Linear Statistical Modeling Methods with SAS

Simple Linear Regression Models

Xuemao Zhang
East Stroudsburg University

February 2, 2024

Outline

- The Relationship Between Two Variables
- Pearson Correlation Coefficient
- Bivariate Normal Distribution
- Inference about Pearson Correlation
- Simple Linear Regression models

The Relationship Between Two Variables

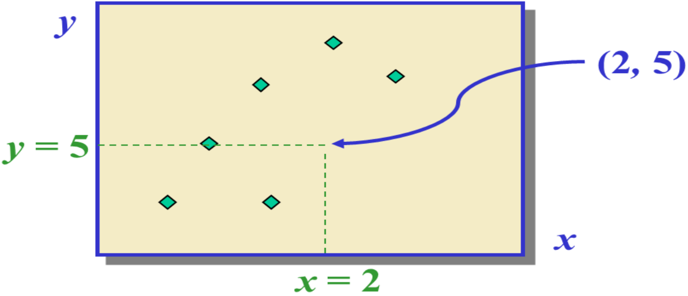
- When two variables are measured on a single experimental unit, the resulting data are called **bivariate data**.
- We can describe each variable individually, and we can also explore the **relationship** between the two variables.
- Bivariate data can be described with
  1. Graphs
  2. Numerical Measures

Two Quantitative Variables

When both of the variables are quantitative, call one variable X and the other Y. And we use x and y to denote the realized value of X and Y, respectively. A single measurement is a pair of numbers (x, y) that can be plotted using a two-dimensional graph called a **scatterplot**.

The Relationship Between Two Variables

The **scatterplot** is the basic tool for graphically displaying bivariate quantitative data.



The Relationship Between Two Variables

Suppose we have the following data:

weight	height	age
64	57	8
71	59	10
53	49	6
67	62	11
55	51	8
58	50	7
77	55	10
57	48	9
56	42	10
51	42	6
76	61	12
68	57	9

The Relationship Between Two Variables

```
data bmi;
input weight height age@@;
datalines;
64 57 8
71 59 10
53 49 6
67 62 11
55 51 8
58 50 7
77 55 10
57 48 9
56 42 10
51 42 6
76 61 12
68 57 9
;

run;
proc print data= bmi;
run;
```

The Relationship Between Two Variables

- To get a scatter plot, we use proc plot in SAS:

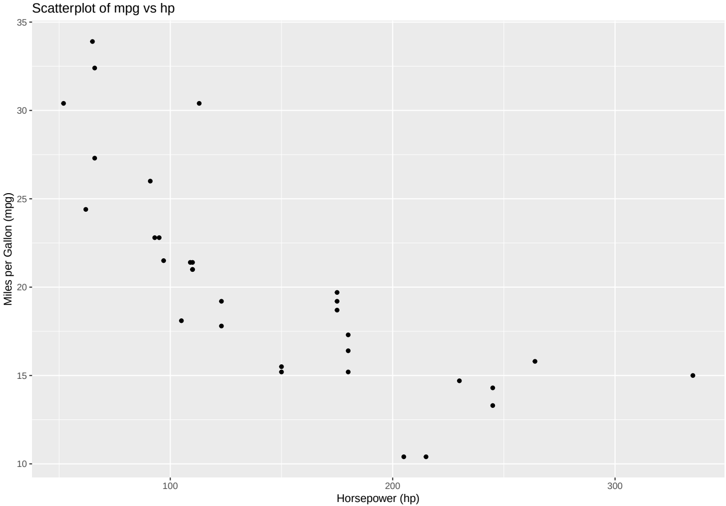
```
proc plot data=bmi;
plot weight*height = 'o';
run;
```

- Or we use proc sgplot

```
proc sgplot data=bmi;
scatter x=height y=weight;
run;
run;
```

The Relationship Between Two Variables

- Another example

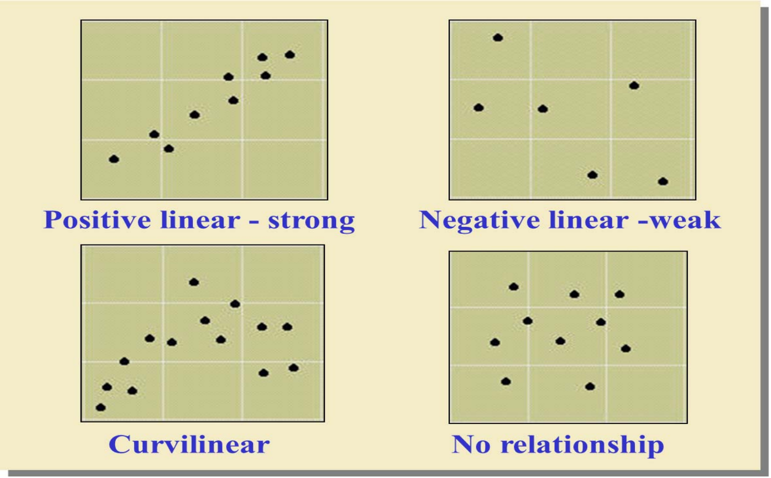


The Relationship Between Two Variables

When analyzing a scatterplot, we should look for:

- **Association.** This is a **pattern** in the scatterplot.
  1. Straight line? or Curve?
  2. No pattern at all?
- **Type of Association.** If there is association, is it:
  - **Linear?**
  - **Nonlinear?**
- **Direction of Association.** Positive or negative?
- **Strength of Association.** Strong or weak?

The Relationship Between Two Variables



Positive linear - strong

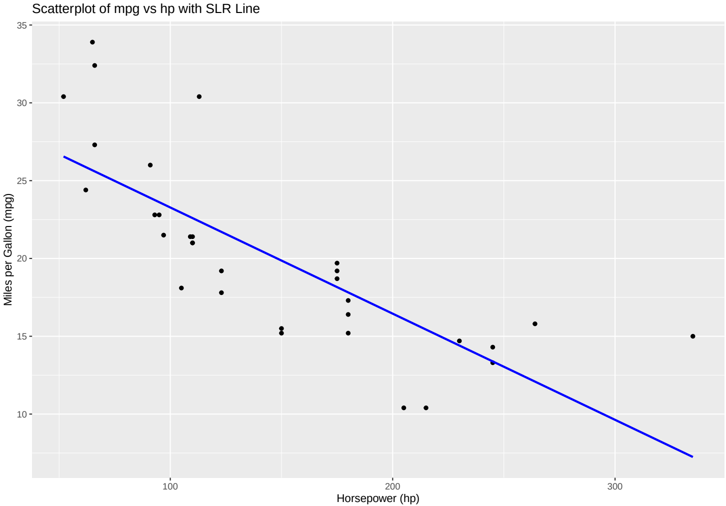
Negative linear -weak

Curvilinear

No relationship

The Relationship Between Two Variables

- For the `mtcars` data, there is association. This shows up as a general negative relation: as we would expect Is the association linear? Look at the scatterplot again.



The Relationship Between Two Variables

- The type of association displayed in the mileage data can be described as **probabilistic**(statistical). Such association is given by a line or curve meant to describe the **central** Y value (usually the **mean**) for each x value. The data values are scattered about the curve in some distributional pattern.
- **Statistical association** is distinct from **deterministic** association. If the association between `hp` and `mpg` were deterministic and linear, all data values would lie precisely on a line.

Correlation for Two Quantitative Variables

- **Correlation** is another name for `linear relationship`
- Let X and Y be two random variables.
- Recall that $\mu_X = \sum_x xP_X(x)$(discrete) or $= \int_{-\infty}^{\infty} xf_X(x)dx$(continuous).
- $\sigma_X^2 = Var(X) = E\left[(X - \mu_X)^2\right] =$

$$\begin{cases} \sum_x (x - \mu_X)^2 P_X(x) & (discrete) \\ \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)dx & (continuous) \end{cases}$$

- **Definition.** The **covariance** of two random variables X and Y is defined by

$$\sigma_{XY}^2 = Cov(X, Y) = E\left[(X - \mu_X)(Y - \mu_Y)\right].$$

- **Definition.**The **correlation coefficient** of two random variables X and Y is defined by

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}.$$

- ρ_{XY} measures the linear relationship between X and Y.

Correlation for Two Quantitative Variables

- Assume that the two random variables X and Y exhibit a **linear pattern** or form.

Suppose n paired measurements, $(x_i,\ y_i),\ i = 1, \ldots, n$ are taken on the variables X and Y (for example, the heights and armspans of n individuals).

We can summarize the location (or center) of each variable by the **sample means**:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

We can summarize the spread of each variable by the **sample standard deviations**:

$$S_x = \sqrt{S_x^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \text{ and } S_y = \sqrt{S_y^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

Pearson Correlation

- However, none of these summary measures says anything about the relationship between the two variables.
  - One measure of the relationship between X and Y is the $\{$**Pearson correlation**$\}$**.**
- The Pearson correlation between X and Y computed from these data is

$$r = \frac{1}{n-1}\sum_{i=1}^{n} x_i' y_i' = \frac{1}{n-1}\frac{S_{xy}}{S_x S_y},$$

where

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}), \text{ and}$$

$$x_i' = \frac{x_i - \overline{x}}{S_x} \text{ and } y_i' = \frac{y_i - \overline{y}}{S_y}$$

are the standardized data.

Pearson Correlation

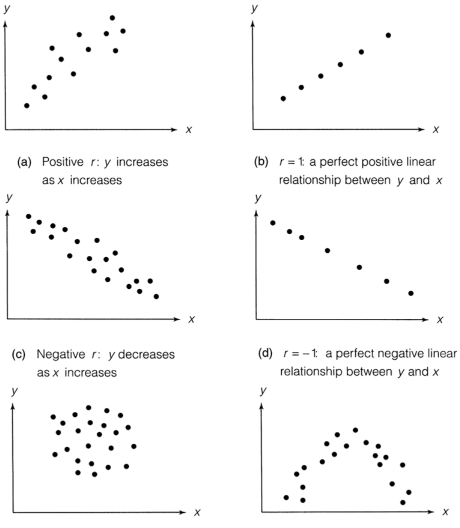
Assume that the two variables X and Y exhibit a **linear pattern** or form.

- $-1 \leq r \leq 1$. Sign of r indicates direction of the linear relationship
- If $r \approx 0$, Weak relationship; random scatter of points
- If $r \approx 1$ or -1, Strong relationship; either positive or negative
- If $r = 1$ or 1, All points fall exactly on a straight line

- Correlation between X and Y is the same as the correlation between Y and X.
- Correlation can never by itself adequately summarize a set of bivariate data. Only when used in conjunction with $\overline{x}$, $\overline{y}$, S_x, and S_y **and a scatterplot** can an adequate summary be obtained.
- The statistical significance of a correlation can only be judged with respect to the sample size. This is not necessarily the same as its practical significance.
- If $S_x = 0$ and/or $S_y = 0$, we define r to be 0. This is because
  - The formula doesn't work in this case (division of 0 by 0)
  - The standard deviation equals 0 if and only if all data values are equal, so
  - There can be no association since there is no variation.

Pearson Correlation

- The following figures illustrate what Pearson correlation measures.



(a) Positive r: y increases as x increases

(b) $r = 1$: a perfect positive linear relationship between y and x

(c) Negative r: y decreases as x increases

(d) $r = -1$: a perfect negative linear relationship between y and x

(e) r near zero: little or no linear relationship between y and x

(f) r near zero: little or no linear relationship between y and x

Bivariate normal distribution

- The Pearson correlation coefficient, r, tells us about the strength of the linear relationship between X and Y in the sample data.
- To make inference about the population correlation coefficient ρ, we perform a hypothesis test of the significance of the correlation coefficient to decide whether the evidence in the sample data is strong enough to indicate a significant linear correlation at the population level.
- Generally, we test the null hypothesis $H_0 : \rho = 0$ against a two-sided alternative $H_a : \rho \neq 0$. That is, we check if the population correlation coefficient ρ is significantly different from 0. To this end, we must assume that the random vector (X, Y) follows a bivariate normal distribution.

It is convenient to introduce the distribution using matrix notation.

Bivariate normal distribution

- **Expectation of a Random Vector**: Suppose we have a *n*-dimensional vector, $\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$. Then the expected value of $\mathbf{Y}$, denoted by $E(\mathbf{Y})$, is defined by

$$E(\mathbf{Y}) = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{bmatrix}.$$

  That is, the expected value of a random vector is a vector whose elements are the expected values of the random variables that are the elements of the random vector.

- **Expectation of a Random Matrix**: Similarly, the expected value of a random matrix is defined to be a matrix whose elements are the expected values of the corresponding random variables in the original matrix.

Bivariate normal distribution

Variance-Covariance Matrix of a Random Vector Suppose we have a n-dimensional vector, $\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$. Then the Variance-Covariance Matrix of $\mathbf{Y}$, denoted by $Var(\mathbf{Y})$, is defined by

$$Var(\mathbf{Y}) = \begin{bmatrix} Var(Y_1) & Cov(Y_1, Y_2) & \cdots & Cov(Y_1, Y_n) \\ Cov(Y_2, Y_1) & Var(Y_2) & \cdots & Cov(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(Y_n, Y_1) & Cov(Y_n, Y_2) & \cdots & Var(Y_n) \end{bmatrix}.$$

Note.

- The Variance-Covariance Matrix $Var(\mathbf{Y})$ is symmetric since $Cov(Y_i, Y_j) = Cov(Y_j, Y_i)$.
- $Var(\mathbf{Y}) = E\{[\mathbf{Y} - E(\mathbf{Y})][\mathbf{Y} - E(\mathbf{Y})]^{'}\} =$ $E\left\{ \begin{bmatrix} Y_1 - E(Y_1) \\ Y_2 - E(Y_2) \\ \vdots \\ Y_n - E(Y_n) \end{bmatrix} [Y_1 - E(Y_1), Y_2 - E(Y_2), \cdots, Y_n - E(Y_n)] \right\}$

Bivariate normal distribution

To use matrix notation, we define the following matrices:

$$\boxed{\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}}$$

Definition. A random vector

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

is said to have a $MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution if its pdf is given by

$$f(\mathbf{y}) = f(y_1, \ldots, y_n) = \left(\frac{1}{2\pi}\right)^{n/2} \left[\frac{1}{\det \boldsymbol{\Sigma}}\right]^{1/2} \exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right].$$

If $n = 2$, the distribution is called Bivariate Normal Distribution.

Bivariate normal distribution

- Let X and Y have a bivariate normal distribution, then

$$\boxed{\boldsymbol{\mu} = (\mu_1, \mu_2)', \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}}$$

- The density function can also be written as

$$f(x, y) = \frac{e^{-Q/2}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}, -\infty < x, y < \infty,$$

where

$$Q = \frac{1}{1-\rho^2}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right].$$

Bivariate normal distribution

Theorem (Marginal distributions). Let $X \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The marginal distribution of any set of component X is multivariate normal with means, variance and covariance obtained by taking the corresponding components of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively.

Theorem (Conditional distributions). Let X be a n-dimensional random vector and Y be an m-dimensional random vector. Suppose

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim MVN_{n+m}\left(\begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{pmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{22} \end{bmatrix}\right) \quad \text{with} \quad \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^T,$$

then

$$X|Y = y \sim MVN_n(\boldsymbol{\mu}_X + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(y - \boldsymbol{\mu}_Y), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

Bivariate normal distribution

Theorem (Marginal distributions). Let X and Y have a bivariate normal distribution. Then

(a). The marginal distribution of X is normal with mean μ_1 and variance σ_1^2.

(b). The marginal distribution of Y is normal with mean μ_2 and variance σ_2^2.

Theorem (Conditional distributions). Let X and Y have a bivariate normal distribution. Then the conditional distribution of X given that $Y = y$ is a normal distribution with mean

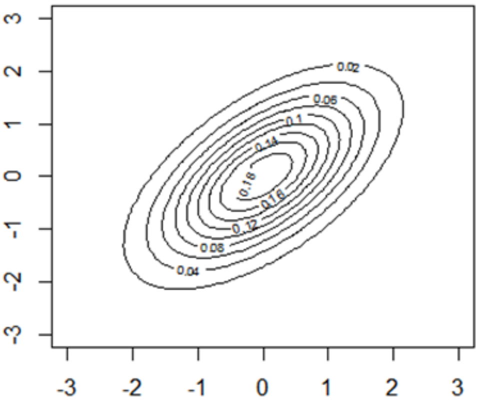
$$\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2)$$

and variance

$$\sigma_1^2(1 - \rho^2).$$

Bivariate normal distribution

- The correlation ρ describes the contour ellipses of the bivariate normal density. If both variables are scaled to have a variance of 1, a correlation of zero corresponds to circular contours, whereas the ellipses become narrower and finally collapse into a line segment as the correlation approaches 1 or -1. The following contour plot of a standard bivariate normal distribution with $\rho = 0.6$ (see SAS code next slide).



Bivariate normal distribution

Density Surface and Contour Plot of a Bivariate Normal Distribution:

- The density surface of the standard bivariate normal distribution (each marginal has mean 0 and standard deviation 1) is produced by the following SAS code.
- The Pearson correlation ρ is specified in the third line of the SAS code (here at 0.6). It would be a good idea to try this program for various values of r between -1 and 1 to explore how the shape of the normal distribution varies with the correlation.
- The correlation ρ describes the contour ellipses of this density. If both variables are scaled to have a variance of 1, a correlation of zero corresponds to circular contours, whereas the ellipses become narrower and finally collapse into a line segment as the correlation approaches 1 or -1.

Bivariate normal distribution

```
title "Bivariate Normal Density";
%let r=0.6;  /*Defines a macro variable r */
data binormal;
pi=3.1416;
do x=-4 to 4 by 0.1;
do y=-4 to 4 by 0.1;
phi=(1/(2*pi))*(1/sqrt(1-&r*&r))*
  exp(-(x**2-2*&r*x*y+y**2)/(2*(1-&r*&r)) );
output;
end;
end;

proc g3d data= binormal;
plot x*y=phi/ rotate=-20;
run;

proc gcontour data=binormal;
plot x*y=phi;
run;
```

Hypothesis Testing for ρ

- If there is a random sample from a **bivariate normal** population, we can consider inference about the population correlation, ρ, using the **Pearson correlation** computed from the sample.
- Let n be the sample size,

$$t = (r - \rho)\sqrt{\frac{n-2}{(1-r^2)(1-\rho^2)}}$$

  has a t_{n-2} distribution.

Note. If the population is not bivariate normal and the sample size n is large, the above statistic has a t_{n-2} distribution approximately.

- Under $H_0 : \rho = \rho_0$, the test statistic

$$t = (r - \rho_0)\sqrt{\frac{n-2}{(1-r^2)(1-\rho_0^2)}}$$

  has a t_{n-2} distribution. We can use this to conduct hypothesis tests.

Hypothesis Testing for ρ

Test of

$$H_0 : \rho = \rho_0.$$

Let t^* is the observed value of the test statistic t calculated from a random sample,

- For $H_{a^+} : \rho > \rho_0$, the p-value is $p^+ = P(t_{n-2} \geq t^*)$;
- For $H_{a_-} : \rho < \rho_0$, the p-value is $p^+ = P(t_{n-2} \leq t^*)$;
- For $H_{a\pm} : \rho \neq \rho_0$, the p-value is $p\pm = 2\min(p_-, p^+)$.

Hypothesis Testing for ρ

Example

weight	height	age
64	57	8
71	59	10
53	49	6
67	62	11
55	51	8
58	50	7
77	55	10
57	48	9
56	42	10
51	42	6
76	61	12
68	57	9

Hypothesis Testing for ρ can be conducted using “corr” procedure in SAS.

Hypothesis Testing for ρ

```
data bmi;
input weight height age@@;
datalines;
64 57 8 71 59 10
53 49 6 67 62 11
55 51 8 58 50 7
77 55 10 57 48 9
56 42 10 51 42 6
76 61 12 68 57 9
;
run;

proc corr data=bmi pearson fisher(rho0=0 TYPE=TWOSIDED);
TITLE "Example of a Pearson Correlation";
   var weight height;
run;
```

SLR model

- Assume that the two random variables X and Y exhibit a **linear pattern** or form.
- There are two numerical measures to describe
  1. The **strength** and **direction** of the relationship between X and Y.
  2. The **form** of the relationship.

Recall that if X and Y have a bivariate normal distribution. Then the conditional distribution of Y given that $X = x$ is a normal distribution with mean

$$\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$

and variance

$$\sigma_Y^2(1 - \rho^2).$$

Notice that

$$E(Y|X = x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X) = \beta_0 + \beta_1 x,$$

where $\beta_0 = \mu_Y - \rho\frac{\sigma_Y}{\sigma_X}\mu_X$ and $\beta_1 = \rho\frac{\sigma_Y}{\sigma_X}$ are unknown.

Instead of modelling a bivariate data set using the bivariate normal distribution, we consider the conditional distribution of $Y|X = x$.

SLR model

- The SLR model attempts to quantify the relationship between a single **independent**(explanatory, or predictor) variable X and a **dependent** (response)variable Y. This reasonably flexible yet simple model has the form

$$Y|_{X=x_i} = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n$$

where ϵ_is are identical independent a random errors having mean 0 and variance σ^2 (often assumed to have a $N(0, \sigma^2)$ distribution) and n is the sample size.

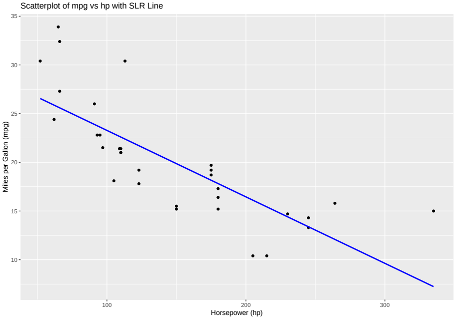
- This model suggests that the means of $Y|X = x_i, i = 1, \ldots, n$ have a common intercept β_0 and common slope β_1.

- For simplicity, the model can be written as

$$Y|x = \beta_0 + \beta_1 x + \epsilon.$$

SLR model

- Recall that the `mtcars` data seems to show a linear relationship between hp and mpg:



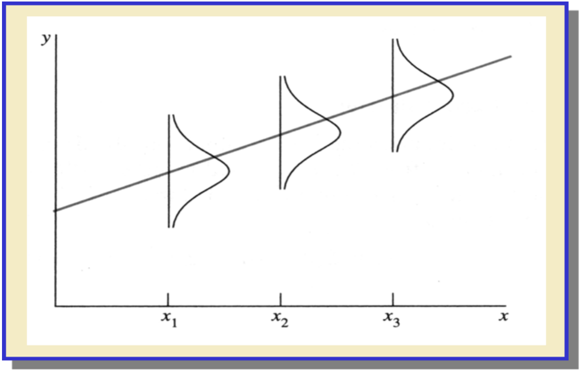
This suggests that we consider the SLR model

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

Where the response Y is mpg, and the predictor x equals hp.

SLR model

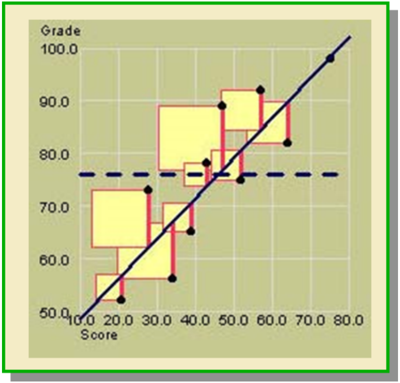
- The line of means, $E(Y_i) = \beta_0 + \beta_1 x_i$, describes average value of Y for any **fixed** value of $X = x_i$. That is, Y_i is a random variable and the $E(Y_i)$ is actually a conditional expectation, $E(Y_i|X = x_i)$.
- The population of measurements is generated as Y deviates from the population line by ϵ. We estimate β_0 and β_1 using sample information $(x_i, y_i),\ i = 1, \ldots, n$.



SLR Model Fitting

The term “model fitting’ ’ refers to using data to estimate model parameters. We will fit the simple linear regression model to a set of data $(x_i, y_i),\ i = 1, \ldots, n$.

It is very natural to minimize the sum of absolute vertical distances, errors of estimation, of the data points from the line.



SLR Model Fitting

Two options for fitting a SLR model are

- **least absolute errors**, which finds values b_0 and b_1 to minimize

$$\mathrm{SAE}(b_0, b_1) = \sum_{i=1}^{n} \mid Y_i - (b_0 + b_1 x_i) \mid,$$

  and

- **least squares**, which finds values $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize

$$\mathrm{SSE}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2.$$

SLR Model Fitting

- To minimize SAE is not an easy task and requires the use of a computer, and the method does not result in closed form solutions.

For now, when fitting regression models, we'll use **least squares only**. Using calculus, we take the derivative of SSE$(\hat{\beta}_0, \hat{\beta}_1)$ with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and set each to 0. Solving these two equations in two unknowns, we find the least squares estimators of β_0 and β_1 to be

$$\begin{aligned}\hat{\beta}_1 =& \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}} \\ =& \frac{\sum_{i=1}^n (x_i - \overline{x})y_i}{S_{xx}}\end{aligned}$$

and

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

SLR Model Fitting

- Because we observed a reasonably linear association between mpg and hp, we will fit a SLR model with `mpg` as the response and `hp` as the regressor.

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

proc sgplot data=mtcars;
scatter x=hp y=mpg;
run;

proc reg data=mtcars;
TITLE "Regression Model fit";
   model mpg = hp;
run;
```

SLR Model Fitting

- We may like to put the scatter plot and the fitted regression line in the same graph:

```
proc sgplot data= mtcars;
   reg y=mpg x=hp;
run;
```

Fitted Values and Residuals

- The **predicted value** of Y at $X = x$ is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

- For $X = x_i$, one of the values in the data set, the predicted value is called a **fitted value** and is written

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

- The **residuals**, $e_i, i = 1, \ldots, n$ are the differences between the observed and fitted values for each data value:

$$e_i = y_i - \hat{Y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i).$$

Fitted Values and Residuals

Recall that we have assumed the responses are produced from the model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

After fitting the model, we have

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i.$$

$\hat{\beta}_0$ is the best guess of β_0, $\hat{\beta}_1$ is the best guess of β_1, and therefore e_i is the best guess of $\epsilon_i, i = 1, \ldots, n$.

- If the model fits well, the residuals should behave as the random errors ϵ_i would: **independent** random selections from the same distribution (usually assumed normal) with mean 0, i.e $N(0, \sigma^2)$.
- If the model does not fit well, a pattern should show up in the distribution of the residuals. This is why we **analyze the residuals** to check the quality of the model fit.

Fitted Values and Residuals

Among the interesting properties of the residuals are these:

- They sum to zero: $\sum_{i=1}^{n} e_i = 0$. This also implies that their mean $\overline{e} = 0$.
- The Pearson correlation between the residuals and the predictors (independent variables) is zero (we say the residuals and the predictors are **uncorrelated**).
- The residuals and fitted values are also **uncorrelated**.

License