

Homework 2

Marat Habibullin

26.10.2014

Task 1

```
library(lattice)
library(MASS)

# creating dataframe
df <- read.table("input-data/teengambling.txt")
df$sex <- as.factor(df$sex)
contrasts(df$sex) <- contr.treatment
contrasts(df$sex)

##      2
## 0 0
## 1 1

# creating model
l <- lm(gamble ~ . ^2, data = df)
laic <- stepAIC(l, trace = 0)
summary(laic)

##
## Call:
## lm(formula = gamble ~ sex + status + income + sex:income + status:income,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -46.499  -6.933  -1.829   4.650  76.906
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -29.8693    18.6280  -1.603  0.116510
## sex2          15.1803    11.2698   1.347  0.185384
## status         0.5899     0.3324   1.774  0.083410 .
## income        13.8884     2.6497   5.241 5.14e-06 ***
## sex2:income   -9.1007     2.1237  -4.285 0.000108 ***
## status:income -0.1698     0.0547  -3.104 0.003451 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.1 on 41 degrees of freedom
## Multiple R-squared:  0.6728, Adjusted R-squared:  0.6329
## F-statistic: 16.86 on 5 and 41 DF,  p-value: 4.936e-09
```

First of all, we have removed the least significant predictors using stepAIC to make our model simpler and, as a result, one can see the Verbal predictor is not significant in our model.

Let's look at the predictor's coefficients. According to encoding we use for Sex qualitative predictor (1=female, 0=male) the models are:

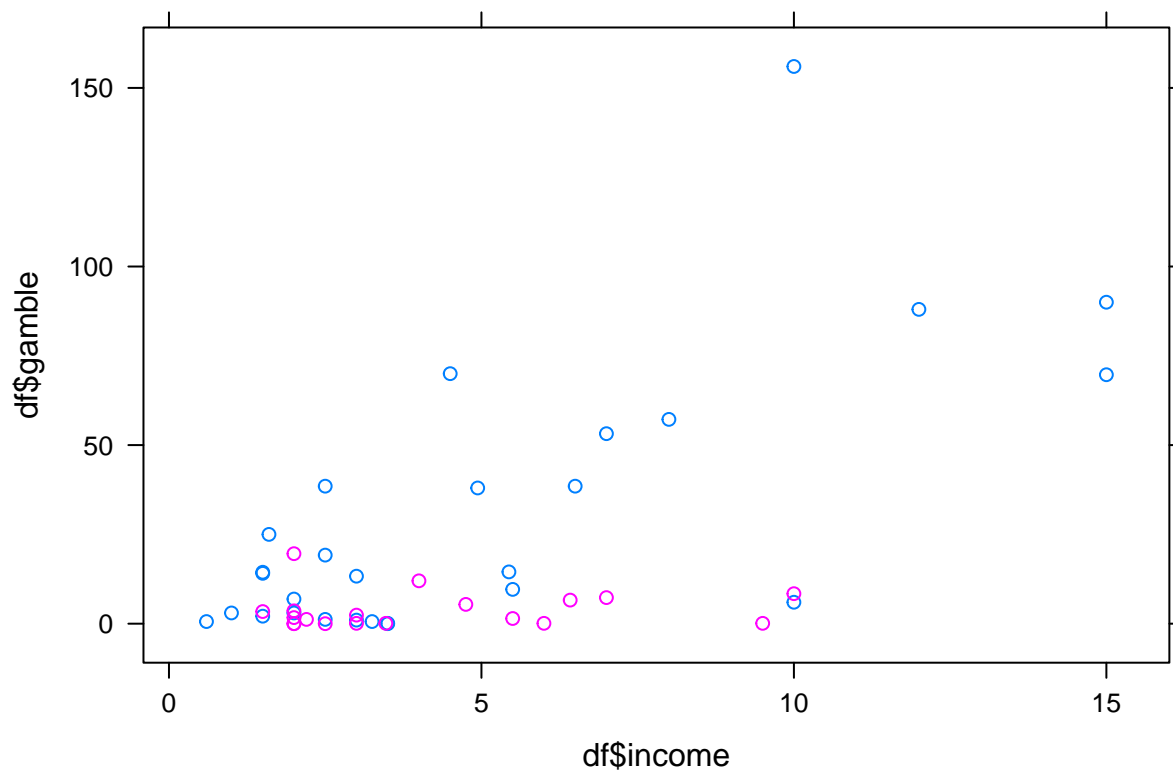
```
gamble = -29.87 + 15.18 + 0.59*status + 13.9*income - 9.1*income - 0.17*status*income =
        = -14.69 + 0.59*status + 4.8*income - 0.17*status*income -- female teens

gamble = -29.87 + 0.59*status + 13.9*income - 0.17*status*income -- male teens
```

The coefficient of Income predictor is the highest among others making this predictor one most affecting the response.

One can notice the models differ in Intercept and Income coefficient. We can give the following interpretation of this fact: having the same income male teens are more inclined to spent money by gambling than female teens (mostly because males' income coefficient is higher). To prove this assumption let's look at the statplot of Gamble vs Income (blue - male, pink - female):

```
xyplot(df$gamble ~ df$income, groups=df$sex)
```



As we can see for almost every fixed value of Income we have higher value of Gambling for male teens than female ones.

Also speaking about Income*Status predictor's coefficient (-0.17) one can say the synergy of this two predictors is detected to reduce the value of gambling. Maybe it is because teens from socioeconomic well-being families better understand risks of gambling even if they earn much money.