IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

REGRESSION

Linear Regression-III

Dr. Ramesh Anbanandam

DEPARTMENT of Management Studies

Learning Objectives

- Understanding Coefficient of Determination
- Test statistical hypotheses and construct confidence intervals on regression model parameters

Coefficient of Determination

- Relationship Among SST, SSR, SSE

SST = SSR + SSE

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$$SS_{yy} = \left(\frac{SS_{xy}{}^2}{SS_{xx}} \right) + \left(SS_{yy} - \frac{SS_{xy}{}^2}{SS_{xx}} \right)$$

where:

SST = total sum of squares

SSR = sum of squares due to regression

SSE = sum of squares due to error

Coefficient of Determination

- The coefficient of determination is:

$$r^2 = SSR/SST$$

where:

SSR = sum of squares due to regression

SST = total sum of squares

Coefficient of Determination

$$r^2 = SSR/SST = 100/114 = .8772$$

The regression relationship is very strong; 88% of the variability in the number of cars sold can be explained by the linear relationship between the number of TV ads and the number of cars sold.

Jupyter code

```
In [5]: t= tbl['TV Ads']
        c= tbl['car Sold']
```

```
In [8]: import statsmodels.api as s
        t = s.add_constant(t)
        model1 = sm.OLS(c,t)
        result1 = model1.fit()
        print(result1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               car Sold   R-squared:                       0.877
Model:                            OLS   Adj. R-squared:                  0.836
Method:                 Least Squares   F-statistic:                     21.43
Date:                Fri, 30 Aug 2019   Prob (F-statistic):             0.0190
Time:                        08:31:20   Log-Likelihood:                -9.6687
No. Observations:                   5   AIC:                             23.34
Df Residuals:                       3   BIC:                             22.56
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         10.0000      2.366      4.226      0.024       2.469      17.531
TV Ads         5.0000      1.080      4.629      0.019       1.563       8.437
==============================================================================
Omnibus:                          nan   Durbin-Watson:                   1.214
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.674
Skew:                           0.256   Prob(JB):                        0.714
Kurtosis:                       1.276   Cond. No.                         6.33
==============================================================================
```

Coefficient of Determination

$$r^2 = SSR/SST = 100/114 = .8772$$

The regression relationship is very strong; 88% of the variability in the number of cars sold can be explained by the linear relationship between the number of TV ads and the number of cars sold.

Sample Correlation Coefficient

$$r_{xy} = (\text{sign of } b_1)\sqrt{\text{Coefficient of Determination}}$$

$$r_{xy} = (\text{sign of } b_1)\sqrt{r^2}$$

$$\hat{y} = b_0 + b_1 x$$

where:

b_1 = the slope of the estimated regression equation

Sample Correlation Coefficient

$$r_{xy} = (\text{sign of } b_1)\sqrt{r^2}$$

The sign of b_1 in the equation $\hat{y} = 10 + 5x$ is "+".

$$r_{xy} = +\sqrt{.8772}$$

r_{xy} = +.9366

Assumptions About the Error Term *e*

1. The error e is a random variable with mean of zero.
2. The variance of e, denoted by e^2, is the same for all values of the independent variable.
3. The values of e are independent.
4. The error e is a normally distributed random variable.

Testing for Significance

- To test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of β_1 is zero.
- Two tests are commonly used:

t Test and *F* Test

- Both the *t* test and *F* test require an estimate of s^2, the variance of e in the regression model.

Estimate of *s*

- An Estimate of *s*

  The mean square error (MSE) provides the estimate of s^2, and the notation s^2 is also used.

$$s^2 = \text{MSE} = \text{SSE}/(n - 2)$$

where:

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

Testing for Significance

- An Estimate of *s*
  - To estimate *s* we take the square root of s^2.
  - The resulting *s* is called the <u>standard error of the estimate</u>.

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n-2}}$$

Testing for Significance

$$S_e = \text{Standard error of the estmiate } (\sigma^2)$$

$$= \frac{SSE}{n-2} = \frac{S_{yy} - \frac{{S_{xy}}^2}{S_{xx}}}{n-2}$$

Testing for Significance: *t* Test

- Hypotheses

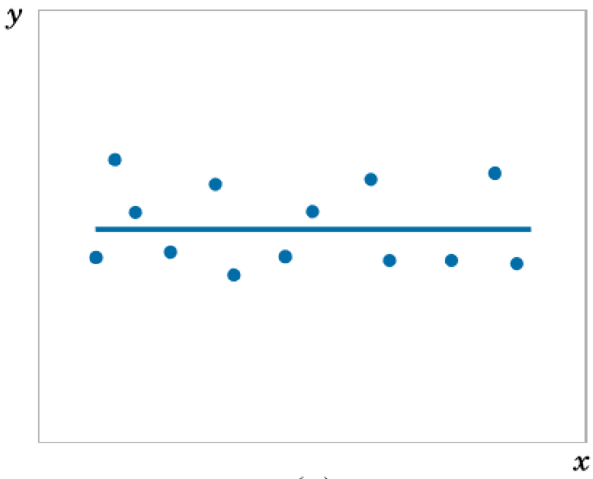
$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

- Test Statistic

$$t = \frac{b_1}{s_{b_1}}$$

Case 1

$$H_0 : \beta_1 = 0$$

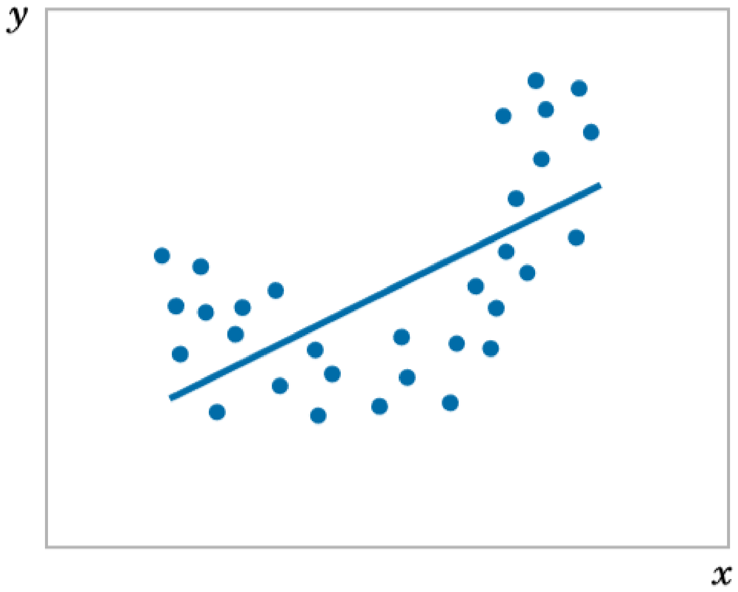


(a)

In this case hypothesis is not rejected

Case 2

$H_a : \beta_1 \neq 0$



(b)

In this case hypothesis is rejected

The Standard Deviation of the Regression Slope

- The standard error of the regression slope coefficient (b_1) is estimated by

$$\boxed{s_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum (x-\bar{x})^2}} = \frac{s_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}}$$

where:

s_{b_1} = Estimate of the standard error of the least squares slope

$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$ = Sample standard error of the estimate

Testing for Significance: t Test

- Rejection Rule

Reject H_0 if p-value $\leq \alpha$

or $t \leq -t_{\alpha/2}$ or $t \geq t_{\alpha/2}$

where:

$t_{\alpha/2}$ is based on a t distribution

with n - 2 degrees of freedom

Testing for Significance: t Test

1. Determine the hypotheses.

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

2. Specify the level of significance. α = .05

3. Select the test statistic.

$$t = \frac{b_1}{s_{b_1}}$$

4. State the rejection rule.

Reject H_0 if p-value $\leq$ .05 or $|t| > 3.182$ (with 3 degrees of freedom)

Testing for Significance: t Test

5. Compute the value of the test statistic.

$$t = \frac{b_1 - \beta_1}{s_{b_1}} = \frac{5}{1.08} = 4.63$$

6. Determine whether to reject H_0.

   t = 4.541 provides an area of .01 in the upper tail. Hence, the p-value is less than .02. (Also, t = 4.63 > 3.182.) We can reject H_0.

Hypothesis Tests for the Slope of the Regression Model

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

$$H_0: \beta_1 \leq 0$$
$$H_1: \beta_1 > 0$$

$$H_0: \beta_1 \geq 0$$
$$H_1: \beta_1 < 0$$

$$t = \frac{b_1 - \beta_1}{S_b}$$

where: $$S_b = \frac{S_e}{\sqrt{SSxx}}$$

$$S_e = \sqrt{\frac{SSE}{n-2}}$$

$$SSxx = \sum X^2 - \frac{(\sum X)^2}{n}$$

$$\beta_1 = \text{the hypothesized slope}$$

$$df = n - 2$$

Confidence Interval for β_1

- We can use a 95% confidence interval for β_1 to test the hypotheses just used in the *t* test.
- H_0 is rejected if the hypothesized value of β_1 is not included in the confidence interval for β_1.

Confidence Interval for β_1

- The form of a confidence interval for β_1 is:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

b_1 is the point estimator

$t_{\alpha/2} s_{b_1}$ is the margin of error

Where $t_{\alpha/2}$ is the t value providing an area of a/2 in the upper tail of a t distribution with n - 2 degrees of freedom

Confidence Interval for β_1

- Rejection Rule

  Reject H_0 if 0 is not included in the confidence interval for β_1.

- 95% Confidence Interval for β_1

  $\boxed{b_1 \pm t_{\alpha/2} s_{b_1}}$ = 5 +/- 3.182(1.08) = 5 +/- 3.44

  or 1.56 to 8.44

- Conclusion

  0 is not included in the confidence interval.

  Reject H_0

Testing for Significance: F Test

- Hypotheses

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

- Test Statistic

$$F = \text{MSR/MSE}$$

F-Test for Significance

- F Test statistic:

$$F = \frac{MSR}{MSE}$$

where

$$MSR = \frac{SSR}{k}$$

$$MSE = \frac{SSE}{n-k-1}$$

where F follows an F distribution with k numerator degrees of freedom and (n - k - 1) denominator degrees of freedom
(k = the number of independent variables in the regression model)

Testing for Significance: F Test

- Rejection Rule

Reject H_0 if
p-value $\leq \alpha$
or $F \geq F_\alpha$

where:

F_α is based on an F distribution with
1 degree of freedom in the numerator and
n - 2 degrees of freedom in the denominator

Testing for Significance: F Test

1. Determine the hypotheses. $H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$

2. Specify the level of significance. $\alpha = .05$

3. Select the test statistic. F = MSR/MSE

4. State the rejection rule. Reject H_0 if p-value $\leq$ .05 or $F \geq$ 10.13 (with 1 d.f. in numerator and 3 d.f. in denominator)

Jupyter Code

In [2]:
```
import numpy as np
import matplotlib.pyplot as plt
```

In [3]:
```
import seaborn as sns
```

In [4]:
```
import pandas as pd
import matplotlib as mpl
import statsmodels.formula.api as sm
from sklearn.linear_model import LinearRegression
from scipy import stats
```

In [5]:
```
tbl = pd.read_excel('C:/Users/Somi/Documents/regr.xlsx')
```

Jupyter code

In [5]:
```
t= tbl['TV Ads']
c= tbl['car Sold']
```

In [8]:
```
import statsmodels.api as s
t = s.add_constant(t)
model1 = sm.OLS(c,t)
result1 = model1.fit()
print(result1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               car Sold   R-squared:                       0.877
Model:                            OLS   Adj. R-squared:                  0.836
Method:                 Least Squares   F-statistic:                     21.43
Date:                Fri, 30 Aug 2019   Prob (F-statistic):             0.0190
Time:                        08:31:20   Log-Likelihood:                -9.6687
No. Observations:                   5   AIC:                             23.34
Df Residuals:                       3   BIC:                             22.56
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         10.0000      2.366      4.226      0.024       2.469      17.531
TV Ads         5.0000      1.080      4.629      0.019       1.563       8.437
==============================================================================
Omnibus:                          nan   Durbin-Watson:                   1.214
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.674
Skew:                           0.256   Prob(JB):                        0.714
Kurtosis:                       1.276   Cond. No.                         6.33
==============================================================================
```

Testing for Significance: F Test

5. Compute the value of the test statistic.

$$F = \text{MSR/MSE} = 100/4.667 = 21.43$$

6. Determine whether to reject H_0.

   F = 17.44 provides an area of .025 in the upper tail. Thus, the p-value corresponding to F = 21.43 is less than 2(.025) = .05. Hence, we reject H_0.

The statistical evidence is sufficient to conclude that we have a significant relationship between the number of TV ads aired and the number of cars sold.

Some Cautions about the Interpretation of Significance Tests

- Rejecting H_0: $\beta_1 = 0$ and concluding that the relationship between *x* and *y* is significant does not enable us to conclude that a cause-and-effect relationship is present between *x* and *y*.
- Just because we are able to reject H_0: $\beta_1 = 0$ and demonstrate statistical significance does not enable us to conclude that there is a linear relationship between *x* and *y*.

Thank You