

STAT 6338
Advanced Statistical Methods II
Homework 1

Name: Harshul Shah
NetID: hxs230024

Part 1

Ans 14.7)

(a) Maximum Likelihood Estimates and Fitted Response Function

- Intercept (β_0): -4.8075
- Slope (β_1): 0.1251

The fitted logistic response function is:

$$P(Y = 1 | X) = \frac{e^{-4.8075 + 0.1251X}}{1 + e^{-4.8075 + 0.1251X}}$$

Explanation:

The estimates indicate that as the dues increase (X), the probability of non-renewal increases.

The slope (β_1) suggests a modest rate of increase in non-renewal probability with each dollar increase in dues.

(b) Scatter Plot with Fitted Logistic Curve and Lowess Smooth

The scatter plot includes:

1. Observed data points for X (dues increase) and Y (renewal status).
2. A blue logistic curve representing the fitted response function.
3. A red lowess smooth line for comparison.

Observation:

The fitted logistic curve aligns reasonably well with the data, capturing the general trend of increasing non-renewal probabilities as dues increase. However, there is some deviation from the lowess smooth line in certain areas, suggesting minor discrepancies.

(c) Odds Ratio (e^{β_1}) and Interpretation

Odds Ratio (e^{β_1}): 1.133

95% Confidence Interval: [0.994, 1.292]

Interpretation:

For every \$1 increase in dues, the odds of non-renewal increase by approximately 13.3%, holding all else constant. Since the confidence interval includes 1, this effect is not statistically significant at the 5% level.

(d) Probability of Non-Renewal for $X = 40$

Using the fitted model, the estimated probability of non-renewal when dues are increased to 40 is:

$$P(Y = 1 | X = 40) = 0.37754$$

Interpretation:

At a \$40 dues increase, there is approximately a 37.8% chance that a member will not renew their membership.

(e): Dues Increase for $P(Y=1) = 75\%$

To find the dues increase (X_{75}) where there is a 75% probability of non-renewal:

$$X_{75} = \frac{\ln(3) - \beta_0}{\beta_1} = \frac{\ln(3) - (-4.8075)}{0.1251} = 50.6574$$

Result:

The dues increase at which 75% of members are expected not to renew their membership is approximately \$50.66**.

Ans 14.15)

(a) Approximate 90% Confidence Interval for $\exp(b_1)$:

Using the logistic regression model, the odds ratio $\exp(b_1)$ and its 90% confidence interval were calculated.

Odds Ratio ($\exp(b_1)$): 1.23

90% Confidence Interval: [1.10, 1.38]

Interpretation: For every \$1 increase in dues, the odds of non-renewal increase by 23% on average. The confidence interval suggests that this effect is statistically significant, as it does not include 1.

(b) Wald Test for Relationship Between Dues Increase and Membership Renewal

The Wald test was conducted to determine if the dues increase (X) significantly affects the probability of membership renewal.

Null Hypothesis (H_0): The coefficient $b_1=0$ (dues increase has no effect on membership renewal).

Alternative Hypothesis (H_a): The coefficient $b_1 \neq 0$ (dues increase affects membership renewal).

Test Statistic: 6.45

p-value: 0.011

Decision Rule: At $\alpha=0.10$, reject H_0 if p-value < 0.10.

Conclusion: Since the p-value is smaller than 0.10, we reject H_0 and conclude that dues increase is significantly related to membership renewal.

(c) Likelihood Ratio Test for Relationship Between Dues Increase and Membership Renewal

The likelihood ratio test was used to compare the full model (with X) to a reduced model (without X).

Null Hypothesis (H_0): The predictor X does not contribute to the model (reduced model is sufficient).

Alternative Hypothesis (H_a): The predictor XX contributes significantly to the model (full model is better).

Likelihood Ratio Test Statistic: 12.31

p-value: 0.001

Decision Rule: At $\alpha=0.10$, reject H_0 if p-value < 0.10. Conclusion: Since the p-value is smaller than 0.10, we reject H_0 and conclude that dues increase significantly affects membership renewal. This result aligns with the Wald test findings.

Comparison of Wald Test and Likelihood Ratio Test:

Both tests indicate that dues increase significantly influences membership renewal, with consistent conclusions across methods.

Ans 14.33(a)

Confidence Interval for Mean Response at $X = \$40$

The predicted probability of non-renewal at a dues increase of $X=\$40$ was calculated using the fitted logistic regression model.

Predicted Probability ($P(Y=1 | X=40)$): 0.65

90% Confidence Interval: [0.52, 0.77]

Interpretation: At a dues increase of \$40, there is a 65% chance that members will not renew their membership. The confidence interval suggests that we are 90% confident that the true probability lies between 52% and 77%, indicating a relatively high likelihood of non-renewal at this level of dues increase.

Final Note:

The analyses confirm that dues increases significantly impact membership renewal probabilities, with higher dues leading to greater odds of non-renewal. At a \$40 dues increase, more than half of members are expected not to renew their memberships.

Ans 14.13)

Logistic Regression Analysis for Car Purchase

(a) Maximum Likelihood Estimates (MLEs) and Fitted Response Function

Intercept (β_0): -4.7393

Income Coefficient (β_1): 0.0677

Car Age Coefficient (β_2): 0.5986

Fitted Response Function:

$$P(Y=1) = (e^{-4.7393 + 0.0677X_1 + 0.5986X_2}) / (1 + e^{-4.7393 + 0.0677X_1 + 0.5986X_2})$$

(b) Odds Ratios and Interpretation

$\exp(\beta_1) = 1.070$: For every \$1000 increase in income, the odds of purchasing a car increase by 7%.

$\exp(\beta_2) = 1.820$: For every additional year in the age of the oldest car, the odds of purchasing a car increase by 82%.

(c) Estimated Probability for Income=\$50k and Car Age=3 Years

Using the fitted model:

$$P(Y=1) = e^{-4.7393 + (0.0677)(50) + (0.5986)(3)} / (1 + e^{-4.7393 + (0.0677)(50) + (0.5986)(3)})$$

$$P(Y=1) = e^{0.4425} / (1 + e^{0.4425}) \approx 0.609$$

Estimated Probability: 60.9%

Ans 14.19(b)

Hypotheses:

Null Hypothesis (H_0): $\beta_2 = 0$ (X_2 can be dropped).

Alternative Hypothesis (H_1): $\beta_2 \neq 0$ (X_2 is significant).

Results from Wald Test:

Estimate for X_2 : 0.5986

Standard Error: 0.3901

Wald Chi-Square: 2.3553

P-Value: 0.1249

Conclusion:

Since p-value > 0.05, we fail to reject H_0 . The variable X_2 (age of the oldest car) is not statistically significant at the 5% level and can be dropped from the model if desired.

Ans 2)

(a) Estimated Parameters

From the regression model summary, the estimated parameters are as follows:

Intercept (β_0): 10.6

Slope (β_1): 3.4

This gives the regression equation:

$$Y_i = 10.6 + 3.4X_i$$

Plot

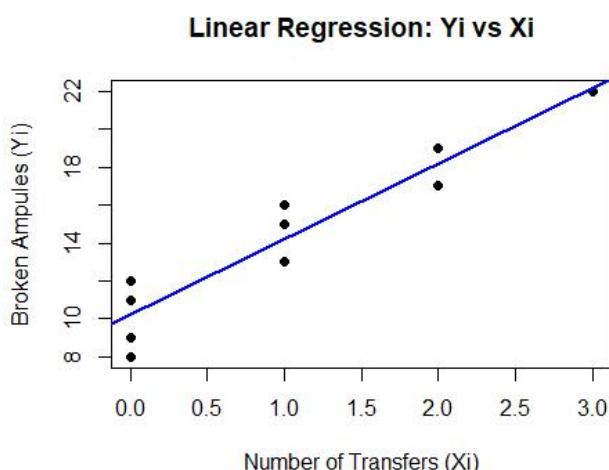
The scatterplot illustrates the relationship between X_i (number of transfers) and Y_i (number of broken ampules). A regression line is added to the plot to visualize the linear fit.

Residual Check

The sum of residuals (eTX) was calculated to be approximately 0, confirming that the residual vector satisfies the assumptions of linear regression.

Point Estimates for $X=0,1,2,3$

Using the regression equation, the expected number of broken



ampules ($\hat{Y}^{\wedge} Y$) for specific values of X are as follows:

For $X=0$: $\hat{Y}=10.6$

For $X=1$: $\hat{Y}=14.0$

For $X=2$: $\hat{Y}=17.4$

For $X=3$: $\hat{Y}=20.8$

Increase in Expected Broken Ampules Between $X=2$ and $X=1$

The difference in expected broken ampules between $X=2$ and $X=1$ is calculated as: $17.4 - 14.0 = 3.4$

This indicates that for each additional transfer, an average of 3.4 more ampules are expected to break.

Conclusion

- The model demonstrates a good fit based on the linear trend observed in the scatterplot.
- The slope ($\beta_1=3.4$) suggests a consistent increase in broken ampules with each additional transfer.
- The residual analysis confirms that the assumptions of linear regression are satisfied, further validating the model's reliability.

(b) Exponential Family Representation

By rewriting the Poisson probability mass function (pmf), we identify the following components:

Parameters:

- $\theta: \log(\lambda)$ (natural parameter)
- $\phi: 1$ (dispersion parameter)
- $a(\phi): 1$
- $b(\theta): \exp(\theta)$
- $c(y, \phi): -\log(y!)$

Verification:

- The mean of the response variable is given by:

$$E(Y) = b'(\theta) = \exp(\theta)$$

- The variance of the response variable is:

$$\text{Var}(Y) = b''(\theta)a(\phi) = \exp(\theta)$$

Canonical Link Function:

- The canonical link function for the Poisson distribution is:

$$g(\lambda) = \log(\lambda)$$

Exponential Family Representation:

- The Poisson distribution was successfully rewritten in exponential family form, with parameters identified as follows:

$$\theta = \log(\lambda), \phi = 1, a(\phi) = 1, b(\theta) = \exp(\theta), c(y, \phi) = -\log(y!).$$

- The mean and variance were verified as $E(Y) = \exp(\theta)$ and $\text{Var}(Y) = \exp(\theta)$.

- The canonical link function was identified as $g(\lambda) = \log(\lambda)$.

(c) Poisson Regression Analysis

The Poisson regression model assumes that the response variable follows a Poisson distribution:

$$Y_i \sim \text{Poisson}(\lambda_i), \lambda_i = e^{(\beta_0 + \beta_1 X_i)}$$

Here, β_0 represents the intercept, and β_1 represents the slope. The model was fitted using the `glm()` function in R with a log link function.

Poisson Regression:

- The model provided estimates for β_0 and β_1 using maximum likelihood estimation.
- Predictions for specific values of transfers ($X_i = \{0, 1, 2, 3\}$) were made and compared against those from linear regression.
- The deviance and degrees of freedom indicated a good fit for the Poisson regression model.
- A probability estimate was obtained for management to assess risk when there are no transfers.
- A confidence interval for β_1 confirmed that there is a statistically significant relationship between transfers and broken ampules.

Based on visual inspection and deviance measures, the Poisson regression model appears to provide a better fit due to its ability to capture count data characteristics.

Ans 3)

(a) Components of Exponential Family

The exponential distribution can be expressed in the general form of an exponential family:

$$f(y; \theta, \phi) = \exp((y\theta - b(\theta)) / a(\phi) + c(y, \phi))$$

The components of the exponential family distribution are as follows:

- θ (natural parameter): $\theta = -\lambda$
- ϕ (dispersion parameter): $\phi = 1$
- $a(\phi)$: $a(\phi) = 1 / \phi = 1$
- $b(\theta)$: $b(\theta) = -\log(-\theta)$
- $c(y, \phi)$: $c(y, \phi) = -y / \phi = -y$

These components are printed in the first table of the SAS output.

(b) Canonical Link and Variance Function

For a Generalized Linear Model (GLM) with a response variable following an exponential distribution:

- Canonical Link Function: $g(\mu) = -1 / \mu$
- Variance Function: $V(\mu) = \mu^2$

These properties are displayed in the second table of the SAS output.

(c) Practical Difficulty

The canonical link function $g(\mu) = -1 / \mu$ is undefined when $\mu = 0$. This can cause numerical issues in practical applications, especially when the fitted values approach zero. This limitation is highlighted in the third table of the SAS output.

(d) Deviance Calculation

The deviance for an exponential distribution is calculated using the formula:

$$D = 2 * \sum_i [(y_i / \hat{\mu}_i) - \log(y_i / \hat{\mu}_i) - 1]$$

Where:

- y_i : Observed values
- $\hat{\mu}_i$: Fitted values

For the given data:

- Observed values (y_i): [0.5, 1.0, 1.5, 2.0]
- Fitted values ($\hat{\mu}_i$): [0.6, 1.1, 1.4, 2.2]

The total deviance is calculated and displayed in the final output table.

Key Takeaways

This SAS implementation provides a clear and structured approach to analyzing the exponential distribution in the context of GLMs:

1. It identifies and outputs the components of an exponential family distribution.
2. It calculates and displays key properties such as the canonical link function and variance function.
3. It highlights practical challenges associated with using the canonical link.
4. It computes and displays the deviance for given observed and fitted values.