Analyze_ab_test_results_notebook

November 12, 2020

0.1 Analyze A/B Test Results

Fatema Buhuligah

This project will assure you have mastered the subjects covered in the statistics lessons.

0.2 Table of Contents

Introduction

A/B tests are very commonly performed by data analysts and data scientists. It is important that you get some practice working with the difficulties of these. For this project, you will be working to understand the results of an A/B test run by an e-commerce website. Your goal is to work through this notebook to help the company understand if they should implement the new page, keep the old page, or perhaps run the experiment longer to make their decision.

As you work through this notebook, follow along in the classroom and answer the corresponding quiz questions associated with each question. The labels for each classroom concept are provided for each question. This will assure you are on the right track as you work through the project, and you can feel more confident in your final submission meeting the criteria. As a final check, assure you meet all the criteria on the RUBRIC.

Part I - Probability

To get started, let's import our libraries.

```
In [1]: import pandas as pd
        import numpy as np
        import random
        import matplotlib.pyplot as plt
        %matplotlib inline
        #We are setting the seed to assure you get the same answers on quizzes as we set up
        random.seed(42)
```

`1.` Now, read in the `ab_data.csv` data. Store it in `df`. **Use your dataframe to answer the questions in Quiz 1 of the classroom.**

a. Read in the dataset and take a look at the top few rows here:

```
In [2]: df = pd.read_csv('ab_data.csv')
        df.head()
```

```
Out[2]:    user_id                   timestamp      group landing_page  converted
        0   851104  2017-01-21 22:11:48.556739    control     old_page          0
        1   804228  2017-01-12 08:01:45.159739    control     old_page          0
        2   661590  2017-01-11 16:55:06.154213  treatment     new_page          0
        3   853541  2017-01-08 18:28:03.143765  treatment     new_page          0
        4   864975  2017-01-21 01:52:26.210827    control     old_page          1
```

b. Use the cell below to find the number of rows in the dataset.

```
In [3]: df.shape[0]
```

```
Out[3]: 294478
```

c. The number of unique users in the dataset.

```
In [4]: df.user_id.nunique()
```

```
Out[4]: 290584
```

d. The proportion of users converted.

```
In [5]: converted = df.query('converted == 1')['user_id'].nunique()
        proportion_converted = converted / df.user_id.nunique()
        proportion_converted
```

```
Out[5]: 0.12104245244060237
```

e. The number of times the `new_page` and `treatment` don't match.

```
In [6]: df.query('(landing_page == "new_page" and group != "treatment") or (landing_page != "new
```

```
Out[6]: 3893
```

f. Do any of the rows have missing values?

```
In [7]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 294478 entries, 0 to 294477
Data columns (total 5 columns):
user_id         294478 non-null int64
timestamp       294478 non-null object
group           294478 non-null object
landing_page    294478 non-null object
converted       294478 non-null int64
dtypes: int64(2), object(3)
memory usage: 11.2+ MB
```

2. For the rows where **treatment** does not match with **new_page** or **control** does not match with **old_page**, we cannot be sure if this row truly received the new or old page. Use **Quiz 2** in the classroom to figure out how we should handle these rows.

a. Now use the answer to the quiz to create a new dataset that meets the specifications from the quiz. Store your new dataframe in **df2**.

```
In [8]: df2 = df.drop(df.query('group == "treatment" & landing_page != "new_page" | group == "co
        df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 290585 entries, 0 to 294477
Data columns (total 5 columns):
user_id         290585 non-null int64
timestamp       290585 non-null object
group           290585 non-null object
landing_page    290585 non-null object
converted       290585 non-null int64
dtypes: int64(2), object(3)
memory usage: 13.3+ MB
```

```
In [9]: # Double Check all of the correct rows were removed - this should be 0
        df2[((df2['group'] == 'treatment') == (df2['landing_page'] == 'new_page')) == False].sha
```

```
Out[9]: 0
```

3. Use **df2** and the cells below to answer questions for **Quiz3** in the classroom.

a. How many unique **user_id**s are in **df2**?

```
In [10]: df2.user_id.nunique()
```

```
Out[10]: 290584
```

b. There is one **user_id** repeated in **df2**. What is it?

```
In [11]: df2[df2.user_id.duplicated()]['user_id']
```

```
Out[11]: 2893    773192
         Name: user_id, dtype: int64
```

c. What is the row information for the repeat **user_id**?

```
In [12]: df2[df2.user_id.duplicated()]
```

Out[12]:

	user_id	timestamp	group	landing_page	converted
2893	773192	2017-01-14 02:55:59.590927	treatment	new_page	0

d. Remove **one** of the rows with a duplicate **user_id**, but keep your dataframe as **df2**.

```
In [13]: df2 = df2.drop(df2[df2.user_id.duplicated()].index)
         df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 290584 entries, 0 to 294477
Data columns (total 5 columns):
user_id         290584 non-null int64
timestamp       290584 non-null object
group           290584 non-null object
landing_page    290584 non-null object
converted       290584 non-null int64
dtypes: int64(2), object(3)
memory usage: 13.3+ MB
```

4. Use **df2** in the cells below to answer the quiz questions related to **Quiz 4** in the classroom.

a. What is the probability of an individual converting regardless of the page they receive?

```
In [14]: df2.query('converted == 1').shape[0] / df2.shape[0]

Out[14]: 0.11959708724499628
```

b. Given that an individual was in the `control` group, what is the probability they converted?

```
In [15]: df2.query('group == "control" & converted == 1').shape[0] / df2.query('group == "contro

Out[15]: 0.1203863045004612
```

c. Given that an individual was in the `treatment` group, what is the probability they converted?

```
In [16]: df2.query('group == "treatment" & converted == 1').shape[0] / df2.query('group == "trea

Out[16]: 0.11880806551510564
```

d. What is the probability that an individual received the new page?

```
In [17]: df2.query('landing_page == "new_page"').shape[0] / df2.shape[0]

Out[17]: 0.5000619442226688
```

e. Consider your results from parts (a) through (d) above, and explain below whether you think there is sufficient evidence to conclude that the new treatment page leads to more conversions.

```
In [18]: df2.query('group == "control" & converted == 1').shape[0] / df2.query('group == "contro

Out[18]: 0.0015782389853555567
```

Since the probability of an individual converting in control group is almost the same as the individual converting in treatment group, (12% & 11.8%) and the difference between these two probabilities is very small (0.15%), I do not think it is a sufficient evidence to conclude that the new treatment page leades to more conversions.

Part II - A/B Test

Notice that because of the time stamp associated with each event, you could technically run a hypothesis test continuously as each observation was observed.

However, then the hard question is do you stop as soon as one page is considered significantly better than another or does it need to happen consistently for a certain amount of time? How long do you run to render a decision that neither page is better than another?

These questions are the difficult parts associated with A/B tests in general.

`1.` For now, consider you need to make the decision just based on all the data provided. If you want to assume that the old page is better unless the new page proves to be definitely better at a Type I error rate of 5%, what should your null and alternative hypotheses be? You can state your hypothesis in terms of words or in terms of p_{old} and p_{new}, which are the converted rates for the old and new pages.

$$H_0 : p_{old} p_{new}$$

$$H_1 : p_{old} < p_{new}$$

`2.` Assume under the null hypothesis, p_{new} and p_{old} both have "true" success rates equal to the **converted** success rate regardless of page - that is p_{new} and p_{old} are equal. Furthermore, assume they are equal to the **converted** rate in **ab_data.csv** regardless of the page.

Use a sample size for each page equal to the ones in **ab_data.csv**.

Perform the sampling distribution for the difference in **converted** between the two pages over 10,000 iterations of calculating an estimate from the null.

Use the cells below to provide the necessary parts of this simulation. If this doesn't make complete sense right now, don't worry - you are going to work through the problems below to complete this problem. You can use **Quiz 5** in the classroom to make sure you are on the right track.

a. What is the **conversion rate** for p_{new} under the null?

```
In [19]: p_new = df2.query('converted == 1').shape[0] / df2.shape[0]
         p_new
```

```
Out[19]: 0.11959708724499628
```

b. What is the **conversion rate** for p_{old} under the null?

```
In [20]: p_old = df2.query('converted == 1').shape[0] / df2.shape[0]
         p_old
```

```
Out[20]: 0.11959708724499628
```

c. What is n_{new}, the number of individuals in the treatment group?

```
In [21]: n_new = df2.query('group == "treatment"').shape[0]
         n_new
```

```
Out[21]: 145310
```

d. What is n_{old}, the number of individuals in the control group?

```
In [22]: n_old = df2.query('group == "control"').shape[0]
         n_old
```

```
Out[22]: 145274
```

e. Simulate n_{new} transactions with a conversion rate of p_{new} under the null. Store these n_{new} 1's and 0's in **new_page_converted**.

```
In [23]: new_page_converted = np.random.choice(2, n_new, replace = True, p=[(1-p_new), p_new])
```

f. Simulate n_{old} transactions with a conversion rate of p_{old} under the null. Store these n_{old} 1's and 0's in **old_page_converted**.

```
In [24]: old_page_converted = np.random.choice(2, n_old, replace = True, p=[(1-p_old), p_old])
```

g. Find p_{new} - p_{old} for your simulated values from part (e) and (f).

```
In [25]: (new_page_converted.sum()/len(new_page_converted)) - (old_page_converted.sum()/len(old_
```

```
Out[25]: 0.0020214299932417995
```

h. Create 10,000 p_{new} - p_{old} values using the same simulation process you used in parts (a) through (g) above. Store all 10,000 values in a NumPy array called **p_diffs**.

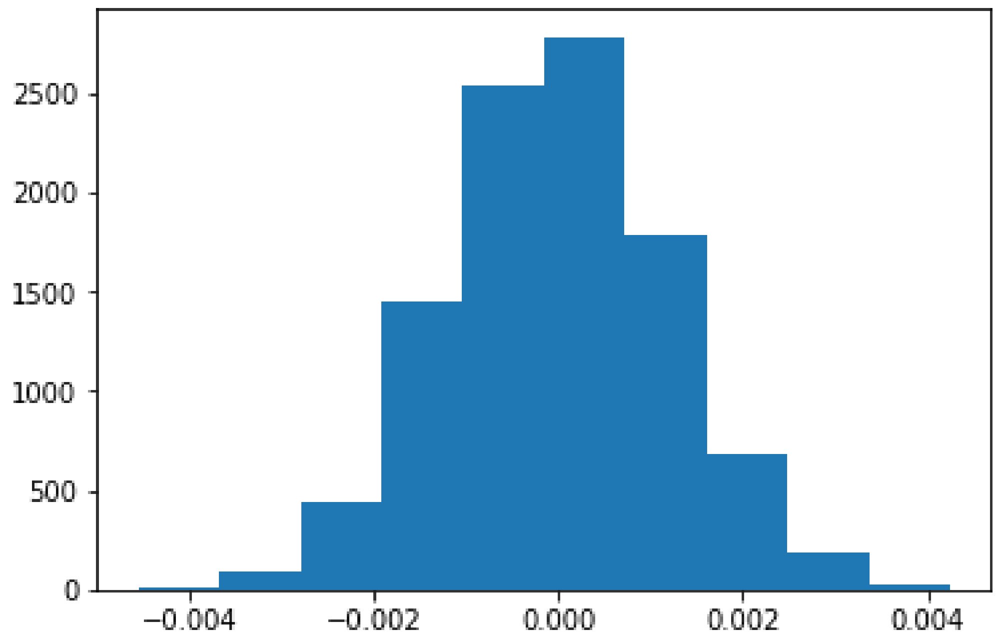
```
In [26]: p_diffs = []

         for _ in range (10000):
             new_page_converted = np.random.choice(2, n_new, replace = True, p=[(1-p_new), p_new
             old_page_converted = np.random.choice(2, n_old, replace = True, p=[(1-p_old), p_old
             p_diffs.append((new_page_converted.sum() / len(new_page_converted)) - (old_page_con
```

```
In [27]: p_diffs = np.array(p_diffs)
```

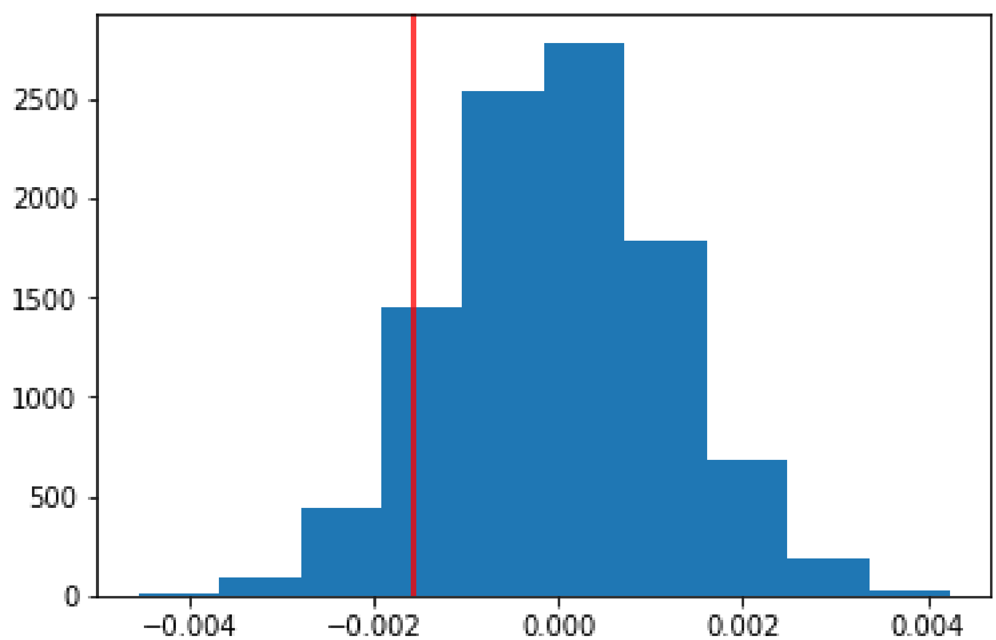
i. Plot a histogram of the **p_diffs**. Does this plot look like what you expected? Use the matching problem in the classroom to assure you fully understand what was computed here.

```
In [28]: plt.hist(p_diffs)
         plt.show()
```



j. What proportion of the **p_diffs** are greater than the actual difference observed in **ab_data.csv**?

In [29]:
```
plt.hist(p_diffs)
actual_diff = (df2.query('group == "treatment" & converted == 1').shape[0] / df2.query(
plt.axvline(x=actual_diff, color='red');
plt.show()
```



```
In [30]: (p_diffs > actual_diff).mean()
```

```
Out[30]: 0.90590000000000004
```

k. Please explain using the vocabulary you've learned in this course what you just computed in part **j.** What is this value called in scientific studies? What does this value mean in terms of whether or not there is a difference between the new and old pages?

The value called P-value. Since 0.95 > p-value > 0.05 and the statistic was in the bulk of the distrbution so it suggested that the statistic was likely come from the null hypothesis. Therefore, I have evidence to fail to reject the null.

l. We could also use a built-in to achieve similar results. Though using the built-in might be easier to code, the above portions are a walkthrough of the ideas that are critical to correctly thinking about statistical significance. Fill in the below to calculate the number of conversions for each page, as well as the number of individuals who received each page. Let `n_old` and `n_new` refer the the number of rows associated with the old page and new pages, respectively.

```
In [31]: import statsmodels.api as sm

         convert_old = df2.query('group == "control" & converted == 1').shape[0];
         convert_new = df2.query('group == "treatment" & converted == 1').shape[0];
         n_old = df2.query('group == "control"').shape[0];


         n_new = df2.query('group == "treatment"').shape[0];
```

```
/opt/conda/lib/python3.6/site-packages/statsmodels/compat/pandas.py:56: FutureWarning: The panda
  from pandas.core import datetools
```

m. Now use `stats.proportions_ztest` to compute your test statistic and p-value. Here is a helpful link on using the built in.

```
In [32]: values = sm.stats.proportions_ztest(np.array([convert_new , convert_old]), np.array([n_
         values
```

```
Out[32]: (-1.3109241984234394, 0.90505831275902449)
```

n. What do the z-score and p-value you computed in the previous question mean for the conversion rates of the old and new pages? Do they agree with the findings in parts **j.** and **k.**?

The critical Z score values when using a 95% confidence level are -1.96 and +1.96 standard deviations and the p-value associated with a 95% confidence level is 0.05 (as mentioned in the below site, and it is the same confidence level I used). The z-score value mean that it is 1.3

standard deviations below the mean. The p-value computed here shown the same value in part j, it is a large value which suggested to stay with the null hypothesis (fail to reject the null hypothesis).

I used this site as a resource to my answer: http://resources.esri.com/help/9.3/arcgisengine/java/gp_toolref

Part III - A regression approach

`1.` In this final part, you will see that the result you achieved in the A/B test in Part II above can also be achieved by performing regression.

a. Since each row is either a conversion or no conversion, what type of regression should you be performing in this case?

Logistic regression is the type of regression shold be used, becuase it used to predict only two possible outcomes.

b. The goal is to use **statsmodels** to fit the regression model you specified in part **a.** to see if there is a significant difference in conversion based on which page a customer receives. However, you first need to create in df2 a column for the intercept, and create a dummy variable column for which page each user received. Add an **intercept** column, as well as an **ab_page** column, which is 1 when an individual receives the **treatment** and 0 if **control**.

```
In [33]: df2['intercept'] = 1
         df2['ab_page'] = pd.get_dummies(df2['group'])['treatment']
```

```
In [34]: df2.head()
```

```
Out[34]:    user_id                   timestamp      group landing_page  converted  \
         0   851104  2017-01-21 22:11:48.556739    control     old_page          0
         1   804228  2017-01-12 08:01:45.159739    control     old_page          0
         2   661590  2017-01-11 16:55:06.154213  treatment     new_page          0
         3   853541  2017-01-08 18:28:03.143765  treatment     new_page          0
         4   864975  2017-01-21 01:52:26.210827    control     old_page          1

            intercept  ab_page
         0          1        0
         1          1        0
         2          1        1
         3          1        1
         4          1        0
```

c. Use **statsmodels** to instantiate your regression model on the two columns you created in part b., then fit the model using the two columns you created in part **b.** to predict whether or not an individual converts.

```
In [35]: log_mod = sm.Logit(df2['converted'], df2[['intercept','ab_page']])
         results = log_mod.fit()
```

```
Optimization terminated successfully.
         Current function value: 0.366118
         Iterations 6
```

d. Provide the summary of your model below, and use it as necessary to answer the following questions.

```
In [36]: results.summary2()
```

```
Out[36]: <class 'statsmodels.iolib.summary2.Summary'>
"""
                          Results: Logit
=================================================================
Model:              Logit            No. Iterations:   6.0000
Dependent Variable: converted        Pseudo R-squared: 0.000
Date:               2020-11-12 17:47 AIC:              212780.3502
No. Observations:   290584           BIC:              212801.5095
Df Model:           1                Log-Likelihood:   -1.0639e+05
Df Residuals:       290582           LL-Null:          -1.0639e+05
Converged:          1.0000           Scale:            1.0000
------------------------------------------------------------------
              Coef.  Std.Err.      z      P>|z|   [0.025   0.975]
------------------------------------------------------------------
intercept   -1.9888    0.0081 -246.6690 0.0000 -2.0046 -1.9730
ab_page     -0.0150    0.0114   -1.3109 0.1899 -0.0374  0.0074
=================================================================

"""
```

e. What is the p-value associated with **ab_page**? Why does it differ from the value you found in **Part II**? **Hint**: What are the null and alternative hypotheses associated with your regression model, and how do they compare to the null and alternative hypotheses in **Part II**?

The p-value associated with the ab_page is 0.189.
The p-value is different than the p-value in part II, since the null and alternative hypothesis are differ in this part from the part II. Based on this site: http://ismayc.github.io/teaching/sample_problems/multiple_logistic.html , the hypothesis associated with the logistic regression are:

$$H_0 : i = 0$$

$$H_1 : i0$$

f. Now, you are considering other things that might influence whether or not an individual converts. Discuss why it is a good idea to consider other factors to add into your regression model. Are there any disadvantages to adding additional terms into your regression model?

Adding more factors will lead to more detailes to the result whether or not an individual converts. However, as a disadvantage of adding more factors we may have a collinearity and we will have a complicated relationship between the variabels, which could be avoided as this will cause over adjustement.

g. Now along with testing if the conversion rate changes for different pages, also add an effect based on which country a user lives in. You will need to read in the **countries.csv** dataset and merge together your datasets on the appropriate rows. Here are the docs for joining tables.

Does it appear that country had an impact on conversion? Don't forget to create dummy variables for these country columns - **Hint: You will need two columns for the three dummy variables.** Provide the statistical output as well as a written response to answer this question.

```
In [59]: df_countries = pd.read_csv('countries.csv')
         new_df = df_countries.set_index('user_id').join(df2.set_index('user_id'))
```

```
In [66]: new_df[['CA','UK','US']] = pd.get_dummies(new_df['country'])
         new_df = new_df.drop(['US'] , axis = 1)
         new_df.head()
```

```
Out[66]:         country                   timestamp      group landing_page  \
         user_id
         834778       UK  2017-01-14 23:08:43.304998    control     old_page
         928468       US  2017-01-23 14:44:16.387854  treatment     new_page
         822059       UK  2017-01-16 14:04:14.719771  treatment     new_page
         711597       UK  2017-01-22 03:14:24.763511    control     old_page
         710616       UK  2017-01-16 13:14:44.000513  treatment     new_page

                  converted  intercept  ab_page  CA  UK
         user_id
         834778           0          1        0   0   1
         928468           0          1        1   0   0
         822059           1          1        1   0   1
         711597           0          1        0   0   1
         710616           0          1        1   0   1
```

```
In [67]: logit = sm.Logit(new_df['converted'], new_df[['intercept', 'CA', 'UK']])
         results = logit.fit()
         results.summary2()
```

```
Optimization terminated successfully.
         Current function value: 0.366116
         Iterations 6
```

```
Out[67]: <class 'statsmodels.iolib.summary2.Summary'>
         """
                                   Results: Logit
         ==================================================================
         Model:              Logit            No. Iterations:    6.0000
         Dependent Variable: converted        Pseudo R-squared: 0.000
         Date:               2020-11-12 19:10 AIC:              212780.8333
         No. Observations:   290584           BIC:              212812.5723
```

```
Df Model:           2                Log-Likelihood:    -1.0639e+05
Df Residuals:       290581           LL-Null:           -1.0639e+05
Converged:          1.0000           Scale:             1.0000
-------------------------------------------------------------------
               Coef.   Std.Err.      z      P>|z|    [0.025   0.975]
-------------------------------------------------------------------
intercept    -1.9967    0.0068  -292.3145  0.0000  -2.0101  -1.9833
CA           -0.0408    0.0269    -1.5178  0.1291  -0.0935   0.0119
UK            0.0099    0.0133     0.7458  0.4558  -0.0161   0.0360
==================================================================

"""
```

In [55]:
```
1/np.exp(-0.0408) , np.exp(0.0099)
```

Out[55]:
```
(1.0416437559600236, 1.0099491671175422)
```

Based on the summary above, it appears that country had an impact on conversion. Moreover, as US baseline, the individual in CA is 1.041 times less likely to convert, whereas in UK is 1.009 times more likely to convert.

h. Though you have now looked at the individual factors of country and page on conversion, we would now like to look at an interaction between page and country to see if there significant effects on conversion. Create the necessary additional columns, and fit the new model.

Provide the summary results, and your conclusions based on the results.

In [68]:
```
new_df['CA_new_page'] = new_df['CA'] * new_df['ab_page']
new_df['UK_new_page'] = new_df['UK'] * new_df['ab_page']

new_df.head()
```

Out[68]:
```
        country                   timestamp      group landing_page  \
user_id
834778       UK  2017-01-14 23:08:43.304998    control     old_page
928468       US  2017-01-23 14:44:16.387854  treatment     new_page
822059       UK  2017-01-16 14:04:14.719771  treatment     new_page
711597       UK  2017-01-22 03:14:24.763511    control     old_page
710616       UK  2017-01-16 13:14:44.000513  treatment     new_page

         converted  intercept  ab_page  CA  UK  CA_new_page  UK_new_page
user_id
834778           0          1        0   0   1            0            0
928468           0          1        1   0   0            0            0
822059           1          1        1   0   1            0            1
711597           0          1        0   0   1            0            0
710616           0          1        1   0   1            0            1
```

In [69]:
```
logit = sm.Logit(new_df['converted'], new_df[['intercept','CA_new_page', 'UK_new_page']
results = logit.fit()
results.summary2()
```

```
Optimization terminated successfully.
         Current function value: 0.366113
         Iterations 6
```

Out[69]:
```
<class 'statsmodels.iolib.summary2.Summary'>
"""
                          Results: Logit
==================================================================
Model:              Logit            No. Iterations:   6.0000
Dependent Variable: converted        Pseudo R-squared: 0.000
Date:               2020-11-12 19:12 AIC:              212779.0384
No. Observations:   290584           BIC:              212810.7773
Df Model:           2                Log-Likelihood:   -1.0639e+05
Df Residuals:       290581           LL-Null:          -1.0639e+05
Converged:          1.0000           Scale:            1.0000
-------------------------------------------------------------------
              Coef.   Std.Err.      z      P>|z|   [0.025   0.975]
-------------------------------------------------------------------
intercept    -1.9963    0.0062  -322.0487  0.0000  -2.0084  -1.9841
CA_new_page  -0.0752    0.0376    -1.9974  0.0458  -0.1489  -0.0014
UK_new_page   0.0149    0.0173     0.8617  0.3888  -0.0190   0.0488
==================================================================

"""
```

In [70]:
```
1/np.exp(-0.0752) , np.exp(0.0149)
```

Out[70]:
```
(1.0780997492739288, 1.0150115583846535)
```

Based on the summary above, as US baseline, the individual in CA that recvives treatment page is 1.078 times less likely to convert, whereas in UK that recvives treatment page is 1.015 times more likely to convert.

In [39]:
```
from subprocess import call
call(['python', '-m', 'nbconvert', 'Analyze_ab_test_results_notebook.ipynb'])
```

Out[39]:
```
0
```