Linear Statistical Modeling Methods with SAS

Multiple Linear Regression Models and Inferences

Xuemao Zhang
East Stroudsburg University

February 28, 2024

Outline

- MLR Models and ANOVA
- Matrix Approach of MLR Models
- Inferences about Regression Parameters

MLR Models

Consider one dependent variable Y and k independent variables, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k$. The data will be in the form of

$$(x_{11}, x_{21}, \cdots, x_{k1}, y_1), \ldots, (x_{1n}, x_{2n}, \cdots, x_{kn}, y_n).$$

Or

$\mathbf{Y}$	$\mathbf{X}_1$	$\cdots$	$\mathbf{X}_k$
y_1	x_{11}	$\cdots$	x_{k1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	x_{1n}	$\cdots$	x_{kn}

Our objective is to use the information provided by the $X_1, X_2, \ldots, X_k$ to predict the value of Y.

MLR Models

Definition. A **multiple linear regression model** relating a random response Y to a set of independent variables $X_1, X_2, \ldots, X_k$ is of the form

$$Y_i|_{X_1=x_{1i},\ldots,X_k=x_{ki}} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i,$$

where x_{li} is the setting of the lth independent variable for the ith observation, $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters, ε_i's are i.i.d random variables with 0 mean and common variance σ^2, $i = 1, \ldots, n$.

If we remove the observation index i, the MLR model is of the form

$$Y|_{X_1=x_1,\ldots,X_k=x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon,$$

where the variables $X_1, X_2, \ldots, X_k$ assume known values.

Since we assume that $E(\varepsilon_i) = 0, i = 1, \ldots, n$, hence

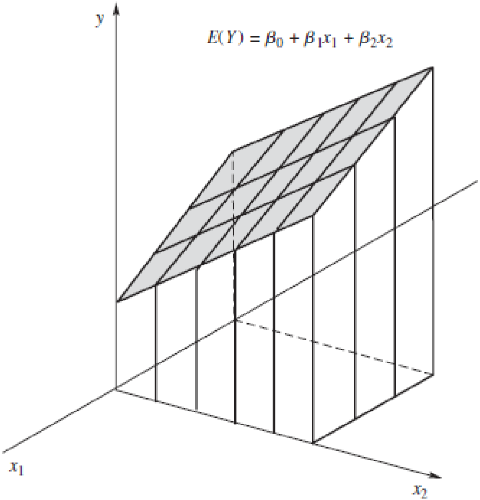
$$E(Y|_{X_1=x_1,\ldots,X_k=x_k}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

The surface (mean response) defined by the deterministic part of the multiple linear regression model in the above, is called the **response surface** of the model.

MLR Models

Example.

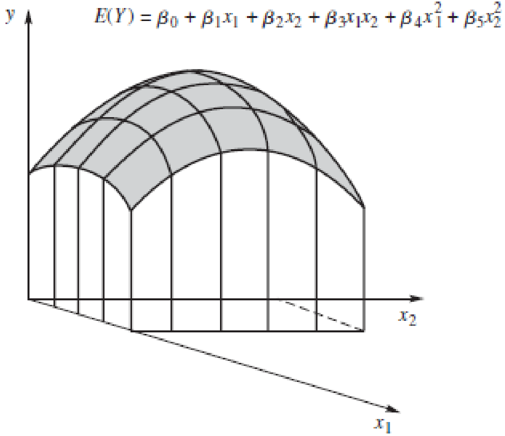
Plot of $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$



MLR Models

Example.

Plot of $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$



Matrix Approach to MLR Models

We now define the following matrices:

$$
\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad X = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}
$$

$$
\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}
$$

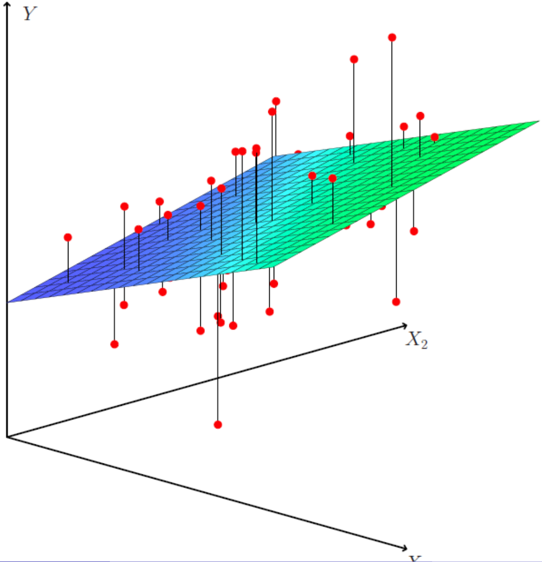
Then the MLR model can be written as

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

where ε has a multivariate distribution with mean $\mathbf{0}$ and variance-covariance matrix $\sigma^2 \mathbf{I}_n$, and $\mathbf{I}_n$ is the n-dimensional identity matrix.

Matrix Approach to MLR Models

Estimation by Least Squares



Matrix Approach to MLR Models

Estimation by Least Squares

- Given estimates $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$, we can make predictions using the formula

$$\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_k x_k.$$

- We estimate $\beta_0, \beta_1, \ldots, \beta_k$ as the values that minimize the sum of squared residuals

$$\begin{aligned} SSE &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots - \widehat{\beta}_k x_{ik})^2 \end{aligned}$$

This is done using statistical software.

Matrix Approach to MLR Models

If we use matrix notation, note that the variance-covariance matrix is given by

$$\begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}.$$

And hence

$$E(\mathbf{Y}) = \mathbf{X}\beta$$

since $E(\varepsilon) = \mathbf{0}$.

Let $\widehat{\beta}$ be the estimator of β by the least square method. Then $\widehat{\beta}$ is obtained by minimizing

$$(\mathbf{Y} - \mathbf{X}\beta)^{'}(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{Y}^{'}\mathbf{Y} - 2\beta^{'}\mathbf{X}^{'}\mathbf{Y} + \beta^{'}\mathbf{X}^{'}\mathbf{X}\beta.$$

Therefore, the least squares normal equations are

$$\mathbf{X}^{'}\mathbf{X}\widehat{\beta} = \mathbf{X}^{'}\mathbf{Y}.$$

And the least squares estimators are:

$$\boxed{\widehat{\beta} = (\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{Y}.}$$

Matrix Approach to MLR Models

Fitted Values In matrix notation, the vector of the fitted values, denoted by $\widehat{\mathbf{Y}}$ is

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta},$$

where $\widehat{\beta} = (\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{Y}$.

Hat Matrix. The hat matrix is defined by

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{X}^{'}.$$

Therefore, the vector of **fitted values** can be written as

$$\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}.$$

Recall: Properties of the Hat matrix

- $\mathbf{H}$ is symmetric.
- Idempotency: $\mathbf{H}\ \mathbf{H} = \mathbf{H}$.

Matrix Approach to MLR Models

Residuals The vector of the residuals is denoted by $\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$. We have

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\beta} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y}.$$

Properties of the $\mathbf{I} - \mathbf{H}$

- $\mathbf{I} - \mathbf{H}$ is symmetric.
- Idempotency: $(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = (\mathbf{I} - \mathbf{H})$.

Variance-Covariance Matrix of Residuals:

$$Var(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H}).$$

Analysis of Variance

$$\boxed{SST = \mathbf{Y}'\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y} = \mathbf{Y}'\left[\mathbf{I} - \frac{1}{n}\mathbf{J}\right]\mathbf{Y},}$$

where $\mathbf{J}$ is the matrix of 1s, $\mathbf{J} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}$.

$$\boxed{\begin{aligned} SSE = &\quad \mathbf{e}'\mathbf{e} = (\mathbf{Y} - \mathbf{X}\widehat{\beta})'(\mathbf{Y} - \mathbf{X}\widehat{\beta}) \\ = &\quad \mathbf{Y}'\mathbf{Y} - \widehat{\beta}'\mathbf{X}'\mathbf{Y} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y} \end{aligned}}$$

$$\boxed{\begin{aligned} SSR = &\quad SST - SSE = \widehat{\beta}'\mathbf{X}'\mathbf{Y} - \tfrac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y} \\ = &\quad \mathbf{Y}'\mathbf{X}\widehat{\beta} - \tfrac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y} = \mathbf{Y}'\left[\mathbf{H} - \frac{1}{n}\mathbf{J}\right]\mathbf{Y}. \end{aligned}}$$

Analysis of Variance

The Analysis of Variance for MLR models can be summarized in the following table.

Source	df	SS	MS	F
Regression	k	SSR	$MSR = SSR/k$	MSR/MSE
Error	n-1-k	SSE	$MSE = SSE/(n-1-k)$	
Total	n-1	SST		

MSE is **unbiased** for the common variance σ^2. Therefore,

$$Var(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

is estimated by

$$MSE(\mathbf{I} - \mathbf{H}).$$

F Test of Usefulness of an MLR Model

- Once we have identified and fit a tentative model, we will want to know if there is a statistically significant relation between the response and the explanatory variables. The tool to answer this question is the **F Test**.

- **The Hypotheses:**

$$\begin{aligned} H_0: &\quad \beta_1 = \beta_2 = \cdots = \beta_k = 0 \text{ (i.e. no relationship)} \\ H_a: &\quad \text{Not } H_0 \end{aligned}$$

- **The Test Statistic:** $F^* = \text{MSR}/\text{MSE}$
- H_0 is rejected only if the calculated test statistic F^* is large: given significance level α, H_0 is rejected only if $F^* \geq F_{df_1=k, df_2=n-1-k, \alpha}$.
- **The p-Value:** $P(F_{k,n-1-k} \geq F^*)$, the area under the density curve of the $F_{k,n-1-k}$ distribution that is greater than or equal to F^*, the observed value of the test statistic.

R^2 and Adjusted R^2

The Coefficient of Multiple Determination. R^2, is defined as

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{SST}.$$

R^2 is

- The proportion of variation in the response explained by the regression.
- The proportion by which the unexplained variation in the response is reduced by the regression.

One problem with using R^2 to measure the quality of model fit, is that it can always be increased by adding another regressor. The **Adjusted Coefficient of Multiple Determination**, R_a^2, is a measure that adjusts R^2 for the number of regressors in the model. It is defined as

$$R_a^2 = 1 - \frac{\text{SSE}/(n-1-k)}{SST/(n-1)}.$$

Properties of the Least-Squares Estimators

Denote the Least-Squares estimator of β by

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{bmatrix}$$

1. $E(\widehat{\beta}_i) = \beta_i$, for $i = 0, 1, 2, \ldots, k$.
2. $V(\widehat{\beta}_i) = c_{i+1,i+1}\sigma^2$, where $c_{i+1,i+1}$ is the element in row $i+1$ and column $i+1$ of $(\mathbf{X}'\mathbf{X})^{-1}$.
3. $Cov(\widehat{\beta}_i, \widehat{\beta}_j) = c_{i+1,j+1}\sigma^2$, where $c_{i+1,j+1}$ is the element in row $i+1$ and column $j+1$ of $(\mathbf{X}'\mathbf{X})^{-1}$; $c_{11} = 1/S_{xx}$.
4. An unbiased estimator of σ^2 is $MSE = SSE/(n-1-k)$, where $SSE = \mathbf{Y}'\mathbf{Y} - \widehat{\beta}'\mathbf{X}'\mathbf{Y}$. (Notice that there are $k+1$ unknown β_i values in the model.)

If, in addition, the ε_i , for $i = 1, 2, \ldots, n$ are normal $N(0, \sigma^2)$,

5. Each $\widehat{\beta}_i$ is normally distributed.
6. The random variable $\dfrac{(n-1-k)MSE}{\sigma^2}$ has a χ^2 distribution with $n-1-k$ df.
7. The statistic MSE is independent of $\widehat{\beta}_i$ for each $i = 0, 1, 2, \ldots, k$.

Properties of the Least-Squares Estimators

Suppose that we wish to make an inference about the linear function

$$a_0\widehat{\beta}_0 + a_1\widehat{\beta}_1 + a_2\widehat{\beta}_2 + \cdots + a_k\widehat{\beta}_k,$$

where $a_0, a_1, a_2, \ldots, a_k$ are constants. In matrix notation, define

$$\mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}.$$

Then

$$a_0\widehat{\beta}_0 + a_1\widehat{\beta}_1 + a_2\widehat{\beta}_2 + \cdots + a_k\widehat{\beta}_k = \mathbf{a}^{'}\widehat{\beta}.$$

THEOREM.

$$\boxed{\mathbf{a}^{'}\widehat{\beta} \sim N(\mathbf{a}^{'}\beta,\ \ [\mathbf{a}^{'}(\mathbf{X}^{'}\mathbf{X})^{-1}a]\sigma^{2}) \text{ if } \varepsilon_i's \text{ are i.i.d. } N(0, \sigma^{2}) \text{ random variables.}}$$

Inferences about Regression Parameters

It can be shown that

$$T = \frac{\mathbf{a}^{'}\widehat{\beta} - (\mathbf{a}^{'}\beta)_0)}{\sqrt{MSE \cdot \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}}$$

possesses a Student's t-distribution under $H_0 : \mathbf{a}^{'}\beta = (\mathbf{a}^{'}\beta)_0$ with $n-1-k$ df, where $(\mathbf{a}^{'}\beta)_0$ is some specified value.

A Test for $\mathbf{a}^{'}\beta$

$H_0 : \mathbf{a}^{'}\beta = (\mathbf{a}^{'}\beta)_0$

$$H_a : \begin{cases} \mathbf{a}^{'}\beta > (\mathbf{a}^{'}\beta)_0 \\ \mathbf{a}^{'}\beta < (\mathbf{a}^{'}\beta)_0 \\ \mathbf{a}^{'}\beta \neq (\mathbf{a}^{'}\beta)_0 \end{cases}$$

Test statistic: $T = \dfrac{\mathbf{a}^{'}\widehat{\beta} - (\mathbf{a}^{'}\beta)_0)}{\sqrt{MSE \cdot \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}}$

Rejection region: $\begin{cases} t \geq t_\alpha \\ t \leq -t_\alpha \\ |t| \geq t_{\alpha/2} \end{cases}$

Here, the t-distribution is based on $n-1-k$ df.

Inferences about Regression Parameters

The corresponding $100(1-\alpha)\%$ confidence interval for $\mathbf{a}^{'}\boldsymbol{\beta}$ is as follows.

$$\boxed{\begin{array}{l}\text{A } 100(1-\alpha)\% \text{ Confidence Interval for } \mathbf{a}^{'}\boldsymbol{\beta}: \\ \qquad \mathbf{a}^{'}\widehat{\boldsymbol{\beta}} \pm t_{\alpha/2}\sqrt{MSE}\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}\end{array}}$$

Remarks.

1. A single β_i can be regarded as a special case of linear combination of $\beta_0, \beta_1, \ldots, \beta_k$, if we choose

$$a_j = \begin{cases} 1, & \text{if } j = i, \\ 0, & \text{if } j \neq i, \end{cases}$$

then $\beta_i = \mathbf{a}^{'}\boldsymbol{\beta}$ for this choice of $\mathbf{a}$.

Inferences about Regression Parameters

Test of Individual Regression Parameters

When we wish to test whether the term $\beta_i x_i$ can be dropped from a multiple regression model, we are interested in testing

$$H_0 : \beta_i = 0, \quad i = 1, 2, \ldots, k.$$

The results derived above can be summarized as what follows.

A Test for individual β_i

$H_0 : \beta_i = 0, \quad i = 1, 2, \ldots, k$

$$H_a : \begin{cases} \beta_i > 0 \\ \beta_i < 0 \\ \beta_i \neq 0 \end{cases}$$

Test statistic: $T = \dfrac{\widehat{\beta}_i - 0}{s.e.(\widehat{\beta}_i)}$

$$\text{Rejection region: } \begin{cases} t \geq t_\alpha \\ t \leq -t_\alpha \\ |t| \geq t_{\alpha/2} \end{cases}$$

Here, the t-distribution is based on $n - 1 - k$ df,
and $s.e.(\widehat{\beta}_i) = \sqrt{MSE}\sqrt{C_{i+1,i+1}}$,
where $C_{i+1,i+1}$ is the (i+1)th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.

Inferences about Regression Parameters

Predicting the Mean Value of Y

2. One useful application of the hypothesis-testing and confidence interval techniques just presented is to solve the problem of estimating the mean $E(Y)$, for fixed values of the independent variables $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_k^*)$. Then

$$E(Y|\mathbf{x} = \mathbf{x}^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_k x_k^*.$$

Notice that $\mathbf{a} = (1, x_1^*, x_2^*, \ldots, x_k^*)'$.

Inferences about Regression Parameters

Predicting a Particular Value of Y

Consider the MLR model

$$Y_i|_{X_1=x_{1i},\ldots,X_k=x_{ki}} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i,$$

where ε_i's are i.i.d Normal random variables with 0 mean and common variance σ^2, $i = 1, \ldots, n$.

Let $\mathbf{x} = \mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_k^*)$ be a fixed vector of the independent variables. Instead of estimating $E(Y)$ value at $\mathbf{x} = \mathbf{x}^*$, we wish to predict the particular (individual) response Y that we will observe if the experiment is run at some time in the future, denoted by Y^*. Then

$$Y^* = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_k x_k^* + \varepsilon.$$

It is natural to estimate Y^* by

$$\widehat{Y^*} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1^* + \widehat{\beta}_2 x_2^* + \cdots \widehat{\beta}_k x_k^* = \mathbf{a}'\beta,$$

where

$$\mathbf{a} = (1, x_1^*, x_2^*, \ldots, x_k^*)'.$$

Theorem. Let $S = \sqrt{MSE}$. Then

$$T = \frac{Y^* - \widehat{Y^*}}{S\sqrt{1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}}$$

possess a Student's t distribution with $n - 1 - k$ df.

A $100(1-\alpha)\%$ Prediction Confidence Interval for Y when $x_1 = x_1^*, x_2 = x_2^*, \ldots, x_k = x_k^*$

$$\boxed{\begin{array}{l} \qquad\qquad \mathbf{a}^{'}\beta \pm t_{\alpha/2, n-1-k} S \sqrt{1 + \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}. \\ \text{where } \mathbf{a} = (1, x_1^*, x_2^*, \ldots, x_k^*)^{'}. \end{array}}$$

Remark. Again, prediction intervals for the actual value of Y are longer than confidence intervals for $E(Y)$ if both confidence levels are the same and both are determined for the same value of $\mathbf{x} = \mathbf{x}^*$.

Example

- Let's see what happens when we identify and fit a model to data in Cars93.csv. The **scatterplot array** on the next slide shows the response, highway mpg (HIGHMPG) and three potential predictors, displacement (DISPLACE), horsepower (HP) and rpm (RPM).

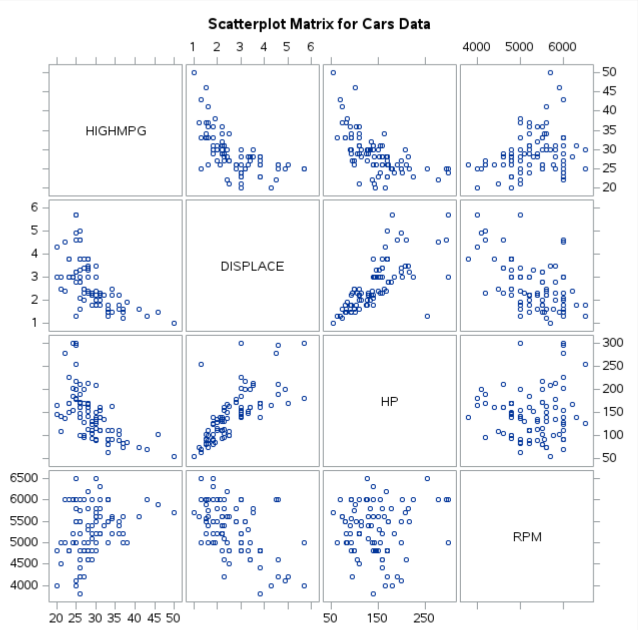
```
proc import
datafile="/home/u5235839/my_shared_file_links/u5235839/cars93.csv"
out=cars93 dbms=csv replace;
getnames=yes;
run;
```

Example

- Scatter plot matrix

```
proc sgscatter data=cars93;
  title "Scatterplot Matrix for Cars Data";
  matrix HIGHMPG DISPLACE  HP  RPM;
run;
```

Example



Scatterplot Matrix for Cars Data
HIGHMPG
DISPLACE
HP
RPM
1 2 3 4 5 6
4000 5000 6000
50 45 40 35 30 25 20
6 5 4 3 2 1
300 250 200 150 100 50
6500 6000 5500 5000 4500 4000
20 25 30 35 40 45 50
50 150 250

Example

- There seems to be a curvilinear relation between the response and each potential predictor, so we will begin by trying a model with linear and squared terms for each predictor. The resulting fitted model is (letting Y denote highway mpg, H denote HP, R denote RPM, and D denote displacement):

$$\begin{aligned}\hat{Y} = & -5.6759 - 6.5411D - 0.1177H + 0.0190R \\ & + 1.0810D^2 + 0.00017H^2 - (1.56 \times 10^{-6})R^2.\end{aligned}$$

```
data cars93new;
set cars93;
D=DISPLACE;
D2=DISPLACE**2;
H=HP;
H2=HP**2;
R=RPM;
R2=RPM**2;
run;
```

```
proc reg simple data=cars93new;
model HIGHMPG = D D2 H H2 R R2;
run;
```

Example

- To include the confidence intervals (99% confidence interval for example) of the regression parameters, we use the following SAS code

```
proc reg data=cars93new;
model HIGHMPG = D D2 H H2 R R2 / clb alpha=0.01;
run;
```

- To get confidence intervals of the mean response and the prediction confidence intervals of an individual response (99% confidence interval for example) of the regression parameters, we use the following SAS code.

```
proc reg data=cars93new;
model HIGHMPG = D D2 H H2 R R2 / clm cli alpha=0.01;
run;
```

Example

- To check model assumptions, we consider QQ plot of the residuals and residual plot:

```
proc reg data=Cars93new;
model HIGHMPG = D D2 H H2 R R2;
output out=new p=predict r=resid;
run;
```

```
proc univariate normal plot data=new;
var resid;
run;
```

```
proc plot data=new;
plot resid*predict='o';
run;
```

License