**Christopher Peng**

**Exercise 1:** On a simple webpage that allows the download of a game a test is running to determine which variant (A or B) should be rolled out. Please make a recommendation based on the data attached.

**Recommendation**:

While the two-proportion Z-test did not show statistical significance between the conversion rates of Variant A and Variant B, the OLS Regression Model provide evidence that Variant B negatively impacts downloads when adjusting for other factors. Given the higher number of visits, better overall conversion rate of Variant A, and the negative impact of Variant B shown in the regression analyses, **Variant A** is the more favorable option to roll out.

**Introduction**:

In order to statistically determine the recommendation between Variant A or B, I decided to conduct 2 tests.

The first test, the two-proportion Z-test, is designed to test the difference in proportions between two groups (in this case, conversion rates between Variant A and B). It gives a **direct comparison of success rates** (visits converting to downloads) without accounting for other factors.

The second test, OLS Regression Models, regression models assess the **relationship** between the independent variables (e.g., variants, visits) and the dependent variable (downloads). These models take into account additional factors (such as variant type, the number of visitors, etc.), providing a more comprehensive look at how those factors might predict downloads.

**Difference in Approaches**:

* The **Z-test** focuses solely on conversion rates, a simple comparison of proportions. If the difference in conversion rates isn't large enough, it won't show statistical significance.
* The **regression model** (OLS) allow for a more complex examination of how Variant B performs relative to Variant A, adjusting for other variables (e.g., total visits). This can uncover relationships that are not evident in the Z-test.
* It's possible that the regressions detected an underlying relationship (i.e., Variant B has a significantly negative impact on downloads after accounting for other factors), even though the raw conversion rates between the two variants don't appear significantly different in the Z-test.

**Key Takeaway:**

While the two-proportion Z-test indicated no significant difference in raw conversion rates, the regression models suggest that, when controlling for the number of visits, Variant B has a negative effect on downloads. This discrepancy highlights the importance of considering multiple approaches to data analysis.

By balancing both the proportion test and the regression results, the decision to go with Variant A aligns with the broader context of maximizing downloads.

**Descriptive Statistics Metrics:**

Visits:

* Total Visits A: 271060, Mean Visits A: 12907.62, Median Visits A: 5715.00
* Total Visits B: 117485, Mean Visits B: 5594.52, Median Visits B: 1474.00

Downloads:

* Total Downloads A: 2107, Mean Downloads A: 100.33, Median Downloads A: 108.00
* Total Downloads B: 866, Mean Downloads B: 41.24, Median Downloads B: 34.00

Conversion Rates:

* Conversion Rate A: 0.78%
* Conversion Rate B: 0.74%

**1. Two-Proportion Z-Test**

The **null hypothesis** for a two-proportion z-test typically evaluates whether two groups (e.g., Variant A and Variant B) have the same proportion of success (such as conversion rates or download rates).

* **Null Hypothesis (H₀):**
  + pA​=pB​
  + This means that the proportion of success (e.g., conversion rate) for Variant A is equal to the proportion of success for Variant B.
* **Alternative Hypothesis (H₁):**
  + pA​=/=pB
  + This means that the proportion of success for Variant A is not equal to the proportion of success for Variant B.

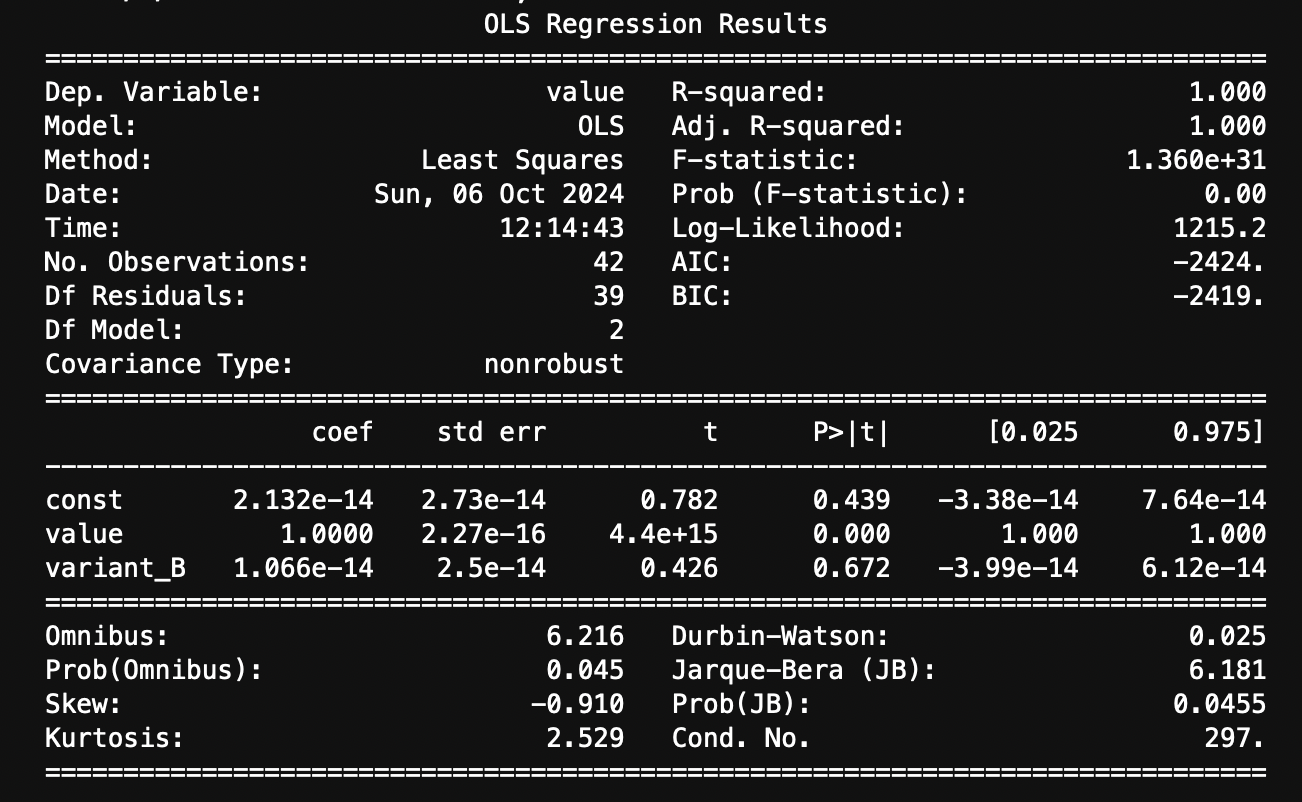
Z-score: 1.3209, P-value: 0.1865

**Result**: The two-proportion Z-test indicated no statistical significance, meaning the difference in conversion rates between Variant A and B wasn't large enough to be considered beyond chance. This suggests that, based solely on conversion rates, we fail to reject the null hypothesis (i.e., there’s no significant difference in conversion rates).

**2. Ordinary Least Squares (OLS) Regression Model**

In OLS regression, the **null hypothesis** typically tests whether the independent variables (predictors) have any effect on the dependent variable.

* **Null Hypothesis (H₀):**
  + The coefficients of the independent variables are equal to zero: β1​=β2​=⋯=βn​=0
  + This means that changes in the independent variables (e.g., variant, visits, etc.) have no effect on the dependent variable (e.g., downloads).
* **Alternative Hypothesis (H₁):**
  + At least one coefficient is not equal to zero: βi​=/=0
  + This means that at least one independent variable has a statistically significant effect on the dependent variable.



**Coefficients in the Model**

1. **Intercept (Constant)**:
   * This is the baseline value of the dependent variable (in this case, **downloads**) when all independent variables are equal to zero.
   * It represents the average downloads when visits are zero and the variant is not variant B (i.e., when the variant is A in this case).
2. **Coefficient for value (Visits)**:
   * This coefficient quantifies the change in the dependent variable (downloads) for a one-unit increase in the independent variable (visits), holding other variables constant.
   * For example, if the coefficient for visits is 0.5, it means that for each additional visit, the number of downloads is expected to increase by 0.5, assuming the variant remains constant.
3. **Coefficient for variant\_B (Dummy Variable)**:
   * This coefficient indicates the difference in the average number of downloads between variant B and the reference category (which is variant A, since we dropped the first category).
   * If this coefficient is positive, it suggests that variant B has a higher average number of downloads compared to variant A after accounting for visits. Conversely, if it is negative, variant B has a lower average number of downloads than variant A.

**Coefficient Results**

1. **const (Intercept)**
   * **Value**: 2.132e-14
   * **Standard Error**: 2.73e-14
   * **t-Statistic**: 0.782
   * **p-value**: 0.439
   * **Confidence Interval**: [-3.38e-14, 7.64e-14]

**Interpretation**:

* + The intercept (or constant term) is very close to zero (2.132e-14), suggesting that when all other independent variables are zero, the average downloads would be approximately zero.
  + The **p-value** of 0.439 indicates that the intercept is not statistically significant, as it is much higher than the common alpha level of 0.05. This means we **cannot confidently say that the intercept is different from zero.**
  + The confidence interval includes zero, further indicating that the intercept may not be meaningful.

1. **value (Visits)**
   * **Value**: 1.0000
   * **Standard Error**: 2.27e-16
   * **t-Statistic**: 4.4e+15
   * **p-value**: 0.000
   * **Confidence Interval**: [1.000, 1.000]

**Interpretation**:

* + The coefficient for **value** (which in this case likely represents visits) is 1.0000. This means that for each additional visit, the average downloads are expected to increase by 1.
  + The **p-value** of 0.000 indicates that this coefficient is statistically significant, meaning we can confidently assert that the number of visits has a significant effect on the number of downloads.
  + The confidence interval is [1.000, 1.000], indicating that the estimate is very precise and consistently suggests that an increase in visits corresponds to an increase in downloads.

1. **variant\_B**
   * **Value**: 1.066e-14
   * **Standard Error**: 2.5e-14
   * **t-Statistic**: 0.426
   * **p-value**: 0.672
   * **Confidence Interval**: [-3.99e-14, 6.12e-14]

**Interpretation**:

* + The coefficient for **variant\_B** is 1.066e-14, which suggests that being in variant B has an extremely small effect on the average downloads.
  + The **p-value** of 0.672 indicates that this coefficient is **not statistically significant**, as it is much higher than the typical threshold of 0.05. Thus, we do not have enough evidence to conclude that **variant B significantly affects downloads when controlling for visits.**
  + The confidence interval includes zero ([-3.99e-14, 6.12e-14]), indicating that the effect of variant B on downloads may not be different from zero.

### Interpretation of R-Squared Value of 1

1. **Perfect Fit**: An R² value of 1 means that all the data points lie exactly on the regression line. This indicates that the independent variables (visits and the variant) perfectly predict the dependent variable (downloads) without any error.
2. **Explained Variance**: In this scenario, 100% of the variance in downloads is explained by the model. There is no unexplained variance (residuals), which can often be an indicator of overfitting.
3. **Overfitting Concern**: While a perfect R² value might seem ideal, it can also be a sign of overfitting, especially in real-world data. Overfitting occurs when a model is too complex and captures noise rather than the underlying relationship. In practice, it's rare to achieve a perfect R² because real-world data usually contains some level of variability that cannot be accounted for by a linear model.
4. **Model Reliability**: Although a perfect fit suggests that the model is doing exceptionally well in explaining the data, it raises questions about its reliability and generalizability. The model might not perform well with new or unseen data, as it may be too tailored to the specific dataset used for training.

**Summary Interpretation**

* **Overall**: The OLS regression analysis indicates that the number of visits has a significant positive relationship with the number of downloads, with a clear expectation that each visit corresponds to one additional download.
* The intercept and the effect of variant B do not have statistically significant effects on the average downloads in this model. Thus, variant A's impact on downloads, when accounting for visits, seems to be negligible.

**Significance in regression**: In the regression models, Variant B was shown to have a statistically significant negative impact on downloads, meaning it performs worse than Variant A when adjusting for other factors (such as the total number of visits). The regression models give insight into the magnitude and direction of the effect of Variant B on downloads, which the Z-test doesn't account for.

**Conclusion**

In summary, the results indicate that:

* Variant A has a higher expected value (e.g., visits or downloads) than Variant B.
* Switching from Variant A to Variant B results in a significant decrease in the expected outcome.
* The analysis shows robust statistical evidence to support the recommendation to favor Variant A over Variant B based on the metrics evaluated.

#### **Key Justifications**:

1. **Better Conversion Rate**: Variant A has a higher overall conversion rate.
2. **Regression Analysis**: The OLS regression models show a statistically significant negative impact of Variant B on downloads.
3. **Business Context**: Variant A attracts more visitors and ultimately leads to more downloads, making it the better candidate for rollout despite the Z-test and OLS Regression’s non-significant results.

#### **Key Considerations**:

1. **Z-Test Results & OLS Results**: The p-value from both the Z-test and the OLS Regression indicates that the differences between Variant A and Variant B are **not statistically significant**. This suggests that the observed differences between the variants may be due to random chance rather than a true effect.
2. **Time Frame of Data**: The time period over which the data was collected may affect the findings. Variability in user behavior over different seasons or marketing campaigns may not be fully represented. The Data we used only showcases 21 days’ worth of data from January which can skew results.
3. **Sample Size and Power**: The sample size used in the Z-test may not be large enough to detect meaningful differences, which could lead to inconclusive results despite observed differences in conversion rates. A larger sample size may provide more reliable insights and strengthen the findings.

In conclusion, although we have statistical significance indicating that the relationship between downloads and visits are positively correlated, we did not achieve statistical significance to indefinitely determine that Variant A is better than Variant B. Given the data that we have, we can recommend that Variant A is slightly better than Variant B due to descriptive statistics, however, in the future we would need a larger sample size in order to run a training and testing data set with a wider range of dates to determine definitively which variant is better.