STA138 Project

3/18/2020

Haolin Li (913838107) | Yiting Kuang (913992425)

Data Description

For this project we have two different datasets for two different problems. The first dataset is about low birth weight of infant and its related information on mother such as age, weight, smoking status during pregnancy, history of pre-mature labor, history of hypertension and number of visits during the first trimester. There is a total of 7 different variables in the dataset. The second dataset is about the number of emergency room visits and the related information regarding the subscribers from the insurance company, such as cost, age, gender, number of interventions, number of tracked drugs prescribed, number of other complications, number of other disease(comorbidities) and number of days of duration of treatment, a total of 9 different variables in the dataset.

Goal of Analysis

The goal for problem 1 is to investigate if the probability of low birth weight of infant is related to the information about the mother, we used binomial distribution for this problem. The statistical methods we used is to first find the summary of the data, then use goodness-of-fit to drop unnecessary variables and interaction terms, then we use the backward stepwise regression method and the AIC criterion to select an appropriate model.

The goal for problem 2 is to model the mean of the number emergency visits as a function of 8 other variables, we will use Poisson distribution for this problem. The procedure of statistical methods is similar to problem 1 but we had transformed and untransformed variables cases to further investigate the model.

Problem 1: Low Birth Rate or Not

Logistic Regression and Data Summary

Full model:

$$\pi' = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_1 X_2 + \beta_8 X_2 X_5 + \beta_9 X_2 X_4$$

Where $X_1 = \text{age}, X_2 = \text{weight}, X_3 = \text{smoke status}, X_4 = \text{pre} - \text{mature labor}, X_5 = \text{hypertension},$ $X_6 = \text{number of visits}, X_1X_2 = \text{age and weight}, X_2X_5 = \text{weight and hypertension}, X_2X_4 = \text{weight and pre} -$ $\text{mature labor}.$

The parameter estimates, standard errors and p-values are shown in the table below:

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.2742473  3.9295522  -0.833    0.405
age            0.1165393  0.1682394   0.693    0.488
weight         0.0258898  0.0304761   0.850    0.396
smokeyes      -0.5087686  0.3545077  -1.435    0.151
preyes        -2.6303649  2.8680721  -0.917    0.359
hypyes        -1.6168753  2.6785514  -0.604    0.546
visits         0.0292054  0.1802465   0.162    0.871
age:weight    -0.0004465  0.0012718  -0.351    0.726
weight:hypyes -0.0009711  0.0171863  -0.057    0.955
weight:preyes  0.0064942  0.0222361   0.292    0.770
```

Goodness-of-fit (Likelihood ratio test: Chi-squared Test)

We want to know if the interaction terms can be dropped. We conduct a hypothesis test: $H_0 = \beta_7 = \beta_8 = \beta_9 = 0$, $H_a = \text{at least one of } \beta_7, \beta_8, \beta_9 \text{ is not } 0$.

```
Analysis of Deviance Table

Model 1: birth ~ age + weight + smoke + pre + hyp + visits
Model 2: birth ~ age + weight + smoke + pre + hyp + visits + age * weight +
    weight * hyp + weight * pre
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       182     202.15
2       179     201.96  3  0.19127    0.979
```

Since the $G^2 = 0.19127$ with $df = 3$ and the p-value of 0.979 is larger than 0.05, we cannot reject H_0. We can conclude that the interaction terms should not be added to the model.

Backward Stepwise Regression and AIC

Next, we will use backward stepwise regression and AIC criterion to obtain an appropriate model. Output is shown below.

```
Call:  glm(formula = birth ~ age + weight + smoke + pre + hyp, family =
"binomial",
    data = baby)

Coefficients:
(Intercept)          age       weight     smokeyes       preyes
   -2.03197      0.06032      0.01615     -0.51837     -1.79404
     hypyes
   -1.78271

Degrees of Freedom: 188 Total (i.e. Null);  183 Residual
Null Deviance:      234.7
Residual Deviance: 202.2        AIC: 214.2
```

We now obtained our final model: $\pi' = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$

Goodness-of-fit (Likelihood ratio test: Chi-squared Test)

We observed that X_6(Visits) is dropped from the stepwise model, we want to know whether it is true that it can be dropped.

Hypothesis Test: $H_0: \beta_6 = 0$, $H_a: \beta_6$ is not 0.

```
Analysis of Deviance Table

Model 1: birth ~ age + weight + smoke + pre + hyp
Model 2: birth ~ age + weight + smoke + pre + hyp + visits
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       183     202.19
2       182     202.15  1 0.032333   0.8573
```

Since the $G^2 = 0.032333$ with $df = 1$ and the p-value of 0.8673 is larger than 0.05, we cannot reject H_0. We can conclude that the term X_6 can be dropped from the model.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.031969   1.111157  -1.829 0.067445 .
age          0.060319   0.036317   1.661 0.096735 .
weight       0.016154   0.006923   2.333 0.019625 *
smokeyes    -0.518366   0.348309  -1.488 0.136688
preyes      -1.794038   0.508841  -3.526 0.000422 ***
hypyes      -1.782710   0.716698  -2.487 0.012868 *
```

From the statistics above, we observed that the p-value for smoke is quite large, therefore we want to know if X_3 can be dropped from the model.

Hypothesis Test: $H_0: \beta_3 = 0$, $H_a: \beta_3$ is not 0.

```
Analysis of Deviance Table

Model 1: birth ~ age + weight + pre + hyp
Model 2: birth ~ age + weight + smoke + pre + hyp + visits
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       184     204.39
2       182     202.15  2   2.2349   0.3271
```

Since the $G^2 = 2.2349$ with $df = 2$ and the p-value of 0.3271 is larger than 0.05, we cannot reject H_0. We can conclude that the term X_3 can be dropped from the model.

Final Model and Results

Finally, we obtained the final model

$$\pi' = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \beta_5 X_5$$

Where $X_1 = \text{age}, X_2 = \text{weight}, X_4 = \text{pre} - \text{mature labor}, X_5 = \text{hypertension}.$ The parameter estimates, standard errors and p-values are shown below.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.30857    1.09224  -2.114 0.034548 *
age          0.06053    0.03584   1.689 0.091209 .
weight       0.01668    0.00698   2.390 0.016850 *
preyes      -1.89383    0.50307  -3.765 0.000167 ***
hypyes      -1.83525    0.73556  -2.495 0.012594 *
```

By using the final model to estimate the percentage of correct classification, we obtained that the correctness of the model is 138/189 = (0.73) = 73%.

Conclusion

Problem 2: Ischemic Heart Disease

Untransformed Predictor Variables

Logistic Regression and Data Summary

Full model:

$$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_3 X_1 + \beta_{10} X_3 X_2 + \beta_{11} X_3 X_4 + \beta_{12} X_3 X_5 + \beta_{13} X_3 X_6 + \beta_{14} X_3 X_7 + \beta_{15} X_3 X_8$$

Where $X_1 = \text{cost}, X_2 = \text{age}, X_3 = \text{gender}, X_4 = \text{inter}, X_5 = \text{drugs}, X_6 = \text{complications}, X_7 = \text{comorbidities}, X_8 = \text{duration}, X_3 X_1 = \text{gender and cost}, \; X_3 X_2 = \text{gender and age}, X_3 X_4 =$

gender and inter, X_3X_5 = gender and drugs, X_3X_6 = gender and complications, X_3X_7 = gender and comorbidities, and X_3X_8 = gender and duration.

The parameter estimates, standard errors and p-values are shown in the table below:

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          4.768e-01  2.052e-01   2.323  0.02016 *
cost                 1.567e-05  3.051e-06   5.136  2.8e-07 ***
age                  6.969e-03  3.457e-03   2.016  0.04383 *
gender               2.332e-01  4.025e-01   0.580  0.56223
inter                1.023e-02  4.177e-03   2.450  0.01427 *
drugs                1.894e-01  1.497e-02  12.653  < 2e-16 ***
complications        2.436e-01  8.310e-02   2.931  0.00338 **
comorbidities        1.453e-03  4.097e-03   0.355  0.72284
duration             2.559e-04  2.210e-04   1.158  0.24688
cost:gender         -7.336e-06  9.080e-06  -0.808  0.41912
age:gender          -9.971e-04  6.803e-03  -0.147  0.88347
gender:inter         7.922e-03  1.139e-02   0.696  0.48660
gender:drugs         6.357e-03  2.865e-02   0.222  0.82442
gender:complications -3.713e-01  1.307e-01  -2.840  0.00452 **
gender:comorbidities -8.660e-03  9.508e-03  -0.911  0.36238
gender:duration      2.701e-04  4.326e-04   0.624  0.53239
```

Goodness-of-fit (Likelihood ratio test: Chi-squared Test)

We want to know if the interaction terms can be dropped, except X_3X_6 because its p-value is low. We want to conduct a hypothesis test :$H_0: \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{14} = \beta_{15} = 0$,
H_a: at *least one* $\beta_9, \beta_{10}, \beta_{11}, \beta_{12}, \beta_{14}, \beta_{15}$ *is not* 0.

```
Analysis of Deviance Table

Model 1: visits ~ cost + age + gender + inter + drugs + complications +
    comorbidities + duration
Model 2: visits ~ cost + age + gender + inter + drugs + complications +
    comorbidities + duration + gender * cost + gender * age +
    gender * inter + gender * drugs + gender * comorbidities +
    gender * duration
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       779     1043.6
2       773     1041.3  6   2.3572   0.8841
```

Since the $G^2 = 2.3572$ with $df = 6$ and the p-value of 0.8841 is larger than 0.05, we cannot reject H_0. We can conclude that the interaction terms except X_3X_6 can be dropped from the model.

Backward Stepwise Regression and AIC

Next, we will use backward stepwise regression and AIC criterion to obtain an appropriate model. Output is shown below.

```
Call:  glm(formula = visits ~ cost + age + gender + inter + drugs +
    complications + duration + gender:complications, family = poisson(),
    data = chem)

Coefficients:
         (Intercept)                  cost                   age                gender
           4.840e-01             1.464e-05             6.751e-03             2.166e-01
               inter                 drugs         complications              duration
           1.074e-02             1.927e-01             2.390e-01             3.190e-04
gender:complications
          -3.566e-01

Degrees of Freedom: 787 Total (i.e. Null);  779 Residual
Null Deviance:      1485
Residual Deviance: 1035         AIC: 3262
```

We now have the model:

$$\log(\mu) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6 + \beta_8X_8 + \beta_{13}X_3X_6$$

Goodness-of-fit (Likelihood ratio test: Chi-squared Test)

We observed that X_7(Comorbidities) is dropped from the full model, we want to know whether it is true that it can be dropped. $H_0: \beta_7 = 0$, $H_a: \beta_7$ is not 0.

```
Analysis of Deviance Table

Model 1: visits ~ cost + age + gender + inter + drugs + complications +
    duration + gender * complications
Model 2: visits ~ cost + age + gender + inter + drugs + complications +
    comorbidities + duration + gender * complications
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       779     1034.9
2       778     1034.8  1 0.019091   0.8901
```

Since the $G^2 = 0.01901$ with df = 1 and the p-value of 0.8901 is larger than 0.05, we cannot reject H_0. We can conclude that the term X_7 can be dropped from the model.

Final Model and Results

Finally, we obtained the final model

$$\log(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_8 X_8 + \beta_{13} X_3 X_6$$

Where $X_1 = \text{cost}, X_2 = \text{age}, X_3 = \text{gender}, X_4 = \text{inter}, X_5 = \text{drugs}, X_6 = \text{complications}, X_8 = \text{duration}, X_3 X_6 = \text{gender and complications}$. The parameter estimates, standard errors and p-values are shown below:

```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.840e-01  1.760e-01   2.750  0.00596 **
cost                  1.464e-05  2.857e-06   5.124 2.99e-07 ***
age                   6.751e-03  2.962e-03   2.280  0.02264 *
gender                2.166e-01  4.522e-02   4.790 1.67e-06 ***
inter                 1.074e-02  3.800e-03   2.827  0.00469 **
drugs                 1.927e-01  1.229e-02  15.680  < 2e-16 ***
complications         2.390e-01  8.263e-02   2.892  0.00383 **
duration              3.190e-04  1.690e-04   1.888  0.05901 .
gender:complications -3.566e-01  1.234e-01  -2.891  0.00384 **
```

Transformed Predictor Variables

Logistic Regression and Data Summary

Full model:

$$\begin{aligned}\log(\mu) = {} & \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 \sqrt{X_1} + \beta_{10} \sqrt{X_2} + \beta_{11} \sqrt{X_4} \\ & + \beta_{12} \sqrt{X_5} + \beta_{13} \sqrt{X_6} + \beta_{14} \sqrt{X_7} + \beta_{15} \sqrt{X_8} + \beta_{16} X_3 \sqrt{X_1} + \beta_{17} X_3 \sqrt{X_2} + \beta_{18} X_3 \sqrt{X_4} + \beta_{19} X_3 \sqrt{X_5} \\ & + \beta_{20} X_3 \sqrt{X_6} + \beta_{21} X_3 \sqrt{X_7} + \beta_{22} X_3 \sqrt{X_8}\end{aligned}$$

Where $X_1 = \text{cost}, X_2 = \text{age}, X_3 = \text{gender}, X_4 = \text{inter}, X_5 = \text{drugs}, X_6 = \text{complications}, X_7 = \text{comorbidities}, X_8 = \text{duration}$ and the transformed terms and interaction terms.

The parameter estimates, standard errors and p-values are shown in the table below:

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-7.472e+00	3.707e+00	-2.016	0.043826	*
cost	-1.157e-05	7.829e-06	-1.478	0.139414	
age	-1.380e-01	6.665e-02	-2.071	0.038354	*
gender	-1.780e-01	8.021e-01	-0.222	0.824386	
inter	1.493e-02	8.658e-03	1.725	0.084559	.
drugs	-5.394e-03	3.807e-02	-0.142	0.887334	
complications	2.482e-02	3.589e-01	0.069	0.944864	
comorbidities	9.550e-03	7.751e-03	1.232	0.217904	
duration	5.172e-04	6.544e-04	0.790	0.429399	
sqrt(cost)	6.105e-03	1.733e-03	3.523	0.000426	***
sqrt(age)	2.138e+00	9.948e-01	2.149	0.031608	*
sqrt(inter)	-5.413e-02	5.035e-02	-1.075	0.282328	
sqrt(drugs)	4.316e-01	7.876e-02	5.480	4.25e-08	***
sqrt(complications)	1.655e-01	3.667e-01	0.451	0.651653	
sqrt(comorbidities)	-6.936e-02	3.921e-02	-1.769	0.076915	.
sqrt(duration)	2.333e-04	1.352e-02	0.017	0.986232	
gender:sqrt(cost)	-6.186e-04	1.637e-03	-0.378	0.705501	
gender:sqrt(age)	5.717e-02	1.051e-01	0.544	0.586354	
gender:sqrt(inter)	4.671e-03	5.501e-02	0.085	0.932323	
gender:sqrt(drugs)	-2.290e-02	6.365e-02	-0.360	0.719058	
gender:sqrt(complications)	-3.346e-01	1.896e-01	-1.765	0.077512	.
gender:sqrt(comorbidities)	-1.197e-02	4.036e-02	-0.297	0.766828	
gender:sqrt(duration)	-1.392e-03	9.197e-03	-0.151	0.879735	

Goodness-of-fit (Likelihood ratio test: Chi-squared Test)

We want to know if the interaction terms can be dropped, except $X_3\sqrt{X_6}$ because its p-value is low. We want to conduct a hypothesis test :$H_0: \beta_{16} = \beta_{17} = \beta_{18} = \beta_{19} = \beta_{21} = \beta_{22} = 0$,
H_a: a$t\ least\ one$ $\beta_{16}, \beta_{17}, \beta_{18}, \beta_{19}, \beta_{21}, \beta_{22}$ $is\ not$ 0.

We have the $G^2 = 1.0462$ with $df = 6$ and the p-value of 0.9838 is larger than 0.05, we cannot reject H_0. We can conclude that the interaction terms except $X_3\sqrt{X_6}$ can be dropped from the model.

Final Model and Result

After using backward stepwise regression and goodness-of-fit, we dropped $X_1, X_5, X_6, X_7, \sqrt{X_4}$ and $\sqrt{X_8}$.

Our final model is:

$$\log(\mu) = \beta_0 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_8X_8 + \beta_9\sqrt{X_1} + \beta_{10}\sqrt{X_2} + \beta_{12}\sqrt{X_5} + \beta_{13}\sqrt{X_6} + \beta_{14}\sqrt{X_7} + \beta_{20}X_3\sqrt{X_6}$$

Where $X_1 = \text{cost}, X_2 = \text{age}, X_3 = \text{gender}, X_4 = \text{inter}, X_5 = \text{drugs}, X_6 = \text{complications}, X_7 = \text{comorbidities}, X_8 = \text{duration}$ and the transformed terms and interaction terms.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-6.7325067	3.6697587	-1.835	0.0666	.
age	-0.1220539	0.0658297	-1.854	0.0637	.
gender	0.2025178	0.0456411	4.437	9.11e-06	***
inter	0.0062730	0.0043873	1.430	0.1528	
duration	0.0004503	0.0002014	2.236	0.0253	*
sqrt(cost)	0.0036843	0.0006470	5.694	1.24e-08	***
sqrt(age)	1.9179322	0.9838798	1.949	0.0513	.
sqrt(drugs)	0.4185867	0.0269356	15.540	< 2e-16	***
sqrt(complications)	0.2079244	0.0826169	2.517	0.0118	*
sqrt(comorbidities)	-0.0263249	0.0169462	-1.553	0.1203	
gender:sqrt(complications)	-0.3642030	0.1491643	-2.442	0.0146	*

Conclusion

According to the results of problem 1 and problem 2, we notice that a more complicated model does not indicate better performance.

In problem 1, we can conclude that the probability of low birth weight of infant is related to the mother's age, weight, history of pre-mature labor, and history of hypertension. The number of visits during the first trimester is not significant to the infant birth weight. Smoking status during pregnancy is a variable that can consider including in the model because it yields a lower AIC; however, based on the given data, dropping the "smoke" variable gives a slightly higher accuracy of classification (about 1% higher).

In problem 2, the final model with transformed variables gives a better result than the one with untransformed variables. Besides, including interaction between gender and other predictor variables for the untransformed case does not generate a lower AIC. Based on the final model with transformed variables, we can conclude that only the subscriber's age, number of interventions, number of other diseases, and number of days of duration of treatment condition are significantly related to having ischemic.

Appendix i: Graphs

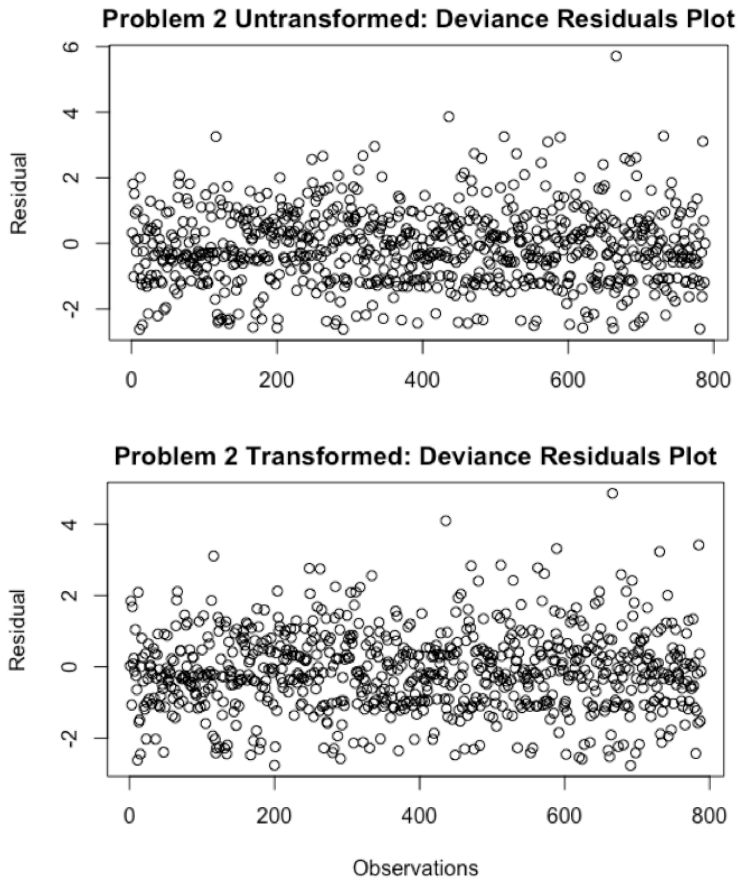

Fig: Plots of deviance residuals against index are shown above for untransformed and transformed variables. There does not seem to have any obvious outliers, and the models seem to fit the data well. However, we can observe that the difference of the residual bounds.

Appendix ii: Code

```
#Problem 1
library("readxl")
baby = read_excel("~/Desktop/Statistics/STA138/baby.xls")

#Statistical Summary
second_baby =
glm(birth~age+weight+smoke+pre+hyp+visits+age*weight+weight*hyp+weight*pre, family
= 'binomial', data = baby)
summary(second_baby)

# Goodness-of-fit
# Drop interaction terms
first_baby = glm(birth~age+weight+smoke+pre+hyp+visits, family = 'binomial', data =
baby)
anova(first_baby,second_baby, test = "Chisq")

#Backward stepwise regression
step(first_baby)

#Goodness-of-fit
#Drop visits
reduced_baby = glm(birth~age+weight+smoke+pre+hyp, family = 'binomial', data =
baby)
anova(reduced_baby,first_baby, test = "Chisq")

#Goodness-of-fit
#Drop smoke and visits
reduced_baby = glm(birth~age+weight+smoke+pre+hyp+visits, family = 'binomial', data
= baby)
check_baby = glm(birth~age+weight+pre+hyp, family = 'binomial', data = baby)
anova(check_baby,reduced_baby, test = "Chisq")

#Statistical Summary
summary(check_baby)

#Estimate the percentage of correct classification
beta0 = -2.30857
beta1 = 0.06053
beta2 = 0.01668
beta4 = -1.89383
beta5 = -1.83525

for (i in 1:189){
  if(beta0+beta1*baby[i,1]*beta2*baby[1,2]+beta4*baby[i,9]+beta5*baby[i,10]
     > 0.5) {
    baby$est_birth[i] = 1
    if(baby$est_birth[i] == baby$birth[i]){
```

```
      baby$accurate[i] = 'yes'
    }
    else{
      baby$accurate[i] = 'no'
    }
  }
  else{
    baby$est_birth[i] = 0
    if(baby$est_birth[i] == baby$birth[i]){
      baby$accurate[i] = 'yes'
    }
    else{
      baby$accurate[i] = 'no'
    }
  }
}
yes_accurate = baby[which(baby$accurate == 'yes'),]
nrow(yes_accurate)/189

#Problem 2
#Untransformed Variables
#Statistical Summary
chem = read_excel("~/Desktop/Statistics/STA138/ischemic.xlsx")
second_chem =
glm(visits~cost+age+gender+inter+drugs+complications+comorbidities+duration+gender*
cost+gender*age+gender*inter+gender*drugs+gender*complications+gender*comorbidities
+gender*duration, family = poisson(), data = chem)
summary(second_chem)

# Goodness-of-fit
# Drop interaction terms except gender*complications
first_chem =
glm(visits~cost+age+gender+inter+drugs+complications+comorbidities+duration, family
= poisson(), data = chem)
third_chem =
glm(visits~cost+age+gender+inter+drugs+complications+comorbidities+duration+gender*
cost+gender*age+gender*inter+gender*drugs+gender*comorbidities+gender*duration,
family = poisson(), data = chem)
anova(first_chem,third_chem, test = "Chisq")

#Backward stepwise regression
step(check2_chem)

# Goodness-of-fit
# Drop comorbidities
fourth_chem =
glm(visits~cost+age+gender+inter+drugs+complications+duration+gender*complications,
family = poisson(), data = chem)
anova(fourth_chem,check2_chem, test = "Chisq")
```

```
#Statistical Summary
summary(fourth_chem)

#Transformed Variables
#Statistical Summary
trans_chem =
glm(visits~cost+age+gender+inter+drugs+complications+comorbidities+duration+sqrt(co
st)+sqrt(age)+sqrt(inter)+sqrt(drugs)+sqrt(complications)+sqrt(comorbidities)+sqrt(
duration)+gender*sqrt(cost)+gender*sqrt(age)+gender*sqrt(inter)+gender*sqrt(drugs)+
gender*sqrt(complications)+gender*sqrt(comorbidities)+gender*sqrt(duration), family
= poisson(), data = chem)
summary(trans_chem)

# Goodness-of-fit
# Drop interaction terms except gender*sqrt(complications)
trans2_chem =
glm(visits~cost+age+gender+inter+drugs+complications+comorbidities+duration+sqrt(co
st)+sqrt(age)+sqrt(inter)+sqrt(drugs)+sqrt(complications)+sqrt(comorbidities)+sqrt(
duration)+gender*sqrt(complications), family = poisson(), data = chem)
anova(trans2_chem,trans_chem, test = "Chisq")

#Backward stepwise regression
step(trans2_chem)

# Goodness-of-fit
# Drop cost,drugs,complications, comorbidities, sqrt(inter),sqrt(duration)
trans3_chem = glm(formula = visits ~ age + gender + inter + duration + sqrt(cost) +
sqrt(age) + sqrt(drugs) + sqrt(complications) + sqrt(comorbidities)
+gender*sqrt(complications), family = poisson(), data = chem)
anova(trans3_chem,trans2_chem, test = "Chisq")

#Statistical Summary
summary(trans3_chem)


#Plot

res_trans_chem = residuals(trans3_chem, type = "deviance")

plot(res_trans_chem, main = "Problem 2 Transformed: Deviance Residuals Plot", xlab
= "Observations", ylab = "Residual" )

res_chem = residuals(fourth_chem, type = "deviance")

plot(res_chem, main = "Problem 2 Untransformed: Deviance Residuals Plot", xlab =
"Observations", ylab = "Residual" )
```