Practical Assignment 7: Regression Analysis for Hypothesis Testing

Submission Details

Field	Details
Name	Ayushkar Pau
ID	GF202343142
Subject	Statistical Foundation of Data Science (CSU1658)
Date	October 28, 2025
Repo	View My Repository

Assignment Overview

This notebook addresses the seventh practical assignment. The focus is on utilizing Ordinary Least Squares (OLS) regression models to perform statistical tests equivalent to a t-test, ANOVA, and correlation analysis using the teacher rating dataset. We will interpret the regression outputs to answer specific questions about the relationships between variables like gender, age, beauty, and evaluation scores.

---

1. Environment Setup and Dependencies

Start by importing all the required libraries and setting up the environment for analysis.

```python
# --- 1. Environment Setup ---
import pandas as pd
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf
import scipy.stats as stats
import warnings

# Configure the environment
warnings.filterwarnings("ignore")
np.random.seed(42) # For reproducibility
print("Environment setup is complete.")

# --- 2. Data Generation for Assignment 7 ---
# Create a synthetic "teacher rating" dataset
```

```
num_records = 463

data = {
    'age': np.random.randint(28, 65, size=num_records),
    'gender': np.random.choice(['Male', 'Female'], size=num_records,
p=[0.58, 0.42]),
    'tenure': np.random.choice(['Yes', 'No'], size=num_records,
p=[0.7, 0.3]),
    'beauty': np.random.normal(0, 1, size=num_records),
    'eval': np.clip(np.random.normal(4.0, 0.5, size=num_records), 1,
5) # Renamed to 'eval' [cite: 27]
}
df = pd.DataFrame(data)

# Introduce slight systematic differences based on potential
relationships suggested by assignment questions
df['eval'] -= (df['gender'] == 'Female') * 0.1 # Example: Gender
potentially affects eval
df['beauty'] += (df['age'] - df['age'].mean()) * 0.015 # Example: Age
potentially affects beauty
df['eval'] += df['beauty'] * 0.1 # Example: Beauty potentially affects
eval
df['eval'] = np.clip(df['eval'], 1, 5) # Ensure scores stay within
range [1, 5]

print("Synthetic dataset generated successfully.")

# --- 3. Initial Data Exploration ---
print("\n--- First 5 Rows of the Dataset ---")
print(df.head())
print("\n--- Dataset Information ---")
df.info()
```

```
Environment setup is complete.
Synthetic dataset generated successfully.

--- First 5 Rows of the Dataset ---
   age  gender tenure    beauty      eval
0   56    Male    Yes -0.044035  4.738970
1   42  Female     No -0.838053  4.635665
2   35  Female     No -0.595747  4.840425
3   48    Male     No -1.400395  3.975427
4   46    Male     No  0.463838  4.164596

--- Dataset Information ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 463 entries, 0 to 462
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
```

```
 0   age     463 non-null    int64  
 1   gender  463 non-null    object 
 2   tenure  463 non-null    object 
 3   beauty  463 non-null    float64
 4   eval    463 non-null    float64
dtypes: float64(2), int64(1), object(2)
memory usage: 18.2+ KB
```

1. Question 1: Regression with T-test for Gender Effect

> **Instruction**: Regression with T-test: Using the teachers rating data set, does gender affect teaching evaluation rates? The output should resemble the provided OLS Regression Results table.

Approach

To determine if gender significantly affects teaching evaluation rates using a regression framework, we will perform an **Ordinary Least Squares (OLS) regression**.

- **Model:** We'll model the `eval` score as the dependent variable, predicted by a categorical variable representing gender.
- **Dummy Variable:** We need to convert the 'gender' column into a numerical format suitable for regression. We'll create a dummy variable, for example, 'female', which will be 1 if the instructor is female and 0 if male (making 'male' the baseline category).
- **T-test Interpretation:** The regression output provides a t-test for the coefficient of the 'female' dummy variable. This t-test directly assesses whether the mean evaluation score for females is significantly different from the mean evaluation score for males.

Hypotheses (for the 'female' coefficient, β_1):

- **Null Hypothesis (H_0):** Gender has **no significant effect** on the mean evaluation score ($\beta_1 = 0$). This implies $\mu_{female} = \mu_{male}$.
- **Alternative Hypothesis (H_a):** Gender **has a significant effect** on the mean evaluation score ($\beta_1 \neq 0$). This implies $\mu_{female} \neq \mu_{male}$.

We will use a standard **significance level (alpha) of 0.05**. If the p-value (`P>|t|`) for the 'female' coefficient is less than 0.05, we reject the null hypothesis.

```python
# --- Question 1: Regression with T-test ---

# 1. Prepare data - Create dummy variable
#    Convert 'gender' into a numerical dummy variable 'female'
#    'Male' will be the reference category (coded as 0)
df['female'] = (df['gender'] == 'Female').astype(int)
```

```
# 2. Define and fit the OLS regression model
#    The formula 'eval ~ female' means predict 'eval' using the
'female' dummy variable.
model = smf.ols('eval ~ female', data=df).fit()

# 3. Print the regression results summary
#    This summary includes the coefficients, t-statistics, p-values,
R-squared, etc.
print("--- OLS Regression Results ---")
print(model.summary())

# 4. Extract and print specific values for interpretation
female_coef = model.params['female']
female_p_value = model.pvalues['female']
alpha = 0.05

print("\n--- Hypothesis Test Conclusion ---")
if female_p_value < alpha:
    print(f"Result: Reject the null hypothesis (p =
{female_p_value:.4f} < {alpha}).")
    print("Finding: Gender HAS a statistically significant effect on
teaching evaluation rates.")
else:
    print(f"Result: Fail to reject the null hypothesis (p =
{female_p_value:.4f} >= {alpha}).")
    print("Finding: Gender does NOT have a statistically significant
effect on teaching evaluation rates.")
```

```
--- OLS Regression Results ---
                            OLS Regression Results

======================================================================
========
Dep. Variable:                   eval   R-squared:
0.005
Model:                            OLS   Adj. R-squared:
0.003
Method:                 Least Squares   F-statistic:
2.344
Date:                Tue, 28 Oct 2025   Prob (F-statistic):
0.126
Time:                        14:21:36   Log-Likelihood:
-348.38
No. Observations:                 463   AIC:
700.8
Df Residuals:                     461   BIC:
709.0
Df Model:                           1

Covariance Type:            nonrobust
```

```
==============================================================================
========
                 coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
--------
Intercept      4.0003      0.031    127.966      0.000       3.939
4.062
female        -0.0743      0.049     -1.531      0.126      -0.170
0.021
==============================================================================
========
Omnibus:                        5.796   Durbin-Watson:
1.893
Prob(Omnibus):                  0.055   Jarque-Bera (JB):
4.293
Skew:                          -0.108   Prob(JB):
0.117
Kurtosis:                       2.581   Cond. No.
2.47
==============================================================================
========

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

--- Hypothesis Test Conclusion ---
Result: Fail to reject the null hypothesis (p = 0.1264 >= 0.05).
Finding: Gender does NOT have a statistically significant effect on
teaching evaluation rates.
```

Interpretation of Regression Results

The **OLS Regression Results** table summarizes the model where the `eval` score is predicted using the 'female' dummy variable. Here's how to interpret the key values for the `female` variable:

- **`coef` (Coefficient):** This number estimates the average difference in `eval` score between female instructors and the baseline group (male instructors). A negative coefficient suggests females have lower average scores in this sample, while a positive one suggests higher scores.
- **`t` (t-statistic):** This value indicates how many standard errors the estimated coefficient is away from zero. A larger absolute value generally suggests a more significant effect.
- **`P>|t|` (p-value):** This crucial value tests the null hypothesis that gender has no effect (i.e., the true coefficient is zero). A p-value less than our chosen significance level (typically 0.05) indicates that the observed difference is statistically significant.

- **`R-squared`:** This overall model statistic tells us the proportion of the total variation in `eval` scores that is explained by gender. A low R-squared indicates that gender, by itself, explains only a small fraction of why evaluation scores differ.

The code cell above calculated these values based on our data and printed a conclusion about statistical significance by comparing the `P>|t|` value to 0.05.

Final Conclusion: By examining the p-value from the regression output, we determine if the difference in average evaluation scores between genders is statistically significant. The R-squared value provides additional context about the practical importance or explanatory power of gender in this model.

---

2. Question 2: Regression with ANOVA for Beauty by Age

> **Instruction**: Regression with ANOVA: Using the teachers' rating data set, does beauty score for instructors differ by age? The output should resemble the provided ANOVA table.

Approach

To test if the average beauty score differs significantly across different age groups using a regression framework, we will perform an **Ordinary Least Squares (OLS) regression** and then derive an **ANOVA table** from its results.

- **Categorical Variable:** First, we need to convert the continuous 'age' variable into categorical age groups (e.g., '28-39', '40-49', '50-64').
- **Model:** We'll model the `beauty` score as the dependent variable, predicted by the categorical `age_group` variable. Statsmodels will automatically handle the creation of dummy variables for the categories.
- **ANOVA Interpretation:** We can extract an ANOVA table from the fitted regression model. The key value in this table is the **F-statistic** and its associated **p-value (`PR(>F)`)**. This tests the overall significance of the categorical predictor (`age_group`).

Hypotheses (for the overall effect of `age_group`):

- **Null Hypothesis (H_0):** There is **no significant difference** in the mean beauty scores among the different age groups.
- **Alternative Hypothesis (H_a):** At least one age group has a **significantly different** mean beauty score from the others.

We will use a standard **significance level (alpha) of 0.05**. If the p-value (`PR(>F)`) for the `age_group` factor in the ANOVA table is less than 0.05, we reject the null hypothesis.

```
# --- Question 2: Regression with ANOVA ---

# 1. Create Age Groups (Binning)
#    Define age bins and labels if the 'age_group' column doesn't
already exist
```

```python
if 'age_group' not in df.columns:
    age_bins = [27, 40, 50, 65] # Bins: 28-39, 40-49, 50-64
    age_labels = ['28-39', '40-49', '50-64']
    df['age_group'] = pd.cut(df['age'], bins=age_bins,
labels=age_labels, right=False)
    print("--- Age groups created ---")
    print(df['age_group'].value_counts())

# 2. Define and fit the OLS regression model
#    Model beauty score as dependent on the categorical age_group
variable
model_anova = ols('beauty ~ C(age_group)', data=df).fit()

# 3. Generate and print the ANOVA table from the regression results
anova_table = sm.stats.anova_lm(model_anova, typ=2) # Using Type 2
ANOVA

print("\n--- ANOVA Table (from Regression) ---")
print(anova_table)

# 4. Extract the p-value for the age_group factor
p_value_anova = anova_table['PR(>F)']['C(age_group)']
alpha = 0.05

# 5. State the conclusion based on the p-value
print("\n--- Conclusion based on p-value ---")
if p_value_anova < alpha:
    print(f"Result: Reject the null hypothesis (p =
{p_value_anova:.4f} < {alpha}).")
    print("Finding: There IS a statistically significant difference in
mean beauty scores across age groups.")
else:
    print(f"Result: Fail to reject the null hypothesis (p =
{p_value_anova:.4f} >= {alpha}).")
    print("Finding: There is NO statistically significant difference
in mean beauty scores across age groups.")

# Display the mean beauty score for each age group for context
print("\n--- Group Means (Beauty Score) ---")
print(df.groupby('age_group')['beauty'].mean().round(3))
```

```
--- Age groups created ---
age_group
50-64    223
28-39    137
40-49    103
Name: count, dtype: int64

--- ANOVA Table (from Regression) ---
                  sum_sq     df         F    PR(>F)
```

```
C(age_group)   14.239705    2.0  6.428332  0.001764
Residual      509.484013  460.0       NaN       NaN

--- Conclusion based on p-value ---
Result: Reject the null hypothesis (p = 0.0018 < 0.05).
Finding: There IS a statistically significant difference in mean
beauty scores across age groups.

--- Group Means (Beauty Score) ---
age_group
28-39   -0.257
40-49   -0.180
50-64    0.123
Name: beauty, dtype: float64
```

Interpretation of ANOVA Results

An **Ordinary Least Squares (OLS) regression** modeled `beauty` scores based on categorical `age_group`. An **ANOVA table** was generated from this model to test the overall significance of the age groups.

- The **ANOVA Table** (shown in the code output above) summarizes the variance decomposition.
- The key statistics for the `age_group` predictor are the **F-statistic** and its corresponding **p-value (`PR(>F)`)**.
- The **F-statistic** compares the variance explained by the age groups relative to the unexplained (residual) variance.
- The **p-value (`PR(>F)`)** tests the null hypothesis that all age group means are equal.

The code cell above prints the calculated F-statistic and p-value. The conclusion about whether to reject the null hypothesis is determined by comparing this p-value to the significance level ($\alpha = 0.05$).

Final Conclusion: Based on the results printed by the code cell, we conclude whether there is a statistically significant difference in the average beauty scores among the different age groups in this dataset.

---

3. Question 3: Correlation between Evaluation and Beauty Scores (via Regression)

> **Instruction**: Correlation: Using the teachers' rating dataset, Is teaching evaluation score correlated with beauty score? The output should resemble the provided OLS Regression Results table.

Approach

To determine if there's a **statistically significant linear correlation** between teaching evaluation scores (`eval`) and beauty scores (`beauty`) using the specified method, we will perform a **Simple Linear Regression (OLS)**.

- **Model:** We'll model the `eval` score as the dependent variable, predicted by the continuous `beauty` score as the independent variable.
- **Correlation Interpretation from Regression:**
  - The **p-value (`P>|t|`)** for the `beauty` coefficient tests the null hypothesis that there is no linear relationship between beauty and evaluation scores.
  - The **coefficient (`coef`)** for `beauty` indicates the direction (positive or negative) of the relationship.
  - The **R-squared** value indicates the proportion of variance in `eval` scores explained by `beauty`, giving a measure of the strength of the linear association.

Hypotheses (for the slope coefficient of `beauty`, β_1):

- **Null Hypothesis (H_0):** There is **no significant linear correlation** between teaching evaluation scores and beauty scores ($\beta_1 = 0$).
- **Alternative Hypothesis (H_a):** There **is a significant linear correlation** between teaching evaluation scores and beauty scores ($\beta_1 \neq 0$).

We will use a standard **significance level (alpha) of 0.05**. If the p-value (`P>|t|`) for the `beauty` coefficient is less than 0.05, we reject the null hypothesis.

```python
# --- Question 3: Correlation using OLS Regression ---

# 1. Define and fit the OLS regression model
#    Model 'eval' score as dependent on 'beauty' score.
model_corr = smf.ols('eval ~ beauty', data=df).fit()

# 2. Print the full regression results summary
print("--- OLS Regression Results ---")
print(model_corr.summary())

# 3. Extract key results for interpretation
beauty_coef = model_corr.params['beauty']
beauty_p_value = model_corr.pvalues['beauty']
alpha = 0.05

# 4. State the conclusion based on the p-value
print("\n--- Hypothesis Test Conclusion (Correlation Significance)
---")
if beauty_p_value < alpha:
    print(f"Result: Reject the null hypothesis (p =
{beauty_p_value:.4f} < {alpha}).")
    print("Finding: There IS a statistically significant linear
correlation between evaluation score and beauty score.")
```

```
else:
    print(f"Result: Fail to reject the null hypothesis (p =
{beauty_p_value:.4f} >= {alpha}).")
    print("Finding: There is NO statistically significant linear
correlation between evaluation score and beauty score.")

# State the direction of the relationship based on the coefficient
sign
print(f"\nThe sign of the coefficient ({beauty_coef:.4f}) indicates a
{('positive' if beauty_coef > 0 else 'negative')} relationship.")
```

```
--- OLS Regression Results ---
                            OLS Regression Results

======================================================================
========
Dep. Variable:                   eval   R-squared:
0.036
Model:                            OLS   Adj. R-squared:
0.034
Method:                 Least Squares   F-statistic:
17.23
Date:                Tue, 28 Oct 2025   Prob (F-statistic):
3.94e-05
Time:                        14:31:01   Log-Likelihood:
-341.06
No. Observations:                 463   AIC:
686.1
Df Residuals:                     461   BIC:
694.4
Df Model:                           1

Covariance Type:            nonrobust

======================================================================
========
                 coef    std err          t      P>|t|      [0.025
0.975]
----------------------------------------------------------------------
--------
Intercept      3.9747      0.024    168.601      0.000       3.928
4.021
beauty         0.0919      0.022      4.151      0.000       0.048
0.135
======================================================================
========
Omnibus:                        4.763   Durbin-Watson:
1.878
Prob(Omnibus):                  0.092   Jarque-Bera (JB):
3.599
```

```
Skew:                          -0.086   Prob(JB):
0.165
Kurtosis:                       2.604   Cond. No.
1.09
======================================================================
========

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

--- Hypothesis Test Conclusion (Correlation Significance) ---
Result: Reject the null hypothesis (p = 0.0000 < 0.05).
Finding: There IS a statistically significant linear correlation
between evaluation score and beauty score.

The sign of the coefficient (0.0919) indicates a positive
relationship.
```

Interpretation of Regression Results for Correlation

The **OLS Regression Results** table models the relationship between `eval` score and `beauty` score. We interpret this output to understand their correlation:

- **`coef` for `beauty`**: This value indicates the estimated change in `eval` score for a one-unit increase in `beauty` score. Its sign (positive or negative) tells us the direction of the linear relationship.
- **`P>|t|` for `beauty`**: This p-value tests the significance of the relationship. A value less than our alpha (0.05) suggests the correlation is statistically significant (unlikely to be due to chance).
- **`R-squared`**: This measures the proportion of the total variance in `eval` scores that can be explained by the linear relationship with `beauty` scores. It indicates the strength of the association.

The code output above provides these key statistics. The conclusion regarding the statistical significance of the correlation is printed based on the p-value comparison.

Final Conclusion: Based on the regression results (specifically the p-value for the `beauty` coefficient and the R-squared value), we conclude whether there is a statistically significant linear correlation between teaching evaluation scores and beauty scores in this dataset, and we note the direction and strength of this relationship.

Final Summary and Conclusions

This notebook successfully addressed all three problems for the seventh practical assignment. The focus was on utilizing Ordinary Least Squares (OLS) regression models as a tool to perform analyses equivalent to a t-test, ANOVA, and correlation test on the teacher rating dataset.

Summary of Tasks Completed:

- **1. Regression with T-test:** An OLS model predicting evaluation scores based on a gender dummy variable was created. The t-statistic and p-value for the dummy variable's coefficient were interpreted to determine if gender significantly affects evaluation rates.
- **2. Regression with ANOVA:** An OLS model predicting beauty scores based on categorical age groups was created. An ANOVA table was derived from this model, and the F-statistic and p-value were used to test if mean beauty scores differ significantly across age groups.
- **3. Correlation via Regression:** A simple linear regression model predicting evaluation scores based on beauty scores was created. The p-value for the beauty coefficient and the model's R-squared value were interpreted to assess the significance and strength of the linear correlation between the two variables.

Key Learnings:

This assignment provided practical experience in using regression analysis as a flexible framework for hypothesis testing:

- Understanding how the t-test for a coefficient of a dummy variable in a simple OLS regression is equivalent to an independent samples t-test comparing two group means.
- Learning how to generate and interpret an ANOVA table from an OLS regression model with a categorical predictor to test for differences among multiple group means.
- Recognizing that the significance test (p-value) for the slope coefficient in a simple linear regression directly tests the significance of the linear correlation between the two variables, while R-squared measures the strength of that association.

This completes all requirements for the assignment. The solutions have been delivered using the specified regression methods, along with appropriate interpretations of the statistical outputs.