7.1 Introduction to regression

Learning goals

- Define input and output features.
- Use scatter plots to identify the direction, form, and strength of a relationship between two features.
- Define regression models.
- Define predicted values, observed values, and residuals.

Input and output features

When analyzing data, some features may be interrelated and have a cause-and-effect relationship. Ex: Data gathered on the weight of cars and fuel efficiency would likely indicate that as weight increases, fuel efficiency decreases.To investigate potential relationships between features, two types are considered:

- An ***input feature*** takes values without being impacted by any other features.
- An ***output feature*** has values that vary in response to variation in some other feature(s).

In an experiment, input features are controlled by researchers and output features are observed. If a cause-and-effect relationship is suspected, the feature that is thought to cause the other feature should be treated as an input.

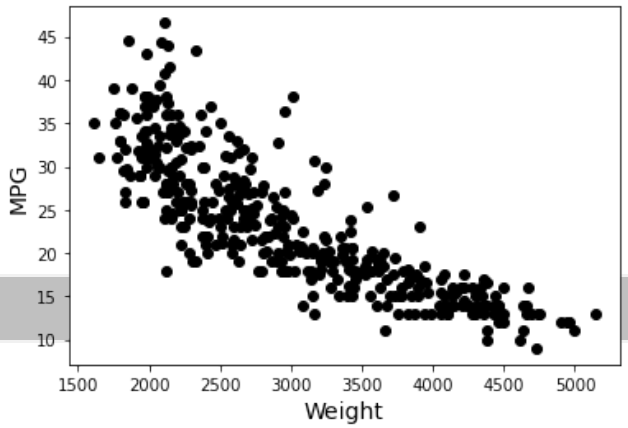
When graphing a scatter plot of two features, the input feature is generally displayed as the x-axis and the output feature as the y-axis.

Visualizing input and output features in the cars dataset.

The cars dataset[1] contains information on cars sold from 1970 through 1982. A scatter plot explores the relationship between two numerical features. In general, cars with more horsepower tend to have higher displacement. But other features could also affect displacement. What does the data say? To create a scatter plot, select horsepower as the input feature and displacement as the output feature.

PARTICIPATION ACTIVITY 7.1.1: Another pair of input and output features in the cars dataset.

Another pair of features that can be considered are weight in pounds and fuel efficiency in miles per gallon (mpg). In general, a lighter car causes better fuel efficiency, with all else being equal.



Identify the input and output features.

If unable to drag and drop, refresh the page.

Weight **Miles per gallon**

Input feature

Output feature

Reset

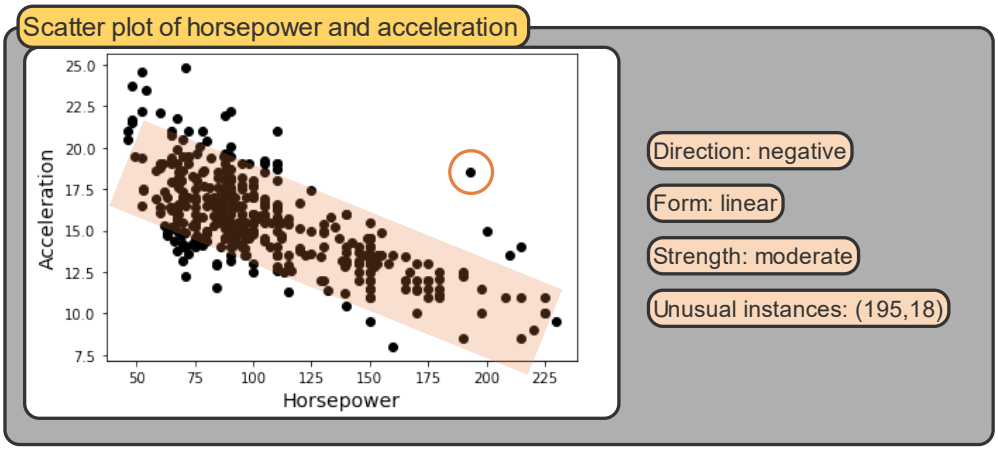
Describing scatter plot relationships

The relationship between the features graphed in a scatter plot can be described in terms of direction, form, and strength:

- The ***direction*** of the relationship is positive if larger values of one feature correspond to larger values of the other feature. The direction of the association is negative if larger values of one feature correspond to smaller values of the other feature.
- The ***form*** of the relationship indicates if the scatter plot follows a linear pattern or a nonlinear pattern, such as a parabola. Sometimes two features may not have an obvious form.
- The ***strength*** of the relationship association indicates how closely the instances in a scatter plot follow the form's pattern.

Some instances in the scatter plot may stand out from the rest of the scatter plot. An ***unusual instance*** is any instance that does not follow the overall pattern. Unusual instances should be noted and may indicate an error in the data collection or give insight to some deeper behavior.

PARTICIPATION ACTIVITY 7.1.2: Properties of the horsepower and acceleration relationship.



Animation content:

Step 1: The scatter plot of horsepower and acceleration appears. A downward-tilted box appears over the points. Most points fall inside the box.
Step 2: The text "Direction: negative" appears.
Step 3: The text "Form: linear" appears
Step 4: The text "Strength: moderate" appears.
Step 5: The point (195,18) is circled. The text "Unusual instances: (195,18)" appears.

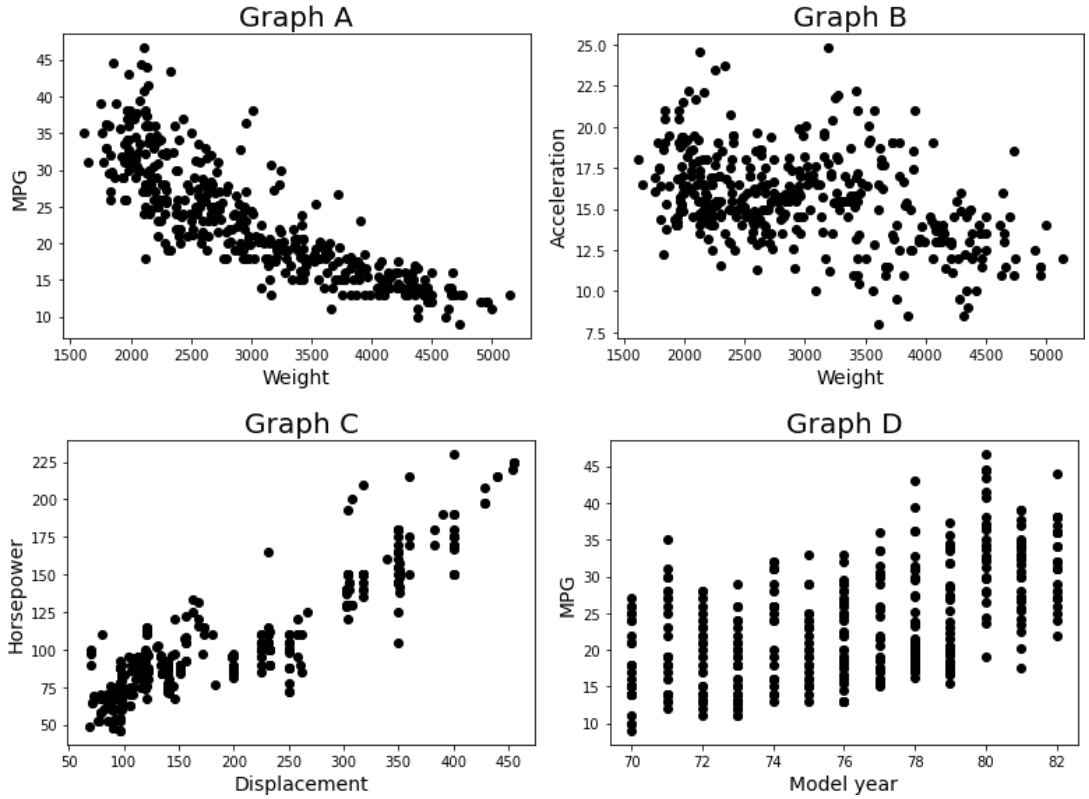
Animation captions:

1. A box can be used to help visualize the relationship between horsepower and acceleration.
2. The box is tilted downward, indicating a negative relationship. In context, as horsepower increases, the amount of time for a car to accelerate from 0 mph to 60 mph decreases.

3. The points seem to follow somewhat of a linear pattern, indicating that the relationship between horsepower and acceleration has a linear form.
4. The points mostly follow the linear form with some deviations, so the linear relationship is moderately strong.
5. The point at (195,18) seems to fall well outside of the overall pattern.

PARTICIPATION ACTIVITY 7.1.3: Describing scatter plots from the cars data.

Graphed below are scatter plots of four pairs of features in the cars dataset.



Match each graph to the corresponding direction, form, and strength.

If unable to drag and drop, refresh the page.

Graph D | Graph A | Graph C | Graph B

	Positive, linear, somewhat weak
	Negative, no obvious form, very weak
	Positive, linear, moderately strong
	Negative, non-linear, moderately strong

Reset

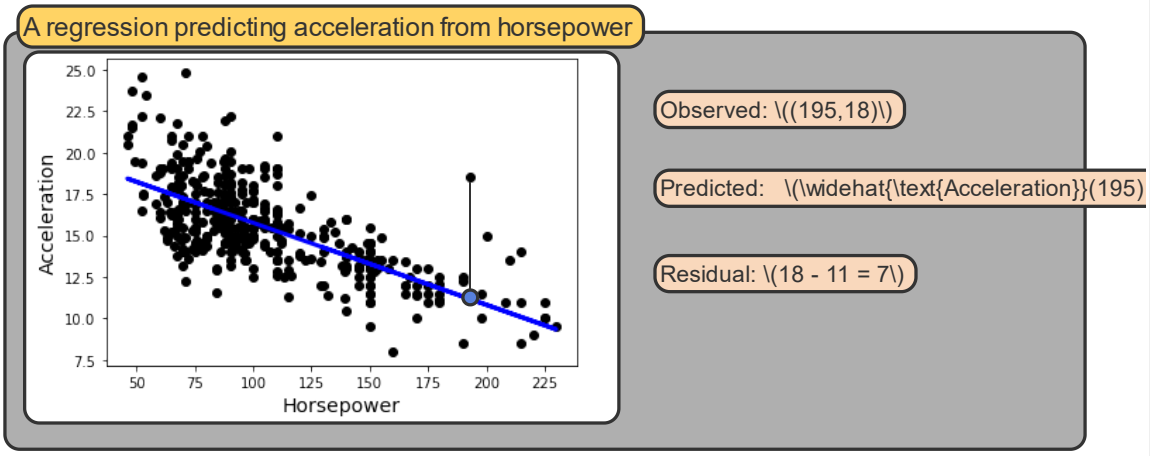
Regression models

A ***model*** for an output feature \(y\) using input feature(s) \(X\) is a function \(f(X)\) that predicts an expected value of \(y\) for a given value of \(X\) as \(\hat{y} = f(X)\). So \(\hat{y}\) is the predicted value for the output feature \(y\) for a given value of the input feature \(X\). Specifically, a ***regression model*** is a model that uses numeric output features. Visually, a regression model for data graphed on a scatter plot traces out a line or a nonlinear curve.

If two features have a strong relationship, predicted values should be close to the observed values. Closeness can be se visually by comparing points to the graph of the model. Closeness can be measured numerically using residuals.The ***residual*** of an instance \((x_i, y_i) \) is the difference between the observed and predicted value \((y_i - \hat{y}_i) \). A well-fitting model should have small residuals on average.

A model is also used to make predictions about values aside from what was observed. Predictions should only be made for values close to the range of the original data. An ***extrapolation*** is a prediction for a value far beyond the range of the original data and is often misleading or inaccurate.

PARTICIPATION ACTIVITY 7.1.4: Predicting acceleration from horsepower.



Animation content:

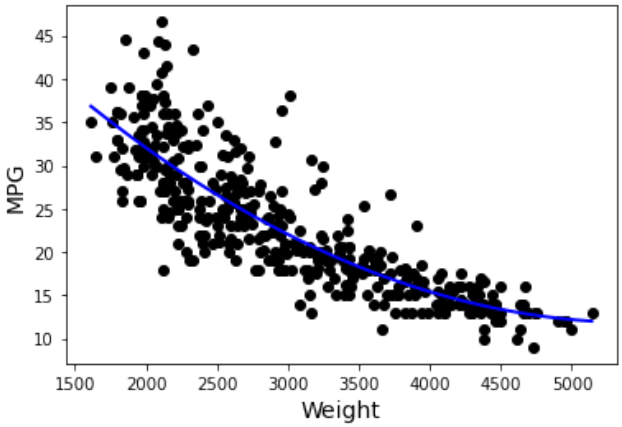
Step 1: The scatter plot of horsepower versus acceleration appears. A regression line appears on top of the scatter plot, matching the overall pattern given by the data.
Step 2: The point at (195,18) is circled and the text "Observed: 18" appears.
Step 3: A point on the curve at (195,11) is highlighted and the text "Predicted: \(\widehat{\text{Acceleration}}(195) = 11\)" appears.
Step 4: A line from the observed point to the predicted point is drawn and the text "Residual: 18 - 11 = 7" appears.

Animation captions:

1. A regression model can predict acceleration from horsepower using a line.
2. The point at (195,18) is an observed value previously considered.
3. The regression line can be used to predict an acceleration from horsepower.
4. The residual of the instance (195,18) can be found by calculating \((\text{observed} - \text{predicted}) \). On the graph, a residual is represented with a line between the observed and predicted values.

PARTICIPATION ACTIVITY 7.1.5: A regression predicting miles per gallon from weight.

The graph below shows a regression predicting miles per gallon from weight using a nonlinear function.



1) The observed miles per gallon value for the car represented by the point at (3050, 39) is _____ miles per gallon.
   - 21
   - 39
   - 3,050

2) The predicted miles per gallon value for a car weighing 3,050 pounds is _____ miles per gallon.
   - 18
   - 21
   - 39

3) The residual for the car represented by the point at (3050, 39) is _____ miles per gallon.
   - 21
   - 18
   - 39

4) Predicting the miles per gallon for a weight of _____ pounds with this regression model would be an extrapolation.
   - 1,500
   - 3,050
   - 7,000

©zyBooks 04/14/24 12:43 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

(*1) Waskom, Michael. 2018. "mwaskom/seaborn-data/mpg.csv". Github repository.https://github.com/mwaskom/seaborn-data/blob/master/mpg.csv

7.2 Simple linear regression

Learning goals

- Define the simple linear regression model.
- Interpret the slope and intercept of a simple linear regression model.
- Calculate predicted values and residuals using simple linear regression.
- Calculate and interpret the correlation coefficient.
- Implement least squares regression using `scikit-learn`.

Simple linear regression

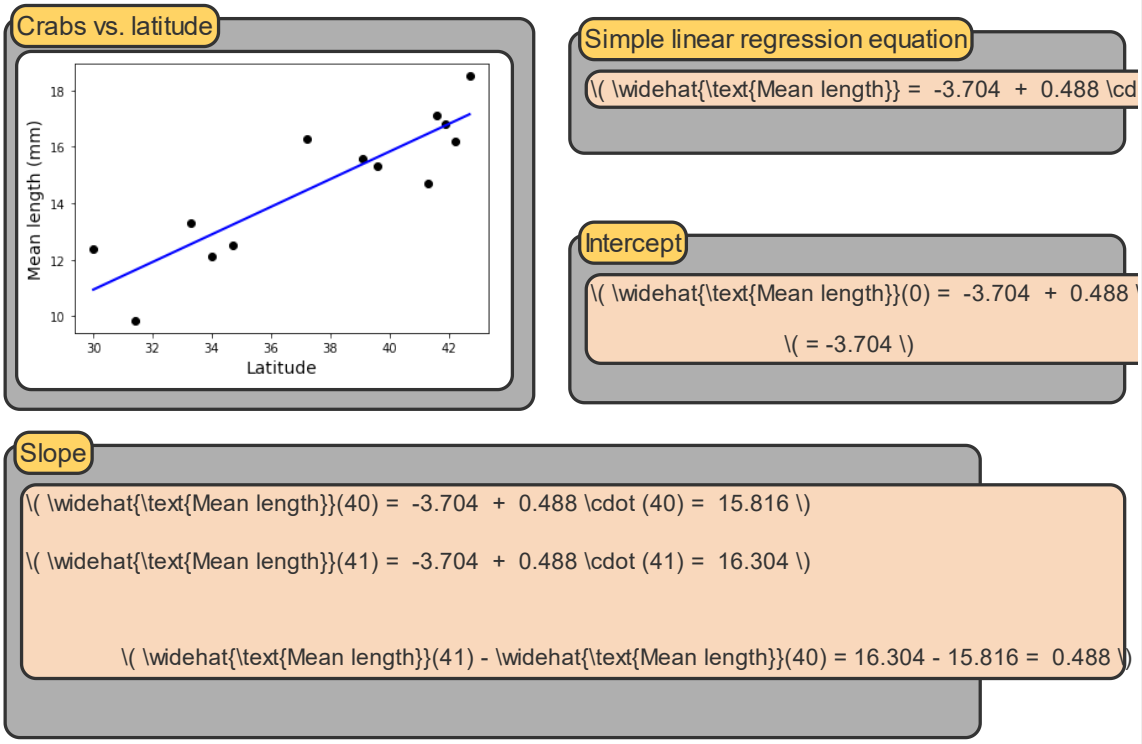
The simplest regression model for a pair of numeric features \((x,y)\) is to consider a model that predicts the output feature \(y\) for a given input feature \(x\) using a line graphed in the scatter plot. A ***simple linear regression*** is a mathematical model of the form \(\hat{y} = b_0 + b_1 x\), where:

- \(x\) is the input feature.
- \(\hat{y}\) is the predicted value of the output feature \(y\) for a given value of \(x\).
- \(b_0 \) is the \(y\)-intercept, representing the predicted value of \(y\) when \(x = 0\).
- \(b_1 \) is the slope, representing how much the predicted value of \(y\) changes for a one-unit change in the value of \(x\).

Figure 7.2.1: Do fiddler crabs that live in higher latitudes tend to be bigger?

Bergmann's rule states that organisms at higher latitudes are generally larger. Does this rule apply to the Atlantic fiddler crab? A study by Johnson found that on average, the shell width of the Atlantic fiddler crab increased by around 0.5 mm for each unit increase in latitude. [12] The study looked at crabs located in 13 locations across the Eastern United States. Samples were collected from each location and the minimum, maximum, mean, and median measurements of the crab shells were recorded.

PARTICIPATION ACTIVITY 7.2.1: A simple linear regression for fiddler crabs.



Animation content:

Step 1: A scatter plot graphing mean length in mm and latitude values appears.
Step 2: A simple linear regression is graphed over the scatter plot. The regression equation \(\widehat{\text{Mean length}} = -3.704 + 0.488 \cdot (\text{latitude}) \) appears.
Step 3: A box labeled "intercept" with the calculation \(\widehat{\text{Mean length}}(0) = -3.704 + 0.488 \cdot (0) = -3.704 \) appears.
Step 4: A box labeled "slope" appears with the calculations \(\widehat{\text{Mean length}}(40) = -3.704 + 0.488 \cdot (40) = 15.816 \), \(\widehat{\text{Mean length}}(41) = -3.704 + 0.488 \cdot (41) = 16.304 \), and \(\widehat{\text{Mean length}}(41) - \widehat{\text{Mean length}}(40) = 16.304 - 15.816 = 0.488 \)

Animation captions:

1. Researchers took body measurements and location data for groups of crabs in the Eastern United States.
2. A simple linear equation predicting fiddler crabs' mean length for a given latitude can be written as \(\hat{y} = -3.704 + 0.488 x \).
3. The intercept term -3.704 is the predicted mean length of fiddler crabs at latitude 0 degrees (the equator).
4. The slope term 0.488 is the amount the mean length of a group of crabs increases when the latitude increases by 1 degree.

PARTICIPATION ACTIVITY 7.2.2: A simple linear regression for fiddler crabs.

1) The prediction \(\widehat{\text{mean length}}(0) = -3.704 \) _____ an extrapolation.
   - is
   - is not

2) The predicted mean length of fiddler crabs at a latitude of +37 degrees is _____.
   - 0.488 mm
   - 14.352 mm
   - 16.3 mm

3) The residual for the instance (37,16.3) is _____.
   - -1.948 mm
   - 14.352 mm
   - 1.948 mm

The least-squares regression line

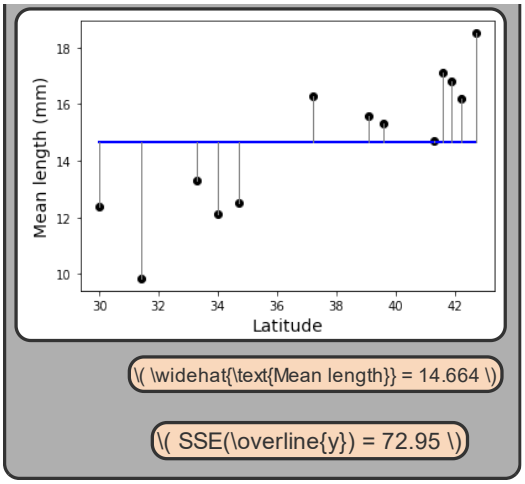
The values of \(b_0\) and \(b_1\) should be chosen so that the predictions \(\hat{y}_i\) are as close as possible, on average, to the observed values \(y_i\). A common method to compute regression coefficients is to minimize some mathematical combination of the residuals \((y_i - \hat{y}_i) \). The ***sum of squared errors*** (\(SSE\)) is the sum of the squares of all residuals and can be found by the formula \(\sum (y_i - \hat{y}_i)^2 \). A ***least-squares regression line*** is a simple linear regression \(\hat{y} = b_0 + b_1 x \) that minimizes the sum of squared errors.

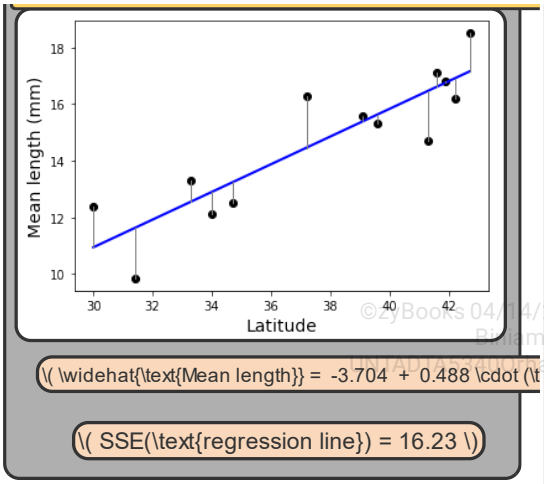
How well the simple linear regression line models the data can be measured by how much closer the predictions from the regression line are to the observed values compared to predictions from a model that assumes no relationship exists between \(x\) and \(y\). Such a "no-relationship" model would be a horizontal line model \(\hat{y} = \overline{y}\) that returns the mean \(\overline{y}\) for every value of \(x\). The proportion of variation explained by a simple linear regression is given by the equation \[\displaystyle \frac{SSE(\overline{y}) - SSE(\text{regression line}) }{SSE(\overline{y})} .\]

PARTICIPATION ACTIVITY 7.2.3: Evaluating the fiddler crab least-squares regression.

"No relationship" model \(\hat{y} = \overline{y} \)

Least squares regression \(\hat{y} = b_0 + b_





Proportion of variation explained by the least-squares regression line

\(\displaystyle \frac{SSE(\overline{y}) - SSE(\text{regression line}) }{SSE(\overline{y})} =

Animation content:

Step 1: Scatter plot of the crab data appears. A "no relationship" model \(\hat{y} = \overline{y}\) appears with the equation: predicted mean length = 14.664. The least-squares regression line model \(\hat{y} = b_0 + b_1 x\) appears with the equation: predicted mean length = -3.704 + 0.488 * (latitude).
Step 2: Lines from the points to the respective regression lines appear. \(SSE(\overline{y}) = 72.95 \) appears under the no relationship model. \(SSE(\text{Regression line}) = 16.23 \) appears under the least-squares regression model.
Step 3: The equation \(\frac{SSE(\overline{y}) - SSE(\text{regression line}) }{SSE(\overline{y})} = \frac{72.95 - 16.23}{72.95} = 0.777 \) appears as the proportion of variation explained by the least-squares regression line.

Animation captions:

1. The least-squares regression line can be compared to the line given by \(\hat{y} = \overline{y}\), which models the situation where \(x\) and \(y\) have no relationship.
2. The sum of squared errors \(SSE\) can be calculated from the residuals \((y_i - \hat{y}_i)\) for both models. The least-squares regression line has the smallest possible \(SSE\) among simple linear regressions.
3. 77.7% of the variation in mean length of a crab can be explained by variation in latitude using the least-squares regression model.

PARTICIPATION ACTIVITY 7.2.4: Another fiddler crab regression.

Consider the regression \(\widehat{\text{Mean length}} = -3.5 + 0.5 \cdot (\text{latitude}) \).
The questions below should be answered without any calculation.

1) The new regression is a _____.

- simple linear regression
- least-squares regression

2) The sum of squared errors \(SSE\) for this new simple linear regression must be _____.

- equal to 72.95
- equal to 16.23
- greater than 16.23

3) The proportion of variation of mean length of crabs explained by variation in latitude under using the new regression model must be _____.

- less than 0.777
- equal to 0.777
- greater than 0.777

The correlation coefficient \(r\) and least-squares regression

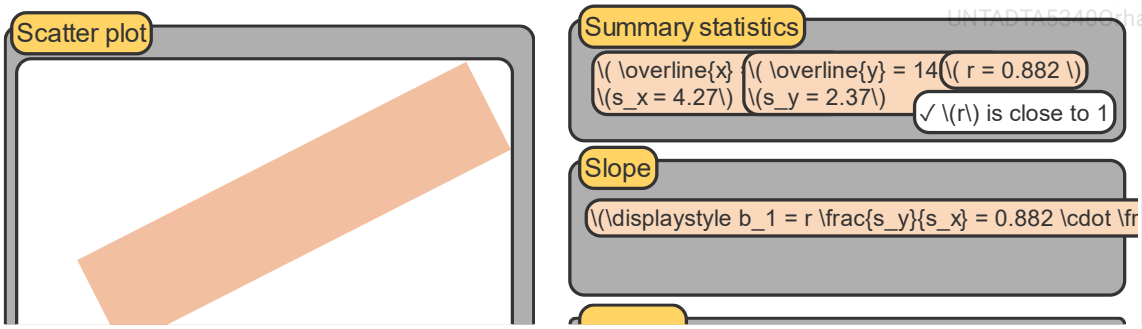
Above, the least-squares regression was given, but was using a simple linear model appropriate in the first place? If so, how were the coefficients computed? The answer to both questions involve the scatter plot and the data's summary statistics. A simple linear regression should only be used as a model when the scatter plot of \((x,y)\) displays a strong linear form. The linearity must be visually confirmed, but the strength can be checked numerically.

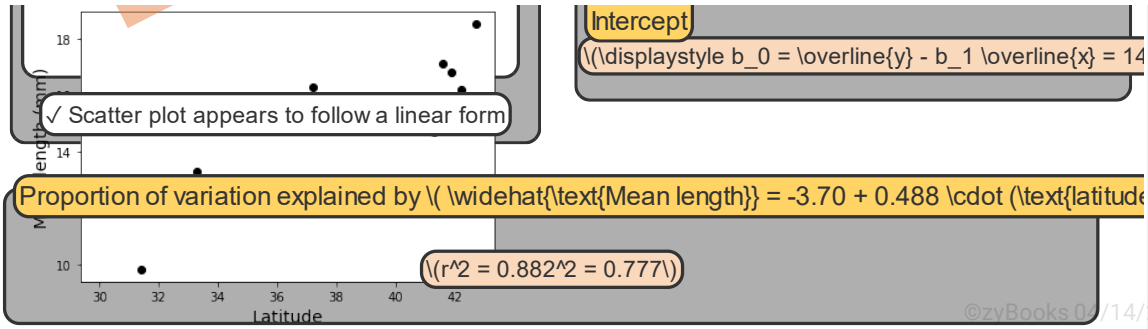
The ***correlation coefficient*** \(r\) measures the direction and strength of a linear relationship as a unitless value between -1 and 1, and can be computed by the formula \[r = \frac{ \sum (x_i - \overline{x}) (y_i - \overline{y}) }{\sqrt{ \sum (x_i - \overline{x})^2 \cdot \sum (y_i - \overline{y})^2 }} = \frac{1}{n-1} \sum \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right),\] where \(\overline{x}\) and \(\overline{y}\) are the sample means, \(s_x\) and \(s_y\) are the sample standard deviations, \(n\) is the sample size, and \((x_i, y_i)\) are the observed values.

A large positive \(r\) value indicates a positive linear association, and a large negative \(r\) value indicates a negative linear association.

If a simple linear regression is appropriate, the slope \(b_1\) and intercept \(b_0\) of the least-squares regression can be computed from the correlation coefficient as \[b_1 = r \frac{s_y}{s_x} \ \ \ \text{ and } \ \ \ b_0 = \overline{y} - b_1 \cdot \overline{x} . \] Furthermore, the proportion of variation explained by the least-squares regression is equal to \(r^2\).

PARTICIPATION ACTIVITY 7.2.5: Examining the fiddler crab least-squares regression in more detail.





Animation content:

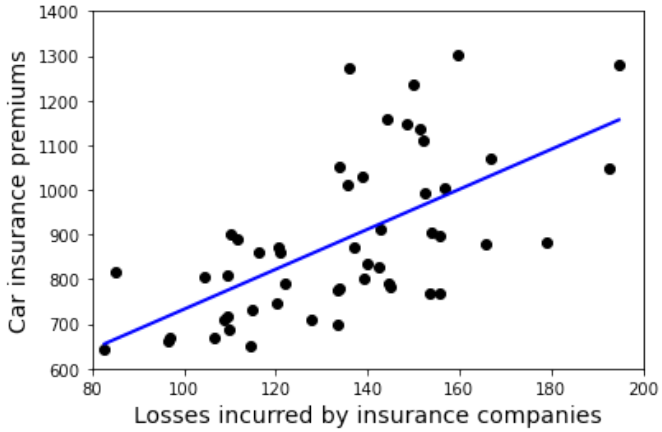
Step 1: A scatter plot graphing mean length in mm and latitude values appears. A rectangular box appears over the points and text appears with a checkmark stating "scatter plot appears to follow a linear form".
Step 2: Summary statistics for the scatter plot appear with the correlation coefficient r = 0.882. Text appears with a checkmark stating "r is close to 1".
Step 3: The summary statistics are plugged into the equations to calculate the slope and intercept of the least-squares regression.
Step 4: The correlation coefficient r is squared to get 0.777, the proportion of variation explained by the least-squares regression.

Animation captions:

1. Examining the scatter plot, the values appear to show a linear form overall.
2. The summary statistics can be used to calculate the correlation coefficient \(r\). A correlation coefficient close to 1 indicates the linear relationship is strong and positive (\(y\) increases as \(x\) increases).
3. Since a simple linear regression is appropriate, the summary statistics and correlation coefficient can be used to calculate the slope and intercept of the least-squares regression.
4. Squaring the correlation coefficient gives the proportion of variation in mean length explained by the linear relationship with latitude given by the least-squares regression.

PARTICIPATION ACTIVITY 7.2.6: A simple linear regression for the bad drivers dataset.

Graphed below is a scatter plot and the least-squares regression for the bad drivers dataset. The dataset contains information on car accidents and car insurance premiums from the National Highway Traffic Safety Administration and the National Association of Insurance Commissioners. [3]



The questions below should be answered based only on the image.

1) The scatter plot appears _____.

- to follow a linear form
- to not follow a linear form

2) The correlation coefficient is _____.

- -0.424
- 0.027
- 0.623
- 4.47

3) The least-squares regression _____ an appropriate model for the bad drivers data.

- is
- is not

PARTICIPATION ACTIVITY 7.2.7: The correlation coefficient.

1) Least squares regression is _____ appropriate for a dataset with a correlation coefficient close to 1.

- always
- sometimes
- never

2) A negative correlation coefficient implies _____.

- simple linear regression is not appropriate
- the data has a negative trend and will have a downward sloping least-squares regression line
- the data has a linear form and will have a negative intercept in the least-squares regression line

Exploring the Atlantic fiddler crab dataset.

Use the tool below to explore each target feature in the crab dataset using latitude as a predictor and examine various summary statistics for each pair of features.

Least squares regression in Python

Python can carry out simple linear regression by fitting a `LinearRegression()` object to `(X,y)` where `X` is an array of the observed values of the input features and `y` is an array of the associated observed values of the output features. The parameters and methods for `LinearRegression()` can be found in the LinearRegression documentation.

Additionally, Python can be used to calculate the coefficient of determination \(r^2\) by squaring the output of the function `r_regression()`. Often the array `y` will need to be reshaped using numpy's ravel function. The parameters for

`r_regression()` can be found in the r_regression documentation.

Simple linear regression in Python.

Full screen

The code below fits and graphs a simple linear regression model object using the crab data. The code then gives summary information about the model.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.
- Understand the use of the methods `fit`, `predict`, `intercept_`, `coef_`, and `score`.
- Note that the three ways to compute the proportion of variation explained by the least-squares regression model all yield the same value.

Your server is starting up.

You will be redirected automatically when it's ready for you.

Event log

PARTICIPATION ACTIVITY 7.2.8: Simple linear regression in Python.

Consider a dataset `A, b`, where `A` is an array of instances of the input features and `b` is an array of instances of the output features.

1) The code that initializes a linear regression object `L` would be _____.

Check Show answer

2) The code that fits a least-squares regression model `L` to the data `A,b` would be _____.

Check Show answer

3) The code that predicts the outcome for an instance with input feature value 15 using the linear regression model `L` would be _____.

Check Show answer

CHALLENGE ACTIVITY 7.2.1: Correlation and linear relationships.

537150.4174434.qx3zqy7

CHALLENGE ACTIVITY 7.2.2: Simple linear regression using scikit-learn.

537150.4174434.qx3zqy7

Start

John F. Kennedy International Airport (JFK) is a major airport serving New York City. JFK wanted to predict the arrival delay incoming flight based on the departure delay. 50 recent flights were randomly selected, and the arrival and departure delay minutes) were recorded.

- Initialize a linear regression model for predicting arrival delay based on departure delay.

The code contains all imports, loads the dataset, fits the regression model, and prints the model's intercept.

main.py | flightsJFK.csv


```
# Import packages and functions
import pandas as pd
from sklearn.linear_model import LinearRegression

# Import flights and remove missing values
flights = pd.read_csv('flightsJFK.csv').dropna()

# Define X and y and convert to proper format
X = flights[['dep_delay']].values.reshape(-1, 1)
y = flights[['arr_delay']].values.reshape(-1, 1)

# Initialize a linear regression model
```

```
linearModel = # Your code goes here

# Fit the linear model
linearModel = linearModel.fit(X, y)

```

1 2 3

Check Next level

(*1) Horst, Allison, and Julien Brun. "lterdatasampler: Educational dataset examples from the Long Term Ecological Research program." 2020. https://github.com/lter/lterdatasampler.

(*2) Johnson, D. 2019. "Fiddler crab body size in salt marshes from Florida to Massachusetts, USA at PIE and VCR LTER and NOAA NERR sites during summer 2016." ver 1. *Environmental Data Initiative.* https://doi.org/10.6073/pasta/4c27d2e778d3325d3830a5142e3839bb.

(*3) Kim, Albert Y., Chester Ismay, and Jennifer Chunn. "The fivethirtyeight R Package: 'Tame Data' Principles for Introductory Statistics and Data Science Courses." *Technology Innovations in Statistics Education 11* (2018). https://escholarship.org/uc/item/0rx1231m.

(*4) Anscombe, F. J. (1973). "Graphs in Statistical Analysis". *American Statistician.* 27 (1): 17-21. https://en.wikipedia.org/wiki/Anscombe's_quartet.

7.3 Linear regression assumptions

Learning goals

- List the assumptions of simple linear regression.
- Assess whether a fitted model satisfies each regression assumption.
- Explain how violated assumptions impact a regression model.
- Use Python to create residual plots.

Model assumptions

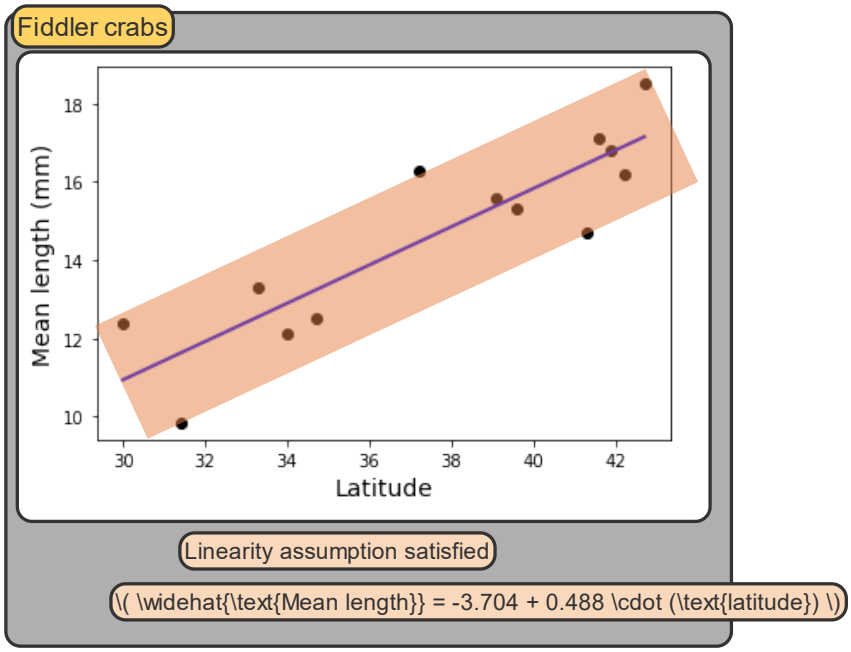
A simple linear regression can be computed for any pair of numeric features, but the resulting line is not necessarily a *good* fit for the relationship between those features. A simple linear regression makes certain assumptions about the relationship between the features x and y. Particularly, a simple linear regression model $\hat{y} = b_0 + b_1 x$ assumes:

- x and y have a linear relationship.
- The residuals of the observations are independent.

- The mean of the residuals is 0 and the variance of the residuals is constant.
- The residuals are approximately normally distributed.

Judging whether \(x\) and \(y\) have a linear relationship is the first step taken before conducting simple linear regression and can be carried out by checking whether the scatter plot of \(x\) and \(y\) shows a linear form.

PARTICIPATION ACTIVITY 7.3.1: Linearity of the fiddler crab dataset.



Animation content:

Step 1: A scatter plot graphing mean length in mm and latitude values appears.
Step 2: An upward tilted rectangular box appears over the points. The points stay within the box overall. The text "linearity assumption satisfied" appears.
Step 3: A line appears on the graph that follows the trend given by the box. The regression equation \(\widehat{\text{Mean length}} = -3.704 + 0.488 \cdot (\text{latitude}) \) is displayed.

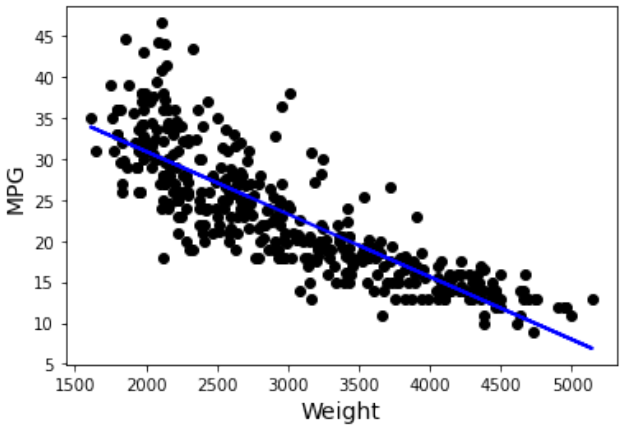
Animation captions:

1. The scatter plot shows location data and body measurements of groups of fiddler crabs.
2. Examining the scatter plot, the values appear to show a linear form overall, meaning the linearity assumption is satisfied.
3. A simple linear regression model can be created predicting mean length from latitude using a least squares regression line.

Data source: This dataset contains body measurements, location, and weather data on fiddler crabs found in the Eastern United States.[1] [2]

PARTICIPATION ACTIVITY 7.3.2: The linearity assumption for weight and miles per gallon.

The cars dataset contains information on weight and miles per gallon (MPG) for 392 cars.[3] Graphed below is the scatter plot displaying weight vs. miles per gallon along with a least squares regression line.



1) The linearity assumption _____.

- appears to hold
- does not appear to hold

2) A simple linear regression model _____ appropriate.

- is
- is not

3) The least squares line appears to model the data _____.

- poorly
- well

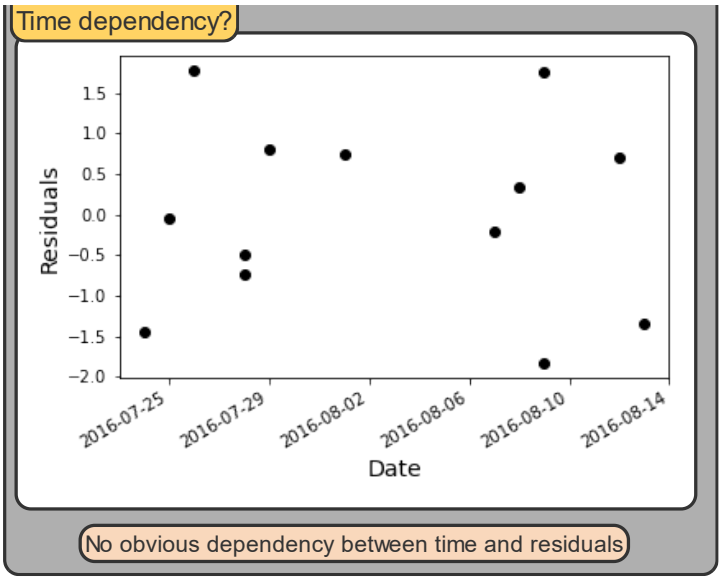
Independence of residuals

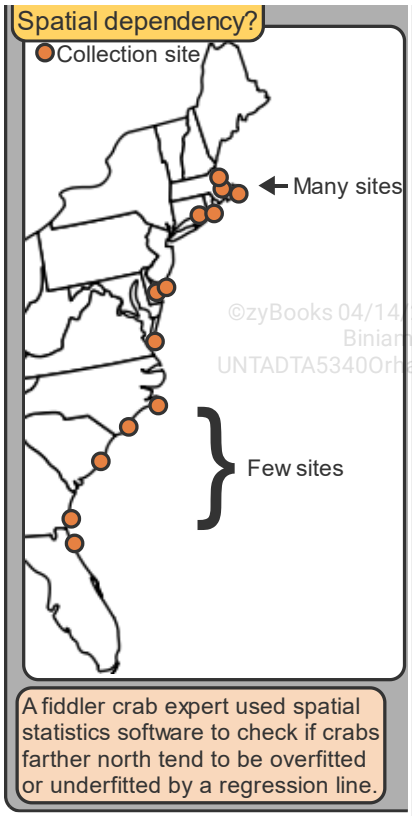
If two features do have a linear relationship, then the distances from the regression line to the points (residuals) should generally be random. A pattern in the residuals indicates the residuals are dependent, violating an assumption of linear regression. Whether residuals are independent is often affected by how the data were collected. Incomplete study design can affect regression models, causing a violation of the assumptions and inaccurate predictions. Ex: A study that aims to understand how individuals respond to different types of lighting, but only samples individuals from the same family, would not have independent residuals. This dependence between instances would likely be reflected in the residuals.

- Time dependence can often be assessed by analyzing the scatter plot of residuals over time.
- Spatial dependence can often be assessed by analyzing a map of where the data was collected along with further inspection of the residuals for spatial patterns.
- Dependencies between observational units must be assessed in context of the study.

For the fiddler crab dataset, the researchers were careful to collect data in a manner that avoided any dependencies between observational units. Possible time and spatial dependencies are investigated in the animation below.

PARTICIPATION ACTIVITY 7.3.3: Independence of residuals for the fiddler crab dataset.





Animation content:

Step 1: A scatter plot graphing the residuals over time appears. The points are spread throughout with no clear pattern.
Step 2: The text "No obvious dependency between time and residuals" appears.
Step 3: A map of the eastern United States appears. 1 dot is in Florida, 1 in Georgia, 1 in South Carolina, 2 in North Carolina, 1 in Virginia, 1 in Delaware, 1 in New Jersey, 1 in Connecticut, 1 in Rhode Island, and 3 in Massachusetts.
Step 4: An arrow appears pointing to the dots in Connecticut, Rhode Island, and Massachusetts with the text "many sites". A bracket appears containing the sites from Georgia to Virginia with the text "few sites". The text "A fiddler crab expert used spatial statistics software to check if crabs farther north tend to be overfitted or underfitted by a regression line" appears.

Animation captions:

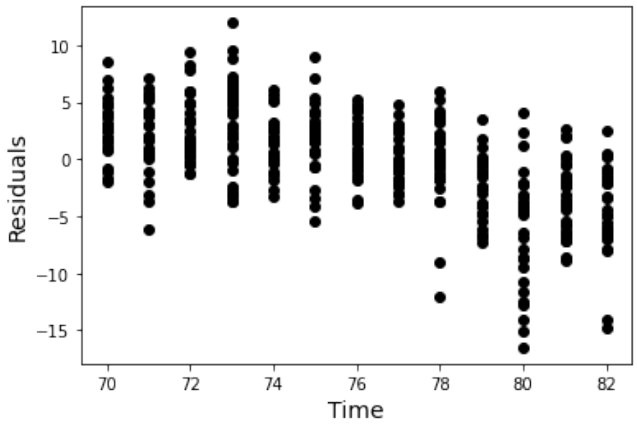
1. The dates of when researchers took crab measures at a given site were recorded. To investigate possible time dependency, a graph of the residuals over time can be analyzed.
2. The points do not appear to follow a pattern, so the data has no obvious time dependency.
3. The researchers also recorded the longitude and latitude of the collection sites. Spatial dependence is difficult to assess visually, but creating a map is a good first step.
4. The graph does raise a minor concern: the northeastern United States had more collection sites than the southeastern United States, so testing for spatial dependence should be explored.

Map image source: Wikimedia Commons

PARTICIPATION ACTIVITY

7.3.4: Independence of residuals for weight and miles per gallon.

The 392 instances in the cars dataset include cars from 1970 to 1982 from several countries. Graphed below is a plot of the residuals over time.



1) The graph of residuals over time indicates that the cars dataset _____ appear to have a time dependency.
   - ○ does
   - ○ does not

2) The cars dataset contains spatial information, so the model's residuals _____ be checked for spatial dependence.
   - ○ should
   - ○ should not

3) The cars dataset contains several instances of the same make and model car from different years. Ex: The dataset has the 1970, 1971, 1972, 1974, 1975, 1978, and 1980 models of Toyota Corona. The repeated use of the same make and model indicates that the cars dataset _____ appear to have dependencies between observational units.
   - ○ does
   - ○ does not

4) The independence of residuals assumption _____.
   - ○ appears to hold
   - ○ does not appear to hold

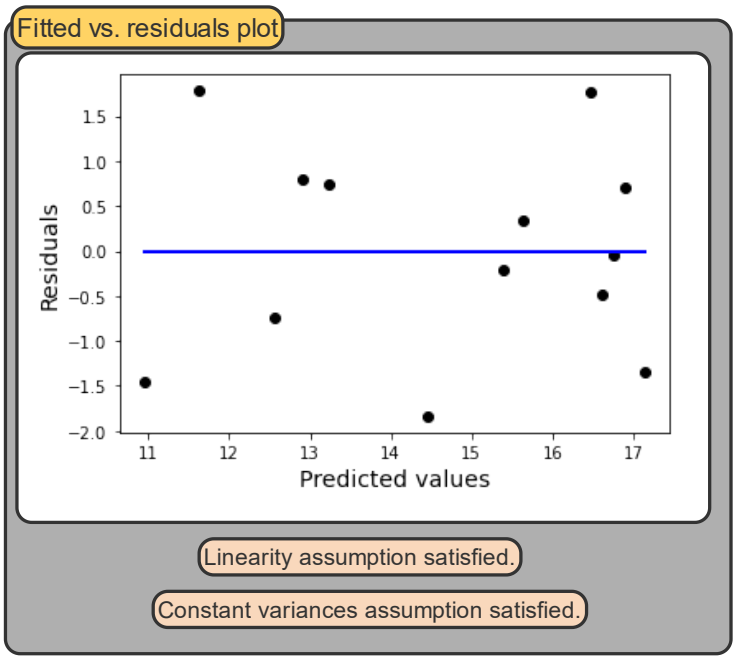
©zyBooks 04/14/24 12:43 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

Mean and variance of residuals

The residuals of a fitted model have a mean of 0 and should have a constant variance for all levels of the input \(x\). Since residuals measure the difference between the observed value and the predicted value for a given \(x\), a mean of 0 can be interpreted as the predicted value equal to the observed value on average. Ex: If several predictions for fiddler crabs' mean length were made, the predictions would be a mix of overestimates and underestimates that cancel each other out when averaging the residuals. Constant variance can be interpreted as predictions made at different levels of \(x\) that are similarly accurate. Ex: If the constant variance assumption holds, a fiddler crab's predicted mean length is, on average, equal to the value predicted from the regression using latitude, and the prediction is equally accurate at every latitude.

The constant variance assumption can be assessed by observing a scatter plot of the residuals. A ***fitted vs. residuals plot*** displays the predicted values \(\hat{y}\) on the horizontal axis and the residuals \((y_i - \hat{y}_i)\) on the vertical axis along with a horizontal line at \((y_i - \hat{y}_i = 0)\). If the residuals have constant variance, the points are similarly spread around the line for all values of \(\hat{y}\) without displaying any other form. Additionally, if the linearity assumption does not hold, then a nonlinear pattern should be present in the fitted vs. residuals graph as well.

PARTICIPATION ACTIVITY 7.3.5: Checking the constant variance assumption for the fiddler crab dataset.



Animation content:

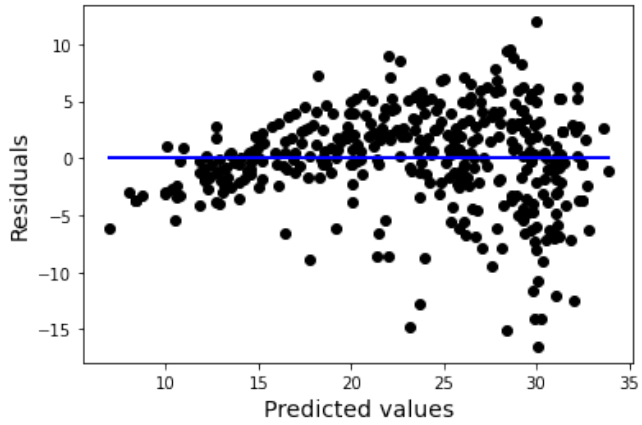
Step 1: A scatter plot graphing the predicted values vs. the residuals appears along with a horizontal line at \(\text{residual} = 0\).
Step 2: The text "Linearity assumption satisfied" appears.
Step 3: The text "Constant variances assumption satisfied" appears.

Animation captions:

1. To investigate the variance of residuals, the fitted vs. residuals plot can be analyzed.
2. A curve in the fitted vs. residuals plot indicates a deviation from linearity, but the mean of the residuals will be 0 because of the model's construction.
3. The points appear to be similarly spread around the line throughout the graph, so the constant variance assumption appears to be satisfied.

PARTICIPATION ACTIVITY 7.3.6: Mean and variance of residuals for weight and miles per gallon.

Graphed below is the fitted vs. residuals plot for the cars dataset.



1) The linearity assumption _____.

- appears to hold
- does not appear to hold

2) The constant variance assumption _____.

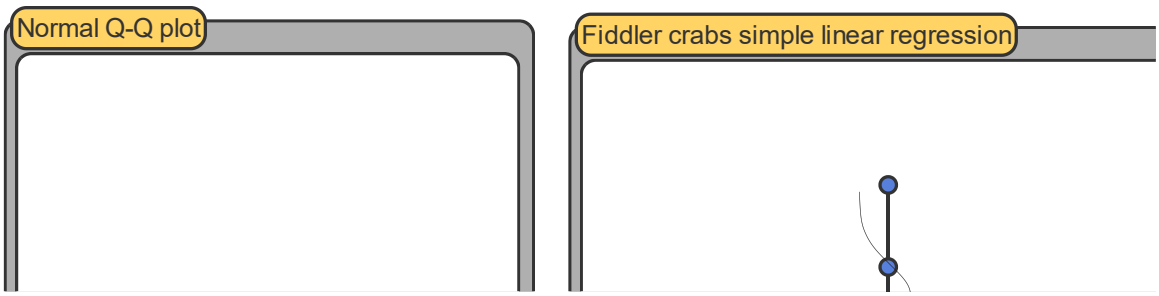
- appears to hold
- does not appear to hold

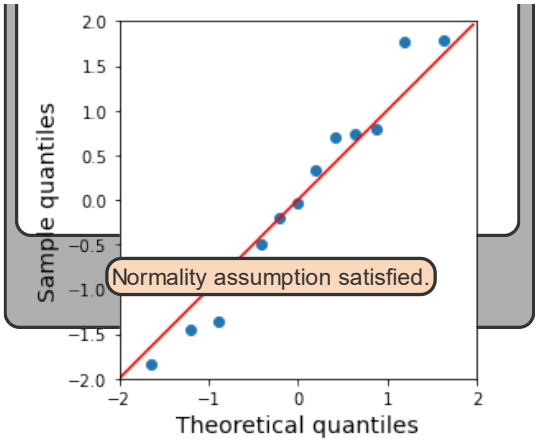
Normality of residuals

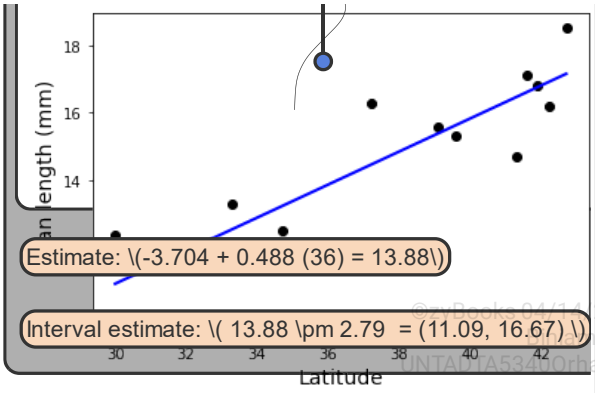
Mathematically, plugging into a linear regression equation tends to give good estimates when the previous assumptions roughly appear to hold. If additional residuals are normally distributed, interval estimates can be constructed as \(\text{estimate} \pm \text{margin of error}\). Furthermore, hypothesis testing can be conducted to measure the strength of evidence in supporting the argument that the linear relationship between \(x\) and \(y\) as modeled by the regression is statistically significant.

The normality assumption can be assessed by observing a graph of the residual quantiles compared to what the quantiles should be for normal residuals. A ***normal Q-Q plot*** displays quantiles the data would have if the data were normal on the horizontal axis, the actual observed quantiles from the sample data on the vertical axis, and a diagonal line where \(\text{theoretical quantile} = \text{sample quantile}\). If the residuals are approximately normal, then the points stay close to the diagonal line.

PARTICIPATION ACTIVITY 7.3.7: Normality of residuals for fiddler crab dataset.







Animation content:

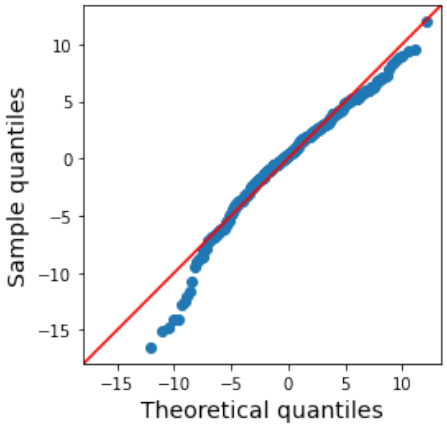
Step 1: A scatter plot graphing the theoretical quantiles versus the sample quantiles appears along with a diagonal line at \(\text{theoretical} = \text{sample}\).
Step 2: The text "Normality assumption satisfied" appears.
Step 3: The scatter plot of latitude vs. mean length appears, with a least squares regression line overlaid. A dot on the line at the point (36, 13.88) appears. The text "Point estimate: \(-3.704 + 0.488 (36) = 13.88\)" appears.
Step 4: A normal distribution appears horizontally, centered at the point (36, 13.88) . Points at (36, 11.09) and (36, 16.67) appear along with the text "Interval estimate: \(13.88 \pm 2.79 = (11.09, 16.67) \)".

Animation captions:

1. To investigate the normality of residuals, the normal Q-Q plot can be analyzed.
2. The points appear to be close to the line at \(\text{theoretical} = \text{sample} \), so the normality assumption appears to be satisfied.
3. The regression equation can provide a good estimation of the mean length of a fiddler crab at a given latitude by plugging in to the regression, regardless of whether the normality assumption is satisfied or not.
4. Since the normality assumption is satisfied, a normal distribution can be used to make interval estimates instead.

PARTICIPATION ACTIVITY 7.3.8: Normality of residuals for weight and miles per gallon.

Graphed below is the normal Q-Q plot for the cars dataset.



1) The points in the normal quantile plot _____ the diagonal line.
   - stay close to
   - do not stay close to to

2) The normality of residuals assumption _____.
   - appears to hold
   - does not appear to hold

3) A normal distribution _____ be used to generate interval estimates predicting miles per gallon from weight.
   - can
   - cannot

Residual plots in Python

In Python, `seaborn` and `statsmodels` can be used to produce residual plots. `sns.regplot(data=df, x='x', y='y')` plots a scatter plot of the features `x` and `y` with the linear regression equation overlaid. The fitted value vs. residual plot can be created using either `sns.regplot()` or `sns.scatterplot()`.

The `statsmodels` module contains functions for fitting and evaluating statistical models. `sm.qqplot(resid, line='45')` plots the standardized residuals against quantiles from a standard normal distribution. Using `line='45'` adds a line with a 45° angle, representing the line where the standardized residuals equal the normal quantiles.

Residual plots with Python.

Full screen

The code below fits a simple linear regression model object using the crab data. Residual plots are created using the fitted model.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.
- Use the plots to evaluate the linear regression assumptions for the crab data.

Your server is starting up.

You will be redirected automatically when it's ready for you.

Event log

PARTICIPATION ACTIVITY 7.3.9: Residual plots.

Suppose `X` and `y` are arrays containing input and output features, respectively. The predicted values of the linear regression model are assigned to `pred` and the residuals are assigned to `resid`. Complete the code to produce each residual plot.

1) Normal Q-Q plot

```
p =
(resid, line='45')
```

Check Show answer

2) Scatter plot with regression line

```
p =
(x=X, y=y)
```

Check Show answer

3) Scatter plot of fitted values vs. residuals

```
p = sns.regplot(
                )
```

Check Show answer

CHALLENGE ACTIVITY 7.3.1: Linear regression assumptions.

537150.4174434.qx3zqy7

(*1) Horst, Allison, and Julien Brun. "Iterdatasampler: Educational dataset examples from the Long Term Ecological Research program." 2020. https://github.com/lter/lterdatasampler.

(*2) Johnson, D. 2019. "Fiddler crab body size in salt marshes from Florida to Massachusetts, USA at PIE and VCR LTER and NOAA NERR sites during summer 2016." ver 1. *Environmental Data Initiative*.https://doi.org/10.6073/pasta/4c27d2e778d3325d3830a5142e3839bb.

(*3) Waskom, Michael. 2018. "mwaskom/seaborn-data/mpg.csv". Github repository.https://github.com/mwaskom/seaborn-data/blob/master/mpg.csv.

7.4 Multiple linear regression

Learning goals

- Define multiple linear regression.
- Use a fitted multiple regression model to make predictions.
- Interpret the intercept and slopes of multiple linear regression.
- Define polynomial regression and interaction terms.
- Implement a multiple regression model in `scikit-learn`.

Multiple linear regression

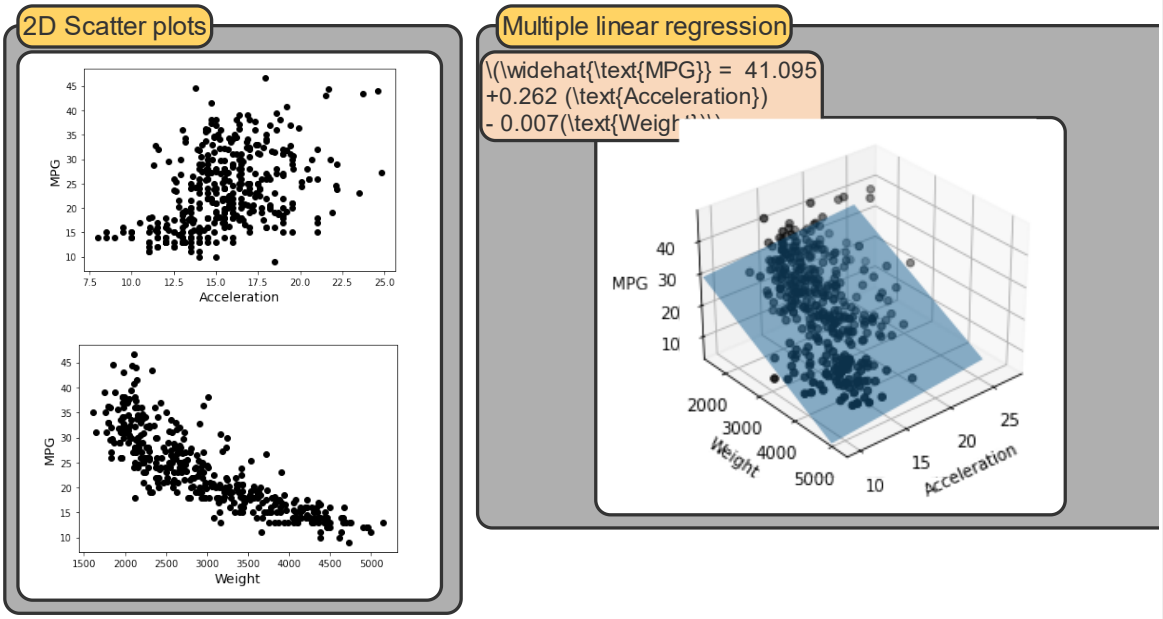
Often, a dataset has multiple input features, each of which could be used individually for simple linear regression. One way to build a better regression model is to incorporate more than one input feature into a single regression equation. A ***multiple linear regression*** is a mathematical model of the form \(\hat{y} = b_0 +b_1 x_1 + \ldots b_k x_k\), in which:

- \(x_1, \ldots , x_k\) are input features.

- $\hat{y}$ is the predicted value of the output feature y for given values of $x_1, \ldots , x_k$.
- b_0 is the y-intercept, representing the predicted value of y when all input features equal zero ($x_1 = \ldots = x_k = 0$).
- $b_1, \ldots, b_k$ are the slopes, representing how much the predicted value of y changes per a one-unit increase in the associated input feature when all other features are held constant .

Selecting which features to use in a multiple linear regression generally requires a method of model evaluation to ensure that only relevant input features are included.

PARTICIPATION ACTIVITY 7.4.1: A multiple linear regression predicting miles per gallon.



Animation content:

Step 1: A scatter plot of acceleration vs miles per gallon appears, displaying a positive relationship. A scatter plot of weight vs miles per gallon appears, displaying a negative relationship between them.
Step 2: The multiple linear regression equation $\widehat{\text{MPG}} = 41.095 + 0.262 (\text{Acceleration}) - 0.007(\text{Weight})$ appears.
Step 3: A 3d graph appears with axes weight, acceleration, and miles per gallon. Black dots representing the observed values appear. A blue plane appears that matches the overall pattern of the dots.

Animation captions:

1. Both the acceleration and weight of a car could each be used to predict miles per gallon.
2. A multiple linear regression using both acceleration and weight yields an equation predicting miles per gallon that takes both features as input.
3. A multiple linear regression of two inputs can be graphed as a 3D plot, in which case the regression traces out a plane.

Data source: This dataset contains information on cars sold from 1970 through 1982. [1]

PARTICIPATION ACTIVITY 7.4.2: Interpreting the multiple linear regression predicting miles per gallon.

The multiple linear regression predicting fuel efficiency (in miles per gallon) from acceleration (in seconds for a car to go from 0 miles per hour to 60 miles per hour) and weight (in pounds) is given by: $$\widehat{\text{MPG}} = 41.095 + 0.262 (\text{Acceleration}) - 0.007(\text{Weight})$$

1) The predicted miles per gallon for a car having an acceleration value of 20 seconds and a weight of 3000 pounds is _____.
   - -15.76
   - 25.335
   - 826.965

2) The coefficient +0.262 represents _____.
   - the positive relationship between acceleration and miles per gallon
   - the predicted miles per gallon of a car with acceleration and weight equal to 0
   - how much the predicted miles per gallon changes for a one-unit change in acceleration

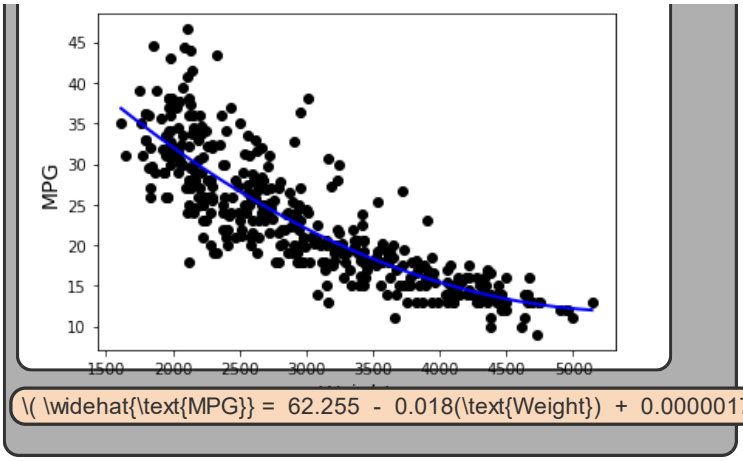
3) A car with a low acceleration and high weight has a relatively _____ predicted miles per gallon.
   - low
   - high

Simple polynomial regression

One special case of multiple linear regression is to include powers of a single features as inputs in the regression equation. When doing so, a regression model is built that has the form of a polynomial equation in x. A ***simple polynomial regression*** is a mathematical model of the form $\hat{y} = b_0 +b_1 x + b_2 x^2 + \ldots b_k x^k$.

A polynomial function is not a linear function of x but the regression is still considered a type of multiple linear regression since $\hat{y}$ is linear *with respect to the coefficients b_i.*

PARTICIPATION ACTIVITY 7.4.3: A simple polynomial regression predicting miles per gallon.



Animation content:

Step 1: A scatter plot of weight versus miles per gallon appears. The points show a quadratic pattern.
Step 2: The equation \(\widehat{\text{MPG}} = 62.255 - 0.018(\text{Weight}) + 0.0000017 (\text{Weight})^2\) appears.
Step 3: A parabola is added to the scatter plot.

Animation captions:

1. The scatter plot of weight vs. miles per gallon shows a non-linear, somewhat quadratic pattern overall.
2. Taking a multiple linear regression using the features \((\text{weight}) \) and \((\text{weight})^2 \) gives a simple polynomial regression predicting miles per gallon using a quadratic function.
3. The simple polynomial regression traces out a parabola in the scatter plot, fitting the overall form.

PARTICIPATION ACTIVITY | 7.4.4: Interpreting the simple polynomial regression predicting miles per gallon.

The simple polynomial regression predicting fuel efficiency (in miles per gallon) from weight (in pounds) is given by: \[\widehat{\text{MPG}} = 62.255 - 0.018(\text{Weight}) + 0.0000017 (\text{Weight})^2 \]

1) The predicted miles per gallon for a car having a weight of 3000 pounds is _____.

- -53.995
- 8.260
- 23.555

2) A simple polynomial regression should be used when a scatter plot displays _____.

- a weak linear form
- a strong linear form
- a nonlinear form

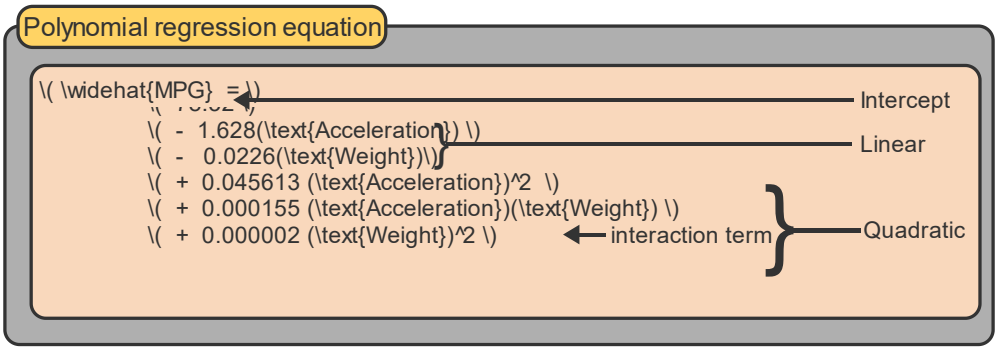
Polynomial regression

A polynomial regression can be considered using multiple input features. A ***polynomial regression*** is a mathematical model of input features that includes all powers and interaction terms of the input features up to a fixed degree. An ***interaction term*** is a term in a regression model that contains multiple input features, such as \((x_1^3 x_2) \). Interaction terms are included to reflect how the prediction changes in non-additive ways as the features vary. The degree of an interaction term is the sum of the powers on the term. Ex: \((x_1^3 x_2) \) is a degree 4 interaction term. Interactive terms can account for dependency between input features.

The general equation of a multiple polynomial equation is cumbersome to write but can be easily displayed for a small degree. Ex: A degree 2 (quadratic) polynomial regression on two input features has the form \(\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 x_1 x_2 + b_5 x_2^2 \).

Like simple polynomial regression, the function used for the regression is not linear, but the regression is said to be linear with respect to the powers and interaction terms.

PARTICIPATION ACTIVITY 7.4.5: A polynomial regression predicting miles per gallon.



Animation content:

Step 1: The equation \(\widehat{\text{MPG}} = 78.32 - 1.628(\text{Acceleration}) - 0.0226(\text{Weight}) + 0.045613 (\text{Acceleration})^2 + 0.000155 (\text{Acceleration}) (\text{Weight}) + 0.000002 (\text{Weight})^2 \) appears.
Step 2: The term \(78.32\) is labeled "intercept".
Step 3: The terms \(- 1.628(\text{Acceleration})\) and \(- 0.0226(\text{Weight})\) are labeled "linear".
Step 4: The term \(+ 0.000155 (\text{Acceleration})(\text{Weight})\) is labeled "interaction term". The terms \(+ 0.045613 (\text{Acceleration})^2\), \(+ 0.000155 (\text{Acceleration})(\text{Weight}) \), and \(+ 0.000002 (\text{Weight})^2\) are labeled "quadratic".

Animation captions:

1. A degree 2 (quadratic) polynomial regression predicting miles per gallon from acceleration and weight has six terms.
2. 78.32 is the intercept term, sometimes called the degree 0 term.
3. The next two terms are linear terms, which are all degree 1 terms.
4. The next three terms are the quadratic terms, which are all degree 2 terms. The quadratic terms include the squares of the input features and an interaction term.

PARTICIPATION ACTIVITY 7.4.6: Interpreting the polynomial regression predicting miles per gallon.

1) If instead a degree 3 (cubic) polynomial regression were used, the terms _____ would be included in addition to the terms in the quadratic regression.

- \((\text{Acceleration})^3\) and \ ((\text{Weight})^3\)
- \((\text{Acceleration})^3\), \((\text{Acceleration})^3(\text{Weight})^3\) and \((\text{Weight})^3\)
- \((\text{Acceleration})^3\), \((\text{Acceleration})^2(\text{Weight})\), \ ((\text{Acceleration})(\text{Weight})^2\), and \((\text{Weight})^3\)

Discovering the mpg data: multiple linear regression.

Use the drop-down menus to select two input features and the model's polynomial degree.

Multiple linear regression in Python

Python carries out multiple regression by fitting a `LinearRegression()` object to `(X,y)` where `X` is an array of the observed values of the input features and `y` is an array of the associated observed values of the output features. This is similar to fitting a simple linear model using least squares regression, except that for multiple linear regression, the array `X` has a column for each input feature.

Python also uses a `LinearRegression()` object to fit a simple or multiple polynomial regression model by first creating an array where the columns are powers of the input features and interaction terms and then feeding that array in as `X`. The array can be generated using `PolynomialFeatures()`. The parameters and methods for `PolynomialFeatures()` can be found in the PolynomialFeatures documentation.

Multiple linear regression in Python.

Full screen

The code below fits three multiple linear regression model objects using the cars data. The first model has two input features, the second model is a simple polynomial regression on one feature, and the third is a polynomial regression on two features. Predictions are made using the models that will vary slightly from the predictions calculated by hand in the section due to the code using many more decimal places throughout.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.
- Understand the use of the methods `fit`, `predict`, `intercept_`, and `coef_` for `LinearModel()`.
- Understand the use of the `fit_transform` method of `PolynomialFeatures()` to create an array of powers and interaction terms of the features to fit a polynomial model.

Your server is starting up.

You will be redirected automatically when it's ready for you.

2024-04-14T17:43:40Z [Normal] Started container git-sync

Event log

©zyBooks 04/14/24 12:43 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY 7.4.7: Multiple linear regression in Python.

1) A `LinearRegression()` object _____ be used to fit polynomial regression models.

- can
- cannot

2) If `poly` was initialized as a degree 2 `PolynomialFeatures()` object and `X` is an array of 3 columns representing features \(a\), \(b\), and \(c\), then `XPoly = poly.fit_transform(X)` will have _____ columns.

- 3
- 6
- 9

CHALLENGE ACTIVITY 7.4.1: Multiple linear regression.

537150.4174434.qx3zqy7

CHALLENGE ACTIVITY 7.4.2: Multiple linear regression using scikit-learn.

537150.4174434.qx3zqy7

Start

Newark Liberty International Airport (EWR) is a major airport serving New York City. EWR wanted to predict the arrival delay incoming flight based on the departure delay. 50 recent flights were randomly selected, and the arrival delays (in minutes) recorded.

- Initialize a multiple regression model for predicting arrival delay based on departure delay and flight distance.

The code contains all imports, loads the dataset, fits the regression model, and prints the model's intercept.

main.py | flightsEWR.csv

```
# Import packages and functions
import pandas as pd
from sklearn.linear_model import LinearRegression

# Import flights and remove missing values
flights = pd.read_csv('flightsEWR.csv').dropna()

# Define X and y and convert to proper format
X = flights[['dep_delay', 'distance']].values.reshape(-1, 2)
y = flights[['arr_delay']].values.reshape(-1, 1)

# Initialize a linear regression model
multipleModel = # Your code goes here

# Fit the linear model
multipleModel = multipleModel.fit(X, y)

```

1 2 3

Check Next level

(*1) Waskom, Michael. 2018. "mwaskom/seaborn-data/mpg.csv". Github repository.https://github.com/mwaskom/seaborn-data/blob/master/mpg.csv.

7.5 Logistic regression

Learning goals

- Define logistic regression.
- Interpret the intercept and slope of a logistic regression model.
- Use a fitted logistic regression model to predict the probability and class of an outcome.
- Use a fitted logistic regression model to make a classification.
- Define and interpret the log-odds and odds ratio.
- Implement a logistic regression model using `scikit-learn`.

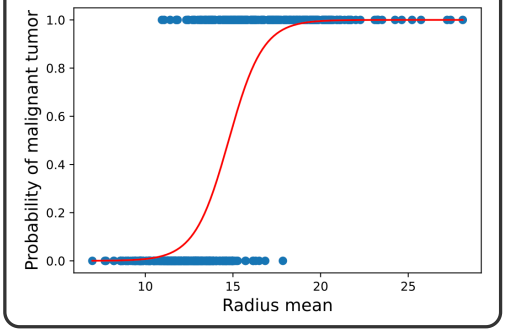
Introduction to logistic regression

A ***logistic regression*** is a model that predicts the probability of an outcome in a binary category using an equation of the form $\hat{p} = \displaystyle \frac{e^{b_0 + b_1 x} }{1 + e^{b_0 + b_1 x }}$, where b_0 and b_1 are values estimated using the data. A logistic regression model graphs an S-shaped curve (a sigmoid). Although a logistic regression model is an example of nonlinear regression, the parameters b_0 and b_1 play a role similar to the parameters of simple linear regression and use the same names.

- b_0 is called the intercept parameter. Unlike linear regression, the b_0 is *not* the y-intercept, but instead *determines* probability of observing a success when all the input features are 0, or baseline success probability. A positive value of b_0 indicates a probability greater than 0.5 while a negative value of b_0 indicates a probability less than 0.5.
- b_1 is the slope parameter that determines how sharply the logistic curve bends as the value of the input feature increases. A positive slope parameter means the curve will increase from 0 to 1 as the input feature increases, while a negative slope parameter indicates the curve will decrease from 1 to 0 as the input feature increases.

PARTICIPATION ACTIVITY 7.5.1: Parameters of the breast cancer logistic regression.

Logistic regression \(\hat{p}\)



Logistic regression equation

\(\hat{p} = \displaystyle \frac{e^{-15.12 + 1.

Intercept parameter \(b_0\)

\(b_0 = -15.12\)

Since \(b_0 < 0\), the intercept is less than \(0.5\).

\(\hat{p}(0) = \displaystyle \frac{e^{-15.12 + 1.02 (0)}}{1 + e^{-15.12 + 1.02 (0)}} \)

\(= 0.0000002713\)

Slope parameter \(b_1\)

\(b_1 = 1.02\)

Since \(b_1 > 0\), the S-curve increases from \(0\) to \(

Animation content:

Step 1: A box titled "Logistic regression \(\hat{p}\)" appears. A logistic regression predicting probability of malignant diagnosis from radius mean appears. All instances are between 0 and 1. The graph increases from 0 to 1 as radius mean increases from 5 to 30. A box titled "Logistic regression equation" appears and the equation \(\hat{p} = \displaystyle \frac{e^{-15.12 + 1.02 (\text{radius mean})}}{1 + e^{-15.12 + 1.02 (\text{radius mean})}} \) appears inside the box.
Step 2: A box titled "Intercept parameter" appears. \(b_0 = -15.12\) appears inside the box along with the text "Since \(b_0\) is less than 0, the intercept is less than 0.5". The equation \(\hat{p}(0) = \displaystyle \frac{e^{-15.12 + 1.02 (0)}}{1 + e^{-15.12 + 1.02 (0)}} = 0.0000002713\) appears.
Step 3: A box titled "Slope parameter" appears. \(b_1 = 1.02\) appears inside the box along with the text "Since \(b_1\) is greater than 0, the S-curve increases from 0 to 1."

Animation captions:

1. A logistic regression predicting the probability a tumor is malignant can be fit from data that recorded radius mean for malignant and benign tumors.
2. The intercept parameter is \(b_0 = -15.12\) and can be used to calculate the intercept as (0.0000002713). The intercept isn't meaningful here, however, as a tumor can't have a radius of 0.
3. The slope parameter is \(b_1 = 1.02\) and indicates that the curve increases from 0 to 1.

Data source: Researchers at the University of Wisconsin collected data from image scans of breast mass cells, recording the tumor radius for each image. Each image came from a biopsy of a benign or malignant breast tumor. [7] [8]

PARTICIPATION ACTIVITY 7.5.2: Interpreting logistic regression parameters.

Consider a logistic regression predicting the probability a tumor is malignant based off of the perimeter mean given by

$$\hat{p} = \displaystyle \frac{e^{-15.71 + 0.16 (\text{perimeter mean})} }{1 + e^{-15.71 + 0.16 (\text{perimeter mean})}}.$$

1) The intercept of the curve should be _____.
   - less than 0.5
   - equal to -15.71
   - greater than 0.5

2) The probability of a tumor being malignant _____.
   - decreases from 1 to 0 as perimeter mean increases
   - increases from 0 to 1 as perimeter mean increases

Classification using a logistic regression

Although more complicated, using nonlinear equations allows a broader class of scenarios to be modeled. One such scenario is classification: a nonlinear regression model can be used to predict the outcome of a categorical feature based on numeric features. In order to do so, categorical features must first be made numeric. ***Hot encoding*** is transforming a categorical feature into numeric feature that equal 0 when an instance is not in the category and 1 when the instance is.

A binary categorical feature hot encoded into a numeric feature p can be interpreted as a probability where the outcomes are either 0% or 100% chance of being in the category. When the predicted probability $\hat{p}$ is greater than some cutoff value, the outcome is predicted to be in the category, otherwise the outcome is predicted to not be in the category. The cutoff value is based on the scenario being modeled, but usually 0.5 is used in classification.

Table 7.5.1: Hot encoding the diagnosis feature from the Wisconsin breast cancer diagnosis dataset.

Malignant tumors are hot-encoded as 1 and benign tumors are hot encoded as 0. The predicted probability $\hat{p}$ for each tumor is obtained using the logistic regression model

$$\hat{p} = \dfrac{e^{-15.12 + 1.02 \text{(radius mean)}}}{1+e^{-15.12 + 1.02 \text{(radius mean)}}}.$$

Using a probability cutoff of 0.5, tumors with predicted probabilities greater than or equal to 0.5 will be classified as 1. Tumors with predicted probabilities less than 0.5 will be classified as 0.

Radius mean	Diagnosis	Observed class, p	Predicted probability, $\hat{p}$	Predicted class
17.99	M	1	0.965	1
20.57	M	1	0.997	1
19.69	M	1	0.994	1
11.42	M	1	0.032	0
...	...	...	...	

20.60	M	1	0.998	1
7.76	B	0	0.001	0

Predicting whether a tumor is malignant or benign using a logistic regression model.

A logistic regression model predicts the probability a tumor is malignant from the radius mean (mm). The probability cutoff determines whether the tumor is classified as malignant or benign. Use the slider to adjust the probability cutoff.

©zyBooks 04/14/24 12:43 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024

PARTICIPATION ACTIVITY | 7.5.3: Predicting whether a tumor is malignant or benign using a logistic regression model.

1) A malignant tumor with radius mean 16 would be represented on the graph as a point at _____.
   - (16, M)
   - (16, 0)
   - (16, 1)

2) Consider a patient who had a tumor with a radius mean of 13. The logistic regression calculated that \(\hat{p}(13) = 0.142\), meaning the tumor is _____.
   - benign
   - 14.2% malignant
   - predicted to be benign

Odds, log-odds, and odds ratio

Writing a logistics model in terms of a linear function can be useful in interpreting the slope and intercept parameters \(b_0\) and \(b_1\). The ***log-odds*** function is obtained by taking the natural logarithm, denoted \(\ln()\), of the odds for the positive outcome of an experiment or study. The resulting expression is a linear function in terms of \(b_0\) and \(b_1\),

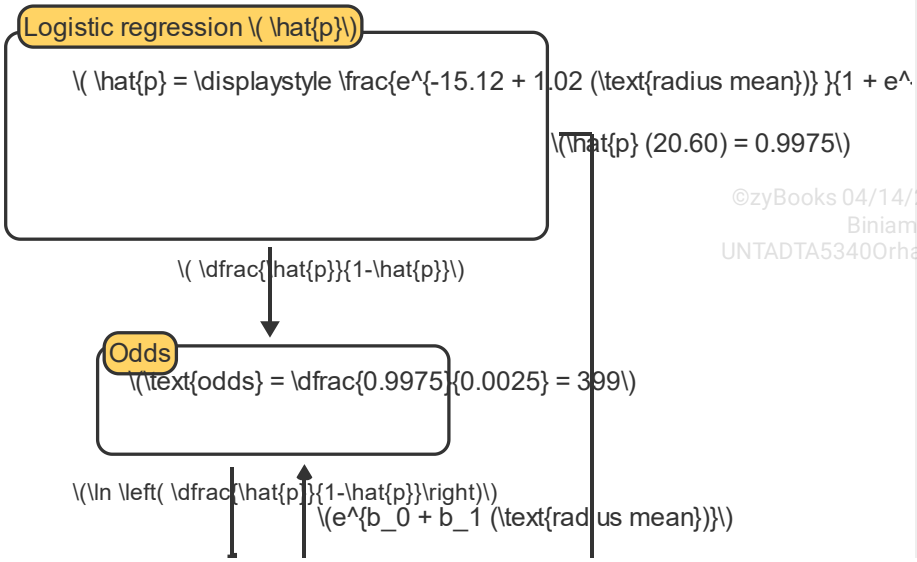
\[\hat{g} = \ln \left(\frac{\hat{p}}{1 - \hat{p}} \right) = b_0 + b_1 x .\]

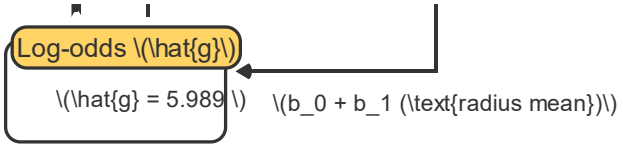
Odds are used in the field of medicine frequently, because identifying relationships between a feature and an outcome is important. The ***odds ratio*** compares the relative odds of an outcome given a feature. An odd's ratio of 1 means that the feature does not affect the outcome.

The following results gives a summary of the interpretation of \(b_0\) and \(b_1\).

- \(b_0\) represents the log-odds when the feature is 0.
- \(e^{b_0}\) represents the odds when the feature is 0.
- \(b_1\) represents the log-odds ratio for a unit change in the feature.
- \(e^{b_1}\) represents the odds ratio for a unit change in the feature.

PARTICIPATION ACTIVITY 7.5.4: Logistic regression and log-odds.





Animation content:

Step 1: A logistic regression model predicts a probability of 0.9975 when the radius mean is 20.60. The logistic regression equation \(\hat{p} = \frac{e^{-15.12+1.02(radius mean)}}{1 + e^{-15.12+1.02(radius mean)}}\) and \(\hat{p}(20.60) = 0.9975\) are displayed.
Step 2: The odds of a tumor being malignant is the probability of a random tumor being malignant \(\hat{p}\) divided by the probability of a random tumor being benign \((1−\hat{p})\). An arrow from the output of the previous equation with the label \(\frac{\hat{p}}{1 - \hat{p}}\) appears. This arrow leads to the equation \(\text{odds} = \frac{0.9975}{0.0025} = 399\).
Step 3: The log-odds \(\hat{g}\) can be found by taking the natural log of the odds. An arrow from the output of the previous equation with the label \(\ln \frac{\hat{p}}{1 - \hat{p}}\) appears. This arrow leads to the value of the log-odds \(\hat{g} = 5.989\).
Step 4: The same value can also be obtained without finding the probability or the odds using the equation \(\hat{g} =−15.12+1.02(radius mean)\). An arrow that connects the logistic regression equation with label \(b_0 + b_1 (radius mean)\) to the log-odds \(\hat{g} = 5.989\) appears.
Step 5: Exponentiating the log-odds gives the odds. Here, \(e^5.989=399\). An arrow that goes from the log-odds \(\hat{g} = 5.989\) to odds \(odds = 399\) with label \(e^{b_0 + b_1 (radius mean)}\) appears.

Animation captions:

1. A logistic regression model predicts a probability of \(0.9975\) when the radius mean is \(20.60\).
2. The odds of a tumor being malignant is the probability of a random tumor being malignant (\(\hat{p}\)) divided by the probability of a random tumor being benign \((1-\hat{p})\).
3. The log-odds \(\hat{g}\) can be found by taking the natural log of the odds.
4. The same value can also be obtained without finding the probability or the odds using the equation \(\hat{g} = -15.12 + 1.02 (\text{radius mean})\).
5. Exponentiating the log-odds gives the odds. Here, \(e^{5.989} = 399.\)

PARTICIPATION ACTIVITY 7.5.5: Log-odds.

1) A slope parameter of \(b_1 = 1.02\) means that a unit increase in radius mean is associated with _____ of a tumor being malignant.

- a 1.02 unit increase in the log-odds
- a 1.02% increase in the probability
- an \(e^{1.02}\)% increase in the probability

2) The resulting odds ratio for a unit change in radius mean is _____.

- 1.02
- 2.02
- 2.77

3) The resulting odds ratio indicates that _____.

- radius mean does not affect the odds that a tumor is malignant
- radius mean is associated with higher odds that a tumor is malignant
- radius mean is associated with lower odds that a tumor is malignant

Logistic regression in Python

Python can carry out logistic regression by fitting a `LogisticRegression()` object to `(X,y)` where `X` is an array of the observed values of the input feature and `y` is an array of 0s and 1s representing the binary categorical variable. The parameters and methods for `LogisticRegression()` can be found in the LogisticRegression documentation. If the categorical variable has not been hot encoded, values can be reassigned from labels to 0 or 1 using `pandas` DataFrame operations.

Logistic regression in Python.

Full screen

The code below fits and graphs a logistic regression model object using the Wisconsin breast cancer data, hot encoding the diagnosis variable along the way. The log-odds linear classifier is also graphed for comparison.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.
- Understand the use of DataFrame manipulation to hot encoding the diagnosis variable.
- Understand the use of the methods `fit` to fit a logistic regression.
- Understand the use of the methods `predict` and `predict_proba` to make predictions from a logistic regression.
- Understand the use of the methods `intercept_` and `coef_` to get the parameter of the logistic regression and log-odds model.

Your server is starting up.

You will be redirected automatically when it's ready for you.

Event log

PARTICIPATION ACTIVITY 7.5.6: Logistic regression in Python.

Consider a dataset `A, b`, in which `A` is an array of instances of the input features and `b` is a numeric array with entries 0 (representing false) and 1 (representing true).

1) The code that initializes a logistic regression object `L` would be _____.

Check Show answer

2) The code that fits a logistic regression model `L` to the data `X, y` would be _____.

Check Show answer

3) The code that predicts the probability that an instance with input value 25 is false/true using the logistic regression model `L` would be _____.

Check Show answer

CHALLENGE ACTIVITY 7.5.1: Logistic regression.

537150.4174434.qx3zqy7

CHALLENGE ACTIVITY 7.5.2: Logistic regression using scikit-learn.

537150.4174434.qx3zqy7

Start

The US Forest Service regularly monitors weather conditions to predict which areas are at risk of wildfires. Data scientists working with the US Forest Service would like to predict whether a wildfire will occur based on humidity.

- Fit the logistic regression model, `logisticModel`, to predict whether a wildfire will occur.

The code contains all imports, loads the dataset, and prints the model coefficients.

main.py | fires.csv

```
# Import packages and functions
import pandas as pd
import numpy as np
from sklearn.linear_model import LogisticRegression

# Load the dataset
fires = pd.read_csv('fires.csv')

# Create input matrix X and output matrix y
X = fires['humidity'].values.reshape(-1, 1)
y = np.ravel(fires['fire'])

# Define and fit the logistic regression model
logisticModel = LogisticRegression()
# Your code goes here

# Print the estimated coefficients
```

1 2 3

Check Next level

(*7) Dua, Dheeru, and Casey Graff. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, 2017. http://archive.ics.uci.edu/ml.

(*8) Street, W.N., W.H. Wolberg and O.L. Mangasarian. "Nuclear feature extraction for breast tumor diagnosis." IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, 1905 (1993): 861-870. https://doi.org/10.1117/12.148698.

7.6 Case study: Energy consumption

Learning goals

- Use simple linear regression to describe relationships and make predictions.
- Use multiple linear regression to describe relationships and make predictions.
- Assess whether a fitted simple linear regression satisfies the model assumptions.
- Implement and evaluate regression models using `scikit-learn` and `seaborn`.

Modeling energy consumption

The US Energy Information Administration estimates that about half of energy consumed in American homes is for two purposes: heating and air conditioning. Energy consumption is highly seasonal, with high demand for heating in the winter and cooling in the summer. About half of residential buildings in the United States rely on natural gas for fueling furnace systems, water heaters, and cooking equipment. Since the demand for natural gas is largely seasonal, so are prices.

Homeowners considering installing home energy sources, such as solar panels, may be interested in their current energy usage. The gas usage dataset contains natural gas usage (in therms), solar energy production, and temperature over an 18-month period from a homeowner in the midwestern United States.

Exploring natural gas use. Full screen

The Python code below imports natural gas and solar panel production data from a homeowner in the Midwest and creates several plots.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below. When does gas usage tend to be highest? When does solar production tend to be highest?

Your server is starting up.

You will be redirected automatically when it's ready for you.

Event log

PARTICIPATION ACTIVITY 7.6.1: Residential natural gas use.

1) What months does natural gas use tend to be highest for this homeowner?

- January-March
- March-May
- May-July

2) At what temperatures does natural gas use tend to be highest for this homeowner?

- Below 20 degrees Fahrenheit
- Between 20 and 40 degrees Fahrenheit
- Above 40 degrees Fahrenheit

3) Which feature appears to be a better predictor of natural gas use: date or temperature?

- Date
- Temperature

Modeling natural gas use based on temperature

A negative association exists between natural gas use and temperature. Since the association is linear, a linear regression model could be used to describe the relationship and predict gas use. After removing days with no natural gas use from the dataset, a linear regression model was fit.

Modeling natural gas use. Full screen

The Python code below imports natural gas and solar panel production data from a homeowner and fits a linear regression model.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.

Your server is starting up.

You will be redirected automatically when it's ready for you.

Event log

PARTICIPATION ACTIVITY 7.6.2: Modeling natural gas use.

1) For a 10-degree increase in temperature, the predicted decrease in natural gas use is _____.
   - -0.0775
   - 0.775
   - -0.775

2) Predict the natural gas use on a 32-degree day.
   - 1.884 therms
   - 3.589 therms
   - 4.364 therms

3) Predict the natural gas use on a 70-degree day.
   - -1.061 therms
   - 0 therms
   - 2.728 therms

Evaluating the regression model

Linear regression models make certain assumptions about the relationship between the input and output features. Once the regression model has been fitted, the next step is to evaluate the regression assumptions.

Evaluating the natural gas model. Full screen

The Python code below imports natural gas data from a homeowner, fits a linear regression model, and creates residual plots for the model.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.

Your server is starting up.

You will be redirected automatically when it's ready for you.

Event log

PARTICIPATION ACTIVITY 7.6.3: Evaluating the natural gas model.

1) The linearity assumption _____.
   - appears to hold
   - does not appear to hold

2) The constant variance assumption _____.
   - appears to hold
   - does not appear to hold

3) The normality of errors assumption _____.
   - appears to hold
   - does not appear to hold

Considering polynomial models

Despite the decreasing trend in natural gas use based on temperature, a linear model does not appear to be the best approach. Since a curve exists in the scatter plot of temperature and natural gas, a polynomial regression model may be a

better choice.

Expanding the natural gas model.

Full screen

The Python code below imports natural gas data from a homeowner and plots several polynomial regression models.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.

Your server is starting up.

You will be redirected automatically when it's ready for you.

Event log

PARTICIPATION ACTIVITY 7.6.4: Multiple regression for gas usage.

1) Which of the following is an advantage of using a polynomial model for predicting gas use?

- Polynomial models avoid extrapolation.
- Polynomial models are easier to interpret.
- Polynomial models are better able to model curved relationships.

2) How else could the regression model be improved?

- Add input features, such as low temperature or weather conditions.
- Remove input features
- Increase the polynomial degree

7.7 Case study: Customer churn

Learning goals

- Compare input features for logistic regression.
- Interpret and calculate predictions using logistic regression with one input feature.
- Interpret and calculate predictions using logistic regression with multiple input features.
- Implement and examine logistic regression models using `scikit-learn` and `seaborn`.

Measuring customer satisfaction

Companies and service providers are interested in growing and retaining a loyal customer base. Keeping customers happy generates more revenue for a company, with less money spent on advertising or incentives. Customer churn, or annual turnover, is a measure of customer satisfaction. Ex: A customer satisfied with their Internet service is not likely to switch providers, but an unhappy customer may look for a new Internet service provider in the next year.

Companies that provide regular services, such as banks, are especially interested in maintaining low customer churn. A European bank randomly sampled 10,000 customers and recorded each customer's age, gender, credit score, country, estimated income, and length of time spent with the bank. The bank also calculated each customer's average bank balance during the previous year, and whether or not the customer closed at least one account (`churn = 1` if a customer closed at least one account in the previous year).

Exploring the customer churn dataset.

Full screen

The Python code below imports the customer churn dataset, and explores the data using tables and plots.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.

Your server is starting up.

You will be redirected automatically when it's ready for you.

Event log

Data source: Keldine Malit. 2018. "Bank Customer Churn Prediction". https://www.kaggle.com/code/kmalit/bank-customer-churn-prediction

PARTICIPATION ACTIVITY 7.7.1: Exploring customer churn.

1) Estimate the proportion of customer churn.

- 20%
- 50%
- 80%

2) Use the box plot to describe the relationship between age and customer churn.

- The mean age is higher for churn=1 compared to churn=0.
- The median age is higher for churn=1 compared to churn=0.
- The maximum age is higher for churn=1 compared to churn=0.

3) Describe the relationship between the number of products a customer holds with a bank and customer churn.

- Customers with fewer bank products are more likely to churn.
- Customers with more bank products are more likely to churn.
- No relationship exists between the number of bank products and customer churn.

Modeling customer churn with logistic regression

Since a non-negligible proportion of customers close their accounts at this bank each year, the bank would like to identify features that can explain or predict customer churn. But, some customer churn is unavoidable. Ex: A customer moving to a new town may close their bank account if no nearby bank locations exist.

Customer churn is a binary categorical feature, which means that simple linear regression will not be suitable. Instead, logistic regression should be used to predict whether or not a customer will close their account. Another advantage of using logistic regression to predict customer churn is the ability to predict a probability, or chance, that a customer may close their account. Ex: If a particular feature is associated with a high chance of customer churn, then the bank may build an early warning system to identify customers based on that feature.

Logistic regression model for customer churn.

Full screen

The Python code below imports the customer churn dataset and fits a logistic regression model using age as the input feature.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.

- Try changing the input feature to credit score or account balance. How does the logistic regression model change?

Your server is starting up.

You will be redirected automatically when it's ready for you.

Event log

PARTICIPATION ACTIVITY 7.7.2: Logistic regression model.

The coefficients from a logistic regression model for predicting whether a customer will churn based on customer age are b_1 = 0.06294447 and b_0 = -3.92858767.

1) Calculate the probability that a 50-year-old customer will churn.
   - 0.314
   - 0.686
   - 0.781

2) Based on the logistic regression model, younger customers are _____ likely to churn.
   - more
   - less
   - equally

3) Based on the logistic regression plots, which feature is the most useful input for predicting customer churn?

- Account balance
- Age
- Credit score

Multiple logistic regression

Multiple linear regression is used to predict a numerical output feature using multiple input features. But, linear regression is not the only model that can take more than one input feature. Multiple logistic regression fits a logistic regression model with two or more input features.

The formula for a multiple logistic regression model is $\hat{p} = \displaystyle \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + ...+ b_k x_k} }{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k}}$, where:

- $x_1, ..., x_k$ are the input features.
- $b_1, ..., b_k$ are the slopes.
- b_0 is the y-intercept, or predicted value when $x_1, ..., x_k$ are all zero.

In the previous logistic regression models, age appeared to be the most useful predictor of whether a customer will churn. Can the logistic regression model be improved by adding more features?

Multiple logistic regression model for customer churn.

Full screen

The Python code below imports the customer churn dataset and fits a multiple logistic regression model using account balance, age, and credit score.

- Click the double right arrow icon to restart the kernel and run all cells.
- Examine the code below.

Your server is starting up.

You will be redirected automatically when it's ready for you.

Event log

PARTICIPATION ACTIVITY 7.7.3: Multiple logistic regression model.

Let x_1 be account balance, x_2 be age, and x_3 be credit score. The estimated coefficients from a multiple logistic regression model are $b_0=-0.00019447$, $b_1=0.0000036$, $b_2=0.0428263$, and $b_3=-0.0051893$.

1) Which feature is negatively associated with churn?

- Account balance
- Age
- Credit score

2) Calculate the estimated probability that a 21-year old customer with an account balance of $1,000 and a credit score of 650 will churn.

- 0.013
- 0.078
- 0.080

3) Two logistic regression models have been fitted: one with a single input feature and one with multiple input features. How could the logistic regression models be compared?

- Compare estimated slopes
- Compare estimated intercepts
- Compare predicted values

7.8 LAB: Creating simple linear regression models

The `nbaallelo_slr` dataset contains information on 126315 NBA games between 1947 and 2015. The columns report the points made by one team, the Elo rating of that team coming into the game, the Elo rating of the team after the game, and the points made by the opposing team. The Elo score measures the relative skill of teams in a league.

- Load the dataset into a data frame.
- Create a new column y in the data frame that is the difference between the points made by the two teams.
- Use `sklearn`'s `LinearRegression()` function to perform a simple linear regression on the y and elo_i columns.
- Compute the proportion of variation explained by the linear regression using the `LinearRegression` object's score method.

Ex: If the Elo rating of the team after the game, `elo_n`, is used instead of `elo_i`, the output is:

```
The intercept of the linear regression line is -59.135.
The slope of the linear regression line is 0.040.
The proportion of variation explained by the linear regression model is 0.111.
```

537150.4174434.qx3zqy7

LAB ACTIVITY 7.8.1: LAB: Creating simple linear regression models 1 / 1

main.py

```
1 Loading latest submission...
```

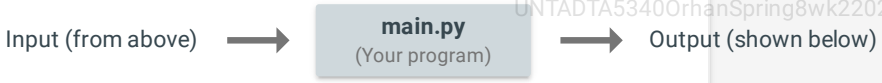
Develop mode | Submit mode

Run your program as often as you'd like, before submitting for grading. Below, type any needed input values in the first box, then click **Run program** and observe the program's output in the second box.

Enter program input (optional)

If your code requires input values, provide them here.

Run program



Program output displayed here

Coding trail of your work What is this?

Retrieving signature

7.9 LAB: Performing logistic regression using LogisticRegression()

The `nbaallelo_log` file contains data on 126314 NBA games from 1947 to 2015. The dataset includes the features `pts`, `elo_i`, `win_equiv`, and `game_result`. Using the csv file `nbaallelo_log.csv` and `scikit-learn`'s `LogisticRegression` function, construct a logistic regression model to classify whether a team will win or lose a game based on the team's elo_i score.

- Hot encode the `game_result` variable as a numeric variable with 0 for L and 1 for W
- Use the `LogisticRegression` function to construct a logistic regression model with `game_result` as the target and `elo_i` as the predictor.
- Predict the probability of a win from an elo_i score of 1310.
- Predict whether a team with an elo_i score of 1310 will win.

Note: Use `ravel()` from `numpy` to flatten the second argument of `LogisticRegression.fit()` into a 1-D array.

Ex: If a elo_i score of 1410 is used instead of 1310, the output is:


```
A team with the given elo_i score has predicted probability:
0.593 losing
0.407 winning
and the overall prediction is 0
```


LAB ACTIVITY | 7.9.1: LAB: Performing logistic regression using LogisticRegression() 1 / 1

main.py

```
1 Loading latest submission...
```


Develop mode | Submit mode

Run your program as often as you'd like, before submitting for grading. Below, type any needed input values in the first box, then click **Run program** and observe the program's output in the second box.

Enter program input (optional)

If your code requires input values, provide them here.

Run program

Input (from above) → **main.py** (Your program) → Output (shown below)

Program output displayed here

Coding trail of your work What is this?

Retrieving signature

©zyBooks 04/14/24 12:43 2087217
Biniam abebe
UNTADTA5340OrhanSpring8wk22024