Multiple Linear Regression
Multiple linear regression

Recall that: a regression model describes how a dependent variable (or response) $Y$ is affected, on average, by one or more independent variables (or factors, or covariates) $X_1, X_2, \ldots, X_k$.

The general equation is

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k.$$  

I shall sometimes write $E(Y)$ as $E(Y|X_1, X_2, \ldots, X_k)$, to emphasize that $E(Y)$ changes with the values of the terms $X_1, X_2, \ldots, X_k$:

$$E(Y|X_1, X_2, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k.$$
As always, we can write

\[ \epsilon = Y - E(Y), \]

or

\[ Y = E(Y) + \epsilon, \]

where the \textit{random error} \( \epsilon \) has expected value zero:

\[ E(\epsilon) = E(\epsilon|X_1, X_2, \ldots, X_k) = 0. \]

So the general equation can also be written

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon. \]
Each term on the right hand side may be an independent variable, or a function of one or more independent variables.

For instance,

$$E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2$$

has two *terms* on the right hand side (not counting the intercept $\beta_0$), but only one *independent variable*.

We write it in the general form as

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

with $X_1 = X$ and $X_2 = X^2$. 

Multiple Linear Regression
Quantitative and qualitative variables

Some variables are measured quantities (i.e., on an interval or ratio scale), and are called *quantitative*. Others are the result of classification into categories (i.e., on a nominal or ordinal scale), and are called *qualitative*.

Some terms may be *functions* of independent variables:
- distance and distance$^2$, or sine and cosine of (month/12).

The simplest case is when all variables are quantitative, and no mathematical functions appear: the *first-order* model.
Interpreting the parameters for first-order models

\( \beta_0 \) is still called the intercept, but now its interpretation is the expected value of \( Y \) when all independent variables are zero:

\[
\beta_0 = E(Y|X_1 = 0, X_2 = 0, \ldots, X_k = 0).
\]

For \( 1 \leq i \leq k \), \( \beta_i \) measures the change in \( E(Y) \) as \( X_i \) increases by 1 with all the other independent variables held fixed.
First-order model: $E(Y) = 1 + 2X_1 + X_2$ for $X_2 = 0, 1, 2$
Example: Grandfather clocks

Dependence of auction price of antique clocks on their age, and the number of bidders at the auction.

- Data for 32 clocks.

Get the data and plot them:

The first-order model is

\[ E(\text{PRICE}) = \beta_0 + \beta_1 \times \text{AGE} + \beta_2 \times \text{NUMBIDS}. \]
Grandfather clocks data: Scatter plots

1. Scatter plot of AGE vs. PRICE
2. Scatter plot of NUMBIDS vs. PRICE
Fitting the model: least squares

As in the case $k = 1$, the most common way of fitting a multiple regression model is by *least squares*.

That is, find $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ so that

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots \hat{\beta}_k X_k$$

minimizes

$$\text{SSE} = \sum(Y_i - \hat{Y}_i)^2.$$ 

As noted earlier, other criteria such as $\sum |Y_i - \hat{Y}_i|$ are sometimes used instead.
Calculus leads to \( k + 1 \) linear equations in the \( k + 1 \) estimates \( \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k \).

These equations are always consistent; that is, they always have a solution when \( n > k \).

Usually, they are also non-singular; that is, the solution is unique.

- If they are singular, we can find a unique solution by either imposing constraints on the parameters or leaving out redundant variables.
Notice:

We will review matrix algebra and use it to solve least squares and derive $E(SSE)$ after having finished Chapter 4.
Model assumptions

No assumptions are needed to find least squares estimates.

To use them to make statistical inferences, we need these assumptions:

- The random errors $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent and have common variance $\sigma^2$;
- For small sample validity, the random errors are normally distributed, at least approximately.
The error variance, $\sigma^2$, is an important measure of model utility.

If $\sigma^2 = 0$:
- Random errors are all 0,
- Model parameters will be estimated without error (zero-length CI),
- $E(Y)$ will be estimated without error (zero-length CI),
- Prediction of $Y$ will have no error (zero-length CI).

If $\sigma^2$ is large:
- Random errors are large (on average in absolute values),
- Model parameters will be estimated with large error (wider CI),
- $E(Y)$ will be estimated with large error (wider CI),
- Prediction of $Y$ will also have large error (wider CI).
Estimation of $\sigma^2$

As before, we estimate $\sigma^2$ using

\[ SSE = \sum (Y_i - \hat{Y}_i)^2. \]

We can show that

\[ E[SSE] = (n - k - 1)\sigma^2, \]

where $k + 1$ is the number of $\beta$s in the model, so the unbiased estimator is

\[ s^2 = \frac{SSE}{df} = \frac{SSE}{n - k - 1} = \frac{SSE}{n - k - 1} \]

$s^2$ is called *Mean Squared Errors* or *MSE*. 

Multiple Linear Regression
Fit the model for Grandfather clocks data

```r
# load in data
setwd("~/Dropbox/teaching/2015Fall/R_datasets/Exercises&Examples")
load("GFCLOCKS.Rdata")

fit = lm(PRICE~AGE + NUMBIDS, data=GFCLOCKS) # linear fit

summary(fit) # display results
```
Testing the utility of a model

Usually, the first test is an overall test of the model:

- \( H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0. \)
- \( H_a : \) at least one \( \beta_i \neq 0. \)

\( H_0 \) asserts that none of the independent variables affects \( Y \); if this hypothesis is not rejected, the model is worthless.

- For instance, its predictions perform no better than \( \bar{y} \).

The test statistic is usually denoted \( F \).
**F** test statistic

\[ F = \frac{(SS_{yy} - SSE)/k}{SSE/(n - k - 1)} = \frac{\text{Mean Square (Model)}}{\text{MSE}}. \]

- **F** is the ratio of the *explained* variability to the *unexplained* variability.
- If \( H_0 \) is true, \( F \) has a \( F \)-distribution with \( k \) and \( n - k - 1 \) degrees of freedom.
- Reject \( H_0 \) if \( F \) is large, i.e., \( F > F_\alpha \), based on \( k \) and \( n - k - 1 \) degrees of freedom.
Rejection region for the global $F$-test

0

$F_{\alpha}$

Rejection region

$F$
F-test in R: the Grandfather clocks data

```r
# load in data
setwd("~/Dropbox/teaching/2015Fall/R_datasets/Exercises&Examples")
load("GFCLOCKS.Rdata")

fit = lm(PRICE~AGE + NUMBIDS, data=GFCLOCKS) #linear fit
summary(fit) #display results
```
\( F \)-test results: the Grandfather clocks data

- \( F \)-statistic: 120.2.
- Degrees of freedom for the \( F \)-test: 2 and 29.
- P-value for the \( F \)-test: \( 9.2 \times 10^{-15} \).
- Conclusion: Reject \( H_0 \) in favor of \( H_1 \).
Testing $\beta$’s

Same as testing in simple linear regression models.

For the grandfathers clocks data, both the age of clocks and the number of bidders are significant predictors.
Multiple coefficient of determination/ R-squared

\[ R^2 = 1 - \frac{\text{SSE}}{\text{SS}_{yy}} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \]

The interpretation of \( R^2 \) is still the fraction of variance “explained” by the regression model.

It measures the correlation between the dependent variable \( Y \) and the independent variables \emph{jointly}. 

Adjusted R-squared

Because the regression model is adapted to the sample data, it tends to explain more variance in the sample data than it will in new data.

Rewrite:

$$1 - R^2 = \frac{\text{SSE}}{\text{SS}_{yy}} = \frac{1}{n} \sum (y_i - \hat{y}_i)^2$$

$$= \frac{1}{n} \sum (y_i - \bar{y})^2$$

Numerator and denominator are *biased* estimators of variance.
Replace $\frac{1}{n}$ with the multipliers that give unbiased variance estimators:

$$\frac{1}{n-p} \sum (y_i - \hat{y}_i)^2 \quad \frac{1}{n-1} \sum (y_i - \bar{y})^2,$$

where $k + 1$ is the number of estimated $\beta$s.

This defines the *adjusted* $R$-squared:

$$R_a^2 = 1 - \frac{1}{n-k-1} \sum (y_i - \hat{y}_i)^2 \quad \frac{1}{n-1} \sum (y_i - \bar{y})^2$$

$$= 1 - \frac{n-1}{n-k-1} \times \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}.$$

$R_a^2 < R^2$, and for a poorly fitting model you may even find $R_a^2 < 0!$