studocu
Search for courses, books or documents
University High School Sign in
Download AI Tools 0 0 Save

Programming for Data Science

Assignment 7, Module 7

Due no later than 11:59 PM CST on **Friday** of the Module

In this assignment, you will begin by revisiting the simple linear regression, then progress to multiple regression including models with and without interaction effects. You will create visualizations to assist in the testing of the assumptions of multiple re[illegible] watched the videos [illegible]

1 out of 14 Download

In a Word document, create a cover sheet with your name, LWC ID, the course name and number, the Assignment Title (Assignment 7) and the due date. Then, on the cover sheet you MUST include the HONOR PLEDGE:

"**On my honor as a student, I have neither given nor received any unauthorized aid on this assignment.**"

When answering the questions, include screenshots of your work for each step, and ensure that the screenshots show all relevant information.

NOTE: You MUST use screenshots from RStudio to document any code or models you use for all questions.

You MUST read the Chapter 10 material (textbook chapter 10 and watch the videos for Multiple Regression Analysis, Multiple Regression – Adding Variables, and Multiple Regression – Model Diagnostics. These form the basis of the assignment. It is STRONGLY recommended that you follow along with the videos as you watch them.

On with the assignment!

Question 1. Simple Linear Regression.

This question is a refresher on simple linear regression. The question uses the gapminder dataframe. First, you must ensure that the tidyverse and gapminder packages are installed.

```
library(tidyverse)
library(gapminder)
view(gapminder)
str(gapminder)
? gapminder
```

Then explore the gapminder dataframe which has six variables:

1. country – factor with 142 levels
2. continent – factor with 5 levels (Africa, Americas, Asia, Europe, Oceania)
3. year – 1952 to 2007 in 5-year increments
4. lifeExp – life expectancy in years, at birth
5. pop – population
6. gdpPerCap – gross domestic product measured in US dollars adjusted for inflation

a. Create a dataframe called africaData which contains the life expectancy and gross domestic product per capita for the continent of Africa. Do not include the continent in the dataset as there is only one continent being used. Document the creation of the dataframe and its contents with screenshots as needed.

b. You will be running a simple linear regression with your two variables: life expectancy and gross domestic product. Explain which variable will be your dependent variable and which will be your independent variable.

c. Create a linear model called africaModel and show your output.

d. In your output, explain the coefficients and their estimates. Are the coefficients statistically significant? If the coefficients are significant, state why this is the case.

e. What is the value for the R^2 value, and what does it tell you?

f. What is the value of the p-value, and what does it tell you?

Question 2. Multiple Regression

In this question, you will use the MASS package and the swiss dataframe. Use the commands in the screenshot below to load the MASS package and examine the swiss dataframe.

```
285 library(MASS)
286 swiss
287
```

287:1 (Top Level)

Console Terminal Background Jobs

R 4.3.3 · C:/Users/DoranP/OneDrive - lindsey.edu/Desktop/DoranRProg/Chap7/

```
> library(MASS)
> swiss
```

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Courtelary	80.2	17.0	15	12	9.96	22.2
Delemont	83.1	45.1	6	9	84.84	22.2
Franches-Mnt	92.5	39.7	5	5	93.40	20.2
Moutier	85.8	36.5	12	7	33.77	20.3
Neuveville	76.9	43.5	17	15	5.16	20.6
Porrentruy	76.1	35.3	9	7	90.57	26.6
Broye	83.8	70.2	16	7	92.85	23.6
Glane	92.4	67.8	14	8	97.16	24.9
Gruyere	82.4	53.3	12	7	97.67	21.0
Sarine	82.9	45.2	16	13	91.38	24.4
Veveyse	87.1	64.5	14	6	98.61	24.5
Aigle	64.1	62.0	21	12	8.52	16.5
Aubonne	66.9	67.5	14	7	2.27	19.1
Avenches	68.9	60.7	19	12	4.43	22.7
Cossonay	61.7	69.3	22	5	2.82	18.7
Echallens	68.3	72.6	18	2	24.20	21.2

Now use the ?swiss command to learn more about the swiss dataframe.

a. Create a multiple regression, without interaction, with Fertility being the dependent variable and Agriculture and Examination being the

independent variables. Store the output of the regression to a dataframe called modelSwiss1. Show your work with screenshots.

b. Explain the coefficients, their estimates, and their Pr(>|t|) column data.

c. In simple linear regression we use the R^2 value to assess the percentage of variation accounted for by a particular variable. In multiple regression we use the Adjusted R^2 value. What is the Adjusted R^2 value and what does it tell you about the model? Is the value considered to be high or low?

d. What is the value of the p-value and what does it tell you about the model?

e. How can we improve the power (Adjusted R^2 Square) value of the model?

Question 3. Multiple Regression

This question continues with the swiss dataframe. We would like to examine the effect that Catholic and Education would have on Fertility.

a. Run the multiple regression without interaction. Save your output in dataframe called swissModel2. Show your code and output with screenshots.

b. Explain the coefficients, their estimates, and their Pr(>|t|) column data.

c. What is the Adjusted R^2 value and what does it tell you about the model? Is the value considered to be high, moderate low?

d. Is this model better (more powerful) than that in Question 2 (use the Adjusted R^2 values)?

e. What is the value of the p-value and what does it tell you about the model?

Document continues below

Discover more from:

Introductory statistics
Stat 133
Egerton University
25 documents
Go to course

116
STAT 133 – Introduction to basic statistics
Introductory stati...
83% (12)

15
Comparison between Brazil and Cuba...
Introductory stati...
100% (1)

2
FORM 2 MATHEMATICS: QUADRATIC...
Introductory stati...
None

7
Probability Problems Solutions for...
Introductory stati...
None

Question 4. Multiple Regression WITH INTERACTION

In this question, you will repeat question 3, but you will use a model with interaction. You do not have to make a very big change. Just change the "+" character in the model code in R to a "*" character.

a. Run the multiple regression with interaction. Save your output in dataframe called swissModel3. Show your code and output with screenshots.

b. Explain the coefficients, their estimates, and their Pr(>|t|) column data.

c. What is the Adjusted R^2 value and what does it tell you about the model? Is the value considered to be high, moderate low?

d. Is this model better (more powerful) than that in Question 3?

e. What is the value of the p-value and what does it tell you about the model?

Question 5. Adding Values to a Model

In this question we will use the penguins dataframe in the palmerpenguins package in the tidyverse. Ensure the tidyverse is loaded in your RStudio, and then examine the penguins dataframe using the following code:

```
library(tidyverse)
library(palmerpenguins)
str(penguins)
?penguins
```

302:1 (Top Level)

Console Terminal Background Jobs

R 4.3.3 · C:/Users/DoranP/OneDrive - lindsey.edu/Desktop/DoranRProg/Chap7/

```
> library(tidyverse)
> library(palmerpenguins)
> str(penguins)
tibble [344 × 8] (S3: tbl_df/tbl/data.frame)
 $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
 $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
 $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
 $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
 $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
 $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

R: Size measurements for adult foraging penguins near Palmer... Find in Topic

A tibble with 344 rows and 8 variables:

species

a factor denoting penguin species (Adélie, Chinstrap and Gentoo)

island

a factor denoting island in Palmer Archipelago, Antarctica (Biscoe, Dream or Torgersen)

bill_length_mm

a number denoting bill length (millimeters)

bill_depth_mm

a number denoting bill depth (millimeters)

flipper_length_mm

an integer denoting flipper length (millimeters)

body_mass_g

an integer denoting body mass (grams)

sex

a factor denoting penguin sex (female, male)

year

an integer denoting the study year (2007, 2008, or 2009)

a. Use the following code template to load all the variables in the penguins dataframe while looking for their effect on body mass in gram (body_mass_g) which is another way of saying "the weight of the penguin". Show your output in a screenshot.

lm(**variable of interest** ~ . , data = **dataframe**) |>

step(direction = "backward"), trace = 0) |>

summary()

The code direction = "backward" means all variables except the variable of

The code direction = "backward" means all variables except the variable of interest will be loaded and then variables will be removed while looking for a higher value of Adjusted R^2. This will continue until the best model has been found. The trace = 0 code tells R to only keep the best model. The summary() command outputs the results to your screen. The items in **bold** are for you to fill in.

b. What does the model formula tell you (look above the residuals in the output for the formula)?

c. Were there any rows which were not included in the model?

d. What is the value of Adjusted R^2, and is it low, moderate, or high?

e. What is the p-value of the model and what does it tell you?

Question 6. The assumptions of linear regression.

This question extends the results of Question 1 which used the africaData dataframe you created. This question uses code (slightly changed) from the video Multiple Regression – Model Diagnostics by Dr. Greg Martin. The video is in the Instructional Materials, and at the end of the video is a link to his pdf files containing code and output for each video. **It is critical that you watch this video before starting this question. Obtaining the pdfs is also a really good idea.**

The linear regression model has four assumptions, and three of them apply to the residuals. The residuals are the difference between the value in the data versus the value predicted by the linear model. flasically, this means that, if a regression line is drawn through a scatterplot, the residuals are the **vertical** distance from each data item to the regression line. The assumptions are:

1. There is a linear relationship between the independent (x) and dependent (y) variable(s) .
2. The residuals are normally distributed.
3. The residuals are homoscedastic (we will get to that term in a moment).
4. The residuals are independent of each other.

Generating a linear model output is very useful, but it is only valid if it passes these four assumption tests. The first step is to generate a plot of the residuals and the regression line. Use the following code to obtain the plot:

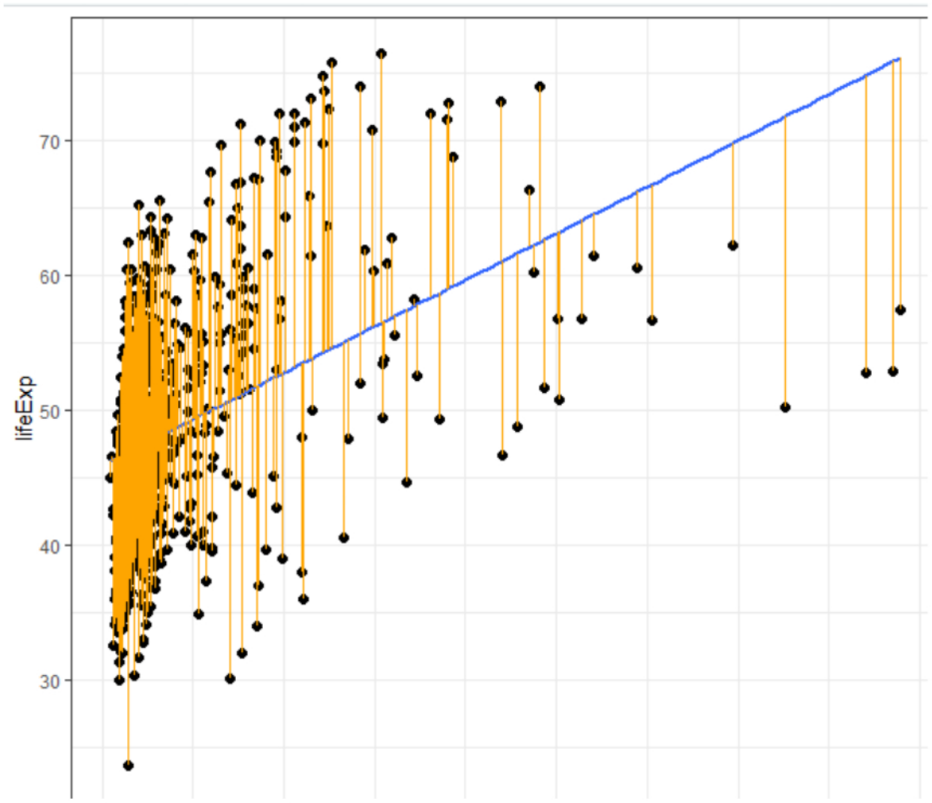
```r
africaData |>
  mutate(predicted = predict(lm(lifeExp ~ gdpPercap, data = africaData))) |>
  ggplot(aes(gdpPercap, lifeExp)) +
  geom_point(size = 2) +
  geom_smooth(method = lm, se = F, size = 1)+
  geom_segment(aes(xend = gdpPercap, yend = predicted), color = "orange") +
  theme_bw()
```

318:1 (Top Level)

Console Terminal Background Jobs

R 4.3.3 · C:/Users/DoranP/OneDrive - lindsey.edu/Desktop/DoranRProg/Chap7/

```
> africaData |>
+   mutate(predicted = predict(lm(lifeExp ~ gdpPercap, data = africaData))) |>
+   ggplot(aes(gdpPercap, lifeExp)) +
+   geom_point(size = 2) +
+   geom_smooth(method = lm, se = F, size = 1)+
+   geom_segment(aes(xend = gdpPercap, yend = predicted), color = "orange") +
+   theme_bw()
`geom_smooth()` using formula = 'y ~ x'
```



The residuals are the vertical lengths of the orange lines. You can see the some are longer than others. We are now ready to tackle the assumptions that underly our model.

a. Use your model from Question 1 and create a dataframe called residModel. The residModule will use the linear model (lm) and will then plot the residModel using the plot() function. **When your code runs, you should see the message "Hit <Return> to see the next plot."**. This is because there are four plots. Use screenshots to show your code and the four plots

there are four plots. Use screenshots to show your code and the four plots.

b. Assumption 1. Linear relationship between dependent (y) and independent (x) variable(s). To assess this assumption, look at the residuals versus fitted values plot. If there is a no pattern in the plot, then you are quite sure that this assumption is correct. Show your diagram here. Does there appear to be no pattern in the plot?

c. There is a test called the Harvey-Collier Multiplier Test for Linearity which you can use to verify if the first assumption (the relationship is linear). However, the null hypothesis (H0) for this test is that the relationship is linear. Therefore, although it will seem strange, we are hoping for a p-value GREATER THAN 0.05 because we want to be unable to reject the null hypotheses. The test is actually quite easy to run, but you need to install the lmtest package

```
install.packages("lmtest")
library(lmtest)
```

324:1 (Top Level) R Script

Console Terminal Background Jobs

R 4.3.3 · C:/Users/DoranP/OneDrive - lindsey.edu/Desktop/DoranRProg/Chap7/

```
> install.packages("lmtest")
WARNING: Rtools is required to build R packages but is not currently installed.
Please download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/DoranP/AppData/Local/R/win-library/4.3'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/lmtest_0.9-40.zip'
Content type 'application/zip' length 406020 bytes (396 KB)
downloaded 396 KB

package 'lmtest' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\DoranP\AppData\Local\Temp\Rtmp44eaTe\downloaded_packages
> library(lmtest)
Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

d. Use the code

harvtest(residModel) to obtain the output. Show your output and determine if the null hypothesis should be rejected. What do you conclude?

e. Assumption 2: The residuals are normally distributed. You can get an idea of whether this assumption is valid by looking at the second of the four plots: the Q-Q Residuals plot. If the plot shows a line rising from left to right then

the model may meet this assumption. Show a copy of your model and state whether you think this assumption is satisfied.

f. There is a test called the Shapiro-Wilks Normality Test to determine if the residuals are normally distributed. The null hypotheses (H0) is that the model residuals are normally distributed, so you are hoping for a p-value greater than 0.05 so that the null hypothesis is **not** rejected. The command to test this is:

   Shapiro.test(residuals(your model name))

   Run the Shapiro test and determine if the assumption is valid or not.

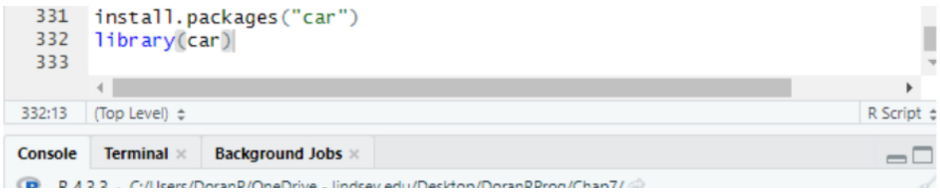
g. Assumption 3. The residuals are homoscedastic. This means that the residual values are constant across all levels of the explanatory (x) variables. The test for this assumption is the flreusch-Pagan test. This is also a part of the lmtest package. The code to run this test is:

   bptest(your model name)

   The null hypothesis assumes homoskedasticity. Therefore, you are hoping for a p-value greater than 0.05.

   Run the test and state your conclusion regarding assumption 3.

h. Assumption 4: The residual values are independent from each other. Sometimes you will have data where the data items, and hence the residuals, are related. For example, in time series data, you might be assessing the changes in strength of athletes over time. However, the residuals for athlete A are all from the same person (Athlete A) and hence will not be independent. The test for this assumption is the Durbin-Watson test and this requires the car package.



```
>
> install.packages("car")
WARNING: Rtools is required to build R packages but is not currently installed.
Please download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/DoranP/AppData/Local/R/win-library/4.3'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.3/car_3.1-2.zip'
Content type 'application/zip' length 1707860 bytes (1.6 MB)
downloaded 1.6 MB

package 'car' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\DoranP\AppData\Local\Temp\Rtmp44eaTe\downloaded_packages
> library(car)
Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode

The following object is masked from 'package:purrr':

    some
```

The code to run the Durbin-Watson test is:

durbinWatsonTest(your model name)

Run the test, show your output, and state whether assumption 4 has been met or not. If the D-W Statistic score is close to 2, this suggests no autocorrelation. A value below 2 indicates positive autocorrelation, and a value above 2 suggests negative autocorrelation. What does your test show?

i. Do you think the model is appropriate for the data. Why or why not?

More from: Introductory statistics (Stat 133)

Introductory statistics
Stat 133
Egerton University
25 documents
Go to course

116 STAT 133 – Introduction to basic statistics
Introductory stati...
83% (12)

15 Comparison between Brazil and Cuba...
Introductory stati...
100% (1)

2 FORM 2 MATHEMATICS: QUADRATIC...
Introductory stati...
None

7 Probability Problems Solutions for...
Introductory stati...
None

More from: moon moan

moon moon

Impact 22

Capella University

Discover more

7

Bank Loan Appraisal Project

Basic Mathematics

None

14

Business Analytics Ideas in Multiple...

Organizational Su...

None

3

Academic Writing Project

Organizational Su...

None

11

1-month Cycle Accounting Project

Organizational Su...

None

Recommended for you

15

Comparison between Brazil and Cuba...

Introductory stati...

100% (1)

116

STAT 133 - Introduction to basic statistics

Introductory stati...

83% (12)

studocu

Company

About Us

Studocu Premium

Academic Integrity

Jobs

Blog

Dutch Website

Study Tools

All Tools

Ask AI

AI Notes

AI Quiz Generator

Contact & Help

F.A.Q.

Contact

Newsroom

Legal

Terms

Privacy Policy

Cookie Settings

Cookie Statement

Copyright & DSA

Download on the App Store

GET IT ON Google Play

Excellent

9,131 reviews on Trustpilot