School of Mathematics and Statistics

Te Kura Mātai Tatauranga

STAT 292 **Assignment 5 Solutions**

Note: Solutions must be either typed or written neatly, and questions must be answered in order.

1. **(20 marks)**

   Table 1 presents a subset of data collected by Väisänen and Järvinen (1977) on bird species in the Krunnit Islands archipelago of Finland. In particular, they reported on the bird species found on each of the islands in 1949 and how many of those bird species were extinct by 1970. It is of interest to understand whether the area of the island (in km^2) is associated with species' survival. The data corresponding to Table 1 are available in the Excel file `Extinction.xlsx`.

		Extinct?	
Island	Area (X)	Yes	No
Ulkokrunni	185.80	5	70
Maakrunni	105.80	3	64
Ristikari	30.70	10	56
Isonkivenletto	8.50	6	45
Hietakraasukka	4.80	3	25
Kraasukka	4.50	4	16
Länsiletto	4.30	8	35

Table 1: Extinction of bird species from 1949 to 1970 on seven islands in the Krunnit Islands archipelago, Finland.

Fit the logistic regression model

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

where X denotes island area and $p(X)$ denotes the probability of extinction.

Figure 1 shows relevant SAS output for the logistic regression model.

(a) **Carry out an appropriate goodness-of-fit test to determine whether the model provides a good fit to the data. State the hypotheses, and give the test statistic and the p-value of the test. What do you conclude at the $\alpha = 0.05$ significance level? (4 marks)**

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	3.7326	5	0.7465	0.5885
Pearson	3.5477	5	0.7095	0.6162

Number of unique profiles: 7

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	246.644	242.463
SC	250.502	250.179
-2 Log L	244.644	238.463

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6.1813	1	0.0129
Score	5.5868	1	0.0181
Wald	5.2804	1	0.0216

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.7022	0.2155	62.4153	<.0001
AREA	1	-0.00667	0.00290	5.2804	0.0216

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
AREA	0.993	0.988	0.999

Figure 1: Summary output for the logistic regression model $\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$.

We wish to test the hypotheses

$$\mathcal{H}_0 : \text{The model provides a good fit.}$$
$$\mathcal{H}_1 : \text{The model does not provide a good fit.}$$

The likelihood ratio goodness-of-fit test statistic is given by

$$G^2 \approx 3.7326,$$

which follows a χ^2_5 distribution asymptotically.

The p-value is given by

$$p\text{-value} \approx P\left(\chi^2_5 > 3.7326\right) \approx 0.5885.$$

As the p-value is greater than $\alpha = 0.05$, we do not reject $\mathcal{H}_0$, and we conclude that the logistic regression model provides a good fit to the data.

(b) **Give estimates of β_0 and β_1 (up to 5dp). (2 marks)**

$$\begin{aligned} \widehat{\beta}_0 &\approx -1.7022 \\ \widehat{\beta}_1 &\approx -0.00667 \end{aligned}$$

(c) **Interpret the association between island area and extinction using the odds ratio (to 3dp). Demonstrate how the odds ratio is calculated from Figure 1. Additionally, provide a 95% confidence interval for the odds ratio. (3 marks)**

$$\exp\left(\widehat{\beta}_1\right) \approx \exp(-0.00667) \approx 0.993$$

An increase in island area by 1 km^2 is associated with an estimated multiplicative change of 0.993 (0.988, 0.999) in the odds of extinction. (Note that this is a *decrease* in the odds of extinction, so larger islands would be expected to have lower odds of bird species going extinct.)

(d) **Find the predicted probability of extinction for an island with an area of 50 km^2 (to 4dp). (2 marks)**

Predicted probabilities corresponding to the model are given by

$$\widehat{p}(X) \approx \frac{\exp(-1.7022 - 0.00667X)}{1 + \exp(-1.7022 - 0.00667X)}.$$

Then

$$\widehat{p}(50) \approx \frac{\exp(-1.7022 - 0.00667 \times 50)}{1 + \exp(-1.7022 - 0.00667 \times 50)} \approx 0.1155$$

Thus, the predicted probability of extinction for a bird species on an island with an area of 50 km^2 is 0.1155.

(e) **Find the fitted count of extinct bird species on the island of Ulkokrunni (to 2dp). Also find the fitted count of non-extinct bird species on Ulkokrunni (to 2dp). (4 marks)**

To find the fitted count of extinct bird species on Ulkokrunni, we must first determine the probability of extinction on Ulkokrunni. The area for Ulkokrunni is 185.80 km^2, so

$$\widehat{p}(185.80) \approx \frac{\exp(-1.7022 - 0.00667 \times 185.80)}{1 + \exp(-1.7022 - 0.00667 \times 185.80)} \approx 0.050167$$

Then the fitted count is given by

$$\begin{aligned} & \text{sample size in the category} \times \text{predicted probability of extinction} \\ & \text{for that category} \\ \approx\ & (5 + 70) \times 0.050167 \approx 3.76 \text{ birds.} \end{aligned}$$

Also, the fitted count of non-extinct bird species on Ulkokrunni is approximately $(5 + 70) - 3.76 \approx 71.24$ birds.

(f) **Test**

$$\begin{aligned} \mathcal{H}_0 : \beta_1 &= 0 \\ \mathcal{H}_1 : \beta_1 &\neq 0 \end{aligned}$$

using the Wald statistic. Give the test statistic and the p-value of the test. What do you conclude at the $\alpha = 0.05$ significance level? (3 marks)

The test statistic is given by

$$Z^2 = 5.2804$$

which is distributed χ^2_1 asymptotically. The p-value is given by

$$p\text{-value} \approx P\left(\chi^2_1 > 5.2804\right) \approx 0.0216$$

As the p-value is less than $\alpha = 0.05$, we reject $\mathcal{H}_0$ and conclude that there is sufficient evidence to suggest that $\beta_1 \neq 0$ (*i.e.*, there is a significant association between island size and risk of extinction for bird species.).

2. **(20 marks)**

Consider data reported by Gilbert (1981) on the relationship between pre-marital sex (*i.e.*, sexual intercourse before marriage), extra-marital sex (*i.e.*, sexual intercourse with someone other than a spouse whilst married), and whether the person had been divorced for a random sample of heterosexual men and women who had been married at least once. These data are presented in Table 2 and are available in the Excel file `Divorce.xlsx`.

Gender (W)	Pre-marital Sex (X)	Extra-marital Sex (Y)	Divorced? (Z) No	Yes
Woman	Yes	Yes	4	17
		No	25	54
	No	Yes	4	36
		No	322	214
Man	Yes	Yes	11	28
		No	42	60
	No	Yes	4	17
		No	130	68

Table 2: Data on reported pre-marital sex, extra-marital sex, and divorce for a random sample of heterosexual men and women.

First, use the backward model selection method to find the simplest model that provides a good fit to the data. Start from the following model, which we will denote by M_2,

$$\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \beta_0 + \beta_i^W + \beta_j^X + \beta_k^Y + \beta_{ij}^{WX} + \beta_{ik}^{WY} + \beta_{jk}^{XY} + \beta_{ijk}^{WXY},$$

where p_{ijk} is the probability of divorce when the gender (W) is at level i, pre-marital sex status (X) is at level j, and extra-marital sex status (Y) is at level k.

Figure 2 shows relevant summary output from SAS.

Summary of Backward Elimination					
Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	GENDER*PREMAR*EXTRAM	1	6	0.1472	0.7012
2	GENDER*PREMARITAL_SE	1	5	0.1434	0.7050
3	GENDER*EXTRAMARITAL_	1	4	0.4027	0.5257

Figure 2: Summary output for the backward selection method applied to the logit model $\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \beta_0 + \beta_i^W + \beta_j^X + \beta_k^Y + \beta_{ij}^{WX} + \beta_{ik}^{WY} + \beta_{jk}^{XY} + \beta_{ijk}^{WXY}$.

(a) **Is model M_2 a saturated model? Why or why not? (3 marks)**

Yes. There are in total 8 logits (corresponding to the $2 \times 2 \times 2$ possible combinations of gender, pre-marital sex, and extra-marital sex levels), and our model estimates

$$1 + (2-1) + (2-1) + (2-1) + (2-1)(2-1) + (2-1)(2-1) + (2-1)(2-1) + (2-1)(2-1)(2-1) = 8$$

non-redundant parameters. Then

$$\begin{aligned}\text{Residual df} &= \text{(no. of logits)} - \text{(no. of non-redundant parameters)} \\ &= 8 - 8 \;=\; 0\end{aligned}$$

As the residual degrees of freedom is 0, the model is saturated.

(b) **What information does Step 1 provide in the SAS output? Write down the test hypotheses. What do you conclude? (4 marks)**

In Step 1, we are comparing the models

Full model (M_2):	$\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \beta_0 + \beta_i^W + \beta_j^X + \beta_k^Y + \beta_{ij}^{WX} + \beta_{ik}^{WY} + \beta_{jk}^{XY} + \beta_{ijk}^{WXY}$
Reduced model (M_1):	$\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \beta_0 + \beta_i^W + \beta_j^X + \beta_k^Y + \beta_{ij}^{WX} + \beta_{ik}^{WY} + \beta_{jk}^{XY}$

(In other words, the reduced model simply removes the three-way interaction between gender, pre-marital sex, and extra-marital sex.)

A test of the hypotheses

$\mathcal{H}_0$: The additional term in M_2 can be deleted.
$\mathcal{H}_1$: The additional term in M_2 cannot be deleted.

produces a test statistic of

$$G^2 \;=\; 0.1472,$$

which is distributed χ_1^2 asymptotically. The p-value is given by

$$p\text{-value} \;\approx\; P\left(\chi_1^2 > 0.1472\right) \;\approx\; 0.7012.$$

As the p-value is larger than $\alpha = 0.05$, we do not reject $\mathcal{H}_0$. This means that the three-way interaction between gender, pre-marital sex, and extra-marital sex (denoted by β_{ijk}^{WXY}) can be removed from the model.

(c) **What is the final model? (2 marks)**

The final model is

$$\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) \;=\; \beta_0 + \beta_i^W + \beta_j^X + \beta_k^Y + \beta_{jk}^{XY}.$$

Now consider the logit model, which we will denote by M_1,

$$\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) \;=\; \beta_0 + \beta_i^W + \beta_j^X + \beta_k^Y + \beta_{jk}^{XY}.$$

which uses a reference level parametrisation for all factors.

Figure 3 shows relevant summary output from SAS.

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	0.6978	3	0.2326	0.8737
Pearson	0.7013	3	0.2338	0.8729

Number of unique profiles: 8

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1435.976	1336.718
SC	1440.919	1361.434
-2 Log L	1433.976	1326.718

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	107.2582	4	<.0001
Score	101.8209	4	<.0001
Wald	87.3775	4	<.0001

Analysis of Maximum Likelihood Estimates							
Parameter			DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept			1	1.3049	0.3150	17.1594	<.0001
GENDER	Man		1	-0.3089	0.1458	4.4870	0.0342
PREMARITAL_SEX	No		1	0.7004	0.4850	2.0851	0.1487
EXTRAMARITAL_SEX	No		1	-0.5962	0.3366	3.1375	0.0765
PREMARITA*EXTRAMARIT	No	No	1	-1.7999	0.5130	12.3119	0.0005

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
GENDER Man vs Woman	0.734	0.552	0.977

Figure 3: Summary output for the logit model $\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \beta_0 + \beta_i^W + \beta_j^X + \beta_k^Y + \beta_{jk}^{XY}$.

(d) **Carry out an appropriate goodness-of-fit test to determine whether model M_1 provides a good fit to the data. State the hypotheses, and give the test statistic and the p-value of the test. What do you conclude at the $\alpha = 0.05$ significance level? (4 marks)**

We wish to test the hypotheses

$$\mathcal{H}_0 : \text{The model provides a good fit.}$$
$$\mathcal{H}_1 : \text{The model does not provide a good fit.}$$

The likelihood ratio goodness-of-fit test statistic is given by

$$G^2 \approx 0.6978,$$

which follows a χ^2_3 distribution asymptotically.

The p-value is given by

$$p\text{-value} \approx P\left(\chi^2_3 > 0.6978\right) \approx 0.8737.$$

As the p-value is greater than $\alpha = 0.05$, we do not reject $\mathcal{H}_0$, and we conclude that the logistic regression model provides a good fit to the data.

(e) **Compare the odds of divorce for men with the odds of divorce for women using an odds ratio, and interpret this odds ratio. Give a 95% confidence interval for the odds ratio. (3 marks)**

$$\exp\left(\widehat{\beta}_1^W\right) \approx \exp(-0.3089) \approx 0.734$$

The odds of divorce for men is estimated to be 0.734 (0.552, 0.977) times the odds of divorce for women, adjusting for pre-marital sex status and extra-marital sex status.