Question 1:

1. Code from R:

temp1<- glm(formula=y~1, data=ds1, family=binomial(logit))

summary(temp1)

BIC(temp1)

AIC number is 3233

1. BIC number is 3239.417
2. Code from R:

summary(ds1$y)

log(521/4000)

The log number of the odds is the same with the intercept of the model. The number is the same because we use only the y column for the model and value in y column is binary (yes or no) only. That’s why the intercept and the log number are equal to each other.

Question 2:

1. Code from R:

temp2<- glm(formula=y~duration, data=ds1, family=binomial(logit))

summary(temp2)

The AIC number reduces 527.2 to 2705.8

1. Duration is a significant predictor for this model. It reduces the number of AIC.
2. With growing duration of the last contact, clients are more likely to subscribe to a term deposit because thee coefficient is 0.0035496 (positive coefficient)
3. Code in R:

exp(-3.255)/(1+exp(-3.255))

When the duration value = 0 then the estimated probability of subscribing to a term deposit for a client is 0.03714763 = 3.7%.

Question 3:

1. The halfway point is 917.268 where:

0.5 = odds/(1+odds)

-3.2559346 + 0.0035496\*duration = log(odds)

=> duration = 917.268.

1. The the slope of the tangent to the regression curve at the halfway point:

917.268\*0.5\*0.5 = 229.317

Question 4:

1. Code in R

temp3<- glm(formula=y~duration+age+marital, data=ds1, family=binomial(logit))

summary(temp3)

ds11<-data.frame(duration=mean(ds1$duration, na.rm=TRUE), age=mean(ds1$age, na.rm=TRUE), marital="married")

predict(temp3, ds11, type="response")

The predictive probability of subscribing to a term deposit for a married client at the mean values of the numeric predictors is: 0.07335345.

1. Code in R:

ds12<-data.frame(duration=300, age=mean(ds1$age, na.rm=TRUE), marital="married")

predict(temp3, ds12, type="response")

The predictive probability of subscribing to a term deposit changes into: 0.08249455

1. Code in R:

ds13<-data.frame(duration=mean(ds1$duration, na.rm=TRUE), age=(mean(ds1$age, na.rm=TRUE)+sd(ds1$age)), marital="married")

predict(temp3, ds13, type="response")

The predictive probability of subscribing to a term deposit changes into: 0.09107385

Question 5:

Code in R:

temp4<- glm(formula=y~duration+campaign+duration\*campaign, data=ds1, family=binomial(logit))

summary(temp4)

The regression coefficients are very small and all the predictors are relevant toward the regression model. All the coefficients look normal with relatively small z-values.

Question 6:

Code in R:

ds2<-ds1

ds2$durc<-ds2$duration-mean(ds2$duration)

ds2$camc<-ds2$duration-mean(ds2$campaign)

temp5<- glm(formula=y~durc+camc+durc\*camc, data=ds2, family=binomial(logit))

summary(temp5)

After centering the predictors, the model has changed slightly but not very significant. The AIC is still the same, the z-values are “almost” identical. The coefficients are relevant for interpretation.

The plot i couldn’t make it because i could not import the library effect into my R version.

Question 7:

Code in R:

temp6<- glm(formula=y~.-day-month, data=ds1, family=binomial(logit))

summary(temp6)

The significant predictors are: jobblue-collar, jobretired, jobunemployed, housingyes, loanyes, contactunknown, duration, campaign, poutcomesuccess, poutcomeother.

The list of significant predictors title doesn’t make much sense! But the AIC number reduces to 2339.2 which is the lowest across all the model we have built so far. The estimated coefficients look normal.

Question 8:

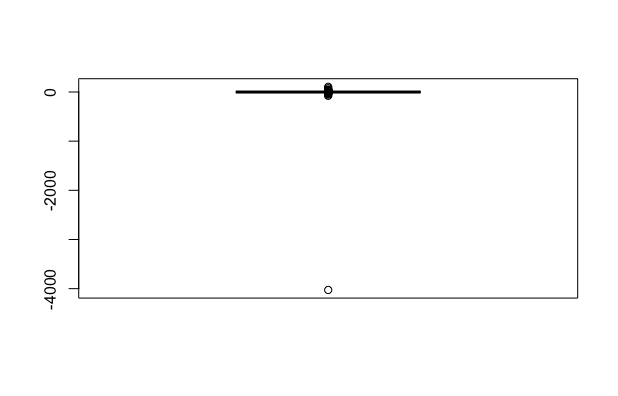
The significant predictors are: jobblue-collar, jobretired, jobunemployed, maritalmarried, housingyes, loanyes, contactunknown, duration, campaign, poutcomesuccess, poutcomeother.

The AIC score of the resulting model is: 2330.4

Question 9:

Code in R:

boxplot(temp7$residuals)



Code in R:

temp8<-glm(formula=y~job+marital+education+housing+loan+

contact + duration + campaign + poutcome, family = binomial(logit),

data = ds14)

summary(temp8)

After left out the outliers the AIC of the model improves to 2313.4. There are also some changes in the significant predictors. Many predictors which were significant in previous model became insignificant at this model.