Solutions to Methods in Fall 2004

1. (a) The regression line for bachelor is

$$Salary = 22.93548 + 0.07944 Month$$

and the regression line for master is

$$\begin{aligned} Salary =& (22.93548 + 11.73176) + (0.707944 - 0.0266) Month \\ =& 34.6672 + 0.6813 Month. \end{aligned}$$

(b) The starting salary for bachelor is 22.93548 and the starting salary for master is 34.67. They are significantly different since the p-value for education is 0.0212.

(c) Yes, since the p-values of month is very small. Since the interaction effect is not significant, the association is not significantly different across the two degree levels.

(d) It is $27 - 4 = 23$.

(e) Let β be the coefficient of the interaction effect. Recall that

$$\frac{SSE_2 - SSE_1}{SSE_1/23} = \frac{\hat{\beta}^2}{std^2(\hat{\beta})^2}.$$

We have

$$SSE_2 = 292.66 + 18.02 = 310.68.$$

2. (a) The estimated value of n_{ij} is

$$\hat{n}_{ij} = \frac{(n_{i1} + n_{i2}) \sum_{i=1}^{5} n_{ij}}{\sum_{i=1}^{5} (n_{i1} + n_{i2})}.$$

The estimated values of n_{ij} are given in Table 1. The value of X^2 is 6.88. Based on 4 degrees of freedom, it is less than 9.49. So, we accept the independence.

(b) The log-likelihood function of $(p_1, \cdots, p_5)$ is

$$\ell(p_1, \cdots, p_5) = \sum_{i=1}^{5} \left[\binom{n_{i1} + n_{i2}}{n_{i1}} p_i^{n_{i1}} (1 - p_i)^{n_{i2}} \right].$$

We have $\hat{p}_i = n_{i1}/(n_{i1} + n_{i2})$. If $p_1 = \cdots = p_5 = p$, we have

$$\hat{p} = \frac{\sum_{i=1}^{5} n_{ij}}{\sum_{i=1}^{5} (n_{i1} + n_{i2})}.$$

So,

$$G^2 = 2[\log(\hat{p}_1, \cdots, \hat{p}_5) - \log(\hat{p})] = 2 \sum_{i=1}^{5} \sum_{j=1}^{2} n_{ij} \log(n_{ij}/\hat{n}_{ij}).$$

Under the data, we have $G^2 = 7.28 < 9.49$. So we still accept the independence.

Table 1: Estimated under Independence

	Independence		Logistic	
	Degree		Degree	
Change	High	Low	High	Low
Worse	3.1837	8.8163	1.6819	10.3181
No Change	17.5102	48.4898	12.8027	53.1973
Little Improvement	15.3878	42.6122	15.2061	42.7939
Moderate Improvement	11.1429	30.8571	14.4525	27.5475
Big Improvement	4.7755	13.2245	7.8569	10.1431

(c) $X^2 = 0.58$ and $G^2 = 0.63$. It shows that the logistic linear model fits very well since they are much low that 3.84 and almost close to 0. Since $0.3897/0.1532 = 2.5438 > 1.96$. We reject the independence.

(d) This model is better than the independent model since the p-values of X^2 and G^2 are much greater than those of independence model, and the difference of the X^2, which is 6.31, and the difference of the G^2, which is 6.65, contribute over 90% on the values.

(e) The low frequency may cause the approximation of χ^2 distribution and normal distribution worse. But it is not a big problem here since all of other frequencies are greater than or equal to 7.

3. (a) This is a typical misleading in statistic. The power can not be evaluated after the data is collected. When the estimated values $\hat{\sigma}^2$ and $\hat{\Delta}$ are used to computed the power function, the power function can not be over $1-\alpha$. Otherwise, the rejection of the null hypothesis is concluded. In addition, the underline distribution also depends on unknown parameters.

(b) We consider the case when C is large or n is large. Suppose it is a t-confidence interval. Then, the null is rejected if

$$-C \leq \hat{\Delta} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \leq \hat{\Delta} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \leq C$$

which is equivalent to

$$t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \leq \hat{\Delta} + C \leq \hat{\Delta} + C + 2t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \leq 2C + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}.$$

When $\Delta = -C$, $\hat{\Delta} + C \sim N(0, \sigma^2/n)$, which gives

$$P(t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \leq \hat{\Delta} + C) = 1 - \alpha/2.$$

However, if we also need to consider

$$P(\hat{\Delta} - C \leq 2C - t_{\alpha,n-1}\frac{s}{\sqrt{n}})$$

If C is large or n is large, the above is almost 1. Thus, we need $\alpha/2 = 0.05$ which is equivalent $\alpha = 0.1$.

(c) Yes. Since in this case, the type I error probability is as we expected. The confidence interval did not use any post-experiment data as the pre-experiment inferences.

4. (a) Both the deviance and its degrees of freedom are 0.

(b) The value of the dispersion parameter is 1.

(c) The degrees of freedom are 4.

(d) The p-value tells us the significance of sex effect in the additive model, that is the significance of sex in the model with only the three main effect term.

(e) Note that party:sex and party:ideology interaction effects are significant. This is the the conditional independence model. It means given party, ideology and sex are independent.

5. (a) When a parameter model is known fitted for the data. The nonparametric method will not be as easy as to be explained as the parametric model. The second is that the result for nonparametric model is not as stable as the parametric model.

(b) If we plot $\log[\hat{S}(x)/(1 - \hat{S}(x))]$, we would be able to see a linear trend. If the model is true, since in this case

$$\log \frac{S(x)}{1 - S(x)} = -x.$$

(c) The generalized linear model is fitted according to the logistic like as

$$\log \frac{S(x)}{1 - S(x)} = \alpha + \beta_i$$

where β_i is the group indicator.

(d) The CDF is

$$F(y) = \int_0^y f(x)dy = \int_0^y \frac{e^\theta \lambda x^{\lambda-1}}{(1 + e^\theta x^\lambda)^2}dx = \frac{e^\theta y^\lambda}{1 + e^\theta y^\lambda}.$$

Thus, the survival function is

$$S(y) = \frac{1}{1 + e^\theta y^\lambda}$$

and the hazard function is

$$h(y) = \frac{f(y)}{S(y)} = \frac{e^{\theta}\lambda y^{\lambda-1}}{1+e^{\theta}y^{\lambda}}.$$

This indicates that the cumulative hazard function is

$$H(y) = \int_0^y h(x)dx = \log(1+e^{\theta}y^{\lambda}).$$

The median survival time x_M solves

$$S(y_M) = \frac{1}{2} \Rightarrow y_M = e^{-\theta/\lambda}.$$

6. (a) Let P, I and A denote in favor, attitude and payment respectively. Let i, j, k index for A, I and P respectively. Then, we have $\bar{y}_{\cdots} = 27.15$. The sum of squares are computed based on the average.

Source	SS	DF	MS	F
A	288.05	2	144.02	28.8
I	1080.00	1	1080.00	216
P	1.20	1	1.20	0.24
A:I	0.35	2	0.18	0.36
A:P	0.35	2	0.18	0.36
I:P	140.83	1	140.83	28.02
A:I:P	0.32	2	0.16	0.06
Error	540	108	5	

(b) The A and I main effects are the significant main effects. The I:P interaction effect are also significant since $F_{0.05,1,108} = 3.08$ and $F_{0.05,1,108} = 3.93$. Thus, all the three factors are significant. Note that a factor is significant means at least one effect related to this factor is significant.

(c) The three significant effects are A, I main effects and I:P interaction effect. P main effect is not significant. Those are based on the F test.

(d) The 95% confidence interval is

$$[27.15 - t_{0.025,108}\sqrt{\frac{MSE}{120}}, 27.15 + t_{0.025,108}\sqrt{\frac{MSE}{120}}] = [26.72, 27.58].$$