Chapter 6: Multiple Regression I

Goals

- Multiple regression models
- Estimation of regression coefficients
- Fitted values and residuals
- Analysis of variance results
- Inference about regression parameters
- Estimation of mean response and prediction of new observation
- Diagnostics and remedial measures
- An example: multiple regression with two predictor variables

6.1 Multiple regression models

- Still have single response variable *Y*
- Now have multiple explanatory variables
- Examples:
  - Blood Pressure vs Age, Weight, Diet, Smoking, Fitness Level
  - Traffic Count vs Time, Location, Population, Month
- Goal: There is a total amount of variation in *Y* (SSTO). We want to explain as much of this variation as possible using a linear model and our predictor variables

First-order multiple linear regression model

Two or more independent variables are used to estimate 1 dependent variable.

$$Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\ldots+\beta_{p-1}X_{i,p-1}+\varepsilon_i$$

Notes:

1) ε_i~independent $N(0,\sigma^2)$
2) $\beta_0, \beta_1, \ldots,\beta_{p-1}$ are parameters
3) $X_{i1},\ldots,X_{i,p-1}$ are known constants. The second subscript on X_{ij} denotes the j^{th} independent variable.
4) i=1,…n, represents the ith trial

Notes:

1) Often when we want to just refer to the first, second,… independent variables, the i subscript is dropped from X_{ij}. Thus, independent variable #1 is X_1, independent variable #2 is X_2,…

2) $Y_i=\beta_0+\beta_1X_{i1}+\ldots+\beta_{p-1}X_{i,p-1}+\varepsilon_i$ is called a "first-order" model since the model is linear in the independent (predictor, explanatory) variables.

3) The term "linear model" refers to the fact that the model is linear in the parameters. For example, $Y_i = \beta_0 + \beta_1X_{i1} + \beta_1^2X_{i2} + \varepsilon_i$ is not a multiple LINEAR regression model.

4) The response function is a **hyperplane**, which is a plane in more than more dimensions.

5) The coefficient on the j^{th} independent variable is β_j. This measures the effect X_j has on Y with the remaining variables in the model held constant.

6) **Qualitative variables** (variables not measured on a numerical scale) can be used in the multiple regression model. For example, let $X_{ij}=1$ to denote female, $X_{ij}=0$ to denote male. More will be done with qualitative variables in Chapter 8.

7) **Polynomial regression models** contain higher than first order terms in the model. For example, $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \varepsilon_i$. More will be done with polynomial regression models in Chapter 8.

8) **Transformed variables** can be used just as in simple linear regression. For example, $\log(Y_i)$ can be taken to be the dependent variable.

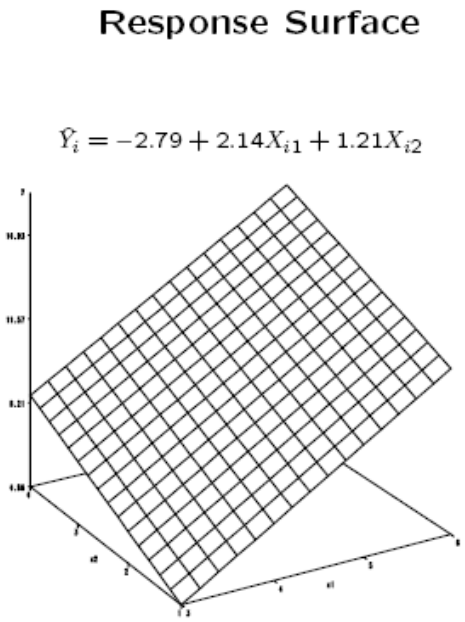
9) If the effect of one independent variable on the dependent variable depends on another independent variable, **interactions** between independent variables can be included in the multiple regression model. For example, $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$. More will be done with interactions in Chapter 8.

10) Combination of cases

Example:

(1) First-order model with two predictors

$Y_i=\beta_0+ \beta_1 X_{i1}+ \beta_2 X_{i2}+\varepsilon_i$



- β_0 is the intercept and β_1 and β_2 are the regression coefficients

- β_1 describes change in mean response per unit increase in X_1 when X_2 is held constant
- β_2 describes change in mean response per unit increase in X_2 when X_1 is held constant
- Variables X_1 and X_2 are additive. Value of X_1 does not affect the change due to X_2. There is no interaction.

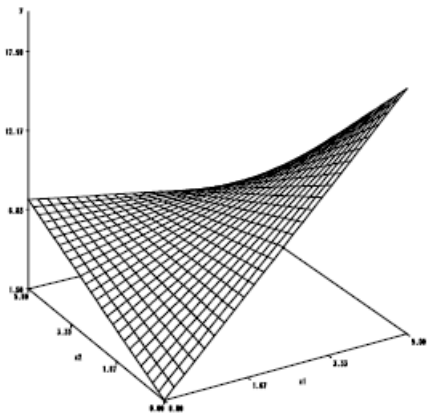
(2) Interaction model

$Y_i=\beta_0+ \beta_1X_{i1}+ \beta_2X_{i2}+ \beta_3X_{i1}X_{i2}+\varepsilon_i$

- Rate of change due to one variable affected by the other.

Interaction Response Surface

$\hat{Y}_i = 1.5 + 3.2X_{i1} + 1.2X_{i2} - .75X_{i1}X_{i2}$



6.2 General linear regression model in matrix terms

Note that the $Y_i=\beta_0+\beta_1X_{i1}+\ldots+\beta_{p-1}X_{i,p-1}+\varepsilon_i$ for i=1,…,n can be written as:

$$
\begin{aligned}
Y_1&=\beta_0+\beta_1X_{11}+\ldots+\beta_{p-1}X_{1,p-1}+\varepsilon_1\\
Y_2&=\beta_0+\beta_1X_{21}+\ldots+\beta_{p-1}X_{2,p-1}+\varepsilon_2\\
Y_3&=\beta_0+\beta_1X_{31}+\ldots+\beta_{p-1}X_{3,p-1}+\varepsilon_3\\
&\vdots\\
Y_n&=\beta_0+\beta_1X_{n1}+\ldots+\beta_{p-1}X_{n,p-1}+\varepsilon_n
\end{aligned}
$$

Let

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X}=\begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \text{ and } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Then $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

Remember that $\boldsymbol{\varepsilon}$ has mean $\mathbf{0}$ and covariance matrix

$$\begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}$$

since the ε_i are independent.

Note that $E(\mathbf{Y}) = E(\mathbf{X\beta} + \boldsymbol{\varepsilon}) = E(\mathbf{X\beta}) = \mathbf{X\beta}$ since $E(\boldsymbol{\varepsilon}) = \mathbf{0}$; and the variance-covariance matrix of $\mathbf{Y}$ is the same as that of $\boldsymbol{\varepsilon}$:

$\boldsymbol{\sigma}^2\{Y\} = \sigma^2\mathbf{I}$

6.3 Estimation of regression coefficients (β_j's)

Parameter estimates are found using the least squares method.

From Chapter 1: The least squares method tries to find the $\hat{\beta}_0$ and $\hat{\beta}_1$ such that SSE = $\Sigma(Y-\hat{Y})^2 = \Sigma(\text{residual})^2$ is minimized. Formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ are derived using calculus.

For multiple regression, the least squares method minimizes SSE = $\Sigma(Y-\hat{Y})^2$ again; however, $\hat{Y}$ is now $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \ldots + \hat{\beta}_{p-1} X_{i,p-1}$

As shown in Section 5.10, the least squares estimators are $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. This holds true for multiple regression with

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix} \text{ and } \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix}$$

Properties of the least squares estimators in simple linear regression (such as: unbiased estimators and minimum variance among unbiased estimators) hold true for multiple linear regression.

6.4 Fitted values and residuals

Let $\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \vdots \\ \hat{Y}_n \end{bmatrix}$.

Then $\hat{\mathbf{Y}} = \mathbf{Xb} = \mathbf{X(X' \ X)^{-1}X'Y} = \mathbf{HY}$,

And the residuals are

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} - \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}$$

Hat matrix: $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{HY}$ where $\mathbf{H}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the hat matrix (We will use this in Chapter 9 to measure the influence of observations on the estimated regression line).

Covariance matrix of the residuals:
$\sigma^2\{\mathbf{e}\} = \sigma^2(\mathbf{I}-\mathbf{H})$ is estimated by $s^2\{\mathbf{e}\} = MSE(\mathbf{I}-\mathbf{H})$.

6.5 Analysis of variance results

ANOVA Table

Source of Variation	SS	df	MS	F
Regression	SSR = $\mathbf{b}'\mathbf{X}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y}$	p-1	MSR = SSR/(p-1)	F^* = MSR/MSE
Error	SSE= $\mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$	n-p	MSE = SSE/(n-p)	
Total	SSTO = $\mathbf{Y}'\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y}$	n-1		

Notes:

1) SSR has p-1 degrees of freedom. Before with 1 independent variable, there was only 1 degree of freedom.
2) SSE has n-p degrees of freedom. Before with 1 independent variable, there were n-2 degrees of freedom.
3) The F test for regression relation:

H_o: $\beta_1=\beta_2=\ldots=\beta_{p-1}=0$
H_a: At least one $\beta\neq 0$

$F^* \sim F(p-1, n-p)$

Reject H_o if $F^* > F(1-\alpha, p-1, n-p)$

Coefficient of determination – R^2

R^2 = SSR/SSTO = (SSTO-SSE)/SSTO

R^2 has the same interpretation as before, but with respect to p-1 independent variables.
$100*R^2$% of the variation in Y can be explained by using the independent variables to estimate Y

Notes:

1) Use R^2 as a measure of fit when the sample size is substantially larger than the number of variables in the model; otherwise, R^2 may be artificially high.
2) As more variables are added to the model, R^2 will always increase even if the additional variables do a poor job of estimating Y.

Solution: Use the Adjusted R^2

$$R_a^2 = 1 - \frac{n-1}{n-p}(1-R^2)$$

- R_a^2 can not be forced to increase as R^2 can be by adding variables.
- R_a^2 adjusts for having "nonsense" variables (additional variables that do a poor job of estimating Y) added to the model that make R^2 increase.
- USE R_a^2 INSTEAD OF R^2
- The interpretation of R_a^2 is about the same as R^2
- $R_a^2 \leq R^2$
- R_a^2 can be less than 0

Coefficient of multiple correlation R is the positive square root of R^2

6.6 Inferences about regression parameters

The least squares and maximum likelihood estimators in b are unbiased:
E{b} = β.

Covariance matrix of **b**:

$$\sigma^2\{\mathbf{b}\} = \sigma^2\{(\mathbf{X'X})^{-1}\mathbf{X'Y}\} = (\mathbf{X'X})^{-1}\mathbf{X'}\sigma^2\{\mathbf{Y}\}[(\mathbf{X'X})^{-1}\mathbf{X'}]' = \sigma^2(\mathbf{X'X})^{-1}$$

Estimated covariance matrix of **b**:

$$s^2\{\mathbf{b}\} = MSE(\mathbf{X'X})^{-1}$$

To determine if the j^{th} independent variable is helpful in predicting the dependent variable, a t-test for $\beta_j=0$ can be conducted.

t-test for β_j in

$$E(Y_i)=\beta_0 + \beta_1 X_{i1} + \ldots + \beta_j X_{ij} + \ldots + \beta_{p-1} X_{i,p-1}$$

1) State H_o and H_a
   H_o: $\beta_j=0$
   H_a: $\beta_j \neq 0$
2) Test statistic: $t^* = \frac{b_j}{s\{b_j\}}$
3) Decision Rule: Conclude H_A if $|t^*| \geq t(1-\alpha/2; n-p)$, otherwise conclude H_0
4) Or use the *P*-value: $2P(t(n-p) > |t^*|)$. Conclude H_A if p-value $\leq \alpha$, otherwise conclude H_0

Interval estimation of β_j

$$b_j \pm t(1-\alpha/2; n-p)s\{b_j\}$$

Joint inferences

The Bonferroni joint confidence intervals can be used toe estimate several regression coefficients simultaneously. If *g* parameters are to be estimated jointly, the confidence limits with family confidence coefficient 1-α are:

$$b_j \pm t(1-\alpha/(2g); n-p)s\{b_j\};$$

Tests concerning subsets of the regression parameters are discussed in Chapter 7.

6.7 Estimation of mean response and prediction of new observation

Interval estimation for $E(Y_h)$ at X_h

Estimate the mean value of Y_h for $\mathbf{X}_h=(1, X_{h1}, X_{h2}, \ldots, X_{h,p-1})'$

As shown in Section 5.13, the estimated variance used in the C.I. for $E(Y_h)$ is

$$\sigma^2\{\hat{Y}_h\} = \sigma^2(\mathbf{X}_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h), \qquad s^2\{\hat{Y}_h\} = MSE(\mathbf{X}_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)$$

The (1-α)100% C.I. for $E(Y_h)$ is

$$\hat{Y}_h \pm t(1-\alpha/2; n-p)s\{\hat{Y}_h\}$$

Notice the degrees of freedom are n-p.

Interval estimation for $Y_{h(new)}$ at X_h

As shown in Section 5.13, the estimated variance used in the P.I. for $Y_{h(new)}$ is

$$s^2\{pred\} = MSE(1+\mathbf{X}_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)$$

The (1-α)100% P.I. for $Y_{h(new)}$ is

$$\hat{Y}_h \pm t(\alpha/2; n-p)s\{pred\}$$

Notice the degrees of freedom are n-p.

Confidence region for regression surface

$\hat{Y}_h \pm Ws\{\hat{Y}_h\}$, where $W^2 = pF(1-\alpha; p, n-p)$

The confidence coefficients 1-□ provides assurance that the region contains the entire regression surface over all combinations of values of the X variables.

Simultaneous confidence intervals

If g different C.I.s or P.I.s were desired with a family confidence coefficient of at least (1-α)100%, the $t(1-\alpha/2, n-p)$ part of the intervals would change to $t(1-\alpha/(2g), n-p)$. This is an application of the Bonferroni procedure.

6.8 Diagnostics and remedial measures

All diagnostic procedures done in Chapter 3 can be done for multiple regression.

1) Residual plots
2) Correlation test for Normality
3) Brown-Forsythe test for constancy of error variance
4) F test for lack of fit: it can be carried over to test whether the multiple regression response function is an appropriate response surface. The difference is that c is the number of groups with distinct sets of levels for the X varialbles.

Notes:

1) To check the linearity of the regression function, plots of e_i vs. X_{ij} should be done for j=1,…,p-1 (i.e., plots for each independent variable).
2) The plot of e_i vs. $\hat{Y}_i$ and e_i vs. X_{ij}
3) More diagnostic procedures are discussed in Chapters 9 and 10.

A scatter plot matrix is a matrix of scatter plots showing the relationship between pairs of variables.

This is often a tool to identify which independent variables are correlated to the dependent variable. In addition, this plot helps identify independent variables that have a strong correlation with each other (more on this discussed later - multicollinearity).

A complement to the scatter plot matrix is the correlation matrix.

6.9 An example – multiple regression with two predictor variables

An example: Dwaine Studios, Inc., operates portrait studios in 21 cities of medium size. These studios specialize in portraits of children. The company is considering an expansion into other cities of medium size and wishes to investigate whether sales (Y) in a community can be

predicted from the number of persons aged 16 or younger in the community (X1) and the per capita disposable personal income in the community (X2). Data on these variables for the most recent year for the 21 cities in which Dwaine Studios is now operating are available. Sales are expressed in thousands of dollars; the number of persons aged 16 or younger is expressed in thousands of persons; and per capita disposable personal income is expressed in thousands of dollars.

Use the first-order regression model $Y_i=\beta_0+ \beta_1X_{i1}+ \beta_2X_{i2}+\varepsilon_i$. Check the scatter plot matrix to see if is reasonable.

1) Estimate the regression function and interpret the parameters.

2) Fitted values and residuals

3) Residual plots. Findings?

4) Normal probability plot. Findings?

5) ANOVA table

6) Test whether sales are related to target population and per capita disposable income.

7) Coefficient of multiple determination and adjusted coefficient of multiple determination.

8) Joint confidence intervals for the regression parameters.

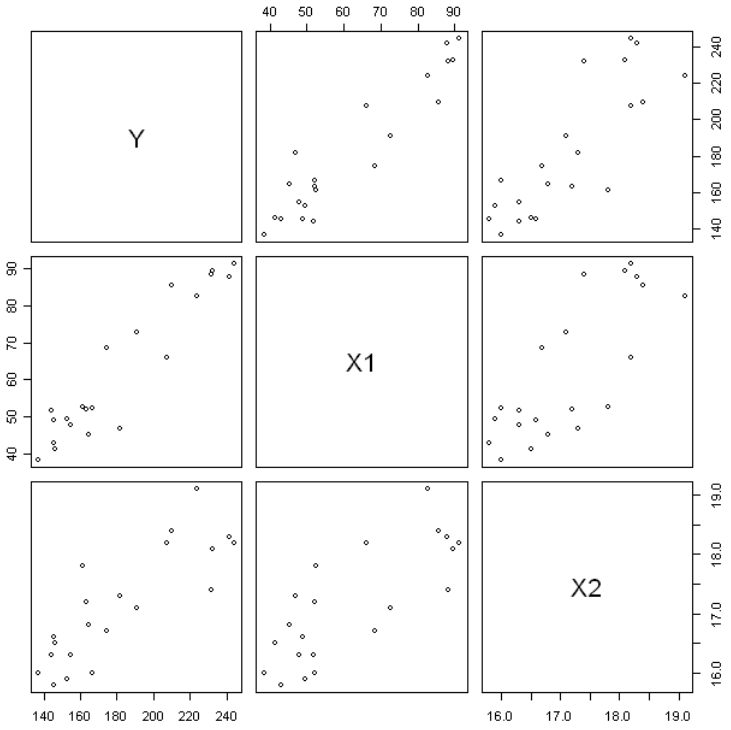
9) Dwaine Studios would like to estimate expected (mean) sales in cities with target population 65.4 thousands of persons aged 16 years or younger and per capita disposable income 17.6 thousand dollars with a 95% confidence interval.

10) Dwaine Studios as part of a possible expansion program would like to predict sales for two new cities, with the following characteristics:

	City A	City B
X_{h1}	65.4	53.1
X_{h2}	17.6	17.7

Prediction intervals with a 90% family confidence coefficient are desired.

Scatter matrix plot from R output



Residual plots