QUALIFYING EXAM SOLUTIONS
Statistical Methods
Spring, 2012

1. (a) Let Y_i be the amount $^{13}CO_2$ (the response) and X_i be the protein used (the explanatory variable). We consider the following model

$$Y_i = f(X_i) + \epsilon_i,$$

where f is a function and $\epsilon_i \sim N(0, \sigma^2)$ is the iid error term. We may use

$$f(x) = \begin{cases} \alpha + \beta x, & \text{if } x \le x_0 \\ \alpha + \beta x_0, & \text{if } x > x_0 \end{cases}$$

From the plot, we may use $x_0 = 0.8$. However, x_0 may also be treated as a parameter to be estimated from the data.

(b) Let y be the vector of resposnse and x be the vector of the independent variable. We can consider the R code as

```
ff <- function(x,x0,ALPHA,BETA){
 SS <- ALPHA+BETA*x0+BETA*(x-x0)*(x<=x0)
 SS
}
```

The function will be used to analyze the data in the next part.

(c) For a given f, the loglikelihood function is

$$\ell(\alpha, \beta, x0, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - f(x_i)]^2.$$

Therefore, we can compute the loglikelihood function in R

```
loglikelihood <- function(ALPHA,BETA,x0,sigmasq,y,x){
 nn <- length(y)
 SS <- -(n/2)*log(2*pi)-(n/2)*log(sigmasq)-(1/(2*sigmasq))*sum((y-ff(x))^2)
 SS
}
gg <- lm(y[x<=0]~x[x<=0])
```

Based on the functions and results, we can derive $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$ for a given $x0$. Then, we can maximize the loglikelihood function.

(d) It may contain the weekly cycle.

(e) The amount of the protein used cannot be mixed with the weekly cycle. It is not necessary to consider the randomization of the order of the subject, because each subject will be assigned the same amount of protein equally, such that the design is balanced.

2. (a) Note that $R^2 = SSM/(SSE + SSM)$. We have

$$SSM = \frac{R^2 \times SSE}{1 - R^2} =$$

Source	DF	SS	MS	F-value
Model	1	244210.3	244210.3	11101.1
Error	10	219.9873	21.99873	
Total	11	244430.3		

(b) The correlation is

$$\hat{\rho} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2 \sum_{i=1}^n (X_i - \bar{X})^2}}.$$

Let the model be

$$Y_i = \alpha + \beta X_i + N(0, \sigma^2).$$

Then

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

and

$$\hat{\beta} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \hat{\rho}\sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}.$$

Therefore,

$$\begin{aligned}\hat{Y}_i =& \alpha + \hat{\beta} X_i \\ =& (\bar{Y} - \hat{\beta}\bar{X}) + \hat{\beta} X_i \\ =& \bar{Y} + \hat{\beta}(X_i - \bar{X})\end{aligned}$$

$$\begin{aligned}SSE =& \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ =& \sum_{i=1}^n [Y_i - (\hat{\alpha} + \hat{\beta} X_i)]^2 \\ =& \sum_{i=1}^n [(Y_i - \bar{Y}) - \hat{\beta}(X_i - \bar{X})]^2 \\ =& \sum_{i=1}^n (Y_i - \bar{Y})^2 + \hat{\beta}^2 \sum_{i=1}^n (X_i - \bar{X})^2 - 2\hat{\beta} \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) \\ =& SST - \hat{\rho}^2 SST.\end{aligned}$$

Thus

$$\hat{\rho}^2 = 1 - SSE/SST = R^2 = 0.9991,$$

which implies $\hat{\rho} = \sqrt{0.9991} = 0.9995$ (because $\hat{\rho} > 0$).

(c) The total 95% confidence interval is $2[\hat{\mu} \pm t_{0.025,10} 1.6163]$, where $\hat{\mu}$ is predicted value of the response at $x = 9$. We do not have enough information to compute $\hat{\mu}$.

3. (a) Let Y_{ij} be the measurement at angle i for $j = 1, 2$. Then, the linear model is

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

where $\epsilon_{ij} \sim^{iid} N(0, \sigma^2)$ is the error term.

(b) We test

$$H_0 : \mu_1 = \mu_2 = \mu_3 \leftrightarrow H_1 : \text{They are not all equal.}$$

We define an F-statistic

$$F^* = \frac{\sum_{i=1}^{3}(\bar{Y_i} - \bar{Y_{..}})^2/3}{\sum_{i=1}^{3}\sum_{j=1}^{2}(Y_{ij} - \bar{Y_i})^2/6}$$

where $\bar{Y_{i\cdot}} = (Y_{i1} + Y_{i2})/2$ and $\bar{Y_{..}} = \sum_{i=1}^{3}\sum_{j=1}^{2} Y_{ij}/6$. Under H_0, $F^* \sim F_{2,3}$. Thus, H_0 is rejected if $F^* > F_{0.05,2,3}$.

4. (a) The cumulative odds ratio is

$$\hat{\theta} = \frac{272(835 + 131)}{454(294 + 49)} = 1.6873.$$

(b) Let π_i be the probability for happy to be the first, second, or third categories, espectively. It assume the count are Poisson distributed, which can be modeled by a multinomial model. The proportional odds model can be used, which is

$$\log \frac{\pi_1}{\pi_2 + \pi_3} = \theta_1 - \beta_2 I_{Income=2} - \beta_3 I_{Income=3}$$

and

$$\log \frac{\pi_1 + \pi_2}{\pi_3} = \theta_2 - \beta_2 I_{Income=2} - \beta_3 I_{Income=3}$$

The estimate of parameters are $\hat{\theta}_1 = -0.2521$, $\hat{\theta}_2 = 2.5643$, $\hat{\beta}_2 = 0.4501$, and $\hat{\beta}_3 = 1.2369$. Then, the odds ratio in part (a) is

$$\hat{\theta} = e^{0.4501} = 1.5685.$$

(c) Let the score of income be 1, 2, and 3, for rows 1, 2, and 3, respectively. Then, the proportional odds model is

$$\log \frac{\pi_1}{\pi_2 + \pi_3} = \theta_1 - \beta Income$$

and

$$\log \frac{\pi_1 + \pi_2}{\pi_3} = \theta_2 - \beta Income.$$

The difference beteen this model and the previous model is the independent variable. In the previous model, income is treated as a factor variable, but in this model it is treated as a continuous variable. This model can interpret the change of odds ratio when income increases a level, which cannot be found in the previous model. Based on the estimates $\hat{\theta}_1 = 0.4789$, $\hat{\theta}_2 = 3.2863$ and $\hat{\beta} = 0.6313$. We have the predict count in the following table:

	Happiness		
Income	1	2	3
1	284.13	290.53	40.34
2	445.23	809.22	165.54
3	179.82	557.25	182.93

Because the deviance goodness of fit is

$$G^2 = 5487.699 - 5482.358 - 5.341$$

which is greater than $\chi^2_{0.05,1} = 3.84$, we conclude that income cannot be treated as a continuous variable.

5. We can use the Pearson test. The result is given in the following table:

Time (hours)	$0-300$	$301-600$	$601-900$	$901-1200$	≥ 1201	Total
Count	206	72	14	6	2	300
$\hat{p}_i$	0.7042	0.2083	0.0616	0.0182	0.0077	1
$\hat{n}_i$	211.27	62.49	18.48	5.47	2.30	300

Then

$$\begin{aligned} X^2 =& \frac{(206-211.27)^2}{211.27} + \frac{(62.49-72)^2}{62.49} + \frac{(18.48-14)^2}{18.48} + \frac{(5.47-6)^2}{5.47} + \frac{(2-2.30)^2}{2.30} \\ =& 2.7566 \end{aligned}$$

which is less than $\chi^2_{0.05,4} = 9.04$. Therefore, we accept H_0 : it is an exponential distribution.

6. (a) The PDF is $f(t) = \lambda e^{-\lambda t}$ and the survival function is $S(t) = e^{-\lambda t}$. Thus, the likelihood function is

$$L(\lambda) = \prod_{i=1}^{n} f^{\delta_i}(t_i) S^{1-\delta_i}(t_i) = \prod_{i=1}^{n} \lambda^{\delta_i} e^{-\lambda t_i},$$

where $\delta_i = 1$ if the i-th case is death and $\delta_i = 0$ if the i-th case is sensored. The loglikelihood function is

$$\ell(\lambda) = (\log \lambda) \sum_{i=1}^{n} \delta_i - \lambda \sum_{i=1}^{n} t_i.$$

Then,

$$\ell'(\lambda) = \frac{\sum_{i=1}^{n} \delta_i}{\lambda} - \sum_{i=1}^{n} t_i \Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i}.$$

Therefore, for females, we have

$$\hat{\lambda} = \frac{13}{1702} = 0.007638$$

and

$$\hat{S}(7) = e^{-0.007638 \times 7} = 0.9479.$$

(b) The model is

$$-\log \lambda_j = 3.65 + 1.23 Sex_j$$

and we have $\hat{\lambda} = 0.02599$ for male and $\hat{\lambda} = 0.007597$ for female. The estimates of the survival functions are $\hat{S}(7) = 0.8337$ for male and $\hat{S}(7) = 0.9482$ for female.