Response and Predictor Variables

The EPA mileage data contains a number of variables and measurements for each vehicle:

- miles per gallon (mpg)
- car or truck?
- cubic inches displacement (cylinder volume)
- rhp rated horsepower
- mfr manufacturer
- city or highway?
- vehicle weight

Response and Predictor Variables

The EPA mileage data contains a number of variables and measurements for each vehicle:

- miles per gallon (mpg)
- car or truck?
- cubic inches displacement (cylinder volume)
- rhp rated horsepower
- mfr manufacturer
- city or highway?
- vehicle weight

Often, we are interested in *predicting* one variable, say mpg, from the others.

Response and Predictor Variables

- miles per gallon (mpg)
- car or truck?
- cubic inches displacement (cylinder volume)
- rhp rated horsepower
- mfr manufacturer
- city or highway?
- vehicle weight

In this situation, mpg is considered a *response* and the others are considered *predictors*

Response and Predictor Variables

- miles per gallon (mpg)
- car or truck?
- cubic inches displacement (cylinder volume)
- rhp rated horsepower
- mfr manufacturer
- city or highway?
- vehicle weight

In this situation, mpg is considered a *response* and the others are considered *predictors*

We want to find some way of using the predictors to estimate value of the response variable

Response and Predictor Variables

Of course, we already know the value of the response variable for the entries in our data matrix.

It might seem odd that we would try to predict something we already know.

Response and Predictor Variables

Of course, we already know the value of the response variable for the entries in our data matrix.

It might seem odd that we would try to predict something we already know.

The real value the model is to predict the response variable for combinations of the predictors that do not appear in our data.

Response and Predictor Variables

Of course, we already know the value of the response variable for the entries in our data matrix.

It might seem odd that we would try to predict something we already know.

The real value the model is to predict the response variable for combinations of the predictors that do not appear in our data.

For example, suppose we were interested in the effect of increasing the displacement of the engine in a certain vehicle by 40 cubic inches.

Response and Predictor Variables

Of course, we already know the value of the response variable for the entries in our data matrix.

It might seem odd that we would try to predict something we already know.

The real value the model is to predict the response variable for combinations of the predictors that do not appear in our data.

For example, suppose we were interested in the effect of increasing the displacement of the engine in a certain vehicle by 40 cubic inches.

If we increase the value of the corresponding predictor variable by 40, and keep the others the same, we can estimate the effect without actually having to build a modified vehicle.

Linear Models

The simplest type of mathematical model we can have is a *linear model*

Linear Models

The simplest type of mathematical model we can have is a *linear model*

A *linear model* estimates the response variable as a weighted average of the predictors.

Linear Models

The simplest type of mathematical model we can have is a *linear model*

A *linear model* estimates the response variable as a weighted average of the predictors.

If we label the response variable Y and the predictors $X_1, X_2, \ldots, X_n$, the general form of a linear model is:

$$Y = \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_n \cdot X_n$$

Linear Models

The simplest type of mathematical model we can have is a *linear model*

A *linear model* estimates the response variable as a weighted average of the predictors.

If we label the response variable Y and the predictors $X_1, X_2, \ldots, X_n$, the general form of a linear model is:

$$Y = \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_n \cdot X_n$$

The β values are constants that we need to assign values to.

Linear Models

The simplest type of mathematical model we can have is a *linear model*

A *linear model* estimates the response variable as a weighted average of the predictors.

If we label the response variable Y and the predictors $X_1, X_2, \ldots, X_n$, the general form of a linear model is:

$$Y = \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_n \cdot X_n$$

The β values are constants that we need to assign values to.

We choose the β values that "best" predict our measured responses from the corresponding predictors.

Linear Models

The term "best" needs some clarification.

Linear Models

The term "best" needs some clarification.

Generally with real data, it will be impossible to find values for the βs that exactly predict our observed response variables.

So the equation

$$Y = \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_n \cdot X_n$$

cannot possibly hold for all of the rows in our data table.

Linear Models

The term "best" needs some clarification.

Generally with real data, it will be impossible to find values for the βs that exactly predict our observed response variables.

So the equation

$$Y = \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_n \cdot X_n$$

cannot possibly hold for all of the rows in our data table.

Since equality is impossible, we try to make the difference between the predicted and actual response variables as small as possible:

$$\text{difference} = Y - (\beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_n \cdot X_n)$$

Linear Models

Actually we determine the β values that make the total of the *squares* of these differences as small as possible.

We use squares of the differences to prevent positive and negative differences from cancelling out.

Linear Models

Actually we determine the β values that make the total of the *squares* of these differences as small as possible.

We use squares of the differences to prevent positive and negative differences from cancelling out.

The term *linear least squares models* or just *linear models* refers to the collection of models based on this technique.

Linear Models

Actually we determine the β values that make the total of the *squares* of these differences as small as possible.

We use squares of the differences to prevent positive and negative differences from cancelling out.

The term *linear least squares models* or just *linear models* refers to the collection of models based on this technique.

The following techniques are special types of linear models analysis:

- Analysis of Variance
- Simple and Multiple Regression
- Analysis of Covariance

Linear Models

- Analysis of Variance
- Simple and Multiple Regression
- Analysis of Covariance

Often these are presented as entirely separate topics, but in fact they are all essentially the same.

Linear Models

- Analysis of Variance
- Simple and Multiple Regression
- Analysis of Covariance

Often these are presented as entirely separate topics, but in fact they are all essentially the same.

Each of them estimates the β values of a linear equation of the form

$$Y = \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_n \cdot X_n$$

Linear Models

- Analysis of Variance
- Simple and Multiple Regression
- Analysis of Covariance

Often these are presented as entirely separate topics, but in fact they are all essentially the same.

Each of them estimates the β values of a linear equation of the form

$$Y = \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_n \cdot X_n$$

The difference is in the nature of the predictor variables X_i (whether they are categorical or continuous).

Linear Models

The relationship between the classical types of analysis and the nature of the predictor variables is:

Classical name	Predictor variables
Analysis of Variance	All predictors are categorical
Regression	All predictors are continuous
Analysis of Covariance	Some categorical, some continuous

Linear Models

The relationship between the classical types of analysis and the nature of the predictor variables is:

Classical name	Predictor variables
Analysis of Variance	All predictors are categorical
Regression	All predictors are continuous
Analysis of Covariance	Some categorical, some continuous

Current statistical software reflects the unified view as most packages provide a *general linear model* routine that handles all three situations.

Linear Models

The relationship between the classical types of analysis and the nature of the predictor variables is:

Classical name	Predictor variables
Analysis of Variance	All predictors are categorical
Regression	All predictors are continuous
Analysis of Covariance	Some categorical, some continuous

Current statistical software reflects the unified view as most packages provide a *general linear model* routine that handles all three situations.

In fact, the set of equations the program needs to solve to find the β values is the same in all three cases.

Linear Models

Returning to our EPA mileage data example, we would probably categorize the variables as follows:

Variable	Type
mpg (miles per gallon)	response
c/t (car or truck)	categorical predictor
cid (displacement)	continuous predictor
rhp (horsepower)	continuous predictor
mfr (manufacturer)	categorical predictor
C/H (city/hwy)	categorical predictor
etw (weight)	continuous predictor
vpc (cylinders)	continuous predictor

Linear Models

Returning to our EPA mileage data example, we would probably categorize the variables as follows:

Variable	Type
mpg (miles per gallon)	response
c/t (car or truck)	categorical predictor
cid (displacement)	continuous predictor
rhp (horsepower)	continuous predictor
mfr (manufacturer)	categorical predictor
C/H (city/hwy)	categorical predictor
etw (weight)	continuous predictor
vpc (cylinders)	continuous predictor

A variable like vpc, which assumes only a few values (4,6,8) may be treated as either categorical or continuous.

Continuous Predictors

Generally there is one β_i parameter for each continuous predictor.

Continuous Predictors

Generally there is one β_i parameter for each continuous predictor.

If we made a table of the response and predictor values for the EPA data, with weight as the predictor, the data/response matrix might look like this:

mpg	weight
18.4	5400
22.1	4400
32.8	3300
17.1	6000
18.2	5600

Continuous Predictors

Actually, there would probably be one additional β known as the *intercept*,

$$Y = \beta_0 + \beta_1 \cdot X$$

Continuous Predictors

Actually, there would probably be one additional β known as the *intercept*,

$$Y = \beta_0 + \beta_1 \cdot X$$

mpg	weight	
18.4	1	5400
22.1	1	4400
32.8	1	3300
17.1	1	6000
18.2	1	5600

Continuous Predictors

Actually, there would probably be one additional β known as the *intercept*,

$$Y = \beta_0 + \beta_1 \cdot X$$

mpg	weight	
18.4	1	5400
22.1	1	4400
32.8	1	3300
17.1	1	6000
18.2	1	5600

The matrix of coefficients in the right hand box is called the *design matrix* for the model.

Continuous Predictors

The interpretation of the design matrix is that the numbers are multipliers for the β values:

Equation:
$18.4 = 1 \cdot \beta_0 + 5400 \cdot \beta_1$
$22.1 = 1 \cdot \beta_0 + 4400 \cdot \beta_1$
$32.8 = 1 \cdot \beta_0 + 3300 \cdot \beta_1$
$17.8 = 1 \cdot \beta_0 + 6000 \cdot \beta_1$
$18.2 = 1 \cdot \beta_0 + 5600 \cdot \beta_1$

Continuous Predictors

The interpretation of the design matrix is that the numbers are multipliers for the β values:

Equation:
$18.4 = 1 \cdot \beta_0 + 5400 \cdot \beta_1$
$22.1 = 1 \cdot \beta_0 + 4400 \cdot \beta_1$
$32.8 = 1 \cdot \beta_0 + 3300 \cdot \beta_1$
$17.8 = 1 \cdot \beta_0 + 6000 \cdot \beta_1$
$18.2 = 1 \cdot \beta_0 + 5600 \cdot \beta_1$

Each data observation produces one equation in this system.

Continuous Predictors

The interpretation of the design matrix is that the numbers are multipliers for the β values:

Equation:
$18.4 = 1 \cdot \beta_0 + 5400 \cdot \beta_1$
$22.1 = 1 \cdot \beta_0 + 4400 \cdot \beta_1$
$32.8 = 1 \cdot \beta_0 + 3300 \cdot \beta_1$
$17.8 = 1 \cdot \beta_0 + 6000 \cdot \beta_1$
$18.2 = 1 \cdot \beta_0 + 5600 \cdot \beta_1$

Each data observation produces one equation in this system.

The system almost never has an exact solution. The "best" solutions minimizes the squared differences between the predicted and actual response values.

Categorical Predictors

Categorical predictors have a separate β parameter for *each value* they assume.

Categorical Predictors

Categorical predictors have a separate β parameter for *each value* they assume.

For the categorical variable *car or truck*, there would be two β values, one for cars and one for trucks.

Categorical Predictors

Categorical predictors have a separate β parameter for *each value* they assume.

For the categorical variable *car or truck*, there would be two β values, one for cars and one for trucks.

The design matrix has one column for each of these two beta values. Suppose entries 1, 4, and 5 are trucks, and β_1 represents cars while β_2 represents trucks. The design matrix is:

mpg		car	truck
18.4	1	0	1
22.1	1	1	0
32.8	1	1	0
17.1	1	0	1
18.2	1	0	1

Categorical Predictors

The interpretation of the design matrix is that the numbers are multipliers for the β values:

Equation:
$18.4 = 1 \cdot \beta_0 + 0 \cdot \beta_1 + 1 \cdot \beta_2$
$22.1 = 1 \cdot \beta_0 + 1 \cdot \beta_1 + 0 \cdot \beta_2$
$32.8 = 1 \cdot \beta_0 + 1 \cdot \beta_1 + 0 \cdot \beta_2$
$17.8 = 1 \cdot \beta_0 + 0 \cdot \beta_1 + 1 \cdot \beta_2$
$18.2 = 1 \cdot \beta_0 + 0 \cdot \beta_1 + 1 \cdot \beta_2$

Categorical Predictors

The interpretation of the design matrix is that the numbers are multipliers for the β values:

Equation:
$18.4 = 1 \cdot \beta_0 + 0 \cdot \beta_1 + 1 \cdot \beta_2$
$22.1 = 1 \cdot \beta_0 + 1 \cdot \beta_1 + 0 \cdot \beta_2$
$32.8 = 1 \cdot \beta_0 + 1 \cdot \beta_1 + 0 \cdot \beta_2$
$17.8 = 1 \cdot \beta_0 + 0 \cdot \beta_1 + 1 \cdot \beta_2$
$18.2 = 1 \cdot \beta_0 + 0 \cdot \beta_1 + 1 \cdot \beta_2$

Again, each data observation produces one equation in this system.

Categorical Predictors

The interpretation of the design matrix is that the numbers are multipliers for the β values:

Equation:
$18.4 = 1 \cdot \beta_0 + 0 \cdot \beta_1 + 1 \cdot \beta_2$
$22.1 = 1 \cdot \beta_0 + 1 \cdot \beta_1 + 0 \cdot \beta_2$
$32.8 = 1 \cdot \beta_0 + 1 \cdot \beta_1 + 0 \cdot \beta_2$
$17.8 = 1 \cdot \beta_0 + 0 \cdot \beta_1 + 1 \cdot \beta_2$
$18.2 = 1 \cdot \beta_0 + 0 \cdot \beta_1 + 1 \cdot \beta_2$

Again, each data observation produces one equation in this system.

The "best" solutions minimizes the squared differences between the predicted and actual response values.