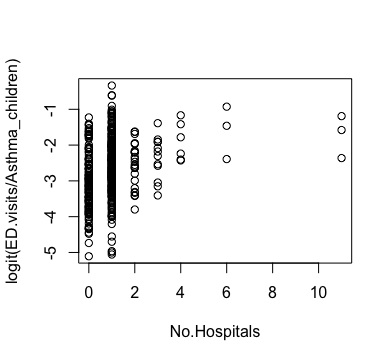
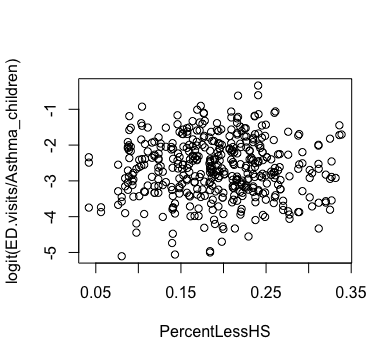
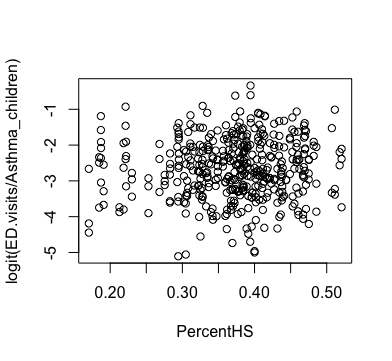
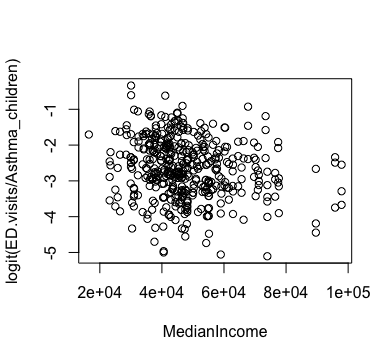
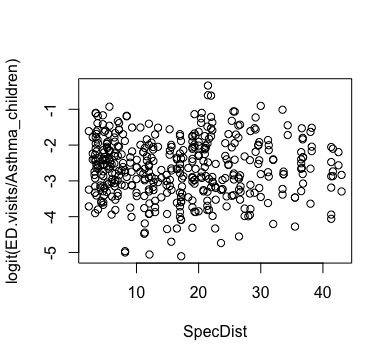
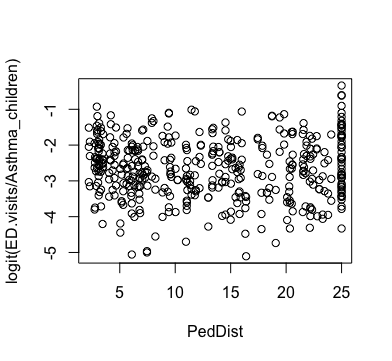
**Question 1: Exploratory Data Analysis**

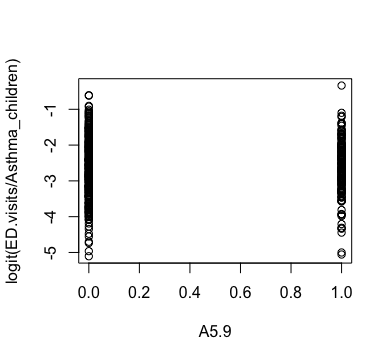
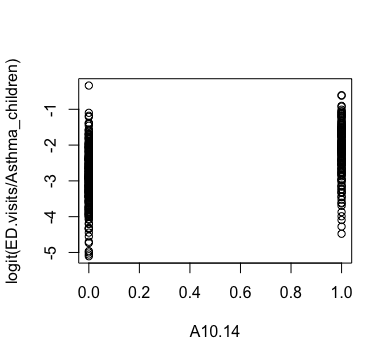
  

* Outside of a generally increasing(positive) trend with respect to No.Hospitals, logit(ED.visits/Asthma\_children) doesn’t seem to have any particular trends with any other variables

Correlation Coefficients:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No.Hospitals | PercentLessHS | PercentHS | MedianIncome | SpecDist | PedDist |
| 0.2652837 | 0.009301958 | 0.05473725 | -0.1750845 | -0.02056114 | -0.004714967 |

* The correlation coefficients agree with the scatterplots; No.Hospitals is somewhat positively correlated to logit(ED.visits/Asthma\_children). MedianIncome also appears to have a weak correlation to logit(ED.visits/Asthma\_children), although this is not very clear in the scatterplot.

Correlation Coefficients:

|  |  |
| --- | --- |
| A5.9 | A10.14 |
| 0.003648238 | 0.4092981 |

* Based on the plot and the correlation coefficient, A10.14 seems to be somewhat positively correlated to logit(ED.visits/Asthma\_children). However, A5.9 appears to be non-correlated.

Based on the analysis above, it is NOT reasonable to assume linear relationships between the log odds of an ED visit and most of the predictors.

**Question 2: Fitting the Logistic Regression Model**

(i) What are the model parameters and their estimates?

The model parameters are the intercept and the predictors(standardized) and their estimates are as follows.

Call:

glm(formula = ED.visits/Asthma\_children ~ ., family = "binomial",

data = newdata, weights = Asthma\_children)

Coefficients: (1 not defined because of singularities)

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

(Intercept) -3.157787 0.024346 -129.706 < 2e-16 \*\*\*

A5.9 0.716228 0.024146 29.663 < 2e-16 \*\*\*

A10.14 1.121537 0.025892 43.316 < 2e-16 \*\*\*

A15.17 NA NA NA NA

sNumHospitals 0.137774 0.003907 35.263 < 2e-16 \*\*\*

sPercentLessHS -0.246375 0.016893 -14.584 < 2e-16 \*\*\*

sPercentHS 0.043294 0.019610 2.208 0.0273 \*

sMedianIncome -0.265967 0.014936 -17.807 < 2e-16 \*\*\*

sAvgDistS 0.067103 0.013686 4.903 9.44e-07 \*\*\*

sAvgDistP -0.093793 0.016163 -5.803 6.52e-09 \*\*\*

---

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

(ii) Write down the equation for the estimated logit transformation of the probability of an ED visit given the predicting variables;

g(p) = -3.157787 + 0.716228\*A5.9 + 1.121537\*A10.14 + 0.137774\*sNumHospitals -0.246375\*sPercentLessHS + 0.043294\*sPercentHS -0.265967\*sMedianIncome + 0.067103\*sAvgDistS – 0.093793\*sAvgDistP

(iii) Interpret the estimated value of the parameters corresponding with A10.14 and sPercentLessHS in the context of the problem.

* For children from age 10 to 14 versus children from age 15 to 17, the log odds of an ED visit by a child with Asthma increases by 1.121537 controlling for other predictors in the model.
* For each unit increase in the standardized percentage of population without a high school degree in the county, the log odds of an ED visit by a child with Asthma in that county decreases by 0.246375 controlling for other predictors in the model.

**Question 3: Testing the significance of linear relationships observed in the data.**

(i) Find a 99% confidence interval for the parameters corresponding to all predictors plus the intercept.

0.5 % 99.5 %

(Intercept) -3.65446252 -3.11200630

A5.9 0.65436066 0.77876074

A10.14 1.05509891 1.18849518

A15.17 NA NA

sNumHospitals 0.12770737 0.14783591

sPercentLessHS -0.29001845 -0.20298957

PercentHS -0.10082069 1.30283735

sMedianIncome -0.30442152 -0.22747354

sAvgDistS 0.03184407 0.10235357

sAvgDistP -0.13547521 -0.05220446

(ii) Which variables are significant at the significance level α = 0.01? Give the p-value for any variable that is not significant.

Coefficients: (1 not defined because of singularities)

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

(Intercept) -3.157787 0.024346 -129.706 < 2e-16 \*\*\*

A5.9 0.716228 0.024146 29.663 < 2e-16 \*\*\*

A10.14 1.121537 0.025892 43.316 < 2e-16 \*\*\*

A15.17 NA NA NA NA

sNumHospitals 0.137774 0.003907 35.263 < 2e-16 \*\*\*

sPercentLessHS -0.246375 0.016893 -14.584 < 2e-16 \*\*\*

sPercentHS 0.043294 0.019610 2.208 0.0273 \*

sMedianIncome -0.265967 0.014936 -17.807 < 2e-16 \*\*\*

sAvgDistS 0.067103 0.013686 4.903 9.44e-07 \*\*\*

sAvgDistP -0.093793 0.016163 -5.803 6.52e-09 \*\*\*

---

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

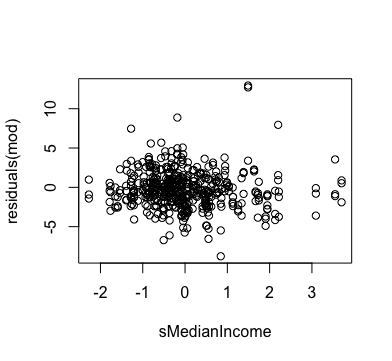
* As we can see in the p-values above, all variables except sPercentHS is significant at α = 0.01.
* The p-value for sPercentHS is 0.0273.

**Question 4: Goodness of Fit.**

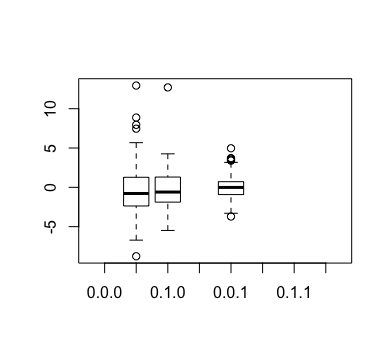
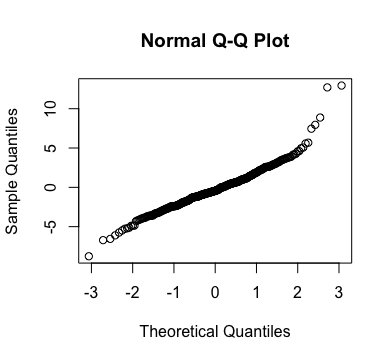
(i) Perform a goodness of fit test for the model? Does the model fit the data well?

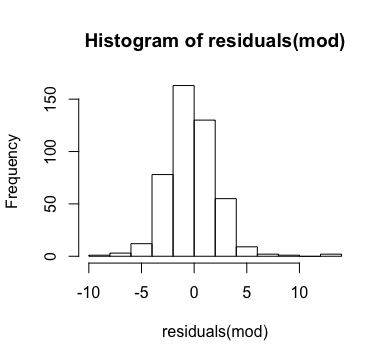
* The deviance test gives a p-value of zero.
* The GOF test using Pearson’s residuals also returns a p-value of zero.
* Thus, we reject the null hypothesis and conclude that the model is NOT a good fit.

(ii) Residual analysis



- Residuals appear uncorrelated to each other with no particular trend.

1.   
   - The left-most column indicates ages 5~9, the middle column ages 10~14, and the right-most column ages 15~17.  
   - The median residual value appears to be similar across different age groups; the variability seems higher for younger ages.
2.   
   - The Normal Q-QPlot shows that the residuals deviate slightly from normality, particularly for higher values.

  
- The histogram shows a bell-curve shape similar to that of a normal distribution.

**Question 6: Prediction**

* g(p) = -3.157787 + 0.137774\* 0.06089483 -0.246375\*0.1073684 + 0.043294\* 0.3555643 -0.265967\*(-0.5471714) + 0.067103\* 0.8022921 – 0.093793\* (-0.1179746) = -2.95
* Note that the given inputs were standardized before entering in the model, since our model was based on standardized variables.
* To estimate the probability, we perform the inverse logit function on g(p) and get 0.04973651.
* Thus, the probability of having an ED visit for a 16 year old asthma patient living in the particular county is 0.04973651 or approximately 4.97%.