HW5_answers

August 30, 2019

1 ARE 106 Summer Session II

2 Homework 5

This homework will be due on **September 9nd, at 4:10pm**

2.1 Name:

2.2 SSID:

Please put your name and SSID in the corresponding cells above.

The homework is worth 13.5 points.

For each of the following questions, show as much of your steps as you can (without going overboard). If you end up getting the wrong answer, but we spot where you made a mistake in the algebra, partial credit will be more readily given. If you only put the final answer, you will be marked either right or wrong.

Answer questions in the correct cell. For problems where you have to input math, make sure that you know that it's a markdown cell (It won't have a `In: []` on the left) and make sure you run the cell by either pressing `Ctrl + Enter` or going to `Cell -> Run Cell`. Alternatively, write all your answers and then go to `Cell -> Run All Cells` after you're done.

Please ignore cells that read `\pagebreak`. These are so your document converts to PDF in a way that will make it possible to grade your homework. Ignore them and only write your answers where it is specified.

When you are finished export your homework to a PDF by going to `File -> Download as -> PDF`.

2.3 Question 1: Probabilities with Normal Distributions

Suppose we have a random variable Y, that follows a normal distribution with mean 0 and standard deviation 1.

Please consult the statistics review and the lecture notes for more information.

For each of the following questions, find their probabilities using the z-table from this link:

http://www.z-table.com/uploads/2/1/7/9/21795380/8573955.png?759

Hint: Remember that the normal distribution is symmetric and adds up to 1

Hint: Remember that the table below gives $Pr(Y \leq z)$ where $z > 0$, so you might have to do some work to get it into a state that you can use the table with.

a. $Pr(Y \leq 1.96)$

b. $Pr(Y > 1.96)$

c. $Pr(Y > 0)$

d. $Pr(-1.96 \leq Y \leq 1.96)$

e. What is $1 - Pr(-1.96 \leq Y \leq 1.96)$? In a hypothesis test, what kind of test (one-tailed, two-tailed) and α level would this correspond to?

Please put your answers for Question 1 here.

- a. 0.9750
- b. 1 - 0.9750 = 0.0250
- c. 0.5
- d. We can think about this using the properties of the normal that we know and love. We know that $Pr(Y \leq 1.96) = Pr(Y > 1.96)$. So if we get $Pr(Y > 1.96)$, we need multiply that by 2 and substract it from 1. From b., we know that $Pr(Y > 1.96) = 0.025$. Multiplying that by two we get: $0.0250 \cdot 2 = 0.05$. Then $1 - 0.05 = .95$.
- e. In this case we can use this number when doing a two-tailed test with $\alpha = 0.05$.

2.4 Question 2: Doing Question 1 with Python

For this question, we'll reuse Question 1 a-d, but we'll do it with Python. The way to do this is first to import the normal distribution from `scipy.stats` and use its CDF:

```
from scipy.stats import norm
```

Then create an instance of a normal distribution with its mean and standard deviation:

```
dist = norm(<mean here>, <standard deviation here>)
```

And now call its CDF, where it takes an argument x and gives you $Pr(Y \leq x)$:

```
dist.cdf(x)
```

Now for Question 1 a-d, find the answers using Python.

[10]:
```
## Put your answers for Question 2 here:

from scipy.stats import norm

dist = norm(0,1)
## a.
dist.cdf(1.96)

## b.
1 - dist.cdf(1.96)
\pagebreak

## c.
dist.cdf(0)

## d.
## Doing same thing as explanation from question 1
1 - 2*(1 -dist.cdf(1.96))

print(dist.ppf(.95))
```

```
1.6448536269514722
```

2.5 Question 3: Test Statistics

For the following questions, first state whether the test needed is a Z-statistic or T-statistic and calculate the test statistic.

Note: Feel free to make a new cell to do some calculations with Python if need be.

- a. The grades on a statistics test at GW University are normally distributed with $\mu = 55$ and $\sigma = 12$. George scored $G = 65$ on the exam. You want to test whether George's score is more than the mean.
- b. The heights of the statisticians working on the Large Hadron Collider project seems to be normally distributed with $\mu = 68$ inches and $\sigma = 5$ inches. Dr. Numbercrackers' height is $H = 61$ inches. You want to test is Dr. Numbercracker's height is different from the mean.
- c. The Bureau on Economic Development conducted a survey of families residing in a small town of Gugelshnackel. 100 residents were surveyed and the survey revealed that $\bar{X} = 45,000$ and $s = 10,538$. You need to test whether mean income is statistically different from a null hypothesis of 40,000.

Please put your answers for Question 3 here.

- a. $\frac{65-55}{12} = \frac{5}{6}$
- b. $\frac{61-68}{5} = \frac{-7}{5} = -1.4$
- c. $\frac{45000-40000}{10538} = \frac{5000}{10538} = .47$

2.6 Question 4: Hypothesis Testing

From your answers to Question 3, a given α level and a particular kind of test, find the critical value and test whether your statistics are statistically significant or not.

For this question you can use `norm.ppf` to find the critical value. Simply put in the area you want and it gives you the associated number:

```
dist.ppf(0.95)
```

Be sure to write down your null and alternative hypotheses.

You can choose to do this using the z-table above or with python.

- a. Using your answer from Question 3 a., run a one tailed test with $\alpha = 0.05$
- b. Using you answer from Question 3 b., run a two-tailed test with $\alpha = 0.05$.
- c. Using you answer from QUestion 3 c., run a two-tailed test with $\alpha = 0.01$.
- d. Calculate the p-value for your answer from Question 3 c.

Please put your answers for Question 4 here

- a.

$H_0 : G > 55$

$H_1 : G \leq 55$

For a one-tailed test, and $\alpha = 0.05$, we need 5% of the area on the right of the distribution so our critical value is 1.648. Since .83 < 1.648, we fail to reject the null. - b.

$H_0 : H = 55$

$H_1 : H \neq 55$

With a two-tailed test and an $\alpha = 0.05$, we have a critical value of 1.96. Since -1.4 > -1.96, we fail to reject the null.

- c.

$H_0 : \bar{X} = 55$

$H_1 : \bar{X} \neq 55$

With a two-tailed test and $\alpha = 0.01$, we need a critical value of 2.33. Since .47 < 2.33, we fail to reject the null.

- d. The p-value would be .63

[26]:
```
## If you need to do calculations in Python, put them here.

## a.
print(\pagebreak
dist.ppf(.95))

## b.
print(dist.ppf(.975))

## c.
print(dist.ppf(.99))

## d.
2*(1- dist.cdf(.47))
```

```
1.6448536269514722
1.959963984540054
2.3263478740408408
```

[26]:
```
0.6383550175651116
```

2.7 Question 5: Nonlinearity

For the given situation, write what you think the best nonlinear transformation to the variables would be to investigate this change. Write your answer in the form of a regression model.

- a. In investigating the effect of IQ on salary, we'd like to see the effect in terms of an elasticity.
- b. We'd like to see how age affects salary. We'd like to investigate whether there is some peak effect of age on salary.
- c. We'd like to investigate whether there is a differential/heterogeneous effect on age across genders.

Write your answer to question 5 here.

- a.

$$log(salary_i) = b_0 + b_1 log(IQ_i) + e_i$$

- b.

$$salary_i = b_0 + b_1 age_i + b_2 age_i^2 + e_i$$

- c.

$$salary_i = b_0 + b_1 age_i + b_2 gender_i + b_3 age_i \cdot gender_i + e_i$$

2.8 Question 6: Testing Significance of our Model

Suppose we run two regressions:

$$(1) : Y_i = b_0 + b_1 X_i + e_i$$

$$(2) : Y_i = b_0 + b_1 X_i + b_2 X_i^2 + e_i$$

The data can be found at:

https://raw.githubusercontent.com/lordflaron/ARE106data/master/HW5.csv

a. Import `pandas` and `statsmodels.formula.api` and load in the data

[14]:
```
## Put your answer for a. here

import pandas as pd
import statsmodels.formula.api as sm\pagebreak


df = pd.read_csv("https://raw.githubusercontent.com/lordflaron/ARE106data/
 ↪master/HW5.csv")
```

b. Now run the first regression.

[20]:
```
## Put your answer for b. here
mod = sm.ols('Y ~ X', data=df)
results = mod.fit()
\pagebreak
results.summary()
```

[20]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.521
Model:                            OLS   Adj. R-squared:                  0.521
Method:                 Least Squares   F-statistic:                     1087.
Date:                Fri, 30 Aug 2019   Prob (F-statistic):          8.45e-162
Time:                        15:12:20   Log-Likelihood:                -4635.6
No. Observations:                1000   AIC:                             9275.
Df Residuals:                     998   BIC:                             9285.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      9.1924      0.975      9.425      0.000       7.278      11.106
X              6.4404      0.195     32.962      0.000       6.057       6.824
==============================================================================
Omnibus:                      381.351   Durbin-Watson:                   2.052
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1447.105
Skew:                           1.822   Prob(JB):                         0.00
Kurtosis:                       7.632   Cond. No.                         6.25
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

c. What is the R^2?

Put your answer here.

0.521

d. Now run the second regression. You can either create the X_i^2 with an assign call or use `np.power(X,2)` in your patsy formula.

[42]:
```
## Put your answer for c. here

mod = sm.ols('Y ~ X + np.power(X,2)', data=df)
results = mod.fit()
results.summary()
```

[42]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.924
Model:                            OLS   Adj. R-squared:                  0.924
Method:                 Least Squares   F-statistic:                     6079.
Date:                Fri, 30 Aug 2019   Prob (F-statistic):               0.00
Time:                        15:44:26   Log-Likelihood:                -3713.9
No. Observations:                1000   AIC:                             7434.
Df Residuals:                     997   BIC:                             7449.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
==
                     coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
--
Intercept          1.9562      0.401      4.881      0.000       1.170
2.743
X                  0.4133      0.114      3.639      0.000       0.190
0.636
np.power(X, 2)     0.9992      0.014     72.811      0.000       0.972
1.026
==============================================================================
Omnibus:                        1.709   Durbin-Watson:                   2.064
Prob(Omnibus):                  0.426   Jarque-Bera (JB):                1.738
Skew:                          -0.066   Prob(JB):                        0.419
Kurtosis:                       2.844   Cond. No.                         53.6
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
"""
```

e. What is the R^2?

Put your answer here.

0.924

f. For the second model you estimated in c., what is the t-statsitic for X? Is it statistically significant at the 0.05 level?

Put your answer here.

3.639. Yes, it is statistically significant at the 0.05 level.

g. Do the hypothesis test for the coefficient for X being 0 using the coefficient and standard error from the table. What do you find? Is it similar to the answer in f?

Put your answer here.

$\frac{.4133-0}{.114} = 3.6254$

It's very similar to the t-statistic from f. It should be the exactly the same, but we dont have all decimal places available so there is some error.

h. Using the R^2 from the second regression you ran, calculate the test statistic for whether Model 2 explains more than Model 1. You can find the correct statistic from the lecture notes. Use the R^2_{alt} as Model 2's R^2 and R^{null} is Model 1's R^2.

Note: K_{alt} is the number of *added* regressors from Model 1 to Model 2. Keep that in mind.

Compare this test statistic against the Chi-squared critical value of a one-tailed test at $\alpha = 0.05$ of 3.841. Is it statistically more informative (i.e. does it reject)?

Put your answer here.

We should use:

$$\frac{R^2_{alt} - R^2_{null}}{(1 - R^2_{alt})/(N - K_{alt} - 1)}$$

[45]:
```
print((.924 - .521)/((1-.924)/(1000 - 1 -1)))
print("We reject the null hypothesis at the 0.05 level.")
```

```
5292.026315789477
We reject the null hypothesis at the 0.05 level.
```