Statistical Method
Bivariate Association

I-Chen Lee, STAT, NCKU

Chapter 4 of
Cleff, T. (2014). Exploratory Data Analysis in Business and Economics.

Chapters 3 of An Introduction to Statistical Learning with R.

October 31, 2023

Overview

Bivariate association

- In the first stage of data analysis we learned how to examine variables and survey traits individually, or univariately. (Univariate Data)
- It is a task how to assess the association between two variables using methods known as bivariate analyses.
- The methods of bivariate analysis depend on the scale of the observed traits or variables.
  - Association between two metric variables
  - Association between two ordinal variables
  - Association Between two ordinal Variables

Textbook: Cleff, T. (2014). Exploratory Data Analysis in Business and Economics.

Bivariate association

Table 4.1 Scale combinations and their measures of association

		Nominal	Ordinal	Metric
Nominal	Dichotomous	Phi; Cramer's V	Biserial rank correlation; Cramer's V	Point-biserial r; classification of metric variables and application of Cramer's V
		[Sect. 4.2]	[Sect. 4.5.2]	[Sect. 4.5.1]
	Non-dichotomous	Cramer's V; contingency coefficient	Cramer's V; contingency coefficient	Classification of metric variables and application of Cramer's V
		[Sect. 4.2]	[Sect. 4.2]	[Sect. 4.2]
Ordinal			Spearman's rho (ρ); Kendall's tau (τ)	Ranking of metric variables and application of ρ or τ
			[Sect. 4.4]	[Sect. 4.4]
Metric				Pearson's correlation (r)
				[Sect. 4.3.2]

Two metric variables

- A metric variable is a variable measured quantitatively. The distance between the values is equal.
- How can one tell whether there's an actual association? If so, what is its strength?
- Graphical method: scatterplot. A scatterplot expresses three aspects of the association between two metric variables.
    - The direction of the relationship: positive or negative.
    - The form of the relationship: linear or non-linear.
    - The strength of the relationship:

Scatterplot

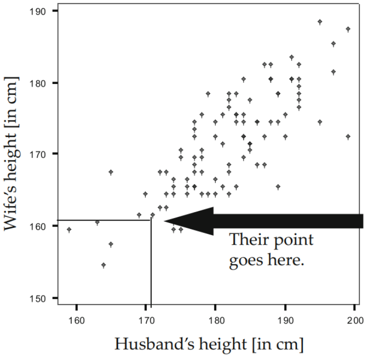
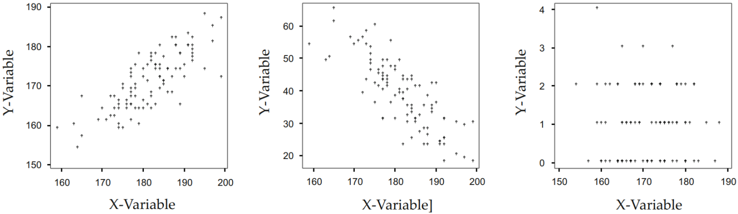


Fig. 4.13 The scatterplot

	names	wheight	hheight
1	John and Judy	1590	1809
2	Carl and Kathryn	1560	1841
3	Craig and Jackie	1620	1659
4	Larry and Susan	1540	1779
5	Scott and Susan	1420	1616
6	John and Margaret	1660	1695
7	Stanley and Patricia	1610	1730
8	David and Lisa	1635	1753
9	Robert and Cathy	1580	1740
10	Larry and Karen	1610	1685
11	Steven and Candice	1590	1735
12	Joseph and Lesley	1610	1713
13	Eric and Ethel	1700	1736

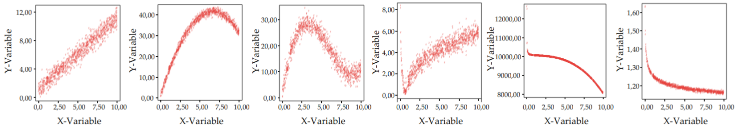
Observation 12:
Joesph (171.3 cm) and Lesley (161.0 cm)

1. The **Direction** of the relationship



positive trend negative trend no clear trend

2. The **form** of the relationship



3. The **strength** of the relationship

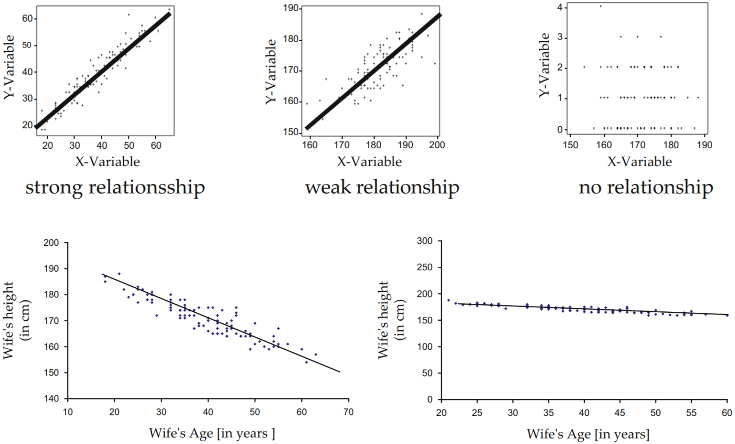


Fig. 4.15 Different representations of the same data (3). . . .

Covariance of two variables

- A correlation coefficient is a measure that gives us the relationship between two metric variables, including the direction (positive or negative) and the strength (from -1 to 1).
- Ideas: Covariance of two variables is the measure of the deviation between each value pair from the bivariate centroid in a scatterplot.

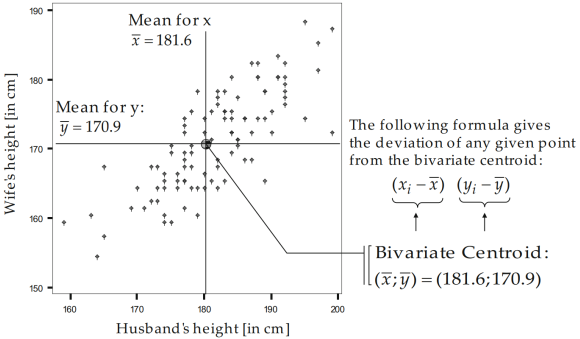


Fig. 4.16 Relationship of heights in married couples

Covariance and correlation (definition in probability theory)

Definition: The covariance of two random variables X and Y, denoted by $Cov(X, Y)$, is defined by

$$Cov(X, Y) = E\left[(X - \mu_x)(Y - \mu_y)\right],$$

where $\mu_x = E(X)$ and $\mu_y = E(Y)$.

- $Cov(X, Y) = E(XY) - E(X)E(Y)$.

Definition: The correlation of two random variables X and Y, denoted by $Corr(X, Y)$, is defined by

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}.$$

Sample covariance and sample correlation

The sample covariance of two random variables X and Y, is defined by

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

The sample correlation r of two random variables X and Y, is defined by

$$r = \frac{\sum_{i=1}^{n}\left[(x_i - \bar{x})((y_i - \bar{y})\right]}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \ -1 \leq r \leq 1,$$

where $\bar{x}$ is the sample mean of X and $\bar{y}$ is the sample mean of Y.
Note that r is called the Pearson's correlation.

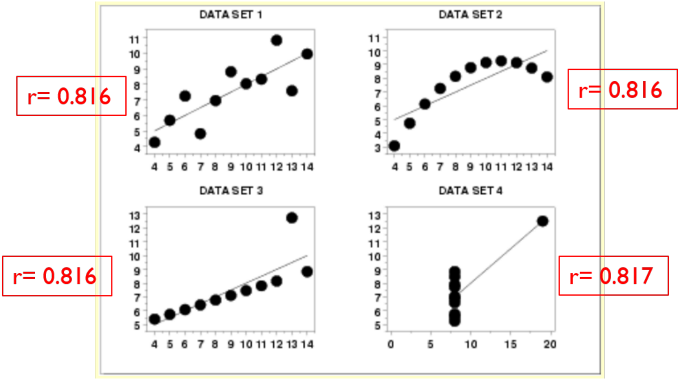
Pearson's correlation

Researchers commonly draw the following distinctions:

- $0 < |r| < 0.3$ negligible association
- $0.3 \leq |r| \leq 0.5$ weak linear association
- $0.5 \leq |r| \leq 0.7$ moderate linear association
- $0.7 \leq |r| \leq 0.9$ high linear association
- $0.9 \leq |r| \leq 1$ strong linear association

Notice!

Higher value of correlation coefficient sometimes is not meaningful. Please still check scatter plots again to see the relationship.
`https://www.itl.nist.gov/div898/handbook/eda/section1/eda16.htm`



Estimation: fitted simple regression

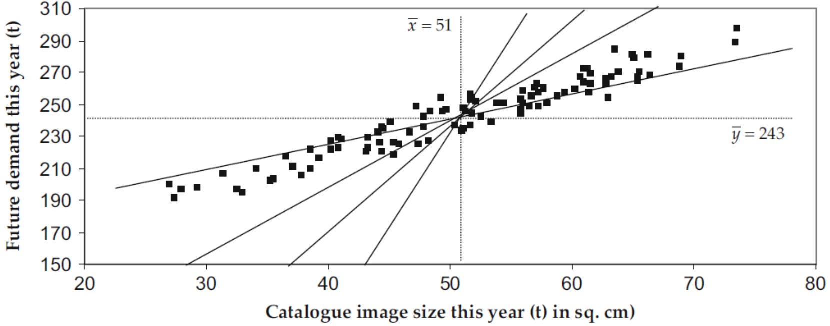


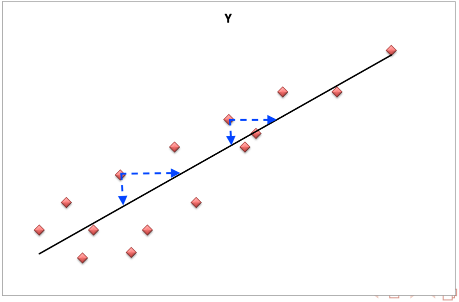
Fig. 5.4 Lines of best fit with a minimum sum of deviations

Introduction to regression

Simple linear regression:

- Y is the dependent variable (response, outcome, output).
- x is the independent variable (covariate, input).
- ε is the error which is the distance between the observation to the true function.

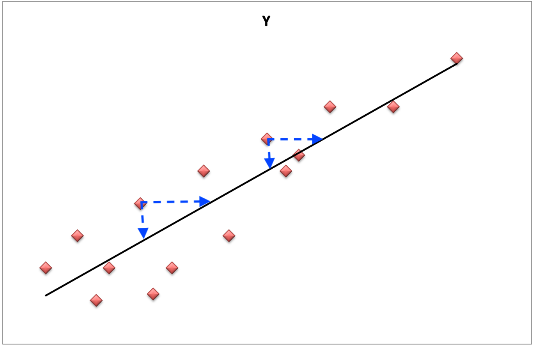
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \;\; i = 1, \ldots, n.$$



Introduction to the multiple regression

If we can estimate the value of $\beta = \{\beta_0, \beta_1, \ldots, \beta_p\}^T$, then the fitted model is

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}.$$



Linear models

- Model assumption:

$$Y_i = f(\mathbf{x}_i) + \epsilon_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \;\; i = 1, \ldots, n,$$

  where $\epsilon_i \sim N(0, \sigma^2)$.
- $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$.
- Let $\mathbf{x}_i = \{x_{i1}, \ldots, x_{ip}\}^T$ and $\boldsymbol{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}^T$.
- The regression function $\mathrm{E}(Y|\boldsymbol{X})$ is linear on β.
- Note that the linear model may be a reasonable approximation.

- Ref: `http://www.estat.me/estat/eStatU/index.html`

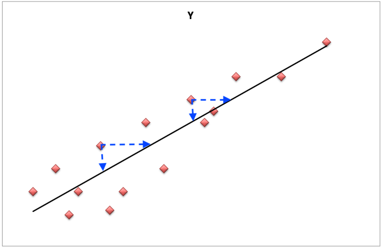
Textbook: Chapters 1-4 of Faraway, *Linear Models Using R*.

Linear models

What are the predictors $\boldsymbol{X}$?

- quantitative inputs;
- transformation of quantitative inputs: $\log(X)$, X^2, $\ldots$;
- polynomial terms or interactions terms;
- qualitative inputs or dummy variables.

Estimation



The common method is called *Least Squares*. The objective function is defined as the residual sum of squares:

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{n} [y_i - f(\mathbf{x}_i)]^2 .$$

Estimation by linear algebra

Let $\boldsymbol{\beta} = \{\beta_0, \beta_1, \ldots, \beta_p\}^T$,

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \text{ and } \boldsymbol{\epsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Then, the model is

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

and the residual sum of squares is

$$\text{RSS}(\beta) = (\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta).$$

Estimation by linear algebra

- Obtain the estimator of β by minimizing $\mathrm{RSS}(\beta)$.
- Take the first derivatives of $\mathrm{RSS}(\beta)$ with respect to β.
- $\dfrac{\partial \mathrm{RSS}(\beta)}{\partial \beta} = -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta) =^{Let} \mathbf{0}$.
- $\hat{\beta} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$.
- The point estimator is

$$\hat{\beta} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

How to read reports?

Use **lm()** in R to run the result of regression.

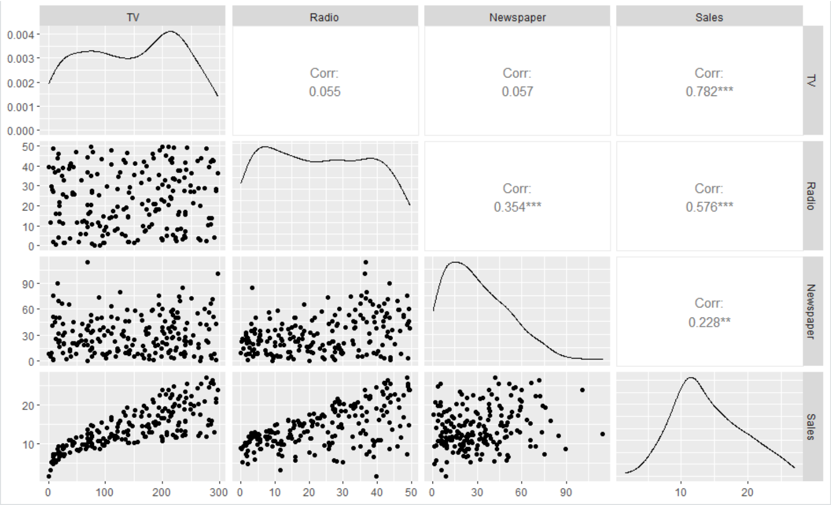
```
Call:
lm(formula = Sales ~ TV + Radio + Newspaper, data = ad)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8277 -0.8908  0.2418  1.1893  2.8292 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,	Adjusted R-squared:  0.8956 
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Other information?



Which model is better?

How to judge a model?

- $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
- Adjusted $R^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$
- Residual sum of square: $\sum_{i=1}^{n}(y_i - \hat{y})^2$
- Mean square error: $\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-p-1}$

Important questions

1. Is at least one of the predictors $x_1, x_2, \ldots, x_p$ useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Question 1: Test for overall predictors

- Null hypothesis:
- Test statistic:
- Under H_0, the distribution of F_0 is
- Reject region:

How to read reports?

Use **lm()** in R to run the result of regression.

```
Call:
lm(formula = Sales ~ TV + Radio + Newspaper, data = ad)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Test for one predictor

Before the testing for significant, the covariance-variance matrix of $\hat{\beta}$ is

$$\mathrm{Cov}(\hat{\beta}) = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\sigma^2.$$

Q: What is the value of σ^2?
We use an unbiased estimator to estimate σ^2:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p-1}.$$

Then,

$$\hat{\beta} \sim N\left(\beta, \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\sigma^2\right),$$

because

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

Test for significant

From above, the variance of $\hat{\beta}_i$ is

$$\mathrm{Var}(\hat{\beta}_i) = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{(i+1)(i+1)} \sigma^2, \;\; i = 0, \ldots p.$$

- Null Hypothesis v.s. Alternative Hypothesis
- $H_0 : \;\; \beta_i = 0$ v.s. $H_A : \;\; \beta_i \neq 0$
- If σ^2 is known, the test statistics under H_0 is

$$Z_i = \frac{\hat{\beta}_i - 0}{\sqrt{\mathrm{Var}(\hat{\beta}_i)}} \sim N(0, 1).$$

- If σ^2 is unknown, the test statistics under H_0 is

$$T_i = \frac{\hat{\beta}_i - 0}{\sqrt{\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{(i+1)(i+1)} \hat{\sigma}^2}} \sim t(n - p - 1).$$

Test for significant

If σ^2 is unknown, the test statistics under H_0 is

$$T_i = \frac{\hat{\beta}_i - 0}{\sqrt{\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{(i+1)(i+1)}\hat{\sigma}^2}} \sim t(n-p-1).$$

Reject H_0 if

$$p\{|T_i| > t_{1-\alpha/2}(n-p-1)\} \leq \alpha.$$

Usually, we set the significant level $\alpha = 0.05$.

How to read reports?

Use **lm()** in R to run the result of regression.

```
Call:
lm(formula = Sales ~ TV + Radio + Newspaper, data = ad)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8277 -0.8908  0.2418  1.1893  2.8292 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,	Adjusted R-squared:  0.8956 
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

How to read reports?

Use **lm()** in SPSS to run the result of regression.

Tests of Between-Subjects Effects

Dependent Variable:Sales

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4860.323[a]	3	1620.108	570.271	.000
Intercept	252.218	1	252.218	88.780	.000
TV	3058.010	1	3058.010	1076.406	.000
Radio	1361.737	1	1361.737	479.325	.000
Newspaper	.089	1	.089	.031	.860
Error	556.825	196	2.841		
Total	44743.250	200			
Corrected Total	5417.149	199			

a. R Squared = .897 (Adjusted R Squared = .896)

Types of sum of squares

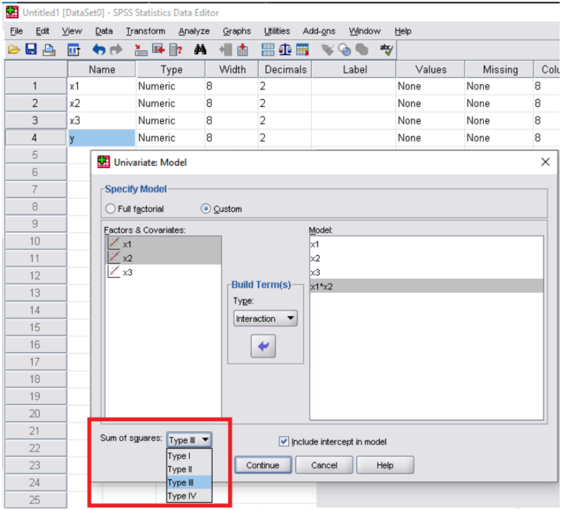
The full model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon.$$

The sources of sum of squares are x_1, x_2, $x_1 x_2$ and error.

- Type I: sequential sum of squares
  $SS(x_1) \to SS(x_2|x_1) \to SS(x_1x_2|x_1, x_2) \to SSE(x_1, x_2, x_3)$.
  $SS(x_2) \to SS(x_1|x_2) \to SS(x_1x_2|x_1, x_2) \to SSE(x_1, x_2, x_3)$.
- Type II: Only main effects
  $SS(x_1|x_2) \to SS(x_2|x_1) \to SS(x_1x_2|x_1, x_2) \to SSE(x_1, x_2, x_3)$.
- Type III: partial sum of squares
  $SS(x_1|x_2, x_1x_2) \to SS(x_2|x_1, x_1x_2) \to SS(x_1x_2|x_1, x_2) \to SSE(x_1, x_2, x_3)$.

SPSS: Types of sum of squares

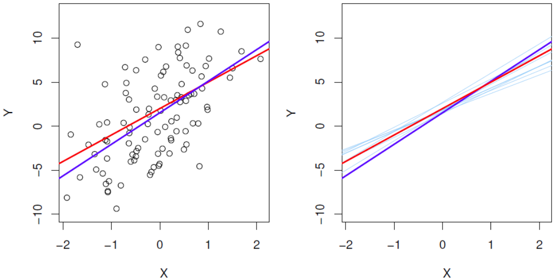


Question 4: Confidence and prediction interval

The least squares plane:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p,$$

is an "estimate" for the true regression plane.
Again, the inaccuracy in the coefficient estimates is related to samples, and We can compute a confidence interval in order to determine how close $\hat{Y}$ will be to $f(X)$. (Red: True regression; Blue: Least squares regression.)



Confidence interval for β_i

Alternative, we can evaluate the confidence interval:

$$CI = \left[\hat{\beta}_i - t_{1-\alpha/2}^{(n-p-1)} \sqrt{\mathrm{Var}(\hat{\beta}_i)}, \hat{\beta}_i + t_{1-\alpha/2}^{(n-p-1)} \sqrt{\mathrm{Var}(\hat{\beta}_i)}\right].$$

Reject H_0 if

$$0 \notin CI,$$

which means the effect of x_i is significant with evidence.

Question 4: Confidence interval for β_i

Parameter Estimates

Dependent Variable:Sales

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	2.939	.312	9.422	.000	2.324	3.554
TV	.046	.001	32.809	.000	.043	.049
Radio	.189	.009	21.893	.000	.172	.206
Newspaper	-.001	.006	-.177	.860	-.013	.011

```
> confint(fit0)
                  2.5 %     97.5 %
(Intercept)  2.32376228 3.55401646
TV           0.04301371 0.04851558
Radio        0.17154745 0.20551259
Newspaper   -0.01261595 0.01054097
```

Confidence interval for the mean response $f(\boldsymbol{x}_0)$ and prediction

Given a set of predictors $\boldsymbol{x}_0$, the fitted mean response is

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p}.$$

The CI for $f(\boldsymbol{x}_0)$ is

$$\left[\hat{y}_0 - t_{1-\alpha/2}^{(n-p-1)} \hat{\sigma} \sqrt{\boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0}, \hat{y}_0 + t_{1-\alpha/2}^{(n-p-1)} \hat{\sigma} \sqrt{\boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0}\right].$$

The CI for a single prediction at $\boldsymbol{x}_0$ is

$$\left[\hat{y}_0 - t_{1-\alpha/2}^{(n-p-1)} \hat{\sigma} \sqrt{1 + \boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0}, \hat{y}_0 + t_{1-\alpha/2}^{(n-p-1)} \hat{\sigma} \sqrt{1 + \boldsymbol{x}_0'(X'X)^{-1}\boldsymbol{x}_0}\right].$$

Prediction interval in R

```
> new.ad <- data.frame(TV = 100, Radio = 20, Newspaper = 30)
> predict(fit0, newdata = new.ad, interval = "confidence")
      fit      lwr      upr
1 13.9922 13.28186 14.70254
> predict(fit0, newdata = new.ad, interval="prediction")
      fit      lwr      upr
1 13.9922 3.924649 24.05975
```

Functions in R

Model fit component to be extracted	Function: `lm()` Package: **stats** Object: `lm.fit` Class: *lm*
Summary	`(summ <- summary(lm.fit))`
Est. method	
$\widehat{\beta}$	`coef(lm.fit)`
$\widehat{\beta}$, se$(\widehat{\beta})$, t-test	`coef(summ)`
$\widehat{\mathrm{Var}}(\widehat{\beta})$	`vcov(lm.fit)`
95% CI for β	`confint(lm.fit)`
$\widehat{\sigma}$	`summ$sigma`
95% CI for σ	

Short summary

Make the steps:

1. EDA
2. Do simple linear regressions.
3. Do multiple regression.
4. Do necessary hypotheses with conclusion.
5. Decide the final fitted model and do the further conclusion with confidence interval.

Example: Advertising data

From p.23, the predictor TV has higher correlation with the response Sales. Do the 3 simple regressions as follows:

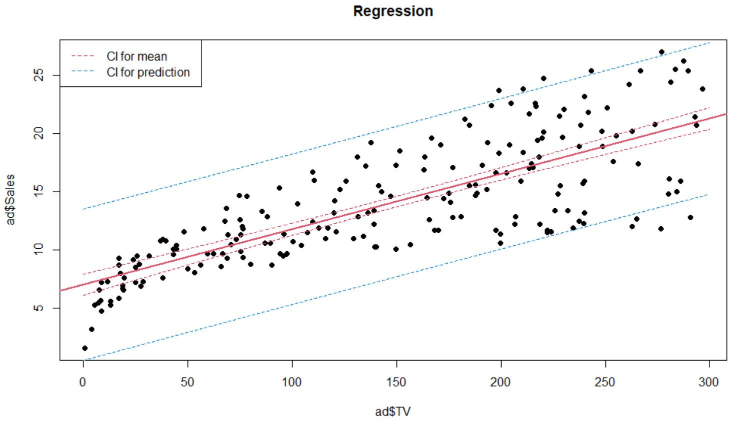
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,    Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.31164    0.56290  16.542   <2e-16 ***
Radio        0.20250    0.02041   9.921   <2e-16 ***
---
Residual standard error: 4.275 on 198 degrees of freedom
Multiple R-squared:  0.332,     Adjusted R-squared:  0.3287
F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
Newspaper    0.05469    0.01658    3.30  0.00115 **
---
Residual standard error: 5.092 on 198 degrees of freedom
Multiple R-squared:  0.05212,   Adjusted R-squared:  0.04733
F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

Example: Advertising data

Take the TV as the predictor, draw the prediction intervals of mean response.



Example: Advertising data

Multiple regression (refer to fit0)

```
Call:
lm(formula = Sales ~ TV + Radio + Newspaper, data = ad)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8277 -0.8908  0.2418  1.1893  2.8292 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,	Adjusted R-squared:  0.8956 
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Question: Why Newspaper is not significant here, but is significant in the simple regression?

Example: Advertising data

Because the predictor Newspaper is not significant, we drop off Newspaper and fit the model again. (Refer to fit12)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.92110    0.29449   9.919   <2e-16 ***
TV           0.04575    0.00139  32.909   <2e-16 ***
Radio        0.18799    0.00804  23.382   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8962
F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

Example: Advertising data

Evidence to say fit12 is good enough compared to fit0? Use the analysis of variance (ANOVA). What is the null hypothesis?

```
> anova(fit12, fit0)
Analysis of Variance Table

Model 1: Sales ~ TV + Radio
Model 2: Sales ~ TV + Radio + Newspaper
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    197 556.91
2    196 556.83  1  0.088717 0.0312 0.8599
```

Question: Can we use "adjusted R^2" or mean square error to show the evidence?

Example: Advertising data

The better model is *Sales TV* + *Radio*. Give the confidence intervals for the observations that (TV, Radio) = (50, 10) and (290, 20).

```
> CI.fit12 <- predict(fit12,
+                     newdata=data.frame(TV = c(50, 290),
+                                        Radio = c(10, 20)),
+                     interval="confidence", level = 0.95)
> CI.fit12
        fit      lwr       upr
1  7.088783  6.68398  7.493586
2 19.949881 19.48780 20.411960
> data.frame(TV = c(50, 290),
+            Radio = c(10, 20),
+            CI.fit12)
   TV Radio       fit      lwr       upr
1  50    10  7.088783  6.68398  7.493586
2 290    20 19.949881 19.48780 20.411960
```

Exercise: Credit data

Analysis the Credit data set. The response is balance (average credit card debt for each individual) and there are predictors: age, cards (number of credit cards), education (years of education), income (in thousands of dollars), limit (credit limit), rating (credit rating), own (house ownership), student (student status), status (marital status), and region (East, West or South).
Following the principle of the regression, fit the more appropriate model for the data set.