Chapter 11

Multiple Regression

Introduction to the Practice of STATISTICS EIGHTH EDITION

Moore / McCabe / Craig

Lecture Presentation Slides

Chapter 11
Multiple Regression

11.1 Inference for Multiple Regression

11.2 A Case Study

11.1 Inference for Multiple Regression

- Population multiple regression model
- Data for multiple regression
- Multiple linear regression model
- Confidence intervals and significance tests
- Squared multiple correlation R^2

Population Multiple Regression Equation

Up to this point, we have considered in detail the linear regression model in which the mean response, μ_y, is related to one explanatory variable x:

$$\mu_y = \beta_0 + \beta_1 x$$

Usually, more complex linear models are needed in practical situations.

There are many problems in which a knowledge of more than one explanatory variable is necessary in order to obtain a better understanding and better prediction of a particular response.

In multiple regression, the response variable y depends on p explanatory variables $x_1, x_2, ..., x_p$:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Data for Multiple Regression

The data for a simple linear regression problem consist of *n* observations (x_i, y_i) of two variables.

Data for multiple linear regression consist of the value of a response variable *y* and *p* explanatory variables $(x_1, x_2, \ldots, x_p)$ on each of *n* cases.

We write the data and enter them into software in the form:

	Variables				
Case	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	…	$\boldsymbol{x_p}$	$\boldsymbol{y}$
1	x_{11}	x_{12}	…	x_{1p}	y_1
2	x_{21}	x_{22}	…	x_{2p}	y_2
…	…	…	…	…	…
n	x_{n1}	x_{n2}	…	x_{np}	y_n

Multiple Linear Regression Model

The **statistical model for multiple linear regression** is

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i}$$

for $i = 1, 2, \ldots, n$.

The **mean response μ_y** is a linear function of the explanatory variables:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

The **deviations ε_i** are independent and Normally distributed $N(0, \sigma)$.

The parameters of the model are $\beta_0, \beta_1, \ldots, \beta_p$ and σ.

The coefficient β_i ($i = 1,\ldots,p$) has the following interpretation: It represents the average change in the response when the variable x_i increases by one unit and *all other* x *variables are held constant.*

Estimation of the Parameters

Select a random sample of n individuals on which p + 1 variables ($x_1,\ldots,x_p,y$) are measured. The least-squares regression method chooses $b_0, b_1, \ldots, b_p$ to minimize the sum of squared deviations $(y_i - \hat{y}_i)^2$, where

$$\hat{y}_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip}$$

As with simple linear regression, the constant b_0 is the y-intercept.

- The regression coefficients ($b_1, \ldots, b_p$) reflect the unique association of each independent variable with the y variable. They are analogous to the slope in simple regression.
- The parameter σ^2 measures the variability of the responses about the population response mean. The estimator of σ^2 is:

$$s^2 = \frac{\sum e_i^2}{n-p-1} = \frac{\sum (y_i - \hat{y}_i)^2}{n-p-1}$$

Confidence Interval for β_j

Estimating the regression parameters β_0, … ,β_j, … ,β_p is a case of one-sample inference with unknown population variance.

We rely on the *t* distribution, with ***n* – *p* – 1 degrees of freedom.**

A **level *C* confidence interval for $\boldsymbol{\beta_j}$** is:

$$b_j \pm t^* SE_{b_j}$$

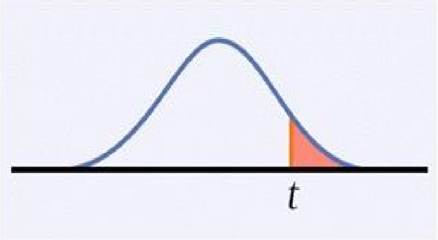
Where SE_{b_j} is the standard error of b_j and t^* is the *t* critical for the $t(n - p - 1)$ distribution with area *C* between $-t^*$ and t^*.

Significance Test for β_j

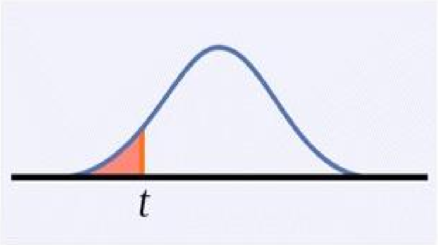
To test the hypothesis H_0: $\beta_j = 0$ versus a one- or two-sided alternative, we calculate the *t* statistic $t = b_j / SE_{b_j}$, which has the *t* ($n - p - 1$) distribution when H_0 is true. The *P*-value of the test is found in the usual way.

Note: *Software typically provides two-sided P-values.*

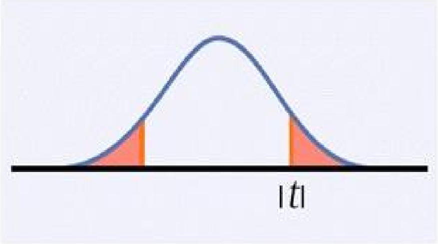
H_a: $\beta_j > 0$ is $P(T \geq t)$



H_a: $\beta_j < 0$ is $P(T \leq t)$



H_a: $\beta_j \neq 0$ is $2P(T \geq |t|)$



Significance Test for β_j

Suppose we test H_0: $\beta_j = 0$ for each j and find that none of the p tests is significant.

Should we then conclude that none of the explanatory variables is related to the response?

No, we should not!

When we fail to reject H_0: $\beta_j = 0$, this means that we probably don't need x_j in the model with all the other variables.

So, failure to reject all such hypotheses merely means that it's safe to throw away at least one of the variables.

Further analysis must be done to see which subset of variables provides the best model.

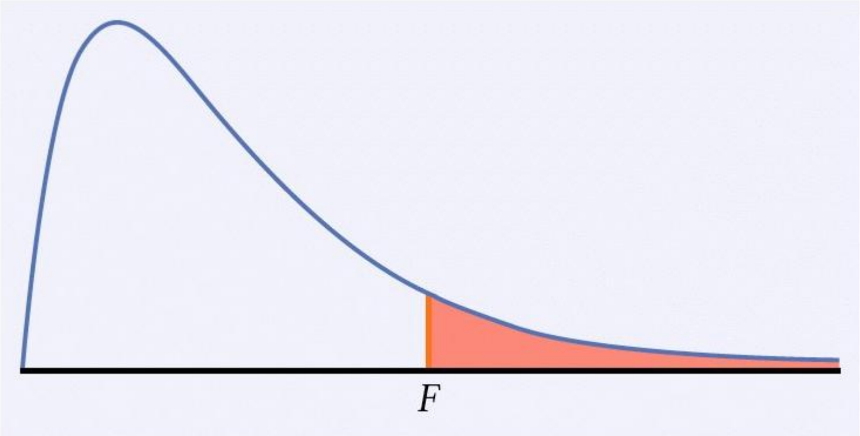
ANOVA F-test for Multiple Regression

In multiple regression, the ANOVA F statistic tests the hypotheses

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0 \quad \text{versus } H_a\text{: at least one } \beta_j \neq 0$$

by computing the F statistic $\boldsymbol{F}$ **= MSM / MSE.**

When H_0 is true, F follows the $F(p, n - p - 1)$ distribution. The P-value is $P(F \geq f)$.



A significant *P*-value doesn't mean that all *p* explanatory variables have a significant influence on *y*—only that at least one does.

ANOVA Table for Multiple Regression

Source	Sum of squares SS	df	Mean square MS	*F*	*P*-value
Model	$\sum(\hat{y}_i - \bar{y})^2$	p	MSM=SSM/DFM	MSM/MSE	Tail area above *F*
Error	$\sum(y_i - \hat{y}_i)^2$	$n - p - 1$	MSE=SSE/DFE		
Total	$\sum(y_i - \bar{y})^2$	$n - 1$			

SSM = model sum of squares

SSE = error sum of squares

SST = total sum of squares

SST = SSM + SSE

DFM = p DFE = $n - p - 1$ DFT = $n - 1$

DFT = DFM + DFE

Squared Multiple Correlation R^2

Just as with simple linear regression, ***R^2*, the squared multiple correlation,** is the proportion of the variation in the response variable *y* that is explained by the model.

$$\boxed{R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{SSM}{SST}}$$

In the particular case of multiple linear regression, the model is all *p* explanatory variables taken together.

The square root of R^2, called the **multiple correlation coefficient,** is the correlation between the observations and the predicted values.

Chapter 11
Multiple Regression

11.1 Inference for Multiple Regression

11.2 A Case Study