PM592: Regression Analysis for Data Science

Name:
Flemming
Wu

HW7

Logistic Regression I

Instructions

- Answer questions directly within this document.
- Upload to Blackboard by the due date & time.
- Clearly indicate your answers to all questions.
- If a question requires analysis, attach all relevant output to this document in the appropriate area. Do not attach superfluous output.
- For the purpose of this assignment, statistical evidence refers to a test statistic and associated p-value.
- If a question requires a table or figure, you must present these in a professionally formatted way (e.g., https://www.researchgate.net/figure/Results-of-Logistic-Regression-Analysis-Unadjusted-and-Adjusted-Odds-Ratios-of_tbl2_47335480)
- If a question requires a conclusion, it must be phrased professionally and coherently.
- There are 3 questions and 30 points possible.

Question 1 [10 points]

Knuckey (2018) discusses the effect of modern and traditional sexism on vote choice in the 2016 presidential election. He used a logistic regression model to determine predictors of voting for Hillary Clinton (vs. Donald Trump).

Y = 1, participant voted for Clinton; 0, participant voted for Trump

X_{MS} = score on an index measuring modern sexism, ranges from 0 to 1; higher scores reflect more sexism

X_{TS} = score on an index measuring traditional sexism, ranges from 0 to 1; higher scores reflect more sexism

$$X_{GEN} = \begin{cases} 1, \text{male} \\ 0, \text{female} \end{cases}$$

The following model is adapted from his paper. All parameter estimates were statistically significant.

$$logit(\hat{\pi}) = 1.8 + 1.5X_{GEN} - 2.1X_{MS} - 2.8X_{TS} - 2.9X_{GEN}X_{MS} + 1.5X_{GEN}X_{TS}$$

1a. [2 points] What is the predicted probability of voting for Clinton for a female that scored a 0 on both scales of sexism?

$$logit(\hat{\pi}) = 1.8 + 1.5(0) - 2.1(0) - 2.8(0) - 2.9(0)(0) + 1.5(0)(0)$$

$$logit(\hat{\pi}) = 1.8$$

$$\hat{\pi} = \frac{e^{1.8}}{1+e^{1.8}} = 0.8581489$$

The predicted probability of voting for Clinton for a female that scored 0 on both scales of sexism is 0.86.

1b. [2 points] What is the predicted probability of voting for Clinton for a male that scored .5 on both scales of sexism?

$$logit(\hat{\pi}) = 1.8 + 1.5(1) - 2.1(.5) - 2.8(.5) - 2.9(1)(.5) + 1.5(1)(.5)$$

$$logit(\hat{\pi}) = 1.8 + 1.5 - 1.05 - 1.4 - 1.45 + 0.75 = 0.15$$

$$\hat{\pi} = \frac{e^{0.5}}{1+e^{0.5}} = 0.5374298$$

The predicted probability of voting for Clinton for a male that scored .5 on both scales of sexism is 0.54.

1c. [2 points] What is the odds ratio for a one-unit increase in modern sexism score, for males?

Holding traditional sexism score constant, the odds ratio for a one-unit increase in modern sexism score for males is given by:

$$e^{(1.8+\ 1.5(1)-\ 2.1(1)-\ 2.8(X_{TS})-\ 2.9(1)(1)+1.5(1)(X_{TS}))-(1.8+\ 1.5(1)-\ 2.1(0)-\ 2.8(X_{TS})-\ 2.9(1)(0)+1.5(1)(X_{TS}))} =$$
$$e^{(-2.1-2.9)} = 0.006737947$$

A one-unit increase in modern sexism score is associated with a 0.007 times the odds of voting for Clinton for males, adjusting for traditional sexism score.

1d. [2 points] What is the odds ratio for a one-unit increase in modern sexism score, for females?

Holding traditional sexism score constant, the odds ratio for a one-unit increase in modern sexism score for females is given by:

$$e^{(1.8+\ 1.5(0)-\ 2.1(1)-\ 2.8(X_{TS})-\ 2.9(0)(1)+1.5(1)(X_{TS}))-(1.8+\ 1.5(0)-\ 2.1(0)-\ 2.8(X_{TS})-\ 2.9(0)(0)+1.5(1)(X_{TS}))} =$$
$$e^{(-2.1)} = 0.1224564$$

A one-unit increase in modern sexism score is associated with a 0.12 times the odds of voting for Clinton for females, adjusting for traditional sexism score.

1e. [2 points] What is the odds ratio comparing a male who scored .75 on both scales of sexism, to a female who scored .25 on both scales of sexism?

The odds ratio comparing a male who scored .75 on both scales of sexism, to a female who scored .25 on both scales of sexism is given by:

$$e^{(1.8+\ 1.5(1)-\ 2.1(0.75)-\ 2.8(0.75)-\ 2.9(1)(0.75)+1.5(1)(0.75))-(1.8+\ 1.5(0)-\ 2.1(0.25)-\ 2.8(0.25)-\ 2.9(0)(0.25)+1.5(0)(0.25))} =$$
$$e^{((1.8+1.5-2.575-2.1-2.175+1.125)-(1.8-0.525-0.7))} = e^{(-2.425-0.575\ )} = 0.04978707$$

A male that scored .75 on both scales of sexism is associated with 0.050 times the odds of voting for Clinton compared to a female that scored .25 on both scales of sexism.

$e^{(-1.425-\ 0.575)} = 0.135$ - a male that scored .75 on both scales of sexism is associated with 0.135 times the odds of voting for Clinton compared to a female that scored .25 on both scales of sexism. Ensure calculations are correct.

Question 2 [12 points]

A study was performed to determine the associations among physical activity cognitive variables and behavior. Dr. Mauvre was interested in correlates of performing physical activity due to intrinsic reasons (i.e., intrinsic enjoyment of the exercise vs. getting some reward). Her primary outcome was exercising for intrinsic enjoyment, and her independent variable was frequency of exercise with child (EWC). This data is located in **intrinsic.dta**.

Y = participant was classified as getting "intrinsic enjoyment" from exercise (1 = intrinsic enjoyment, 0 = no intrinsic enjoyment; derived from a survey)

X_{EWC} = # of days exercised with child in past week (range: 0-4)

- Run a logistic regression with ewc linearly related to enjoyex.

2a. [1 point] Is ewc related to enjoyex? Provide statistical evidence to justify your answer.

```
> m <- glm(enjoyex ~ ewc, data = intrinsic)
> summary(m)

Call:
glm(formula = enjoyex ~ ewc, data = intrinsic)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.33840    0.03482   9.718  < 2e-16 ***
ewc          0.07619    0.02114   3.603 0.000355 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.2379147)

    Null deviance: 95.162  on 388  degrees of freedom
Residual deviance: 92.073  on 387  degrees of freedom
AIC: 549.39

Number of Fisher Scoring iterations: 2
```

Yes, number of days exercised with child in past week (ewc) is related to intrinsic enjoyment (enjoyex) (t=3.6, p=.00036).

Need to specify family = binomial to run a logistic regression.

```
Call:
glm(formula = enjoyex ~ ewc, family = binomial, data = intrinsic)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.66631    0.14941  -4.459 8.22e-06 ***
ewc          0.31403    0.08957   3.506 0.000455 ***
---
```

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

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 530.89  on 388  degrees of freedom
Residual deviance: 518.25  on 387  degrees of freedom
AIC: 522.25

Number of Fisher Scoring iterations: 4
```

The number of days exercised with child in past week is related to intrinsic enjoyment (z=3.506, p<0.001).

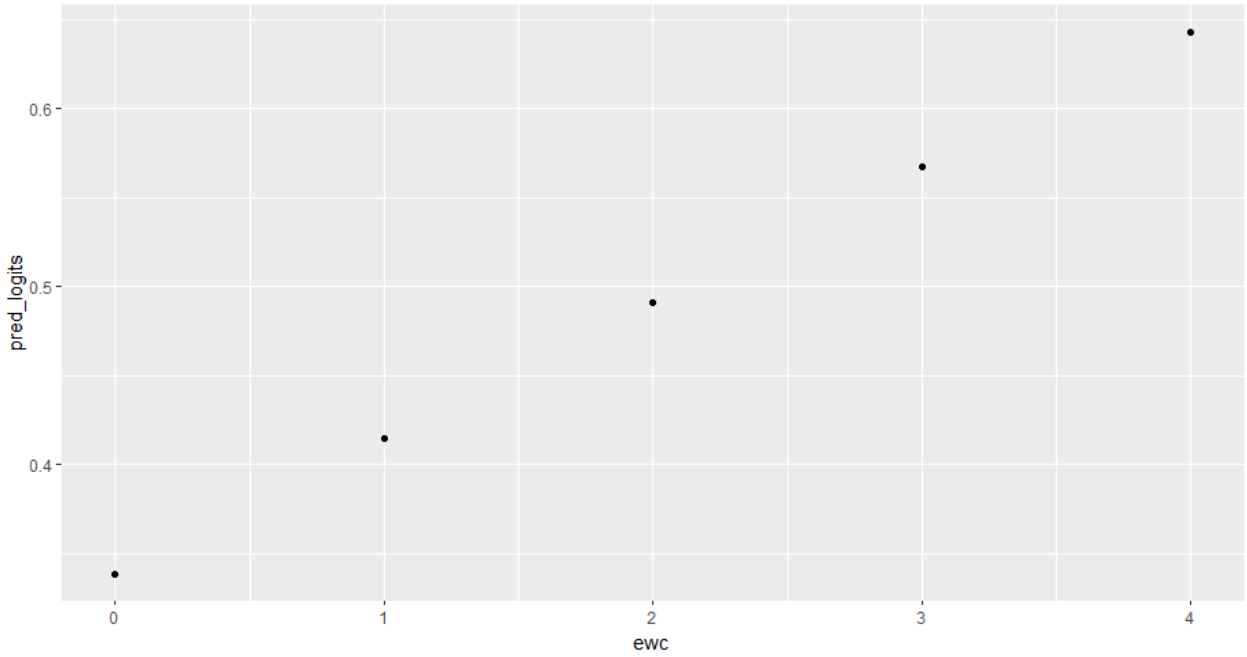
2b. [1 point] Report and interpret the odds ratio associated with ewc.

The odds ratio is given by $e^{\widehat{\beta}} = e^{0.07619} = 1.079168$, and indicates that a one-day increase in ewc is associated with 1.08 times the odds of intrinsic enjoyment from exercise.

The odds ratio is $e^{0.314} = 1.37$. A one-day increase in exercise with child is associated with 1.37 times the odds of enjoying exercise, or a 37% increase in the odds.

2c. [2 points] Graph the relationship between the model-predicted logit of enjoyex and ewc. The relationship was constrained to be linear based on your coding scheme.

```
> intrinsic$pred_logits <- predict(m, type = "link")
> ggplot(intrinsic, aes(x=ewc, y=pred_logits)) + geom_point()
```



- Run a similar logistic regression, instead treating ewc as a set of dummy predictors.

2d. [1 point] Is ewc related to enjoyex? Provide statistical evidence to justify your answer.

```
> m.2 <- glm(enjoyex ~ factor(ewc), data = intrinsic)
> summary(m.2)

Call:
glm(formula = enjoyex ~ factor(ewc), data = intrinsic)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.28378    0.03985   7.121  5.3e-12 ***
factor(ewc)1  0.22080    0.06119   3.608 0.000349 ***
factor(ewc)2  0.24563    0.07102   3.458 0.000604 ***
factor(ewc)3  0.23622    0.07930   2.979 0.003078 **
factor(ewc)4  0.21622    0.13556   1.595 0.111529
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.2350259)

    Null deviance: 95.162  on 388  degrees of freedom
Residual deviance: 90.250  on 384  degrees of freedom
AIC: 547.61

Number of Fisher Scoring iterations: 2
```

```
> anova(m.2, test = "LRT")
Analysis of Deviance Table

Model: gaussian, link: identity

Response: enjoyex

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         388     95.162
factor(ewc)  4    4.912       384     90.250 0.0003315 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Yes, ewc as a dummy variable set is related to intrinsic enjoyment ($\chi^2_4 = 4.912$, p=0.00033), the addition of the variables does significantly contribute to the model.

Again, need to specify family = binomial for a logistic regression.

```
Call:
glm(formula = enjoyex ~ factor(ewc), family = binomial, data = intrinsic)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.9258     0.1823  -5.077 3.82e-07 ***
factor(ewc)1   0.9441     0.2645   3.570 0.000357 ***
factor(ewc)2   1.0436     0.3038   3.435 0.000592 ***
factor(ewc)3   1.0058     0.3367   2.987 0.002815 **
factor(ewc)4   0.9258     0.5648   1.639 0.101168
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 530.89  on 388  degrees of freedom
Residual deviance: 510.33  on 384  degrees of freedom
AIC: 520.33

Number of Fisher Scoring iterations: 4

> anova(m.2, test = "LRT")
Analysis of Deviance Table

Model: binomial, link: logit

Response: enjoyex

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
```

```
NULL                          388     530.89
factor(ewc)  4   20.552       384     510.33 0.0003884 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

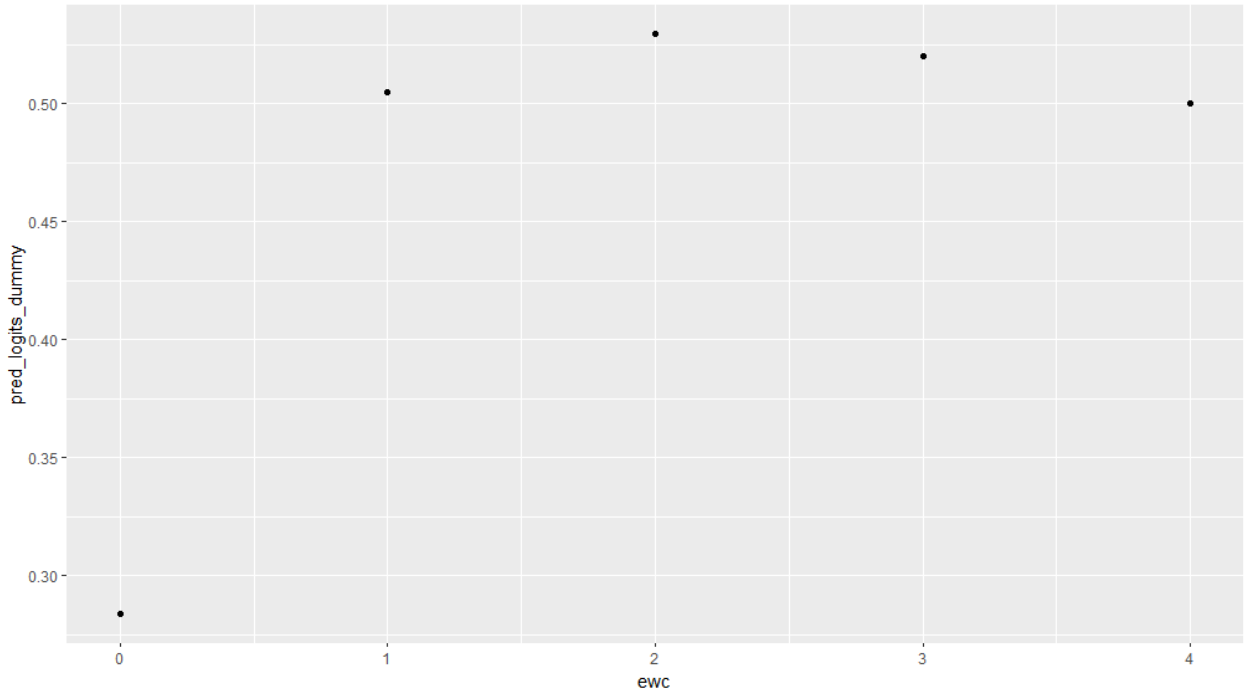
Yes, ewc as a dummy variable set is related to intrinsic enjoyment ($\chi^2_4 = 20.6$, p<0.001), the addition of the variables does significantly contribute to the model.

2e. [2 points] Report and interpret the odds ratios produced by this dummy variable set.

1 day of exercise with child is associated with a ~~$e^{0.2208} = 1.25$~~ $e^{0.9441} = 2.57$ times the odds of intrinsic enjoyment compared to 0 days of exercise with child. 2 days of exercise with child is associated with a ~~$e^{0.24563} = 1.28$~~ $e^{1.0436} = 2.84$ times the odds of intrinsic enjoyment compared to 0 days of exercise with child. 3 days of exercise with child is associated with a ~~$e^{0.23622} = 1.27$~~ times the odds of intrinsic enjoyment compared to 0 days of exercise with child. 4 days of exercise with child is associated with a ~~$e^{0.21622} = 1.24$~~ times the odds of intrinsic enjoyment compared to 0 days of exercise with child.

2f. [2 points] Graph the relationship between the model-predicted logit of `enjoyex` and `ewc`. Does the relationship appear linear?

```
> intrinsic$pred_logits_dummy = predict(m.2, type = "link")
> ggplot(intrinsic, aes(x=ewc, y=pred_logits_dummy)) + geom_point()
```



The relationship actually appears to be quadratic, with the peak at 2 days ewc. Additionally, it seems that 0 days of ewc is the most different from all other ewc values, with 1, 2, 3, and 4 days ewc being similar to each other.

- Evaluate which coding scheme best reflects the data.

2g. [2 points] Use the likelihood ratio test to determine if the dummy variable coding scheme for ewc fits better than the linear coding scheme. Report the associated p-value and make a decision on which coding scheme to use.

```
> anova(m, m.2, test = "LRT")
Analysis of Deviance Table

Model 1: enjoyex ~ ewc
Model 2: enjoyex ~ factor(ewc)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       387     92.073
2       384     90.250  3    1.823  0.05132 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The coding of ewc as a dummy variable does not seem to improve model fit compared to constraining ewc as a linear variable ($\chi^2_3 = 1.823$, p=0.051).

```
> anova(m, m.2, test = "LRT")
Analysis of Deviance Table

Model 1: enjoyex ~ ewc
Model 2: enjoyex ~ factor(ewc)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       387     518.25
2       384     510.33  3   7.9135  0.04783 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Yes, p=0.048.

2h. [1 point] Based on what you found, do you believe that any ewc categories could be collapsed in this analysis?

Based on the findings from modeling intrinsic enjoyment on ewc, I found that the predicted logits for having at least one day of exercise with child are similar, all higher than the predicted logit for no days of exercise with child. Therefore, I believe that ewc could be collapsed into a binary variable indicating whether or not one had any days of exercise with child.

Question 3 [8 points]

Dr. Mauvre additionally collected information on exercise self-efficacy and participants' BMI, which was classified into normal weight, overweight, and obese.

Y = participant was classified as getting "intrinsic enjoyment" from exercise (1 = intrinsic enjoyment, 0 = no intrinsic enjoyment)

X_{SE} = self-efficacy for exercise, measured on a 10-point scale (higher values = more self-efficacy)

$$X_{OW} = \begin{cases} 1, \text{overweight BMI} \\ 0, \text{otherwise} \end{cases}$$

$$X_{OB} = \begin{cases} 1, \text{obsese BMI} \\ 0, \text{otherwise} \end{cases}$$

She fit the following models to the data:

1) $logit(\pi) = -2.35 + 0.63X_{SE}$, LL = -355.10

2) $logit(\pi) = -1.96 + 0.67X_{SE} - 0.24X_{OW} - 1.31X_{OB}$, LL = -336.55

3) $logit(\pi) = -2.07 + 0.66X_{SE} - 1.17X_{OB}$, LL = -337.09

3a. [2 points] What is the interpretation of β_{OB} in Model 2?

β_{OB} in Model 2 gives the expected change in the log odds of intrinsic enjoyment for obese participants compared to ~~non-obese~~ normal weight participants holding self-efficacy constant. It can also be interpreted as indicating that obese participants are associated with a $e^{-1.31}$ = 0.2698201 times the odds of intrinsic enjoyment of non-obese participants, adjusting for self-efficacy.
Model 2 is a categorical (factor) encoding of the weight variable. Since overweight and obese are listed, the one missing, normal weight, is the reference category.

3b. [2 points] What is the interpretation of β_{OB} in Model 3?

β_{OB} in Model 3 gives the expected change in the log odds of intrinsic enjoyment for obese participants compared to non-obese participants holding self-efficacy and overweight status constant. It can also be interpreted as indicating that obese participants are associated with a $e^{-1.17}$ = 0.3103669 times the odds of intrinsic enjoyment of non-obese participants, adjusting for self-efficacy and overweight status.

3c. [2 points] Does the addition of weight status in Model 2 improve model fit compared to Model 1? Justify your answer.

```
> q = (-2*-355.10) - (-2*-336.55)
> pchisq(q=q, df=2, lower.tail = FALSE)
[1] 8.786934e-09
```

The likelihood ratio test can be done as a chi-squared test with -2 times the difference between the two log-likelihoods of the models on 2 degrees of freedom, since model 2 had 2 extra parameters. According to my test, the addition of weight status in Model 2 did improve model fit compared to Model 1 ($\chi_2^2 = 37.1$, p<.001).

3d. [2 points] Does Model 2 have statistically better fit compared to Model 3? Justify your answer.

```
> q = (-2*-337.09) - (-2*-336.55)
> pchisq(q=q, df=1, lower.tail = FALSE)
[1] 0.2986976
```

No, the addition of the overweight category variable in Model 2 did not improve the fit from Model 3 ($\chi_1^2 = 1.08$, p=0.30).