BIOST HW1

Shuyang Wu

1/29/2017

Question 1

(a) The marginal distribution of quality of life:

good: 0.44 moderate: 0.3 poor: 0.26

Conditional probability of qualiy of life given "high" grade histology:

good: 0.118 moderate: 0.294 poor: 0.588

Conditional probability of qualiy of life given "low" grade histology:

good: 0.606 moderate: 0.303 poor: 0.091

The conditional distributions of quality of life are different. This makes sense, because intuitively patients with more severe cancer prognosis are more likely to have a poor quality of life (0.588) comparing to patients with low grade histology (0.091).

(b) Missing completely at random means that whether or not participants answered the questions regarding their quality of life is completely random and independent of other events (for example histology).

Missing at random means that the propensity for a data point to be missing is related to some of the observed data such as histology and not related to the quality of life answer itself.

Missing not at random means that the missing of entries for quality of life is dependent on other un-observed/un-measured variables such as patients' education level etc.

(c) Table 2: Missing not at random – all estimates are Biased The marginal Distribution of quality of life:

good: 0.605 moderate: 0.275 poor: 0.119

Conditional probability of qualiy of life given "high" grade histology:

good: 0.231 moderate: 0.385 poor: 0.385

Conditional probability of qualiy of life given "low" grade histology:

good: 0.723 moderate: 0.241 poor: 0.036

Table 3: Missing Completely at random – all estimates are Unbiased The marginal Distribution of quality of life:

good: 0.44 moderate: 0.3 poor: 0.26

Conditional probability of qualiy of life given "high" grade histology:

good: 0.118 moderate: 0.294 poor: 0.588

Conditional probability of qualiy of life given "low" grade histology:

good: 0.606 moderate: 0.303 poor: 0.091

Table 4: Missing at random The marginal Distribution of quality of life: – (biased)

good: 0.481 moderate: 0.301 poor: 0.218

Conditional probability of qualiy of life given "high" grade histology: – (unbiased)

good: 0.118 moderate: 0.294 poor: 0.588

Conditional probability of qualiy of life given “low” grade histology: –(unbiased)

good: 0.606 moderate: 0.303 poor: 0.091

(d)

```
#good
(60*132/(132-49) + 6*68/(68-42))/200
```

```
## [1] 0.55557
```

```
#moderate
(20*132/(132-49) + 10*68/(68-42))/200
```

```
## [1] 0.2898054
```

```
#poor
(3*132/(132-49) + 10*68/(68-42))/200
```

```
## [1] 0.1546247
```

Table 2 The marginal distribution of quality of life:

good: 0.555 moderate: 0.290 poor: 0.155

```
#good
(60*132/(132-33) + 6*68/(68-17))/200
```

```
## [1] 0.44
```

```
#moderate
(30*132/(132-33) + 15*68/(68-17))/200
```

```
## [1] 0.3
```

```
#poor
(9*132/(132-33) + 30*68/(68-17))/200
```

```
## [1] 0.26
```

Table 3: The marginal Distribution of quality of life:

good: 0.44 moderate: 0.3 poor: 0.26

```
#good
(60*132/(132-33) + 4*68/(68-34))/200
```

```
## [1] 0.44
```

```
#moderate
(30*132/(132-33) + 10*68/(68-34))/200
```

```
## [1] 0.3
```

```
#poor
(9*132/(132-33) + 20*68/(68-34))/200
```

```
## [1] 0.26
```

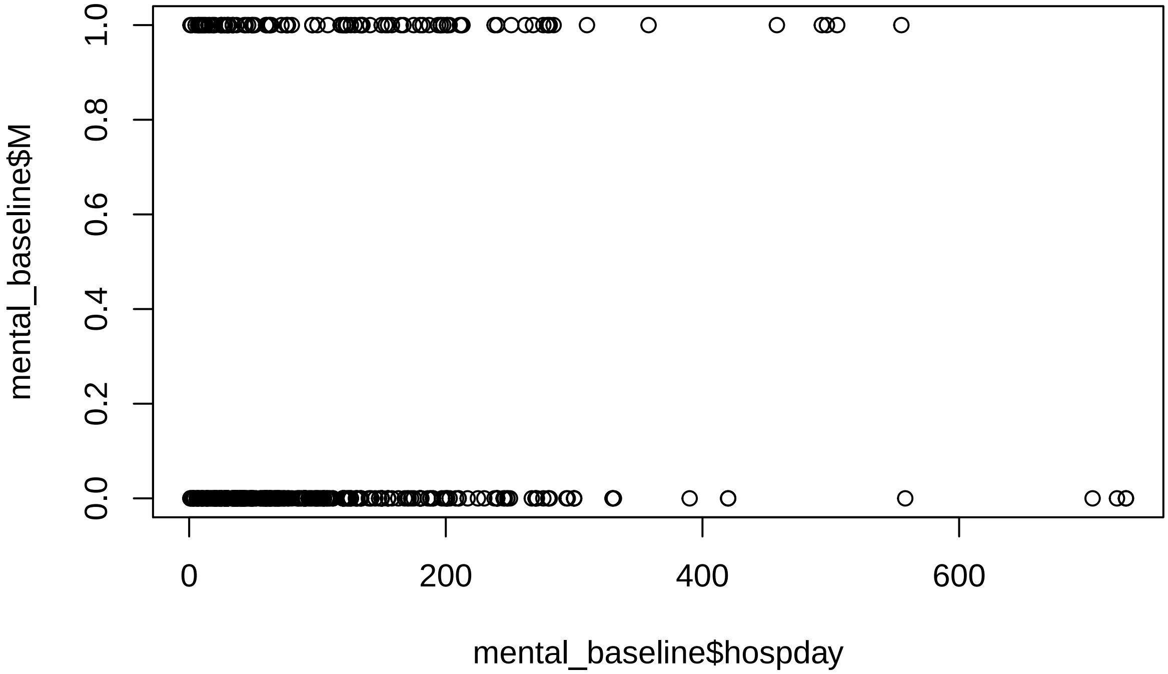
Table 3: The marginal Distribution of quality of life:

good: 0.44 moderate: 0.3 poor: 0.26

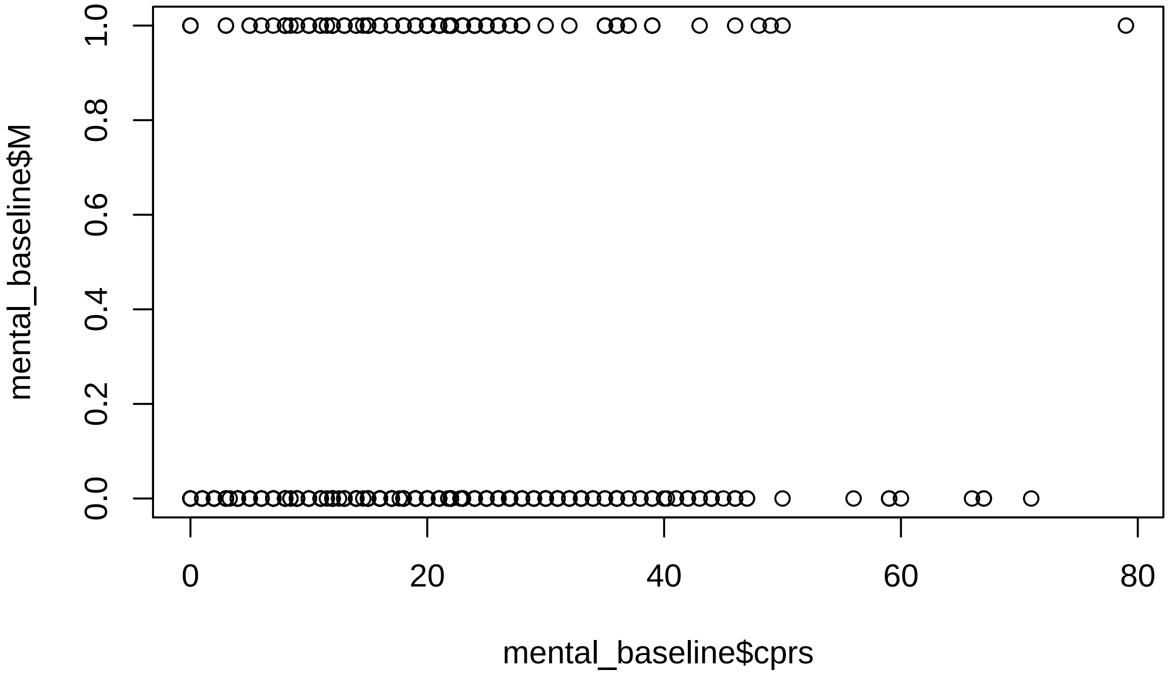
Question 2

```
mental_baseline <- read.csv("~/Downloads/mental_baseline.csv")
mental_baseline$M <- rep(0, 500)
mental_baseline$M[is.na(mental_baseline$sat)] <- 1

#look at missing pattern
plot(mental_baseline$hospday,mental_baseline$M)
```



```
plot(mental_baseline$cprs,mental_baseline$M)
```



```r
#check for dependencies and calculate propensity scores
propensity <- glm(M ~ cprs + hospday, family=binomial, data=mental_baseline)
summary(propensity)
```

```
##
## Call:
## glm(formula = M ~ cprs + hospday, family = binomial, data = mental_baseline)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2057  -0.6804  -0.6157  -0.5563   2.0143
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.9243390  0.2348155  -8.195  2.5e-16 ***
## cprs         0.0144267  0.0083485   1.728   0.0840 .
## hospday      0.0022943  0.0009143   2.509   0.0121 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 503.16  on 499  degrees of freedom
## Residual deviance: 494.15  on 497  degrees of freedom
## AIC: 500.15
##
## Number of Fisher Scoring iterations: 4
```

```r
ps <- propensity$fitted.values

#complete case analysis
cc <- na.omit(mental_baseline$sat)
mean(cc)
```

```
## [1] 19.1037
```

```r
#IPW
weighted.mean(mental_baseline$sat, 1/ps, na.rm = T)
```

```
## [1] 18.92381
```

```r
ipw.mean=function(data1,i){
  fit.ps=glm(M~ hospday + cprs,data=data1[i,],family=binomial)
  tmp.ps=fit.ps$fitted.value
  weighted.mean(data1$sat[i],1/tmp.ps,na.rm=T)

}

ipw.mean(mental_baseline, 1:nrow(mental_baseline))
```

```
## [1] 18.92381
```

```r
boot.out = boot(mental_baseline, ipw.mean, R = 100)

boot.ci(boot.out, index = 1, type = c("norm", "perc"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 100 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.out, type = c("norm", "perc"), index = 1)
##
## Intervals :
## Level      Normal              Percentile
## 95%   (18.49, 19.38 )   (18.46, 19.37 )
## Calculations and Intervals on Original Scale
## Some percentile intervals may be unstable
```

(a) Missing data pattern: Missing Completely at Random

(b) Based on the logistic regression model output, hospday has a little significant effect with on the missingness of sat, cprs has none. I think the missingness mechanism is missing completely at random.

(c) Complete case analysis result: 19.1037 IPW: 18.92381 95% confidence interval: (18.41, 19.37)

(d) The two estimate are similar because the missing pattern is Missing Completely at Random, meaning that the missingness is not dependent on any observed variables but itself. Therefore, using both inverse probability weighting and complete case analysis shouldn't make a difference on the results obtained.

Question 3

```r
#(a)
bc <- read.csv("~/Downloads/bc.csv")
propensity <- glm(treatment ~ tgr + age, family=binomial, data=bc)
summary(propensity)
```

```
##
## Call:
## glm(formula = treatment ~ tgr + age, family = binomial, data = bc)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3018  -1.0975   0.5953   0.8130   1.3301
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.685824   0.542856   8.632  < 2e-16 ***
## tgr         -0.466160   0.234627  -1.987   0.0469 *
## age         -0.058612   0.008742  -6.705 2.02e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 738.37  on 645  degrees of freedom
## Residual deviance: 682.16  on 643  degrees of freedom
## AIC: 688.16
##
## Number of Fisher Scoring iterations: 4
```

```r
#(b) treatment effect
#empirical mean difference
mean(bc$tgr) - mean(bc$tmass)
```

```
## [1] -13.0031
```

```
#IPW
ipw.ATE=function(data1,i){
  fit.ps=glm(treatment ~ tgr + age,data=data1[i,],family=binomial)
  tmp.ps=fit.ps$fitted.value
  mean(data1[i,2]*data1[i,1]/tmp.ps)-mean((1-data1[i,2])*data1[i,1]/(1-tmp.ps)) #mean difference (E(Y1)
}

ipw.ATE(bc,1:nrow(bc))
```

```
## [1] 0.04724655
```

```
#non-parametric calibration
Y<-bc[,1]
treat<-bc[,2]
X<-as.matrix(bc[,-c(1,2)])
fit1<- ATE(Y,treat,X)
fit1
```

```
## Call:
## ATE(Y = Y, Ti = treat, X = X)
##
## The analysis was completed for a simple study design with binary treatment.
##
## Point Estimates:
##     E[Y(1)]     E[Y(0)]         ATE
## 13.70481064 13.79887357 -0.09406293
```

```
summary(fit1)
```

```
## Call:
## ATE(Y = Y, Ti = treat, X = X)
##
##          Estimate   StdErr 95%.Lower 95%.Upper Z.value p.value
## E[Y(1)] 13.704811  0.181065 13.349931 14.059691 75.6902  <2e-16 ***
## E[Y(0)] 13.798874  0.227354 13.353269 14.244478 60.6935  <2e-16 ***
## ATE     -0.094063  0.223817 -0.532735  0.344609 -0.4203  0.6743
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(c) empirical mean difference: 13.0031 IPW: 0.04724655 non-parametric calibration: 0.09406 Because the treatment assignment is largely dependent on age, empirical mean difference does not take into account the dependency, its value is relatively larger than those from IPW and non-parametric calibration.

Question 4

(a) I have a dataset at hand, it includes head and neck cancer patients' clinical information, but some of the data were undetermined, for example, tumor stage, HPV status, and metastastic levels. The missingness for tumor stage is missing completely at random; Missing at random or possibily not at random for metastastic levels and HPV status.