Testing the Overall Model

- It is common in regression analysis to compute an F test to determine the overall significance of the model.
- Most computer software packages include the F test and its associated ANOVA table as standard regression output.
- This test determines whether at least one of the regression coefficients (from multiple predictors) is different from zero.
- Simple regression provides only one predictor and only one regression coefficient to test.

Testing the Overall Model

Here are the five steps of the overall F-test for regression:

Step I: State the null and alternative hypotheses:

The null hypothesis states that the model with no independent variables fits the data as well as your model. Whereas the alternative hypothesis says that your model fits the data better than the intercept-only model.

The hypotheses being tested in simple regression by the F test for overall significance are:

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

And for a multiple regression model with intercept, we want to test the following null hypothesis and alternative hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \ldots\ldots\ldots = \beta_{k-1} = 0$$
$$H_1 : \beta_j \neq 0, \text{ For at least one value of j.}$$

Testing the Overall Model

Here are the five steps of the overall F-test for regression:

Step II: Compute the test statistic F:

1. n is the number of observations, k is the number of independent variables.

2. Sum of Squares for Error = $SSE = \sum (y - \hat{y})^2$

3. Corrected Sum of Squares Total = $SST = \sum (y - \bar{y})^2$

4. Corrected Sum of Squares for Model = SSR = $SSM = \sum (\hat{y} - \bar{y})^2$

Testing the Overall Model

Here are the five steps of the overall F-test for regression:

Step II: Compute the test statistic F:

5. SSM+SSE=SST
6. Mean of Squares for Model: MSM = SSM / k
7. Mean of Squares for Error: MSE = SSE / (n-k-1)
8. Mean of Squares Total: MST = SST /n-1
9. $F = \frac{MSM}{MSE} = \frac{SSM/k}{SSE/n-k-1}$

Testing the Overall Model

Here are the five steps of the overall F-test for regression:

Step III: Find a (1 - α) 100% confidence interval for degrees of freedom using an F-table or statistical software.

Step IV: Decide whether to accept or reject the null hypothesis.

If Cal Value(P value)< Tabulated Value: Accept Null Hypothesis otherwise reject Null Hypothesis.

Testing the Overall Model

Degrees of freedom (df):

- Regression df is the number of independent variables in our regression model=p=k-1
- Residual df is the total number of observations (rows) of the dataset subtracted by the number of variables being estimated=n-p=n-k-1
- Total df =n-1

p-value < los (0.05)
Reject Ho

p Value:

- To determine whether any of the differences between the means are statistically significant, compare the p-value to your significance level to assess the null hypothesis.
- If the p-value is less than or equal to the significance level, you reject the null hypothesis.
- If the p-value is greater than the significance level, you do not have enough evidence to reject the null hypothesis

p-value > los (0.05)
accept Ho

Testing the Overall Model

ANOVA for Significance of Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio	P Value
Regression /Model	SSR	(k)	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$	
Error /Residual	SSE	(n-(k+1)) =(n-k-1)	$MSE = \frac{SSE}{(n-(k+1))}$		
Total	SST	(n-1)	$MST = \frac{SST}{(n-1)}$		

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$F = \frac{R^2}{(1-R^2)} \frac{(n-(k+1))}{(k)}$$

$$\bar{R}^2 = 1 - \frac{\frac{SSE}{(n-(k+1))}}{\frac{SST}{(n-1)}} = \frac{MSE}{MST}$$

F-Tab

$\bar{R}^2$

R^2

F-cal

Adjusted R^2

Interpreting Multiple Regression Computer Output

Many of the concepts discussed thus far are highlighted.

Note the following items:

1. The equation of the regression model
2. The ANOVA table with the F value for the overall test of the model
3. The t ratios, which test the significance of the regression coefficients
4. The value of SSE and MSE
5. The value of standard error
6. The value of R^2
7. The value of adjusted R^2

Interpreting Multiple Regression Computer Output

Interpreting Multiple Regression Computer Output

1. The multiple linear regression equation :

The multiple linear regression equation is just an extension of the simple linear regression equation – it has an “x” for each explanatory variable and a coefficient for each “x”.

2. Interpretation of the coefficients:

In the multiple linear regression equation The coefficients in the equation are the numbers in front of the x’s. Each “x” has a coefficient. We should understand what these values mean in the context of the problem.

Interpreting Multiple Regression Computer Output

(4.9 , 5.6)

3.Confidence intervals for the coefficients:

In the multiple linear regression equation A confidence interval is of the form of best estimate ± margin of error In general: Formula for confidence interval for a coefficient (βi):

t* = t cal

Formula for confidence interval for a coefficient (β_i):

$$b_i \pm (t^*_{n-v-1})(SE(b_i))$$

↓ β - coefficient

Note 1:
The degrees of freedom for the t* critical value is the DFE in the Analysis of Variance table.
(Recall, DFE = n – k – 1 where k = the number of explanatory variables)

Note 2:
The subscript "i" in the formula are for the specific explanatory variable.

Interpreting Multiple Regression Computer Output

4. Using the Multiple Linear Regression equation for prediction:

One of the uses of a regression analysis is for prediction. Predicting using a multiple linear regression equation is just an extension of predicting with a simple linear regression equation. We just have to make sure to put the right values in for the right x's.

5. Determining a final model – how to choose "significant" predictors of the response variable:

Another reason for performing a multiple linear regression analysis is to determine which (if any) of the explanatory variables are significant predictors of the response variable. It means to identify explanatory variables which are useful predictors of the response variable (i.e. help to "explain" the response variable That is, does each explanatory variable

Case I – Understanding the Output

A health researcher wants to be able to predict "VO2max", an indicator of fitness and health. Normally, to perform this procedure requires expensive laboratory equipment and necessitates that an individual exercise to their maximum (i.e., until they can longer continue exercising due to physical exhaustion). This can put off those individuals who are not very active/fit and those individuals who might be at higher risk of ill health (e.g., older unfit subjects). For these reasons, it has been desirable to find a way of predicting an individual's VO2max based on attributes that can be measured more easily and cheaply. To this end, a researcher recruited 100 participants to perform a maximum VO2max test, but also recorded their "age", "weight", "heart rate" and "gender". Heart rate is the average of the last 5 minutes of a 20 minute, much easier, lower workload cycling test. The researcher's goal is to be able to predict VO2 max based on these four attributes: age, weight, heart rate and gender.

We have six variables:

1. VO_2max (y) → dependent variable
2. Age (x_1)
3. Weight (x_2)
4. Heart Rate (x_3)
5. Gender (x_4)
6. CaseNo

→ ind. variables

Case I – Understanding the Output

Interpreting and Reporting the Output:

The first table of interest is the **Model Summary** table. This table provides the *R*, R^2, adjusted R^2, and the standard error of the estimate, which can be used to determine how well a regression model fits the data:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.760[a]	.577	.559	5.69097

a. Predictors: (Constant), gender, age, heart_rate, weight

The "**R**" column represents the value of *R*, the ***multiple correlation coefficient***. *R* can be considered to be one measure of the quality of the prediction of the dependent variable; in this case, VO_2max. A value of 0.760, in this example, indicates a good level of prediction. The "**R Square**" column represents the R^2 value (also called the coefficient of determination), which is the proportion of variance in the dependent variable that can be explained by the independent variables (technically, it is the proportion of variation accounted for by the regression model above and beyond the mean model). You can see from our value of 0.577 that our independent variables explain 57.7% of the variability of our dependent variable, VO_2max. However, you also need to be able to interpret "**Adjusted R Square**" (*adj.* R^2) to accurately report your data.

Case I – Understanding the Output

Statistical Significance

The *F*-ratio in the **ANOVA** table (see below) tests whether the overall regression model is a good fit for the data. The table shows that the independent variables statistically significantly predict the dependent variable, $F(4, 95) = 32.393$, $p < .0005$ (i.e., the regression model is a good fit of the data).

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4196.483	4	1049.121	32.393	.000[b]
	Residual	3076.778	95	32.387		
	Total	7273.261	99			

a. Dependent Variable: VO2max

b. Predictors: (Constant), gender, age, heart_rate, weight

SSM

SSE

SST

p-value

p-value < LOS(0.05)

Fcal > Ftab (4, 95)

32.393 > = 2.247

Case I – Understanding the Output

Estimated Model Coefficients

The general form of the equation to predict VO2max from age, weight, heart_rate, gender, is:

$\hat{y}$ β_0 $\beta_1 x_{age}$ $\beta_2 x_{weigh}$ $\beta_3 x_{hr}$ $\beta_4 x_{gender}$

Predicted VO2max = 87.83 – (0.165 x age) – (0.385 x weight) – (0.118 x heart_rate) + (13.208 x gender)

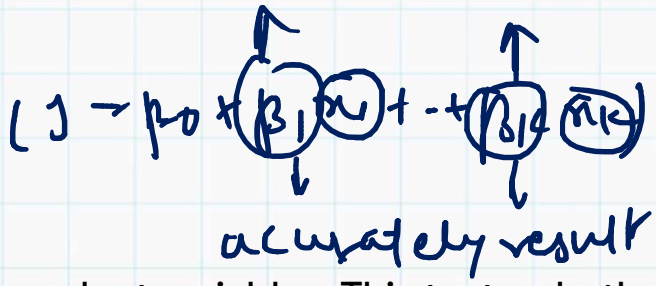
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	**87.830**	6.385		13.756	.000	75.155	100.506
	age	**-.165**	.063	-.176	-2.633	.010	-.290	-.041
	weight	**-.385**	.043	-.677	-8.877	.000	-.471	-.299
	heart_rate	**-.118**	.032	-.252	-3.667	.000	-.182	-.054
	gender	**13.208**	1.344	.748	9.824	.000	10.539	15.877

a. Dependent Variable: VO2max

Unstandardized coefficients indicate how much the dependent variable varies with an independent variable when all other independent variables are held constant. Consider the effect of age in this example. The unstandardized coefficient, B_1, for age is equal to -0.165 (see **Coefficients** table). This means that for each one year increase in age, there is a decrease in VO_2max of 0.165 ml/min/kg.

Case I – Understanding the Output



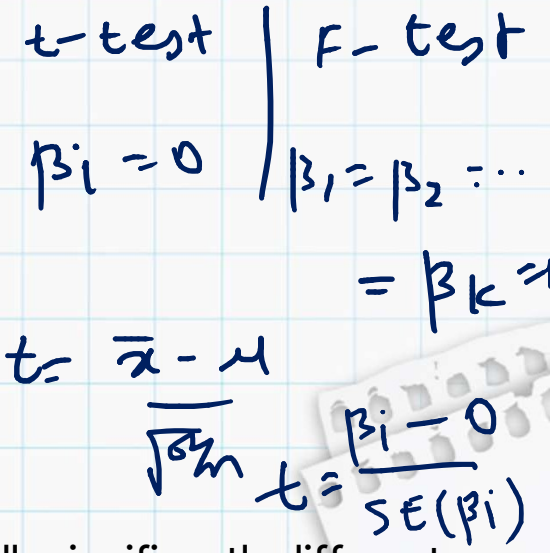
Statistical significance of the independent variables.

You can test for the statistical significance of each of the independent variables. This tests whether the unstandardized (or standardized) coefficients are equal to 0 (zero) in the population. If p < .05, you can conclude that the coefficients are statistically significantly different to 0 (zero). The t-value and corresponding p-value are located in the "t" and "Sig." columns, respectively, as highlighted below:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	87.830	6.385		**13.756**	**.000**	75.155	100.506
	age	-.165	.063	-.176	**-2.633**	**.010**	-.290	-.041
	weight	-.385	.043	-.677	**-8.877**	**.000**	-.471	-.299
	heart_rate	-.118	.032	-.252	**-3.667**	**.000**	-.182	-.054
	gender	13.208	1.344	.748	**9.824**	**.000**	10.539	15.877

a. Dependent Variable: VO2max



You can see from the "**Sig.**" column that all independent variable coefficients are statistically significantly different from 0 (zero). Although the intercept, B_0, is tested for statistical significance, this is rarely an important or interesting finding.

Case I – Understanding the Output

Overall Conclusion:

A multiple regression was run to predict VO_2max from gender, age, weight and heart rate. These variables statistically significantly predicted VO_2max, $F(4, 95) = 32.393$, $p < .0005$, $R^2 = .577$. All four variables added statistically significantly to the prediction, $p < .05$.

Question: Predict the VO2 max for 40 years age, 55kg weight, heart rate 70 beats per minute of a female.

Predicted VO2max = 87.83 – (0.165 x age) – (0.385 x weight) – (0.118 x heart_rate) + (13.208 x gender)

$VO_2 max = 51.79$

A	B	C
1	0	
0	1	

Case I – Understanding the Output

Case II – Literacy Rate Example

Literacy rate is a reflection of the educational facilities and quality of education available in a country, and mass communication plays a large part in the educational process. In an effort to relate the literacy rate of a country to various mass communication outlets, a demographer has proposed to relate literacy rate to the following variables: number of daily newspaper copies (per 1000 population), number of radios (per 1000 population), and number of TV sets (per 1000 population). Here are the data for a sample of 10 countries:

Case II – Literacy Rate Example

Country	newspapers	radios	tv sets	literacy rate
Czech Republic / Slovakia	280	266	228	0.98
Italy	142	230	201	0.93
Kenya	10	114	2	0.25
Norway	391	313	227	0.99
Panama	86	329	82	0.79
Philippines	17	42	11	0.72
Tunisia	21	49	16	0.32
USA	314	1695	472	0.99
Russia	333	430	185	0.99
Venezuela	91	182	89	0.82

Case II – Literacy Rate Example

Below is the Minitab output from a Multiple Linear Regression analysis.

$x_1 = 200$, $x_2 = 800$, $x_3 = 250$

β

Predictor	Coef	SE Coef	T	P
Constant	0.51486	0.09368	5.50	0.002
newspaper copies	0.0005421	0.0008653	0.63	0.554
radios	-0.0003535	0.0003285	-1.08	0.323
television sets	0.001988	0.001550	1.28	0.247

S = 0.186455 R-Sq = 69.9% R-Sq(adj) = 54.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	0.48397	0.16132	4.64	0.053
Residual Error	6	0.20859	0.03477		
Total	9	0.69256			

Case II – Literacy Rate Example

Question 1: What is the response variable? What are the explanatory variables?

Response variable: Y = literacy rate.

Explanatory variables:

1. X1 = number of daily newspaper copies,
2. X2 = number of radios, and
3. X3 = number of TV sets
   (all per 1000 people in the population of the country).

Case II – Literacy Rate Example

Question 2: Write the least-squares regression equation for this problem. Explain what each term in the regression equation represents in terms of the problem

$$\hat{y} = 0.51486 + 0.00054x1 - 0.00035x2 + 0.00199x3$$

where

$\hat{y}$= predicted literacy rate

x1 = the number of daily newspaper copies in the country (per 1000 people)

x2 = the number of radios in the country (per 1000 people)

x3 = the number of TV sets in the country (per 1000 people)

Case II – Literacy Rate Example

Interpretation of the coefficients in the multiple linear regression equation

- Let's start with the interpretation of the coefficient for newspaper copies (x1). Like the slope in simple linear regression, it tells us that we predict the literacy rate to increase by 0.00054 for every additional daily newspaper copy in that country (per 1000 people in the population).

- But, there is more. To properly interpret the coefficient of daily newspaper copies, the other two variables can't be changing – only the number of daily newspaper copies increases by 1. So, a way to interpret the coefficient of number of daily newspaper copies is as follows:

- For every additional daily newspaper copy per 1000 people in a population, literacy rate is predicted to increase by 0.00054, keeping the number of radios and TV sets the same.

Case II – Literacy Rate Example

Interpretation of the coefficients in the multiple linear regression equation

- Since the coefficient is negative, we'd expect the literacy rate to be lower for every additional radio per 1000 people in the population (for countries with the same number of daily newspaper copies and TV sets per 1000 people in the population).

Case II – Literacy Rate Example

Question 3: What are the degrees of freedom for the t* value in this problem?

- Recall, the degrees of freedom for any hypothesis test or confidence interval that involves a t-statistic is DFE = n – v – 1,
- where v = the number of explanatory variables in the model.
- In our problem, n = 10 and v = 3.
- Therefore, the degrees of freedom for the t* critical value is 10 – 3- 1 = 6.

Case II – Literacy Rate Example

Question 4: Determine the lower and upper bounds for the 95% confidence interval for .β3

Formula for confidence interval for a coefficient (β_i):

$$b_i \pm (t^*_{n-v-1})(SE(b_i))$$

β-coeff p-value S.E

b3 = 0.00199, SE(b3) = 0.00155, and t = 2.447.

Therefore, the lower bound = (0.00199) – (2.447)(0.00155) = -0.00180.

The upper bound = (0.00199) + (2.447)(0.00155) = 0.00578.

0.00199

We write the 95% confidence interval for B3 as (-0.00180, 0.00578).

Case II – Literacy Rate Example

Question 5: Determine the lower and upper bounds for the 95% confidence interval for $\beta 1$ and $\beta 2$.

Case II – Literacy Rate Example

Question 6: Predict literacy rate for a country that has 200 daily newspaper copies (per 1000 in the population), 800 radios (per 1000 in the population), and 250 TV sets (per 1000 in the population).

$\hat{y} = 0.8436$

Case II – Literacy Rate Example

Question 7: Verify that the F-statistic in the output above equals MSM / MSE.

MSM = 0.16132

and

MSE = 0.03477.

0.16132/0.03477 = 4.6396 or 4.64 rounded to two decimal places.

Case II – Literacy Rate Example

Question 8: Conclusion on R square and adjusted R square value

→ ∴ R^2 = 69.9 %. We can say that our independent variables x_1, x_2, x_3 explains total 69.9 % of variation of our dependent variable ie. Literacy rate.

Case II – Literacy Rate Example

Question 9: What are the degrees of freedom for this F-statistic?

numerator df = DFM = # explanatory variables = 3.

denominator df = n – v – 1 = 10 – 3 – 1 = 6.

Case II – Literacy Rate Example

Question 10: State a conclusion in the context of the problem.

There is suggestive, but weak, evidence to indicate that at least one of number of daily newspaper copies, number of radios, and/or number of TV sets help to explain a country's literacy rate (p-value = 0.053).

Some notes:

1) Even though the evidence is weak, we should continue the analysis to find out for sure if there is at least one explanatory variable that is a significant predictor of literacy rate and, if so, which one or ones. Anytime the p-value is less than 0.1 for the F-test, we should continue the analysis.

2) Remember, the conclusion states that there is suggestive evidence that at least one explanatory variable is a significant predictor of literacy rate. It does NOT tell us how many or which one or ones are significant predictors of literacy rate – only that there is at least one that is.

3) If the F-test indicates no evidence to reject the null hypothesis, then there is no need to continue the analysis as there is no evidence to indicate that any of the explanatory variables are helpful in explaining the response variable. However, if there is even the slightest bit of evidence to reject the null hypothesis from the F-test (i.e., p-value < 0.10), we should continue the analysis. This will involve doing t-tests on each explanatory variable, as we will see below.

Case II – Literacy Rate Example

Question 11: Calculate the t-statistic (with degrees of freedom) for newspaper copies.

t =0.0005421 / 0.0008653 = 0.6265.

the degrees of freedom for the t-test is DFE = n – v – 1 = 6.

Calculate the t-statistic (with degrees of freedom) for radio.

Calculate the t-statistic (with degrees of freedom) for Tv sets.

Case II – Literacy Rate Example

Question 12: What are the degrees of freedom for the t-tests in the final model?

Recall, the degrees of freedom for the t-test is DFE = n – v – 1. There are only 2 explanatory variables left in the model, so the degrees of freedom for the t-tests = 10 – 2 – 1 = 7.

Case II – Literacy Rate Example

Question 13: Which of the following is true?

A] TV sets would remain in the model because we always need to have at least one explanatory variable in the model.

B] TV sets would remain in the model since its p-value is less than 0.05.

A is not correct because it is possible that a backwards selection process will eliminate all variables. But, remember that we'll stop eliminating variables once all remaining variables have p-values less than 0.05, which is the case here. Therefore, C is also incorrect.