Oblig 1, Sanders

## Komment

Since we did not have lectures in crossvalidation, a persion aksed the teacher what we shuld do with thise tasks, and the teachers answer was that we will not need to do them. So this is my solution for the oblig, where I have not compleated what was needed from the dalayed lecture. If it is not approved I will do the tasks.

## Problem 1

# Getting data  
res\_bodyfat <- read.csv("res\_bodyfat/res\_bodyfat.csv")  
attach(res\_bodyfat)

Aim: Find how well bmi, easily computed as the ratio between weight and squared hight (so mesured in kg/m^2), can be used to predict pbfm, whose measurment involves instead a bioelectrical impedance analysis.

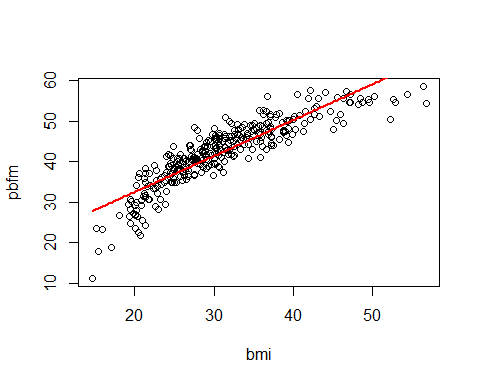
### a

Plotting **bnp** against **pbfm**, and fitting a simple linear model with a summary.

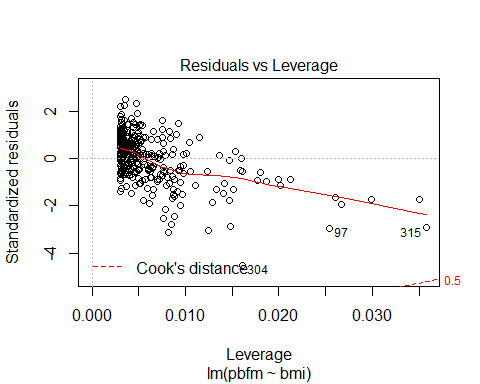
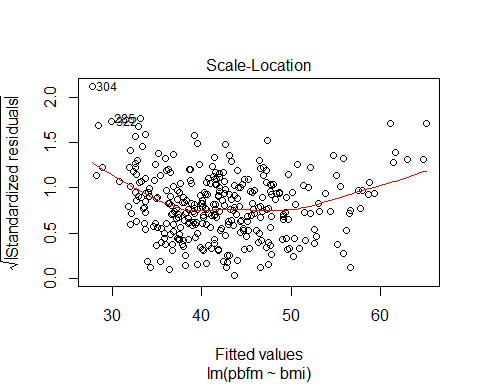
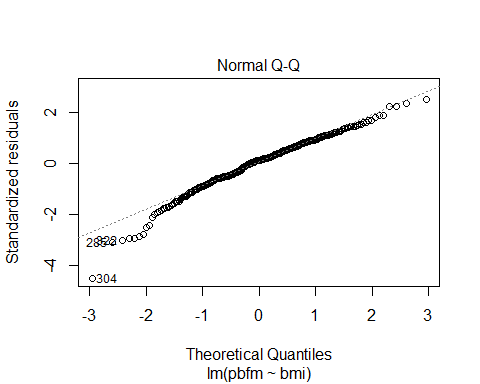
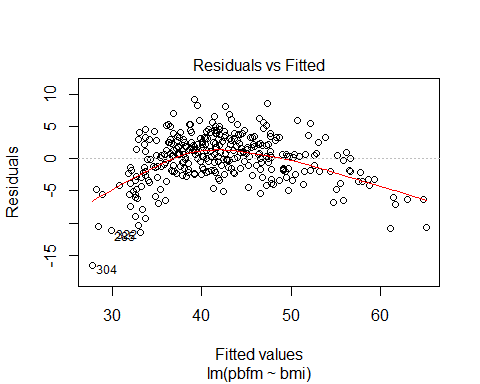
plot(bmi, pbfm)  
fit.simpleLinear <- lm(pbfm ~ bmi)  
print(summary(fit.simpleLinear))

##   
## Call:  
## lm(formula = pbfm ~ bmi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.5116 -2.0714 0.4083 2.4994 9.1758   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14.82772 0.82671 17.94 <2e-16 \*\*\*  
## bmi 0.88481 0.02589 34.17 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.702 on 325 degrees of freedom  
## Multiple R-squared: 0.7823, Adjusted R-squared: 0.7816   
## F-statistic: 1168 on 1 and 325 DF, p-value: < 2.2e-16

x = (seq(min(bmi), max(bmi), length = 200))  
beta = coef(fit.simpleLinear)  
lines(x, beta[1] + beta[2]\* x, col = 2, lty = 1, lwd = 2)

 A linear model seems ok in the center of the bmi. But it is quite off in the lower 1/4 and in the upper 1/4.We can see that the intercept and bmi is strongly significant, indicating a strong correlation between bmi and pdfm. The R-squared is siginificant, but not super high, so the model can probably be improved. To see this clearly we plott indicitative plotts on the fitness.

plot(fit.simpleLinear)



#### Residuals vs Fitted

Shows the Anscombe plot, wich would idealy present a eaven scattering with no patterns in the points. In the plott we see a clear pattern, and is indicating possiple violation of homoscedasicity. Meaning that must be independent of the index.

#### Normal Q-Q

This plot shows quantile-quantile plot. This plot shows the values of the resiuals in increasing order. To see if the residuals folows a normal distribution. For the assumption of normality to be valid we would expect the values to folow the line in the plot. In the plot a heavy tail and head can be observed. So the assumtion of normality is not strong.

#### Scale-Location

#### Residual vs Leverage

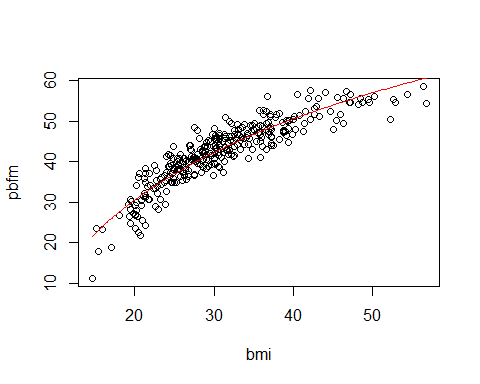
## b

We can see that we dont have any negative number, in such cases it is often benefital to use a logarithmic transformation.

fit.log <- lm(pbfm ~ log(bmi))  
print(summary(fit.log))

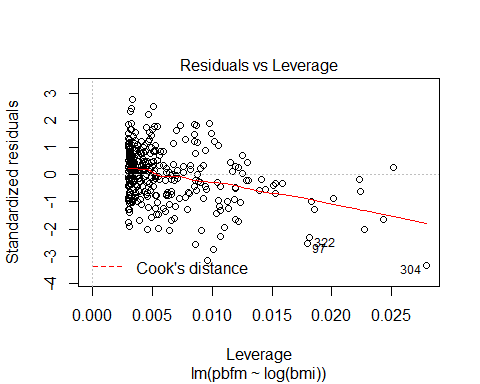
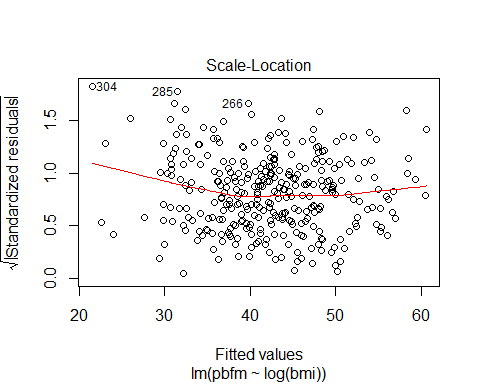
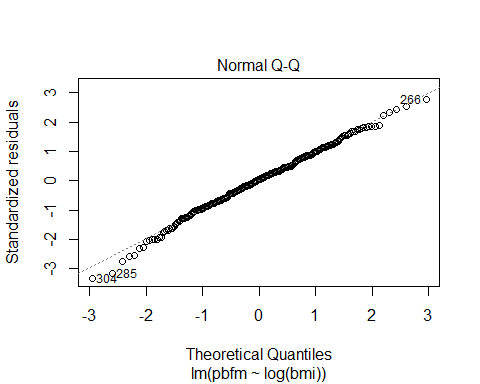
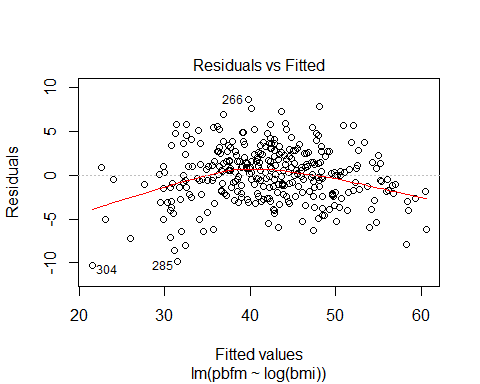
##   
## Call:  
## lm(formula = pbfm ~ log(bmi))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.2548 -2.0453 0.1026 2.1238 8.6029   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -55.7327 2.3337 -23.88 <2e-16 \*\*\*  
## log(bmi) 28.8031 0.6845 42.08 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.125 on 325 degrees of freedom  
## Multiple R-squared: 0.8449, Adjusted R-squared: 0.8444   
## F-statistic: 1771 on 1 and 325 DF, p-value: < 2.2e-16

beta = coef(fit.log)  
  
x = log(seq(min(bmi), max(bmi), length = 200))  
plot(bmi, pbfm)  
  
lines(exp(x),( beta[1] + beta[2]\* x), col = 2)



And to see how good the model is we plot some helping plots

plot(fit.log)

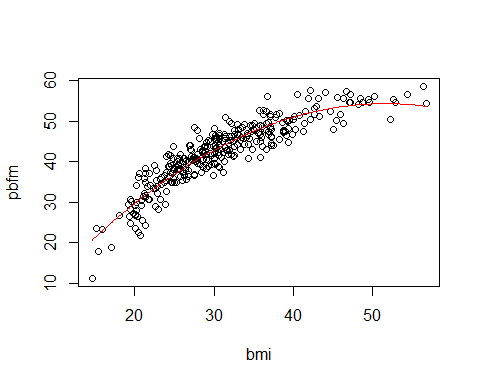


It was also suggested to use a quadratic model.

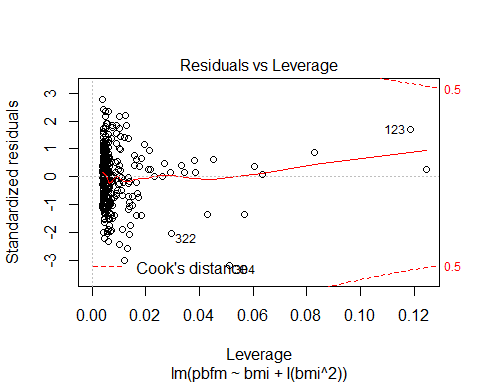
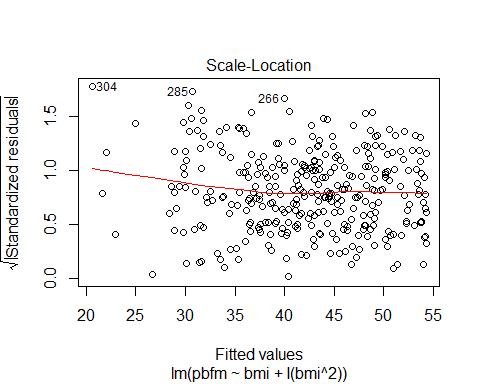
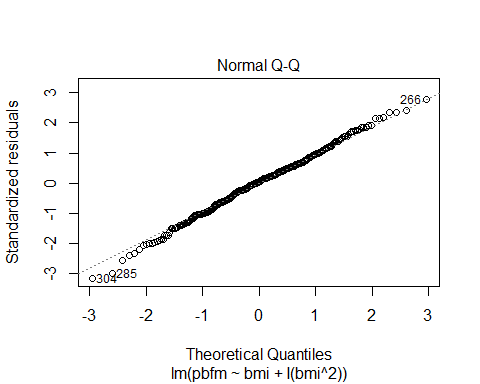
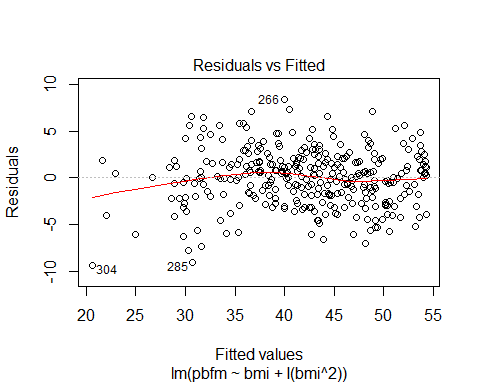
fit.quad<- lm(pbfm ~ bmi + I(bmi^2))  
print(summary(fit.quad))

##   
## Call:  
## lm(formula = pbfm ~ bmi + I(bmi^2))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.3403 -1.9246 0.1433 1.8665 8.3780   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11.17790 2.16977 -5.152 4.49e-07 \*\*\*  
## bmi 2.53223 0.13229 19.142 < 2e-16 \*\*\*  
## I(bmi^2) -0.02448 0.00194 -12.617 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.036 on 324 degrees of freedom  
## Multiple R-squared: 0.854, Adjusted R-squared: 0.8531   
## F-statistic: 947.8 on 2 and 324 DF, p-value: < 2.2e-16

beta = coef(fit.quad)  
  
x = (seq(min(bmi), max(bmi), length = 200))  
plot(bmi, pbfm)  
#points(bmi, pbfm)  
lines(x, beta[1] + beta[2]\*x + beta[3]\*x^2, col=2)

 And the acopening helping plots

plot(fit.quad)



## C - Leave-p-out cross-validation (<https://en.wikipedia.org/wiki/Cross-validation_(statistics)>)

#max.p = 10 #n = 4 # n = len bmi #shuffled = sample(1:length(bmi)) #subset.size = length(bmi) %/% n #indices<-matrix(list(), nrow=n, ncol=1) # indices[[1,1]] = shuffled[1:subset.size] # indices[[2,1]] = shuffled[subset.size +1:subset.size\*2] # indices[[3,1]] = shuffled[(subset.size*2 + 1):subset.size*3] # indices[[4,1]] = shuffled[(subset.size\*3 + 1):length(bmi)]

#for(i in 0:n-1){ # a = (subset.size*i) + 1 # b = subset.size*(i+1) # print(c(a,b)) # indices[[i+1,1]] = shuffled[a : b] #}

#for (i in 1:max.p){ # for(j in 0:n){ # x = bmi[c(indices[[j %% n]], indices[[(j+1) %% n]], indices[[(j+2) %% n]])] # y = pbfm[c(indices[[j %% n]], indices[[(j+1) %% n]], indices[[(j+2) %% n]])] # fit = lm(pbfm ~ poly(x, i), x = TRUE) # x.test = bmi[] # } #}

## Problem 2

oral\_ca <- read.csv("oral\_ca/oral\_ca.csv")  
attach(oral\_ca)  
print(summary(oral\_ca))

## drinks ccstatus cigs age   
## Min. : 0.00 Min. :0.0000 Min. : 0.00 Min. :21   
## 1st Qu.: 1.50 1st Qu.:0.0000 1st Qu.: 3.00 1st Qu.:48   
## Median : 15.75 Median :0.0000 Median :20.00 Median :56   
## Mean : 31.40 Mean :0.4887 Mean :16.36 Mean :56   
## 3rd Qu.: 48.00 3rd Qu.:1.0000 3rd Qu.:20.00 3rd Qu.:65   
## Max. :140.00 Max. :1.0000 Max. :60.00 Max. :80   
## sex M\_drinks M\_cigs   
## Min. :0.0000 Min. :0.000000 Min. :0.00000   
## 1st Qu.:0.0000 1st Qu.:0.000000 1st Qu.:0.00000   
## Median :0.0000 Median :0.000000 Median :0.00000   
## Mean :0.2771 Mean :0.005038 Mean :0.01511   
## 3rd Qu.:1.0000 3rd Qu.:0.000000 3rd Qu.:0.00000   
## Max. :1.0000 Max. :1.000000 Max. :1.00000

Aim of the study was to evaluate the risk of Oral cancer based on the variables **drinks** (number og 1oz ethanol-equivalent drinks consumed per week), **sex**, **age** and **cigs** (number of cigaretts smoked per day).

### We are firs interested in the effect of smoking alone

#### (a)

Q:

Dichotomize the variable cigs in two categories, smokers and not smokers, and create a table with the observed frequencies for cases and controls. In addition, provide the estimated probabilities (including their standard errors) of experiencing an oral cancer (i.e., being a case) for the two sub-populations.

A:

N = length(oral\_ca$ccstatus)  
smoker.nonSmoker = cigs == 0  
smokers = subset( oral\_ca, oral\_ca$cigs > 0 )#smokers = oral\_ca[ oral\_ca$cigs > 0]  
nonSmokers = subset(oral\_ca, oral\_ca$cigs == 0) # nonSmokers = oral\_ca[ oral\_ca$cigs == 0 ]  
number.of.cc = nrow( subset(oral\_ca,oral\_ca$ccstatus == 1) )  
number.of.cc.smokers = nrow( subset(oral\_ca, oral\_ca$ccstatus == 1 & oral\_ca$cigs > 0) )  
number.of.cc.nonSmokers = nrow(subset(oral\_ca, oral\_ca$ccstatus == 1 & oral\_ca$cigs == 0) )  
n.smoker = nrow( subset( oral\_ca, oral\_ca$cigs > 0) )  
n.nonSmoker = nrow( subset( oral\_ca, oral\_ca$cigs == 0 ))  
  
# Binomial - Mean  
pi.common.hat = number.of.cc / N # Number of ppl with cancer devided on ppl (people in the query )  
pi.smoker.hat = number.of.cc.smokers / n.smoker # Number of ppl with cancer and smokes, devided on number of ppl who smokes.  
pi.nonSmoker.hat = number.of.cc.nonSmokers / n.nonSmoker# Number of ppl with cancer and does not smoke, devided on number of ppl who does not soke.  
  
# Binomial Vairance - Var(X) = n\*p\*(1-p)  
se.pi.common = sqrt(pi.common.hat \* (1 - pi.common.hat) / N)  
se.pi.smoker = sqrt(pi.smoker.hat \* (1 - pi.nonSmoker.hat) /n.smoker)  
se.pi.nonSmoker = sqrt(pi.nonSmoker.hat \* (1 - pi.nonSmoker.hat) /n.nonSmoker)  
  
cc.smoke.df = data.frame(   
 pi = c(pi.common.hat, pi.smoker.hat, pi.nonSmoker.hat),   
 se = c(se.pi.common, se.pi.smoker, se.pi.nonSmoker),  
 row.names = c("General", "Smoker", "NonSmoker")  
 )

#### (b)

Q:

Test the hypothesis that the two probabilities are equal and comment on the result.

A:

Hypothesis: H0: probSmokers == probNonSmokers against H1: probSmokers != probNonSmokers. Computing the p-value

# compute the likelihood ratio statistics test  
llik\_pi.smoker.hat\_pi.nonSmoker.hat = sum(dbinom(smoker.nonSmoker[ smoker.nonSmoker == TRUE], 1, pi.smoker.hat, log = TRUE)) +   
 sum(dbinom(smoker.nonSmoker[ smoker.nonSmoker == FALSE], 1, pi.nonSmoker.hat, log = TRUE))  
llik\_pi.common.hat = sum(dbinom(smoker.nonSmoker, 1, pi.common.hat, log=TRUE))  
  
w = 2 \* (llik\_pi.smoker.hat\_pi.nonSmoker.hat - llik\_pi.common.hat)  
p.val = 1 - pchisq(w, df = 1)

We can se that the p-val is zero, so we can drop H0.

#### (c)

Q:

Fit a linear logistic model using the dichotomized variable as explanatory variable and comment on the result: does being a smoker increase or decrease the risk of experiencing oral cancer? How much, in terms of log-odds?

A:

y = oral\_ca$ccstatus == 1  
x = oral\_ca$cigs > 0  
  
mod = glm(y ~ x, family = 'binomial')  
print(summary(mod))

##   
## Call:  
## glm(formula = y ~ x, family = "binomial")  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.285 -1.285 -0.744 1.073 1.685   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.1431 0.2448 -4.669 3.03e-06 \*\*\*  
## xTRUE 1.3927 0.2706 5.147 2.65e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 550.15 on 396 degrees of freedom  
## Residual deviance: 520.14 on 395 degrees of freedom  
## AIC: 524.14  
##   
## Number of Fisher Scoring iterations: 4

beta = mod$coefficients  
print(beta)

## (Intercept) xTRUE   
## -1.143064 1.392719

pi.nonSmoker = exp(beta[1]) / (1 + exp(beta[1]))  
pi.smoker = exp(beta[1] + beta[2]) / (1 + exp(beta[1] + beta[2]))  
  
#odds  
odds.nonSmoker = pi.nonSmoker.hat / (1 - pi.nonSmoker.hat)  
odds.smoker = pi.smoker.hat / (1 - pi.smoker.hat)  
#log odds ratio  
log.odds.ratio = (odds.smoker / odds.nonSmoker)  
  
log.odds.ratio

## [1] 4.02578

We see that the odds of cc for a smoker against a non somoker is aprox 4, aka. 4 to 1.

#### (d)

Q:

Repeat the analysis of point (c) by considering, now, the number of cigarettes as a continuous variable. Comment on the result: What does the regression coefficient for this variable mean now? Why does the value of the intercept change with respect to point (c), although they are both related to the odds for non-smokers?

A:

x = cigs  
y = ccstatus  
mod.continuous = glm(y ~ x, family = 'binomial')  
print(summary(mod.continuous))

##   
## Call:  
## glm(formula = y ~ x, family = "binomial")  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.1923 -1.0237 -0.8228 1.1088 1.5796   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.909057 0.171766 -5.292 1.21e-07 \*\*\*  
## x 0.053624 0.008614 6.225 4.81e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 550.15 on 396 degrees of freedom  
## Residual deviance: 504.39 on 395 degrees of freedom  
## AIC: 508.39  
##   
## Number of Fisher Scoring iterations: 4

The regression coefficient now tells us the increase in odds of cancer per cigarets smoked per day. The intercept is changed since smoking one sigaret per day is closer to zero than smoking 20. In the non continius model the biggest smoker and the rest is given the same odds.

### Consider now the other three variables (drinks, sex and age) as well:

#### (e)

Q:

Fit a linear logistic model including all the explanatory variables and report the result. What is the increase in terms of log-odds for an increasing number of cigarettes per day smoked estimated by this model? Why did it change from the one obtained in model ﬁtted in point (c)?

mod.all = glm(y ~ oral\_ca$drinks + oral\_ca$cigs + oral\_ca$age + oral\_ca$sex, family = 'binomial')  
print(summary(mod.all))

##   
## Call:  
## glm(formula = y ~ oral\_ca$drinks + oral\_ca$cigs + oral\_ca$age +   
## oral\_ca$sex, family = "binomial")  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.7185 -0.8589 -0.5832 0.9644 1.9776   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.966071 0.620756 -3.167 0.00154 \*\*   
## oral\_ca$drinks 0.029623 0.004643 6.380 1.77e-10 \*\*\*  
## oral\_ca$cigs 0.035480 0.009571 3.707 0.00021 \*\*\*  
## oral\_ca$age 0.006529 0.009960 0.656 0.51213   
## oral\_ca$sex 0.594499 0.272752 2.180 0.02928 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 550.15 on 396 degrees of freedom  
## Residual deviance: 443.84 on 392 degrees of freedom  
## AIC: 453.84  
##   
## Number of Fisher Scoring iterations: 5

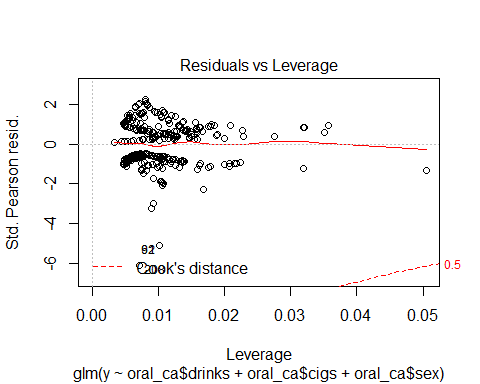
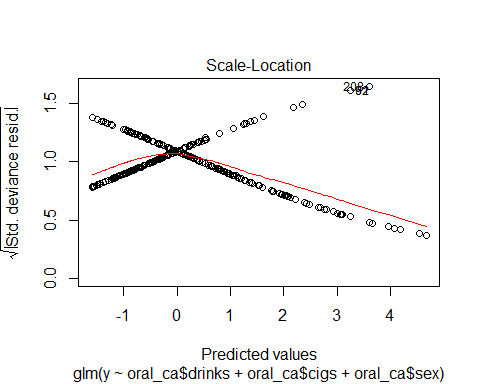
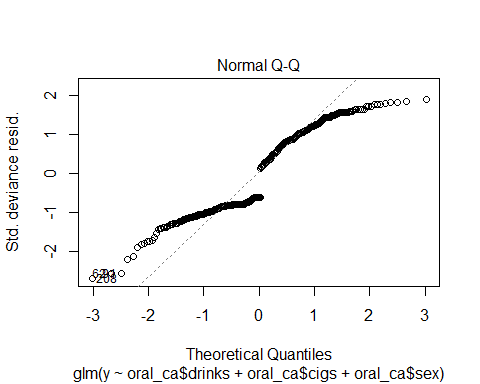
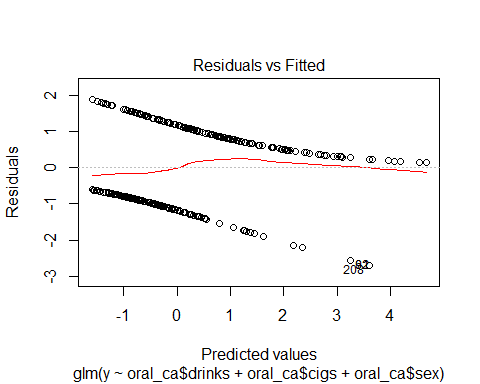
The increase in terms of log-odds for an increasing number of cigaretts par day by this model is 0.035, wich is a little lower than in the prewius model. This can be because of some of the affects of other facors was tried explained by one facotr. Now we have other factors to explain this change. Wich will change the value.

#### (e)

mod.without.Age = glm(y ~ oral\_ca$drinks + oral\_ca$cigs + oral\_ca$sex, family = 'binomial')  
print(summary(mod.without.Age))

##   
## Call:  
## glm(formula = y ~ oral\_ca$drinks + oral\_ca$cigs + oral\_ca$sex,   
## family = "binomial")  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6943 -0.8596 -0.6084 0.9607 1.8857   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.592919 0.238787 -6.671 2.54e-11 \*\*\*  
## oral\_ca$drinks 0.029498 0.004638 6.360 2.01e-10 \*\*\*  
## oral\_ca$cigs 0.035536 0.009565 3.715 0.000203 \*\*\*  
## oral\_ca$sex 0.582183 0.271756 2.142 0.032169 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 550.15 on 396 degrees of freedom  
## Residual deviance: 444.27 on 393 degrees of freedom  
## AIC: 452.27  
##   
## Number of Fisher Scoring iterations: 5

plot(mod.without.Age)



Since the explanatory variable age has a large p-value can we not say it gives significant value to the model. So it does not seem like there is a difference in age. Lookin at the deviance the age variable did noting significant for the model ether. So i would choose the one with less explanatory variables, aka the one without Age

#### (g)

Q:

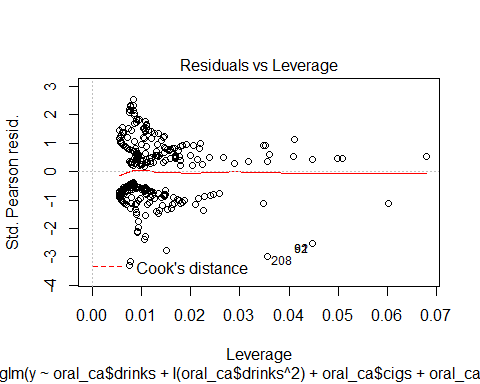
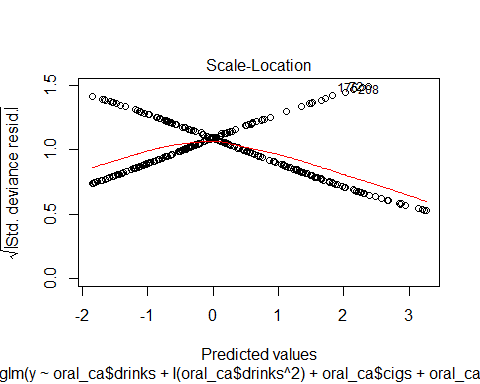
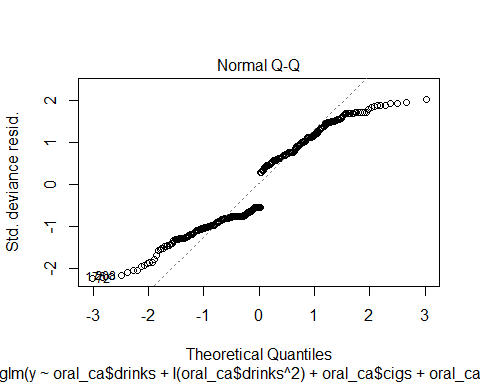
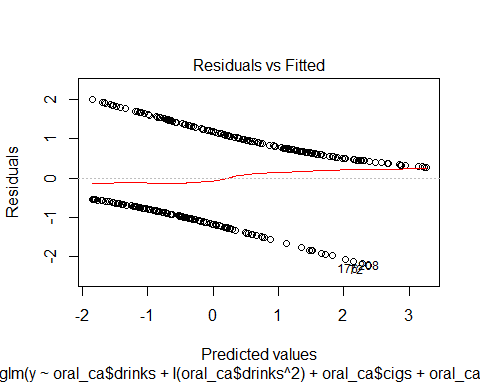
Use a polynomial of degree 2 to model the effect of drinks. Does it improve the model?

A:

mod.quad.drinks = glm(y ~ oral\_ca$drinks + I(oral\_ca$drinks^2) + oral\_ca$cigs + oral\_ca$sex, family = 'binomial')  
print(summary(mod.quad.drinks))

##   
## Call:  
## glm(formula = y ~ oral\_ca$drinks + I(oral\_ca$drinks^2) + oral\_ca$cigs +   
## oral\_ca$sex, family = "binomial")  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2192 -0.8379 -0.5405 0.8756 1.9980   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.850e+00 2.667e-01 -6.936 4.03e-12 \*\*\*  
## oral\_ca$drinks 5.362e-02 1.052e-02 5.099 3.42e-07 \*\*\*  
## I(oral\_ca$drinks^2) -2.295e-04 8.419e-05 -2.726 0.006405 \*\*   
## oral\_ca$cigs 3.301e-02 9.642e-03 3.424 0.000618 \*\*\*  
## oral\_ca$sex 7.256e-01 2.854e-01 2.542 0.011007 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 550.15 on 396 degrees of freedom  
## Residual deviance: 437.62 on 392 degrees of freedom  
## AIC: 447.62  
##   
## Number of Fisher Scoring iterations: 4

plot(mod.quad.drinks)



The change to a quadratic polynomial for the drinks varibale did not seem to change much, it is a small variable comaring to the other once, but have a sigificant p-value. The p-value can be caused by the hig similarity between drinks and drinks^2, this theory is strengthen by the increase in the p-value of drinks.

#### (h)

Q:

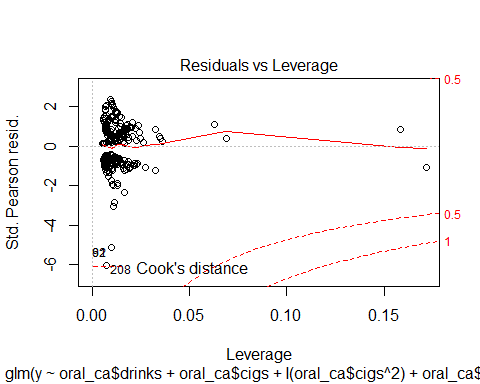
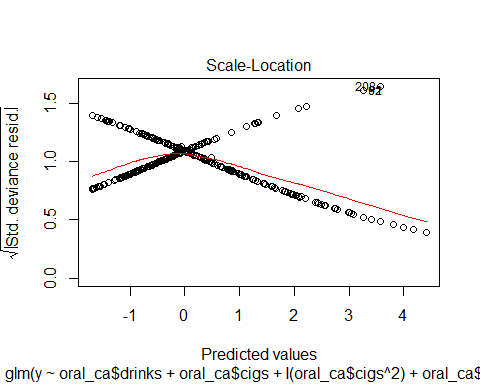
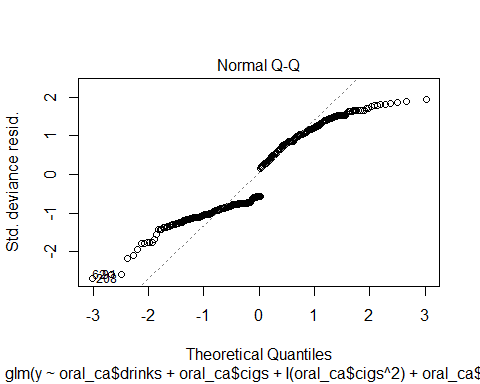
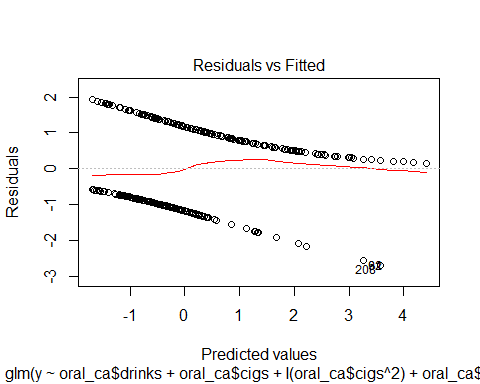
What about doing that for cigs instead of drinks? What is the effect on the model?

A:

mod.quad.smokes = glm(y ~ oral\_ca$drinks + oral\_ca$cigs+ I(oral\_ca$cigs^2) + oral\_ca$sex, family = 'binomial')  
print(summary(mod.quad.smokes))

##   
## Call:  
## glm(formula = y ~ oral\_ca$drinks + oral\_ca$cigs + I(oral\_ca$cigs^2) +   
## oral\_ca$sex, family = "binomial")  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6870 -0.8764 -0.5808 0.9695 1.9303   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.6943120 0.2697349 -6.281 3.36e-10 \*\*\*  
## oral\_ca$drinks 0.0289791 0.0046347 6.253 4.04e-10 \*\*\*  
## oral\_ca$cigs 0.0541177 0.0238775 2.266 0.0234 \*   
## I(oral\_ca$cigs^2) -0.0004485 0.0005225 -0.858 0.3907   
## oral\_ca$sex 0.6019008 0.2742844 2.194 0.0282 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 550.15 on 396 degrees of freedom  
## Residual deviance: 443.55 on 392 degrees of freedom  
## AIC: 453.55  
##   
## Number of Fisher Scoring iterations: 5

plot(mod.quad.smokes)



Adding cigs^2 as a variable also did not change the model much. Here the cigs^2 variable eaven has a high p-value, showing the insignificance of the variable.

#### (i)

Adding the quadratic variables did not seem to give the biggest effect. It seems to make the none suared variable (cigs, drinks) less significant with a compareble degree the quadratic values give to the model. So they might be better but not significantly. No significant difference in the residuals ether.

#### (j)

If the model fit a quareatic scale, that means that two drinks is more then twice as bad than one drink. It does not necessarily make it worse to drink 10 units, if the model fits with linear or quadratic wariables, it will just change the coefficients. A way to explain this is with the functions

f(x = 10) = beta \* x = a, and f(x = 10) = beta1 \* x + beta2 \* x^2 = a.

The point of the function is to show that if you drink 10 drinks and the relation is linear can be the same as drinking 10 drinks in a quadradic model. This is correct in all polynomials.

#### (k)

…