Question2

First we read in the data, take a first look(in appendix) and then do the mean centering:

```
library("pROC")
library("arm")
ldata <- read.table('lalondedata', header = TRUE, sep = ',')
ldata$age.c = ldata$age - mean(ldata$age)
ldata$educ.c = ldata$educ - mean(ldata$educ)
```

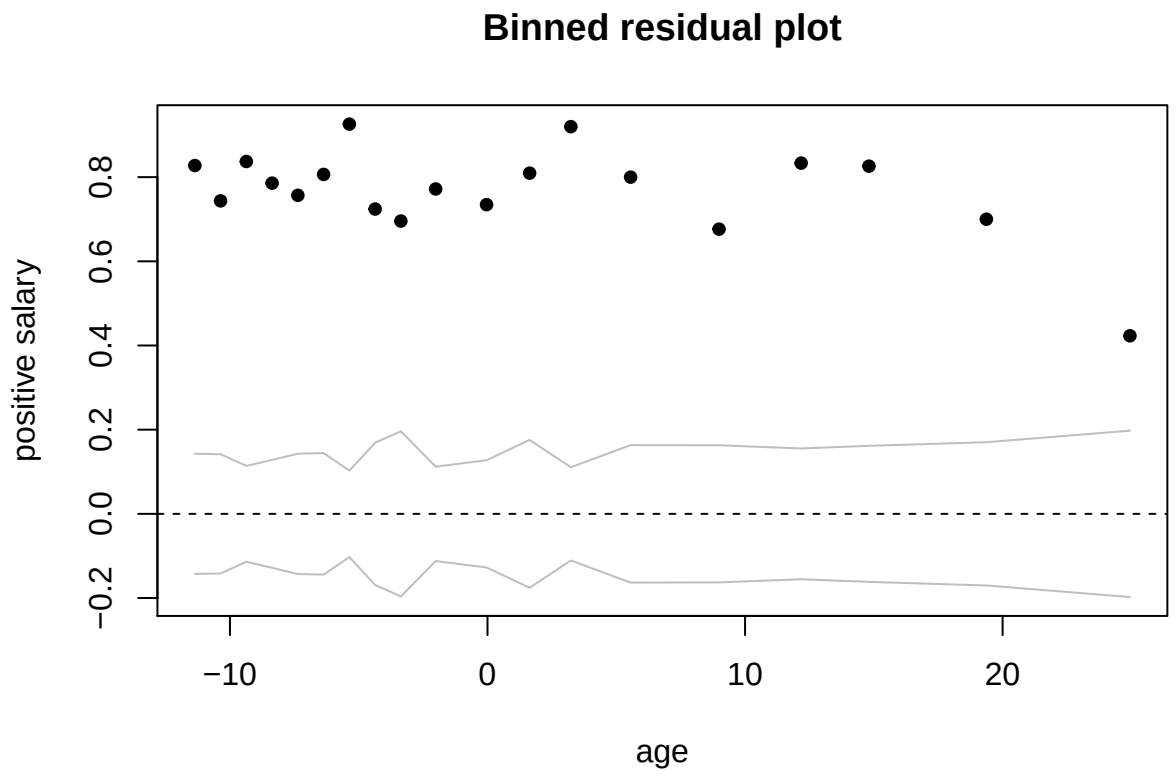
We started from making a dummy variable to show if salary is positive in 1978, and creating two dummy varible as predictor variable: employed_74 and employed_75

```
ldata$positive_sal78 = rep(0, nrow(ldata))
ldata$positive_sal78[ldata$re78 != 0] = 1
ldata$employed_74 = rep(0, nrow(ldata))
ldata$employed_75 = rep(0, nrow(ldata))
ldata$employed_74[ldata$re74 != 0] = 1
ldata$employed_75[ldata$re75 != 0] = 1
```

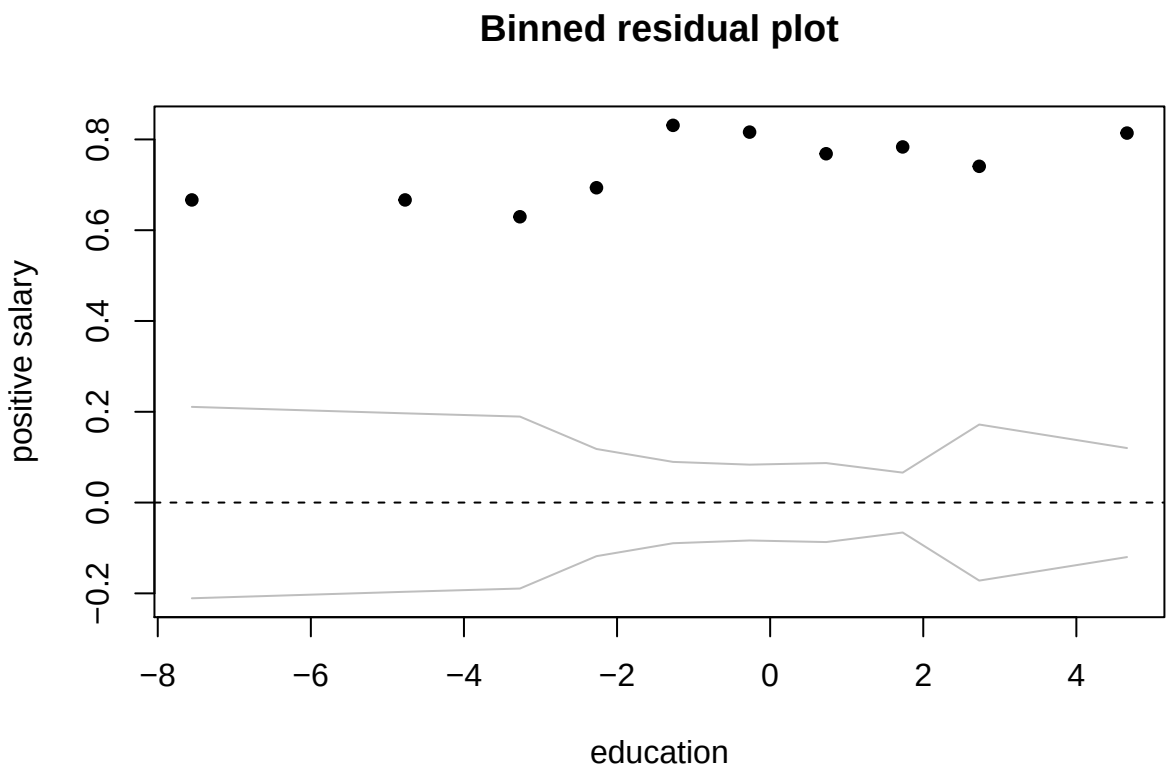
Exploratory data analysis:

Now we are looking at binnedplots of continuous predictors versus zero: All things seem to be good except Age variable seems to have a quadratic trend.

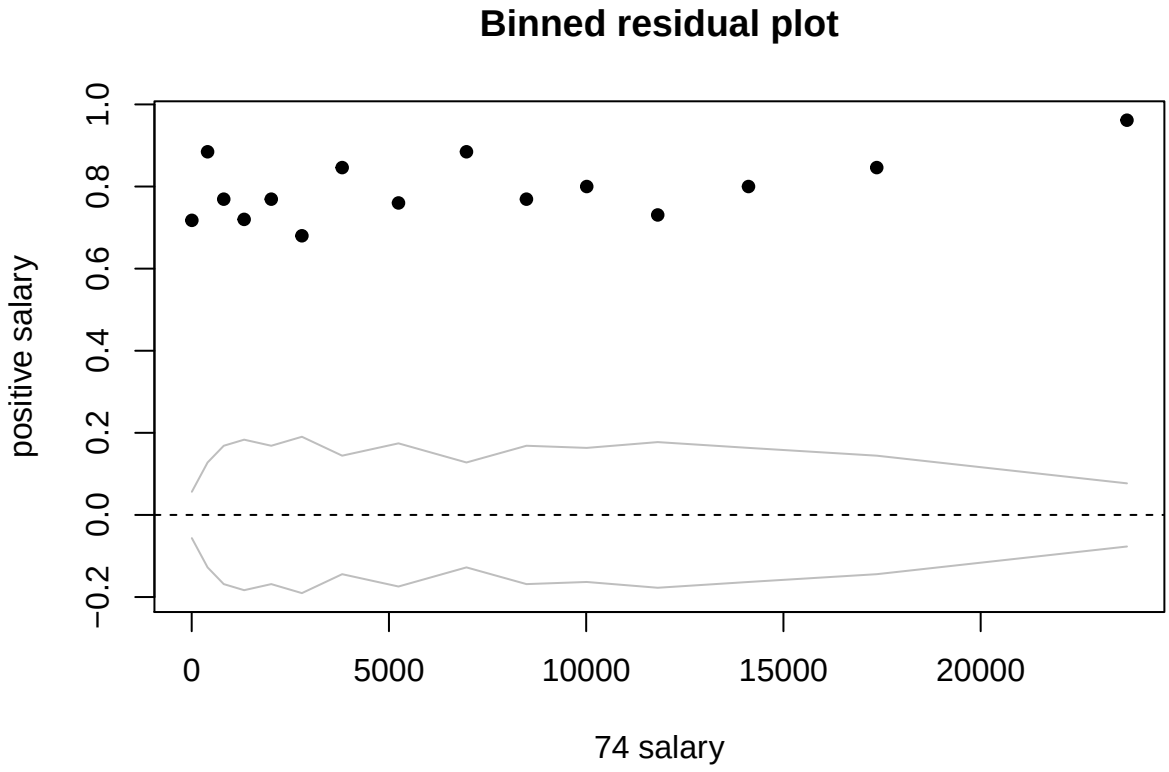
```
binnedplot(ldata$age.c, y=ldata$positive_sal78, xlab = "age", ylab = "positive salary")
```



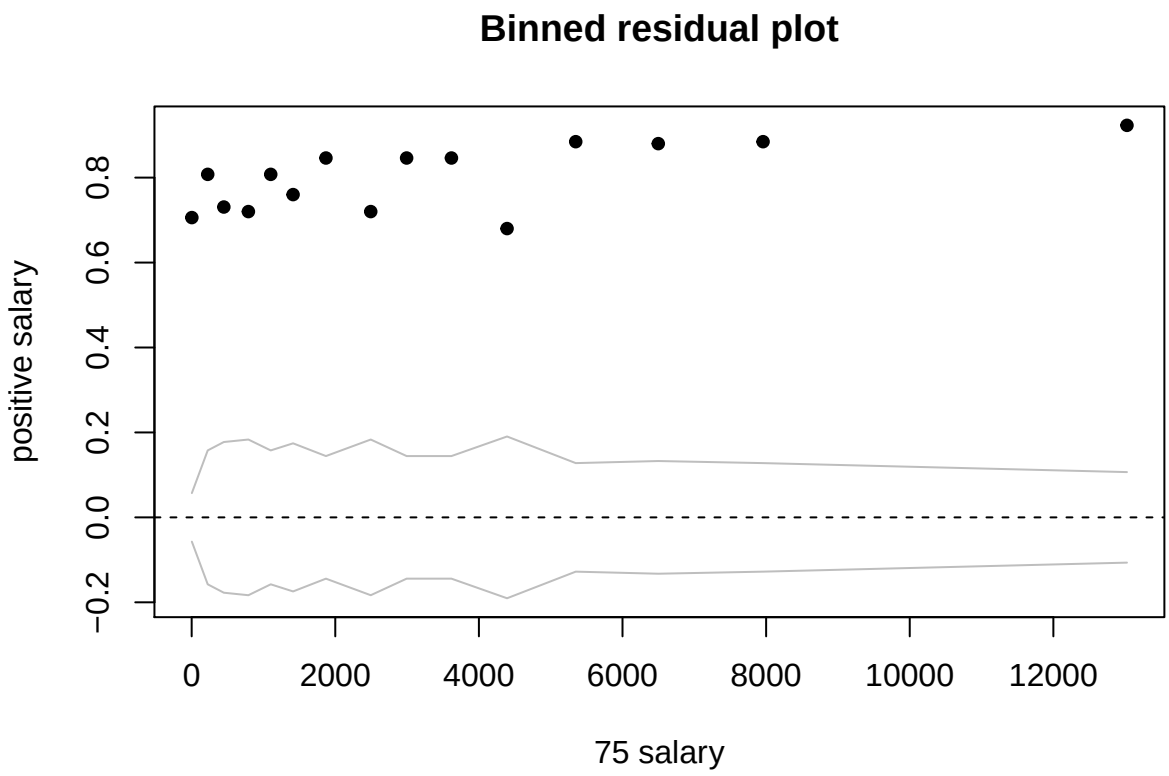

```
binnedplot(ldata$educ.c, y=ldata$positive_sal78, xlab = "education", ylab = "positive salary")
```




```
binnedplot(ldata$re74, y=ldata$positive_sal78, xlab = "74 salary", ylab = "positive salary")
```




```
binnedplot(ldata$re75, y=ldata$positive_sal78, xlab = "75 salary", ylab = "positive salary")
```



Fit the first model

After some Exploratory data analysis, we then try a logistic regression that has a main effect for every variable and linear predictors:

```
q2reg1 = glm(positive_sal78 ~ age.c + educ.c + re74 + re75 + as.factor(employed_74) +
               as.factor(employed_75) + as.factor(black) + as.factor(hispan) +
               as.factor(married) + as.factor(nodegree) + as.factor(treat),
               data = ldata, family = binomial)
summary(q2reg1)
```

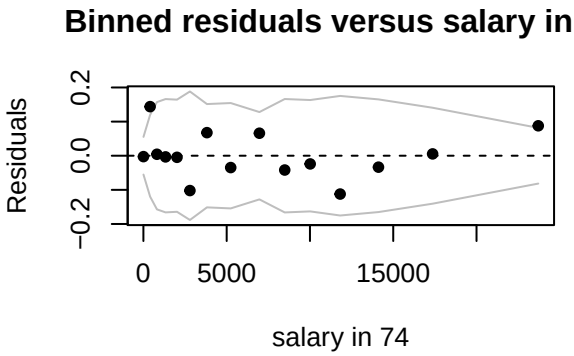
```
##
## Call:
## glm(formula = positive_sal78 ~ age.c + educ.c + re74 + re75 +
##     as.factor(employed_74) + as.factor(employed_75) + as.factor(black) +
##     as.factor(hispan) + as.factor(married) + as.factor(nodegree) +
##     as.factor(treat), family = binomial, data = ldata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2941   0.3665   0.6260   0.7641   1.4589
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              9.141e-01  3.191e-01   2.865 0.004176 **
## age.c                   -3.640e-02  1.093e-02  -3.329 0.000871 ***
## educ.c                   4.804e-02  5.346e-02   0.899 0.368828
## re74                     3.926e-05  2.423e-05   1.620 0.105147
## re75                     9.221e-05  5.204e-05   1.772 0.076445 .
```

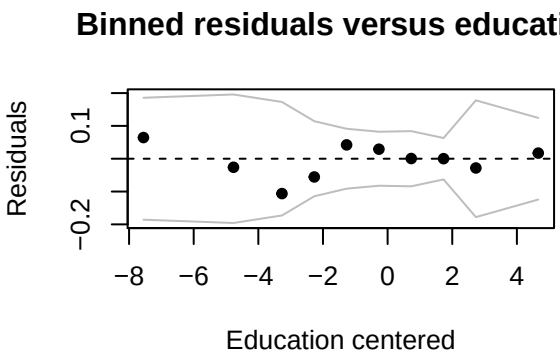
```
## as.factor(employed_74)1 -8.309e-02  2.834e-01  -0.293 0.769372
## as.factor(employed_75)1  6.602e-02  2.692e-01   0.245 0.806292
## as.factor(black)1       -5.356e-01  2.659e-01  -2.014 0.044020 *
## as.factor(hispan)1       2.138e-01  3.643e-01   0.587 0.557325
## as.factor(married)1      1.624e-02  2.412e-01   0.067 0.946305
## as.factor(nodegree)1     1.081e-01  2.986e-01   0.362 0.717449
## as.factor(treat)1        3.722e-01  2.787e-01   1.336 0.181598
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 666.50  on 613  degrees of freedom
## Residual deviance: 629.76  on 602  degrees of freedom
## AIC: 653.76
##
## Number of Fisher Scoring iterations: 4
```

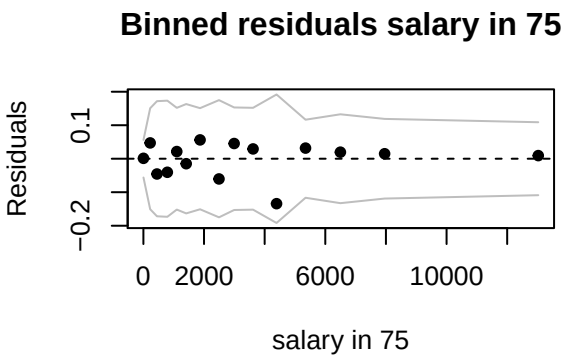
Model diagnostics

We first do binned residual plots for numeric variables: We don't find any noticeable patterns.

```
par(mfcol = c(2,2))
rawresid1 = ldata$positive_sal78 - fitted(q2reg1)
binnedplot(x=ldata$age.c, y = rawresid1, xlab = "Age centered", ylab = "Residuals",
main = "Binned residuals versus age")

binnedplot(x=ldata$educ.c, y = rawresid1, xlab = "Education centered", ylab = "Residuals",
main = "Binned residuals versus education")

binnedplot(x=ldata$re74, y = rawresid1, xlab = "salary in 74", ylab = "Residuals",
main = "Binned residuals versus salary in 74")

binnedplot(x=ldata$re75, y = rawresid1, xlab = "salary in 75", ylab = "Residuals",
main = "Binned residuals salary in 75")
```







Then look at average residuals by dummy variables using the tapply command:

```
tapply(rawresid1, ldata$black, mean)
```

```
##             0             1
## 1.029411e-09 8.307175e-10
```

```
tapply(rawresid1, ldata$hispan, mean)
```

```
##             0             1
## 9.315583e-10 1.095431e-09
```

```
tapply(rawresid1, ldata$married, mean)
```

```
##             0             1
## 2.972536e-10 1.870830e-09
```

```
tapply(rawresid1, ldata$nodegree, mean)
```

```
##             0             1
## 9.866986e-10 9.297030e-10
```

```
tapply(rawresid1, ldata$treat, mean)
```

```
##             0             1
## 1.073914e-09 6.652248e-10
```

```
tapply(rawresid1, ldata$employed_74, mean)
```

```
##             0             1
## 1.008136e-10 1.507488e-09
```

```
tapply(rawresid1, ldata$employed_75, mean)
```

```
##            0            1
## 6.417200e-12 1.577787e-09
```

Confusion matrix with .5 threshold and .6 threshold:

```
threshold = 0.6
table(ldata$positive_sal78, q2reg1$fitted > threshold)
```

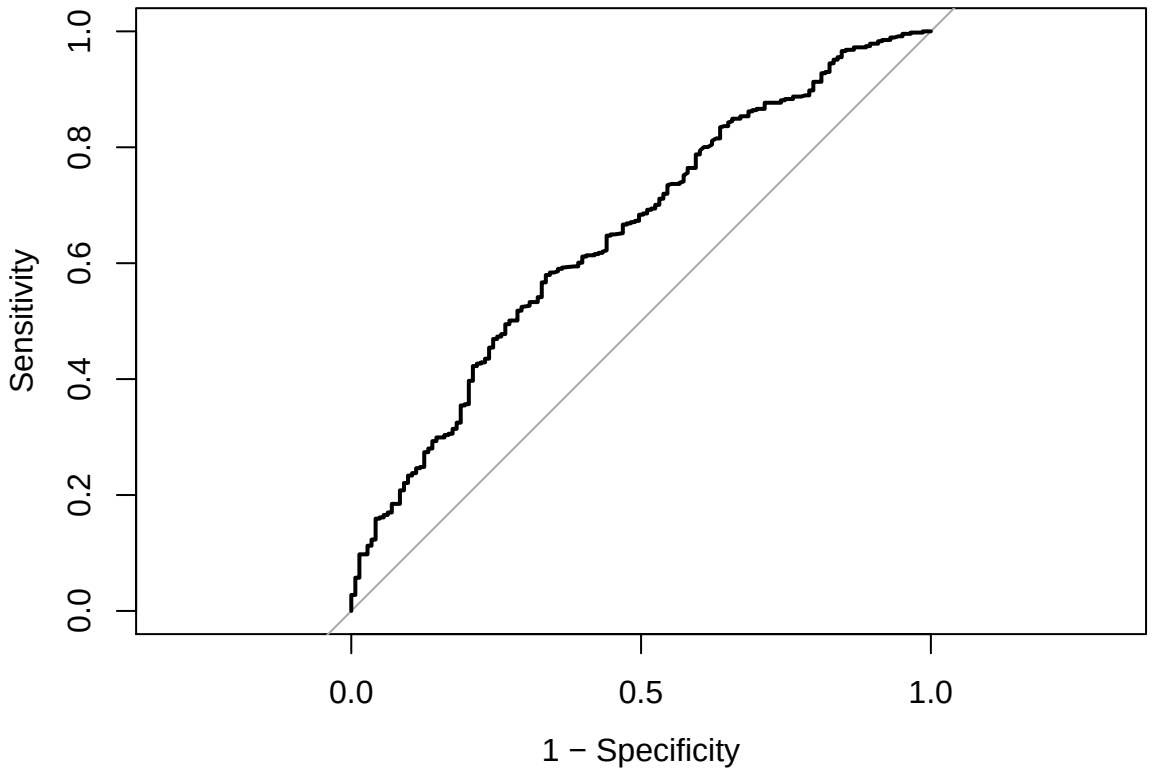
```
##
##     FALSE TRUE
##   0    22  121
##   1    19  452
```

```
threshold = 0.5
table(ldata$positive_sal78, q2reg1$fitted > threshold)
```

```
##
##     FALSE TRUE
##   0    10  133
##   1     6  465
```

Then we look at ROC curve: We didn't find specific pattern from the model diagnostics, so we decide not to do transformations. We got area under the curve value: 0.6501.

```
roc(ldata$positive_sal78, fitted(q2reg1), plot=T, legacy.axes=T)
```




```
##
## Call:
## roc.default(response = ldata$positive_sal78, predictor = fitted(q2reg1),     plot = T, legacy.axes =
##
## Data: fitted(q2reg1) in 143 controls (ldata$positive_sal78 0) < 471 cases (ldata$positive_sal78 1).
## Area under the curve: 0.6501
```

Analytics for interaction I'm interested in:

We're interested in education level and want to make sure whether the effect changes by demographic groups of different education level, so we remove Education variable and do the change in deviance test, and we get a p value of 0.3695, so education level might not be so important, but we will keep it in the model.

```
q2reg3 = glm(positive_sal78 ~ age.c + re74 + re75 + as.factor(employed_74)
             + as.factor(employed_75) + as.factor(black) + as.factor(hispan)
             + as.factor(married) + as.factor(nodegree) + as.factor(treat),
             data = ldata, family = binomial)
anova(q2reg1, q2reg3, test= "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: positive_sal78 ~ age.c + educ.c + re74 + re75 + as.factor(employed_74) +
##     as.factor(employed_75) + as.factor(black) + as.factor(hispan) +
##     as.factor(married) + as.factor(nodegree) + as.factor(treat)
## Model 2: positive_sal78 ~ age.c + re74 + re75 + as.factor(employed_74) +
##     as.factor(employed_75) + as.factor(black) + as.factor(hispan) +
##     as.factor(married) + as.factor(nodegree) + as.factor(treat)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       602     629.76
## 2       603     630.56 -1 -0.80539   0.3695
```