_k$. We can use this fact when forming confidence intervals (CIs) or doing hypothesis tests on $\boldsymbol{\beta}$.