Hyppthesis_Testing_With_Python

September 11, 2024

1 Hypothesis testing with Python

1.1 Introduction

As you've been learning, analysis of variance (commonly called ANOVA) is a group of statistical techniques that test the difference of means among three or more groups. It's a powerful tool for determining whether population means are different across groups and for answering a wide range of business questions.

In this activity, you are a data professional working with historical marketing promotion data. You will use the data to run a one-way ANOVA and a post hoc ANOVA test. Then, you will communicate your results to stakeholders. These experiences will help you make more confident recommendations in a professional setting.

In your dataset, each row corresponds to an independent marketing promotion, where your business uses TV, social media, radio, and influencer promotions to increase sales. You have previously provided insights about how different promotion types affect sales; now stakeholders want to know if sales are significantly different among various TV and influencer promotion types.

To address this request, a one-way ANOVA test will enable you to determine if there is a statistically significant difference in sales among groups. This includes: * Using plots and descriptive statistics to select a categorical independent variable * Creating and fitting a linear regression model with the selected categorical independent variable * Checking model assumptions * Performing and interpreting a one-way ANOVA test * Comparing pairs of groups using an ANOVA post hoc test * Interpreting model outputs and communicating the results to nontechnical stakeholders

1.2 Step 1: Imports

Import pandas, pyplot from matplotlib, seaborn, api from statsmodels, ols from statsmodels.formula.api, and pairwise_tukeyhsd from statsmodels.stats.multicomp.

[15]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.multicomp import pairwise_tukeyhsd
```

Pandas was used to load the dataset marketing_sales_data.csv as data, now display the first five rows. The variables in the dataset have been adjusted to suit the objectives of this lab. As shown in this cell, the dataset has been automatically loaded in for you. You do not need to download the .csv file, or provide more code, in order to access the dataset and proceed with this lab. Please continue with this activity by completing the following instructions.

```
[27]: data = pd.read_csv('marketing_sales_data.csv')
      data.head()
```

```
[27]:      TV      Radio  Social Media Influencer       Sales
      0     Low   1.218354      1.270444      Micro   90.054222
      1  Medium  14.949791      0.274451      Macro  222.741668
      2     Low  10.377258      0.061984       Mega  102.774790
      3    High  26.469274      7.070945      Micro  328.239378
      4    High  36.876302      7.618605       Mega  351.807328
```

The features in the data are: * TV promotion budget (in Low, Medium, and High categories) * Social media promotion budget (in millions of dollars) * Radio promotion budget (in millions of dollars) * Sales (in millions of dollars) * Influencer size (in Mega, Macro, Nano, and Micro categories)

Question: Why is it useful to perform exploratory data analysis before constructing a linear regression model?

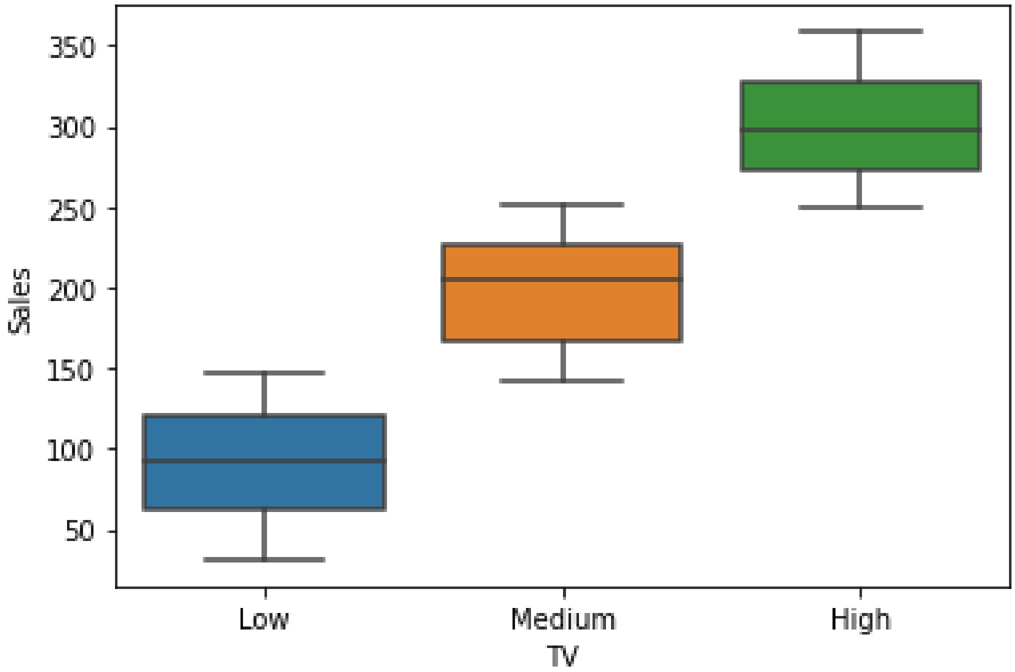
- To understand which variables are present in the data
- To consider the distribution of features, such as minimum, mean, and maximum values
- To plot the relationship between the independent and dependent variables and visualize which features have a linear relationship
- To identify issues with the data, such as incorrect or missing values.

1.3 Step 2: Data exploration

First, use a boxplot to determine how Sales vary based on the TV promotion budget category.

```
[4]: sns.boxplot(x= 'TV', y='Sales', data = data)
```

```
[4]: <matplotlib.axes._subplots.AxesSubplot at 0x70c0b7a3e0d0>
```



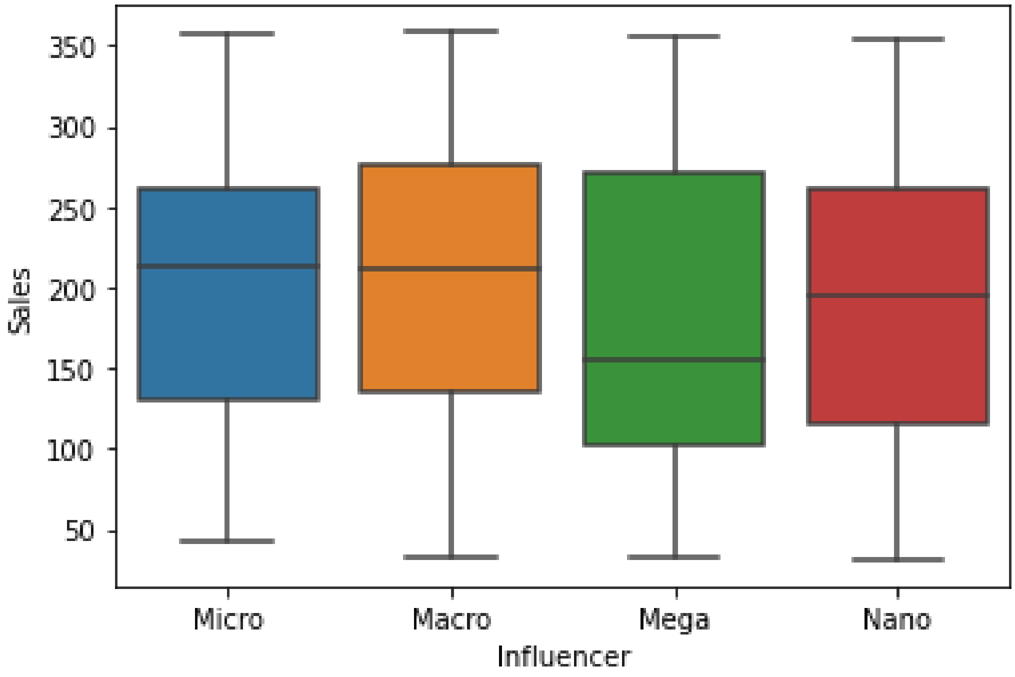
Question: Is there variation in `Sales` based off the `TV` promotion budget?

There is considerable variation in `Sales` across the `TV` groups. The significance of these differences can be tested with a one-way ANOVA.

Now, use a boxplot to determine how `Sales` vary based on the `Influencer` size category.

```
[7]: sns.boxplot(x= 'Influencer', y='Sales', data = data)
```

```
[7]: <matplotlib.axes._subplots.AxesSubplot at 0x70c0b575f450>
```



Question: Is there variation in `Sales` based off the `Influencer` size?

There is some variation in `Sales` across the `Influencer` groups, but it may not be significant.

1.3.1 Remove missing data

You may recall from prior labs that this dataset contains rows with missing values. To correct this, drop these rows. Then, confirm the data contains no missing values.

```
[11]: data = data.dropna(axis=0)

      data.isnull().sum(axis=0)
```

```
[11]: TV              0
      Radio           0
      Social Media    0
      Influencer      0
      Sales           0
      dtype: int64
```

1.4 Step 3: Model building

Fit a linear regression model that predicts `Sales` using one of the independent categorical variables in `data`. Refer to your previous code for defining and fitting a linear regression model.

[13]:

```python
ols_formula = 'Sales ~ C(TV)'

OLS = ols(formula = ols_formula, data = data)

model = OLS.fit()

model_summary = model.summary()

model_summary
```

[13]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.874
Model:                            OLS   Adj. R-squared:                  0.874
Method:                 Least Squares   F-statistic:                     1971.
Date:                Tue, 13 Aug 2024   Prob (F-statistic):          8.81e-256
Time:                        18:22:55   Log-Likelihood:                -2778.9
No. Observations:                 569   AIC:                             5564.
Df Residuals:                     566   BIC:                             5577.
Df Model:                           2
Covariance Type:            nonrobust
================================================================================
===
                      coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
---
Intercept         300.5296      2.417    124.360      0.000     295.783
305.276
C(TV)[T.Low]     -208.8133      3.329    -62.720      0.000    -215.353
-202.274
C(TV)[T.Medium]  -101.5061      3.325    -30.526      0.000    -108.038
-94.975
==============================================================================
Omnibus:                      450.714   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               35.763
Skew:                          -0.044   Prob(JB):                     1.71e-08
Kurtosis:                       1.775   Cond. No.                         3.86
==============================================================================

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

Question: Which categorical variable did you choose for the model? Why?

- TV was selected as the preceding analysis showed a strong relationship between the TV promotion budget and the average Sales.
- Influencer was not selected because it did not show a strong relationship to Sales in the analysis.

1.4.1 Check model assumptions

Now, check the four linear regression assumptions are upheld for your model.

Question: Is the linearity assumption met?

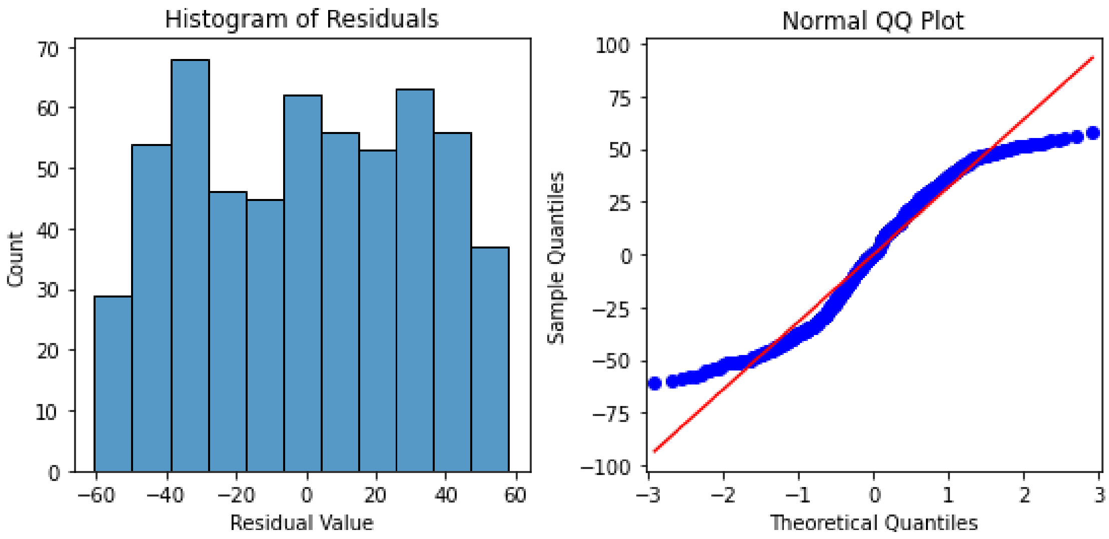
Because the model does not have any continuous independent variables, the linearity assumption is not required.

The independent observation assumption states that each observation in the dataset is independent. As each marketing promotion (row) is independent from one another, the independence assumption is not violated.

Next, verify that the normality assumption is upheld for the model.

[16]:

```
residuals = model.resid

fig, axes = plt.subplots(1, 2, figsize = (8,4))

sns.histplot(residuals, ax=axes[0])

axes[0].set_xlabel("Residual Value")

axes[0].set_title("Histogram of Residuals")



sm.qqplot(residuals, line='s',ax = axes[1])

axes[1].set_title("Normal QQ Plot")

plt.tight_layout()

plt.show()
```



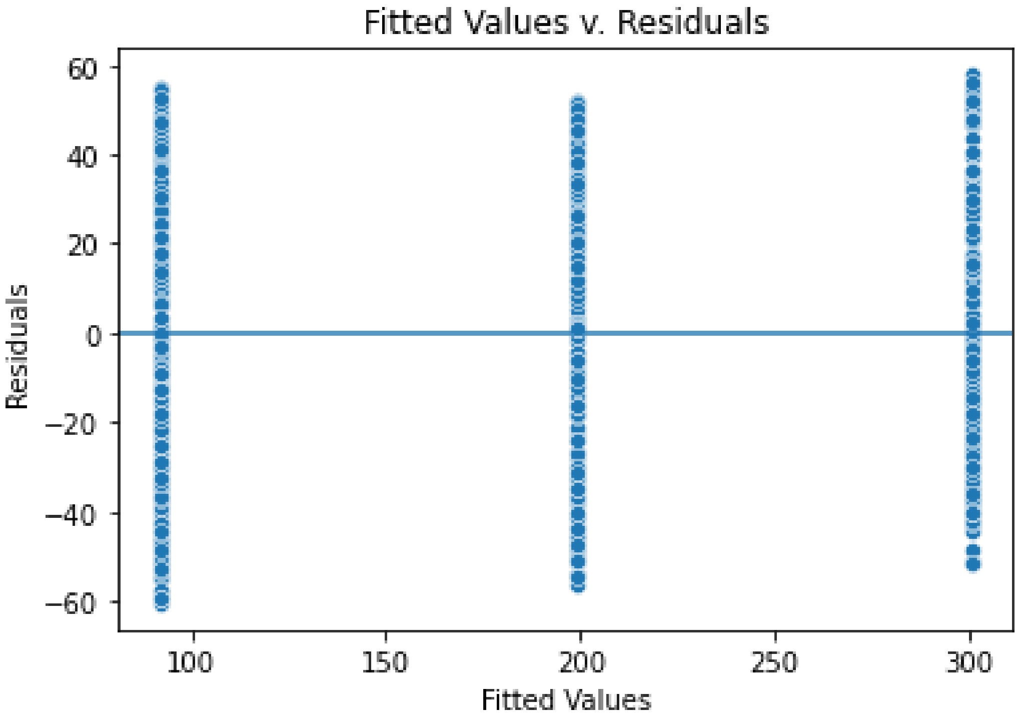
Question: Is the normality assumption met?

There is reasonable concern that the normality assumption is not met when `TV` is used as the independent variable predicting `Sales`. The normal q-q forms an 'S' that deviates off the red diagonal line, which is not desired behavior.

Now, verify the constant variance (homoscedasticity) assumption is met for this model.

[17]:

```
fig = sns.scatterplot(x = model.fittedvalues, y = model.resid)

fig.set_xlabel("Fitted Values")

fig.set_ylabel("Residuals")

fig.set_title("Fitted Values v. Residuals")

fig.axhline(0)

plt.show()
```



Question: Is the constant variance (homoscedasticity) assumption met?

The variance where there are fitted values is similarly distributed, validating that the constant variance assumption is met.

1.5 Step 4: Results and evaluation

First, display the OLS regression results.

[19]:
```
model.summary()
```

[19]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.874
Model:                            OLS   Adj. R-squared:                  0.874
Method:                 Least Squares   F-statistic:                     1971.
Date:                Tue, 13 Aug 2024   Prob (F-statistic):          8.81e-256
Time:                        18:30:34   Log-Likelihood:                -2778.9
No. Observations:                 569   AIC:                             5564.
Df Residuals:                     566   BIC:                             5577.
Df Model:                           2
Covariance Type:            nonrobust
```

```
================================================================================
===
                      coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
---
Intercept         300.5296      2.417    124.360      0.000     295.783
305.276
C(TV)[T.Low]     -208.8133      3.329    -62.720      0.000    -215.353
-202.274
C(TV)[T.Medium]  -101.5061      3.325    -30.526      0.000    -108.038
-94.975
==============================================================================
Omnibus:                      450.714   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               35.763
Skew:                          -0.044   Prob(JB):                     1.71e-08
Kurtosis:                       1.775   Cond. No.                         3.86
==============================================================================

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

Question: What is your interpretation of the model's R-squared?

Using TV as the independent variable results in a linear regression model with $R^2 = 0.874$. In other words, the model explains 87.4% of the variation in Sales. This makes the model an effective predictor of Sales.

Question: What is your intepretation of the coefficient estimates? Are the coefficients statistically significant?

The default TV category for the model is High, because there are coefficients for the other two TV categories, Medium and Low. According to the model, Sales with a Medium or Low TV category are lower on average than Sales with a High TV category. For example, the model predicts that a Low TV promotion would be 208.813 (in millions of dollars) lower in Sales on average than a High TV promotion.

The p-value for all coefficients is 0.000, meaning all coefficients are statistically significant at $p = 0.05$. The 95% confidence intervals for each coefficient should be reported when presenting results to stakeholders. For instance, there is a 95% chance the interval $[-215.353, -202.274]$ contains the true parameter of the slope of β_{TVLow}, which is the estimated difference in promotion sales when a Low TV promotion is chosen instead of a High TV promotion.

Question: Do you think your model could be improved? Why or why not? How?

Given how accurate TV was as a predictor, the model could be improved with a more granular view of the TV promotions, such as additional categories or the actual TV promotion budgets. Further, additional variables, such as the location of the marketing campaign or the time of year, may increase model accuracy.

1.5.1 Perform a one-way ANOVA test

With the model fit, run a one-way ANOVA test to determine whether there is a statistically significant difference in Sales among groups.

[20]:

```
sm.stats.anova_lm(model, type = 2)
```

[20]:

```
              df        sum_sq       mean_sq            F         PR(>F)
C(TV)        2.0  4.052692e+06  2.026346e+06  1971.455737  8.805550e-256
Residual   566.0  5.817589e+05  1.027843e+03          NaN            NaN
```

Question: What are the null and alternative hypotheses for the ANOVA test?

The null hypothesis is that there is no difference in Sales based on the TV promotion budget.

The alternative hypothesis is that there is a difference in Sales based on the TV promotion budget.

Question: What is your conclusion from the one-way ANOVA test?

The results of the one-way ANOVA test indicate that you can reject the null hypothesis in favor of the alternative hypothesis. There is a statistically significant difference in Sales among TV groups.

Question: What did the ANOVA test tell you?

The results of the one-way ANOVA test indicate that you can reject the null hypothesis in favor of the alternative hypothesis. There is a statistically significant difference in Sales among TV groups.

1.5.2 Perform an ANOVA post hoc test

If you have significant results from the one-way ANOVA test, you can apply ANOVA post hoc tests such as the Tukey's HSD post hoc test.

Run the Tukey's HSD post hoc test to compare if there is a significant difference between each pair of categories for TV.

[26]:

```
tukey_oneway = pairwise_tukeyhsd(endog = data['Sales'], groups = data['TV'])

tukey_oneway.summary()
```

[26]:

```
<class 'statsmodels.iolib.table.SimpleTable'>
```

Question: What is your interpretation of the Tukey HSD test?

The first row, which compares the High and Low TV groups, indicates that you can reject the null hypothesis that there is no significant difference between the Sales of these two groups.

You can also reject the null hypotheses for the two other pairwise comparisons that compare High to Medium and Low to Medium.

Question: What did the post hoc tell you?**

A post hoc test was conducted to determine which `TV` groups are different and how many are different from each other. This provides more detail than the one-way ANOVA results, which can at most determine that at least one group is different. Further, using the Tukey HSD controls for the increasing probability of incorrectly rejecting a null hypothesis from peforming multiple tests.

The results were that `Sales` is not the same between any pair of `TV` groups.

1.6 Considerations

What are some key takeaways that you learned during this lab?

[Write your response here. Double-click (or enter) to edit.]

What summary would you provide to stakeholders? Consider the statistical significance of key relationships and differences in distribution.

[Write your response here. Double-click (or enter) to edit.]

Reference Saragih, H.S. *Dummy Marketing and Sales Data*

Congratulations! You've completed this lab. However, you may not notice a green check mark next to this item on Coursera's platform. Please continue your progress regardless of the check mark. Just click on the "save" icon at the top of this notebook to ensure your work has been logged.