STA 4504/5503 - Practice set 3 (with solutions)

True or False

1. Subjects suffering from mental depression are measured after 1 week of treatment, 2 weeks of treatment, and 4 weeks of treatment in terms of a (normal, abnormal) response outcome. Covariates are severity of condition at original diagnosis (1 = severe, 0 = mild) and treatment used (1 = new, 0 = standard). Since each subject contributes three observations to the analysis, we can use the GEE (generalized estimating equations) method to fit the model. To use this method, we must choose a working correlation matrix for the form of the dependence among the three responses, but the method is robust in the sense that it still gives appropriate estimates and standard errors for large n even if the actual correlation structure is somewhat different from the one we assumed.
   TRUE

2. With a single categorical response variable, logistic regression models are more appropriate than loglinear models.
   TRUE

3. GEE models are also called a *subject-specific models*.
   FALSE

4. When a GEE and a GLMM model is fitted on the same dataset, the coefficient β_1 associated with predictor x_1 is interpreted the same for both models.
   FALSE

5. McNemar's exact test as in `mcnemar.exact{exact2x2}` calculates p-values using the χ^2 distribution.
   FALSE

Open Ended Problems

1. You decide to use GEE methods to handle dependent observations because of repeated measurement or clustering of some type.

   (a) Explain what is meant by an exchangeable "working correlation matrix".
   Answer: The *exchangeable* correlation structure assumes that Y_i and Y_j have correlation ρ for all i, j within the same cluster

$$\text{Cor}(\boldsymbol{Y}) = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

but for i and j in different clusters, correlation is 0. E.g. Consider Y_1, Y_2 to be in same cluster and Y_3, Y_4 to be a different (independent) cluster. Then,

$$\text{Cor}(\boldsymbol{Y}) = \begin{pmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{pmatrix}$$

(b) Describe one other type of "dependence matrix" you can use (except independence).
Answer: The *autoregressive* correlation structure assumes that Y_i and Y_j have correlation $\rho^{|i-j|}$. This is commonly used when $(Y_1, \ldots, Y_T)$ represent repeated observations over time where outcomes closer in time are more correlated.

$$\text{Cor}(\boldsymbol{Y}) = \begin{pmatrix} 1 & \rho & \cdots & \cdots & \rho^{T-1} \\ \rho & 1 & \cdots & \cdots & \rho^{T-2} \\ \rho^2 & \rho & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \cdots & \rho & 1 \end{pmatrix}$$

(c) If you ignore the dependence, will there be bias in your (i) parameter estimates, (ii) standard error estimates?
Answer: (i) Fitting method gives estimates that are consistent even if correlation structure is miss-specified. (ii) Adjusts standard errors to reflect actual observed dependence. Therefore, overly complicated structures are not encouraged. However, if dependence was neglected by fitting a GLM (not a GEE) here in `R` estimation of standard errors would be biased.

2. Consider the loglinear model of independence for a two-way contingency table. This has equation for expected frequencies $\{\mu_{ij}\}$ in an $I \times J$ contingency table,

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$$

Motivate this model, by showing how the definition of statistical independence of two categorical variables implies that a loglinear model of this form holds.
Answer:

$$\mu_{ij} = n\pi_{ij} \overset{\text{ind.}}{=} n\pi_{i+}\pi_{+j} \Rightarrow \log(\mu_{ij}) = \underbrace{\log(n)}_{\lambda} + \underbrace{\log \pi_{i+}}_{\lambda_i^X} + \underbrace{\log \pi_{+j}}_{\lambda_j^Y}$$

3. To allow for association between X and Y, this model is extended to

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

(a) For a 2×2 contingency table, express the log odds ratio in terms of expected frequencies, and use it to show that the odds ratio for this model equals $\exp(\lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}$. (Hence the two-factor interaction parameters provide information about the XY association.)
Answer:

$$\begin{aligned}\log\left(\frac{\mu_{ij}\mu_{i'j'}}{\mu_{ij'}\mu_{i'j}}\right) =& \log(\mu_{ij}) + \log(\mu_{i'j'}) - \log(\mu_{ij'}) - \log(\mu_{i'j}) \\ =&(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}) + (\lambda + \lambda_{i'}^X + \lambda_{j'}^Y + \lambda_{i'j'}^{XY}) \\ &- (\lambda + \lambda_i^X + \lambda_{j'}^Y + \lambda_{ij'}^{XY}) - (\lambda + \lambda_{i'}^X + \lambda_j^Y + \lambda_{i'j}^{XY}) \\ =&\lambda_{ij}^{XY} + \lambda_{i'j'}^{XY} - \lambda_{ij'}^{XY} - \lambda_{i'j}^{XY}\end{aligned}$$

(b) What will the degrees of freedom and Deviance be for this model?
Answer: Saturated model so 0 and 0.

4. Two tests are developed to detect movement related pain at the shoulder.

	Test 2				
Test 1	No pain	Mild pain	Moderate pain	Severe pain	Total
No pain	15	3	1	1	20
Mild pain	4	18	3	2	27
Moderate pain	4	5	16	4	29
Severe pain	1	2	4	17	24
Total	24	28	24	24	100

(a) For rater agreement, calculate Cohen's (unweighted) Kappa κ.
Answer:

```
> pain=matrix(c(15,4,4,1,
+               3,18,5,2,
+               1,3,16,4,
+               1,2,4,17),4,4,byrow=F,
+               dimnames=list(c("No","Mild","Moderate","Severe"),
+                                    c("No","Mild","Moderate","Severe")))
> library(psych)
> cohen.kappa(pain)

                 lower estimate upper
unweighted kappa  0.42     0.55  0.67
weighted kappa    0.53     0.67  0.81
```

(b) Should a weighted Kappa be used instead in this case? Why, and if so provide the weighting scheme.
Answer: Yes since the responses are ordered. The default weight scheme used was

```
> cohen.kappa(pain)$weight
```

```
                No      Mild  Moderate    Severe
No       1.0000000 0.8888889 0.5555556 0.0000000
Mild     0.8888889 1.0000000 0.8888889 0.5555556
Moderate 0.5555556 0.8888889 1.0000000 0.8888889
Severe   0.0000000 0.5555556 0.8888889 1.0000000
```

(c) What conclusion can you make?
Answer: We can create a 95% CI using

```
> sqrt(cohen.kappa(pain)$var.kappa)
[1] 0.06323179
```

to get

$$\hat{\kappa} \mp 1.96(0.0632)$$

but that is already done. For the weighted kappa we have (0.67, 0.81) so that significantly greater than 0.5, so a moderate positive rater agreement.