Problem 1

(a)

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-0.2926	0.3721	-1.0220	0.4367	-0.79	0.4316
trt	-1.0702	0.5447	-2.1377	-0.0027	-1.96	0.0494
time	0.3356	0.1696	0.0031	0.6680	1.98	0.0479
trt*time	1.1109	0.2751	0.5718	1.6500	4.04	<.0001

From the output above and the equation from the exam:

$\hat{\beta}_0 = -0.2926, \hat{\beta}_1 = -1.0702, \hat{\beta}_2 = 0.3356, \hat{\beta}_3 = 1.1109$

$logit\{\pi(trt, t\} = -0.2926 - 1.0702trt + 0.3356time + 1.1109time * trt$

(b)

New treatment (trt = 1): $logit\{\pi(trt = 1, t\} = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 time + \hat{\beta}_3 time$

Standard treatment (trt = 0): $logit\{\pi(trt = 0, t\} = \hat{\beta}_0 + \hat{\beta}_2 time$

$\theta(time) = e^{logit\{\pi(trt=1,t\} - logit\{\pi(trt=0,t\}} = e^{\hat{\beta}_1 + \hat{\beta}_3 time} = e^{-1.0702 + 1.1109 * time}$

Time = 1: $\hat{\theta}(time) = e^{-1.0702+1.1109*1} = 1.0415$

Time = 2: $\hat{\theta}(time) = e^{-1.0702+1.1109*2} = 3.1633$

Time = 3: $\hat{\theta}(time) = e^{-1.0702+1.1109*3} = 9.6071$

Explanation: In the first month (time =1), the odds ratio between the new trt and standard trt is 1.0415, which indicates there is not big difference (almost same) on the effect of new trt and standard trt. The second month odds ratio increase to 3.1633, which means the odds ratio of new trt is around 2 times better than the standard trt. At the third month (time =3), the odds ratio of new trt to standard trt increased to 9.6071, it indicates there is a good effect on new trt compare to standard trt.

(c)

If use an exchangeable working correlation matrix, we still get valid inference. From the GEEs key features, the validity of the inference does not depend on the whether or not the specification of the correlation structure is correct. GEE gives us a robust inference on the regression coefficients, which is valid regardless whether or not the correlation structure specified is right.

(d)

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	-0.3159	0.3589	148	-0.88	0.3802
trt	-1.1786	0.5869	298	-2.01	0.0455
time	0.3562	0.1686	298	2.11	0.0354
trt*time	1.1850	0.3243	298	3.65	0.0003

From the output above and the equation from the exam:

$\hat{\beta}_0 = -0.3159, \hat{\beta}_1 = -1.1786, \hat{\beta}_2 = 0.3562, \hat{\beta}_3 = 1.1850$

$logit\{\pi(trt, t\} = -0.3159 + b_i - 1.1786trt + 0.3562time + 1.1850time$

For subject i, if standard trt assigned, trt = 0: $logit\{\pi(trt,t\} = -0.3159 + b_i + 0.3562time$; So the odds ratio of standard trt with one month increase is $e^{0.3562} = 1.4279$
If new trt assigned, trt = 1: $logit\{\pi(trt,t\} = -1.4945 + b_i + 1.5412time$; Therefore, odds ratio of new trt with one month increase is $e^{1.5412}$ =4.6702.
Therefore, the new trt is better.

(e)

Covariance Parameter Estimates			
Cov Parm	Subject	Estimate	Standard Error
Intercept	id	0.2152	0.3039

The estimate of $\sigma^2$from the previous model is 0.2152.
Interpretation: after adjusting trt and time, the variance of between-subject is 0.2152.
Explanation: As the variance is relatively small, indicates a small difference between subjects, so we can conclude that the estimates of beta from the above two models are close.

Code:

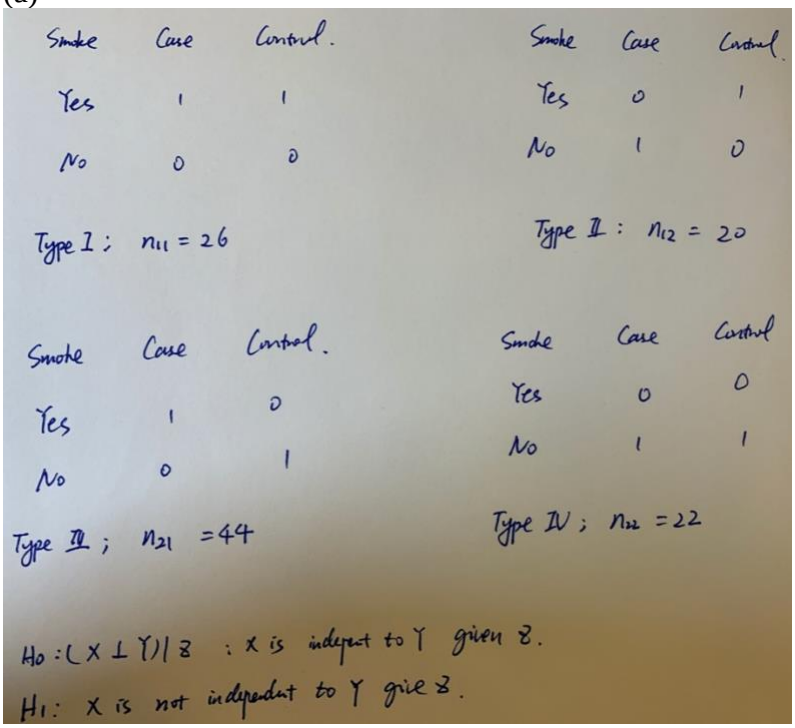
```
data problem1;
input id y trt time;
datalines;
… *(copy from txt file on moodle)*
;

proc genmod descending data=problem1;
class id;
model y = trt time trt*time/dist=bin link=logit;
repeated subject = id / type = un corrw;
run;

proc glimmix method=quad(qpoints=10);
class id;
model y = trt time trt*time/dist=bin link=logit s;
random int/subject = id type=vc;
run;
```

Problem 2

(a)



CMH test statistics: $\chi^2 = \frac{[n_{11}(1-1)+n_{12}(1-0.5)+n_{21}(0-0.5)+n_{22}(0-0)]^2}{n_{11}\times 0+n_{12}\times 0.25+n_{21}\times 0.25+n_{22}\times 0} = 9$

As $\chi^2_{0.05,1} = 3.841$, as our test statistic is greater than 3.841, we reject H0 at 0.05 significance level, so we conclude the X and Y are not independent given Z.

(b)

	Case		
Control	Yes	No	Total
Yes	26	20	46
No	44	22	66
Total	70	42	112

$H_0: \pi_{12} = \pi_{21}$

McNemar's test statistics: $\chi^2 = \frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}} = 9$

As $\chi^2_{0.05,1} = 3.841$, as our test statistic is greater than 3.841, we reject H0 at 0.05 significance level, so we conclude the $\pi_{12} \neq \pi_{21}$.

(c)

$\hat{\theta}_{XY} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{70\times 66}{42\times 46} = 2.3913$ $Var(log\hat{\theta}_{XY}) = \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right) = 0.07499$

95% CI for $\log\hat{\theta}_{XY}$: $log\hat{\theta}_{XY} \pm 1.96 \times \sqrt{Var(log\hat{\theta}_{XY})}$ = (0.3351,1.4086)

95% CI for $\hat{\theta}_{XY}$: $(e^{0.3351}, e^{1.4086}) = (1.3981,4.0902)$

(d)
Because in part c, it assumes the all units are independent to each other. However, from the question we know that the patients were matched to each other (paired). When we put the data in part c, we split the matched data into new form, so they should have some correlations, that's why the estimates from part c is not valid.

(e)

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
smoke	1	0.7885	0.2697	8.5479	0.0035

From the output and the equation from the question, we can conclude that the $\hat{\beta}$ =0.7885
95% $\hat{\beta}$ CI: $0.7885 \pm 1.96 \times 0.2697 = (0.2599, 1.3171)$
$e^{\hat{\beta}} = 2.2001$ and 95% CI for $e^{\hat{\beta}}$ is $e^{0.2599}, e^{1.3171}) = (1.2968, 3.7326)$
Interpretation: The odds ratio of lung cancer patients for smokers is 2.2 to non-smokers.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	9.2238	1	0.0024
Score	9.0000	1	0.0027
Wald	8.5479	1	0.0035

From the output above, we can see the Wald test correspond p-value is 0.0035 which is smaller than 0.05, we reject null hypothesis, X is not independent to Y, consistent with the CMH test.

Code:

```
data problem2;
input smoke_c x1 x2;
cards;
1 26 20
0 44 22
;
run;

data problem2;
set problem2;
array temp{2} x1-x2;
do i=1 to 2;
count = temp(i);
smoke_t = 2-i;
output;
end;
run;

data problem2a;
set problem2;
retain id;
```

```
if _n_=1 then id=0;
do j=1 to count;
id =id + 1;
do k=0 to 1;
if k=0 then
smoke=smoke_c;
else
smoke=smoke_t;
output;
end;
end;
run;

proc logistic descending data=problem2a;
class id;
model k = smoke/ link=logit;
strata id;
run;
```

Problem 3

(a)

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	0	0.0000	0.0000	0.0000	0.0000	.	.
x1	1	0.5936	0.5200	-0.4256	1.6127	1.30	0.2537
x2	1	1.4051	0.5549	0.3175	2.4926	6.41	0.0113
x3	1	-0.1556	0.6781	-1.4846	1.1733	0.05	0.8184
Scale	0	1.0000	0.0000	1.0000	1.0000		

From the output above, $\log\left\{\frac{\Pi_{ij}}{\Pi_{ji}}\right\} = \beta_i - \beta_j \ (i < j)$

$\hat{\beta}_1 = 0.5936, \hat{\beta}_2 = 1.4051, \hat{\beta}_3 = -0.1556, \hat{\beta}_4 = 0$

Criterion	DF	Value	Value/DF
Deviance	3	7.2346	2.4115
Scaled Deviance	3	7.2346	2.4115
Pearson Chi-Square	3	6.4230	2.1410
Scaled Pearson X2	3	6.4230	2.1410

From the output above, Deviance is 7.2346 with df =3. As $\chi^2_{0.05,3} = 7.815$. So, deviance is smaller than $\chi^2_{0.05,3}$. Therefore, we fail to reject the null hypothesis here, the null model fits the data well.

$df = \binom{4}{2} - 4 + 1 = 3$

(b)

$\log\left\{\frac{\Pi_{ij}}{\Pi_{ji}}\right\} = \beta_i - \beta_j \ (i < j)$, if $\Pi_{ij} > \Pi_{ji}, \beta_i - \beta_j > 0$

$\log\left\{\frac{\Pi_{jk}}{\Pi_{kj}}\right\} = \beta_j - \beta_k (j < k)$ if $\Pi_{jk} > \Pi_{kj}, \beta_j - \beta_k > 0$

Therefore, $\beta_i > \beta_j > \beta_k$; $\log\left\{\frac{\Pi_{ik}}{\Pi_{ki}}\right\} = \beta_i - \beta_k > 0$, so $\Pi_{ik} > \Pi_{ki}$

Overall, if i is better than j, and j is better than k, so the i is better than k.

(c)

$\hat{\beta}_1 = 0.5936, \hat{\beta}_2 = 1.4051, \hat{\beta}_3 = -0.1556, \hat{\beta}_4 = 0$

$\hat{\beta}_2 > \hat{\beta}_1 > \hat{\beta}_4 > \hat{\beta}_3$

Therefore, the ranking is B > A > D > C

(d)

Estimated Covariance Matrix			
	Prm2	Prm3	Prm4
Prm2	0.27038	0.14673	0.20466
Prm3	0.14673	0.30790	0.15930
Prm4	0.20466	0.15930	0.45977

$\widehat{\Pi}_{14} = \frac{e^{\hat{\beta}_1 - e^{\hat{\beta}_4}}}{1+e^{\hat{\beta}_1 - \hat{\beta}_4}} = \frac{e^{0.5936}}{1+e^{0.5936}} = 0.6442$

$\widehat{Var}(\hat{\beta}_1 - \hat{\beta}_4) = \widehat{Var}(\hat{\beta}_1) + \widehat{Var}(\hat{\beta}_4) - 2\widehat{Cov}(\hat{\beta}_1, \hat{\beta}_4) = 0.27038$

95% CI for $\hat{\beta}_1 - \hat{\beta}_4$: $0.5936 \pm 1.96 \times \sqrt{0.27038} = (-0.4256, 1.6128)$

95% CI for $\widehat{\Pi}_{14}$ $\left(\frac{e^{-0.4256}}{1+e^{-0.4256}}, \frac{e^{1.6128}}{1+e^{1.6128}}\right) = (0.3952, 0.8338)$

(e)

Estimated Covariance Matrix			
	Prm2	Prm3	Prm4
Prm2	0.27038	0.14673	0.20466
Prm3	0.14673	0.30790	0.15930
Prm4	0.20466	0.15930	0.45977

$\widehat{\Pi}_{13} = \frac{e^{\hat{\beta}_1} - e^{\hat{\beta}_3}}{1+e^{\hat{\beta}_1-\hat{\beta}_4}} = \frac{e^{0.5936+0.1556}}{1+e^{0.5936+0.1556}} = 0.6790$

$\widehat{Var}(\hat{\beta}_1 - \hat{\beta}_3) = \widehat{Var}(\hat{\beta}_1) + \widehat{Var}(\hat{\beta}_3) - 2\widehat{Cov}(\hat{\beta}_1, \hat{\beta}_3) = 0.27038 + 0.45977 - 2 \times 0.2046$
$= 0.32083$

95% CI for $\hat{\beta}_1 - \hat{\beta}_4$: $(0.5936 + 0.1556) \pm 1.96 \times \sqrt{0.32083} = (-0.3610, 1.8594)$

95% CI for $\widehat{\Pi}_{14}$ $\left(\frac{e^{-0.3610}}{1+e^{-0.3610}}, \frac{e^{1.8594}}{1+e^{1.8594}}\right) = (0.4107, 0.8652)$

Code:

```
data problem3;
input winner player $ y1-y4;
cards;
1 A . 1 7 8
2 B 8 . 3 9
3 C 3 2 . 0
4 D 2 3 2 .
;
data problem3;
set problem3;
array temp{4} y1-y4;
do loser =1 to 4;
count = temp(loser);
output;
end;
run;

data problem3;
set problem3;
if winner=loser then delete;
if winner<loser then do;
y=1; ind1=winner; ind2=loser;
end;

else do;
y=0; ind1=loser; ind2=winner;
end;

array x{4};
do k=1 to 4;
```

```
if k=ind1 then
x[k]=1;
else if k=ind2 then
x[k]=-1;

else
x[k]=0;
end;
drop y1-y4 k;
run;

proc sort data=problem3;
by ind1 ind2 descending y;

proc genmod desc data=problem3;
freq count;
model y = x1 x2 x3/ dist=bin link=logit aggregate noint covb;
run;
```

Problem 4

(a)

Source	DF	Chi-Square	Pr > ChiSq
Intercept	3	376.48	<.0001
time	3	39.82	<.0001
Residual	0	.	.

From the output above, we can see the chi-square for time is 39.82 with p-value smaller than 0.0001, we reject the null hypothesis aka reject the marginal homogeneity at level 0.05.

(b)

Empirical Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept1	-1.0516	0.1841	-1.4125	-0.6908	-5.71	<.0001
Intercept2	0.0199	0.1820	-0.3368	0.3766	0.11	0.9128
Intercept3	1.2996	0.1994	0.9087	1.6905	6.52	<.0001
x	-0.9795	0.1665	-1.3059	-0.6531	-5.88	<.0001

From the output above, p-value on x which tests the homogeneity here is smaller than 0.0001, so we reject the null hypothesis, we reject the marginal homogeneity at level 0.05.

(c)

Empirical Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	37.3750	2.1262	33.2077	41.5423	17.58	<.0001
x	12.9583	2.0535	8.9335	16.9832	6.31	<.0001

From the output above, output of the $Y_1 - Y_2$ shows that the difference is significant different at level 0.05 due to the p-value is smaller than 0.0001 < 0.05.
From the output above, we can also conclude that the difference of the mean of $Y_1 - Y_2$ is 12.9583, so there is a decrease in time of falling asleep and the average decrease is 12.9583.

(d)

$$\chi^2 = \sum_{i<j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}} = \frac{(4-14)^2}{4+14} + \frac{(2-6)^2}{2+6} + \frac{(2-14)^2}{2+14} + \frac{(0-11)^2}{0+11} + \frac{(1-4)^2}{1+4} + \frac{(1-9)^2}{1+9} = 35.7556$$

Pearson chi-square test statistic is 35.7556, which is larger than the $\chi^2_{0.05,6} = 12.592$, in this situation, we reject null hypothesis. Therefore, the symmetry is not good, not symmetric.

(e)

We can estimate the probability by $\hat{\pi} = p = \frac{\#\ of\ Y_2 < Y_1}{total\ number} = \frac{14+6+4+9+11+14}{120} \approx 0.48$

$Var(\hat{\pi}) = \frac{p(1-p)}{n} = 0.00208$

95% CI of $\hat{\pi}$: $p \pm 1.96 \times \sqrt{0.00208} = (0.39, 0.57)$

Code:

```
data problem4;
input y1 v1-v4 ;
```

```
cards;
1 7 4 2 1
2 14 5 1 0
3 6 9 18 2
4 4 11 14 22
;

data problem4;
set problem4;
array temp{4} v1-v4;
do y2=1 to 4;
count = temp(y2);
output;
end;
run;

proc catmod data = problem4;
weight count;
response marginals;
model y1*y2 = _response_;
repeated time 2;
run;

data problem4a;
set problem4;
retain id;
if _n_=1 then id=0;
do i=1 to count;
id = id + 1;
do t= 1 to 2;
x = 2-t;
if t = 1 then y = y1;
if t = 2 then y = y2;
output;
end;
end;
run;

proc genmod data= problem4a;
class id;
model y = x/ dist = multinomial link = clogit;
repeated subject = id/ type = ind;
run;

data problem4b;
set problem4a;
```

```
if y = 1 then ttfa = 10;
else if y=2 then ttfa = 25;
else if y=3 then ttfa = 45;
else ttfa = 75;
run;

proc genmod data= problem4b;
class id;
model ttfa = x/ dist = normal link = identity;
repeated subject = id/ type = un;
run;
```