Brian Tenneson

ADS 534

08/12/2024

Statement regarding academic integrity: Large Language Models (LLMs) were utilized alongside human oversight to perform data analysis and generate insights across various contexts and datasets.

1

1(a)

The formula for expected counts in a contingency table for a Chi-square test is:

Expected Count=(Row Total)×(Column Total)/Grand Total. Using this formula four times, we obtain the following values for a, b, c, and d:

a (Expected count for Democrats, White): approximately 418.50

b (Expected count for Democrats, Black): approximately 209.76

c (Expected count for Republicans, White): approximately 421.30

d (Expected count for Republicans, Black): approximately 211.16.

Therefore, the completed table of expected counts is given by:

	White	Black	Hispanic	Asian	Total
Democrat	418.50	209.76	171.71	96.19	896
Republican	421.30	211.16	172.96	96.89	902
Independent	386.75	193.81	158.73	88.92	828
Total	1227	615	503	282	2627

1(b)

Null Hypothesis (H_0): There is no association between registered political party and race. This implies that the party affiliation is independent of race.

Alternative hypothesis (H_A): There is an association between registered political party and race. This implies that the party affiliation depends on race, or vice versa.

1(c)

Using python, we get a chi-square test statistic $\chi^2 = 16.50$ with a p-value of 0.011. The degrees of freedom for this χ^2 test is $6 = (R-1)(C-1)$ where $R = 3$ is the number of rows and $C = 4$ is the number of columns. This all means we would reject the null hypothesis. The critical value for χ^2 I found to be 12.6 and given that $16.50 > 12.6$, with a p-value under 0.05, we reject the null hypothesis so there is an association between registered political party and race.

2

2(a) Here is some python code related to the inquiry:

```
# Calculation of expected counts
R_Yes = 45
R_No = 102
C_Female = 99
Grand_Total = 147

# Expected count for women having symptom X
Expected_Yes_Female = (R_Yes * C_Female) / Grand_Total

# Expected count for women not having symptom X
Expected_No_Female = (R_No * C_Female) / Grand_Total

Expected_Yes_Female, Expected_No_Female
```

Result

(30.306122448979593, 68.6938775510204)

For women having symptom X: approximately 30.31

For women not having symptom X: approximately 68.69.

For men, we use a similar technique:

```
# Recalculating the expected counts for men

# Previously defined constants for recalculating
R_Yes = 45
R_No = 102
```

```
Grand_Total = 147
C_Male = 48

# Expected count for men having symptom X
Expected_Yes_Male = (R_Yes * C_Male) / Grand_Total

# Expected count for men not having symptom X
Expected_No_Male = (R_No * C_Male) / Grand_Total

Expected_Yes_Male, Expected_No_Male
```

Result

(14.693877551020408, 33.30612244897959)

For men having symptom X: approximately 14.69

For men not having symptom X: approximately 33.31.

The expectancy table including both genders is then given by

symptom X?	Yes	No
Male Expected	14.69	33.31
Female Expected	30.31	68.69

2(b)

We use code in two stages:

```
# Observed and expected values for each cell
O_Male_Yes = 9
O_Male_No = 39
O_Female_Yes = 36
O_Female_No = 63
```

```
E_Male_Yes = 14.69
E_Male_No = 33.31
E_Female_Yes = 30.31
E_Female_No = 68.69

# Calculating Chi-square statistic
chi_square_stat = ((O_Male_Yes - E_Male_Yes)**2 / E_Male_Yes) + \
                  ((O_Male_No - E_Male_No)**2 / E_Male_No) + \
                  ((O_Female_Yes - E_Female_Yes)**2 / E_Female_Yes) + \
                  ((O_Female_No - E_Female_No)**2 / E_Female_No)

chi_square_stat
```

Result:

$\chi^2 = 4.72$

The result was derived from the following equation

$$\chi^2 = \sum \frac{(O-E)^2}{E}.$$

The degrees of freedom is $(R-1)(C-1) = (1)(1) = 1$.

```
from scipy.stats import chi2

# Degrees of freedom
df = 1
```

```
# Critical value for chi-square test at 0.05 level
critical_value = chi2.ppf(0.95, df)

critical_value
```

Result

3.84

Commentary: Since our Chi-square statistic 4.72 is greater than the critical value 3.84, we can conclude that the difference observed in the contingency table is statistically significant. This suggests that there is likely an association between gender and the presence of symptom X at the 0.05 level of significance.

2(c)

To calculate the odds ratio (OR) of having symptom X for men relative to women, we use the data provided in the contingency table:

Symptom X?	Male	Female
Yes	9	36
No	39	63

.

The odds of men having symptom X is the ratio of men with the symptom to those without:

Odds_Men=9/39.

The odds of women having symptom X is the ratio of women with the symptom to those without:

Odds_Women=36/63.

The odds ratio (OR) is then calculated by dividing the odds for men by the odds for women:

OR=Odds_Men/Odds_Women=(9/39)/(36/63) which is approximately 0.40.

Commentary: This means that men have about 0.404 times the odds of having symptom X compared to women, indicating that women are more likely to have symptom X than men.

2(d)

I used the following code:

```
import numpy as np

# Given values
a = 9  # men with symptom X
b = 39 # men without symptom X
c = 36 # women with symptom X
d = 63 # women without symptom X

# Calculate Odds Ratio (OR)
OR = (a / b) / (c / d)

# Calculate the standard error (SE) of log(OR)
SE_log_OR = np.sqrt(1/a + 1/b + 1/c + 1/d)

# Calculate the 95% confidence interval for log(OR)
z_score = 1.96
lower_log_OR = np.log(OR) - z_score * SE_log_OR
upper_log_OR = np.log(OR) + z_score * SE_log_OR

# Exponentiate to get the CI for the OR
CI_OR = (np.exp(lower_log_OR), np.exp(upper_log_OR))
OR, CI_OR
```

Result: The 95% confidence interval for this Odds Ratio is (0.176, 0.928). Note that 1 is not in this interval, making it appear very likely that women will have symptom X more frequently than men.

3

3(a)

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 \times \text{apgar5}$$

Here:

p is the probability of a germinal matrix hemorrhage occurring (grmhem $= 1$).

β_0 is the intercept.

β_1 is the regression coefficient for the apgar5 score, indicating the effect of the apgar5 score on the log-odds of experiencing a hemorrhage.

3(b)

```
Optimization terminated successfully.
         Current function value: 0.394634
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                 grmhem   No. Observations:                  100
Model:                          Logit   Df Residuals:                       98
Method:                           MLE   Df Model:                            1
Date:                Fri, 09 Aug 2024   Pseudo R-squ.:                 0.06642
Time:                        13:19:47   Log-Likelihood:                -39.463
converged:                       True   LL-Null:                       -42.271
Covariance Type:            nonrobust   LLR p-value:                   0.01781
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.3037      0.619     -0.491      0.624      -1.517       0.910
apgar5        -0.2496      0.104     -2.392      0.017      -0.454      -0.045
==============================================================================
```

Then $\hat{\beta}_1$ is -0.2496 and it indicates the effect of the apgar5 score on the log-odds of experiencing hemorrhage. A negative coefficient indicates that higher apgar5 scores are associated with lower log odds of having a hemorrhage. This means that as the apgar5 score increases, the probability of a hemorrhage decreases.

3(c) The probability I found is 0.258745. So, if a particular child has a five-minute apgar5 score of 3, they would face a one in four probability of having a brain hemorrhage.

3(d)

```
Optimization terminated successfully.
         Current function value: 0.394634
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                  grmhem   No. Observations:                  100
Model:                           Logit   Df Residuals:                       98
Method:                            MLE   Df Model:                            1
Date:                 Fri, 09 Aug 2024   Pseudo R-squ.:                 0.06642
Time:                         14:01:57   Log-Likelihood:                -39.463
converged:                       True   LL-Null:                       -42.271
Covariance Type:            nonrobust   LLR p-value:                   0.01781
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.3037      0.619     -0.491      0.624      -1.517       0.910
apgar5        -0.2496      0.104     -2.392      0.017      -0.454      -0.045
==============================================================================
Estimated odds ratio for a one-unit increase in five-minute Apgar score: 0.7791
```

Since $\hat{\beta}_1 < 0$, the OR $\left(OR = e^{\hat{\beta}_1}\right)$ will be less than 1. This indicates that an increase in the apgar5 score is associated with a decrease in the odds of a hemorrhage, suggesting that better apgar5 scores (which reflect better neonatal health) are protective against hemorrhages.

3(e) We see $3\left(\hat{\beta}_1\right) \approx -0.7488$ and note that $e^{-0.7488} \approx 0.473$. This odds ratio 0.473 tells us that the odds of suffering a germinal matrix hemorrhage decrease by a factor of about 0.473 with each 3-unit increase in the apgar5 score, which is a 52.7% decrease in odds.

3(f)

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_2 \times \text{tox}$$

3(g)

```
Optimization terminated successfully.
         Current function value: 0.409247
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                  grmhem   No. Observations:                  100
Model:                           Logit   Df Residuals:                       98
Method:                            MLE   Df Model:                            1
Date:                 Sat, 10 Aug 2024   Pseudo R-squ.:                 0.03185
Time:                         09:20:14   Log-Likelihood:                -40.925
converged:                       True   LL-Null:                       -42.271
Covariance Type:            nonrobust   LLR p-value:                    0.1008
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept     -1.5353      0.295     -5.211      0.000      -2.113      -0.958
tox           -1.4604      1.066     -1.370      0.171      -3.550       0.629
==============================================================================
```

The coefficient $\hat{\beta}_2 = -1.4604$ for toxemia status suggests that the presence of toxemia is associated with a decrease in the log-odds of suffering a germinal matrix hemorrhage. Specifically, this negative coefficient indicates that patients with toxemia have lower odds of having a germinal matrix hemorrhage compared to those without toxemia, according to the data. However, the p-value of 0.171 suggests that this effect is not statistically significant at common significance levels (e.g., 0.05). Thus, while the model indicates a protective trend of toxemia against germinal matrix hemorrhage, this finding is not statistically robust based on the provided data.

3(h) I used the following code:

```
import numpy as np

# Calculate the predicted probability of germinal matrix hemorrhage for a child whose mother had toxemia
beta_0 = -1.5353
beta_2 = -1.4604
```

```
tox = 1  # Mother was diagnosed with toxemia

# Calculate the probability using the logistic function
predicted_probability = 1 / (1 + np.exp(-(beta_0 + beta_2 * tox)))
predicted_probability
```

Result:

approximately 0.0476

So for a child whose mother was diagnosed with toxemia, its predicted probability of experiencing a germinal brain hemorrhage is 0.048 or one in twenty (roughly).

3(i) I used the following code to get the result afterwards:

```
# Calculate the odds ratio for toxemia status
odds_ratio_tox = np.exp(beta_2)
odds_ratio_tox
```

Result (approximately)

0.232

Commentary: The estimated odds ratio is approximately 0.232. This means that the odds of suffering a germinal matrix hemorrhage for children whose mothers were diagnosed with toxemia are about 23.2% of the odds for children whose mothers were not diagnosed with toxemia. In other words, the presence of toxemia significantly decreases the odds of a germinal matrix hemorrhage occurring in the child.