A logo of a company

Description automatically generated

Data Advanced Analytics Google Certification

## Course 5 **Regression Analysis: Simplify Complex Data Relationships**

**Module 1:**

**Regression analysis or Regression models**

A group of statistical techniques that use existing data to estimate the relationships between a single dependent variable and one or more independent variables.

**What you’ll learn**

* Modeling variable relationships

**Helpful resources and tips**

As a learner, you can choose to complete one or multiple courses in this program. However, to obtain the Google Advanced Data Analytics Certificate, you must complete all of the courses. This reading describes what is required to obtain a certificate and best practices for you to have a good learning experience on Coursera.

**Obtain the Google Advanced Data Analytics Certificate**

To receive your official Google Advanced Data Analytics Certificate, you must:

* Pass all graded assignments in all 7 courses of the certificate program. Each graded assignment is part of a cumulative graded score for the course, and the passing grade for each course is 80%.

AND **one**of the following:

* Pay the [course certificate fee](https://www.coursera.support/s/article/209818963-Payments-on-Coursera?language=en_US),
* Be approved for [Coursera Financial Aid](https://www.coursera.support/s/article/209819033-Apply-for-Financial-Aid-or-a-Scholarship?language=en_US), **or**
* Complete the certificate through an educational institution, employer, or agency that's sponsoring your participation.

**Healthy habits for course completion**

Here is a list of best practices that will help you complete the courses in the program in a timely manner:

* **Plan your time:** Setting regular study times and following them each week can help you make learning a part of your routine. Use a calendar or timetable to create a schedule, and list what you plan to do each day in order to set achievable goals. Find a space that allows you to focus when you watch the videos, review the readings, and complete the activities.
* **Work at your own pace:** Everyone learns differently, so this program has been designed to let you work at your own pace. Although your personalized deadlines start when you enroll, feel free to progress through the program at the speed that works best for you. There is no penalty for late assignments; to earn your certificate, all you have to do is complete all of the work. You can extend your deadlines at any time by going to **Overview** in the navigation panel and selecting **Switch Sessions**. If you have already missed previous deadlines, select **Reset my deadlines** instead.
* **Be curious:** If you find an idea that gets you excited, act on it! Ask questions, search for more details online, explore the links that interest you, and take notes on your discoveries. The steps you take to support your learning along the way will advance your knowledge, create more opportunities in this high-growth field, and help you qualify for jobs.
* **Take notes:** Notes will help you remember important information in the future, especially as you’re preparing to enter a new job field. In addition, taking notes is an effective way to make connections between topics and gain a better understanding of those topics.
* **Review exemplars:** Exemplars are completed assignments that fully meet an activity's criteria. Many activities in this program have exemplars for you to compare to your own work. Although there are often many ways to complete an assignment, exemplars offer you guidance and inspiration about how to complete the activity.
* **Chat (responsibly) with other learners:** If you have a question, chances are, you’re not alone. Use the [discussion forums](https://www.coursera.org/learn/regression-analysis-simplify-complex-data-relationships/discussions) to ask for help from other learners taking this program. You can also visit Coursera’s [Global Online Community](https://coursera.community/). Other important things to know while learning with others can be found in the [Coursera Honor Code](https://learner.coursera.help/hc/en-us/articles/209818863-Coursera-Honor-Code) and [Code of Conduct](https://learner.coursera.help/hc/en-us/articles/208280036-Coursera-Code-of-Conduct).
* **Update your profile:** Consider [updating your profile](https://www.coursera.org/account/profile) on Coursera to include your photo, career goals, and more. When other learners find you in the discussion forums, they can click on your name to access your profile and get to know you better.

Documents, spreadsheets, presentations, and labs for course activities

To complete certain activities in the program, you will need to use digital documents, spreadsheets, presentations, and/or labs. Data analytics professionals use these software applications to collaborate within their teams and organizations. If you need more information about using a particular tool, refer to these resources:

* [Microsoft Word: Help and learning](https://support.microsoft.com/en-us/word): Microsoft Support page for Word
* [Google Docs](https://support.google.com/docs/topic/9046002?hl=en&ref_topic=1382883): Help Center page for Google Docs
* [Microsoft Excel: Help and learning](https://support.microsoft.com/en-us/excel): Microsoft Support page for Excel
* [Google Sheets](https://support.google.com/docs/topic/9054603?hl=en&ref_topic=1382883): Help Center page for Google Sheets
* [Microsoft PowerPoint: Help and learning](https://support.microsoft.com/en-us/powerpoint): Microsoft Support page for PowerPoint
* [How to use Google Slides](https://support.google.com/docs/answer/2763168?hl=en&co=GENIE.Platform%3DDesktop): Help Center page for Google Slides
* [Common problems with labs](https://support.google.com/qwiklabs/answer/9133560?hl=en&ref_topic=9134804): Troubleshooting help for Qwiklabs activities

**Module, course, and certificate glossaries**

This program covers a lot of terms and concepts, some of which you may already know and some of which may be unfamiliar to you. To review terms and help you prepare for graded quizzes, refer to the following glossaries:

* **Module glossaries**: At the end of each module’s content, you can review a glossary of terms from that module. Each module’s glossary builds upon the terms from the previous modules in that course. The module glossaries are not downloadable; however, all of the terms and definitions are included in the course and certificate glossaries, which are downloadable.
* **Course glossaries**: At the end of each course, you can access and download a glossary that covers all of the terms in that course.
* **Certificate glossary**: The certificate glossary includes all of the terms in the entire certificate program and is a helpful resource that you can reference throughout the program or at any time in the future.

You can access and download the certificate glossaries and save them on your computer. You can always find the course and certificate glossaries using the course’s [Resources](https://www.coursera.org/learn/regression-analysis-simplify-complex-data-relationships/resources/h7UO6) tab. To access the **Advanced Data Analytics Certificate glossary**, click the following link and select *Use Template*.

* Link to the glossary: [Advanced Data Analytics Certificate glossary](https://docs.google.com/document/d/193-AtS7MlB2w4buwiCyPjBoOhIbbByKgHWPpYnSR9VI/template/preview)

OR

* If you don’t have a Google account, you can download the glossary directly from the following attachment.

[Advanced Data Analytics Certificate glossary](https://d3c33hcgiwev3.cloudfront.net/AiPoIzIKSpq2RjyiTqOmUQ_24121a299b8b49b295ebf41e4ba3f9f1_Advanced-Data-Analytics-Certificate-glossary.docx?Expires=1722384000&Signature=ExbpfzEZ6Py2CmS5LnEFmP85cCWemllXuo8mPqXFzFai9gaOYjGjoQ8ScvDewK0txr7mu-Ncacf4YpRkd1xA-McKIOlbdSLkYmDgbv3GV7lHOwTSCXhMIw~DFtehtqAm768~PoRMrU9dn-jYgJNv3iHGM3ZpGpyrSrZtgFLfuNQ_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

[DOCX File](https://d3c33hcgiwev3.cloudfront.net/AiPoIzIKSpq2RjyiTqOmUQ_24121a299b8b49b295ebf41e4ba3f9f1_Advanced-Data-Analytics-Certificate-glossary.docx?Expires=1722384000&Signature=ExbpfzEZ6Py2CmS5LnEFmP85cCWemllXuo8mPqXFzFai9gaOYjGjoQ8ScvDewK0txr7mu-Ncacf4YpRkd1xA-McKIOlbdSLkYmDgbv3GV7lHOwTSCXhMIw~DFtehtqAm768~PoRMrU9dn-jYgJNv3iHGM3ZpGpyrSrZtgFLfuNQ_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A" \t "_blank)

**Data Analytics Certificate glossary**

If you completed the original [Google Data Analytics Certificate](https://www.coursera.org/professional-certificates/google-data-analytics?utm_source=google&utm_medium=institutions&utm_campaign=gwgsite-gDigital-paidha-sem-bk-gen-exa-glp-br-null&_ga=2.170664992.1625030801.1661901112-1742325342.1661901112), you may recognize some overlap with several of the glossary terms in this program. Refer to the Data Analytics Certificate glossary, linked in the [Resources](https://www.coursera.org/learn/regression-analysis-simplify-complex-data-relationships/resources/4nTfn) tab, to review these foundational terms and concepts. The definitions of some terms in the Data Analytics Certificate glossary differ from the definitions of the same terms in this program since the Advanced Data Analytics Certificate builds upon the concepts taught in the previous program.

**Course feedback**

Providing feedback on videos, readings, and other materials is easy. With the resource open in your browser, you can find the thumbs-up and thumbs-down symbols.

* Click **thumbs-up** for materials you find helpful.
* Click **thumbs-down** for materials that you do not find helpful.

If you want to flag a specific issue with an item, click the flag icon, select a category, and enter an explanation in the text box. This feedback goes back to the course development team and isn’t visible to other learners. All feedback received helps to create even better certificate programs in the future.

For technical help, visit the [Learner Help Center](https://learner.coursera.help/hc/en-us).

“Regression models are super important, powerful tools. You’ll be able to answer a wide variety of questions using different types of regression models.”

“I love being able to take large amounts of data and being able to solve massive problems.”

“There is a massive opportunity for innovation right now in the analytics field in marketing.”

**Course 5 overview**

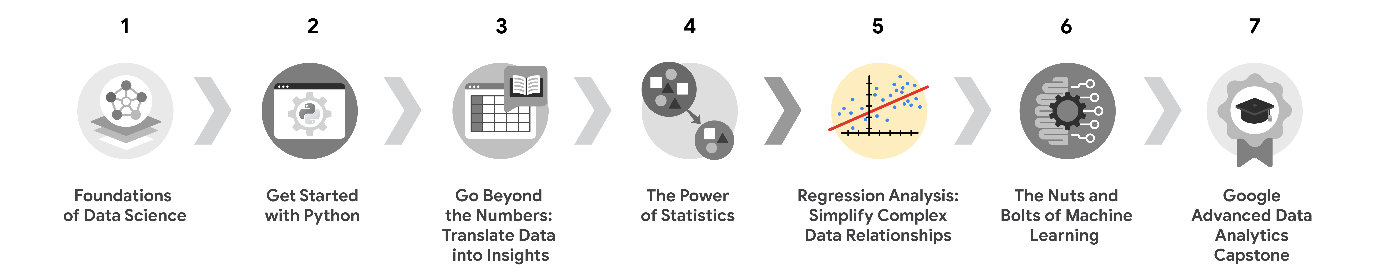


Hello, and welcome to **Regression Analysis: Simplify Complex Data Relationships**, the fifth course in the Google Advanced Data Analytics Certificate. You’re making great progress on this exciting journey!

In previous courses, you learned how data professionals contribute to the success of an organization, the basic syntax and functions of the Python programming language, the main stages of exploratory data analysis (EDA), and how to use statistics to analyze and interpret data. In this course, you’ll explore how to simplify complex data relationships using linear and logistic regression models.

**Course descriptions**

The Google Advanced Data Analytics Certificate has seven courses. **Regression Analysis: Simplify Complex Data Relationships** is the fifth course.

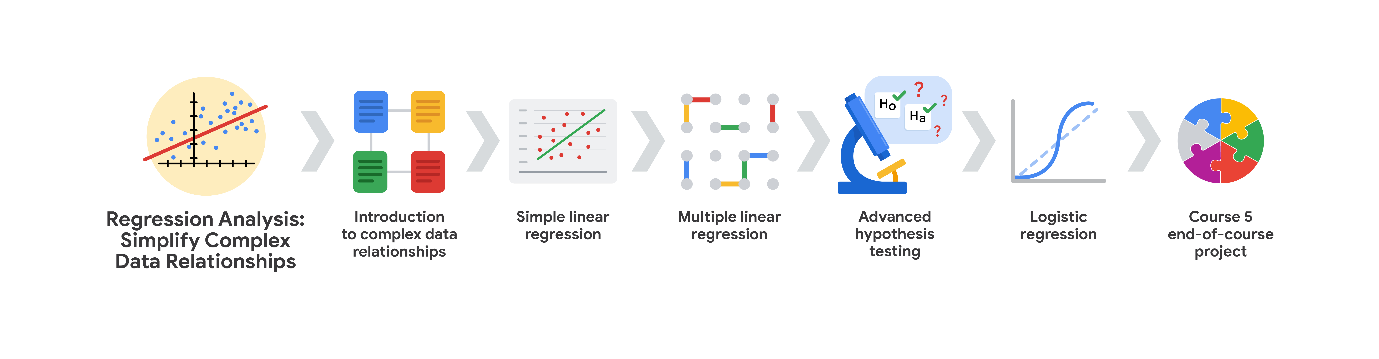


1. [**Foundations of Data Science**](https://www.coursera.org/learn/foundations-of-data-science/home/week/1) — Learnhow data professionals operate in the workplace and how different roles in the field of data science contribute to an organization’s vision of the future. Then, explore data science roles, communication skills, and data ethics.
2. [**Get Started with Python**](https://www.coursera.org/learn/get-started-with-python/home/week/1) —Discover how the programming language Python can power your data analysis. Learn core Python concepts, such as data types, functions, conditional statements, loops, and data structures.
3. [**Go Beyond the Numbers: Translate Data into Insights**](https://www.coursera.org/learn/go-beyond-the-numbers-translate-data-into-insight/home/week/1) — Learn the fundamentals of data cleaning and visualizations and how to reveal the important stories that live within data.
4. [**The Power of Statistics**](https://www.coursera.org/learn/the-power-of-statistics/home/week/1) — Explore descriptive and inferential statistics, basic probability and probability distributions, sampling, confidence intervals, and hypothesis testing.
5. [**Regression Analysis: Simplify Complex Data Relationships**](https://www.coursera.org/learn/regression-analysis-simplify-complex-data-relationships/home/week/1) — *(current course)* Learn to model variable relationships, focusing on linear and logistic regression.
6. [**The Nuts and Bolts of Machine Learning**](https://www.coursera.org/learn/the-nuts-and-bolts-of-machine-learning/home/week/1) — Learn unsupervised machine learning techniques and how to apply them to organizational data.
7. [**Google Advanced Data Analytics Capstone**](https://www.coursera.org/learn/google-advanced-data-analytics-capstone/home/week/1) — Complete a hands-on project designed to demonstrate the skills and competencies you acquire in the program.

**Course 5 content**

Each course of this certificate program is broken into modules. You can complete courses at your own pace, but the module breakdowns are designed to help you finish the entire Google Advanced Data Analytics Certificate in about six months.

What’s to come? Here’s a quick overview of the skills you’ll learn in each module of this course.



**Module 1: Introduction to complex data relationships**

In this part of the course, you’ll examine how PACE can serve as a solid foundation for regression analysis. Then, you’ll learn about linear and logistic regression, how they’re similar and different, and when to apply each approach. Finally, you’ll use real data to learn how to apply estimation techniques in Python and study more use-cases of regression.

**Module 2: Simple linear regression**

You will explore examples of ordinary least squares estimation in Python and learn the four main model assumptions of simple linear regression: linearity, normality of residuals, independent observations, and homoscedasticity. You’ll build a model in Python and learn how to evaluate model fit using R squared, holdout samples, and measures of uncertainty, like confidence intervals and p-values. Then, you'll begin to develop a story using numbers and statistics and learn how to communicate the results of your simple linear regression model using Python visualizations.

**Module 3: Multiple linear regression**

In this section, you will build on your knowledge of simple linear regression by extending the concepts into multiple regression. You’ll learn how to test assumptions using math, built-in Python functions, and visualizations during the EDA process. Next, you’ll interpret and evaluate multiple regression results, examine the idea of overfitting, and discuss ways to overcome the effects of overfitting using regularization techniques.

**Module 4: Advanced hypothesis testing**

In this part of the course, you’ll examine a variety of hypothesis tests: Chi-squared, ANOVA, ANCOVA, MANOVA, and MANCOVA. You will also investigate null and alternative hypotheses and how they help articulate testing results.

**Module 5: Logistic regression**

You will explore logistic regression models, with a heavy emphasis on logit logistic regression. You’ll build binomial, multinomial, and ordinal logistic regression models, and log-linear Poisson regression models. Next, you’ll discover important classification model metrics such as ROC and AUC, precision, recall, and type I and type II errors.

**Module 6: Course 5 end-of-course project**

As you conclude this part of the Advanced Data Analytics Certificate, you will put everything you’ve learned into one end-of-course project. You’ll be tasked with solving a business problem using the provided data. The concepts and skills you will learn in this part of the course will be critical to your success as a data professional.

**What to expect**

Each course offers many types of learning opportunities:

* **Videos** led by Google instructors teach new concepts, introduce the use of relevant tools, offer career support, and provide inspirational personal stories.
* **Readings** build on the topics discussed in the videos, introduce related concepts, share useful resources, and describe case studies.
* **Discussion prompts** explore course topics for better understanding and allow you to chat and exchange ideas with other learners in the [**discussion forums**](https://www.coursera.org/learn/regression-analysis-simplify-complex-data-relationships/discussions).
* **Self-review activities** and **labs** give you hands-on practice in applying the skills you are learning and allow you to assess your own work by comparing it to a completed example.
* **Interactive plug-ins** encourage you to practice specific tasks and help you integrate knowledge you have gained in the course.
* **In-video quizzes** help you check your comprehension as you progress through each video.
* **Practice quizzes** allow you to check your understanding of key concepts and provide valuable feedback.
* **Graded quizzes** demonstrate your understanding of the main concepts of a course. You must score 80% or higher on each graded quiz to obtain a certificate, and you can take a graded quiz multiple times to achieve a passing score.

**Tips for success**

* It is strongly recommended that you go through the items in each lesson in the order they appear because new information and concepts build on previous knowledge.
* Participate in all learning opportunities to gain as much knowledge and experience as possible.
* If something is confusing, don’t hesitate to replay a video, review a reading, or repeat a self-review activity.
* Use the additional resources that are referenced in this course. They are designed to support your learning. You can find all of these resources in the [**Resources**](https://www.coursera.org/learn/regression-analysis-simplify-complex-data-relationships/resources/yHPhW) tab.
* When you encounter useful links in this course, bookmark them so you can refer to the information later for study or review.
* Understand and follow the [Coursera Code of Conduct](https://www.coursera.support/s/article/208280036-Coursera-Code-of-Conduct?) to ensure that the learning community remains a welcoming, friendly, and supportive place for all members.

**What you’ll learn**

* Regression analysis
* Simple linear regression
* Multiple linear regression
* Hypothesis testing
* Logistic regression

PACE (Regression Modelling):

* **Plan: Understand your data in the problem context**, Contextualize & understand the data and the problem
* **Analyze: EDA, check model assumptions & select model**

Determine if we should move forward with building the model!

* **Construct: Construct & evaluate model**, Determine how well your model fits the data!
* **Evaluate: Interpret model and share story**, Descriptions must take into account the context of the data!

**Model assumptions**

Statements about the data that must be true to justify the use of particular data science techniques.

**Question**

What are model assumptions?

The processes associated with converting model statistics into statements describing the relationships between the variables in the data

Ways to measure how well a model fits the data

Statements about the data that must be true to justify the use of particular data science techniques

The processes associated with building a model

Correct

Model assumptions are statements about the data that must be true to justify the use of particular data science techniques. Data professionals use model assumptions to add validity to their conclusions. If model assumptions are true, then they can have more confidence in the results of a model.

**Test your knowledge: PACE in regression analysis**

**1.**

**Question 1**

**In regression modeling, which statement describes the PACE plan stage?**

**1 / 1 point**

**Understanding the data in the context of a problem**

**Preparing formal results and visualizations for stakeholders**

**Building the regression model in a coding language**

**Examining data more closely to choose an appropriate model**

**Correct**

**In regression modeling, understanding the data in the context of a problem describes the PACE plan stage. During the plan stage, a data professional considers what data they have access to, how the data was collected, and what the business needs are.**

**2.**

**Question 2**

**In which PACE stage does a data professional initially check the model assumptions?**

**1 / 1 point**

**Construct**

**Plan**

**Analyze**

**Execute**

**Correct**

**During the analyze stage, a data professional initially checks the model assumptions.**

**3.**

**Question 3**

**What three tasks typically occur during the PACE construct stage? Select all that apply.**

**1 / 1 point**

**Evaluate the model results**

**Correct**

**A data professional builds the model, rechecks and confirms model assumptions, and evaluates results in the construct stage.**

**Re-check and confirm the model assumptions**

**Correct**

**A data professional builds the model, rechecks and confirms model assumptions, and evaluates results in the construct stage.**

**Present the visualizations to stakeholders**

**Build the model**

**Correct**

**A data professional builds the model, rechecks and confirms model assumptions, and evaluates results in the construct stage.**

**Line**

A collection of an infinite number of points extending in two opposite directions

**Linear regression**

A technique that estimates the linear relationship between a **continuous** dependent variable and one or more independent variables

**Dependent variable (Y)**

The variable a given model estimates, also referred to as a response or outcome variable.

**Independent variable (X)**

The variable that explains trends in the dependent variable, also referred to as an explanatory or predictor variable.

**Slope**

The amount that y increases or decreases per one-unit increase of x.

**Intercept**

The value of y, the dependent variable, when x, the independent variable, equals to 0

**Positive correlation**

A relationship between two variables that trend to increase or decrease together.

**Negative correlation**

A relationship between two variables where when one variable increases, the other variable tends to decrease, and vice versa.

**Causation**

A cause-and-effect relationship where one variable directly causes the other to change in a particular way

**Linear regression overview**

* Linear regression is a way to model linear relationships
* Dependent variables vary according to independent variables
* The slope identifies how much the dependent variable changes per one-unit change in the independent variable
* Correlation describes linear relationships between variables
* Correlation is not causation

**Observed values (actual values)**

The existing sample of data

Each data point in the sample is represented by an observed value of the independent variable and an observed value of the independent variable.

**Regression coefficients**

The estimated betas in a regression model. Represented as beta\_{i} hat.

**Ordinary Least Squares Estimation (OLS)**

Common way to calculate linear regression coefficients (beta hat)\_{n}

**Loss function**

A function that measures the distance between the observed values and the model’s estimated values.

**Question**

It is often not possible to calculate the true values of parameters.

True

Parameters are properties of populations and not samples, so it is often impossible to calculate their true value because it is usually the case that the whole population cannot be observed. In these cases, it is possible to calculate *estimates* of parameters using sample data.

**Test your knowledge: Linear regression**

**1.**

**Question 1**

**What technique estimates the linear relationship between a continuous dependent variable and one or more independent variables?**

**1 / 1 point**

**Causation**

**Model validation**

**Linear regression**

**Intercept**

**Correct**

**Linear regression estimates the linear relationship between a continuous dependent variable and one or more independent variables.**

**2.**

**Question 2**

**Which of the following statements accurately describe dependent and independent variables? Select all that apply.**

**1 / 1 point**

**The dependent variable is the variable the given model estimates.**

**Correct**

**The dependent variable is the variable the given model estimates. It tends to vary based on the values of independent variables. Independent variables are also referred to as explanatory or predictor variables.**

**The dependent variable tends to vary based on the values of independent variables.**

**Correct**

**The dependent variable is the variable the given model estimates. It tends to vary based on the values of independent variables. Independent variables are also referred to as explanatory or predictor variables.**

**Independent variables are also referred to as explanatory or predictor variables.**

**Correct**

**The dependent variable is the variable the given model estimates. It tends to vary based on the values of independent variables. Independent variables are also referred to as explanatory or predictor variables.**

**The independent variable tends to vary based on the values of dependent variables.**

**3.**

**Question 3**

**What term describes an inverse relationship between two variables?**

**1 / 1 point**

**Intercept**

**Negative correlation**

**Slope**

**Positive correlation**

**Correct**

**With negative correlation, when one variable increases, the other variable tends to decrease. The reverse is also true.**

**4.**

**Question 4**

**Fill in the blank: The goal of regression analysis is to use math to define the \_\_\_\_\_ between the sample X’s and Y’s in order to understand how the variables interact.**

**1 / 1 point**

**relationship**

**independence**

**value**

**model**

**Correct**

**The goal of regression analysis is to define a relationship mathematically between the sample X’s and Y’s in order to understand how the variables interact.**

**Logistic regression**

A technique that models a categorical dependent variable based on one or more independent variables.

**Question**

**What technique models a categorical variable based on one or more independent variables?**

Logistic regression

**Correct**

Logistic regression models a categorical variable based on one or more independent variables. The dependent variable in a logistic regression can have two or more possible discrete values.

**Link function**

A nonlinear function that connects or links the dependent variable to the independent variables mathematically

**Categorize: Linear and logistic regression**

**Assumes a linear relationship between the dependent and independent variable(s)**

Linear regression

**Foundational for data science**

Both

**The outcome variable (Y) is continuous**

Linear regression

**Can use multiple predictor/independent variables (X)**

Both

**A link function is used to model the outcome variable (Y)**

Logistic regression

**The dependent and independent variable(s) must be correlated**

Linear regression

**The outcome variable (Y) is categorical.**

Logistic regression

**Allows us to estimate relationships between observed variables**

Both

**Everyday examples of logistic regression**

So far, you have learned about two regression models: linear and logistic regression. **Regression models** are a group of statistical techniques that use existing data to estimate the relationships between a single dependent variable and one or more independent variables. Before exploring each model in depth, let’s examine how regression can help you answer questions you encounter every day. Doing so will help you become more comfortable applying your knowledge as a data professional.

For this discussion prompt, consider the following:

* What are some relationships you observe at a store that can be modeled using linear regression? For example, sales of toothpaste and toothbrushes may be linearly related.
* What about logistic regression? What outcomes have you observed that could be modeled using this model? For example, the length of a volunteer application could be related to the odds that someone volunteers for an organization. The number of pages the volunteer application has might be related to the chances of that person actually volunteering.

**Linear Regression**

Linear regression can be used to model relationships where the dependent variable changes continuously with the independent variable. For example, in a retail store, the relationship between advertising spend and weekly sales revenue can be modeled using linear regression. As the amount spent on advertising increases, the sales revenue might also increase proportionally. Another example is the relationship between the number of employees working at a store and the amount of revenue generated. If more employees are present, it can lead to better customer service, resulting in higher sales. Additionally, the price of a product and the quantity sold can often be modeled linearly, where a decrease in price might lead to an increase in sales volume.

**Logistic Regression**

Logistic regression is suitable for modeling binary outcomes. For instance, in a store, the probability of a customer making a purchase (yes/no) can be modeled based on various factors such as the time spent in the store, the number of items they viewed, or whether they interacted with a sales associate. Another example is modeling whether a customer will return to the store (returning customer vs. non-returning customer) based on their previous shopping behavior and satisfaction ratings. Lastly, logistic regression can be used to predict whether a customer will use a loyalty card during their purchase based on factors like purchase amount, time of the day, and the type of products bought.

**Test your knowledge: Logistic regression**

**1.**

Question 1

What is a nonlinear function that connects or links a dependent variable to the independent variables mathematically?

1 / 1 point

Relationship function

Link function

Regression function

Loss function

Correct

The link function connects, or links, a dependent variable to the independent variables mathematically. Data professionals use the link function to express the relationship between the X’s and the probability that Y equals some outcome.

**2.**

Question 2

What type of regression models a categorical variable based on one or more independent variables?

1 / 1 point

Ordinary regression

Logistic regression

Coefficient regression

Linear regression

Correct

Logistic regression models a categorical variable based on one or more independent variables. The dependent variable can have two or more possible discrete values.

**Glossary terms from module 1**

**Terms and definitions from Course 5, Module 1**

**Absolute values**: (Refer to **observed values**)

**Causation**: A cause-and-effect relationship where one variable directly causes the other to change in a particular way

**Dependent variable (Y)**: The variable a given model estimates

**Explanatory variable**: (Refer to **independent variable**)

**Independent variable (X)**: A variable whose trends are associated with the dependent variable

**Intercept (constant 𝐵**0**)**: The y value of the point on the regression line where it intersects with the y-axis

**Line**: A collection of an infinite number of points extending in two opposite directions

**Linear regression**: A technique that estimates the linear relationship between a continuous dependent variable and one or more independent variables

**Link function**: A nonlinear function that connects or links the dependent variable to the independent variables mathematically

**Logistic regression**: A technique that models a categorical dependent variable based on one or more independent variables

**Loss function**: A function that measures the distance between the observed values and the model’s estimated values

**Model assumptions**: Statements about the data that must be true to justify the use of a particular modeling technique

**Negative correlation**: An inverse relationship between two variables, where when one variable increases, the other variable tends to decrease, and vice versa

**Observed values:** The existing sample of data, where each data point in the sample is represented by an observed value of the dependent variable and an observed value of the independent variable

**Outcome variable**: (Refer to **dependent variable**)

**Positive correlation**: A relationship between two variables that tend to increase or decrease together

**Predictor variable**: (Refer to **independent variable**)

**Regression analysis**: A group of statistical techniques that use existing data to estimate the relationships between a single dependent variable and one or more independent variables

**Regression coefficient**: The estimated betas in a regression model

**Regression models**: (Refer to **regression analysis**)

**Response variable**: (Refer to **dependent variable**)

**Slope**: The amount that y increases or decreases per one-unit increase of x

**Module 1 challenge**

**1.**

Question 1

Fill in the blank: Regression models are groups of \_\_\_\_\_ techniques that use data to estimate the relationships between a single dependent variable and one or more independent variables.

statistical

**2.**

Question 2

\_\_\_\_\_ finds the mean of Y given a particular value of X.

Simple linear regression

**3.**

Question 3 (NOT FINAL)

A data professional creates a model in Python and rechecks the model assumptions. What PACE stage are they working in?

Analyze

Review [the video about PACE in regression analysis.](https://www.coursera.org/learn/ada-c5/lecture/SCmP4/pace-in-regression-analysis-v-458)

**4.**

Question 4

What technique estimates the relationship between a continuous dependent variable and one or more independent variables?

Linear regression

**5.**

Question 5

Which of the following statements accurately describe dependent and independent variables? Select all that apply.

An independent variable is often represented by X.

Correct

A dependent variable is the variable a given model estimates.

Correct

**6.**

Question 6

What is an inverse relationship between two variables, where one variable increases, the other variable tends to decrease?

Negative correlation

**7.**

Question 7

A data professional creates a linear regression equation and reviews the properties of populations, sometimes referred to as Mu of y and the betas. What term describes this portion of the equation?

Parameters

**8.**

Question 8

A veterinary practice wants to determine whether most new patients will choose to return for follow-up care. A data analyst for the practice investigates this issue by modeling a categorical variable based on one or more independent variables. What technique do they use?

Logistic regression

**9.**

Question 9

A data professional wants to connect the dependent variable and independent variable mathematically. What function can enable them to make this connection?

Link function

**Module 2:**

**Simple linear regression**

A technique that estimates the linear relationship between one independent variable X, and one continuous dependent variable, Y

**Linear regression**

* Python programming
* Exploratory data analysis (EDA)
* Statistics

**What you’ll learn**

* Linear regression equation
* Ordinary least squares estimation technique
* Four linear regression assumptions
* Build a linear regression model in Python
* Evaluate linear regression model using metrics
* Interpreting results

“Mentorships have been a huge part of my progression in my career.”

“Look for groups that are focused on the areas that you’re interested in. If we’ve thinking about BI and analytics, there’s so many groups of folks who are business intelligence professionals, and there’s niche groups that are focused on data science or predictive analytics.”

“I’ve actually found that LinkedIn is a really powerful tool. It takes work, but it’s a very simple entry point for networking.”

**Best fit line**

The line that fits the data best by minimizing some loss function or error

**Predicted values**

The estimated Y values for each X calculated by a model

**Residual**

The difference between observed or actual values and the predicted values of the regression line

The sum of the residuals is always equal to 0 for OLS estimators

**Sum of Squared Residuals (SSR)**

The sum of the squared differences between each observed value and its associated predicted value

**Ordinary Least Squares (OLS)**

A method that minimizes the sum of squared residuals to estimate parameters in a linear regression model

**Question**

What term describes the difference between observed or actual values and the predicted values of the regression line?

Residuals

Residuals describe the difference between observed or actual values and the predicted values of the regression line. Residual equals observed value minus predicted value.

**Explore ordinary least squares**

As previously mentioned, one way for finding the best fit line in regression modeling is to try different models until you find the best one. But for simple linear regression, the formulas for the best beta coefficients have been derived. In this reading, you will go through an example to gain a better understanding of how the sum of squared residuals can change as 𝛽^0*β*^​0​ and 𝛽^1*β*^​1​ change. There will be resources for further exploration if you’re interested in deriving the formulas for estimating the coefficients using ordinary least squares. In this reading, we will cover:

* Formula and notation review
* Minimizing the sum of squared residuals (SSR)
* Estimating beta coefficients

**Formula and notation review**

Earlier, you learned about simple linear regression as a method for estimating the linear relationship between a continuous dependent variable and one independent variable. An estimate based on simple linear regression can be represented mathematically as 𝑦^=𝛽^0+𝛽^1𝑋*y*^​=*β*^​0​+*β*^​1​*X* .

Remember that the hat symbol indicates that the beta coefficients are just estimates. As a result, the y-values derived from the regression model are also just estimates.

A common technique for calculating the coefficients of a linear regression model is called ordinary least squares, or OLS. Ordinary least squares estimates the beta coefficients in a linear regression model by minimizing a measure of error called the sum of squared residuals.

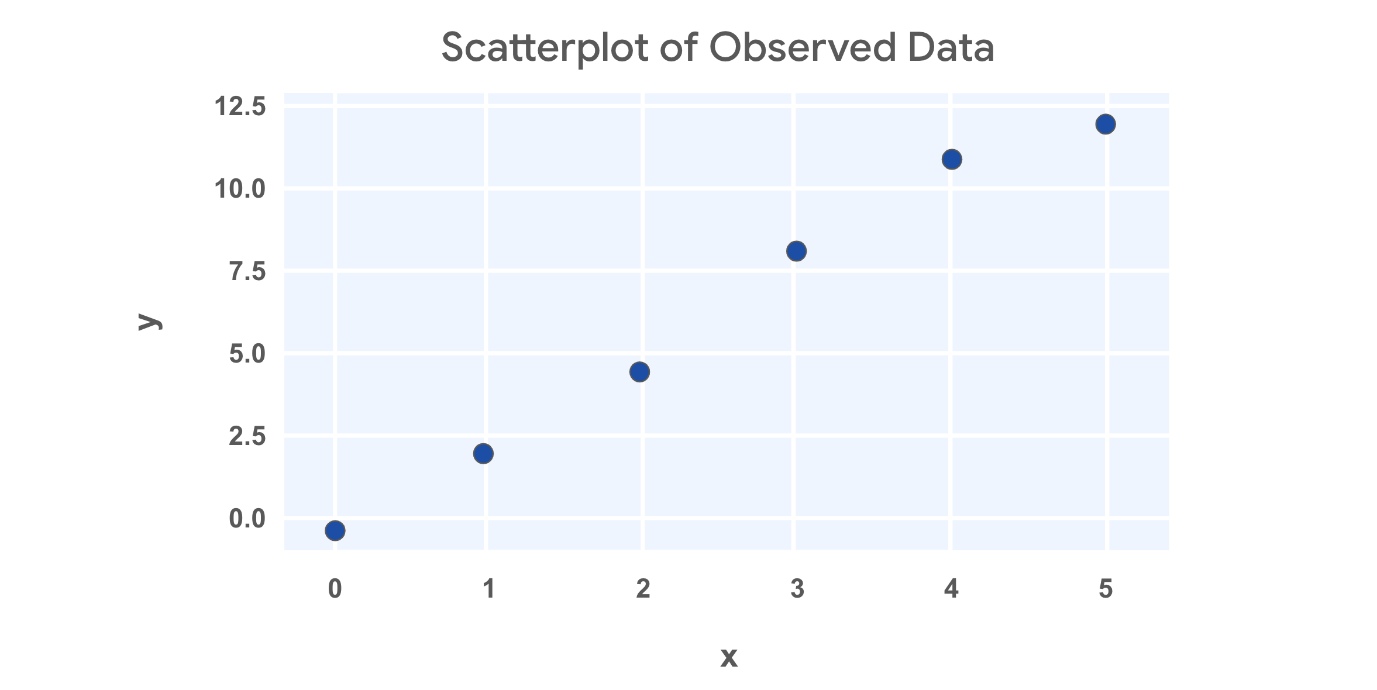
You can calculate the sum of squared residuals via this formula: ∑𝑖=1𝑁(𝑂𝑏𝑠𝑒𝑟𝑣𝑒𝑑−𝑃𝑟𝑒𝑑𝑖𝑐𝑡𝑒𝑑)2∑*i*=1*N*​(*Observed*−*Predicted*)2, which can be rewritten using mathematical notation as: ∑𝑖=1𝑁(𝑦𝑖−𝑦^𝑖)2∑*i*=1*N*​(*yi*​−*y*^​*i*​)2.

The large E shaped symbol is the capital Greek letter, sigma, and it denotes a sum. So the sum of squared residuals is the sum of the squared differences between the observed values and the values predicted by the regression model.

**Minimizing the sum of squared residuals (SSR)**

For the purposes of this reading, assume that you have a dataset of 6 observations: (0, -1), (1, 2), (2, 4), (3, 8), (4, 11), and (5, 12). These can be plotted on a 2-dimensional X-Y coordinate plane.

| **X (observed)** | **Y (observed)** |
| --- | --- |
| 0 | -1 |
| 1 | 2 |
| 2 | 4 |
| 3 | 8 |
| 4 | 11 |
| 5 | 12 |



**Line 1: 𝑦^=−0.5+3𝑥*y*^​=−0.5+3*x***

Next, let’s assume some values for 𝛽^0*β*^​0​ and 𝛽^1*β*^​1​and calculate the sum of squared residuals. For the first attempt, let’s assume 𝛽^0=−0.5*β*^​0​=−0.5 and 𝛽^1=3*β*^​1​=3. Then the linear equation would be 𝑦^=−0.5+3𝑥*y*^​=−0.5+3*x*. Since you now have the equation for 𝑦*y*, you can calculate the predicted values by plugging in each value of 𝑥*x*.

For example, if 𝑥=0*x*=0, then 𝑦^=−0.5+3∗0=−0.5*y*^​=−0.5+3∗0=−0.5. If 𝑥=1*x*=1, then 𝑦^=−0.5+3∗1=2.5*y*^​=−0.5+3∗1=2.5. So after calculating all the predicted values, then you can calculate the residual for each data point.

| **X (observed)** | **Y (actual)** | **Y (predicted) = -0.5 + 3x** | **Residual** |
| --- | --- | --- | --- |
| 0 | -1 | -0.5 | -1 - (-0.5) = -1+0.5 = -0.5 |
| 1 | 2 | 2.5 | 2 - 2.5 = -0.5 |
| 2 | 4 | 5.5 | 4 - 5.5 = -1.5 |
| 3 | 8 | 8.5 | 8 - 8.5 = -0.5 |
| 4 | 11 | 11.5 | 11 - 11.5 = -0.5 |
| 5 | 12 | 14.5 | 12 - 14.5 = -2.5 |

Then you can square each of the residuals by multiplying them by themselves, and then adding them all together to calculate the sum of squared residuals.

| **Residual** | **Squared Residual** |
| --- | --- |
| -1 - (-0.5) = -1+0.5 = -0.5 | 0.25 |
| 2 - 2.5 = -0.5 | 0.25 |
| 4 - 5.5 = -1.5 | 2.25 |
| 8 - 8.5 = -0.5 | 0.25 |
| 11 - 11.5 = -0.5 | 0.25 |
| 12 - 14.5 = -2.5 | 6.25 |

Sum of squared residuals =0.25+0.25+2.25+0.25+0.25+6.25=9.5

**Line 2: 𝑦^=−0.5+2.5𝑥*y*^​=−0.5+2.5*x***

Next, let’s adjust just the slope from the prior example. So 𝛽^0=−0.5*β*^​0​=−0.5 but 𝛽^1=2.5*β*^​1​=2.5. Then the linear equation would be 𝑦^=−0.5+2.5𝑥*y*^​=−0.5+2.5*x*. You can plug in values for 𝑥*x* just like last time to calculate the predicted values and get the squared residuals.

| **X (observed)** | **Y (actual)** | **Y (predicted) = -0.5 + 2.5x** | **Residual** | **Squared Residuals** |
| --- | --- | --- | --- | --- |
| 0 | -1 | -0.5 | -0.5 | 0.25 |
| 1 | 2 | 2 | 0 | 0 |
| 2 | 4 | 4.5 | -0.5 | 0.25 |
| 3 | 8 | 7 | 1 | 1 |
| 4 | 11 | 9.5 | 1.5 | 2.25 |
| 5 | 12 | 12 | 0 | 0 |

Sum of squared residuals =0.25+0+0.25+1+2.25+0=3.75=0.25+0+0.25+1+2.25+0=3.75.

Great! This estimate is way better!

**Estimating beta coefficients**

You could keep adjusting the slope and intercept, and then calculating the predicted values, residuals, and squared residuals. But there’s really no way to be sure you’ve found the best fit line. Through advanced math, some formulas have been derived to find the beta coefficients that minimize error.

There are multiple ways to write out the formulas for finding the beta coefficients. For simple linear regression, one way to write the formulas is as follows:

* 𝛽^1=∑𝑖=1𝑛(𝑋𝑖−𝑋‾)(𝑌𝑖−𝑌‾)∑𝑖=1𝑛(𝑋𝑖−𝑋‾)2*β*^​1​=∑*i*=1*n*​(*Xi*​−*X*)2∑*i*=1*n*​(*Xi*​−*X*)(*Yi*​−*Y*)​
* 𝛽^0=𝑌‾−𝛽^1𝑋‾*β*^​0​=*Y*−*β*^​1​*X*

You won’t be asked to calculate beta coefficients without help from a computer, but it can be interesting to explore if you desire. We’ve provided additional resources in case you’re interested.

**Key takeaways**

Given a sample of data, you can try out different lines that could fit your data. You could calculate the sum of squared residuals for each line to determine which fits your data best. As a data professional, it’s important to understand what the sum of squared residuals represents, and how to calculate it on your own. Thankfully, we have computers and programming languages that can calculate the sum of squared residuals and perform OLS for us. You can explore the deeper math behind OLS and SSR on your own if you wish!

**Resources**

* [Parameter Estimation - Ordinary Least Squares Method](https://www.geo.fu-berlin.de/en/v/soga-py/Basics-of-statistics/Linear-Regression/Simple-Linear-Regression/Parameter-Estimation/index.html): *Rudolph, A., Krois, J., Hartmann, K. (2023): Statistics and Geodata Analysis using Python (*[*SOGA-Py*](https://www.geo.fu-berlin.de/soga-py)*). Department of Earth Sciences, Freie Universitaet Berlin.*

**Correlation and the intuition behind simple linear regression**

So far you’ve learned that simple linear regression is a technique that estimates the linear relationship between one independent variable, X, and one continuous dependent variable, Y. You’ve also learned about ordinary least squares estimation (OLS), which is a common way to determine the coefficients of the regression line—the line of “best fit” through the data. In this reading, you’ll explore the meaning of correlation; learn about *r*, or the “correlation coefficient;” and discover how to determine the regression equation. This knowledge will help you better understand relationships between variables, and thus how linear regression works.

**Correlation**

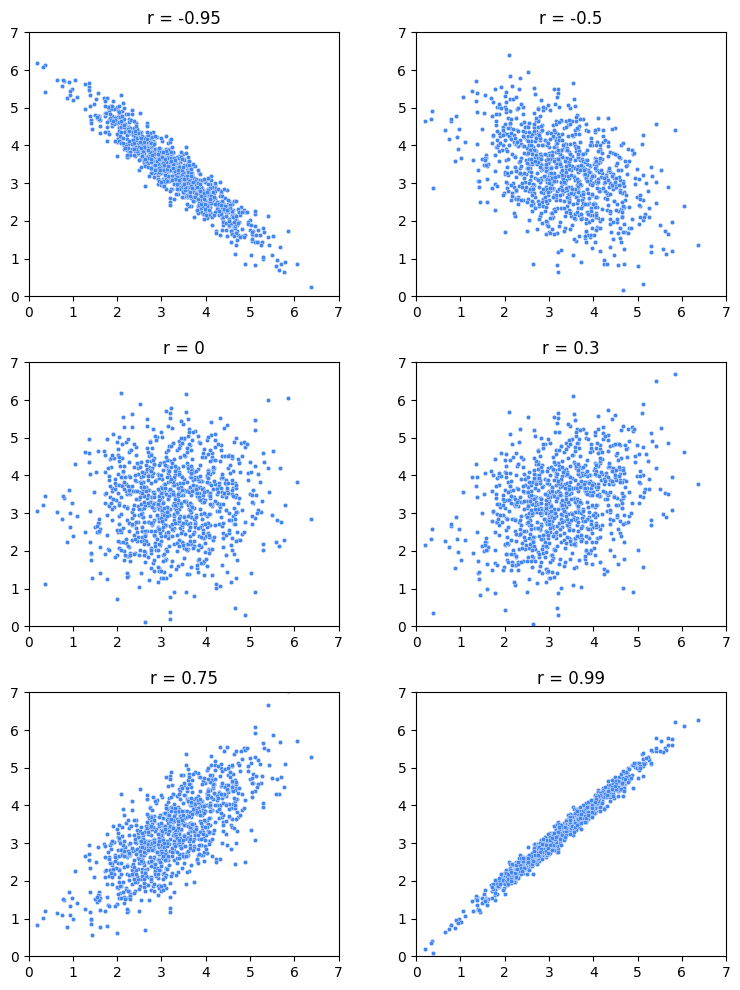
Correlation is a measurement of the way two variables move together. If there is a strong correlation between the variables, then knowing one will be very helpful to predict the other. However, if there is a weak correlation between two variables, then knowing the value of one will not tell you much about the value of the other. In the context of linear regression, correlation refers to *linear* correlation: as one variable changes, so does the other at a constant rate.

In the statistics course, you learned that a continuous variable can be summarized using some basic numbers. Two of these summary statistics are:

* **Average:** A measurement of central tendency (mean, median, or mode)
* **Standard deviation:** A measurement of spread

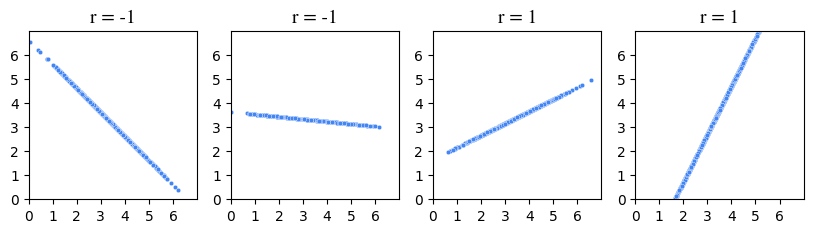
When two variables are summarized together, there is another relevant statistic called ***r***, **Pearson’s correlation coefficient** (named after the person who helped develop it), or simply the linear **correlation coefficient.** The correlation coefficient quantifies the strength of the linear relationship between two variables. It always falls in the range of [-1, 1]. When *r* is negative, there is a negative correlation between the variables: as one increases, the other decreases. When *r* is positive, there is a positive correlation between the variables: as one increases, so too does the other. When *r* = 0, there is no *linear* correlation between the variables. Note that there are cases where one variable might be precisely determined by another—like y=x2 or y=sin(x)—but the value of the *linear* correlation between X and Y would nonetheless be low or zero because their relationship is non-linear.

The following figure depicts scatterplots of bivariate (bi = “two”, variate = “variables”) data where each variable has the same mean and standard deviation and only the correlation coefficient varies.



Notice that the closer to -1 or 1 *r* is, the more linear the data appears. When *r* is exactly 1 or exactly -1, then the variables are perfectly correlated, and their graph is a line. When *r* is zero, there is no correlation between the variables, and, in this example, the data appears as a shapeless cloud of points.

However, *r* only tells you the strength of the linear correlation between the variables; it does not tell you anything about the magnitude of the slope of the relationship between the variables aside from its sign. For example, variables with *r*=1 wouldn't tell you if increasing X by one would lead to Y increasing by 10, 100, 0.1, or something else. It would only tell you that you can be sure that it *would* increase. This fact is illustrated in the following figure, where even though the slopes of the lines are all different, *r* is only either -1 or 1. If the line is perfectly horizontal or perfectly vertical, then *r* is undefined. (If you’re wondering why, refer to the equation below. One of the terms in the denominator would equal zero, which would make the whole denominator equal zero, which would result in an undefined solution.)



**Calculate *r***

The formula for *r* is:

𝑟=𝑐𝑜𝑣𝑎𝑟𝑖𝑎𝑛𝑐𝑒(𝑋 ⁣,𝑌)(𝑆 ⁣𝐷 𝑋)(𝑆 ⁣𝐷 𝑌)*r*=(*SD* *X*)(*SD* *Y*)*covariance*(*X*,*Y*)​

where:

𝑐𝑜𝑣𝑎𝑟𝑖𝑎𝑛𝑐𝑒(𝑋 ⁣,𝑌)=∑𝑖=1𝑛(𝑥𝑖−𝑥ˉ)(𝑦𝑖−𝑦ˉ)𝑛*covariance*(*X*,*Y*)=*ni*=1∑*n*​(*xi*​−*x*ˉ)(*yi*​−*y*ˉ​)​

**Note:** The formulas for *r* and covariance given here represent those used for entire populations. For samples, the denominator of the covariance formula is *n - 1* and, similarly, the standard deviations in the formula for *r* are calculated using *n - 1* instead of *n.* For simplicity, this reading will use the population formulas in its demonstrations.

An easier way of thinking about this calculation is: the numerator—the covariance—represents the extent to which X and Y vary together from their respective means. When this value is positive, it suggests that high values of X tend to be associated with high values of Y, indicating a positive correlation. Conversely, if the value is negative, it suggests that high values of X tend to be associated with low values of Y and vice versa, indicating a negative correlation.

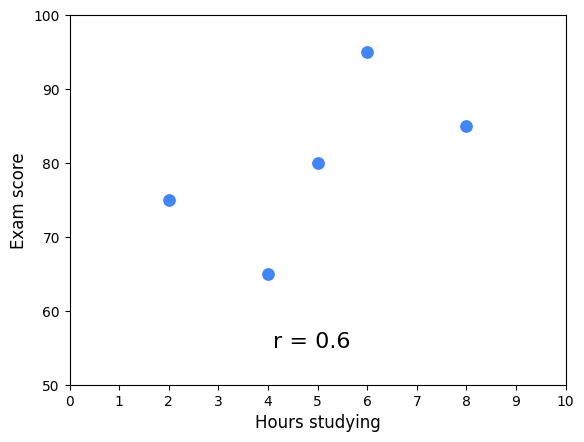
The denominator—the product of the standard deviations—standardizes the units of the numerator. It adjusts for the inherent variability of the individual variables. This makes *r* a statistic without a unit. It is a pure number, without dimension.

An equivalent way to calculate *r* is to convert each data point in each variable to standard units (subtract the mean, divide by the standard deviation), then take the average of the products.

Here’s an example. Suppose five students took an exam and you recorded how many hours they spent studying and also their grade. The following table breaks out the calculation of *r.*

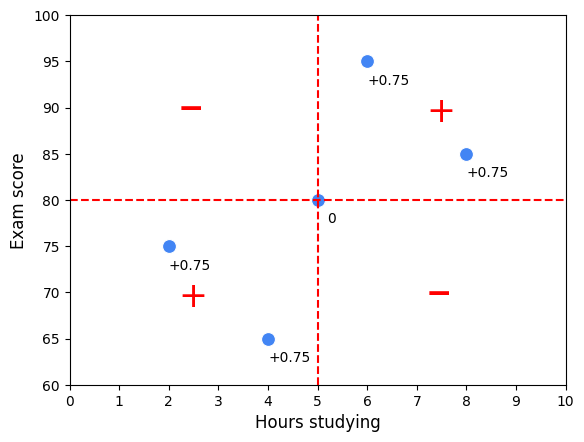
| **Hours studying (X)** | **Exam grade (Y)** | **X in standard units** | **Y in standard units** | **Product of standard units** |
| --- | --- | --- | --- | --- |
| 2 | 75 | -1.5 | -0.5 | 0.75 |
| 4 | 65 | -0.5 | -1.5 | 0.75 |
| 5 | 80 | 0 | 0 | 0 |
| 6 | 95 | 0.5 | 1.5 | 0.75 |
| 8 | 85 | 1.5 | 0.5 | 0.75 |
| **mean X = 5**  **SD X = 2** | **mean Y = 80**  **SD Y = 10** |  |  | **mean of products (r) = 0.6** |

The correlation coefficient is 0.6. Here is a graph of this data:



Notice that the cloud of points slopes upwards. This corresponds with *r* being positive. The correlation coefficient works as an indicator of association because it uses the product of each variable’s deviation from its mean. When the product is positive, it means *both* the X and the Y values are either below their respective means (negative standard units) or above their respective means (positive standard units). They vary together. However, when this product is negative, it means one of the values is above its mean and the other is below it. They vary in opposing directions relative to their respective means.

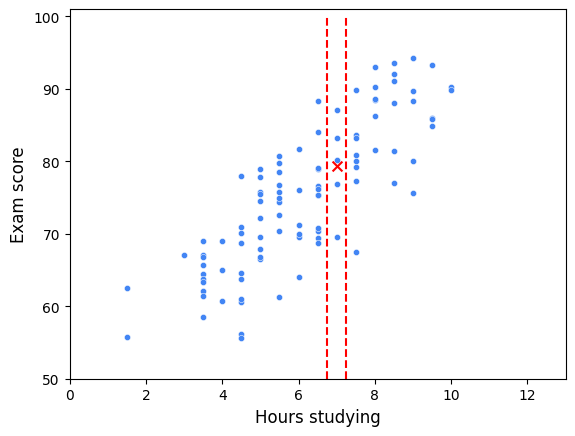
The following figure illustrates this idea. The figure is divided into quadrants. The vertical line represents the mean X value and the horizontal line represents the mean Y value. Each point is labeled with the product of its standardized scores (refer to the table above). The average of these scores is *r*. When *r* is positive, more points will tend to be in the positive quadrants, and vice versa.



**Regression**

In the absence of any other information, if you had to guess a randomly selected student’s exam score, the best way for you to minimize your error would be to guess the average of all the students’ scores. But what if you also knew how many hours that student studied? Now, your best guess might be the average score of only the students who studied for that many hours.

Here is an example using a sample of 100 students with study times rounded to the nearest half hour. Suppose you were told a student studied for seven hours. To guess their exam score, one way to minimize error is to guess the average of only the students who studied for seven hours.



In this scatterplot, all of the students who studied for seven hours fall between the two vertical lines. Their mean exam score is represented by an X. Linear regression expands on this concept. A regression line represents the estimated average value of Y for every value of X, given the assumptions and limitations of a linear model. In other words, the actual average Y values for each X might not lie exactly on the regression line if the relationship between X and Y is not perfectly linear or if there are other factors influencing Y that are not included in the model. The regression line attempts to balance out these influences to find a straight-line relationship that best fits the data as a whole. It’s an estimation of the central tendency of Y, given X.

**The regression equation**

Now that you know about *r* and you better understand the concept of regression, you’re ready to put everything together to find the line of best fit through the data. The formula for this line is known as the regression equation. There are two keys to this step.

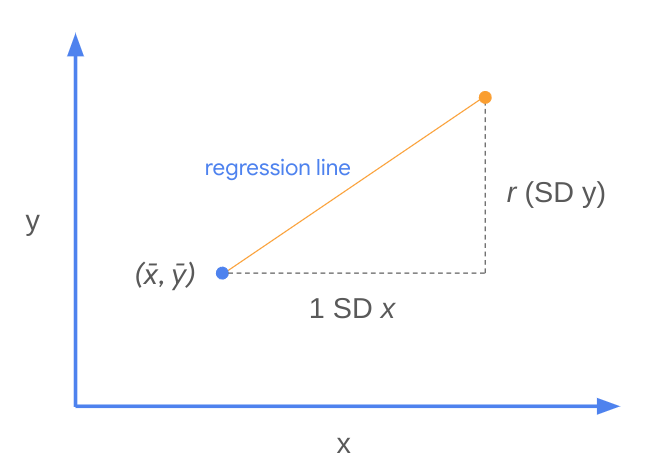
The first is:

* The mean value of X and the mean value of Y (i.e., point (*x̄*, *ȳ*)) will always fall on the regression line.

The second is to understand what *r* means:

* For each increase of one standard deviation in X, there is an expected increase of *r* standard deviations in Y, on average over X.

The following figure illustrates how these concepts work together to determine the regression line.



In other words, the slope of the regression line is:

𝑚=𝑟(𝑆 ⁣𝐷 𝑦)𝑆 ⁣𝐷 𝑥*m*=*SD* *xr*(*SD* *y*)​

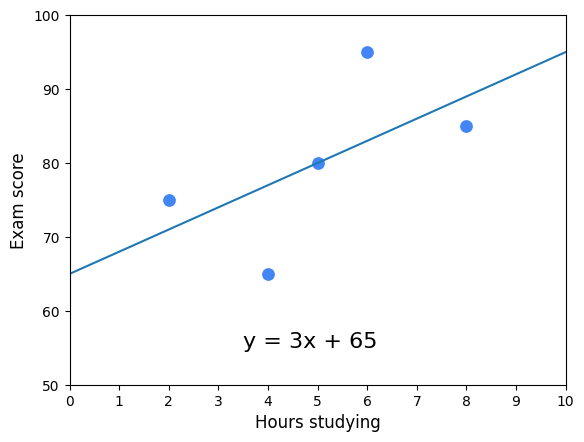
This is *m* in the formula for a line: *y* = *mx* + *b*. The intercept, represented by *b,* is therefore: *b* = *y* - *mx*. Because you know that point (*x̄*, *ȳ*) is always on the regression line, you can plug in the *x* and *y* values from this point to calculate the intercept. Here’s an example using the original sample of five students.

|  | **Hours studying (X)** | **Exam grade (Y)** |
| --- | --- | --- |
| **mean:** | 5 | 80 |
| **SD:** | 2 | 10 |
| **r:** | **0.6** |  |

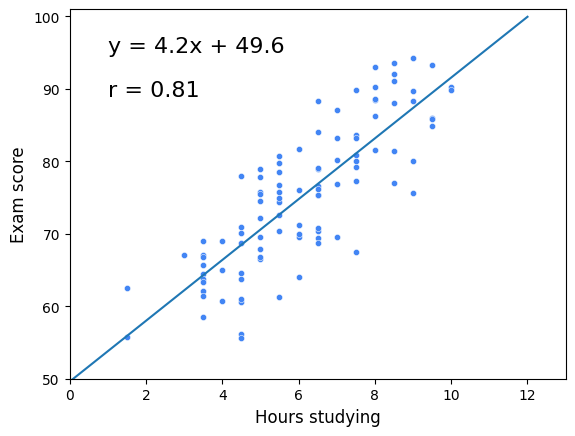
Broken into steps:

1. Calculate slope: 𝑚=𝑟(𝑆 ⁣𝐷 𝑦)𝑆 ⁣𝐷 𝑥=0.6(10)2=3.*m*=*SD* *xr*(*SD* *y*)​=20.6(10)​=3.
2. Calculate the intercept: Substitute *x̄*, *ȳ,* and *m* into the equation *y* = *mx* + *b*:   80 = 3(5) + *b*  →   *b =* 65.
3. Generalize to get the regression equation: *y* = 3*x* + 65.

Here is the regression line overlaid onto the data:



This is referred to as "the regression of Y on X." Here is the regression line for all 100 students:



**Key Takeaways**

Linear regression is one of the most important tools that data professionals use to analyze data. Understanding the fundamental building blocks of simple linear regression will help you as you continue learning about more complex methods of regression analysis. Here are some key points to keep in mind:

* Correlation is a measurement of the way two variables move together.
* *r (a.k.a.* Pearson’s correlation coefficient, a.k.a. correlation coefficient) quantifies the strength of the linear relationship between two variables.
  + It always falls in the range of [-1, 1].
  + Variables that tend to vary together from their means are positively correlated. Conversely, variables that tend to vary in opposite ways to their respective means are negatively correlated.
* The regression line estimates the average y value for each *x* value. It minimizes the error when estimating *y*, given x.
* The slope of the regression line is 𝑟(𝑆 ⁣𝐷 𝑦)𝑆 ⁣𝐷 𝑥*SD* *xr*(*SD* *y*)​.
* The point (*x̄*, *ȳ*) is always on the regression line.

**Test your knowledge: Foundations of linear regression**

Review Learning Objectives

**1.**

Question 1

Fill in the blank: The best fit line is the line that fits the data best by minimizing some \_\_\_\_\_.

loss function

Correct

The best fit line is the line that fits the data best by minimizing some loss function. To find the best fit line, it’s necessary to measure error, which is the difference between the observed values and the predicted values generated by the model.

**2.**

Question 2

What is the sum of the squared differences between each observed value and the associated predicted value?

Sum of squared residuals

Correct

The sum of squared residuals is the sum of the squared differences between each observed value and the associated predicted value. Data professionals use this sum to capture a summary of total error in the model.

**3.**

Question 3

What does the circumflex symbol, or "hat" (^), indicate when used over a coefficient?

The coefficient is an estimate or predicted value

**Model assumptions**

Statements about the data that must be true in order to justify the use of a particular modelling techniques

**Linear regression assumptions**

* Linearity
* Normality
* Independent observations
* Homoscedasticity

**Linearity assumption**

Each predictor variable (X\_i) is linearly related to the outcome variable (Y).

**Normality assumption**

The residuals or errors are normally distributed

**Independent observation assumption**

Each observation in the dataset is independent

**Homoscedasticity assumption**

The variation of the residuals (errors) is constant or similar across the model

**Question**

There are four assumptions of simple linear regression, including linearity, normality, and independent observations. What is the fourth assumption?

Homoscedasticity

Correct

The four assumptions of simple linear regression are linearity, normality, independent observations, and homoscedasticity. Linearity assumes that each predictor variable *Xi* is linearly related to the outcome variable *Y*. Normality assumes that the residual values are normally distributed. Independent observation assumes that each observation in the dataset is independent. And homoscedasticity assumes the values have the same variance.

**The four main assumptions of simple linear regression**

In this reading, you will review the four main assumptions of simple linear regression, how to check that the assumptions are met, and what to do if an assumption is not met. You can use the additional resources to replicate the graphs and explore assumptions on your own. If there are any terms not defined in this reading, refer to the glossary of terms available throughout the course at the end of each module. This reading will cover:

* Simple linear regression assumptions
* How to check the validity of the assumptions
* What to do if an assumption is violated

**Simple linear regression assumptions**

To recap, there are four assumptions of simple linear regression:

1. **Linearity:** Each predictor variable (Xi) is linearly related to the outcome variable (Y).
2. **Normality:** The errors are normally distributed.**\***
3. **Independent Observations:** Each observation in the dataset is independent.
4. **Homoscedasticity:** The variance of the errors is constant or similar across the model.**\***

**\*Note on errors and residuals**

This course has rather interchangeably used the terms "errors" and "residuals" in connection with regression. You may see this in other online resources and materials throughout your time as a data professional. In actuality, there is a difference:

* **Residuals** are the difference between the predicted and observed values. You can calculate residuals after you build a regression model by subtracting the predicted values from the observed values.
* **Errors** are the natural noise assumed to be in the model.
* Residuals are used to estimate errors when checking the normality and homoscedasticity assumptions of linear regression.

**How to check the validity of the assumptions**

As previously reviewed, many of the simple linear regression assumptions can be checked through data visualizations. Some assumptions can be checked before a model is built, and others can only be checked after the model is constructed, and predicted values are calculated.

**Linearity**

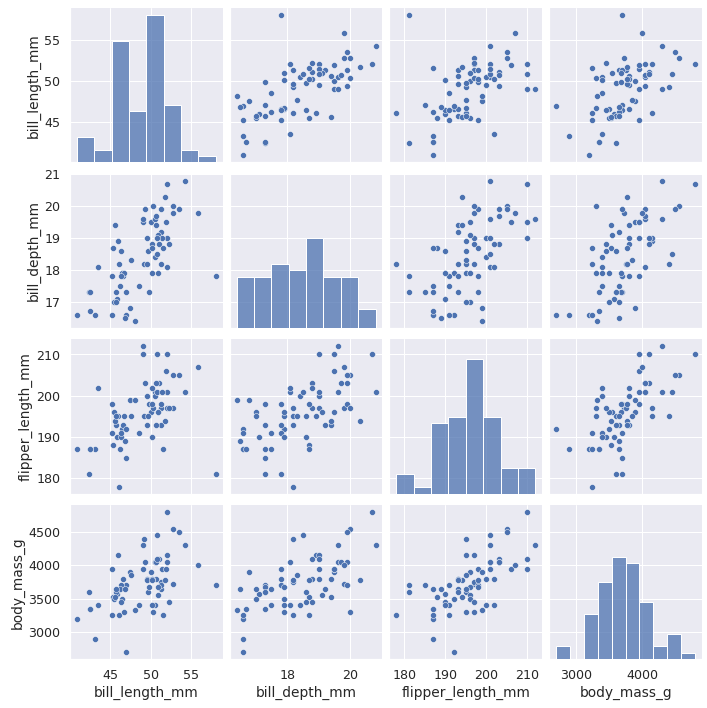
In order to assess whether or not there is a linear relationship between the independent and dependent variables, it is easiest to create a scatterplot of the dataset. The independent variable would be on the x-axis, and the dependent variable would be on the y-axis. There are a number of different Python functions that you can use to read in the data and to create a scatterplot. Some packages used for data visualizations include Matplotlib, seaborn, and Plotly. Testing the linearity assumption should occur before the model is built.

1

2

# Create pairwise scatterplots of Chinstrap penguins data

sns.pairplot(chinstrap\_penguins)



**Normality**

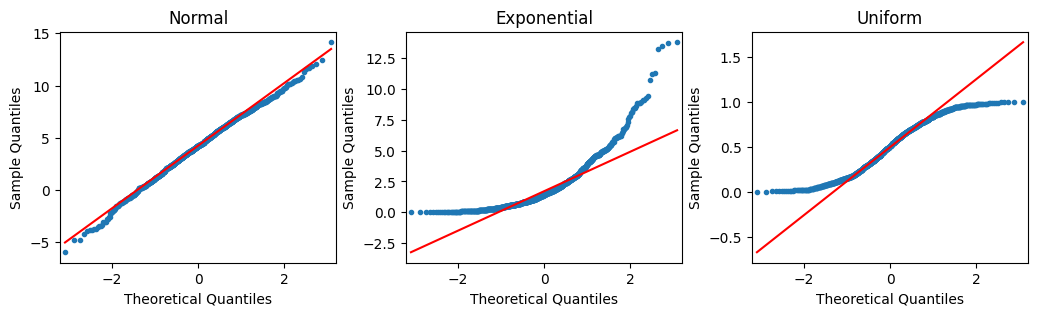
The normality assumption **focuses on the errors**, which can be estimated by the residuals, or the difference between the observed values in the data and the values predicted by the regression model. For that reason, the normality assumption can only be confirmed **after** a model is built and predicted values are calculated. Once the model has been built, you can either create a QQ-plot to check that the residuals are normally distributed, or create a histogram of the residuals. Whether the assumption is met is up to some level of interpretation.

**Quantile-quantile plot**

The **quantile-quantile plot** (**Q-Q plot**) is a graphical tool used to compare two probability distributions by plotting their quantiles against each other. Data professionals often prefer Q-Q plots to histograms to gauge the normality of a distribution because it’s easier to discern whether a plot adheres to a straight line than it is to determine how closely a histogram follows a normal curve. Here’s how Q-Q plots work when assessing the normality of a model’s residuals:

1. **Rank-order the residuals**. Sort your *n* residuals from least to greatest. For each one, calculate what percentage of the data falls at or below this rank. These are the *n* quantiles of your data.
2. **Compare to a normal distribution.** Divide a standard normal distribution into *n*+1 equal areas (i.e., slice it *n* times). If the residuals are normally distributed, the quantile of each residual (i.e., what percentage of the data falls below each ranked residual) will align closely with the corresponding z-scores of each of the *n* cuts on the standard normal distribution (these can be found in a normal z-score table or, more commonly, using statistical software).
3. **Construct a plot.** A Q-Q plot has the known quantile values of a standard normal distribution along its x-axis and the rank-ordered residual values on its y-axis. If the residuals are normally distributed, the quantile values of the residuals will correspond with those of the standardized normal distribution, and both will increase linearly. If you first standardize your residuals (convert to z-scores by subtracting the mean and dividing by the standard deviation), the two axes will be on identical scales, and, if the residuals are indeed normally distributed, the line will be at a 45° angle. However, standardizing the residuals is not a requirement of a Q-Q plot. In either case, if the resulting plot is not linear, the residuals are not normally distributed.

In the following figure, the first Q-Q plot depicts data that was taken from a normal distribution. It forms a line when plotted against the quantiles of a standard normal distribution. The second plot depicts data that was drawn from an exponential distribution. The third plot uses data drawn from a uniform distribution. Notice how the second and third plots don’t adhere to a line.



**How to code a Q-Q plot**

Thankfully, you don’t have to manually perform the steps outlined previously. There are computing libraries to handle that. One way to create a Q-Q plot is to use the statsmodels library. If you import **statsmodels.api**, you can use the [qqplot()](https://www.statsmodels.org/stable/generated/statsmodels.graphics.gofplots.qqplot.html) function directly. The example below uses the residuals from a **statsmodels ols** model object. The model regresses penguins’ flipper length on their bill depth (Y on X).

1

2

3

4

5

6

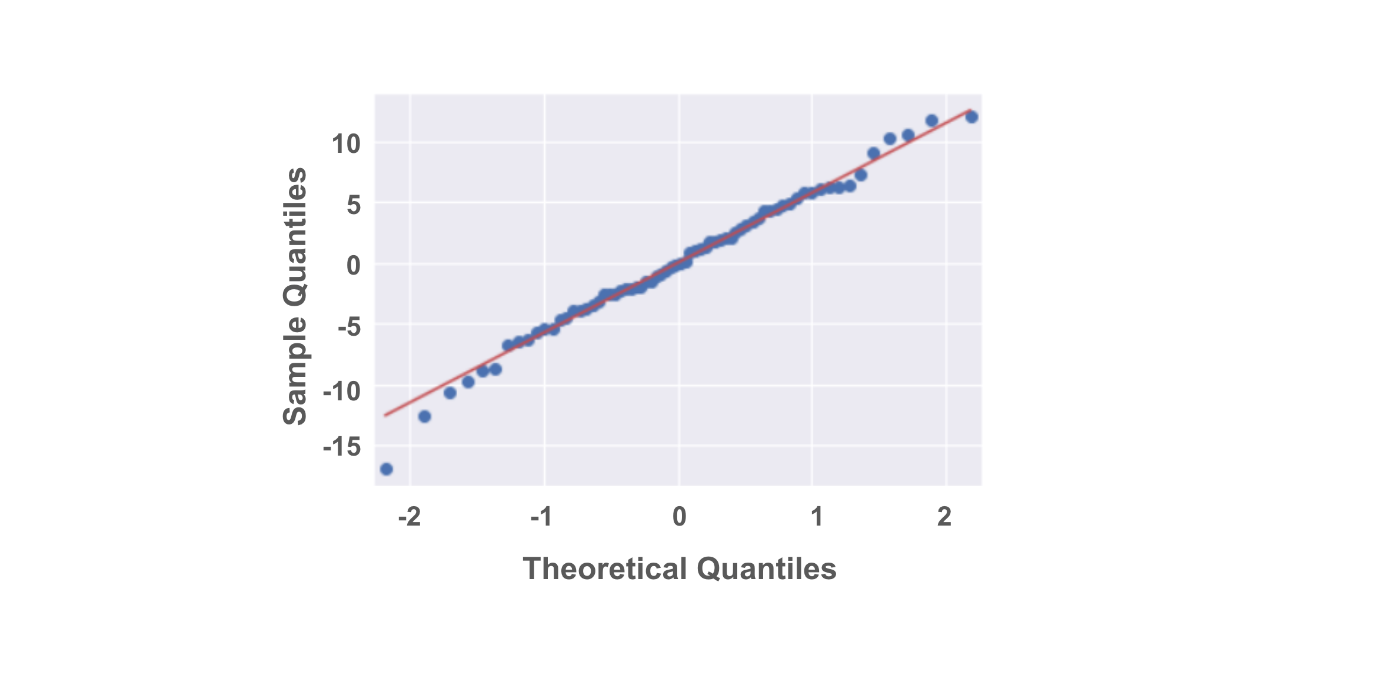
import statsmodels.api as sm

import matplotlib.pyplot as plt

residuals = model.resid

fig = sm.qqplot(residuals, line = 's')

plt.show()



And here is a histogram of the same data:

1

2

3

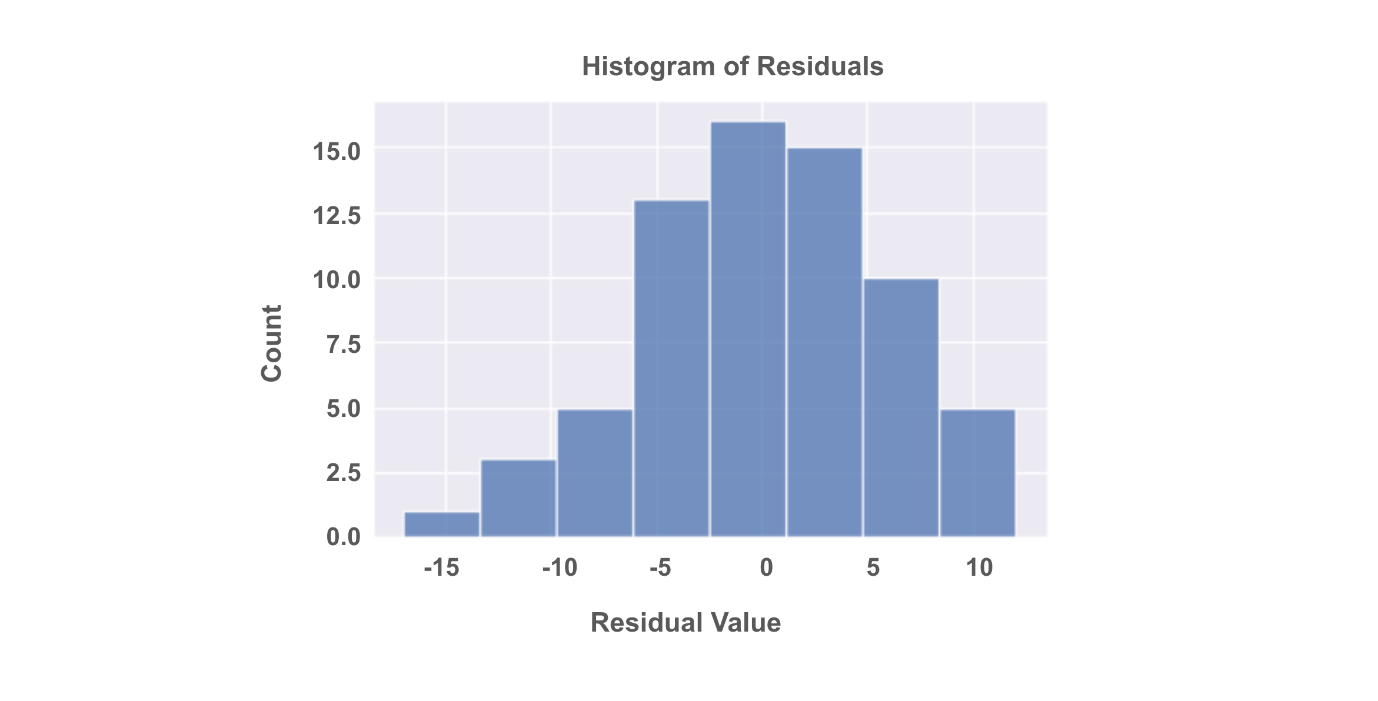
4

fig = sns.histplot(residuals)

fig.set\_xlabel("Residual Value")

fig.set\_title("Histogram of Residuals")

plt.show()



**Independent Observations**

Whether or not observations are independent is dependent on understanding your data. Asking questions like:

* How was the data collected?
* What does each data point represent?
* Based on the data collection process, is it likely that the value of one data point impacts the value of another data point?

An objective review of these questions, which would include soliciting insights from others who might notice things you don't, can help you determine whether or not the independent observations assumption is violated. This in turn will allow you to determine your next steps in working with the dataset at hand.

**Homoscedasticity**

Like the normality assumption, the homoscedasticity assumption concerns the residuals of a model, so it can only be evaluated after a regression model has already been constructed. A scatterplot of the fitted values (i.e., the model’s predicted Y values) versus the residuals can help determine whether the homoscedasticity assumption is violated.

1

2

3

4

5

6

7

import matplotlib.pyplot as plt

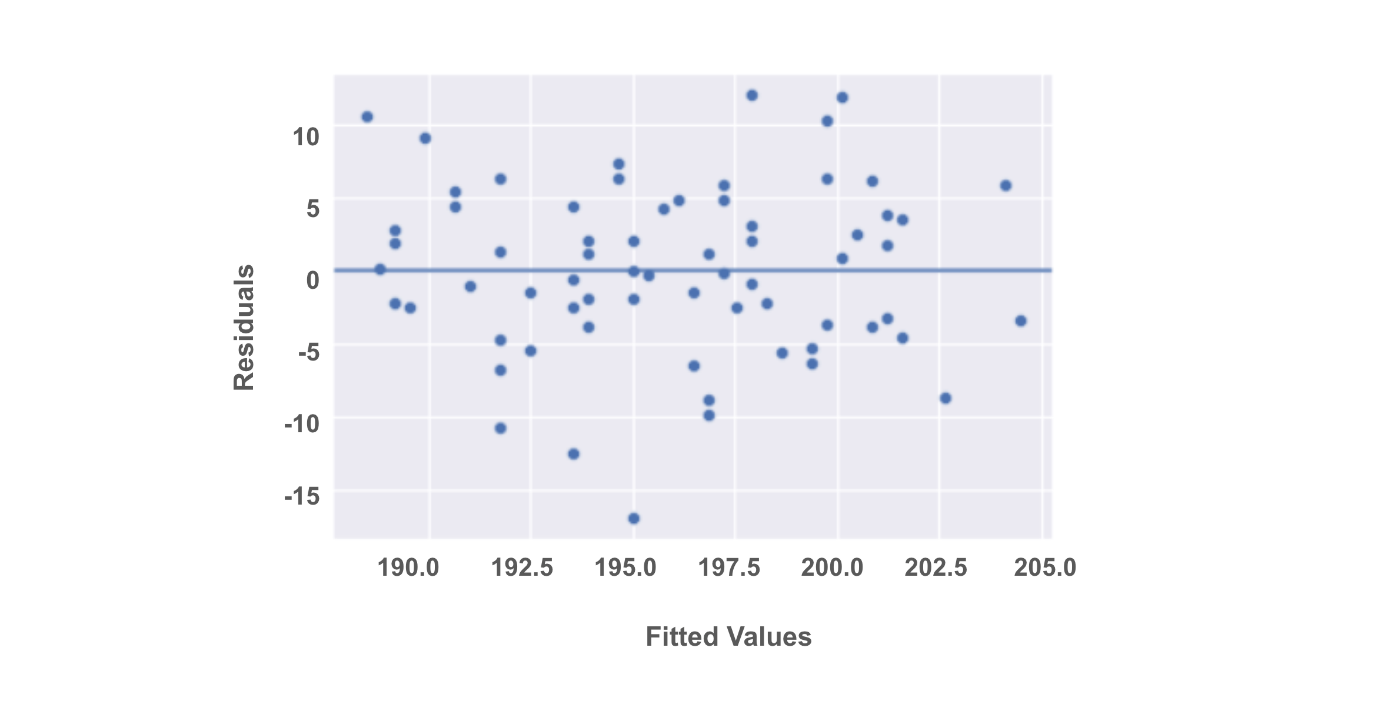
fig = sns.scatterplot(fitted\_values, residuals)

fig.axhline(0)

fig.set\_xlabel("Fitted Values")

fig.set\_ylabel("Residuals")

plt.show()



**What to do if an assumption is violated**

Now that you've reviewed the four assumptions and how to test for their violations, it’s time to discuss some common next steps you can take once an assumption is violated. Keep in mind that if you transform the data, this might change how you interpret the results. Additionally, if these potential solutions don’t work for your data, you have to consider trying a different kind of model.

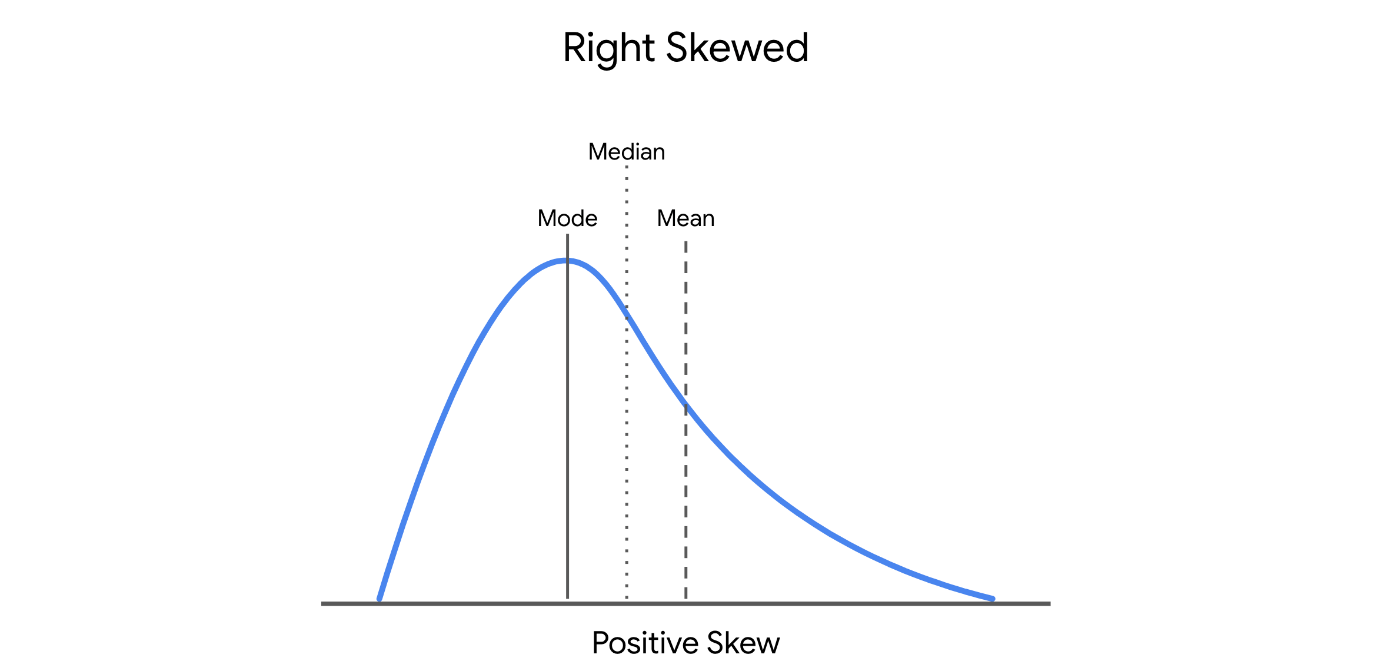
For now, focus on a few essential approaches to get you started!

**Linearity**

* Transform one or both of the variables, such as taking the logarithm.
  + For example, if you are measuring the relationship between years of education and income, you can take the logarithm of the income variable and check if that helps the linear relationship.

**Normality**

* Transform one or both variables. Most commonly, this would involve taking the logarithm of the outcome variable.
  + When the outcome variable is right skewed, such as income, the normality of the residuals can be affected. So, taking the logarithm of the outcome variable can sometimes help with this assumption.
  + If you transform a variable, you will need to reconstruct the model and then recheck the normality assumption to be sure. If the assumption is still not satisfied, you’ll have to continue troubleshooting the issue.



**Independent observations**

* Take just a subset of the available data.
  + If, for example, you are conducting a survey and get responses from people in the same household, their responses may be correlated. You can correct for this by just keeping the data of one person in each household.
  + Another example is if you are collecting data over a time period. Let’s say you are researching data on bike rentals. If you collect your data every 15 minutes, the number of bikes rented out at 8:00 a.m. might correlate with the number of bikes rented out at 8:15 a.m. But, perhaps the number of bikes rented out is independent if the data is taken once every 2 hours, instead of once every 15 minutes.

**Homoscedasticity**

* Define a different outcome variable.
  + If you are interested in understanding how a city’s population correlates with the number of restaurants in a city, you know that some cities are much more populous than others. You can then redefine the outcome variable as the ratio of population to restaurants.
* Transform the Y variable.
  + As with the above assumptions, sometimes taking the logarithm or transforming the Y variable in another way can potentially fix inconsistencies with the homoscedasticity assumption.

**Key takeaways**

* There are four key assumptions for simple linear regression: linearity, normality, independent observations, and homoscedasticity.
* There are different ways to check the validity of each assumption. Some assumptions can be checked before the model is built, while some can be checked after the model is built.
* There are ways to work with the data that can correct for violations of model assumptions.
* Changing the variables will change the interpretation.
* If the assumptions are violated, even after data transformations, you should consider other models for your data.

**Resources for more information**

* [Download the seaborn penguins dataset here](https://raw.githubusercontent.com/mwaskom/seaborn-data/master/penguins.csv)
* More information about the penguins dataset: [Introduction to Palmer penguins](https://allisonhorst.github.io/palmerpenguins/articles/intro.html)
* More information about Q-Q plots: [Normal Quantile-Quantile Plots (video from jbstatistics)](https://www.youtube.com/watch?v=X9_ISJ0YpGw)

**Scatterplot matrix**

A series of scatterplots that show the relationships between pairs or variables.

**Code functions and documentation**

In this reading, you will review some of the code from the videos using a different subset of the penguin data. This reading will also share some tips when approaching the statsmodels documentation. This is a good opportunity to review Python functionality in conjunction with exploratory data analysis, basic data cleaning, and model construction.

**Review functions from video**

**Load the dataset**

The first few lines of code set up the coding environment and loaded the data. As you might be familiar with, you can call on the **import** function to import any necessary packages. You should use conventional aliases as needed. The example below references a dataset on penguins available through the **seaborn** package.

1

2

3

4

5

6

7

8

9

# Import packages

import pandas as pd

import seaborn as sns

# Load dataset

penguins = sns.load\_dataset("penguins")

# Examine first 5 rows of dataset

penguins.head()

**Clean data**

After loading the data, the data was cleaned up to create a subset of data for the purposes of our course. The example isolates just the Chinstrap penguins from the dataset and drops rows with missing data.

The index of the dataframe is reset using the **reset\_index()** [function](https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.reset_index.html). When you subset a dataframe, the original row indices are retained. For example, let’s say there were Adelie or Gentoo penguins in rows 2 and 3. By subsetting the data just for Chinstrap penguins, your new dataframe would be listed as row 1 and then row 4, as rows 2 and 3 were removed. By resetting the index of the dataframe, the row numbers become rows 1, 2, 3, etc. The data frame becomes easier to work with in the future.

Review the code below. You are encouraged to run the code in your own notebook.

1

2

3

4

5

# Subset just Chinstrap penguins from data set

chinstrap\_penguins = penguins[penguins["species"] == "Chinstrap"]

# Reset index of dataframe

chinstrap\_penguins.reset\_index(inplace = True, drop = True)

**Setup for model construction**

Now that the data is clean, you are able to plot the data and construct a linear regression model. First, extract the one X variable, **bill\_depth\_mm**, and the one Y variable, **flipper\_length\_mm**, that you are targeting.

2

ols\_data = chinstrap\_penguins[["bill\_depth\_mm", "flipper\_length\_mm"]]

Because this example is using statsmodels, save the ordinary least squares formula as a string so the computer can understand how to run the regression. The Y variable, **flipper\_length\_mm** comes first, followed by a tilde and the name for the X variable, **bill\_depth\_mm**.

1

2

# Write out formula

ols\_formula = "flipper\_length\_mm ~ bill\_depth\_mm"

**Construct the model**

In order to construct the model, you’ll first need to import the **ols** function from the **statsmodels.formula.api** interface.

1

2

# Import ols function

from statsmodels.formula.api import ols

Next, plug in the formula and the saved data into the **ols** function. Then, use the **fit** method to fit the model to the data. Lastly, use the **summary** method to get the results from the regression model.

1

2

3

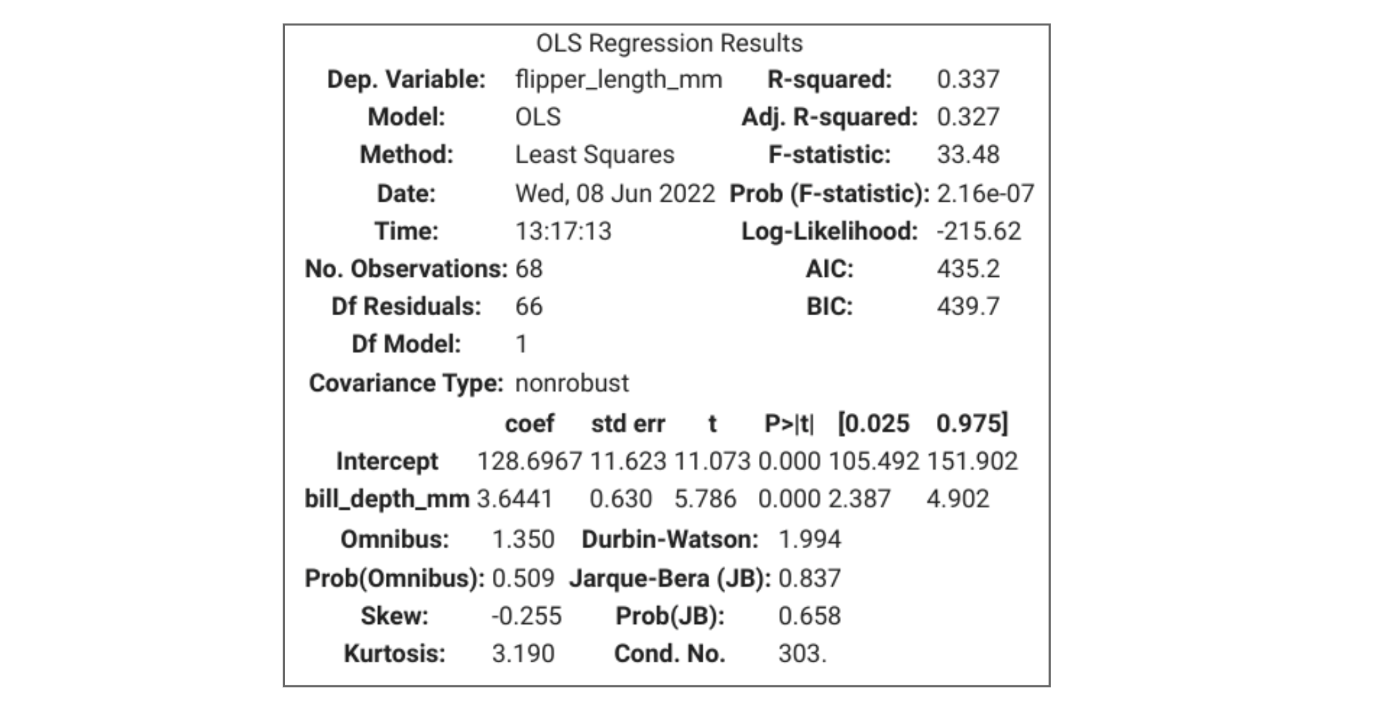
4

# Build OLS, fit model to data

OLS = ols(formula = ols\_formula, data = ols\_data)

model = OLS.fit()

model.summary()



**Model predictions and residuals**

You can access the predictions and residuals from a fitted [statsmodels.regression.linear\_model.OLS](https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.html) or [statsmodels.regression.linear\_model.OLSResults](https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLSResults.html) object as follows.

**Predictions**

Use the model’s [predict()](https://www.statsmodels.org/stable/generated/statsmodels.regression.linear_model.OLS.predict.html#statsmodels.regression.linear_model.OLS.predict) method, passing to it an array containing the values of the independent variable(s):

1

predictions = model.predict(chinstrap\_penguins[["bill\_depth\_mm"]])

**Residuals**

Use the model’s **resid** attribute:

1

residuals = model.resid

**Navigating statsmodels documentation**

It can require significant work to approach a new Python package or a new set of Python functions, especially when first coding. The benefit of Python being an open source programming language is that there is a strong Python community asking and answering questions. Part of being a successful data professional is knowing how to make your code work and troubleshooting when your code breaks. One way to do this is to go directly to the source, or the official documentation of a particular package.

You’ve been using the **statsmodels** package to build simple linear regression models. The [statsmodels documentation](https://www.statsmodels.org/devel/api.html) is available online and is updated regularly. Specifically, you are using the [statsmodels.formula.api interface](https://www.statsmodels.org/devel/api.html#statsmodels-formula-api) to perform ordinary least squares estimation.

Examining the page on the **ols** [function](https://www.statsmodels.org/devel/generated/statsmodels.formula.api.ols.html#statsmodels.formula.api.ols) or the function that performs OLS estimation, you will observe the different function parameters that are allowed, with some notes about each.

Unfortunately, at this time, the statsmodels documentation does not include code examples of how to use the function. If you find documentation that doesn’t provide as many examples as you need, or documentation that doesn’t provide examples that you need to troubleshoot your code, remember that you can always search online for the function you’re trying to use and explore how others in the Python community have handled comparable problems.

**Key takeaways**

* You can review this reading to refresh your memory about the code in the corresponding video.
* You can view the statsmodels (or any other package’s) documentation as needed.
* If the documentation does not hold the answer you’re looking for, you can always turn to the internet to check out other people’s work.

**Test your knowledge: Assumptions and construction in Python**

**1.**

Question 1

How does a data professional determine if a linearity assumption is met?

1 / 1 point

They confirm whether data on the X-Y coordinate falls along a straight line.

They confirm whether data on the X-Y coordinate falls along a downward curved line.

They confirm whether data on the X-Y coordinate falls along an upward curved line.

They confirm whether data on the X-Y coordinate resembles a random cloud.

Correct

A data professional determines if a linearity assumption is met by confirming whether data on the X-Y coordinate falls along a straight line. A linearity assumption is passed when each predictor variable X is linearly related to the outcome variable Y.

**2.**

Question 2

Which of the following statements accurately describes the normality assumption?

1 / 1 point

The normality assumption can only be confirmed while a model is being built.

The normality assumption can be confirmed anytime during model building.

The normality assumption can only be confirmed after a model is built.

The normality assumption can only be confirmed before a model is built.

Correct

The normality assumption can only be confirmed after a model is built. It focuses on the model errors, which can be estimated by the residuals.

**3.**

Question 3

A data professional uses a scatterplot to plot residuals and predicted values from a regression model to check for homoscedasticity and finds that this assumption is met. What shape do the points in the scatterplot appear as?

1 / 1 point

Random cloud

Curved line

Straight line

Cone

Correct

The residuals appear as a random cloud of points, which confirms the variation of residuals is consistent or similar across the model and satisfies the assumption of homoscedasticity.

**4.**

Question 4

What type of visualization uses a series of scatterplots that show the relationships between pairs of variables?

1 / 1 point

Scatterplot matrix

Residual matrix

Linear matrix

Scatterplot residuals

Correct

A scatterplot matrix uses a series of scatterplots that show the relationships between pairs of variables. This helps data professionals assess whether there is a linear relationship between the independent and dependent variables.

**Confidence band**

The area surrounding the line that describes the uncertainty around the predicted outcome at every value of X

**Question**

In a linear regression model, what is the area surrounding the regression line that describes the uncertainty around the predicted outcome at every value of X?

confidence band

Correct

In a linear regression model, a confidence band is the area surrounding the regression line that describes the uncertainty around the predicted outcome at every value of X. A confidence band reveals the confidence interval for each point on a regression line. It is another way to report findings responsibly.

**Interpret measures of uncertainty in regression**

**Goal of Reading**

In this reading, we will continue exploring uncertainty in regression analysis, specifically through confidence intervals, confidence bands, and p-values. Together, we will:

* Review key concepts
* Discuss how to interpret measures of uncertainty
* Review sample graphs

**Review of Concepts**

Recall that we can represent a simple linear regression line as y=β0+β1X*y*=*β*0​+*β*1​*X*.

Since regression analysis utilizes **estimation** techniques, there is always a level of uncertainty surrounding the predictions made by regression models. To represent the error, we can actually rewrite the equation to include an error term, represented by the letter ϵ*ϵ* (pronounced “epsilon”): y=β0+β1X+ϵ*y*=*β*0​+*β*1​*X*+*ϵ*.

There is one residual, also known as the difference between the predicted and actual value, for each data point in the dataset used to construct the model. We can then quantify how uncertain the entire model is through a few measures of uncertainty:

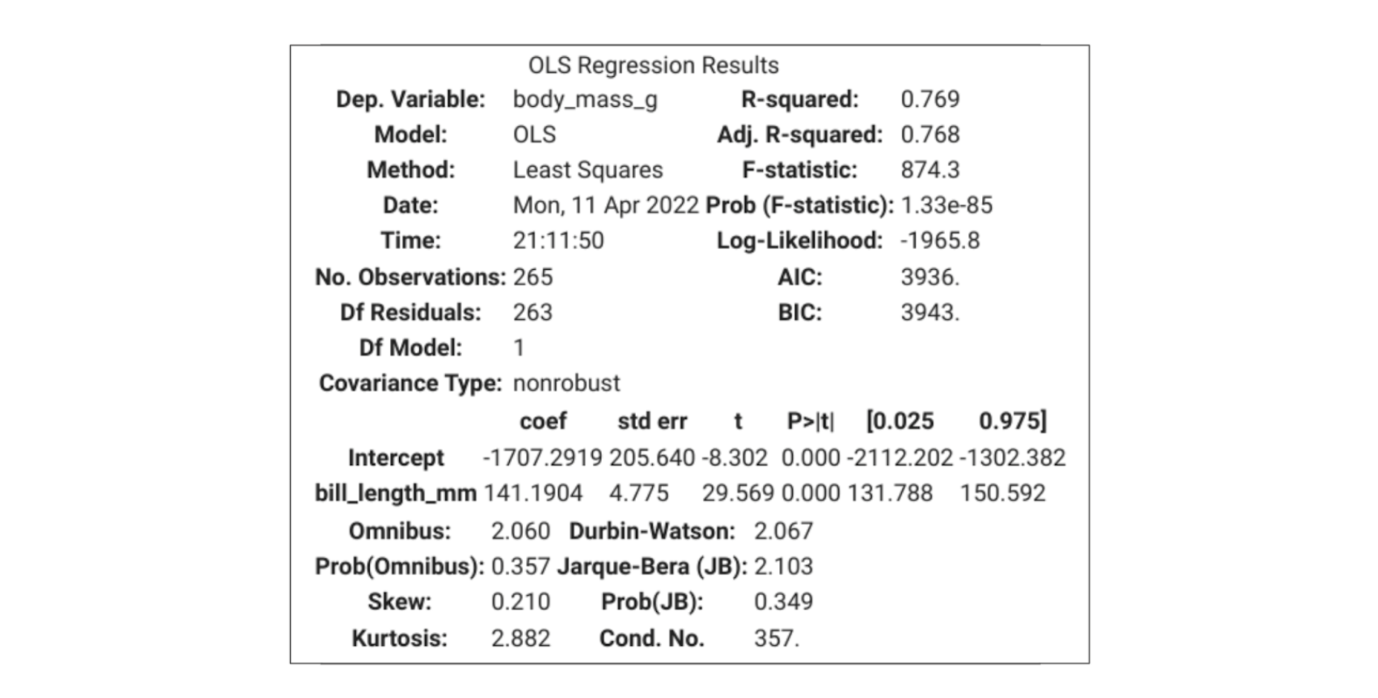
* **Confidence intervals** around beta coefficients
* **P-values** for the beta coefficients
* **Confidence band** around the regression line

You can refer to the glossary of terms to check any key terms and definitions, but we’ve provided the two key terms here:

* **Confidence interval:** a range of values that describes the uncertainty surrounding an estimate
* **P-value:** the probability of observing results as extreme as those observed when the null hypothesis is true

**Interpreting Uncertainty**

Let’s first revisit the summary of results from the linear regression model we created together in prior videos:



According to the simple linear regression model we built, β1^*β*1​^​ is 141.1904. So for every one-millimeter increase in the bill length of a penguin, we would expect a penguin to have about 141.1904 more grams in body mass. The estimate has a p-value of 0.000, which is less than 0.05, meaning that the coefficient is “statistically significant.” Additionally our estimate has a 95% confidence interval of 131.788 and 150.592. Let’s review these short sentences a bit more.

Previously you may have learned about p-values and confidence intervals within the context of hypothesis testing. Even though it may seem unintuitive, even in regression analysis we are testing hypotheses.

**P-values**

When running regression analysis, you want to know if X is really correlated with y or not. So we do a hypothesis test on the regression results. In regression analysis, for each beta coefficient, we are testing the following set of null and alternative hypotheses:

* H₀ (null hypothesis): β1=0*β*1​=0
* H₁ (alternative hypothesis): β1≠0*β*1​​=0

In our example, because the p-value is less than 0.05, we can reject the null hypothesis that β1*β*1​ is equal to 0, and state that the coefficient is statistically significant, which means that a difference in bill length of a penguin is truly correlated with a difference in body mass.

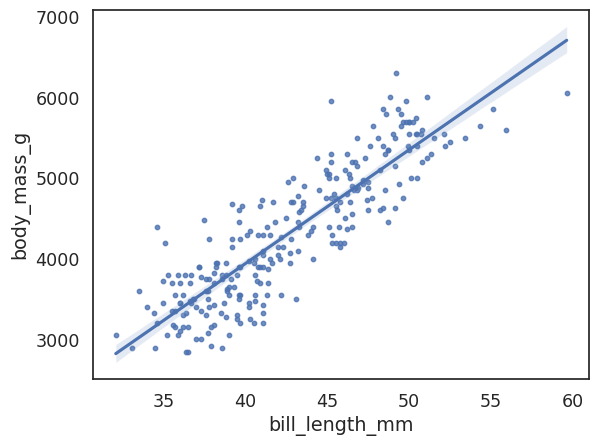
**Confidence Intervals**

Each beta coefficient also has a confidence interval associated with its estimate. A 95% interval means the interval itself has a 95% chance of containing the true parameter value of the coefficient. So there is 5% chance that our confidence interval [131.788, 150.592] does not contain the true value of β1*β*1​. More precisely, this means that if you were to repeat this experiment many times, 95% of the confidence intervals would contain the true value of β1*β*1​.

But, since there is uncertainty in both of the estimated beta coefficients, then the estimated y values also have uncertainty. This is where confidence bands become useful.

**Example Graph**

* **Confidence band:** the area surrounding the line that describes the uncertainty around the predicted outcome. You can think of the confidence band as representing the confidence interval surrounding each point estimate of y. Since there is uncertainty at every point in the line, we use the confidence band to summarize the confidence intervals across the regression model. The confidence band is always narrowest towards the mean of the sample and widest at the extremities.



**Key Takeaways**

* Regression analysis utilizes **estimation** techniques, so there is always uncertainty around the predictions.
* We can measure uncertainty using confidence intervals, p-values, and confidence bands.
* For every coefficient estimate, we are testing the hypothesis that the coefficient equals 0.

**Common evaluation metrics**

* R^2
* Mean Square Error (MSE)
* Mean Absolute Error (MAE)

**R^2 (The coefficient of determination)**

Measures the proportion of variation in the dependent variable, Y, explained by the independent variable(s), X

**Hold-out sample**

A random sample of observed data that is not used to fit the model

**Evaluation metrics for simple linear regression**

In this reading, we’ll provide a more comprehensive overview about evaluation metrics for simple linear regression. In a prior video we covered R², and mentioned a few other metrics, MAE and MSE. In this reading, we will review the metrics we’ve previously mentioned, and introduce a few more as well that you may encounter throughout your career as a data professional.

**Review of R², MSE, and MAE**

The main evaluation metric for linear regression is R², or the coefficient of determination.

**R²: The coefficient of determination**

**R²** measures the proportion of variation in the dependent variable, Y, explained by the independent variable(s), X.

* This is calculated by subtracting the sum of squared residuals divided by the total sum of squares from 1.

R2=1−Sum of squared residualsTotal sum of squares*R*2=1−Total sum of squaresSum of squared residuals​

R² ranges from 0 to 1. So, if a model has an R² of 0.85, that means that the X variables explain about 85% of the variation in the Y variable. Although R² is a highly interpretable and commonly used metric, you may also encounter mean squared error (MSE) and mean absolute error (MAE) when R² is insufficient in evaluating model performance.

**MSE: Mean squared error**

**MSE (mean squared error)** is the average of the squared difference between the predicted and actual values.

* Because of how MSE is calculated, MSE is very sensitive to large errors.

**MAE: Mean absolute error**

**MAE (mean absolute error)** is the average of the absolute difference between the predicted and actual values.

* If your data has outliers that you want to ignore, you can use MAE, as it is not sensitive to large errors.

**Other evaluation metrics**

Beyond the three metrics listed above, you may also encounter [**AIC (Akaike information criterion) and BIC (Bayesian information criterion)**](https://machinelearningmastery.com/probabilistic-model-selection-measures/).

Lastly, there is **adjusted R²**, which will be addressed in more detail in upcoming videos. It is a variation of R² that accounts for having multiple independent variables present in a linear regression model.

**Key takeaways**

* There are many evaluation metrics to choose from with regard to simple linear regression.
* The most common evaluation metric you’ll encounter is probably R². But, there are times when R² is insufficient or inappropriate to use.
* Based on your experiences and the particulars of a metric, you can use your best judgment to select an appropriate metric to evaluate a regression model.

# Report regression results

Recently, you have learned how to build a simple linear regression models and report the results of a regression model. You’ve also learned there are various ways to quantify uncertainty in regression results. From confidence bands to p-values, there are different ways to measure certainty about the results of a regression model. These metrics of certainty help us understand how to communicate our findings to inform business decisions and insights.

For this discussion prompt, consider the following:

* Why is it important to report uncertainty around the results of regression models?
* What is an example where reporting uncertainty might be particularly important?

Submit 3–5 sentences (100–200 words total). Then, visit the [discussion forum](https://www.coursera.org/learn/regression-analysis-simplify-complex-data-relationships/discussions) to read what other learners have written, and respond to at least one of them building on the original points made with your own thoughts or relevant examples.

**Response:**

Reporting uncertainty around the results of regression models is crucial because it provides context for the reliability and precision of the estimates. Uncertainty measures, such as confidence intervals and p-values, help stakeholders understand the range within which the true parameter values likely fall and the statistical significance of the findings. This transparency is essential for informed decision-making, as it prevents overconfidence in the results and highlights the potential for variability.

An example where reporting uncertainty is particularly important is in medical research, such as the development of a new drug. When estimating the effectiveness of the drug, it's vital to communicate the confidence intervals around the estimated effect size. This ensures that healthcare providers and patients understand the potential range of outcomes and the degree of certainty in the drug's efficacy, ultimately leading to more cautious and informed healthcare decisions.

**Test your knowledge: Evaluate a linear regression model**

**1.**

Question 1

What is the area surrounding a regression line, which describes the uncertainty around the predicted outcome at every value of X?

1 / 1 point

Ordinary least squares

R squared

Confidence interval

Confidence band

Correct

The confidence band is the area surrounding a regression line, which describes the uncertainty around the predicted outcome at every value of X. The confidence band reveals the confidence interval for each point on a regression line.

**2.**

Question 2

Fill in the blank: R squared measures the \_\_\_\_\_ in the dependent variable, Y, that is explained by the independent variable, X.

1 / 1 point

proportion of accuracy

coefficient of variation

proportion of variation

coefficient of accuracy

Correct

R squared measures the proportion of variation in the dependent variable, Y, that is explained by the independent variable, X. It is calculated by subtracting the sum of squared residuals (explained variance) divided by the total variance from 1.

**3.**

Question 3

Which linear regression evaluation metric is sensitive to large errors?

1 / 1 point

Mean squared error (MSE)

Mean absolute error (MAE)

The coefficient of determination

Adjusted R squared

Correct

Mean squared error (MSE) is sensitive to large errors. The MSE is the average of the squared difference between the predicted and actual values.

**Slope**

The amount that Y increases or decreases per one-unit increase of X

Y = intercept + slope\*X

**Causation**

A cause-and-effect relationship where one variable directly causes the other to change in a particular way

**Libraries**

Matplotlib, seaborn

**Programs**

* Tableau
* Powerpoint
* Google slides

**Correlation versus causation: Interpret regression results**

In previous videos, you learned that correlation is not causation. In this reading, you will continue to explore the differences between correlation and causation so that you will be prepared to report regression results responsibly, honestly, and effectively.

**What is correlation?**

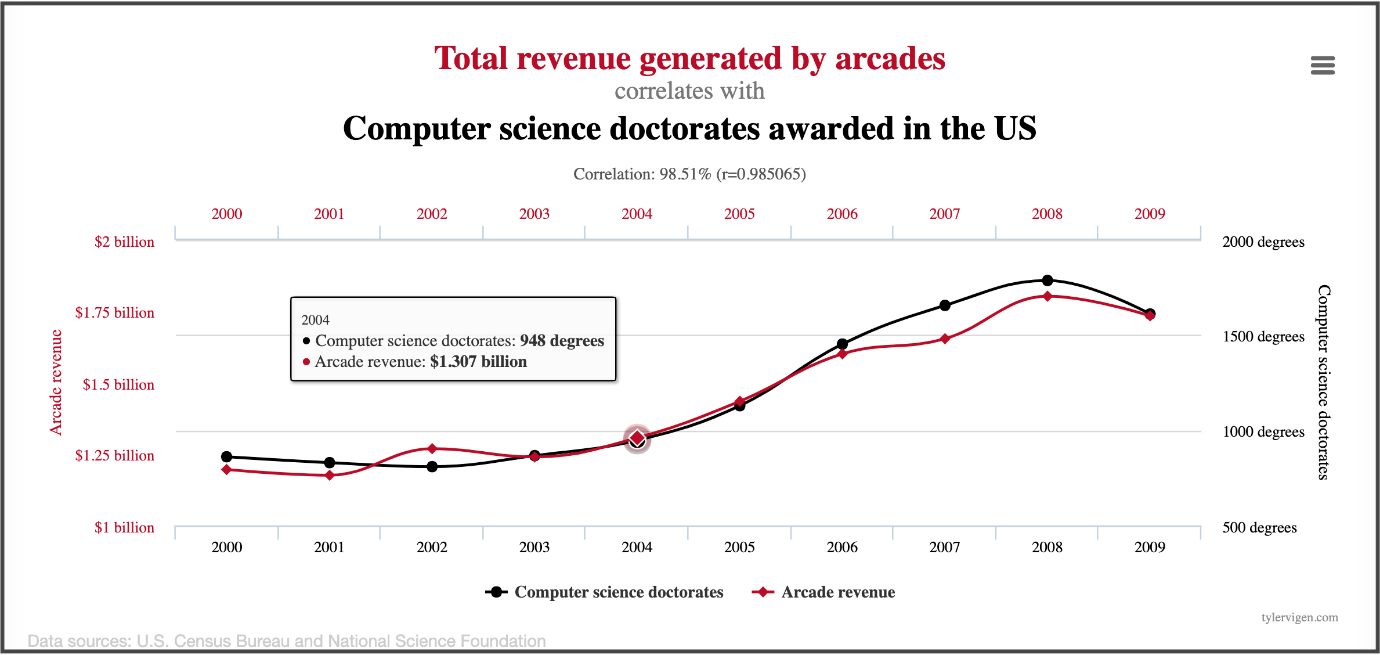
You might recall that there are two main kinds of correlation: positive and negative correlation.

* **Positive correlation** is a relationship between two variables that tend to increase or decrease together.
* **Negative correlation** is an inverse relationship between two variables, where when one variable increases, the other variable tends to decrease, and vice versa.

To generalize, **correlation** measures the way two variables tend to change together. There is a metric called the [Pearson correlation coefficient](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html) that ranges from -1 to 1 that can measure the relationship between two variables.

Note that correlation is just observational. Two variables can be correlated–they tend to change together without one variable causing the other variable to change. In fact, there is an entire website and book, [*Spurious Correlations*](https://www.tylervigen.com/spurious-correlations), devoted to documenting interesting and unexpected correlations between variables.

For example, here is a graph illustrating the correlation between total revenue generated by arcades and computer science doctorates awarded in the United States. Over time, computer science doctorates and arcade revenue increase at about the same rate. So, the graph definitely shows a correlation between the two variables, but it’s pretty hard to argue that one causes the other.



**It’s difficult to claim causation**

Previously you learned that **causation** describes a cause-and-effect relationship where one variable directly causes the other to change in a particular way. Although this is an intuitive definition, proving causation requires a lot of particular circumstances to be met.

To argue for causation between variables, in general, you must run a [**randomized controlled experiment**](https://www.urban.org/research/data-methods/data-analysis/quantitative-data-analysis/impact-analysis/experiments). The following are some key components of a proper randomized controlled experiment:

* You must control for every factor in the experiment.
* You must have a control group under certain conditions.
* You must have at least one treatment group under certain conditions.
* The difference(s) between the control and treatment groups must be observable and measurable.

Setting up a randomized controlled experiment is quite laborious and intensive. There are a number of requirements and factors not included in this reading, but there is a lot of information online and in academic research that you can explore on your own. Understanding the basics of causal claims allows you to responsibly report the results of your data analysis.

**Correlation leads to interesting insights**

When working as a data professional, you often do not have complete control of how the data is collected. You or your team might not be able to run a randomized controlled experiment. But, even if you cannot make causal claims, correlational research can still yield interesting results that have meaningful business implications.

**Scenario 1: Optimizing athletic performance**

Suppose a runner is training for a race. There are many ways to track health data—from built-in apps to paid-for apps. But, there are also so many factors that can contribute to the runner’s performance—how much water they drank, how sore their muscles are on a given day, the weather, how much sleep they got, what equipment they are using on race day, and the clothes they are wearing. It’s very hard to claim that any one factor would make or break their race time. But, over time, one might observe patterns in how sleep, soreness, water, clothing, and other factors tend to correlate with performance. This is why athletes can be so particular about brands of equipment, their diet, and pre-race day routines.

**Claims you can make (correlation)**

* When the runner drinks more water the day before a race, they tend to have more stamina.
* When the runner doesn’t run long distances the week before a race, they tend to feel better on race day.

**Claims you cannot make (causation)**

* Drinking more water the day before a race causes the runner to run faster.
* Not running long distances the week before a race causes the runner to run faster.

**Scenario 2: Improving food quality**

Perhaps you’re a new chef at a restaurant or you’re cooking for yourself or family. Every time you make a dish, there are a lot of variables: what pan was used, when the ingredients were purchased, if the ingredients are in season, and how hungry everyone was. Any one of these factors can change how “good” the dish is. But, this data is valuable regardless of causal claims. Over time, you can hone your cooking skills for this particular dish to ensure better food quality.

These are just two examples where gathering data to understand correlation between factors can drastically improve outcomes. The same principles can be applied on a larger scale, with big data, and in different industries, depending on the desired outcome you want to optimize.

**Claims you can make (correlation)**

* When I use fresher ingredients, the final dish tends to taste better.
* When I am very hungry, the final dish tends to taste better.

**Claims you cannot make (causation)**

* Using fresher ingredients makes the dish taste better.
* Being hungrier makes the dish taste better.

**Key takeaways**

* Claiming causation requires specific circumstances that are often not within your control.
* Correlation analyses are an incredibly useful tool for data professionals, and can provide interesting insights and actionable next steps.

# Test your knowledge: Interpret linear regression results

**1.**

Question 1

Which of the following are best practices when communicating linear regression results? Select all that apply.

0.75 / 1 point

Use data visualizations to present the results.

Correct

When communicating linear regression results, best practices include the following: Provide measures of uncertainty around estimated results, make the findings quickly understood without technical terms, and use data visualizations to present the results.

Always extrapolate to a larger or different group any data insights that apply only to a specific, smaller population.

Provide measures of uncertainty around estimated results.

Correct

When communicating linear regression results, best practices include the following: Provide measures of uncertainty around estimated results, make the findings quickly understood without technical terms, and use data visualizations to present the results.

Make the findings quickly understood without technical terms.

You didn’t select all the correct answers

**2.**

Question 2

Which of the following statements accurately describe coefficients and p-values for regression model interpretation? Select all that apply.

1 / 1 point

P-values demonstrate whether coefficients are statistically significant.

Correct

Coefficients determine how changes in the independent variables are associated with changes in the dependent variable. P-values demonstrate whether coefficients are statistically significant.

P-values determine how changes in the independent variables are associated with changes in the dependent variable.

Coefficients determine how changes in the independent variables are associated with changes in the dependent variable.

Correct

Coefficients determine how changes in the independent variables are associated with changes in the dependent variable. P-values demonstrate whether coefficients are statistically significant.

Coefficients demonstrate whether P-values are statistically significant.

**Review**

* PACE in simple linear regression
* Ordinary least squares
* Linearity assumptions
* Normality assumption
* Independent observations assumption
* Homoscedasticity assumption

**Glossary terms from module 2**

**Terms and definitions from Course 5, Module 2**

**Adjusted R**2: A variation of R2 that accounts for having multiple independent variables present in a linear regression model

**Best fit line**: The line that fits the data best by minimizing some loss function or error

**Causation**: Describes a cause-and-effect relationship where one variable directly causes the other to change in a particular way

**Confidence band**: The area surrounding a line that describes the uncertainty around the predicted outcome at every value of X

**Confidence interval**: A range of values that describes the uncertainty surrounding an estimate

**Correlation**: Measures the way two variables tend to change together

**Dependent variable (Y)**: The variable a given model estimates

**Errors**: The natural noise assumed to be in a regression model

**Hold-out sample**: A random sample of observed data that is not used to fit the model

**Homoscedasticity assumption**: An assumption of simple linear regression stating that the variation of the residuals (errors) is constant or similar across the model

**Independent observation assumption**: An assumption of simple linear regression stating that each observation in the dataset is independent

**Independent variable (X)**: A variable whose trends are associated with the dependent variable

**Linear regression**: A technique that estimates the linear relationship between a continuous dependent variable and one or more independent variables

**Linearity assumption**: An assumption of simple linear regression stating that each predictor variable (Xi) is linearly related to the outcome variable (Y)

**MAE (Mean Absolute Error)**: The average of the absolute difference between the predicted and actual values

**Model assumptions**: Statements about the data that must be true in order to justify the use of a particular modeling technique

**MSE (Mean Squared Error)**: The average of the squared difference between the predicted and actual values

**Negative correlation**: An inverse relationship between two variables, where when one variable increases, the other variable tends to decrease, and vice versa.

**Normality assumption**: An assumption of simple linear regression stating that the residual values or errors are normally distributed

**Ordinary least squares (OLS)**: A method that minimizes the sum of squared residuals to estimate parameters in a linear regression model

**Outcome variable (Y)**: (Refer to **dependent variable**)

**P-value**: The probability of observing results as extreme as those observed when the null hypothesis is true

**Positive correlation**: A relationship between two variables that tend to increase or decrease together.

**Predicted values**: The estimated Y values for each X calculated by a model

**R**2 **(The Coefficient of Determination)**: Measures the proportion of variation in the dependent variable, Y, explained by the independent variable(s), X

**Residual**: The difference between observed or actual values and the predicted values of the regression line

**Scatterplot matrix**: A series of scatter plots that demonstrate the relationships between pairs of variables

**Simple linear regression**: A technique that estimates the linear relationship between one independent variable, X, and one continuous dependent variable, Y

**Slope**: The amount that y increases or decreases per one-unit increase of x

**Sum of squared residuals (SSR)**: The sum of the squared difference between each observed value and its associated predicted value

**Terms and definitions from the previous module**

**A**

**Absolute values**: (Refer to **observed values**)

**C**

**Causation**: A cause-and-effect relationship where one variable directly causes the other to change in a particular way

**D**

**Dependent variable (Y)**: The variable a given model estimates

**E**

**Explanatory variable**: (Refer to **independent variable**)

**I**

**Independent variable (X)**: A variable whose trends are associated with the dependent variable

**Intercept (constant 𝐵**0**)**: The y value of the point on the regression line where it intersects with the y-axis

**L**

**Line**: A collection of an infinite number of points extending in two opposite directions

**Linear regression**: A technique that estimates the linear relationship between a continuous dependent variable and one or more independent variables

**Link function**: A nonlinear function that connects or links the dependent variable to the independent variables mathematically

**Logistic regression**: A technique that models a categorical dependent variable based on one or more independent variables

**Loss function**: A function that measures the distance between the observed values and the model’s estimated values

**M**

**Model assumptions**: Statements about the data that must be true to justify the use of a particular modeling technique

**N**

**Negative correlation**: An inverse relationship between two variables, where when one variable increases, the other variable tends to decrease, and vice versa

**O**

**Observed values:** The existing sample of data, where each data point in the sample is represented by an observed value of the dependent variable and an observed value of the independent variable

**Outcome variable**: (Refer to **dependent variable**)

**P**

**Positive correlation**: A relationship between two variables that tend to increase or decrease together

**Predictor variable**: (Refer to **independent variable**)

**R**

**Regression analysis**: A group of statistical techniques that use existing data to estimate the relationships between a single dependent variable and one or more independent variables

**Regression coefficient**: The estimated betas in a regression model

**Regression models**: (Refer to **regression analysis**)

**Response variable:** (Refer to **dependent variable)**

**S**

**Slope**: The amount that y increases or decreases per one-unit increase of x

**Module 2 challenge**

**1.**

Question 1

What is the difference between observed or actual values and the predicted values of a regression line?

Residual

**2.**

Question 2

In linear regression, what mathematical technique is used to calculate beta zero hat and beta one hat?

Ordinary least squares

**3.**

Question 3

A data professional testing for linear regression assumptions plots their dependent variable against their independent variable and notices that the graph appears as an upward curve. Which model assumption does this invalidate?

Linearity

**4.**

Question 4

FIll in the blank: A scatterplot \_\_\_\_\_ is a series of scatterplots that show the relationships between pairs of variables.

matrix

**5.**

Question 5

A data analytics professional working for a storage facility checks model assumptions while determining optimal storage space sizes. They notice that the model's residuals appear in a cone-shaped pattern when plotted against the independent variable. Which model assumption does this invalidate?

Homoscedasticity

**6.**

Question 6

Fill in the blank: A \_\_\_\_\_ is the area surrounding a line that describes the uncertainty around the predicted outcome at every value of X.

confidence band

**7.**

Question 7

A data professional determines how much of the variation in the X variable explains the variation in the Y variable. Which model evaluation metric enables this determination?

R squared

**8.**

Question 8 (not FINAL)

Which of the following statements accurately describe running a randomized, controlled experiment? Select all that apply.

It is a study design that randomly assigns participants into groups.

Correct

It cannot have a control group.

It is typically used when arguing for causation between variables.

To be successful, data professionals must control for every factor in the experiment.

Correct

You didn’t select all the correct answers

**Module 3:**

**What you’ll learn**

* Variable selection
* Regularization

**Multiple linear regression or multiple regression**

A technique that estimates the relationship between one continuous dependent variable and two or more independent variables.

**Full multiple regression equation**

Y = beta\_0 + beta\_1 \* x\_1 + beta\_2 \* x\_2 + …….+ beta\_n \* x\_n

**What you’ll learn**

* One hot encoding
* Interaction terms

**Multiple linear regression scenarios**

**Goals of Reading**

Now that you have learned what multiple linear regression is, in this reading, you will explore three scenarios in which multiple regression models can help a company or organization understand a business problem. The goal of the reading is to understand the versatility of multiple linear regression, and to get you thinking about various applications of this powerful and flexible regression technique.

**Scenario 1: Selling graphic design services**

Let’s say that you’re a data professional working at a company that sells graphic design services. The company you work for might be interested in understanding the factors related to customer satisfaction and retention. There are many ways you can measure this, and you can use any of the following factors to develop a promising multiple linear regression model.

**Potential dependent variables (Y)**

* Customer satisfaction
* Number of returning customers
* [Net Promoter Score](https://www.qualtrics.com/experience-management/customer/net-promoter-score/)
* Satisfaction with customer service

**Potential independent variables (X)**

* Cost of services
* Customer service response time
* Adding new graphic design packages
* Changing page layout

**Scenario 2: Running a restaurant**

Imagine that you are working at a restaurant, and you want to determine how to improve the success of your business. Like any other client-facing business, you want to keep your costs down, your revenue high, and your customers happy. Similar to the prior example, there are many ways to measure the restaurant’s success. There are also a number of variables that could be correlated with the chosen metric of success.

**Potential dependent variables (Y)**

* Total revenue
* Number of reviews online
* Number of five-star reviews online
* Number of reservations per week

**Potential independent variables (X)**

* Spending on advertising/marketing
* Operational costs
* Size of menu
* Foot traffic
* Cancellation of reservations
* Business partnerships (ex: delivery apps, farmers’ markets, community organizations)

**Scenario 3: Agricultural production**

Suppose you are working in agricultural production, perhaps on a farm or a ranch. Even though this is a very different environment from a restaurant or online service, multiple regression can still be helpful. For example, let’s say that you are trying to predict crop yield, revenue for the season, or amount of crops sold. From the weather to soil conditions to labor and resource usage, there are many factors that could contribute to a good year or a bad year for a farm or any kind of agricultural production. Multiple regression can be used to help better plan and predict for worse years.

**Potential dependent variables (Y)**

* Crop yield
* Revenue
* Crops sold

**Potential independent variables (X)**

* Weather (rainfall, temperature)
* Nutrients in soil
* Historic crop yield
* Cost of fertilizer
* Cost of fuel, water, or energy used to maintain crops
* Cost of labor
* Partnerships with local restaurants or, grocery stores

**Key Takeaways**

* Multiple regression is a versatile and effective way to understand and describe more complex relationships between variables.
* Multiple regression can be used in a variety of industries and contexts.

**Resources for more information**

* [“Multiple Regression: Definition, Uses, and 5 Examples.” *Indeed Editorial Team*](https://www.indeed.com/career-advice/career-development/multiple-regression).
* [“Multivariate Regression Analysis | STATA Data Analysis Examples.” *UCLA: Statistical Consulting Group.*](https://stats.oarc.ucla.edu/stata/dae/multivariate-regression-analysis/)

**Test your knowledge: Understand multiple linear regression**

1.

Question 1

Fill in the blank: \_\_\_\_\_ is a technique that estimates the linear relationship between one continuous dependent variable and two or more independent variables.

Multiple linear regression

Correct

Multiple linear regression is a technique that estimates the linear relationship between one continuous dependent variable and two or more independent variables. The multiple regression technique can yield highly interpretable and communicable results.

2.

Question 2

What concept refers to how two independent variables together affect the *y* dependent variable?

Interaction terms

**One hot encoding**

A data transformation technique that turns one categorical variable into several binary variables.

**Question**

What data transformation technique turns one categorical variable into several binary variables?

One hot encoding

Correct

Feedback: One hot encoding is a data transformation technique that turns one categorical variable into several binary variables. Data professionals use one hot encoding when there is a categorical independent variable and they need to represent the category as numbers.

**Independent observation assumption**

Each observation in the dataset is independent

**Normality assumption**

The residuals are normally distributed

**Homoscedasticity assumption**

The variation of the residuals (errors) is constant or similar across the model

**No multicollinearity assumption**

No two independent variables (Xi and Xj) can be highly correlated with each other.

**Question**

Fill in the blank: The \_\_\_\_\_ states that no two independent variables (X−i*X*−*i* and X−j*X*−*j*) can be highly correlated with each other.

no multicollinearity assumption

Correct

Feedback: The no multicollinearity assumption states that no two independent variables (X−i*X*−*i* and X−j*X*−*j*) can be highly correlated with each other.  This means that X−i*X*−*i* and X−j*X*−*j* cannot be linearly related to each other.

**Variance Inflation Factors (VIF)**

Quantifies how correlated each independent variable is will all of the other independent variables.

**How to handle data with multicollinearity**

* Drop one or more variables that have high multicollinearity
* Create new variables using existing data

**Multiple linear regression assumptions and multicollinearity**

In prior videos, you have learned about linear regression assumptions. In this reading, you will build off that knowledge base to extend your understanding of multiple linear regression assumptions. This reading will help you review assumptions that apply to both simple linear regression and multiple linear regression, and will then focus more heavily on the concept of multicollinearity.

**Multiple linear regression assumptions**

Recall that simple linear regression has four main assumptions that provide validity to the results derived from the analysis. To this list of four assumptions, we add the no multicollinearity assumption when working with multiple linear regression.

1. **Linearity:** Each predictor variable (Xi*Xi*​) is linearly related to the outcome variable (Y).
2. **(Multivariate) normality:** The errors are normally distributed.**\***
3. **Independent observations:** Each observation in the dataset is independent.
4. **Homoscedasticity:** The variation of the errors is constant or similar across the model.**\***
5. **No multicollinearity:** No two independent variables (Xi*Xi*​ and Xj*Xj*​) can be highly correlated with each other.

**\*Note on errors and residuals**

As noted earlier, “residuals” and “errors” are sometimes used interchangeably, but there is a difference. We use residuals to estimate errors when we are checking the normality and homoscedasticity assumptions of linear regression.

* **Residuals** are the difference between the predicted and observed values. You can calculate residuals after you build a regression model by subtracting the predicted values from the observed values.
* **Errors** are the natural noise assumed to be in the model.

**Extending prior assumptions**

Much of what you learned about the first four assumptions with regard to simple linear regression can be directly applied to multiple linear regression. The code might be slightly different or longer, but the rationale is the same.

**Linearity**

* With multiple linear regression, you need to consider whether each *x* variable has a linear relationship with the *y* variable.
* You can make multiple scatterplots instead of just one, using seaborn’s [pairplot()](https://seaborn.pydata.org/generated/seaborn.pairplot.html) function, or the [scatterplot()](https://seaborn.pydata.org/generated/seaborn.scatterplot.html) function multiple times. Other libraries with plotting capabilities will have similar functions.

**Independent observations**

* The independent observations assumption is still primarily focused on data collection.
* You can check the validity of the assumption in the same way you would with simple linear regression.

**(Multivariate) Normality**

* Just as with simple linear regression, you can construct the model, and then create a Q-Q plot of the residuals.
* If you observe a straight diagonal line on the Q-Q plot, then you can proceed in your analysis. You can also plot a histogram of the residuals and check if you observe a normal distribution that way.
* **Note:** It’s a common misunderstanding that the independent and/or dependent variables must be normally distributed when performing linear regression. This is not the case. Only the model’s residuals are assumed to be normal.

**Homoscedasticity**

* As with simple linear regression, for multiple linear regression, just create a plot of the residuals vs. fitted values.
* If the data points seem to be scattered randomly across the line where residuals equal 0, then you can proceed.

**How to check the no multicollinearity assumption**

The no multicollinearity assumption is unique to multiple linear regression as it focuses on potential relationships between different independent (X) variables. When assessing the no multicollinearity assumption, you’re interested in identifying any linear relationships between the independent (X) variables. X variables that are linearly related could muddle the interpretation of the model’s results. If there are X variables that are linearly related, it is usually best to remove some independent variables from the model.

Note, however, that the assumption of no multicollinearity is most important when you are using your regression model to make inferences about your data, because the inclusion of collinear data increases the standard errors of the model’s beta parameter estimates. But there may be times when the primary purpose of your model is to make predictions and the need for accurate predictions outweighs the need to make inferences about your data. In this case, including the collinear independent variables may be justified because it’s possible that their inclusion would result in better predictions.

There are a few ways to check the no multicollinearity assumption. This reading will cover two of them. One is purely visual, and the other is numerical in nature. Both can be done prior to building the linear regression model.

**Scatterplots or Scatterplot Matrix**

A visual way to identify multicollinearity between independent (X) variables is using scatterplots or scatterplot matrices. The process is the same as when you checked the linearity assumption, except now you’re just focusing on the X variables, not the relationship between the X variables and the Y variable. If you’re using the seaborn library, you can use the **pairplot** function, or the **scatterplot** function multiple times.

**Variance Inflation Factors (VIF)**

Calculating the variance inflation factor, or VIF, for each independent (X) variable is a way to quantify how much the variance of each variable is “inflated” due to correlation with other X variables. You can read more about VIFs on the [Pennsylvania State University’s Eberly College of Science](https://online.stat.psu.edu/stat462/node/180/) website or on the website for Vilnius University’s e-book on [*Practical Econometrics and Data Science*](http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/4-5-Multiple-collinearity.html). The details of calculating VIF are beyond the scope of this course, but it's helpful to know that VIFi*VIFi*​​ represents the amount that the standard error of coefficient βi*βi*​ increases relative to a situation in which all of the predictor variables are uncorrelated.

To calculate the VIF for each predictor variable, you can use the [variance\_inflation\_factor()](https://www.statsmodels.org/dev/generated/statsmodels.stats.outliers_influence.variance_inflation_factor.html) function from the **statsmodels** package. Here is an example of how you might obtain VIFs for your predictor variables.

1

2

3

4

5

6

from statsmodels.stats.outliers\_influence import variance\_inflation\_factor

X = df[['col\_1', 'col\_2', 'col\_3']]

vif = [variance\_inflation\_factor(X.values, i) for i in range(X.shape[1])]

vif = zip(X, vif)

print(list(vif))

The smallest value a VIF can take on is 1, which would indicate 0 correlation between the X variable in question and the other predictor variables in the model. A high VIF, such as 5 and above, according to the [statsmodels documentation](https://www.statsmodels.org/dev/generated/statsmodels.stats.outliers_influence.variance_inflation_factor.html#:~:text=The%20variance%20inflation%20factor%20is,of%20the%20design%20matrix%2C%20exog.), can indicate the presence of multicollinearity.

**What to do if there is multicollinearity in your model**

**Variable Selection**

The easiest way to handle multicollinearity is simply to only use a subset of independent variables in your model. For example, if your multiple linear regression model is something like this:

y=β0+β1X1+β2X2+β3X3*y*=*β*0​+*β*1​*X*1​+*β*2​*X*2​+*β*3​*X*3​

But if X1*X*1​ and X3*X*3​ are highly correlated, then you can choose to include only X1*X*1​ or X3*X*3​ in your final model, but not both.

There are a few specific statistical techniques you can use to select variables strategically. You’ll learn about these more in future videos:

* Forward selection
* Backward elimination

**Advanced Techniques**

In addition to the techniques listed above that will be covered in-depth in this course, there are more advanced techniques that you may come across in your career as a data professional, such as:

* Ridge regression
* Lasso regression
* Principal component analysis (PCA)

These techniques can result in more accurate and predictive models, but can complicate the interpretation of regression results.

**Key Takeaways**

* Many of the assumptions of simple linear regression extend readily to multiple linear regression.
* You can use scatterplots and variance inflation factors to check for multicollinearity in a regression model.
* There are different techniques for variable selection to remove multicollinearity in a model.

**Identify: Multiple regression assumptions**

**X1 and Y are not correlated**

Linearity

**The errors of X2 are autocorrelated.**

Independent observations

**The errors of X2 are not normally distributed**

Multivariate normality

**As X2 increases, the variability of the errors increases.**

Homoscedasticity

**X1 and X2 are highly correlated.**

No multicollinearity

**X1 and X3 are highly correlated.**

No multicollinearity

**X3 and Y are correlated.**

Does not violate any assumption

**The errors of X3 are normally distributed.**

Does not violate any assumption

**Test your knowledge: Model assumptions revisited**

**1.**

Question 1

Which of the following statements is true? Select all that apply.

0.5 / 1 point

One hot encoding allows data professionals to turn one categorical variable into several binary variables.

One hot encoding is a data transformation technique.

Correct

One hot encoding is a data transformation technique that allows data professionals to turn one categorical variable into several binary variables.

One hot encoding is for ordinal variables.

One hot encoding allows data professionals to turn several categorical variables into one binary variable.

This should not be selected

One hot encoding is a data transformation technique that allows data professionals to turn one categorical variable into several binary variables.

**2.**

Question 2

What is the definition of the no multicollinearity assumption?

1 / 1 point

No two independent variables can be highly correlated with each other.

No observation in the dataset can be independent.

No predictor variable can be linearly related to the outcome variable.

Variation of the residual must be constant or similar across the model.

Correct

Multicollinearity states that no two independent variables can be highly correlated with each other. This means that Xi and Xj cannot be linearly related.

**3.**

Question 3

In what ways might a data professional handle data with multicollinearity? Select all that apply.

1 / 1 point

Square the variables that have high multicollinearity.

Turn one categorical variable into several binary variables.

Drop one or more variables that have high multicollinearity.

Correct

 A data professional might handle data with multicollinearity by dropping one or more variables that have high multicollinearity. They might also create new variables using existing data.

Create new variables using existing data.

Correct

 A data professional might handle data with multicollinearity by dropping one or more variables that have high multicollinearity. They might also create new variables using existing data.

**Interaction term**

A term that represents how the relationship between two independent variables is associated with changes in the mean of the dependent variable.

**Question**

Fill in the blank: An interaction term represents the relationship between two independent variables and the change in the mean of the \_\_\_\_\_ variable.

dependant

Correct

An interaction term represents the relationship between two independent variables and the change in the mean of the dependent variable. Typically, it is the product of the two independent variables.

**Annotated follow-along resource: Interpret multiple regression results with Python**

Launch lab

Instructions

All of the instructional videos with onscreen coding demonstrations have a corresponding follow-along guide that is available to you. The follow-along guide is an annotated Jupyter notebook organized to match the content from each module. It contains the same code shown in the videos for that module. This guide is provided for your reference; you do not need to add any text or run the code yourself. If you would like to run the code, you will need to run each cell sequentially for the code to function as intended.

In addition to content that is identical to what is covered in the videos, you’ll often find additional information throughout the guide to explain the purpose of each concept covered, why the code is written in a certain way, and tips for running the code.

The landing page for each follow-along notebook also provides information about data sources (when relevant) and tips on how to access and use these guides.

**Data dictionary**

This notebook uses one of seaborn's built-in datasets called "Palmer's penguins," which contains information about penguins. The data was collected from Palmer Station in Antarctica.

The dataset contains 344 rows, each representing data collected from a unique penguin.

There are seven columns:

| **Column name** | **Type** | **Description** |
| --- | --- | --- |
| species | str | The species of penguin |
| island | str | The name of the island on which the penguin was found |
| bill\_length\_mm | float64 | The length of the penguin's bill, in millimeters |
| bill\_depth\_mm | float64 | The depth of the penguin's bill, in millimeters |
| flipper\_length\_mm | float64 | The length of the penguin's flipper, in millimeters |
| body\_mass\_g | float64 | The weight of the penguin, in grams |
| sex | str | The sex of the penguin |

More information about this dataset can be found at the [Palmer's Penguins GitHub page](https://github.com/allisonhorst/palmerpenguins) and the [GitHub page for seaborn's datasets](https://github.com/mwaskom/seaborn-data).

**Set up a split screen**

While watching the videos that follow this notebook, you may find it helpful to track the instructor’s progress by following along in your own Jupyter Notebook. It can be helpful to review the code notebook alongside the video, especially if you’re new to coding in Python.

To do so:

1. Open the video in one browser window.
2. Then, open the annotated follow-along guide in a separate window.
3. Arrange your screen so that the video and the follow-along guide are both visible.

A screenshot of a computer

Description automatically generated

**Tips for working with the follow-along guide**

Follow these suggestions to enhance your learning experience:

* Reference the [Jupyter notebooks reading](https://www.coursera.org/learn/get-started-with-python/supplement/2poER/create-upload-and-edit-jupyter-notebooks) before starting if you need more information on working with the Jupyter notebooks.
* Go to the section of the follow-along guide for the current module's content. The follow-along guide has different sections for each video included in the module's content. The in-video message will direct you to the relevant section in the guide for the specific video you are viewing.
* Follow along in the notebook as the instructor discusses the code.

**Test your knowledge: Model interpretation**

**1.**

Question 1

Fill in the blank: An interaction term represents how the relationship between two independent variables is associated with the changes in the \_\_\_\_\_ of the dependent variable.

1 / 1 point

rate of change

multicollinearity

category

mean

Correct

An interaction term represents how the relationship between two independent variables is associated with the changes in the mean of the dependent variable. Typically, data professionals represent an interaction term as the product of the two independent variables in question.

**2.**

Question 2

Which of the following relevant statistics can be found by using statsmodel’s OLS function? Select all that apply.

1 / 1 point

Variance inflation factors

Standard errors

Correct

Coefficients, standard errors, p-values, and t-statistics can be found by using statsmodel’s OLS function.

P-values

Correct

Coefficients, standard errors, p-values, and t-statistics can be found by using statsmodel’s OLS function.

Coefficients

Correct

Coefficients, standard errors, p-values, and t-statistics can be found by using statsmodel’s OLS function.

**Overfitting**

When a model fits the observed or training data too specifically, and is unable to generate suitable estimates for the general population

**Adjusted R^2**

A variation of the R^2 regression evaluation metric that penalizes unnecessary explanatory variables.

**Adjusted R^2 vs. R^2**

* Adjusted R^2 is used to compare models of varying complexity

Determine if you should add another variable or not

* R^2 is more easily interpretable

Determine how much variation in the dependent variable is explained by the model

**Underfitting and overfitting**

As you have been learning, a multiple regression model is built using sample data from the population of interest with the goal of applying the model to unseen data from the population and getting reliable results. **Underfitting** and **overfitting** are two obstacles that the multiple regression model must mitigate so it can be applicable. In this reading, you will gain a general understanding of underfitting and get a closer look at overfitting.

**The two ways a model can be unreliable**

**Underfitting**

In the case of underfitting, a multiple regression model fails to capture the underlying pattern in the outcome variable. An underfitting model has a low R-squared value.

A model can underfit the data for a variety of reasons. The independent variables in the model might not have a strong relationship with the outcome variable. In this situation, different or additional predictors are needed. It could be the case that the sample dataset is too small, and this prevents the model from being able to learn the relationship between the predictors and the outcome. Using more sample data to build the model might reduce the problem of underfitting.

Consider the example of a multiple regression model that predicts the resale price of a pre-owned car. This model has two predictors: the color of the car and the year it was manufactured. The model’s R-squared value is quite low. This indicates that the model is underfitting because the current predictors do not have a strong relationship with the car’s resale price. There are likely other important predictors missing from the multiple regression model, like the mileage on the car or the make of the car.

There are additional reasons that a multiple regression model might underfit the data, and the methods used to reduce this obstacle depend on the specific context. Because an underfitting multiple regression model is not able to capture the relationship between predictors and outcome in the sample data, this model will also not be able to produce reliable results when it is used on unseen data from the population.

**The difference between training data and test data**

Before you learn more about overfitting, it is important to cover a step data scientists take before building a multiple regression model. They divide the sample data into two categories called **training data** and **test data**. Training data is used to build the model, and test data is used to evaluate the model’s performance after it has been built. Splitting the sample data in this way is also called **holdout sampling**, with the holdout sample being the test data. Holdout sampling allows data scientists to evaluate how a model performs on data it has not experienced yet.

The holdout sample might also be called the **validation data**. Regardless, the general idea remains the same: this is the data that is used to evaluate the model.

Data scientists obtain the training and test data by randomly splitting the sample dataset so that each record exclusively belongs to one of the two categories. This way, some records are used as the training data and other records are used as the test data.

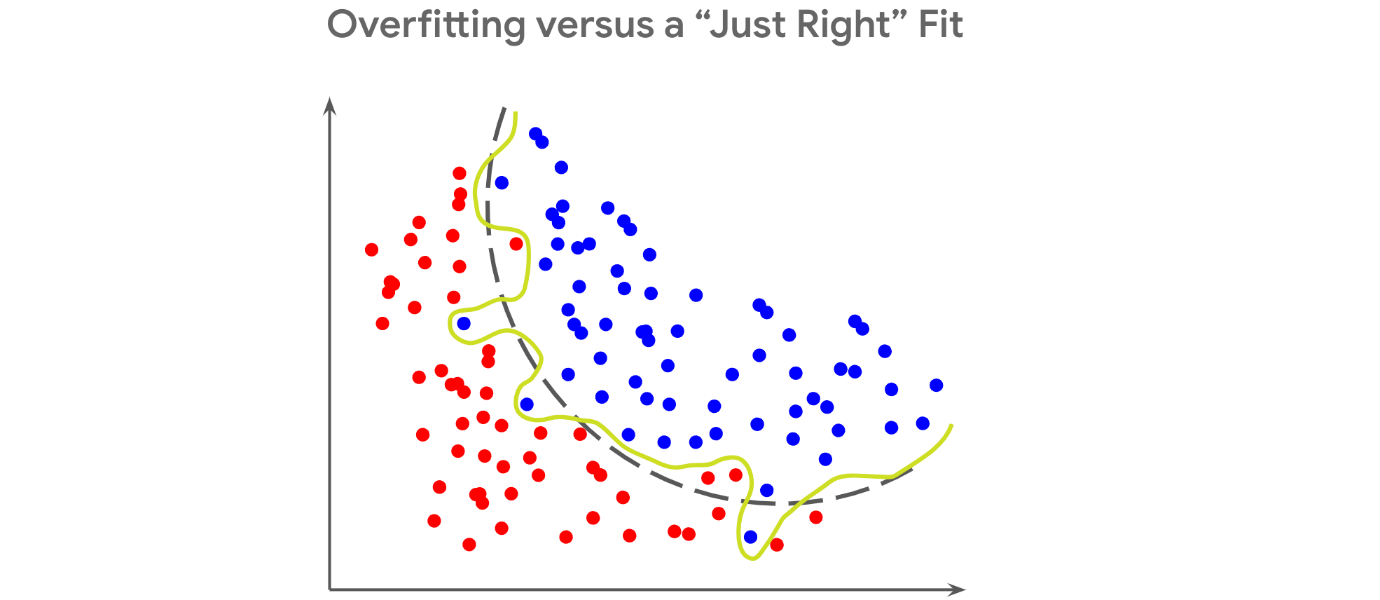
**Overfitting**

Underfitting causes a multiple regression model to perform poorly on the training data, which indicates that the model performance on test data will also be substandard. In contrast, overfitting causes a model to perform well on training data, but its performance is considerably worse when evaluated using the unseen test data. That’s why data scientists compare model performance on training data versus test data to identify overfitting.

Why is there a discrepancy between an overfitting model’s performance on training data versus test data?

An overfitting model fits the observed or training data too specifically, making the model unable to generate suitable estimates for the general population. This multiple regression model has captured the **signal** (i.e. the relationship between the predictors and the outcome variable) *and* the **noise** (i.e. the randomness in the dataset that is not part of that relationship). You cannot use an overfitting model to draw conclusions for the population because this model **only** applies to the data used to build it.

In the plot below, the dashed black line represents an optimal multiple regression model that performs well in distinguishing between the red and blue dots without overfitting the data. In contrast, the squiggly yellow line represents a model that overfits the data. Although this line might even do a slightly better job of separating the blue dots from the red ones, it is too specific to this data and will not perform well on another sample from the same population. In contrast, the black line will be continuously reliable in distinguishing between the two colors.



**Why does overfitting result in a higher R-squared value?**

Earlier you learned that R-squared is a goodness of fit measure because it tells you the proportion of variance in the outcome variable that is captured by the independent variables in the multiple regression model. However, as you add more independent variables to a model, the associated R-squared value will increase regardless of whether or not those predictors have a strong relationship with the outcome variable.

In the example of the multiple regression model that predicts car resale price, you could continue to add more independent variables to the model, such as the number of letters in the name of the person selling the car and the favorite food of the person who bought the car (if you had this data, of course). These predictors are very unlikely to have a relationship with the resale price of the car, but if you add them to your multiple regression model, the R-squared value would still increase. Although this could lead you to think that the model with more predictors is performing better, the inflated R-squared value is a false sign of improvement.

Generally, R-squared will continue to increase with more predictors because the model will become overly specific to the data it was built on even if the predictors do not have a strong relationship with the outcome variable. This is why a high R-squared value is not enough by itself to indicate that the model will perform well and might instead be a sign of overfitting.

**When to use adjusted R-squared instead**

Along with the R-squared value, a multiple regression model also has an associated **adjusted R-squared value**. The adjusted R-squared penalizes the addition of more independent variables to the multiple regression model. Additionally, the adjusted R-squared only captures the proportion of variation explained by the independent variables that show a significant relationship with the outcome variable. These differences prevent the adjusted R-squared value from becoming inflated like the R-squared value.

When comparing between multiple regression models with varying numbers of predictors, you might find that models with more predictors have a higher R-squared value. This could be a result of overfitting. To avoid selecting an overfitting model with an inflated R-squared, use the adjusted R-squared metric to select the optimal model.

**Bias versus variance**

A model that underfits the sample data is described as having a high **bias** whereas a model that does not perform well on new data is described as having high **variance**. In data science, there is a phenomenon known as the **bias versus variance tradeoff**. This tradeoff is a dilemma that data scientists face when building any machine learning model because an ideal model should have low bias and low variance. This is another way of saying that it should neither underfit nor overfit. However, as you try to lower bias, variance inevitably increases and vice versa.

This is why you can never fully resolve the problems of underfitting and overfitting. Instead, focus on reducing these problems in your multiple regression model as much as possible.

**Key takeaways**

Both underfitting and overfitting are obstacles to building a reliable multiple regression model. Although underfitting can be identified by model performance on training data, you must evaluate both training and test performance to identify overfitting. Because overfitting will result in an inflated R-squared value, use the adjusted R-squared value when comparing among multiple regression models with varying numbers of predictors. Although you cannot fully eliminate underfitting and overfitting from the model because of the bias versus variance tradeoff, you can significantly reduce these problems after they have been identified.

**Resources for more information**

* [A detailed description of underfitting and how to mitigate it](https://www.ibm.com/cloud/learn/underfitting)
* [Scikit-learn library documentation for the train\_test\_split function](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html)
* [A blog discussing multiple, adjusted, and predicted R-squared values](https://blog.minitab.com/en/adventures-in-statistics-2/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables)

**Forward selection**

A stepwise variable selection process that begins with the null model, with 0 independent variables, considers all possible variables to add. It incorporates that independent variable that contributes the most explanatory power to the mode.

**Backward selection**

A stepwise variable selection process that begins with the full model, with all independent variables, and removes the independent variable that adds the least explanatory power to the model.

**Extra-sum-of-squares F-test**

Quantifies the difference between the amount of variance that is left unexplained by a reduced model that is explained by the full model

**Question**

What is the process of determining which variables or features to include in a given model?

Variable selection

Correct

Variable selection is the process of determining which variables or features to include in a given model. Variable selection is iterative.

**Bias-variance tradeoff**

Balance between two model qualities, bias and variance, to minimize overall error for unobserved data.

**Bias**

Simplifies the model predictions by making assumptions about the variable relationships. A highly biased model may oversimplify the relationship, under fitting to the observed data and generating inaccurate estimates.

**Variance**

Model flexibility and complexity, so the model can learn from existing data, but a model with high variants can over fit to the observed data and generate inaccurate estimates for unseen data.

**Regularization**

A set of regression techniques that shrinks regression coefficient estimates toward zero, adding in bias, to reduce variance.

**Regularized regression**

* Lasso regression
* Ridge regression
* Elastic-net regression

**Multicollinearity and model minimalism**

Previously, we discussed the concepts of multicollinearity and variable selection. As you continue learning about and applying multiple linear regression and these concepts, it is important to consider different scenarios. In your work as a data professional, you will encounter cases where you’ll have plenty of data, but you’ll have to consider the context and/or use advanced techniques to find the best model that only includes some of the data available.

For this discussion prompt, consider the following:

* Why is it important to not include every variable possible in a multiple linear regression?
* What are some scenarios where you might have a lot of variables available? Are there variables you think would be important? Are there variables you think might be less important? Why would you exclude them?

Submit 1–2 scenarios (200–400 words total) describing the potential regression problem, variables you might be able to exclude, and variables you might be able to include.  Then, visit the [discussion forums](https://www.coursera.org/learn/regression-analysis-simplify-complex-data-relationships/discussions) to read what other learners have written, and respond to at least one post with your own thoughts.

**Importance of Not Including Every Variable in a Multiple Linear Regression**

Including every possible variable in a multiple linear regression can lead to several problems. One of the primary concerns is multicollinearity, where two or more predictor variables are highly correlated. This can inflate the variance of the coefficient estimates and make the model unstable, leading to unreliable predictions and interpretations. Additionally, a model with too many variables can become overfit, meaning it captures the noise in the training data rather than the underlying pattern. This reduces the model's generalizability to new data. Moreover, including unnecessary variables can make the model more complex, harder to interpret, and computationally expensive.

**Scenario 1: Predicting Housing Prices**

**Potential Regression Problem:** In a project to predict housing prices in a large metropolitan area, you might have access to a wide range of variables, including:

* House characteristics (size, number of bedrooms, number of bathrooms)
* Location specifics (proximity to schools, crime rate, neighborhood amenities)
* Market factors (interest rates, local economic indicators)
* Temporal factors (seasonality, year of construction)

**Variables to Include:**

* Size of the house: A fundamental determinant of price.
* Number of bedrooms and bathrooms: Directly impacts usability and market demand.
* Location specifics: Proximity to good schools and low crime rates generally increase property value.

**Variables to Exclude:**

* Highly specific market factors: Some economic indicators might not significantly impact housing prices at a granular level or could be highly collinear with other variables like interest rates.
* Temporal factors: Year of construction could be less important if it does not correlate strongly with the current state of the property.

**Reason for Exclusion:** Excluding highly collinear variables or those with minimal impact on the price helps simplify the model and improves its predictive power. For example, interest rates might already be reflected in market factors, making some specific economic indicators redundant.

**Scenario 2: Customer Churn Prediction for a Subscription Service**

**Potential Regression Problem:** In predicting customer churn for a subscription-based service, you might have access to numerous variables, including:

* Customer demographics (age, gender, income)
* Service usage metrics (frequency of use, duration of subscription)
* Customer interaction data (customer service interactions, feedback)
* Competitor actions (promotions, pricing changes)

**Variables to Include:**

* Service usage metrics: Direct indicators of customer engagement and satisfaction.
* Customer interaction data: Reflects customer experience and potential dissatisfaction.

**Variables to Exclude:**

* Competitor actions: While potentially relevant, these are often harder to quantify accurately and might introduce noise if not directly impacting the customer's decision.
* Some demographic variables: Age and gender might be less predictive compared to direct service usage metrics.

**Reason for Exclusion:** Excluding competitor actions can help avoid introducing noise, especially if their impact is indirect or hard to measure. Similarly, demographic variables like age and gender might not provide additional predictive power beyond what is already captured by service usage and interaction metrics.

**Conclusion**

In both scenarios, carefully selecting variables helps create a parsimonious model that balances complexity and predictive accuracy. By focusing on variables that directly impact the outcome and avoiding those that introduce multicollinearity or noise, you can build a more robust and interpretable multiple linear regression model.

**Test your knowledge: Variable selection and model evaluation**

**1.**

Question 1

Fill in the blank: Adjusted R squared is a variation of the R squared regression evaluation metric that \_\_\_\_\_ unnecessary explanatory variables.

1 / 1 point

eliminates

penalizes

adds

rewards

Correct

Adjusted R squared is a variation of the R squared regression evaluation metric that penalizes unnecessary explanatory variables. Similar to R squared, adjusted R squared varies from less than 0 to 1.

**2.**

Question 2

Which of the following statements accurately describe the differences between adjusted R squared and R squared? Select all that apply.

1 / 1 point

R squared is more easily interpretable.

Correct

R squared determines how much variation in the dependent variable is explained by the model. Another difference is adjusted R squared is used to compare models of varying complexity.

R squared is used to compare models of varying complexity.

Adjusted R squared is more easily interpretable.

Adjusted R squared is used to compare models of varying complexity.

Correct

Adjusted R squared is used to compare models of varying complexity. R squared is more easily interpretable.

**3.**

Question 3

What variable selection process begins with the full model that has all possible independent variables?

1 / 1 point

Forward selection

F-test

Extra-sum-of Squares

Backward elimination

Correct

The backward elimination variable section process begins with the full model.

**4.**

Question 4

Which of the following are regularized regression techniques? Select all that apply.

0.5 / 1 point

Ridge regression

F-test regression

This should not be selected

Lasso regression, ridge regression, and elastic-net regression are regularized regression techniques. The extra-sum-of-squares F-test is a test to determine the threshold for when to add or remove variables.

Lasso regression

Correct

Lasso regression, ridge regression, and elastic-net regression are regularized regression techniques.

Elastic-net regression

Correct

Lasso regression, ridge regression, and elastic-net regression are regularized regression techniques.

**Glossary terms from module 3**

**Terms and definitions from Course 5, Module 3**

**Adjusted R**2: A variation of R2 that accounts for having multiple independent variables present in a linear regression model

**Backward elimination**: A stepwise variable selection process that begins with the full model, with all possible independent variables, and removes the independent variable that adds the least explanatory power to the model

**Bias**: Refers to simplifying the model predictions by making assumptions about the variable relationships

**Bias-variance trade-off**: Balance between two model qualities, bias and variance, to minimize overall error for unobserved data

**Errors**: The natural noise assumed to be in a regression model

**Extra Sum of Squares F-test**: Quantifies the difference between the amount of variance that is left unexplained by a reduced model that is explained by the full model

**Feature selection**: (Refer to **variable selection**)

**Forward selection**: A stepwise variable selection process that begins with the null mode—with 0 independent variables—which considers all possible variables to add; it incorporates the independent variable that contributes the most explanatory power to the model

**Homoscedasticity assumption**: An assumption of simple linear regression stating that the variation of the residuals (errors) is constant or similar across the model

**Independent observation assumption**: An assumption of simple linear regression stating that each observation in the dataset is independent

**Interaction term**: Represents how the relationship between two independent variables is associated with changes in the mean of the dependent variable

**Linearity assumption**: An assumption of simple linear regression stating that each predictor variable (Xi) is linearly related to the outcome variable (Y)

**Multiple linear regression**: A technique that estimates the relationship between one continuous dependent variable and two or more independent variables

**Multiple regression**: (Refer to **multiple linear regression**)

**No multicollinearity assumption**: An assumption of multiple linear regression stating that no two independent variables (Xi and Xj) can be highly correlated with each other

**Normality assumption**: An assumption of simple linear regression stating that the residuals are normally distributed

**One hot encoding**: A data transformation technique that turns one categorical variable into several binary variables

**Overfitting**: When a model fits the observed or training data too specifically and is unable to generate suitable estimates for the general population

**R**2 **(The Coefficient of Determination)**: The proportion of variance of the dependent variable, Y, explained by the independent variable or variables, X

**Regularization**: A set of regression techniques that shrinks regression coefficient estimates towards zero, adding in bias, to reduce variance

**Variable selection**: The process of determining which variables or features to include in a given model

**Variance**: Refers to model flexibility and complexity, so the model learns from existing data

**Variance inflation factors (VIF)**: Quantifies how correlated each independent variable is with all of the other independent variables

**Terms and definitions from previous modules**

**A**

**Absolute values**: (Refer to **observed values**)

**Adjusted R**2: A variation of R2 that accounts for having multiple independent variables present in a linear regression model

**B**

**Best fit line**: The line that fits the data best by minimizing some loss function or error

**C**

**Causation**: Describes a cause-and-effect relationship where one variable directly causes the other to change in a particular way

**Confidence band**: The area surrounding a line that describes the uncertainty around the predicted outcome at every value of X

**Confidence interval**: A range of values that describes the uncertainty surrounding an estimate

**Correlation**: Measures the way two variables tend to change together

**D**

**Dependent variable (Y)**: The variable a given model estimates

**E**

**Errors**: In a regression model, the natural noise assumed to be in a model

**Explanatory variable**: (Refer to **independent variable**)

**H**

**Hold-out sample**: A random sample of observed data that is not used to fit the model

**Homoscedasticity assumption**: The fourth assumption of simple linear regression, where the variation of the residuals (errors) is constant or similar across the model

**I**

**Independent observation assumption**: The third assumption of simple linear regression, where each observation in the dataset is independent

**Independent variable (X)**: A variable that explains trends in the dependent variable

**Intercept (constant 𝐵**0**)**: The y value of the point on the regression line where it intersects with the y-axis

**L**

**Line**: A collection of an infinite number of points extending in two opposite directions

**Linearity assumption**: The first assumption of simple linear regression, where each predictor variable (Xi) is linearly related to the outcome variable (Y)

**Linear regression**: A technique that estimates the linear relationship between a continuous dependent variable and one or more independent variables

**Link function**: A nonlinear function that connects or links the dependent variable to the independent variables mathematically

**Logistic regression**: A technique that models a categorical dependent variable based on one or more independent variables

**Loss function**: A function that measures the distance between the observed values and the model’s estimated values

**M**

**MAE (Mean Absolute Error)**: The average of the absolute difference between the predicted and actual values

**Model assumptions**: Statements about the data that must be true in order to justify the use of a particular modeling technique

**MSE (Mean Squared Error)**: The average of the squared difference between the predicted and actual values

**N**

**Negative correlation**: An inverse relationship between two variables, where when one variable increases, the other variable tends to decrease, and vice versa

**Normality assumption**: The second assumption of simple linear regression, where the residual values or errors are normally distributed

**O**

**Observed values:** The existing sample of data, where each data point in the sample is represented by an observed value of the dependent variable and an observed value of the independent variable

**Ordinary least squares (OLS)**: A method that minimizes the sum of squared residuals to estimate parameters in a linear regression model

**Outcome variable (Y)**: (Refer to **dependent variable**)

**P**

**P-value**: The probability of observing results as extreme as those observed when the null hypothesis is true

**Positive correlation**: A relationship between two variables that tend to increase or decrease together.

**Predicted values**: The estimated Y values for each X calculated by a model

**Predictor variable**: (Refer to **independent variable**)

**R**

**R**2 **(The Coefficient of Determination)**: Measures the proportion of variation in the dependent variable, Y, explained by the independent variable(s), X

**Residual**: The difference between observed or actual values and the predicted values of the regression line

**Regression analysis**: A group of statistical techniques that use existing data to estimate the relationships between a single dependent variable and one or more independent variables

**Regression coefficient**: The estimated betas in a regression model

**Regression models**: (Refer to **regression analysis**)

**Response variable: (**Refer to **dependent variable)**

**S**

**Scatterplot matrix**: A series of scatterplots that demonstrate the relationships between pairs of variables

**Simple linear regression**: A technique that estimates the linear relationship between one independent variable, X, and one continuous dependent variable, Y

**Slope**: The amount that y increases or decreases per one-unit increase of x

**Sum of squared residuals (SSR)**: The sum of the squared difference between each observed value and its associated predicted value

**Module 3 challenge**

Question 1

A data professional at a car rental agency uses a regression technique to learn about how customers engage with various sections of the company website. They estimate the linear relationship between one continuous dependent variable and three independent variables. What technique are they using?

Multiple linear regression

Question 2

Which of the following are examples of categorical variables? Select all that apply.

* Shirt size
* Shirt type
* Shirt country of manufacture

Question 3

Fill in the blank: The no multicollinearity assumption states that no two \_\_\_\_\_ variables can be highly correlated with each other.

independent

Question 4

What term represents how the relationship between two independent variables is associated with changes in the mean of the dependent variable?

Interaction term

Question 5

Which of the following statements accurately describe adjusted R squared? Select all that apply.

It can vary from 0 to 1.

It penalizes unnecessary explanatory variables.

It is a regression evaluation metric.

Question 6

What stepwise variable selection process begins with the null model and zero independent variables?

Forward selection

Question 7

A data professional reviews model predictions for a human resources project. They discover that the model performs poorly on both the training data and the test holdout data, consistently predicting figures that are too low. This leads to inaccurate estimates about employee retention. What quality does this model have too much of?

Bias

Question 8 (NOT FINAL)

What regularization technique is recommended when working with large datasets and when there is uncertainty as to whether variables should drop out of the model?

Ridge regression

**Module 4:**

Chi-squared [x^2] tests will help us determine if two categorical variables are associated with one another, and whether a categorical variable follows an expected distribution.

**X^2 Test for independence**

Determine whether or not two categorical variables are associated with each other.

**Question**

Fill in the blank: The chi-squared test for independence determines whether \_\_\_\_\_ categorical variables are associated with each other.

two

**Chi-squared tests: Goodness of fit versus independence**

In the previous course, you learned how hypothesis tests are used to see significant differences among groups. Chi-squared tests are used to determine whether one or more observed categorical variables follow expected distribution(s). For example, you may expect that 50% more movie goers attend movies on weekends in comparison to weekdays. After observing movie goers attendance for a month, you then can perform a chi-squared test to see if your initial hypothesis was correct.

This reading will cover the two main chi-squared tests—goodness of fit and test for independence—which can be used to test your expected hypothesis against what actually occurred. Data professionals perform these hypothesis tests to offer organizations actionable insights that drive decision making.

**The Chi-squared goodness of fit test**

**Chi-squared (χ²) goodness of fit test** is a hypothesis test that determines whether an observed categorical variable with more than two possible levels follows an expected distribution. The null hypothesis (H0) of the test is that the categorical variable follows the expected distribution. The alternative hypothesis (Ha) is that the categorical variable does not follow the expected distribution. Consider the scenario in this reading that will define the null and alternative hypotheses based on the scenario, set up a Goodness of Fit test, evaluate the test results, and draw a conclusion.

**Chi-squared goodness of fit scenario**

Imagine that you work as a data professional for an online clothing company. Your boss tells you that they expect the number of website visitors to be the same for each day of the week. You decide to test your boss’s hypothesis and pull data every day for the next week and record the number of website visitors in the table below:

| **Day of the Week** | **Observed Values** |
| --- | --- |
| Sunday | 650 |
| Monday | 570 |
| Tuesday | 420 |
| Wednesday | 480 |
| Thursday | 510 |
| Friday | 380 |
| Saturday | 490 |
| Total | 3,500 |

Here are the main steps you will take:

1. Identify the Null and Alternative Hypotheses
2. Calculate the chi-square test statistic (**𝛘**2)
3. Calculate the p-value
4. Make a conclusion

**Step 1: Identify the null and alternative hypotheses**

The first step in performing a chi-squared goodness of fit test is to determine your null and alternative hypothesis. Since you are testing if the number of website visitors follows your boss’s expectations, the below are your null and alternative hypotheses :

H0: The week you observed follows your boss’s expectations that the number of website visitors is equal on any given day

Ha: The week you observed does not follow your boss’s expectations; therefore, the number of website visitors is not equal across the days of the week

**Step 2: Calculate the chi-squared test statistic (𝛘2)**

Next, calculate a test statistic to determine if you should reject or fail to reject your null hypothesis. This test statistic is known as the chi-squared statistic and is calculated based on the following formula:

Σ(observed−expected)2expectedΣ*expected*(*observed*−*expected*)2​

The intuition behind this formula is that it should quantify the extent of any discrepancies between observed frequencies and expected frequencies for each categorical level. Squaring these differences does two things. First, it ensures that all discrepancies between observed and expected contribute positively to the chi-squared statistic. Second, it penalizes larger discrepancies. Dividing the sum of the squared differences by the expected frequency of each category level standardizes the differences. In other words, it accounts for the fact that larger discrepancies are more significant when the expected frequencies are small, and less so when the expected frequencies are large.

Returning to the example, since there were a total of 3,500 website visitors you observed; your boss’s expectation is that 500 visitors would visit each day (3,500/7). In the formula above, 500 would serve as the “expected” value. A column has been added to your original table to include the test statistic calculation for each weekday:

| **Day of the Week** | **Observed Values** | **Chi-Squared Test Statistic** |
| --- | --- | --- |
| Sunday | 650 | (650−500)2500=45500(650−500)2​=45 |
| Monday | 570 | (570−500)2500=9.8500(570−500)2​=9.8 |
| Tuesday | 420 | (420−500)2500=12.8500(420−500)2​=12.8 |
| Wednesday | 480 | (480−500)2500=0.8500(480−500)2​=0.8 |
| Thursday | 510 | (510−500)2500=0.2500(510−500)2​=0.2 |
| Friday | 380 | (380−500)2500=28.8500(380−500)2​=28.8 |
| Saturday | 490 | (490−500)2500=0.2500(490−500)2​=0.2 |

The 𝛸2 statistic would be the sum of the third column above:

𝛸2 = 45 + 9.8 + 12.8 + 0.8 + 0.2 + 28.8 + 0.2

𝛸2 = 97.6

Note that the 𝛸2 goodness of fit test does not produce reliable results when there are any expected values of less than five.

**Step 3: Find the p-value**

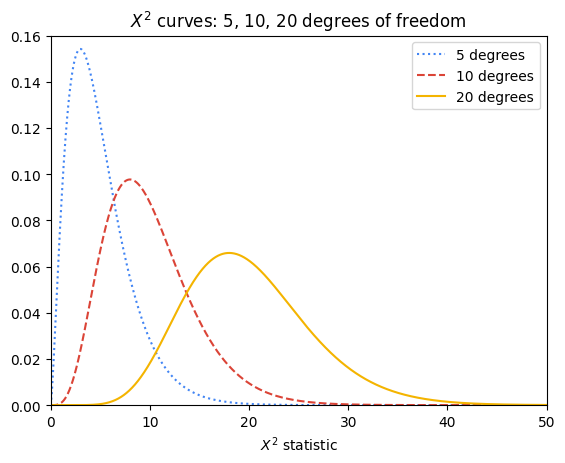
Now that you have calculated the 𝛸2 statistic, consider the following question: What is the probability of obtaining a 𝛸2 statistic of 97.6 or greater when examining 3,500 website visits if they null hypothesis is true? This is the question that the p-value—or “observed significance level"—will answer.

A long time ago, Pearson realized that p-values for 𝛸2 statistics very closely corresponded with areas under certain curves, known as 𝛸2 curves. 𝛸2 curves represent probability density functions, and their shapes vary based on how many degrees of freedom are present in the experiment. Degrees of freedom are determined by the model, not by the data. This means that, in the website traffic example, the degrees of freedom are determined by how many different days a given visit can occur on—not by how many visits are sampled nor by the daily frequencies of the samples themselves. When the model is fully specified (i.e., you know all the possible categorical levels), then:

**degrees of freedom = number of categorical levels – 1**.

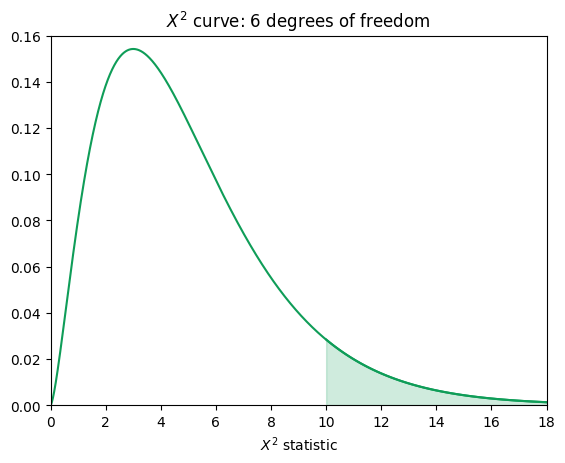
In this example, there are seven categorical levels (one for each day of the week). Therefore, there are six degrees of freedom. This is because the counts of each level (day) are free to fluctuate, but once you know the counts for six days, the seventh day cannot vary. It must result in a total of 3,500 when summed with the other six days.

The following figure depicts the 𝛸2 curves for three different degrees of freedom: five, 10, and 20.



The p-value for a given 𝛸2 test statistic is very closely approximated by the area to its right beneath the 𝛸2 curve of the appropriate degrees of freedom. Notice that the more degrees of freedom there are in the experiment, the greater the area under the right tail of the curve for any given 𝛸2 test statistic, and therefore the greater the probability of getting a given 𝛸2 test statistic if the null hypothesis is true.

The following figure contains the 𝛸2 curve for six degrees of freedom. For a 𝛸2 test statistic of, for instance, 10, the value of P is approximated by the shaded area under the curve where x ≥ 10.



In the case of the website visits, there are six degrees of freedom, but the 𝛸2 test statistic is 97.6—far along the x-axis in the right-skewed tail of the curve. The area under this interval is miniscule: 7.94e-19. In other words, the chances of getting a 𝛸2 test statistic ≥ 97.6 from 3,500 website visits if the null hypothesis were true are 7.94e-17%—practically zero.

**Step 4: Make a conclusion**

Since the p-value is far less than 0.05, there is sufficient evidence to suggest that the number of visitors is not equal per day.

**Coding**

Thankfully, you don’t need to manually calculate your 𝛸2 test statistic or determine P by hand. You can use the [chisquare() function](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html) from Python’s **scipy.stats** package to do this. The following code uses your observed and expected values to calculate the chi-squared test statistic and the p-value. Note that the degrees of freedom are set to the number of observed frequencies minus one. This can be adjusted using the **ddof** parameter, but note that this parameter represents **k - 1 - ddof** degrees of freedom, where **k** is the number of observed frequencies. So, by default, **ddof=0** when you call the function, and setting **ddof=1** means that your degrees of freedom are reduced by two.

1

2

3

4

5

import scipy.stats as stats

observations = [650, 570, 420, 480, 510, 380, 490]

expectations = [500, 500, 500, 500, 500, 500, 500]

result = stats.chisquare(f\_obs=observations, f\_exp=expectations)

result

RunReset

The output confirms your calculation of the chi-square test statistic in Step 2 and also gives you the associated p-value. Because the p-value is less than the significance level of 5%, you can reject the null hypothesis.

**The Chi-squared test for independence**

**Chi-squared (χ²) Test for Independence** is a hypothesis test that determines whether or not two categorical variables are associated with each other. It is valid when your data comes from a random sample and you want to make an inference about the general population. The null hypothesis (H0) of the test is that two categorical variables are independent. The alternative hypothesis (Ha) is that two categorical variables are not independent.

**Chi-squared test for independence scenario**

Now suppose that you have been asked to expand your analysis to look at the relationship between the device that a website user used and their membership status. To do this, you must use the 𝛸2 test for independence. In this example, the 𝛸2 test of independence determines whether the type of device a visitor uses to visit the website (Mac or PC) is dependent on whether he or she has a membership account or browses as a guest (member or guest).

**Step 1: Identify the null and alternative hypotheses**

Just like the goodness of fit scenario, the first step is to determine your null and alternative hypotheses. You are comparing if the device used to visit your clothing store (Mac or PC) is independent from the visitor’s membership status (member or guest). From that information you can determine that your null and alternative hypotheses are as follows:

H0: The type of device a website visitor uses to visit the website is independent of the visitor’s membership status.

Ha: The type of device a website visitor uses to visit the website is not independent of the visitor’s membership status.

**Step 2: Calculate the chi-squared test statistic (𝛘2)**

To calculate the 𝛸2 test statistic, arrange the data as a table that contains *m* x *n* values, where *m* and *n* are the number of possible levels contained within each respective categorical variable. The following table below breaks down the website visitors based on the device they used and their membership status.

| **Observed Values** | **Member** | **Guest** | **Total** |
| --- | --- | --- | --- |
| **Mac** | 850 | 450 | 1,300 |
| **PC** | 1,300 | 900 | 2,200 |
| **Total** | 2,150 | 1,350 | 3,500 |

Notice that the table starts with 2 x 2 known values (two levels for each category), from which totals are derived. These totals can be used to calculate the expected values, which are necessary to get the 𝛸2 test statistic.

To calculate the expected values, use the following formula:

expected value = column total ∗ row totaloverall total*expected* *value* = *overall* *totalcolumn* *total* ∗ *row* *total*​

For example, the expected value for a Mac member would be:

expected value = 2,150 ∗ 1,3003,500 =799*expected* *value* = 3,5002,150 ∗ 1,300​ =799

The logic of this calculation is as follows: if device and membership status are truly independent, then the rate of Mac users who are members should be the same as the rate of Mac users who are guests. The percentage of users who use Macs out of *all* the users is 1,300 / 3,500 = 0.371 \* 100 = 37.1%. Accordingly, 37.1% of members and 37.1% of guests would be expected to use Macs. So, 0.371 \* 2,150 members ≈ 799.

The following table contains all the expected values:

| **Expected Values** | **Member** | **Guest** |
| --- | --- | --- |
| **Mac** | 799 | 501 |
| **PC** | 1,351 | 849 |

**Step 3: Find the p-value**

Finding the p-value associated with a particular 𝛸2 test statistic is similar to the process outlined already for the goodness of fit test. The only minor difference is how to determine the number of degrees of freedom. For an independence test with two categorical variables with *m* x *n* possible levels, there are (*m* – 1) (*n* – 1) degrees of freedom, assuming there are no other constraints on the probabilities. So, in the working example, this means there is (2 – 1)(2 – 1) = 1 degree of freedom. The p-value in this example is 0.00022. It was determined using Python.

**Step 4: Make a conclusion**

Because the p-value is 0.00022, you can reject the null hypothesis in favor of the alternative. You conclude that the type of device a website user uses is not independent of his or her membership status. You may recommend to your boss to dive into the reasons behind why visitors sign up for paid memberships more on a particular device. Perhaps the sign-up button appears differently on a particular device. Or maybe there are device-specific bugs that need to be fixed. These are just a couple examples of things you might consider for further exploration.

**Coding**

You can use the **scipy.stats** package’s [chi2\_contingency() function](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html) to obtain the 𝛸2 test statistic and p-value of a 𝛸2 independence test. The **chi2\_contingency()** function only needs the observed values; it will calculate the expected values for you. Here is the Python code:

1

2

3

4

5

import numpy as np

import scipy.stats as stats

observations = np.array([[850, 450],[1300, 900]])

result = stats.contingency.chi2\_contingency(observations, correction=False)

result

RunReset

The output above is in the following order: the 𝛸2 statistic, p-value, degrees of freedom, and expected values in array format. One thing to note is that when degrees of freedom = 1 (i.e., you have a 2 x 2 table), the default behavior of the **stats.chi2\_contingency()** function is to apply [Yates’ correction for continuity](https://en.wikipedia.org/wiki/Yates's_correction_for_continuity). This is to make it less likely that small discrepancies will result in significant 𝛸2 values. It is designed to be used when it’s possible for an expected frequency in the table to be small (generally < 5). In the given example, it is known that the expected values are all well over five. Therefore, the **correction** parameter was set to False.

**Key takeaways**

* The 𝛸2 goodness of fit test is used to test if an observed categorical variable follows a particular expected distribution.
* The 𝛸2 test for independence is used to test if two categorical variables are independent of each other or not (when samples are drawn at random and you want to make an inference about the whole population).
* Both 𝛸2 tests follow the same hypothesis testing steps to determine whether you should reject or fail to reject the null hypothesis to drive decision making, as you have explored elsewhere in this program.

**Test your knowledge: The chi-squared test**

**1.**

Question 1

The chi-squared goodness of fit test determines whether an observed categorical variable follows an expected distribution.

1 / 1 point

True

Correct

The chi-squared goodness of fit test determines whether an observed categorical variable follows an expected distribution. The test’s null hypothesis states that the variable follows the expected distribution. The alternative hypothesis states that the variable doesn’t follow the expected distribution.

Question 2

Which test determines whether two categorical variables are associated with each other?

1 / 1 point

Chi-squared test for independence

Correct

The chi-squared test for independence determines whether two categorical variables are associated with each other. The test’s null hypothesis is that the variables are independent. The alternative hypothesis states that the variables are not independent and are therefore associated with each other.

Question 3

Fill in the blank: The chi-squared statistic equals the sum of the observed number minus the expected number, squared, divided by the \_\_\_\_\_ number.

1 / 1 point

 expected

Correct

The chi-squared statistic equals the sum of the observed number minus the expected number, squared, divided by the expected number

**Analysis of Variance (ANOVA)**

A group of statistical techniques that test the difference of means between three or more groups

**One-way ANOVA**

Compares the means of one continuous dependent variable based on three or more groups of one categorical variable

**Question**

Fill in the blank: Analysis of variance is a group of statistical techniques that test the difference of means between \_\_\_\_\_ groups.

three or more

Correct

Analysis of variance is a group of statistical techniques that test the difference of means between three or more groups. This is an extension of t-tests, which tests the means between several groups.

**Two-way ANOVA**

Compares the means of one continuous dependent variable based on three or more groups of two categorical variables.

**More about ANOVA**

You’ve learned that analysis of variance—ANOVA—is a group of statistical techniques that test the difference of means between groups. ANOVA testing is useful when you want to test a hypothesis about group differences based on categorical independent variables.  For example, if you wanted to determine whether changes in people’s weight when following different diets are statistically significant or due to chance, you could use ANOVA to analyze the results. Data professionals routinely must ascertain if there are meaningful differences between groups in their data. This reading will examine more closely the intuition behind ANOVA using a worked example. Later in the program, you will learn how to implement ANOVA in Python.

**An overview of ANOVA**

The intuition behind ANOVA is to compare the variability between different groups with the variability within the groups. If they are comparable, then the differences between groups are more likely to be due to sampling variability. On the other hand, if the variability between groups is much larger than the variability expected from the samples within their respective groups, then those groups are probably drawn from significantly different subpopulations.

The variation between groups and within groups is calculated as sums of squares, which are then expressed as a ratio. This ratio is known as the F-statistic. The formula for each component of these calculations is presented in the worked example that follows.

Previously, you learned about one-way and two-way ANOVA. To review:

* **One-way ANOVA:** Compares the means of one continuous dependent variable based on three or more groups of **one** categorical variable
* **Two-way ANOVA:** Compares the means of one continuous dependent variable based on three or more groups of **two** categorical variables

To help you understand the intuition behind ANOVA, this reading will break down a worked example of a simple one-way ANOVA test.

**One-way ANOVA**

**5 steps**

There are five steps in performing a one-way ANOVA test:

1. Calculate group means and grand (overall) mean
2. Calculate the sum of squares between groups (SSB) and the sum of squares within groups (SSW)
3. Calculate mean squares for both SSB and SSW
4. Compute the F-statistic
5. Use the F-distribution and the F-statistic to get a p-value, which you use to decide whether to reject the null hypothesis

**Example**

Return to the example of students studying for an exam. Suppose that in this case you wanted to compare three different studying programs, A, B, and C to determine whether they have an effect on exam score. Here is the data:

| **Student** | **Study program (X)** | **Exam score (Y)** |
| --- | --- | --- |
| 1 | A | 88 |
| 2 | A | 79 |
| 3 | A | 86 |
| 4 | A | 90 |
| 5 | B | 94 |
| 6 | B | 84 |
| 7 | B | 87 |
| 8 | B | 89 |
| 9 | C | 85 |
| 10 | C | 76 |
| 11 | C | 81 |
| 12 | C | 78 |

First, state your hypotheses:

H0: 𝜇A =  𝜇B = 𝜇C

The mean score of group A = the mean score of group B = the mean score of group C.

H1: NOT (𝜇A =  𝜇B = 𝜇C)

The means of each group are not all equal. Remember, even if only one mean differs, that is sufficient evidence to reject the null hypothesis.

Next, determine your confidence level—the threshold above which you will reject the null hypothesis. This value is dependent on your situation and usually requires some domain knowledge. A common threshold is 95%.

Now, begin the steps of ANOVA.

**Step 1**

**Calculate group means and grand mean. The grand mean is the overall mean of all samples in all groups.**

The following table restructures the data in the previous table such that the scores for each study group are contained in their own column. Additionally, the mean score of each group has been calculated.

| **Study program A scores** | **Study program B scores** | **Study program C scores** |
| --- | --- | --- |
| 88 | 94 | 85 |
| 79 | 84 | 76 |
| 86 | 87 | 81 |
| 90 | 89 | 78 |
| **Mean: 85.75** | **Mean: 88.5** | **Mean: 80** |

**Grand mean (M**G**) = 84.75**

**Step 2**

**A. Calculate the sum of squares between groups (SSB).**

SSB=∑g=1ng(Mg−MG)2SSB=*g*=1∑​*ng*​(*Mg*​−*MG*​)2

where:

ng*ng*​ = the number of samples in the *g*th group

Mg*Mg*​ = mean of the *g*th group

MG*MG*​ = grand mean

→ **SSB** = [4(85.75 – 84.75)² + 4(88.5 – 84.75)² + 4(80 – 84.75)²]

= 4 + 56.25 + 90.25

**= 150.5**

**B. Calculate the sum of squares within groups (SSW).**

SSW=∑g=1∑i=1(xgi−Mg)2SSW=*g*=1∑​*i*=1∑​(*xgi*​−*Mg*​)2

where:

xgi*xgi*​ = sample *i* of the *g*th group

Mg*Mg*​ = mean of the *g*th group

The double summation acts like a nested loop. The outer loop is for each group and the inner loop is for all the samples in that group. So, for each sample in group 1, subtract from it the group's mean and square the result. Then, do the same thing with the samples in group 2, using the group 2 mean. Continue this way for all the groups and sum all the results.

The following table shows the squared difference between each observation and its group mean. It also contains the sums of these squared differences for each of the three study groups, A, B, and C.

| **Program A** | **Program B** | **Program C** | **(xAi−MA)2(*xAi*​−*MA*​)2** | **(xBi−MB)2(*xBi*​−*MB*​)2** | **(xCi−MC)2(*xCi*​−*MC*​)2** |
| --- | --- | --- | --- | --- | --- |
| 88 | 94 | 85 | 5.06 | 30.25 | 25 |
| 79 | 84 | 76 | 45.56 | 20.25 | 16 |
| 86 | 87 | 81 | 0.06 | 2.25 | 1 |
| 90 | 89 | 78 | 18.06 | 0.25 | 4 |
| MA*MA*​ = 85.75 | MB*MB*​ = 88.5 | MC*MC*​ = 80 | **Sum: 68.75** | **Sum: 53** | **Sum: 46** |

→ **SSW** = 68.75 + 53 + 46

**= 167.75**

**Step 3**

**Calculate mean squares between groups and within groups. The mean square is the sum of squares divided by the degrees of freedom, respectively.**

**A. Mean squares between groups (MSSB):**

MSSB=SSBk−1MSSB=*k*−1SSB​

where:

k*k* = the number of groups

Note: k−1*k*−1 represents the degrees of freedom between groups

→ **MSSB** = 150.53−13−1150.5​

**= 75.25**

**B.Mean squares within groups (MSSW):**

MSSW=SSWn−kMSSW=*n*−*k*SSW​

where:

n*n* = the total number of samples in all groups

k*k* = the number of groups

Note: n−k*n*−*k* represents the degrees of freedom within groups

→ **MSSW** = 167.7512−312−3167.75​

**= 18.64**

**Step 4**

**Compute the F-statistic.**

The F-statistic is the ratio of the mean sum of squares between groups (MSSB) to the mean sum of squares within groups (MSSW):

F-statistic=MSSBMSSWF-statistic=MSSWMSSB​

→ **F-statistic** = 75.2518.64 18.6475.25​

**= 4.04**

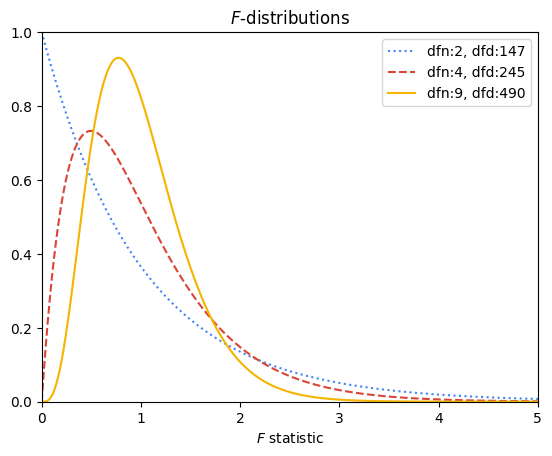
A higher F-statistic indicates a greater variability between group means relative to the variability within groups, suggesting that at least one group mean is significantly different from the others.

**Step 5**

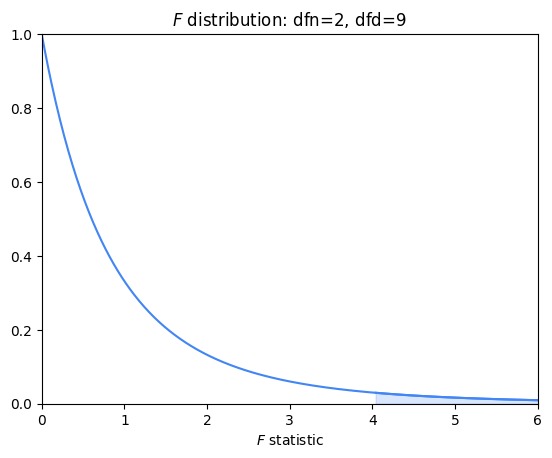
**Use the F-distribution and the F-statistic to get a p-value, which you use to decide whether to reject the null hypothesis.**

Similar to t-tests and 𝛸2 tests, ANOVA testing finds the area under a particular probability distribution curve—the F-distribution—of the null hypothesis to determine a p-value. The larger the F-statistic, the lesser the area beneath the curve and the more evidence against the null hypothesis, thus resulting in a lower p-value.

The shape of the F-distribution curve is determined by the degrees of freedom between and within groups. Here is a graph depicting F-distributions for three, five, and 10 groups—each group containing 50 samples. Note that “dfn” represents the degrees of freedom in the numerator (between groups) and “dfd” represents the degrees of freedom in the denominator (within groups). Notice how the degrees of freedom affect the shape of the curve.



Similar to 𝛸2 curves, F-distributions help determine the probability of falsely rejecting the null hypothesis. In the case of ANOVA, this probability is represented by the area of the F-distribution beneath the curve where x ≥ your F-statistic. For example, the following graph depicts the F-distribution for the exam scores example. It has two degrees of freedom in the numerator and nine degrees of freedom in the denominator. The area beneath the curve where x ≥ 4.04 (the computed F-statistic) is shaded.



You can use statistical software to calculate this area. You will learn how to do this in a later activity. In this case, the area beneath the F-distribution to the right of 4.04 is 0.05604. This is the probability of observing an F-statistic greater than 4.04 if the null hypothesis were true. Whether this is sufficient to reject the null hypothesis is a decision you make at the beginning of your hypothesis test. For example, if you decided that you wanted a confidence level of 95% or greater, you cannot reject the null hypothesis that the means of the distributions for each program of study are all the same, because the p-value is 0.056.

**Assumptions of ANOVA**

ANOVA will only work if the following assumptions are true:

1. The dependent values for each group come from normal distributions
   * Note that this assumption does NOT mean that *all* of the dependent values, taken together, must be normally distributed. Instead, it means that *within each group*, the dependent values should be normally distributed.
   * ANOVA is generally robust to violations of normality, especially when sample sizes are large or similar across groups, due to the central limit theorem. However, significant violations can lead to incorrect conclusions.
2. The variances across groups are equal
   * ANOVA compares means across groups and assumes that the variance around these means is the same for all groups. If the variances are unequal (i.e., heteroscedastic), it could lead to incorrect conclusions
3. Observations are independent of each other
   * ANOVA assumes that one observation does not influence or predict any other observation. If there is autocorrelation among the observations, the results of the ANOVA test could be biased.

**Key takeaways**

* ANOVA tests are statistical tests that examine whether or not the means of a continuous dependent variable are significantly different from one another based on the different levels of one or more independent categorical variables.
* It is sufficient for one group’s mean to be significantly different from the others to reject the null hypothesis; however, ANOVA testing is limited in that it doesn’t tell you *which* group is different. To make such a determination, other tests are necessary.
* ANOVA works by comparing the variance between each group to the variance within each group. The greater the ratio of variance between groups to variance within groups, the greater the likelihood of rejecting the null hypothesis.
* ANOVA depends on certain assumptions, so it is important to check that your data meets them in order to avoid drawing false conclusions. At the very least, if your data does not meet all of them, identify these violations.

**Null hypothesis (H\_0)**

There is no difference in the price of diamonds based on color grade.

**Alternative hypothesis (H\_1)**

There is a difference in the price of diamonds based on color grade.

**Post hoc test**

Performs a pairwise comparison between all available groups while controlling for the error rate.

“If you’re going to enter this field, professional development is something that will always be a part of your life.”

“I made the decision of investing on myself.”

Practice Quiz

**Test your knowledge: Analysis of variance**

**1.**

Question 1

Which of the following statements accurately describe t-tests and analyses of variance? Select all that apply.

1 / 1 point

A t-test can only test the difference of mean between two groups.

Correct

An analysis of variance test can test means between several groups.

Correct

A t-test can only test the difference of mean between two groups. An analysis of variance test can test means between several groups.

An analysis of variance test can only test the difference of mean between two groups.

**2.**

Question 2

Which of the following are analysis of variance (ANOVA) tests? Select all that apply.

1 / 1 point

Two-way ANOVA

Correct

One-way ANOVA and two-way ANOVA are types of analysis of variance tests. Analysis of variance, commonly called ANOVA, is a group of statistical techniques that test the difference of means between three or more groups.

One-way ANOVA

Correct

One-way ANOVA and two-way ANOVA are types of analysis of variance tests. Analysis of variance, commonly called ANOVA, is a group of statistical techniques that test the difference of means between three or more groups.

**3.**

Question 3

Fill in the blank: A post hoc test performs a pairwise comparison between all available groups while controlling for the \_\_\_\_\_.

1 / 1 point

error rate

Correct

A post hoc test performs a pairwise comparison between all available groups while controlling for the error rate. There is always a small chance that the null hypothesis is falsely rejected purely based on probability. The post hoc ANOVA test controls for that increasing probability.

**ANCOVA (Analysis of Covariance)**

A statistical technique that tests the difference of means between three or more groups while controlling for the effects of covariates, or variable(s) irrelevant to your test.

**Question**

Covariates are the variables that are directly relevant to the question to be answered in an analysis of covariance test.

False

**MANCOVA (Multivariate Analysis of Covariance)**

An extension of ANCOVA and MANOVA that compares how two or more continuous outcome variables vary according to categorical independent variables, while controlling for covariates.

**Test your knowledge: ANCOVA, MANOVA, and MANCOVA**

**1.**

Question 1

Which statistical technique better isolates the relationship between a single categorical variable of interest and the Y variable?

1 / 1 point

One-way ANOVA

Multivariate analysis of variance (MANOVA)

Multivariate analysis of covariance (MANCOVA)

Analysis of covariance (ANCOVA)

Correct

Analysis of covariance (ANCOVA) better isolates the relationship between a single categorical variable of interest and the Y variable. By taking the covariate into account, the ANCOVA technique allows data professionals to draw more accurate conclusions about the relationships among variables.

**2.**

Question 2

Which of the following statements accurately describe ANCOVA and linear regression? Select all that apply.

1 / 1 point

ANCOVA allows for continuous and categorical independent variables

Correct

ANCOVA includes covariates to gain a more clear understanding of the categorical variable. Linear regression helps predict the Y variable for unrecognized data.

Linear regression focuses on a continuous Y variable

Correct

ANCOVA includes covariates to gain a more clear understanding of the categorical variable. Linear regression helps predict the Y variable for unrecognized data.

ANCOVA includes covariates to gain a more clear understanding of the categorical variable.

Correct

ANCOVA includes covariates to gain a more clear understanding of the categorical variable. Linear regression helps predict the Y variable for unrecognized data.

Linear regression helps predict the Y variable for unrecognized data.

Correct

ANCOVA includes covariates to gain a more clear understanding of the categorical variable. Linear regression helps predict the Y variable for unrecognized data.

**3.**

Question 3

What is the key difference between MANCOVA and MANOVA?

1 / 1 point

MANOVA has two or more continuous variables.

MANCOVA controls for covariates.

MANCOVA includes a null hypothesis.

MANOVA includes a categorical variable.

Correct

The key difference between MANCOVA and MANOVA is that MANCOVA controls for covariates. If a data professional is only interested in one categorical variable and they want to control for another variable, they can use MANCOVA.

**Glossary terms from module 4**

**Terms and definitions from Course 5, Module 4**

**Analysis of Variance (ANOVA)**: A group of statistical techniques that test the difference of means between three or more groups

**ANCOVA (Analysis of Covariance)**: A statistical technique that tests the difference of means between three or more groups while controlling for the effects of covariates, or variable(s) irrelevant to the test

**Chi-squared (χ²) Goodness of Fit Test**: A hypothesis test that determines whether an observed categorical variable follows an expected distribution

**Chi-squared (χ²) Test for Independence**: A hypothesis test that determines whether or not two categorical variables are associated with each other

**Hypothesis testing**: A statistical procedure that uses sample data to evaluate an assumption about a population parameter

**One-Way ANOVA**: A type of statistical testing that compares the means of one continuous dependent variable based on three or more groups of one categorical variable

**MANCOVA (Multivariate Analysis of Covariance)**: An extension of ANCOVA and MANOVA that compares how two or more continuous outcome variables vary according to categorical independent variables, while controlling for covariates

**MANOVA (Multivariate Analysis of Variance)**: An extension of ANOVA that compares how two or more continuous outcome variables vary according to categorical independent variables

**Post hoc test**: An ANOVA test that performs a pairwise comparison between all available groups while controlling for the error rate

**Two-Way ANOVA**: A type of statistical testing that compares the means of one continuous dependent variable based on three or more groups of two categorical variables

**Terms and definitions from previous modules**

**A**

**Absolute values**: (Refer to **observed values**)

**Adjusted R**2: A variation of R2 that accounts for having multiple independent variables present in a linear regression model

**B**

**Backward elimination**: A stepwise variable selection process that begins with the full model, with all possible independent variables, and removes the independent variable that adds the least explanatory power to the model

**Best fit line**: The line that fits the data best by minimizing some loss function or error

**Bias**: Refers to simplifying the model predictions by making assumptions about the variable relationships

**Bias-variance trade-off**: Balance between two model qualities, bias and variance, to minimize overall error for unobserved data

**C**

**Causation**: Describes a cause-and-effect relationship where one variable directly causes the other to change in a particular way

**Confidence band**: The area surrounding a line that describes the uncertainty around the predicted outcome at every value of X

**Confidence interval**: A range of values that describes the uncertainty surrounding an estimate

**Correlation**: Measures the way two variables tend to change together

**D**

**Dependent variable (Y)**: The variable a given model estimates

**E**

**Errors**: In a regression model, the natural noise assumed to be in a model

**Explanatory variable**: (Refer to **independent variable**)

**Extra Sum of Squares F-test:** Quantifies the difference between the amount of variance that is left unexplained by a reduced model that is explained by the full model

**F**

**Feature selection**: (Refer to **variable selection**)

**Forward selection**: A stepwise variable selection process that begins with the null mode—with 0 independent variables—which considers all possible variables to add; it incorporates the independent variable that contributes the most explanatory power to the model

**H**

**Hold-out sample**: A random sample of observed data that is not used to fit the model

**Homoscedasticity assumption**: An assumption of simple linear regression stating that the variation of the residuals (errors) is constant or similar across the model

**I**

**Independent observation assumption**: An assumption of simple linear regression stating that each observation in the dataset is independent

**Independent variable (X)**: The variable whose trends are associated with the dependent variable

**Interaction term**: Represents how the relationship between two independent variables is associated with changes in the mean of the dependent variable

**Intercept (constant 𝐵**0**)**: The y value of the point on the regression line where it intersects with the y-axis

**L**

**Line**: A collection of an infinite number of points extending in two opposite directions

**Linear regression**: A technique that estimates the linear relationship between a continuous dependent variable and one or more independent variables

**Linearity assumption**: An assumption of simple linear regression stating that each predictor variable (Xi) is linearly related to the outcome variable (Y)

**Link function**: A nonlinear function that connects or links the dependent variable to the independent variables mathematically

**Logistic regression**: A technique that models a categorical dependent variable based on one or more independent variables

**Loss function**: A function that measures the distance between the observed values and the model’s estimated values

**M**

**MAE (Mean Absolute Error)**: The average of the absolute difference between the predicted and actual values

**Model assumptions**: Statements about the data that must be true in order to justify the use of a particular modeling technique

**MSE (Mean Squared Error)**: The average of the squared difference between the predicted and actual values

**Multiple linear regression**: A technique that estimates the relationship between one continuous dependent variable and two or more independent variables

**Multiple regression**: (Refer to **multiple linear regression**)

**N**

**Negative correlation**: An inverse relationship between two variables, where when one variable increases, the other variable tends to decrease, and vice versa

**Normality assumption**: An assumption of simple linear regression stating that the residuals are normally distributed

**No multicollinearity assumption**: An assumption of simple linear regression stating that no two independent variables (Xi and Xj) can be highly correlated with each other

**O**

**Observed values:** The existing sample of data, where each data point in the sample is represented by an observed value of the dependent variable and an observed value of the independent variable

**One hot encoding**: A data transformation technique that turns one categorical variable into several binary variables

**Ordinary least squares estimation (OLS)**: A common way to calculate linear regression coefficients

**Outcome variable (Y)**: (Refer to **dependent variable**)

**Overfitting**: When a model fits the observed or training data too specifically and is unable to generate suitable estimates for the general population

**P**

**P-value**: The probability of observing results as extreme as those observed when the null hypothesis is true

**Positive correlation**: A relationship between two variables that tend to increase or decrease together.

**Predicted values**: The estimated Y values for each X calculated by a model

**Predictor variable**: (Refer to **independent variable**)

**R**

**R**2 **(The Coefficient of Determination)**: The proportion of variance of the dependent variable, Y, explained by the independent variable or variables, X

**Regression analysis**: A group of statistical techniques that use existing data to estimate the relationships between a single dependent variable and one or more independent variables

**Regression coefficient**: The estimated betas in a regression model

**Regression models**: (Refer to **regression analysis**)

**Regularization**: A set of regression techniques that shrinks regression coefficient estimates towards zero, adding in bias, to reduce variance

**Residual**: The difference between observed or actual values and the predicted values of the regression line

**Response variable:** (Refer to **dependent variable**)

**S**

**Scatterplot matrix**: A series of scatterplots that demonstrate the relationships between pairs of variables

**Simple linear regression**: A technique that estimates the linear relationship between one independent variable, X, and one continuous dependent variable, Y

**Slope**: The amount that y increases or decreases per one-unit increase of x

**Sum of squared residuals (SSR)**: The sum of the squared difference between each observed value and its associated predicted value

**V**

**Variable selection**: The process of determining which variables or features to include in a given model

**Variance**: Refers to model flexibility and complexity, so the model learns from existing data

**Variance inflation factors (VIF)**: Quantifies how correlated each independent variable is with all of the other independent variables

**Module 4 challenge**

**1.**

Question 1

Fill in the blank: The chi-squared \_\_\_\_\_ of fit test determines whether an observed categorical variable follows an expected distribution.

goodness

**2.**

Question 2

What examines the relationship between categorical variables and continuous variables?

Analysis of variance

**3.**

Question 3

A junior data analyst at a fabric supplier works to identify the expected outcome of a new product introduction. They compare the means of one continuous dependent variable based on four groups of two categorical variables. What type of test does this scenario describe?

Two-way analysis of variance

**4.**

Question 4

The post hoc test performs a pairwise comparison between all available groups while controlling for what?

error rate

**5.**

Question 5

A data professional at an automotive manufacturer is asked to find a solution to a common manufacturing defect. They research the relationship between categorical and continuous variables to ensure all variables are relevant to the specific defect. What type of statistical technique do they use?

Analysis of covariance

**6.**

Question 6

A data professional compares how two or more continuous variables vary according to categorical independent variables. What statistical technique are they using?

Multivariate analysis of variance

**7.**

Question 7

A researcher wants to evaluate the effectiveness of different job training programs on various skill outcomes. She has two continuous dependent variables: a technical skills score and a soft skills score. Her independent variable is the training program, which can be either in-person instruction or online instruction. What type of analysis should she use?

(NOT FINAL)

MANOVA

**Module 5:**

**Binomial logistic regression linearly assumptions**

There should be a linear relationship between each X variable and the logit of the probability that Y equals to 1.

**Logit (log-odds)**

The logarithm of the odds of a given probability. So the logit of probability p is equal to the logarithm of p divided by 1 minus p.

**Maximum likelihood estimation (MLE)**

A technique for estimating the beta parameters that maximize the likelihood of the model producing the observed data.

**Likelihood**

The probability of observing the actual data, given some set of beta parameters.

**Binomial logistic regression assumptions**

* Linearity
* Independent observations

**Binomial logistic regression assumptions**

* Linearity
* Independent observations
* No multicollinearity
* No extreme outliners

# Test your knowledge: Foundations of logistic regression

**1.**

Question 1

No extreme outliers is one of the four main binomial logistic regression assumptions. What are the other three? Select all that apply.

1 / 1 point

Independent observations

Correct

The four main binomial logistic regression assumptions are: linearity, independent observations, no multicollinearity, and no extreme outliers.

No multicollinearity

Correct

The four main binomial logistic regression assumptions are: linearity, independent observations, no multicollinearity, and no extreme outliers.

Linearity

Correct

The four main binomial logistic regression assumptions are: linearity, independent observations, no multicollinearity, and no extreme outliers.

Homoscedasticity

**2.**

Question 2

Logit is the logarithm of the odds of a given probability.

1 / 1 point

True

False

Correct

Logit is the logarithm of the odds of a given probability. It is the most common link function used to linearly relate the X variables to the probability of Y.

**3.**

Question 3

Fill in the blank: The maximum likelihood estimation is a technique used for estimating the beta parameters that \_\_\_\_\_ the likelihood of a model producing the observed data.

1 / 1 point

control

reduce

maximize

balance

Correct

The maximum likelihood estimation is a technique used for estimating the beta parameters that maximize the likelihood of a model producing the observed data.

**Test your knowledge: Logistic regression with Python**

**1.**

Question 1

When building a logistic regression model, what does CLF stand for?

Classifier

**2.**

Question 2

Which package do you use to create a plot of your model to visualize its results?

Seaborn package

**Confusion matrix**

A graphical representation of how accurate a classifier is at predicting the labels for a categorical variable

**Precision**

The proportion of positive predictions that were true positives.

**Recall**

The proportion of positives the model was able to identify correctly.

**Question**

Precision measures the proportion of positive predictions that were false positives.

False

Correct

Precision measures the proportion of positive predictions that were true positives. Precision is equal to the number of true positives, divided by the sum of true positives and false positives.

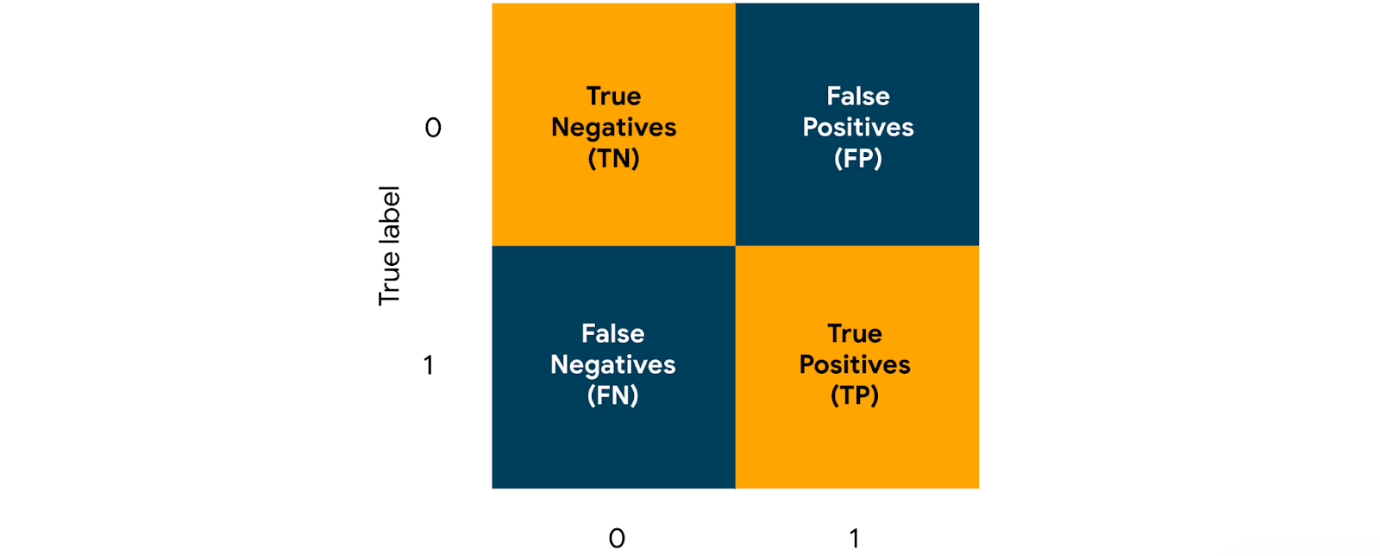
**Common logistic regression metrics in Python**

Logistic regression is a powerful technique for categorical prediction tasks in data science. Data professionals often use metrics such as precision, recall, and accuracy, as well as visualizations such as ROC curves, to gauge the performance of their logistic regression models. It is important to evaluate the performance of a model, as this shows how well the model can make predictions. The results from applying metrics can be used to report how well a model performs to relevant stakeholders.

In this reading, you will review parts of a confusion matrix and understand how to compute and visualize metrics for evaluating logistic regression through code in Python.

**Parts of a confusion matrix**

A confusion matrix helps summarize the performance of a classifier. The components of a confusion matrix are used to compute metrics for evaluating logistic regression classifiers.



The four key parts of a confusion matrix, in the context of binary classification, are the following:

1. **True negatives:**

The count of observations that a classifier correctly predicted as False (0)

2. **True positives:**

The count of observations that a classifier correctly predicted as True (1)

3. **False positives:**

The count of observations that a classifier incorrectly predicted as True (1)

4. **False negatives:**

The count of observations that a classifier incorrectly predicted as False (0)

These counts are useful in computing metrics such as precision, recall, accuracy, and ROC for evaluating logistic regression classifiers.

**Precision**

One of the major metrics for evaluating a logistic regression classifier is **precision**. Precision measures the proportion of data points predicted as True that are actually True. Imagine that you have built a logistic regression classifier for email spam detection, trained this classifier on a relevant dataset, and used this classifier to generate predictions for a set of emails. The predictions consist of True and False values. True represents an email predicted as spam, and False represents an email predicted as not spam. The precision for this classifier would convey the proportion of emails that are actually spam, out of all the emails that have been predicted as spam.

The formula for precision is as follows:

Precision=True PositivesTrue Positives + False Positives*Precision*=*True* *Positives* + *False* *PositivesTrue* *Positives*​

To compute precision in Python, you can use the **precision\_score()** function from the **metrics** module in the **sklearn** library. You can start with the following import statement.

**import sklearn.metrics as metrics**

The **precision\_score()** function takes in true values and predicted values as arguments and returns the precision score. Assume that **y\_test** and **y\_pred** are variables that contain true values and predicted values respectively. You can use the following code to compute the precision.

**metrics.precision\_score(y\_test,y\_pred)**

In the context of email spam detection, if precision\_score() returns 0.91, that means 91% of the emails predicted as spam are indeed spam.

**Recall**

Another major metric for evaluating a logistic regression classifier is **recall**. Recall measures the proportion of data points that are predicted as True, out of all the data points that are actually True. Imagine that you have built a logistic regression classifier for fraud detection and generated predictions. In the predictions, True represents a credit card transaction predicted as fraudulent, and False represents a credit card transaction predicted as not fraudulent. The recall for this classifier would convey the proportion of fraudulent credit card transactions that the classifier correctly identified as such.

The formula for recall is as follows:

Recall=True PositivesTrue Positives + False Negatives*Recall*=*True* *Positives* + *False* *NegativesTrue* *Positives*​

To compute recall in Python, you can use the **recall\_score()** function from the metrics module. The function takes in true values and predicted values as arguments and returns the recall score. You can use the following code to compute the recall.

**metrics.recall\_score(y\_test,y\_pred)**

In the context of fraud detection among credit card transactions, if the **recall\_score()** function returns 0.87, that means 87% of the fraudulent credit card transactions are correctly detected as fraudulent.

**Accuracy**

Another important metric for evaluating logistic regression is **accuracy**. Accuracy measures the proportion of data points that are correctly classified. Imagine that you have built a logistic regression classifier for loan approval prediction. In the predictions, True represents a prediction that the loan will be approved, and False represents a prediction that the loan will not be approved. The accuracy score for this classifier would convey the proportion of loans that have been correctly classified.

The formula for accuracy is as follows:

Accuracy=True Positives+True NegativesTotal Predictions*Accuracy*=*Total* *PredictionsTrue* *Positives*+*True* *Negatives*​

To compute accuracy in Python, you can use the **accuracy\_score()** function from the metrics module. The function takes in true values and predicted values as arguments and returns the accuracy score. You can use the following code to compute the accuracy.

**metrics.accuracy\_score(y\_test,y\_pred)**

In the context of loan approval prediction, if the **accuracy\_score()** function returns 0.90, that means 90% of the loans are correctly predicted, with the predictions being either will be approved or will not be approved.

**ROC curves**

An **ROC curve** helps in visualizing the performance of a logistic regression classifier. ROC curve stands for receiver operating characteristic curve. To visualize the performance of a classifier at different classification thresholds, you can graph an ROC curve. In the context of binary classification, a classification threshold is a cutoff for differentiating the positive class from the negative class.

An ROC curve plots two key concepts

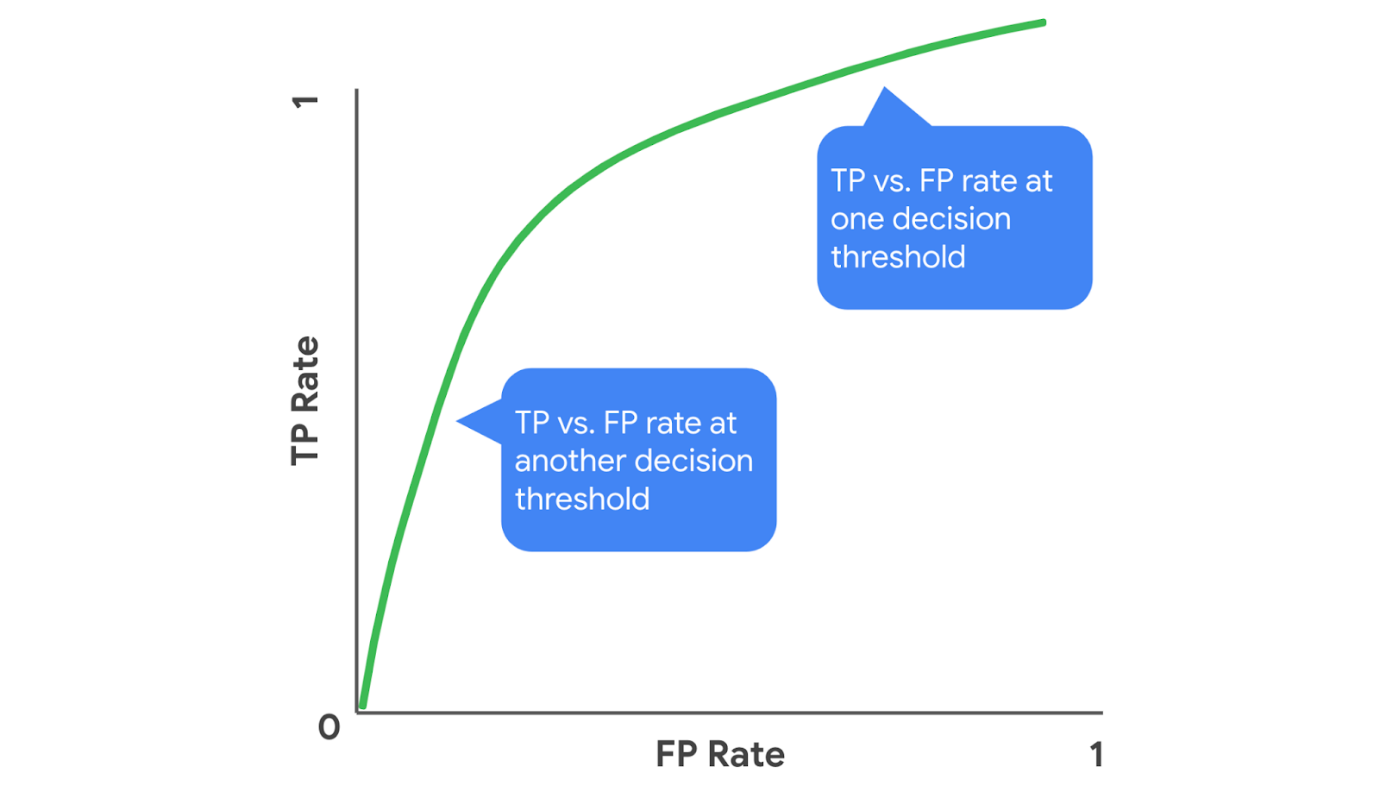
1. **True Positive Rate:** equivalent to **Recall.** The formula for True Positive Rate is as follows:

True Positive Rate=True PositivesTrue Positives + False Negatives*True* *Positive* *Rate*=*True* *Positives* + *False* *NegativesTrue* *Positives*​

2. **False Positive Rate:**The ratio between the False Positives and the total count of observations that should be predicted as False. The formula for False Positive Rate is as follows:

False Positive Rate=False PositivesFalse Positives + True Negatives*False* *Positive* *Rate*=*False* *Positives* + *True* *NegativesFalse* *Positives*​

For each point on the curve, the x and y coordinates represent the False Positive Rate and the True Positive Rate respectively at the corresponding threshold.

An example of an ROC curve. For each point on the curve, the x and y coordinates represent the False Positive Rate and the True Positive Rate respectively at the corresponding threshold.

You can examine an ROC curve to observe how the False Positive Rate and True Positive Rate change together over the different thresholds. In the ROC curve for an ideal model, there would exist a threshold at which the True Positive Rate is high and the False Positive Rate is low. The more that the ROC curve hugs the top left corner of the plot, the better the model does at classifying the data.

You can use the following steps to graph an ROC curve in Python.

Start by importing the necessary modules as follows.

**import matplotlib.pyplot as plt**

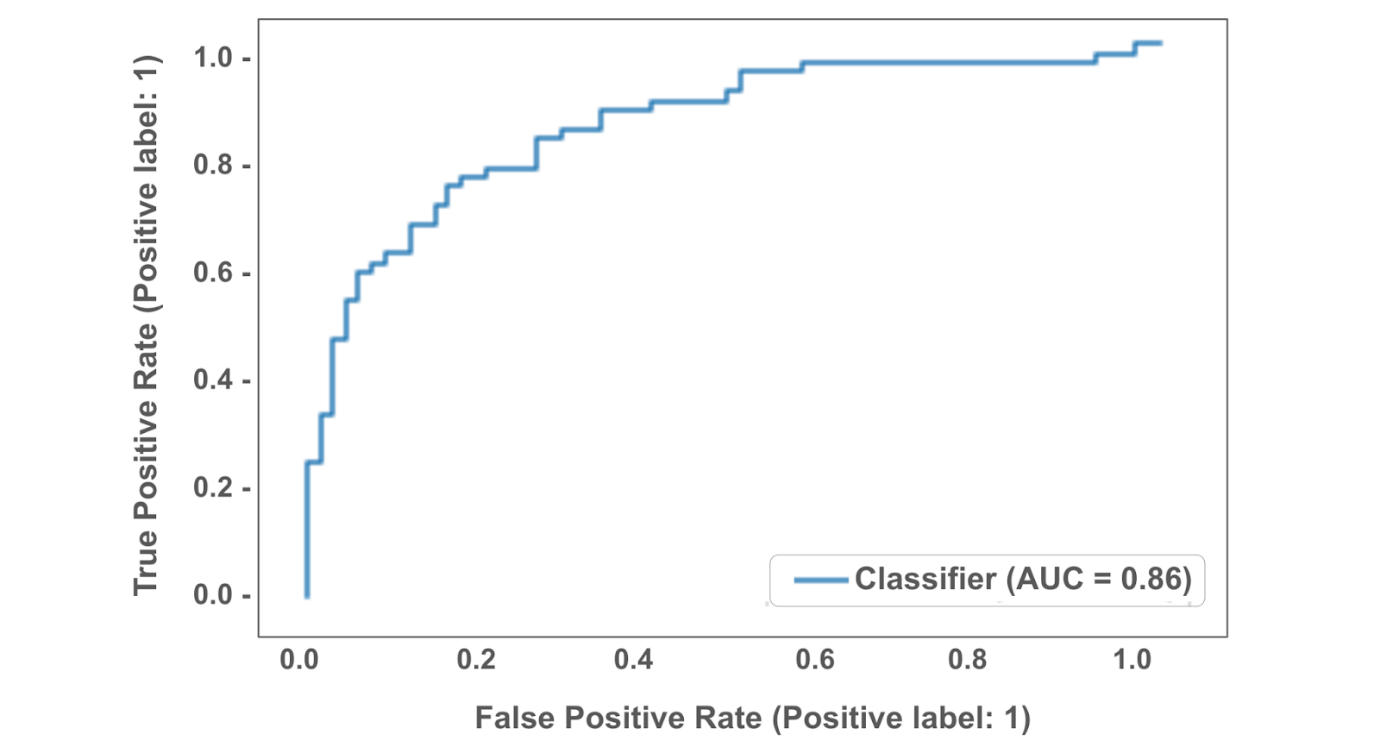
**from sklearn.metrics import RocCurveDisplay**

Then use the following code to plot the ROC curve.

**RocCurveDisplay.from\_predictions(y\_test, y\_pred)**

**plt.show()**

Using these steps to generate an ROC curve could result in a graph.

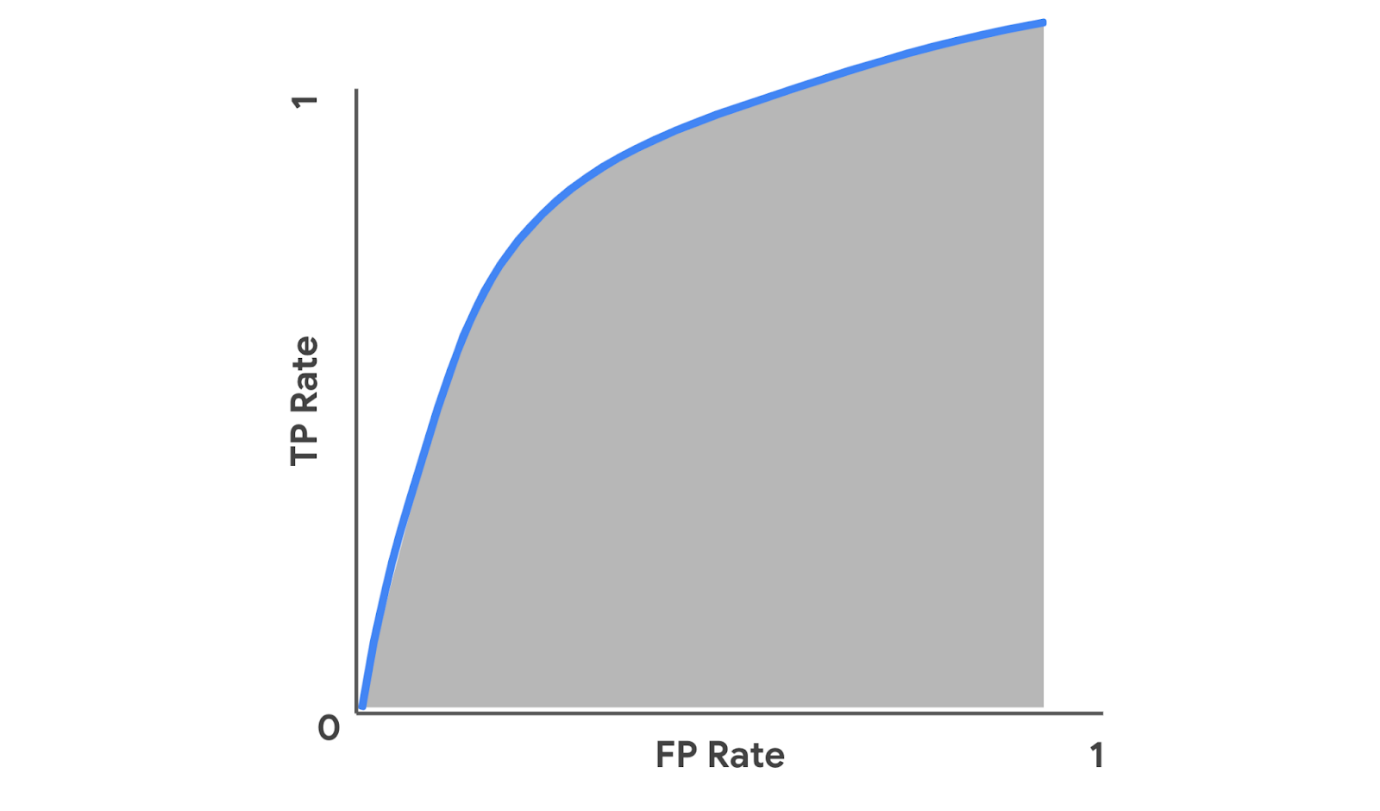


In this graph, the ROC curve indicates that the corresponding classifier performs decently well.

**AUC**

**AUC** stands for area under the ROC curve. AUC provides an aggregate measure of performance across all possible classification thresholds. AUC ranges in value from 0.0 to 1.0. A model whose predictions are 100% wrong has an AUC of 0.0, and a model whose predictions are 100% correct has an AUC of 1.0. An AUC smaller than 0.5 indicates that the model performs worse than a random classifier (i.e. a classifier that randomly assigns each example to True or False), and an AUC larger than 0.5 indicates that the model performs better than a random classifier.

In the following visualization, AUC is the area of the shaded region.



To compute AUC in Python, you can use the roc\_auc\_score() function from the metrics module. The function takes in true values and predicted values as arguments and returns the accuracy score. You can use the following code to compute the AUC.

**metrics.roc\_auc\_score(y\_test,y\_pred)**

For example, in the context of a logistic regression classifier for email spam detection, if the **roc\_auc\_score()** function returns 0.99, that means 99% of the classifier’s predictions are correct across all classification thresholds.

**Key takeaways**

* Precision, Recall, and Accuracy are common metrics for evaluating the performance of a logistic regression classifier.
* You can use functions from the metrics module in the sklearn library to compute these metrics in Python.
* Graphing an ROC curve helps visualize how a classifier performs across different classification thresholds.
* Computing AUC helps aggregate a classifier’s performance across thresholds into one measure.

**Resources for more information**

* [precision\_score](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html): Documentation on the precision\_score() function
* [recall\_score](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html?highlight=recall_score#sklearn.metrics.recall_score): Documentation on the recall\_score() function
* [accuracy\_score](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html?highlight=accuracy_score#sklearn.metrics.accuracy_score): Documentation on the accuracy\_score() function
* [RocCurveDisplay.from\_predictions](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.RocCurveDisplay.html#sklearn.metrics.RocCurveDisplay.from_predictions): Documentation on the RocCurveDisplay.from\_predictions() function
* [roc\_auc\_score](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html#sklearn.metrics.roc_auc_score): Documentation on the roc\_auc\_score() function

**Interpret logistic regression models**

Interpreting a logistic regression model involves examining coefficients and computing metrics. After you fit your logistic regression model to training data, you can access the coefficient estimates from the model using code in Python. You can then use those values to understand how the model makes predictions. This reading will show you an example of how to interpret coefficients from a logistic regression model, as well as things to consider when choosing metrics for model evaluation.

**Coefficients from the model**

To understand how a logistic regression model works, it is important to start with the equation that describes the relationship between the variables. That equation is also called the logit function.

**The logit function**

When the logit function is written in terms of the independent variables, it conveys the following: there is a linear relationship between each independent variable, X*X*, and the logit of the probability that the dependent variable, Y*Y*, equals 1. The logit of that probability is the logarithm of the odds of that probability.

The equation for the logit function in binomial logistic regression is shown below. This involves the probability that Y*Y*equals 1, because 1 is the typical outcome of interest in binary classification, where the possible values of Y*Y* are 1 and 0.

logit(p)=log(p1 −p)=β0+β1X1    where p=P(Y=1) *logit*(*p*)=*log*(1 −*pp*​)=*β*0​+*β*1​*X*1​    *where* *p*=*P*(*Y*=1)

**Interpret coefficients**

Imagine you have built a binomial logistic regression model for predicting emails as spam or non-spam. The dependent variable, Y*Y*, is whether an email is spam (1) or non-spam (0). The independent variable, X1*X*1​, is the message length.  Assume that **clf** is the classifier you fitted to training data.

You can use the following code to access the coefficient β1*β*1​estimated by the model:

**clf.coef\_**

If the estimated β1*β*1​ is 0.186, for example, that means a one-unit increase in message length is associated with a 0.186 increase in the log odds of p*p*. To interpret change in odds of Y as a percentage, you can exponentiate β1*β*1​, as follows.

eβ1=e0.186≈1.204*eβ*1​=*e*0.186≈1.204

So, for every one-unit increase in message length, you can expect that the odds the email is spam increases by 1.204, or 20.4%.

**Things to consider when choosing metrics**

The next important step after examining the coefficients from a logistic regression model is evaluating the model through metrics. The most commonly used metrics include precision, recall, and accuracy. The following sections describe things to keep in mind when choosing between these.

**When to use precision**

Using precision as an evaluation metric is especially helpful in contexts where the cost of a false positive is quite high and much higher than the cost of a false negative. For example, in the context of email spam detection, a false positive (predicting a non-spam email as spam) would be more costly than a false negative (predicting a spam email as non-spam). A non-spam email that is misclassified could contain important information, such as project status updates from a vendor to a client or assignment deadline announcements from an instructor to a class of students.

**When to use recall**

Using recall as an evaluation metric is especially helpful in contexts where the cost of a false negative is quite high and much higher than the cost of a false positive. For example, in the context of fraud detection among credit card transactions, a false negative (predicting a fraudulent credit card charge as non-fraudulent) would be more costly than a false positive (predicting a non-fraudulent credit card charge as fraudulent). A fraudulent credit card charge that is misclassified could lead to the customer losing money, undetected.

**When to use accuracy**

It is helpful to use accuracy as an evaluation metric when you specifically want to know how much of the data at hand has been correctly categorized by the classifier. Another scenario to consider: accuracy is an appropriate metric to use when the data is balanced, in other words, when the data has a roughly equal number of positive examples and negative examples. Otherwise, accuracy can be biased. For example, imagine that 95% of a dataset contains positive examples, and the remaining 5% contains negative examples. Then you train a logistic regression classifier on this data and use this classifier predict on this data. If you get an accuracy of 95%, that does not necessarily indicate that this classifier is effective. Since there is a much larger proportion of positive examples than negative examples, the classifier may be biased towards the majority class (positive) and thus the accuracy metric in this context may not be meaningful. When the data you are working with is imbalanced, consider either transforming it to be balanced or using a different evaluation metric other than accuracy.

**Key takeaways**

* Examine the beta coefficients from a model to understand how the model predicts the dependent variable.
* When determining which metrics are meaningful for evaluating a logistic regression classifier, consider the context of the data involved, how the predictions will be used, and how impactful False Positives versus False Negatives are in that context.

**Resources for more information**

* [LogisticRegression](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html): Documentation for implementing Logistic Regression models using **sklearn** and accessing intercept and coefficients from a model

**Test your knowledge: Interpret logistic regression results**

**1.**

Question 1

The confusion matrix is a graphical representation of how accurate a classifier is at predicting what for a categorical variable?

1 / 1 point

Errors

Validity

Labels

Precision

Correct

The confusion matrix is a graphical representation of how accurate a classifier is at predicting the labels for a categorical variable. It displays how many data points were accurately categorized by the classifier for each category. The other squares in a grid convey how many data points were misclassified.

**2.**

Question 2

Fill in the blank: \_\_\_\_\_ measures the proportion of positive predictions that were true positives.

1 / 1 point

Precision

Accuracy

Recall

Validity

Correct

Precision measures the proportion of positive predictions that were true positives.

**3.**

Question 3

Which of the following provide additional information about the likelihood of a result being merely by chance? Select all that apply.

0.5 / 1 point

Confidence intervals

Maximum likelihood estimation

This should not be selected

The p-value and confidence intervals provide additional information about the likelihood of a result being merely by chance.

Logit

P-value

Correct

The p-value and confidence intervals provide additional information about the likelihood of a result being merely by chance.

**Linear regression**

* Accessible interpretation
* Explain which factors impact the outcome variables
* Check model assumptions

**Hypothesis testing**

* Outcome variable is continuous
* Focus on comparing different groups

**Evaluating logistic regression**

* P-value
* Confusion matrices
* Precision
* Recall
* Accuracy
* ROC/AUC
* AIC
* BIC

Logistic regression coefficients report in percentages how much a factor increases or decreases the **likelihood of an outcome.**

**Prediction with different types of regression**

As you have been learning, key regression techniques that you will encounter in your work as a data professional include linear regression, hypothesis testing, and logistic regression. When your goal is to make predictions with data, it is important to consider these different approaches and think about which approach will best help you achieve your task. In this reading, you will learn more about how to choose the most relevant regression technique for a project, based on the question you want to answer, the outcome variable, and how it is measured.

**How to choose a regression technique**

When choosing a regression technique, it is important to consider the data you are working with and the question you want to address.

**Things to consider**

1. What is the question you want to answer? In other words, what do you want to predict?
2. Which variable in your data can be the outcome variable?
3. How is the outcome variable measured? If the outcome variable is continuous, it is more likely that either linear regression or hypothesis testing will be most appropriate. However, if the outcome variable is binary, you will find logistic regression to be more useful.

**Example contexts for regression**

The following examples demonstrate how the questions about prediction, outcome variable, and measurement can be navigated in order to choose a regression technique.

**Example context: User engagement**

In your work as a data professional, imagine that you are interested in making predictions about user engagement for a mobile app.

First, you might ask, what is the question you want to answer?

One possible question could be “How much does each in-app feature influence user engagement?” The in-app features might include a live chat with customer support, an FAQ section that updates weekly, and a community space to connect with other users. Next, you might ask, which variable in your data can be the outcome variable? If you have access to data about users’ session lengths (in other words, how long users spend in the app each time they open it), the outcome variable can be session length. Your next question might be: how is the outcome variable measured? Session length can be measured by number of minutes, which is continuous. Because the outcome variable is continuous, and you are interested in how much each feature influences the outcome variable, you could proceed with linear regression and check the relevant model assumptions. If there is only one feature of interest, you would build a simple linear regression model. If there are multiple features of interest, you would build a multiple linear regression model.

Another question of interest could be “Does a dynamic landing page versus a static landing page make a difference in user engagement?” The outcome variable can be session length, measured by number of minutes, for this example, too. Since the outcome variable is continuous, and the target question is about whether there is a difference in user engagement when one type of landing page is used over the other, you could proceed with hypothesis testing. You can then frame the hypotheses, which could be the following:

* Null hypothesis (H0): Users spend approximately the same amount of time in the app when the landing page is dynamic versus when it is static.
* Alternative hypothesis (H1): Users do NOT spend approximately the same amount of time in the app when the landing page is dynamic versus when it is static.

Another question you might be interested in is “Will a user engage with the new line of products in-app?” Next, you might ask, which variable in your data can be the outcome variable? If you have access to data about whether a user clicks to view the new line of products, that could be the outcome variable. The next question is: how is the outcome variable measured? Whether a user clicks to view that content can be represented as a binary variable, with 1 indicating they clicked to view the content and 0 indicating that they did not click to view that content. Since this outcome variable is binary, you could proceed with binomial logistic regression.

**Example context: Patient response**

Now imagine that you are tasked with making predictions about patient responses to medical treatments.

You can start by asking, what is the question you want to answer?

A possible question could be “How much does each factor influence a patient’s response to a medical treatment?” If the goal of the treatment is to improve white blood cell (WBC) count and you have access to that data, WBC count can be the outcome variable. The outcome variable is a continuous measure, and you could use linear regression to address this task.

Another question of interest could be “Will Treatment A, Treatment B, or Treatment C have a stronger impact on a patient’s WBC count?” The outcome variable in this case would also be WBC count, which is continuous. Since the target question is about comparing different treatments, it would be best to proceed with hypothesis testing. You can then form the hypotheses, which could be the following.

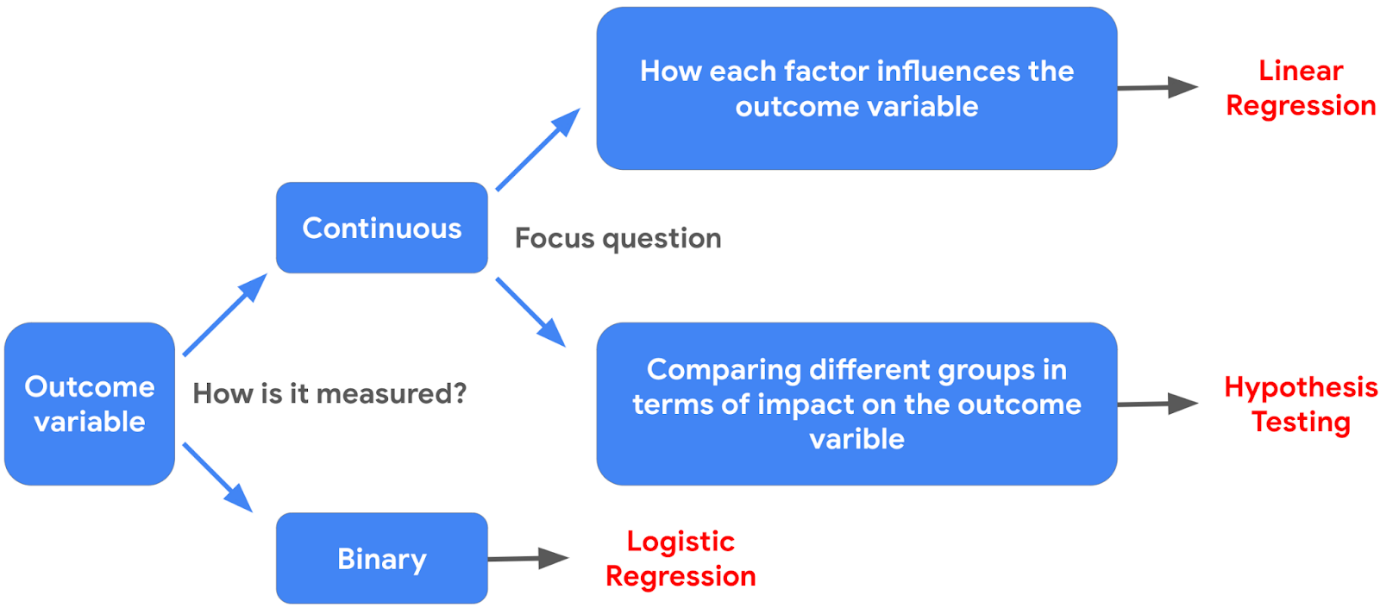
* Null hypothesis (H0): Patients have approximately the same white blood cell count with each treatment.
* Alternative hypothesis (H1): Patients do NOT have approximately the same white blood cell count with each treatment.

A different question you might be interested in: “With Treatment A, will a patient’s WBC count reach the ideal range?” If you have access to the associated data, the outcome variable would be whether a patient’s WBC count reaches the ideal range or not, which is a binary variable: 1 indicating that their WBC count falls within the ideal range and 0 indicating that it does not. You could build a logistic regression model to make predictions in this scenario.

**Key takeaways**

* Consider the question you want to answer and the data you have access to when choosing a regression technique for making predictions.
* Identifying the outcome variable of interest and how it is measured will help you decide which regression technique is most suitable for your task.

The following flowchart captures a high-level approach for choosing a regression technique, starting from the outcome variable, as discussed in this reading. Also note that hypothesis testing is connected to regression analysis. For example, in linear regression, the process of testing whether there is a correlation between two variables (in other words, determining if the coefficients are statistically significant in the linear model) involves a hypothesis test.



**Resources for more information**

* If you want to learn more about different types of regression models, you can check out [this article](https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/) about different types of regression models, covering linear regression, logistic regression, and more.
* If you want to learn more about hypothesis testing, you can check out [this article](https://towardsdatascience.com/hypothesis-testing-for-data-scientists-everything-you-need-to-know-8c36ddde4cd2) that provides an overview of the key steps for approaching hypothesis testing in data science.

**Test your knowledge: Compare regression models**

**1.**

Question 1

Which model might a data professional consider first if the outcome variable is binary?

1 / 1 point

Single linear regression

Hypothesis testing

Binomial logistic regression

Multiple linear regression

Correct

If the outcome variable is binary, a data professional might consider a binomial logistic regression model. After building the model, the best way to test if logistic regression is the right choice is to evaluate the model with metrics.

**2.**

Question 2

A data professional can use recall to evaluate a logistic regression model. What other metrics can be used to meet this goal? Select all that apply.

0.75 / 1 point

Precision

Correct

To evaluate a logistic regression model, a data professional can use recall, p-value, confusion matrices, and precision

R squared

Confusion matrices

P-value

Correct

To evaluate a logistic regression model, a data professional can use recall, p-value, confusion matrices, and precision.

You didn’t select all the correct answers

**Glossary terms from module 5**

**Terms and definitions from Course 5, Module 5**

**Accuracy**: Refers to the proportion of data points that were correctly categorized

**Binomial logistic regression**: A technique that models the probability of an observation falling into one of two categories, based on one or more independent variables

**Binomial logistic regression linearity assumption**: An assumption stating that there should be a linear relationship between each X variable and the logit of the probability that Y equals one

**Confusion matrix**: A graphical representation of how accurate a classifier is at predicting the labels for a categorical variable

**Likelihood**: The probability of observing the actual data, given some set of beta parameters

**Logistic regression**: A technique that models a categorical dependent variable (Y) based on one or more independent variables (X)

**Log-odds function**: (Refer to **logit**)

**Logit**: The logarithm of the odds of a given probability

**Maximum Likelihood Estimation (MLE)**: A technique for estimating the beta parameters that maximize the likelihood of the model producing the observed data

**Precision**: The proportion of positive predictions that were true positives

**Recall**: The proportion of positives the model was able to identify correctly

**Terms and definitions from previous modules**

**A**

**Absolute values**: (Refer to **observed values**)

**Adjusted R**2: A variation of R2 that accounts for having multiple independent variables present in a linear regression model

**Analysis of Variance (ANOVA)**: A group of statistical techniques that test the difference of means between three or more groups

**ANCOVA (Analysis of Covariance)**: A statistical technique that tests the difference of means between three or more groups while controlling for the effects of covariates, or variable(s) irrelevant to the test

**B**

**Backward elimination**: A stepwise variable selection process that begins with the full model, with all possible independent variables, and removes the independent variable that adds the least explanatory power to the model

**Best fit line**: The line that fits the data best by minimizing some loss function or error

**Bias**: Refers to simplifying the model predictions by making assumptions about the variable relationships

**Bias-variance trade-off**: Balance between two model qualities, bias and variance, to minimize overall error for unobserved data

**C**

**Causation**: Describes a cause-and-effect relationship where one variable directly causes the other to change in a particular way

**Chi-squared (χ²) Goodness of Fit Test**: A hypothesis test that determines whether an observed categorical variable follows an expected distribution

**Chi-squared (χ²) Test for Independence**: A hypothesis test that determines whether or not two categorical variables are associated with each other

**Confidence band**: The area surrounding a line that describes the uncertainty around the predicted outcome at every value of X

**Confidence interval**: A range of values that describes the uncertainty surrounding an estimate

**Correlation**: Measures the way two variables tend to change together

**D**

**Dependent variable (Y)**: The variable a given model estimates

**E**

**Errors**: In a regression model, the natural noise assumed to be in a model

**Explanatory variable**: (Refer to **independent variable**)

**Extra Sum of Squares F-test**: Quantifies the difference between the amount of variance that is left unexplained by a reduced model that is explained by the full model

**F**

**Feature selection**: (Refer to **variable selection**)

**Forward selection**: A stepwise variable selection process that begins with the null mode—with 0 independent variables—that considers all possible variables to add; it incorporates the independent variable that contributes the most explanatory power to the model

**H**

**Hold-out sample**: A random sample of observed data that is not used to fit the model

**Homoscedasticity assumption**: An assumption of simple linear regression stating that the variation of the residuals (errors) is constant or similar across the model

**Hypothesis testing**: A statistical procedure that uses sample data to evaluate an assumption about a population parameter

**I**

**Independent observation assumption**: An assumption of simple linear regression stating that each observation in the dataset is independent

**Independent variable (X)**: The variable whose trends are associated with the dependent variable

**Interaction term**: Represents how the relationship between two independent variables is associated with changes in the mean of the dependent variable

**Intercept (constant 𝐵**0**)**: The y value of the point on the regression line where it intersects with the y-axis

**L**

**Line**: A collection of an infinite number of points extending in two opposite directions

**Linearity assumption**: An assumption of simple linear regression stating that each predictor variable (Xi) is linearly related to the outcome variable (Y)

**Linear regression**: A technique that estimates the linear relationship between a continuous dependent variable and one or more independent variables

**Link function**: A nonlinear function that connects or links the dependent variable to the independent variables mathematically

**Logistic regression**: A technique that models a categorical dependent variable based on one or more independent variables

**Loss function**: A function that measures the distance between the observed values and the model’s estimated values

**M**

**MAE (Mean Absolute Error)**: The average of the absolute difference between the predicted and actual values

**MANCOVA (Multivariate Analysis of Covariance)**: An extension of ANCOVA and MANOVA that compares how two or more continuous outcome variables vary according to categorical independent variables, while controlling for covariates

**MANOVA (Multivariate Analysis of Variance)**: An extension of ANOVA that compares how two or more continuous outcome variables vary according to categorical independent variables

**Model assumptions**: Statements about the data that must be true in order to justify the use of a particular modeling technique

**MSE (Mean Squared Error)**: The average of the squared difference between the predicted and actual values

**Multiple linear regression**: A technique that estimates the relationship between one continuous dependent variable and two or more independent variables

**Multiple regression**: (Refer to **multiple linear regression**)

**N**

**Negative correlation**: An inverse relationship between two variables, where when one variable increases, the other variable tends to decrease, and vice versa

**Normality assumption**: An assumption of simple linear regression stating that the residuals are normally distributed

**No multicollinearity assumption**: An assumption of simple linear regression stating that no two independent variables (Xi and Xj) can be highly correlated with each other

**O**

**Observed values:** The existing sample of data, where each data point in the sample is represented by an observed value of the dependent variable and an observed value of the independent variable

**One hot encoding**: A data transformation technique that turns one categorical variable into several binary variables

**One-Way ANOVA**: A type of statistical testing that compares the means of one continuous dependent variable based on three or more groups of one categorical variable

**Ordinary least squares estimation (OLS)**: A common way to calculate linear regression coefficients

**Outcome variable (Y)**: (Refer to **dependent variable**)

**Overfitting**: When a model fits the observed or training data too specifically and is unable to generate suitable estimates for the general population

**P**

**P-value**: The probability of observing results as extreme as those observed when the null hypothesis is true

**Positive correlation**: A relationship between two variables that tend to increase or decrease together.

**Post hoc test**: An ANOVA test that performs a pairwise comparison between all available groups while controlling for the error rate

**Predicted values**: The estimated Y values for each X calculated by a model

**Predictor variable**: (Refer to **independent variable**)

**R**

**R**2 **(The Coefficient of Determination)**: Measures the proportion of variation in the dependent variable, Y, explained by the independent variable(s), X

**Regression analysis**: A group of statistical techniques that use existing data to estimate the relationships between a single dependent variable and one or more independent variables

**Regression coefficient**: The estimated betas in a regression model

**Regression models**: (Refer to **regression analysis**)

**Regularization**: A set of regression techniques that shrinks regression coefficient estimates towards zero, adding in bias, to reduce variance

**Residual**: The difference between observed or actual values and the predicted values of the regression line

**Response variable:** (Refer to **dependent variable**)

**S**

**Scatterplot matrix**: A series of scatterplots that demonstrate the relationships between pairs of variables

**Simple linear regression**: A technique that estimates the linear relationship between one independent variable, X, and one continuous dependent variable, Y

**Slope**: The amount that y increases or decreases per one-unit increase of x

**Sum of squared residuals (SSR)**: The sum of the squared difference between each observed value and its associated predicted value

**T**

**Two-Way ANOVA**: A type of statistical testing that compares the means of one continuous dependent variable based on three or more groups of two categorical variables

**V**

**Variable selection**: The process of determining which variables or features to include in a given model

**Variance**: Refers to model flexibility and complexity, so the model learns from existing data

**Variance inflation factors (VIF)**: Quantifies how correlated each independent variable is with all of the other independent variables

**Module 5 challenge**

**1.**

Question 1

Fill in the blank: Binomial logistic regression is a technique that models the \_\_\_\_\_ of an observation falling into one of two categories, based on one or more independent variables.

1 / 1 point

probability

**2.**

Question 2

Logit includes which other probability formula?

1 / 1 point

Odds

**3.**

Question 3

What technique estimates the beta parameters that increase the likelihood of the model producing observed data?

Maximum likelihood estimation

**4.**

Question 4

Which regression assumption states that, if multiple X variables are in a model, they should not be highly correlated with one another?

No multicollinearity

**5.**

Question 5

What graphical representation demonstrates a classifier’s accuracy at predicting the labels for a categorical variable?

1 / 1 point

Confusion matrix

**6.**

Question 6

A data professional calculates precision in logistic regression results. They have 89 true positives, 83 true negatives, 3 false positives, and 1 false negative. What is the calculation for precision?

89 / (89 + 3)

**7.**

Question 7

A data professional calculates accuracy in logistic regression results. They have 99 true positives, 91 true negatives, and 248 total predictions. What is the calculation for accuracy?

(99 + 91) / 248

**8.**

Question 8

A data professional calculates recall in logistic regression results. They have 91 true positives, 84 true negatives, 6 false positives, and 5 false negatives. What is the calculation for recall?

91 / (91 + 5)

**Module 6:**

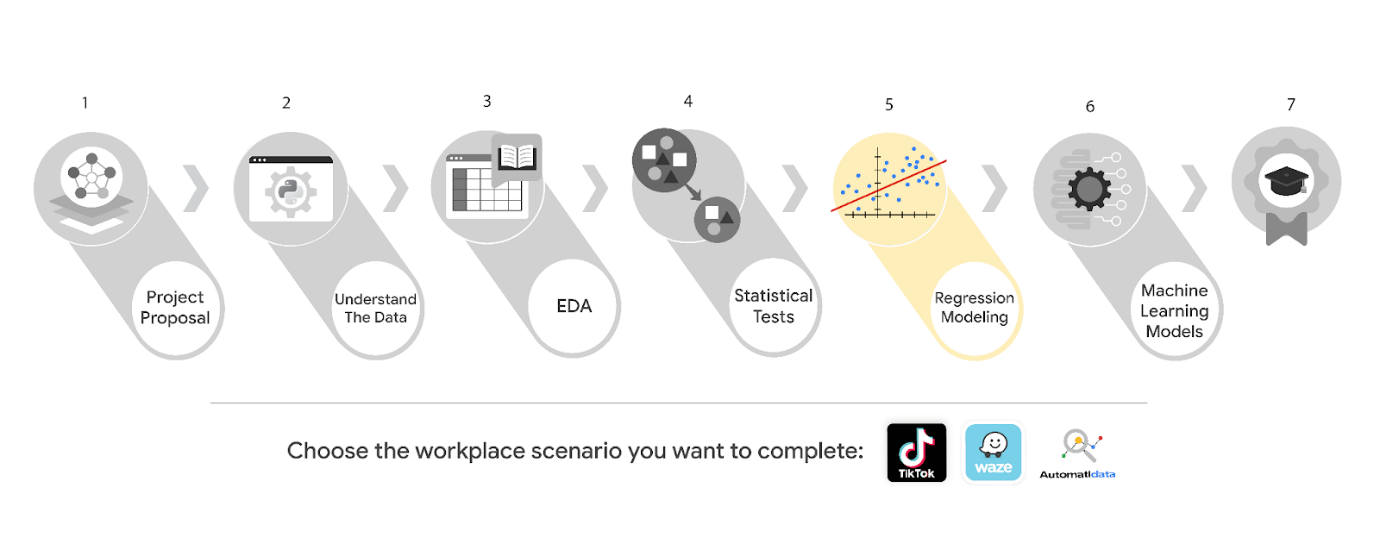
**Explore your Course 5 workplace scenarios**

**Overview**

This certificate offers you a choice of several different workplace scenarios to use when completing each end-of-course project:

* Automatidata, featuring a fictional data consulting firm
* TikTok, created in partnership with the short-form video hosting company
* Waze, created in partnership with the realtime driving directions app

Each scenario offers you an opportunity to apply your skills and create work samples to share when applying for jobs; so, you will be practicing similar skills regardless of the workplace scenario. It is recommended that you work with the same scenario for each end-of-course project to have a cohesive experience. However, you are welcome to investigate any of the workplace scenarios you are interested in as you progress through the program.



***Reminder:*** *We recommend that you choose one workplace scenario to follow for all end-of-course projects to ensure end-to-end project development.*

The minimum requirement to earn your Advanced Data Analytics Certificate is to complete the end-of-course project, using one workplace scenario, for each course. You may complete the project for as many of the workplace scenarios as you wish. Completing the project for more than one workplace scenario in a single course offers you additional practice and work examples you can add to your portfolio and share with prospective employers during your job search.

This reading offers an overview of all available workplace scenarios. Before moving on, identify the scenario you would like to complete for the Course 5 end-of-course project.

**Course 5 workplace scenarios**

**Automatidata**



**Project goal:**

In this fictional scenario, the New York City Taxi and Limousine Commission (TLC) has approached the data consulting firm Automatidata to develop an app that enables TLC riders to estimate the taxi fares in advance of their ride.

**Background:**

Since 1971, TLC has been regulating and overseeing the licensing of New York City's taxi cabs, for-hire vehicles, commuter vans, and paratransit vehicles.

**Scenario:**

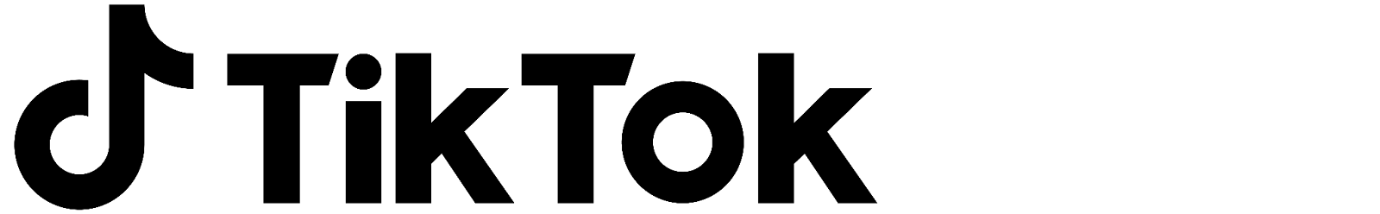
The relationship between fare amounts and payment type has been analyzed. The operations manager with New York City TLC is seeking more insight through regression modeling. The team’s next milestone is to run a regression model for taxi fares based on variables in the dataset.

**Course 5 tasks:**

* Compute descriptive statistics
* Create a regression model from the New York City TLC dataset
* Create an executive summary for the Automatidata data team before sharing the results with the client

***Note:*** *The story, all names, characters, and incidents portrayed in this project are fictitious. No identification with actual persons (living or deceased) is intended or should be inferred. And, the data shared in this project has been created for pedagogical purposes.*

**TikTok**



**Project goal:**

The TikTok data team is developing a machine learning model for classifying claims made in videos submitted to the platform.

**Background:**

TikTok is the leading destination for short-form mobile video. The platform is built to help imaginations thrive. TikTok's mission is to create a place for inclusive, joyful, and authentic content–where people can safely discover, create, and connect.

**Scenario:**

The data team at TikTok is close to their goal of building a model to assist in the classification of claims in videos. The next step is to use the project data to create a regression model. As a member of TikTok’s data team, you'll determine the type of regression model that is needed and develop one using TikTok's claim classification data.

**Course 5 tasks:**

* Import relevant packages and TikTok data
* Exploratory data analysis and check model assumptions
* Determine the correct modeling approach
* Build the regression model
* Finish checking model assumptions
* Evaluate the model
* Interpret model results and summarize findings for cross-departmental stakeholders within TikTok

***Note:*** *The story, all names, characters, and incidents portrayed in this project are fictitious. No identification with actual persons (living or deceased) is intended or should be inferred. And, the data shared in this project has been created for pedagogical purposes.*

**Waze**



**Project goal:**

Waze leadership has asked your data team to develop a machine learning model to predict user churn. Churn quantifies the number of users who have uninstalled the Waze app or stopped using the app. This project focuses on monthly user churn. An accurate model will help prevent churn, improve user retention, and grow Waze’s business.

**Background:**

Waze’s free navigation app makes it easier for drivers around the world to get to where they want to go. Waze’s community of map editors, beta testers, translators, partners, and users helps make each drive better and safer.

**Scenario:**

Your team is more than halfway through their user churn project. Earlier you completed a project proposal, used Python to analyze and visualize Waze’s user data, and conducted a hypothesis test. As a next step, leadership asks your team to build a regression model to predict user churn based on a variety of variables.

**Course 5 tasks:**

* Check model assumptions
* Build a binomial logistic regression model
* Evaluate the model
* Share an executive summary with the Waze leadership team

***Note:*** *The story, all names, characters, and incidents portrayed in this project are fictitious. No identification with actual persons (living or deceased) is intended or should be inferred. And, the data shared in this project has been created for pedagogical purposes.*

**Key Takeaways**

In Course 5, Regression Analysis: Simplify Complex Data Relationships, you practiced modeling variable relationships, and investigated linear and logistic regression to better understand data modeling. Additionally, you reviewed model assumptions and evaluation techniques that will help you interpret and articulate relationships in datasets.

**Course 5 skills:**

* Conduct statistical analysis
* Conduct regression modeling
* Create predictive models
* Expand Python coding
* Share Insights and Ideas with stakeholders

**Course 5 end-of-course project:**

* Regression model within a Python notebook
* Executive summary with results of model and insights

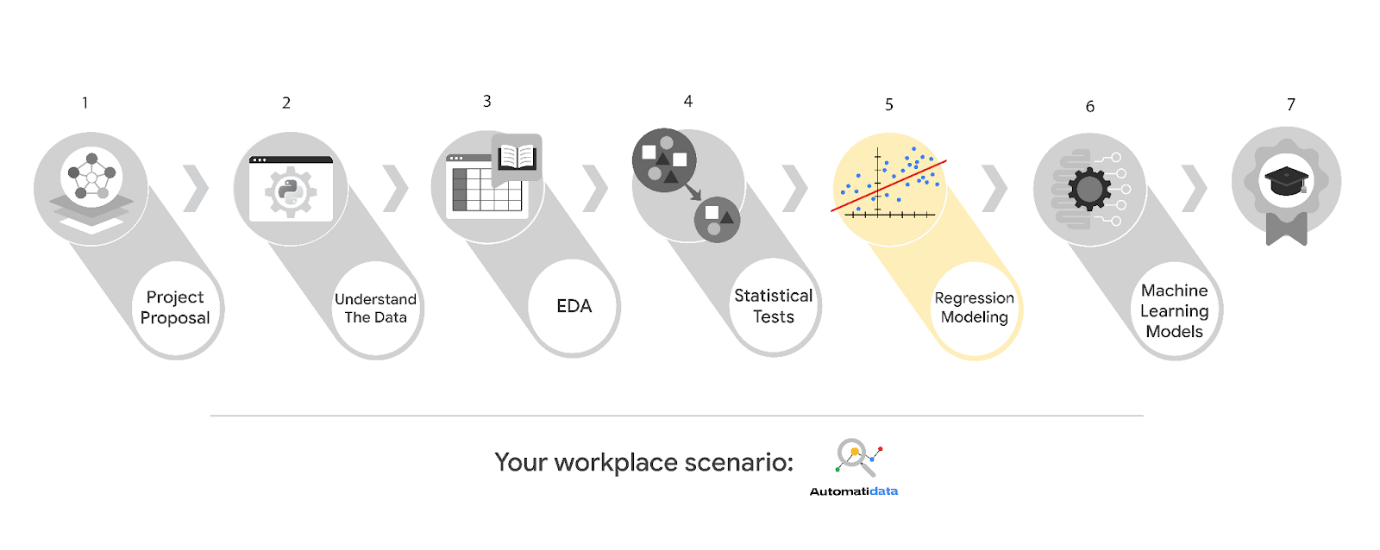
The end-of-course portfolio projects are designed for you to apply your data analytical skills within a workplace scenario. No matter which scenario you work with, you will practice your ability to discuss data analytic topics with coworkers, internal team members, and external clients.

As a reminder, you are required to complete one project for each course. To gain additional practice, or to add more samples to your portfolio, you may complete as many of the scenarios as you wish.

**Course 5 end-of-course portfolio project overview: Automatidata**

**Learn about the Course 5 Automatidata workplace scenario!**

The end-of-course project in Course 5 focuses on your ability to build regression models using Python. The end-of-course projects were designed with you in mind, offering an opportunity for you to practice and apply your data analytic skills. The materials provided here will guide you through discussions with co-workers, internal team members, and external stakeholders.



Learn more about the project, your role, and expectations in this reading.

**Background on the Automatidata scenario**

Automatidata works with its clients to transform their unused and stored data into useful solutions, such as performance dashboards, customer-facing tools, strategic business insights, and more. They specialize in identifying a client’s business needs and utilizing their data to meet those business needs.

Automatidata is consulting for the New York City Taxi and Limousine Commission (TLC). New York City TLC is an agency responsible for licensing and regulating New York City's taxi cabs and for-hire vehicles. The agency has partnered with Automatidata to develop a regression model that helps estimate taxi fares before the ride, based on data that TLC has gathered.

The TLC data comes from over 200,000 taxi and limousine licensees, making approximately one million combined trips per day.

***Note:*** *This project's dataset was created for pedagogical purposes and may not be indicative of New York City taxi cab riders' behavior.*

**Project background**

Automatidata is near the end of the TLC project. The following tasks are needed at this stage of the project:

* Determine the correct modeling approach
* Build a regression model
* Finish checking model assumptions
* Evaluate the model
* Interpret model results and summarize findings for stakeholders within TLC

**Your assignment**

You will create a regression model. Determine the type of regression model that is needed and develop one using the TLC data.

**Team members of Automatidata and the New York City TLC**

**Automatidata Team Members**

* Udo Bankole, Director of Data Analysis
* Deshawn Washington, Data Analysis Manager
* Luana Rodriquez, Senior Data Analyst
* Uli King, Senior Project Manager

Your teammates at Automatidata have technical experience with data analysis and data science. However, you should always be sure to keep summaries and messages to these team members concise and to the point.

**New York City TLC Team Members**

* Juliana Soto, Finance and Administration Department Head
* Titus Nelson, Operations Manager

***Note:*** *The story, all names, characters, and incidents portrayed in this project are fictitious. No identification with actual persons (living or deceased) is intended or should be inferred. The data shared in this project has been altered for pedagogical purposes.*

The TLC team members are program managers who oversee operations at the organization. Their roles are not highly technical, so be sure to adjust your language and explanation accordingly.

**Specific project deliverables**

In this end-of-course project, you will gain valuable practice of your new skills as you complete the following deliverables:

* Complete a PACE Strategy Document to consider questions, details, and action items for each stage of the project scenario
* Answer the questions in the Jupyter notebook project file
* Build a regression model in Python
* Report the results in an executive summary

Good luck in your role! Automatidata looks forward to seeing how you communicate your creative work and approach problem-solving!

**Key takeaways**

The end-of-course project is designed for you to practice and apply course skills in a fictional workplace scenario. By completing each course’s end-of-course project, you will have work examples that will enhance your portfolio and showcase your skills for future employers.

**Get started on the next course**

Congratulations on completing another course in the Google Advanced Data Analytics certificate! In this part of the program, you learned more about regression models, and practiced creating and analyzing multiple linear regression models.

The entire program has seven courses:

1. **Foundations of Data Science –** This course introduces the fundamentals of data science, how different data professionals operate in the workplace, and how these roles contribute to an organization’s vision of their future. The data science workflow PACE (plan, analyze, construct, enhance) is introduced to help you better understand how to navigate the technical and workplace expectations of this career.
2. **Getting Started with Python –** In this course, you will get started with Python for data analytics by developing an understanding of Python syntax, logic, data types, objects, and object-oriented programming.
3. **Go Beyond the Numbers: Translate Data into Insights –** Learn the fundamentals of data cleaning and visualizations and how to uncover meaningful stories in the data.
4. **The Power of Statistics –** Learn descriptive and inferential statistics, basic probability and probability distributions, sampling, confidence intervals, and hypothesis testing.
5. **Regression Analysis: Simplifying Complex Data Relationships** **–** In this course, you will apply your knowledge to modeling variable relationships, with a focus on linear regression, analysis of variance (ANOVA), and logistic regression. From model assumptions to evaluation and interpretation, you will understand relationships in datasets based on PACE. **(This is the course you just completed. Well done!)**
6. **The Nuts and Bolts of Machine Learning –** This course covers the fundamentals of supervised machine learning, and introduces learners to unsupervised learning through K-means and other clustering models. Learners will use different classification techniques such as decision trees, random forests, and gradient boosting to approach a realistic business problem.
7. **Google Advanced Data Analytics Capstone –** This course presents the capstone project for the Advanced Data Analytics certificate, which incorporates key concepts from each of the six preceding courses. The capstone project will yield data-driven suggestions including visualizations and models to provide insight for the business problem.

Now that you have completed this course, you are ready to move on to the next course: [**The Nuts and Bolts of Machine Learning**](https://www.coursera.org/learn/the-nuts-and-bolts-of-machine-learning/home/week/1).

Keep up the great work!

Mark as completed

Like

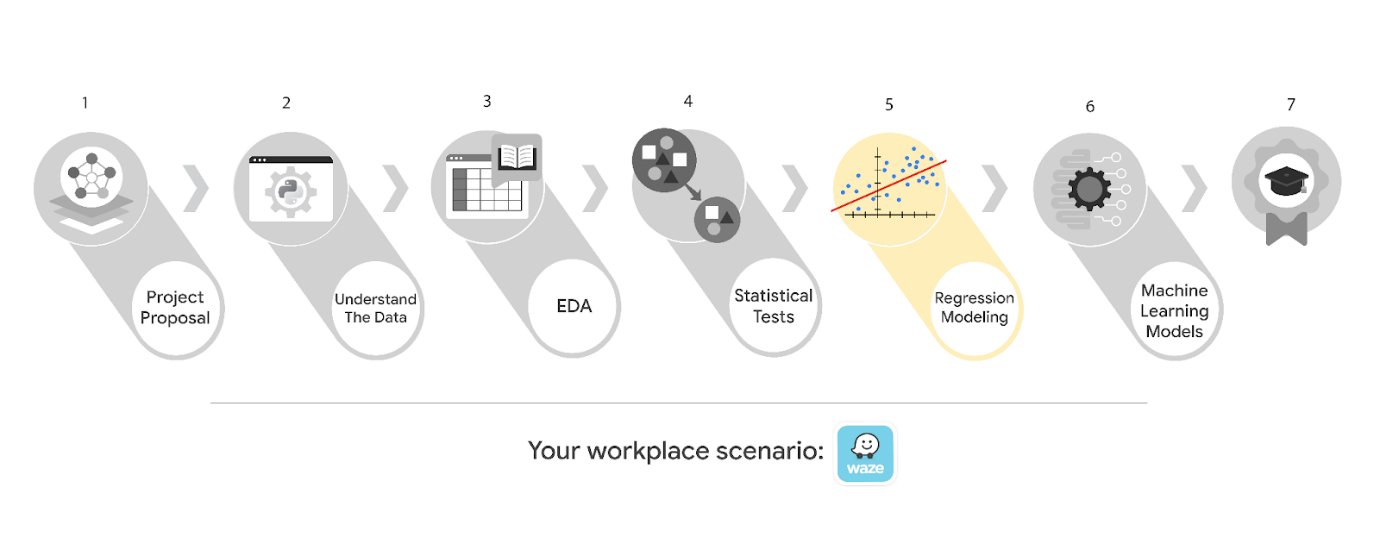
Dislike

Report an issue

**Course 5 end-of-course portfolio project overview: Waze**

**Learn about the Course 5 Waze workplace scenario!**

The end-of-course project in Course 5 focuses on your ability to build regression models using Python. As a reminder, in Course 1 you developed a project proposal that outlined milestones, which progress with each of the end-of-course projects. A visual representation is provided in the graphic shown here:



Learn more about the project, your role, and expectations in this reading.

**Background on the Waze scenario**

Waze’s free navigation app makes it easier for drivers around the world to get to where they want to go. Waze’s community of map editors, beta testers, translators, partners, and users helps make each drive better and safer. Waze partners with cities, transportation authorities, broadcasters, businesses, and first responders to help as many people as possible travel more efficiently and safely.

You’ll collaborate with your Waze teammates to analyze and interpret data, generate valuable insights, and help leadership make informed business decisions. Your team is about to start a new project to help prevent user churn on the Waze app. Churn quantifies the number of users who have uninstalled the Waze app or stopped using the app. This project focuses on monthly user churn.

This project is part of a larger effort at Waze to increase growth. Typically, high retention rates indicate satisfied users who repeatedly use the Waze app over time. Developing a churn prediction model will help prevent churn, improve user retention, and grow Waze’s business. An accurate model can also help identify specific factors that contribute to churn and answer questions such as:

* Who are the users most likely to churn?
* Why do users churn?
* When do users churn?

For example, if Waze can identify a segment of users who are at high risk of churning, Waze can proactively engage these users with special offers to try and retain them. Otherwise, Waze may lose these users without knowing why.

Your insights will help Waze leadership optimize the company’s retention strategy, enhance user experience, and make data-driven decisions about product development.

**Project background**

Waze’s data team is working on the churn project. The following tasks are needed at this stage of the project:

* Determine the correct modeling approach
* Build a regression model
* Finish checking model assumptions
* Evaluate the model
* Interpret model results and summarize findings for cross-departmental stakeholders within Waze

**Your assignment**

You will create a regression model for the churn project. You'll determine the type of regression model that is needed and develop one using Waze's churn project data.

**Team members at Waze**

**Data team roles**

* Harriet Hadzic - Director of Data Analysis
* May Santner - Data Analysis Manager
* Chidi Ga - Senior Data Analyst
* Sylvester Esperanza - Senior Project Manager

Data team members have technical experience with data analysis and data science. However, you should always be sure to keep summaries and messages to these team members concise and to the point.

**Cross-functional team members**

* Emrick Larson - Finance and Administration Department Head
* Ursula Sayo - Operations Manager

Your Waze team includes several managers overseeing operations. It is important to adapt your communication to their roles since their responsibilities are less technical.

***Note:*** *The story, all names, characters, and incidents portrayed in this project are fictitious. No identification with actual persons (living or deceased) is intended or should be inferred. And, the data shared in this project has been created for pedagogical purposes.*

**Specific project deliverables**

With this end-of-course project, you will gain valuable practice and apply your new skills as you complete the following:

* Complete the questions in the Course 5 PACE strategy document
* Answer the questions in the Jupyter notebook project file
* Build a binomial logistic regression model
* Create an executive summary to share your results

Good luck with this project! Waze looks forward to seeing how you communicate your creative work and approach problem-solving!

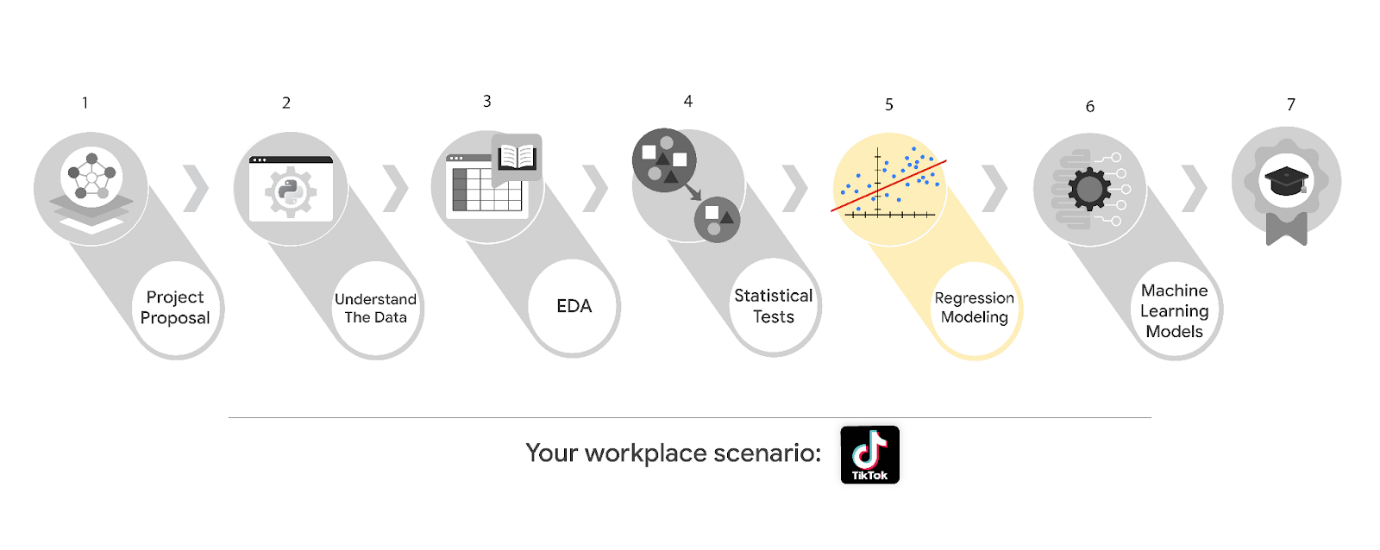
**Key takeaways**

The Google Advanced Data Analytics Certificate end-of-course project is designed for you to practice and apply course skills in a fictional workplace scenario. By completing each course’s end-of-course project, you will have work examples that will enhance your portfolio and showcase your skills for future employers.

**Course 5 end-of-course portfolio project overview: TikTok**

**Learn about the Course 5 TikTok workplace scenario!**

The end-of-course project in Course 3 focuses on your ability to use exploratory data analysis to organize and understand the data within a project. As a reminder, in Course 1 you developed a project proposal that outlined milestones, which progress with each of the end-of-course projects. A visual representation is provided in the graphic shown here:



Learn more about the project, your role, and expectations in this reading.

**Background on the TikTok scenario**

At TikTok, our mission is to inspire creativity and bring joy. Our employees lead with curiosity and move at the speed of culture. Combined with our company's flat structure, you'll be given dynamic opportunities to make a real impact on a rapidly expanding company, and grow your career.

TikTok users have the ability to submit reports that identify videos and comments that contain user claims. These reports identify content that needs to be reviewed by moderators. The process generates a large number of user reports that are challenging to consider in a timely manner.

TikTok is working on the development of a predictive model that can determine whether a video contains a claim or offers an opinion. With a successful prediction model, TikTok can reduce the backlog of user reports and prioritize them more efficiently.

**Project background**

TikTok’s data team is working on the claims classification project. The following tasks are needed at this stage of the project:

* Determine the correct modeling approach
* Build a regression model
* Finish checking model assumptions
* Evaluate the model
* Interpret model results and summarize findings for cross-departmental stakeholders within TikTok

**Your assignment**

You will create a regression model for the claims classification data. You'll determine the type of regression model that is needed and develop one using TikTok's claim classification data.

**Team members at TikTok**

**Data team roles**

* Willow Jaffey- Data Science Lead
* Rosie Mae Bradshaw- Data Science Manager
* Orion Rainier- Data Scientist

The members of the data team at TikTok are well versed in data analysis and data science. Messages to these more technical coworkers should be concise and specific.

**Cross-functional team members**

* Mary Joanna Rodgers- Project Management Officer
* Margery Adebowale- Finance Lead, Americas
* Maika Abadi- Operations Lead

Your TikTok team includes several managers, who oversee operations. It is important to adjust your general correspondence appropriately to their roles, given that their responsibilities are less technical in nature.

***Note:*** *The story, all names, characters, and incidents portrayed in this project are fictitious. No identification with actual persons (living or deceased) is intended or should be inferred. And, the data shared in this project has been created for pedagogical purposes.*

**Specific project deliverables**

With this end-of-course project, you will gain valuable practice and apply your new skills as you complete the following:

* Course 5 PACE Strategy Document to consider questions, details, and action items for each stage of the project scenario
* Answer the questions in the Jupyter notebook project file
* Create a regression model
* Evaluate the model
* Create an executive summary to share your results

**Key takeaways**

The Google Advanced Data Analytics Certificate end-of-course project is designed for you to practice and apply course skills in a fictional workplace scenario. By completing each course’s end-of-course project, you will have work examples that will enhance your portfolio and showcase your skills for future employers.

**END OF THE COURSE**