Statistical Modeling with R - Fall 2016
Homework 3

DUE IN: Tuesday, 04.10.2016 at 11:59 a.m.,

HOW: electronically in pdf-format via submission to `www.turnitin.com`

Class id: 13494794

enrollment password: Ti20Ta16Nic

Please register for the class on turnitin ahead of time.

GROUP WORK: is allowed with a maximum of 3 persons per group. PLEASE stay within the same group throughout the semester. Only one solution is accepted and graded per group. Please include the names of all group members on each assignment.

HOW MANY: There will be a total of six homework assignments in this semester. We will do a random selection of questions to be graded. Each week a total of eight points can be gained. Only the five best homeworks will be counted.

DUE DATES: 20.09., 27.09., 04.10., 11.10., 18.10., 25.10. (tentatively, subject to change)

FORMAT: Please do the required analyses and provide answers in complete sentences. **Provide the R syntax for the commands.** Just report those statistics that are relevant; do not copy complete R output. Integrate requested figures or tables into your document and give a brief verbal comment/caption on them.

Wells in Bangladesh

Many of the wells used for drinking water in South Asia are contaminated with natural arsenic. Exposure to arsenic, a cumulative poison, increases the risk of cancer and other diseases. The assoiciated risks are estimated to be proportional to the exposure. This assignment aims at modeling the decisions of households in Bangladesh to change their source of drinking water by switching the well they have been using.

Data `data(Wells, package="effects")`

Source A. Gelman and J. Hill (2007) Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press.

Variables Description of variables:

switch whether or not the household switched to another well from an unsafe well: `no` or `yes`

arsenic the level of arsenic contamination in the household's original well, in hundreds of micrograms per liter; all are above 0.5, which was the level identified as "safe".

distance in meters to the closest known safe well.

education in years of the head of the household.

associations whether or not any members of the household participated in any community organizations: `no` or `yes`.

```
library(car)
library(MASS)
data(Wells, package="effects")
```

1. Create a logistic regression model using `switch` as dependent variable and no predictor, just the intercept.

   (a) (half a point) How large is the AIC score for this model?

```
wells.1 <- glm(switch ~ 1 ,
  family=binomial(logit), data=Wells)
summary(wells.1)

##
## Call:
## glm(formula = switch ~ 1, family = binomial(logit), data = Wells)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.308  -1.308   1.052   1.052   1.052
##
## Coefficients:
##             Estimate Std. Error z value            Pr(>|z|)
## (Intercept)  0.30296    0.03681    8.23 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 4118.1  on 3019  degrees of freedom
## AIC: 4120.1
##
## Number of Fisher Scoring iterations: 4

AIC(wells.1)

## [1] 4120.099
```

   The AIC for the naive model is 4120.0992171.

   (b) (half a point) How large is the BIC score for this model?

```
BIC(wells.1)

## [1] 4126.112
```

The BIC for the naive model is 4126.1122292.

(c) (1 point) Compute the log odds for the mean response and compare it to the coefficient estimate for the intercept in this model [hint: Be careful about the coding of the variable `switch` in the data set].

```
logodds <- log(mean(as.numeric(Wells$switch)-1)/(2-mean(as.numeric(Wells$switch))))
```

The two numbers are the same, i.e. $0.3029584 = 0.3029584$.

2. Add `distance` as a predictor to the model.

```
wells.dist <- update(wells.1, .~. + distance)
summary(wells.dist)

##
## Call:
## glm(formula = switch ~ distance, family = binomial(logit), data = Wells)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4406  -1.3058   0.9669   1.0308   1.6603
##
## Coefficients:
##               Estimate Std. Error z value             Pr(>|z|)
## (Intercept)  0.6059594  0.0603102  10.047 < 0.0000000000000002 ***
## distance    -0.0062188  0.0009743  -6.383       0.000000000174 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 4076.2  on 3018  degrees of freedom
## AIC: 4080.2
##
## Number of Fisher Scoring iterations: 4

AIC(wells.dist)

## [1] 4080.238
```

(a) (half a point) In comparison to the naive model, by how much has the AIC changed?
The AIC has changed by $4120.0992171 - 4080.2378243 = 39.8613928$.

(b) (half a point) Is distance a significant predictor in this model for modeling the probability of switching to a safe well?

```
Anova(wells.dist)

## Analysis of Deviance Table (Type II tests)
##
## Response: switch
##          LR Chisq Df       Pr(>Chisq)
## distance   41.861  1 0.00000000009798 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes, distance is a significant predictor.

(c) (half a point) With growing distance to a safe well are families less likely or are they more likely to switch?

Since the coefficient is negative, they are less likely to switch.

(d) (half a point) According to the second model, what is the estimated probability of switching for a household that is next to an existing safe well (i.e. distance equals 0)?

The estimated probability at distance equals 0 is given by the inverse of the logit function evaluated at the intercept, i.e. $\text{logit}^{-1}(0.606) = \frac{e^{0.606}}{1+e^{0.606}}$. For a household that is next to an existing safe well the estimated probability of switching amounts to 0.647.

3. Using the second model with `distance` as predictor,

(a) (1 point) compute the halfway point, i.e the distance at which the estimated probability of switching equals 0.5

The halfway point has the property $p = 0.5$, i.e. $\log(\frac{p}{1-p}) = \log(1) = 0$. Hence the estimated probability of switching becomes 0.5, when the linear predictor becomes 0. Thus,

$$\begin{aligned} p &= 0.5 \\ \Leftrightarrow 0.606 - 0.0062 \cdot \text{distance} &= 0 \\ \Leftrightarrow \text{distance} &= \frac{0.606}{0.0062} \\ \Leftrightarrow \text{distance} &= 97.4396. \end{aligned}$$

The halfway point is at a distance of 97.44 meters between household and safe well.

(b) (1 point) Compute the slope of the tangent to the regression curve at the halfway point.

The slope of the tangent to the logistic regression curve equals $\beta \cdot p \cdot (1-p)$. Hence, for the halfway point this means $\frac{\beta}{4}$. This yields a slope of -0.0016.

4. Create a logistic regression model using `switch` as dependent variable and all available predictors.

```
wells.main <- glm(switch ~ education + association + distance + arsenic,
  family=binomial(logit), data=Wells)
summary(wells.main)
```

```
##
## Call:
## glm(formula = switch ~ education + association + distance + arsenic,
##     family = binomial(logit), data = Wells)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5942  -1.1976   0.7541   1.0632   1.6739
##
## Coefficients:
##                 Estimate Std. Error z value             Pr(>|z|)
## (Intercept)    -0.156712   0.099601  -1.573                0.116
## education       0.042447   0.009588   4.427          0.00000955 ***
## associationyes -0.124300   0.076966  -1.615                0.106
## distance       -0.008961   0.001046  -8.569 < 0.0000000000000002 ***
## arsenic         0.467022   0.041602  11.226 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 3907.8  on 3015  degrees of freedom
## AIC: 3917.8
##
## Number of Fisher Scoring iterations: 4
```

(a) (1 point) Which predictors are significant? Do the estimated coeffciients make common sense to you?

All predictors except `associations` are significant at the 5% level. The signs of the coefficients make common sense (distance is negative, arsenic level positive, education positive).

(b) (half a point) Starting with this model use the automatic backward/forward selection method to derive a suitable model. Report the significant predictors and the AIC score of the model.

```
wells.best <- stepAIC(wells.main, scope=list(upper=wells.main,lower=~1),direction='bo
```

```
## Start:  AIC=3917.83
## switch ~ education + association + distance + arsenic
##
##               Df Deviance    AIC
```

```
## <none>              3907.8 3917.8
## - association  1    3910.4 3918.4
## - education    1    3927.7 3935.7
## - distance     1    3985.2 3993.2
## - arsenic      1    4056.1 4064.1
```

```
summary(wells.best)
```

```
##
## Call:
## glm(formula = switch ~ education + association + distance + arsenic,
##     family = binomial(logit), data = Wells)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5942  -1.1976   0.7541   1.0632   1.6739
##
## Coefficients:
##                 Estimate Std. Error z value             Pr(>|z|)
## (Intercept)    -0.156712   0.099601  -1.573                0.116
## education       0.042447   0.009588   4.427          0.00000955 ***
## associationyes -0.124300   0.076966  -1.615                0.106
## distance       -0.008961   0.001046  -8.569 < 0.0000000000000002 ***
## arsenic         0.467022   0.041602  11.226 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 3907.8  on 3015  degrees of freedom
## AIC: 3917.8
##
## Number of Fisher Scoring iterations: 4
```

```
AIC(wells.best)
```

```
## [1] 3917.826
```

(c) (half a point) Draw a components-residual plot and assess whether any quadratic effects should be added.

```
crPlots(wells.best)
```

The component/residual plots arranged in Fig. **??** reveal an indication for a (negative) quadratic effect of the variable *arsenic*. This quadratic effect also makes common sense.

5. (2 points) Using, the best model derived in Question 4b calculate two models: one which also includes a quadratic effect of `arsenic`, and one that uses a logarithmic transformation of

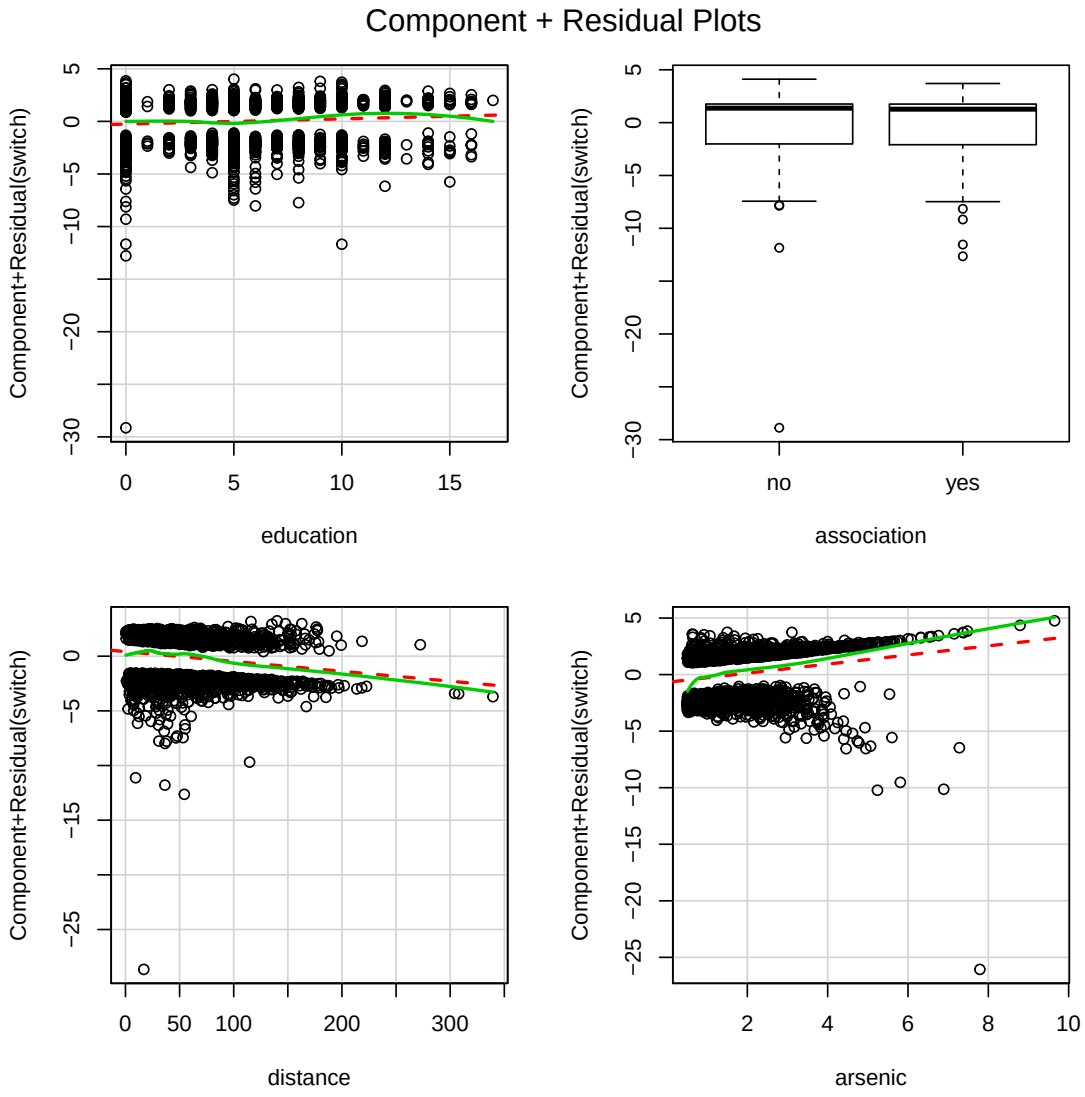

Figure 1: The component/residual plots reveal an indication for a (negative) quadratic effect of extslarsenic.

`arsenic`. Use the natural logarithm. Compare the two models using the AIC and a χ^2-test!

```
wells.best.aq <- update(wells.best, .~. + I(arsenic^2), data=Wells)
Wells$larsen <- with(Wells, log(arsenic))
wells.best.ln <- glm(switch ~ education +  distance + larsen,
  family=binomial(logit), data=Wells)
anova(wells.best.aq, wells.best.ln,  test="Chisq")

## Analysis of Deviance Table
##
## Model 1: switch ~ education + association + distance + arsenic + I(arsenic^2)
## Model 2: switch ~ education + distance + larsen
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3014     3888.9
## 2      3016     3878.2 -2   10.726

AIC(wells.best.aq,wells.best.ln)

##                df      AIC
## wells.best.aq  6 3900.880
## wells.best.ln  4 3886.154
```

6. Using the model that was found better in question 5.

   (a) (1 point) Calculate the predictive probability of switching for a household at the mean values of all predictors.

   ```
   pred.means <- data.frame(education = mean(Wells$education, na.rm = TRUE),
   distance = mean(Wells$distance, na.rm = TRUE), larsen = mean(Wells$larsen,
   na.rm = TRUE))
   p.pred.means <- predict(wells.best.ln, pred.means, type="response")
   ```

   The baseline predicted probability of switching, i.e. the predicted probability of switching for a household at the mean values of all predictors equals 0.582.

   (b) Calculate the effect on the probability of switching when keeping all other predictors constant and

      i. (half a point) changing distance from the mean score to 100 meters;

      ```
      pred.d100 <- data.frame(education = mean(Wells$education, na.rm = TRUE),
      distance = 100, larsen = mean(Wells$larsen,
      na.rm = TRUE))
      p.pred.d100 <- predict(wells.best.ln, pred.d100, type="response")
      ```

      All other predictors constant, when changing the distance from the mean score to 100 meters the probability of switching decreases from $p_{mean} = 0.582$ to $p_{d100} = 0.4564$.

ii. (half a point) changing education from the mean score to one standard deviation above the mean score;

```
pred.edu1 <- data.frame(education = (mean(Wells$education, na.rm = TRUE) +
sd(Wells$education, na.rm=TRUE)), distance = mean(Wells$distance,
na.rm = TRUE), larsen = mean(Wells$larsen, na.rm = TRUE))
p.pred.edu1 <- predict(wells.best.ln, pred.edu1, type="response")
```

All other predictors constant, when changing education from the mean score to one standard deviation above the mean score the probability of switching increases from $p_{mean} = 0.582$ to $p_{edu1} = 0.6235$.

7. (2 points) Compute a logistic regression model for the decision to switch the well using distance, arsenic concentration (in its original scale) and the interaction between the two. How do you interpret the regression coefficients? Are these interpretations meaningful? Give reasons for your answer!

```
wells.ad <- glm(switch ~ distance*arsenic,
  family=binomial(logit), data=Wells)
summary(wells.ad)
```

```
##
## Call:
## glm(formula = switch ~ distance * arsenic, family = binomial(logit),
##     data = Wells)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7823  -1.2004   0.7696   1.0816   1.8476
##
## Coefficients:
##                   Estimate Std. Error z value             Pr(>|z|)
## (Intercept)      -0.147868   0.117538  -1.258              0.20838
## distance         -0.005772   0.002092  -2.759              0.00579 **
## arsenic           0.555977   0.069319   8.021 0.00000000000000105 ***
## distance:arsenic -0.001789   0.001023  -1.748              0.08040 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 3927.6  on 3016  degrees of freedom
## AIC: 3935.6
##
## Number of Fisher Scoring iterations: 4
```

The interpretation follows the usual system, e.g. the inverse logit of the intercept gives the probability of a household to switch which is next to a safe well and where the currently used

well has an arsenic level of 0. Since all wells in the data set have an arsenic level higher than 0.5, this makes no sense.

8. (2 points) Center the variables `distance` and `arsenic` and re-build the model built in question 7. How do you interpret the regression coefficients? Are these interpretations meaningful? Give reasons for your answer! Draw the effects plot for this model and interpret!

```
Wells$arsenic.c <- with(Wells, arsenic-mean(arsenic,na.rm=TRUE))
Wells$distance.c <- with(Wells, distance-mean(distance,na.rm=TRUE))
wells.ad.c <- glm(switch ~ distance.c * arsenic.c,
  family=binomial(logit), data=Wells)
summary(wells.ad.c)
```

```
##
## Call:
## glm(formula = switch ~ distance.c * arsenic.c, family = binomial(logit),
##     data = Wells)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7823  -1.2004   0.7696   1.0816   1.8476
##
## Coefficients:
##                       Estimate Std. Error z value             Pr(>|z|)
## (Intercept)           0.351094   0.039852   8.810 <0.0000000000000002 ***
## distance.c           -0.008737   0.001048  -8.337 <0.0000000000000002 ***
## arsenic.c             0.469508   0.042074  11.159 <0.0000000000000002 ***
## distance.c:arsenic.c -0.001789   0.001023  -1.748               0.0804 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4118.1  on 3019  degrees of freedom
## Residual deviance: 3927.6  on 3016  degrees of freedom
## AIC: 3935.6
##
## Number of Fisher Scoring iterations: 4
```

Now, the coefficients can be interpreted in the usual manner, since they measure the effect obtained by changing one predictor by one unit keeping the other at the mean.