Psychol. Statistics using R

TAIPEI MEDICAL UNIVERSITY

臺北醫學大學

TAIPEI MEDICAL UNIVERSITY

Regression

Changwei W. Wu, Ph.D.

Graduate Institute of Mind, Brain and Consciousness
Research Center of Brain and Consciousness
Taipei Medical University

Statistics

1. **Linear Regression**
   - Concept and Concerns
2. **Multiple Regression**
   - Model comparison and interpretation
   - Assumptions
3. **[R] Assumption check**

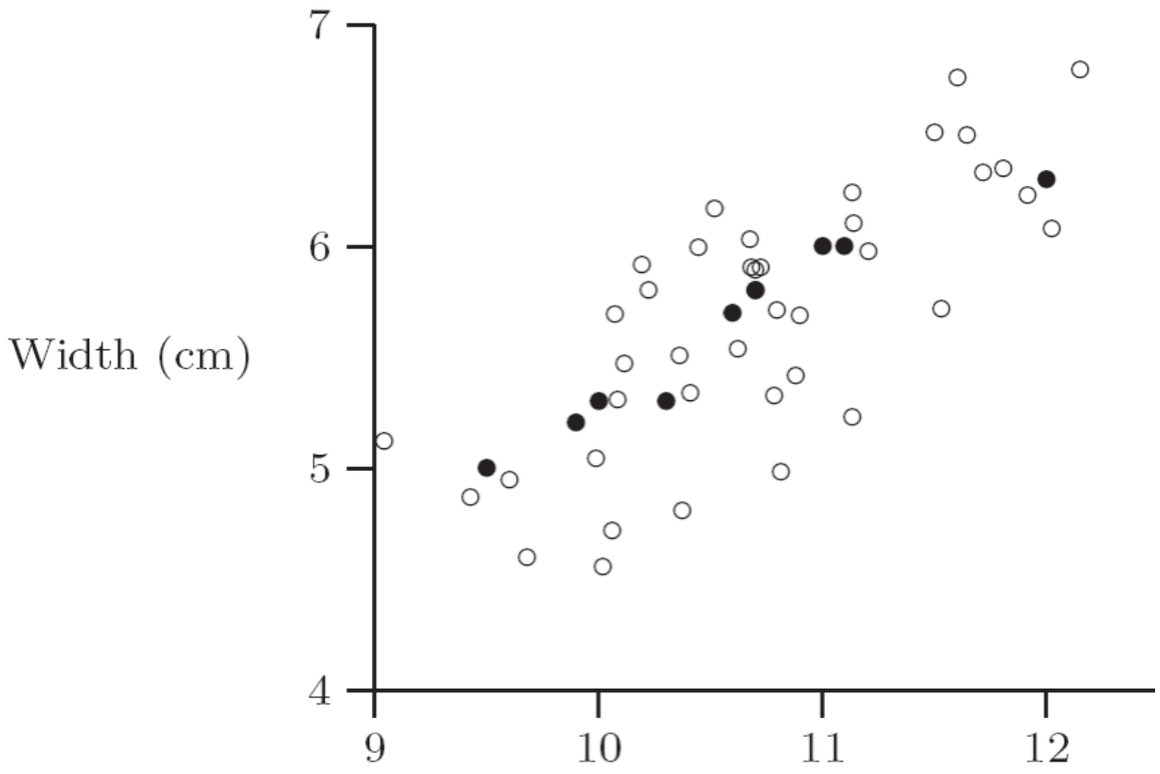
Theories

Practice

Assignment

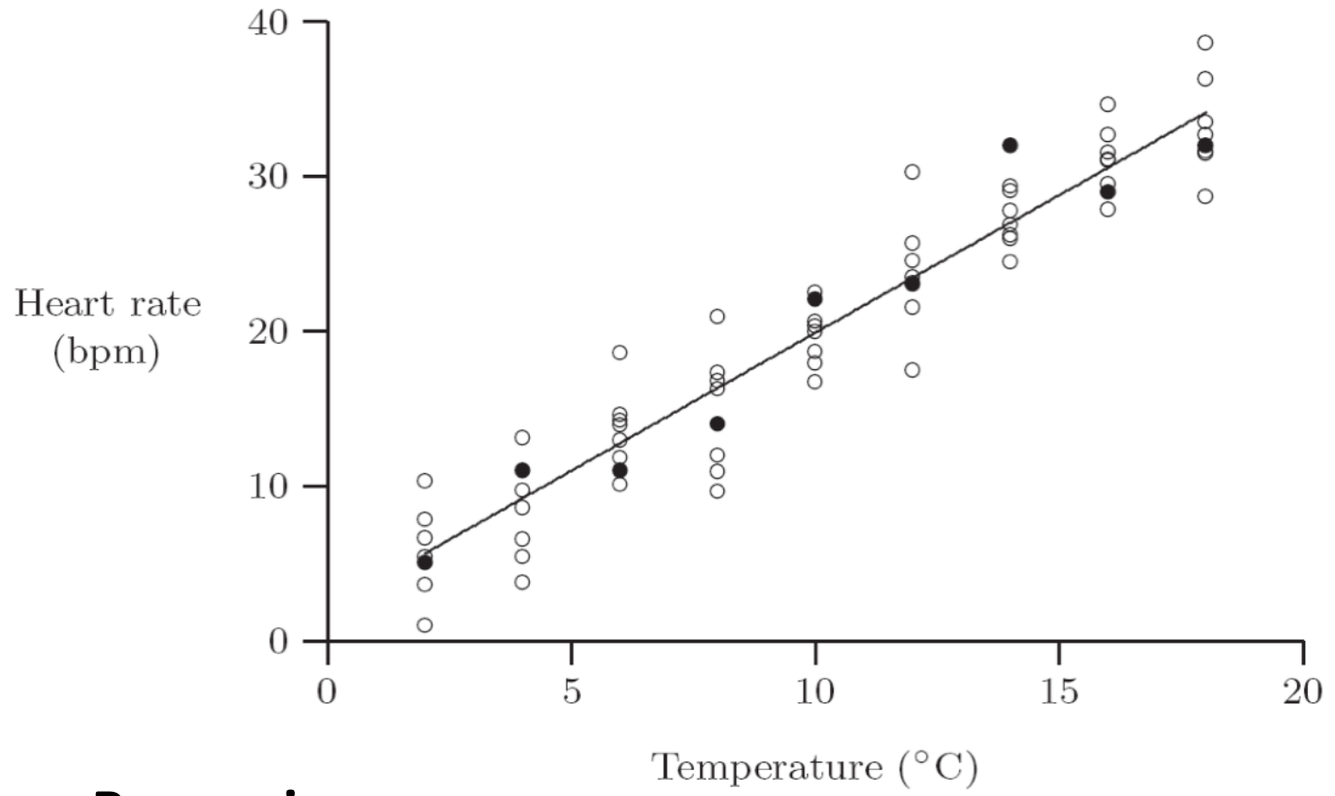
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

Correlation vs. Regression



Correlation:
X & Y are not controlled.
Observe their relations.
(Not to estimate residuals)

Animal	Length (cm)	Width (cm)
1	10.7	5.8
2	11.0	6.0
3	9.5	5.0
4	11.1	6.0
5	10.3	5.3
6	10.7	5.8
7	9.9	5.2
8	10.6	5.7
9	10.0	5.3
10	12.0	6.3



Regression:
Using **X** (controlled) to predict **Y**.
(to **estimate unexplained error**)

Recording number	X Temperature (°Celsius)	Y Heart rate (bpm)
1	2	5
2	4	11
3	6	11
4	8	14
5	10	22
6	12	23
7	14	32
8	16	29
9	18	32

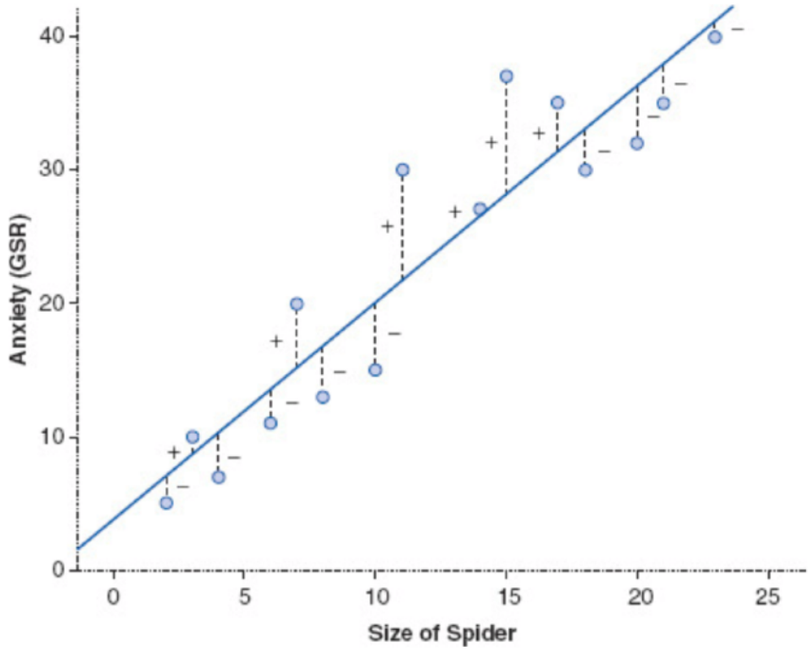
TAIPEI MEDICAL UNIVERSITY 1960 臺北醫學 TAIPEI MEDICAL UI

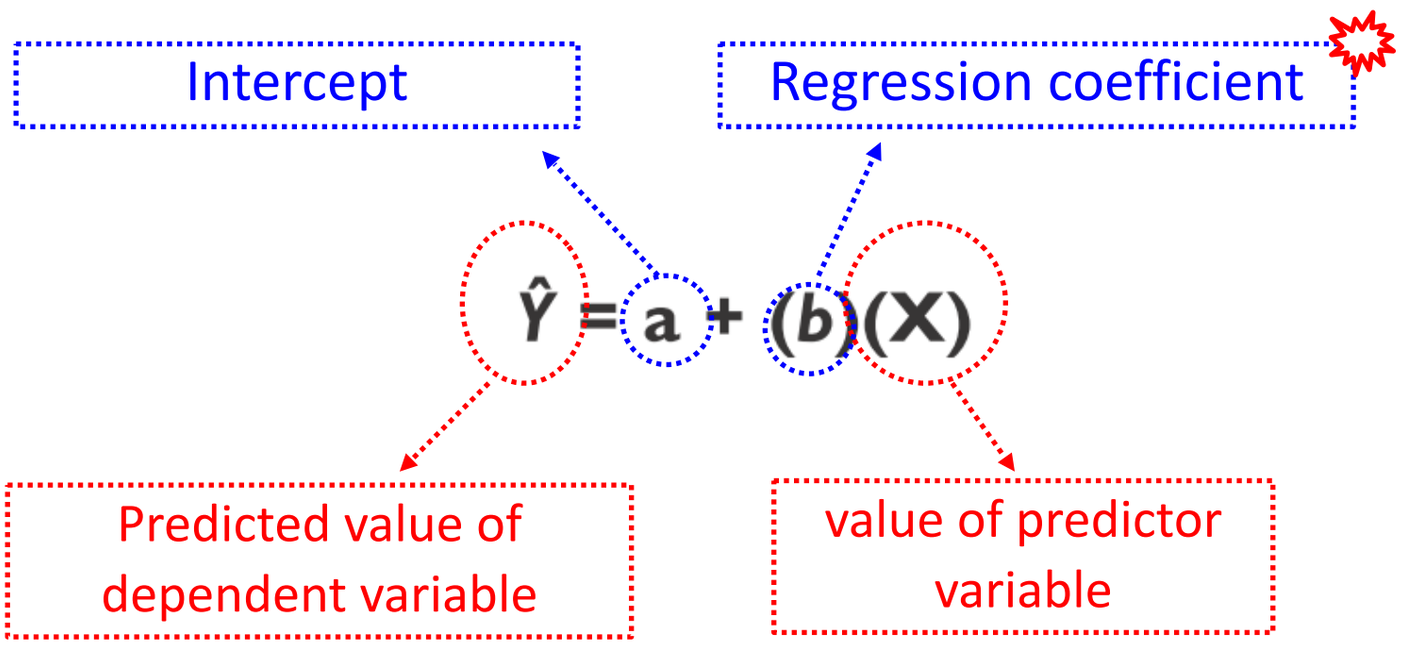
Regression (Prediction)

- Express **the linear relationship between the two variables** (simple bivariate regression).

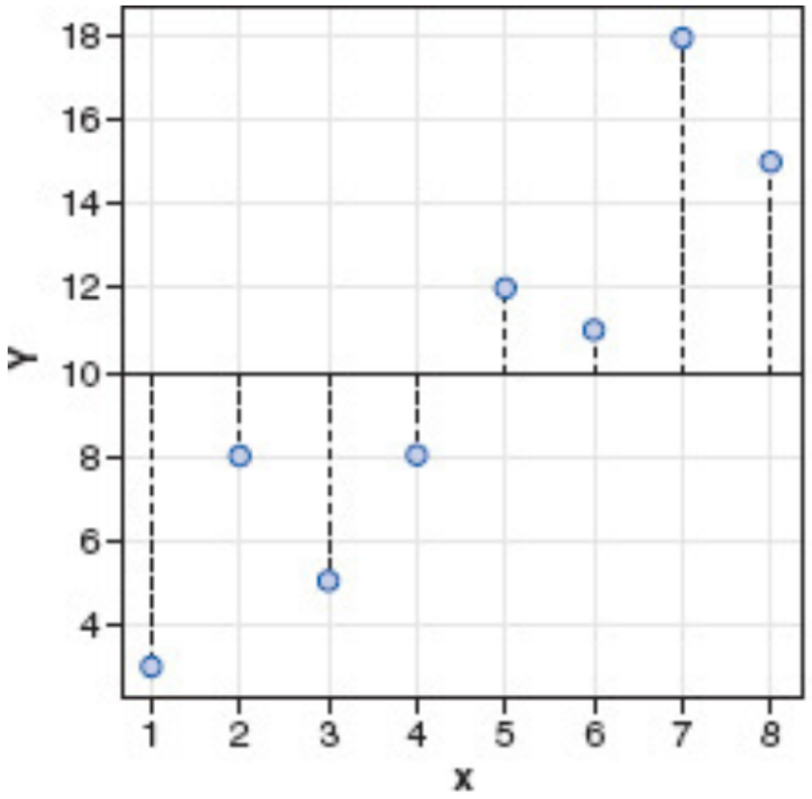
- This formula can be used to predict a value for Y from a value for X.
  - ➢Predictor variable (X):
    variable that is used to make a prediction (variable predicted from)
  - ➢Dependent variable (Y):
    variable that is predicted (variable predicted to)

$$Y = X\beta + \epsilon$$

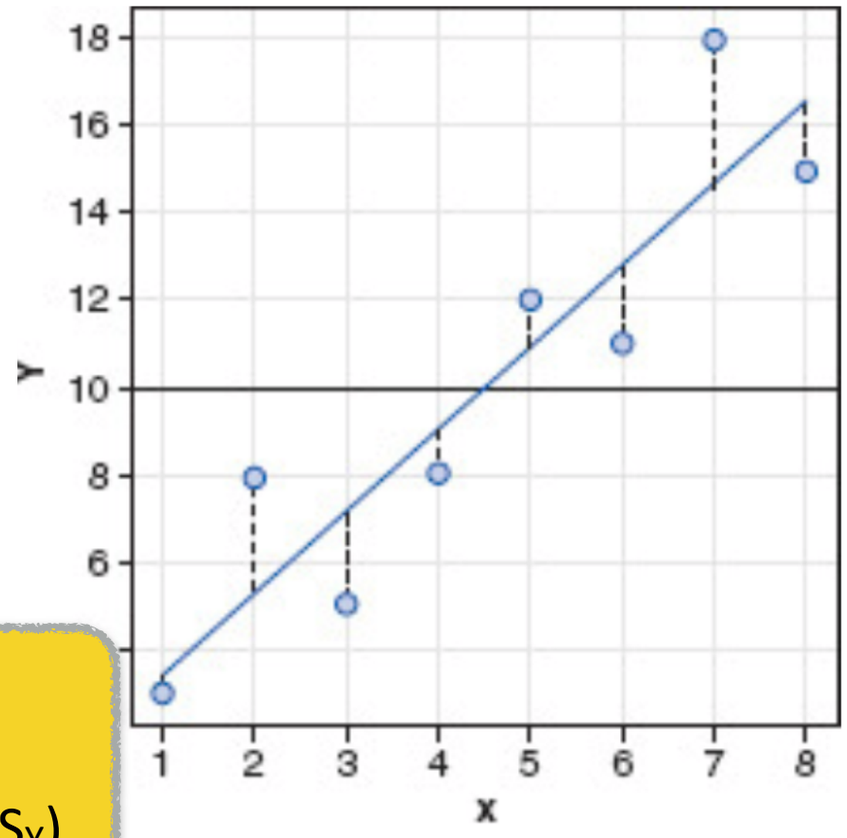




Regression



SS_T uses the differences between the observed data and the mean value of *Y*



SS_R uses the differences between the observed data and the regression line

SS_T

Total variance in the data (SS_Y)

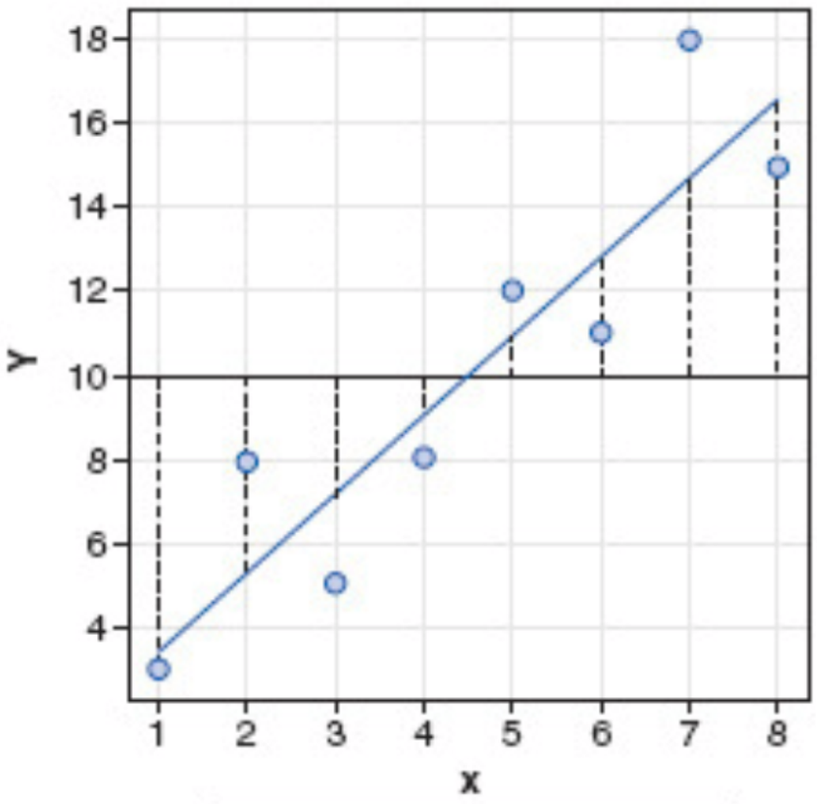
SS_R

Variance explained by regression model

SS_E

Unexplained variance (Residue)

$$Y = X\beta + \epsilon$$



SS_M uses the differences between the mean value of *Y* and the regression line

.1. The Sum of Squares (total) is

$$\sum (Y_i - \overline{Y})^2 = \sum (\hat{Y}_i - \overline{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

Sum of Squares total SS_{Total}	=	Sum of Squares due to regression SS_R	+	Sum of Squares residual or error SS_E.

Regression Alternative

SS_T

Total variance in the data (SS_Y)

SS_R

Variance explained by regression model

SS_E

Unexplained variance (Error)

.1. The Sum of Squares (total) is

$$\sum (Y_i - \overline{Y})^2 = \sum (\hat{Y}_i - \overline{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

Sum of Squares total	=	Sum of Squares due to regression	+	Sum of Squares residual or error
SS_{Total}		SS_R		SS_E.

$$SS_R = b^2 SS_X = b SS_{XY}$$

$$SS_E = \sum (Y_i - \hat{Y}_i)^2 = SS_{Total} - SS_R$$

$$b = \frac{SS_{XY}}{SS_X}$$

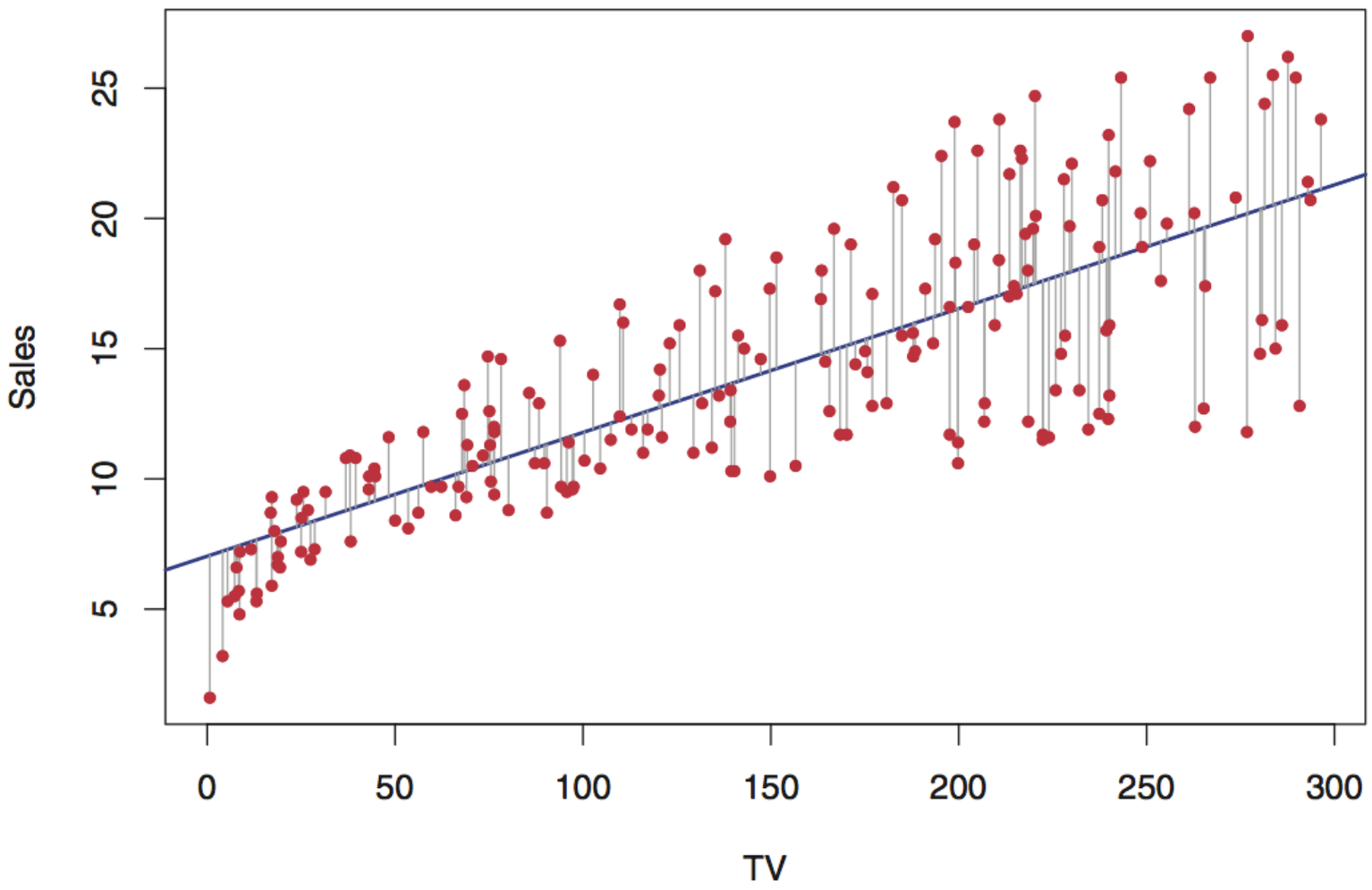
<u>Coefficient of Determination</u>:

The proportion of how much total variance can be explained by the regression model.

$$r^2 = \frac{SS_R}{SS_{Total}} = \frac{(SS_{XY})^2}{SS_X \cdot SS_Y}$$

臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

Linear Regression



$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}.$$

- ***Is there a relationship between advertising budget and sales?***
- *How strong is the relationship between advertising budget and sales?*
- *How accurately can we predict future sales?*

$H_0 : \beta_1 = 0$ There is no relationship between X and Y

$H_a : \beta_1 \neq 0,$ There is some relationship between X and Y.

For each predictor

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

Significance of Regression

The global test of significance for regression is an ANOVA. The hypotheses are

- H_0: The variation in Y is *not* explained by a linear model, i.e., $\beta = 0$.
- H_a: A significant portion of the variation in Y is *explained* by a linear model, i.e., $\beta \neq 0$.

For the entire model (global):

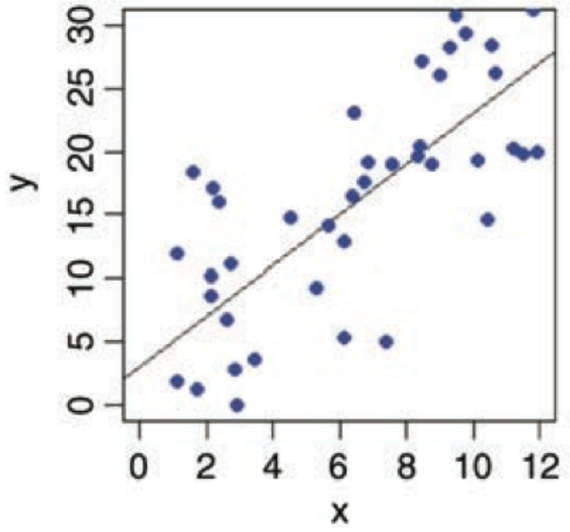
Source of variation		SS	DF	MS	E(MS)	F
Regression	Predictor	$\mathrm{SS_R}$	1	$\mathrm{MS_R}$	$\sigma_Y^2 + \beta^2 \mathrm{SS}_X$	$\frac{\mathrm{MS_R}}{\mathrm{MS_E}}$
Error	Residuals	$\mathrm{SS_E}$	$n-2$	$\mathrm{MS_E}$	σ_Y^2	
Total		$\mathrm{SS_{Total}}$	$n-1$			

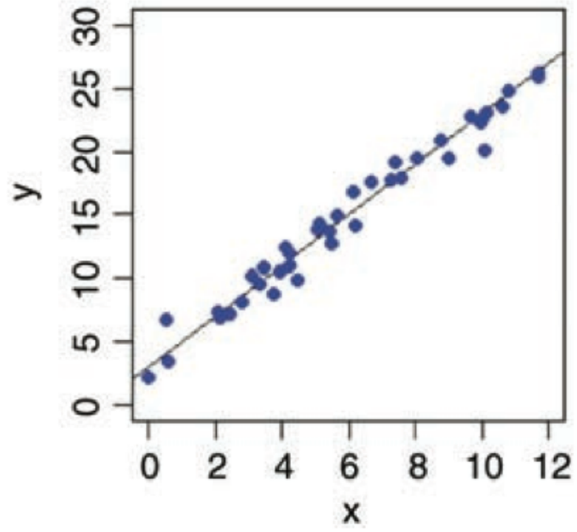
Accuracy of Regression

.1. The Sum of Squares (total) is

$$\sum (Y_i - \overline{Y})^2 = \sum (\hat{Y}_i - \overline{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

Sum of Squares total SS_{Total}	=	Sum of Squares due to regression SS_R	+	Sum of Squares residual or error SS_E.





Questions	*Verification*	*Practical solution*
1. Does the model fit the data well?	***Check the coefficient of determination***	$r^2 = \frac{SS_R}{SS_{Total}}$
2. Is it influenced by small number of cases?	***Case-wise diagnostics***	Detect outliers by case-wise **influence measures**
3. Can the model generalize to other samples?	***Meet the Assumptions of Regression***	*Multiple strategies* for generalizing the results.

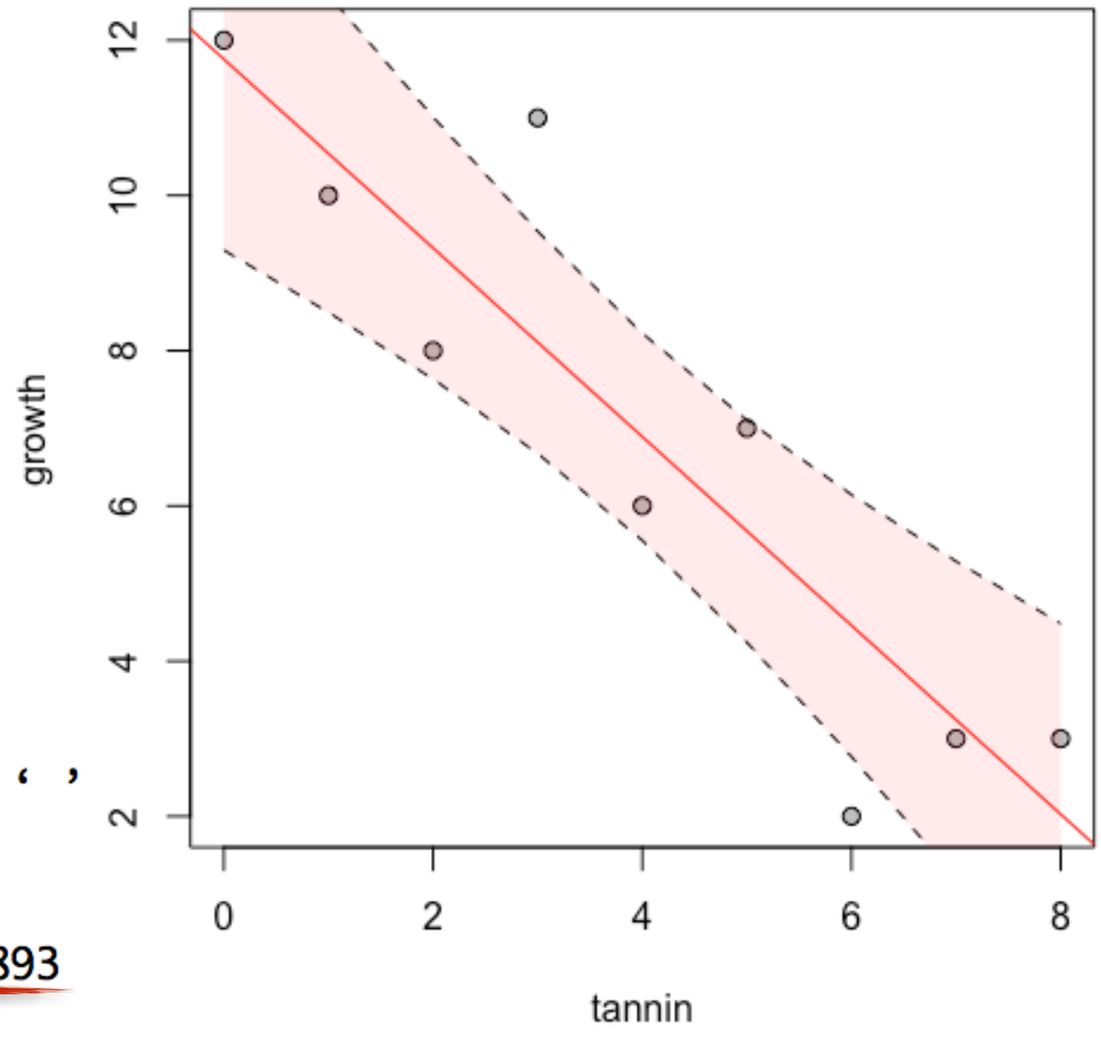
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

DEMO

Tannin vs. Growth

③ Testing

- **Practice:**
  - *model <- lm(growth~tannin)*
  - *summary(model)*
  - *summary.aov(model)*



- **Results:**

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.7556     1.0408  11.295 9.54e-06 ***
tannin       -1.2167     0.2186  -5.565 0.000846 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 1.693 on 7 degrees of freedom
Multiple R-squared:  0.8157,    Adjusted R-squared:  0.7893
F-statistic: 30.97 on 1 and 7 DF,  p-value: 0.0008461
```

TAIPEI MEDICAL UNIVERSITY 臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

Assumptions of Regression

- linear relationship between Y and predictors.

$$\hat{Y} = a + (b)(X)$$

◉ ***Types of variable (X, Y)***

- Predictors X should be quantitative or categorical (no need to be normally distributed);
- Outcome variables Y must be quantitative, continuous and independently sampled.

◉ ***without Multicollinearity (X)***

➡ Predictors should NOT be perfectly correlated with each other. (VIF)

◉ ***Homoscedasticity (σ_ϵ):***

➡ At each level of predictor (X), the variance of Residuals should be constant.

◉ ***Lack of Autocorrelation (σ_ϵ):***

➡ For any 2 observations, the residuals are uncorrelated.

◉ ***Errors are Normally distributed (σ_ϵ):***

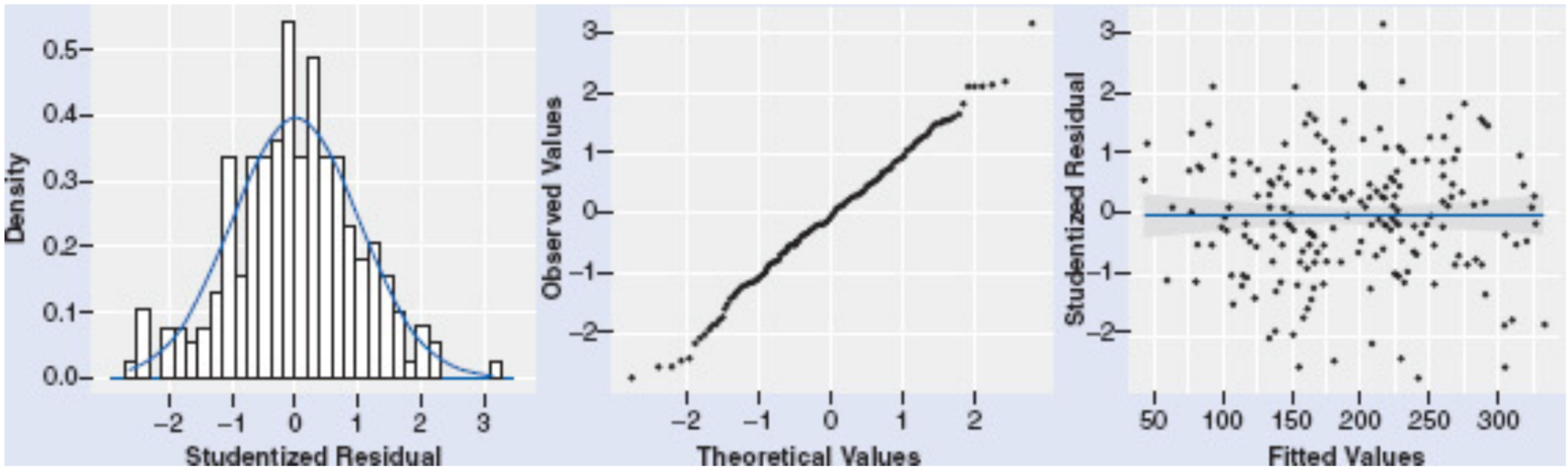
➡ Residuals (not predictor) are random, normally distributed with a mean of 0.

Normally distributed residual error (σ)

$$\epsilon \sim N(0, \sigma^2)$$

Assumption Check of Regression

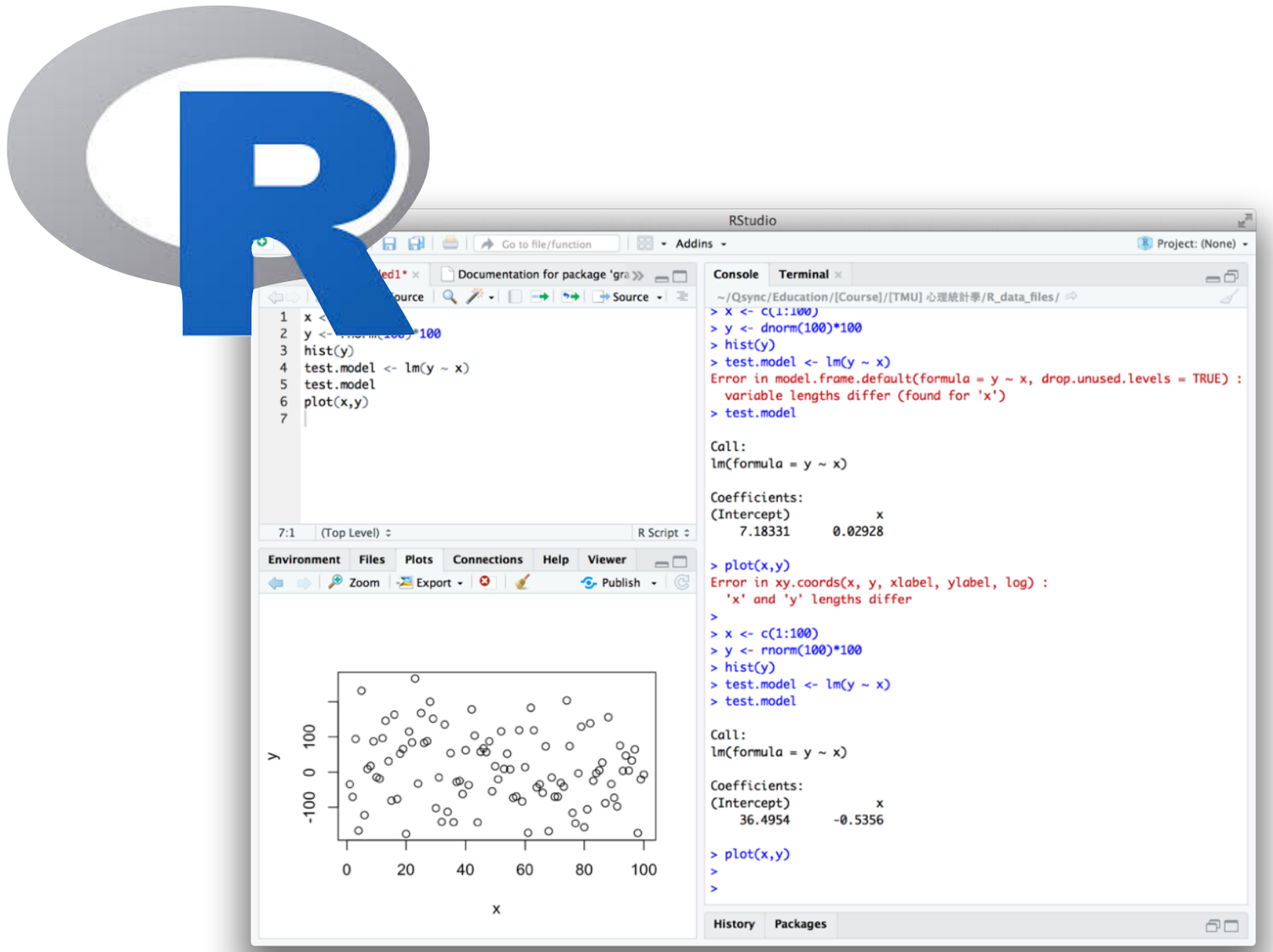
- **Remember to check the assumptions to make sure your model generalizes beyond your sample:**
- Look at the graph of the standardized residuals plotted against the fitted values. If it looks like random dots then this is good. If not, be aware of the issues in *linearity* and *variance homogeneity*.
- Look at the histogram of the residuals. If it looks like a normal distribution then this is good. If not, see if it reaches significance (esp. for small sample size).



TAIPEI MEDICAL UNIVERSITY

lm
summary
summary.aov
abline
predict
lines

① SIMPLE REGRESSION



TAIPEI MEDICAL UNIVERSITY
1960
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

DEMO

Growth Reduction by Tannins

- **Background**: Tannin-induced proline-rich salivary proteins (PRPs) diminish the anti-nutritional effects of dietary polyphenolics in rats; therefore the growth in size may be associated with dietary tannin concentration.

Outcome measure: Growth in rat size

Predictor: dietary Tannin concentration

- **Set up Hypothesis**:

① Hypothesis

$H_0: \beta = 0$

$H_a: \beta \neq 0$

Load: **tannin.csv**

- **Data import**:
  - *reg <- read.csv("tannin.csv", header = TRUE)*

臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

DEMO

Growth Reduction by Tannins

② Assumption

- **Assumption check:**
  - *Here we assume the data fulfil all assumptions for Regression.*

③ Testing

General form:

- ***lm(outcome ~ predictor, data=dataframe, na.action=action)***
- ***predict(model, newdata, interval="type")***

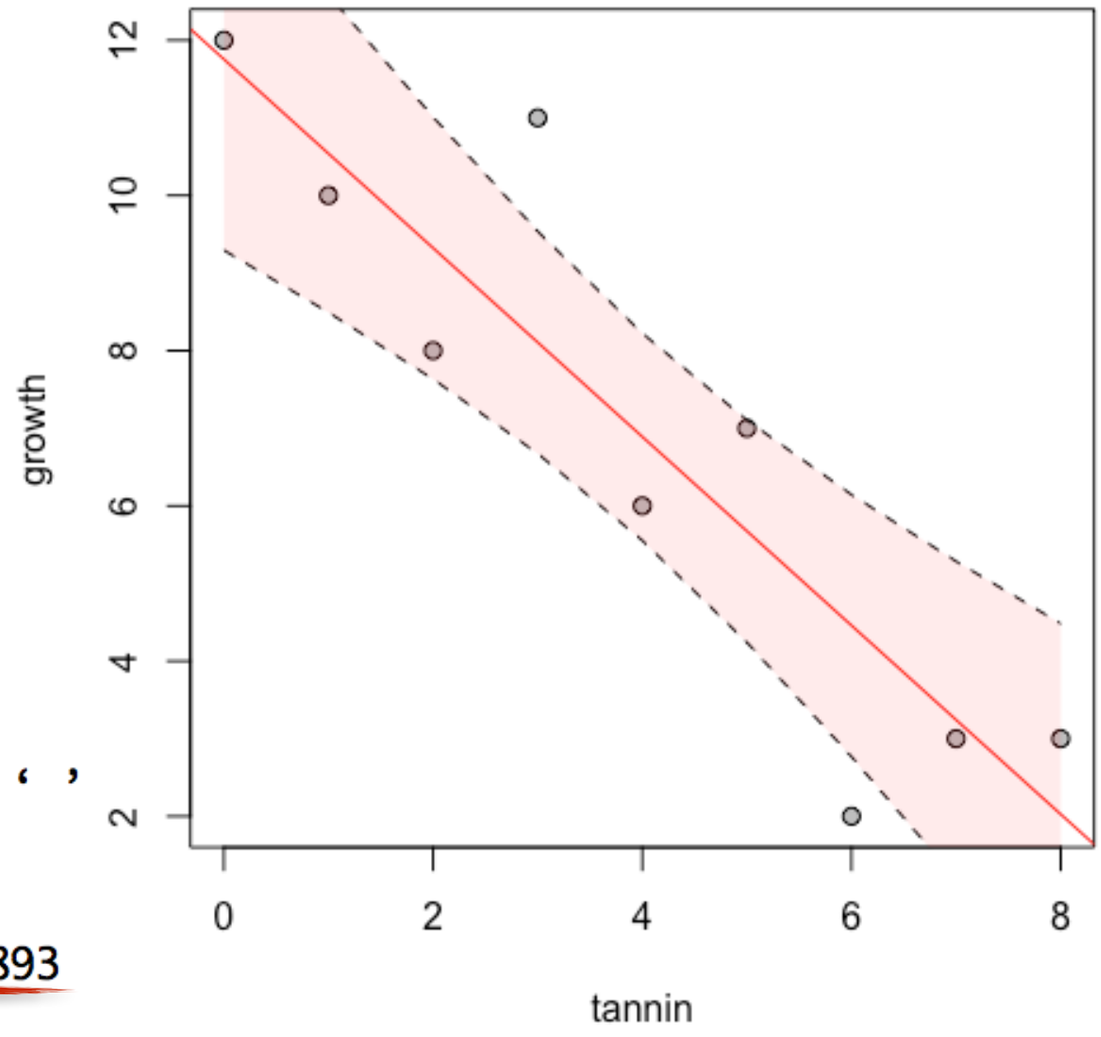
- **Parameters**:
  - *outcome*: the variable to be predicted (dependent variable).
  - *predictor(s)*: the controlled variable (independent variable).
  - *dataframe*: the name of data frame you stored your indices.
  - *na.action=(**na.fail** / na.omit / na.exclude)*: choosing strategy for the missing/ not-available data.
  - *interval=(**confidence** / prediction)*: for C.I. estimations.

DEMO

Tannin vs. Growth

③ Testing

- **Practice:**
  - *model <- lm(growth~tannin)*
  - *summary(model)*
  - *summary.aov(model)*



- **Results:**

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.7556     1.0408  11.295 9.54e-06 ***
tannin       -1.2167     0.2186  -5.565 0.000846 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 1.693 on 7 degrees of freedom
Multiple R-squared:  0.8157,    Adjusted R-squared:  0.7893
F-statistic: 30.97 on 1 and 7 DF,  p-value: 0.0008461
```

TAIPEI MEDICAL UNIVERSITY 臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

Validation of Regression

- ***Adjusted R^2:***
  - **R^2** : how much of the Y variance is accounted for by the regression model from one sample data.
  - **Adjusted R^2** : how much variances in Y would be accounted for if the model had been derived from the population where the sample was taken.
  - ***Stein's formula*** *(2002)*:

$$adjusted\ R^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right)\left(\frac{n-2}{n-k-2}\right)\left(\frac{n+1}{n}\right)\right](1-R^2)$$

- ***Data splitting:***
  - Randomly splitting your dataset, computing a regression equation on both havens of the data and then comparing resulting models.
  - Comparing ***R^2*** and **β** values in the two samples, you can tell how well the original model generalizes.

k: predictor number
n: total sample size

DEMO

Tannin vs. Growth

④ Effect Size

- **Effect size**:

```
Residual standard error: 1.693 on 7 degrees of freedom
Multiple R-squared:  0.8157,    Adjusted R-squared:  0.7893
F-statistic: 30.97 on 1 and 7 DF,  p-value: 0.0008461
```

⑤ Decision

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.7556     1.0408  11.295 9.54e-06 ***
tannin       -1.2167     0.2186  -5.565 0.000846 ***
```

- **Reporting decision:**

  The rat growth was negatively associated with tannin concentration (*p* < 0.001). The regression model is y = -1.22*x+11.76, which explained 79% of the total variation in rat growth.

臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

1. **Is there any outlier?**
2. **How about the assumption check?**

DEMO

Tannin vs. Growth

② Assumption

1.Is there any outlier?

```
Influence measures of
         lm(formula = growth ~ tannin) :

   dfb.1_  dfb.tnnn   dffit cov.r  cook.d   hat inf
1  0.1323 -1.11e-01  0.1323 2.167 0.01017 0.378   *
2 -0.2038  1.56e-01 -0.2058 1.771 0.02422 0.261
3 -0.3698  2.40e-01 -0.3921 1.323 0.08016 0.178
4  0.7267 -3.24e-01  0.8981 0.424 0.24536 0.128
5 -0.1011 -1.55e-17 -0.1864 1.399 0.01937 0.111
6  0.0635  1.13e-01  0.3137 1.262 0.05163 0.128
7  0.0741 -5.29e-01 -0.8642 0.667 0.27648 0.178
8  0.0256 -6.86e-02 -0.0905 1.828 0.00476 0.261
9 -0.2263  4.62e-01  0.5495 1.865 0.16267 0.378   *
```

DEMO

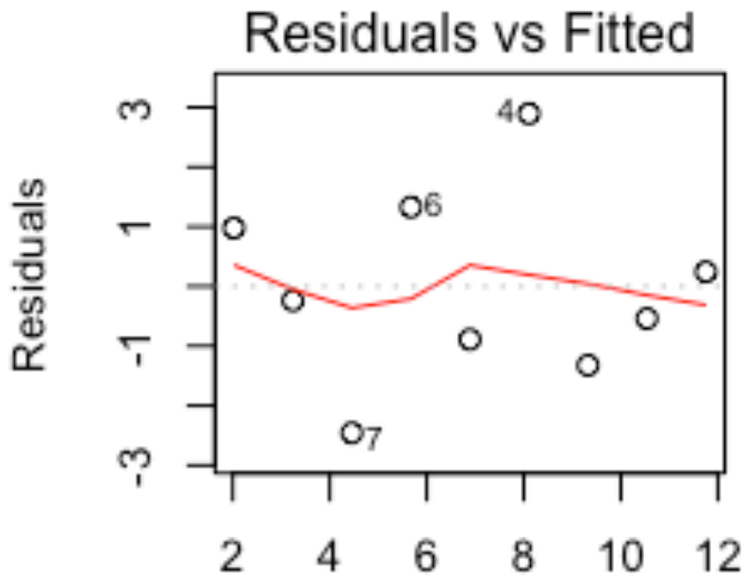
Tannin vs. Growth

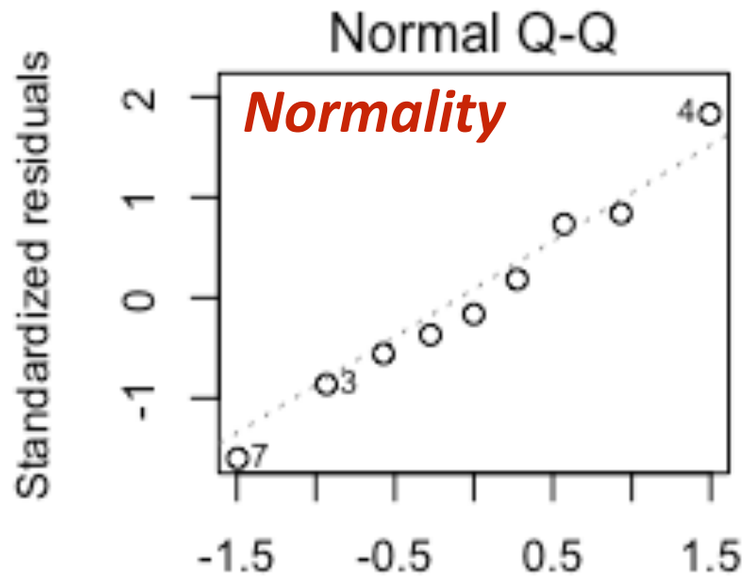
② Assumption

2.the assumption check

- **Visual inspection of model assumptions:**
  - ***plot(model)***

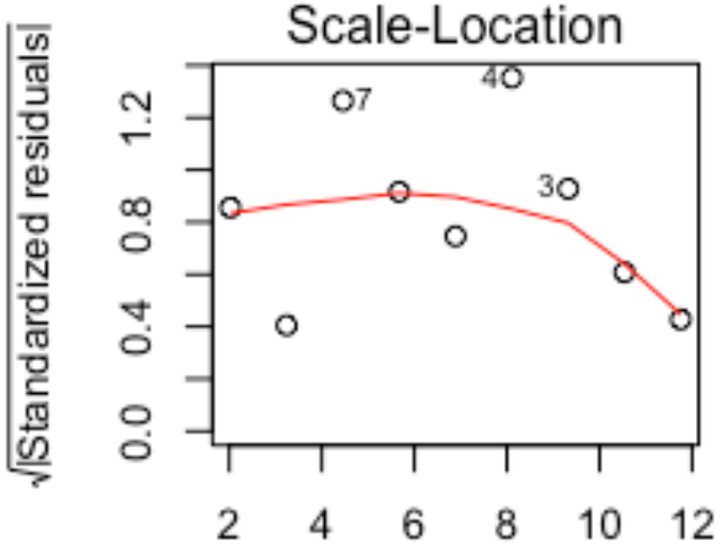
Heteroscedasticity

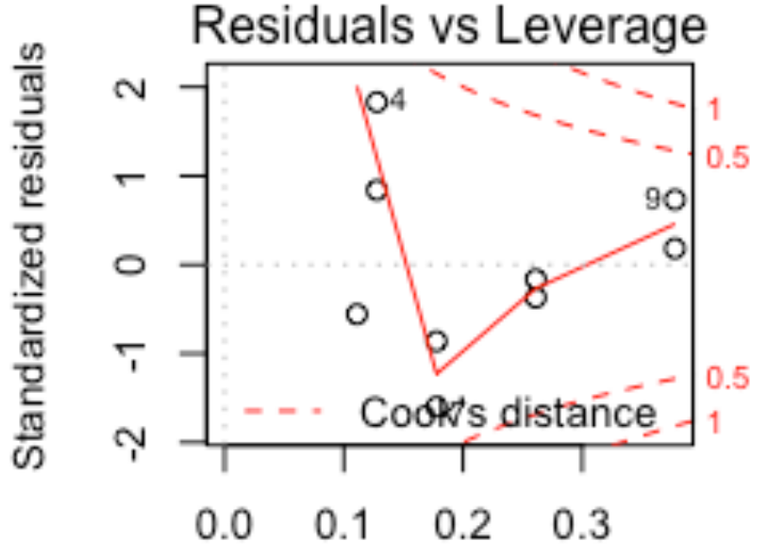




Outlier

Heteroscedasticity + Outlier

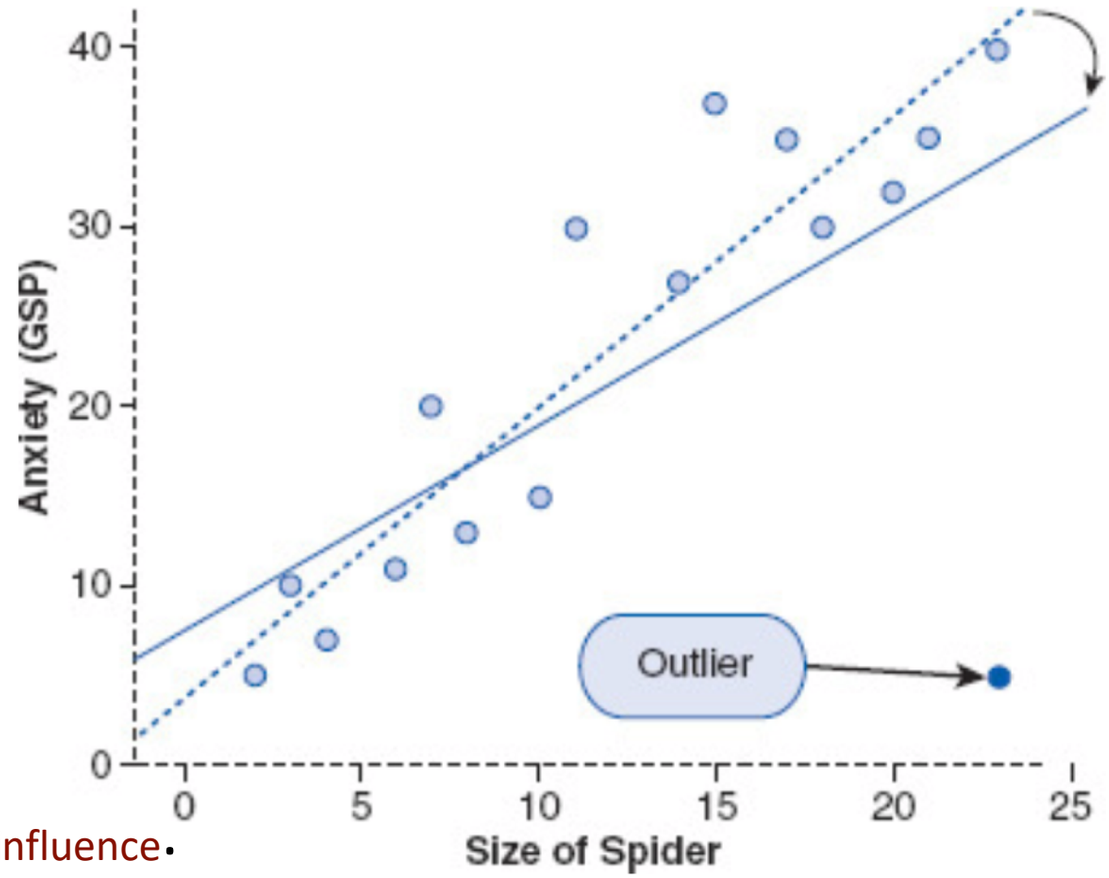




Outlier & Influence

- **<u>Unstandardized residuals</u>** (σ_ε): remaining errors.
- **<u>Standardized residual</u>** (σ_ε **/ SD**):
  - no more than 5% of cases > 2
  - no more than 1% > 2.5.
- **<u>Cook's distance</u>**:
  - Cook and Weisberg (1982)
  - **Values > 1** may be a cause of concern.
- **<u>Leverage (hat values)</u>**:
  - Leverage values:
  - Leverage lies between 0no influence and 1complete influence.

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$$



Parameter	Case 30 Included	Case 30 Excluded	Difference
Constant (intercept)	29.00	31.00	-2.00
Predictor (slope)	-0.90	-1.00	0.1 **dfb**
Model (regression line)	Y=(-0.9)X + 29	Y=(-1)X + 31	
Predicted Y	28.10	30.00	-1.90 **dffit**

DEMO

London Pubs

- The Mayor of London at the turn of the last century was interested in how drinking affected mortality.
- London is divided up into eight regions, so he can measure the number of pubs and the number of deaths over a period of time in the eight regions.
- Plot the scatter plot and see if you can find out any particular region.

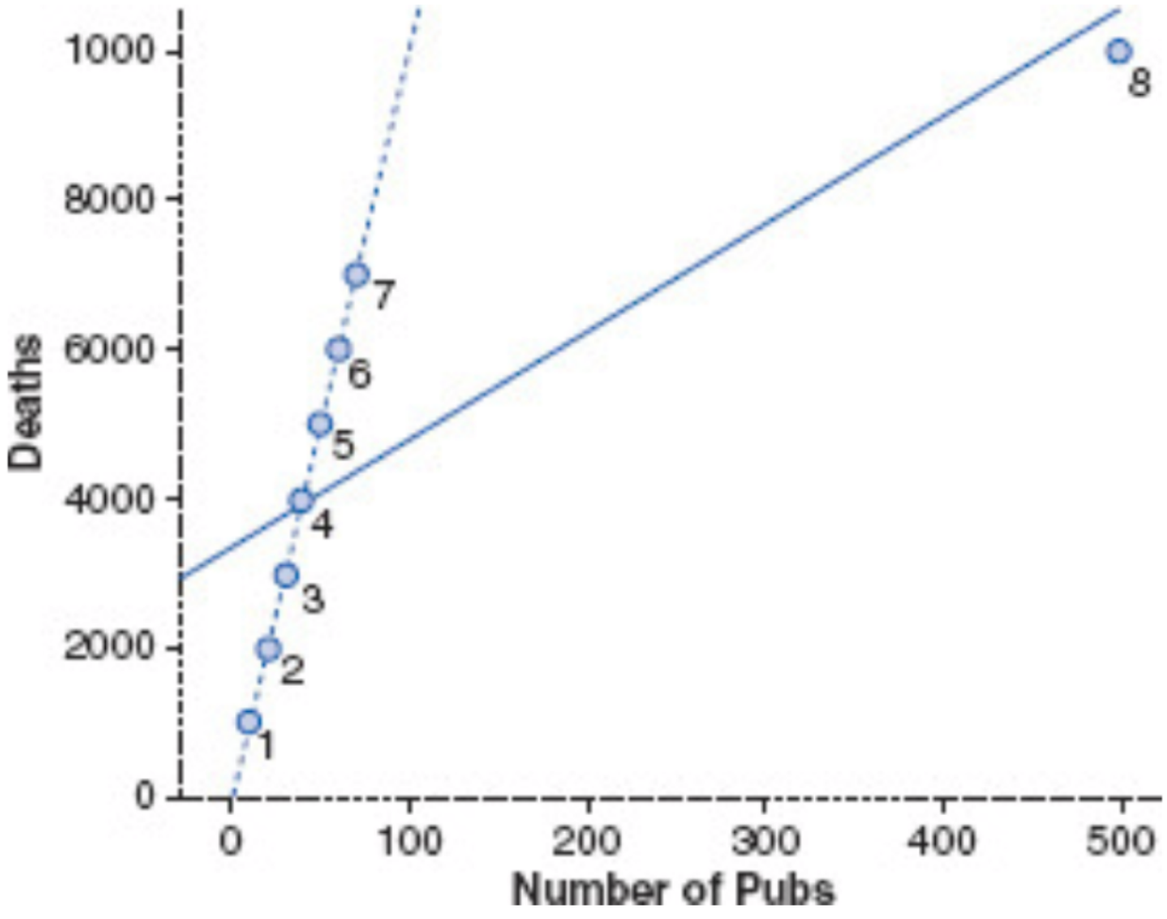
Load: **pubs.dat**

Use '**influence.measures**'

TAIPEI MEDICAL UNIVERSITY 臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

DEMO

London Pubs




```
#------Outlier---------
pubs <- read.delim("pubs.dat", header = TRUE)
pubReg <- lm(mortality ~ pubs, data = pubs)
summary(pubReg)

rstandard(pubReg)  → generate standardized residuals
influence.measures(pubReg)
```

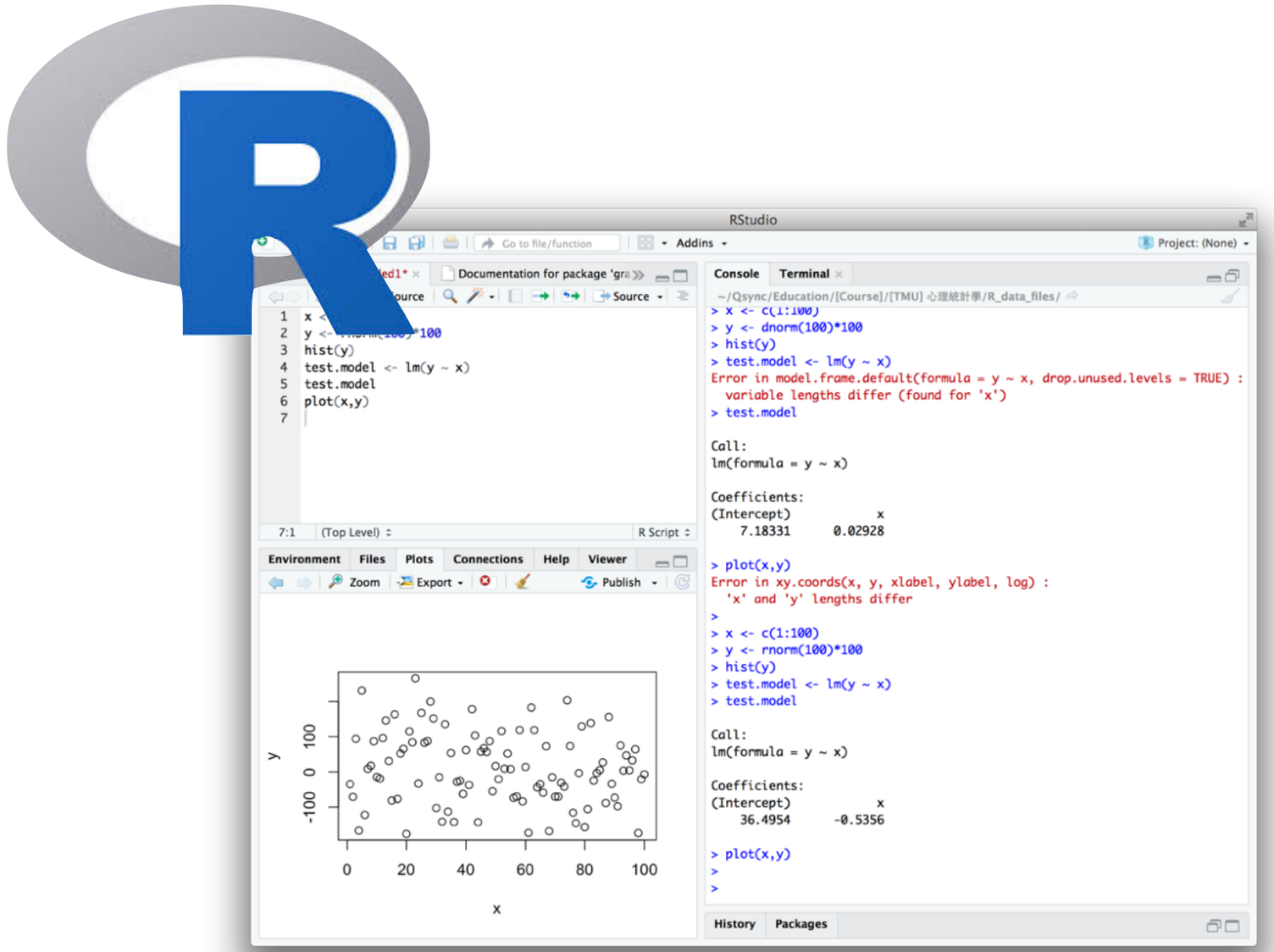
```
Influence measures of
         lm(formula = mortality ~ pubs, data = pubs) :

   dfb.1_ dfb.pubs   dffit cov.r   cook.d   hat inf
1 -0.7432  0.36886 -0.7440 0.712 2.13e-01 0.166
2 -0.4077  0.18484 -0.4096 1.226 8.53e-02 0.157
3 -0.1749  0.07132 -0.1770 1.578 1.81e-02 0.149
4  0.0157 -0.00564  0.0161 1.678 1.55e-04 0.143
5  0.1934 -0.05933  0.2004 1.512 2.29e-02 0.137
6  0.3833 -0.09618  0.4047 1.125 8.09e-02 0.132
7  0.6300 -0.12023  0.6808 0.625 1.71e-01 0.129
8     NaN      NaN     NaN   NaN 2.27e+02 0.987   *
```

臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

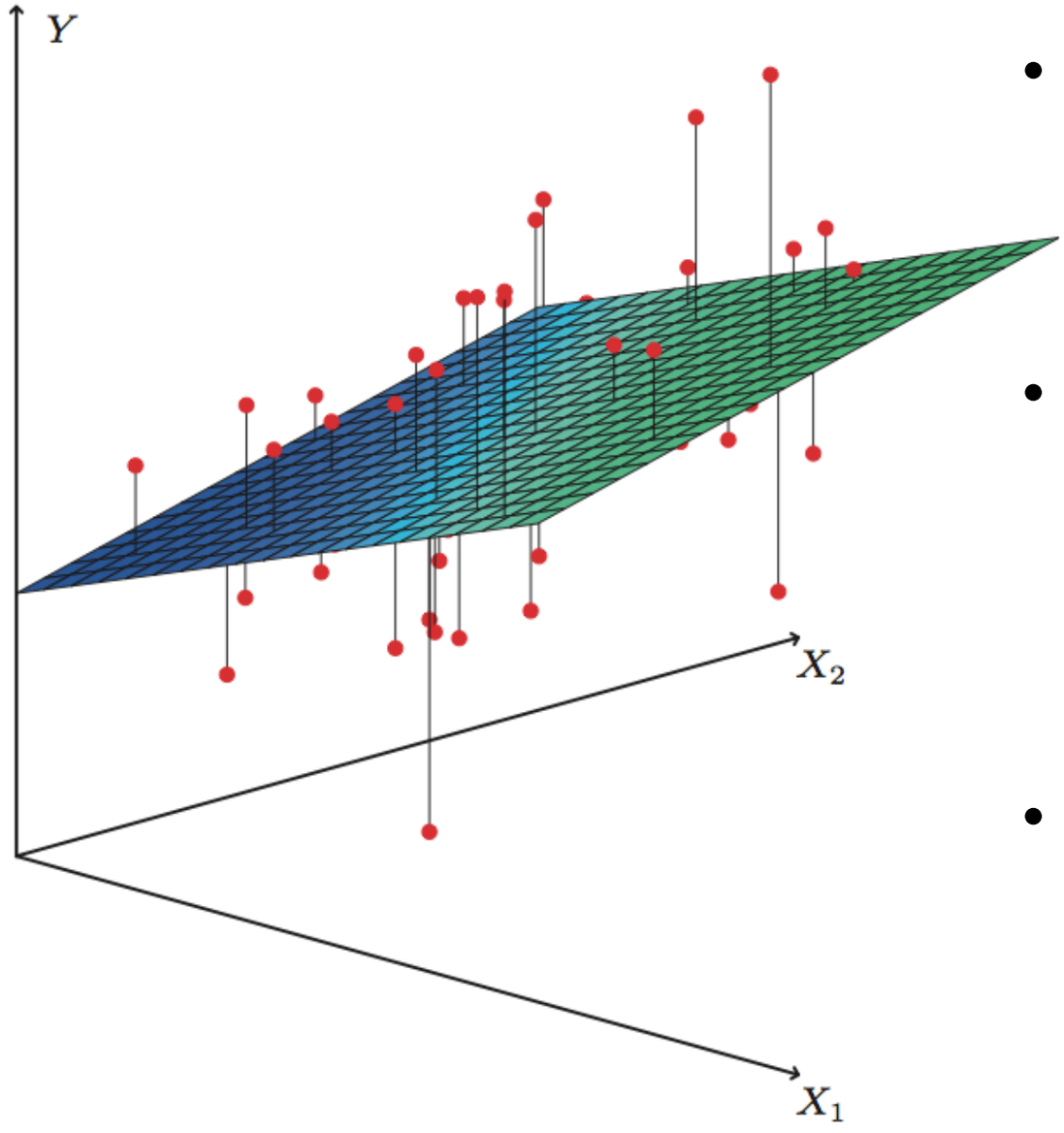
anova
AIC
confint

② MULTIPLE REGRESSION



TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

Multiple Regression



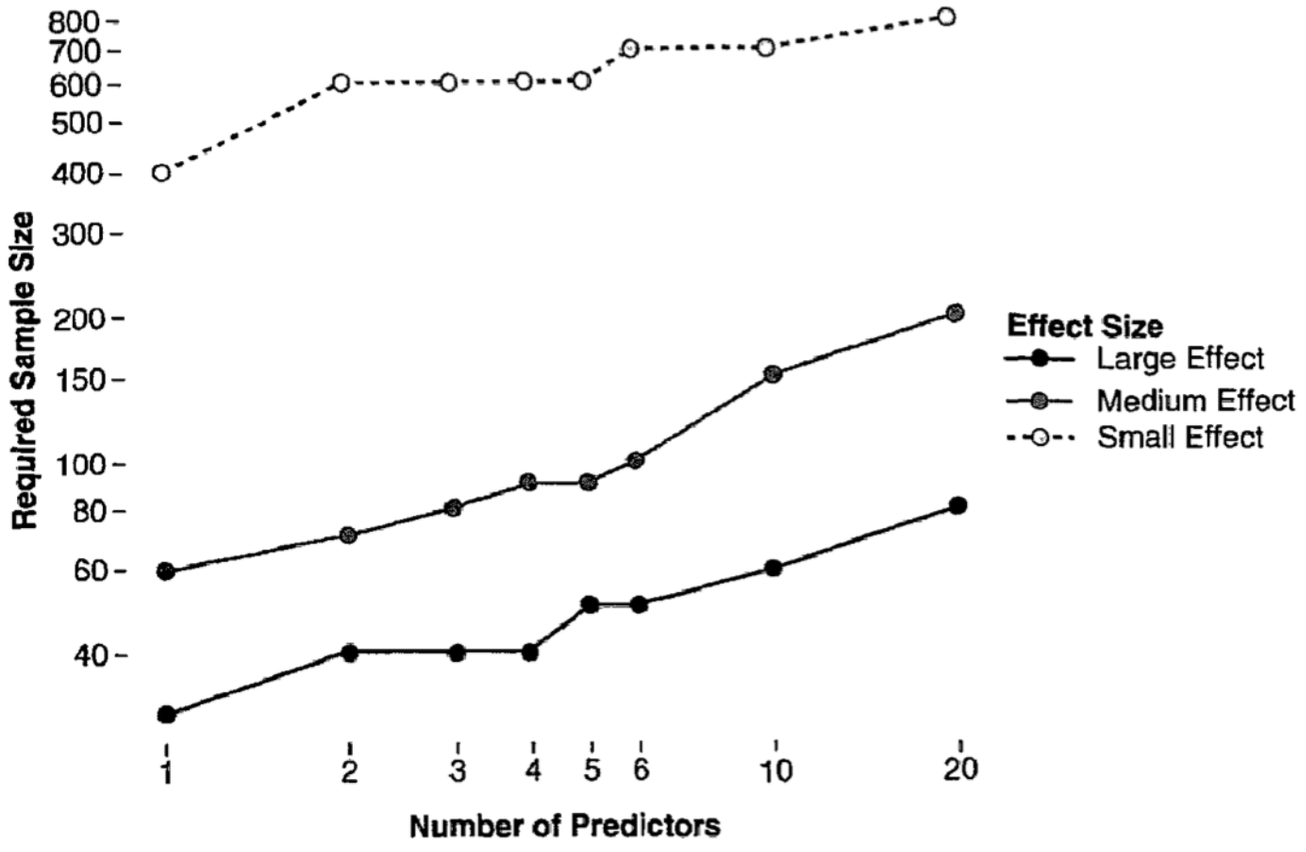
- To learn about the relationship between **several** independent variables (X) and **one** dependent variable (Y).
- Widely used in research for General question:
  - "**Which is the best predictor of …**" among several possible predictors.
- Drawbacks: no causal information in all types of regression.

Sample Size & Predictors

k: predictor number
N: total sample size

- **<u>Sample size in Regression</u>**:
  - The more, the better (for sure).
  - ∵ the expected correlation is: **k/(N-1)** . [expected ***r*** = 0 for random data]
  - It depends on the **effect size** of interest(s).

- **<u>Predictor numbers</u>** (for multiple regression):
  - It's important not to include too many predictors (X).
  - **Degree of freedom**: **N-k-1** .
  - Inclusion of the predictors when you have strong theoretical grounding.



臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

DEMO

Sales vs. Advertisements

- **Background**: The record company executive was now extending the model of album sales to incorporate possible factors. He thought three factors may influence the album sales: (1) investment on the advertisement (**advert**); (2) the number of times songs from the album are played on air the week before release (**airplay**); (2) the attractiveness of the band (**attract**). Which is the best predictor for the final album sale?

Load:
AlbumSales.dat

Outcome measure: Album sales

Predictor: advert, airplay, attract.

- **Set up Hypothesis**:

① Hypothesis

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

DEMO

Sales vs. Advertisements

③ Testing

- **Practice:**
  - *albumSales.3<-lm(sales ~ adverts + airplay + attract, data = album)*
  - *summary(albumSales.3)*

Simple Regression

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.341e+02  7.537e+00  17.799   <2e-16 ***
adverts     9.612e-02  9.632e-03   9.979   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 65.99
Multiple R-squared:  0.3346,
F-statistic: 99.59 on 1 and 19
```

Multiple Regression

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -26.612958  17.350001  -1.534    0.127
adverts       0.084885   0.006923  12.261  < 2e-16 ***
airplay       3.367425   0.277771  12.123  < 2e-16 ***
attract      11.086335   2.437849   4.548 9.49e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.09 on 196 degrees of freedom
Multiple R-squared:  0.6647,    Adjusted R-squared:  0.6595
F-statistic: 129.5 on 3 and 196 DF,  p-value: < 2.2e-16
```

DEMO

Comparing Models

- For R^2 index, adding more predictors will increase R^2

  ➜ How to define good model?

k: predictor number
n: total sample size

- **<u>Akaike information criterion (AIC)</u>**:

$$AIC = n\ln\left(\frac{SSE}{n}\right) + 2k$$

  - A larger value of the AIC indicates worse fit, corrected for the k value.
  - **AIC(model1, model2)**

```
                df      AIC
albumSales.1     3 2247.375
albumSales.3     5 2114.337
```

- **<u>Use F-ratio for comparison</u>**:
  - to compare the mean square error between models.
  - **anova(model1, model2)**

providing significance

```
Model 1: sales ~ adverts
Model 2: sales ~ adverts + airplay + attract
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    198 862264
2    196 434575  2    427690 96.447 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
```

DEMO

Sales vs. Advertisements

④ Effect Size

- **Effect size**:

```
Residual standard error: 47.09 on 196 degrees of freedom
Multiple R-squared:  0.6647,    Adjusted R-squared:  0.6595
F-statistic: 129.5 on 3 and 196 DF,  p-value: < 2.2e-16
```

⑤ Decision

- **Reporting decision:**

  According to the American Psychological Association (APA) guideline.

albumSales.3	β	SE(β)	Studentized β	*t-value*	*p-value*
Intercept	*-26.61*	*17.35*		*-1.53*	*0.127*
Advertising budget	*0.09*	*0.01*	*0.51*	*12.26*	*< .001*
Plays on Radio	*3.37*	*0.28*	*0.51*	*12.12*	*< .001*
Attractiveness	*11.09*	*2.44*	*0.19*	*4.59*	*< .001*

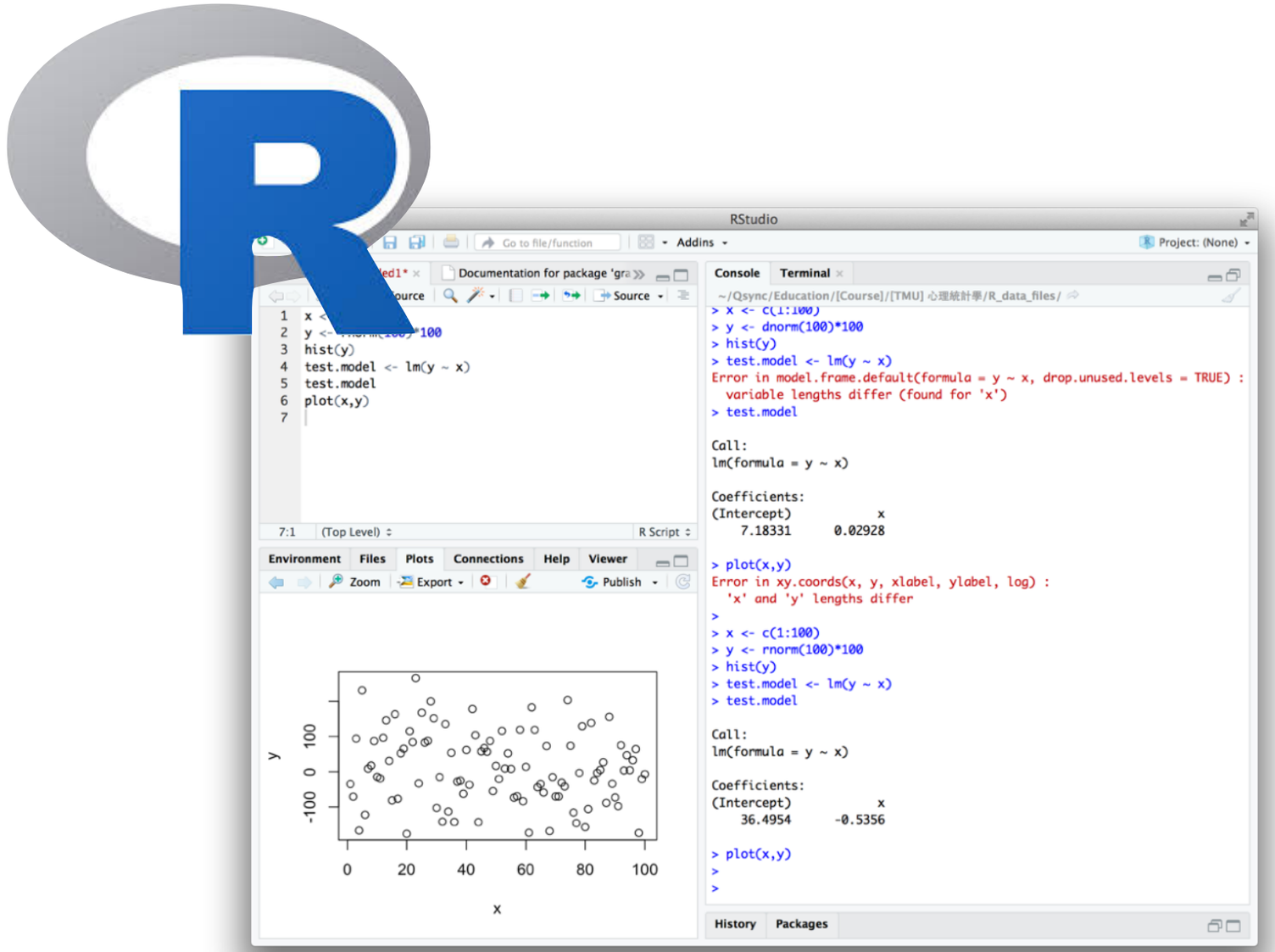
Residual type	R naming convention	Mathematical formula
Standardized by σ		$\hat{\varepsilon}_i/\sigma$
Internally studentized[a]	Standardized	$\hat{\varepsilon}_i/\hat{\sigma}$
Externally studentized[b]	Studentized	$\hat{\varepsilon}_i/\hat{\sigma}_{(-i)}$

[a] $\hat{\sigma}$ is an estimate of σ based on all observations,
[b] $\hat{\sigma}_{(-i)}$ is an estimate of σ obtained after excluding the i-th observation.

TAIPEI MEDICAL UNIVERSITY 臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

lm.beta
influence.measures
durbinWatsonTest (dwt)
vif

③ ASSUMPTIONS OF REGRESSION MODELS



TAIPEI MEDICAL UNIVERSITY
臺北醫學大學
TAIPEI MEDICAL UNIVERSITY

Assumptions of Regression

$$\hat{Y} = a + (b)(X)$$

- ***Types of variable (X, Y)***
  - Predictors should be quantitative or categorical (no need to be normally distributed);
  - Outcome variables must be quantitative, continuous and independently sampled.
- ***Without Multicollinearity (X)***
  - ➡ should be NO perfect linear relationship between two or more predictors
- ***Homoscedasticity (σ_ε):***
  - ➡ Residuals at each level of predictors should have constant variance.
- ***Lack of Autocorrelation (σ_ε):***
  - ➡ Residuals are uncorrelated.
- ***Errors are Normally distributed (σ_ε):***
  - ➡ Residuals (not predictor) are random, normally distributed with a mean of 0.

Independence (Y, σ)

Normal distribution (σ)

Consistency (X,Y)

Assumptions of Regression

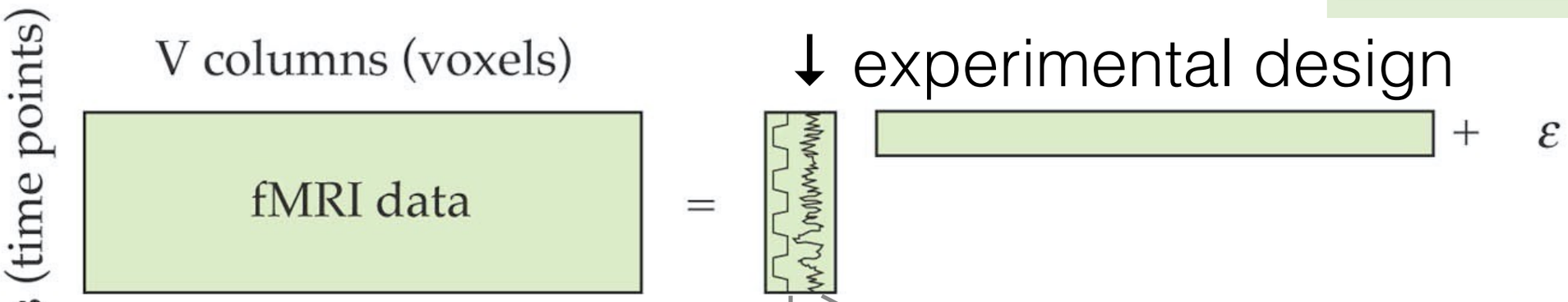
$$\hat{Y} = a + (b)(X)$$

- **If the assumptions are violated, you can NOT generalize your findings beyond the samples.**

Multicollinearity of fMRI Predictors

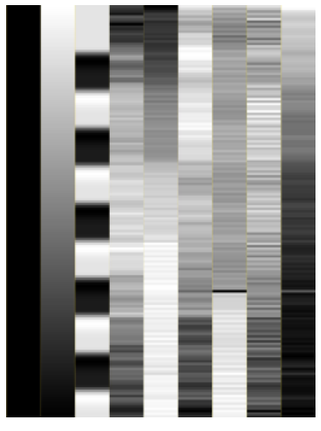
- **Collinearity** means that there is a strong correlation between two or more predictors in a multiple regression model (only for multiple regression).
- If there is perfect collinearity between predictors (r=1), it becomes impossible to obtain unique estimates of the regression coefficients (β).

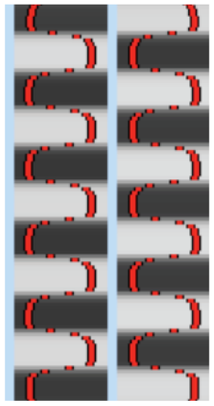
$$Y = X\beta + \epsilon$$



↓ experimental design

① No collinearity
in this design matrix





② Collinearity
in this design matrix

Multicollinearity of Predictors

- **Collinearity** means that there is a <u>strong correlation between two or more predictors</u> in a multiple regression model (only for multiple regression).
- If there is perfect collinearity between predictors (r=1), it becomes impossible to obtain unique estimates of the regression coefficients (β).

- **<u>Untrustworthy βs</u>** :
  - standard error of β increases.
  - less likely to present the population.
- **<u>Limits the size of correlation</u>** :
  - like partial correlation, the 2nd predictor accounts for little variance.
- **<u>Unable to assess importance of predictor</u>**:
  - which predictor is the most important?

Variance inflation factor (VIF)

$$\mathrm{VIF} = \frac{1}{1 - R_i^2} \qquad \mathrm{VIF} = \frac{\sum_{i=1}^{k} \overline{\mathrm{VIF}}}{k}$$

R_i is the multiple correlation of the regression between X_i and the remaining k-1 predictors.

VIF=1: not correlated.
1 < VIF < 5: moderately correlated.
VIF > 5: highly correlated.

Bowerman & O'Connell, 1990

http://www.statisticshowto.com/variance-inflation-factor/

DEMO

Multicollinearity

Variance inflation factor (VIF)

$$\mathrm{VIF} = \frac{1}{1 - R_i^2} \qquad \mathrm{VIF} = \frac{\sum_{i=1}^{k} \overline{\mathrm{VIF}}}{k}$$

R_i is the multiple correlation of the regression between X_i and the remaining k-1 predictors.

VIF=1: not correlated.
1 < VIF < 5: moderately correlated.
VIF > 5: highly correlated.

Bowerman & O'Connell, 1990

Compare β of different regressors

```
Coefficients:
             Estimate Std. Error t value
(Intercept) -26.612958  17.350001  -1.534
adverts       0.084885   0.006923  12.261
airplay       3.367425   0.277771  12.123
attract      11.086335   2.437849   4.548
```

VIF values:

```
  adverts  airplay  attract
1.014593 1.042504 1.038455
```

Standardized β:

```
> QuantPsyc::lm.beta(albumSales.3)
   adverts    airplay    attract
0.5108462 0.5119881 0.1916834
```

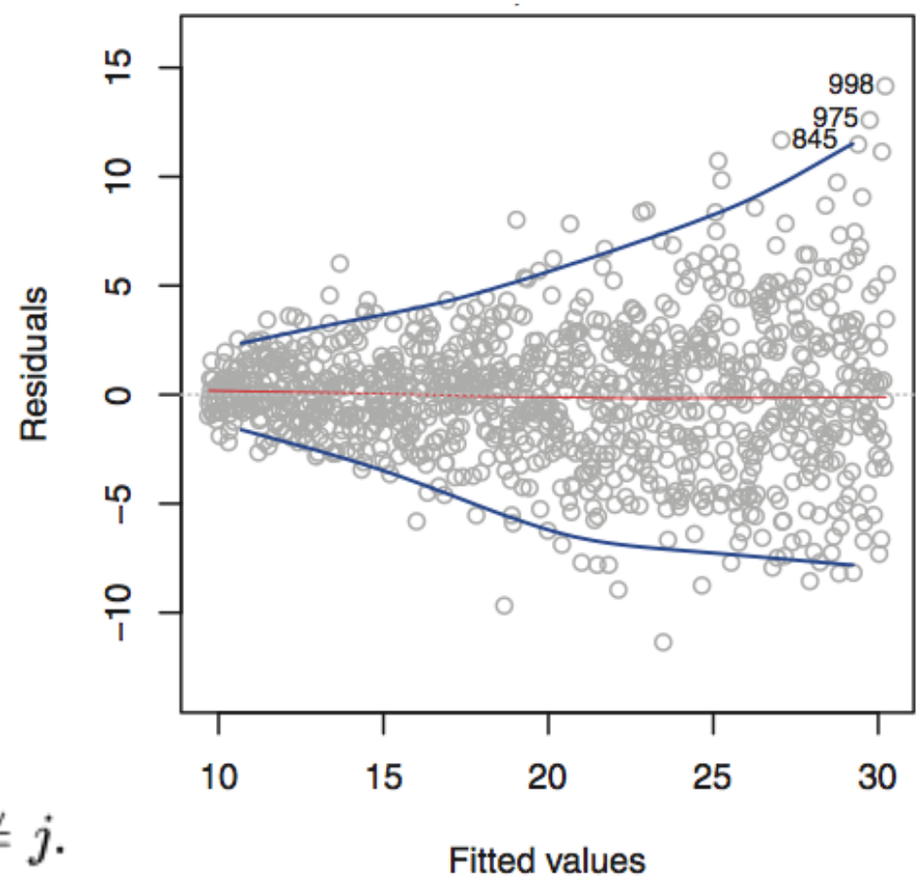
as a results of one 'standard deviation' change

臺北醫學大學 TAIPEI MEDICAL UNIVERSITY

Considerations of Residuals

1. **<u>Normally distributed errors</u>**:
   - the error term has uniform variance.
2. **<u>Homoscedasticity</u>:**
   - the error term has uniform variance.
   - *visual inspection*
3. **<u>Independent errors</u>**:
   - Zero expectation without autocorrelations.
   - ***Gauss-Markov theorem***

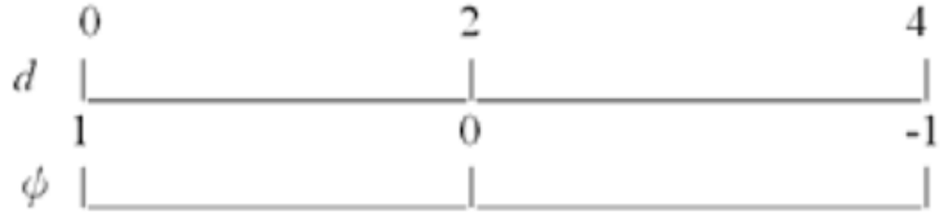
$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j.$$



Durbin-Watson Test

$$H_0 : \rho = 0$$
$$H_1 : \rho > 0$$

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$



https://en.wikipedia.org/wiki/Gauss-Markov_theorem

library("performance")
check_autocorrelation(Model)

DEMO

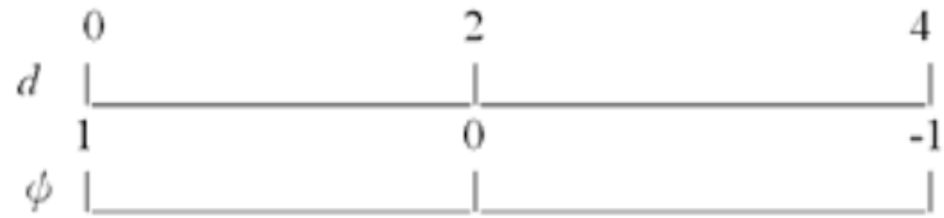
Considerations of Residuals

Given $e_t = \rho e_{t-1} + \nu_t,$ the Durbin-Watson statistic states that null hypothesis: $\rho = 0$.

Durbin-Watson Test

$$H_0 : \rho = 0$$
$$H_1 : \rho > 0$$

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$



d is approximately equal to 2(1 – ρ), where ρ is the sample autocorrelation of the residuals

```
> car::dwt(albumSales.3)
 lag Autocorrelation D-W Statistic p-value
   1       0.0026951      1.949819   0.736
 Alternative hypothesis: rho != 0
```

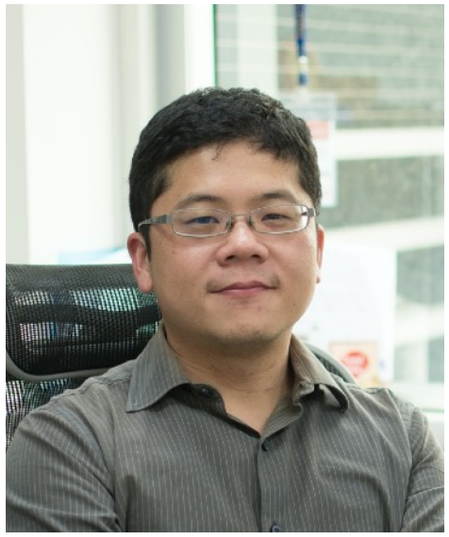
Notice: it depends on the order (i→i+1) of data. If you change the order, d will change.

Discussion

1. **Linear Regression**
   - Concept and assumptions

2. **Multiple Regression**
   - Model comparison and interpretation

3. **[R] Assumption check & Nonlinear**



THANK YOU FOR YOUR ATTENTION

E-mail: sleepbrain@tmu.edu.tw