DATA ANALYSIS WITH R

Hypothesis tests and linear regression

Introduction

- There are two branches in statistics: descriptive statistics and inferential statistics.
- Data exploration and data visualization fall under descriptive statistics.
- In this section, we will cover more on inferential statistics including hypothesis tests and linear regression.

Hypothesis test

- It is a statistical method to determine which of two hypotheses (null hypothesis, H_0 and alternative hypothesis, H_1) has more evidence.

- Example:
  - Test whether two population has same mean.
  - Test whether the correlation between two variables are non-zero.

- Requires test statistic and p-value (or critical value) for making decision on hypothesis.

Hypothesis test

- Three main steps in hypothesis testing:
  1. State the null and alternative hypothesis.
  2. Calculate the test statistics.
  3. Calculate the p-value using the test statistics and make conclusion.

- Alternatively, we can compare the test statistics and critical value based on our chosen significant level.

Hypothesis test

Universiti Kebangsaan Malaysia
The National University of Malaysia

- Concluding a hypothesis test:
  - If p-value is smaller than significance level:
    - This indicates that there is evidence against the null hypothesis.
    - The null hypothesis is rejected. (or alternative hypothesis supported)
  - If p-value is bigger than significance level:
    - This indicates that there is not enough evidence against the null hypothesis.
    - The null hypothesis is not rejected. (or alternative hypothesis not supported)

Hypothesis test for means

Hypothesis test for means

- Assume that the population is normally distributed with unknown parameters.

- If you recall back, there are many cases for hypothesis test involving population means.

- Previously, we have said that, if the sample size is large, we can use the sample variance as an approximation of the population variance and use the normal distribution for the critical value or p-value.

- But in practice, we will almost always use t-distribution even if the sample size is large.

One sample, known variance

- Use z-test.
- There are no function for z-test in base R, but you can use `z.test()` function from `TeachingDemos` package.
- Hypotheses tested:
  - Null hypothesis, $H_0: \mu = \mu_0$
  - Alternative hypothesis:
    - $H_1: \mu \neq \mu_0$ (two-tailed test)
    - $H_1: \mu < \mu_0$ (left-tailed test)
    - $H_1: \mu > \mu_0$ (right-tailed test)

One sample, known variance

- Syntax: `z.test(x, mu, sd, alternative, ...)`
  - `x`: the data in vector form.
  - `mu`: hypothesized mean value.
  - `sd`: known standard deviation value.
  - `alternative`: direction of the alternative hypothesis. It can either be `"two.sided"`, `"less"`, or `"greater"`.

Example

- `x <- c(1.05, 1.11, 1.19, 1.21, 1.22, 1.29, 1.31, 1.32, 1.33, 1.37, 1.41, 1.45, 1.46, 1.65, 1.78)`

- Assuming the variance $\sigma^2 = 0.2^2$, test the following hypotheses:
  - $H_0: \mu = 1.5$ vs $H_1: \mu \neq 1.5$.
  - $H_0: \mu = 1.5$ vs $H_1: \mu > 1.5$.

Example

```
> library(TeachingDemos)
> x <- c(1.05, 1.11, 1.19, 1.21, 1.22, 1.29, 1.31,
1.32, 1.33, 1.37, 1.41, 1.45, 1.46, 1.65, 1.78)
> z.test(x=x, mu=1.5, sd=0.2, alternative="two.sided")

        One Sample z-test

data:   x
z = -3.0338, n = 15.00000, Std. Dev. = 0.20000,
Std. Dev. of the sample mean = 0.05164, p-value =
0.002415
alternative hypothesis: true mean is not equal to 1.5
95 percent confidence interval:
 1.242121 1.444545
sample estimates:
mean of x
 1.343333
```

Output of hypothesis tests in R

- In hypothesis tests, we will calculate the test statistic based on the specific test that we do.
- Then to make conclusion, we compare the test statistic with the critical value, or we calculate the p-value using the test statistic.
- When using R, we will mostly use the p-value approach when making conclusion.
- If the p-value is less than the critical value, then we will reject the null hypothesis. Otherwise, we do not reject the null hypothesis.

One sample, unknown variance

- Use t-test and the function `t.test()` in base R.

- Hypotheses tested (two sided):
  - Null hypothesis, $H_0: \mu = \mu_0$
  - Alternative hypothesis, $H_1: \mu \neq \mu_0$

- Syntax: `t.test(x, mu, alternative, ...)`
  - `x`: the data in vector form.
  - `mu`: hypothesized mean value.
  - `alternative`: direction of the alternative hypothesis. It can either be `"two.sided"`, `"less"`, or `"greater"`.

Example

- `x <- c(1.05, 1.11, 1.19, 1.21, 1.22, 1.29, 1.31, 1.32, 1.33, 1.37, 1.41, 1.45, 1.46, 1.65, 1.78)`

- Test the following hypotheses:
  - $H_0: \mu = 1.5$ vs $H_1: \mu \neq 1.5$.
  - $H_0: \mu = 1.5$ vs $H_1: \mu < 1.5$.

Example

```
> x <- c(1.05, 1.11, 1.19, 1.21, 1.22, 1.29, 1.31,
1.32, 1.33, 1.37, 1.41, 1.45, 1.46, 1.65, 1.78)
> t.test(x=x, mu=1.5, alternative="two.sided")

        One Sample t-test

data:  x
t = -3.1589, df = 14, p-value = 0.006967
alternative hypothesis: true mean is not equal to 1.5
95 percent confidence interval:
 1.236962 1.449704
sample estimates:
mean of x
 1.343333
```

Comparing means of two samples

- Use `t.test()` again, but specify the `x` and `y` value.

- Hypotheses tested (two sided):
  - Null hypothesis, $H_0: \mu_1 = \mu_2$
  - Alternative hypothesis, $H_1: \mu_1 \neq \mu_2$

Comparing means of two samples

- Syntax: `t.test(x, y, mu, alternative, paired, var.equal, ...)`
  - `x`: the data in vector form for first sample.
  - `y`: the data in vector form for second sample.
  - `mu`: hypothesized mean value difference.
  - `alternative`: direction of the alternative hypothesis. It can either be `"two.sided"`, `"less"`, or `"greater"`.
  - `paired`: `TRUE` if it is a paired data, `FALSE` otherwise.
  - `var.equal`: `TRUE` if assume the two samples have equal variances, `FALSE` otherwise.

Example

- Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process. Specifically, catalyst 1 is currently used; but catalyst 2 is acceptable but cheaper. A test is run to check if catalyst 2 does not change the process yield. Is there any difference in the mean yields? Use $\alpha = 0.05$, and assume the data is normally distributed with equal variances.

- `cat1 <- c(91.50, 94.18, 92.18, 95.39, 91.79, 89.07, 94.72, 89.21)`
- `cat2 <- c(89.19, 90.95, 90.46, 93.21, 97.19, 97.04, 91.07, 92.75)`

Example

```
> cat1 <- c(91.50, 94.18, 92.18, 95.39, 91.79, 89.07, 
94.72, 89.21)
> cat2 <- c(89.19, 90.95, 90.46, 93.21, 97.19, 97.04, 
91.07, 92.75)
> t.test(x=cat1, y=cat2, alternative="two.sided", 
mu=0, var.equal=TRUE)

        Two Sample t-test

data:  cat1 and cat2
t = -0.35359, df = 14, p-value = 0.7289
alternative hypothesis: true difference in means is 
not equal to 0
95 percent confidence interval:
 -3.373886  2.418886
sample estimates:
mean of x mean of y 
  92.2550   92.7325
```

Testing mean equality for more than two samples

- Assume (1) normality, (2) equal variance, and (3) independent samples.
- Use one-way ANOVA (analysis of variance) and the `aov()` function in base R.
- Hypotheses tested:
  - Null hypothesis, H_0: The means for each category/group is the same ($\mu_1 = \mu_2 = \cdots = \mu_k$)
  - Alternative hypothesis, H_1: There are at least two categories/groups with different means ($\mu_i \neq \mu_j$ for some i and j)

Testing mean equality for more than two samples

- Syntax: `aov(formula, data, ...)`
  - `formula`: the formula specifying model.
  - `data`: the data frame which the variable in the formula is from.

- The `aov()` function gives the ANOVA table.

- To get the p-value, we will have to use the `summary()` function.
  - Eg: `summary(aov(...))`

Example

- Using the `iris` dataset in R (`data(iris)`), test whether the three species (*setosa*, *versicolor* and *virginica*) have the same mean sepal length.

Example

```
> data(iris)
> aov(Sepal.Length~Species,data=iris)
Call:
   aov(formula = Sepal.Length ~ Species, data = iris)

Terms:
                 Species Residuals
Sum of Squares  63.21213  38.95620
Deg. of Freedom        2       147

Residual standard error: 0.5147894
Estimated effects may be unbalanced
> summary(aov(Sepal.Length~Species,data=iris))
             Df Sum Sq Mean Sq F value Pr(>F)
Species       2  63.21  31.606   119.3 <2e-16 ***
Residuals   147  38.96   0.265
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
```

Hypothesis test for variances

One sample

- Assume population is normally distributed
- Use chi-squared test for variance.
- Hypotheses tested (two sided):
  - Null hypothesis, $H_0: \sigma^2 = \sigma_0^2$
  - Alternative hypothesis, $H_1: \sigma^2 \neq \sigma_0^2$
- Unfortunately, there are no built-in function for chi-square test for variance in base R. But we can use `varTest()` function in `EnvStats` package.

One sample

- Syntax: `varTest(x, alternative, sigma.squared, ...)`
  - `x`: the data in vector form.
  - `alternative`: direction of the alternative hypothesis. It can either be `"two.sided"`, `"less"`, or `"greater"`.
  - `sigma.squared`: hypothesized value for σ^2.

Example

- `x <- c(1.05, 1.11, 1.19, 1.21, 1.22, 1.29, 1.31, 1.32, 1.33, 1.37, 1.41, 1.45, 1.46, 1.65, 1.78)`

- Test the following hypotheses:
  - $H_0: \sigma^2 = 0.04$ vs $H_1: \sigma^2 \neq 0.04$.

Example

```
> library(EnvStats)
> x <- c(1.05, 1.11, 1.19, 1.21, 1.22, 1.29, 1.31,
1.32, 1.33, 1.37, 1.41, 1.45, 1.46, 1.65, 1.78)
> print(varTest(x=x, alternative="two.sided",
sigma.squared=0.04))

Results of Hypothesis Test
--------------------------

...

Test Statistic:                  Chi-Squared = 12.91333

Test Statistic Parameter:        df = 14

P-value:                         0.9332821

...
```

Two samples

- Assume the populations are normally distributed.

- Use F-test of equality of variances, and the `var.test()` function in base R.

- Hypotheses tested (two sided):
  - Null hypothesis, $H_0: \sigma_1^2 = \sigma_2^2$
  - Alternative hypothesis, $H_1: \sigma_1^2 \neq \sigma_2^2$

Two samples

- Syntax: `var.test(x, y, alternative, ...)`
  - `x`: the data in vector form for first sample.
  - `y`: the data in vector form for second sample.
  - `alternative`: direction of the alternative hypothesis. It can either be `"two.sided"`, `"less"`, or `"greater"`.

Example

- Using the catalysts data in previous slides, does the two populations have equal variance?

- `cat1 <- c(91.50, 94.18, 92.18, 95.39, 91.79, 89.07, 94.72, 89.21)`
- `cat2 <- c(89.19, 90.95, 90.46, 93.21, 97.19, 97.04, 91.07, 92.75)`

Example

```
> cat1 <- c(91.50, 94.18, 92.18, 95.39, 91.79, 89.07,
94.72, 89.21)
> cat2 <- c(89.19, 90.95, 90.46, 93.21, 97.19, 97.04,
91.07, 92.75)
> var.test(x=cat1, y=cat2, alternative="two.sided")

      F test to compare two variances

data:  cat1 and cat2
F = 0.63907, num df = 7, denom df = 7, p-value =
0.5691
alternative hypothesis: true ratio of variances is not
equal to 1
95 percent confidence interval:
 0.1279433 3.1920724
sample estimates:
ratio of variances
         0.6390651
```

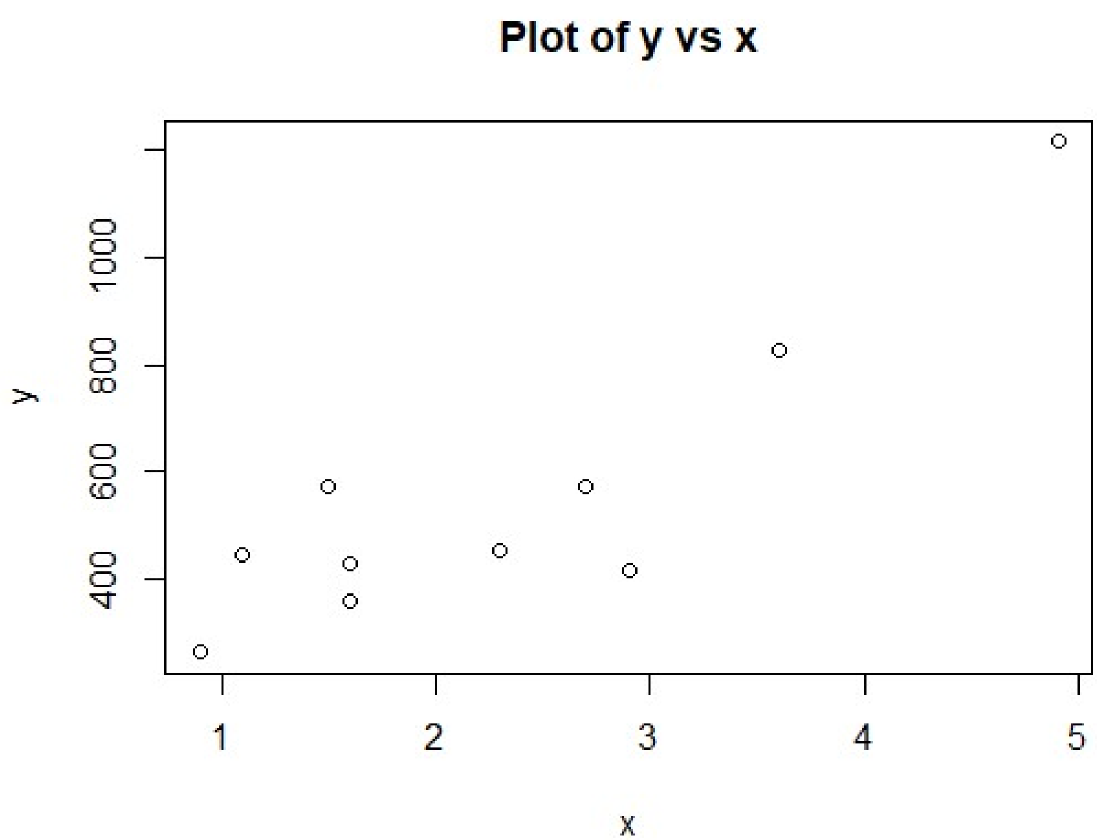
Linear regression analysis

Introduction to linear regression

- An economist want to determine whether there is a linear relationship between a country's gross domestic product (GDP) and carbon dioxide (CO2) emissions. The data are shown in the table.

GDP (trillions of \$), x	CO_2 emission (millions of metric tons), y
1.6	428.2
3.6	828.8
4.9	1214.2
1.1	444.6
0.9	264
2.9	415.3
2.7	571.8
2.3	454.9
1.6	358.7
1.5	573.5

Introduction to linear regression



Introduction to linear regression

- In regression analysis, there are 2 types of variables:
  1) Dependent variable, y (response/outcome variable)
  2) Independent variable, x (predictor/regressor/explanatory variable)

- The most basic type of regression, is the linear regression.

- For linear regression, we assume that there is an underlying linear relationship between the dependent variable y and the independent variable x.

Introduction to linear regression

- Regression analysis is a statistical method that is used to study:
  - Relationship – among the variables (2 or more)
  - Forecast/predict – predict the value of variable interest
- Useful in many areas of study, such as in economics, physics, biology, social science, engineering, technology and business management.

Simple linear regression

Simple linear regression

- In simple linear regression, we only have one regressor or independent variable x.

- The simple linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

for an observation i with response variable y_i and explanatory variable x_i, where ε_i is the random error term.

- We assume that the error terms are (1) independent, (2) have zero mean, (3) have constant variance and (4) normally distributed.

Linear regression in R

- We can use the `lm()` function in R to fit the linear regression to the data.

- Syntax: `lm(formula, data)`
  - `formula`: the formula specifying model.
  - `data`: the data frame which the variable in the formula is from.

Example

- Using the GDP and CO_2 data presented in the previous slides, fit a linear model to the data.

GDP (trillions of \$), x	CO_2 emission (millions of metric tons), y
1.6	428.2
3.6	828.8
4.9	1214.2
1.1	444.6
0.9	264
2.9	415.3
2.7	571.8
2.3	454.9
1.6	358.7
1.5	573.5

Example

```
> x <- c(1.6,3.6,4.9,1.1,0.9,2.9,2.7,2.3,1.6,1.5)
> y <- c(428.2,828.8,1214.2,444.6,264,415.3,571.8,454.9,358.7,573.5)
> GDP <- data.frame(x=x,y=y)
> fit <- lm(y~x, data=GDP)
> summary(fit)

Call:
lm(formula = y ~ x, data = GDP)

Residuals:
     Min        1Q    Median        3Q       Max
-255.830  -59.432   -1.379   99.999  176.983

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   102.29      95.93   1.066 0.317416
x             196.15      36.96   5.306 0.000723 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 138.3 on 8 degrees of freedom
Multiple R-squared:  0.7788, Adjusted R-squared:  0.7511
F-statistic: 28.16 on 1 and 8 DF,  p-value: 0.0007227
```

Example

```
> summary(fit)

Call:
lm(formula = y ~ x, data = GDP)

Residuals:
     Min       1Q   Median       3Q      Max
-255.830  -59.432   -1.379   99.999  176.983

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   102.29      95.93   1.066 0.317416
x             196.15      36.96   5.306 0.000723 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 138.3 on 8 degrees of freedom
Multiple R-squared:  0.7788,     Adjusted R-squared:  0.7511
F-statistic: 28.16 on 1 and 8 DF,  p-value: 0.0007227
```

Least squares estimates:

$$\hat{\beta}_0 = 102.29$$

$$\hat{\beta}_1 = 196.15$$

p-values for testing the parameters $\beta = 0$ vs $\beta \neq 0$

Value of R^2

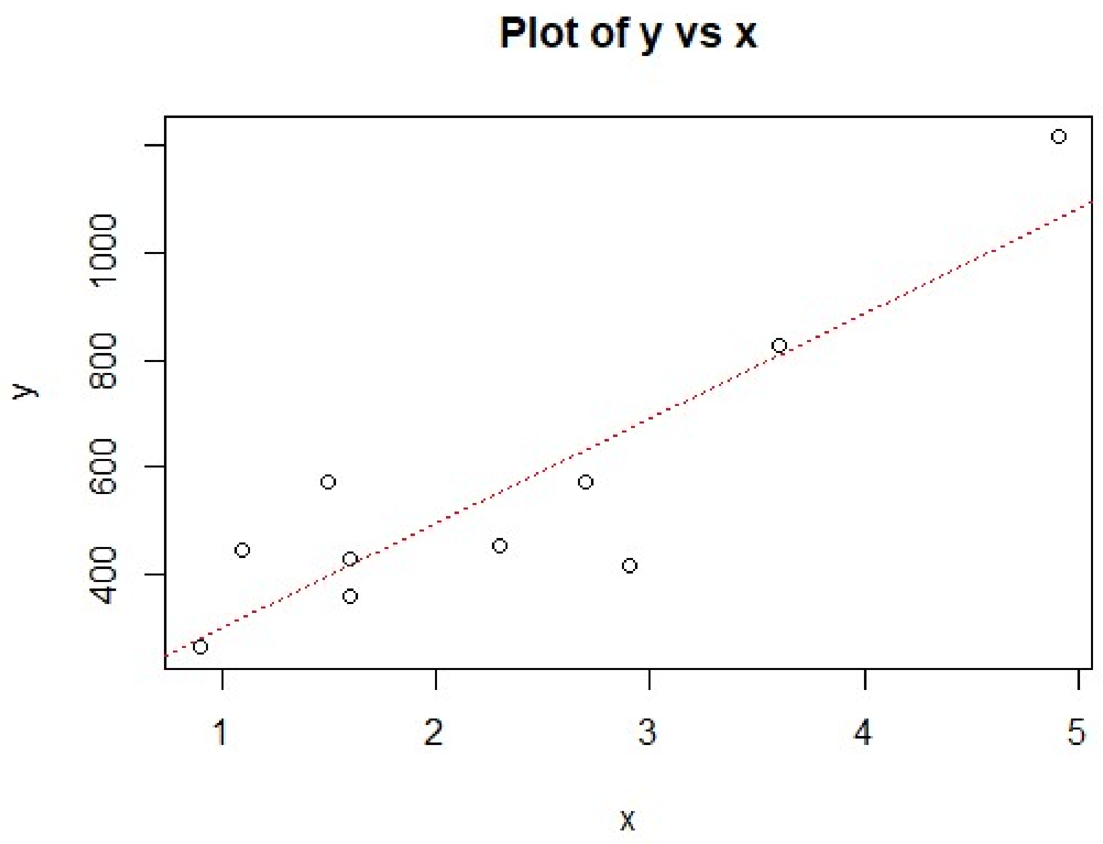
A few notes

- For simple linear regression, the hypothesis test $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$ relates to the significance of regression.
- IF H_0 is not rejected (i.e. $\beta_1 = 0$) $\Longrightarrow$ there is no linear relationship between x and y.
- In previous example, it appears that x is significant in the linear model ($\beta_1 \neq 0$) and that the linear relationship is significant.
- Also, from the previous example, the fitted regression line is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 102.29 + 196.15x$$

Plotting the regression line

```
fit <- lm(y~x, data=GDP)
plot(y~x, data=GDP, main="Plot of y vs x")
abline(fit, col="red", lty=3)
```



Universiti Kebangsaan Malaysia
The National University of Malaysia

Multiple linear regression

Multiple linear regression

- Multiple linear regression is similar to simple linear regression, except we have more than one regressor or independent variables.

- The model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$
$$= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i$$

  for $i = 1, \ldots, n$ and $j = 1, \ldots, k$ is called a multiple linear regression with k regressors or independent variables.

- The parameter β_j represents the expected change in response y per unit change in x_j when all remaining regressor variables are held constant.

Multiple linear regression in R

- The `lm()` function can be used for multiple linear regression as well. We just need to add more variables in the `formula`.

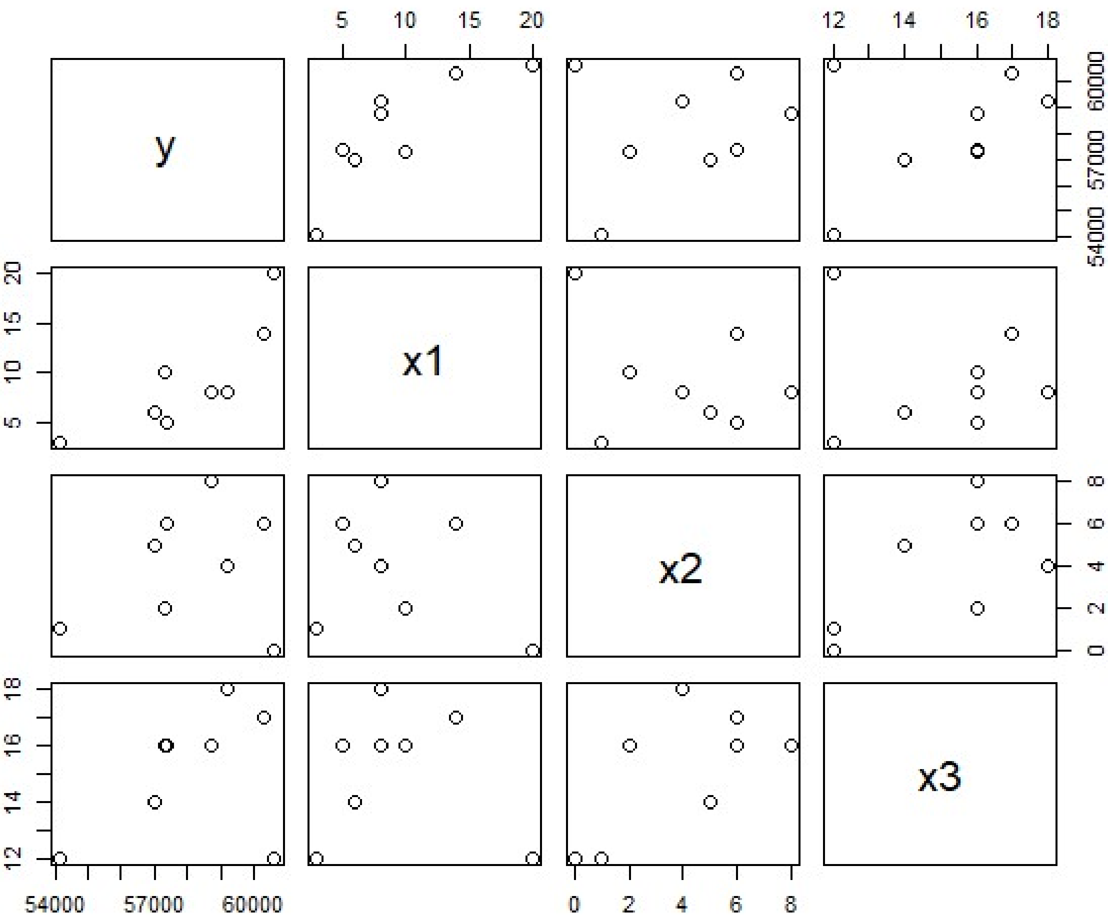
- Example:
  - `fit <- lm(y~x1+x2+x3, data=dataset)`

Example

- A researcher wants to determine how employee salaries at a certain company are related to the length of employment, previous experience, and education. The researcher selects eight employees from the company and obtains the data.

Employee	Salary, y	Employment (yrs), x_1	Experience (yrs), x_2	Education (yrs), x_3
A	57,310	10	2	16
B	57,380	5	6	16
C	54,135	3	1	12
D	56,985	6	5	14
E	58,715	8	8	16
F	60,620	20	0	12
G	59,200	8	4	18
H	60,320	14	6	17

Example



Example

```
> y <- c(57310,57380,54135,56985,58715,60620,59200,60320)
> x1 <- c(10,5,3,6,8,20,8,14)
> x2 <- c(2,6,1,5,8,0,4,6)
> x3 <- c(16,16,12,14,16,12,18,17)
> Employment <- data.frame(y=y,x1=x1,x2=x2,x3=x3)
> fit <- lm(y~x1+x2+x3, data=Employment)
> summary(fit)

Call:
lm(formula = y ~ x1 + x2 + x3, data = Employment)

Residuals:
      1       2       3       4       5       6       7       8 
-824.76  156.82 -153.52  158.90  -56.65  364.09  804.95 -449.82 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 49764.45    1981.35  25.116 1.49e-05 ***
x1            364.41      48.32   7.542  0.00166 ** 
x2            227.62     123.84   1.838  0.13991    
x3            266.94     147.36   1.812  0.14430    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 659.5 on 4 degrees of freedom
Multiple R-squared:  0.9438, Adjusted R-squared:  0.9017
F-statistic:  22.4 on 3 and 4 DF,  p-value: 0.005804
```

Note

- In the previous example, only the variable x_1 is significant to the model.

- For x_1, the p-value is less than 0.01. So, there is strong evidence against the hypothesis that the parameter associated with it is zero.

- The other variables have p-values greater than 0.1. There is no evidence to reject the hypothesis that the parameter associated with them to be zero.

Dummy variable

Example

- Using the previous example, suppose we have the following data with gender.
- Now, suppose we only want to regress the salary using employment duration (x_1) and gender (x_4).

Employee	Salary, y	Employment (yrs), x_1	Experience (yrs), x_2	Education (yrs), x_3	Gender, x_4
A	57,310	10	2	16	Male
B	57,380	5	6	16	Male
C	54,135	3	1	12	Female
D	56,985	6	5	14	Female
E	58,715	8	8	16	Female
F	60,620	20	0	12	Male
G	59,200	8	4	18	Male
H	60,320	14	6	17	Female

Dummy variable model

- In the example, the categorical variable can be entered into the regression model through dummy or indicator variables.

- Variable levels coded 0 and 1.

- E.g. If there are two categories:

$$x_4 = \begin{cases} 0, & \text{if observation from category Female} \\ 1, & \text{if observation from category Male} \end{cases}$$

Example

- The model: $y = \beta_0 + \beta_1 x_1 + \beta_4 x_4 + \varepsilon$

- For female:

$$y = \beta_0 + \beta_1 x_1 + \beta_4(0) = \beta_0 + \beta_1 x_1$$

- For male:

$$y = \beta_0 + \beta_1 x_1 + \beta_4(1) = \beta_0 + \beta_1 x_1 + \beta_4$$

- β_4 is the average difference in Y between the two categories, when other variables (in this case x_1) is the same.

Example

```
> y <- c(57310,57380,54135,56985,58715,60620,59200,60320)
> x1 <- c(10,5,3,6,8,20,8,14)
> x2 <- c(2,6,1,5,8,0,4,6)
> x3 <- c(16,16,12,14,16,12,18,17)
> x4 <- c("Male","Male","Female","Female","Female","Male","Male",
          "Female")
> Employment <- data.frame(y=y,x1=x1,x2=x2,x3=x3,x4=x4)
> fit <- lm(y~x1+x4, data=Employment)
> summary(fit)

Call:
lm(formula = y ~ x1 + x4, data = Employment)

Residuals:
        1         2         3         4         5         6
-1083.174   549.000 -1919.685    -6.989  1098.141  -897.521
        7         8
 1431.695   828.533

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  55117.4     1056.8  52.154  4.9e-08 ***
x1             312.4      101.7   3.071   0.0277 *
x4Male         151.4     1041.1   0.145   0.8900

...
```

3 or more categories

- For 3 or more categories, the qualitative variable entered the model using more dummy variables.

- Example:

$$E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where

$$x_1 = \begin{cases} 1, & \text{if Group B} \\ 0, & \text{otherwise} \end{cases}, \qquad x_2 = \begin{cases} 1, & \text{if Group C} \\ 0, & \text{otherwise} \end{cases}$$

x_1	x_2	
0	0	If the observation is from Group A
1	0	If the observation is from Group B
0	1	If the observation is from Group C

Functions & descriptions

Function	Description
`z.test()`	Used to run a z-test for population mean. Available in `TeachingDemos` package.
`t.test()`	Used to run t-test for population means. It can be used for one sample, two samples or paired two samples.
`aov()`	It can be used to run one-way ANOVA to test for equality of means. Use `summary()` for more details including p-value.
`varTest()`	Used to run chi-squared test for variance (one population). Available in `EnvStats` package.
`var.test()`	Used to run F-test for two population variances.
`prop.test()`	Used to run proportion test (not covered in this lecture).
`lm()`	Used to fit a linear model to a data. Use `summary()` to the `lm` object for more details.