IIT ROORKEE

FREE ONLINE EDUCATION swayam शिक्षित भारत, उन्नत भारत

NPTEL ONLINE CERTIFICATION COURSE

MULTIPLE REGRESSION MODEL-II

Dr. A. Ramesh

DEPARTMENT OF MANAGEMENT STUDIES

Agenda

- Testing for significance
  - F Test
  - t Test
- Python Demo for multiple regression

Testing for Significance

- The F test is used to determine whether a significant relationship exists between the dependent variable and the set of all the independent variables; we will refer to the F test as the test for overall significance.
- If the F test shows an overall significance, the t test is used to determine whether each of the individual independent variables is significant.
- A separate t test is conducted for each of the independent variables in the model; we refer to each of these t tests as a test for individual significance.

F Test

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

The hypotheses for the F test involve the parameters of the multiple regression model.

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_a: One or more of the parameters is not equal to zero

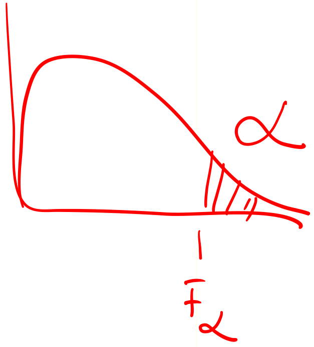
F test significance

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

H_a: One or more of the parameters is not equal to zero

TEST STATISTIC

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{SSR/p}{SSE/(n-p-1)}$$



REJECTION RULE

p-value approach:	Reject H_0 if p-value $\leq \alpha$
Critical value approach:	Reject H_0 if $F \geq F_\alpha$

where F_α is based on an F distribution with p degrees of freedom in the numerator and $n - p - 1$ degrees of freedom in the denominator.

F test significance

$$F = \frac{10.8}{.328} = 32.9$$

F Test

In [15]:
```
from statsmodels.formula.api import ols
model = ols('travel_time ~ x1+n_of_deliveries ', data=df1).fit()
model.summary()
```

```
C:\Users\HP\Anaconda3\lib\site-packages\scipy\stats\stats.py:1390: UserWa
g anyway, n=10
  "anyway, n=%i" % int(n))
```

Out[15]:

OLS Regression Results

Dep. Variable:	travel_time	**R-squared:**	0.904
Model:	OLS	**Adj. R-squared:**	0.876
Method:	Least Squares	**F-statistic:**	32.88
Date:	Fri, 06 Sep 2019	**Prob (F-statistic):**	0.000276
Time:	11:16:53	**Log-Likelihood:**	-6.8398
No. Observations:	10	**AIC:**	19.68
Df Residuals:	7	**BIC:**	20.59
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.8687	0.952	-0.913	0.392	-3.119	1.381
x1	0.0611	0.010	6.182	0.000	0.038	0.085
n_of_deliveries	0.9234	0.221	4.176	0.004	0.401	1.446

Omnibus:	0.039	**Durbin-Watson:**	2.515
Prob(Omnibus):	0.981	**Jarque-Bera (JB):**	0.151
Skew:	0.074	**Prob(JB):**	0.927
Kurtosis:	2.418	**Cond. No.**	435.

ANOVA table

Source	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	SSR	p	$\text{MSR} = \dfrac{\text{SSR}}{p}$	$F = \dfrac{\text{MSR}}{\text{MSE}}$
Error	SSE	$n - p - 1$	$\text{MSE} = \dfrac{\text{SSE}}{n - p - 1}$	
Total	SST	$n - 1$		

t Test for individual significance

For any parameter β_i

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

TEST STATISTIC

$$t = \frac{b_i - \beta_i}{s_{b_i}} = \frac{b_i}{s_{b_i}}$$

REJECTION RULE

p-value approach:	Reject H_0 if p-value $\leq \alpha$
Critical value approach:	Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

where $t_{\alpha/2}$ is based on a t distribution with $n - p - 1$ degrees of freedom.

t Test for individual significance

$$b_1 = .061135 \quad s_{b_1} = .009888$$
$$b_2 = .9234 \quad s_{b_2} = .2211$$

$$t = .061135/.009888 = 6.18$$
$$t = .9234/.2211 = 4.18$$

t Test for individual significance

In [15]:
```
from statsmodels.formula.api import ols
model = ols('travel_time ~ x1+n_of_deliveries ', data=df1).fit()
model.summary()
```

```
C:\Users\HP\Anaconda3\lib\site-packages\scipy\stats\stats.py:1390: UserWa
g anyway, n=10
  "anyway, n=%i" % int(n))
```

Out[15]:

OLS Regression Results

Dep. Variable:	travel_time	**R-squared:**	0.904
Model:	OLS	**Adj. R-squared:**	0.876
Method:	Least Squares	**F-statistic:**	32.88
Date:	Fri, 06 Sep 2019	**Prob (F-statistic):**	0.000276
Time:	11:16:53	**Log-Likelihood:**	-6.8398
No. Observations:	10	**AIC:**	19.68
Df Residuals:	7	**BIC:**	20.59
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-0.8687	0.952	-0.913	0.392	-3.119	1.381
x1	0.0611	0.010	6.182	0.000	0.038	0.085
n_of_deliveries	0.9234	0.221	4.176	0.004	0.401	1.446

Omnibus:	0.039	**Durbin-Watson:**	2.515
Prob(Omnibus):	0.981	**Jarque-Bera (JB):**	0.151
Skew:	0.074	**Prob(JB):**	0.927
Kurtosis:	2.418	**Cond. No.**	435.

Regression Approach to ANOVA

Regression Approach to ANOVA

- Three different assembly methods, referred to as methods A, B, and C, have been proposed.
- Managers at Chemitech want to determine which assembly method can produce the greatest number of filtration systems per week

$H_0: \mu_A = \mu_B = \mu_C$

$H_a: \mu_A \neq \mu_B \neq \mu_C$

A	B	C
58	58	48
64	69	57
55	71	59
66	64	47
67	68	49

ANOVA

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
A	5	310	62	27.5		
B	5	330	66	26.5		
C	5	260	52	31		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	520	2	260	9.176471	0.003818	3.885294
Within Groups	340	12	28.33333			
Total	860	14				

Dummy variables for the chemitech experiment

A	B	
1	0	Observation is associated with assembly method A
0	1	Observation is associated with assembly method B
0	0	Observation is associated with assembly method C

3-1

Dummy variables for the chemitech experiment

$$E(y) = \text{Expected value of the number of units produced per week}$$
$$= \beta_0 + \beta_1 \text{A} + \beta_2 \text{B}$$

- If we are interested in the expected value of the number of units assembled per week for an employee who uses method C, our procedure for assigning numerical values to the dummy variables would result in setting A = B= $E(y) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$
- The multiple regression equation then reduces to

Dummy variables for the chemitech experiment

- For method A the values of the dummy variables are A = 1 and B = 0, and

$$E(y) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$$

- For method B we s $E(y) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$

SUMMARY OUTPUT									
Regression Statistics									
Multiple R	0.777593186								
R Square	0.604651163								
Adjusted R Square	0.53875969								
Standard Error	5.322906474								
Observations	15								
ANOVA									
	df	SS	MS	F	Significance F				
Regression	2	520	260	9.176471	0.003818412				
Residual	12	340	28.33333						
Total	14	860							
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%	
Intercept	52	2.380476143	21.84437	4.97E-11	46.81338804	57.18661196	46.81338804	57.18661196	
A	10	3.366501646	2.970443	0.011692	2.665023022	17.33497698	2.665023022	17.33497698	
B	14	3.366501646	4.15862	0.001326	6.665023022	21.33497698	6.665023022	21.33497698	

Estimation of E(y)

- $b_0 = 52$
- $b_1 = 10$
- $b_2 = 14$

Assembly Method	Estimation of E(y)
A	$b_0+b_1 = 52+10=62$
B	$b_0+b_2 = 52+14 = 66$
C	52

Testing the significance

$$H_0 : \beta_1 = \beta_2 = 0$$

Thank You