Stat 402 – Lecture Seven
Professor Esfandiari

The objective of this lecture is to…

- Introduce you to logistic regression as a special case of "GLM" or general linear model.
- Show you how to conduct logistic regression in R and interpret the output.
- Discuss interaction and its interpretation in logistic regression.
- Discuss the concept of deviance in logistic regression.
- Discuss "goodness of fit" tests in logistic regression.
- Discuss concept of residuals and relevant residual plots.
- Discuss influential points and mmp plots

I Introduction

In multiple linear regression, the outcome variable was numerical and the predictors could be categorical, numerical, or a combination of categorical with categorical and categorical with numerical.

One of the central advances in statistics during the second half of the twentieth century was the development of the ***general liner models (GLMS), and their most important special case, logistic regression. The well understood linear model was extended to problems in which the response is categorical or discrete rather than a continuous numerical variable.***

The structure of GLM is very similar to that of the linear model we discussed before. We are interested in understanding how the mean of Y varies as the values of the predictor change. **In multiple linear regression the outcome can take any values from minus infinity (-∞) to Plus infinity (+ ∞). Whereas, in logistic regression the outcome variable can only assume values between zero and one; with "1" = "success", and "0" = "failure".**

The other difference between multiple linear regression and logistic regression is that in multiple linear regression we use **ordinary least square regression (OLS)** to estimate the unknown parameters, with the goal of minimizing the sum of the square of the residuals or the differences between the observed and the predicted responses. **Whereas, in logistic regression maximum likelihood equations are derived from the probability distribution of the dependent variables** and solved using the Newton- Raphson method for nonlinear systems of equations. For details see the paper by Scott Czepiel posted on week seven of CCLE.

II An example of logistic regression with a single numerical predictor

So far we have talked about situations in which the outcome variable is quantitative. In this lecture we are going to talk about a situation in which the outcome variable is binomial. **When the outcome variable is binomial, the regression method used is referred to as "logistic regression".**

We are going to predict a binomial random variable Y based on a single predictor variable x via logistic regression. The outcome variable is binomial.

Properties of a binomial distribution

- There are m identical trials
- Each trial results in only two possibilities; either a success "S" or failure "F"
- The probability of success or (θ) is the same for all of the trials
- Trails are independent of each other

$\boldsymbol{Y \sim Bin(m, \theta)}$

The probability that Y takes the integer values (j = 1, 2, …..,m) is given by…

$$\mathbf{P(Y = j) = \binom{m}{j} \theta^{j}(1-\theta)^{m-j} = \frac{m!}{j!\,(m-j)!}\theta^{j}(1-\theta)^{m-j} \quad (j = 1, 2, \dots, m)}$$

The mean and variance of Y (i.e., the binomial distribution) is given by

$$\mathbf{E(Y) = m\theta}$$

$$\mathbf{Var(Y) = m\theta(1-\theta)}$$

In the logistic linear setting, we consider…

- **θ to be "success"**
- $\boldsymbol{(1-\theta)}$ ***to be the*** **failure.**

Example. Suppose the binomial event is randomly guessing the answers to a multiple-choice test about which you know nothing; such as taking a test in another language. Additionally, assume that the test has four questions and each question has four options only one of which is correct. The possible list of outcomes and their relevant probabilities is given below.

m = 4, P(S) or P(θ)= ¼, P(F) or P(1 - θ)= 3/4

j	$\binom{m}{j}$	$\theta^j(1-\theta)^{m-j}$	$\frac{m!}{j!\,(m-j)!}\theta^j(1-\theta)^{m-j}$
0	$\binom{4}{0}=1$	$1/4^0(3/4)^4$=81/256	1* 81/256 = 81/256
1	$\binom{4}{1}=4$	$1/4^1(3/4)^3$=27/256	4* 27/256 =108/256
2	$\binom{4}{2}=6$	$1/4^2(3/4)^2$=9/256	6* 9/256 = 54/256
3	$\binom{4}{3}=4$	$1/4^3(3/4)^1$=3/256	4 * 3/256 = 12/256
4	$\binom{4}{4}=1$	$1/4^4(3/4)^0$=1/256	1 * 1/256 = 1/256

Total probability of all the possible evens =

81/256 + 108/256 + 54/256 + 12/256 + 1/256 = 256/256 = 1

The mean and variance of the above distribution are calculated as follows:

$E(Y) = m\theta = 4 * \frac{1}{4} = 1$

The variance of a binomial distribution is

$Var(Y) = m\theta(1-\theta) = 4 * \frac{1}{4} * \frac{3}{4} = 0.75$

Logistic Regression

In logistic regression, we model Y as a function of series of predictors including $x_1,\ x_2\ , \ldots , x_p$ which could be qualitative, quantitative, or a combined effect of qualitative with qualitative and qualitative with quantitative.

We will consider a case in which we only have a single predictor or x.

$$(Y|x_i) \sim Bin(m_i, \theta(x_i)) \quad i = 1, 2, \ldots, n$$

The sample proportion or "success" at each i is given by y_i/m_i.

$$E(y_i/m_i \,|x_i) = \theta(x_i)$$

$$Var((y_i/m_i \,|x_i) = \frac{\theta(x_i * (1-\theta(x_i))}{m_i}$$

Example from Sheather.

We are going to consider 164 French restaurants in New York that were included in the 2006 Zagat rating (Sheather page 265 – see chapter 8 – week seven). We are wondering if these restaurants are included in the Michelin Guide. You can find this data on week seven of CCLE.

The variables are

x_i(Food rating or the predictor variable). The rating is given by customers. This is a quantitative variable; the maximum is thirty

y_i = number of restaurants included in Michelin Guide
m_i = Total number of restaurants with a Zagat rating (164)

$m_i - y_i$ = number not included in Michelin Guide

$\frac{y_i}{m_i}$ = probability of success or probability of being included in the Michelin Guide

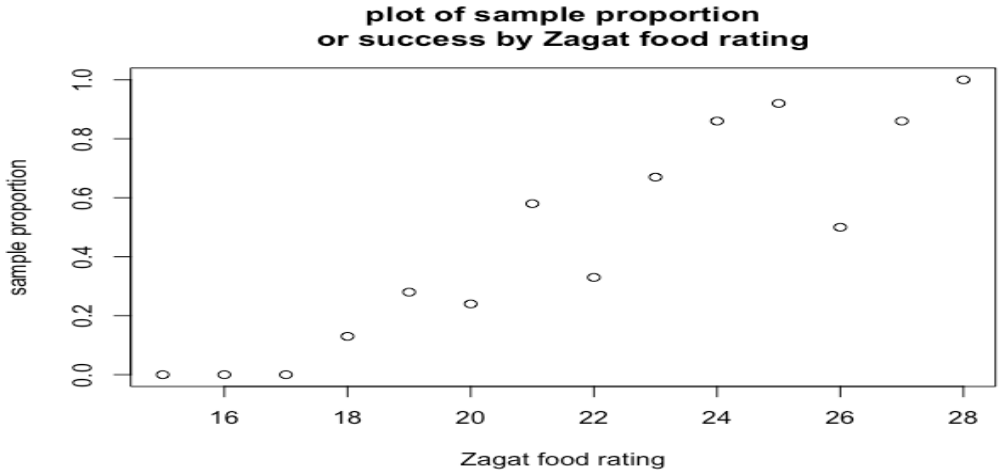
Example

- m_i = 12 restaurants that got a rating of (X = 22)
- y_i = 4; It means four restaurants were included in the Michelin Guide.
- $m_i - y_i = 8$; means eight restuaruatns were not included in the Michelin Guide
- $\frac{y_i}{m_i} = \frac{4}{12}$ = 0.33; means 4 out of 12 restaurants were listed in Michelin Guide
- (This is the observed probability of success or being rated in the Michelin Guide)

Table one. French restaurants in the Michelin Guide broken down by food

x_i Food rating predictor	y_i in Michelin	$m_i - y_i$	m_i Total number of restaurants	$\frac{y_i}{m_i} = probability\ of$ success
15	0	1	1	0.00
16	0	1	1	0.00
17	0	8	8	0.00
18	2	13	15	2/15 = 0.13
19	5	13	18	0.28
20	8	25	33	0.24
21	15	11	26	0.58
22	**4**	**8**	**12**	**0.33**
23	12	6	18	0.67
24	6	1	7	0.86
25	11	1	12	0.92
26	1	1	2	0.50
27	6	1	7	0.86
28	4	0	4	1.00
NA	Total=74	Total = 90	Total =164	NA

Using the data saved as "michelinfood.csv" posted on week seven of CCLE, we can draw the plot of Y (probability of success or probability of being included in the Michelin rating) as a function of Zagat food rating. This plot is given below.



As it is clear from the plot displayed above, **the shape of the underlying function, θ(x) is not a straight line. It is rather S-shaped with very small values of X associated with zero probability of "success" and high values of X associated with probability of "success" close to one.**

Logistic Function as the log of odds

It can be shown that…

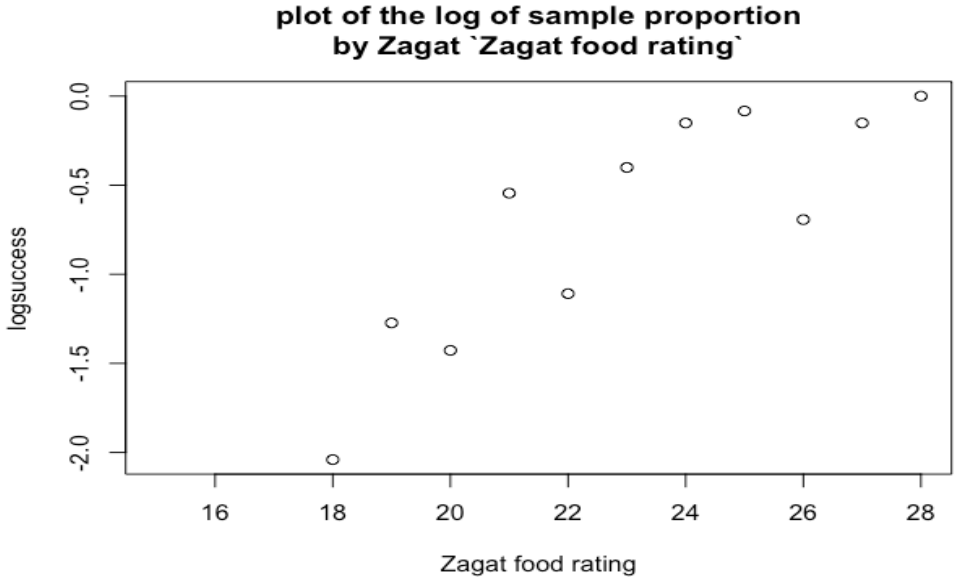
$$\left[log\left(\frac{\theta(x)}{1-\theta(x)}\right)\right] = \beta_0 + \beta_1 x$$

$$log\left(\frac{\theta(x)}{1-\theta(x)}\right) \text{ is called LOGIT}$$

If the chosen function is correct, then…

the plot of $log\left(\frac{\theta(x)}{1-\theta(x)}\right)$ against x will create a straight line. See below:

```
> logsuccess=log(`sample proportion`)
> plot(`Zagat food rating`,logsuccess, main = "plot of the log of sample proportion
+ by Zagat `Zagat food rating`")
```



The quantity $\frac{\theta(x)}{(1-\theta(X)}$ is known as odds

odds in favor of success = $\frac{P(success)}{(1-P(success)} = \frac{\theta}{1-\theta}$

odds against success = $\frac{1-P(success)}{P(success)} = \frac{1-\theta}{\theta}$

THUS, ODDS IN LOGISTIC REGRESSION ARE IN THE FORM OF ODDS IN FAVOR OF A "SUCCESS".

Let…

X = Zagat food rating

$\theta(x)$ ***The prability that restaurnt be included in the*** **Michelin guide**

Then, the logistic regression model for the response variable

$$\theta(x) \text{ as a function of } X \text{ is given by}$$

$$\theta(x) = \frac{1}{1 + exp(-\{\beta_0 + \beta_1 x)\}}$$

We will now use R to run the logistic regression model for the prediction of probability of being included in the Michelin guide or "success" as a function of Zagat food rating.

> m1<-glm(InMichelin~Food, family=binomial)
> summary(m1)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.84154	1.86236	-5.821	5.84e-09 ***
Food	**0.50124**	**0.08768**	**5.717**	**1.08e-08 *****

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 61.427 on 13 degrees of freedom
Residual deviance: 11.368 on 12 degrees of freedom
AIC: 41.491

Number of Fisher Scoring iterations: 4

$$\hat{\theta}_x = \frac{1}{1+exp(-\{\hat{\beta}_0+\hat{\beta}_1 x)\}} = \frac{1}{1+exp(-10.84+0.50124x\}}$$

$$log\left(\frac{\hat{\theta}_x}{1-\hat{\theta}_x}\right) = \hat{\beta}_0 + \hat{\beta}_1 x = -10.84 + 0.50124x$$

$$\left(\frac{\hat{\theta}_x}{1-\hat{\theta}_x}\right) = exp(\hat{\beta}_0 + \hat{\beta}_1 x) = \exp(-10.84 + 0.50124x)$$

Notice that the log of odds is a liner function of X.

Interpretation of the slope for Zagat food rating…

- **For one unit of increase in Zagat food rating, the log of the odds for being included in the Michelin guide increases by 0.50124.**

- **However, the log of odds is not easy to interpret. That is why we exponentiate it to get the odds ratio.**

```
> exp(0.50124)
[1] 1.650767
```

- **For one unit of increase in Zagat food rating, the odds of being included in the Michelin guide increases by 1.65.**

```
> exp(0.50124*5)
[1] 12.25826
```

- **For five units of increase in Zagat food rating, the odds of being included in the Michelin guide increases by 12.26.**

III Testing the parameters in logistic regression with one quantitative predictor

The standard approach to testing

$$\mathbf{H_0: \beta_1 = 0}$$

We will use the Wald test statistic to test the above null hypothesis

$$\mathbf{Z = \frac{\hat{\beta_1}}{\text{estimated standaard error of } (\hat{\beta_1})}}$$

The Wald statistic is then compared to the standard normal distribution to test for statistical significance.

For the above example

$$\mathbf{Z = \frac{0.50124}{0.08768} = 5.716697}$$

```
> 0.50124/0.08768
[1] 5.716697
```

```
> pnorm(5.72)
[1] 1
```

P-value is almost zero

In the example given above the Z value associated with the test of the slope is statistically significant (Z = 5.717, P ~ 0.000). **The interpretation was that with one unit of increase in X or Zagat food rating the odds of being included in the Michelin Guide increases by 1.7 and this is statistically significant.**

The confidence interval based on the Wald Statistic are given below

$$\hat{\beta_1} +/- Z_{1-\alpha/2} * \text{estimated standard error of } (\hat{\beta_1})$$

For the given example, the 95% confidence interval will be:

0.50124 +/- 1.96*0.08768 = (0.329, 0.673)

> 1.96*0.08768
[1] 0.1718528
> 0.50124-0.1718528
[1] 0.3293872
> 0.50124+0.1718528
[1] 0.6730928

We now exponentiate the 95% confidence interval

> exp(0.329)
[1] 1.389578
> exp(0.673)
[1] 1.960109

Based on odds ratios, the 95% confidence interval would be (1.40,1.96)

- **We are 95% confident that for one unit of increase in X or Zagat food rating the odds of being included in Michelin guide increase 1.40 to 1.96.**

- **Since the 95% confidence interval does not include one, we reject the null and conclude that the odds of being included in the Michelin guide is related to Zagat food rating.**

We can use the following command to estimate the coefficients as well as the confidence interval for the partial slopes. The reported results are similar to what was calculated within rounding error.

> round(exp(cbind(Estimate=coef(m1), confint(m1))),2)

	Estimate	**2.5 %**	**97.5 %**
(Intercept)	0.00	0.00	0.00
food	1.65	1.41	1.99

IV Example of Logistic Regression with multiple predictors and interaction effect between two categorical predictors

Research has shown that diabetic type II is related to age, smoking, and hypertension. We are going to be using a logistic regression model to predict the odds of having diabetic type II as a function of age, hypertension, smoking, and the combined effect smoking and hypertension. This data set is posted on week eight of CCLE.

Exploratory Data Analysis

> table(Diabetes.new)

```
Diabetes.new
  0  (No)                1 (Yes)
8143                      1805
```

> table(HypertensionDX)

```
        HypertensionDX

 No                       Yes
5776                      4172
```

> table(SmokingStatus_NISTCode)

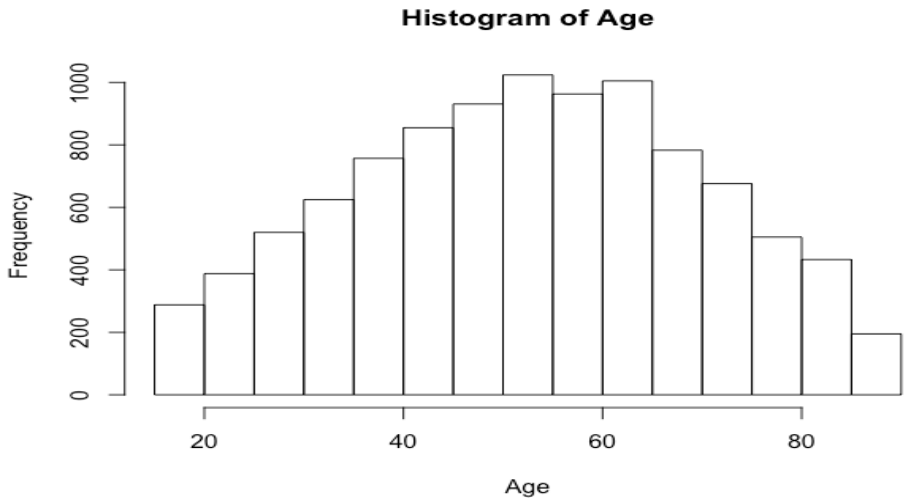
```
SmokingStatus_NISTCode
 FALSE              FORMER              TRUE
  1904               1888               611
```

> table(HypertensionDX,SmokingStatus_NISTCode)

```
                              SmokingStatus_NISTCode
HypertensionDX      FALSE              FORMER              TRUE
        no          987                996                 318
        yes          917                892                 293
```

Based on the tables given above, the frequencies in the given tables look reasonable.

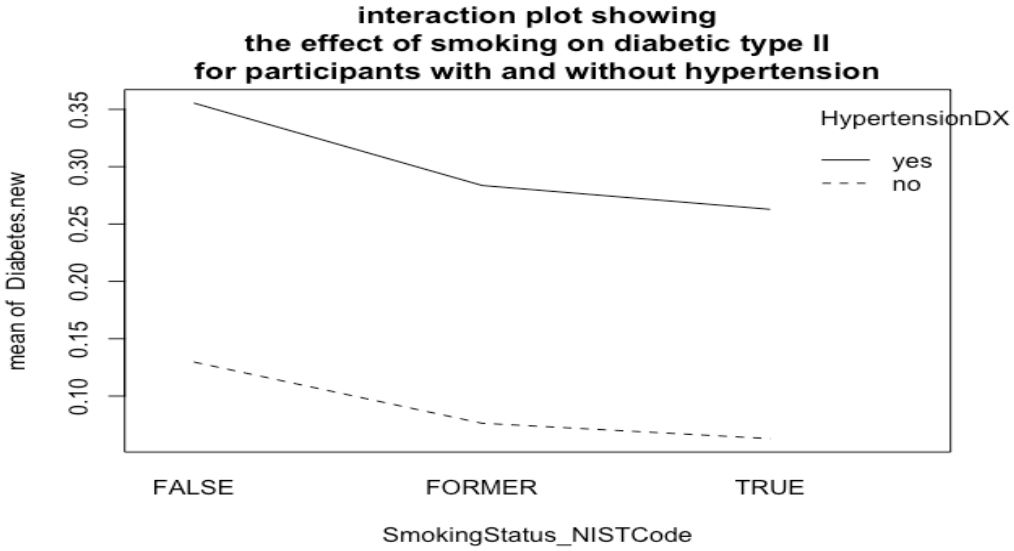
The histogram of age also looks good and no transformation needs to be made. See next page.




```
> m2<-
glm(Diabetes.new~Age+HypertensionDX+SmokingStatus_NISTCode+HypertensionDX*Smoki
ngStatus_NISTCode,
+ family="binomial")
> summary(m1)

Coefficients:
                                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                                       -3.317642   0.185652 -17.870  < 2e-16 ***
Age                                               0.025680   0.002741   9.370  < 2e-16 ***
HypertensionDXyes                                 1.045690   0.120654   8.667  < 2e-16 ***
SmokingStatus_NISTCodeFORMER                      -0.439191   0.154575  -2.841  0.00449 **
SmokingStatus_NISTCodeTRUE                        -0.591187   0.252091  -2.345  0.01902 *
HypertensionDXyes:SmokingStatus_NISTCodeFORMER  0.124006   0.185511   0.668  0.50384
HypertensionDXyes:SmokingStatus_NISTCodeTRUE    0.358845   0.292932   1.225  0.22057
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4404.9  on 4402  degrees of freedom
Residual deviance: 3951.8  on 4396  degrees of freedom
  (5545 observations deleted due to missingness)
AIC: 3965.8
Number of Fisher Scoring iterations: 5
```

Based on the above model, all of the main effects are significant, but, none of the interaction effects are significant.

> interaction.plot(SmokingStatus_NISTCode,HypertensionDX,Diabetes.new,main= "interaction plot showing
+ the effect of smoking on diabetic type II
+ for participants with and without hypertension")



The p-value for the interaction effect is more than 0.05. The plot also has parallel lines indicating that the effect of smoking on diabetic type II is similar for participants with and without hypertension.

We could use the following command to calculate the odd ratios associated with the coefficients of model two (m2) as well as the confidence intervals.

> round(exp(cbind(Estimate=coef(m1), confint(m2))),2)

	Estimate	2.5 %	97.5 %
(Intercept)	0.04	0.03	0.05
Age	**1.03**	**1.02**	**1.03**
HypertensionDXyes	**2.85**	2.25	3.61
SmokingStatus_NISTCodeFORMER	**0.64**	0.47	0.87
SmokingStatus_NISTCodeTRUE	**0.55**	0.33	0.89
HypertensionDXyes:SmokingStatus_NISTCodeFORMER	1.13	0.79	1.63
HypertensionDXyes:SmokingStatus_NISTCodeTRUE	1.43	0.82	2.59

Based on the above model (m2), age, hypertension, smoking, and hypertension are all significant predictors of diabetic type II

Interpretation of odds ratios for model two (m2):

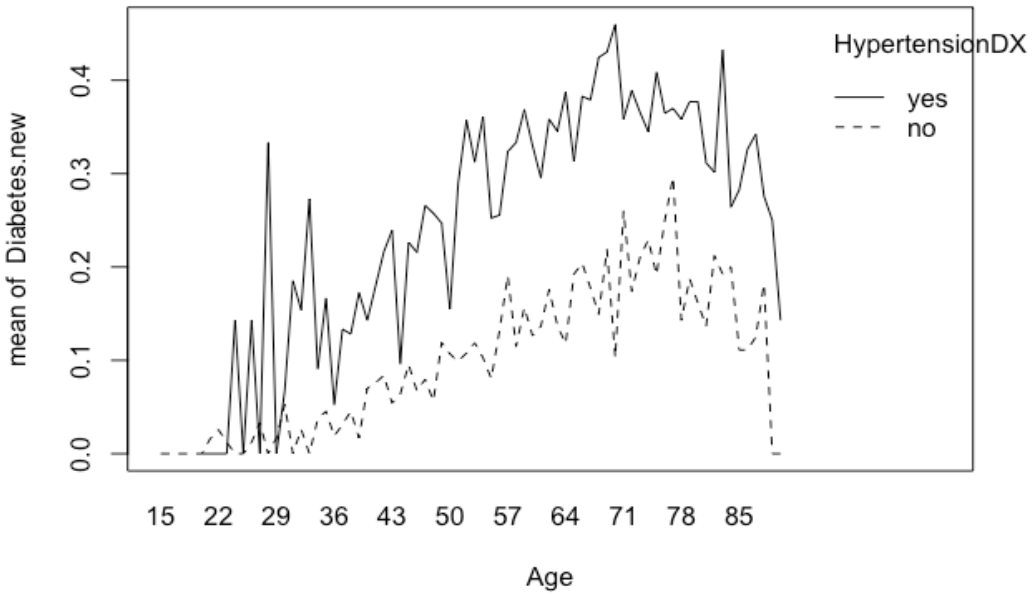
Keeping all else constant…

- The odds of having diabetic type II is 2.85 times more for participants with hypertension.
    - Smoking has three levels, false (non-smoker), smoker, and former smoker. R makes non-smoker the base.
- The odds of having diabetic type II is 36% less for former smokers than non-smokers.
- The odds of having diabetic type II is 45% less for smokers than non-smokers

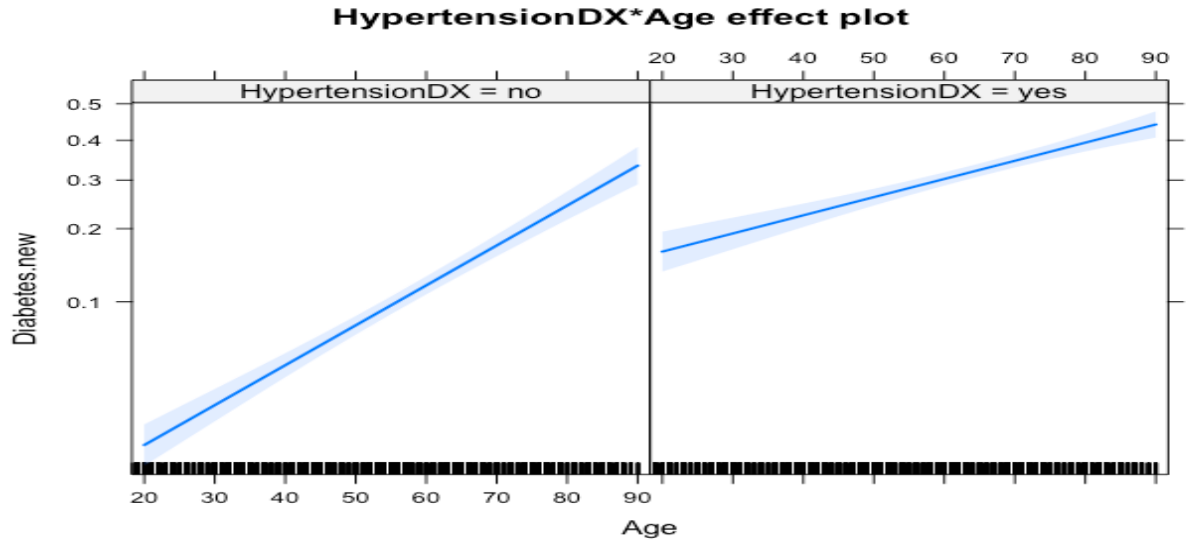
IV Example of logistic regression with interaction between a numerical and a categorical variable

```
> m4<-lm(Diabetes.new~HypertensionDX+Age+Age*HypertensionDX)
> summary(m4)

Coefficients:
                        Estimate    Std. Error   t value    Pr(>|t|)
(Intercept)             -0.0777346  0.0139599    -5.568     2.64e-08 ***
HypertensionDXyes        0.1300913  0.0292128     4.453     8.55e-06 ***
Age                      0.0034984   0.0002807  12.463      < 2e-16 ***
HypertensionDXyes:Age   0.0007287   0.0004923   1.480       0.139
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3638 on 9944 degrees of freedom
Multiple R-squared:  0.1092,  Adjusted R-squared:  0.1089
F-statistic: 406.1 on 3 and 9944 DF,  p-value: < 2.2e-16
```

> interaction.plot(Age,HypertensionDX,Diabetes.new)



> HypertensionDX<-factor(HypertensionDX)
> Diabetes.new<-factor(Diabetes.new)
> Age=as.numeric(Age)
> library(effects)
> plot(allEffects(m4),ask=FALSE)



> round(exp(cbind(Estimate=coef(m4), confint(m4))),4)

	Estimate	2.5 %	97.5 %
(Intercept)	0.9252	0.9002	0.9509
HypertensionDXyes	1.1389	1.0755	1.2061
Age	1.0035	1.0030	1.0041
HypertensionDXyes:Age	1.0007	0.9998	1.0017

Intepretations:

- Keeping all else constant, the odds of having diabetic type II is 14% more for people who have hypertension.

It makes sense to exponentiate age by ten points.

- **exp(0.0034984*10)**

[1] 1.035603

- Keeping all else constant, for each ten years of increase in age, the odds of having diabetic type II increases by 3.5%.

- The effect of age on having diabetic type II does not vary with hypertension.

Sometimes the interaction plots between a numerical and a categorical variable are not easy to interpret and do not look very clear. In such cases, one can always change the numerical variable to a factor. Of course, dividing the numerical variable to a factor should make sense mathematically and from the theoretical point of view.

We will now make age a factor.

> agecategorical<-cut(Age, breaks = c(0,40,60,90),lables=c("less than 40","40 to 60", "60 to 90"),right=FALSE)
> table(agecategorical)

agecategorical

[0,40]	[40,60]	[60,90]
2400	3759	3778

> m5<-
glm(Diabetes.new~agecategorical+HypertensionDX+agecategorical*HypertensionDX,family="binomial")
> summary(m5)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.9580	0.1596	-24.799	< 2e-16	***
agecategorical[40,60]	1.7231	0.1747	9.865	< 2e-16	***
agecategorical[60,90]	2.3902	0.1750	13.661	< 2e-16	***
HypertensionDXyes	2.0709	0.2416	8.571	< 2e-16	***
agecategorical[40,60):HypertensionDXyes	-0.8286	0.2585	-3.205	0.00135	**
agecategorical[60,90):HypertensionDXyes	-1.0727	0.2556	-4.198	2.7e-05	***

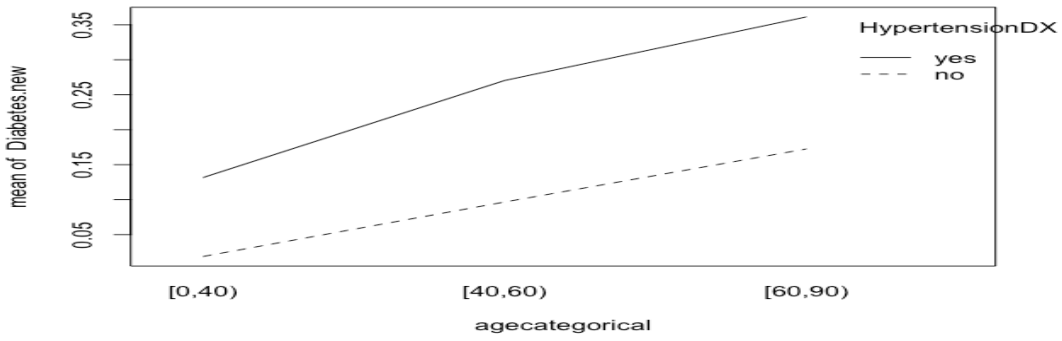
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 9414.8 on 9936 degrees of freedom
Residual deviance: 8195.5 on 9931 degrees of freedom
(11 observations deleted due to missingness)
AIC: 8207.5
Number of Fisher Scoring iterations: 6

> round(exp(cbind(Estimate=coef(m5), confint(m5))),4)

	Estimate	2.5 %	97.5 %
(Intercept)	0.0191	0.0137	0.0257
agecategorical[40,60]	5.6017	4.0250	7.9968
agecategorical[60,90]	10.9159	7.8390	15.5932
HypertensionDXyes	7.9318	4.9218	12.7338
agecategorical[40,60]:HypertensionDXyes	0.4367	0.2632	0.7271
agecategorical[60,90]:HypertensionDXyes	0.3421	0.2074	0.5664

interaction.plot(agecategorical,HypertensionDX,Diabetes.new)



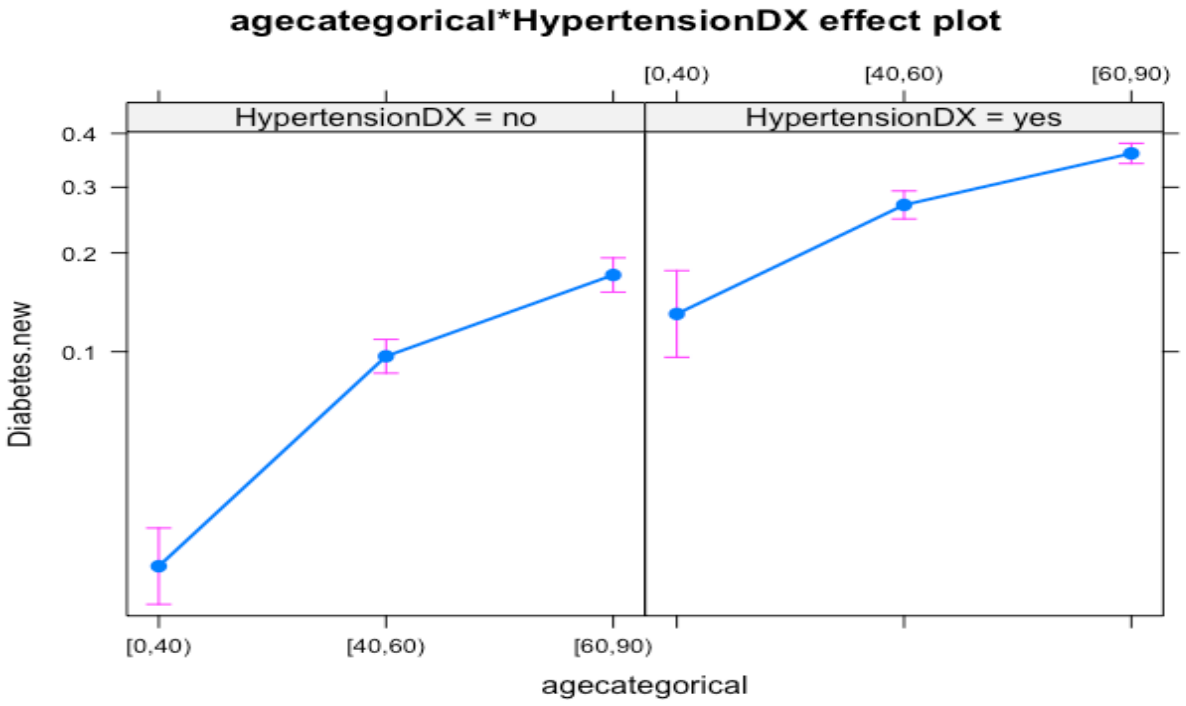
Visual interpretation of the interaction effect

- Overall, participants with high hypertension have a higher odd of having diabetic type II and this is true across all three age groups.

- The effect of age on having diabetic type II is not similar for different age groups. As people grow older the odds of having diabetic type II becomes more for people who have hypertension.

Using the effects library to draw the interaction effect

```
> agecategorical<-factor(agecategorical)
> HypertensionDX<-factor(HypertensionDX)
> Diabetes.new<-factor(Diabetes.new)
> m5<-
glm(Diabetes.new~agecategorical+HypertensionDX+agecategorical*HypertensionDX,family="
binomial")
> library(effects)
> plot(allEffects(m5),ask=FALSE)
```



Should we use age as categorical or numerical?

The answer to this question depends on what the physicians find more informative and makes more sense clinically. The interaction effect which was not significant when age was treated as a numerical variable is now significant. The interpretations related to the odds of diabetic type II as a function of age are also different.

Based on model five (m5), the interpretations are as follows:

Keeping all else constant…

- The odds of diabetic type two is 5.6 times more for the 40-60 age group than less than 40.
- The odds of diabetic type II is 10.9 times higher for 60-90 age group than less than 40.
- The odds of diabetic type II is 7.9 times more for individuals with than without hypertension.

All of the above differences are statistically significant. From the above, you can see the type of information that age provides when it is treated as a numerical or categorical variable.

Interpretation of interaction effects between two categorical variables in logistic regression

In order to interpret the interaction effect associated with the interaction effect between age and hypertension, we will …

- Generate the relevant contingency tables,
- Calculate the conditional odds ratios, and
- Show how the odds ratios of the interaction effect is the ratio of two conditional odds ratios.

> table(Diabetes.new,agecategorical,HypertensionDX)

, , HypertensionDX = no

	agecategorical		
Diabetes.new	[0,40]	[40,60]	[60,90]
0	2094	2056	1127
1	40	220	235

FOR PARTICIPANTS WHO DO NOT HAVE HYPERTENSION…

- The odds of having diabetic type II is 40/2094 for less than 40
- The odds of having diabetic type II is 220/2056 for the 40 to 60 year olds
- **The odds of having diabetic type II is $\frac{220/2056}{40/2094} = \frac{220*2094}{40*2056}$ = 5.60 times higher for 40-60 year olds compare to less than 40.**

, , HypertensionDX = yes

Diabetes.new	agecategorical [0,40]	[40,60]	[60,90]
0	231	1082	1543
1	35	401	873

FOR PARTICIPANTS WHO HAVE HYPERTENSION…

- The odds of having diabetic type II is 35/231 for less than 40
- The odds of having diabetic type II is 401/1082 for the 40 to 60 year olds
- **The odds of having diabetic type II is $\frac{401/1082}{35/231}=\frac{401*231}{35*1082}$ = 2.45 times higher for 40-60 year olds compare to less than 40.**

THE ODDS RATIO FOR THE INTERACTION EFFECT IS THE RATIO OF THE TWO CONDITIONAL ODDS RATIOS.

2.45/5.60 = 0.4375 (THE VALUE REPORTED BY R IS 0.4367 – VERY CLOSE WITHIN ROUNDING ERROR)

agecategorical[40,60]:HypertensionDXyes	**0.4367**	0.2632	0.7271

INTERPRETATION OF THE INTERACTION EFFECT

- **WE HAVE TWO BASELINES HERE AGE (<40) AND (HYPERTENSION = NO)**
- **THE ODDS OF 40 TO 60 YEAR OLDS COMPARED TO FORTY YEAR OLDS TO HAVE DIABETIC TYPE II IS 64% LESS FOR 40-60 YEAR OLDS WHO DO NOT HAVE HYPERTENSION COMPARED TO THOSE WHO DO.**

We seem to have an issue with the sample size in the cell for participants with hypertension and age less than forty. What we need to do in cases like this is to seek the opinion of an expert to figure out the best cut off for age. What makes sense statistically does not always make sense medically. This is potentially the reason for the significance of interaction in the case of treating as a categorical variable and the lack of its significance when age is treated as a numerical variable. We could try cutting at the median and seeing what happens. But, that may not make sense clinically. Facts like this speak for the importance of communication between the client and the statistician.

, , HypertensionDX = no

	agecategorical		
Diabetes.new	[0,40]	[40,60]	[60,90]
0	2094	2056	1127
1	40	220	235

, , HypertensionDX = yes

	agecategorical		
Diabetes.new	[0,40]	[40,60]	[60,90]
0	231	1082	1543
1			

V Analysis of Deviance for Logistic Regression

We will now delete the interaction effect from model two given on page 12 and see how the results will change.

```
> m3<-update(m1, .~.-HypertensionDX:SmokingStatus_NISTCode)
> summary(m3)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.380674	0.174617	-19.361	< 2e-16	***
Age	0.025801	0.002736	9.430	< 2e-16	***
HypertensionDXyes	1.129735	0.090237	12.520	< 2e-16	***
SmokingStatus_NISTCodeFORMER	-0.355067	0.085528	-4.151	3.3e-05	***
SmokingStatus_NISTCodeTRUE	-0.334447	0.130085	-2.571	0.0101	*

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 4404.9 on 4402 degrees of freedom
Residual deviance: 3953.5 on 4398 degrees of freedom
(5545 observations deleted due to missingness)
AIC: 3963.5
Number of Fisher Scoring iterations: 5

As it is evident from model three, after removing the interaction term, null deviance did not change at all showing that the parameters in the two models explain the data equally well.

As is true in the linear models, the ANOVA function can be used to compare two or more "GLM" models that differ by one or more terms. For example, in the above we removed the interaction term from model two (m2) to create model three (m3) and test the null hypothesis that having diabetic type II does not depend on the combined effect of smoking and hypertension vs. the alternative hypothesis that it does.

> anova(m3,m2,test="Chisq")
Analysis of Deviance Table

Model 1: Diabetes.new ~ Age + HypertensionDX + SmokingStatus_NISTCode
Model 2: Diabetes.new ~ Age + HypertensionDX + SmokingStatus_NISTCode +
HypertensionDX * SmokingStatus_NISTCode

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	4398	3953.5			
2	4396	3951.8	**2**	**1.6718**	**0.4335**

The test statistic is the change in deviance between the two fitted models. The p-value is computed by comparing this value with the Chi-square distribution with degrees of freedom (df) equal to the change in the degree of freedom for the two models. The change in deviance is 1.6718 with two degrees of freedom with the two regressors related to the interaction effect removed from the model. When we compare this value with the chi-square distribution with two degrees of freedom, we find the p-value to be 0.4335.

> pchisq(1.6718,2)
[1] 0.5665158

In the chi-square distribution with two degrees of freedom, the area below 1.6718 is 0.5665 and the area above is 0.4335; reflecting the p-value. **Thus, we fail to reject the null and conclude that diabetic type II does not depend on the combined effect of smoking and hypertension. Because this test is based on deviance rather than variance, the output is called "analysis of deviance" table.**

TYPE II TESTS AND THE ANOVA FUNCTION

The ANOVA function in the car package can be used for GLM. Each line in the analysis of deviance table provides a "likelihood" ratio test based on the change in the deviance when comparing the two models. **The type II likelihood ratio statistics tests the same hypothesis tested by the Wald statistic in the summary output for the model.**

> library(car)
> Anova(m2)

Analysis of Deviance Table (Type II tests)

Response: Diabetes.new

	LR Chisq	Df	Pr(>Chisq)
Age	91.222	1	< 2.2e-16 ***
HypertensionDX	169.188	1	< 2.2e-16 ***
SmokingStatus_NISTCode	19.160	2	6.909e-05 ***
HypertensionDX:SmokingStatus_NISTCode	1.672	2	0.4335

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

The results obtained from the analysis of deviance table are similar to the results reported in the GLM model. The interaction effect is not found to be statistically significant. HOWEVER, THE WALD TEST RESULTS ARE NOT IDENTICAL TO THE CORRESPONDING LIKELIHOOD RATIO TESTS.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.317642	0.185652	-17.870	< 2e-16 ***
Age	0.025680	0.002741	9.370	< 2e-16 ***
HypertensionDXyes	1.045690	0.120654	8.667	< 2e-16 ***
SmokingStatus_NISTCodeFORMER	-0.439191	0.154575	-2.841	0.00449 **
SmokingStatus_NISTCodeTRUE	-0.591187	0.252091	-2.345	0.01902 *
HypertensionDXyes:SmokingStatus_NISTCodeFORMER	0.124006	0.185511	0.668	0.50384
HypertensionDXyes:SmokingStatus_NISTCodeTRUE	0.358845	0.292932	1.225	0.22057

VI Goodness of fit test for logistic regression

Overall performance of the fitted model can be measured by several different goodness-of-fit tests. Two tests that are used frequently include **Pearson chi-square goodness-of-fit test** and the **deviance goodness-of-fit test** (analogous to the multiple linear regression lack-of-fit F-test). Both of these tests have statistics that are approximately chi-square distributed with $c - p$ degrees of freedom, where c is the number of distinct combinations of the predictor variables. **When a test is rejected, there is a statistically significant lack of fit. Otherwise, there is no evidence of lack of fit.**

Using Differences in Deviance Values to Compare Models

The difference in the deviance of the "intercept only model" and the "model of interest" can be used to test the goodness of fit of the model.
We will show how this works in the case of our final model.

```
> m3<-update(m1, .~.-HypertensionDX:SmokingStatus_NISTCode)
> summary(m3)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.380674	0.174617	-19.361	< 2e-16	***
Age	0.025801	0.002736	9.430	< 2e-16	***
HypertensionDXyes	1.129735	0.090237	12.520	< 2e-16	***
SmokingStatus_NISTCodeFORMER	-0.355067	0.085528	-4.151	3.3e-05	***
SmokingStatus_NISTCodeTRUE	-0.334447	0.130085	-2.571	0.0101	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Null deviance: 4404.9 on 4402 degrees of freedom
Residual deviance: 3953.5 on 4398 degrees of freedom
(5545 observations deleted due to missingness)
AIC: 3963.5
Number of Fisher Scoring iterations: 5

$\boldsymbol{H_0: \theta(x) = \frac{1}{1+exp(-\beta_0)}}$ **(i.e.** $\boldsymbol{\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0}$**)**

$$\boldsymbol{H_a = \frac{1}{1+exp\{-\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4\}}}$$

(i.e. at least one $\boldsymbol{\beta\ does\ not\ equal\ to\ zero}$**)**

The difference in these two deviances is given by:

$G^2_{H_0}$-$G^2_{H_a}$= 4404.9 - 3953.5 = 451.4

We need to compare this difference with the chi-square distribution with df_{H_0}-df_{H_a}=4402-4398 = 4
Which is equal to the number of regressors in the model

```
> pchisq(451.4,4)
[1] 1
```

The area below 451.4 is one which means the area above it is almost zero. This means that our model has less error than intercept only model and explains some of the variance in the outcome variable.

R^2 for Logistic Regression Model

Recall, that in multiple linear regression model:

$R^2 = 1 - \frac{RSS}{SST}$

RSS = residual sum of squares

SST = Sum of square of total

$\boldsymbol{R^2_{deviance} = 1 - \frac{G^2_{Ha}}{G^2_{H_0}} = 1 - \frac{3953.5}{4404.9} = 1 - 0.8975232 = 0.102476}$

Pearson goodness of fit test for logistic linear regression

Pearson's Chi-square is another alternative for measuring the goodness of fit for logistic regression. The degrees of freedom associated with the Pearson goodness-of-fit statistics is the same as the degrees of freedom for residual deviance, namely

df = n – p -1

N = sample size
p = number of predictors
In the following we will calculate, Pearson goodness of fit test for our final model on the diabetic data.

This was our final model

```
> m3<-
glm(Diabetes.new~Age+HypertensionDX+SmokingStatus_NISTCode,family="binomial")
> summary(m3)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -3.380674   0.174617 -19.361  < 2e-16 ***
Age                           0.025801   0.002736   9.430  < 2e-16 ***
HypertensionDXyes             1.129735   0.090237  12.520  < 2e-16 ***
SmokingStatus_NISTCodeFORMER -0.355067   0.085528  -4.151  3.3e-05 ***
SmokingStatus_NISTCodeTRUE   -0.334447   0.130085  -2.571   0.0101 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4404.9  on 4402  degrees of freedom
Residual deviance: 3953.5  on 4398  degrees of freedom
  (5545 observations deleted due to missingness)
AIC: 3963.5

Number of Fisher Scoring iterations: 5
```

R command and output for calculation of Pearson goodness-of-fit test

```
> #Logistic regression output on page 274


> print(paste("Pearson's X^2 =",round(sum(residuals(m1,type="pearson")^2),3)))
[1] "Pearson's X^2 = 4177.186"
```

The null hypothesis is that the logistic linear model is a good fit for the data. In order to see if we reject or fail to reject the null, we need to compare the calculated Chi-square value of Pearson with the critical value from the Chi-square table.

The critical value of Chi-square at a = 0.05 is:

```
> qchisq(0.95,4398)
[1] 4553.395
```

4177.186< 4553.395

Conclusion: Since the calculated value of Pearson Chi square (**4177.186)** is less than the critical value (**4553.395**), we fail to reject the null hypothesis, and conclude that the logistic model fits the data.

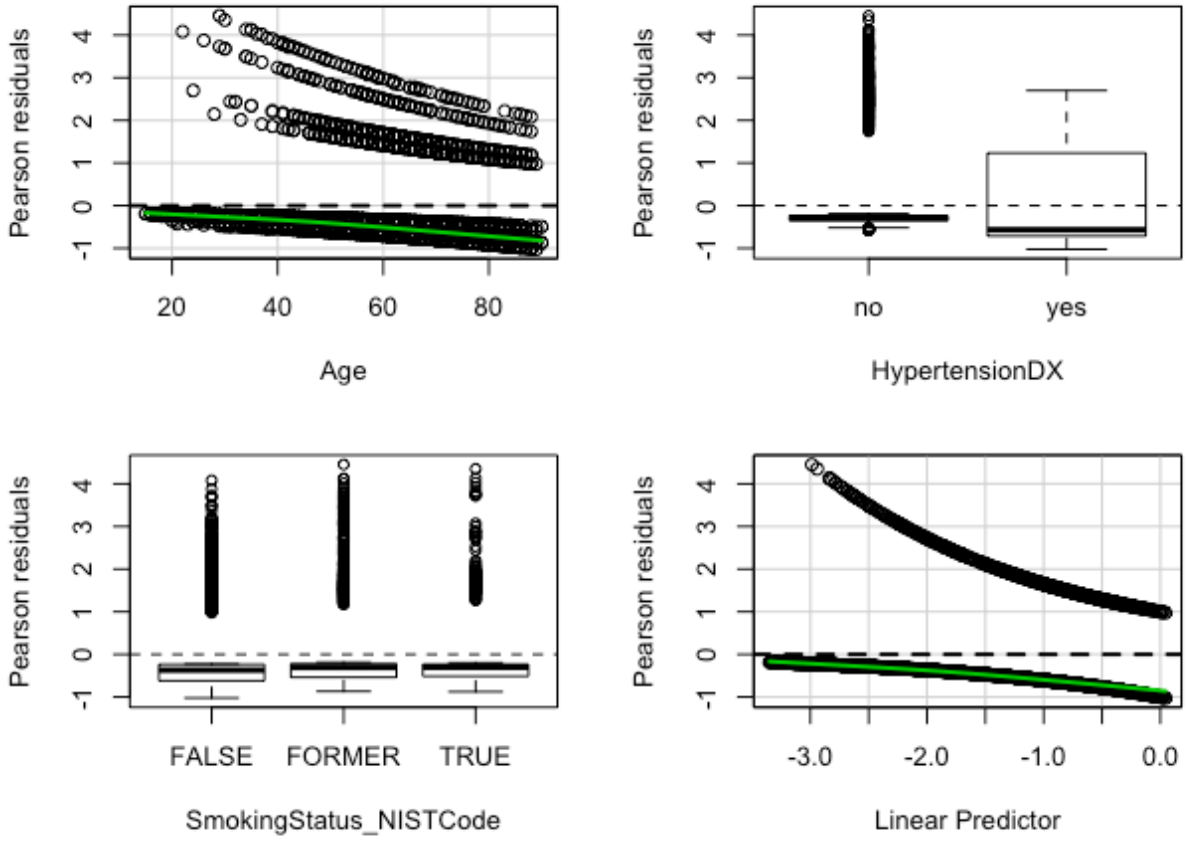
VII Plot of residuals in, influential points, and mmp plots in logistic regression

> residualPlots(m3)	Test stat	Pr(>|t|)
Age	73.435	0
HypertensionDX	NA	NA

The lack of fit test is only provided for the numerical predictor of age and not for the categorical predictors of hypertension and smoking or the linear predictor. The significance of the lack of fit

for > residualPlots(m2)	Test stat	Pr(>|t|)
Age	73.435	0
HypertensionDX	NA	NA

e shows that this plot indicates lack of fit.

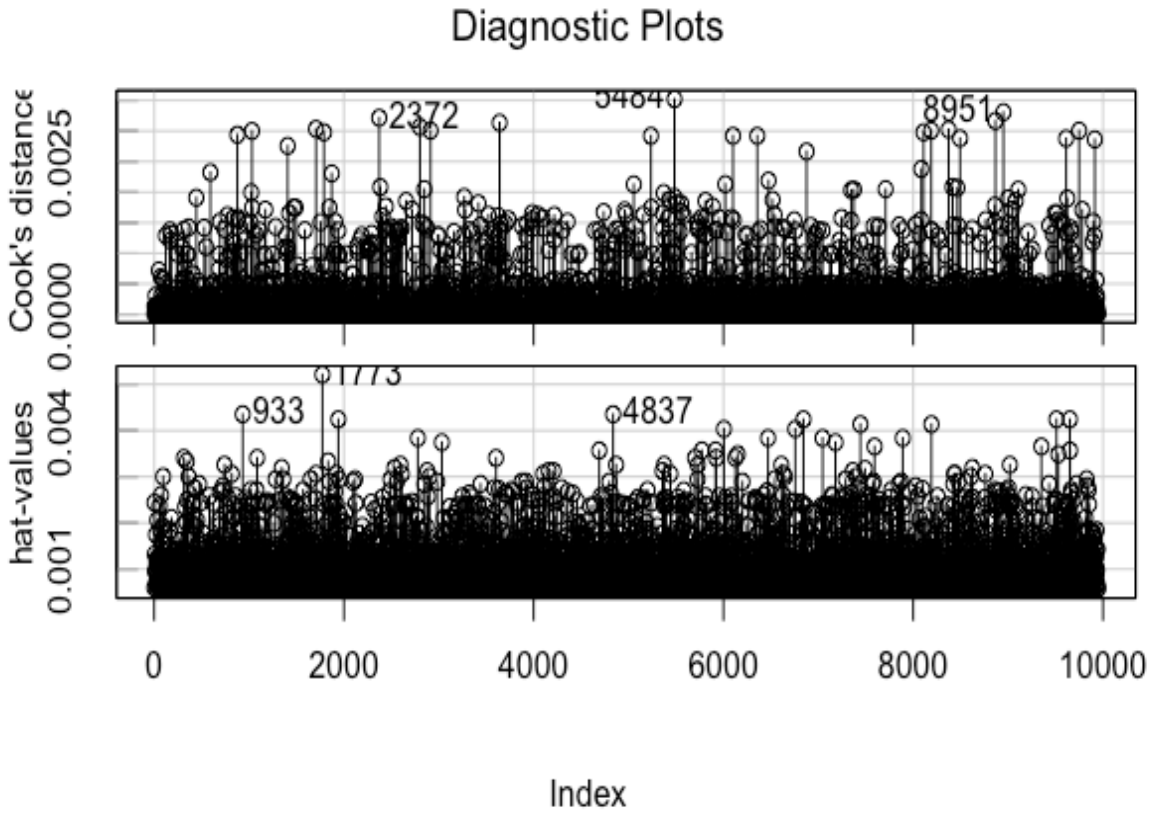


As we see in logistic regression, the plot of Pearson residuals is strongly patterned; specially the plot of Pearson residuals against the linear predictor, where the residuals can only take two values, depending on whether the response is equal to zero or to one. In the plot against age, we see a little more variety in the residuals. In the case of hypertension and smoke, we see two and three boxplots for Pearson residuals because they are factors with two and three levels.

Drawing influence plots in logistic regression

influenceIndexPlot(m3, vars=c("Cook","hat"), id.n=3)

The diagnostic plot given below, shows the cook distance and hat values. The reason that we get only Cook's distance and hat value is that by setting variables equal to cook and hat, we have limited the number of diagnostics to two. Observations 2372, 5404, 8951, 933, 1773, 4837 have been identified as high leverage. We will remove these points to see how the coefficients change. However, since the sample size is large, removing these points may not change the coefficients considerably.



Removing the Observations with High Cook's Distance

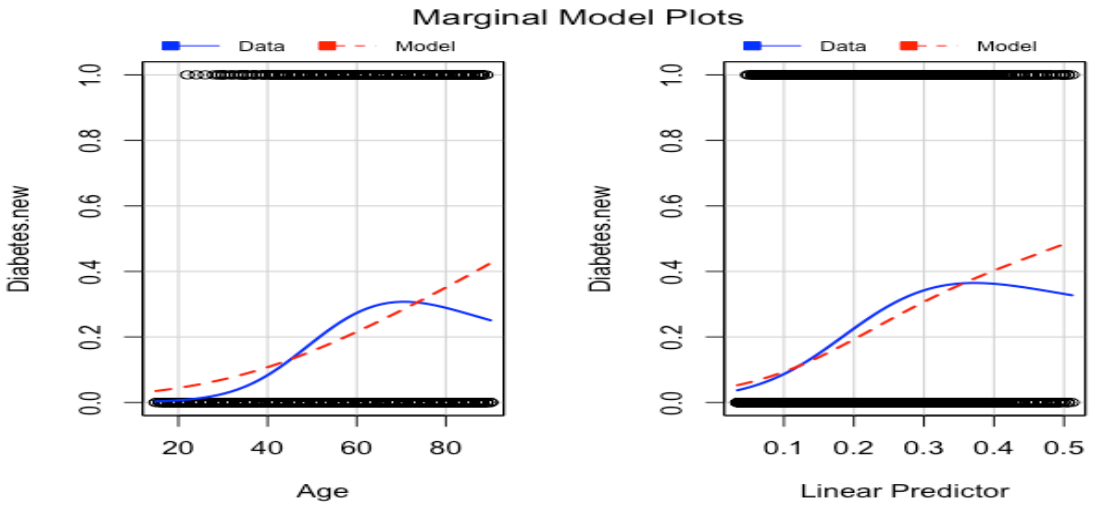
```
> influenceIndexPlot(m3, vars=c("Cook","hat"), id.n=3)
> compareCoefs(m2, update(m3, subset=-c(2372,5404,8951,933,1773,4837)))

Call:
1: glm(formula = Diabetes.new ~ Age + HypertensionDX + SmokingStatus_NISTCode,
family =
  "binomial")
2: glm(formula = Diabetes.new ~ Age + HypertensionDX + SmokingStatus_NISTCode,
family =
  "binomial", subset = -c(2372, 5404, 8951, 933, 1773, 4837))
```

	Est. 1	SE 1	Est. 2	SE 2
(Intercept)	-3.38067	0.17462	-3.41322	0.17572
Age	**0.02580**	0.00274	**0.02623**	0.00275
HypertensionDXyes	**1.12974**	0.09024	**1.13776**	0.09043
SmokingStatus_NISTCodeFORMER	**-0.35507**	0.08553	**-0.35472**	0.08561
SmokingStatus_NISTCodeTRUE	**-0.33445**	0.13009	**-0.34037**	0.13113

It is clear from the above, the removal of the points with high leverage has changed neither the coefficients and nor the standard error.

Marginal Model Plots

Library(car)
Library(alr3)
Mmps(name of the model)



The marginal model plots skip the interaction effect plot. Up to about the age of 70, the plot based on the model (red) and the plot based on the data or moving averages are similar.