5.1 Introduction to simple linear regression (SLR)

Regression lines

A ***simple linear regression*** is a way to model the linear relationship between two quantitative variables, using a line drawn through those variables' data points, known as a ***regression line***. Ex: The animation below shows a scatterplot with several data points that represent house price and house size for 5 houses in Eugene, Oregon (Source: Victoria Whitman's 2005 house sales), with a regression line drawn through the points. A common use of a regression line is to make predictions.

PARTICIPATION ACTIVITY 5.1.1: House price and house size scatterplot.

Animation captions:

1. A particular house with 1683 square feet sold for a price of $259,900.
2. Other houses with the indicated square feet sold for other prices.
3. The trend suggests that as house square feet increase, house prices also increase. The relationship appears to be linear.
4. The simple linear regression line summarizes the relationship, but individual data points are typically above or below the line.
5. A common use of a linear regression model is to make predictions. Ex: The model predicts that a 2200 square foot house would cost about $280,000.

PARTICIPATION ACTIVITY 5.1.2: Linear relationship between house price and size.

Refer to the animation above.

1) The house with 1922 square feet sold for $269,900. Is this house's data point above or below the regression line?

- ○ Above
- ○ Below

2) Did the house with 1922 square feet sell for more or less than one might expect based on the regression line?

- More
- Less

3) Given a house with 2100 square feet and a sale price of $\$269,000$, would this house's point lie above or below the regression line?

- Above
- Below

4) The regression line only represents the linear relationship for points that lie on the line.

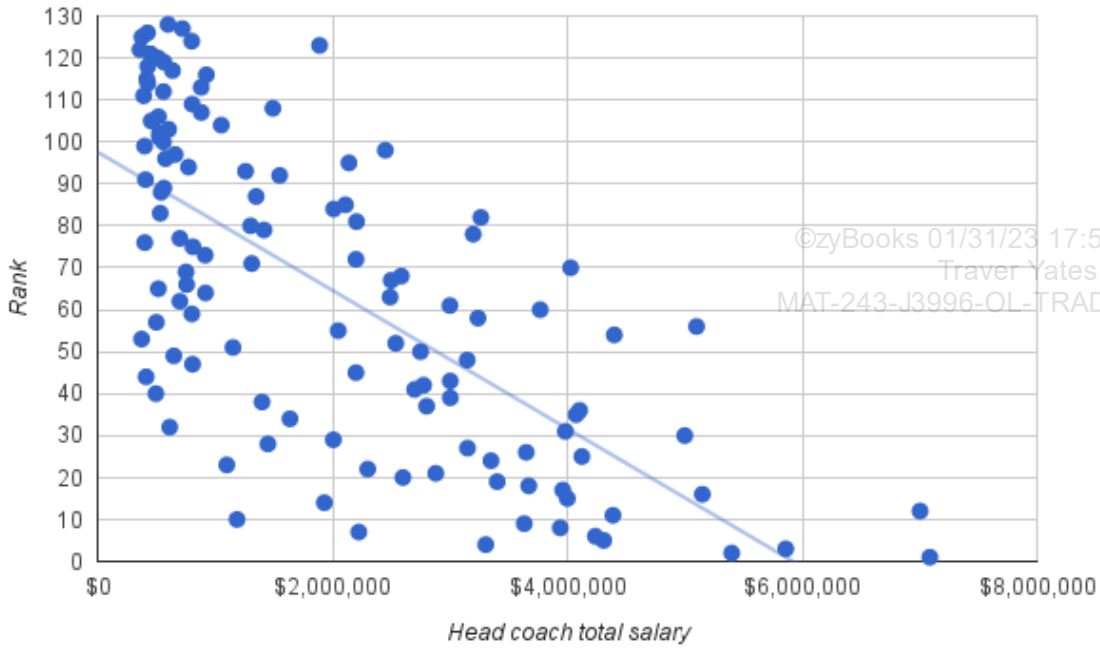
- True
- False

5) Using the linear regression model, what is the predicted price of an 1800 sq ft house?

- $\$260,000$
- $\$280,000$

Below is a scatterplot showing all 128 college football team rankings and the total salary for each team's head coach[1] [2]. The regression line shows that higher salaries tend to be related to better rank.

Figure 5.1.1: College football rankings vs. head coach salary.



PARTICIPATION ACTIVITY 5.1.3: College football rank and coach salaries.

Refer to the figure above.

1) The #1 ranked team's coach earns what approximate salary?

- $\$6,000,000$
- $\$7,000,000$

2) About nine coaches have salaries very near to $\$4$ million. What is the approximate range of ranks for those coaches' nine teams?

- $10 - 70$
- $1 - 130$

3) Assuming salary predicts rank, if a school hired a coach and paid the coach $\$4$ million, what rank might the school expect?

- 10
- 32

- 70

4) Do most of the 128 data points lie on the regression line?
   - Yes
   - No

5) A team hires a coach with a salary of $4 million, but ends up with a rank of 100. Is the regression line wrong?
   - Yes
   - No

Response and predictor variables

Price is linearly related to size in the house example above. In the scatterplot with price on the vertical Y axis and size on the horizontal X axis, the regression line slopes upwards. This upward slope means larger houses tend to cost more than smaller houses. Without additional information, a person would predict that a larger house would cost more than a small house. Here, house price is known as the response variable, while house size is the predictor variable:

In a linear regression involving two variables, the ***response variable*** is the variable being modeled or predicted, while the ***predictor variable*** is the variable used to predict the response.

The response variable often responds in some way to a change in the predictor. Knowing the value of a predictor often allows one to predict the response. In a linear regression, the response variable is sometimes called the ***dependent variable*** or ***output*** or ***outcome***. Also, in a linear regression, a predictor variable is sometimes called an ***independent variable*** or ***input*** or ***covariate***.

PARTICIPATION ACTIVITY 5.1.4: Identifying response and predictor variables.

1) An analyst examines a baseball team's batting average and the number of games won over a season. What is the predictor variable?
   - Batting average
   - Games won

2) A researcher claims that a country's electricity consumption is higher given a greater gross domestic product (GDP). What is the predictor variable?

- Electricity consumption
- GDP

3) Consider predicting a car's fuel efficiency in miles per gallon from the car's engine size. What is the response variable?

- Fuel efficiency
- Engine size

4) A restaurant owner wishes to model the linear relationship between mean daily costs and the number of customers served, to help manage future costs. What is the response variable?

- Daily costs
- Customers served

Creating a simple linear regression line

A scatterplot can display all the information of a simple linear regression model. The response variable goes on the vertical Y axis, while the predictor variable goes on the horizontal X axis. The plotted points show the data. The simple linear regression line shows the overall linear relationship among the plotted points [Regress]. Each position along the line represents where one would expect Y to be given a particular X, based on the linear relationship. Assuming the data represents a population, the line can be created as follows:

The ***population simple linear regression function*** is $\mathrm{E}(Y) = \beta_0 + \beta_1 X$, where β_0 and β_1 are regression parameters.

In the linear regression function, the E stands for ***expected value***, so $E(Y)$ is read as "the expected value of Y".

PARTICIPATION ACTIVITY 5.1.5: Simple linear regression line.

1) What sign does the slope of the simple linear regression line have if $\beta_1 > 0$?

- Positive
- Negative

2) What is the sign of Y when $X = 0$ on the simple linear regression line if $\beta_0 > 0$?

- Positive
- Negative

A particular data point has an *actual* value for Y. Relative to the regression line, that point may be:

- *above* the line: $Y > \mathrm{E}(Y)$
- *below* the line: $Y < \mathrm{E}(Y)$, or
- *on* the line: $Y = \mathrm{E}(Y)$

For each actual data point Y, a (positive or negative) value can be added to the expected value $E(Y)$ to achieve Y, as below.

The ***simple linear regression model*** is $Y = \beta_0 + \beta_1 X + \varepsilon$, where ε is the regression error term.

This error is not an error in the sense of a mistake. A ***regression error***, $\varepsilon = Y - \mathrm{E}(Y)$, is a statistical error modeled as a random variable with a normal distribution that has zero mean and constant variance. The regression error is:

- positive for points that lie above the regression line,
- negative for points that lie below the regression line, and
- zero for points that lie on the regression line.

Given a linear regression model, the model is commonly used to make predictions. A later section discusses how to estimate the regression parameters, β_0 and β_1.

PARTICIPATION ACTIVITY 5.1.6: Making predictions using a simple linear regression model between house price and size.

Refer to the simple linear regression model above. Suppose that the simple linear regression model for the house price example has $\beta_0 = 190,000$ and $\beta_1 = 40$.

1) What would one expect a house with 2000 square feet to sell for?

Check Show answer

2) Given a house with 2000 square feet and a sale price of $\$271,000$, what is the regression error for this house?

Check Show answer

3) What would one expect a house with 2200 square feet to sell for?

Check Show answer

4) Given a house with 2200 square feet and a sale price of $\$275,000$, what is the regression error for this house?

Check Show answer

Minimizing absolute error

The regression parameters are typically unknown because the entire population dataset is not often observable. Instead, a random sample of data points is obtained. Statistical inference is used to draw inferences about the population based on what can be observed in the sample. To apply statistical inference to make a prediction using regression requires estimating the population regression parameters β_0 and β_1 from sample data. Thus, given sample data points, finding the "best fitting" regression line for those data points provides estimates of β_0 and β_1.

One method for estimating β_0 and β_1 minimizes the sum of the absolute errors for the sample points.

PARTICIPATION ACTIVITY 5.1.7: A best fitting regression line minimizes the errors.

Animation captions:

1. A sample of 4 data points is randomly selected from a population.
2. A horizontal regression line has large errors and does not fit well. Each error, ε, is the vertical distance between a point and the line.
3. The sum of absolute errors, SAE, is 48.
4. Increasing the slope of the regression line improves the fit. SAE is reduced to 34.
5. As the slope of the regression line increases further, SAE is reduced to 20.
6. As the slope of the regression line increases further, SAE decreases to 14.

7. As the slope of the regression line increases further, SAE increases to 20.
8. The line that appears to minimizes the sum of absolute errors has an SAE of 14.

PARTICIPATION ACTIVITY 5.1.8: Calculating errors for a regression line.

Refer to the animation above.

1) What are the regression errors for the horizontal regression line at $Y = 17$?
   - $12, 12, 10, 14$
   - $-12, -12, 10, 14$
   - $12, 12, -10, -14$

2) What is the sum of errors for the horizontal regression line at $Y = 17$?
   - 0
   - 48

3) What is the sum of *absolute* errors for the horizontal regression line at $Y = 17$?
   - 0
   - 48

4) For the regression line given by $Y = 2 + 3X$, what are the Y values along that line that correspond to the sample X values: $0, 3, 7, 10$?
   - $5, 5, 27, 31$
   - $2, 11, 23, 32$

5) What are the regression errors for the regression line given by $Y = 2 + 3X$?
   - $3, -6, 4, -1$
   - $-3, 6, -4, 1$

6) What is the sum of absolute errors for the regression line given by $Y = 2 + 3X$?
   - 0
   - 14

Least squares method

Summing the *absolute* errors is one method to measure how far the line is from the points. Another method to measure error is by computing the *sum of squared errors*. The sum of squared errors is the sum of the differences between the Y values of the data points and the values obtained from the regression line.

Example 5.1.1: Computing the sum of squared errors.

Given the data points below, compute the sum of squared errors for the regression equation $Y = 7 + 2X$.

X	0	3	7	10
Y	5	5	27	31

Solution

For each value of X, the regression values from the equation $Y = 7 + 2X$ are obtained. The errors are the difference between the Y values of the data points and the regression values.

X	0	3	7	10
Y	5	5	27	31
$Y = 7 + 2X$	7	13	21	27
Errors	-2	-8	6	4
Square errors	4	64	36	16

Thus, the sum of the square errors is $4 + 64 + 36 + 16 = 120$.

PARTICIPATION ACTIVITY 5.1.9: Calculating sum of squared errors for a regression line.

Given the data points below, compute the sum of squared errors for the regression equation $Y = 2 + 3X$.

X	0	3	7	10

Y	5	5	27	31

1) What are the squared errors for the regression line given by $Y = 2 + 3X$? Type as a comma-separated list: #, #, #, #

Check Show answer

2) What is the sum of squared errors for the regression line give by $Y = 2 + 3X$?

Check Show answer

The ***method of least squares*** derives a linear regression model by minimizing the sum of squared errors. For a sample with n observations, the sum of squared errors is $\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$.

The ***sample simple linear regression function*** is $\hat{Y} = b_0 + b_1 X$, where $\hat{Y}$ are the predicted or fitted response values based on the simple linear regression model, and the regression parameter estimators, b_0 and b_1 are the values of the regression parameters, β_0 and β_1 that minimize the sum of squared errors.

The "hat" notation in $\hat{Y}$ is a statistical convention that denotes a sample estimate. $\hat{Y} = b_0 + b_1 X$ is the sample simple linear regression line that estimates the population simple linear regression line $\mathrm{E}(Y) = \beta_0 + \beta_1 X$.

A ***simple linear regression fitted value***, $\hat{Y}_i = b_0 + b_1 X_i$, is the predicted value of Y for the ith sample value of X based on the sample simple linear regression line.

A ***simple linear regression residual***, $\varepsilon_i = Y_i - \hat{Y}_i$, is the ith estimated regression error based on the sample simple linear regression line.

For the data above, $b_0 = 2$ and $b_1 = 3$ minimize the sum of squared errors. Thus, the sample simple linear regression line is $\hat{Y} = 2 + 3X$. The fitted value when $X_1 = 0$ is $\hat{Y}_1 = 2 + 3(0) = 2$. The corresponding regression residual is $\varepsilon_1 = Y_1 - \hat{Y}_1 = 5 - 2 = 3$.

PARTICIPATION ACTIVITY 5.1.10: Calculating fitted values and residuals for a sample simple linear regression line.

Use the sample simple linear regression line $\hat{Y} = 2 + 3X$ for the data points below to answer the following questions.

X	0	3	7	10
Y	5	5	27	31

1) What is the fitted value when $X_2 = 3$?

Check Show answer

2) What is the regression residual when $X_2 = 3$?

Check Show answer

Python-Function 5.1.1: linregress(x,y).

The intercept parameter estimator b_0 and the slope parameter estimator b_1 can be obtained by using the `linregress(x,y)` function, which takes in two arrays of equal length as input and returns b_0, b_1, the correlation coefficient r, the p-value for the correlation coefficient t-test, and the corresponding standard error. The `linregress(x,y)` function uses the scipy.stats library.

```
# The numpy library is imported to build two arrays
import numpy as np
import scipy.stats as st

x = np.array([0, 3, 7, 10])
y = np.array([5, 5, 27, 31])

print(st.linregress(x,y))
```



```
LinregressResult(slope=3.0, intercept=2.0, rvalue=0.94542880030087728,
pvalue=0.054571199699122705, stderr=0.73108327748669655)
```

Run example

References

(*1) "NCAA College Football Predictive Rankings and Ratings." *TeamRankings*. www.teamrankings.com/college-football/ranking/predictive-by-other.

(*2) "NCAA Salaries." *USA Today*. sports.usatoday.com/ncaa/salaries/

(*Regress) The unusual name "regression line" stems from 19th century genetics researcher Francis Galton, who noticed that heights of descendants of tall people tended to "regress" back towards typical heights. Regress means to go back to a former state.

5.2 SLR assumptions

Simple linear regression assumptions

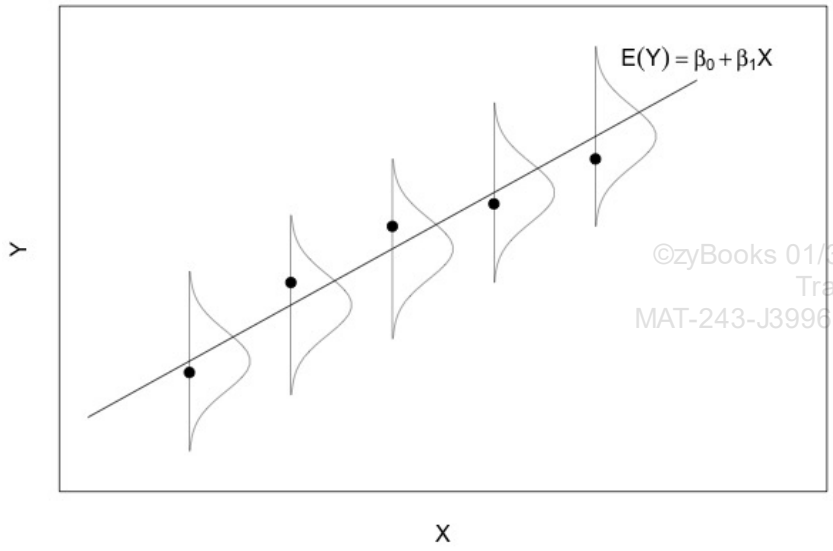
The simple linear regression model assumes that at each value of the predictor, X, the probability distribution of the regression error, $\varepsilon = Y - \mathrm{E}(Y) = Y - \beta_0 - \beta_1 X$:

- has a mean of zero,
- has constant variance, and
- is normal.

At each value of X in the figure below, the density curve shows the probability distribution of the regression error to be:

- centered on the regression line (where the error is zero),
- equally wide (equally variance), and
- bell-shaped and symmetric (normal).

Figure 5.2.1: Scatterplot illustrating regression error assumptions.



In addition, the value of ε for one observation is assumed to be independent of the value of ε for any other observation.

PARTICIPATION ACTIVITY 5.2.1: Simple linear regression assumptions.

Assume the simple linear regression model assumptions hold.

1) For a particular value of X, where can the expected value of Y be found with respect to the simple linear regression line?
   - Above the regression line
   - Below the regression line
   - On the regression line

2) As X increases, what happens to the variability of Y values around the regression line?
   - Increases
   - Decreases
   - Remains constant

3) What information about the value of other regression errors is known when the value of one regression error is known?

- None
- Some
- All

Checking simple linear regression assumptions

An earlier section stated that the simple linear regression model assumes that at each value of the predictor, X, the probability distribution of the regression error, $\varepsilon = Y - \mathrm{E}(Y) = Y - \beta_0 - \beta_1 X$:

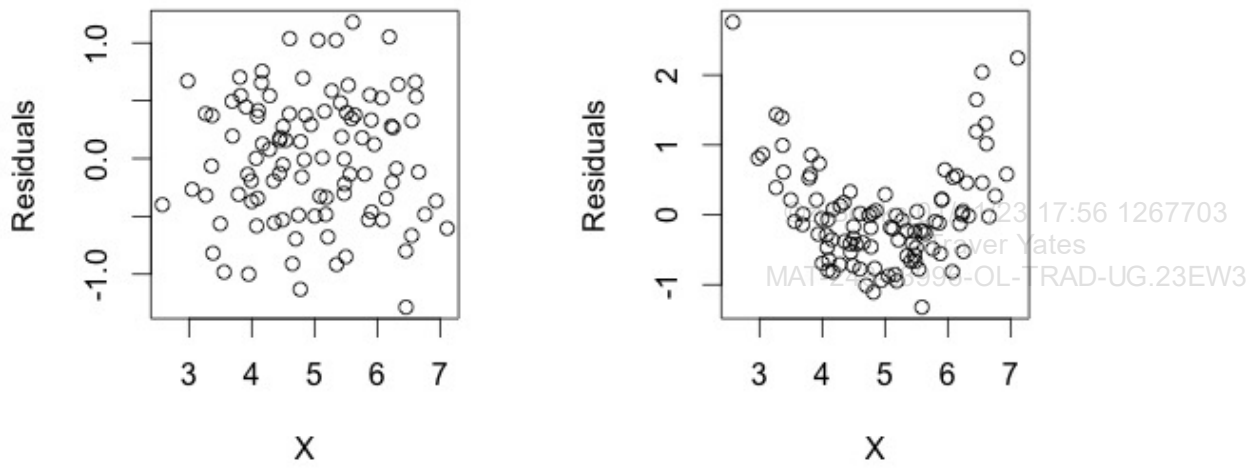
- has a mean of zero,
- has constant variance, and
- is normal.

In addition, the value of ε for one observation is assumed to be independent of the value of ε for any other observation.

A simple linear regression model should generally only be used for prediction if a reasonable chance exists that these assumptions hold in the population. Population regression errors are not observable, so to assess the assumptions, the sample residuals, $\varepsilon_i = Y_i - \hat{Y}_i$, are used. Statistical tests for the residuals are available (not discussed here). Alternatively, graphical methods can be used.

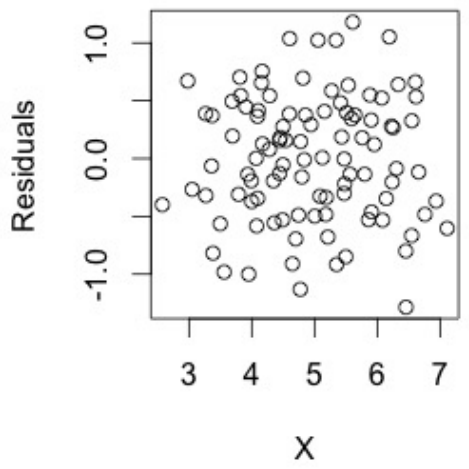
Assessing the "mean of zero" assumption: Create a scatterplot with the residuals on the vertical axis and X on the horizontal axis. Scan the plot from left to right and visually estimate whether the mean of the residuals for each fixed value of X is approximately zero. If the mean is approximately zero all the way across the plot then conclude that the "mean of zero" assumption holds. Ex: The following plot on the left has a residual mean that is approximately zero all the way across the plot, so the "mean of zero" assumption holds. The residuals in the left-hand plot look random. By contrast, the residuals in the plot on the right have a non-random "falling-rising" pattern. The residual mean is positive on the left of the plot when X is low, negative in the middle of the plot, and positive on the right of the plot when X is high. The "mean of zero" assumption does not hold for the data in the right-hand plot. Other patterns that would indicate the "mean of zero" assumption does not hold include a "rising-falling" pattern or any strong linear trend.

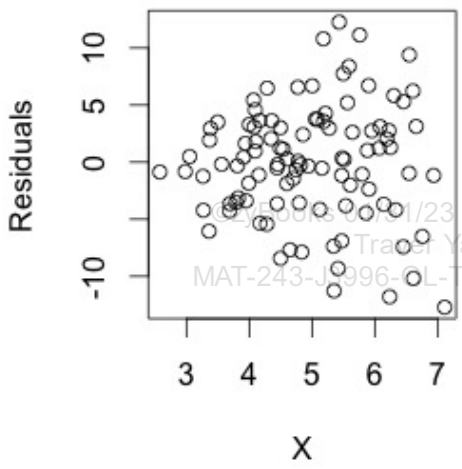
Figure 5.2.2: Residual plots for assessing the mean of zero assumption.



Assessing the "constant variance" assumption: Create a scatterplot with the residuals on the vertical axis and X on the horizontal axis. Scan the plot from left to right and visually estimate whether the variance of the residuals for each fixed value of X is approximately constant. If the variance is approximately constant all the way across the plot then conclude that the "constant variance" assumption holds. Ex: The following plot on the left has a residual variance that is approximately constant all the way across the plot, so the "constant variance" assumption holds. The residuals in the left-hand plot look random. By contrast, the residuals in the plot on the right have a non-random "megaphone" pattern. The residual variance is low on the left of the plot when X is low, medium in the middle of the plot, and high on the right of the plot when X is high. The "constant variance" assumption does not hold for the data in the right-hand plot. Another pattern that would indicate the "constant variance" assumption does not hold is a "funnel" pattern where the residual variance is high on the left of the plot and low on the right of the plot.

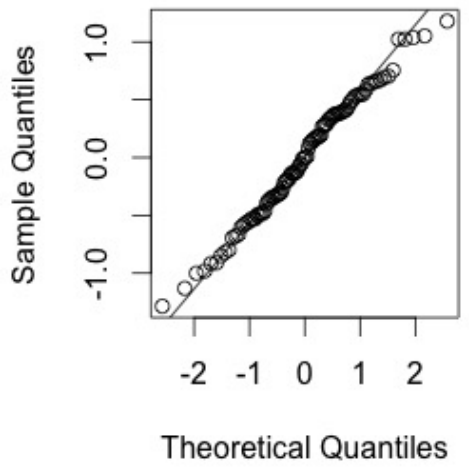
Figure 5.2.3: Residual plots for assessing the constant variance assumption.

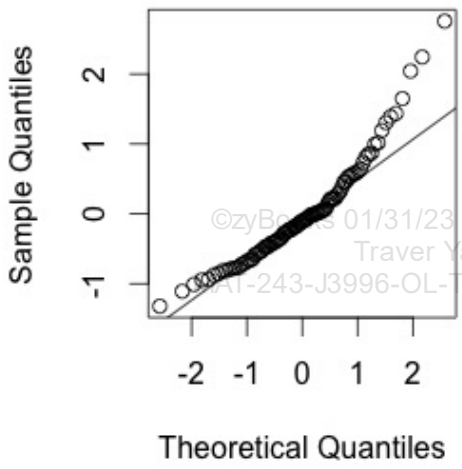




Assessing the "normality" assumption: Create a normal probability plot of the residuals, which has the ordered residuals plotted on the vertical axis and theoretical normal quantiles plotted on the horizontal axis. If the plotted points lie reasonably close to the diagonal line on the plot then conclude that the "normality" assumption holds. Ex: The following plot on the left has plotted points that lie reasonably close to the diagonal line, so the "normality" assumption holds. By contrast, the plotted points in the plot on the right do not lie close to the diagonal line. The "normality" assumption does not hold for the data in the right-hand plot. Another pattern that would indicate the "normality" assumption does not hold is a nonlinear pattern curving down instead of up.

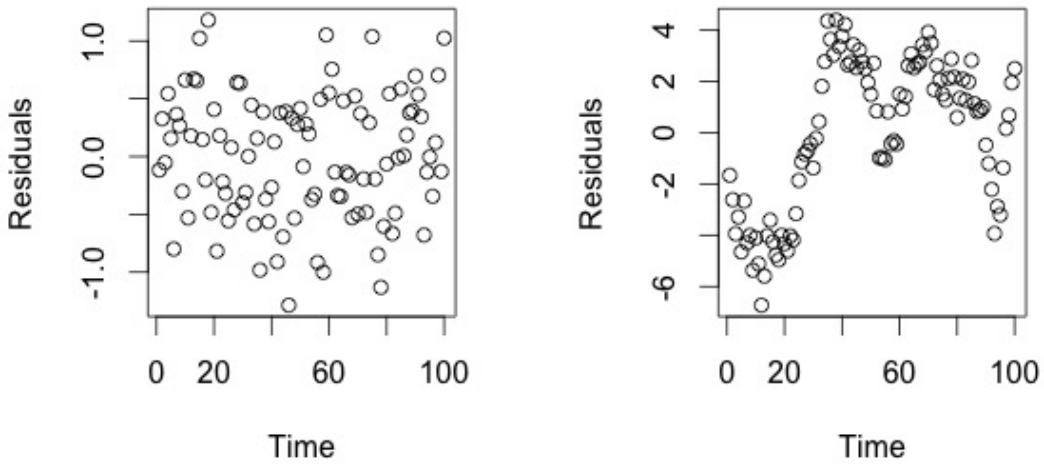
Figure 5.2.4: Residual normal probability plots for assessing the normality assumption.





Assessing the "independence" assumption: One form of dependence that can affect regression models arises through poor study design. Ex: failing to adjust for the same family members in a study of cholesterol. Another common form of dependence is time dependence. Such time dependence can be assessed if the data were collected in time order and that information is available. Otherwise, the independence assumption is difficult to assess. However, if a variable that tracks time is available, create a scatterplot with the residuals on the vertical axis and time order on the horizontal axis. Scan the plot from left to right and visually estimate whether the residuals tend to track one another no more closely than would be expected by chance. If the residuals seem random then conclude that the "independence" assumption holds. Ex: The following plot on the left has residuals with no discernible pattern all the way across the plot, so the "independence" assumption holds. In contrast, the residuals in the plot on the right tend to track one another more closely than would be expected by chance. The "independence" assumption does not hold for the data in the right-hand plot.

Figure 5.2.5: Time order residual plots for assessing the independence assumption.



PARTICIPATION ACTIVITY 5.2.2: Residual plots for assessing regression assumptions.

1) A residual scatterplot with fitted values on the horizontal axis has a clear positive linear trend. Which regression assumption does such a pattern cast doubt upon?

   ○ Mean of zero

- Constant variance
- Normality
- Independence

2) A residual scatterplot with fitted values on the horizontal axis has vertical variation that increases from low on the left side of the plot to high on the right side of the plot. Which regression assumption does such a pattern cast doubt upon?

- Mean of zero
- Constant variance
- Normality
- Independence

3) A residual scatterplot with time order on the horizontal axis has points that seem to track one another more than would be expected by chance. Which regression assumption does such a pattern cast doubt upon?

- Mean of zero
- Constant variance
- Normality
- Independence

4) Which regression assumption is not generally assessed using a residual scatterplot?

- Mean of zero
- Constant variance
- Normality
- Independence

5) How do outliers stand out on a residual scatterplot?

- As isolated points on the far left or far right of the plot

- As isolated points at the top or bottom of the plot
- As isolated points anywhere on the edge of the plot

The following precautions should be taken when checking the simple linear regression assumptions graphically:

- Assessing whether a particular plot supports an assumption is subjective. One should question an assumption when a clear non-random pattern exists in a plot, but not be overly concerned with weak patterns.
- Checking assumptions graphically requires a reasonably large sample size, typically at least 30.
- Ideally, all four assumptions should hold for a model to be valid.
- That said, simple linear regression models are reasonably robust to mild violations of the assumptions.
- Severe violations of the assumptions can be addressed through more complex models, some of which are considered in a later section.

5.3 Correlation and coefficient of determination

Correlation

Correlation describes the association or dependence between two variables. A ***positive correlation*** between two variables means that as one variable increases, the other variable increases as well. A ***negative correlation*** between two variables means that as one variable increases, the other variable decreases. The strength of correlation between a predictor variable and a response variable can be measured by the *correlation coefficient*. The population correlation coefficient is denoted by ρ and the sample correlation coefficient is denoted by R.

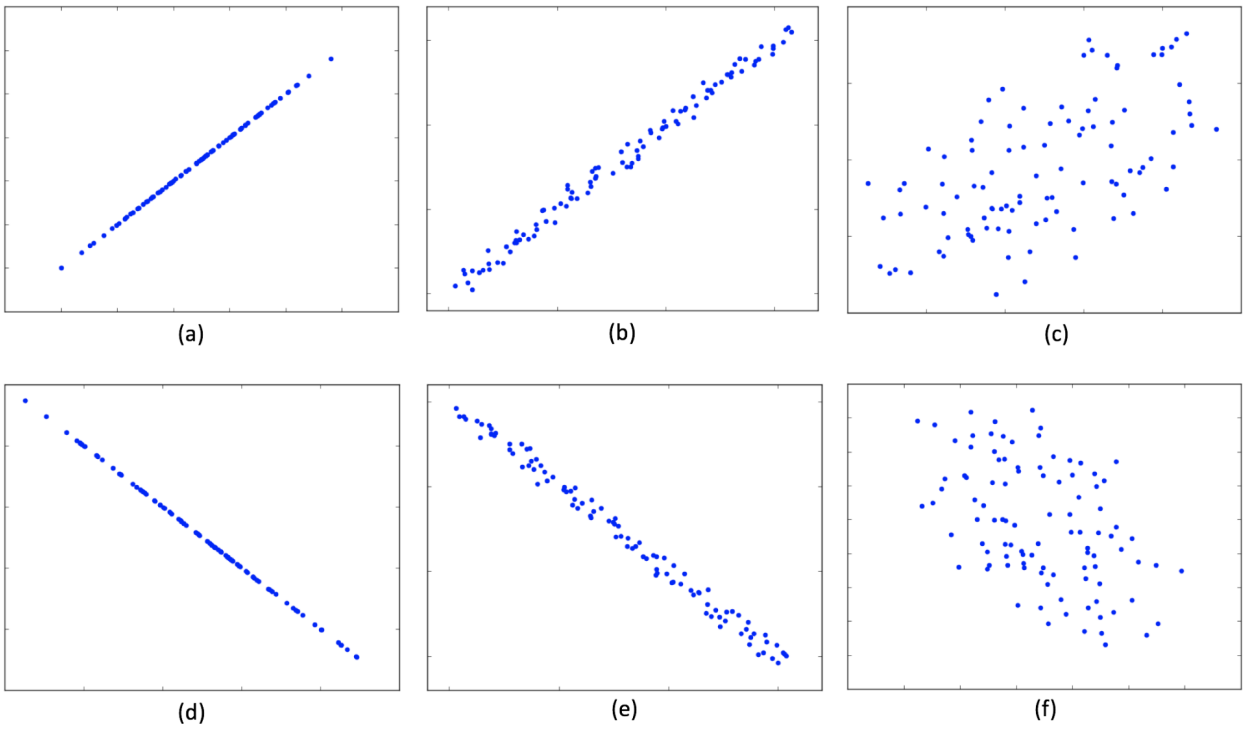
The strength of correlation can be described by the absolute value of R. The table below gives a rough guideline.

Table 5.3.1: Strength of correlation.

Value of $\lvert R \rvert$	Strength of correlation
$0 < \lvert R \rvert \leq 0.40$	Weak

$0.40 < \lvert R \rvert \leq 0.80$	Moderate
$0.80 < \lvert R \rvert \leq 1.00$	Strong

Figure 5.3.1: (a) Perfect positive correlation, (b) strong positive correlation, (c) weak positive correlation, (d) perfect negative correlation, (e) strong negative correlation, (f) weak negative correlation.



Several methods to find the correlation coefficient exist. The formula to calculate the *Pearson correlation coefficient* R is given below

$$R = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\left(\sum x^2\right) - (\sum x)^2}\sqrt{n\left(\sum y^2\right) - (\sum y)^2}}$$

Calculating the correlation coefficient using the formula is almost never done by hand, especially when dealing with large datasets. Commonly, a programming language like R or Python is used. The example below illustrates the use of the formula. However, built-in functions will be used in further examples.

Example 5.3.1: Calculating the correlation coefficient using the formula.

Consider the paired datasets $X = \{1, 2, 4, 5\}$ and $Y = \{1, 3, 5, 7\}$.

a. Find the correlation coefficient between X and Y.
b. Determine the strength and direction of the correlation between X and Y.

Solution

a. The table below is useful in obtaining $\sum xy$, $\sum x$, $\sum y$, $\sum x^2$, and $\sum y^2$.

x	y	xy	x^2	y^2
1	1	1	1	1
2	3	6	4	9
4	5	20	16	25
5	7	35	25	49
$\sum x = 12$	$\sum y = 16$	$\sum xy = 62$	$\sum x^2 = 46$	$\sum y^2 = 84$

Note that $n = 4$ because each dataset has 4 elements. Thus,

$$R = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\left(\sum x^2\right) - (\sum x)^2}\sqrt{n\left(\sum y^2\right) - (\sum y)^2}}$$

$$= \frac{4(62) - (12)(16)}{\sqrt{4(46) - (12)^2}\sqrt{4(84) - (16)^2}}$$

$$= 0.990$$

b. Since the value of $R = 0.990 > 0.8$, X and Y are positively and strongly correlated.

Often, applications deal with more than one variable. A ***correlation matrix*** is a table that shows the correlation coefficients between each pair of variables. A correlation matrix for a collection of n variables $X_1, \ldots, X_n$ is an $n \times n$ matrix where the ijth entry is the correlation coefficient of X_i and X_j. Note that the diagonal entries are always 1 because the correlation between a variable and itself is 1. Further, the correlation matrix is symmetric. That is, the ijth entry is the same as the jith entry.

Python-Function 5.3.1: corr().

The correlation function `DataFrame.corr()` calculates the pairwise correlation of the columns in the dataframe and outputs a correlation matrix.

The code below uses the Exam1 and Exam2 columns of the ExamScore dataset and produces a 2 × 2 correlation matrix.

```
import pandas as pd
scores = pd.read_csv("http://data-analytics.zybooks.com/ExamScores.csv")
print(scores[['Exam1','Exam2']].corr())
```

```
          Exam1     Exam2
Exam1  1.000000  0.078613
Exam2  0.078613  1.000000
```


The code below uses the Exam1, Exam2, Exam3, and Exam4 columns of the ExamScore dataset and produces a 4 × 4 correlation matrix.

```
print(scores[['Exam1','Exam2','Exam3','Exam4']].corr())
```

```
          Exam1     Exam2     Exam3     Exam4
Exam1  1.000000  0.078613  0.256859  0.261306
Exam2  0.078613  1.000000  0.271642  0.318124
Exam3  0.256859  0.271642  1.000000  0.277656
Exam4  0.261306  0.318124  0.277656  1.000000
```

Run example

PARTICIPATION ACTIVITY 5.3.1: Finding the correlation between the variables.

Use the coefficient matrix below to answer the questions below about the correlation between exam scores in the ExamScores dataset.

```
          Exam1     Exam2     Exam3     Exam4
Exam1  1.000000  0.078613  0.256859  0.261306
Exam2  0.078613  1.000000  0.271642  0.318124
Exam3  0.256859  0.271642  1.000000  0.277656
Exam4  0.261306  0.318124  0.277656  1.000000
```


1) What is the correlation coefficient between Exam1 and Exam4?

- 0.277656
- 0.261306
- 0.078613

2) What is the sign or direction of the correlation between Exam1 and Exam4?

- Positive
- Negative

3) What is the strength of the correlation between Exam1 and Exam4?

- Weak
- Moderate
- Strong

t-test for the population correlation coefficient

The distribution of the Pearson correlation coefficients for samples of size n follows a t-distribution with $n - 2$ degrees of freedom. When determining whether a linear relationship or association exist, the t-test for population correlation coefficient is useful.

The t-test for the population correlation coefficient is performed as follows.

Procedure 5.3.1: The t-test.

1. Set the null and alternative hypotheses

$$H_0 : \rho = 0 \quad H_a : \rho > 0 \text{ (right-tailed)} \quad H_a : \rho < 0 \text{ (left-tailed)} \quad H_a : \rho \neq 0 \text{ (two-tailed)}$$

   where ρ is the population correlation coefficient.

2. Use statistical software to find the test-statistic

$$t = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

   where the degrees of freedom is $n - 2$.

3. Use statistical software to find the p-value that corresponds to t.
4. Make a decision given a previously selected significance level α, typically 0.05.
   - If the p-value is less than the significance level, sufficient evidence exists to reject the null hypothesis H_0 in favor of the alternative hypothesis H_a. That is, statistically significant evidence exists to support a correlation or association between the variables.

- If the p-value is greater than or equal to the significance level, insufficient evidence exists to reject the null hypothesis H_0. That is, no statistically significant evidence exists to support a correlation or association between the variables.

Example 5.3.2: Using the t-test for population correlation coefficient.

The teacher of this year's class of 50 students plans to use the same exam the following year. The teacher has reason to believe that a positive linear relationship exist between student scores in Exam1 and Exam4 from the ExamScores dataset. Does sufficient evidence exist at the $\alpha = 0.05$ significance level to support the teacher's claim?

Solution

The null hypothesis is that the population correlation coefficient is zero, which implies that no linear relationship exists between Exam1 and Exam4. The alternative hypothesis is that the population correlation coefficient is positive, which implies a linear relationship between Exam1 and Exam4. Mathematically,

$$H_0 : \rho = 0 \quad H_a : \rho > 0$$

Since the correlation coefficient between Exam1 and Exam4 is $R = 0.261306$, the t test statistic is

$$t = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} = \frac{0.261306\sqrt{50-2}}{\sqrt{1-0.261306^2}} \approx 1.876$$

The degrees of freedom is $df = 50 - 2 = 48$. Thus, the p-value is $P(T \geq 1.876) = 0.033$.

Since the p-value is less than the significance level $\alpha = 0.05$, sufficient evidence exists that a positive linear relationship exists between Exam1 and Exam4.

Python-Function 5.3.2: scipy.stats.pearsonr(x,y).

The Pearson correlation coefficient function `pearsonr(x,y)` takes in two arrays or DataFrames x and y and outputs the Pearson correlation coefficient and the corresponding two-tailed p-value. The function requires the `scipy.stats` library.

The code below uses the Exam1 and Exam4 columns of the ExamScore dataset and outputs the correlation coefficient and the two-tailed p-value obtained.

```
import pandas as pd
import scipy.stats as st
scores = pd.read_csv("http://data-analytics.zybooks.com/ExamScores.csv")
st.pearsonr(scores['Exam1'],scores['Exam4'])
```

```
(0.26130551001268937, 0.066807681266337585)
```


Note that this two-tailed p-value (0.067) is twice the one-tailed p-value obtained in the example above (0.033).

Run example

PARTICIPATION ACTIVITY 5.3.2: Using the t-test for population correlation coefficient.

A researcher claims that the packed cell volume, the percent volume of red blood cells in blood, has a linear relationship with hemoglobin, a protein in red blood cells used to transport oxygen. Using the `pearsonr()` function on the Hb and PCV columns of the PackedCellVolume dataset as shown below, determine whether sufficient evidence exist to support the researcher's claim at the $\alpha = 0.01$ significance level.

```
import pandas as pd
import scipy.stats
scores = pd.read_csv("http://data-analytics.zybooks.com/PackedCellVolume.csv")
scipy.stats.pearsonr(scores['Hb'],scores['PCV'])
```

```
(0.86988551147299953, 5.2622218922693876e-05)
```

Run example

1) What is the null hypothesis?

- $R = 0$
- $\rho = 0$
- $\rho \neq 0$

2) What is the alternative hypothesis?

- $R \neq 0$
- $\rho = 0$
- $\rho \neq 0$

3) What is the p-value?

- 5.262×10^{-5}
- 2.631×10^{-5}

4) What is the conclusion for the test?

- Reject the null hypothesis
- Fail to reject the null hypothesis

Coefficient of determination

Another quantity that shows how well a regression equation represents the data is the *coefficient of determination*. The ***coefficient of determination***, denoted by R^2, gives the ratio of the variance in the response variable explained by the predictor variable. Conceptually, the coefficient of determination is a measure of how closely the regression line follows the pattern of the data. The farther the actual data points are from the regression line, the less useful the line actually is in predicting the value of the response variable.

The coefficient of determination is obtained by the following equation as illustrated in the animation below.

$$R^2 = \frac{\sum \left(\hat{Y}_i - \bar{Y} \right)^2}{\sum \left(Y_i - \bar{Y} \right)^2}$$

where $\hat{Y}_i$ is the estimated value of Y_i and $\bar{Y}$ is the mean of the Y_is.

Since the coefficient of determination is the square of the correlation coefficient, $0 \leq R^2 \leq 1$.

PARTICIPATION ACTIVITY 5.3.3: Coefficient of determination.

Animation content:

undefined

Animation captions:

1. The coefficient of determination can be interpreted from the distance of the Y values of the points from the fitted values $\hat{Y}$ and the distance of the fitted values from the mean $\bar{Y}$.

2. The total deviation for each data point is the distance from the Y value to the mean $\bar{Y}$.
3. The explained deviation for each data point is the distance from the fitted line to the mean, $\hat{Y} - \bar{Y}$.
4. The unexplained deviation is the distance from the Y value to the fitted value, $Y - \hat{Y}$. Thus, the total deviation is the sum of the explained deviation and unexplained deviation.
5. Adding up the squares of the explained deviation for all points gives the explained variance. Adding up the squares of the total deviation for all points gives the total variance.
6. The coefficient of determination R^2 is the ratio between the explained variance and the total variance.

Python-Practice 5.3.1: Coefficient of determination.

Another way to find the regression coefficients is to use the `statsmodels` package. The code `ols('y ~ x', data=dataframe).fit()` where y is the dataframe column that represents the response variable and x is the dataframe column that represents the predictor variable. The output gives other quantities including the coefficient of determination.

```
# The necessary packages are imported
import pandas as pd
from statsmodels.formula.api import ols

# The ExamScores dataset is loaded
scores = pd.read_csv('http://data-analytics.zybooks.com/ExamScores.csv')

# Creates a linear regression model
results = ols('Exam4 ~ Exam1', data=scores).fit()

# Prints the results
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Exam4   R-squared:                       0.068
Model:                            OLS   Adj. R-squared:                  0.049
Method:                 Least Squares   F-statistic:                     3.518
Date:                Fri, 21 Jun 2019   Prob (F-statistic):             0.0668
Time:                        14:09:42   Log-Likelihood:                -173.00
No. Observations:                  50   AIC:                             350.0
Df Residuals:                      48   BIC:                             353.8
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     57.7627     10.052      5.746      0.000      37.552      77.973
Exam1          0.2266      0.121      1.876      0.067      -0.016       0.469
==============================================================================
Omnibus:                        3.859   Durbin-Watson:                   1.723
Prob(Omnibus):                  0.145   Jarque-Bera (JB):                2.809
Skew:                           0.428   Prob(JB):                        0.245
Kurtosis:                       3.784   Cond. No.                         753.
==============================================================================
```


The explained and unexplained variance can be obtained from the analysis of variance table created using the code below.

```
# The necessary packages are imported
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols

# The ExamScores dataset is loaded
scores = pd.read_csv('http://data-analytics.zybooks.com/ExamScores.csv')

# Creates a linear regression model
results = ols('Exam4 ~ Exam1', data=scores).fit()

# Creates an analysis of variance table
aov_table = sm.stats.anova_lm(results, typ=2)

# Prints the analysis of variance table
print(aov_table)
```

```
               sum_sq    df         F    PR(>F)
Exam1      217.166351   1.0  3.517655  0.066808
Residual  2963.333649  48.0       NaN       NaN
```

In the output, the explained variance is given by the sum of squares for Exam1, which is 217.166351, while the unexplained variance is given by the sum of squares for the Residual, which is 2963.333649. Using these values, the value for R^2 can also be computed as follows.

$$
\begin{aligned}
R^2 &= \frac{\text{explained variance}}{\text{total variance}} \\
&= \frac{217.166351}{217.166351 + 2963.333649} \\
&= 0.068
\end{aligned}
$$

Suppose a teacher wants to see whether Exam4 scores can be predicted using Exam1 scores. The value for R^2 means that only 6.8% of the variance in Exam4 scores can be explained by the variance in Exam1 scores. The low value for R^2 suggests that Exam1 is not a good predictor for how well a student would do in Exam4.

Run example

PARTICIPATION ACTIVITY 5.3.4: Coefficient of determination.

A researcher examines how well the amount of hemoglobin (Hb) in blood can predict packed cell volume (PCV). Use the analysis of variance table below to answer the following questions.

	sum_sq	df	F	PR(>F)
Hb	0.008725	1.0	37.321988	0.000053
Residual	0.002805	12.0	NaN	NaN

Run example

1) What is the explained variance? Type as: #.######

Check Show answer

2) What is the total variance? Type as: #.######

Check Show answer

3) What percent of the variance in PCV can be explained from the variance in Hb? Type as: ##.#%

Check Show answer

5.4 Interpreting SLR models

Interpreting regression parameters

The estimates for the population regression parameters, b_0 and b_1, can be interpreted as follows:

- b_0 represents the ***estimated simple linear regression Y-intercept***, which is the fitted value of Y when $X = 0$.
- b_1 represents the ***estimated simple linear regression slope***, which is the change in the fitted value of Y per unit change in X.

The Y-intercept, b_0, has a practical meaning for datasets in which the predictor variable X can take on a value of 0 or values near 0. The measurement unit of b_0 is the same as the measurement unit of the response Y variable. Ex: In an experiment where the response variable is blood pressure and the predictor variable is drug dose, b_0 has a practical meaning because a dose of 0 means that the patient did not take the drug. On the other hand, in a linear model where the house price is the response variable and the house size is the predictor variable, b_0 has no practical meaning because a house cannot have an area of 0.

The slope, b_1, always has a practical meaning. The measurement unit of b_1 is measurement unit of Y per measurement unit of X. Ex: In a linear model where house price (measured in dollars) is the response variable and house size (measured in square feet) is the predictor variable, b_1 is measured in dollars per square foot.

The animation below uses the Reaction dataset that records the reaction times (in milliseconds) of 6 adults after consuming alcohol (in standard drink units). Reaction time is the response variable and alcohol consumption is the predictor variable. The code and output below are used to obtain the sample linear regression equation in the animation.

```
import pandas as pd
import scipy.stats as st
df = pd.read_csv('http://data-analytics.zybooks.com/Reaction.csv')
print(st.linregress(df['Drinks'],df['Reaction']))
```

```
LinregressResult(slope=6.0000000000000009, intercept=3.9999999999999964,
 rvalue=0.97281665268828232, pvalue=0.0010983582017795293, stderr=0.71414284285428509)
```

Run example

PARTICIPATION ACTIVITY 5.4.1: Estimates for the slope and intercept regression parameters.

Animation captions:

1. The sample simple linear regression line represents the linear relationship between Y and X.

2. b_0 represents the Y-intercept of the line, or the fitted value of Y when $X = 0$.
3. The fitted value of Y increases from 28 to 34 and X increases from 4 to 5. The slope is the change in Y, denoted by ΔY divided by the change in X, denoted by ΔX.
4. The b_1 represents the slope of the line, which is 6.

PARTICIPATION ACTIVITY 5.4.2: Interpreting simple linear regression parameters.

Refer to the animation above.

1) What is the Y-intercept?
   - 4
   - 6

2) What is the measurement unit for the Y-intercept?
   - Standard drink units
   - Milliseconds

3) What is the measurement unit for the slope?
   - Standard drink units per millisecond
   - Milliseconds per standard drink unit

Interpreting residual standard error

A number of quantities use the residuals, $\varepsilon_i = Y_i - \hat{Y}_i$:

- The ***residual sum of squares*** is the sum of squared residuals for the sample, $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$. The residual sum of squares is typically denoted SSE because the residuals are estimated errors. The notation SSR is used for regression sum of squares, which is defined in a later section.
- The ***residual degrees of freedom*** is $n - p$, where p is the number of regression parameters. Ex: Simple linear regression has $p = 2$ regression parameters, so the residual degrees of freedom is $n - 2$. Residual degrees of freedom is often called error degrees of freedom.
- The ***residual mean square*** is the residual sum of squares divided by the residual degrees of freedom, $MSE = \frac{SSE}{n-p}$. Ex: For simple linear regression, $MSE = \frac{SSE}{n-2}$.

- The ***residual standard error*** is the square root of the residual mean square, $s = \sqrt{MSE}$. The residual standard error, s, estimates the standard deviation of the residuals. The measurement unit of the residual standard error is the same as the measurement unit of the response Y variable.

The residual sum of squares, residual degrees of freedom, and residual mean square are used in an Analysis of Variance table, which is often abbreviated as "ANOVA table." A later section shows how an ANOVA table is used in model assessment.

The residual standard error is a measure of the precision of a model prediction. The sample simple linear regression line can be used to predict a future value of Y for a fixed value of X. A relatively small residual standard error indicates that the actual future value of Y is likely to be relatively close to the predicted value. Therefore, less residual standard error is better.

PARTICIPATION ACTIVITY 5.4.3: Interpreting the residual standard error.

Use the table below to answer questions about the residual standard error of the linear model, $\hat{Y} = 4 + 6X$, for the Reaction dataset.

Alcohol consumption, X	0	1	3	6	6	8
Reaction time, Y	6	12	18	33	42	57
Fitted values, $\hat{Y}$	4	10	22	40	40	52

1) What are the residual degrees of freedom?
   - 4
   - 5

2) What is the residual sum of squares?
   - 20
   - 102

3) What is the residual mean square?
   - 20.4
   - 25.5

4) What is the residual standard error?
   - 5.05
   - 25.5

5) What is the measurement unit for the residual standard error in this example?

- Standard drink units
- Milliseconds

5.5 Testing SLR parameters

Population and sample regression lines

The sample regression parameter estimators, b_0 and b_1, are the values of the population regression parameters, β_0 and β_1, that minimize the sum of squared errors for the model $\mathrm{E}(Y) = \beta_0 + \beta_1 X$, where β_0 is the intercept parameter and β_1 is the slope parameter. The estimator, b_1, is an unbiased estimate of β_1, which means that under repeated sampling, the expected value of b_1 equals β_1. However, the value of b_1 for a particular sample will generally not be equal to β_1. The value of b_1 differs from β_1 because of sampling variation, as illustrated in the following animation.

PARTICIPATION ACTIVITY 5.5.1: Visualizing population and sample regression lines.

Animation captions:

1. Suppose no population linear relationship exists between Y and X. Therefore, the population regression line has a zero slope.
2. Sample data from such a population may have a linear relationship, which implies that the sample regression line has a non-zero slope.
3. However, a sample with a regression line that has a slope close to zero may exist.
4. Alternatively, a negative population linear relationship exists between Y and X.
5. Sample data from such a population will likely have a negative linear relationship, which implies that a sample regression line would have a slope that is far from zero.
6. In practice, the population regression line is not known, so inference is made on the population regression line based on the sample regression line.

Testing the linear regression slope parameter

The estimator b_1 is a single number used to estimate β_1. If β_1 were equal to 0, then no linear relationship would exist between Y and X. However, in such a circumstance, as illustrated in the

animation above, sampling uncertainty could lead to a non-zero estimator, b_1. To determine whether b_1 is sufficiently far from 0 to reject the possibility that $\beta_1 = 0$ in the population, the t-test can be used.

Procedure 5.5.1: t-test for the slope parameter.

1. Set the null hypothesis, $H_0: \beta_1 = 0$, versus the alternative hypothesis, $H_a: \beta_1 \neq 0$.
2. Use statistical software to find the t-statistic, which is b_1 divided by the standard error of b_1.
3. Use statistical software to find the p-value that corresponds to the t-statistic. The p-value is the probability of observing a t-statistic at least as far from 0 as the one observed, if the null hypothesis were true. The reference t-distribution has $n - p$ degrees of freedom, where n is the sample size and p is the number of regression parameters (typically the number of predictor variables plus one for the intercept).
4. Make a decision based on a previously selected significance level, typically 0.05:
    - If the p-value is less than the significance level, reject the null hypothesis, $H_0: \beta_1 = 0$, in favor of the alternative hypothesis, $H_a: \beta_1 \neq 0$. Conclude that β_1 is statistically significantly different from 0, which means that a statistically significant linear relationship exists between Y and X.
    - If the p-value is greater than or equal to the significance level, fail to reject the null hypothesis, $H_0: \beta_1 = 0$, in favor of the alternative hypothesis, $H_a: \beta_1 \neq 0$. Conclude that β_1 is not statistically significantly different from 0, which means that a statistically significant linear relationship does not exist between Y and X.

Python-Function 5.5.1: ols(), fit(), and summary().

The `ols()` function performs ordinary linear regression, and the `fit()` function fits the data to the regression line. These functions require the statsmodels.formula.api library to be imported. `ols()` takes two parameters. The first parameter is in the form 'Y ~ X', where Y is the response variable and X is the predictor variable. The second parameter is the dataset that contains the variables.

The `summary()` function returns a summary including estimates for the parameters, t-statistics, p-values, and confidence intervals.

All of the above functions require the statsmodels.formula.api library to be imported.

```
import pandas as pd
import statsmodels.formula.api as smf
scores = pd.read_csv('http://data-analytics.zybooks.com/ExamScores.csv')

model = smf.ols('Exam4 ~ Exam2', scores).fit()
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Exam4   R-squared:                       0.101
Model:                            OLS   Adj. R-squared:                  0.082
Method:                 Least Squares   F-statistic:                     5.405
Date:                Wed, 25 Oct 2017   Prob (F-statistic):             0.0244
Time:                        21:59:27   Log-Likelihood:                -172.10
No. Observations:                  50   AIC:                             348.2
Df Residuals:                      48   BIC:                             352.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     62.3017      6.204     10.042      0.000      49.828      74.776
Exam2          0.1788      0.077      2.325      0.024       0.024       0.333
==============================================================================
Omnibus:                        4.604   Durbin-Watson:                   1.676
Prob(Omnibus):                  0.100   Jarque-Bera (JB):                3.507
Skew:                           0.509   Prob(JB):                        0.173
Kurtosis:                       3.805   Cond. No.                         459.
==============================================================================
```


Run example

Example 5.5.1: Testing a simple linear regression slope.

The ExamScores dataset is a record of 4 exam scores for 50 students. $Y = \text{Exam4}$ is the response variable and $X = \text{Exam2}$ is the predictor variable. The teacher believes that a linear relationship exists between Exam4 scores and Exam2 scores. Does sufficient evidence exist to support the teacher's claim at the $\alpha = 0.05$ significance level? Use the partial output below.

```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     62.3017      6.204     10.042      0.000      49.828      74.776
Exam2          0.1788      0.077      2.325      0.024       0.024       0.333
```


Solution

The null hypothesis is that no linear relationship exists between Exam4 scores and Exam2 scores. The alternative hypothesis is that a linear relationship exists between scores from both exams. Mathematically,

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

The t-statistic is the estimate for the slope parameter divided by the standard error for the estimate. That is,

$$t = \frac{b_1}{SE_{b_1}} = \frac{0.1788}{0.077} = 2.325$$

The p-value is $P(t \leq -2.325 \text{ or } t \geq 2.325) = 0.024$.

Since the p-value is less than the significance level $(0.024 < 0.05)$, the null hypothesis is rejected. Thus, statistically significant evidence exists to support the teacher's claim that exam 4 scores and exam 2 scores are linearly dependent.

PARTICIPATION ACTIVITY 5.5.2: Testing a simple linear regression slope.

The Disease dataset is a record of disease marker levels and treatment times for 12 patients. $Y = \text{Disease}$ is the response variable that measures the level of the disease marker and $X = \text{Time}$ is the predictor variable that gives the treatment time in months for each patient. A researcher claims that a linear relationship exists between disease marker levels and treatment times.

```
import pandas as pd
import statsmodels.formula.api as smf
df = pd.read_csv('http://data-analytics.zybooks.com/Disease.csv')

model = smf.ols('Disease ~ Time', df).fit()
print(model.summary())

# warning appears because the sample size is less than 20
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Disease   R-squared:                       0.708
Model:                            OLS   Adj. R-squared:                  0.679
Method:                 Least Squares   F-statistic:                     24.24
Date:                Sat, 28 Oct 2017   Prob (F-statistic):           0.000602
Time:                        06:53:06   Log-Likelihood:                -31.415
No. Observations:                  12   AIC:                             66.83
Df Residuals:                      10   BIC:                             67.80
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     21.0000      2.286      9.187      0.000      15.907      26.093
Time          -2.0000      0.406     -4.924      0.001      -2.905      -1.095
==============================================================================
Omnibus:                        1.751   Durbin-Watson:                   3.500
Prob(Omnibus):                  0.417   Jarque-Bera (JB):                0.833
Skew:                           0.027   Prob(JB):                        0.659
Kurtosis:                       1.711   Cond. No.                         12.6
==============================================================================
```

Run example

1) What is the null hypothesis for testing whether a linear relationship exists between disease level and treatment time?

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

2) What is the alternative hypothesis for testing whether a linear relationship exists between disease level and treatment time?

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

3) Given that the p-value is 0.001, what is the decision based on a significance level of $\alpha = 0.05$?

- Reject H_0
- Fail to reject H_0

4) What is the conclusion about whether a linear relationship exists between disease marker level and treatment time?

- A statistically significant linear relationship exists
- A statistically significant linear relationship does not exist

The slope test above is a two-tailed test, which means that the null hypothesis can be rejected for t-statistics that are either negative and far from zero, or positive and far from zero. Occasionally, a one-tailed test is used when a specific reason exists to test for a positive (or negative) linear relationship.

PARTICIPATION ACTIVITY 5.5.3: One-tailed test for the slope.

Another researcher claims that as treatment times increase, disease level markers decrease. Does statistically significant evidence exist to support the researcher's claim? Use the linear regression model for the Disease dataset below.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Disease   R-squared:                       0.708
Model:                            OLS   Adj. R-squared:                  0.679
Method:                 Least Squares   F-statistic:                     24.24
Date:                Sat, 28 Oct 2017   Prob (F-statistic):           0.000602
Time:                        06:53:06   Log-Likelihood:                -31.415
No. Observations:                  12   AIC:                             66.83
Df Residuals:                      10   BIC:                             67.80
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     21.0000      2.286      9.187      0.000      15.907      26.093
Time          -2.0000      0.406     -4.924      0.001      -2.905      -1.095
==============================================================================
Omnibus:                        1.751   Durbin-Watson:                   3.500
Prob(Omnibus):                  0.417   Jarque-Bera (JB):                0.833
Skew:                           0.027   Prob(JB):                        0.659
Kurtosis:                       1.711   Cond. No.                         12.6
==============================================================================
```


1) What is the alternative hypothesis for a one-tailed test for a negative linear relationship between disease level and treatment time?

   - $H_a: \beta_1 < 0$
   - $H_a: \beta_1 > 0$

2) The t-statistic for the two-tailed slope test for this example is $t = -\frac{2}{0.4062} = -4.924$. What is the t-statistic for a one-tailed test for a negative linear relationship between disease level and treatment time?

   - $t = -\frac{4.924}{2} = -2.462$
   - $t = -4.924$
   - $t = -4.924 \cdot 2 = -9.848$

3) What is the p-value for a one-tailed test for a negative linear relationship between treatment time and disease level?

   - $\frac{0.001}{2} = 0.0005$
   - 0.001
   - 0.000

4) What is the conclusion about whether a negative linear relationship exists between disease level and treatment time, based on a significance level of 0.05?

- A statistically significant negative linear relationship exists.
- A statistically significant negative linear relationship does not exist.

5) What would be the conclusion about whether a *positive* linear relationship exists between disease level and treatment time, based on a significance level of 0.05? Hint: The p-value for testing for a positive relationship is the area under the probability curve to the right of the test statistic, which is $1 - 0.0005 = 0.9995$ in this case.

- A statistically significant positive linear relationship exists.
- A statistically significant positive linear relationship does not exist.

Testing the intercept parameter

Rarely, a reason may exist to test the possibility that the intercept parameter, β_0, is zero in the population. Testing $\beta_0 = 0$ is only warranted if the response variable Y could possibly be zero when the predictor variable X is zero. Ex: In the treatment time and disease example above, the disease level cannot possibly be zero when treatment time is zero. Thus, testing $\beta_0 = 0$ is not warranted for this example. In examples where testing $\beta_0 = 0$ is warranted, the intercept test is performed similarly as the slope test above.

PARTICIPATION ACTIVITY 5.5.4: Testing a simple linear regression intercept.

In a particular manufacturing process, X is the predictor variable that represents the number of items produced during a 1-hour time period and Y is the response variable that represents

the number of rejected items during that time. The t-statistic for the simple linear regression intercept, β_0, fit to these data is $t = \frac{1}{1.830} = 0.546$ with a p-value of 0.300.

1) Is testing $\beta_0 = 0$ warranted for this example?
   - No
   - Yes

2) What are the hypotheses for testing whether zero rejected items could be present on average when zero items were produced?
   - $H_0: \beta_0 = 0$ versus $H_a: \beta_0 \neq 0$
   - $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$

3) Given that the p-value is 0.300, what is the decision based on a significance level of 0.05?
   - Reject H_0
   - Fail to reject H_0

4) What is the conclusion about whether the intercept is not statistically significantly different from zero, in other words, whether zero rejected items could be present on average when zero items are produced?
   - The intercept is statistically significantly different from zero.
   - The intercept is not statistically significantly different from zero.

Confidence intervals for regression parameters

The slope estimator, b_1, provides a single number to estimate β_1. To quantify sampling uncertainty about a single number used to estimate β_1, one can calculate an interval around the number. A ***confidence interval for the slope*** is an interval around b_1 that quantifies sampling uncertainty when b_1 is used to estimate β_1. The confidence interval is given by

$$[b_1 - t^*(SE), b_1 + t^*(SE)]$$

where SE is the standard error, and t^* depends on the degrees of freedom and the confidence level of interest, and can be found from a t-distribution table. Although far less common, the confidence interval for the intercept can also be calculated.

Example 5.5.2: Finding a confidence interval for the slope.

The ExamScores dataset is a record of 4 exam scores for 50 students. $Y = \text{Exam4}$ is the response variable and $X = \text{Exam2}$ is the predictor variable. The teacher believes that a linear relationship exists between Exam4 scores and Exam2 scores. Find a 99% confidence level for the slope. Use the partial output below.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	62.3017	6.204	10.042	0.000	49.828	74.776
Exam2	0.1788	0.077	2.325	0.024	0.024	0.333

Solution

The confidence interval in the output is a 95% confidence interval. The regression line has two parameters, the slope and intercept parameters, so $p = 2$. At the 99% confidence level with degrees of freedom $n - p = 50 - 2 = 48$, the critical value is $t^* = 2.682$. Thus, the 99% confidence interval for the slope is,

$$[0.1788 - (2.682)(0.077), 0.1788 + (2.682)(0.077)] = [-0.028, 0.385]$$

PARTICIPATION ACTIVITY 5.5.5: Finding a confidence interval for the slope.

Use the linear regression model for the Disease dataset below to answer the following questions.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Disease   R-squared:                       0.708
Model:                            OLS   Adj. R-squared:                  0.679
Method:                 Least Squares   F-statistic:                     24.24
Date:                Sat, 28 Oct 2017   Prob (F-statistic):           0.000602
Time:                        06:53:06   Log-Likelihood:                -31.415
No. Observations:                  12   AIC:                             66.83
Df Residuals:                      10   BIC:                             67.80
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     21.0000      2.286      9.187      0.000      15.907      26.093
Time          -2.0000      0.406     -4.924      0.001      -2.905      -1.095
==============================================================================
Omnibus:                        1.751   Durbin-Watson:                   3.500
Prob(Omnibus):                  0.417   Jarque-Bera (JB):                0.833
Skew:                           0.027   Prob(JB):                        0.659
Kurtosis:                       1.711   Cond. No.                         12.6
==============================================================================
```


1) What is the interpretation of the 95% confidence interval for the slope?

- One is 95% confident that treatment time decreases by between 1.095 and 2.905 months for each one unit increase in disease level.
- One is 95% confident that disease level decreases by between 1.095 and 2.905 units for each additional month of treatment.

2) What is the most likely value for the slope?

- -2.905
- -2
- -1.095

3) Is the population slope likely to be 0?

- No
- Yes

©zyBooks 01/31/23 17:56 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

Analysis of Variance F-test

Since the population regression line is given by $E(Y) = \beta_0 + \beta_1 X$, determining whether an association exists between X and Y is equivalent to determining whether $\beta_1 \neq 0$. The association between two variables can be tested using the ANOVA F-test. The ***simple linear regression ANOVA F-test*** is a method for testing $H_0: \beta_1 = 0$ versus $H_a: \beta_1 \neq 0$. For simple linear regression, the ANOVA F-test is equivalent to the slope t-test, which is covered in an earlier section. However, when testing for association involving multiple predictor variables, the ANOVA F-test and slope t-tests are not equivalent.

The ANOVA table includes the following quantities:

- The ***residual sum of squares*** is $SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$, where $\hat{Y}_i$ is the ith fitted response variable.
- The ***residual degrees of freedom*** is $n - p$.
- The ***residual mean square*** is $MSE = \frac{SSE}{n-p}$.
- The ***regression sum of squares*** is $SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$, where $\hat{Y}_i$ is the ith fitted response variable and $\bar{Y}$ is the sample mean of the response values.
- The ***regression degrees of freedom*** is $p - 1$, where p is the number of regression parameters. Ex: Simple linear regression has $p = 2$ regression parameters, so the regression degrees of freedom is 1.
- The ***regression mean square*** is the regression sum of squares divided by the regression degrees of freedom, $MSR = \frac{SSR}{p-1}$. Ex: For simple linear regression, $MSR = \frac{SSR}{1} = SSR$.
- The ***total sum of squares*** is $SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$. Note that $SSTO = SSR + SSE$.
- The ***total degrees of freedom*** is $n - 1$. Here n is the sample size. Note that $n - 1 = (p - 1) + (n - p)$, so total degrees of freedom = regression degrees of freedom + residual degrees of freedom.

The quantities above can be summarized using an ANOVA table.

PARTICIPATION ACTIVITY 5.5.6: Composition of the ANOVA table.

Animation content:

undefined

Animation captions:

1. Total sum of squares is partitioned into regression and residual sums of squares. $SSTO = SSR + SSE$.
2. Total degrees of freedom is partitioned into regression and residual degrees of freedom. n − 1 = (p − 1) + (n − p).
3. Regression mean square is the sum of squares divided by the degrees of freedom. MSR = SSR / (p − 1).

4. Residual mean square is the sum of squares divided by the degrees of freedom. MSE = SSE / (n – p).
5. ANOVA F-statistic is the ratio of the regression mean square to the residual mean square. F = MSR / MSE.
6. The p-value is the probability of observing an F-statistic at least as large as the one observed, if the null hypothesis were true.

Procedure 5.5.2: Simple linear regression ANOVA F-test.

The steps for the ANOVA F-test are as follows:

- Set the null hypothesis, $H_0: \beta_1 = 0$, versus the alternative hypothesis, $H_a: \beta_1 \neq 0$.
- Use statistical software to find the F-statistic, which is the regression mean square divided by the residual mean square. The ANOVA F-statistic is the same as the square of the slope t-statistic. $F = \frac{MSR}{MSE}$
- Use statistical software to find the p-value that corresponds to the F-statistic. The p-value is the probability of observing an F-statistic at least as far from 0 as the one observed, if the null hypothesis were true. The reference F-distribution has p-1 numerator degrees of freedom and n-p denominator degrees of freedom. The ANOVA F-test has the same p-value as the slope t-test.
- Make a decision based on a previously selected significance level, typically 0.05:
  - If the p-value is less than the significance level, reject the null hypothesis, $H_0: \beta_1 = 0$, in favor of the alternative hypothesis, $H_a: \beta_1 \neq 0$. Conclude that β_1 is significantly different from 0, which means that a statistically significant linear relationship exists between Y and X.
  - If the p-value is greater than or equal to the significance level, fail to reject the null hypothesis, $H_0: \beta_1 = 0$, in favor of the alternative hypothesis, $H_a: \beta_1 \neq 0$. Conclude that β_1 is not significantly different from 0, which means that a statistically significant linear relationship does not exist between Y and X.

Python-Practice 5.5.1: Using ANOVA to test the correlation between two variables.

The `ols()` function takes two parameters, 'Y ~ X1' where Y is the response variable and X1 is a predictor variable, and a data frame. The sm.stats.anova_lm command takes in a linear model and an ANOVA type as parameters and outputs an ANOVA table. The different types

of ANOVA are beyond the scope of this material. To give the correct output, the parameter for the ANOVA type, denoted as typ, is set to 2.

In the code below, Exam4 is the response variable and Exam1 is the predictor variable.

```
from statsmodels.formula.api import ols
import statsmodels.api as sm
import pandas as pd
df=pd.read_csv("http://data-analytics.zybooks.com/ExamScores.csv")
mod = ols('Exam4 ~ Exam1',df).fit()
print(sm.stats.anova_lm(mod, typ=2))
```



```
              sum_sq    df         F    PR(>F)
Exam1      217.166351   1.0  3.517655  0.066808
Residual  2963.333649  48.0       NaN       NaN
```

Run example

Example 5.5.3: Using the ANOVA table and performing the F-test for simple linear regression.

The teacher of a statistics class with 50 students believes that scores in first exam predict how well students do in fourth exam. The ANOVA table where Exam4 is the response variable and Exam1 is the predictor variable is given below.

```
              sum_sq    df         F    PR(>F)
Exam1      217.166351   1.0  3.517655  0.066808
Residual  2963.333649  48.0       NaN       NaN
```

a. Find the regression mean square MSR and the residual mean square MSE.

b. Perform an ANOVA F-test to determine whether sufficient evidence exists to support the teacher's belief at the \alpha = 0.10 significance level.

Solution

a. To find MSR, the regression sum of squares is divided by the regression degrees of freedom (p-1).

MSR = \displaystyle\frac{SSR}{p - 1} = \displaystyle\frac{217.166}{1} = 217.166

To find MSE, the residual sum of squares is divided by the residual degrees of freedom (n-p).

MSE = \displaystyle\frac{SSE}{n - p} = \displaystyle\frac{2963.334}{48} = 61.736

b. The null hypothesis is that no association exists between Exam4 and Exam1. The alternative hypothesis is that an association exists between scores in the first exam and the fourth exam. Mathematically, the null hypothesis is that the slope parameter is zero and the alternative hypothesis is that the slope parameter is not equal to zero as given below.

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

From the ANOVA table,

$F = 3.518$

The p-value is

$p\text{-value} = P(F \geq 3.518) \approx 0.067$

Since the p-value is less than the significance level $\alpha = 0.10$, sufficient evidence exists to support the hypothesis that an association exists between Exam4 and Exam1.

PARTICIPATION ACTIVITY 5.5.7: Using the ANOVA table.

The height and grade point average of 35 randomly selected students from a high school are recorded. Use the ANOVA table below to match each quantity to the correct value. The grade point average, denoted by gpa, is the response variable and the height is the predictor variable.

```
from statsmodels.formula.api import ols
import statsmodels.api as sm
import pandas as pd
df=pd.read_csv("http://data-analytics.zybooks.com/gpa.csv")
mod = ols('gpa ~ height',df).fit()
print(sm.stats.anova_lm(mod, typ=2))
```

```
            sum_sq    df         F    PR(>F)
height    0.027590   1.0  0.093809  0.761314
Residual  9.705507  33.0       NaN       NaN
```

Run example

Select the definition that matches each term

1) 1

- Regression degrees of freedom.
- ANOVA p-value.
- ANOVA F-statistic.
- Residual mean square.

- Regression mean square.
- Residual degrees of freedom.

2) 33

- Regression degrees of freedom.
- ANOVA p-value.
- ANOVA F-statistic.
- Residual mean square.
- Regression mean square.
- Residual degrees of freedom.

3) 0.027590

- Regression degrees of freedom.
- ANOVA p-value.
- ANOVA F-statistic.
- Residual mean square.
- Regression mean square.
- Residual degrees of freedom.

4) 0.29411

- Regression degrees of freedom.
- ANOVA p-value.
- ANOVA F-statistic.
- Residual mean square.
- Regression mean square.
- Residual degrees of freedom.

5) 0.093809

- Regression degrees of freedom.
- ANOVA p-value.
- ANOVA F-statistic.
- Residual mean square.
- Regression mean square.
- Residual degrees of freedom.

6) 0.761314

- Regression degrees of freedom.
- ANOVA p-value.
- ANOVA F-statistic.
- Residual mean square.
- Regression mean square.
- Residual degrees of freedom.

Reset

PARTICIPATION ACTIVITY 5.5.8: Performing the ANOVA F-test for simple linear regression.

A student in the same high school claims that taller students tend to have lower gpas. At the \alpha = 0.05 significance level, does statistically significant evidence exists to support the student's claim? Use the data from the ANOVA table below.

	sum_sq	df	F	PR(>F)
height	0.027590	1.0	0.093809	0.761314
Residual	9.705507	33.0	NaN	NaN

1) What is the null hypothesis H_0?

- \beta_0 = 0
- F =0.094
- \beta_1 = 0

2) What is the alternative hypothesis H_a?

- \beta_1 \neq 0
- \beta_1 = 0.761
- \beta_0 \neq 0

3) What is the result of the F-test?

- Reject the null hypothesis
- Fail to reject the null hypothesis

4) What is the conclusion of the F-test within the context of the problem?

- No statistically significant evidence exists to support the claim that tallness causes lower gpas.
- No statistically significant evidence exist to support the claim that height and gpa are associated.

The coefficient of determination

The ***coefficient of determination***, denoted by R^2 is another measure of correlation. The coefficient of determination is useful because the quantity measures the proportion of total variation in the response variable, Y, that is accounted for by the linear regression model. Intuitively, the value of R^2 can be viewed as a quantitative way of measuring certainty when making predictions from a model. Although R^2 can easily be calculated by squaring the Pearson correlation coefficient, R^2 can also be calculated from the values in the ANOVA table: $R^2=\frac{SSTO-SSE}{SSTO}=\frac{SSR}{SSTO}$.

$SSTO$ and SSE measure different quantities:

- The best prediction of Y if one ignores X is the sample mean of Y, $\bar{Y}$. Thus, the total sum of squares, $SSTO=\sum_{i=1}^n (Y_i-\bar{Y})^2$ measures the variation in Y ignoring X.
- The best prediction of Y based on the simple linear regression model is $\hat{Y}$. Thus, the residual sum of squares, $SSE=\sum_{i=1}^n (Y_i-\hat{Y})^2$, measures the variation remaining in Y after using X to predict Y.

The value of R^2 is typically expressed as a percentage between 0% and 100%:

- If a strong linear relationship exists between Y and X, using X to predict Y will leave little variation remaining in Y. Then SSE will be small and R^2 will be high (generally greater than 90%).
- Conversely, if a linear relationship does not exist between Y and X, using X to predict Y will leave a lot of variation remaining in Y. Then SSE will be close to $SSTO$ and R^2 will be low (close to 0%).

Example 5.5.4: Finding and interpreting the coefficient of determination.

Consider the data involving the statistics class of 50 students where Exam1 is the predictor variable and Exam4 is the response variable.

Find and interpret the quantities involved in calculating the coefficient of determination. Use the data from the ANOVA table below.

	sum_sq	df	F	PR(>F)
Exam1	217.166351	1.0	3.517655	0.066808
Residual	2963.333649	48.0	NaN	NaN

Solution

SSR = 217.166351 measures the variation between fitted values and the sample mean of Exam4 scores.

SSE = 2963.333649 measures the variation remaining in Exam4 after using Exam1 to predict Exam4 scores.

The total sum of squares SSTO is calculated by adding SSR and SSE. SSTO represents the variation in Exam4 scores ignoring Exam1 scores.

SSTO = SSR + SSE = 217.166351 + 2963.333649 = 3180.5

To calculate R^2, SSR is divided by SSTO.

R^2 = \displaystyle\frac{SSR}{SSTO} = \displaystyle\frac{217.166351}{3180.5} \approx 0.068

Thus, approximately 6.8\% of the total variation in Exam4 scores is accounted for by the linear regression model with Exam1 as a predictor.

PARTICIPATION ACTIVITY 5.5.9: Finding and Interpreting the coefficient of determination.

Consider the data involving 35 students in a high school where height is the predictor variable and gpa is the response variable. Use the ANOVA table below.

	sum_sq	df	F	PR(>F)
height	0.027590	1.0	0.093809	0.761314
Residual	9.705507	33.0	NaN	NaN

1) What is the variation in gpa ignoring height?

- 0.027590
- 9.705507
- 9.733097

2) What is the variation remaining in gpa after using height to predict gpa?

- 0.027590
- 9.705507
- 9.733097

3) What is the value of R^2?

- -0.3\%
- 99.7\%
- 0.3\%

4) What proportion of the total variation in gpa is accounted for by the linear regression model with height as the predictor?

- 99.7\%
- 0.3\%

©zyBooks 01/31/23 17:56 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3