STAT2401 Analysis of Experiments

Lecture Week 7 Dr Darfiana Nur

THE UNIVERSITY OF WESTERN AUSTRALIA

Aims of Lecture Week 7

- **Aim 1 Matrix approach to Linear Regression**

(Sheather Ch 5.2, Moore et al Ch 11)

1.1 SLR in a matrix form

1.2 Least Squares method

- **Aim 2 Multiple Linear Regression (MLR)** (Sheather Ch 5.2, Moore et al Ch 11)

2.1 The model - MLR

2.2 Structure

- **Aim 3 Parameter Estimation – MLR** (Sheather Ch 5.2, Moore et al Ch 11)

3.1 The detail

3.2 Some examples

Aim 1 Matrix approach to linear regression

- **Vector-matrix notation** simplifies presentation of least squares regression
- Once the problem is written and solved in matrix terms, solution can be applied to **any linear regression problem**
- Notation:
  - **Vectors** represented by lowercase bold letters: $\boldsymbol{y}, \hat{\boldsymbol{e}}, \boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}$
  - **Matrices** represented by uppercase bold letters: $\boldsymbol{X}, \boldsymbol{I}$
  - **Scalars** represented by italicized letters: y_i, x_i
- Will be using basic vector-matrix operations: multiplication, transpose, inverse

We use a matrix to contain the explanatory variable(s)

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}$$

The matrix **X** contains a column of 1's and a column for the single explanatory variable

It is an (nx2) matrix. n rows, 2 columns

Addition of matrices

$$A = \begin{pmatrix} 1 & 2 \\ 4 & 3 \end{pmatrix}; B = \begin{pmatrix} 5 & 2 \\ 1 & -3 \end{pmatrix}$$

Just add the corresponding elements together

$$A + B = \begin{pmatrix} 1+5 & 2+2 \\ 4+1 & 3+(-3) \end{pmatrix} = \begin{pmatrix} 6 & 4 \\ 5 & 0 \end{pmatrix}$$

We cannot add matrices of different dimensions

Multiplication of matrices

For example

$$A = \begin{pmatrix} 1 & 2 \\ 4 & 3 \end{pmatrix}; B = \begin{pmatrix} 5 \\ 1 \end{pmatrix}$$

What about *BA*?
this is undefined!

$$AB = \begin{pmatrix} 1 & 2 \\ 4 & 3 \end{pmatrix}\begin{pmatrix} 5 \\ 1 \end{pmatrix} = \begin{pmatrix} 1\times 5 + 2\times 1 \\ 4\times 5 + 3\times 1 \end{pmatrix} = \begin{pmatrix} 7 \\ 23 \end{pmatrix}$$

dimensions **2** x 2 2 x **1** **2** x **1**

these must be equal

For example

$$A = \begin{pmatrix} 3 & -1 & 7 \\ 2 & 4 & 1 \end{pmatrix}; B = \begin{pmatrix} 5 \\ 1 \end{pmatrix}$$

A is a (2 x **3**) matrix and **B** is a (**2** x 1) vector. As $3 \neq 2$ we **cannot** multiply these together. We can't add them either.

$$AB = \begin{pmatrix} 3 & -1 & 7 \\ 2 & 4 & 1 \end{pmatrix}\begin{pmatrix} 5 \\ 1 \end{pmatrix} = \begin{pmatrix} 3\times 5 - 1\times 1 + 7\times ? \\ 2\times 5 + 4\times 1 + 1\times ? \end{pmatrix} = \begin{pmatrix} ? \\ ? \end{pmatrix}$$

Aim 1.1 Simple Linear Regression in a matrix form

Consider the linear model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i ; Var(\varepsilon_i) = \sigma^2$$

We can write it in matrix notation as follows. Let,

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}; \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

2x1 vector

We have used vectors to contain the observations, the residuals and the regression parameters.

We use a matrix to contain the explanatory variable(s)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i ; Var(\varepsilon_i) = \sigma^2$$

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ : & : \\ 1 & X_n \end{pmatrix}$$

The matrix **X** contains a column of 1's and a column for the single explanatory variable

It is an (*n*x2) matrix. *n* rows, 2 columns

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ : \\ y_n \end{pmatrix} ; \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ : \\ \varepsilon_n \end{pmatrix} ; \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$\mathbf{X\beta} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ : & : \\ 1 & X_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ : \\ \beta_0 + \beta_1 X_n \end{pmatrix}$$

The simple linear model in matrix notation is

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

Often we need sums of squares terms in regression

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ : \\ \varepsilon_n \end{pmatrix}; \boldsymbol{\varepsilon}' = (\varepsilon_1 \quad \varepsilon_2 \quad .. \quad \varepsilon_n)$$

$$\underset{1xn}{\boldsymbol{\varepsilon}'}\underset{nx1}{\boldsymbol{\varepsilon}} = \varepsilon_1{}^2 + \varepsilon_2{}^2 + .. + \varepsilon_n{}^2 = \underset{1x1}{\sum \varepsilon_i^2}$$

For any vector *a*, *a'a* represents the sum of the squares of the elements of *a*. It is a 1x1 scalar number

Means and variances

We know that

$$E(Y_1 \mid X) = \beta_0 + \beta_1 X_1$$

$$E(Y_2 \mid X) = \beta_0 + \beta_1 X_2$$

$$\vdots$$

$$E(Y_n \mid X) = \beta_0 + \beta_1 X_n$$

This is equivalent to

$$\begin{aligned} E(\mathbf{Y} \mid \mathbf{X}) &= E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= E(\mathbf{X}\boldsymbol{\beta}) + E(\boldsymbol{\varepsilon}) \\ &= \mathbf{X}\boldsymbol{\beta} \end{aligned}$$

Variances and covariances

- For any set of observations Y we have $\hat{Y}$

$$Var(Y) = E[(Y - E(Y))^2] \approx \tfrac{1}{n-c} \sum (Y_i - \hat{Y}_i)^2$$

where c (or p) is the number of parameters estimated.

- To find Covariance between any two variables Y and X is defined as follows

$$Cov(Y, X) = E[(Y - E(Y))(X - E(X))]$$

$$= E(XY) - E(X)E(Y)$$

$$\approx \tfrac{1}{n} \sum (Y_i - \overline{Y}_i)(X_i - \overline{X}_i) = \tfrac{1}{n} \left(\sum X_i Y_i - n\overline{X}\overline{Y} \right)$$

Vectors

For vectors, variances and covariances are a little different. Consider the residual vector ε. The assumptions of the linear model are that

$$E(\varepsilon_i) = 0; Var(\varepsilon_i) = \sigma^2; Cov(\varepsilon_i, \varepsilon_j) = 0$$

In matrix form we can write this as

$$E(\boldsymbol{\varepsilon}) = \mathbf{0}; Var(\boldsymbol{\varepsilon}) = \sigma^2 I_n$$

For a vector, the term *Var* refers to the matrix containing the variance of each element of the vector plus all the possible covariances between elements in the vector

Example: For a vector ε of dimension 2x1

$$E(\boldsymbol{\varepsilon}) = \begin{pmatrix} E(\varepsilon_1) \\ E(\varepsilon_2) \end{pmatrix}; Var(\boldsymbol{\varepsilon}) = \begin{pmatrix} Var(\varepsilon_1) & Cov(\varepsilon_1, \varepsilon_2) \\ Cov(\varepsilon_2, \varepsilon_2) & Var(\varepsilon_2) \end{pmatrix}$$

For the linear model

$$E(\boldsymbol{\varepsilon}) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}; Var(\boldsymbol{\varepsilon}) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$$

For general *n*

$$E(\boldsymbol{\varepsilon}) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}; Var(\boldsymbol{\varepsilon}) = \sigma^2 I_n = \sigma^2 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & .. & 1 \end{pmatrix}$$

So, for any vector **Y**

$$Var(\mathbf{Y}) = E[(\mathbf{Y} - E(\mathbf{Y}))(\mathbf{Y} - E(\mathbf{Y}))']$$

$$Var(a\mathbf{Y}) = a^2 Var(\mathbf{Y})$$

$$Var(\mathbf{AY}) = \mathbf{A} Var(\mathbf{Y}) \mathbf{A}'$$

$$E(\mathbf{AY}) = \mathbf{A} E(\mathbf{Y})$$

$$Cov(\mathbf{X}, \mathbf{Y}) = E[(\mathbf{X} - E(\mathbf{X}))'(\mathbf{Y} - E(\mathbf{Y}))]$$

Aim 1.2 Least squares estimation

- To estimate the vector β we must minimise the *SSE*

$$SSE = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

- Recall that to find the LSE we have to solve the normal equations:

$$\sum Y_i = n\beta_0 + \beta_1 \sum X_i$$

$$\sum X_i Y_i = \beta_0 \sum X_i + \beta_1 \sum X_i^{\,2}$$

- It is not hard to show that, for the SLR model

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix}; \mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix}$$

- These terms are exactly the ones needed to solve the normal equations. In fact the normal equations are equivalent to writing

$$\begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix} = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^{2} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

OR $$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

Least squares estimates

- Solving this matrix equation for β is straightforward

$$\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \hat{\boldsymbol{\beta}}$$

- These are exactly the same least squares estimators from before but now we don't have to worry about a lot of messy detail.

- We can see that the least squares estimates are just a linear combination of the observations, **Y**

- **Unbiased estimates**

$$E\left(\hat{\boldsymbol{\beta}}\right) = E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y})$$
$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y})$$
$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$
$$= \boldsymbol{\beta}$$

The LS estimates are unbiased.

- **Predicted values and Residuals**

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}; \hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}}$$
$$= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \mathbf{H}\mathbf{Y}\end{aligned}$$

- Here we explicitly see that the predicted values are just a linear combination of the observations **Y**

- The matrix **H** is called the hat matrix. It transforms the observations **Y** into the predicted values. It has the property that **HH = H.**

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} &= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}\end{aligned}$$

These are the sort of things that a computer can easily be programmed to do

- To find the variances of the parameter estimates

$$
\begin{aligned}
Var(\hat{\boldsymbol{\beta}}) &= Var((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) = Var(\mathbf{A}\mathbf{Y}) \\
&= \mathbf{A}\,Var(\mathbf{Y})\mathbf{A}' \\
&= \sigma^2 \mathbf{A}\mathbf{A}' \\
&= \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}
$$

- This matrix will contain the variances exactly as before, but also gives the term $Cov(\hat{\beta}_0, \hat{\beta}_1)$

We don’t need to know the exact form of this term,
only that it can be extracted from the matrix above and that
we can calculate this matrix if we need to.

Confidence interval for the regression line

We know that
$$Var(\hat{Y} \mid X_0) = \sigma^2 \left(\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum (X_i - \overline{X})^2} \right)$$

For a given point we know that X_0

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_0 = \begin{pmatrix} 1 & X_0 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$
$$= \mathbf{X}_0' \hat{\boldsymbol{\beta}}$$

So

$$Var(\hat{Y} \mid X_0) = Var(\mathbf{X}_0' \hat{\beta})$$
$$= \mathbf{X}_0' Var(\hat{\beta}) X_0$$
$$= \sigma^2 \mathbf{X}_0' (\mathbf{X}'\mathbf{X})^{-1} X_0$$

Again this is the same expression but in matrix terms

Prediction and forecasting individual observations

- We know that

$$Var(Y \mid X_0) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum (X_i - \overline{X})^2} \right)$$

- Again this is the same expression as before but in matrix terms

$$\begin{aligned} Var(Y \mid X_0) &= Var(\mathbf{X_0'}\hat{\beta} + \hat{\varepsilon}) \\ &= \mathbf{X_0'} Var(\hat{\beta}) X_0 + \sigma^2 \\ &= \sigma^2 [1 + \mathbf{X_0'}(\mathbf{X'X})^{-1} X_0] \end{aligned}$$

Thinking in terms of matrices

- Write down the Normal Equations for regression:
- Write these as a matrix equation:
- Consider how each element of the matrix and vector that don't involve $\underline{\beta}$ can be written as an inner product:
- What vectors are involved in these inner products?
- Define a matrix X and a vector $\underline{Y}$ so that you can write the Normal Equations as $X^T X B = X^T \underline{Y}$:
- Use the same X and $\underline{Y}$ to write out the original equations relating responses to covariates:

Aim 2 Multiple regression

- For many real data sets, we wish to examine the relationship between a response and a range of explanatory variables.

- When predicting fire damage, we may predict better if we consider not just distance from fire station but also
  - the income of the family
  - how much the contents are insured for
  - whether the house has a smoke alarm
  - do the occupants own a fire extinguisher

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 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 $x_1, x_2, \ldots x_p$:

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

Data for Multiple Regression

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

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

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

	Variables				
Case	x_1	x_2	…	x_p	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

Aim 2.1 Multiple Linear Regression Model

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

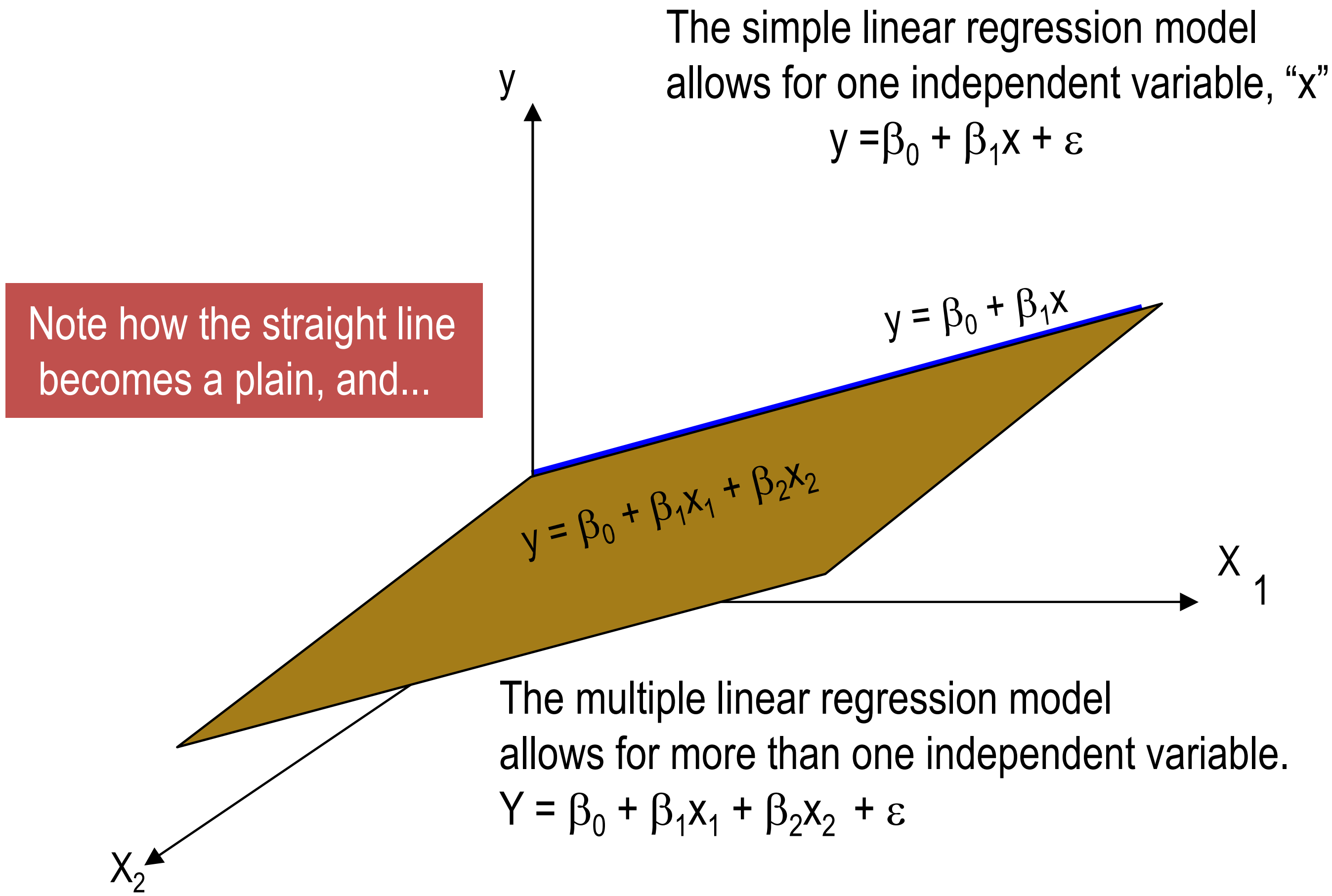
$$\boldsymbol{y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i}$$

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.*



The simple linear regression model
allows for one independent variable, “x”
$y = \beta_0 + \beta_1 x + \varepsilon$
Note how the straight line
becomes a plain, and...
y
$y = \beta_0 + \beta_1 x$
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
X_1
X_2
The multiple linear regression model
allows for more than one independent variable.
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

- **Required conditions for the error variable ε**
  - The error ε have mean equal to zero and a constant variance σ^2 (independent of any value of *X*). σ^2 is unknown.
  - The errors are independent of each other.

    (no pattern in the residual plot)

Multiple linear regression

- In simple linear regression, the mean function is modelled as
$$E(Y|x) = \beta_0 + \beta_1 x_1$$
- It is common for the response y to be influenced by more than one explanatory variable, so we add additional explanatory variables:
$$E(Y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$
- For example, we might go from $\beta_0 + \beta_1 x_1$ to $\beta_0 + \beta_1 x_1 + \beta_2 x_2$, i.e., add the variable x_2 in order to explain variability in Y that is not already explained by x_1

The Multiple Linear Regression Model

- **Matrix formulation in general**

$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \ldots + \beta_p X_{1p} + E_1$

$Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \ldots + \beta_p X_{2p} + E_2$

...

$Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \ldots + \beta_p X_{np} + E_n$

- To summarise, form nx1 vectors: $Y' = (Y_1, \ldots, Y_n)$, $X_j' = (x_{1j}, \ldots, x_{nj})$, $E' = (E_1, \ldots, E_n)$, $1' = (1, \ldots, 1)$.
- Write parameters as a (p+1)x1 vector: $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$
- Then the set of equations above become

$Y = \beta_0 1 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + E = [1, X_1, X_2, \ldots, X_p]\, \beta + E$

- Write matrix $X = [1, X_1, X_2, \ldots, X_p]'$; this is nx(p+1), and has one column for each of the different predictor variables.

Aim 2.2 The whole model:

$$Y = X\beta + E$$

- We assume n>p (more observations than predictors) so that the set of solutions of equations has no solution.

(Recall from linear algebra:
if n<p, possibly infinitely many solutions;
if n=p, a solution, if it exists, is unique;
if n>p, no solution.)

- The method of least squares minimises the sum of squares of errors.

Aim 3 Parameter estimation

- 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

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

Properties of least squares estimates: MLR

- $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$ and we have now have p predictors, with

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

- The covariance matrix of the LS estimates is

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$$

and as before, we estimate σ^2 from RSS, i.e.,

$$s^2 = \frac{RSS}{n-p-1} = \frac{1}{n-p-1}\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$$

- Hence, for carrying out a t-test for testing $H_0{:}\,\beta_i = 0$, we use

$$\frac{\hat{\beta}_i - 0}{\mathrm{se}(\hat{\beta}_i)} \sim t_{n-p-1}$$

- We can obtain $\mathrm{se}(\hat{\beta}_i)$ as the square root of the ith diagonal element of $\mathrm{var}(\widehat{\boldsymbol{\beta}})$

Aim 3.1 Parameter estimation: MLR

- If we have p predictors, we can write that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$, and the least squares estimate is
$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$
- Hence, the fitted values can be written as $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, or $\widehat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, and the matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is known as the 'hat' matrix*
- Residuals are $\hat{\boldsymbol{e}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$, and RSS can be written as
$$RSS = (\boldsymbol{y} - \widehat{\boldsymbol{y}})'(\boldsymbol{y} - \widehat{\boldsymbol{y}})$$
and as before, we estimate σ^2 from RSS, i.e.,
$$s^2 = \frac{RSS}{n-p-1} = \frac{1}{n-p-1}\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$$
- Note that the number of degrees of freedom is $n - p - 1$

Confidence Interval for β_j

- Estimating the regression parameters $\beta_0, \ldots, \beta_j, \ldots, \beta_p$ 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 β_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^*.

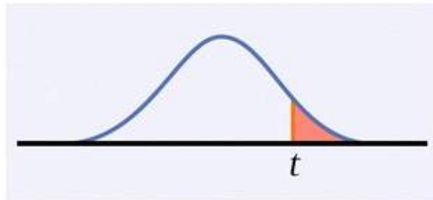
Significance Test for β_j , j=0,1,…,p

- To test the hypothesis H_0: $\beta_j = 0$ versus a one- or two-sided alternative, we calculate the t statistic

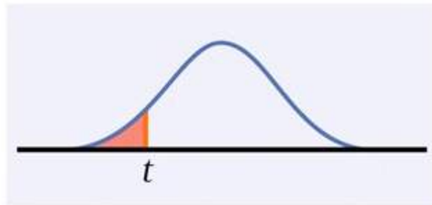
$$t = b_j / SE_{b_j} \;\sim\; t\,(n - p - 1)$$

distribution when H_0 is true. The P-value of the test is found in the usual way.

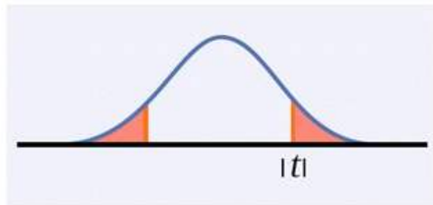
H_a: $\beta_j > 0$ is $P(T \geq t)$



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



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



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

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 do not need x_j in the model with all the other variables.

- So, failure to reject all such hypotheses merely means that it is 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.**

Estimate of σ^2

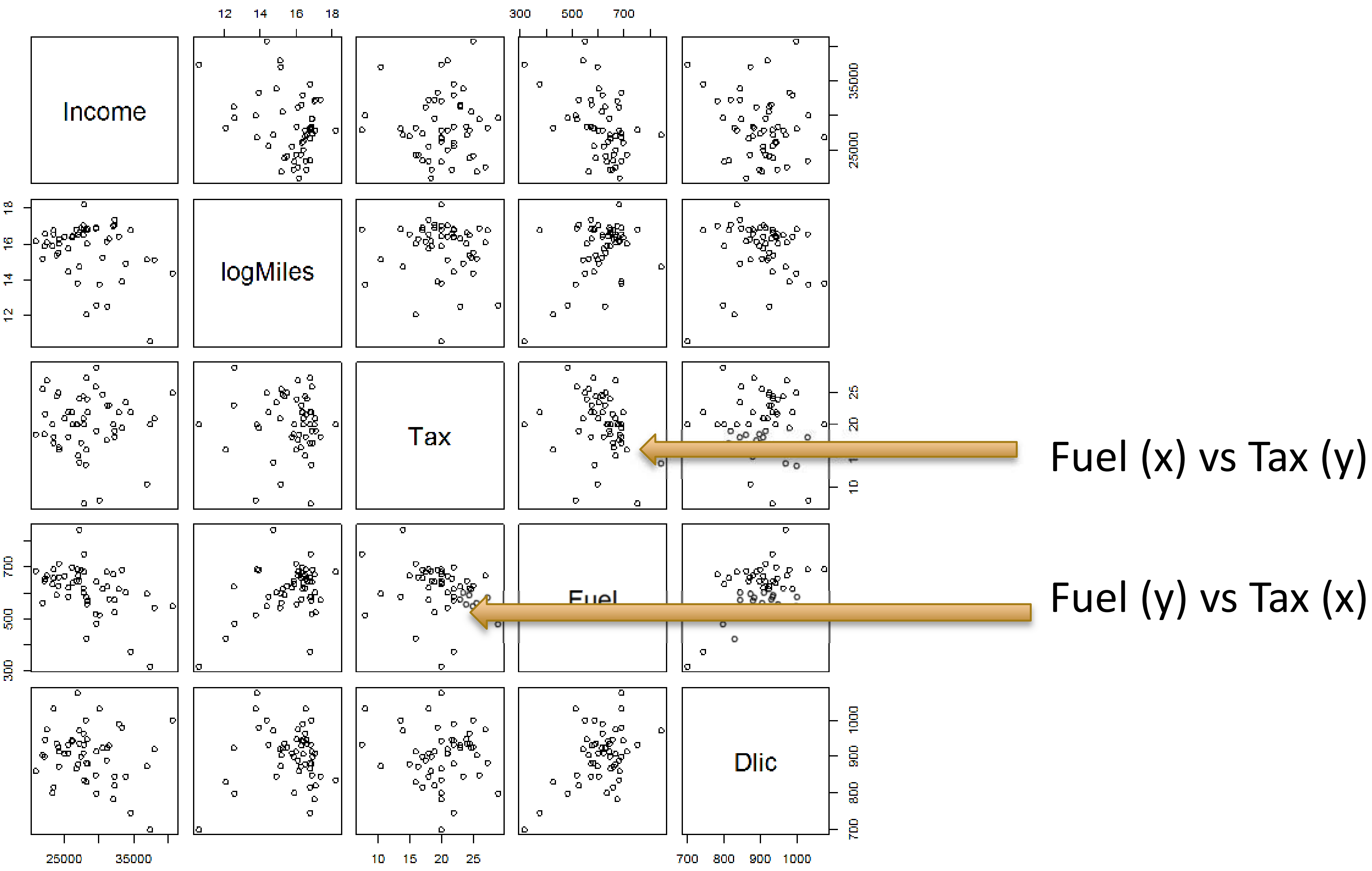
- The covariance matrix of the LS estimates is
$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$$
and as before, we estimate σ^2 using s^2
- Hence, for carrying out a t-test for testing $H_0: \beta_i = 0$, we use
$$\frac{\hat{\beta}_i - \beta_i}{\mathrm{se}(\hat{\beta}_i)} \sim t_{n-p-1}$$

Aim 3.2 Example 1: Fuel consumption in US states

- **Objective:** to understand how **fuel consumption** varies across the 50 US states and DC, and particular, to understand the effect on fuel consumption of state gasoline tax
- Note transformations of explanatory variables

`Income`	Average personal income for 2000
`logMiles`	log2 of miles of Federal highways
`Tax`	State gasoline tax (cents per gallon)
`Fuel`	Fuel sold per thousand licensed drivers
`Dlic`	Licensed drivers per thousand people

Scatterplot matrix



Scatterplot matrix – notes

- Scatterplot matrix is a 2D array of scatterplots
- Each plot is relevant to a particular one-predictor regression of the variable on the vertical axis, given the variable on the horizontal axis
- For example, if we were regressing fuel consumption on tax, we might produce the plot in the 4th row and 3rd column and then proceed to fit a linear model
  - `Fuel` decreases as `Tax` increases, but lots of variability!
  - Similar qualitative judgments about each of the regressions of Fuel on other variable

Scatterplot matrix – notes

- What information can (and can't) we glean from the scatterplot matrix?
  - The marginal (individual) relationships between the response and each of the variables are not sufficient to understand the **joint** relationship between the response and the predictors
  - We also need to take into account the relationship **between** the predictors
  - Can view pairwise relationships between predictors in other cells
    - Some weak, some strong relationships between predictors, e.g., `logMiles/Fuel` and `logMiles/Dlic`

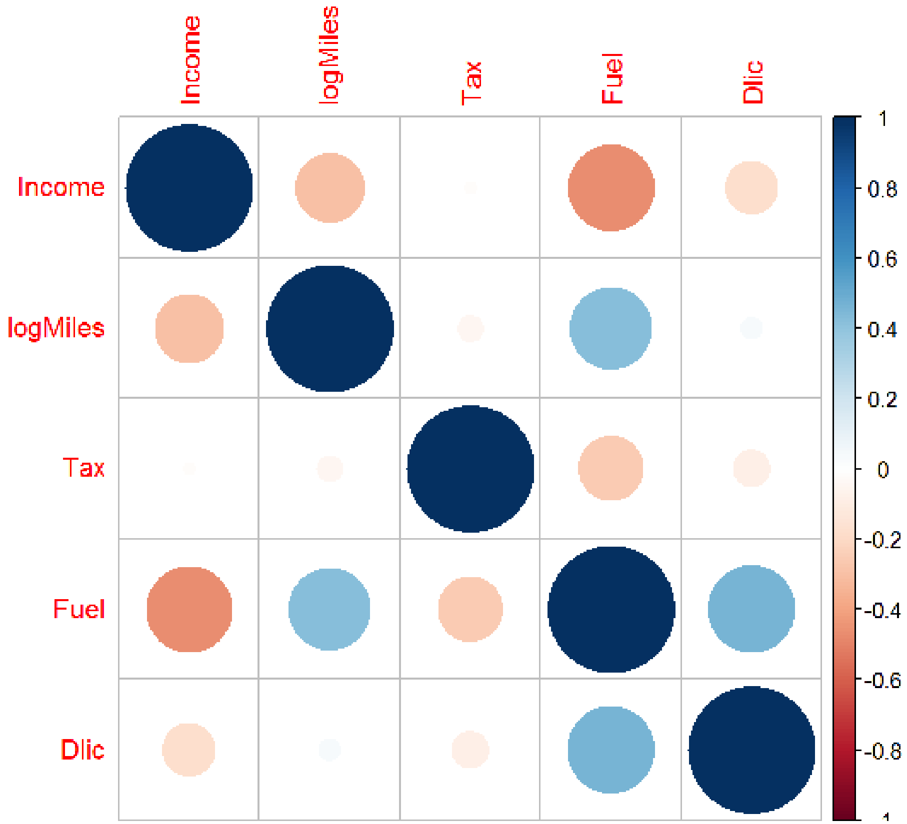
Additional exploratory analyses

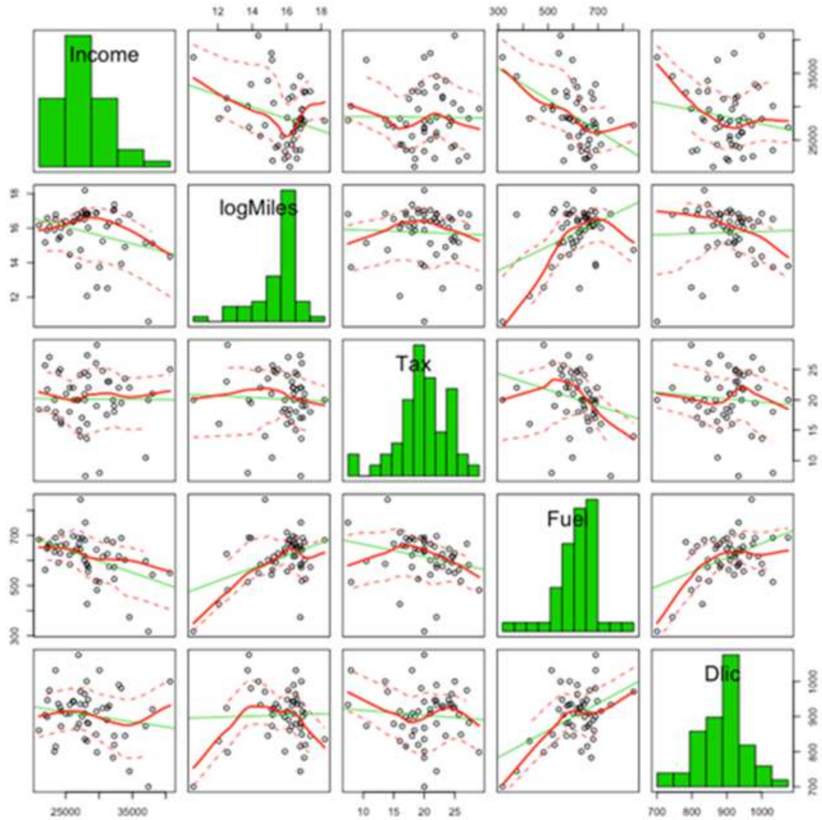
- Univariate summaries, multivariate summaries such as correlation matrix (caution!):

	Income	logMiles	Tax	Fuel	Dlic
Income	1.0000	-0.2959	-0.0107	-0.4644	-0.1760
logMiles	-0.2959	1.0000	-0.0437	0.4220	0.0306
Tax	-0.0107	-0.0437	1.0000	-0.2594	-0.0858
Fuel	-0.4644	0.4220	-0.2594	1.0000	0.4685
Dlic	-0.1760	0.0306	-0.0858	0.4685	1.0000

In R: round(cor(Fuel2001), 4)

Additional exploratory analyses






```
require(corrplot)
corrplot(cor(Fuel2001))
```

Multiple linear regression

- Consider the model
  $$E(Fuel \mid X) = \beta_0 + \beta_1 Tax + \beta_2 Dlic + \beta_3 Income + \beta_4 logMiles$$
- Number of explanatory variables is $p = 4$, but including the intercept, there are $p + 1 = 5$ coefficients to estimate
- The 'X-matrix' will be 51×5
- *R* syntax for fitting MLR:

```
Fuel.lm1 <- lm(Fuel ~ Tax + Dlic + Income + logMiles, data =
Fuel2001)
```

  - The first argument to lm is a 'formula' object, corresponding to the linear model that we're fitting
  - If the data frame doesn't contain any other predictors, then we can use `Fuel.lm1 <- lm(Fuel ~ ., data = Fuel2001)`

Model summary

```
> summary(Fuel.lm1)


Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 154.192845 194.906161   0.791 0.432938
Tax          -4.227983   2.030121  -2.083 0.042873 *
Dlic          0.471871   0.128513   3.672 0.000626 ***
Income       -0.006135   0.002194  -2.797 0.007508 **
logMiles     18.545275   6.472174   2.865 0.006259 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 64.89 on 46 degrees of freedom
Multiple R-squared:  0.5105,   Adjusted R-squared:  0.4679
F-statistic: 11.99 on 4 and 46 DF,  p-value: 9.331e-07
```

Tax: $\widehat{\boldsymbol{\beta}}\mathbf{1}$=-4.228, se($\widehat{\boldsymbol{\beta}}\mathbf{1}$)=2.030, t=-2.083; pval=0.042

Dlic: $\widehat{\boldsymbol{\beta}}\mathbf{2}$=0.472, se($\widehat{\boldsymbol{\beta}}\mathbf{2}$)=0.126, t=3.672; pval=0.000626

ETC

Parameter estimates

- Recall that $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$
- In *R*, we first need to construct the 'X'-matrix:

```
> X <- as.matrix(Fuel2001[, -4]) # Remove fuel consumption
> X <- cbind(1, X)
> head(X)
   1 Income logMiles  Tax      Dlic
AL 1  23471 16.52711 18.0 1031.3801
AK 1  30064 13.73429  8.0 1031.6411
AZ 1  25578 15.75356 18.0  908.5972
AR 1  22257 16.58244 21.7  946.5706
CA 1  32275 17.36471 18.0  844.7033
CO 1  32949 16.38960 22.0  989.6062
```

Parameter estimates

- $(\boldsymbol{X}'\boldsymbol{X})$ is `t(X) %*% X`
- $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is `solve(t(X) %*% X)`
- $(\boldsymbol{X}'\boldsymbol{y})$ is `t(X)%*% Fuel2001[, 4] # response 4th column`
- Putting it all together, we get

```
> betahat <- solve(t(X) %*% X) %*% t(X)%*% Fuel2001[, 4]
```

```
> round(betahat, 6)
                [,1]
          154.192845
Income     -0.006135
logMiles   18.545275
Tax        -4.227983
Dlic        0.471871
```

Covariance matrix of $\widehat{\boldsymbol{\beta}}$

- The covariance matrix of the LS estimates is
$$\operatorname{var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$$
where we replace σ^2 by its estimate s^2
- It is a $(p+1) \times (p+1)$ matrix whose **diagonal elements** give us the variances of the coefficient estimates

```
> print(solve(t(X) %*% X), digits = 4) ##   This is (X'X)^-1
                          Income    logMiles        Tax        Dlic
          9.022e+00 -5.981e-05 -1.932e-01 -2.852e-02 -4.080e-03
Income   -5.981e-05  1.143e-09  1.000e-06  4.263e-08  1.189e-08(X^' X)^(-1)
logMiles -1.932e-01  1.000e-06  9.948e-03  1.602e-04  5.402e-06
Tax      -2.852e-02  4.263e-08  1.602e-04  9.788e-04  5.599e-06
Dlic     -4.080e-03  1.189e-08  5.402e-06  5.599e-06  3.922e-06
```

Covariance matrix of $\widehat{\beta}$

- The summary table gives us an estimate of

$$s = \sqrt{RSS/(n-p-1)}$$

- But we can also extract it from the summary object as

```
> s <- summary(Fuel.lm1)$sigma
> s
[1] 64.89122
```

Putting it all together

```
> round(s * sqrt(diag(solve(t(X) %*% X))), 6) ## Std. Error
               Income   logMiles        Tax       Dlic
194.906161   0.002194   6.472174   2.030121   0.128513
```

Example 2: Menu Pricing in a New Italian Restaurant in New York City

- (Sheather Ch 1) This example highlights the use of multiple regression in a practical business setting.
- The aims of the restaurant are to provide the highest quality Italian food utilizing state-of-the art décor while setting a new standard for high-quality service in Manhattan. The data are in the form of the average of customer views on
  - Y = Price = the price (in $US) of dinner (including one drink & a tip)
  - X_1 = Food = customer rating of the food (out of 30)
  - X_2 = Décor = customer rating of the decor (out of 30)
  - X_3 = Service = customer rating of the service (out of 30)
  - X_4 = East = dummy variable = 1 (0) if the restaurant is east (west) of Fifth Avenue

Example 2: Menu Pricing in a New Italian Restaurant in New York City

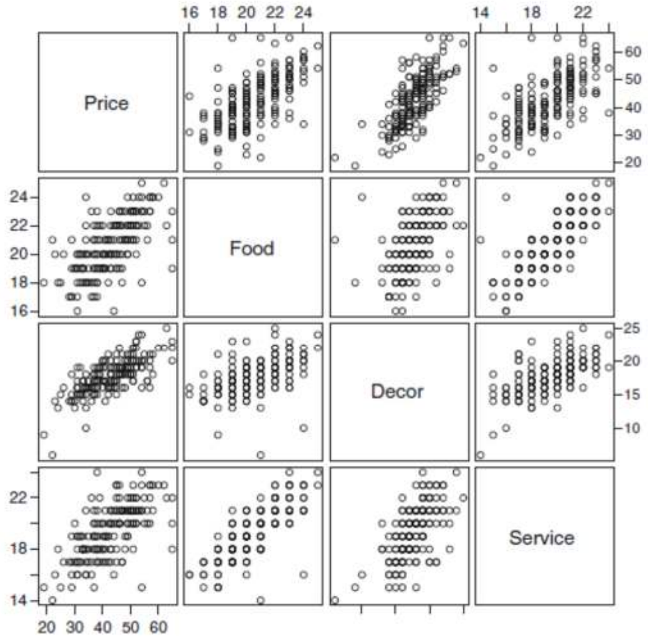


Figure 1.5 Matrix plot of Price, Food, I

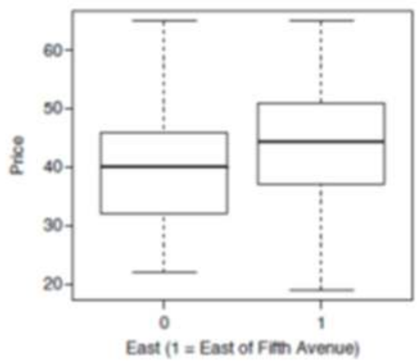


Figure 1.6 Box plots of Price for the two levels of the dummy variable East

- In particular you have been asked to:
- Develop a regression model that directly predicts the price of dinner (in dollars) using a subset or all of the four potential predictor variables listed above.
- Determine which of the predictor variables Food, Décor and Service has the largest estimated effect on Price? Is this effect also the most statistically significant?

Example 2: Menu Pricing in a New Italian Restaurant in New York City

- To predict price, we consider the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + E$$

- At this point we shall assume that all the necessary assumptions hold. In particular, we shall assume that the model is a valid model for the data.
- We shall check these assumptions for this example later and along with identify any outliers.

```
lm(formula = Price ~ Food + Decor + Service + East, data = nyc)

Residuals:
    Min      1Q  Median      3Q     Max
-14.0465 -3.8837   0.0373   3.3942  17.7491

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.023800   4.708359  -5.102 9.24e-07 ***
Food          1.538120   0.368951   4.169 4.96e-05 ***
Decor         1.910087   0.217005   8.802 1.87e-15 ***
Service      -0.002727   0.396232  -0.007   0.9945
East          2.068050   0.946739   2.184   0.0304 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.738 on 163 degrees of freedom
Multiple R-squared:  0.6279,  Adjusted R-squared:  0.6187
F-statistic: 68.76 on 4 and 163 DF,  p-value: < 2.2e-16
```

Example 2: Menu Pricing in a New Italian Restaurant in New York City

- The initial regression model is

Price = – 24.02 + 1.54 Food + 1.91 Decor – 0.003 Service + 2.07 East

At this point we shall leave the variable Service in the model even though its regression coefficient is not statistically significant.

- The variable Décor has the largest effect on Price since its regression coefficient is largest.
- Note that Food, Décor and Service are each measured on the same 0 to 30 scale and so it is meaningful to compare regression coefficients.
- The variable Décor is also the most statistically significant since its *p* -value is the smallest of the three.
- In order that the price achieved for dinner is maximized, the new restaurant should be on the east of Fifth Avenue since the coefficient of the dummy variable is statistically significantly larger than 0.
- It does not seem possible to achieve a price premium for "setting a new standard for high quality service in Manhattan" for Italian restaurants since the regression coefficient of Service is not statistically significantly greater than zero.

```
Call:
lm(formula = Price ~ Food + Decor + East, data = nyc)

Residuals:
    Min      1Q  Median      3Q     Max
-14.0451 -3.8809  0.0389  3.3918 17.7557

Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) -24.0269   4.6727     -5.142    7.67e-07 ***
Food          1.5363   0.2632      5.838    2.76e-08 ***
Decor         1.9094   0.1900     10.049    < 2e-16 ***
East          2.0670   0.9318      2.218    0.0279 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.72 on 164 degrees of freedom
Multiple R-squared:  0.6279,  Adjusted R-squared:  0.6211
F-statistic: 92.24 on 3 and 164 DF,  p-value: < 2.2e-16
```

- Dropping the predictor Service from the initial model
- The regression coefficients for the variables in both models are very similar.