Lab 5: HPV Vaccination Completion Rates (Logistic Regression)

Merhan Ghandehari

March 13, 2016

Introduction

Gardasil is a vaccine that has recommended for all women aged 9–26 in order to be protected from HPV (human papillomavirus) virus. This vaccin should be given three times within a 6–12 month timespan. In this research we are going to investigate the effects of different factors that can result in failure to complete the three-shot sequence. That is, we want to find "good" predictors for regimen completion in order to answer the following question:

- Which groups of patients appear to have a higher rate of completion? (this groups are defined by age, race and urban or suburban clinics)
- Which variables and patient characteristics (e.g., location of clinic, insurance type, and type of practice) best predict Gardasil vaccination completion?

Theoretically race and socioeconomic status can have a strong effect on the completion rate. For example, it seems that women who leave in poor, minority-heavy communities and among those lacking health insurance are less likely to complete the three-shot sequence within the 12-month period. So patient demographics, socioeconomic status, and care physician characteristics are some factors that can be examined in this research. Also we want to examin this hypotheis that whether patients who receive medical assistance and/or go to urban clinics are more likely to fail to complete the regimen than those who have some sort of insurance and/or go to suburban clinics.

Methods

Our dependent variable in this lab (i.e., completion of the HPV vaccination) was a binary variable (yes/no outcome). So, regression model was used as an appropriate method for modeling a binary dependent variable. All of the independent variales, except age, were categorical and we converted them to factor. In addition, we created dummy variables bacuase our categorical variabels have several sub-categories, and we are interested to to assess each subcategory and if significant use it in the multivariate logistic regression. Also we noticed that there is a descrepency between the variable "completed"" and "number of shots". So we decided to correct the completed vaiable based upon the number of shots that the patients completed.

First of all the impact of our categorical variables, as independent variables, on vaccination completion was examined by using tables, plots, and running a simple logistic regression. For each categorical variable, we created a table (e.g., `table(gardasil$Completed,gardasil$Location)`), and then we calculated the percentages for each column (`prop.table(locs, 2) *100`). This table helps a lot to find the most significant elements of each variable. After making simple logistic regression, probabilities, odds ratios and their confidence interval were calculated and assessed. Also anova test was run. I also ploted the fitted values (e.g., `plot(gardasil$Location, fitted(glmLocation))`) that was quite useful to find what element of each variable is statisticly significant. I did not run a univariate logistic regression on each individual dummy variable, because it was quite streightforward to find the most effective elements of each group based on the above-mentioned methods and I would explian the details in the result section.

After finding the significant variables, I fit a multivariate logistic regressions to the most interesting variables I found in the previous step. First I compared this model to the simple regression models. Although some of

the variables seems significant in the context of a simple model, but they are not signficant in a full model. So we can conclude that a simple model cannot explain the whole variation in out dependent variable. Then I eliminated the non-significant variables and assesed the reduced moldels. I used `AIC`, `anova(fit.reduced1, test = "LRT")`, `drop1(fit.reduced2, test = "LRT")`, and `anova(fit.reduced1, fit.reduced2, test = "Chisq")` to assess the models. I make four reduced model to arrive at my final model.

Finally I examined my model for any possible interaction affects. I tested individual interactions to see which one improve the model using summray of logistic model (AIC) and annova test. Then I added some interaction terms to my model.

Results:

This section includes some of the R code, tables, and plots that best summarize the outputs.

1 - We calulated the total percentage of completed versus incompleted vaccines. The resulsts shows that `66%` of wemen have not gotten the full vaccination and only `33%` have succeded to complete the vaccination.

```
# percentage of completed vs incompleted vaccines
summary(gardasil$Completed) / sum(summary(gardasil$Completed))
#  No       Yes
# 0.6680821 0.3319179
```

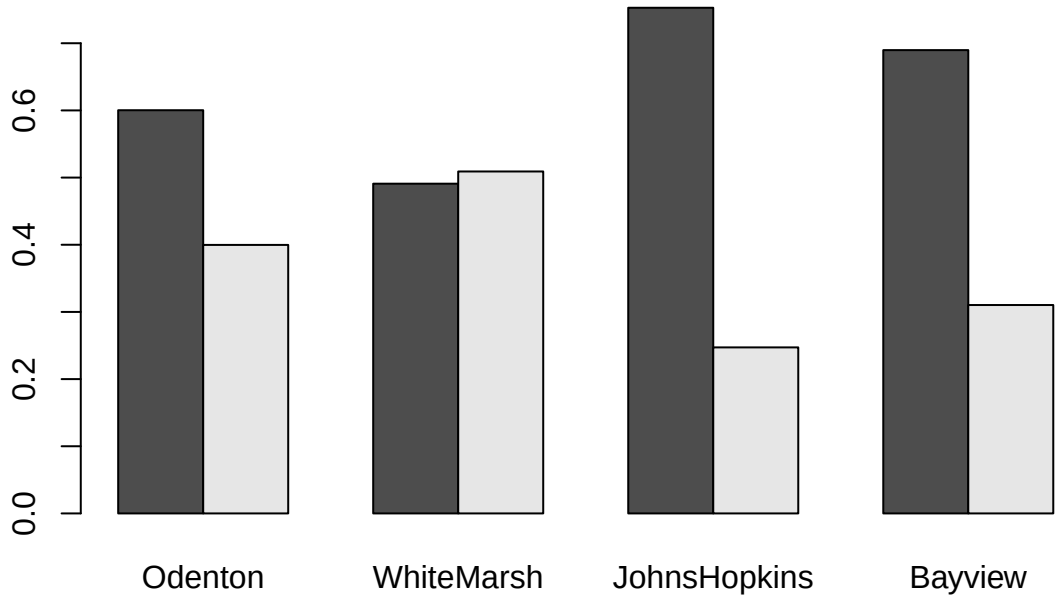
2 - Here we have the Bivariate Logistic Regression Models for each variable:

2-1- completion rates by location

```
#completion rates by location
locs= table(gardasil$Completed,gardasil$Location)
kable(prop.table(locs, 2) *100, digits = 1) # column percentages
```

	Odenton	WhiteMarsh	JohnsHopkins	Bayview
No	60	49.1	75.3	69
Yes	40	50.9	24.7	31

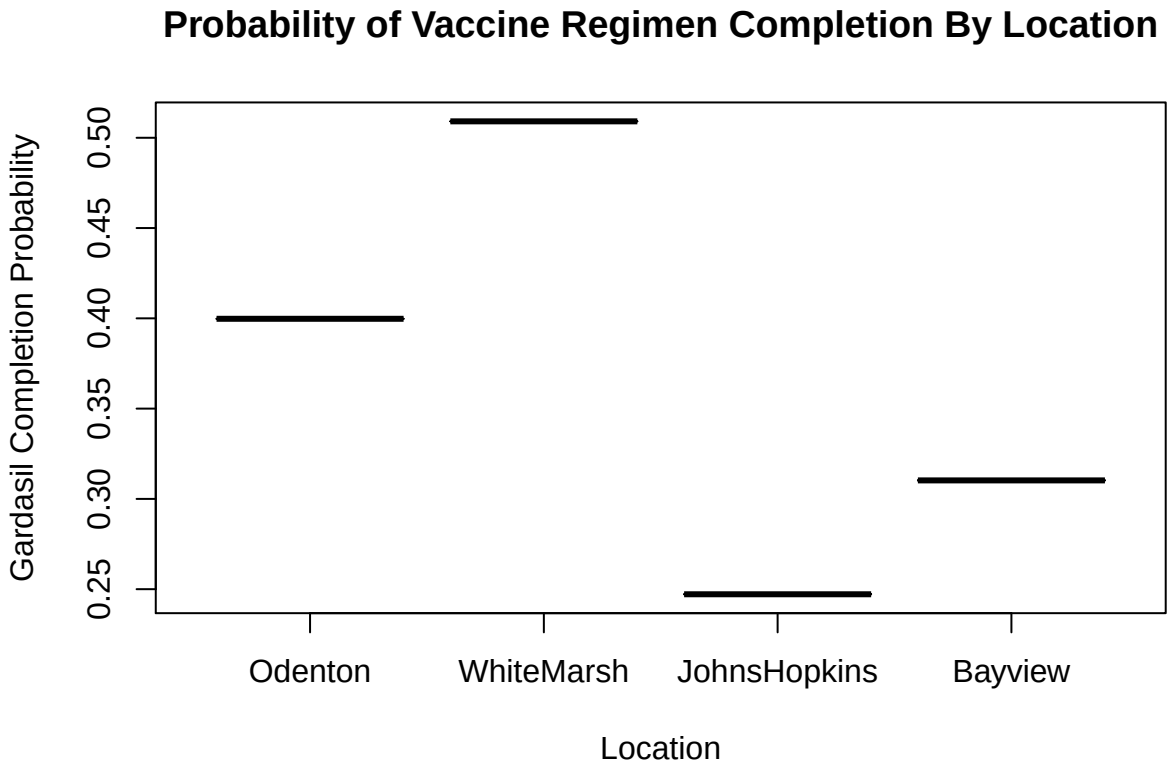
```
barplot(prop.table(locs, 2), beside=TRUE)
```




```
glmLocation=glm(gardasil$Completed~gardasil$Location, family=binomial)
kable(exp(cbind(OR=coef(glmLocation),confint(glmLocation))), digits = 4)
```

	OR	2.5 %	97.5 %
(Intercept)	0.6660	0.5776	0.7668
gardasil$LocationWhiteMarsh	1.5572	1.1122	2.1818
gardasil$LocationJohnsHopkins	0.4931	0.2924	0.8018
gardasil$LocationBayview	0.6754	0.5174	0.8780

```
#summary(glmLocation)
#anova(glmLocation, test = "LRT")
plot(gardasil$Location, fitted(glmLocation),
     main = "Probability of Vaccine Regimen Completion By Location",
     xlab = "Location", ylab = "Gardasil Completion Probability")
```



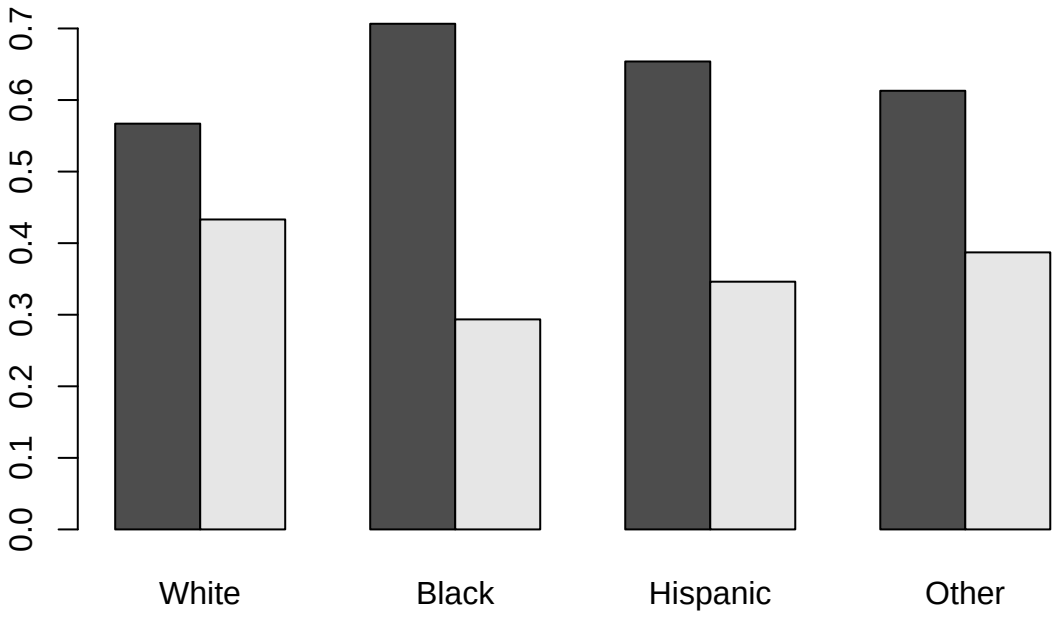

```
# Based on the results we can conclude that patients who go to the Johns Hopkins clinic have the lowest
```

2-2- Completion rates by Race

```
race= table(gardasil$Completed,gardasil$Race)
kable(prop.table(race, 2) *100, digits = 1) # column percentages
```

	White	Black	Hispanic	Other
No	56.7	70.7	65.4	61.3
Yes	43.3	29.3	34.6	38.7

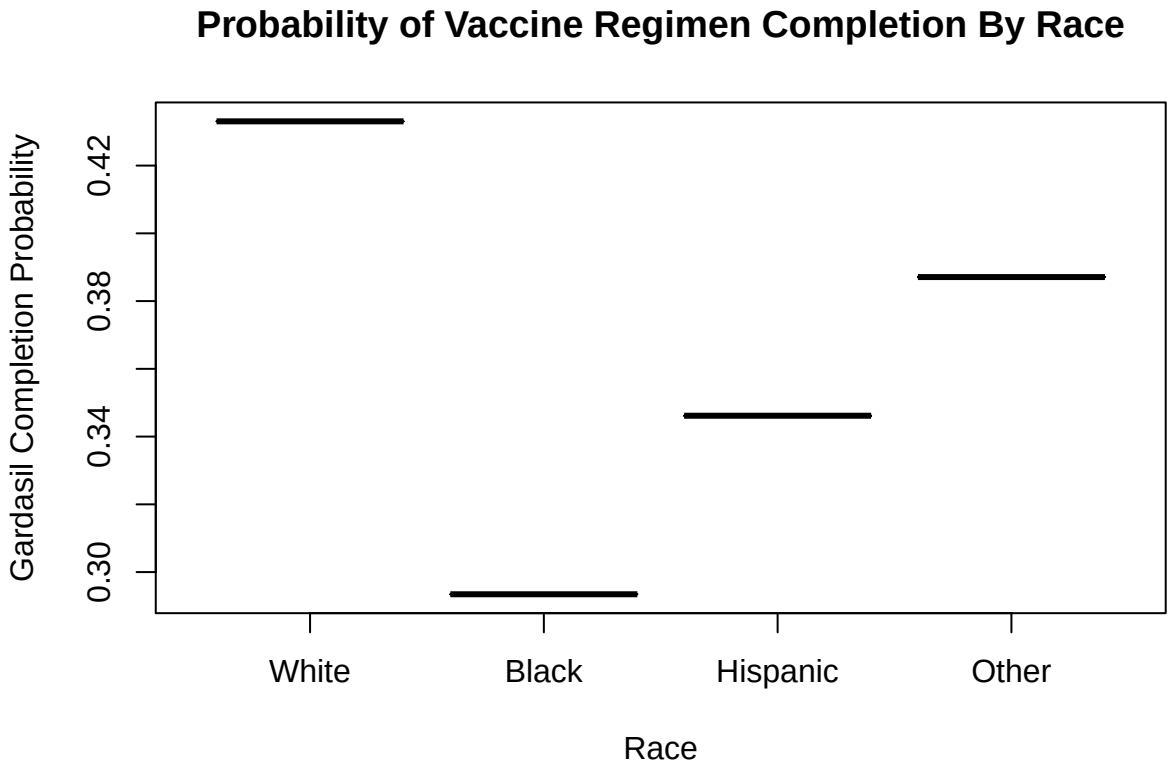
```
barplot(prop.table(race, 2), beside=TRUE)
```




```
glmRace=glm(gardasil$Completed~gardasil$Race, family=binomial)
kable(exp(cbind(OR=coef(glmRace),confint(glmRace))), digits = 4)
```

	OR	2.5 %	97.5 %
(Intercept)	0.7639	0.6596	0.8837
gardasil$RaceBlack	0.5437	0.4221	0.6980
gardasil$RaceHispanic	0.6931	0.3768	1.2349
gardasil$RaceOther	0.8268	0.5931	1.1468

```
#summary(glmRace)
plot(gardasil$Race, fitted(glmRace),
     main = "Probability of Vaccine Regimen Completion By Race",
     xlab = "Race", ylab = "Gardasil Completion Probability")
```



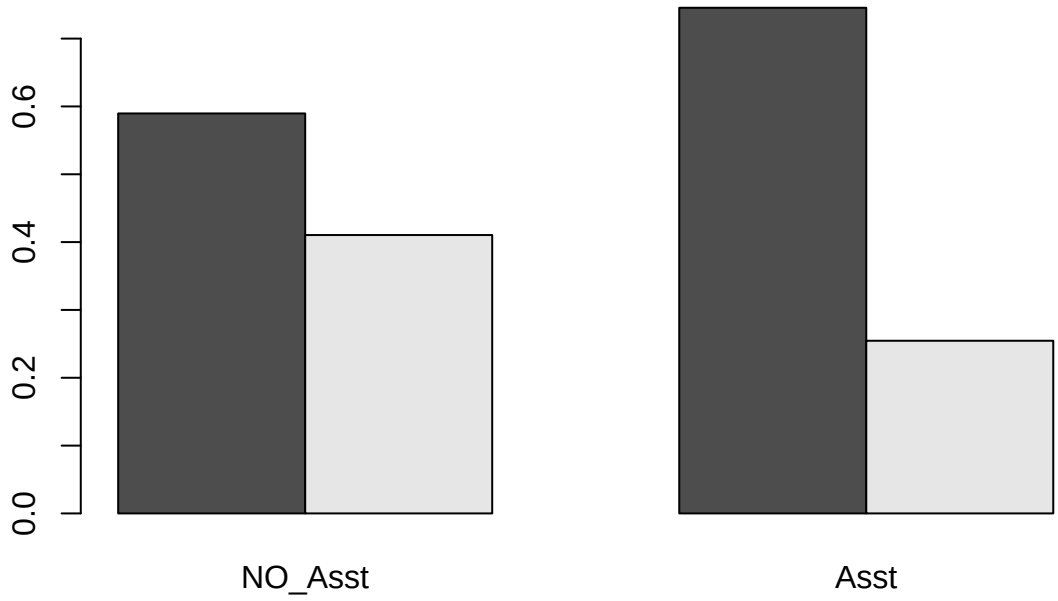

```
# Based on the results, the race black is a significant predictor. The odds ratio of black is .54 and t
```

2-3- Completion rates by Medical Assistance

```
MedAssist= table(gardasil$Completed,gardasil$MedAssist)
kable(prop.table(MedAssist, 2) *100, digits = 1) # column percentages
```

	NO_Asst	Asst
No	59	74.5
Yes	41	25.5

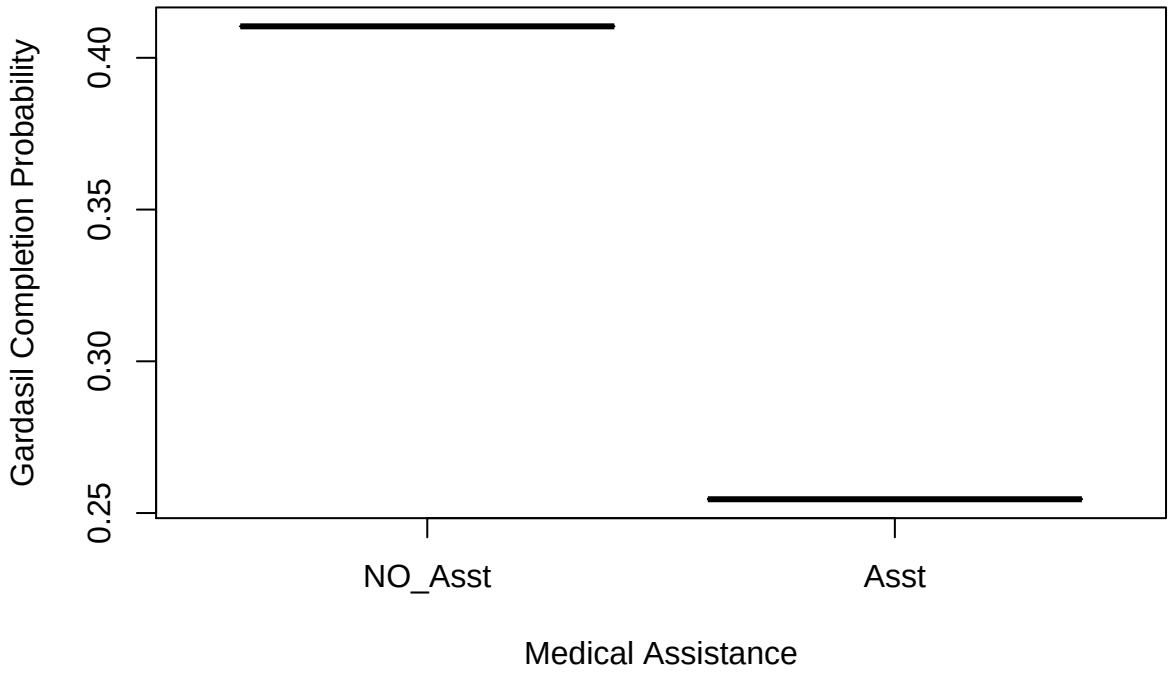
```
barplot(prop.table(MedAssist, 2), beside=TRUE)
```




```
glmMedAssist=glm(gardasil$Completed~gardasil$MedAssist , family=binomial)
kable(exp(cbind(OR=coef(glmMedAssist),confint(glmMedAssist))), digits = 4)
```

	OR	2.5 %	97.5 %
(Intercept)	0.6960	0.6182	0.7829
gardasil$MedAssistAsst	0.4906	0.3630	0.6565

```
#summary(glmMedAssist)
plot(gardasil$MedAssist, fitted(glmMedAssist),
     main = "Probability of Vaccine Regimen Completion by Medical Assistance",
     xlab = "Medical Assistance", ylab = "Gardasil Completion Probability")
```




```
# It seems that the patients who use the medical assistace are less likely to finish the Gardasil regim
```

2-4- Completion rates by Location Type

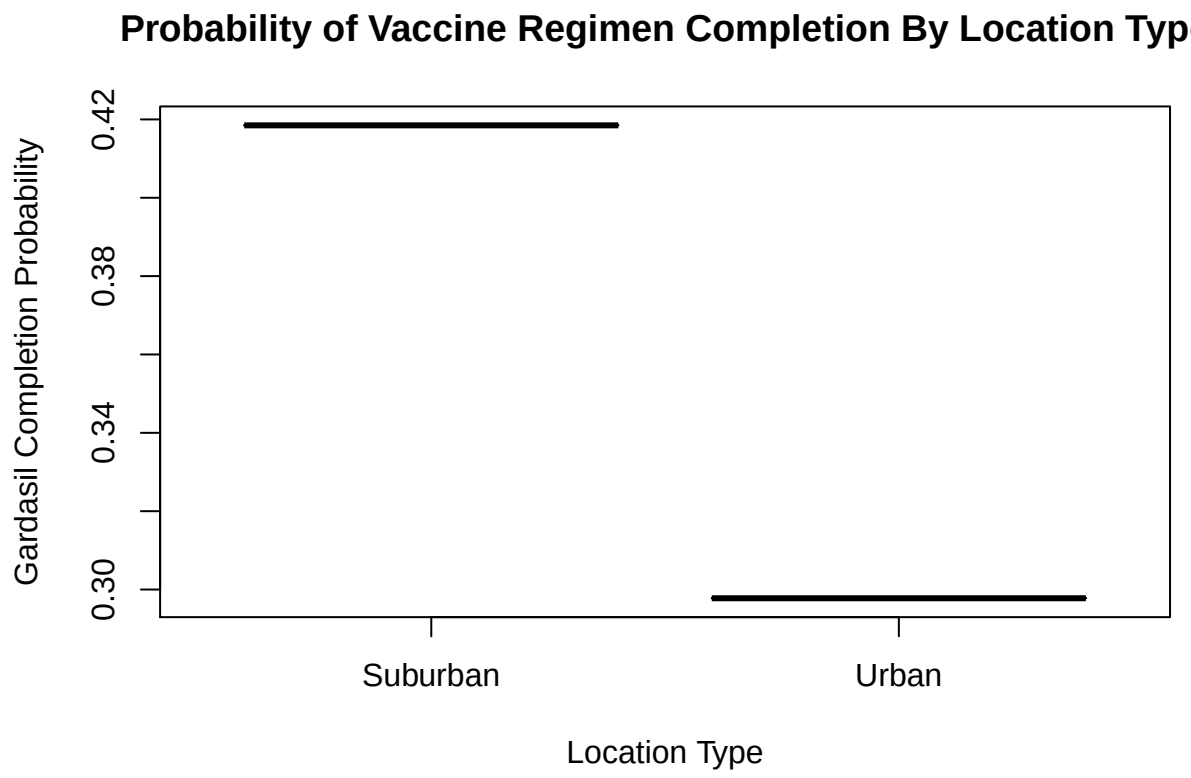
```
LocationType= table(gardasil$Completed,gardasil$LocationType)
kable(prop.table(LocationType, 2) *100, digits = 1) # column percentages
```

	Suburban	Urban
No	58.2	70.2
Yes	41.8	29.8

```
glmLocationType=glm(gardasil$Completed~gardasil$LocationType , family=binomial)
kable(exp(cbind(OR=coef(glmLocationType),confint(glmLocationType))), digits = 4)
```

	OR	2.5 %	97.5 %
(Intercept)	0.7196	0.6329	0.8176
gardasil$LocationTypeUrban	0.5893	0.4630	0.7472

```
#summary(glmLocationType)
plot(gardasil$LocationType, fitted(glmLocationType),
     main = "Probability of Vaccine Regimen Completion By Location Type",
     xlab = "Location Type", ylab = "Gardasil Completion Probability")
```




```
# Location type is also a significant predictor. Those who go to to the urban clinics are less likely f
```

2-5- Completion rates by Insurance Type

```
InsuranceType= table(gardasil$Completed,gardasil$InsuranceType)
kable(prop.table(InsuranceType, 2) *100, digits = 1) # column percentages
```

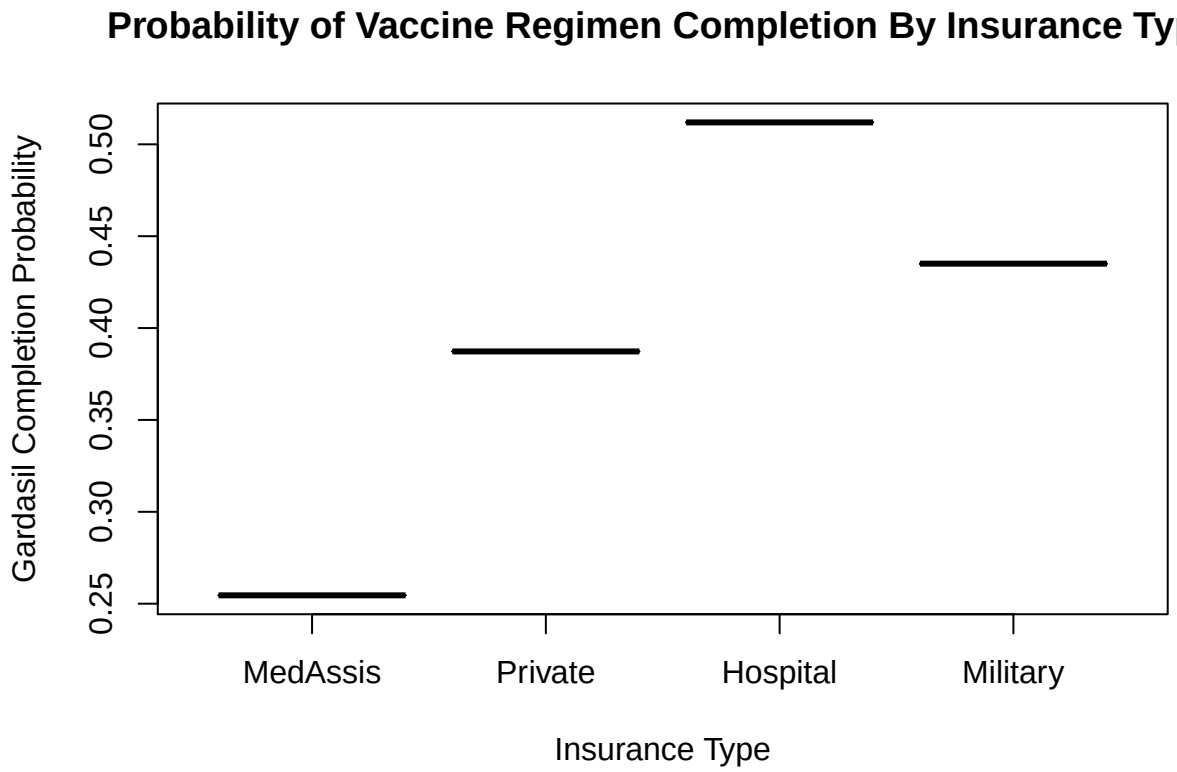
	MedAssis	Private	Hospital	Military
No	74.5	61.3	48.8	56.5
Yes	25.5	38.7	51.2	43.5

```
glmInsuranceType=glm(gardasil$Completed~gardasil$InsuranceType , family=binomial)
kable(exp(cbind(OR=coef(glmInsuranceType),confint(glmInsuranceType))), digits = 4)
```

	OR	2.5 %	97.5 %
(Intercept)	0.3415	0.2586	0.4454
gardasil$InsuranceTypePrivate	1.8510	1.3634	2.5355
gardasil$InsuranceTypeHospital	3.0714	1.8528	5.1144
gardasil$InsuranceTypeMilitary	2.2552	1.5977	3.2042

```
#summary(glmInsuranceType)
plot(gardasil$InsuranceType, fitted(glmInsuranceType),
     main = "Probability of Vaccine Regimen Completion By Insurance Type",
```

```
    xlab = "Insurance Type", ylab = "Gardasil Completion Probability")
```




```
# As we saw before Medical Assistance is a significant predictor. It seems that PrivatePayer is also ca
```

2-6- Completion rates by Practice Type

```
#completion rates by Practice Type
PracticeType= table(gardasil$Completed,gardasil$PracticeType)
kable(prop.table(PracticeType, 2) *100, digits = 1) # column percentages
```

	Pediatric	FamilyPrac	OB_GYN
No	67.1	58.9	NaN
Yes	32.9	41.1	NaN

```
glmPracticeType=glm(gardasil$Completed~gardasil$PracticeType , family=binomial)
kable(exp(cbind(OR=coef(glmPracticeType),confint(glmPracticeType))), digits = 4)
```

	OR	2.5 %	97.5 %
(Intercept)	0.4898	0.3925	0.6078
gardasil$PracticeTypeFamilyPrac	1.4240	1.0793	1.8838

```
#summary(glmPracticeType)

# only 32% of the Pediatric group complete the vaccination. So, Pediatric group can be an intersting du
```

2-7- Completion rates by AgeGroup

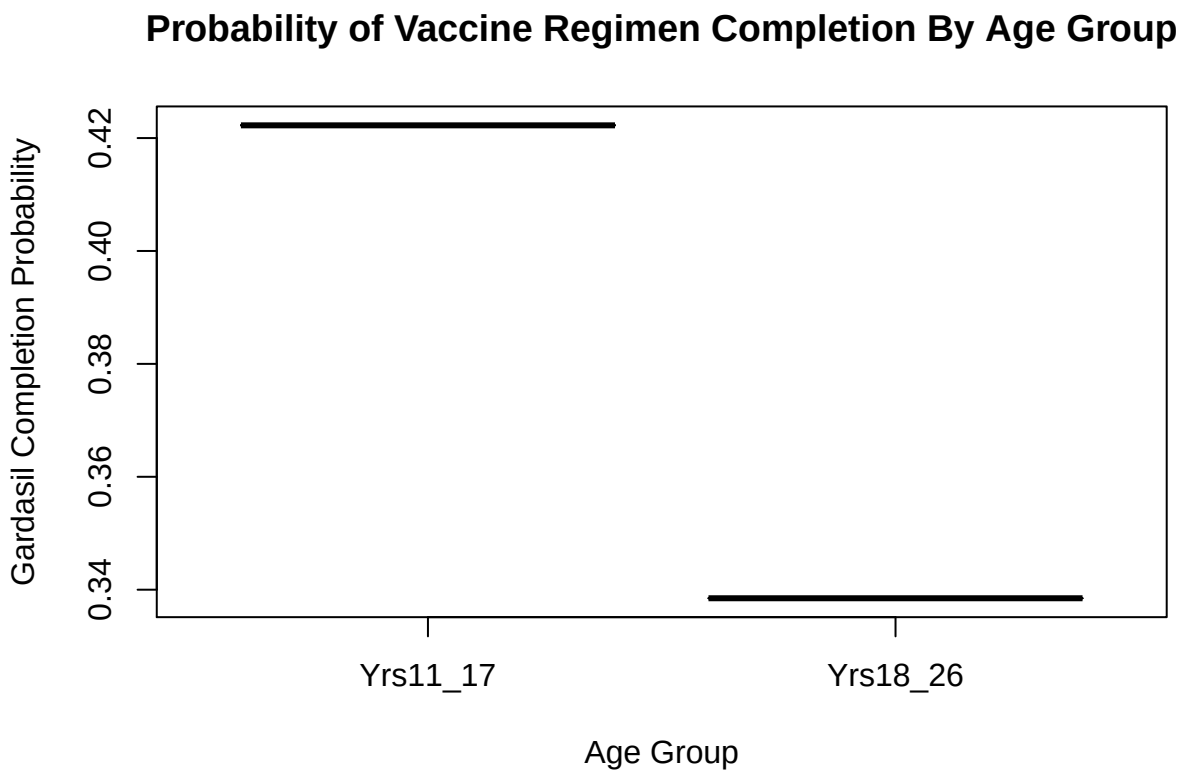
```
AgeGroup= table(gardasil$Completed,gardasil$AgeGroup)
kable(prop.table(AgeGroup, 2) *100, digits = 1) # column percentages
```

	Yrs11_17	Yrs18_26
No	57.8	66.2
Yes	42.2	33.8

```
glmAG=glm(gardasil$Completed~gardasil$AgeGroup, family=binomial)
kable(exp(cbind(OR=coef(glmAG),confint(glmAG))), digits = 4)
```

	OR	2.5 %	97.5 %
(Intercept)	0.7309	0.6287	0.8486
gardasil$AgeGroupYrs18_26	0.7001	0.5640	0.8684

```
#summary(glmAG)
plot(gardasil$AgeGroup, fitted(glmAG),
     main = "Probability of Vaccine Regimen Completion By Age Group",
     xlab = "Age Group", ylab = "Gardasil Completion Probability")
```




```
# The Probability of Gardasil Vaccine Regimen Completion does not appear to be significantly affected b
```

2-8- Completion rates by Age

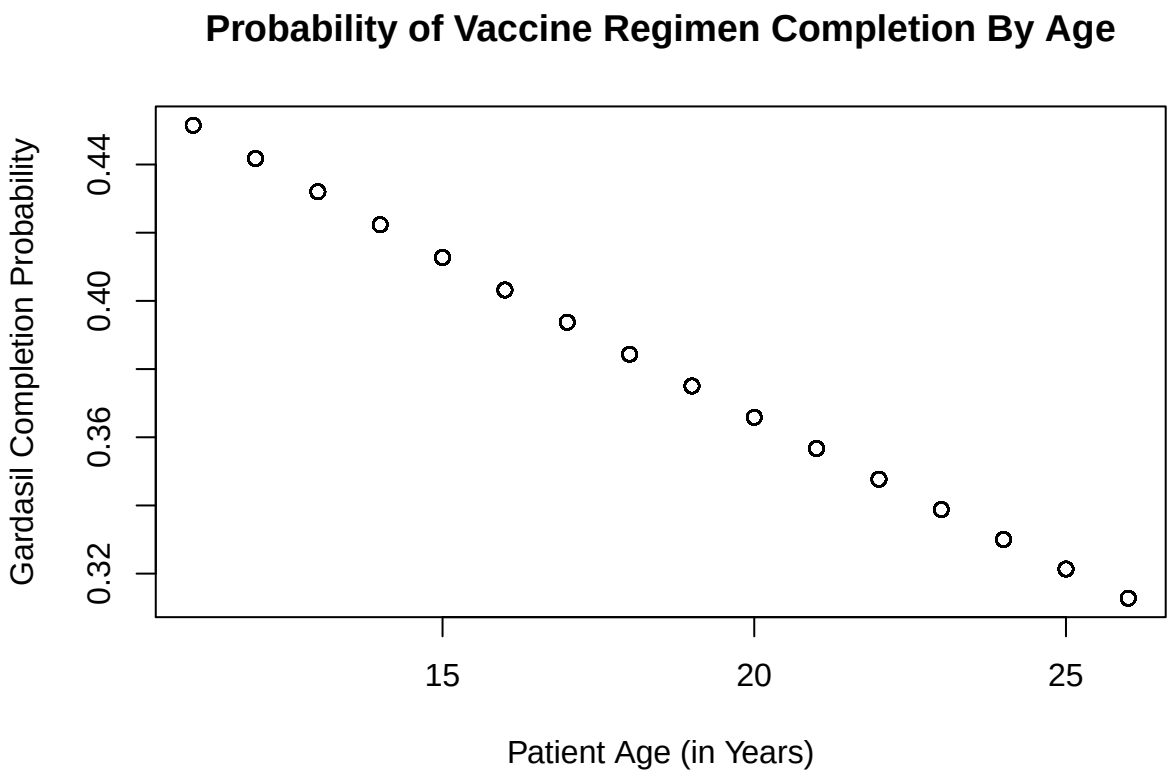
```
Age= table(gardasil$Completed,gardasil$Age)
kable(prop.table(Age, 2) *100, digits = 1) # column percentages
```

	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
No	53.3	60	55.4	50.6	56.4	53	64.1	65.7	70.5	71.7	66.3	65.6	60	64.9	69.8	62.2
Yes	46.7	40	44.6	49.4	43.6	47	35.9	34.3	29.5	28.3	33.7	34.4	40	35.1	30.2	37.8

```
glmAge=glm(gardasil$Completed~gardasil$Age, family=binomial)
kable(exp(cbind(OR=coef(glmAge),confint(glmAge))), digits = 4)
```

	OR	2.5 %	97.5 %
(Intercept)	1.2711	0.7841	2.0618
gardasil$Age	0.9613	0.9368	0.9862

```
#summary(glmAge)
plot(gardasil$Age, fitted(glmAge), main = "Probability of Vaccine Regimen Completion By Age",
     xlab = "Patient Age (in Years)", ylab = "Gardasil Completion Probability")
```




```
# Probability of completion does  appear to be affected by age. There is a negative (or inverse) relati
```

3- Multivariate Logistic Regression Models: We begin with a model with all of the parametes that were significant in our bivariate logistic regression models.

```
# Null hypothesis: the probability of a patient's vaccine completion is no better than the average prob
fit.full = glm(Completed ~ MedAssist + Black + LocationType + Pediatric + PrivatePayer + gardasil$Age +
                HospitalBased + Hispanic + FamilyPractice, data=gardasil, family="binomial")
#summary (fit.full)
#anova(fit.full, test = "LRT")

#AIC: 1806.9

# because both Johns Hopkins and Bayview were significant and urban, I just used LocationType (urban =

#From the p-values for the regression coefficients (last column), you can see that PrivatePayer and His

#lmtest:::lrtest(fit.full)
#The chi-square of 89.825 with 9 degrees of freedom and an associated p-value of significantly less tha


fit.reduced1 = glm(Completed ~ MedAssist + Black + LocationType + gardasil$Age +
                HospitalBased + FamilyPractice + Pediatric, data=gardasil, family="binomial")
#summary (fit.reduced1)
#anova(fit.reduced1, test = "LRT")
#anova(fit.full, fit.reduced1, test = "Chisq")
# AIC: 1803
# Pediatric is the least significant for both the Wald and LRT tests. We Will eliminate them.
```

```r
#The nonsignificant chi-square value (p = 0.9384) suggests that the reduced model fits as well as the f


fit.reduced2 = glm(Completed ~ MedAssist + Black + LocationType + gardasil$Age +
                HospitalBased + FamilyPractice, data=gardasil, family="binomial")
#summary (fit.reduced2)
#anova(fit.reduced2, test = "LRT")
# AIC: 1803.6
# Each regression coefficient in this reduced model is statistically significant
# The fit.reduced2 model has a similar AIC to mylogit2, so keep it.

#drop1(fit.reduced2, test = "LRT")
# Null hypothesis: There is no difference in the deviance between the two models.
#anova(fit.reduced1, fit.reduced2, test = "Chisq")
# All of the variables are sygnificant. So, we do not drop any variabels from our model. Also anova tes

#interaction model
fit.reduced3 = glm(Completed ~ MedAssist + Black + Age +
                 FamilyPractice + HospitalBased * LocationType, data=gardasil, family="binomial")
#summary (fit.reduced3)
#anova(fit.reduced3, test = "LRT")
#AIC: 1802.8
#anova(fit.reduced2, fit.reduced3, test = "Chisq")

fit.reduced4 = glm(Completed ~ MedAssist + Black + Age * LocationType +
                     FamilyPractice + HospitalBased * LocationType, data=gardasil, family="binomial")
summary (fit.reduced4)
```

```
##
## Call:
## glm(formula = Completed ~ MedAssist + Black + Age * LocationType +
##     FamilyPractice + HospitalBased * LocationType, family = "binomial",
##     data = gardasil)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8051  -0.9435  -0.7345   1.1905   1.8924
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       2.05222    0.33605   6.107 1.02e-09 ***
## MedAssistAsst                    -0.39745    0.19770  -2.010 0.044390 *
## Black1                           -0.50138    0.13082  -3.833 0.000127 ***
## Age                              -0.10963    0.01702  -6.442 1.18e-10 ***
## LocationTypeUrban                -3.18285    0.57927  -5.495 3.92e-08 ***
## FamilyPractice1                  -0.54782    0.14266  -3.840 0.000123 ***
## HospitalBased1                    1.11289    0.36554   3.045 0.002330 **
## Age:LocationTypeUrban             0.14794    0.03046   4.857 1.19e-06 ***
## LocationTypeUrban:HospitalBased1 -1.01336    0.49823  -2.034 0.041958 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 1876.7  on 1412  degrees of freedom
## Residual deviance: 1763.3  on 1404  degrees of freedom
## AIC: 1781.3
##
## Number of Fisher Scoring iterations: 4
```

```
anova(fit.reduced3, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Completed
##
## Terms added sequentially (first to last)
##
##
##                            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                         1412     1876.7
## MedAssist                   1  23.8681      1411     1852.8 1.032e-06 ***
## Black                       1  12.8110      1410     1840.0 0.0003446 ***
## Age                         1  21.7741      1409     1818.2 3.067e-06 ***
## FamilyPractice              1  10.8816      1408     1807.4 0.0009712 ***
## HospitalBased               1   3.1822      1407     1804.2 0.0744458 .
## LocationType                1  14.5680      1406     1789.6 0.0001352 ***
## HospitalBased:LocationType  1   2.8256      1405     1786.8 0.0927747 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
kable(exp(cbind(OR=coef(fit.reduced4),confint(fit.reduced4))), digits = 4)
```

	OR	2.5 %	97.5 %
(Intercept)	7.7852	4.0491	15.1304
MedAssistAsst	0.6720	0.4552	0.9891
Black1	0.6057	0.4678	0.7815
Age	0.8962	0.8665	0.9263
LocationTypeUrban	0.0415	0.0132	0.1285
FamilyPractice1	0.5782	0.4364	0.7637
HospitalBased1	3.0431	1.5160	6.4360
Age:LocationTypeUrban	1.1594	1.0923	1.2310
LocationTypeUrban:HospitalBased1	0.3630	0.1336	0.9492

```
"AIC: 1781.3
The fit.reduced4 model has the smallest AIC. Adding the (HospitalBased * LocationType) interaction term
This model shows that Race and Insurance Type are good predictors of Gardasil completion
Based on our final model, we can say that MedAssist, Black, Age, LocationType, FamilyPractice, and Hosp
```

```
## [1] "AIC: 1781.3\nThe fit.reduced4 model has the smallest AIC. Adding the (HospitalBased * LocationTy
```

Discussion:

The results of a bunch of univariate and multivariate logistic regressions illustarates that the probability (likelihood) of completed HPV vaccination is much higher in wemen who are white, see an OB-GYN, go to suburban clinics, and do not have medical assistance. Also having a hospital insurance and being in the age group of 11 to 17 are effective parameters that help to complete HPV vaccination. On the other hand, wemen who are black or hispanic, have medical assistance or private Insurance, go to urban clinics, and go to pediatric are less likely to complete HPV vaccination. Therfore, we can conclude that patients of color and low socioeconomic status are less likely to have a completed HPV vaccination. This issue is very obvious when we can see that the white wemon are twice as likely to complete the HPV vaccine than black wemen. I think all of the variables in the final model makes sense as a strong predictor of our dependent variable and I do not see any unpredictable predictor in our model.

Another intesting result was younger females have a higher completion rate than older females. The only reason that I can guess for this observation is that parents have significant role in completion rate by taking care of their children's health. Also the higher rate of completion in suburban clinics than urban clinics is due to the bias that we see in race and socioeconomic status. socioeconomic status can the main reason of low completion rate in medical assistance insurance. But, the low quality of medical assistance insurance is another point that worth to mention.

One of the problems that I have for making inference out of the data that we have in this lab was the unequal number of wemen in each group (sample size). For example, the number reported Hispanic wemen was mush less than the other gropus of wemen, and changing the size of this group may completly change the result of our investigation.