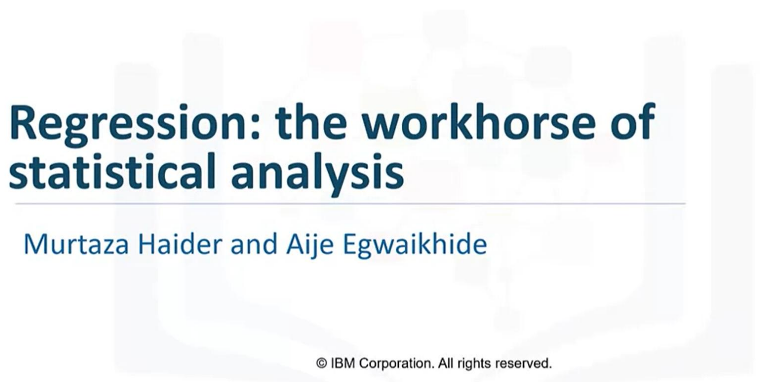
Week 5 - Regression Analysis

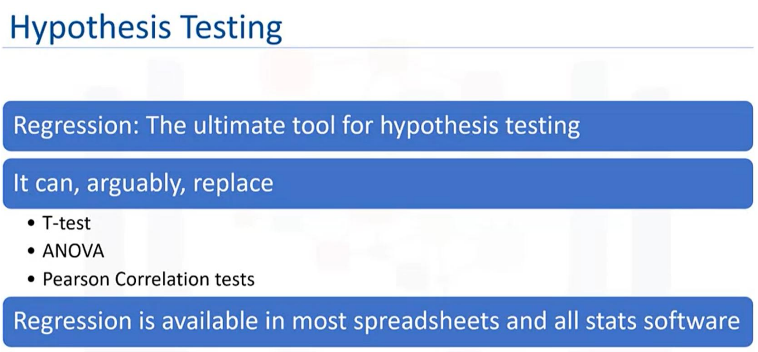
1. Regression - the workhorse of statistical analysis



Regression: the workhorse of statistical analysis

Murtaza Haider and Aije Egwaikhide

© IBM Corporation. All rights reserved.



Hypothesis Testing

Regression: The ultimate tool for hypothesis testing

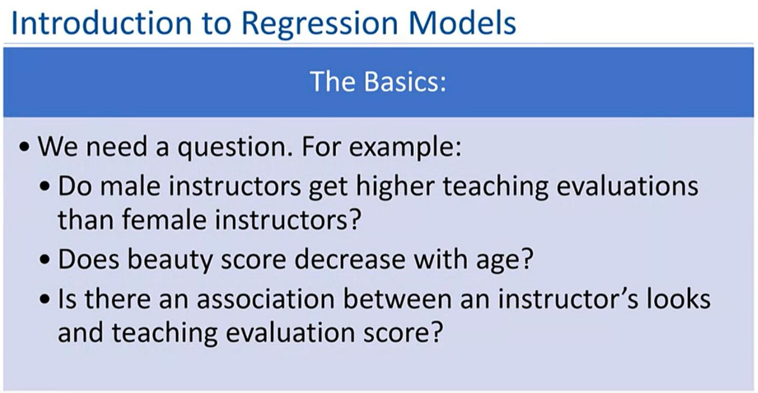
It can, arguably, replace

- T-test
- ANOVA
- Pearson Correlation tests

Regression is available in most spreadsheets and all stats software

In this video, we will introduce the Fundamentals of regression analysis, which we believe is the workhorse of statistical analysis.

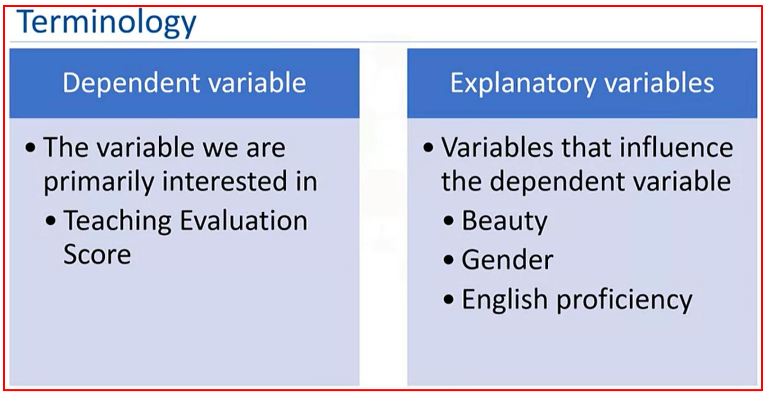
Now, in terms of hypothesis testing, these tests measure the strength of relationship between two or more variables. And you have to run them independently. But if you know how to run regression, we say, as a practical data scientist, you can forego these tests and go straight to regression, which is available in most spreadsheets and also in all statistical software.



Introduction to Regression Models

The Basics:

- We need a question. For example:
  - Do male instructors get higher teaching evaluations than female instructors?
  - Does beauty score decrease with age?
  - Is there an association between an instructor's looks and teaching evaluation score?



Terminology

Dependent variable

- The variable we are primarily interested in
  - Teaching Evaluation Score

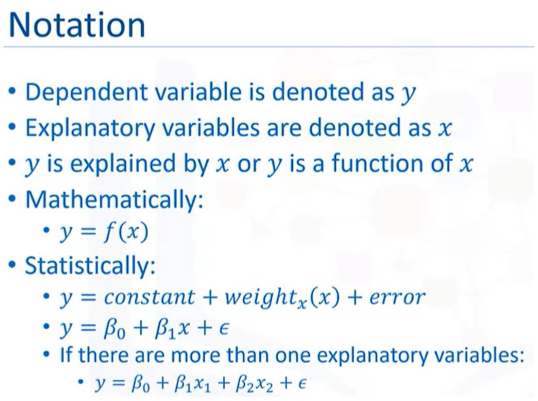
Explanatory variables

- Variables that influence the dependent variable
  - Beauty
  - Gender
  - English proficiency

So here the fundamental basics of regression model. First of all, you need a question to answer using regression model. For instance, do male instructors get higher teaching evaluations than female instructors? Or does the beauty score decrease with the aid of the individual instructor? Or is there an association between an instructor's looks and the teaching evaluation score that we see? Do good-looking professors get hired teaching evaluation scores?

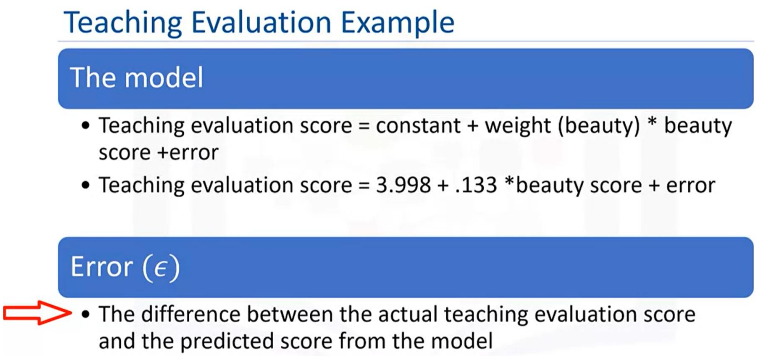
So with these questions in mind, we focus now on the terminology of regression model. So there are two types of regression variables that we use.

- One is a **dependent variable**, that is the variable that we are really interested in. For example, the teaching evaluation score of an individual instructor.
- And the **explanatory variable**s that explain the variance or differences of values of the dependent variable. So for example, teaching evaluation score could be explained by the looks, or the gender, or the English language proficiency of an individual instructor. So you have two types of variables, dependent and explanatory.



Notation

- Dependent variable is denoted as y
- Explanatory variables are denoted as x
- y is explained by x or y is a function of x
- Mathematically:
  - $y = f(x)$
- Statistically:
  - $y = constant + weight_x(x) + error$
  - $y = \beta_0 + \beta_1 x + \epsilon$
  - If there are more than one explanatory variables:
    - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$



Teaching Evaluation Example

The model

- Teaching evaluation score = constant + weight (beauty) * beauty score +error
- Teaching evaluation score = 3.998 + .133 *beauty score + error

Error (ϵ)

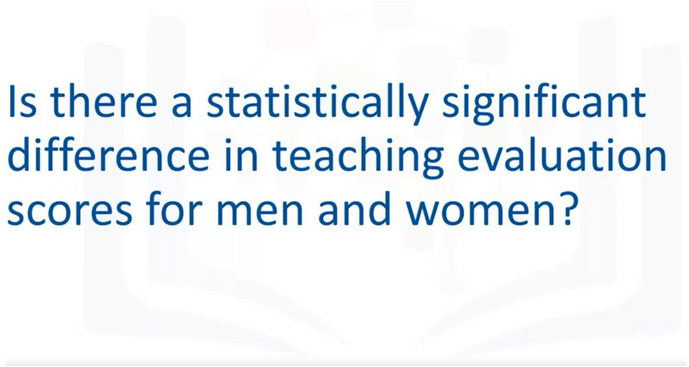
- The difference between the actual teaching evaluation score and the predicted score from the model

Statistically, if you run them, an estimate, a regression model, Y is equal to some constant and then a weighting factor for the variable X. If it's a beauty score, then the weighting factor for the beauty score and the error term. An error term is whatever we cannot explain by the model, that goes into error term. And I will explain this a little more in a minute. And the error term which we represent as epsilon.

总结：

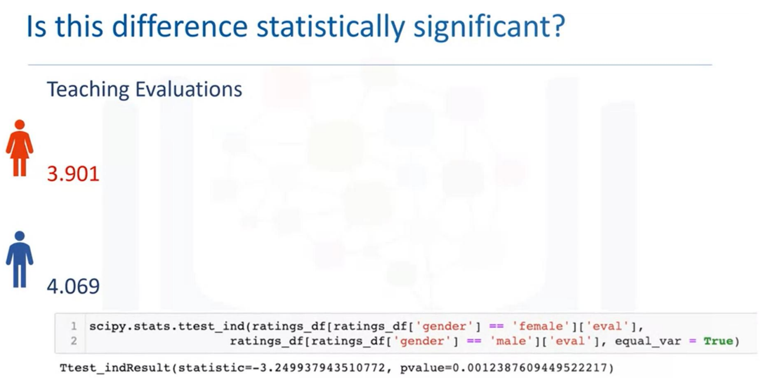
2. Regression in place of t - test

Regression in place of a T-Test



Is there a statistically significant difference in teaching evaluation scores for men and women?

In this video we will illustrate how to use regression analysis in place of at t-test. We will begin with a question, and the question is, is there a statistically significant difference in teaching evaluation scores for men and women?

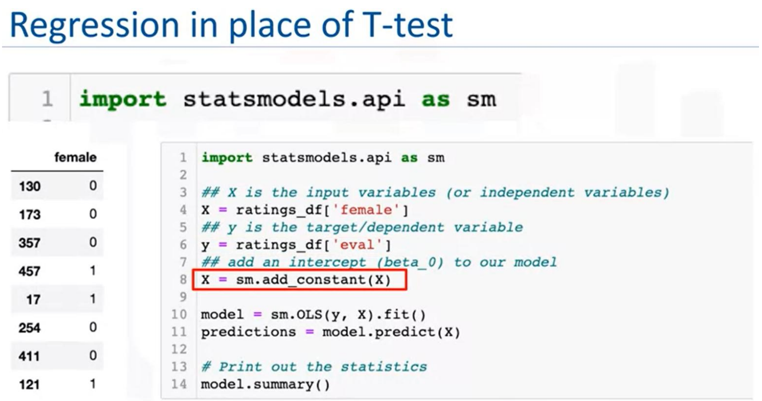


Is this difference statistically significant?

Teaching Evaluations

```
scipy.stats.ttest_ind(ratings_df[ratings_df['gender'] == 'female']['eval'],
                      ratings_df[ratings_df['gender'] == 'male']['eval'], equal_var = True)
```

```
Ttest_indResult(statistic=-3.249937943510772, pvalue=0.0012387609449522217)
```



Regression in place of T-test

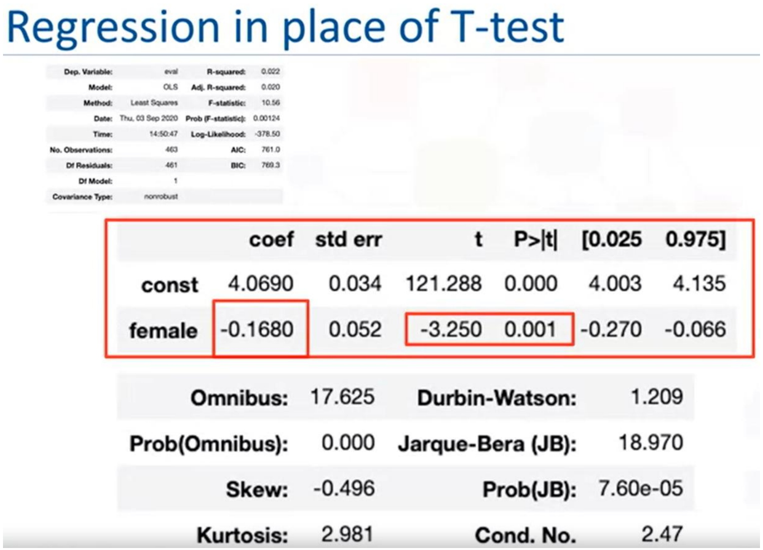
```
import statsmodels.api as sm
```

	female
130	0
173	0
357	0
457	1
17	1
254	0
411	0
121	1

```
import statsmodels.api as sm

## X is the input variables (or independent variables)
X = ratings_df['female']
## y is the target/dependent variable
y = ratings_df['eval']
## add an intercept (beta_0) to our model
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```

When we compute, the averages while using the teaching evaluation data set, we find that the teaching evaluation score for women is around 3.9, and for men, it's around 4.06. The question is, is this difference, even though it's small, statistically significant? We can run at t-test using Python and compute the statistical significance for the t-test. Here, our conclusion is that the teaching evaluation scores difference between men and women, is statistically significant. What if we were to do the same thing with the regression model?

We will do the linear regression in Python. We will be using the statsmodel library. We will create a list for the independent variable, that is the female variable, which has been turned to a binary variable, where 1 equals female and 0 is male. We will also create a list for the dependent variable, teaching evaluation score. We will manually add the constant beta zero, then we will fit and make predictions, and print out the model summary.



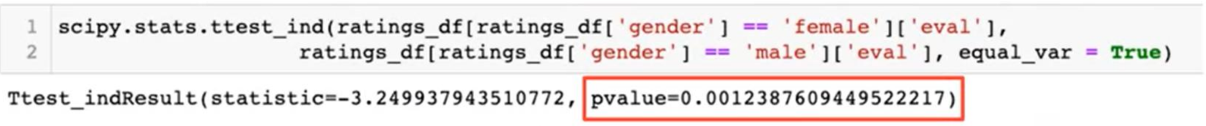
Regression in place of T-test

Dep. Variable:	eval	R-squared:	0.022
Model:	OLS	Adj. R-squared:	0.020
Method:	Least Squares	F-statistic:	10.56
Date:	Thu, 03 Sep 2020	Prob (F-statistic):	0.00124
Time:	14:50:47	Log-Likelihood:	-378.50
No. Observations:	463	AIC:	761.0
Df Residuals:	461	BIC:	769.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.0690	0.034	121.288	0.000	4.003	4.135
female	-0.1680	0.052	-3.250	0.001	-0.270	-0.066

Omnibus:	17.625	Durbin-Watson:	1.209
Prob(Omnibus):	0.000	Jarque-Bera (JB):	18.970
Skew:	-0.496	Prob(JB):	7.60e-05
Kurtosis:	2.981	Cond. No.	2.47

The model summary will print out a table like this. But we are only interested in this part of this table for the t-test, it prints out the coefficient error, t statistics and pin value. We can see the t statistics for the female variable is negative 3.25, and the P value is less than 0.05. That means that there is a statistical difference in mean values for male and female instructors. The coefficient means that you are most likely to lose about 0.17 marks for being a female. We can see that the results from using a regression model, and the conclusion is identical, if we run a t-test.



```
scipy.stats.ttest_ind(ratings_df[ratings_df['gender'] == 'female']['eval'],
                      ratings_df[ratings_df['gender'] == 'male']['eval'], equal_var = True)
```

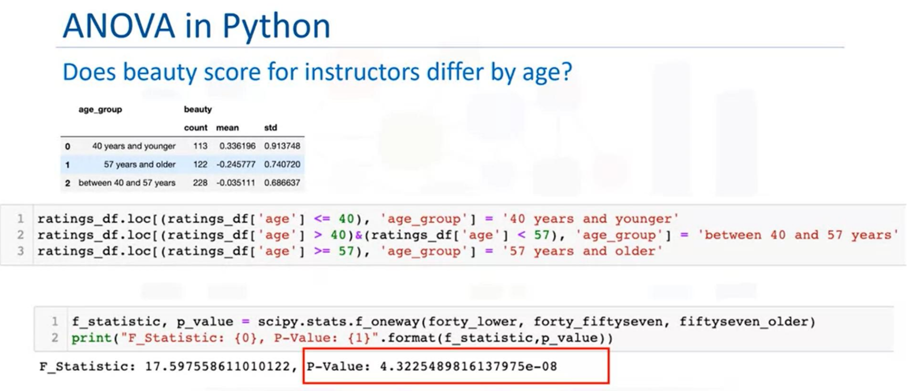
```
Ttest_indResult(statistic=-3.249937943510772, pvalue=0.0012387609449522217)
```

总结：

3. Regression in place of ANOVA

Regression in place of ANOVA

© IBM Corporation. All rights reserved.



ANOVA in Python

Does beauty score for instructors differ by age?

	age_group	beauty		
		count	mean	std
0	40 years and younger	113	0.336196	0.913748
1	57 years and older	122	-0.245777	0.740720
2	between 40 and 57 years	228	-0.035111	0.686637

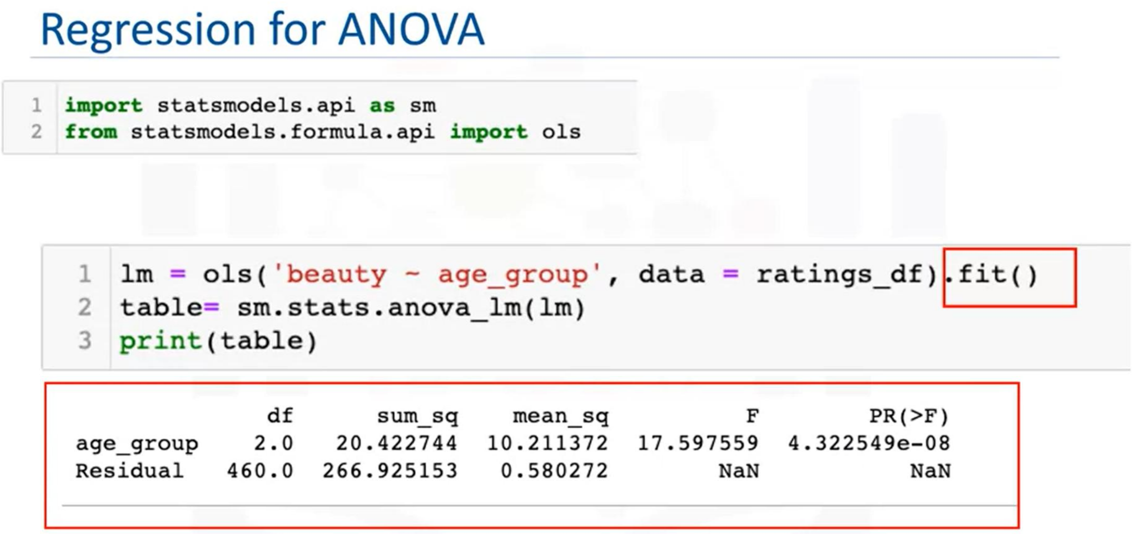
```
ratings_df.loc[(ratings_df['age'] <= 40), 'age_group'] = '40 years and younger'
ratings_df.loc[(ratings_df['age'] > 40)&(ratings_df['age'] < 57), 'age_group'] = 'between 40 and 57 years'
ratings_df.loc[(ratings_df['age'] >= 57), 'age_group'] = '57 years and older'
```

```
f_statistic, p_value = scipy.stats.f_oneway(forty_lower, forty_fiftyseven, fiftyseven_older)
print("F_Statistic: {0}, P-Value: {1}".format(f_statistic,p_value))
```

```
F_Statistic: 17.597558611010122, P-Value: 4.3225489816137975e-08
```

When we are comparing the difference in means or when we are comparing the averages between groups that are more than two, we will use ANOVA or analysis of variance. We know that if there are only two groups, we can use the t-test, but when we are comparing averages for more than two groups, we use analysis of variance.

Working with our teaching evaluation dataset, we took the teaching evaluation scores and then we wanted to see what would happen if we took the instructors and divide them into three groups, 40 years and younger, those between 40 and 57 years of age and those that are 57 years or older. We computed the average value for teaching evaluation score for the three groups. We wanted to determine if the three mean values were statistically different. To recap, we ran the analysis of variance test, which uses F-distribution. The p-value is less than 0.05. We reject the null hypothesis that averages of the group are equal and concluded that the differences are statistically significan.



Regression for ANOVA

```
import statsmodels.api as sm
from statsmodels.formula.api import ols
```

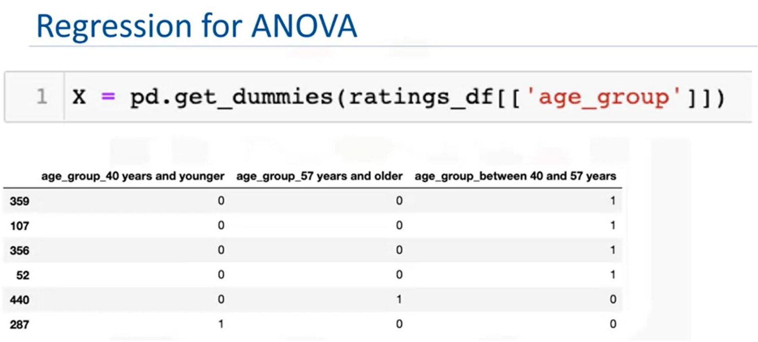
```
lm = ols('beauty ~ age_group', data = ratings_df).fit()
table= sm.stats.anova_lm(lm)
print(table)
```

```
              df      sum_sq    mean_sq          F        PR(>F)
age_group    2.0   20.422744  10.211372  17.597559  4.322549e-08
Residual   460.0  266.925153   0.580272        NaN           NaN
```

Now, let us do this with the regression model. We will use the statsmodel library and also import the **ols** function. We will create or initiate a linear model of the beauty score, which is our y-variable. Please note that when dealing with a linear regression model, the y-variable has to be a continuous variable. Otherwise results will not be accurate. Now, create the linear model and fit it using the fit function. Use the **anova_lm** function to create a table that prints out the results of the test statistics. The results will look like this.

It will print out the degree of freedom, the sum of square, F statistic and the p-value. Like ANOVA from this api package, we get the same results, which is that we will reject the null hypothesis, the averages of the group are equal and conclude that the differences are statistically significant.

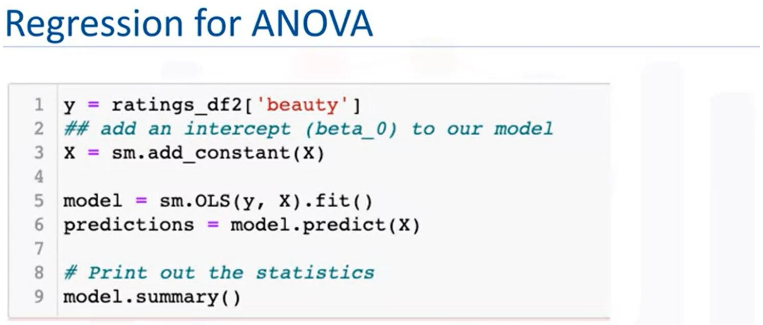
总结：



Regression for ANOVA

```python
X = pd.get_dummies(ratings_df[['age_group']])
```

	age_group_40 years and younger	age_group_57 years and older	age_group_between 40 and 57 years
359	0	0	1
107	0	0	1
356	0	0	1
52	0	0	1
440	0	1	0
287	1	0	0

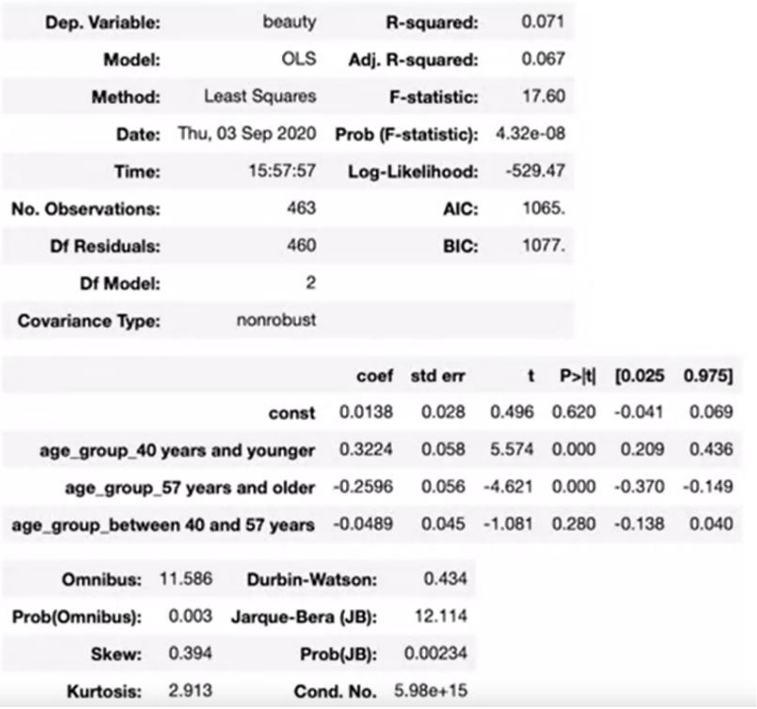


Regression for ANOVA

```python
y = ratings_df2['beauty']
## add an intercept (beta_0) to our model
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```

You can also turn the age group values into dummy values and run it like you run the regression for t-test. To do that, you will need to create dummy variables for the age groups using the get_dummies function in pandas. It will look like this, where one means they belong to that group and zero means otherwise. Just like a binary variable, values can only belong to one group.

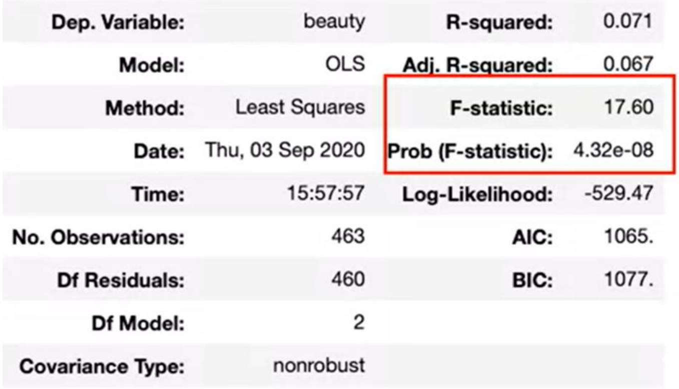
Run the same as you did for the t-test by fitting the variables into an OLS function, predict, and print out the model summary. We will get results like this (下图). Taking a closer look, we can see the same results for the F statistic and the p-value.



Dep. Variable:	beauty	R-squared:	0.071
Model:	OLS	Adj. R-squared:	0.067
Method:	Least Squares	F-statistic:	17.60
Date:	Thu, 03 Sep 2020	Prob (F-statistic):	4.32e-08
Time:	15:57:57	Log-Likelihood:	-529.47
No. Observations:	463	AIC:	1065.
Df Residuals:	460	BIC:	1077.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0138	0.028	0.496	0.620	-0.041	0.069
age_group_40 years and younger	0.3224	0.058	5.574	0.000	0.209	0.436
age_group_57 years and older	-0.2596	0.056	-4.621	0.000	-0.370	-0.149
age_group_between 40 and 57 years	-0.0489	0.045	-1.081	0.280	-0.138	0.040

Omnibus:	11.586	Durbin-Watson:	0.434
Prob(Omnibus):	0.003	Jarque-Bera (JB):	12.114
Skew:	0.394	Prob(JB):	0.00234
Kurtosis:	2.913	Cond. No.	5.98e+15

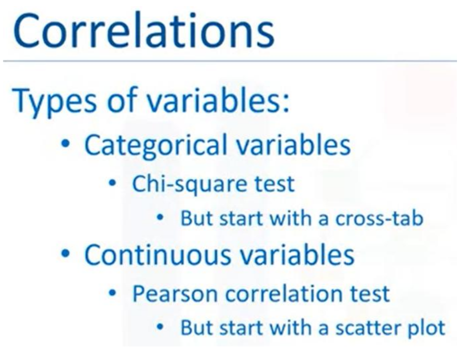


Dep. Variable:	beauty	R-squared:	0.071
Model:	OLS	Adj. R-squared:	0.067
Method:	Least Squares	F-statistic:	17.60
Date:	Thu, 03 Sep 2020	Prob (F-statistic):	4.32e-08
Time:	15:57:57	Log-Likelihood:	-529.47
No. Observations:	463	AIC:	1065.
Df Residuals:	460	BIC:	1077.
Df Model:	2		
Covariance Type:	nonrobust		

总结：

4. Regression in place of Correlation

Regression in place of Correlation

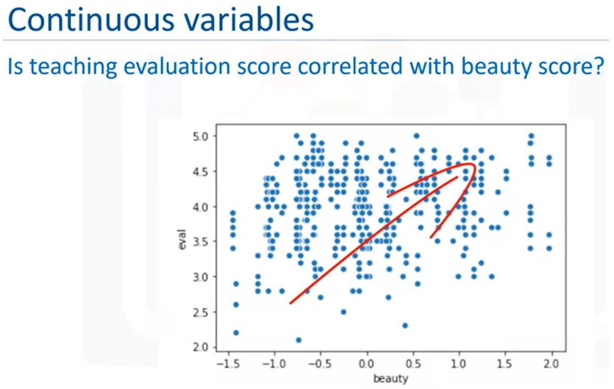


Correlations

Types of variables:

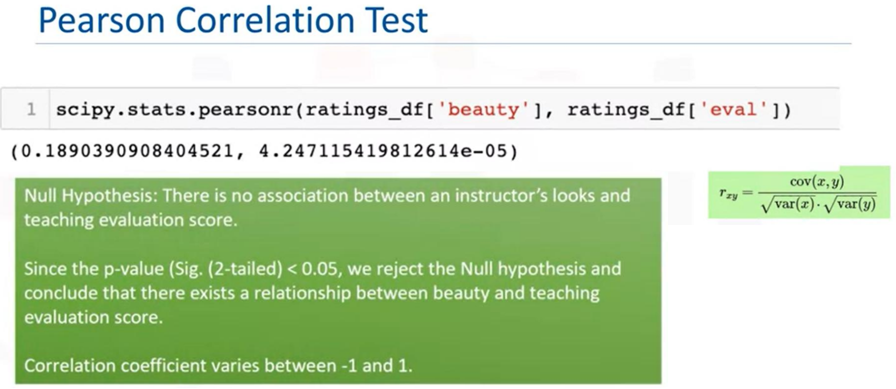
- Categorical variables
  - Chi-square test
    - But start with a cross-tab
- Continuous variables
  - Pearson correlation test
    - But start with a scatter plot

In this video we will illustrate how one can use regression models in place of tests conducted for correlation analysis. We will return to the basics. There are two types, or mostly two types of variables. First, are the categorical variables for which we use chi-square tests to determine if there is an association between the two. And second are the categorical variables, or we could have continuous variables where we use the Pearson correlation test. We will focus on just the continuous variables.



Continuous variables

Is teaching evaluation score correlated with beauty score?



Pearson Correlation Test

```
scipy.stats.pearsonr(ratings_df['beauty'], ratings_df['eval'])
```

```
(0.1890390908404521, 4.247115419812614e-05)
```

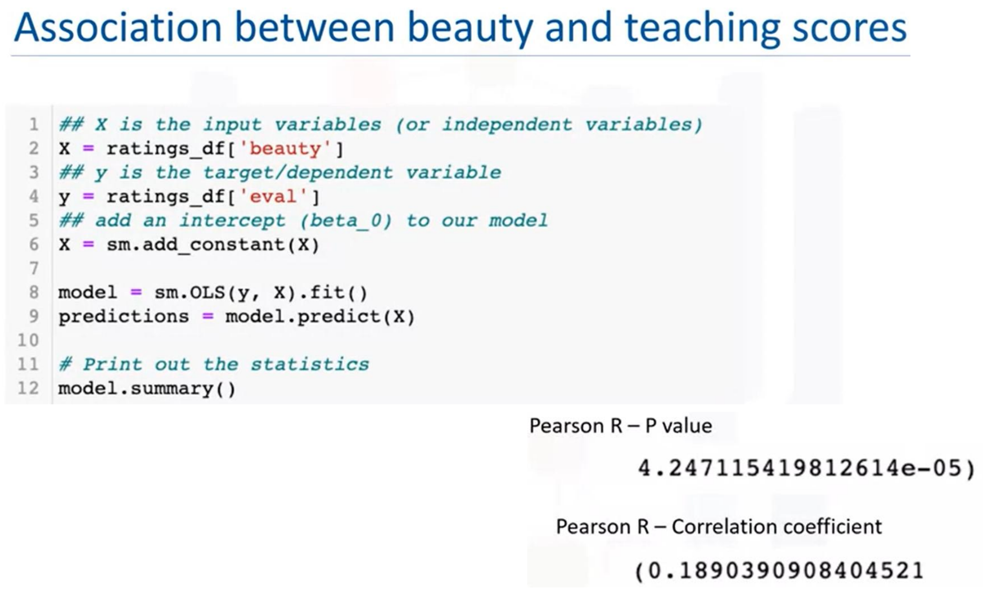
Null Hypothesis: There is no association between an instructor's looks and teaching evaluation score.

Since the p-value (Sig. (2-tailed) < 0.05, we reject the Null hypothesis and conclude that there exists a relationship between beauty and teaching evaluation score.

Correlation coefficient varies between -1 and 1.

$$r_{xy} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)} \cdot \sqrt{\text{var}(y)}}$$

We can plot two continuous variables in a scatter plot. The teaching evaluation scores are on the Y axis and the normalized beauty scores are on the X axis. You could sort of see a relationship between the two variables. It's an upward sloping type of a relationship. We see that as the beauty score increases, so does the teaching evaluation score. Remember, we used the Pearson correlation test to determine the relationship and its significance level.



Association between beauty and teaching scores

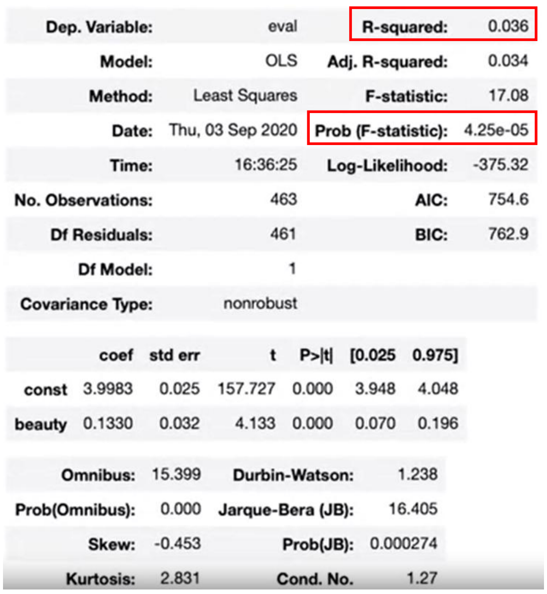
```
## X is the input variables (or independent variables)
X = ratings_df['beauty']
## y is the target/dependent variable
y = ratings_df['eval']
## add an intercept (beta_0) to our model
X = sm.add_constant(X)

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

# Print out the statistics
model.summary()
```

Pearson R – P value

```
4.247115419812614e-05)
```

Pearson R – Correlation coefficient

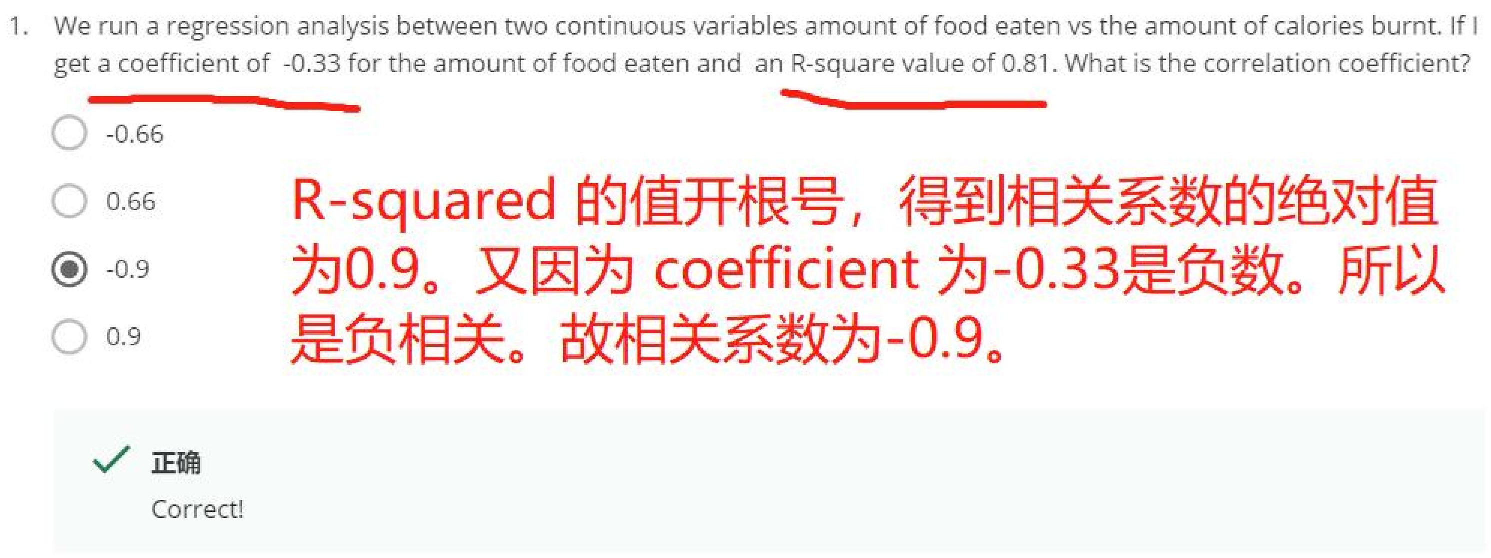
```
(0.1890390908404521
```



```
   Dep. Variable:               eval        R-squared:     0.036
           Model:                OLS    Adj. R-squared:    0.034
          Method:      Least Squares       F-statistic:    17.08
            Date:   Thu, 03 Sep 2020 Prob (F-statistic):   4.25e-05
            Time:           16:36:25    Log-Likelihood:   -375.32
No. Observations:                463               AIC:    754.6
    Df Residuals:                461               BIC:    762.9
        Df Model:                  1
 Covariance Type:          nonrobust

           coef  std err        t  P>|t|  [0.025  0.975]
  const  3.9983    0.025  157.727  0.000   3.948   4.048
 beauty  0.1330    0.032    4.133  0.000   0.070   0.196

      Omnibus:  15.399   Durbin-Watson:     1.238
Prob(Omnibus):   0.000  Jarque-Bera (JB):  16.405
         Skew:  -0.453        Prob(JB):  0.000274
     Kurtosis:   2.831       Cond. No.      1.27
```

Now let's do the same in regression. Just like we did with the T test and the F test, we will fit a linear model for both the beauty, an evaluation score, values and print out the models summary. Taking a closer look, it prints out a P value of 4.25 * 10 raised to power negative 5, which is less than 0.05.. That is very similar to when we run the Pearson r function. It will also give us the R-squared value, that is, if we **took square root** 0.036. It will give us 0.189 which is the some value as the correlation coefficient from computing the Pearson R.

总结：



1. We run a regression analysis between two continuous variables amount of food eaten vs the amount of calories burnt. If I get a coefficient of -0.33 for the amount of food eaten and an R-square value of 0.81. What is the correlation coefficient?

○ -0.66

○ 0.66

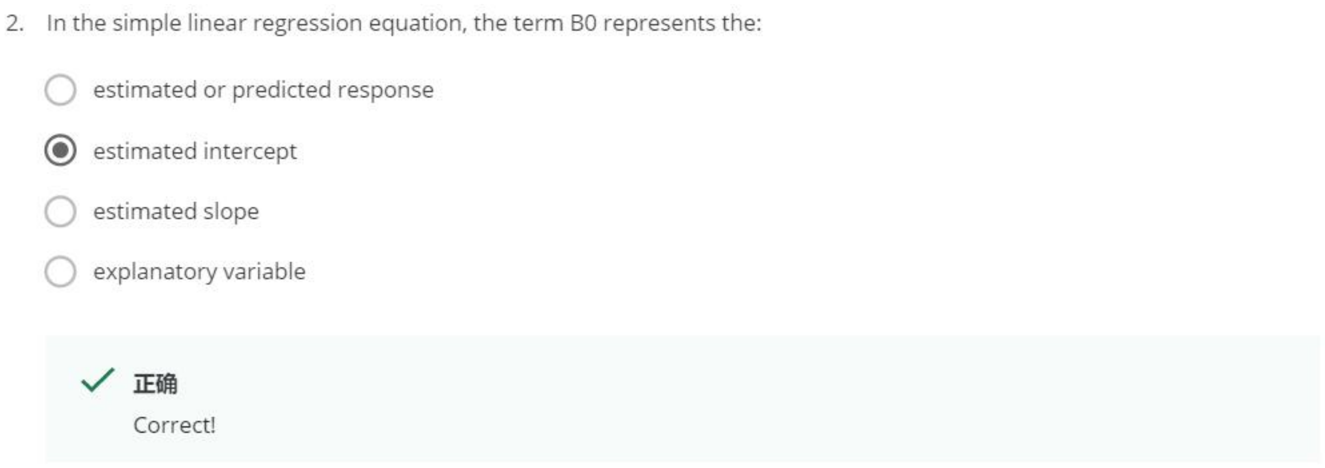
● -0.9

○ 0.9

R-squared 的值开根号，得到相关系数的绝对值为0.9。又因为 coefficient 为-0.33是负数。所以是负相关。故相关系数为-0.9。

✓ **正确**

Correct!



2. In the simple linear regression equation, the term B0 represents the:

○ estimated or predicted response

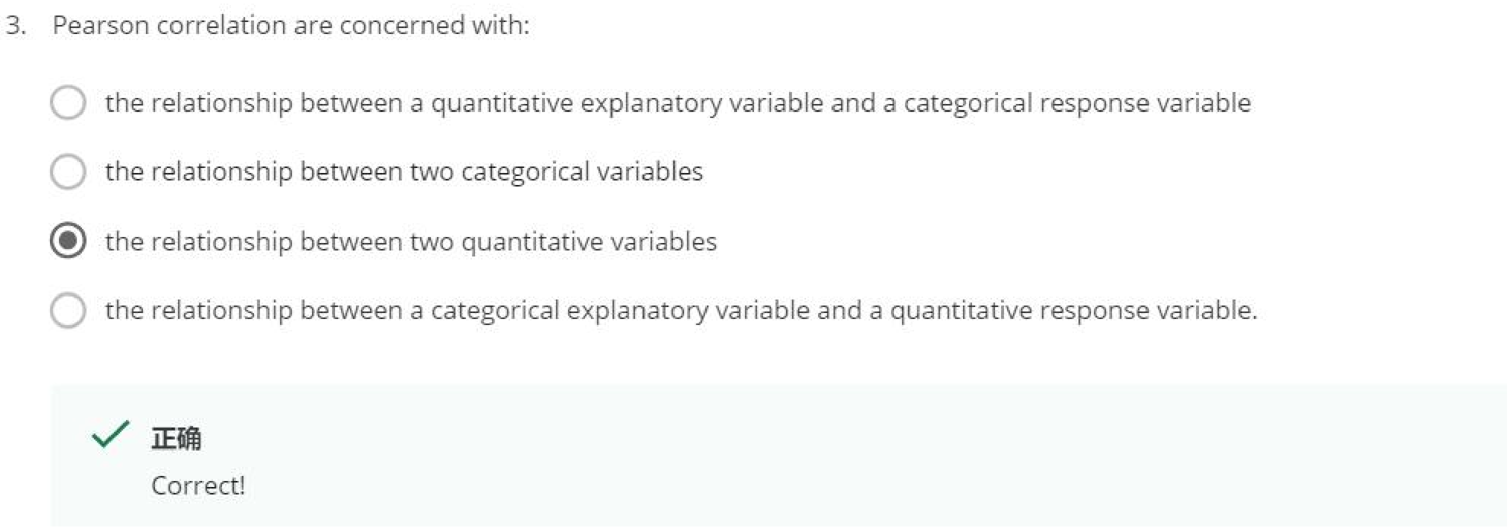
● estimated intercept

○ estimated slope

○ explanatory variable

✓ **正确**

Correct!



3. Pearson correlation are concerned with:

○ the relationship between a quantitative explanatory variable and a categorical response variable

○ the relationship between two categorical variables

● the relationship between two quantitative variables

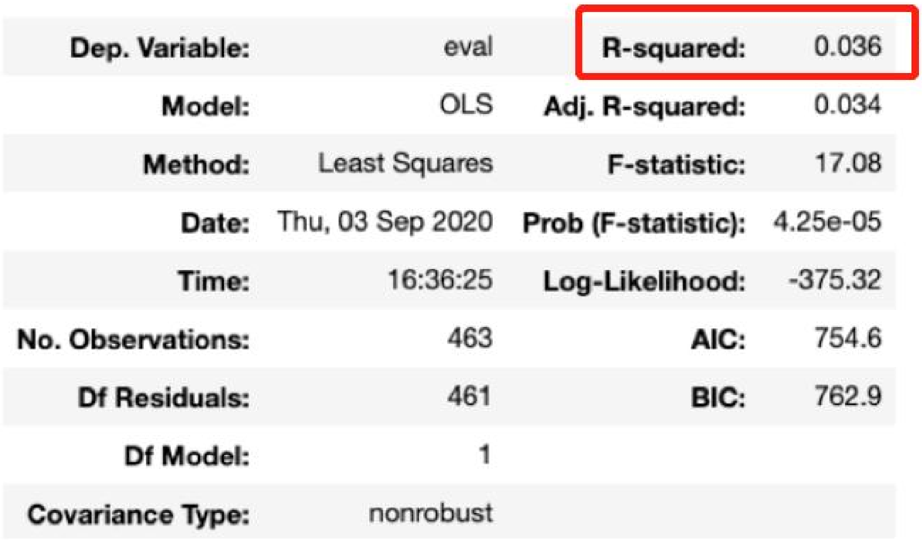
○ the relationship between a categorical explanatory variable and a quantitative response variable.

✓ **正确**

Correct!

总结：

2. Give the results of the regression analysis below, what is the correlation coefficient?



Dep. Variable:	eval	R-squared:	0.036
Model:	OLS	Adj. R-squared:	0.034
Method:	Least Squares	F-statistic:	17.08
Date:	Thu, 03 Sep 2020	Prob (F-statistic):	4.25e-05
Time:	16:36:25	Log-Likelihood:	-375.32
No. Observations:	463	AIC:	754.6
Df Residuals:	461	BIC:	762.9
Df Model:	1		
Covariance Type:	nonrobust		

开根号即是 correlation coefficient

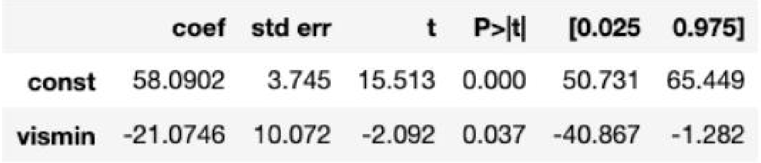
● 0.19 √ 正确

○ 0.036

○ 0.034

○ 17.08

4. We run a regression analysis in place of a t-test to test if there is a difference in number of students enrolled in classes with professors who are visible minority(vismin = 1) vs professors who are not (vismin = 0). The table is shown below. What does the coefficient for vismin mean?



	coef	std err	t	P>|t|	[0.025	0.975]
const	58.0902	3.745	15.513	0.000	50.731	65.449
vismin	-21.0746	10.072	-2.092	0.037	-40.867	-1.282

○ Professors who are visible minority get about 21 students more on average that professors who aren't visible minority.

○ We can't conclude because the error is too large and if factored could change the conclusion of the tests.

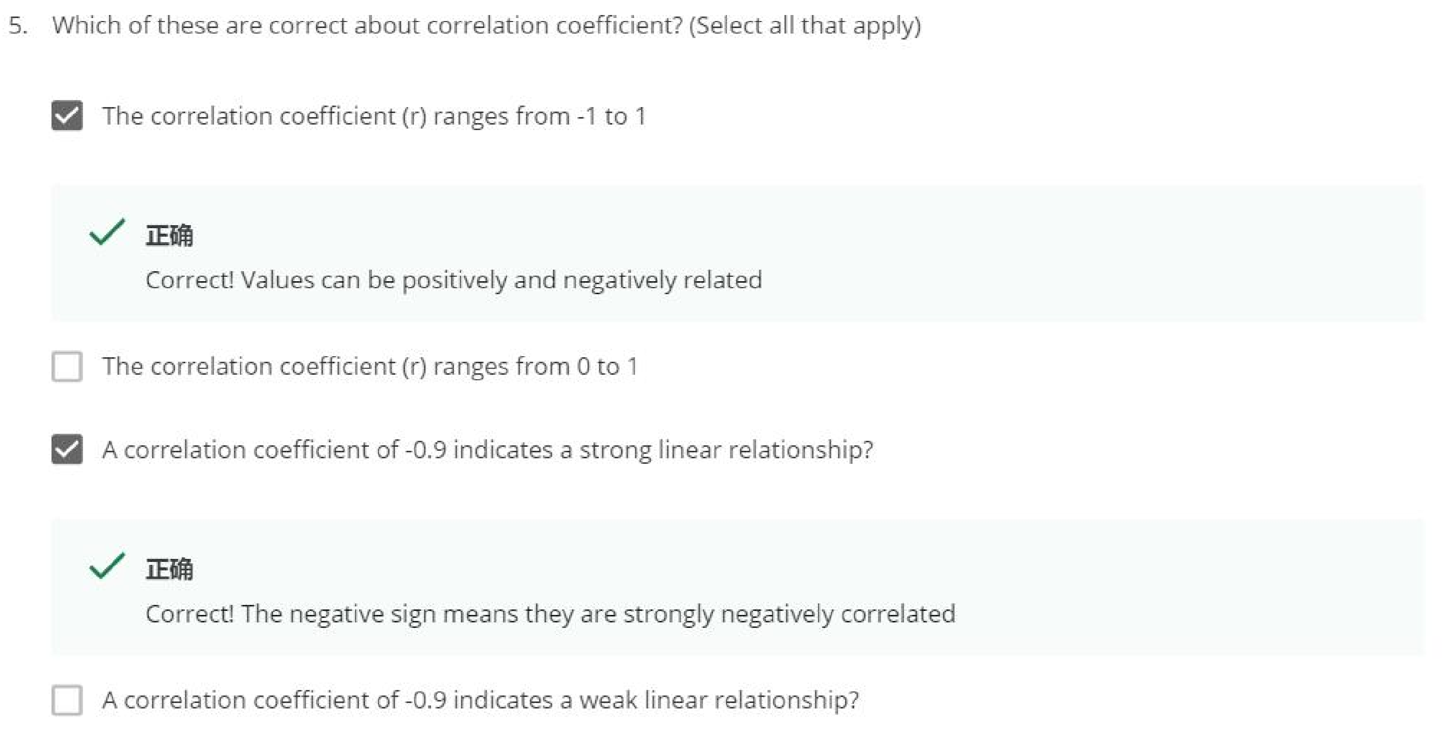
● Professors who are visible minority get about 21 students less on average that professors who aren't visible minority.

○ Professors who are visible minority get about 58 students less on average that professors who aren't visible minority.

✓ **正确**

Correct!

总结：



5. Which of these are correct about correlation coefficient? (Select all that apply)

☑ The correlation coefficient (r) ranges from -1 to 1

✓ 正确

Correct! Values can be positively and negatively related

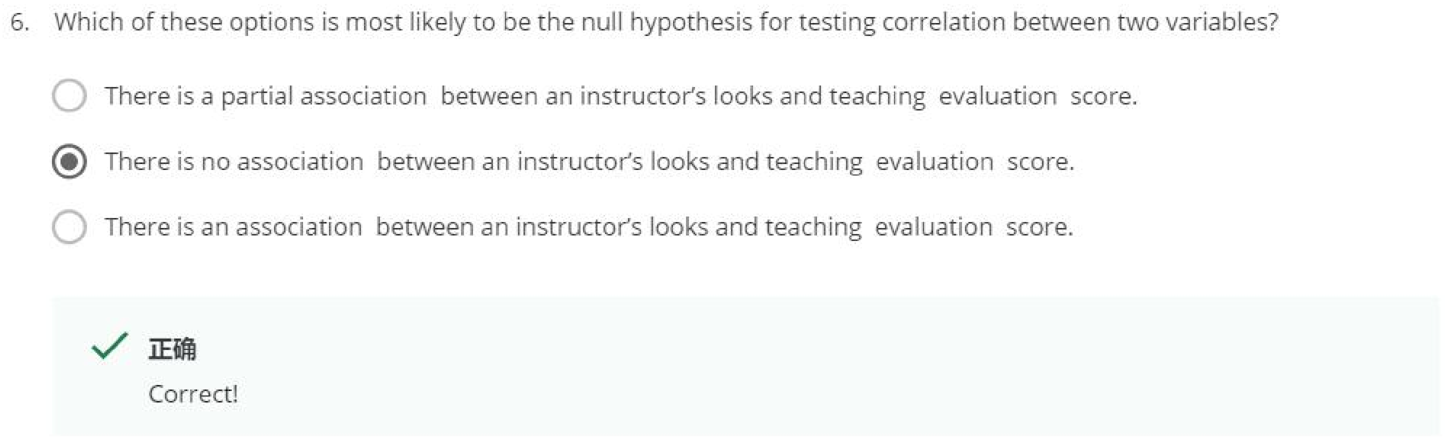
□ The correlation coefficient (r) ranges from 0 to 1

☑ A correlation coefficient of -0.9 indicates a strong linear relationship?

✓ 正确

Correct! The negative sign means they are strongly negatively correlated

□ A correlation coefficient of -0.9 indicates a weak linear relationship?



6. Which of these options is most likely to be the null hypothesis for testing correlation between two variables?

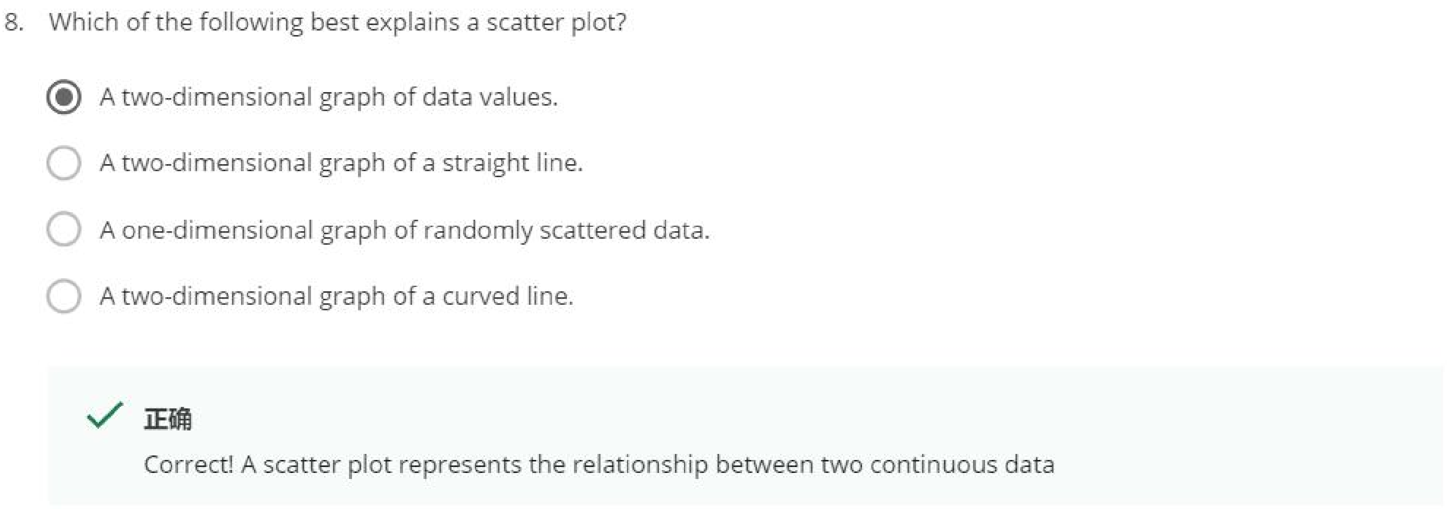
○ There is a partial association between an instructor's looks and teaching evaluation score.

● There is no association between an instructor's looks and teaching evaluation score.

○ There is an association between an instructor's looks and teaching evaluation score.

✓ 正确

Correct!



8. Which of the following best explains a scatter plot?

● A two-dimensional graph of data values.

○ A two-dimensional graph of a straight line.

○ A one-dimensional graph of randomly scattered data.

○ A two-dimensional graph of a curved line.

✓ 正确

Correct! A scatter plot represents the relationship between two continuous data

总结：