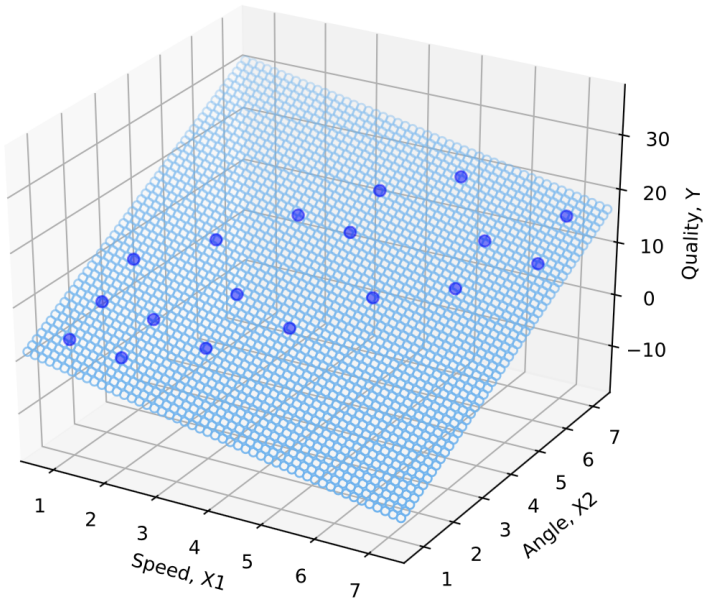
6.1 Introduction to multiple regression

Introduction to multiple regression

A ***multiple regression*** is a way to model the linear relationship between one quantitative response variable and more than one predictor variable. Ex: A multiple regression model can be used to model the linear relationship between the quality of a car component and manufacturing inputs, such as machine speed and cutting angle. $Y = \text{Quality}$ is the response variable, while $X_1 = \text{Speed}$ and $X_2 = \text{Angle}$ are predictor variables represented in the figure below. The ***response variable*** is the variable being modeled or predicted, while the ***predictor variables*** are the variables used to predict the response.

Figure 6.1.1: A 3D scatterplot and plane representing a multiple regression model for car quality with two predictor variables $X_1 =$ speed and $X_2 =$ angle.



A multiple regression model has two parts. The first part is the population linear regression function, which represents the expected value of Y given a particular set of values for $X_1, X_2, \ldots, X_n$. The ***population linear regression function*** is $\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$, where $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ are regression parameters. The second part is the regression error term, ϵ, which represents the difference between the actual value of Y and the expected value of Y given a particular set of values for $X_1, X_2, \ldots, X_n$.

The ***multiple regression model*** is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$, which is the sum of the population regression function and the regression error term.

Example 6.1.1: Elements of the multiple regression model.

The body fat dataset gives the amount of body fat (%), the triceps skinfold circumference (mm), the midarm circumference (cm), and the thigh circumference (cm) of 20 healthy women aged $25 - 34$ years. A researcher wishes to use multiple regression to understand how body fat depends on these body measurements.

a. What is the response variable?
b. What are the predictor variables?
c. What is the expected value of the response variable?
d. What is the multiple regression model?

Solution

a. Since the researcher is interested in predicting body fat based on the three body measurements, the response variable is body fat.
b. Since the three body measurements are used to predict body fat, the predictor variables are the three variables triceps skinfold circumference, the midarm circumference, and thigh circumference.
c. Let Y be body fat, X_1 be triceps skinfold circumference, X_2 be midarm circumference, and X_3 be thigh circumference. Then the expected value of Y is $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$, where β_0, β_1, β_2, and β_3 are the population regression parameters.
d. The multiple regression model is the population regression function plus the regression error term, or $Y = E(Y) + \epsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$.

PARTICIPATION ACTIVITY 6.1.1: Identifying elements of a multiple regression model.

Suppose a car buyer wants to evaluate the quality of a vehicle based on speed and steering angle based on a cars dataset compiled from expert car reviews.

1) What type of variable is the quality of the vehicle?
   - ○ Response variable
   - ○ Predictor variable
2) What type of variable is steering angle?

- Response variable
- Predictor variable

3) How many response variables can be analyzed in one multiple regression model?

- One
- Multiple

4) How many predictor variables can be analyzed in one multiple regression model?

- One
- Multiple

5) What is the multiple regression model?

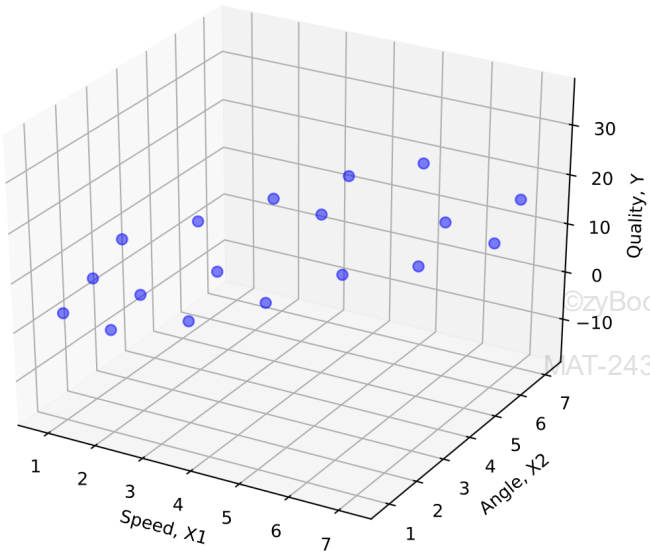
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
- $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Python-Practice 6.1.1: 3D scatterplot.

The mplot3d toolkit of matplotlib allows plotting of 3D objects such as scatterplots and 2D projections.

```
from mpl_toolkits.mplot3d import Axes3D
import matplotlib.pyplot as plt
import pandas as pd

df = pd.read_csv("http://data-analytics.zybooks.com/Cars.csv")

fig = plt.figure()

#mplot3d is needed for projection='3d'
ax = fig.add_subplot(111, projection='3d')

ax.scatter(df['Speed'], df['Quality'], df['Angle'], c='b', marker='o')

ax.set_xlabel('Speed, X1')
ax.set_ylabel('Angle, X2')
ax.set_zlabel('Quality, Y')

plt.show()
```


The 3D scatterplot output is shown below.



Run example

Estimating the multiple regression model

Once a population linear regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$ is proposed, the next step is to estimate the model using sample data to find a sample multiple regression function. The most common method to estimate the model is called "least squares." Intuitively, the least squares method finds the values of the regression parameters, $\beta_0, \beta_1, \ldots, \beta_n$ that minimize the sum of squared regression errors. The ***sample multiple regression function*** is $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_n X_n$, where $\hat{Y}$ is the fitted response value and $b_1, b_2, \ldots, b_n$ are the estimates for $\beta_0, \beta_1, \ldots, \beta_n$ that minimize the sum of squared errors. The "hat" notation in $\hat{Y}$ is a statistical convention that denotes a sample estimate.

Python-Practice 6.1.2: Multiple regression using Python.

The body fat dataset gives the amount of body fat, the triceps skinfold circumference, the midarm circumference, and the thigh circumference of 20 research subjects. Suppose a researcher wishes to use multiple regression to model the linear relationship between the amount of body fat for an individual and these body measurements. Let Y = body fat be the response variable and X_1 = triceps skinfold circumference, X_2 = midarm circumference, and X_3 = thigh circumference be the three predictor variables.

Similar to simple linear regression, `ols()`, `fit()`, and `summary()` are used to perform multiple regression, fit the data to the regression line, and display a summary. All of these functions require the `statsmodels.formula.api` module to be imported.

To perform multiple regression, the predictor variables are joined with `+` in the `ols()` function. Finally, the `fit()` method is used.

```
import pandas as pd
import statsmodels.formula.api as sms

fat = pd.read_csv('https://static-resources.zybooks.com/static/fat.csv')

# Response variable
Y = fat['body_fat_percent']

# Generates the linear regression model
# Multiple predictor variables are joined with +
model = sms.ols('Y ~ triceps_skinfold_thickness_mm + midarm_circumference_cm +
thigh_circumference_cm', data = fat).fit()

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



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.801
Model:                            OLS   Adj. R-squared:                  0.764
Method:                 Least Squares   F-statistic:                     21.52
Date:                Mon, 15 Jul 2019   Prob (F-statistic):           7.34e-06
Time:                        18:46:57   Log-Likelihood:                -44.312
No. Observations:                  20   AIC:                             96.62
Df Residuals:                      16   BIC:                             100.6
Df Model:                           3
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                      117.0847     99.782      1.173      0.258     -94.445     328.614
triceps_skinfold_thickness_mm    4.3341      3.016      1.437      0.170      -2.059      10.727
midarm_circumference_cm         -2.1861      1.595     -1.370      0.190      -5.568       1.196
thigh_circumference_cm          -2.8568      2.582     -1.106      0.285      -8.330       2.617
==============================================================================
Omnibus:                        1.200   Durbin-Watson:                   2.243
Prob(Omnibus):                  0.549   Jarque-Bera (JB):                0.830
Skew:                          -0.085   Prob(JB):                        0.660
Kurtosis:                       2.016   Cond. No.                     1.15e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.15e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The values in the `coef` column in the output correspond to the β population regression parameters of the multiple regression model. The `Intercept coef` is β_0. The subsequent entries in the `coef` column correspond to β coefficients of each predictor variable in the order entered in the `ols()` function. Thus,

$$\hat{Y} = 117.0847 + 4.3341X_1 - 2.1861X_2 - 2.8568X_3 .$$

Run example

PARTICIPATION ACTIVITY 6.1.2: Estimating the multiple regression function.

Multiple regression was performed on the cars dataset with Y = quality as the response variable and X_1 = speed and X_2 = angle as the predictor variables. The following output was obtained.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.978
Model:                            OLS   Adj. R-squared:                  0.975
Method:                 Least Squares   F-statistic:                     332.2
Date:                Mon, 15 Jul 2019   Prob (F-statistic):           3.80e-13
Time:                        20:48:21   Log-Likelihood:                -21.142
No. Observations:                  18   AIC:                             48.28
Df Residuals:                      15   BIC:                             50.95
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.5382      0.473      1.137      0.273      -0.471       1.547
Speed         -1.9046      0.176    -10.834      0.000      -2.279      -1.530
Angle          4.0280      0.178     22.574      0.000       3.648       4.408
==============================================================================
Omnibus:                        4.358   Durbin-Watson:                   2.121
Prob(Omnibus):                  0.113   Jarque-Bera (JB):                1.414
Skew:                           0.082   Prob(JB):                        0.493
Kurtosis:                       1.637   Cond. No.                         14.4
==============================================================================
```


Run example

1) What is the sample multiple regression function for this dataset?

- $\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- $\hat{Y} = 0.5382 - 1.9046 X_1 + 4.028 X_2$
- $\hat{Y} = 0.5382 - 1.9046 X_1 + 4.028 X_2 + \epsilon$

2) What is the sample linear regression function for a model that only includes X_1 and excludes X_2?

- $\hat{Y} = 0.5382 - 1.9046 X_1 + 4.028 X_2$
- $\hat{Y} = 0.5382 - 1.9046 X_1 + X_2$
- Unknown

3) What is the predicted quality of a car similar to those in the sample with speed 5 and steering angle 6?

- 14.645
- 15.183
- $15.183 + \epsilon$

Fitted values and residuals

A ***multiple regression fitted value***, $\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + \cdots + b_nX_{ni}$, is the predicted value of Y for the ith sample values of $X_1, X_2, X_3, \ldots$ based on the sample multiple regression function. Ex: The first sample observation in the body fat example has predictor values $X_1 = 19.5, X_2 = 29.1, X_3 = 43.1$. The first regression fitted value based on the sample multiple regression function $\hat{Y} = 117.085 + 4.334X_1 - 2.186X_2 - 2.857X_3$ is $\hat{Y}_1 = 117.085 + 4.334(19.5) - 2.186(29.1) - 2.857(43.1) = 14.8$.

A ***multiple regression residual***, $e_i = Y_i - \hat{Y}_i$, is the ith estimated regression error based on the sample multiple regression function. Ex: The first sample observation in the body fat example has response $Y = 11.9$. The first regression residual is $e_1 = Y_1 - \hat{Y}_1 = 11.9 - 14.8 = -2.9$.

Python-Practice 6.1.3: Multiple regression residuals.

The model object generated by `statsmodels` also contains the fitted values and residuals for each sample. To display the fitted values, the `fittedvalues` property of the model is called.

```
import pandas as pd
import statsmodels.formula.api as sms

fat = pd.read_csv('https://static-resources.zybooks.com/static/fat.csv')

# Response variable
Y = fat['body_fat_percent']

# Generates the linear regression model
# Multiple predictor variables are joined with +
model = sms.ols('Y ~ triceps_skinfold_thickness_mm + midarm_circumference_cm +
thigh_circumference_cm', data = fat).fit()

# Prints a list of the fitted values for each sample
print(model.fittedvalues)
```

```
0     14.854990
1     20.218841
2     20.986682
3     23.127320
4     11.757607
5     22.243718
6     25.714317
7     22.270641
8     19.594818
9     20.548382
10    24.595555
11    24.992309
12    15.009401
13    13.672305
14    11.811948
15    23.727468
16    22.973604
17    26.785902
18    18.526280
19    20.487912
dtype: float64
```


To display the residuals, the **resid** property of the model is called.

```
# Prints a list of the residuals for each sample
print(model.resid)
```

```
0    -2.954990
1     2.581159
2    -2.286682
3    -3.027320
4     1.142393
5    -0.543718
6     1.385683
7     3.129359
8     1.705182
9    -1.248382
10    0.804445
11    2.207691
12   -3.309401
13    4.127695
14    0.988052
15    0.172532
16   -0.373604
17   -1.385902
18   -3.726280
19    0.612088
dtype: float64
```

Run example

PARTICIPATION ACTIVITY 6.1.3: Calculating fitted values and residuals.

Multiple regression was performed on the cars dataset with quality as the response variable and speed and angle as the predictor variables. The following fitted values and residuals were obtained.

```
import pandas as pd
import statsmodels.formula.api as sms

cars = pd.read_csv('http://data-analytics.zybooks.com/Cars.csv')

# Response variable
Y = cars['Quality']

# Generates the linear regression model
# Multiple predictor variables are joined with +
model = sms.ols('Y ~ Speed + Angle', cars).fit()

print(model.fittedvalues)
```



```
0      0.756997
1      2.661578
2      2.880407
3      5.003817
4      4.784987
5      7.127226
6      6.689567
7      6.908397
8      9.250636
9     10.717557
10    13.059796
11    11.374046
12    12.840967
13    13.278626
14    14.964377
15    17.087786
16    19.211196
17    15.402036
dtype: float64
```

```
print(model.resid)
```

```
0     1.243003
1     0.338422
2    -0.880407
3    -1.003817
4     0.215013
5     0.872774
6    -0.689567
7     1.091603
8    -1.250636
9    -0.717557
10    0.940204
11   -0.374046
12   -0.840967
13   -0.278626
14    0.035623
15    0.912214
16   -0.211196
17    0.597964
dtype: float64
```


Run example

1) Based on the output, what is the second regression fitted value?
Type as: #.######

Check Show answer

2) Based on the output, what is the fifth regression residual? Type as: #.######

Check Show answer

6.2 Multiple regression assumptions and diagnostics

Assumptions of the multiple regression model

A multiple regression model is considered valid only if the following assumptions can be made about the population. Since population regression errors are not observable, the sample residuals $e_i = Y_i - \hat{Y}_i$ are used to determine whether each assumption is violated.

- Mean of zero: The mean of each residual for each set of values for the predictor variables is zero. Equivalently, this assumption says that the response variable is a linear function of each of the predictor variables.
- Independence: The residuals are independent. This condition can be difficult to assess. A common way to determine independence is by plotting residuals with respect to the time in which the data is collected. If a trend exists, then the independence assumption is potentially violated.
- Normality: The residuals of each set of values for the predictor variables form a normal distribution. If the plotted points lie reasonably close to the diagonal line on the plot then one can conclude that the normality assumption holds.
- Constant variance: The residuals of each set of values for the predictor variables should have equal or similar variance. A common term for this condition is *homoscedasticity*. If the variance does not remain constant throughout the plot, then the model exhibits *heteroscedasticity*.

PARTICIPATION ACTIVITY 6.2.1: Residual plots for assessing regression assumptions.

Animation content:

undefined

Animation captions:

1. Mean of zero assumption: The mean of the residuals should be approximately zero for each fixed value on the horizontal axis, i.e., no strong trends.
2. Any clear linear or nonlinear trend indicates the mean of zero assumption may not hold. Ex: A strong nonlinear rising-falling trend indicates the assumption does not hold.
3. Constant variance assumption: The variance of the residuals should be approximately constant for each fixed value on the horizontal axis, i.e., like a horizontal band.
4. Any clear change in variability indicates the constant variance assumption may not hold. Ex: A strong funnel pattern indicates the assumption does not hold.
5. An outlier is an observation with an extreme residual, which appears as an isolated point at the top or bottom of a residual plot.
6. Independence assumption: the value of one error should be independent of the value of any other error. Ex: No patterns in a plot of residuals versus time should exist.
7. A strong relationship between one residual and the next indicates the independence assumption may not hold. Ex: Residuals that closely follow one another.
8. Normality assumption: the errors are assumed to be normally distributed, which is assessed using a normal probability plot of the residuals, not a residual scatterplot.

PARTICIPATION ACTIVITY 6.2.2: Residual plots for assessing regression assumptions.

Refer to the animation above.

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

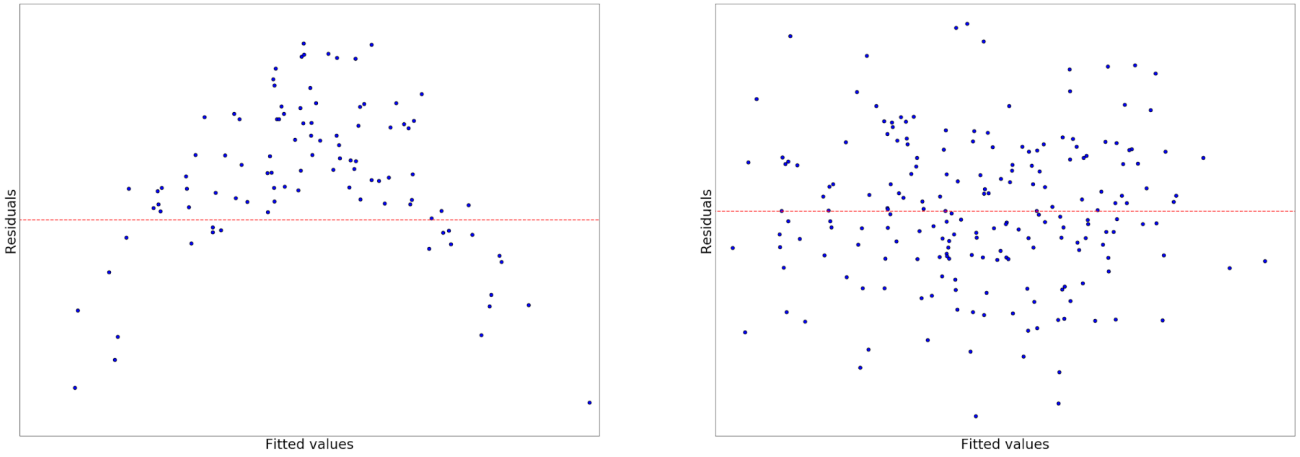
Using graphs to validate multiple regression assumptions

The following residual plots can be used to assess whether a model that predicts Y = percentage body fat using predictor variables X_1 = triceps skinfold thickness, X_2 = midarm circumference,

and X_3 = thigh circumference satisfies the multiple regression assumptions. The sample size of 20 is smaller than is typically useful, but is used nevertheless for illustrative purposes.

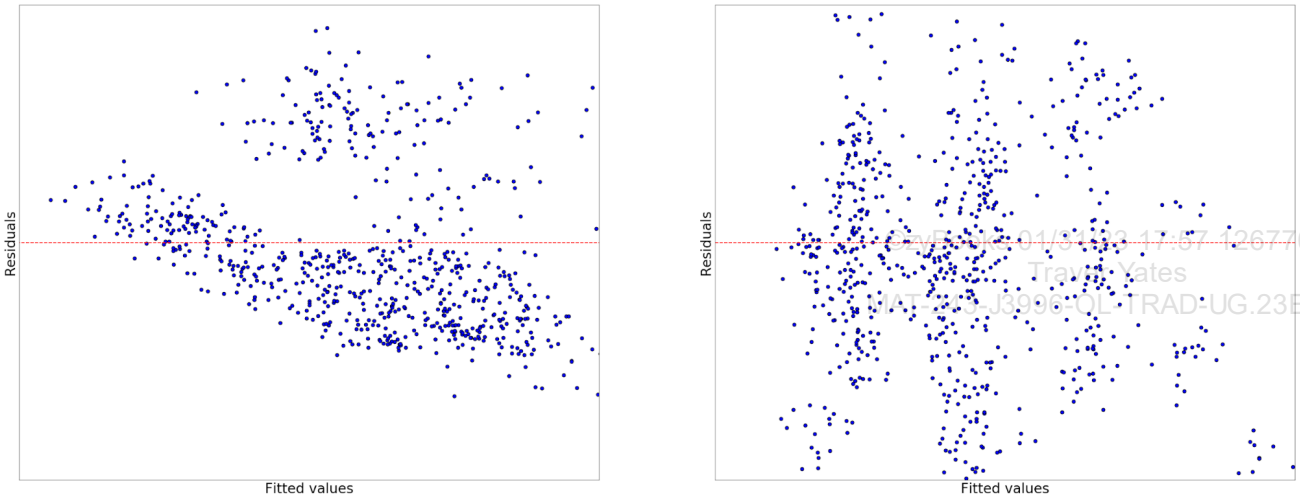
Residual plots can be used to visually evaluate the assumptions of linearity and homoscedasticity. A residual plot with a non-linear shape suggests violation of the linearity assumption.

Figure 6.2.1: A nonlinear dataset (left) and a dataset with no evidence of nonlinearity (right).



A residual plot with a fan-shaped distribution of data points suggests that some data points have a smaller variance than others. Since the variance is not constant, such a plot suggests violation of the homoscedasticity assumption.

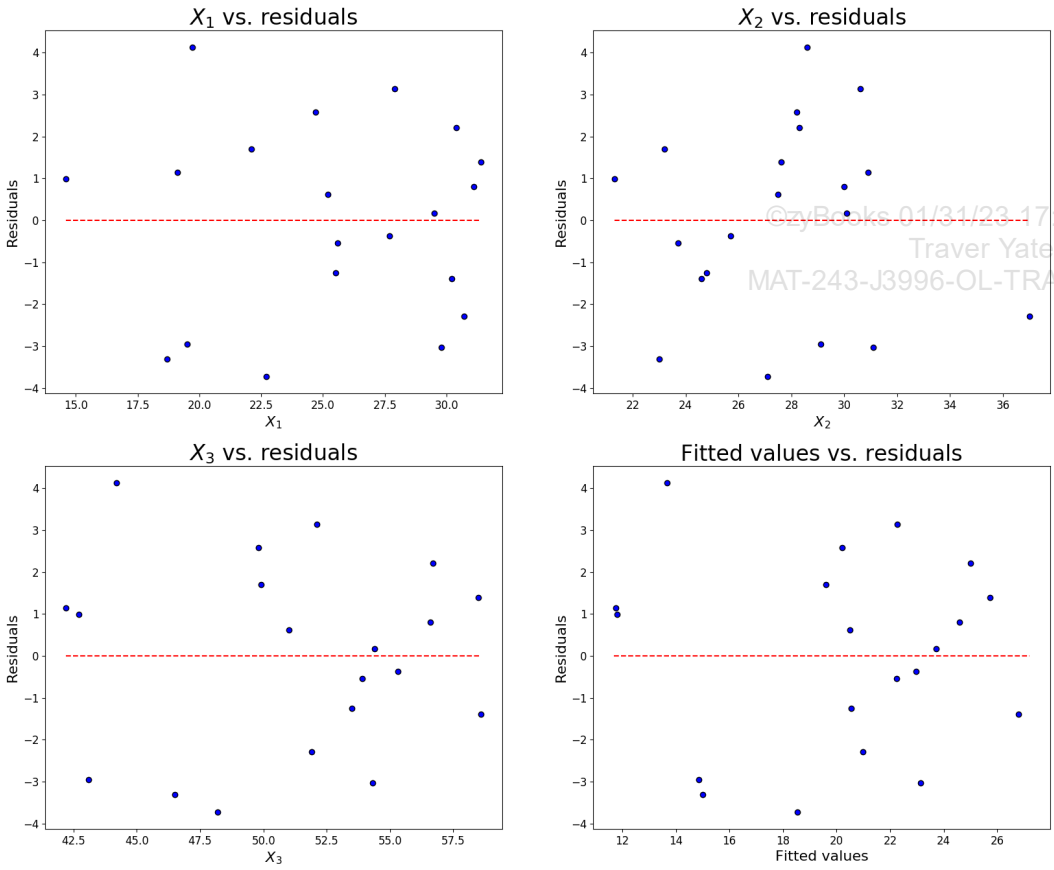
Figure 6.2.2: A heteroscedastic dataset (left) and a homoscedastic dataset (right).



Python-Practice 6.2.1: Residual plots.

©zyBooks 01/31/23 17:57 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3

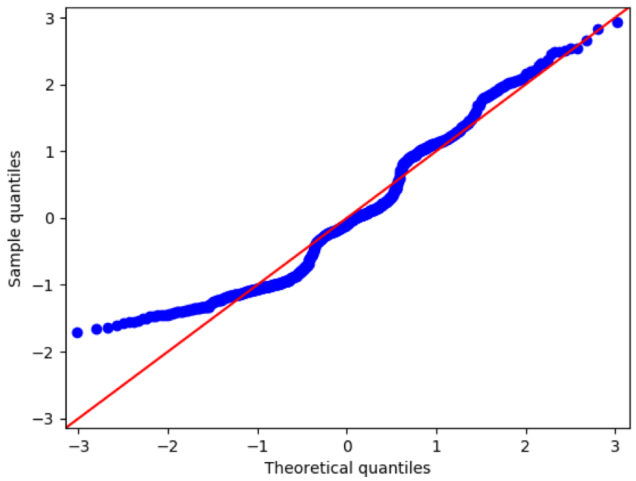
```
import pandas as pd
import statsmodels.formula.api as sms
import matplotlib.pyplot as plt

fat = pd.read_csv('https://static-resources.zybooks.com/static/fat.csv')

# Response variable
Y = fat['body_fat_percent']

# Generates the linear regression model
# Multiple predictor variables are joined with +
model = sms.ols('Y ~ triceps_skinfold_thickness_mm + midarm_circumference_cm +
thigh_circumference_cm', data = fat).fit()

plt.figure(figsize = (20, 16))
plt.tight_layout()

plt.subplot(2, 2, 1)
plt.scatter(x = fat['triceps_skinfold_thickness_mm'], y = model.resid, color = 'blue',
edgecolor = 'k')
xmin = min(fat['triceps_skinfold_thickness_mm'])
xmax = max(fat['triceps_skinfold_thickness_mm'])
plt.hlines(y = 0, xmin = xmin, xmax = xmax, color = 'red', linestyle = '--')
plt.xlabel('$X_1$', fontsize = 16)
plt.ylabel('Residuals', fontsize = 16)
plt.xticks(fontsize = 12)
plt.yticks(fontsize = 12)
plt.title('$X_1$ vs. residuals', fontsize = 24)

plt.subplot(2, 2, 2)
plt.scatter(x = fat['midarm_circumference_cm'], y = model.resid, color = 'blue', edgecolor
= 'k')
xmin = min(fat['midarm_circumference_cm'])
xmax = max(fat['midarm_circumference_cm'])
plt.hlines(y = 0, xmin = xmin, xmax = xmax, color = 'red', linestyle = '--')
plt.xlabel('$X_2$', fontsize = 16)
plt.ylabel('Residuals', fontsize = 16)
plt.xticks(fontsize = 12)
plt.yticks(fontsize = 12)
plt.title('$X_2$ vs. residuals', fontsize = 24)

plt.subplot(2, 2, 3)
plt.scatter(x = fat['thigh_circumference_cm'], y = model.resid, color = 'blue', edgecolor =
'k')
xmin = min(fat['thigh_circumference_cm'])
xmax = max(fat['thigh_circumference_cm'])
plt.hlines(y = 0, xmin = xmin, xmax = xmax, color = 'red', linestyle = '--')
plt.xlabel('$X_3$', fontsize = 16)
plt.ylabel('Residuals', fontsize = 16)
plt.xticks(fontsize = 12)
plt.yticks(fontsize = 12)
plt.title('$X_3$ vs. residuals', fontsize = 24)

plt.subplot(2, 2, 4)
plt.scatter(x = model.fittedvalues, y = model.resid, color = 'blue', edgecolor = 'k')
xmin = min(Y)
xmax = max(Y)
plt.hlines(y = 0, xmin = xmin, xmax = xmax, color = 'red', linestyle = '--')
plt.xlabel('Fitted values', fontsize = 16)
plt.ylabel('Residuals', fontsize = 16)
plt.xticks(fontsize = 12)
plt.yticks(fontsize = 12)
plt.title('Fitted values vs. residuals', fontsize = 24)
plt.show()
```

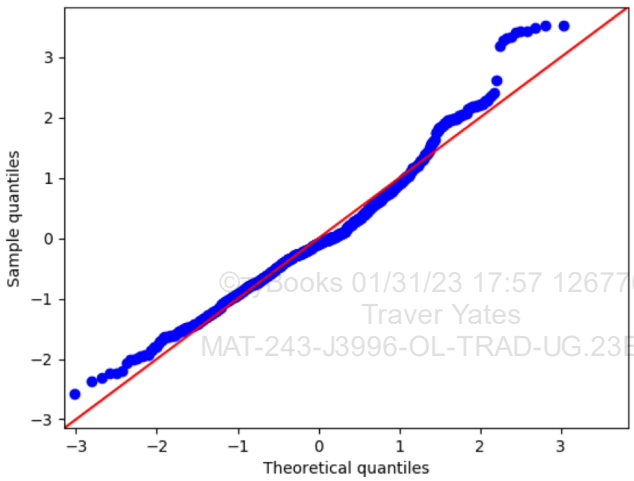


Run example

A Q-Q plot can be used to visually evaluate the assumption of normality. A Q-Q plot in which the points deviate significantly from the diagonal line suggests violation of the normality assumption.

Figure 6.2.3: Q-Q plots of datasets with non-normally distributed residuals (left) and normally distributed residuals (right).

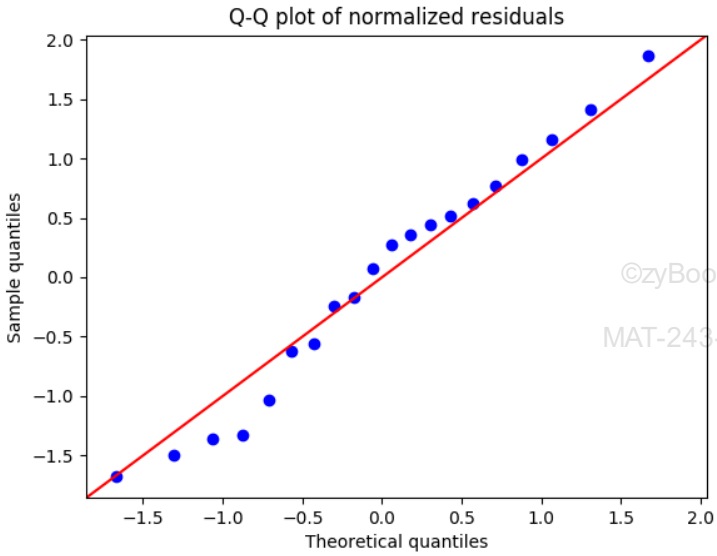




Python-Practice 6.2.2: qqplot().

The **qqplot** method is used to generate a Q-Q plot.

```
import pandas as pd
import statsmodels.graphics.gofplots as smg
import statsmodels.formula.api as sms
import matplotlib.pyplot as plt

fat = pd.read_csv('https://static-resources.zybooks.com/static/fat.csv')

# Response variable
Y = fat['body_fat_percent']

# Generates the linear regression model
# Multiple predictor variables are joined with +
model = sms.ols('Y ~ triceps_skinfold_thickness_mm + midarm_circumference_cm +
thigh_circumference_cm', data = fat).fit()

plt.figure(figsize = (8 ,5))
plt.subplot(1, 2, 1)
fig = smg.qqplot(model.resid, line = '45', fit = 'True')

plt.xlabel('Theoretical quantiles')
plt.ylabel('Sample quantiles')
plt.title('Q-Q plot of normalized residuals')
plt.show()
```



Run example

PARTICIPATION ACTIVITY 6.2.3: Checking the multiple regression assumptions.

Refer to the plots above.

1) Based on the residual scatterplots, does the "mean of zero" assumption appear to hold?
   - ○ Yes
   - ○ No

2) Based on the residual scatterplots, does the "constant variance" assumption appear to hold?
   - ○ Yes
   - ○ No

3) Based on the normal probability residual plot, does the "normality" assumption appear to hold?
   - ○ Yes
   - ○ No

4) Can one assess whether time dependence exists between the errors in any of the plots shown? (Such error

time dependence was discussed earlier.)

- Yes
- No

Care must be taken when checking the multiple 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, multiple 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.

Issues with multiple regression models

Individual data observations can have a relatively strong influence on multiple regression models because of the presence of *outliers* and *high-leverage observations*.

An ***outlier*** is an observation in a multiple regression analysis that has an extreme residual. For an outlier to be extreme, the observation's Y value is either much larger than predicted by the model or much smaller than predicted by the model.

A ***high leverage observation*** is an observation in a multiple regression analysis that has an extreme combination of predictor values. An extreme combination of values might be particularly high or low values for all the predictors or a combination of predictor values that is relatively unusual. A high leverage observation has the potential to be highly influential on the results of the multiple regression analysis.

Another potential problem is the presence of *multicollinearity*. ***Multicollinearity*** occurs when two or more predictors in a multiple regression model are so highly correlated that the estimated model becomes unstable meaning that the regression parameter estimates become unreliable with inflated standard errors. Interpreting a multiple regression model in the presence of multicollinearity can be challenging. However, estimation and prediction for a multiple regression model in the presence of multicollinearity is relatively unaffected.

PARTICIPATION ACTIVITY 6.2.4: Influential observations and multicollinearity in multiple regression.

Select the definition that matches each term

1) Outlier

- An observation with an extreme residual
- An observation with an extreme combination of predictor values
- A condition where two or more predictors are so highly correlated that the estimated model becomes unstable

2) High leverage observation

- An observation with an extreme residual
- An observation with an extreme combination of predictor values
- A condition where two or more predictors are so highly correlated that the estimated model becomes unstable

3) Multicollinearity

- An observation with an extreme residual
- An observation with an extreme combination of predictor values
- A condition where two or more predictors are so highly correlated that the estimated model becomes unstable

Reset

6.3 Coefficient of multiple determination

Coefficient of multiple determination

In the simple linear regression chapter, the coefficient of determination was defined as the percent variance in the response variable explained by the predictor variable. A similar quantity called the *coefficient of multiple determination* exists for models with two or more predictor variables. The ***coefficient of multiple determination***, denoted by R^2, measures the ratio of total variance in the response variable, Y, that is explained by the predictor variables $X_1, \ldots, X_n$.

The coefficient of determination is given by the formula

$$R^2 = \frac{SSR}{SSTO} = \frac{\sum_{i=1}^{N} \left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{N} (Y_i - \overline{Y})^2}$$

where SSR denotes the regression sum of squares or the variance explained by the predictor variables $X_1, \ldots, X_n$, $SSTO$ denotes the total sum of squares or the total variance, and N is the sample size.

Python-Practice 6.3.1: Coefficient of multiple determination.

Consider the body fat dataset and a model where the response variable Y is percent body fat and the predictor variable $X1$ is triceps skinfold thickness in millimeters. The model is constructed using the code below.

`R-squared: 0.711`, which means that 71.1% of the variance in percent body fat can be explained by the variance in triceps skinfold thickness.

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm

fat = pd.read_csv('https://static-resources.zybooks.com/static/fat.csv')

# Response variable
Y = fat['body_fat_percent']

m01 = ols('Y ~ triceps_skinfold_thickness_mm', data = fat).fit()
print(m01.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.711
Model:                            OLS   Adj. R-squared:                  0.695
Method:                 Least Squares   F-statistic:                     44.30
Date:                Mon, 15 Jul 2019   Prob (F-statistic):           3.02e-06
Time:                        21:16:05   Log-Likelihood:                -48.058
No. Observations:                  20   AIC:                             100.1
Df Residuals:                      18   BIC:                             102.1
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept                     -1.4961      3.319     -0.451      0.658      -8.470       5.477
triceps_skinfold_thickness_mm     0.8572      0.129      6.656      0.000       0.587       1.128
==============================================================================
Omnibus:                        1.783   Durban-Watson:                   1.928
Prob(Omnibus):                  0.410   Jarque-Bera (JB):                1.511
Skew:                          -0.600   Prob(JB):                        0.470
Kurtosis:                       2.389   Cond. No.                         136.
==============================================================================
```


Adding another predictor variable, midarm circumference X_2, to the model increases the value of R^2 to 0.786 or 78.6%. In other words, using both triceps skinfold thickness and midarm circumference, instead of just triceps skinfold thickness, improved the model's ability to predict a person's percent body fat. The addition of X_2 made the model more accurate and increased the explained variance by $78.6\% - 71.1\% = 7.5\%$.

```
m12 = ols('Y ~ triceps_skinfold_thickness_mm + midarm_circumference_cm', data = fat).fit()
print(m12.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.786
Model:                            OLS   Adj. R-squared:                  0.761
Method:                 Least Squares   F-statistic:                     31.25
Date:                Mon, 15 Jul 2019   Prob (F-statistic):           2.02e-06
Time:                        21:16:05   Log-Likelihood:                -45.050
No. Observations:                  20   AIC:                             96.10
Df Residuals:                      17   BIC:                             99.09
Df Model:                           2
Covariance Type:            nonrobust
================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept                      6.7916      4.488      1.513      0.149      -2.678      16.261
triceps_skinfold_thickness_mm     1.0006      0.128      7.803      0.000       0.730       1.271
midarm_circumference_cm       -0.4314      0.177     -2.443      0.026      -0.804      -0.059
==============================================================================
Omnibus:                        1.363   Durban-Watson:                   2.371
Prob(Omnibus):                  0.506   Jarque-Bera (JB):                0.873
Skew:                           0.068   Prob(JB):                        0.646
Kurtosis:                       1.985   Cond. No.                         304.
==============================================================================
```


Adding thigh circumference X_3 to a model with two predictor variables further increases the value of R^2 to 0.801 as shown in the output below.

```
m = ols('Y ~ triceps_skinfold_thickness_mm + midarm_circumference_cm + thigh_circumference_cm',
data = fat).fit()
print(m.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.801
Model:                            OLS   Adj. R-squared:                  0.764
Method:                 Least Squares   F-statistic:                     21.52
Date:                Mon, 15 Jul 2019   Prob (F-statistic):           7.34e-06
Time:                        21:16:05   Log-Likelihood:                -44.312
No. Observations:                  20   AIC:                             96.62
Df Residuals:                      16   BIC:                             100.6
Df Model:                           3
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                      117.0847     99.782      1.173      0.258     -94.445     328.614
triceps_skinfold_thickness_mm    4.3341      3.016      1.437      0.170      -2.059      10.727
midarm_circumference_cm         -2.1861      1.595     -1.370      0.190      -5.568       1.196
thigh_circumference_cm          -2.8568      2.582     -1.106      0.285      -8.330       2.617
==============================================================================
Omnibus:                        1.200   Durbin-Watson:                   2.243
Prob(Omnibus):                  0.549   Jarque-Bera (JB):                0.830
Skew:                          -0.085   Prob(JB):                        0.660
Kurtosis:                       2.016   Cond. No.                     1.15e+04
==============================================================================
```

Alternatively, the value of R^2 for the model above with three predictor variables can be computed from the analysis of variance table.

```
X = fat[ [ 'triceps_skinfold_thickness_mm',
           'midarm_circumference_cm', 'thigh_circumference_cm']]
model = ols('Y ~ X', data = fat).fit()
print(sm.stats.anova_lm(model, typ=2))
```

```
              sum_sq    df          F    PR(>F)
X         396.984612   3.0  21.515712  0.000007
Residual   98.404888  16.0        NaN       NaN
```

The variance explained by X_1, X_2, and X_3 is 396.984612 and the total variance is $396.984612 + 98.404888 = 495.3895$. Thus,

$$R^2 = \frac{396.984612}{495.3895} = 0.801$$

Run example

PARTICIPATION ACTIVITY 6.3.1: Coefficient of multiple determination.

Consider a body fat model with two predictor variables, X_1 = triceps skinfold thickness and X_3 = thigh circumference. The output of the `ols lm` regression model is given below.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.778
Model:                            OLS   Adj. R-squared:                  0.752
Method:                 Least Squares   F-statistic:                     29.80
Date:                Mon, 15 Jul 2019   Prob (F-statistic):           2.77e-06
Time:                        23:04:08   Log-Likelihood:                -45.422
No. Observations:                  20   AIC:                             96.84
Df Residuals:                      17   BIC:                             99.83
Df Model:                           2
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                      -19.1742      8.361     -2.293      0.035     -36.814      -1.535
triceps_skinfold_thickness_mm    0.2224      0.303      0.733      0.474      -0.418       0.863
thigh_circumference_cm           0.6594      0.291      2.265      0.037       0.045       1.274
==============================================================================
Omnibus:                        1.263   Durbin-Watson:                   2.359
Prob(Omnibus):                  0.532   Jarque-Bera (JB):                0.843
Skew:                           0.063   Prob(JB):                        0.656
Kurtosis:                       2.002   Cond. No.                         846.
==============================================================================
```


Run example

1) What is the value for the coefficient of multiple determination R^2?

   - 0.752
   - 0.778

2) What is the interpretation of R^2 for this model?

   - Percent of the total variance explained by X_1
   - Percent of the total variance explained by X_3
   - Percent of the total variance explained by X_1 and X_3

PARTICIPATION ACTIVITY 6.3.2: Coefficient of determination.

Consider a body fat model with two predictor variables, X_2 = midarm circumference and X_3 = thigh circumference. The analysis of variance table is given below.

```
              sum_sq    df         F    PR(>F)
X23       384.279719   2.0  29.39775  0.000003
Residual  111.109781  17.0       NaN       NaN
```

Run example

1) What is the variance explained by X_2 and X_3? Type as: ###.######

Check Show answer

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

Check Show answer

3) What percent of the variance in body fat can be explained from the variance in X_2 and X_3? Type as: ##.#%

Check Show answer

Adjusted coefficient of multiple determination

The ***adjusted coefficient of multiple determination***, denoted by R^2_{adj}, is an adjustment to R^2 that allows alternative models for the same response variable to be compared. The formula for the adjusted coefficient of multiple determination is given by

$$R^2_{adj} = 1 - (1 - R^2)\left[\frac{N-1}{N-(k+1)}\right]$$

where N is the sample size and k is the number of predictor variables.

The value of R^2_{adj} only tends to increase when a worthwhile predictor is added to the model. In general, a model with a higher R^2_{adj} value is preferable to a model with a lower value. However, R^2_{adj} should not be used in isolation, but rather in conjunction with other criteria such as whether the model provides useful answers to the research questions and gives relatively precise predictions. These issues are beyond the scope of this material.

Example 6.3.1: Finding the adjusted coefficient of determination.

Consider a simple case of the body fat model with 1 predictor variable $X_1 =$ triceps skinfold thickness. The coefficient of determination for this model is $R^2 = 0.711$. Find the adjusted coefficient of determination using the formula above given that the sample size is 20.

Solution

Since the sample size is 20 and the model contains 1 predictor variable, $N = 20$ and $k = 1$. The formula $R^2_{adj} = 1 - (1 - R^2)\left[\dfrac{N-1}{N-(k+1)}\right]$ is used in the calculation below.

$$R^2_{adj} = 1 - (1 - 0.711)\left[\frac{20-1}{20-(1+1)}\right] = 0.695$$

The resulting adjusted coefficient of determination $R^2_{adj} = 0.695$ is consistent with the `ols` output in the earlier subsection.

PARTICIPATION ACTIVITY 6.3.3: Finding the adjusted coefficient of multiple determination.

Consider the model with 2 predictor variables $X_2 =$ midarm circumference and $X_3 =$ thigh circumference. The coefficient of multiple determination for this model with a sample size of 20 is $R^2 = 0.776$.

1) What is the value of R^2_{adj}? Type as: #.###

Check Show answer

PARTICIPATION ACTIVITY 6.3.4: Interpreting values for the adjusted coefficient of multiple determination.

Consider two body fat models.

The first body fat model contains one predictor variable $X_1 =$ triceps skinfold thickness. The output of the `ols` . function for this model is given below.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.711
Model:                            OLS   Adj. R-squared:                  0.695
Method:                 Least Squares   F-statistic:                     44.30
Date:                Mon, 15 Jul 2019   Prob (F-statistic):           3.02e-06
Time:                        21:16:05   Log-Likelihood:                -48.058
No. Observations:                  20   AIC:                             100.1
Df Residuals:                      18   BIC:                             102.1
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept                     -1.4961      3.319     -0.451      0.658      -8.470       5.477
triceps_skinfold_thickness_mm     0.8572      0.129      6.656      0.000       0.587       1.128
==============================================================================
Omnibus:                        1.783   Durbin-Watson:                   1.928
Prob(Omnibus):                  0.410   Jarque-Bera (JB):                1.511
Skew:                          -0.600   Prob(JB):                        0.470
Kurtosis:                       2.389   Cond. No.                         136.
==============================================================================
```


The second body fat model contains two predictor variables, X_1 = triceps skinfold thickness and X_3 = thigh circumference. The output of the `ols` function for this model is given below.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.778
Model:                            OLS   Adj. R-squared:                  0.752
Method:                 Least Squares   F-statistic:                     29.80
Date:                Mon, 15 Jul 2019   Prob (F-statistic):           2.77e-06
Time:                        23:04:08   Log-Likelihood:                -45.422
No. Observations:                  20   AIC:                             96.84
Df Residuals:                      17   BIC:                             99.83
Df Model:                           2
Covariance Type:            nonrobust
================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept                    -19.1742      8.361     -2.293      0.035     -36.814      -1.535
triceps_skinfold_thickness_mm     0.2224      0.303      0.733      0.474      -0.418       0.863
thigh_circumference_cm         0.6594      0.291      2.265      0.037       0.045       1.274
==============================================================================
Omnibus:                        1.263   Durbin-Watson:                   2.359
Prob(Omnibus):                  0.532   Jarque-Bera (JB):                0.843
Skew:                           0.063   Prob(JB):                        0.656
Kurtosis:                       2.002   Cond. No.                         846.
==============================================================================
```

1) What percentage of the total variance in body fat is explained by the multiple regression model with triceps skinfold thickness and thigh circumference as the predictors?

- ○ 71.1%
- ○ 75.2%
- ○ 77.8%

2) Is the model with two predictor variables better than the model with one predictor variable?

- Yes
- No
- Perhaps

©zyBooks 01/31/23 17:57 1267703
Traver Yates
MAT-243-J3996-OL-TRAD-UG.23EW3