Regression Analysis

- **Regression analysis** is a statistical technique for investigating and modeling the relationship between variables, like X and Y.
- Using per capita income (X) to estimate life expectancy of individuals within a country (Y).
- Using the size of a crab claw (X) to estimate the closing force that it can exert (Y).
- Using the height of a person (X) to estimate the weight of that one (Y).
- The straight line, we usually write as

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

- A simple regression model is to model the relationship between the dependent variable, Y and the independent variable, X as a straight line.

Example

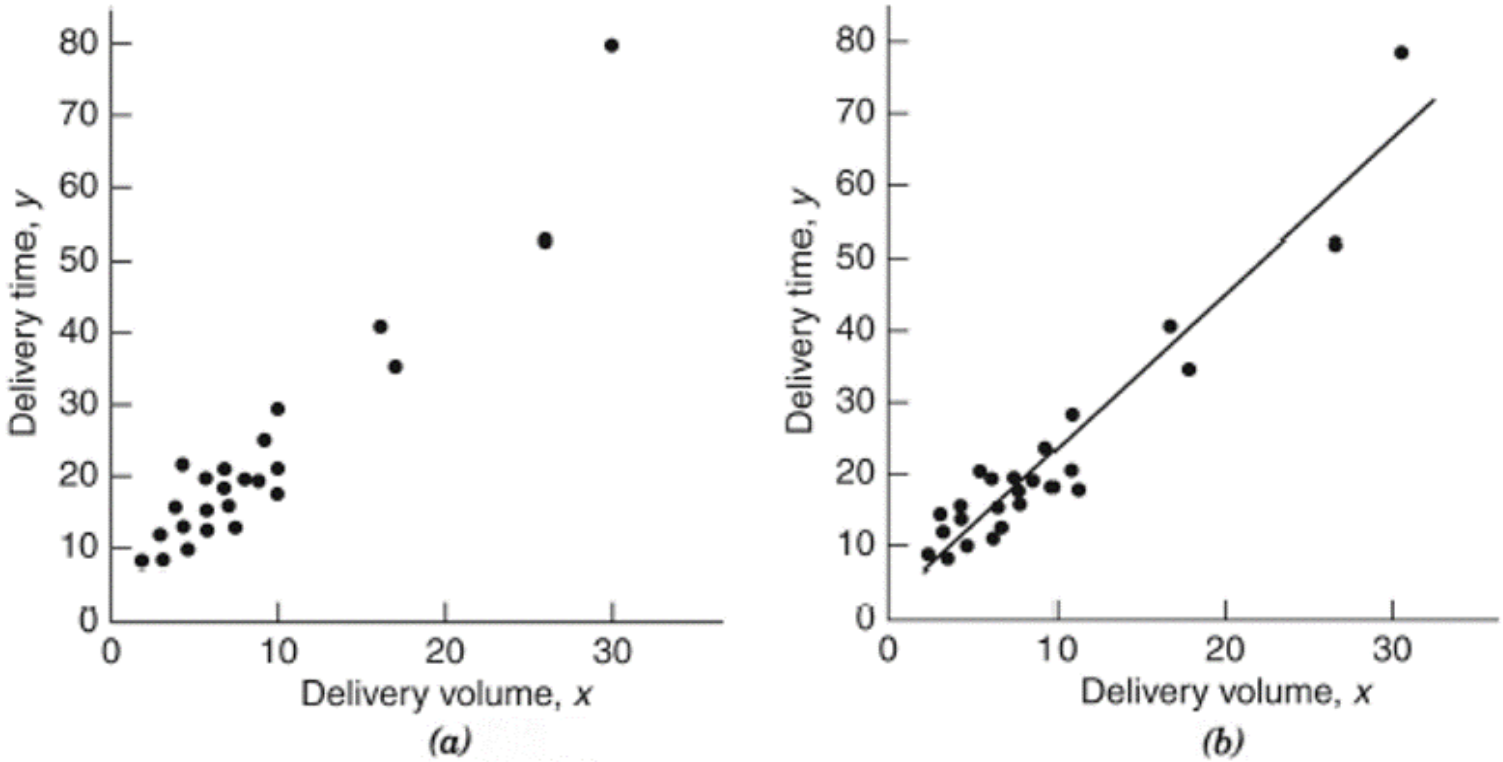


Figure 1.1 (*a*) Scatter diagram for delivery volume. (*b*) Straight-line relationship between delivery time and delivery volume.

Pearson Correlation Coefficient

- Correlation coefficient measures the strength of an association between two variables, like X and Y.
- The commonly used correlation coefficient is the Pearson correlation coefficient.
- Given two continuous variables X and Y, the **Pearson correlation coefficient** ρ is defined as:

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E[(X-\mu_X)^2]}\sqrt{E[(Y-\mu_Y)^2]}}.$$

  where $Cov(X,Y)$ is the covariance of X and Y.
- It can be shown that $-1 \leq \rho \leq 1$.
- Positive ρ means X and Y move in the same direction. Negative ρ means X and Y move in the opposite directions.

- If X and Y have a random sample, of size n: $(x_1, y_1), ..., (x_n, y_n)$, then the Pearson correlation coefficient ρ can be estimated by the **sample** correlation coefficient, r, defined as:

$$r = \frac{Cov(x, y)}{\sqrt{Var(x)}\sqrt{Var(y)}}$$

where $Cov(x, y)$ is the sample covariance of X and Y:

$$Cov(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}),$$

$Var(x)$ and $Var(y)$ are sample variance of X and Y respectively:

$$\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{and} \quad \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

- r indicates the strength of a relationship between two variables.
- r does not tell us the what the relationship is.
- We cannot use r to predict one variable from the other variable.
- We need to fit a model to the data so that we can establish a relationship between X and Y.
- We will assume the target variable (response) is Y and the regressor (explanatory) is X.

Correlation Coefficients: in R

Consider the following data (**Example 1**) which details the sex, height, weight and age of a group of people.

```
> data = read.table ("C:/Data/ex10_1.txt" ,header=T)
> data
  gender height weight age
1      M     68    155  23
2      F     61     99  20
3      F     63    115  21
4      M     70    205  45
5      M     69    170  38
6      F     65    125  30
7      M     72    220  48
> attach(data)
> cor(cbind(height, weight, age), method="pearson")
          height    weight       age
height 1.0000000 0.9716498 0.8731010
weight 0.9716498 1.0000000 0.9239653
age    0.8731010 0.9239653 1.0000000
```

Correlation Coefficients: in Python

```
data = pd.read_csv (r"C:\Data\ex10_1.txt", sep = " ")
print(data)
```

```
  gender  height  weight  age
0      M      68     155   23
1      F      61      99   20
2      F      63     115   21
3      M      70     205   45
4      M      69     170   38
5      F      65     125   30
6      M      72     220   48
```

```
#correlation between weight and height
print( scipy.stats.pearsonr(data['weight'], data['height']) )
```

```
(0.9716498256927286, 0.00025598539082999355)
```

```
#correlation between weight nad height also can be computed a
print( np.corrcoef(data['weight'], data['height']) )
```

```
array([[1.        , 0.97164983],
       [0.97164983, 1.        ]])
```

Correlation Coefficients: in SAS

```
* Correlation values;
proc corr data=example1 nosimple;
title "Example of a correlation matrix";
var height weight age;
run;
*nosimple is used to suppress the descriptive statistics;
```

Example of a correlation matrix

The CORR Procedure

3 Variables:	height weight age

Pearson Correlation Coefficients, N = 7
Prob > |r| under H0: Rho=0

	height	weight	age
height	1.00000	0.97165 0.0003	0.87310 0.0103
weight	0.97165 0.0003	1.00000	0.92397 0.0029
age	0.87310 0.0103	0.92397 0.0029	1.00000

Scatter Plots in SAS (1)

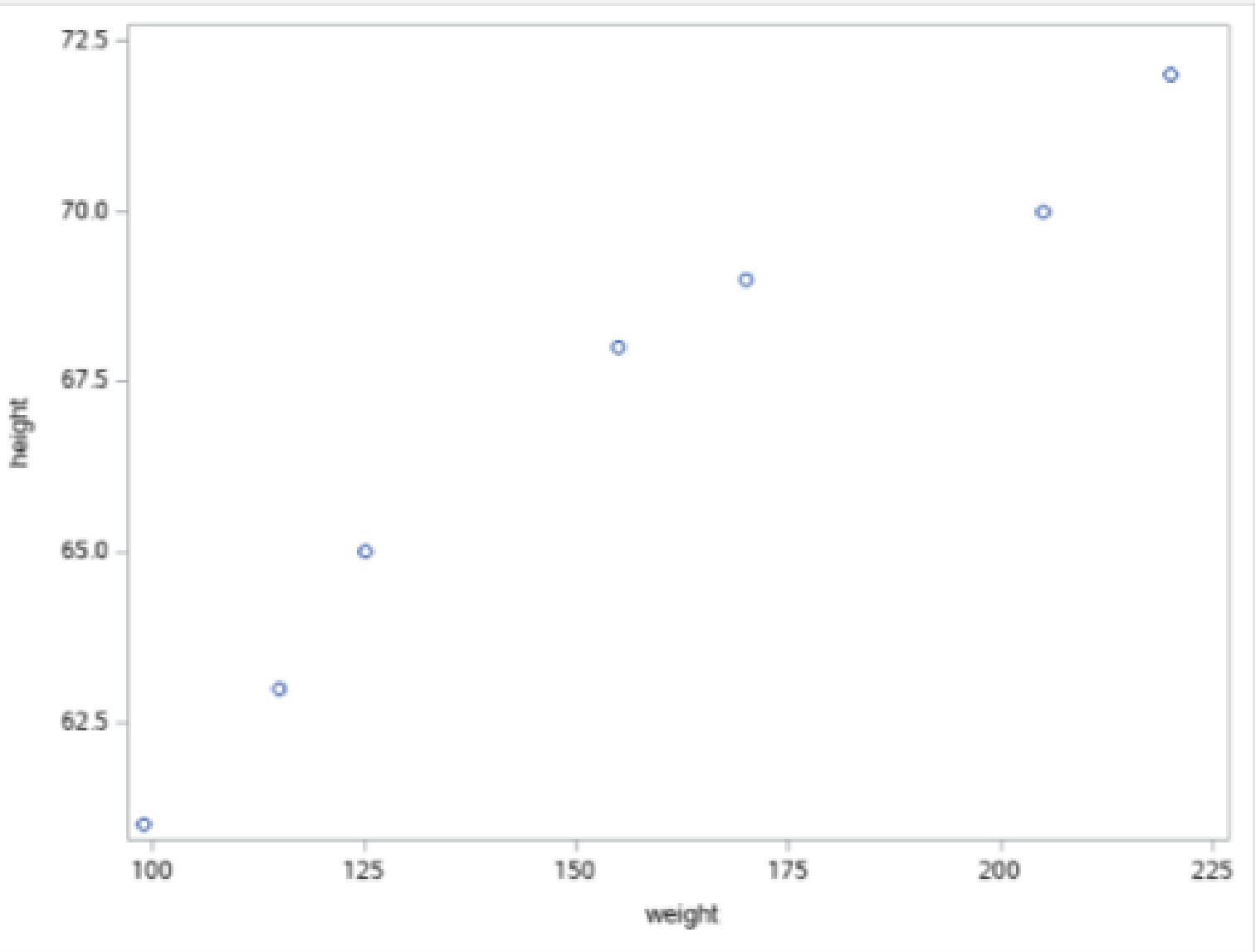
The code is:

```
* Scatter plot of height vs weight;
proc sgscatter  data = example1;
   plot height * weight;
run;

* Scatter plot of height vs weight classified by gender;
proc sgscatter  data = example1;
   plot height * weight
   / datalabel = gender group = gender;
run;
```

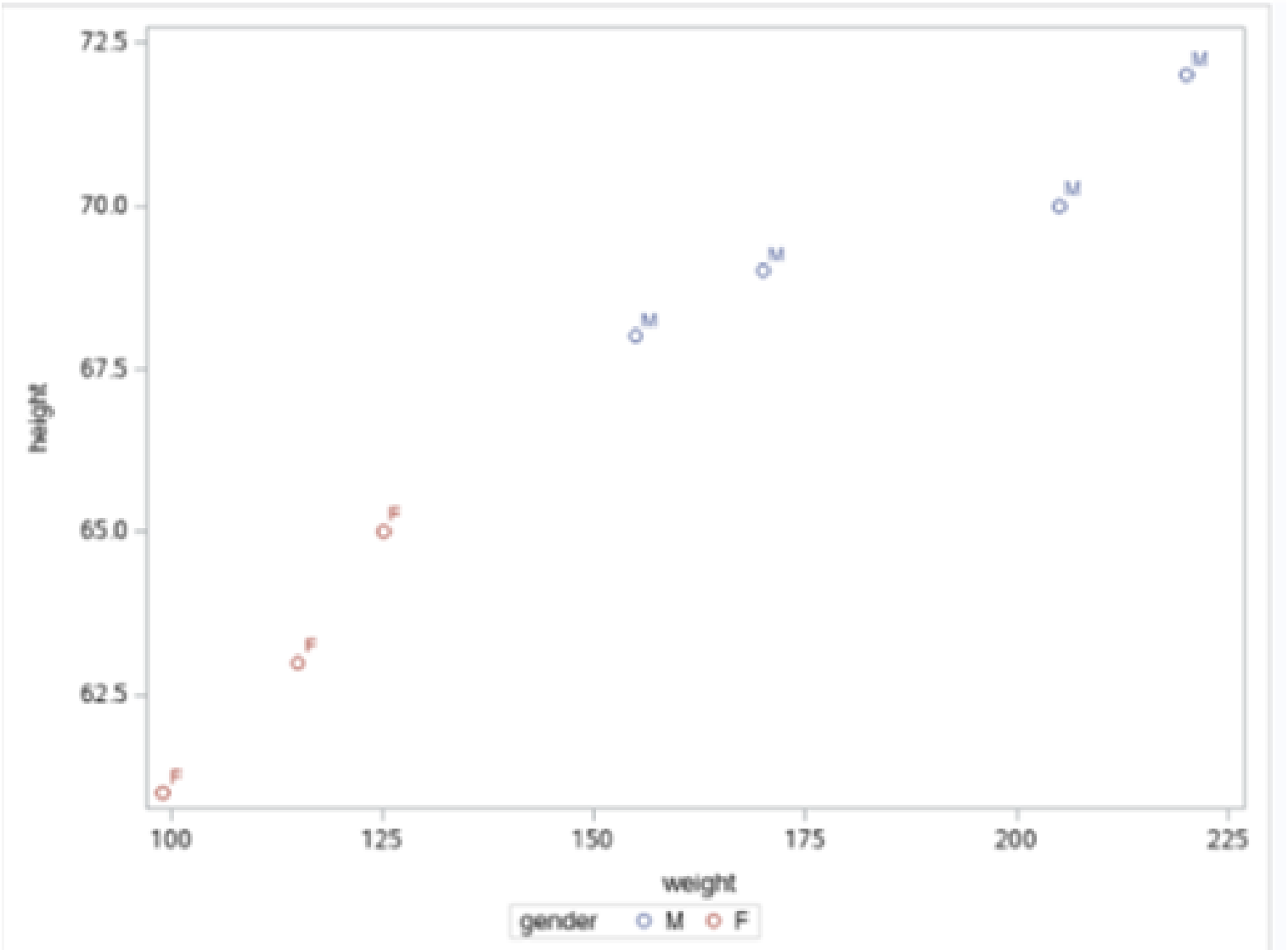
Scatter Plots in SAS (2)

The output (1):



Scatter Plots in SAS (3)

The output (2):



Assumptions

- With only 2 variables X and Y with sample $(x_1, y_1), ..., (x_n, y_n)$, the form of a simple model is:

$$y = \beta_0 + \beta_1 x + \epsilon \quad \text{or}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, ..., n$$

  where β_0 is called intercept and β_1 is called slope.

- The first assumption that we need is the relation ship of X and Y are linear. We can check the sample, to see if the relationship is (approximately) linear or not (the checking can be done using a scatter plot of x and y).
- The second assumption that we need is: $\epsilon_i \sim N(0, \sigma^2)$ where σ^2 is a constant (most of the time it is unknown). We refer to this as **normality assumption** and **constant variance assumption**.
- The third assumption is: ϵ_i and ϵ_j are uncorrelated for all $i \neq j$.
- When model has more than one regressor, we will further assume that the regressors are uncorrelated.

For a given x_i, the assumptions made previously implies:

- $E(y_i) = \beta_0 + \beta_1 x_i, i = 1, ..., n.$
- $Var(y_i) = Var(\epsilon_i) = \sigma^2, i = 1, ..., n.$
- y_i's are independent.
- y_i's follow normal distributions with the above means and variances.
- All y_i's have the same variance.

- The simple model is:

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

- After using the data to estimate the parameters β_0, β_1, we have **fitted model**:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

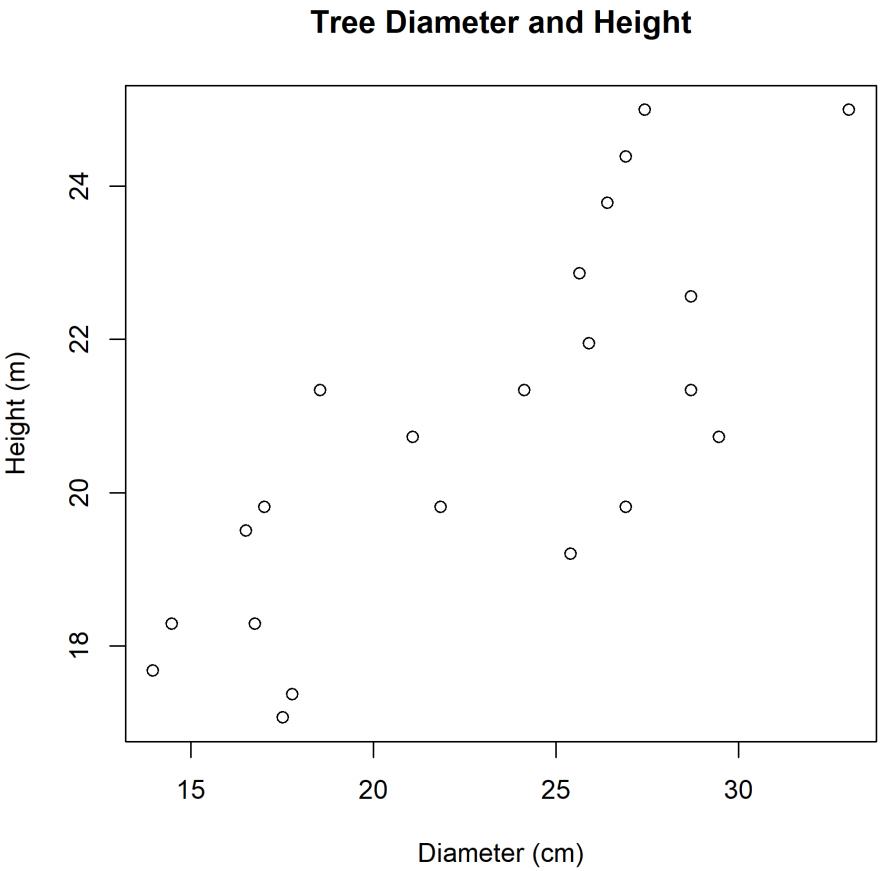
where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of β_0 and β_1.

- However, how to estimate β_0 and β_1? There are few methods: ordinary least squares (OLS) method or maximum likelihood estimation (MLE) method are the popular methods.

- In the scope of this course, we only introduce the OLS method.

Least Squares Estimation (1)

Example (Tree Diameter and Heights)



- Consider the data on the left, which consists of measurements of tree diameter at breast height, and the corresponding heights of those trees.
- The x variable is diameter, and the y variable is tree height.
- If we are willing to assume a simple linear regression of y on x, then we need to find the best fitting line through this scatter plot.
- This is done by using least squares estimation.

Least Squares Estimation (2)

- In least squares, we consider all possible candidate lines.
- For each line, we compute the sum of the squared residuals.
- The line that minimizes the term sum of the squared residuals is picked as the line of best-fit.

Estimation of β_0 and β_1 (1)

- Given a dataset of n observations: $(y_1, x_1), ..., (y_n, x_n)$.

- From the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, ..., n$$

(which is called **sample regression model**), we derive the sum of squares of error:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimators that minimize $S(\beta_0, \beta_1)$ with respect to two parameters β_0 and β_1.

Estimation of β_0 and β_1 (2)

- The estimators/minimizers can be derived by solving:

$$\left[\frac{\partial S}{\partial \beta_0}\right]_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left[\frac{\partial S}{\partial \beta_1}\right]_{\hat{\beta}_0,\hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0$$

- Which lead to

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad (1)$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \quad (2)$$

- (1) and (2) are called the **least-squares normal equations**.

Estimation of β_0 and β_1 (3)

- Denote $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$.

- From (1), we have $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, and replace this $\hat{\beta}_0$ into (2), we have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \dfrac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

Some Frequently Used Notations (1)

- Denote

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- And denote

$$S_{xy} = \sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n} = \sum_{i=1}^{n} y_i (x_i - \bar{x})$$

- Then we have

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

- From the fitted model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, each observed y_i has a corresponding fitted value $\hat{y}_i$. The difference between them is called **raw residual**:

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

Some Frequently Used Notations (2)

- After having the fitted model, we can get the raw residuals, and the residuals sum of squares is defined as

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

- Denote $SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$.

- If we denote $SS_R = \hat{\beta}_1 S_{xy}$ then it can be shown that

$$SS_T = SS_R + SS_{res}$$

where SS_R is the regression sum of squares and has 1 degree of freedom in a simple model. In multiple model with k coefficients (not counting the intercept) then SS_R has $df = k$.

Properties of the OLS estimators

- It can be shown that under the given assumptions, we have

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}}) \quad \text{and}$$

$$\hat{\beta}_0 \sim N\Big(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})\Big).$$

- $\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0.$
- $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i.$
- Sum of the residuals weighted by the corresponding regressor's values (x_i) or the fitted values $(\hat{y}_i)$ is zero:

$$\sum_{i=1}^{n} x_i e_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} \hat{y}_i e_i = 0.$$

- The estimate of σ^2 is obtained from the **residual sum of squares**,

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

- An estimator of σ^2 is usually denoted as $\hat{\sigma}^2$.

- One choice for $\hat{\sigma}^2$ is $SS_{Res}/(n-2)$, which is denoted as MS_{res} - residual mean square. This MS_{res} is an unbiased estimator of σ^2.

- $\sqrt{MS_{res}}$ is called **residual standard error** or standard error of regression.

- The quantity $n-2$ is the number of degrees of freedom for the residual sum of squares SS_{res} (because the two degrees of freedom are associated with the estimate of $\hat{\beta}_0$ and $\hat{\beta}_1$).

Test the Significance of Model (1)

- (1) **Assumptions**: all assumptions as mentioned in slide 11.
- (2) **Hypotheses**:

$$H_0 : \text{all coefficients in the model are 0} \quad \text{vs}$$

$$H_1 : \text{at least one coefficient is non-zero.}$$

Where the **intercept is exclusive** (not counted). Equivalently, we have the hypotheses as:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0.$$

- This tests the significance of the model, that is, is there a linear relationship between the response and the regressor?
- Keep in mind that, $H_0 : \beta_1 = 0$ is rejected meaning: the straight line model is adequate, or a better model could be obtained by adding higher order term(s) of x.

Test the Significance of Model (2)

- (3) In simple model, SS_R has $df = 1$ and SS_{res} has $df = n - 2$, hence the **test statistic** is:

$$F_0 = \frac{SS_R/1}{SS_{Res}/(n-2)} = \frac{MS_R}{MS_{Res}} \sim F_{1,n-2}.$$

- (4) **p-value** is found by the **right area of F_0 under the distribution of** $F_{1,n-2}$.

- (5) At significance level α, $H_0 : \beta_1 = 0$ is rejected if $F_0 > F_{1,n-2}(\alpha)$ or equivalently if p-value is less than α.

- Failing to reject $H_0 : \beta_1 = 0$ implies that there is no linear relationship between y and x.

- Alternatively, if $H_0 : \beta_1 = 0$ is rejected, it implies that x helps in explaining the variability in y.

ANOVA

The test to test the significance of model is called F-test, or Anova test, which usually can be seen from output that has the format similar as below.

Source of Variation	Sum of Squares	DF	MS	F_0
Regression	SS_R	1	MS_R	MS_R/MS_{Res}
Residual	SS_{Res}	n-2	MS_{Res}	
Total	SS_T	n-1		

Test the Significance of a Regressor in Model

- t-test can be used to test the significance of a regressor in model.
- Consider regressor x in the model, which has coefficient β_1. The hypotheses to test the significance of this regressor is:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0. \qquad (1)$$

- The test statistic is

$$t_0 = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{MS_{res}}}.$$

- Under $H_0 : \beta_1 = 0$, then $t_0 \sim t_{n-2}$.
- Denote $t_{n-2}(\alpha/2)$ be the percentile point of a t-distribution with $df = (n-2)$ such that the right area of this point is $\alpha/2$, then the two-sided test (1) rejects H_0 if p-value is less than α or equivalently if

$$|t_0| > t_{n-2}(\alpha/2).$$

- In some other texts, $t_{n-2}(\alpha/2)$ can be denoted by $t_{n-2,\alpha/2}$ or $t_{\alpha/2,n-2}$.

F-test vs t-test in Simple Model

- In simple model, the F-test for the significance of model and the t-test to test the significance of a regressor have the same hypotheses and produce the same p-value (though the test statistics are different).

- The t-test for the significance of the slope with the null $H_0 : \beta_1 = 0$ has test statistic

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{MS_{Res}/S_{xx}}}$$

which has

$$t_0^2 = \frac{\hat{\beta}_1{}^2 S_{xx}}{MS_{Res}} = \frac{\hat{\beta}_1 S_{xy}}{MS_{Res}} = \frac{MS_R}{MS_{Res}} = F_0.$$

- Hence, the t-test and the F-test are identical in the **situation of simple model**.

Coefficient of Determination R^2

- The quantity
$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$
is called the **coefficient of determination**
- SS_T is a measure of the variability in y without considering the effect of the regressor variable x and SS_{Res} is a measure of the variability in y remaining after x has been considered, hence R^2 is often called the proportion of variation explained by the regressor x.

- Since $0 \leq SS_{Res} \leq SS_T$, we have $0 \leq R^2 \leq 1$.

- The larger R^2, the better the fitted model is.

R^2 can be misleading!

- Simply adding more terms to the model will increase R^2
- As the range of the regressor variable increases (decreases), R^2 generally increases (decreases).
- R^2 does not indicate the appropriateness of a linear model

Example 1

Consider Example 1 again, where variable `height` is considered as a regressor and `weight` as the response.
We'll fit a simple model and derive some statistics as following, using R, Python and SAS.

- A fitted model (where the estimate of coefficients are reported)
- The estimate of σ^2
- An F-test for the significance of model
- A t-test for the significance of regressor height, and then verify that the p-value of the F-test and t-test are the same.
- Value of R^2.

Example 1 in R (1)

```
> model1 <- lm(weight ~ height, data = data)
> summary(model1)

Call:
lm(formula = weight ~ height, data = data)

Residuals:
      1       2       3       4       5       6       7
-13.361   8.977   2.595  14.256  -9.553  -9.788   6.873

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -592.645     81.542  -7.268 0.000771 ***
height        11.191      1.218   9.190 0.000256 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 11.86 on 5 degrees of freedom
Multiple R-squared:  0.9441,        Adjusted R-squared:  0.9329
F-statistic: 84.45 on 1 and 5 DF,  p-value: 0.000256
```

Example 1 in R (2)

```
> anova(model1)
Analysis of Variance Table

Response: weight
          Df  Sum Sq Mean Sq F value    Pr(>F)
height     1 11880.3 11880.3  84.451 0.000256 ***
Residuals  5   703.4   140.7
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Example 1 in Python (1)

```
# simple model weight ~ height
import statsmodels.api as sm
weight = data['weight']
n = len(weight)
inter = [1]*n
X = np.column_stack((inter, data['height']))
model1 = sm.OLS(weight, X )
results1 = model1.fit()
print(results1.summary())
```

Example 1 in Python (2)

The output

OLS Regression Results

Dep. Variable:	weight	R-squared:	0.944
Model:	OLS	Adj. R-squared:	0.933
Method:	Least Squares	F-statistic:	84.45
Date:	Sat, 10 Oct 2020	Prob (F-statistic):	0.000256
Time:	00:03:13	Log-Likelihood:	-26.068
No. Observations:	7	AIC:	56.14
Df Residuals:	5	BIC:	56.03
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-592.6446	81.542	-7.268	0.001	-802.255	-383.034
x1	11.1913	1.218	9.190	0.000	8.061	14.322

Omnibus:	nan	Durbin-Watson:	2.161
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.741
Skew:	-0.033	Prob(JB):	0.690
Kurtosis:	1.407	Cond. No.	1.22e+03

Example 1 in Python (3)

The Anova table

```
from statsmodels.formula.api import ols

mod1 = ols('weight ~ height', data=data).fit()
anova1 = sm.stats.anova_lm(mod1, typ=2)
print(anova1)
```

```
                sum_sq   df           F    PR(>F)
height     11880.327238  1.0  84.450853  0.000256
Residual     703.387048  5.0         NaN       NaN
```

Example 1 in SAS: Simple Model (1)

The code to form a model:

```
*Simple model: weight~height;
proc reg data=example1;
  model weight = height;
  output out=analysis P =yhat R =residual STUDENT = resid cookd= cooks H = leverage;
*a dataset named "analysis" is stored with all information about data and the above variables;
*P = fitted, R = raw residuals, student is standardized residuals ;
run;
quit;
```

Example 1 in SAS: Simple Model (2)

The output(1):

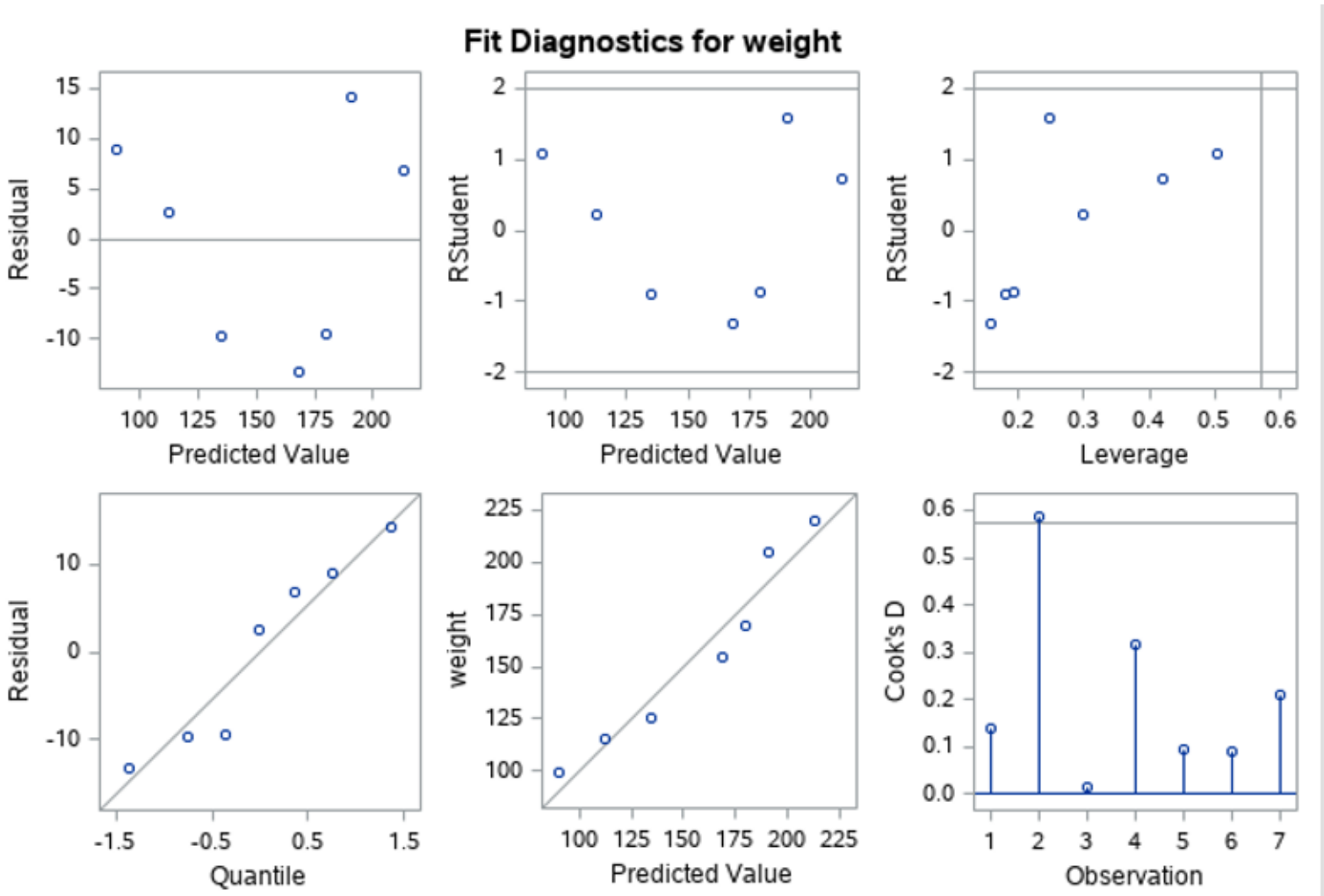
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	11880	11880	84.45	0.0003
Error	5	703.38705	140.67741		
Corrected Total	6	12584			

Root MSE	11.86075	R-Square	0.9441
Dependent Mean	155.57143	Adj R-Sq	0.9329
Coeff Var	7.62399		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-592.64458	81.54217	-7.27	0.0008
height	1	11.19127	1.21780	9.19	0.0003

Example 1 in SAS: Simple Model (3)

The output also delivers the residual plots which will be used in later section to analyze model.



Model with k Regressors (1)

TABLE 3.1 Data for Multiple Linear Regression

Observation, i	Response, y	Regressors			
		x_1	x_2	$\cdots$	x_k
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

Notations

- n: sample size or number of observations available
- k: number of regressors
- y_i: ith observed response
- x_{ij}: ith observation of regressor x_j.
- The regression model can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k.$$

Model with k Regressors (2)

- The model also can be written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \quad i = 1, ..., n.$$

- In a matrix format, if we use

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

  then the regression model above can be written as

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

- We wish to find the estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

- The method to find $\hat{\boldsymbol{\beta}}$ is similar as in simple model, which is OLS or MLE.

LSE of the Regression Coefficients

- The OLS estimation (LSE) of $\boldsymbol{\beta}$ is

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

  provided that the inverse matrix $(\mathbf{X}'\mathbf{X})^{-1}$ exists.

- $(\mathbf{X}'\mathbf{X})^{-1}$ always exists if the regressors are linearly independent, that is, no column of the $\mathbf{X}$ matrix is a linear combination of the other columns.

- The fitted model is then

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}.$$

- The n residuals then can be conveniently written in matrix notation as

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}.$$

Mean and Variance of $\hat{\boldsymbol{\beta}}$

- $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$, that means

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}.$$

- The variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

- Denote $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$, then $\text{Var}(\hat{\beta}_j) = \sigma^2 C_{jj}$ and $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}$.

Estimation of σ^2 (1)

- Similar as in the simple linear regression, an estimator of σ^2 can be derived from the residual sum of squares

$$SS_{Res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e}.$$

- The residual mean square is

$$MS_{Res} = \frac{SS_{Res}}{n-p} \quad \text{which has} \quad E(MS_{Res}) = \sigma^2.$$

- Hence, $\hat{\sigma}^2 = MS_{Res}$ is an unbiased estimator of σ^2.

R^2 and Adjusted R^2

- R^2 is calculated exactly as in simple linear regression, and its meaning remains the same.
- R^2 can be inflated simply by adding more terms to the model (even insignificant terms).
- However, for the similar accuracy, a simpler model is preferred, hence we have adjusted R^2, denoted as R^2_{Adj}-which penalizes you for added terms to the model that are not significant.

$$R^2_{Adj} = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)}.$$

Test the Significance of Model: ANOVA (1)

- The test for significance is a test to determine if there is a linear relationship between the response and **any** of the regressor variables

- The hypotheses are

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0, \quad \text{for at least one } j$$

- Rejection of H_0 implies that at least one of the regressor $x_1, x_2, ..., x_k$ contributes significantly to the model.

- The test procedure is a generalization of the anova used in simple linear regression:

$$SS_T = SS_R + SS_{Res}$$

$$F_0 = \frac{SS_R/k}{SS_{Res}/(n-k-1)} = \frac{MS_R}{MS_{Res}}$$

Test the Significance of Model: ANOVA (2)

- It can be shown that **under H_0** we have

$$F_0 \sim F_{k,n-k-1},$$

hence, the p-value of this test is the right tail probability of F_0 under $F_{k,n-k-1}$ distribution.

- At significance level α, reject H_0 if $F_0 > F_{k,n-k-1}(\alpha)$ or if p-value is less than α.

TABLE 3.4 Analysis of Variance for Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	MS_R/MS_{Res}
Residual	SS_{Res}	$n-k-1$	MS_{Res}	
Total	SS_T	$n-1$		

Test the Significance of Regressors: t-Test

- The hypotheses for testing the significance of any **one regressor** or the coefficient β_j are
$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$$
- Not rejecting H_0 indicates that the regressor x_j can be deleted from the model.
- The test statistics
$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$
where C_{jj} is the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to $\hat{\beta}_j$
- At significance level α, H_0 is rejected if $|t_0| > t_{n-k-1}(\alpha/2)$.
- This test is a partial or marginal test because $\hat{\beta}_j$ depends on all other regressors in the model. Thus, **this is a test of the contribution of x_j given the other regressors in the model**.

Test the Significance of Regressors: F-Test (1)

- There are k regressors in the model

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

and one may want to test **the significance of a subset of regressors** at once.

- Consider a subset of r regressors where $r < k$. Denote $\boldsymbol{\beta}_1$ is $(p - r) \times 1$ and $\boldsymbol{\beta}_2$ is $r \times 1$ are the partition of the regression coefficients:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$$

- We want to test if the group of r regressors with the coefficients in $\boldsymbol{\beta}_2$ is significance or not. The hypotheses should be

$$H_0 : \boldsymbol{\beta}_2 = 0 \quad \text{vs} \quad H_1 : \boldsymbol{\beta}_2 \neq 0.$$

Test the Significance of Regressors: F-Test (2)

- The full model (will full coefficients $\boldsymbol{\beta}$) has the regression sum of squares denoted by $SS_R(\boldsymbol{\beta})$.
- The first group of regressors ($\boldsymbol{\beta}_1$) is added into model, which has regression sum of squares $SS_R(\boldsymbol{\beta}_1)$.
- Given the first group of regressors is already added into the model, the contribution of second group of regressors ($\boldsymbol{\beta}_2$) is then

$$SS_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) = SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta}_1) \quad \text{with } df = r.$$

- The test of significance of r regressors with coefficients $\boldsymbol{\beta}_2$ then has test statistic

$$F_0 = \frac{SS_R(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)/r}{MS_{Res}} \sim F_{r,n-p} \text{ under } H_0$$

  where MS_{Res} is derived from the full model.
- At significance level α, $H_0 : \boldsymbol{\beta}_2 = 0$ is rejected if $F_0 > F_{r,n-p}(\alpha)$ or if p-value is less than α.

Example 1

Consider Example 1 again, where variable `height` (x_1 which has coefficient β_1) and age (x_2 which has coefficient β_2) are considered as regressors and `weight` as the response.
We'll fit a model and derive some statistics as following, using R, Python and SAS.

- A fitted model (where the estimate of coefficients are reported)
- The estimate of σ^2
- An F-test for the significance of model
- A t-test for the significance of regressor height or age given other regressor included or excluded in the model
- Value of R^2 and adjusted R^2, R_a^2.

Example 1 in R (1)

```
> model2 <- lm(weight ~ height +age, data = data)
> summary(model2)
Call:
lm(formula = weight ~ height + age, data = data)

Residuals:
      1       2       3       4       5       6       7
  1.753   5.457   4.220   8.197 -10.038 -13.043   3.454

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -419.0416   121.2740  -3.455   0.0259 *
height         7.9921     2.1078   3.792   0.0192 *
age            1.2532     0.7209   1.738   0.1571
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.01 on 4 degrees of freedom
Multiple R-squared:  0.9682,        Adjusted R-squared:  0.9522
F-statistic: 60.81 on 2 and 4 DF,  p-value: 0.001014
```

Example 1 in R (2)

```
> anova(model2)

Analysis of Variance Table

Response: weight
          Df  Sum Sq Mean Sq F value    Pr(>F)
height     1 11880.3 11880.3 118.603 0.0004036 ***
age        1   302.7   302.7   3.022 0.1571338
Residuals  4   400.7   100.2
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

$SSR(\beta_1|\beta_0) = 11880.3$
$SSR(\beta_2|\beta_0, \beta_1) = 302.7$
$SS_{res} = 400.7.$
The test of significance of age when height is **already added** into model has test statistic $F = 3.022$ and p-value of 0.157.
That test is different from the test to test the significance of age when height is **not added** into model yet.

Example 1 in Python (1)

```
# model weight ~ height + age
import statsmodels.api as sm
X = np.column_stack((inter, data['height'],data['age']))

model2 = sm.OLS(weight, X )
results2 = model2.fit()
print(results2.summary())
```

Example 1 in Python (2)

```
from statsmodels.formula.api import ols

mod2 = ols('weight ~ height + age', data=data).fit()
anova2 = sm.stats.anova_lm(mod2, typ=2)
print(anova2)
```

```
                sum_sq   df          F    PR(>F)
height     1440.170532  1.0  14.377388  0.019240
age         302.710505  1.0   3.021994  0.157134
Residual    400.676543  4.0        NaN       NaN
```

Example 1 in SAS

The code to form model for weight with height and age as regressors with the sum of squares available.

```
* Multiple model: weight~height + age;
proc reg data=example1;
  model weight = height age/SS1;
run;
quit;
*SS1 is the sequential SS_R (Type I SS) as in anova table in R;
```

Example 1 in SAS

The output:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	12183	6091.51887	60.81	0.0010
Error	4	400.67654	100.16914		
Corrected Total	6	12584			

Root MSE	10.00845	R-Square	0.9682
Dependent Mean	155.57143	Adj R-Sq	0.9522
Coeff Var	6.43335		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	-419.04161	121.27400	-3.46	0.0259	169417
height	1	7.99212	2.10777	3.79	0.0192	11880
age	1	1.25323	0.72092	1.74	0.1571	302.71050

The General Concept of Indicator Variables (1)

- Qualitative variables - also known as categorical variables. Qualitative variables do not have a scale of measurement.
- Example of categorical variable: gender (M & F), religion (Buddhism,Taoism, Christianity, Islam, Hinduism), house type (HDB, EC, Condo, Landed), employment status (employed and unemployed), shifts (day, evening, night),...
- We must assign a set of levels to a categorical variable to account for the effect that the variable may have on the response.
- Indicator variables - a variable that assigns levels to the qualitative variable (also known as dummy variables).

Example 1: Variable Gender

Consider variable gender in the Example 1, which has 2 categories F and M. We'll add this variable into the model. Hence, the model we are building now has 3 regressors: height, age and gender.

- We use an indicator variable that takes on the values 0 and 1 to identify the gender's categories. Let

$$x_3 = \begin{cases} 0 & \text{if the observation is from a female, F} \\ 1 & \text{if the observation is from a male, M} \end{cases}$$

- The choice of 0 and 1 to identify the levels of a categorical variable is arbitrary, and we normally just let the software choose.

- Any two distinct values for x_3 (since x_3 has two levels) should be satisfactory, however 0 and 1 are most popular.

Model with Indicator Variable

- The model is
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon.$$
- Consider females, for which $x_3 = 0$, the regression model becomes
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$
- For males, for which $x_3 = 1$ this model becomes:
$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 + \varepsilon \\ &= (\beta_0 + \beta_3) + \beta_1 x_1 + \beta_2 x_2 + \varepsilon. \end{aligned}$$
- Hence, changing from female to male induces a change in the intercept, increasing by β_3 (other coefficients are unchanged and identical).
- Note that we assume that the variance is equal for all levels of the categorical variable.

The General Concept of Indicator Variables (2)

- For categorical variables with a levels, we would need $a-1$ indicator variables.

- For example, in the data, we have response weight, first regressor height (x_1), second regressor age (x_2) and third regressor race which has three categories, A, B, and C. Then two indicator variables (called x_3 and x_4) will be needed:

$$x_3 = \begin{cases} 1 & \text{if the observation is from race A} \\ 0 & \text{if otherwise} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{if the observation is from race B} \\ 0 & \text{if otherwise} \end{cases}$$

The general regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon.$$

From this model, we can have 3 different models for 3 races.

Interaction Term

- Now, suppose that the height (x_1) may relates to the response weight (y) differently for different gender (x_2). Then we may consider to add the interaction term of height and gender to the model. The model $y \sim x_1 + x_2 + x_3 + x_1 x_3$ is then

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \varepsilon.$$

- The model for females is then

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

and the model for males is then

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 + \beta_4 x_1 + \varepsilon = (\beta_0 + \beta_3) + (\beta_1 + \beta_4) x_1 + \beta_2 x_2 + \varepsilon.$$

Example 1 with Gender: in R (1)

```
> data$gender = as.factor(data$gender)
> model3 <- lm(weight ~ height +age + gender, data = data)
> summary(model3)

Call:
lm(formula = weight ~ height + age + gender, data = data)

Residuals:
     1      2      3      4      5      6      7
 1.252  2.281  6.232  5.877 -12.783 -8.513  5.654

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -254.1068   277.3604  -0.916    0.427
height         5.2309     4.6831   1.117    0.345
age            1.5870     0.9204   1.724    0.183
genderM       15.6511    23.2239   0.674    0.549

Residual standard error: 10.77 on 3 degrees of freedom
Multiple R-squared:  0.9723,        Adjusted R-squared:  0.9447
F-statistic: 35.16 on 3 and 3 DF,  p-value: 0.007742
```

Example 1 with Gender: in R (2)

```
> model4 <- lm(weight ~ height +age + gender + height*gender)
> summary(model4)
Call:
lm(formula = weight ~ height + age + gender + height * gender)

Residuals:
     1      2      3      4      5      6      7
 1.434 -2.095  4.190  9.133 -8.001 -2.095 -2.566

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -186.6079   250.9837  -0.744    0.535
height            4.4472     4.1931   1.061    0.400
age               0.8211     0.9945   0.826    0.496
genderM        -502.7463   385.2683  -1.305    0.322
height:genderM    7.6709     5.6928   1.347    0.310

Residual standard error: 9.55 on 2 degrees of freedom
Multiple R-squared:  0.9855,        Adjusted R-squared:  0.9565
F-statistic: 33.99 on 4 and 2 DF,  p-value: 0.02878
```

Example 1 with Gender in Python

```
#create indicator variable for gender:
dummy = pd.get_dummies(data['gender']).values
print(dummy)
print(dummy[:, 1]) # choose this in the model if we want M = 1 and F = 0.
#print(dummy[:, 0]) # choose this in the model if we want F = 1 and M = 0.
```

```
[[0 1]
 [1 0]
 [1 0]
 [0 1]
 [0 1]
 [1 0]
 [0 1]]
[1 0 0 1 1 0 1]
```

```
# model weight ~ height + age + gender
import statsmodels.api as sm
X = np.column_stack((inter, data['height'],data['age'], dummy[:, 1]))

model3 = sm.OLS(weight, X )
results3 = model3.fit()
print(results3.summary())
```

Example 1 with Interaction Term in Python

```
#Model with interaction term height*gender
import statsmodels.api as sm
HG = data['height']*dummy[:, 1]
X = np.column_stack((inter, data['height'],data['age'], dummy[:, 1], HG))

model4 = sm.OLS(weight, X )
results4 = model4.fit()
print(results4.summary())
```

Example 1 with Gender in SAS (1)

To create an indicator for variable gender:

```
*give variable gender an indicator (dummy) variable;
data example1;
set example1;
if gender = "M" then gen= 1; *choose male = 1;
if gender = "F" then gen= 0; *and choose female = 0;
run;
*when fitting a model, variable gen should be used instead of gender;
```

Example 1 with Gender in SAS (2)

Model with indicator for gender and report the partial SS_R.

```
* Multiple model: weight~height + age + gen;
proc reg data=example1;
  model weight = height age gen/SS1;
run;
quit;
```

Example 1 with Gender in SAS (3)

The output:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	12236	4078.57346	35.16	0.0077
Error	3	347.99390	115.99797		
Corrected Total	6	12584			

Root MSE	10.77024	R-Square	0.9723
Dependent Mean	155.57143	Adj R-Sq	0.9447
Coeff Var	6.92302		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	-254.10682	277.36037	-0.92	0.4271	169417
height	1	5.23093	4.68315	1.12	0.3454	11880
age	1	1.58698	0.92038	1.72	0.1831	302.71050
gen	1	15.65107	23.22391	0.67	0.5486	52.68265

Example 1 with Gender in SAS (4)

Model with interaction term of height and gender.

```
*** Model with interaction term;
* we need to create a variable for the interaction term first;
data example1;
set example1;
height_gen = height*gen;
run;
* now can fit a model with the created interaction term above;
proc reg data=example1;
  model weight = height age gen height_gen;
run;
quit;
```

Model Checking Is Needed

- We have made some assumptions about the model, however, we can not guarantee these assumptions are all met at the first step when building the model.
- We only can check for the linearity assumption between y and the regressors, and check the correlation between regressors (by correlation values, by scatter plots).
- After building the model, we need to check if the fitted model satisfying the assumptions made for the error term.
- We usually cannot detect the "problem" by examination of the standard summary statistics, such as the t or F statistics, or R^2, these statistics do not ensure model adequacy.
- Investigate the residuals of the fitted model can help us to detect the assumptions violation.

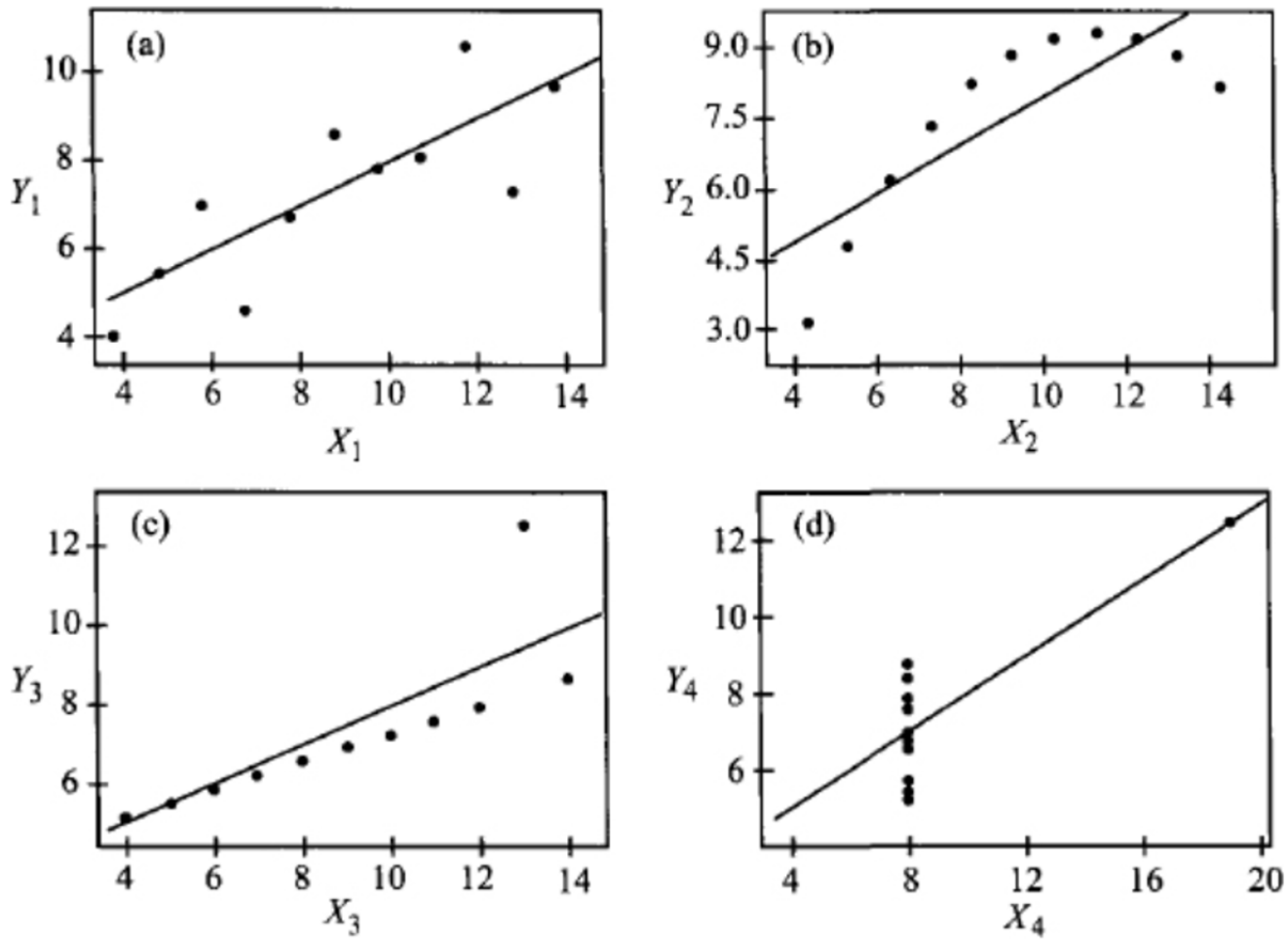


Figure: Four different datasets have same size, which form the same fitted model with same R^2 and same F value.

Mean of Residuals

- We have made assumption about the error term in the model, such that: the errors are uncorrelated; follow $N(0, \sigma^2)$ where σ^2 is a constant.
- After building a model, the raw residuals that we can get from this model is

$$e_i = y_i - \hat{y}_i, \quad , i = 1, ..., n.$$

- The residuals have zero mean and their approximate average variance can be estimated by

$$\frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_{Res}}{n-p} = MS_{Res}.$$

- The residuals are in fact not independent (since they should sum up to 0), however when $p << n$ the nonindependence has little effect on their use for model adequacy checking.

Variance of Residuals

- We define matrix $\mathbf{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and have

$$\hat{\boldsymbol{y}} = \mathbf{H}\boldsymbol{y}.$$

  Matrix **H** is called the hat matrix or projection matrix, which is an important matrix in regression analysis.

- The diagonal value h_{ii} is called the **leverage** value for the ith observation, it is the weight given to y_i in determining the ith fitted values $\hat{y}_i$.

- It is shown that, the **variance for each residual** actually is

$$\text{Var}(e_i) = \sigma^2(1 - h_{ii}), \quad \text{and} \quad \text{Cov}(e_i, e_j) = -\sigma^2 h_{ij}.$$

Standardized Residuals

- **Standardized residuals** is defined as

$$\frac{e_i}{\sigma\sqrt{(1-h_{ii})}}, \quad i = 1, ..., n.$$

- Since σ is unknown, hence we can estimate it by $\sqrt{MS_{Res}}$.

- Substitute $\sqrt{MS_{Res}}$ into the standardized residuals, we get

$$r_i = \frac{e_i}{\sqrt{MS_{Res}}\sqrt{(1-h_{ii})}}, \quad i = 1, ..., n$$

  which is called internally studentized residuals, or commonly called by standardized residuals (SR).

- If all the assumptions made for the model are met, we **expect the SR independently follow** $N(0, 1)$. Hence, the plots of these SR can help use to check if the assumptions are violated or not.

QQ Plot of Standardized Residuals

- We can obtain the QQ plot of the SR to check if the SR follow normal distribution.
- **We expect**: this plot resembles a (nearly) straight line with an intercept of 0, and a slope 1.
- A common defect that shows up on the normal probability plot is the occurrence of some large residuals. This may indicate that the corresponding observations are outliers.

Scatter plot of SR vs Fitted Values

- We have assumed that the residuals are uncorrelated with the fitted values, hence **we expect**: the points are scattered randomly around 0, within the horizontal band of (-3,3), somehow resembles the Figure SRvF(a).
- Plots of residuals vs fitted values that resemble any of the patterns in the Figure SRvF(b-d) are indicating model deficiencies.
- **Figure SRvF(b-c) indicate that the variance of the errors is not constant**. The outward-opening pattern in (b) implies that the variance is an increasing function of y (an inward-opening funnel indicates the decreasing trend).
- The double-bow pattern in Figure SRvF(c) often occurs when y is a proportion between 0 and 1. The variance of a binomial proportion near 0.5 is greater than one near 0 or 1.
- A curved plot such as in the **Figure SRvF(d) indicates nonlinearity**. This could mean that a higher order term of the regressor variables is needed in the model. Transformation on the regressor and/or the response y may also be helpful in these cases.

Possible Scatter Plot of S.R vs Fitted Values

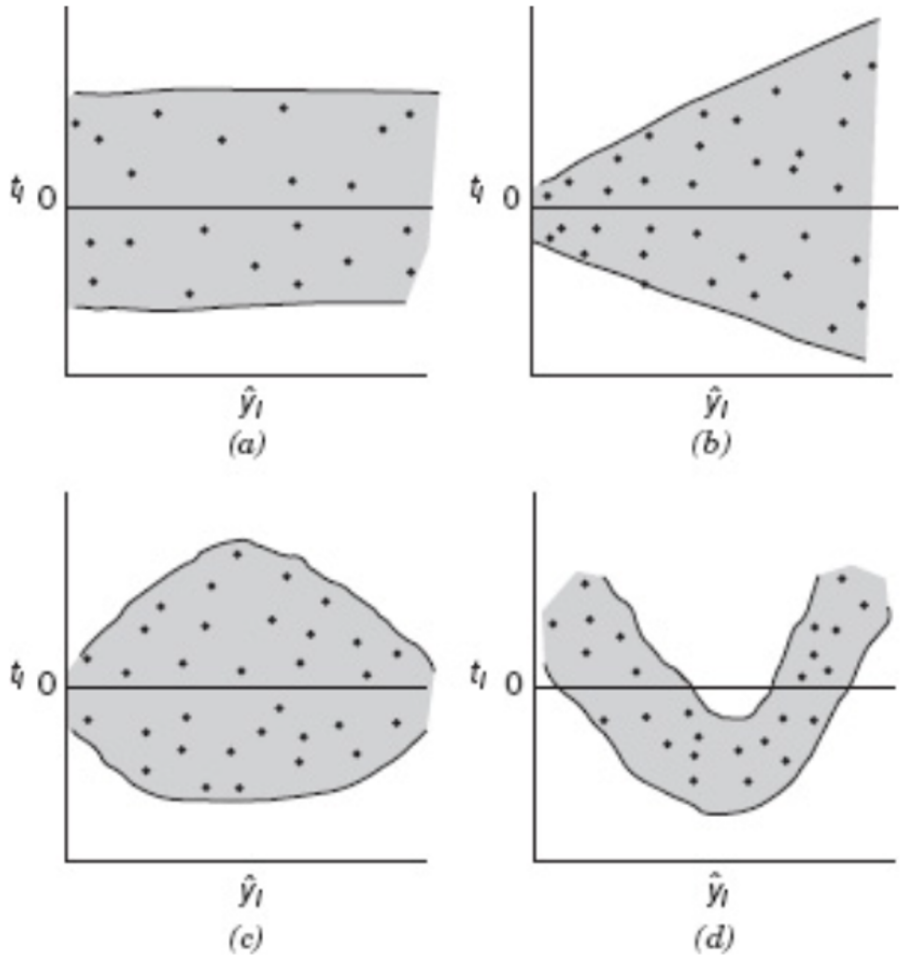


Figure: SRvF: Patterns for residuals plots: (a) satisfactory; (b) funnel; (c) double bow; (d) nonlinear.

Outliers and Treatment

- An outlier is an extreme observation, one that is considerably different from the majority of the data.
- Residuals that are considerably larger in absolute value than the others (say 3 or 4 standard deviations from the mean) indicate potential y space outliers.
- Residual plots against fitted values and the normal probability plot are helpful in detecting outliers.
- The outliers could be the result of faulty measurement or analysis, incorrect recording data, or failure of measuring instrument. We call them "bad" outliers.
- For bad outliers, we can correct them if possible, or delete them from the data set.
- Sometimes the outlier is an unusual but perfectly plausible observation. Deleting these points to improve the fit of the model can be dangerous, as it can give us a false sense of precision in estimation/prediction. Hence, we should keep them in the data.

Influential Points

- Some observations may have more influence to the model building than other observations. If the influence is significant, we call those points as influential points.

- To detect the influential points, one may consider the hat diagonal values h_{ii} in conjunction with the standardized residuals. Observations with large hat diagonals ($h_{ii} > 2(k+1)/n$) and large residuals are likely to be influential.

- One may use Cook's distance to measure the influence of a data point, where the Cook's distance for ith observation is

$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}.$$

- Any data point that has Cook's distance larger than 1 can be considered as an influential point.

Residuals Analysis in R (1)

We consider Example 1, with a model for response weight depends on height (x_1), which was named as "model1" in previous R code.

```
> model1$res #raw residuals of model1
         1          2          3          4          5          6
-13.361446   8.977410   2.594880  14.256024  -9.552711  -9.787651
> rstandard(model1) #standardized residuals of model1
         1          2          3          4          5          6
-1.2266800  1.0752884  0.2614343  1.3851132 -0.8955944 -0.9108609  0
>
```

Residuals Analysis in R (2)

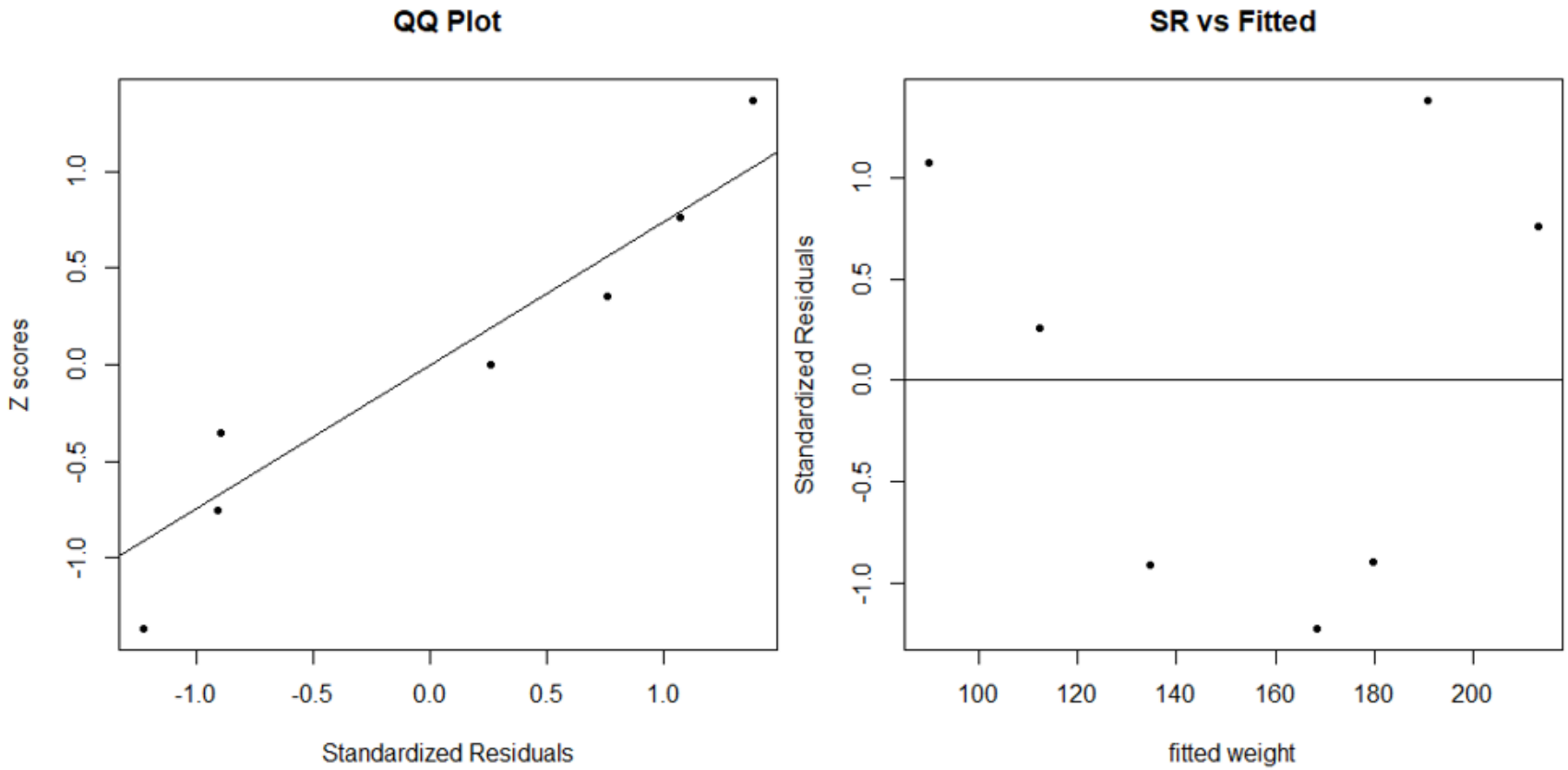


Figure: The QQ plot of SR (left) and the plot of fitted values vs SR (right)

Residuals Analysis in R (3)

```
> p = 2 # simple model weight ~ height has p = 2
> n = length(weight) # sample size
> x<-cbind(c(rep(1,n)),height)
> hat<-x%*%solve(t(x)%*%x)%*%t(x) # hat matrix
> diag(hat)
[1] 0.1566265 0.5045181 0.2996988 0.2469880 0.1912651 0.1792169 0.42
> which(diag(hat)>2*p/n) # to find if there is leverage point
integer(0)
> cooks.distance(model1) # Cooks distance
          1           2           3           4           5           6
0.13972622 0.58866582 0.01462498 0.31464034 0.09484668 0.09057839 0.
```

Residuals Analysis for Example 1

- There is no large SR, hence data do not have outlier.
- QQ plot of SR is quite good.
- The plot of fitted value vs SR is within the horizontal band of (-2,2) which is good. However, it is not clear if the points follow a quadratic shape since data size is very small ($n = 7$). One may try to add the second order term of height ($height^2$) into the model to see if the plot improves.
- From the calculation of h_{ii}, data do not have any leverage point.
- The Cook's distance values are all small. Therefore, data do not have any influential point.

Predicted Values and Raw Residuals in Python

```
#predicted values of model1:
print( results1.fittedvalues)
```

```
0    168.361446
1     90.022590
2    112.405120
3    190.743976
4    179.552711
5    134.787651
6    213.126506
dtype: float64
```

```
# the raw residuals of model1
print(results1.resid)
```

```
0   -13.361446
1     8.977410
2     2.594880
3    14.256024
4    -9.552711
5    -9.787651
6     6.873494
dtype: float64
```

Residuals, Leverage and Cook's Distance in Python

```
analysis = results1.get_influence()
# many information about model stored in this object.
#the standardized residuals:
SR = analysis.resid_studentized_internal
print(SR)
```

```
[-1.22668004  1.07528838  0.26143431  1.38511323 -0.89559435 -0.91086091
  0.76205059]
```

```
#Leverage (hat values)
leverage = analysis.hat_matrix_diag
print(leverage)
```

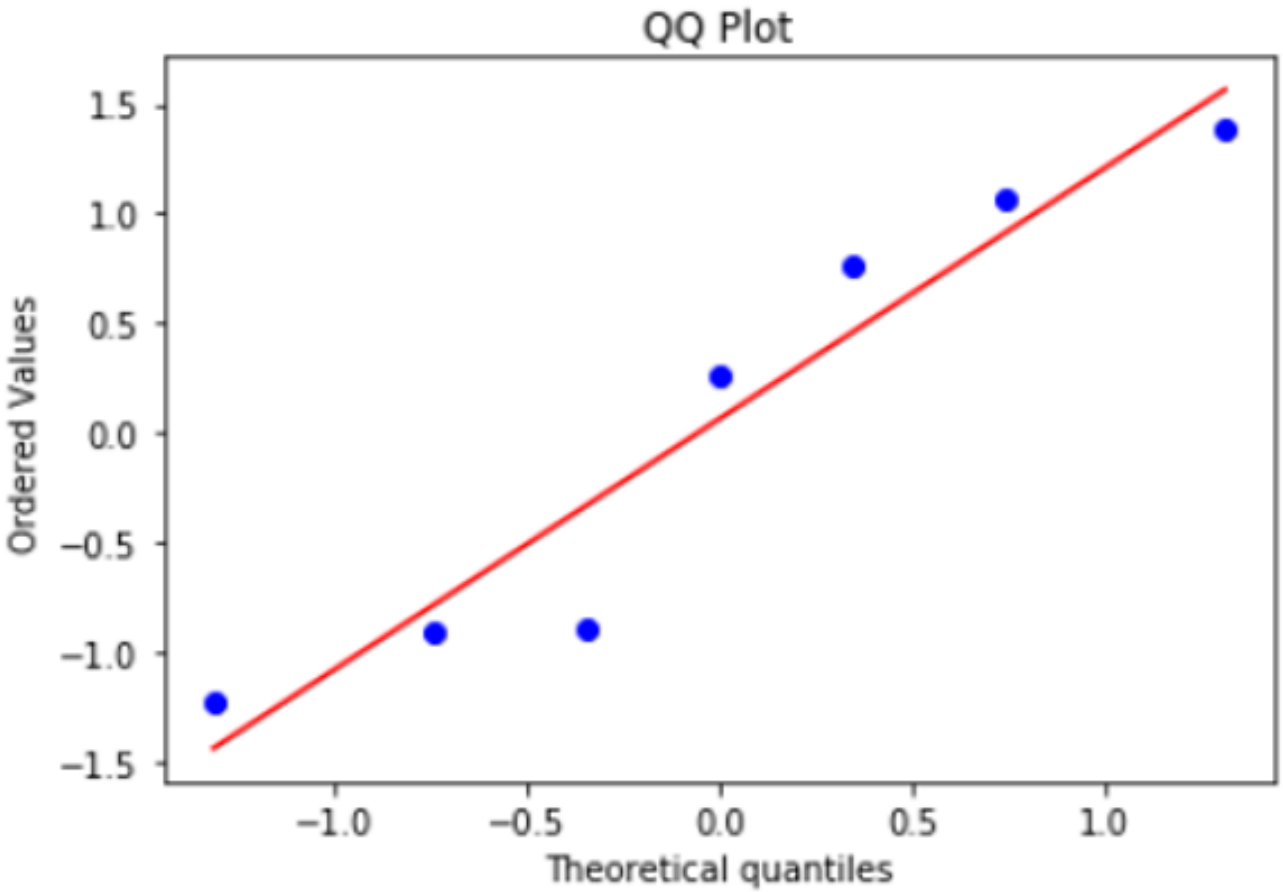
```
[0.15662651 0.50451807 0.2996988  0.24698795 0.19126506 0.17921687
 0.42168675]
```

```
#Cook's D values (and p-values) as tuple of arrays
cooks_d, p = analysis.cooks_distance
print(cooks_d)
```

```
[0.13972622 0.58866582 0.01462498 0.31464034 0.09484668 0.09057839
 0.21172123]
```

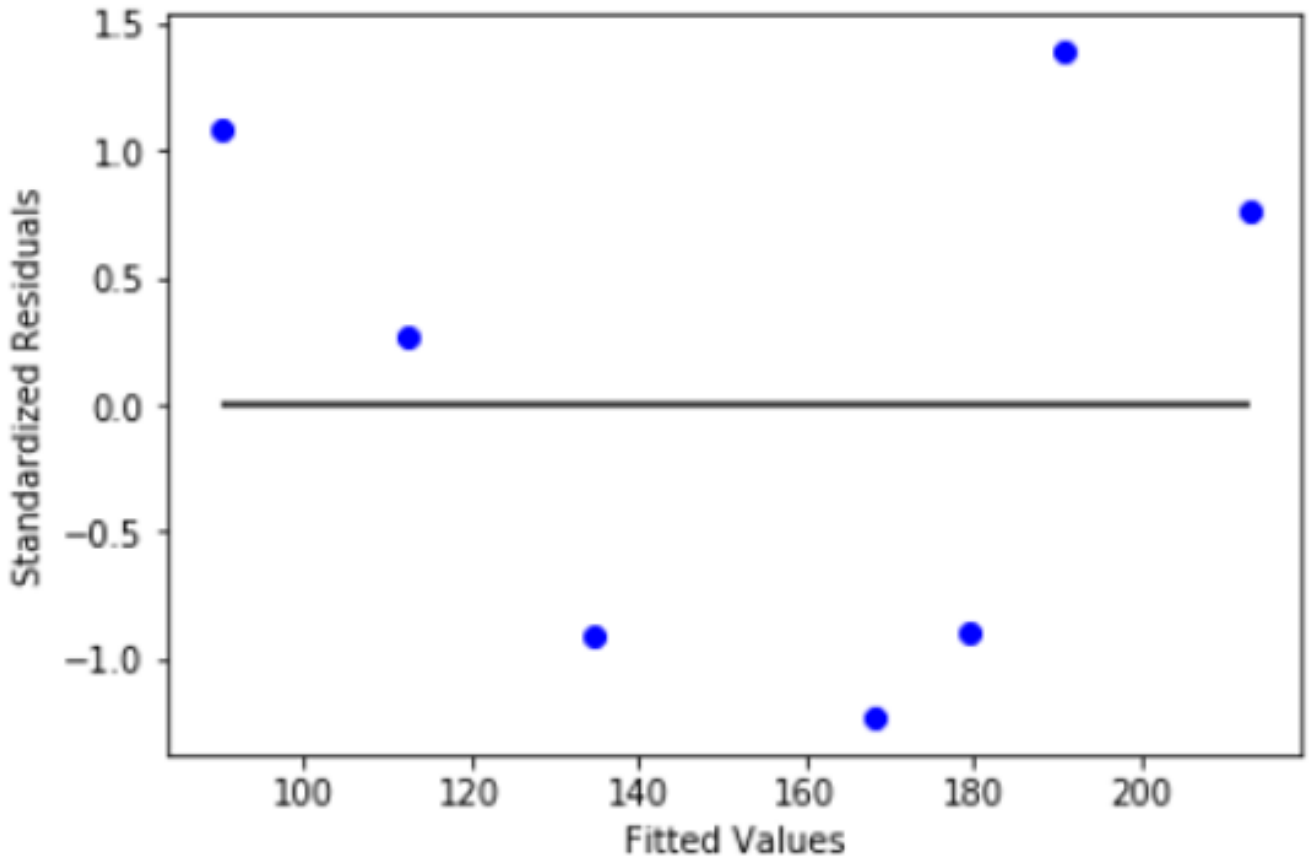
QQ Plot of Residuals in Python

```
########################  QQ PLOTS of standardized residuals
scipy.stats.probplot(SR, dist="norm", plot=pyplot)
pyplot.title('QQ Plot')
pyplot.show()
```



Plot of Residuals vs Fitted Values in Python

```
import matplotlib.pyplot as pyplot
pyplot.scatter(results1.fittedvalues, SR,color='b')
pyplot.xlabel('Fitted Values')
pyplot.ylabel('Standardized Residuals')
pyplot.hlines(0, xmin = min(fitted), xmax = max(fitted) )
pyplot.title('')
pyplot.show()
```



Residuals Analysis in SAS

- Refer back to the slide of the code to create a simple model for weight depending on height, which shows how to get the standardized residuals, the leverage values and the Cook's distance for each observation.

- When a model is fitted, the residual plots (QQ plot of SR; SR vs fitted values; SR vs regressors) are created by default.

Possible Transformations

- The assumptions might be violated, such as: the response might not have a linear association with the regressor x (the scatter plot or the plot of SR vs fitted shows curvy shape); The variance might not be constant (funnel shape is shown in the plot of SR vs fitted).

- Applying transformation in the model might help in those cases, such as:
  - Transform the regressor x (take $\log x$ or $1/x$ be the regressor instead of x,...);
  - Transform the response (take $\log y$, or y^{λ} with some suitable λ be the response instead of original y).
  - Adding higher order term of x also can help.