UNIVERSITY OF CALGARY & DEPARTMENT OF MATHEMATICS AND STATISTICS

STATISTICS 635 Assignment 3

Qian Wang

December 14, 2018

1 PROBLEM 1

(a) To combine three categories, we have 32 covariate pattern. The rate p of claims is calculated in Table 1.1. Based on that, we can plot the scatter plot of rate & age, rate &car, rate & district in Figure 1.1.

We can see from the plot, age, car, district may all influence the claims rate. Only from the figure, we can not consider which one is the main effects.

(b) Fit a poisson model:

$$\log(\mu) = \log(n) + \beta_0 + \beta_{car} x_1 + \beta_{age} x_2 + \beta_{dis} x_3 + \beta_{CA} x_1 x_2 + \beta_{CD} x_1 x_3 + \beta_{AD} x_2 x_3$$

where car $x_1 = 2, 3, 4$, and $x_1 = 1$ is the reference; age $x_2 = 2, 3, 4$, and x_2 =1 is the reference; district $x_3 = 1$, and $x_3 = 0$ is the reference. By R program *Problem_1*, we can get the estimate result in Table 1.2. From Table 1.2, we can find that the coefficient of age is the most significant. (i.e. p-value is the smallest). So the age is the main effect.

(c) Fit a poisson model:

$$\log(\mu) = \log(n) + \beta_0 + \beta_{car} x_1 + \beta_{age} x_2 + \beta_{dis} x_3$$

where car is x_1, age is x_2, and x_1 and x_2 are continuous variable. District $x_3 = 1$, and $x_3 = 0$ is the reference. By R program *Problem_1*, we can get the estimate result in Table 1.3.

So the fitted model is

$$\log(\mu) = 3.466 - 1.8525 + 0.1978 x_1 - 0.1767 x_2 + 0.2186 x_3$$

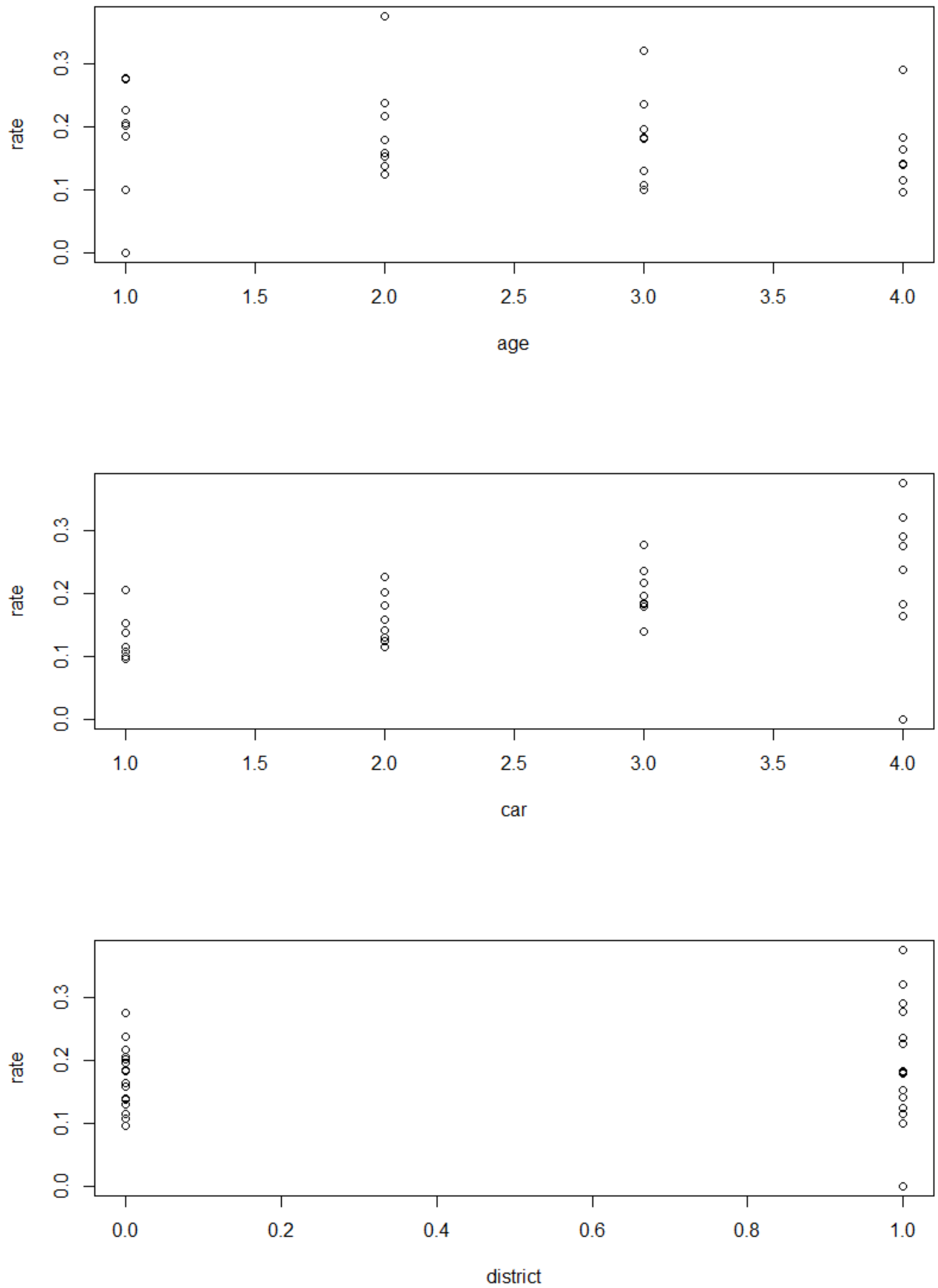

Figure 1.1: Scatter plots of rate by car, age, and district

car	age	district	y	n	p	fitted value of model_2
1	1	0	65	317	0.2050	50.7747
1	2	0	65	476	0.1366	63.8908
1	3	0	52	486	0.1070	54.6651
1	4	0	310	3259	0.0951	307.1859
2	1	0	98	486	0.2016	94.8669
2	2	0	159	1004	0.1584	164.2310
2	3	0	175	1355	0.1292	185.7391
2	4	0	877	7660	0.1145	879.9048
3	1	0	41	223	0.1839	53.0485
3	2	0	117	539	0.2171	107.4484
3	3	0	137	697	0.1966	116.4359
3	4	0	477	3442	0.1386	481.8455
4	1	0	11	40	0.2750	11.5963
4	2	0	35	148	0.2365	35.9553
4	3	0	39	214	0.1822	43.5670
4	4	0	167	1019	0.1639	173.8446
1	1	1	2	20	0.1000	3.9864
1	2	1	5	33	0.1515	5.5119
1	3	1	4	40	0.1000	5.5987
1	4	1	36	316	0.1139	37.0647
2	1	1	7	31	0.2258	7.5300
2	2	1	10	81	0.1235	16.4878
2	3	1	22	122	0.1803	20.8104
2	4	1	102	724	0.1409	103.4910
3	1	1	5	18	0.2778	5.3284
3	2	1	7	39	0.1795	9.6746
3	3	1	16	68	0.2353	14.1358
3	4	1	63	344	0.1831	59.9256
4	1	1	0	3	0.0000	1.0823
4	2	1	6	16	0.3750	4.8370
4	3	1	8	25	0.3200	6.3335
4	4	1	33	114	0.2895	24.2019

Table 1.1: The rate of claims for each category

Then we can calculate the goodness of fit statistics:

$$\chi^2 = \sum_{i=1}^{n} = \frac{(o-e)^2}{e} = 23.5$$

with degree of freedom 28, p-value 0.7078.

$$D = -2(l_0 - l_{max}) = 24.69$$

	Estimate Std.	Error	z value	Pr(>|z|)	
(Intercept)	-1.6083	0.1232	-13.05	<2e-16	***
district1	-0.127	0.3028	-0.42	0.67482	
car2	0.0169	0.1567	0.11	0.91429	
car3	-0.0477	0.1919	-0.25	0.80366	
car4	0.2209	0.326	0.68	0.49791	
age2	-0.3644	0.1719	-2.12	0.03401	*
age3	-0.643	0.1822	-3.53	0.00042	***
age4	-0.7406	0.1347	-5.5	3.80E-08	***
district1:car2	0.0817	0.1773	0.46	0.64475	
district1:car3	0.125	0.1899	0.66	0.51038	
district1:car4	0.4263	0.2216	1.92	0.05439	.
district1:age2	-0.0636	0.3397	-0.19	0.85143	
district1:age3	0.2663	0.3157	0.84	0.39891	
district1:age4	0.2709	0.2853	0.95	0.34242	
car2:age2	0.1041	0.2113	0.49	0.6224	
car3:age2	0.4854	0.243	2	0.0458	*
car4:age2	0.3398	0.3805	0.89	0.37187	
car2:age3	0.1996	0.2178	0.92	0.3594	
car3:age3	0.663	0.2473	2.68	0.00734	**
car4:age3	0.3275	0.3813	0.86	0.39047	
car2:age4	0.1628	0.1686	0.97	0.33421	
car3:age4	0.4214	0.2037	2.07	0.0386	*
car4:age4	0.3191	0.3377	0.95	0.34463	

Table 1.2: Estimated parameter value of poisson model(with interaction)

	Estimate Std.	Error	z value	Pr(>|z|)	
(Intercept)	-1.8525	0.0799	-23.18	<2e-16	***
car	0.1978	0.0208	9.51	<2e-16	***
age	-0.1767	0.0185	-9.56	<2e-16	***
district1	0.2186	0.0585	3.74	0.00019	***

Table 1.3: Estimated parameter value of poisson model(without interaction)

with degree of freedom 28, p-value 0.6449. Both are not significant, so we have evidence to say that the model fits data well. And this model is simplier than the model in (b), therefore this model in (c) is better.

2 PROBLEM 2

(a) Since the marginal row are fixed, the minimal model is:

$$log(\mu_{ij}) = \mu + \text{treatment} \tag{2.1}$$

Now we are interested in such two model:

$$log(\mu_{ij}) = \mu + \text{treatment} + \text{response} \tag{2.2}$$

$$log(\mu_{ij}) = \mu + \text{treatment} + \text{response} + \text{treatment} * \text{response} \tag{2.3}$$

To simplify the notation, let *treatment* = T, *response* = R. Then we can test:

$$H_0 : \lambda^{T*R} = 0 \qquad H_1 : \lambda^{T*R} \neq 0$$

Compare model (2.3) and model (2.2), we have

$$\Delta D = 18.6425 - 1.7764e - 15 = 18.6425$$

The df = $6-4=2$. The p-value = $8.95e-5$. It is significant, and we should refuse the non-hypothesis, which means the distribution of responses is different for the placebo and vaccine groups.

(b) Using the model (2.3), we can get the fitted values in Table 2.1.

treatment	**response**	**frequency**	**fitted values**	**deviance residual**	**pearson residual**
placebo	small	25	16.13699	2.040115	2.206329
placebo	moderate	8	13.53425	-1.62972	-1.50432
placebo	large	5	8.328767	-1.2469	-1.15343
vaccine	small	6	14.86301	-2.61546	-2.29894
vaccine	moderate	18	12.46575	1.468817	1.56747
vaccine	large	11	7.671233	1.127679	1.201852

Table 2.1: The fitted value, deviance and pearson deviance of model (2.1)

$$\chi^2 = 18.6425$$

$$\text{D} = 18.6425$$

From Table 2.1, we can see that the cell "vaccine-small" contribute most to the χ^2.

(c) We fit the proportional odds model as follow:

$$\text{logit}(P \geqslant j) = \alpha_j + \beta_1 X_1, \quad j = 1,2$$

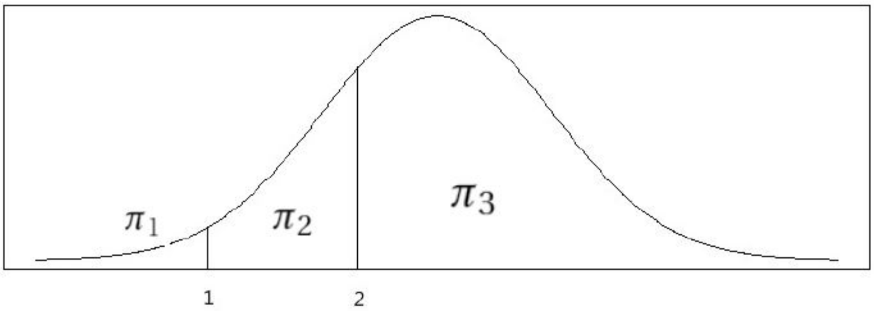

Figure 2.1: Diagram for placebo

where X_1 = treatment (vaccine = 1; placebo = 0). By *VGAM* in R, we can get the fitted model:

$$\text{logit}(P \geqslant 1) = -2.4408 + 1.8373X_1$$

$$\text{logit}(P \geqslant 2) = -0.5650 + 1.8373X_1$$

Then a location shift between the two treatment groups is

$$\Delta = -2.4408 - (-0.5650) = -1.8758$$

The fitted value is in Table 2:

	large	moderate	small
placebo	0.0801	0.2823	0.6376
vaccine	0.3536	0.4276	0.2189

Table 2.2: fitted value of the proportional odds model for flu vaccine trial data

For placebo, $\pi_1 = 0.0801$, $\pi_2 = 0.2823$, $\pi_3 = 0.6376$. For vaccine, $\pi_1 = 0.3536$, $\pi_2 = 0.4276$, $\pi_3 = 0.2189$.

The diagrams are in the Figure 2.1 and Figure 2.2.

3 Problem 3

(a) The plot is in Figure 3.1. The weight increase of group C slowed down from the second week, far less than group A and group B. And group A and group B have almost the same rate of weight increase and they are both showing linear growth.

(b) The normal linear model with different intercepts and different slopes for the three treatment groups is:

$$E(Y_{ijk}) = \alpha_i + \beta_i t_k + \epsilon_{ijk}$$

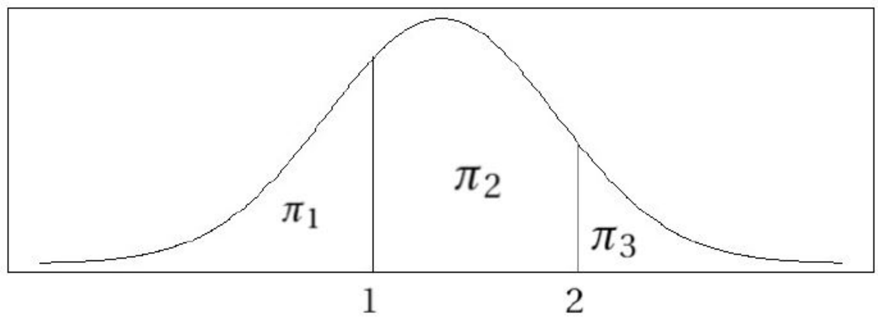

Figure 2.2: Diagram for vaccine

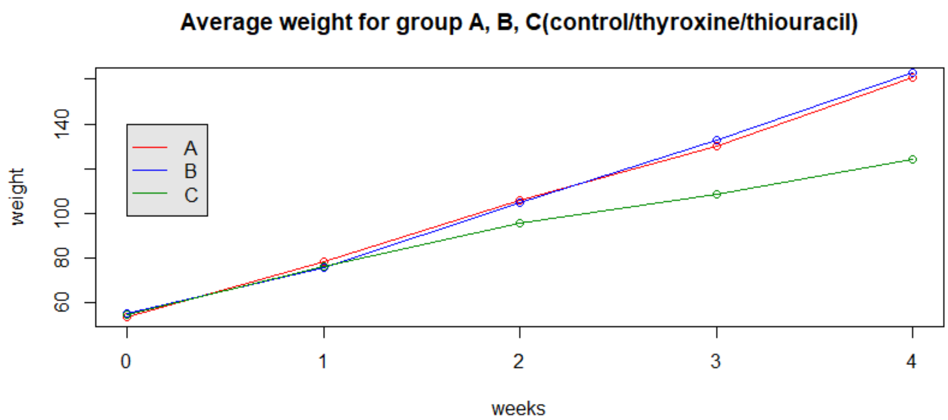

Figure 3.1: Average weights over weeks for different groups

where Y_{ijk} is the weights at time t_k $(k = 0, \cdots, 4)$ for patient j $(j = 1, \cdots, 27)$ in group i(where $i = 1$ for group A, $i = 2$ for group B and $i = 3$ for group C). The result is in Table 3.1. We can see from the result, we do not have evidence to say the intercept is

	Estimate Std.	Error	t value	Pr(>|t|)	
(Intercept) α_1	52.88	2.655	19.92	<2e-16	***
weeks β_1	26.48	1.084	24.43	<2e-16	***
treatthiouracil $\alpha_3 - \alpha_1$	4.78	3.754	1.27	0.21	
treatthyroxine $\alpha_2 - \alpha_1$	-0.794	4.137	-0.19	0.85	
weeks:treatthiouracil $\beta_3 - \beta_1$	-9.37	1.533	-6.11	1.1E-08	***
weeks:treatthyroxine $\beta_2 - \beta_1$	0.663	1.689	0.39	0.7	

Table 3.1: Results of naive analyses of weights

different. And it also shows that there is no differences between β_1 and β_2, whereas we have strong evidence to say β_3 is different from β_1.

(c) In stage I, we fit a linear regression for each subject:

$$y_{ij} = b_{i0} + b_{i1} t_{ij} + \epsilon_{ij}$$

then we can get the estimated result in Table 3.2.

In stage II, we fit a regression for three groups:

$$\hat{b_{i0}} = \alpha_1 + \alpha_2 x_{i2} + \alpha_3 x_{i3} + e_{i0}$$

$$\hat{b_{i1}} = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_{i0}$$

then the estimate result is in Table 3.3 and Table 3.4.

Change from group A to group B corresponds to -1.46 unit decrease in baseline weight. Change from group A to group C corresponds to 14.15 unit increase in baseline weight. The intercept of A and C is quite different.

Change from group A to group B corresponds to 0.663 more in weight change rate. Change from group A to group C corresponds to -9.37 less in weight change rate. The slope of A and C is quite different. The result is consistent with that shown in the Figure 3.1.

(d) By *Program_3*, we can conduct Welch Two Sample t-test. For group A and B, we have

$$t = -0.06, \quad df = 70, \quad p - value = 1$$

The test is not significant, i.e., true difference in means is equal to 0, which means the weight in A and B is the same.
For group A and C, we have

$$t = 2, \quad df = 80, \quad p - value = 0.04$$

The test is slightly significant, i.e., true difference in means is not equal to 0, which means the weights in A and B tend to be different.

	subject	group	b_0	b_1
1	1	A	57	28.3
2	2	A	62.4	28.7
3	3	A	47.2	33.3
4	4	A	43.4	29.2
5	5	A	56.6	23
6	6	A	45.4	27.5
7	7	A	49.6	21.9
8	8	A	65.8	22.1
9	9	A	46.2	22.7
10	10	A	55.2	28.1
11	11	B	57.4	30.5
12	12	B	51.2	20.7
13	13	B	47.4	34.2
14	14	B	57.2	29.9
15	15	B	53.6	22.2
16	16	B	51.8	21.9
17	17	B	46	30.6
18	18	C	67	17
19	19	C	63.2	15.7
20	20	C	56.8	18.7
21	21	C	66.2	14.9
22	22	C	52.8	22.6
23	23	C	55.2	16.1
24	24	C	62.2	12.9
25	25	C	53	21.1
26	26	C	46.2	15.1
27	27	C	54	17

Table 3.2: Estimates of intercepts and slopes for each subject

source	**df**	**mean square**	**F**	**p value**	
group	2	167	1.9	0.17	
residual	24	43.8			
parameter	**estimate**	**Std.error**	**t**	**p value**	
α_1	52.88	2.094	25.26	<2e-16	***
$\alpha_2 - \alpha_1$	-0.794	3.263	-0.24	0.81	
$\alpha_3 - \alpha_1$	4.780	2.961	1.61	0.12	**

Table 3.3: Analysis of variance of intercept estimates in Table 3.2

source	df	mean square	F	p value	
group	2	294	18.3	1.50E-05	***
residual	24	16			
parameter	**estimate**	**Std.error**	**t**	**p value**	
β_1	26.48	1.266	20.92	< 2e-16	***
$\beta_2 - \beta_1$	0.663	1.973	0.34	0.74	
$\beta_3 - \beta_1$	-9.37	1.79	-5.23	2.3e-05	***

Table 3.4: Analysis of variance of slope estimates in Table 3.2

(e) We use the following model to estimate the fixed effects:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_{C_2} g_B + \beta_{C_3} g_C + \beta_{G_2} g_B t_{ij} + \beta_{G_3} g_C t_{ij} + \epsilon_{ij}$$

The result is in Table 3.5.

	Estimate	Std.err	Wald	Pr(>|W|)	
(Intercept) β_0	53.91	2.36	524.03	<2e-16	***
treatthiouracil β_{C_3}	4.41	3.07	2.06	0.15	
treatthyroxine β_{C_2}	-1.38	2.9	0.22	0.64	
weeks β_1	25.63	1.26	413.63	<2e-16	***
treatthiouracil:weeks β_{G_3}	-9.38	1.59	34.62	4e-09	***
treatthyroxine:weeks β_{G_2}	1.33	2.43	0.3	0.58	

Table 3.5: Estimate results of GEE methods

With model-based standard errors, we can see the wald test result in Tbale 3.5. The intercept difference between group A and B, or A and C is the same.(Test not significant.) And the slope difference between group A and B is the same.(Test not significant.) However, the slope difference between group A and C is different.(Test is significant.)

(f) For random slope and random intercept, we use the following model to estimate the fixed effects:

$$y_{ij} = \alpha_1 + \beta_1 t_{ij} + \alpha_2 g_B + \alpha_3 g_C + \beta_2 g_B t_{ij} + \beta_3 g_C t_{ij} + a_{i0} + a_{i1} t_{ij} + b_{i0} + b_{i1} t_{ij} + \epsilon_{ij}$$

The results is in Table 3.6.

To compare the estimate result of these four methods(Naive, Two-satge, GEE, REML), we can summarize them to Table 3.7:

4 Problem 4

(a) Consider the logistic model as follow:

$$\log(\pi_{ij}) = \alpha + \beta_1 z_{i2} + \beta_2 z_{i3} + \beta_3 z_{i4}$$

	Value	Std.Error	DF	t-value	p-value
(Intercept) α_1	52.9	2.09	105	25.26	0
treatthiouracil $\alpha_3 - \alpha_1$	4.8	2.96	24	1.61	0.119
treatthyroxine $\alpha_2 - \alpha_1$	-0.8	3.26	24	-0.24	0.81
weeks β_1	26.5	1.27	105	20.92	0
treatthiouracil:weeks $\beta_3 - \beta_1$	-9.4	1.79	105	-5.23	0
treatthyroxine:weeks $\alpha_2 - \alpha_1$	0.7	1.97	105	0.34	0.738

Table 3.6: Estimate results of REML method

	Naive		Two-stage		GEE		REML	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
α_1	52.88	2.655	52.88	2.094	53.91	2.36	52.9	2.09
$\alpha_2 - \alpha_1$	-0.794	4.137	-0.794	3.263	-1.38	2.9	-0.8	2.96
$\alpha_3 - \alpha_1$	4.78	3.754	4.780	2.961	4.41	3.07	4.8	3.26
β_1	26.48	1.084	26.480	1.266	25.63	1.26	26.5	1.27
$\beta_2 - \beta_1$	0.663	1.689	0.663	1.973	1.33	2.43	0.7	1.79
$\beta_3 - \beta_1$	-9.37	1.533	-9.37	1.79	-9.38	1.59	-9.4	1.97

Table 3.7: Comparison of analyses of the ratdrink data using various different methods

Using *glm* in R, we can fit the model as follow:

$$\log(\pi_{ij}) = 1.144 - 3.323 z_{i2} - 4.476 z_{i3} - 4.130 z_{i4}$$

The goodness of fit statistics are:

$$\chi^2 = \sum_{i=1}^{n} = \frac{(o-e)^2}{e} = 154.7070$$

with degree of freedom 54, p-value $1.19e-11$.

$$D = -2(l_0 - l_{max}) = 173.4532$$

with degree of freedom 54, p-value $1.88e-14$. Both are significant, therefore, we consider the model here doesn't fit data well.

(b) Consider the following model:

$$\text{logit}(\pi_{ij}) = \alpha + \beta_1(\text{GRP} = 2) + \beta_2(\text{GRP} = 3) + \beta_3(\text{GRP}=4)$$

The estimate results are as follow in Table 4.1:
When using the exchangeable structure as a working correlation matrix, we can get the correlation coefficient $\rho = 0.185$.

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	1.21	0.225	5.4	0.27	4.49
GRP2	-3.37	0.566	-5.95	0.43	-7.83
GRP3	-4.58	1.309	-3.5	0.624	-7.35
GRP4	-4.25	0.853	-4.98	0.605	-7.02

Table 4.1: The estimated effects of treatment group based on GEE method

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.809	0.362	5	5.60E-07	***
as.factor(GRP)2	-4.54	0.735	-6.18	6.40E-10	***
as.factor(GRP)3	-5.883	1.175	-5.01	5.60E-07	***
as.factor(GRP)4	-5.606	0.908	-6.18	6.50E-10	***
	Variance σ^2	Std.Dev.			
random effects	2.28	1.51			

Table 4.2: The estimated effects of treatment group based on GLIMM method

(c) Consider the logistic model as follow:

$$\log(\pi_{ij}) = \alpha + u_i + \beta_1 z_{i2} + \beta_2 z_{i3} + \beta_3 z_{i4}$$

The estimate results are as follow in Table 4.2:

Using standard errors of fixed effects for inference, we can see from the Table 4.2, all estimates are significant.

(d) The results are summarized in the following Table 4.3:

	Binomial ML		GEE		GLMM	
	Estimate	SE	Estimate	SE	Estimate	SE
(Intercept)	1.144	0.129	1.21	0.270	1.809	0.362
GRP2	-3.323	0.331	-3.37	0.430	-4.540	0.735
GRP3	-4.476	0.731	-4.58	0.624	-5.883	1.175
GRP4	-4.130	0.476	-4.25	0.605	-5.606	0.908

Table 4.3: Summary of the results based on three methods

Compare the standard error of three models, standard error of GLMM is bigger than the ohter two. GEE considers the random effects, which can make the model more robust.