Assignment 4 DSC 423

Alex Teboul

3/7/2019

Class

DSC 423 - Data Analysis and Regression
Total points: 43pts

Problem 1 Churn analysis [10 pts for undergrad/16 pts for grad students]

Given the large number of competitors, cell phone carriers are very interested in analyzing and predicting customer retention and churn. The primary goal of churn analysis is to identify those customers that are most likely to discontinue using your service or product. The dataset churn_train.csv contains information about a random sample of customers of a cell phone company. For each customer, company recorded the following variables:

1. CHURN: 1 if customer switched provider, 0 if customer did not switch
2. GENDER: M, F
3. EDUCATION (categorical): code 1 to 6 depending on education levels
   LAST_PRICE_PLAN_CHNG_DAY_CNT: No. of days since last price plan change
4. TOT_ACTV_SRV_CNT: Total no. of active services
5. AGE: customer age
6. PCT_CHNG_IB_SMS_CNT: Percent change of latest 2 months incoming SMS wrt previous 4 months incoming SMS
7. PCT_CHNG_BILL_AMT: Percent change of latest 2 months bill amount wrt previous 4 months bill amount
8. COMPLAINT: 1 if there was at least a customer's complaint in the two months, 0 no complaints

The company is interested in a churn predictive model that identifies the most important predictors affecting probability of switching to a different mobile phone company (churn = 1). Answer the following questions:

Problem 1 a)

Get data and assign variables

```
# load in the data from file
myd=read.csv("/Users/alexteboul/Desktop/churn_train1.csv", header=T)
myd[1,]
```

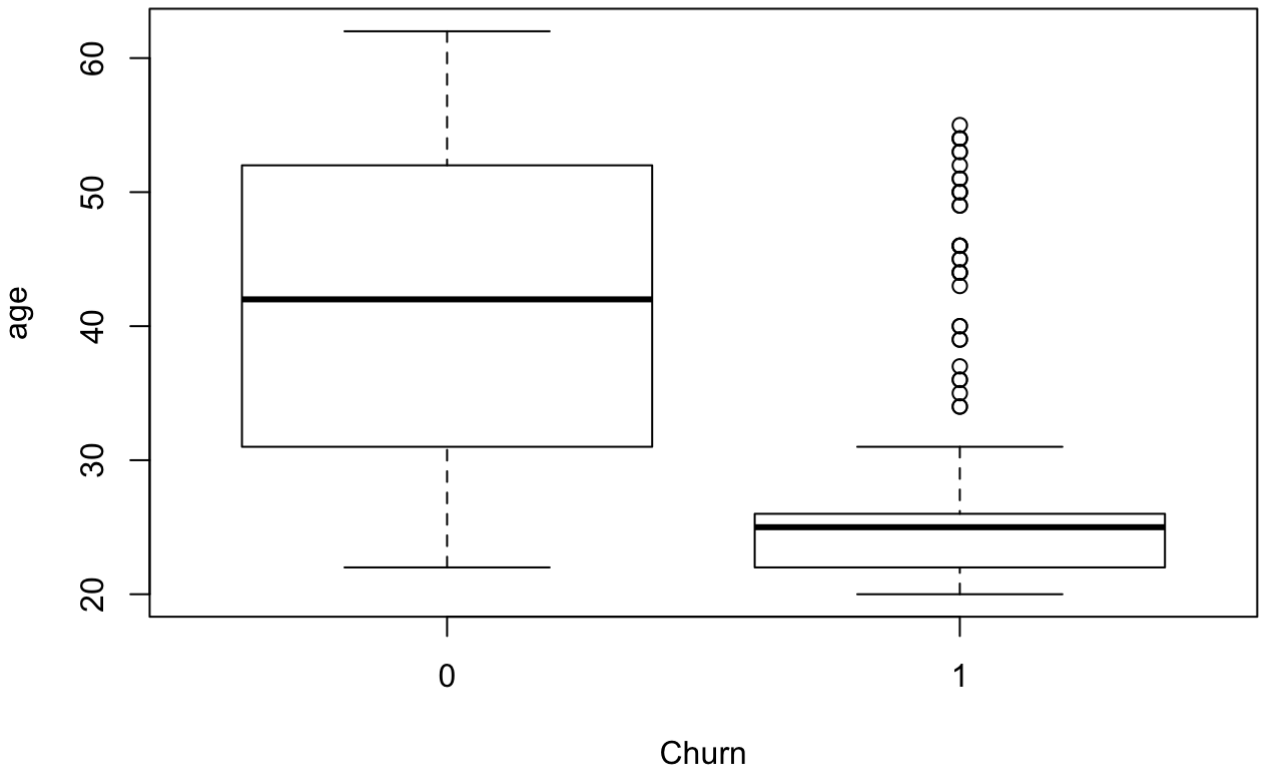
```
##   GENDER EDUCATION LAST_PRICE_PLAN_CHNG_DAY_CNT TOT_ACTV_SRV_CNT AGE
## 1      M         2                            0                1  36
##   PCT_CHNG_IB_SMS_CNT PCT_CHNG_BILL_AMT CHURN COMPLAINT
## 1           0.8421053         0.5709716     0         0
```

```
#create dummy variables for sector and status;
#gender (M,F)
myd$gender=(myd$GENDER=='M')*1
#Education (.,1,2,3,4,5,6) I ASSUME THE "."'s mean something in the EDUCATION column....
myd$ed1=(myd$EDUCATION==1)*1
myd$ed2=(myd$EDUCATION==2)*1
myd$ed3=(myd$EDUCATION==3)*1
myd$ed4=(myd$EDUCATION==4)*1
myd$ed5=(myd$EDUCATION==5)*1
myd$ed6=(myd$EDUCATION==6)*1
#Complaint(1,0)
#myd$complaint=(myd$COMPLAINT)
#Get variables
churn = myd$CHURN

lpcdaycount = myd$LAST_PRICE_PLAN_CHNG_DAY_CNT
totalactive = myd$TOT_ACTV_SRV_CNT
age = myd$AGE
percentchangeCNT = myd$PCT_CHNG_IB_SMS_CNT
percentchangeBILL = myd$PCT_CHNG_BILL_AMT
complaint = myd$COMPLAINT
gender = myd$gender
ed1 = myd$ed1
ed2 = myd$ed2
ed3 = myd$ed3
ed4 = myd$ed4
ed5 = myd$ed5
ed6 = myd$ed6
```

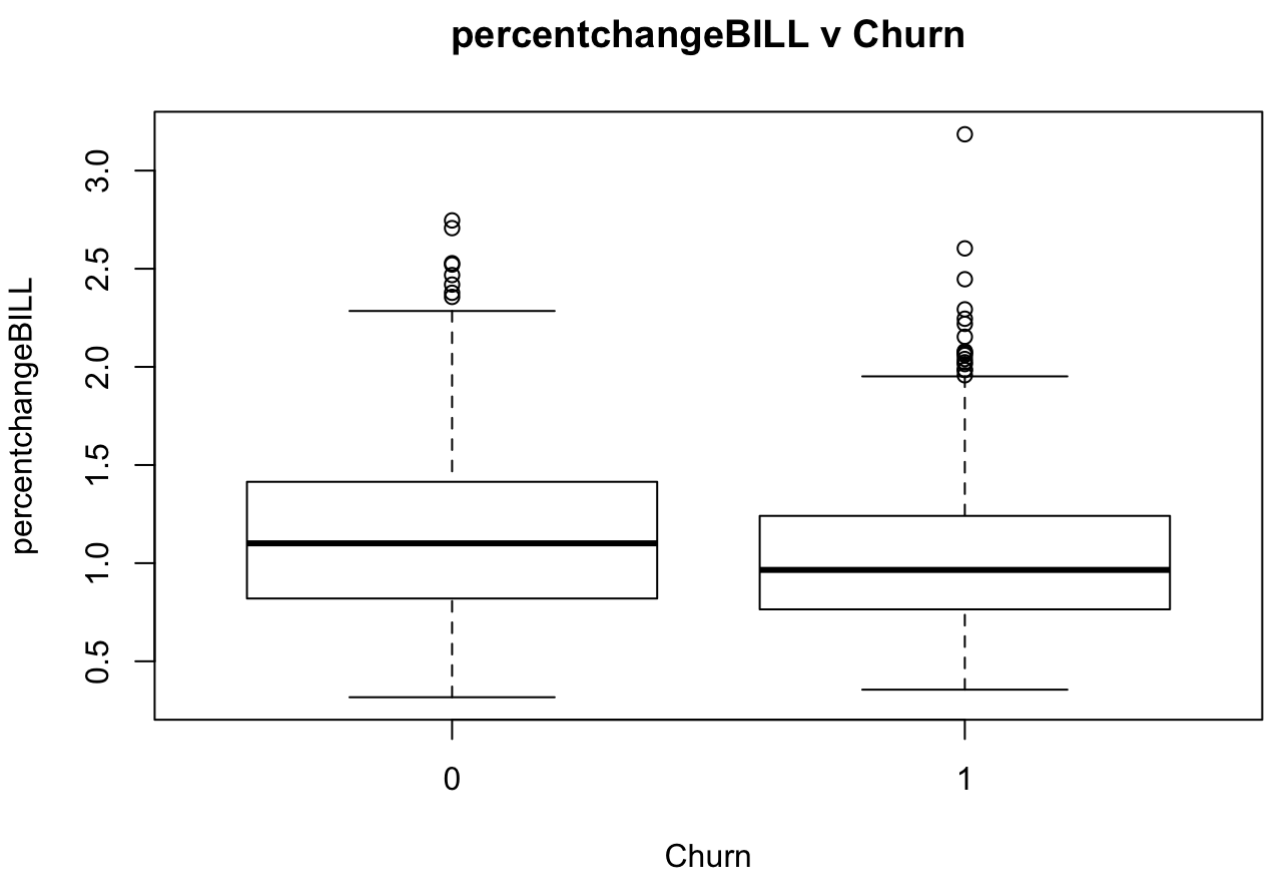
a) Create two boxplots to analyze the observed values of age and PCT_CHNG_BILL_AMT by churn value. Analyze the boxplots and discuss how customer age and changes in bill amount affect churn probabilities. [1 pt for R code for boxplots, 1 pt for analysis = 2 pts]

```
#age v churn
boxplot(age~churn,data=myd, main="age v Churn", xlab="Churn",ylab="age")
```



Answer: It would appear that younger customers are more likely to switch providers as indicated by the distribution of ages for churn probability 1. Most customers who switch providers are between the ages of 20 and 30. Customers over 30 are less likely to switch provider.

```
#percentchangeBILL v chrun
boxplot(percentchangeBILL~churn,data=myd, main="percentchangeBILL v Churn", xlab="Churn"
,ylab="percentchangeBILL")
```



Answer: This suggests that changes in a customer's bill amount actually have very little influence on whether or not a customer switches providers. This is indicated by the nearly identical distributions of percentchangeBILL for the two churn probabilities (0 or 1).

In summary, age appears to be a bigger factor in customer churn than percent change in bill amount.

Problem 1 b)

b) Fit a logistic regression model to predict the churn probability using the data in the dataset (Churn is the response variable and the remaining variables are the independent x-variables). Remove x-variables that are not significant using alpha=0.05. Write down the expression of the fitted model. (HINT: probability of interest is p = pr(churn = 1) [1 pt R code for model, 1 pt non-significant x-variables, 1 pt expression = 3 pts]

```
# logistic regression model fitted using glm() function with family=binomial
full_1 <- glm(churn~lpcdaycount + totalactive + age + percentchangeBILL + percentchangeC
NT + complaint + gender + ed1+ed2+ed3+ed4+ed5+ed6, data=myd, family=binomial())
summary(full_1) # display results
```

```
##
## Call:
## glm(formula = churn ~ lpcdaycount + totalactive + age + percentchangeBILL +
##     percentchangeCNT + complaint + gender + ed1 + ed2 + ed3 +
##     ed4 + ed5 + ed6, family = binomial(), data = myd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2097  -0.4284  -0.0700   0.5364   3.2293
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        6.80494    0.57512  11.832  < 2e-16 ***
## lpcdaycount        0.21604    0.56435   0.383  0.70186
## totalactive       -0.55313    0.06355  -8.704  < 2e-16 ***
## age               -0.17773    0.01272 -13.974  < 2e-16 ***
## percentchangeBILL -0.41560    0.22317  -1.862  0.06257 .
## percentchangeCNT  -0.39120    0.14421  -2.713  0.00667 **
## complaint          0.52043    0.22677   2.295  0.02174 *
## gender            -0.08550    0.20442  -0.418  0.67574
## ed1                0.48134    0.25057   1.921  0.05473 .
## ed2                0.36725    0.25193   1.458  0.14492
## ed3                0.80434    0.62350   1.290  0.19704
## ed4                1.17089    0.98969   1.183  0.23677
## ed5               12.88189  623.77735   0.021  0.98352
## ed6                1.09267    1.75600   0.622  0.53378
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1360.48  on 982  degrees of freedom
## Residual deviance:  714.69  on 969  degrees of freedom
## AIC: 742.69
##
## Number of Fisher Scoring iterations: 13
```

Answer: The following variables are insignificant at the 0.05 significance level: gender, education dummy variables(ed1,ed2,ed3,ed4,ed5,ed6), lpcdaycount, and also percentchangeBILL.

The following variabels are significant:totalactive, age, percentchangeCNT, and complaint

```
# logistic regression model with insignificant variables removed
fittest <- glm(churn ~ totalactive + age + percentchangeCNT  + complaint, data=myd, fami
ly=binomial())
summary(fittest) # display results
```

```
##
## Call:
## glm(formula = churn ~ totalactive + age + percentchangeCNT +
##     complaint, family = binomial(), data = myd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2892  -0.4282  -0.0730   0.5460   3.2765
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       6.72099    0.48041  13.990  < 2e-16 ***
## totalactive      -0.54745    0.06249  -8.760  < 2e-16 ***
## age              -0.17921    0.01275 -14.051  < 2e-16 ***
## percentchangeCNT -0.41796    0.14377  -2.907  0.00365 **
## complaint         0.50512    0.22278   2.267  0.02337 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1360.48  on 982  degrees of freedom
## Residual deviance:  724.17  on 978  degrees of freedom
## AIC: 734.17
##
## Number of Fisher Scoring iterations: 6
```

Answer:

Expression: CHURN = 6.72099 - 0.54745(totalactive) - 0.17921(age) - 0.41796(percentchangeCNT) + 0.50512(complaint)

```
# DOUBLE CHECK*** backward selection procedure for variable selection
step(full_1, direction=c("backward"), alpha=0.05, trace=F)
```

```
##
## Call:  glm(formula = churn ~ totalactive + age + percentchangeBILL +
##     percentchangeCNT + complaint, family = binomial(), data = myd)
##
## Coefficients:
##       (Intercept)        totalactive                age
##            7.1140            -0.5489            -0.1778
## percentchangeBILL   percentchangeCNT          complaint
##           -0.3991            -0.4123             0.5049
##
## Degrees of Freedom: 982 Total (i.e. Null);  977 Residual
## Null Deviance:       1360
## Residual Deviance: 720.9     AIC: 732.9
```

```
# logistic regression model with insignificant variables removed
fit <- glm(churn~ totalactive + age + percentchangeBILL + percentchangeCNT + complaint,
 data=myd, family=binomial())
summary(fit) # display results
```

```
##
## Call:
## glm(formula = churn ~ totalactive + age + percentchangeBILL +
##     percentchangeCNT + complaint, family = binomial(), data = myd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2631  -0.4345  -0.0717   0.5555   3.2691
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        7.11401    0.53506  13.296   <2e-16 ***
## totalactive       -0.54892    0.06282  -8.738   <2e-16 ***
## age               -0.17781    0.01270 -14.002   <2e-16 ***
## percentchangeBILL -0.39914    0.22057  -1.810   0.0704 .
## percentchangeCNT  -0.41230    0.14403  -2.863   0.0042 **
## complaint          0.50489    0.22283   2.266   0.0235 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##

## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1360.48  on 982  degrees of freedom
## Residual deviance:  720.87  on 977  degrees of freedom
## AIC: 732.87
##
## Number of Fisher Scoring iterations: 6
```

Answer: The coefficients/variables are different when doing backward selection, indicating that the I should not have removed percentchangeBILL from my model. I will include it now, as the alpha=0.05 setting confirms that it will be signifance at the level we desire.

__***FINAL Expression: CHURN = 7.1140 - 0.5489(totalactive) - 0.1778(age) - 0.3991(percentchangeBILL) - 0.4123(percentchangeCNT) + 0.5049(complaint)__

Problem 1 c)

c) Analyze the final logistic regression model and discuss the effect of each variable on the churn probability. Discuss results in terms of odds ratios. [1 pt residual plot, 1 pt odds ratio, 1 pt discussion = 3 pts]

```
library(lmtest)
```

```
## Loading required package: zoo
```

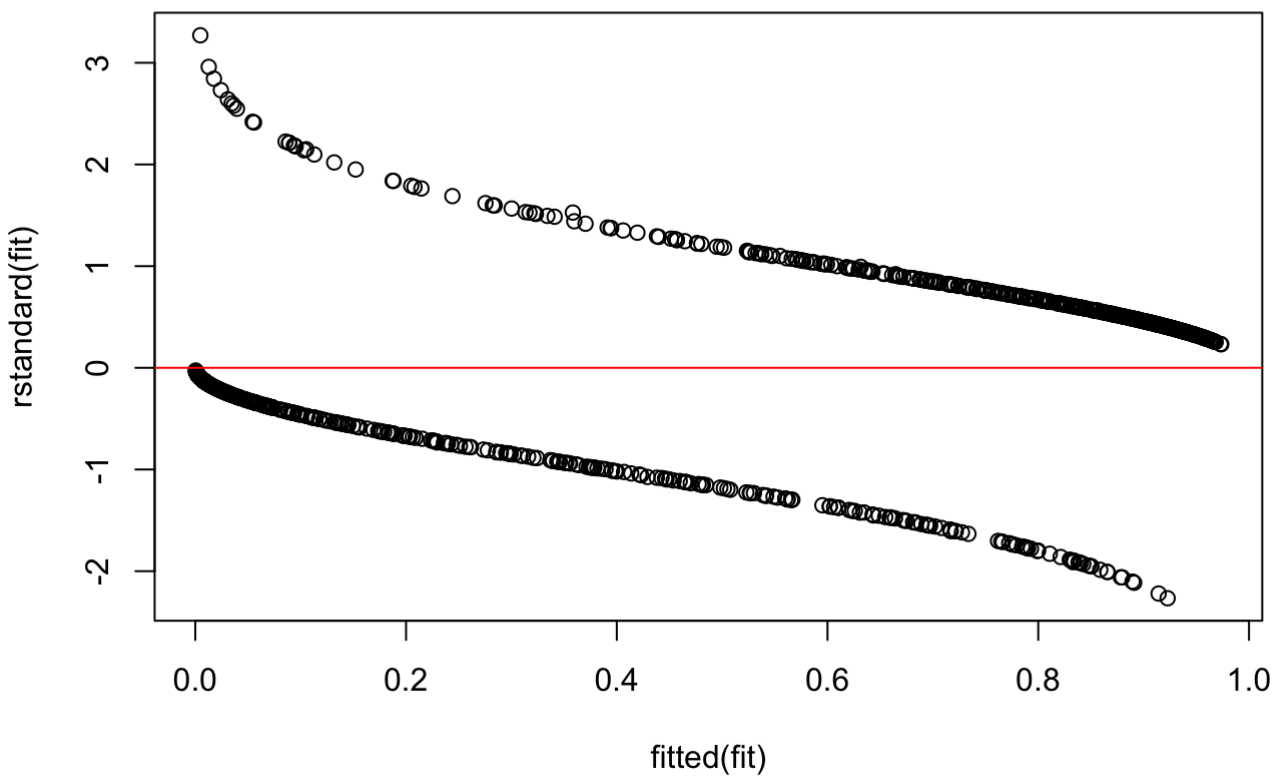
```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

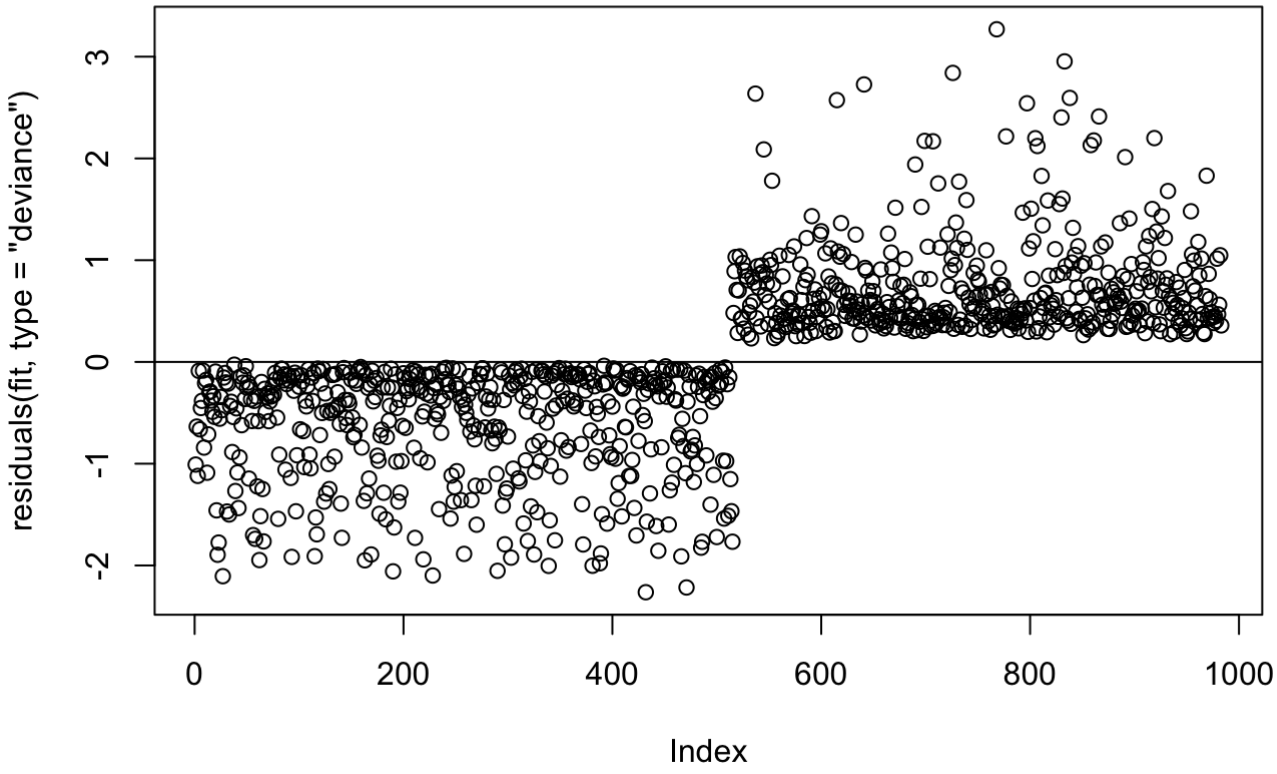
```
lrtest(fit)
```

```
## Likelihood ratio test
##
## Model 1: churn ~ totalactive + age + percentchangeBILL + percentchangeCNT +
##     complaint
## Model 2: churn ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -360.43
## 2   1 -680.24 -5 639.61  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#residual plots
#Plot residuals vs predicted values
plot( fitted(fit), rstandard(fit), main="Predicted vs Residuals plot") #one way to inter
pret
abline(a=0, b=0, col='red') #add zero line
```




```
plot(residuals(fit,type="deviance")) #another interpretation for logistic regression
abline(a=0, b=0)
```




```
confint(fit) # 95% CI for the coefficients
```

```
## Waiting for profiling to be done...
```

```
##                          2.5 %      97.5 %
## (Intercept)         6.09966918  8.19977779
## totalactive        -0.67490903 -0.42833722
## age                -0.20390711 -0.15404370
## percentchangeBILL -0.83456268  0.03125915
## percentchangeCNT  -0.69783228 -0.13041254
## complaint           0.06795635  0.94269152
```

```
exp(coef(fit)) # compute exp(coefficients) to analyze change in odds for changes in X
```

```
##       (Intercept)       totalactive               age percentchangeBILL
##      1229.0707704         0.5775714         0.8370999         0.6708988
##  percentchangeCNT         complaint
##         0.6621248         1.6567979
```

```
exp(coef(fit))-1 # compute exp(coefficients) to analyze change in odds for changes in X
```

```
##       (Intercept)       totalactive               age percentchangeBILL
##      1228.0707704        -0.4224286        -0.1629001        -0.3291012
##  percentchangeCNT         complaint
##        -0.3378752         0.6567979
```

```
#exp(confint(fit)) # 95% CI for exp(coefficients), that is change in odds
#predict(fit, type="response") # predicted values
#residuals(fit, type="deviance") # residuals
```

Answer:

residuals: Based on the residuals plot you can see a fairly normal distribution for both churn=1 and churn=0, as expected.

1) totalactive: The odds p/(1-p) of churn **decrease by about 42%**, for each additional service the customer has. This means customers with more services are more locked in to the provider and less likely to switch.

2) age: The odds p/(1-p) of churn **decrease by about 16%**, for each additional year older a customer is. This means that older customers are more likely to stay with the same provider and less likely to switch (churn)

3) percentchangeBILL: The odds p/(1-p) of churn **decrease by about 33%**, for each additional percent change of the latest 2 month bill amount with respect to the previous 4 months bill amount.

4) percentchangeCNT: The odds p/(1-p) of churn **decrease by about 34%**, for each additional percent change of the latest 2 months incoming SMS with respect to the previous 4 months incoming SMS. This means that a customer who is receiving more SMS messages is less likely to churn and switch providers.

5) complaint: The odds p/(1-p) of churn **increase by about 66%**, if there was a customer complaint in the last two months. No surprise here, if a customer is complaining then they're more likely to switch providers.

Problem 1 d)

d) Compute the predicted churn probability and the prediction interval for a male customer who is 43 years old, and has the following information LAST_PRICE_PLAN_CHNG_DAY_CNT=0, TOT_ACTV_SRV_CN=4, PCT_CHNG_IB_SMS_CNT= 1.04, PCT_CHNG_BILL_AMT= 1.19, and COMPLAINT =1. [1 pt computing predicted probability, 1 pt prediction interval = 2 pts]

```
newd = data.frame(age=c(43), lpcdaycount=c(0), totalactive=c(4), percentchangeCNT=c(1.04
), percentchangeBILL=c(1.19), complaint=c(1))
predict(fit,newdata=newd,type="response", se.fit=T) #type="response" for probabilities
```

```
## $fit
##          1
## 0.04202857
##
## $se.fit
##          1
## 0.01015331
##
## $residual.scale
## [1] 1
```

```
pr=predict(fit, type="response") # predicted values
```

Answer:

The predicted probability for a 20 year old individual living in sector 2 of the city is computed as p_hat = 0.042 with standard error of 0.01. Therefore the 95% prediction interval is 0.042 +/- 1.96*0.01 or (0.022,0.062).

This suggests that the churn probability is very low, between about 2% and 6% with estimate of 4%. This makes sense from the first look done at the boxplot showing older individuals less frequently churning.

Problem 1 e)

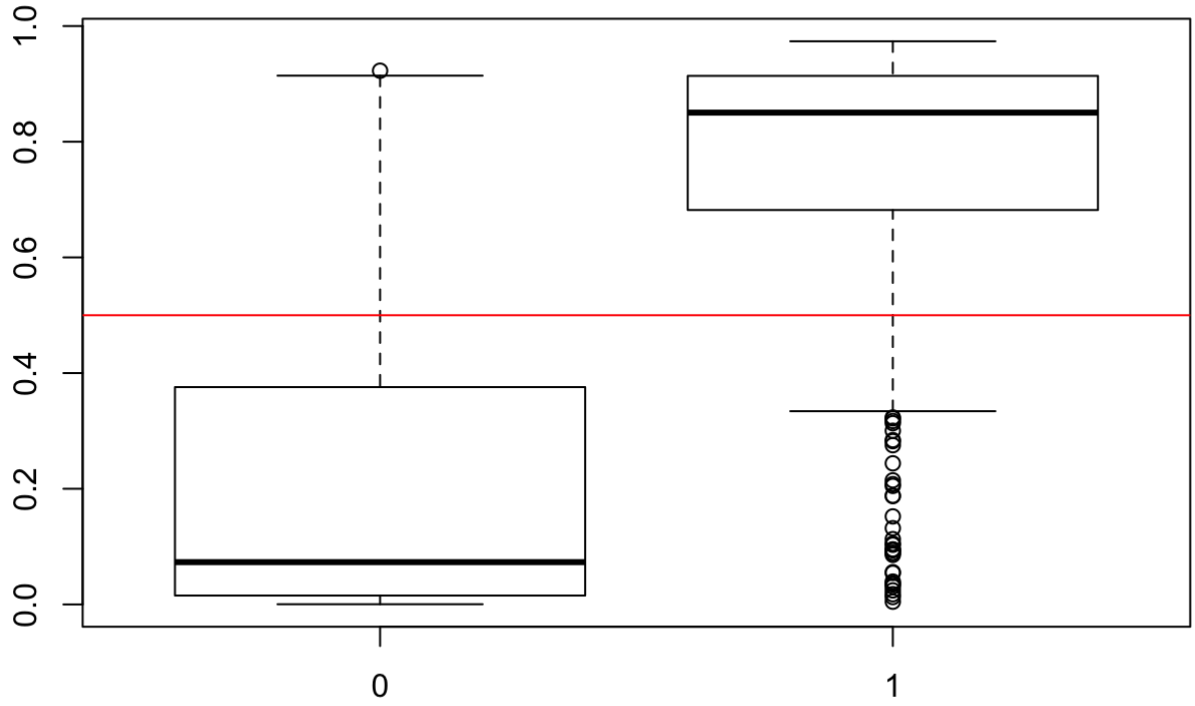
e) OPTIONAL CHALLENGE - ONLY FOR GRADUATE STUDENTS: The dataset churn_test.csv contains a new set of customers, and can be used to test the validity of the churn predictive model. Apply the methods discussed in week 9 lecture to identify a threshold T for the predicted churn probability in order to define a classification rule for customers, so that: - predicted probability p(churn) > = T, then customer is a "likely churn", and - predicted probability p(churn) < T, then customer is a "unlikely churn". Compute the optimal T value, and create the classification matrix summarizing classification results. Hint: You can use the Classify_functions.R in your solution. [1 pt for R code for getting range of thresholds to choose optimal T, 1 pt for computing predicted churn outcomes corresponding to predicted probabilities, 1 pt for computing confusion matrix and metrics, 1 pt for selecting the P* that optimizes a certain metric, 1 pt for computing predicted churn outcomes for a separate testing set (churn_test.csv), 1 pt for computing the confusion matrix that summarizes classification results and metrics = 6 pts]

```
#1 pt for R code for getting range of thresholds to choose optimal T
#1 pt for computing predicted churn outcomes corresponding to predicted probabilities
#1 pt for computing confusion matrix and metrics
#1 pt for selecting the P* that optimizes a certain metric
#1 pt for computing predicted churn outcomes for a separate testing set (churn_test.csv)
#1 pt for computing the confusion matrix that summarizes classification results and metr
ics

# load in the data from file
mydtest=read.csv("/Users/alexteboul/Desktop/churn_test.csv", header=T)
# CHURN = 7.1140 - 0.5489(totalactive) - 0.1778(age) - 0.3991(percentchangeBILL) - 0.412
3(percentchangeCNT) + 0.5049(complaint)__
#Get variables
CHURN = mydtest$CHURN
TOT_ACTV_SRV_CNT = mydtest$TOT_ACTV_SRV_CNT
AGE = mydtest$AGE
PCT_CHNG_IB_SMS_CNT = mydtest$PCT_CHNG_IB_SMS_CNT
PCT_CHNG_BILL_AMT = mydtest$PCT_CHNG_BILL_AMT
COMPLAINT = mydtest$COMPLAINT
```

Answer: The data is now loaded in…

```
#1 pt for R code for getting range of thresholds to choose optimal T
#boxplot of predicted probabilities by task success
# useful visualization for classification purposes
boxplot(fitted(fit)~churn, data = myd, names=c("0", "1"))
abline(a=0.5, b=0,col='red') #threshold line
```



Answer: Based on this boxplot, I'm going to say the threshold is at about 0.5 because that splits well between the two. No overlap.

```
M_fit <- glm(myd$CHURN~ myd$TOT_ACTV_SRV_CNT + myd$AGE + myd$PCT_CHNG_BILL_AMT + myd$PCT
_CHNG_IB_SMS_CNT + myd$COMPLAINT, data=mydtest, family=binomial())

#1 pt for computing predicted churn outcomes corresponding to predicted probabilities
source("Classify_functions1.R")

#compute the predicted outcome based on probability threshold 0.5
y.test = mydtest$CHURN
#predicted outcomes in testing set:
predz=as.vector(predict(M_fit,mydtest, type="response"))
```

```
## Warning: 'newdata' had 98 rows but variables found have 983 rows
```

```
ypred = classify(predz,0.5)
ypred
```

```
##   [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 1 0 0 0 1 0 1 0 0
##  [36] 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 1 0 1 1 0 0 0 0
##  [71] 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0
## [106] 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 1
## [141] 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 1 0 0 0 0 0 0
## [176] 0 1 0 0 0 1 0 1 0 0 0 0 0 0 1 1 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [211] 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1
## [246] 0 0 1 1 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0
## [281] 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1
## [316] 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0 0 0
## [351] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 0 0 0 0
## [386] 0 0 1 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
## [421] 1 0 1 0 0 0 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 1 1 1
## [456] 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0
## [491] 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1
## [526] 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1
## [561] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1 1 1 1
## [596] 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1
## [631] 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1
## [666] 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 0 1 1 0 1
## [701] 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1 0 1 1 0 1 1 1
## [736] 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1
## [771] 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 1 1 1 0 1 0 1 0
## [806] 1 0 1 1 1 1 0 0 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1 0 0 1 0 1 1 1 1 1 0 1 1
## [841] 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1 1
## [876] 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [911] 1 1 1 0 1 1 0 1 0 1 0 1 1 1 1 1 0 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1
## [946] 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
## [981] 1 1 1
```

```
#m=compare(ypred,y.test)
```

Answer:

'newdata' had 98 rows but variables found have 983 rowsError in if (pred == yvar[i]) if (yvar[i] == 1) tp = tp + 1 else tn = tn + : missing value where TRUE/FALSE needed

Problem 2 FOR EXTRA CREDIT [11 pts]

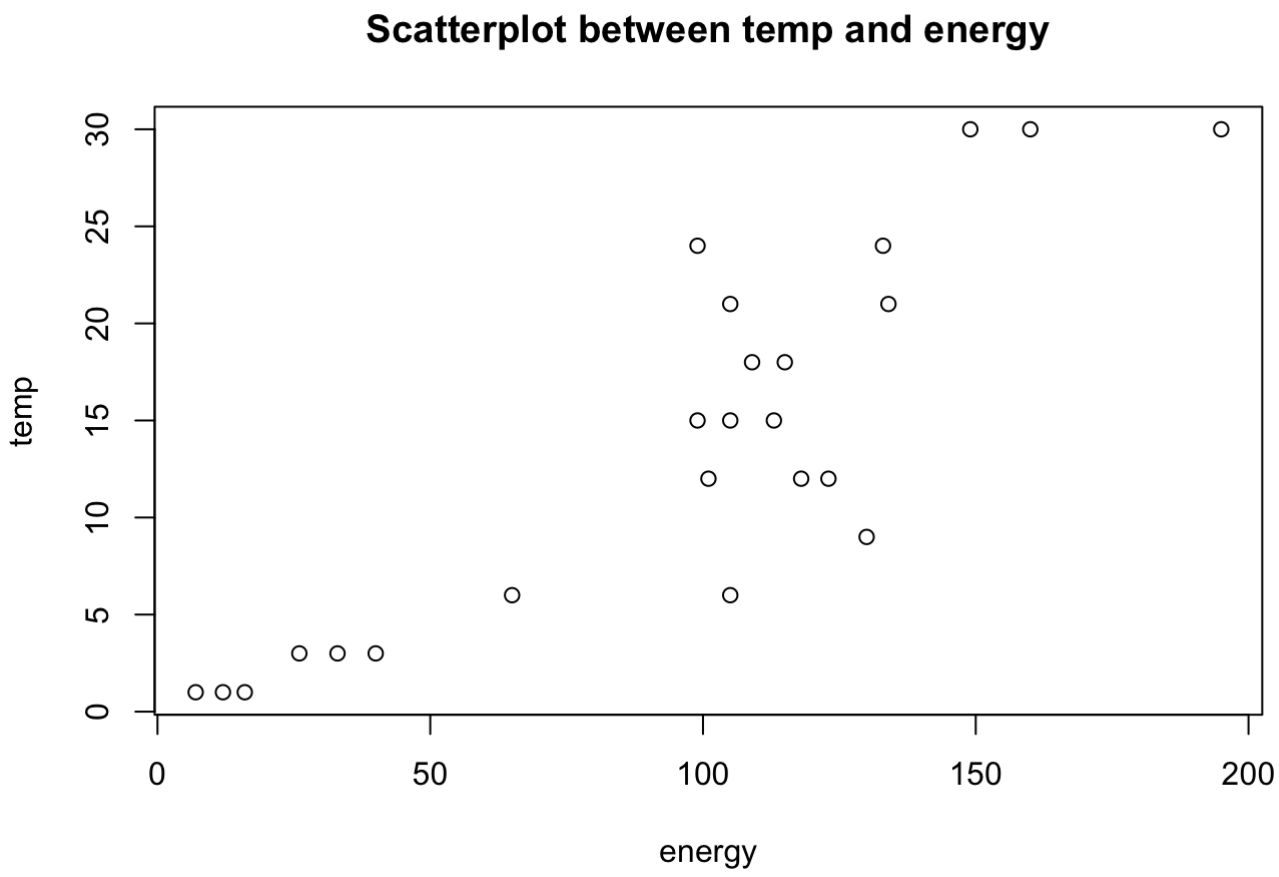
A researcher is interested in evaluating the relationship between energy consumption by the homeowner and the difference between the internal and external temperatures. A sample of 30 homes was used in the study. During an extended period of time, the average temperature difference (in o F) (TEMPD) inside and outside the homes was recorded. The average energy consumption (ENERGY) was also recorded for each home. The data are stored in the energytemp.txt data file.

Problem 2 a)

a) Create a scatterplot of ENERGY (y) versus TEMPD (x) to visualize the association between the two variables. Analyze the association displayed by the scatterplot. [1 pt for scatterplot, 1 pt for analysis = 2 pts]

```
## load in the data from file
myd2=read.table("energytemp.txt", header=T, sep="\t")
## get the variables
energy=myd2$energy
temp=myd2$temp

# scatterplot between energy and temp
plot(energy, temp, main="Scatterplot between temp and energy", xlab="energy", ylab="tem
p")
```



Answer: The scatterplot appears to indicate that there is a positive relationship between energy consumption in a home and the magnitude of the difference in temperatures between the inside and outside of a home (energy and temp). This makes sense because when it's really cold out a home make consume more energy to keep residents warm, so high energy and high temp value. There is not sufficient evidence to conclude that the relationship is linear though, just positive.

Problem 2 b)

b) Fit a cubic model (HINT: create two new variables TEMP2 and TEMP3: y= B0 + B1x + B2x^2 + B3x^3 + e. In R use the code: tempd2 = tempd^2; tempd3 = tempd^3; Include the new variables in the regression model) [1 pt]

```
#new variables
temp2 = temp^2
temp3 = temp^3

fitextra <- lm(energy ~ temp + temp2 + temp3, data = myd2 )
summary(fitextra)
```

```
##
## Call:
## lm(formula = energy ~ temp + temp2 + temp3, data = myd2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.159 -11.257  -2.377   9.784  26.841
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.036232  10.115284  -1.684    0.108
## temp         24.523999   3.371636   7.274 4.91e-07 ***
## temp2        -1.490029   0.266166  -5.598 1.77e-05 ***
## temp3         0.029278   0.005643   5.188 4.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.73 on 20 degrees of freedom
## Multiple R-squared:  0.9137, Adjusted R-squared:  0.9008
## F-statistic: 70.62 on 3 and 20 DF,  p-value: 8.105e-11
```

```
#other method
#f2 <- lm(energy ~ poly(temp,3))
#summary(f2)
```

Answer: Expression: energy = -17.036232 + 24.523999(temp) - 1.490029(temp2) + 0.029278(temp3) + e

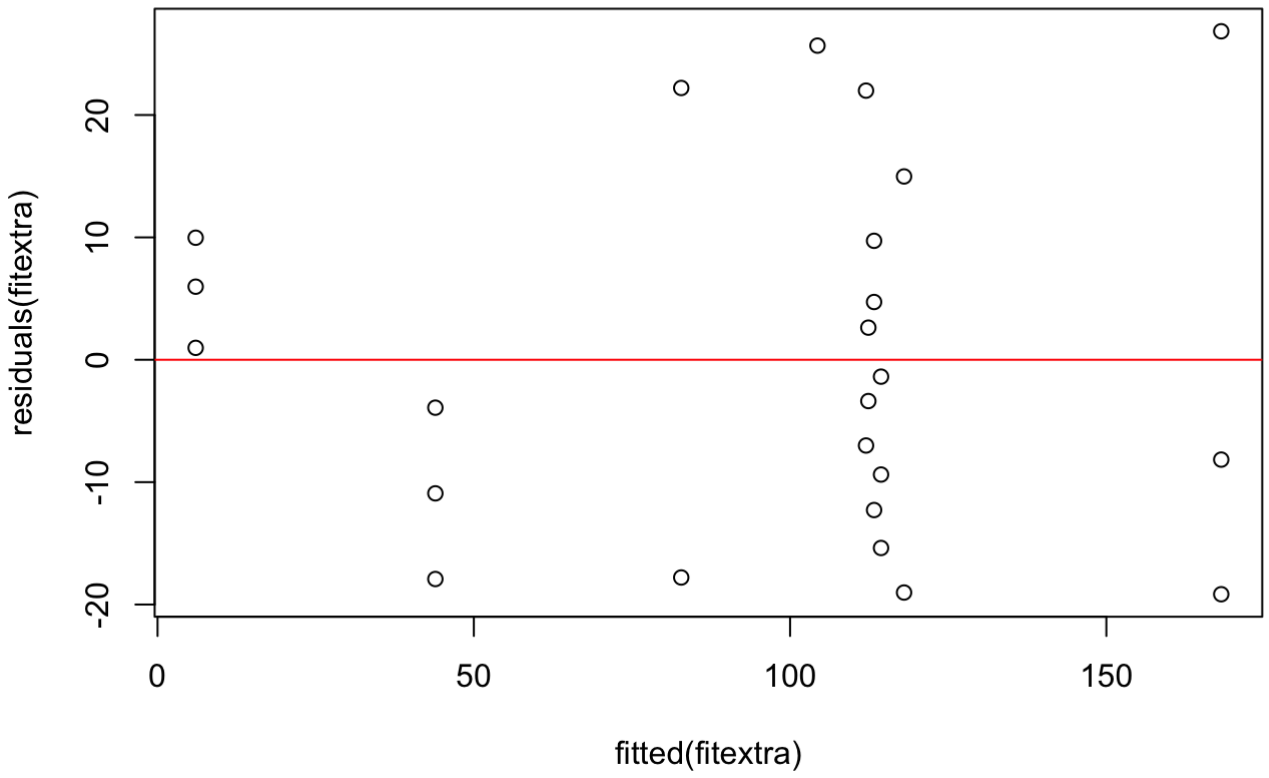
Problem 2 c)

c) Are all variables in the model significant? [1 pt] Answer: Yes all the variables in the model are significant. temp, temp2, and temp3 are significant at at least the 0.05 significance level so they're good to keep.

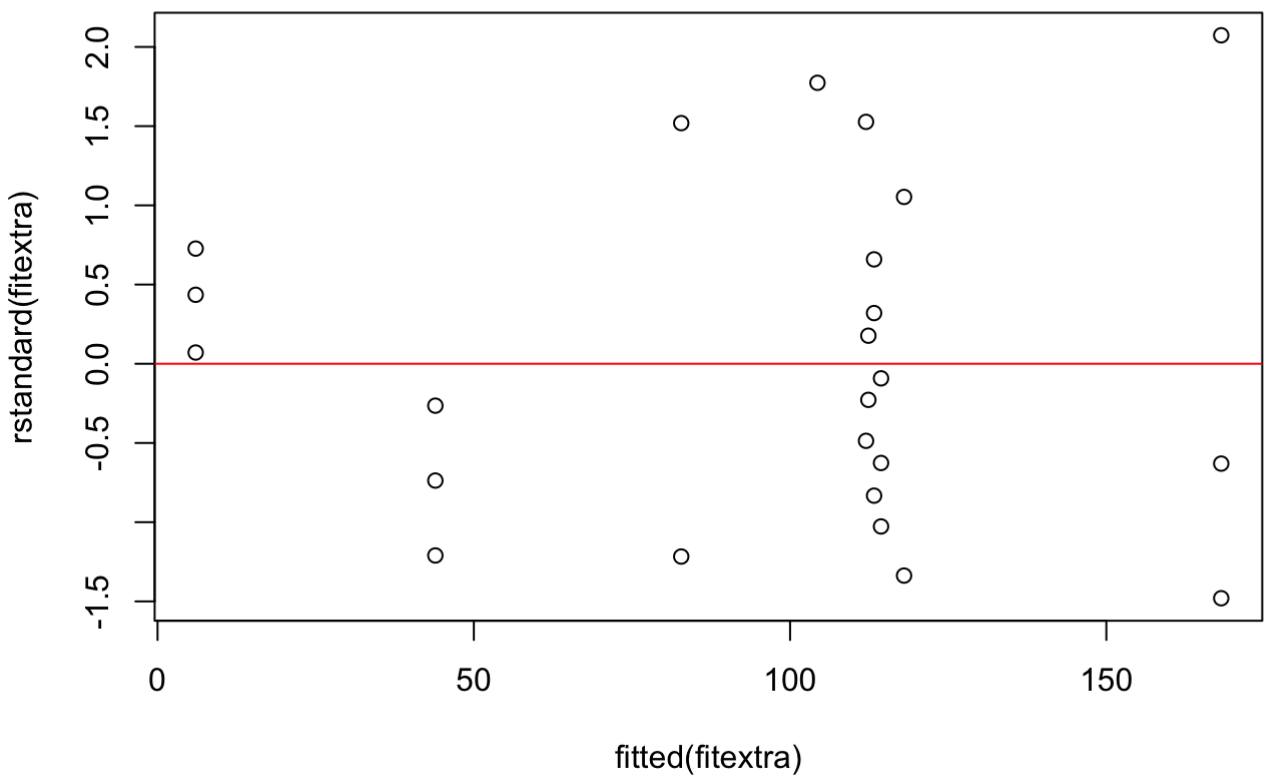
Part 2 d)

d) Create the residual plots (residuals vs predicted; residuals vs x variable; and normal plot of residuals). Analyze residual plots to evaluate the normality and constant variance assumptions. Discuss your findings. [3 pts for residual plots, 1 pt for analysis = 4 pts]

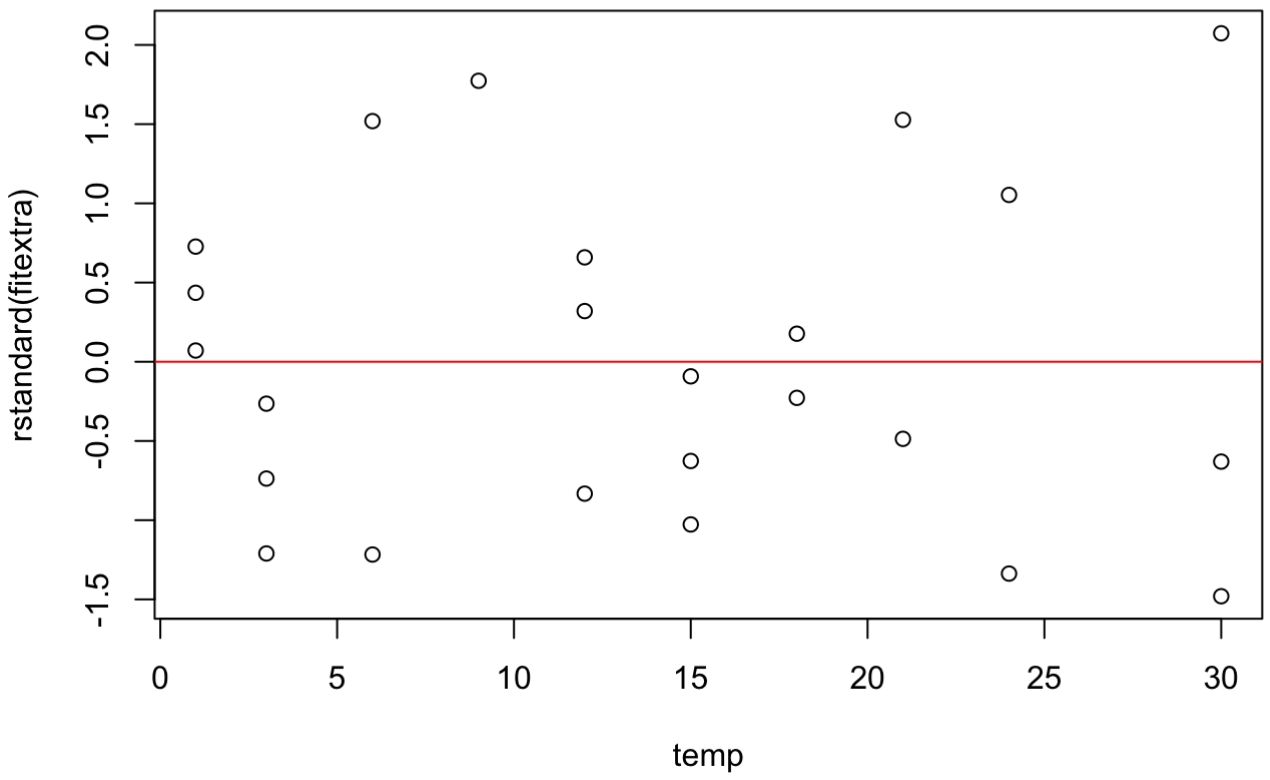
```
#residuals vs predicted
plot(fitted(fitextra),residuals(fitextra))
abline(a=0, b=0, col='red') #add zero line
```



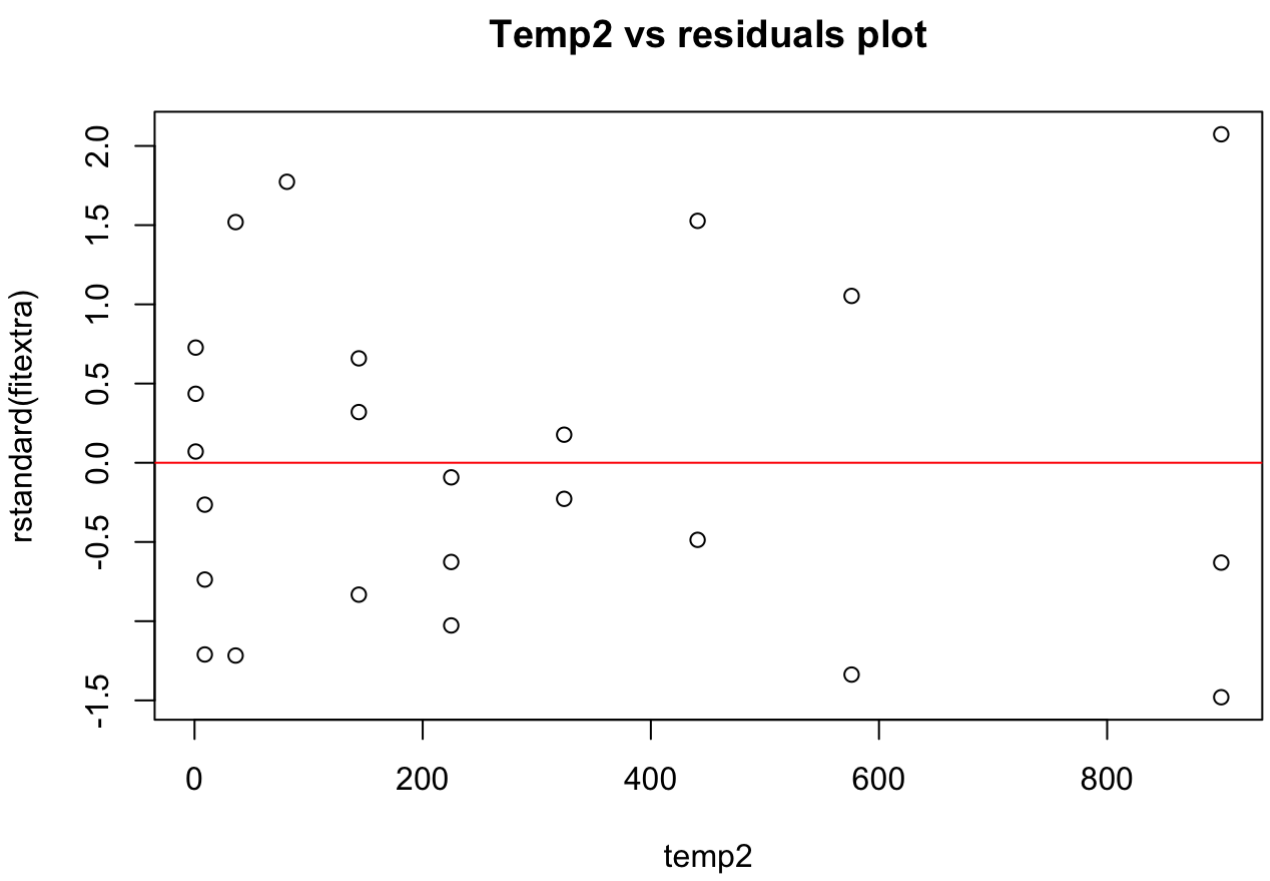

```
#residual plots
#Plot residuals vs predicted values
plot( fitted(fitextra), rstandard(fitextra), main="Predicted vs Residuals plot")
abline(a=0, b=0, col='red') #add zero line
```



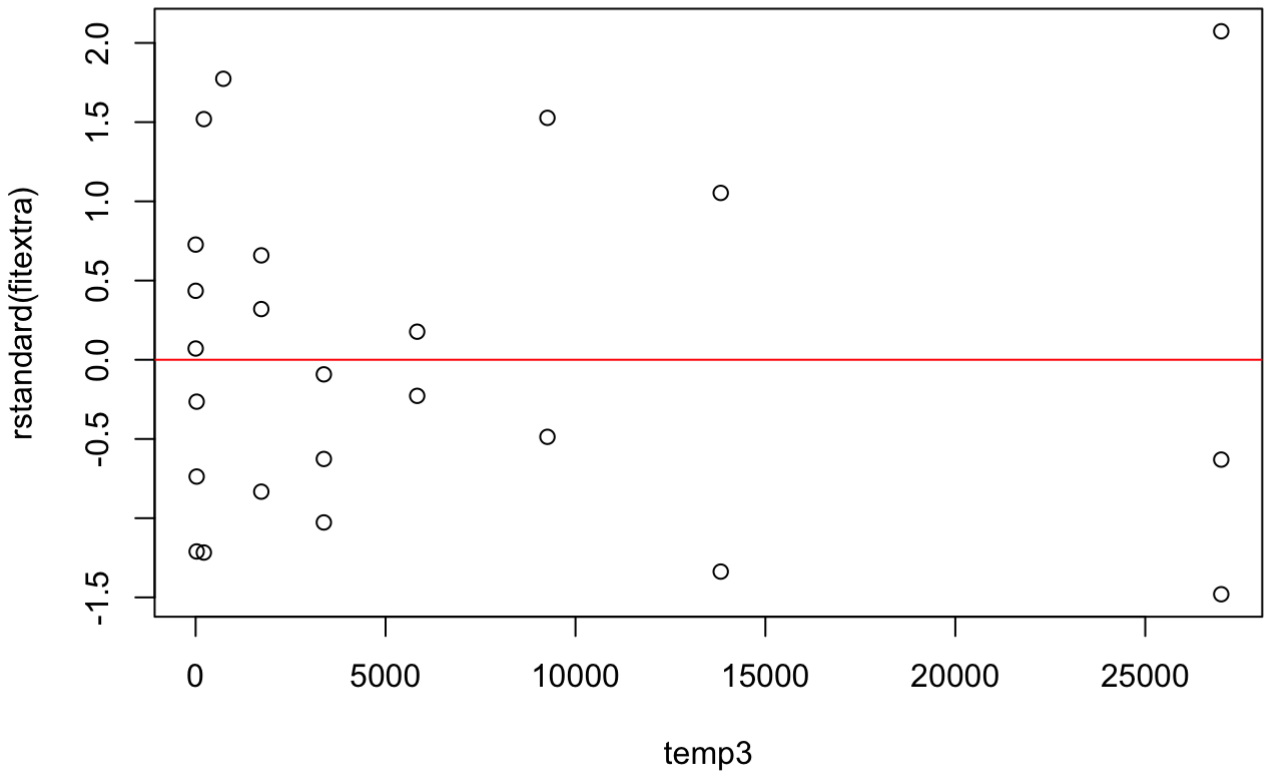

```
#residuals vs x variable
#Plot residuals vs temp:
plot(temp, rstandard(fitextra), main="Temp vs residuals plot")
abline(a=0, b=0,col='red')
```



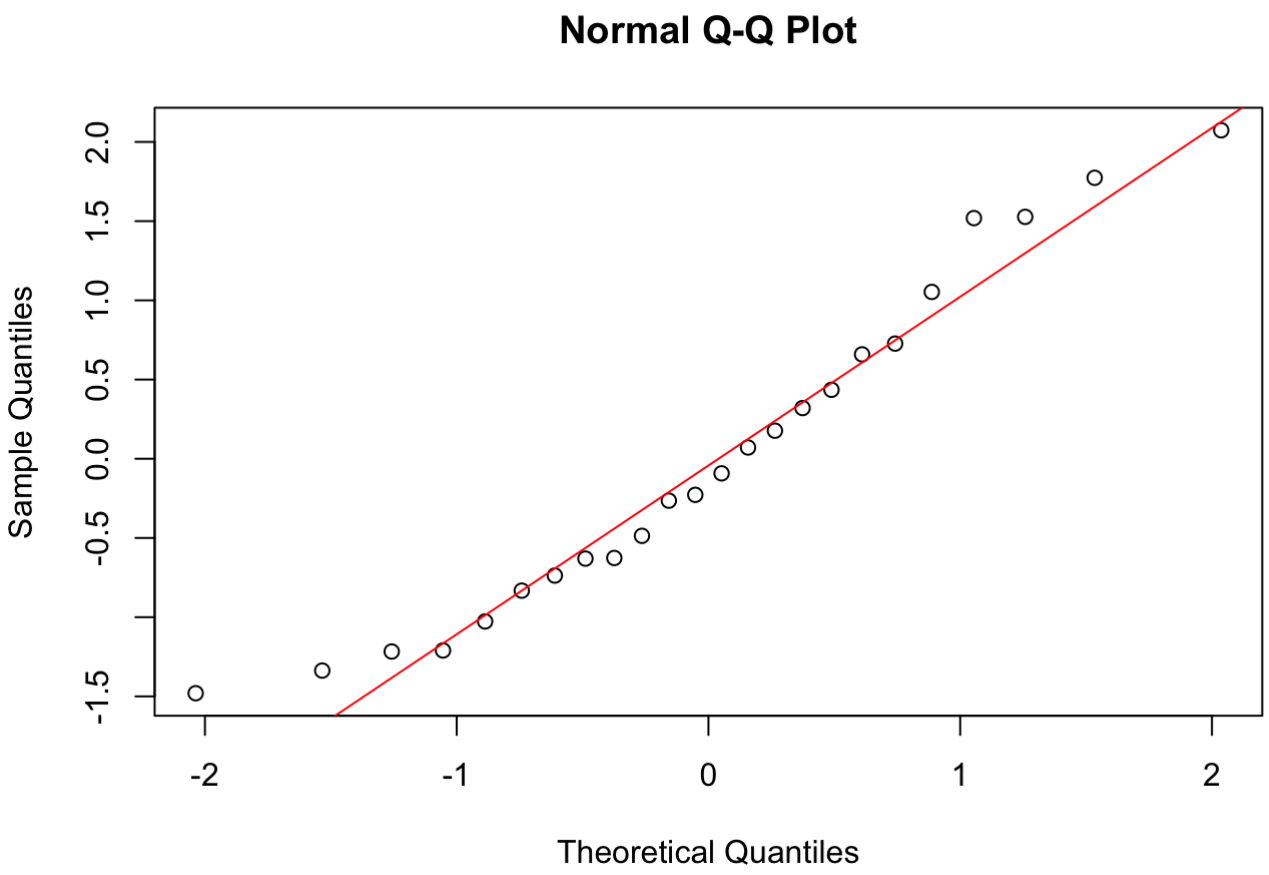

```
#Plot residuals vs temp2:
plot(temp2, rstandard(fitextra), main="Temp2 vs residuals plot")
abline(a=0, b=0,col='red')
```




```
#Plot residuals vs each temp3:
plot(temp3, rstandard(fitextra), main="Temp3 vs residuals plot")
abline(a=0, b=0,col='red')
```




```
#normal plot of residuals
#normal probability plot of residuals
qqnorm(rstandard(fitextra))
qqline(rstandard(fitextra), col = 2)
```



Answer: Residuals – Based on the residual plots, I do believe the model assumptions are met by the data. Standardized residuals vs predicted: The first plot, predicted v. residuals, does not appear to show a random scatter. That said, there are only 30 data points. The other residual plot v each variable appear fairly randomly scattered at least.

NormalQQplot – Good, points close to line indicating normal distribution of errors. A few potentially influential points, but nothing too bad considering such a small dataset.

Problem 2 e)

e) If you are satisfied with the fitted regression model, write down its expression. [1 pt] Answer:
Expression: energy = -17.036232 + 24.523999(temp) - 1.490029(temp2) + 0.029278(temp3) + e

Problem 2 f)

f) Use the fitted regression model to predict the average energy consumption for an average difference in temperature equal to TEMPD=10. (HINT: In R use the following code: new <- data.frame(tempd=c(10), tempd2=c(100), tempd3=c(1000)) then use the predict() function with the fitted regression model as explained in the document under week 5. [1 pt for R code, 1 pt for answer = 2 pts]

```
new <- data.frame(temp=c(10), temp2=c(100), temp3=c(1000))
# compute average response value and confidence interval
predict(fitextra, new, interval="confidence",level=0.95)
```

```
##        fit      lwr      upr
## 1 108.4787 95.96589 120.9915
```

Answer: So the average energy consumption for those values would be 108.48 units based on my model. The lower bound is 95.97 and upper bound is 120.99 for the 95% confidence interval. These values make sense when referencing the scatterplof of energy v temp.

“Reflection” Problem [2 pts]

Post a message in the “Reflection for Assignment 4” thread on the discussion board indicating which question in this assignment you found to be the easiest, the one you found to be the hardest, and why. Alex Teboul Answer:

Easiest Question: Extra credit was generally easiest. Part a with the scatterplot was specifically the easiest question because it did not involve much deep thinking or more in depth coding. Everything about EC was straightforward.

Hardest Question: Hardest question was Part e of Problem 1 d, because we didn’t really cover how to do problems of that type in class. Didn’t know how to find the Threshold and move from there. Overall good assignment though