PM592: Regression Analysis for Health Data Science
Lab 9 – Logistic Regression Assumptions & Diagnostics
Data Needed: *vote_mhealth.csv*

This lab is devoted entirely to the exercise.

Lab 9 Exercises

Objective(s):	Assess the linearity assumption for logistic regression using 3 techniques, translate the concepts of confounding and effect modification to logistic regression, assess logistic regression model diagnostics and goodness of fit.
Datasets Required:	`vote_mhealth`

Research has shown that participation in voting is higher for those with greater resources, such as time, money, and social status. It was largely unknown whether mental health status had an effect on the likelihood of voting. A working hypothesis is that individuals who experience more depression also experienced more feelings of hopelessness and decreased efficacy. This is compounded by physical correlates of depression, such as lethargy and physical aches, that must also be dealt with.

Use this data set to explore whether mental health is related to the likelihood of voting. Examine age, education, and gender as possible confounders and effect modifiers.

```
vote96
1 if the respondent voted in the 1996 presidential election, 0 otherwise

age
Age of respondent

educ
Number of years of formal education completed by the respondent

female
1 if respondent is female, 0 if male

mhealth
Index variable which assesses the respondent's mental health, ranging from 0
(an individual with no depressed mood) and 9 (an individual with the most
severe depressed mood).
```

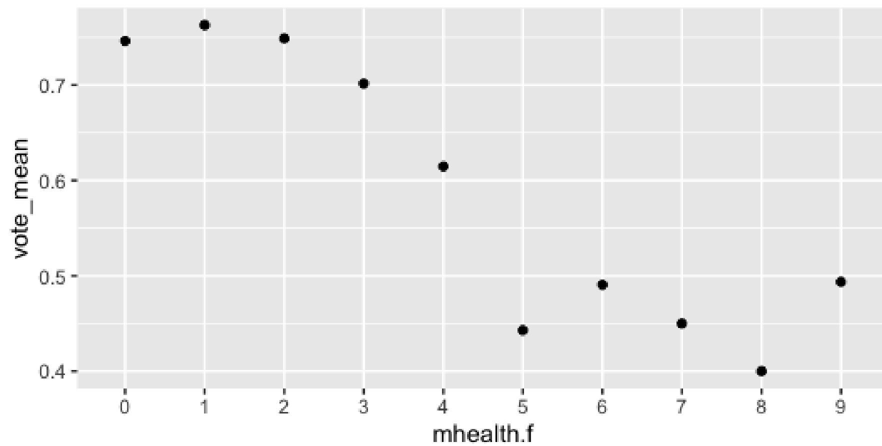
1. Before you begin, determine whether you would like to center any of the variables.

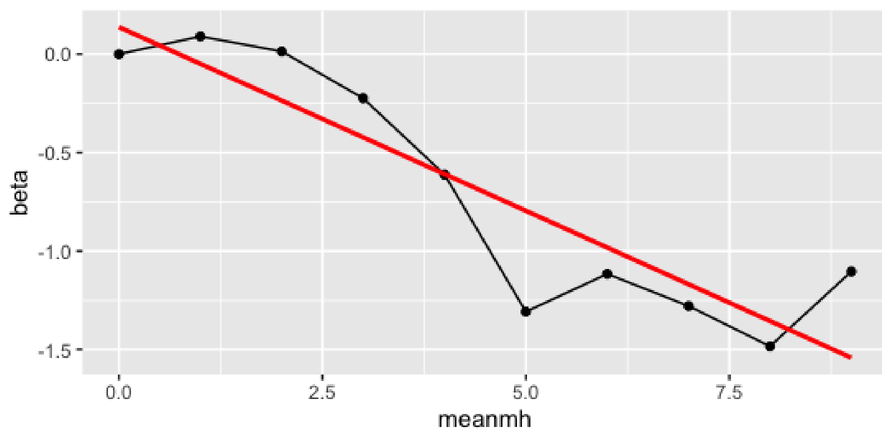
   Centering age and education for better interpretability of coefficient estimates. Especially important to center age, since a 0 value for age wouldn't make sense, and the minimum age to vote is 18. Everyone has likely gotten some years of education.

```
> votemh <-
+   votemh %>%
+   mutate(age.c = age - mean(age), educ.c = educ - mean(educ))
```

2. Examine the form of the relationship between mental health on voting.

   a. Use the grouped smooth method to assess the linearity of mental health and voting. Provide the likelihood ratio test statistic and p-value for the categorical model vs. the ordinal/linear model.

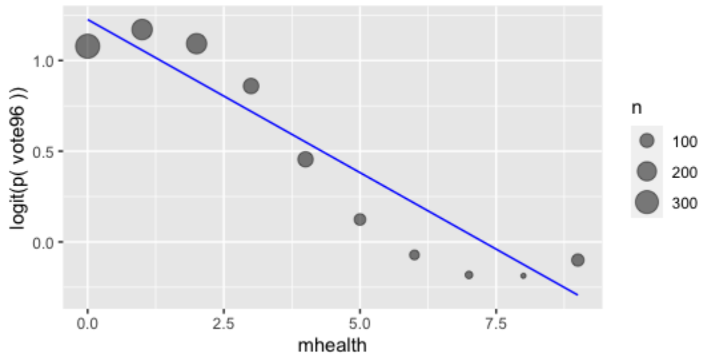





```
Analysis of Deviance Table

Model 1: vote96 ~ mhealth
Model 2: vote96 ~ factor(mhealth)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1315     1598.6
2      1307     1584.5  8   14.106  0.07906 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Compared to the linear model, the dummy variable encoding of mhealth does not improve model fit ($\chi^2_8 = 14.106$, p=0.08)

   b. Use the LOESS method to assess the linearity of mental health and voting. Provide a graph showing the relationship between mental health and the logit of the outcome, based on the LOESS smoother.



The relationship between mental health and the logit of voting does appear to be linear.

c. Use the FP method to assess the linearity of mental health and voting.

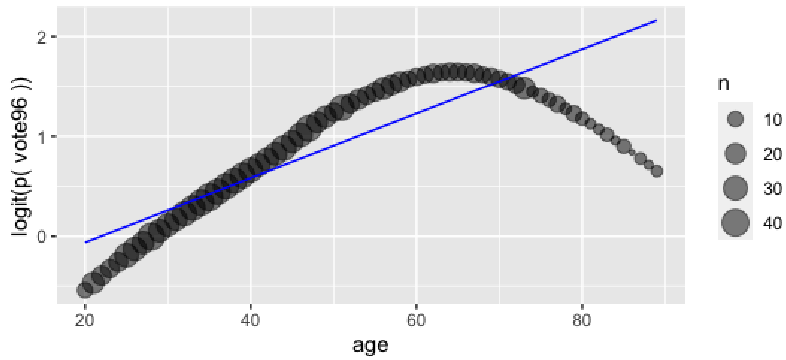
```
> mfp(vote96 ~ fp(mhealth), data = votemh, family = binomial)
Call:
mfp(formula = vote96 ~ fp(mhealth), data = votemh, family = binomial)


Deviance table:
                 Resid. Dev
Null model       1663.824
Linear model     1598.606
Final model      1589.697

Fractional polynomials:
        df.initial select alpha df.final power1 power2
mhealth          4      1  0.05        4      3      3


Transformations of covariates:
                                                                   formula
mhealth I(((mhealth+1)/10)^3)+I(((mhealth+1)/10)^3*log(((mhealth+1)/10)))

Coefficients:
Intercept  mhealth.1  mhealth.2
    1.233     -1.295      7.474

Degrees of Freedom: 1316 Total (i.e. Null);  1314 Residual
Null Deviance:      1664
Residual Deviance: 1590        AIC: 1596
```

```
> 1-pchisq(q=(1598.606-1589.697), df=3)
[1] 0.03052557
```

The fractional polynomials approach suggests that a polynomial term of $x^3 + x^3\ln(x)$ offers the best fit for mental health score and voting.

3. Examine the form of the relationship between all covariates (age, educ, female) and voting using whichever measure you'd like.

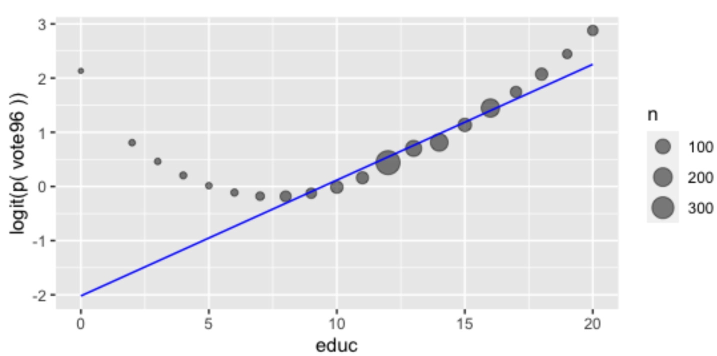
   a. Determine how age is related to the logit.




```
> mfp(vote96 ~ fp(age), data = votemh, family = binomial)
Call:
mfp(formula = vote96 ~ fp(age), data = votemh, family = binomial)


Deviance table:
                 Resid. Dev
Null model       1663.824
Linear model     1580.41
Final model      1555.575

Fractional polynomials:
    df.initial select alpha df.final power1 power2
age          4      1  0.05        4      2      3


Transformations of covariates:
                          formula
age I((age/100)^2)+I((age/100)^3)

Coefficients:
Intercept      age.1      age.2
   -1.047     18.224    -18.317

Degrees of Freedom: 1316 Total (i.e. Null);  1314 Residual
Null Deviance:      1664
Residual Deviance: 1556         AIC: 1562
```

The relationship between age and the logit of the outcome seems to be nonlinear – the fractional polynomials approach suggests a $x^2 + x^3$ polynomial term.

b. Determine how education is related to the logit.

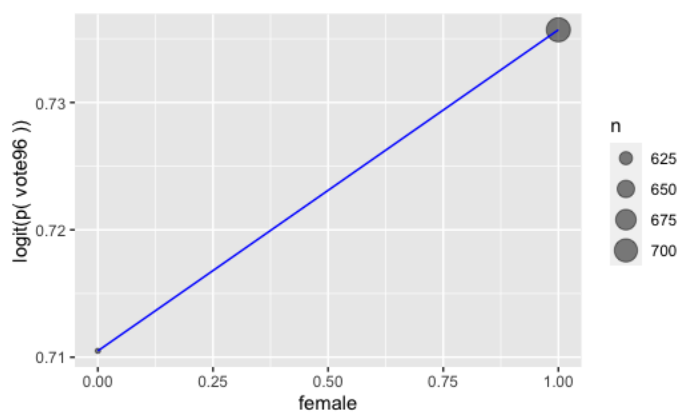


```
> glm(vote96 ~ educ + I(educ^2) + I(educ^3),
+     data = votemh, family = binomial) %>%
+   anova(test = "LRT")
Analysis of Deviance Table

Model: binomial, link: logit

Response: vote96

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       1316     1663.8
educ       1   87.811      1315     1576.0 < 2.2e-16 ***
I(educ^2)  1   15.857      1314     1560.2 6.832e-05 ***
I(educ^3)  1    1.571      1313     1558.6      0.21
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the LOESS method, it appears that education is quadratically related to the logit of the outcome, and the traditional polynomials approach also suggests that a quadratic term is significant in the model.

c. Determine how gender is related to the logit.

Satisfies the linearity assumption because there are only two variables. The logit is higher for females than for males. Additionally, there are many more females than males in the data set.



4. Determine your preliminary final model.

a. Re-assess the linearity of mental health with education and age in the model.

```
Call:
mfp(formula = vote96 ~ fp(mhealth) + female + educ.c + educ.c^2 +
    age_sq + age_cube, data = votemh, family = binomial)


Deviance table:
                 Resid. Dev
Null model       1663.824
Linear model     1407.971
Final model      1407.971

Fractional polynomials:
         df.initial select alpha df.final power1 power2
educ.c            1      1  0.05        1      1      .
age_sq            1      1  0.05        1      1      .
age_cube          1      1  0.05        1      1      .
mhealth           4      1  0.05        1      1      .
female            1      1  0.05        1      1      .


Transformations of covariates:
                      formula
mhealth  I(((mhealth+1)/10)^1)
female                 female
educ.c                 educ.c
age_sq                 age_sq
age_cube             age_cube

Coefficients:
 Intercept    educ.c.1    age_sq.1  age_cube.1   mhealth.1
 -0.551019    0.249391   16.224914  -14.572729   -1.152216
  female.1
 -0.007692

Degrees of Freedom: 1316 Total (i.e. Null);  1311 Residual
Null Deviance:      1664
Residual Deviance: 1408         AIC: 1420
```

After including all covariates in the model, the suggested polynomial term for mental health is linear. ` I(((mhealth+1)/10)^1)`

b. Assess confounding of each covariate on the effect of mental health on voting. You can choose how to do this:

i. One-by-one (good when you're doing exploratory analyses for confounders)

	vote96			vote96			vote96			vote96			vote96		
Predictors	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*	*Log-Odds*	*CI*	*p*
(Intercept)	1.22	1.05 – 1.40	**<0.001**	1.07	0.88 – 1.26	**<0.001**	-0.47	-0.89 – -0.04	**0.031**	1.21	1.00 – 1.43	**<0.001**	-0.72	-1.19 – -0.25	**0.003**
mhealth	-0.18	-0.22 – -0.13	**<0.001**	-0.15	-0.19 – -0.10	**<0.001**	-0.16	-0.20 – -0.11	**<0.001**	-0.18	-0.22 – -0.13	**<0.001**	-0.12	-0.16 – -0.07	**<0.001**
educ c				0.22	0.16 – 0.27	**<0.001**							0.28	0.22 – 0.34	**<0.001**
educ c^2				0.02	0.01 – 0.03	**<0.001**							0.01	0.00 – 0.02	**0.014**
(age/100)^2							16.58	11.83 – 21.34	**<0.001**				16.32	11.38 – 21.27	**<0.001**
(age/100)^3							-16.49	-22.05 – -10.91	**<0.001**				-14.88	-20.65 – -9.07	**<0.001**
female										0.02	-0.21 – 0.26	0.847	-0.01	-0.27 – 0.25	0.932
Observations	1317			1317			1317			1317			1317		
R^2 Tjur	0.051			0.104			0.115			0.051			0.186		

It looks like all covariates are significant (p<0.001) except for gender (p=0.847).

ii. All-at-once (good when you are certain about the set of confounders you want to examine)

c. Write your preliminary final model.

$$\hat{\pi} = -0.73 - 0.12X_{mhealth} + 0.28X_{educ.c} + 0.012X_{educ.}^2 + 16.33\frac{X_{age}^2}{100} - 14.90\frac{X_{age}^3}{100}$$

5. Assess your preliminary final model.

a. What is the pseudo R^2 for this model?

```
> DescTools::PseudoR2(final_model)
 McFadden
0.1573957
```

The model explains 16% of the variance in voting probability.

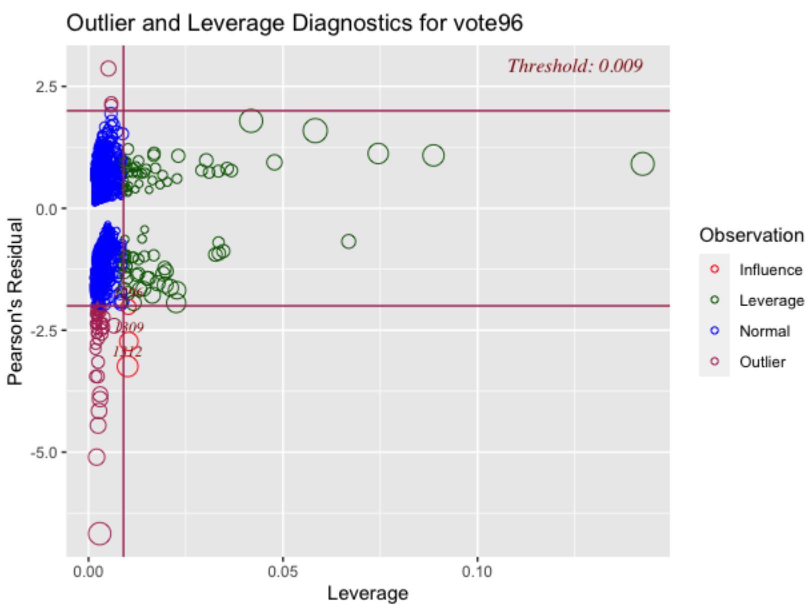
b. How many covariate patterns are there? Based on this, would you trust the Pearson's or Hosmer-Lemeshow test for goodness of fit? Compute the test statistic and p-value for GOF.

There are many covariate patterns, especially because age is continuous. Based on this, I will use the Hosmer-Lemeshow test for goodness of fit.

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  final_model$y, fitted(final_model)
X-squared = 19.283, df = 18, p-value = 0.3746
```

There is no evidence that suggests lack of good fit (p=0.37).

c. List a few covariate patterns that might concern you. Why do they not fit well?



Covariate patterns 1312, 1309, and 1296 have leverage and are outliers, making them influential observations. However, their leverage is not extremely high and the large sample size makes each point less likely to change the parameter estimates by much.

d. Are you confident in your model? Or do you need to re-assess?

I am confident that the set of variables that I have included in the model fits the data well.

6. Present your final model.

a. Present the results of your model in a professionally formatted table. Include the unadjusted and adjusted models.

	vote96			vote96		
Predictors	*Odds Ratios*	*CI*	*p*	*Odds Ratios*	*CI*	*p*
(Intercept)	3.40	2.86 – 4.05	**<0.001**	0.48	0.31 – 0.75	**0.001**
mhealth	0.84	0.80 – 0.88	**<0.001**	0.89	0.85 – 0.93	**<0.001**
educ c				1.32	1.25 – 1.40	**<0.001**
educ c^2				1.01	1.00 – 1.02	**0.014**
(age/100)^2				12345078.30	89859.10 – 1732998153.47	**<0.001**
(age/100)^3				0.00	0.00 – 0.00	**<0.001**
Observations	1317			1317		
R^2 Tjur	0.051			0.186		

b. Write a conclusion that briefly describes your modeling approach and explains the effect of mental health on voting. Include relevant odds ratios, confidence intervals, and p-values.

The following steps were performed to assess the effect of mental health on voting in the ’96 election. First the distributions of each of the covariates of interest were assessed and age and education were centered on their means. Then, the form of the relationship between voting and each of the variables of interest – age, education, and mental health – were assessed. It was determined that mental health alone was not linearly related to the logit of voting. Additionally, the relationship between voting and education and age were nonlinear, so polynomial terms were added for age and education. Then, the form of the relationship between mental health and voting was assessed again, adjusting for the other covariates, and was found to be linearly related to the logit of voting. Next, the set of covariates were each individually assessed for confounding, and gender was eliminated from the model as it did not seem to be a confounder. The equation of the final model is: $\hat{\pi} = -0.73 - 0.12X_{mhealth} + 0.28X_{educ.c} + 0.012X_{educ.}^2 + 16.33\frac{X_{age}^2}{100} - 14.90\frac{X_{age}^3}{100}$. The odds ratio for mental health indicates that a one-unit increase in mental health score is associated with 0.89 times the risk of voting (p<0.001), adjusting for age and education. The confidence interval for the odds ratio is (0.85, 0.93).