# bm2_hw8

*Siyan Chen*

*4/22/2019*

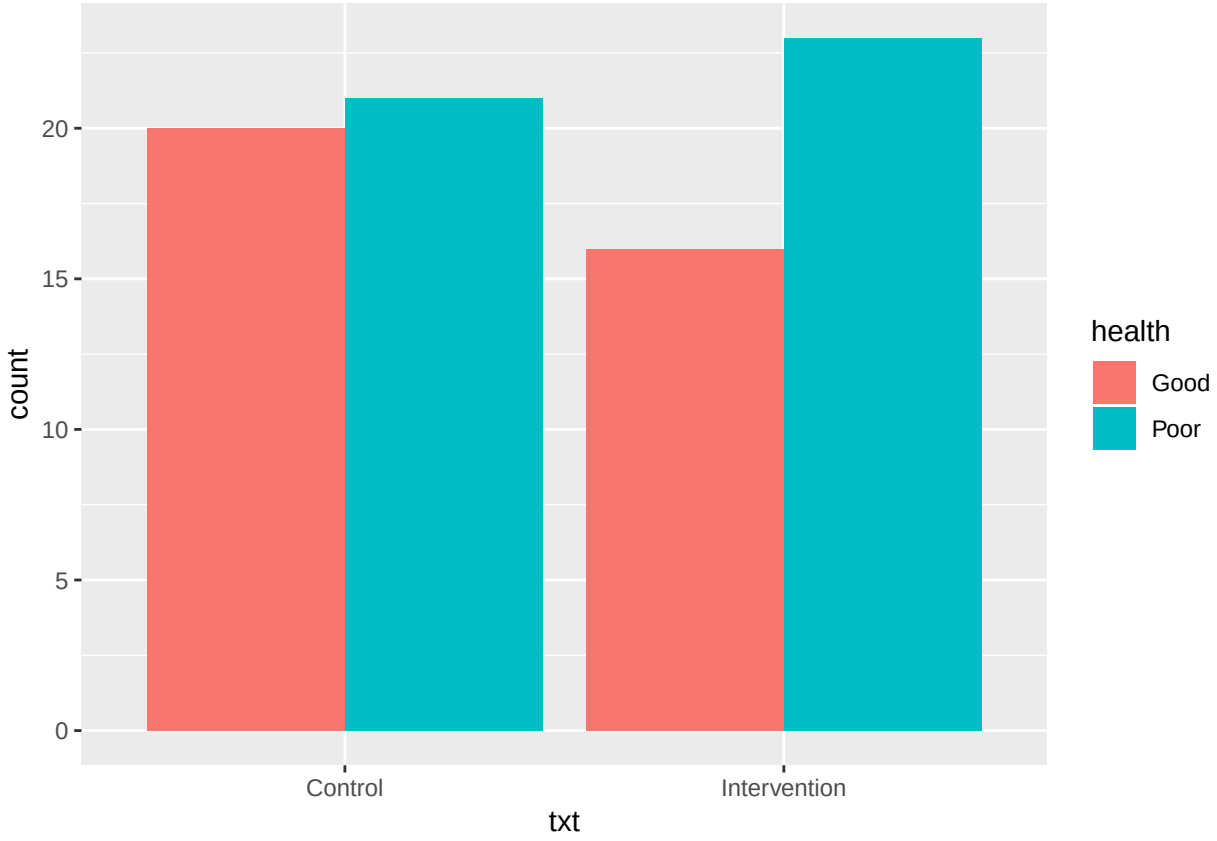
```
### data manipulation
df1 = df %>%
  filter(time == 1) %>%
  select(id, health)

names(df1) = c("id", "health_baseline")

resp1 = left_join(df, df1, by = "id") %>%
  mutate(agegroup = as.factor(agegroup),
         health = as.numeric(health == "Good"),
         health_baseline = as.factor(health_baseline)) %>%
  mutate(time = recode(time, "1" = "1","2" = "3", "3" = "6", "4" = "12")) %>%
  mutate(time = as.numeric(time))
resp = subset(resp1,time>1)
```

**(a)**

```
df %>%
  filter(time == 1) %>%
  ggplot(aes(x = txt, fill =  health)) + geom_bar(position = "dodge")
```




```
resp_r = resp1 %>%
  filter(time == 1)


glm1 = glm(health ~ txt, data = resp_r, family = "binomial")
summary(glm1)
```

```
##
## Call:
## glm(formula = health ~ txt, family = "binomial", data = resp_r)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.157  -1.157  -1.028   1.198   1.335
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.04879    0.31244  -0.156    0.876
## txtIntervention -0.31412    0.45122  -0.696    0.486
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.10  on 79  degrees of freedom
## Residual deviance: 109.62  on 78  degrees of freedom
## AIC: 113.62
##
## Number of Fisher Scoring iterations: 4
```

According to the plot, when subjects are assigned to intervention treatment group, participants self_rated level of health tend to be good. For control group, the proportion of self_rated level of health to be good are appromately same as that of self_rated level of health to be bad.

According to the model, p value of txt is not significant.Therefore, there is no significant relationship between randomized group and health self-rating.

**b)**

```
gee1 = gee(health ~ health_baseline + txt + agegroup + time, data = resp, family = "binomial", id = id,
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27

## running glm to get initial regression estimate

##         (Intercept) health_baselinePoor     txtIntervention
##          0.18528086         -1.71063852          1.99669985
##       agegroup25-34         agegroup35+                time
##          1.19749448          1.39742621          0.02536275
```

```
summary(gee1)
```

```
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
##  Link:                      Logit
##  Variance to Mean Relation: Binomial
##  Correlation Structure:     Unstructured
##
## Call:
## gee(formula = health ~ health_baseline + txt + agegroup + time,
##     id = id, data = resp, family = "binomial", corstr = "unstructured",
##     scale.fix = TRUE, scale.value = 1)
##
## Summary of Residuals:
##         Min          1Q      Median          3Q         Max
## -0.98144969 -0.18317233  0.08914345  0.17159228  0.83093959
##
##
## Coefficients:
##                         Estimate Naive S.E.    Naive z Robust S.E.
## (Intercept)          0.12457924 0.47137316  0.2642901  0.51374172
## health_baselinePoor -1.81418056 0.48958528 -3.7055456  0.50961334
## txtIntervention      2.10225898 0.48779381  4.3097286  0.53777951
## agegroup25-34        1.35250468 0.48130172  2.8100973  0.50420159
## agegroup35+          1.42052166 0.79781620  1.7805124  0.78372968
## time                 0.03243343 0.03665686  0.8847848  0.04755408
##                       Robust z
## (Intercept)          0.2424939
## health_baselinePoor -3.5599158
## txtIntervention      3.9091467
## agegroup25-34        2.6824681
## agegroup35+          1.8125148
```

```
## time                 0.6820326
##
## Estimated Scale Parameter:  1
## Number of Iterations:  5
##
## Working Correlation
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.1719328 0.5859907
## [2,] 0.1719328 1.0000000 0.2013998
## [3,] 0.5859907 0.2013998 1.0000000
```

Coefficient interpretation:
$\beta 0$: The log odds ratio of self_rating health status to be good is 0.125 on average for subpopulation in 15-24 age and control group with health_baseline to be good.
$\beta_{health_baselinePoor}$: The log odds rato of self_rating health status to be good is -1.81 on average for group of healt_baseline to be poor versus group of healt_baseline to be good adjusting for other variables.
$\beta_{txtIntervention}$ The log odds rato of self_rating health status to be good is 2.10 on average for intervention group versus control group adjusting for other variables.
$\beta_{agegroup25-34}$: The log odds rato of self_rating health status to be good is 1.35 on average for age group 25-34 versus age group 15-24 adjusting for other variables.
$\beta_{agegroup35+}$: The log odds rato of self_rating health status to be good is 1.42 on average for age group 35+ versus age group 15-24 adjusting for other variables.
$\beta_{time}$:The log odds rato of self_rating health status to be good is 0.032 on average for one unit change in time adjusting for other variables.

**c)**

```
glmm.fit = glmer(health ~ health_baseline + txt + agegroup + time + (1|id), family = "binomial", data =
summary(glmm.fit)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: health ~ health_baseline + txt + agegroup + time + (1 | id)
##    Data: resp
##
##      AIC      BIC   logLik deviance df.resid
##    185.0    208.0    -85.5    171.0      192
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.6112 -0.2327  0.1402  0.2982  1.8239
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 5.721    2.392
## Number of obs: 199, groups:  id, 78
##
## Fixed effects:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.19521    0.87019   0.224  0.82250
## health_baselinePoor -2.77610    0.98381  -2.822  0.00478 **
```

```
## txtIntervention      3.41325    1.07267   3.182  0.00146 **
## agegroup25-34        2.25651    1.00877   2.237  0.02529 *
## agegroup35+          1.98229    1.38118   1.435  0.15123
## time                 0.03718    0.06933   0.536  0.59176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) hlth_P txtInt a25-34 agg35+
## hlth_bslnPr -0.374
## txtIntrvntn -0.256 -0.449
## agegrp25-34 -0.319 -0.379  0.395
## agegroup35+ -0.195 -0.274  0.206  0.390
## time        -0.472 -0.016  0.047  0.007 -0.007
```

The GLMM model is $logit(E(Y_ij|b_i)) = (b_i + \beta_1) + X_{ij}^T\beta$

Coefficient interpretation:
$\beta_{health_baselinePoor}$: cannot interprete.
$\beta_{txtIntervention}$: cannot interprete
$\beta_{agegroup25-34}$ cannot interprete
$\beta_{agegroup35+}$ cannot interprete
$\beta_{time}$ The log odds ratio of self-rating of health to be good is 0.03718 for one unit change in time.

For GEE model, all coefficient can be explianed because we consider the subpopulation situation while for GLMM model, some coefficient cannot be explained because it is on individuals levels. For example, the treatment group is predetermined for specidic individuals, so the coefficient cannot be explianed.