Chapter 2

Sri Seshadri

9/24/2017

Repreoducing the example in page 39

Let's read in the data and prepare the data for analysis.

```
GSS96 <- read.csv(file = 'gss96.csv')
```

Relevant EDA plot

We find that the response variable is not continuous variable. But still pretend that it is... Why?? because the because the book does ... :).

```
# plot the response variable
hist(GSS96$ATTEND)
```

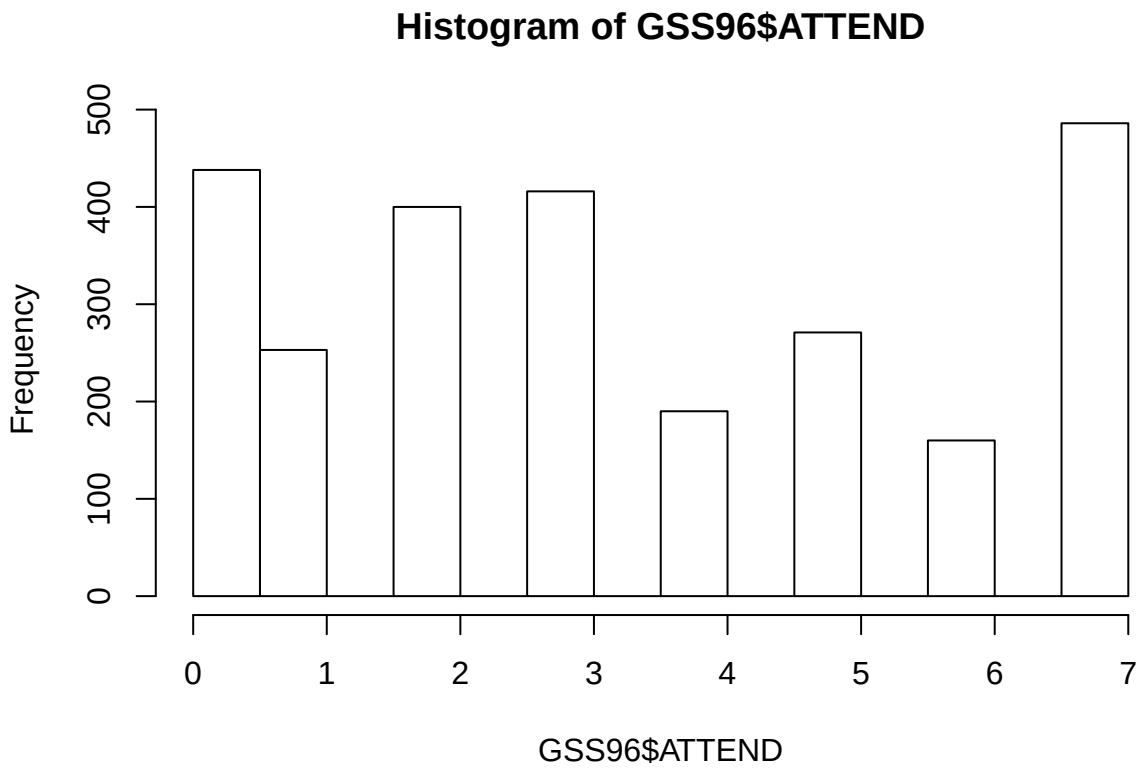

Figure 1: Histogram of Attendance

Also, we see the data has a bunch of missing values, wonder how did the book example dealt with it? See table below. The number of the observations in Example 2.1 in page 39 has 2614. This does not add up with the one in the data set provided in the course shell. Are we using the same data set?

```
library(mosaic)
sanitycheck <- do.call(rbind,dfapply(GSS96,favstats, select = is.numeric))
```

```r
# Round
sanitycheck <- round(sanitycheck,2)
knitr::kable(sanitycheck, caption = "Data sanity check for GSS96")
```

Table 1: Data sanity check for GSS96

	min	Q1	median	Q3	max	mean	sd	n	missing
ID	1.00	726.75	1452.50	2178.25	2904.00	1452.50	838.46	2904	0
MARITAL	1.00	1.00	2.00	4.00	5.00	2.44	1.63	2903	1
DIVORCE	1.00	2.00	2.00	2.00	2.00	1.77	0.42	1658	1246
CHILDS	0.00	0.00	2.00	3.00	8.00	1.83	1.68	2889	15
AGE	18.00	32.00	42.00	55.00	89.00	44.78	16.87	2898	6
INCOME	1.00	9.00	11.00	12.00	12.00	9.86	2.99	1947	957
POLVIEWS	1.00	3.00	4.00	5.00	7.00	4.19	1.36	2743	161
FUND	1.00	1.00	2.00	3.00	3.00	1.97	0.79	2732	172
ATTEND	0.00	1.00	3.00	5.00	7.00	3.36	2.44	2614	290
SPANKING	1.00	1.00	2.00	3.00	4.00	2.09	0.89	1923	981
TOTRELIG	2.00	100.00	300.00	1000.00	99996.00	1249.69	5119.16	527	2377
SEI	17.10	32.30	38.90	63.50	97.20	47.85	18.99	2781	123
PASEI	17.10	32.30	38.45	63.50	97.20	47.56	18.61	2326	578
VOLTEER	0.00	0.00	0.00	0.00	9.00	0.33	0.89	2904	0
GENDER	0.00	0.00	1.00	1.00	1.00	0.56	0.50	2904	0
RACE	0.00	0.00	0.00	0.00	1.00	0.19	0.39	2904	0
PRAYER	1.00	4.00	4.00	5.00	6.00	4.24	1.09	2904	0
EDUCATE	0.00	12.00	13.00	16.00	20.00	13.36	2.93	2895	9
VOLRELIG	0.00	0.00	0.00	0.00	1.00	0.07	0.26	2904	0
ZEDUCATE	-4.56	-0.47	-0.12	0.90	2.27	0.00	1.00	2895	9
ZAGE	-1.59	-0.76	-0.16	0.61	2.62	0.00	1.00	2898	6
POLVIEW1	1.00	1.00	2.00	3.00	3.00	2.11	0.78	2743	161

Data manipulation.

Well, I'll remove the missing values from the data... I'll use the complete cases for undertanding. Still feeling uneasy what the book and the data I picked doesn't agree.

```r
# Get rid of ID column.
GSS96 <- GSS96[complete.cases(GSS96),-1]

# Get the predictor matrix created
predictors <- model.matrix(ATTEND ~ ., GSS96)[,-1]
ATTENDANCE <- GSS96$ATTEND

# The output doesn't look good when I have the predictors and the response as matrices
GSS96 <- data.frame(cbind(predictors,ATTENDANCE))
```

Modeling

Attempting to prove to myself that GLM with gaussian distribution and identity link assumption, is same as ther OLS and try interpret the output of the GLM model. Compare the Residuals deviance (GLM) and residual Sum of squares (OLS).

```r
# Let do the glm model with link function being identity

glm.identity <- glm(ATTENDANCE ~ ., family = gaussian(), data = GSS96)

summary(glm.identity)
```

```
##
## Call:
## glm(formula = ATTENDANCE ~ ., family = gaussian(), data = GSS96)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.32806  -0.67072  -0.05302   0.78566   1.88522
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.142e+01  1.672e+00  -6.828 1.02e-09 ***
## MARITAL     -5.400e-01  4.757e-01  -1.135 0.259416
## DIVORCE      3.728e-01  2.681e-01   1.391 0.167806
## CHILDS      -5.080e-02  9.324e-02  -0.545 0.587257
## AGE          8.631e-03  1.058e-02   0.816 0.416882
## INCOME       3.767e-02  5.112e-02   0.737 0.463065
## POLVIEWS     6.967e-02  2.563e-01   0.272 0.786338
## FUND         7.730e-01  1.779e-01   4.346 3.67e-05 ***
## SPANKING     5.088e-01  1.336e-01   3.808 0.000257 ***
## TOTRELIG     3.558e-06  6.197e-05   0.057 0.954350
## SEI         -5.424e-03  7.502e-03  -0.723 0.471561
## PASEI       -3.359e-03  6.523e-03  -0.515 0.607836
## VOLTEER     -4.783e-02  1.065e-01  -0.449 0.654575
## GENDER       6.343e-02  2.438e-01   0.260 0.795370
## RACE        -9.471e-01  5.035e-01  -1.881 0.063230 .
## PRAYER       2.872e+00  2.146e-01  13.385  < 2e-16 ***
## EDUCATE     -3.390e-03  5.802e-02  -0.058 0.953536
## VOLRELIG     5.820e-01  3.233e-01   1.800 0.075195 .
## ZEDUCATE            NA         NA      NA       NA
## ZAGE                NA         NA      NA       NA
## POLVIEW1    -4.507e-02  4.129e-01  -0.109 0.913313
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.108044)
##
##     Null deviance: 475.880  on 107  degrees of freedom
## Residual deviance:  98.616  on  89  degrees of freedom
## AIC: 336.67
##
## Number of Fisher Scoring iterations: 2
```

```r
anova(glm.identity)
```

```
## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
```

```
## Response: ATTENDANCE
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev
## NULL                        107     475.88
## MARITAL    1    2.807       106     473.07
## DIVORCE    1    4.197       105     468.88
## CHILDS     1    9.272       104     459.60
## AGE        1    0.032       103     459.57
## INCOME     1    2.721       102     456.85
## POLVIEWS   1   11.847       101     445.00
## FUND       1    7.181       100     437.82
## SPANKING   1    4.897        99     432.92
## TOTRELIG   1   62.278        98     370.65
## SEI        1    0.615        97     370.03
## PASEI      1   17.459        96     352.57
## VOLTEER    1   20.772        95     331.80
## GENDER     1   11.017        94     320.78
## RACE       1    2.745        93     318.04
## PRAYER     1  215.825        92     102.21
## EDUCATE    1    0.000        91     102.21
## VOLRELIG   1    3.583        90      98.63
## ZEDUCATE   0    0.000        90      98.63
## ZAGE       0    0.000        90      98.63
## POLVIEW1   1    0.013        89      98.62
```

```
OLS <- lm(data = GSS96, formula = ATTENDANCE ~ .)
summary(OLS)
```

```
##
## Call:
## lm(formula = ATTENDANCE ~ ., data = GSS96)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32806 -0.67072 -0.05302  0.78566  1.88522
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.142e+01  1.672e+00  -6.828 1.02e-09 ***
## MARITAL     -5.400e-01  4.757e-01  -1.135 0.259416
## DIVORCE      3.728e-01  2.681e-01   1.391 0.167806
## CHILDS      -5.080e-02  9.324e-02  -0.545 0.587257
## AGE          8.631e-03  1.058e-02   0.816 0.416882
## INCOME       3.767e-02  5.112e-02   0.737 0.463065
## POLVIEWS     6.967e-02  2.563e-01   0.272 0.786338
## FUND         7.730e-01  1.779e-01   4.346 3.67e-05 ***
## SPANKING     5.088e-01  1.336e-01   3.808 0.000257 ***
## TOTRELIG     3.558e-06  6.197e-05   0.057 0.954350
## SEI         -5.424e-03  7.502e-03  -0.723 0.471561
## PASEI       -3.359e-03  6.523e-03  -0.515 0.607836
## VOLTEER     -4.783e-02  1.065e-01  -0.449 0.654575
## GENDER       6.343e-02  2.438e-01   0.260 0.795370
```

```
## RACE        -9.471e-01  5.035e-01  -1.881 0.063230 .
## PRAYER       2.872e+00  2.146e-01  13.385  < 2e-16 ***
## EDUCATE     -3.390e-03  5.802e-02  -0.058 0.953536
## VOLRELIG     5.820e-01  3.233e-01   1.800 0.075195 .
## ZEDUCATE            NA         NA      NA       NA
## ZAGE                NA         NA      NA       NA
## POLVIEW1    -4.507e-02  4.129e-01  -0.109 0.913313
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.053 on 89 degrees of freedom
## Multiple R-squared:  0.7928, Adjusted R-squared:  0.7509
## F-statistic: 18.92 on 18 and 89 DF,  p-value: < 2.2e-16
```

```
anova(OLS)
```

```
## Analysis of Variance Table
##
## Response: ATTENDANCE
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## MARITAL    1   2.807   2.807   2.5337 0.1149817
## DIVORCE    1   4.197   4.197   3.7873 0.0547971 .
## CHILDS     1   9.272   9.272   8.3682 0.0047998 **
## AGE        1   0.032   0.032   0.0290 0.8652367
## INCOME     1   2.721   2.721   2.4559 0.1206320
## POLVIEWS   1  11.847  11.847  10.6919 0.0015311 **
## FUND       1   7.181   7.181   6.4810 0.0126239 *
## SPANKING   1   4.897   4.897   4.4196 0.0383519 *
## TOTRELIG   1  62.278  62.278  56.2054 4.618e-11 ***
## SEI        1   0.615   0.615   0.5552 0.4581690
## PASEI      1  17.459  17.459  15.7570 0.0001457 ***
## VOLTEER    1  20.772  20.772  18.7465 3.903e-05 ***
## GENDER     1  11.017  11.017   9.9426 0.0022020 **
## RACE       1   2.745   2.745   2.4776 0.1190281
## PRAYER     1 215.825 215.825 194.7803 < 2.2e-16 ***
## EDUCATE    1   0.000   0.000   0.0001 0.9909210
## VOLRELIG   1   3.583   3.583   3.2340 0.0755155 .
## POLVIEW1   1   0.013   0.013   0.0119 0.9133133
## Residuals 89  98.616   1.108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let's use gender, race and educate as the predictors

```
model1 <- glm(ATTENDANCE ~ GENDER + RACE + EDUCATE, data = GSS96, family = gaussian())
summary(model1)
```

```
##
## Call:
## glm(formula = ATTENDANCE ~ GENDER + RACE + EDUCATE, family = gaussian(),
##     data = GSS96)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -5.2991  -1.5843   0.4183   1.7215   2.4441
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.659057   1.281452   3.636 0.000433 ***
## GENDER       0.701917   0.418440   1.677 0.096456 .
## RACE         0.539406   0.826593   0.653 0.515476
## EDUCATE     -0.005156   0.083764  -0.062 0.951035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 4.431455)
##
##     Null deviance: 475.88  on 107  degrees of freedom
## Residual deviance: 460.87  on 104  degrees of freedom
## AIC: 473.2
##
## Number of Fisher Scoring iterations: 2
```

```
paste('BIC is :', BIC(model1))
```

```
## [1] "BIC is : 486.608040653577"
```

Adding prayer as predictor to model1

```
model2 <- glm(ATTENDANCE ~ GENDER + RACE + EDUCATE + PRAYER, data = GSS96, family = gaussian())
summary(model2)
```

```
##
## Call:
## glm(formula = ATTENDANCE ~ GENDER + RACE + EDUCATE + PRAYER,
##     family = gaussian(), data = GSS96)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3273  -0.8733   0.1119   1.0387   2.5853
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.47209    1.16153  -6.433    4e-09 ***
## GENDER       0.02958    0.25342   0.117   0.9073
## RACE        -0.85873    0.50160  -1.712   0.0899 .
## EDUCATE      0.05848    0.05000   1.170   0.2449
## PRAYER       2.51703    0.18195  13.834   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.565619)
##
##     Null deviance: 475.88  on 107  degrees of freedom
## Residual deviance: 161.26  on 103  degrees of freedom
## AIC: 361.79
##
## Number of Fisher Scoring iterations: 2
```

```r
paste('BIC is :', BIC(model2))
```

```
## [1] "BIC is : 377.87843704311"
```

Finally the intercept only mode

```r
model3 <- glm(ATTENDANCE ~ 1, data = GSS96, family = gaussian())
summary(model3)
```

```
##
## Call:
## glm(formula = ATTENDANCE ~ 1, family = gaussian(), data = GSS96)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.8981  -1.8981   0.1019   2.1019   2.1019
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.8981     0.2029   24.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 4.447473)
##
##     Null deviance: 475.88  on 107  degrees of freedom
## Residual deviance: 475.88  on 107  degrees of freedom
## AIC: 470.66
##
## Number of Fisher Scoring iterations: 2
```