QUALIFYING EXAM SOLUTIONS
Statistical Methods
Winter 2013

1. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the estimates of β_0 and β_1. Then, the residual is $\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$.

   (a) Note that $Cov(\hat{\epsilon}, \hat{Y}_i) = 0$. We have

   $$\begin{aligned} Cov(\hat{\epsilon}_i, Y_i) =& Cov(\hat{\epsilon}_i, Y_i - \hat{\epsilon}_i) + Cov(\hat{\epsilon}_i, \hat{\epsilon}_i) \\ =& Cov(\hat{\epsilon}_i, \hat{\epsilon}_i) \\ =& V(\hat{\epsilon}_i). \geq 0. \end{aligned}$$

   Therefore, the plot is expected.

   (b) Straightforwardly, the sample correlation is

   $$\begin{aligned} \frac{\sum_{i=1}^n \hat{\epsilon}_i (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n \hat{\epsilon}_i^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} =& \frac{\sum_{i=1}^n \hat{\epsilon}_i (\hat{\epsilon}_i - \bar{Y})}{\sqrt{\sum_{i=1}^n \hat{\epsilon}_i^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \\ =& \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{\sqrt{\sum_{i=1}^n \hat{\epsilon}_i^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \\ =& \sqrt{\frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \\ =& \sqrt{1 - R^2}. \end{aligned}$$

   (c) We need to look at the plot between $\hat{Y}_i$ and $\hat{\epsilon}_i$.

2. I would like to choose the first one because it is easy in the analysis. In 2, it only gives three good and two bad. Like 0 or 1 value. In 3, in only gives one is better than the other. We have have $A > B$, $B > C$, $C > A$ cases. After we choose 1, we can fit an ANOVA model as

   $$y_{ij} = \mu_i + \epsilon_{ij}$$

   where $\epsilon_{ij} \sim N(0, \sigma^2)$.

3. (a) It is known that $R^2 = 0.8455$ and $\hat{\sigma}^2 = 5.208$. Complete the following ANOVA table.

	DF	SS	MS	F-value
x_1	1	401.87	401.87	77.2
x_2	1	310.30	310.30	59.609
Error	25	130.14	5.208	
Total	27	842.31		

   (b) Compute $V(\hat{\beta}_1)$ and $V(\hat{\beta}_2)$.

   $$V(\hat{\beta}_1) = 0.03133\hat{\sigma}^2 = 0.03133 \times 5.208 = 0.1632$$

   and

   $$V(\hat{\beta}_2) = 0.0001908 \times 5.208 = 0.0009926.$$

(c) Compute the 99% confidence intervals for the mean value and observation, respectively, when $x_1 = 10$ and $x_2 = 20$.

$$\hat{y} = 14.6702 + 2.1353 \times 10 + 0.2433 \times 20 = 40.89$$

and

$$V(\hat{y}) = \hat{\sigma}^2 \begin{pmatrix} 1 & 10 & 20 \end{pmatrix} \begin{pmatrix} 3.77929203 & -0.3367167938 & 0.0052268597 \\ -0.33671679 & 0.0313348449 & -0.0009087777 \\ 0.00522686 & -0.0009087777 & 0.0001907578 \end{pmatrix} \begin{pmatrix} 1 \\ 10 \\ 20 \end{pmatrix} = 0.1003.$$

Therefore, the 99% confidence interval for the mean value is

$$40.89 \pm 2.787 \times \sqrt{0.1003} = [40.01, 41.77]$$

and the 99% confidence interval for the observation

$$40.89 \pm 2.787 \times \sqrt{0.1003 + 5.208} = [34.469, 47.311].$$

(d) The R^2 value of the model with x_1 and x_2 interaction effect is $R^2 = 0.9024$. Compute the F-value of the interaction effect.

The SSE is

$$SSE = SST \times R^2 = 842.31 \times (1 - 0.9024) = 82.21.$$

Then,

$$\hat{\sigma}^2 = \frac{82.21}{24} = 3.425.$$

Thus, the F-value of the interaction effect is

$$F^* = \frac{130.14 - 82.21}{3.425} = 13.99.$$

(e) Complete the following ANOVA table.

	DF	SS	MS	F-value
x_1	1	401.87	401.87	117.33
x_2	1	310.30	310.30	90.60
$x_1 : x_2$	1	47.93	47.93	13.99
Error	24	82.21	3.425	
Total	27	842.31		

4. The data reported the counts between income and job satisfaction of 2699 employees.

	Job Satisfaction			
Income	Very Low	Low	High	Very High
Very Low	59(43.17)	116(97.43)	257(275.27)	637(??)
Low	30(27.34)	63(61.71)	186(174.33)	398(??)
High	10(19.75)	41(44.57)	127(125.92)	311(??)
Very High	10(??)	26(??)	125(??)	303(??)

```
> summary(gg)
glm(formula = yy ~ factor(Income) + factor(Job) + ll, family = poisson)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.90044    0.10071  38.731  < 2e-16 ***
factor(Income)2  0.63111    0.12101   5.215 1.84e-07 ***
factor(Income)3  1.47641    0.13012  11.347  < 2e-16 ***
factor(Income)4  2.13600    0.16057  13.303  < 2e-16 ***
factor(Job)2    -0.78263    0.08884  -8.810  < 2e-16 ***
factor(Job)3    -1.43993    0.16044  -8.975  < 2e-16 ***
factor(Job)4    -1.82992    0.23671  -7.731 1.07e-14 ***
ll               0.09578    0.02198   4.359 1.31e-05 ***
    Null deviance: 2452.442  on 15  degrees of freedom
Residual deviance:   12.615  on  8  degrees of freedom
```

(a) Compute the odds ratio and its 95% confidence interval of Job Satisfaction Low and High between Income Low versus Income High (i.e. the second and third rows versus the second and third columns).

$$\hat{\theta} = \frac{63 \times 127}{41 \times 186} = 1.0492$$

and

$$\sigma_{\log \hat{\theta}} = \sqrt{\frac{1}{63} + \frac{1}{127} + \frac{1}{41} + \frac{1}{186}} = 0.2313.$$

The 95% confidence interval is

$$1.0492 e^{-1.96 \times 0.2313} = [0.6668, 1.6510].$$

(b) The fitted counts of the main effect loglinear model are given inside parentheses, but some of fitted values are missing. Complete the table. Compute the deviance goodness-of-fit and Pearson goodness-of-fit. Test whether this model fits the data.

Note that the row total and column total are the same for the observed counts and the fitted counts. The table can be completed as

	Job Satisfaction			
Income	Very Low	Low	High	Very High
Very Low	59(43.17)	116(97.43)	257(275.27)	637(653.12)
Low	30(27.34)	63(61.71)	186(174.33)	398(413.62)
High	10(19.75)	41(44.57)	127(125.92)	311(298.76)
Very High	10(18.74)	26(42.29)	125(119.48)	303(283.49)

The goodness-of-fit statistics are

$$G^2 = 2 \sum_{i=1}^{4} \sum_{j=1}^{4} n_{ij} \log \frac{n_{ij}}{\hat{n}_{ij}} = 32.27$$

and

$$X^2 = \sum_{i=1}^{4}\sum_{j=1}^{4} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = 30.17.$$

Because they are greater than $\chi^2_{0.05,9} = 16.92$, we conclude that the main effect model does not fit the data.

(c) A linear-by-linear term is defined by taking the product of score values of income and job satisfaction, where the score values are 1, 2, 3, and 4 for levels from low to high, respectively. Compute the fitted values when both Income and Job Satisfaction are "Very High" in the linear-by-linear association model. Test whether this model fits the data.

The fitted count is

$$\hat{n}_{44} = e^{3.90044+2.136-1.82992+0.09578\times 16} = 310.75.$$

Because $G^2 = 12.615 < \chi^2_{0.05,8} = 15.51$, we conclude the model fits the data.

(d) The column-effect model, which uses the score values of "Income" in the interaction term, had 11.448 residual deviance value. Test whether the column-effect model can be reduced to the linear-by-linear association model. State the null hypothesis, test statistic, and the conclusion.

The column-effect model is

$$\log \lambda_{ij} = \mu + \alpha_i + \beta_j + i\gamma_j$$

for $i = 1, 2, 3, 4$ and $j = 1, 2, 3, 4$, where μ, α_i, β_j, and γ_j are parameters, with constant $\gamma_1 = 0$. The column-effect model reduces to the linear-by-linear association model is

$$H_0 : \gamma_4 - \gamma_3 = \gamma_3 - \gamma_2 = \gamma_2 - \gamma_1.$$

The test statistic is the difference of the G^2 value as

$$12.615 - 11.448 < \chi^2_{0.05,2} = 5.99.$$

We conclude that the column-effect model can reduce to the linear-by-linear association model.

5. The following table reports the data of the survival time in weeks of stomach lung patients.

Placebo	3 4 4+ 5 7 7 7+ 9+ 10 10+ 10+ 12 12 14 15+ 16 17+ 17+ 18 20
Treatment	7 9 9+ 10 10+ 11 11 13 14+ 15 16 19 19+ 21 22+ 23 24 24+ 26 27+

(a) Compute the Kaplan-Meier estimate of the survival function for the placebo group for $t \leq 10$.

The survival function is

$$\hat{S}(t) = \begin{cases} 1 & \text{when } t < 3 \\ 19/20 = 0.950 & \text{when } 3 \le t < 4 \\ 0.950(1 - 1/19) = 0.9 & \text{when } 4 \le t < 5 \\ 0.950(1 - 1/17) = 0.847 & \text{when } 5 \le t < 7 \\ 0.847(1 - 2/16) = 0.741 & \text{when } 7 \le t < 10 \\ 0.741(1 - 1/12) = 0.679 & \text{when } 10 \le t < 12. \end{cases}$$

(b) Assume the survival time follows the exponential distribution. Estimate the survival functions and compute the expected survival time for placebo and treatment groups, respectively.

For the placebo group, we have

$$\hat{\lambda}_1 = \frac{\sum_{i=1}^{20} \delta_i}{\sum_{i=1}^{20} t_i} = 0.0553.$$

The expected survival time is $1/0.0533 = 18.762$. For the treatment group, we have

$$\hat{\lambda}_2 = \frac{\sum_{i=1}^{20} \delta_i}{\sum_{i=1}^{20} t_i} = 0.0394.$$

The expected survival time is $1/0.0394 = 25.381$.

(c) Assume the survival time follows the Weibull distribution and the **R** output is given below. Write down the fitted survival functions. Test whether the Weibull distribution can reduce to the exponential distribution.

```
survreg(formula = Surv(weeks, censor) ~ factor(treat), data = lung,
    dist = "weibull")
                Value Std. Error     z         p
(Intercept)     2.725      0.119 22.92 2.73e-116
factor(treat)2  0.348      0.165  2.11  3.46e-02
Log(scale)     -0.888      0.163 -5.44  5.24e-08
```

The scale parameter is

$$\hat{\alpha} = e^{-0.888} = 0.4115.$$

The survival function for the placebo group is

$$\hat{S}(t) = e^{-e^{-2.725} t^{0.4115}} = e^{0.0655 t^{0.4115}}$$

and the survival function for the treatment group is

$$\hat{S}(t) = e^{-e^{-2.725-0.348} t^{0.4115}} = e^{-0.0463 t^{0.4115}}.$$

Because the p-value of Log(scale) is small, we conclude the Weibull distribution cannot reduce to the exponential distribution.

(d) Suppose the cox proportional hazard model is used and the fitted survival functions are given in the following table. Complete the table.

Time	Placebo	Treatment
3	0.9652	0.9856
4	0.9303	0.9708
5	0.8941	0.9551
7	0.7865	0.9062
9	0.7496	0.8885
10	0.6732	0.8502
11	0.5889	0.8047
12	0.5095	0.7583
13	0.4698	0.7335
14	0.4320	0.7087