CHAPTER 1

Simple Linear Regression

1.1 The Simple Linear Regression Model

1.1 The Simple Linear Regression (SLR) Model

Main Idea: We will be studying model that look at the relationship between a quantitative response variable (Y) and a quantitative explanatory variable (X).

- How strongly related are they?

- In the future, if we know value of one, can we predict the other?

The regression model used to serve three major purposes:

- Description.
- Control.
- Prediction.

Example 1:

1. How is the price of a used car related to the number of miles it’s been driven?

2. Is the number of doctors in a city related to the number of hospitals?

3. How can we predict the price of a textbook from the number of pages?

Algebra Review for Linear Equation:

Equation for a straight line:

$$y = mx + c$$

which can be written also as

$$y = \beta_0 + \beta_1 x$$

where

$\beta_0 (or\ c)$: y-intercept, the value of Y when X = 0,

$\beta_1 (or\ m)$: slope, the increase in Y when X goes up by 1 unit

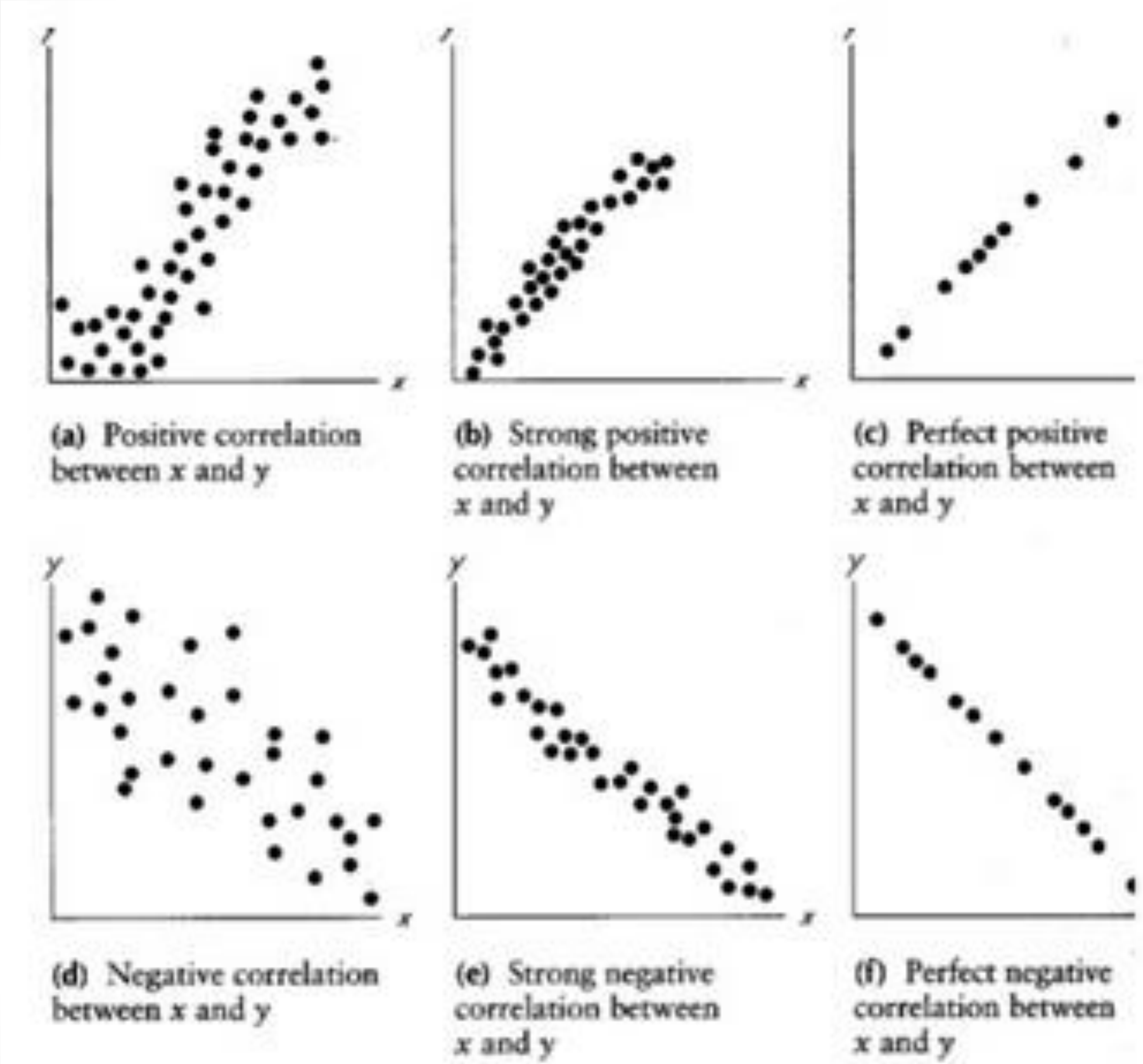
Data for Simple Linear Regression:

- Observe pairs of variables: (X_i, Y_i)
- $i = 1, \cdots, n$ (n is often called the sample size)
- Y_i is the value of the response variable for the i^{th} case.
- X_i is the value of the explanatory variable for the i^{th} case.

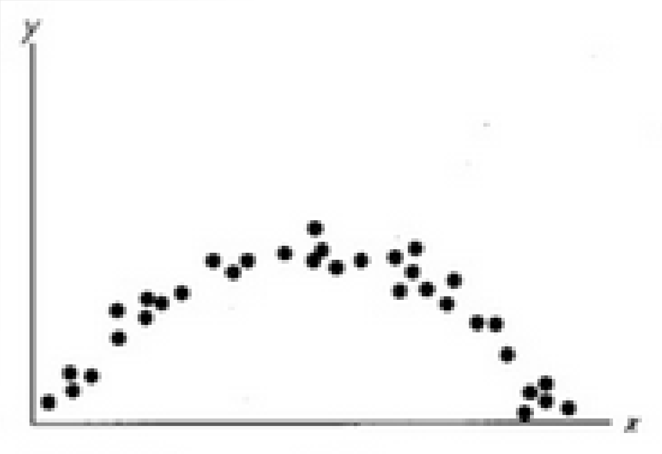
Scatterplot:

A scatterplot shows the relationship between two quantitative variables measured for the same individuals. The values of one variable on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as a point on the graph.

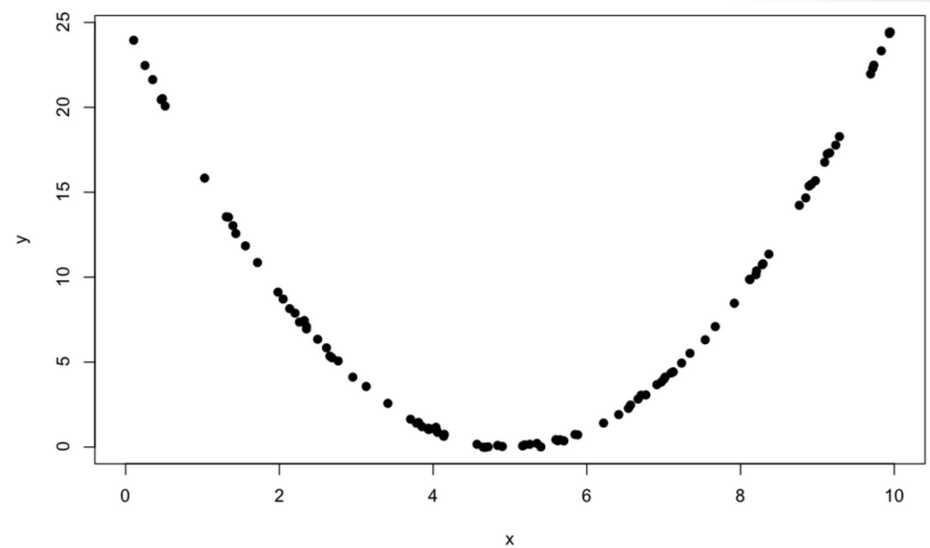
- Scatterplot are used to demonstrate association between two quantitative variables.
- Association (relationship) can be classified into three categories: Linear, Nonlinear, No correlation.
- Two variables may be correlated but not through a linear model (nonlinear).

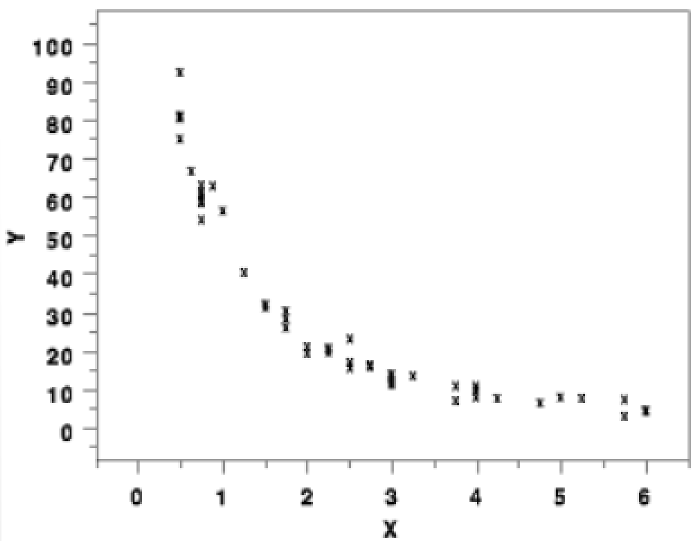


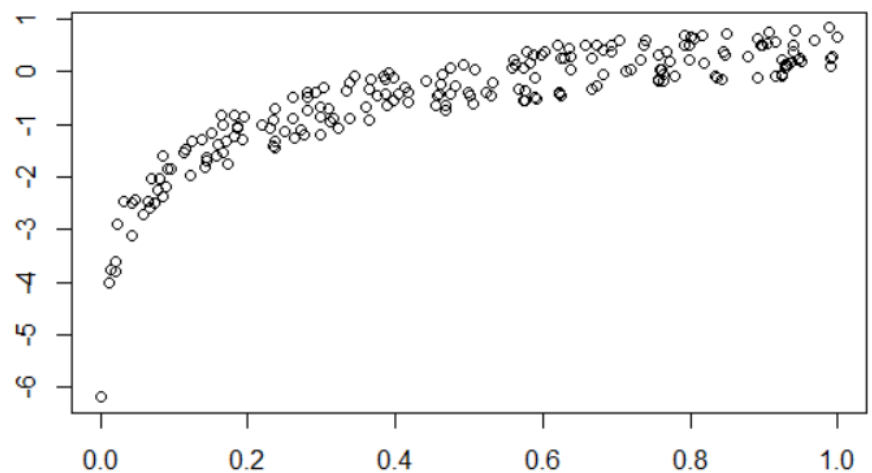
Linear Relationship



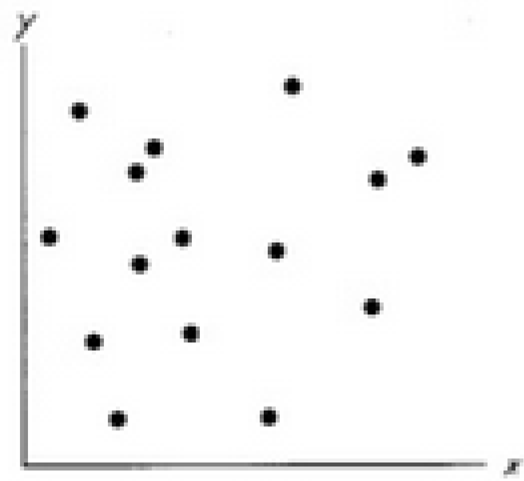
(h) Nonlinear correlation between x and y







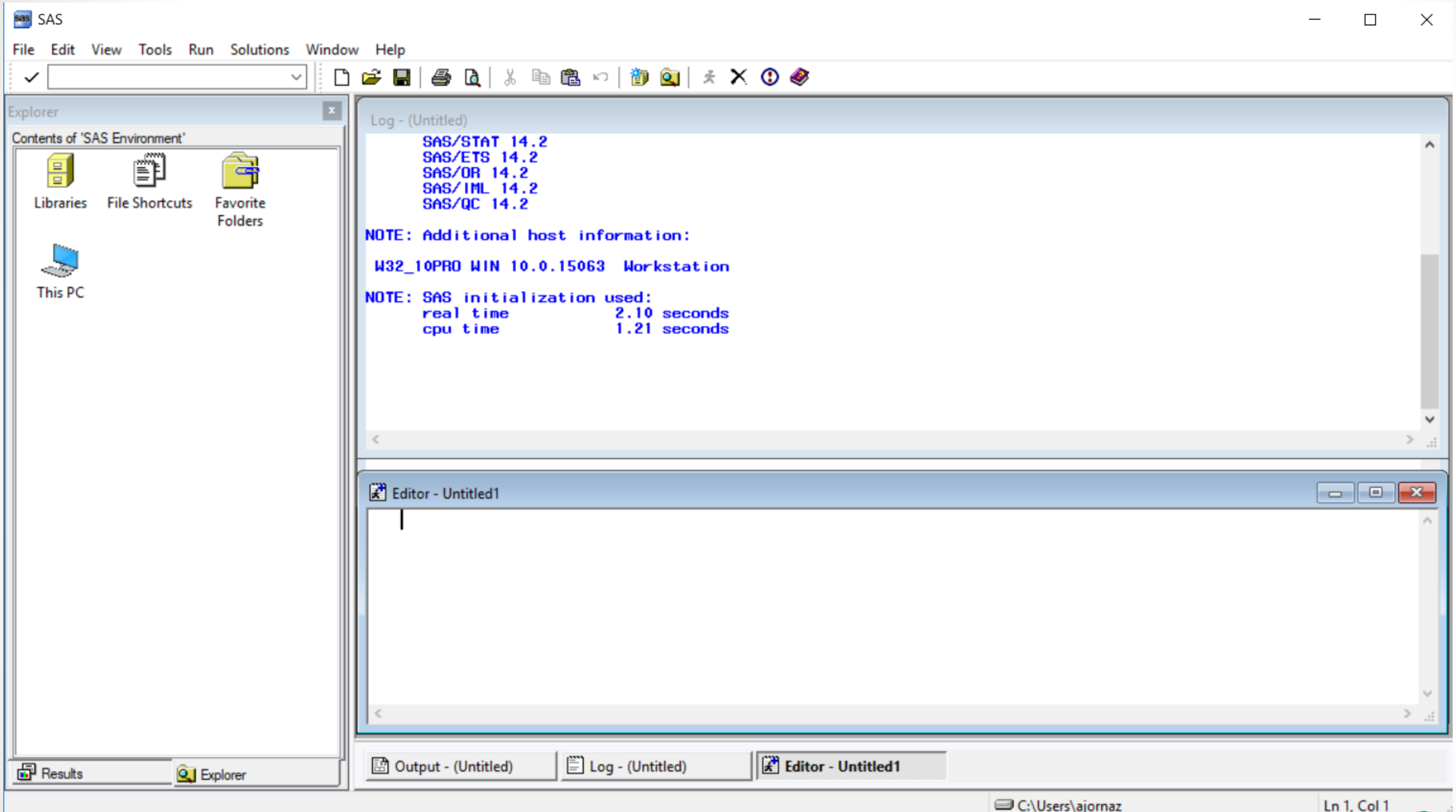
Non-linear Relationship



(g) No correlation between x and y

No Relationship

Starting with SAS (Statistical Analysis Software):



Getting Your Data into SAS:

There are different ways to getting data into SAS. The easy way is entering the data manually,

Dataset name

```
DATA uspresidents;
INPUT President $ Party $ Number;
DATALINES;
Adams     F  2
Lincoln   R 16
Grant     R 18
Kennedy D 35
;
RUN;
```

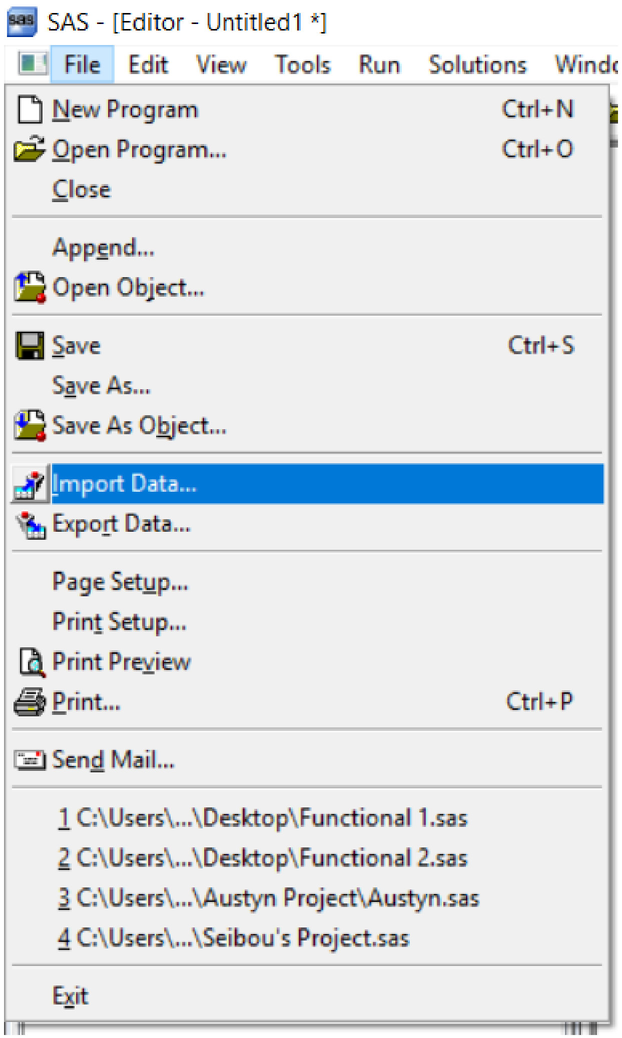
Variable name

Dataset

You need semicolon (;) at the end of each line in SAS.
The ($) means that variable is a categorical variable.

Import Data to SAS:

The easy way is select Import Data from the drop down file menu.



Import Data to SAS:

We can also use PROC IMPORT.

```
proc import datafile = “the file link”
out=dataset name dbms=csv replace;
getnames=yes;
run;
```

Printing the Results from SAS:

We use PROC PRINT.

```
PROC PRINT DATA=dataset name;
RUN;
```

We can use (var statement) to print a specific variable.

```
PROC PRINT DATA=uspresidents;
VAR President;
RUN;
```

Example 1: *(Porsche prices)*

Suppose that we are interested in purchasing a Porsche sports car. *Porscheprices.csv* has three variables which are price, age and mileage.

1. Identify the response variable and the explanatory variable(s).

2. Import the dataset, and print out.

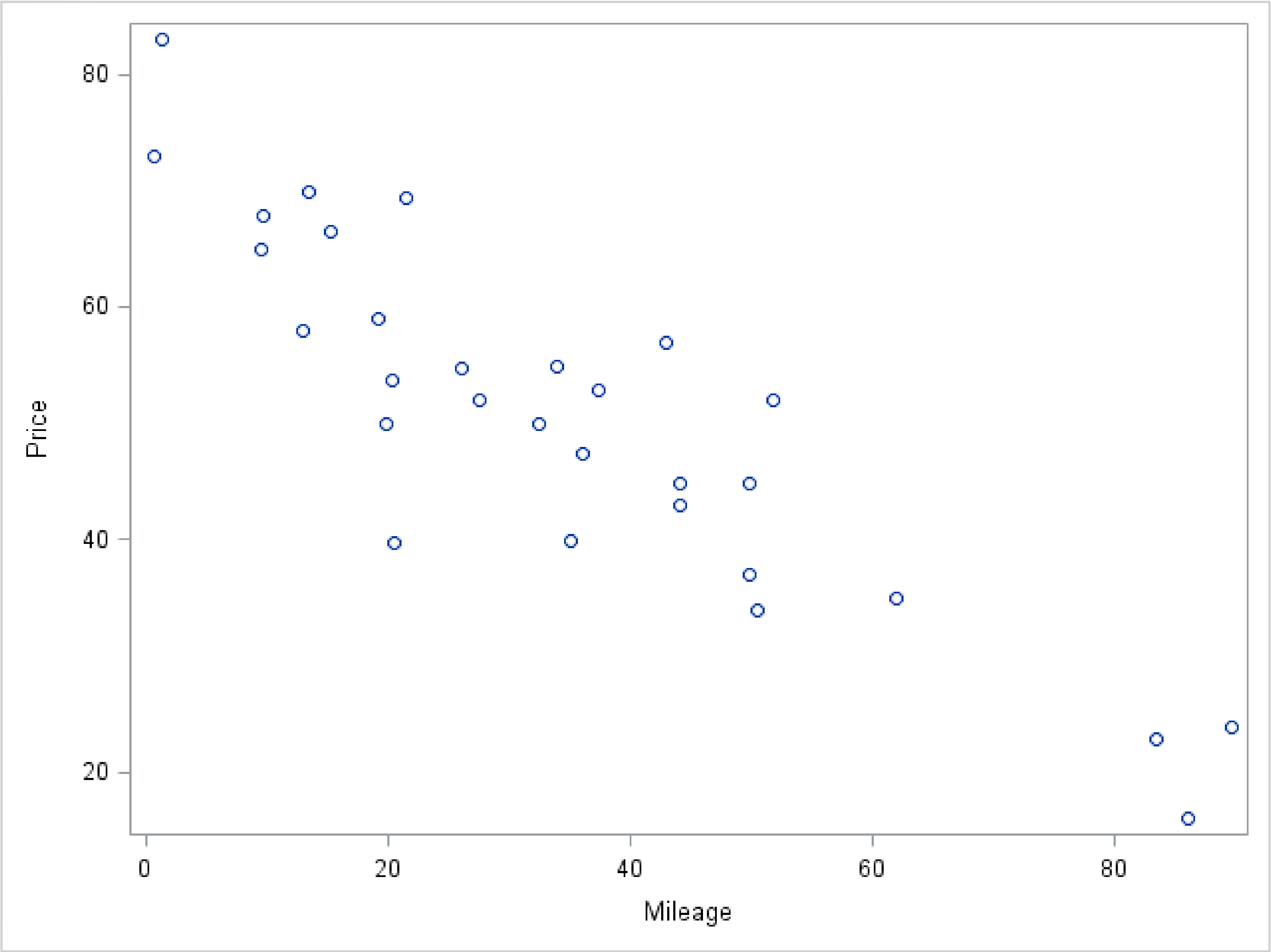
The SAS System

Obs	Price	Age	Mileage
1	69.4	3	21.5
2	56.9	3	43
3	49.9	2	19.9
4	47.4	4	36
5	42.9	4	44

3. Use scatterplot to determine the relationship between the price and mileage.

There are different ways to graph the scatterplot in SAS, one of them using PROC SGPLOT.

```
proc sgplot data = PorschePrice;
scatter x = Mileage y = Price;
run;
```



Price
80
60
40
20
0
20
40
60
80
Mileage

Simple Linear Regression Model:

Statement of model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$$Y = \boxed{\beta_0 + \beta_1 X} + \varepsilon$$

$$\text{Data} = \boxed{\text{Model}} + \text{Error}$$

where

- β_0 is the intercept (the value of Y when X = 0).
- β_1 is the slope (the increase or decrease in Y when X goes up by 1 unit).
- ϵ_i is the i^{th} random error term.

https://istats.shinyapps.io/ExploreLinReg/

Fitting a Simple Linear Model:

The basic idea is minimize how far off we are when we used the line to **predict** Y $(\hat{Y})$ by comparing to **actual** Y. In other ward, minimize the error term (ϵ_i). This approach is called the ordinary least squares (OLS).

- For individual in the data:

$$\boldsymbol{Residual} = \epsilon = y - \hat{y} = observed\ y - predicted\ y$$

- The least square regression line is the line that minimizes the sum of the squared residuals for all points in the dataset. The sum of squared errors (SSE) is that minimum sum.

Example 2: *(Porsche prices)*

For the same dataset in example 1.

1. Using SAS, fit the regression model for price and mileage.

We use PROC REG to fit the regression model.

```
proc reg data=Porscheprices;
model price = mileage;
run;
```

Example 2: *(Porsche prices)*

The regression results:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	71.09045	2.36986	30.00	<.0001
Mileage	1	-0.58940	0.05665	-10.40	<.0001

The simple regression model is:

$$\widehat{Price} = 71.09 - 0.5894.\,Mileage$$

Example 2: *(Porsche prices)*

2. Interpret the intercept and the slope.

   - **Intercept:** The predicted price of a new car (0 mile) is $71,090.
   - **Slope:** For every additional 1000 miles on a used Porsche, the predicted price goes down by $589

Note: in many cases, the intercept lies far from the data used to fit the model and has no practical interpretation.

To display the **predicted** values and **residuals**, we need to use output statement in PROC REG .

```
proc reg data=PorschePrice;
model price = mileage;
output out=residual p=yhat r=res;
run;
```

Create a new data to storage the results

Predicted values

Residuals

```
proc print data=residual;
run;
```

Example 2: *(Porsche prices)*

3. What is the fitted value of the price corresponding to 21,500 (21.5) miles?

$$Price = 71.09 - 0.5894(21.500) = \$58.42$$

4. What is the residual corresponding 21,500 miles?

$$residual = 69.4 - 58.42 = 10.98$$

Obs	Price	Age	Mileage	yhat	res
1	69.4	3	21.5	58.4183	10.9817
2	56.9	3	43	45.7462	11.1538
3	49.9	2	19.9	59.3614	-9.4614
4	47.4	4	36	49.8720	-2.4720
5	42.9	4	44	45.1568	-2.2568
6	36.9	6	49.8	41.7383	-4.8383
7	83	0	1.3	70.3242	12.6758
8	72.9	0	0.67	70.6956	2.2044
9	69.9	2	13.4	63.1925	6.7075
10	67.9	0	9.7	65.3733	2.5267
11	66.5	2	15.3	62.0726	4.4274
12	64.9	2	9.5	65.4911	-0.5911
13	58.9	4	19.1	59.8329	-0.9329
14	57.9	3	12.9	63.4872	-5.5872
15	54.9	10	33.9	51.1098	3.7902
16	54.7	11	26	55.7660	-1.0660
17	53.7	4	20.4	59.0667	-5.3667
18	51.9	4	27.5	54.8819	-2.9819
19	51.9	10	51.7	40.6184	11.2816
20	49.9	3	32.4	51.9939	-2.0939
21	44.9	4	44.1	45.0979	-0.1979
22	44.8	13	49.8	41.7383	3.0617
23	39.9	6	35	50.4614	-10.5614
24	39.7	6	20.5	59.0077	-19.3077
25	34.9	8	62	34.5476	0.3524
26	33.9	7	50.4	41.3846	-7.4846
27	23.9	20	89.6	18.2801	5.6199
28	22.9	22	83.4	21.9344	0.9656
29	16	20	86	20.4020	-4.4020
30	52.9	3	37.4	49.0469	3.8531

Reading Assignment

Read section 1.1