**Biostatistics Nonlinear regression project**

Zhiqiang Shu and Pegah Jafarinasabian

5/1/2013

**Project:**

In epidemiology studies, a common focus is on assessing changes in a response distribution with a continuous exposure, adjusting for covariates. For example, Longnecker in 2001 studied the relationship between the DDT metabolite DDE and preterm delivery. The substance DDT is effective against malaria-transmitting mosquitoes, and so is widely used in malaria-endemic areas in spite of growing evidence of health risks. The Longnecker et al. (2001) study measured DDE in mother’s serum during the third trimester of pregnancy, while also recording the gestational age at delivery, GAD, and demographic factors, such as age. Following standard practice in reproductive epidemiology, Longnecker et al. (2001) dichotomized GAD using a 37-week cut-off, so that deliveries occurring prior to 37 weeks of completed gestation were classified as preterm. In this project, the goal is to fit two non-linear regression models:

1. Using the raw gestational ages and using cubic polynomial in each variable as the regressor
2. Dichotomizing the gestational age at delivery as Longnecker did and then fit a logistic and probit regression in R using cubic polynomial in each variable as the regressor.

Divide the dataset into training and test and find out which one of these two models give better prediction.

**1. Data description**

DDE has been shown to have an impact on human health and it has been reported to influence gestational age for pregnant women. In this report, we use DDE amount, gestational age, and age data to determine the relationship between these variables.

We are given a 'dde' dataset. In the 'dde' file, 18 columns of data are recorded: NINDB number, total cholesterol amount, triglycerides amount, pp DDE amount, pp DDE lipid-adj amount, pp DDE source, maternal age, weeks gestation between LMP and delivery, DDE serum, race. Concerning this project, we only need three datasets: pp DDE amount, maternal age, and weeks of gestation between LMP and delivery (GAD in the introduction). There are 2380 samples followed in this study. We divide 2380 samples into train dataset (1500 samples) and test dataset (880 samples).

**2. Goals of the project**

The goal is to find out the relationship between DDT amount and gestational age at delivery. We will try to fit non-linear regressions (cubic polynomial, logit, and probit regression) to 1500 sample train dataset to describe this relationship and compare the fitness using test dataset.

**3. Data pre-processing**

1. Identify the column meanings by referring to the project keys
2. Change the names of columns to identifiable ones, e.g. 'DDE\_A' changes to 'DDE\_A pp.DDE amount', 'V\_MAGE' changes to 'V\_MAGE Maternal age', 'GEST\_DEL' changes to 'GEST\_DEL Weeks gestation between LMP delivery'.
3. Reorganize the table in R (only keep NINDB number, pp DDE amount, maternal age, and GAD)

*setwd("C:/Users/Zhiqiang/Desktop/Biostat project")*

*# set working directory to appropriate place*

*dde=read.csv ("dde (modify).csv", header=T)*

*# read datastr(dde)*

*str(dde)*

*# check the structure of data*

*dde.train = dde[1:1500, ]*

*# assign first 1500 rows of dde to train dataset*

*dde.test= dde[1501:2380,]*

*# assign the rest 880 rows to test dataset*

*attach (dde)*

*# for exploratory analysis purpose, attach dde data*

*dde.r = cbind (NINDB, DDE\_A..pp.DDE..amount.,V\_MAGE..Maternal.age., GEST\_DEL..Weeks.gestation.between.LMP...delivery. )*

*# for better explore data*

**4. Preliminary exploratory analysis**

*summary(dde.r)*

*NINDB DDE\_A..pp.DDE..amount. V\_MAGE..Maternal.age.*

*Min. : 5100461 Min. : 2.50 Min. :13.00*

*1st Qu.:10123136 1st Qu.: 17.09 1st Qu.:20.00*

*Median :45124701 Median : 24.70 Median :23.00*

*Mean :41832100 Mean : 30.18 Mean :24.24*

*3rd Qu.:66123969 3rd Qu.: 36.50 3rd Qu.:28.00*

*Max. :82134511 Max. :178.06 Max. :45.00*

*GEST\_DEL..Weeks.gestation.between.LMP...delivery.*

*Min. :27.70*

*1st Qu.:38.10*

*Median :39.70*

*Mean :39.54*

*3rd Qu.:41.00*

*Max. :90.90*

*> sd(DDE\_A..pp.DDE..amount.)*

*[1] 19.88323*

*> mean(DDE\_A..pp.DDE..amount.)*

*[1] 30.18173*

*> median(DDE\_A..pp.DDE..amount.)*

*[1] 24.695*

*> range(DDE\_A..pp.DDE..amount.)*

*[1] 2.50 178.06*

*> sd(DDE\_A..pp.DDE..amount.)*

*[1] 19.88323*

*> mean( V\_MAGE..Maternal.age.)*

*[1] 24.23782*

*> median( V\_MAGE..Maternal.age.)*

*[1] 23*

*> range( V\_MAGE..Maternal.age.)*

*[1] 13 45*

*> sd( V\_MAGE..Maternal.age.)*

*[1] 6.148963*

*> mean(GEST\_DEL..Weeks.gestation.between.LMP...delivery.)*

*[1] 39.54134*

*> median(GEST\_DEL..Weeks.gestation.between.LMP...delivery.)*

*[1] 39.7*

*> range(GEST\_DEL..Weeks.gestation.between.LMP...delivery.)*

*[1] 27.7 90.9*

*> sd(GEST\_DEL..Weeks.gestation.between.LMP...delivery.)*

*[1] 3.328083*

*# check the distribution of each variable*

*>hist(DDE\_A..pp.DDE..amount.)*



*>hist(V\_MAGE..Maternal.age.)*



>*hist(GEST\_DEL..Weeks.gestation.between.LMP...delivery.)*



*#plot the relationship between gestation time and age or DDE amout*

*>plot (GEST\_DEL..Weeks.gestation.between.LMP...delivery. ~DDE\_A..pp.DDE..amount.)*



*>plot (GEST\_DEL..Weeks.gestation.between.LMP...delivery. ~ V\_MAGE..Maternal.age.)*



*# next part is using train dataset to fit different regressions*

*>detach(dde)*

*>attach(dde.train)*

**5. Regression analysis**

1. Cubic polynomial regression

*>lm.c = lm(GEST\_DEL..Weeks.gestation.between.LMP...delivery.~ DDE\_A..pp.DDE..amount.+ I (DDE\_A..pp.DDE..amount.^2) + I (DDE\_A..pp.DDE..amount.^3) + V\_MAGE..Maternal.age. + I(V\_MAGE..Maternal.age.^2) + I(V\_MAGE..Maternal.age.^3))*

*>summary(lm.c)*

*Call:*

*lm(formula = GEST\_DEL..Weeks.gestation.between.LMP...delivery. ~*

*DDE\_A..pp.DDE..amount. + I(DDE\_A..pp.DDE..amount.^2) + I(DDE\_A..pp.DDE..amount.^3) +*

*V\_MAGE..Maternal.age. + I(V\_MAGE..Maternal.age.^2) +*

*I(V\_MAGE..Maternal.age.^3))*

*Residuals:*

*Min 1Q Median 3Q Max*

*-11.413 -1.456 0.070 1.434 51.217*

*Coefficients:*

*Estimate Std. Error*

*(Intercept) 3.952e+01 4.964e+00*

*DDE\_A..pp.DDE..amount. -5.157e-02 2.704e-02*

*I(DDE\_A..pp.DDE..amount.^2) 7.136e-04 4.763e-04*

*I(DDE\_A..pp.DDE..amount.^3) -3.199e-06 2.198e-06*

*V\_MAGE..Maternal.age. 5.541e-03 5.783e-01*

*I(V\_MAGE..Maternal.age.^2) 3.499e-03 2.184e-02*

*I(V\_MAGE..Maternal.age.^3) -8.321e-05 2.656e-04*

*t value Pr(>|t|)*

*(Intercept) 7.961 3.35e-15 \*\*\**

*DDE\_A..pp.DDE..amount. -1.907 0.0567 .*

*I(DDE\_A..pp.DDE..amount.^2) 1.498 0.1343*

*I(DDE\_A..pp.DDE..amount.^3) -1.455 0.1458*

*V\_MAGE..Maternal.age. 0.010 0.9924*

*I(V\_MAGE..Maternal.age.^2) 0.160 0.8727*

*I(V\_MAGE..Maternal.age.^3) -0.313 0.7541*

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 3.602 on 1493 degrees of freedom*

*Multiple R-squared: 0.008245, Adjusted R-squared: 0.00426*

*F-statistic: 2.069 on 6 and 1493 DF, p-value: 0.05405*

*# but actually later I found linear regression fits better*

*>lm.c1= lm(GEST\_DEL..Weeks.gestation.between.LMP...delivery.~ DDE\_A..pp.DDE..amount. + V\_MAGE..Maternal.age.)*

*>summary(lm.c1)*

*Call:*

*lm(formula = GEST\_DEL..Weeks.gestation.between.LMP...delivery. ~*

*DDE\_A..pp.DDE..amount. + V\_MAGE..Maternal.age.)*

*Residuals:*

*Min 1Q Median 3Q Max*

*-11.516 -1.459 0.097 1.463 51.290*

*Coefficients:*

*Estimate Std. Error t value*

*(Intercept) 39.722843 0.411174 96.608*

*DDE\_A..pp.DDE..amount. -0.012344 0.004606 -2.680*

*V\_MAGE..Maternal.age. 0.007305 0.015235 0.479*

*Pr(>|t|)*

*(Intercept) < 2e-16 \*\*\**

*DDE\_A..pp.DDE..amount. 0.00745 \*\**

*V\_MAGE..Maternal.age. 0.63166*

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*Residual standard error: 3.603 on 1497 degrees of freedom*

*Multiple R-squared: 0.004999, Adjusted R-squared: 0.003669*

*F-statistic: 3.76 on 2 and 1497 DF, p-value: 0.0235*

*# next step is use the model to predict the Y's in terms of the variable in test dataset*

*# Once I get the predicted Y's, Y is binary, I can calculate proportion of mismatches: n(mismatch)/880*

*#Y is real number, I can calculaye the mean squared error (MSE): 1/880 \sum\_{i=1}^880 (Y^{True}\_i - #Y^{predict}\_i)^2*

*> detach (dde.train)*

*>attach (dde.test)*

*# use 'test' to place two variables in test dataset*

*>test= data.frame(DDE\_A..pp.DDE..amount., V\_MAGE..Maternal.age.)*

*# use predict function to get the prediction Y's*

*>lm.predict = predict (lm.c, newdata=test)*

*>lm.predict1= predict(lm.c1, newdata=test)*

*# use mean squared error to test the fitness of the model*

*# 'hydroGPF' package has mse function*

*>library("hydroGOF", lib.loc="D:/R-3.0.0/library")*

*>DDE\_A..pp.DDE..amount.=data.frame(DDE\_A..pp.DDE..amount.)*

*>lm.predict=data.frame(lm.predict)*

*>lm.mse = mse(lm.predict, DDE\_A..pp.DDE..amount.)*

*>lm.mse*

*lm.predict*

*467.9698*

*# similarly, calculate the MSE of linear regression*

*>lm.predict1= data.frame(lm.predict1)*

*>lm.mse1 =mse(lm.predict1, DDE\_A..pp.DDE..amount.)*

*>lm.mse1*

*lm.predict1*

*469.0193*

*# Very interesting, the MSE of model lm.predict (cubic regression) is a little bit smaller than that of model lm.predict1(linear regression), though p-values of linear regression is significant.*

*#assign binary to Y*

*> dde.train[,8]=ifelse(dde.train[,8]<37,1,0)*

*>dde.test[,8]=ifelse(dde.test[,8]<37,1,0)*

*# fit logistic regression on train dataset*

*>detach(dde.test)*

*>attach(dde.train)*

*> dde.logr = glm(GEST\_DEL..Weeks.gestation.between.LMP...delivery.~ DDE\_A..pp.DDE..amount. + V\_MAGE..Maternal.age., family=binomial ("logit"))*

*> summary(dde.logr)*

*Call:*

*glm(formula = GEST\_DEL..Weeks.gestation.between.LMP...delivery. ~*

*DDE\_A..pp.DDE..amount. + V\_MAGE..Maternal.age., family = binomial("logit"))*

*Deviance Residuals:*

*Min 1Q Median 3Q Max*

*-1.2403 -0.5970 -0.5545 -0.5210 2.0891*

*Coefficients:*

*Estimate Std. Error z value Pr(>|z|)*

*(Intercept) -1.874874 0.311290 -6.023 1.71e-09 \*\*\**

*DDE\_A..pp.DDE..amount. 0.013501 0.003000 4.501 6.78e-06 \*\*\**

*V\_MAGE..Maternal.age. -0.008534 0.011769 -0.725 0.468*

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*(Dispersion parameter for binomial family taken to be 1)*

*Null deviance: 1325.6 on 1499 degrees of freedom*

*Residual deviance: 1305.9 on 1497 degrees of freedom*

*AIC: 1311.9*

*Number of Fisher Scoring iterations: 4*

*# Try to combine logistic and cubic regression*

*> dde.logr1 = glm(GEST\_DEL..Weeks.gestation.between.LMP...delivery.~ DDE\_A..pp.DDE..amount.+ I (DDE\_A..pp.DDE..amount.^2) + I (DDE\_A..pp.DDE..amount.^3) + V\_MAGE..Maternal.age. + I(V\_MAGE..Maternal.age.^2) + I(V\_MAGE..Maternal.age.^3), family=binomial ("logit"))*

*> summary (dde.logr1)*

*Call:*

*glm(formula = GEST\_DEL..Weeks.gestation.between.LMP...delivery. ~*

*DDE\_A..pp.DDE..amount. + I(DDE\_A..pp.DDE..amount.^2) + I(DDE\_A..pp.DDE..amount.^3) +*

*V\_MAGE..Maternal.age. + I(V\_MAGE..Maternal.age.^2) +*

*I(V\_MAGE..Maternal.age.^3), family = binomial("logit"))*

*Deviance Residuals:*

*Min 1Q Median 3Q Max*

*-0.9921 -0.6287 -0.5366 -0.4553 2.3709*

*Coefficients:*

*Estimate Std. Error z value Pr(>|z|)*

*(Intercept) -6.882e-01 3.459e+00 -0.199 0.8423*

*DDE\_A..pp.DDE..amount. 5.039e-02 2.030e-02 2.482 0.0131 \**

*I(DDE\_A..pp.DDE..amount.^2) -5.116e-04 3.365e-04 -1.520 0.1284*

*I(DDE\_A..pp.DDE..amount.^3) 1.725e-06 1.484e-06 1.162 0.2451*

*V\_MAGE..Maternal.age. -1.111e-01 4.040e-01 -0.275 0.7832*

*I(V\_MAGE..Maternal.age.^2) -4.846e-04 1.527e-02 -0.032 0.9747*

*I(V\_MAGE..Maternal.age.^3) 5.786e-05 1.852e-04 0.312 0.7548*

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*(Dispersion parameter for binomial family taken to be 1)*

*Null deviance: 1325.6 on 1499 degrees of freedom*

*Residual deviance: 1292.1 on 1493 degrees of freedom*

*AIC: 1306.1*

*Number of Fisher Scoring iterations: 4*

*# predict on the test dataset by logistic regression*

*>detach (dde.train)*

*>attach (dde.test)*

*>logr.predict = predict (dde.logr, newdata=test)*

*>logr.predict1=predict(dde.logr1, newdata=test)*

*#calculate mismatch proportion*

*>logrCompare <- ifelse(dde.test[,8]==logr.predict, 1,0)*

*>logrMismatch = sum(logrCompare == 0)/880*

*>logrMismatch*

*#similarly, it is easy to get mismatch for model logr1*

*>logrCompare 1<- ifelse(dde.test[,8]==logr.predict1, 1,0)*

*>logrMismatch1 = sum(logrCompare1 == 0)/880*

*>logrMismatch1*

*0.1314*

*#similarly, we can perform probit regression*

*# fit probit regression on train dataset*

*>attach(dde.train)*

*>dde.pror = glm(as.factor(GEST\_DEL..Weeks.gestation.between.LMP...delivery.)~ DDE\_A..pp.DDE..amount. + V\_MAGE..Maternal.age., family=binomial ("probit"))*

*>summary(dde.pror)*

*Call:*

*glm(formula = as.factor(GEST\_DEL..Weeks.gestation.between.LMP...delivery.) ~*

*DDE\_A..pp.DDE..amount. + V\_MAGE..Maternal.age., family = binomial("probit"))*

*Deviance Residuals:*

*Min 1Q Median 3Q Max*

*-1.2267 -0.5983 -0.5526 -0.5162 2.1046*

*Coefficients:*

*Estimate Std. Error z value Pr(>|z|)*

*(Intercept) -1.124433 0.172444 -6.521 7.00e-11 \*\*\**

*DDE\_A..pp.DDE..amount. 0.007972 0.001765 4.518 6.25e-06 \*\*\**

*V\_MAGE..Maternal.age. -0.004876 0.006473 -0.750.451*

*---*

*Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1*

*(Dispersion parameter for binomial family taken to be 1)*

*Null deviance: 1325.6 on 1499 degrees of freedom*

*Residual deviance: 1305.2 on 1497 degrees of freedom*

*AIC: 1311.2*

*Number of Fisher Scoring iterations: 4*

*#predict on test dataset by probit*

*>pror.predict = predict (dde.pror, newdata=test)*

*#calculate mismatch proportion*

*>prorCompare = ifelse (dde.test[,8]==dde.pror.predict, 1,0)*

*>prorMismatch = sum(logrCompare == 0)/880*

*> prorMistmatch*

*0.1428*

**6. Summary and discussion**

In this report, we first looked at the question and clarified the goals in statistics view. Later, we took a look at the given data and identified usable datasets. Next step, we explored basic statistical features of the datasets, such as mean, median, range, standard deviation. Before doing regression, we planned to divide dataset (2380 samples) into training group (1500 samples), which are used for fitting regression and testing group (880 samples), which are used to test the prediction of the regression. We tried to fit a non-linear regression to the data, using gestational age at delivery as dependent variable (Y), using age and DDE amount as two independent variables. We first fitted a cubic polynomial regression and found that the p-values of coefficient of 'age' is very large, meaning 'age' might not be a good independent variable to predict gestational age at delivery. As a comparison, we also fitted a linear regression and surprisingly found that the p-value for this regression had reached significant level. To explain this, we think that the p-values of coefficients are not necessarily related to the reliability of the model because the null hypotheses for these p-values are that these coefficients equal to 0. To test the prediction ability of our model, we utilized test group. We compared Y values from test group and Y values calculated from the models using mean squared error (MSE). In the processes of fitting regressions and test predictions, we found sometimes it got confused by different names in different files, so we had to attach and detach 'dde', 'dde.train', and 'dde.test' back and forth. We think this can be improved by some commands in R but we have not found them. We found that MSE of cubic polynomial model is slightly smaller than linear regression. So we concluded that cubic polynomial regression has a better predictive ability that linear regression. However, we thought that it is not clear how powerful and precise this cubic polynomial regression is, because it is possible that both models are not satisfactory. To test different non-linear regressions, we further tried to fit logistic and probit regression. According to Longnecker's method, we assign '1' to gestational age < 37 weeks, meaning they are events of preterm; and assign '0' to gestational age >=37 weeks, meaning these are not events. After fitting the logit regression and probit regression by using training dataset, we also used testing dataset to validate the reliability of the models. For binary data, we can use mismatch proportion to measure the fitness of these models. It comes out that about 90% of predictions match with the real binary data. The results of logit and probit regression are very similar. In summary, these non-linear regressions basically give us an idea of the relationship between preterm delivery and DDE amount as well as the actual age. One problem for cubic polynomial regession is that we have no idea how large MSE should be, so we cannot conclude whether this non-linear regression is good or not. Meanwhile, we cannot compare cubic polynomial regression and logit regression because they use different methods to determine the fitness (MSE). Furthermore, to fit the logit and probit regression, we have to dichotomize the data into binary, which leads to a huge loss of information. In clinical, though 37 weeks can be used as a criterion to determine whether or not it is a preterm delivery, the actual time of gestational age is also very important in terms of the health of the woman and the fetus. We believe there are more advanced statistical tests and convenient R codes there, but with the limitation of knowledge we cannot apply and test all of them one by one. This class and this project gave us a good opportunity to learn and explore how to apply statistics to biological and clinical studies. We will gain more knowledge and power as we study more advanced statistics and get more comfortable using R language.

Appendix

*setwd("C:/Users/Zhiqiang/Desktop/Biostat project")*

*dde = read.csv ("dde (modify).csv", header=T)*

*str(dde)*

*# divide into train dataset and test dataset*

*dde.train = dde[1:1500, ]*

*dde.test= dde[1501:2380,]*

*attach(dde)*

*# explortory stat analysis*

*hist(DDE\_A..pp.DDE..amount.)*

*hist(V\_MAGE..Maternal.age.)*

*hist(GEST\_DEL..Weeks.gestation.between.LMP...delivery.)*

*plot (GEST\_DEL..Weeks.gestation.between.LMP...delivery. ~DDE\_A..pp.DDE..amount.)*

*plot (GEST\_DEL..Weeks.gestation.between.LMP...delivery. ~ V\_MAGE..Maternal.age.)*

*mean(DDE\_A..pp.DDE..amount.)*

*median(DDE\_A..pp.DDE..amount.)*

*range(DDE\_A..pp.DDE..amount.)*

*sd(DDE\_A..pp.DDE..amount.)*

*mean( V\_MAGE..Maternal.age.)*

*median( V\_MAGE..Maternal.age.)*

*range( V\_MAGE..Maternal.age.)*

*sd( V\_MAGE..Maternal.age.)*

*mean(GEST\_DEL..Weeks.gestation.between.LMP...delivery.)*

*median(GEST\_DEL..Weeks.gestation.between.LMP...delivery.)*

*range(GEST\_DEL..Weeks.gestation.between.LMP...delivery.)*

*sd(GEST\_DEL..Weeks.gestation.between.LMP...delivery.)*

*detach(dde)*

*#Focusing on train dataset*

*attach (dde.train)*

*# fit cubic regression on train dataset*

*lm.c = lm(GEST\_DEL..Weeks.gestation.between.LMP...delivery.~ DDE\_A..pp.DDE..amount.+ I (DDE\_A..pp.DDE..amount.^2) + I (DDE\_A..pp.DDE..amount.^3) + V\_MAGE..Maternal.age. + I(V\_MAGE..Maternal.age.^2) + I(V\_MAGE..Maternal.age.^3))*

*lm.c1= lm(GEST\_DEL..Weeks.gestation.between.LMP...delivery.~ DDE\_A..pp.DDE..amount. + V\_MAGE..Maternal.age.)*

*summary(lm.c)*

*summary(lm,c1)*

*# predict on test dataset by cubic regression*

*detach (dde.train)*

*attach (dde.test)*

*test= data.frame(DDE\_A..pp.DDE..amount., V\_MAGE..Maternal.age.)*

*lm.predict = predict (lm.c, newdata=test)*

*lm.predict1= predict (lm.c1, newdata=test)*

*#calculate mean squared error between test dataset and predict ones*

*#lm.mse =1/880 (sum\_{i=1}^880 (dde.test$GEST\_DEL..Weeks.gestation.between.LMP...delivery.\_i - lm.predict\_i)^2)*

*library("hydroGOF", lib.loc="D:/R-3.0.0/library")*

*DDE\_A..pp.DDE..amount.=data.frame(DDE\_A..pp.DDE..amount.)*

*lm.predict=data.frame(lm.predict)*

*lm.mse = mse(lm.predict, DDE\_A..pp.DDE..amount.)*

*lm.predict1=data.frame(lm.predict1)*

*lm.mse1 = mse(lm.predict1, DDE\_A..pp.DDE..amount.)*

*#assign binary to Y*

*dde.train[,8]=ifelse(dde.train[,8]<37,1,0)*

*dde.test[,8]=ifelse(dde.test[,8]<37,1,0)*

*# fit logistic regression on train dataset*

*attach (dde.train)*

*dde.logr = glm(GEST\_DEL..Weeks.gestation.between.LMP...delivery.~ DDE\_A..pp.DDE..amount.+ V\_MAGE..Maternal.age. , family=binomial ("logit"))*

*summary(dde.logr)*

*dde.logr1 = glm(GEST\_DEL..Weeks.gestation.between.LMP...delivery.~ DDE\_A..pp.DDE..amount.+ I (DDE\_A..pp.DDE..amount.^2) + I (DDE\_A..pp.DDE..amount.^3) + V\_MAGE..Maternal.age. + I(V\_MAGE..Maternal.age.^2) + I(V\_MAGE..Maternal.age.^3), family=binomial ("logit"))*

*summary(dde.logr1)*

*#predict on test dataset by logit*

*detach (dde.train)*

*attach (dde.test)*

*logr.predict = predict (dde.logr, newdata=test)*

*logr.predict1 = predict (dde.logr1, newdata=test)*

*#calculate mismatch proportion*

*logrCompare <- ifelse(dde.test[,8]==logr.predict, 1,0)*

*logrMismatch = sum(logrCompare == 0)/880*

*logrMismatch*

*logrCompare1 <- ifelse(dde.test[,8]==logr.predict1, 1,0)*

*logrMismatch1 = sum(logrCompare1 == 0)/880*

*logrMismatch1*

*# fit probit regression on train dataset*

*dde.pror = glm(as.factor(GEST\_DEL..Weeks.gestation.between.LMP...delivery.)~ DDE\_A..pp.DDE..amount. + V\_MAGE..Maternal.age., family=binomial ("probit"))*

*summary(dde.pror)*

*#predict on test dataset by logit*

*pror.predict = predict (dde.pror, newdata=test)*

*#calculate mismatch proportion*

*prorCompare <- ifelse(dde.test[,8]==pror.predict, 1,0)*

*prorMismatch = sum(prorCompare == 0)/880*

*prorMismatch*