PM592: Regression Analysis for Data Science

Name:
Flemming
Wu

HW8

Logistic Regression II

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 2 questions and 30 points possible.

Question 1 [19 points]

Shanahan et al. (2016) studied the effect of exposure to nature and health outcomes in an urban Australian population. A questionnaire was administered to assess health outcomes and the amount of time each participant typically spends in green spaces (e.g., parks). The data is located in **green.csv**.

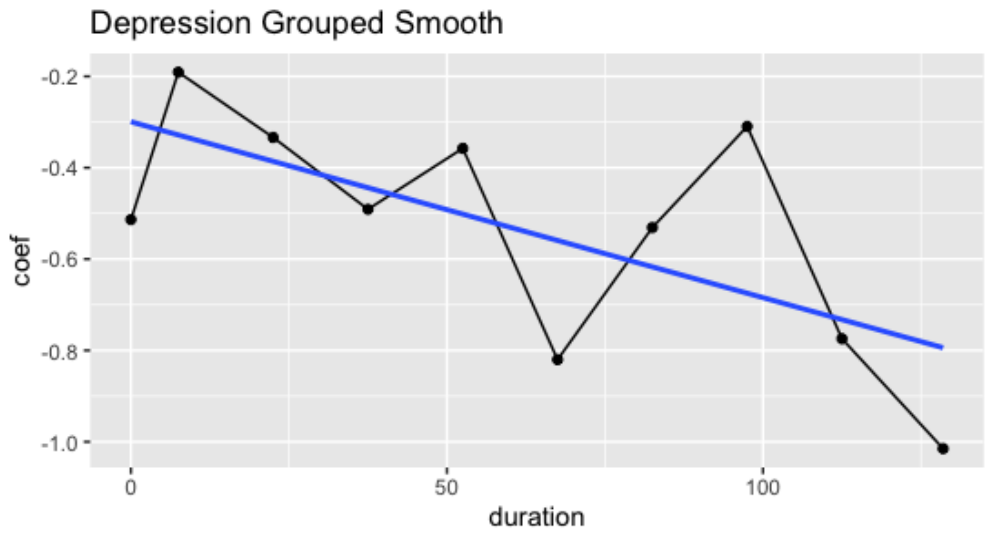
depres = Subject was classified as having depression

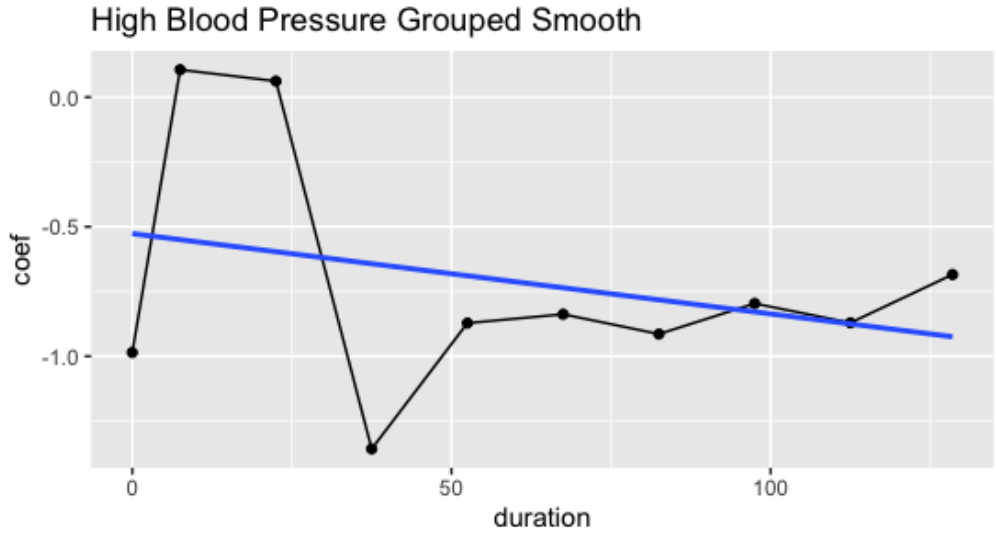
hibp = Subject was classified as having high blood pressure

dur = Duration (minutes) of each visit to green spaces
(Note that response options included 0, 1-15, 16-30, 31-45, …, 106-120, >120 minutes. To approximate these categorical responses as continuous, the midpoint of each interval was taken as the value for "dur." Individuals responding >120 minutes had a recorded "dur" value of 128.5.)

1a. [4 points] Assess the assumption of linearity for the effect of duration of visit to green spaces on 1) depression and 2) high blood pressure using the grouped smooth approach.

```
> green$dur.f <- factor(green$dur)
>
> depres.m <- glm(depres ~ dur.f, family = binomial, data = green)
>
> tibble(
+     duration = as.numeric(levels(green$dur.f)),
+     coef = depres.m$coefficients
+ ) %>%
+   ggplot(aes(x=duration, y=coef)) +
+   geom_point() +
+   geom_line() +
+   geom_smooth(method = "glm", se = F, formula = "y ~ x") +
+   ggtitle("Depression Grouped Smooth")
>
> hibp.m <- glm(hibp ~ dur.f, family = binomial, data = green)
>
> tibble(
+   duration = as.numeric(levels(green$dur.f)),
+   coef = hibp.m$coefficients
+ ) %>%
+   ggplot(aes(x=duration, y=coef)) +
+   geom_point() +
+   geom_line() +
+   geom_smooth(method = "glm", se = F, formula = "y ~ x") +
+   ggtitle("High Blood Pressure Grouped Smooth")
```

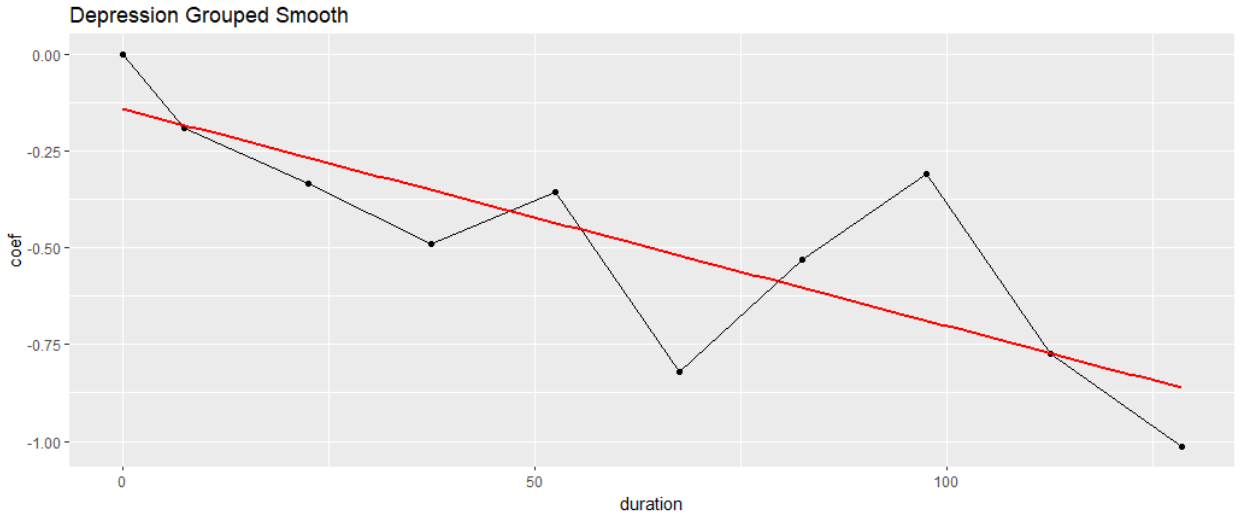




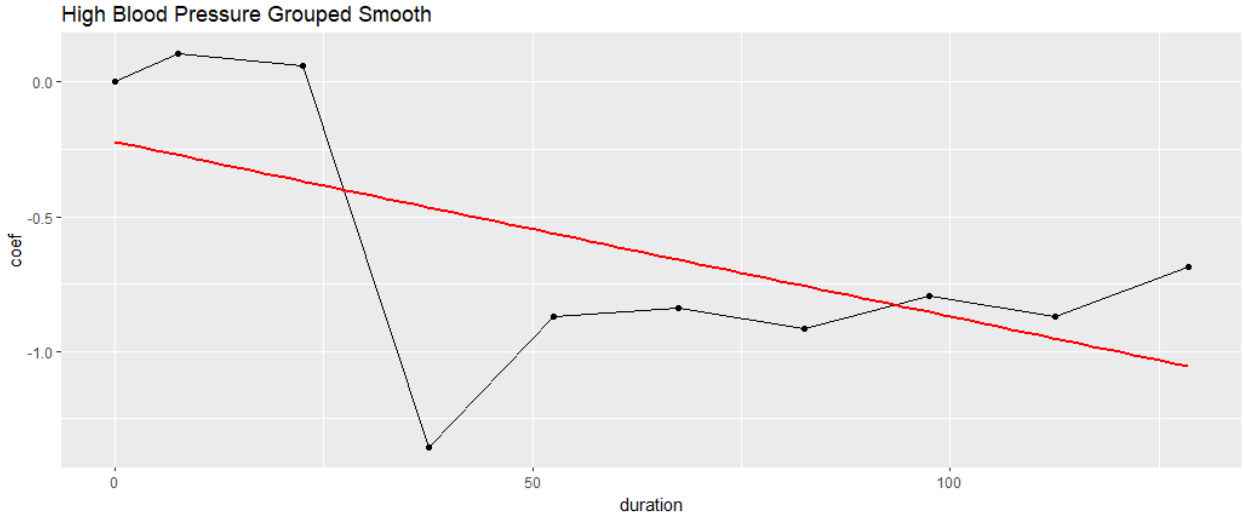
Since there were only 10 duration times in the dataset, I used them as-is instead of converting them into quartiles. I converted duration times to a factor variable, and plotted them against the predicted logits. According to the grouped smooth approach, it appears that depression is linearly related to the probability of the effects of duration to green spaces. The top graph shows an approximately linear relationship. As for the relationship between high blood pressure and the effects of duration to green spaces, it appears that there is a large spike at lower durations, then a drop, and then finally flattens out to a line. I would like to assess other methods to get a better idea of this relationship as it is not very clear in the grouped smooth approach.

When plotting grouped smooth, start the plot at 0 (the pred logit for baseline category)

```r
tibble(
    duration = as.numeric(levels(green$dur.f)),
    coef = c(0, depres.m$coefficients[2:length(depres.m$coefficients)]) # the pred logit should start at 0 for the reference group
) %>%
  ggplot(aes(x=duration, y=coef)) +
  geom_point() +
  geom_line() +
  geom_smooth(method = "glm", se = F, formula = "y ~ x", color="red") +
  ggtitle("Depression Grouped Smooth")
```

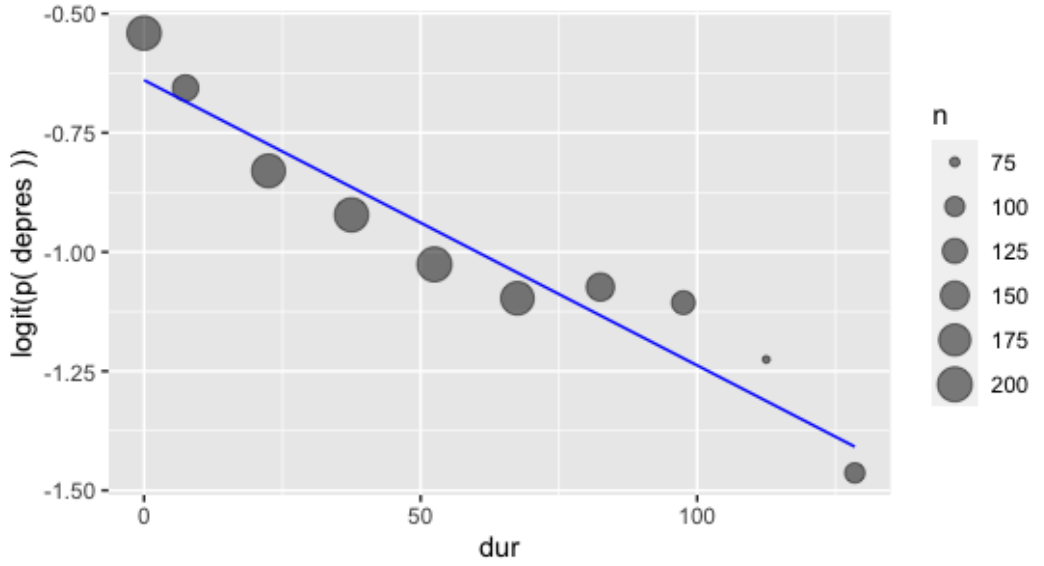


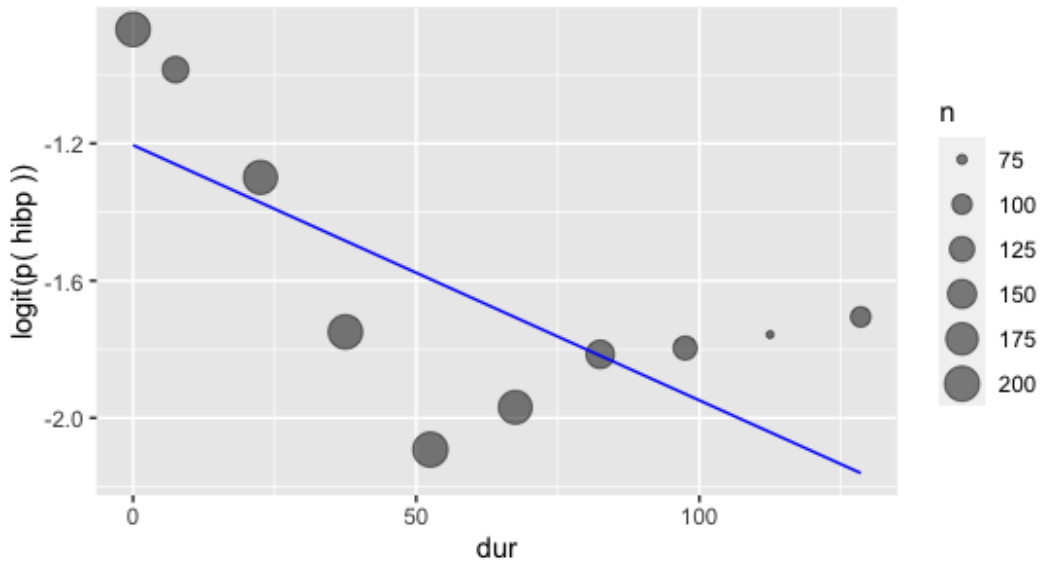

```r
tibble(
  duration = as.numeric(levels(green$dur.f)),
  coef = c(0, hibp.m$coefficients[2:length(hibp.m$coefficients)])
) %>%
  ggplot(aes(x=duration, y=coef)) +
  geom_point() +
  geom_line() +
  geom_smooth(method = "glm", se = F, formula = "y ~ x", color = "red") +
  ggtitle("High Blood Pressure Grouped Smooth")
```



1b. [4 points] Assess the assumption of linearity for the effect of duration of visit to green spaces on 1) depression and 2) high blood pressure using the LOESS smoothing approach.

```
> # 1b
> logit_plot("dur", "depres", green)
`geom_smooth()` using formula = 'y ~ x'
> logit_plot("dur", "hibp", green)
`geom_smooth()` using formula = 'y ~ x'
```





According to the LOESS smoothing approach, the relationship between duration of visit to green spaces and depression is again linear. However, for the relationship between duration and high blood pressure, the relationship appears to be nonlinear. The predicted logit declines sharply until a duration of about 50 minutes, and then slowly increases afterwards.

1c. [4 points] Assess the assumption of linearity for the effect of duration of visit to green spaces on 1) depression and 2) high blood pressure using the fractional polynomials approach.

```
> mfp(depres ~ fp(dur), data = green, family = binomial)
Call:
mfp(formula = depres ~ fp(dur), data = green, family = binomial)


Deviance table:
                Resid. Dev
Null model      1826.487
Linear model    1811.981
Final model     1811.981

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


Transformations of covariates:
                formula
dur I(((dur+7.5)/100)^1)

Coefficients:
Intercept      dur.1
  -0.6092    -0.5706

Degrees of Freedom: 1537 Total (i.e. Null);  1536 Residual
Null Deviance:      1826
Residual Deviance: 1812         AIC: 1816
```

```
> mfp(hibp ~ fp(dur), data = green, family = binomial)
Call:
mfp(formula = hibp ~ fp(dur), data = green, family = binomial)


Deviance table:
                Resid. Dev
Null model      1453.518
Linear model    1428.991
Final model     1422.184

Fractional polynomials:
    df.initial select alpha df.final power1 power2
dur          4      1  0.05        2      0      .


Transformations of covariates:
                formula
dur log(((dur+7.5)/100))

Coefficients:
Intercept      dur.1
   -1.878     -0.398

Degrees of Freedom: 1537 Total (i.e. Null);  1536 Residual
Null Deviance:      1454
Residual Deviance: 1422         AIC: 1426
```

According to the fractional polynomials approach, the best way to model the relationship between duration and depression is linear. Interestingly, it says that the best way to model the relationship between duration and high blood pressure is by taking the logarithm of duration.

1d. [3 points] Write a short (3-5 sentences) methods section describing how you evaluated the assumption of linearity in these models and providing your rationale for modeling the variables in the way you did.

To assess the assumption of linearity for the relationship between duration of visit to green spaces and the response variables depression and high blood pressure, three methods were used: the grouped smooth approach, the LOESS smooth approach, and the fractional polynomials approach. First, for the grouped smooth method, the predicted logits were obtained for each of the ten duration levels and plotted, which revealed a linear relationship for depression and duration and an unclear, but suspiciously nonlinear relationship for high blood pressure and duration. Next, the LOESS smoothing approach further confirmed a linear relationship for depression and duration and a nonlinear relationship for high blood pressure and duration. Finally, the fractional polynomials approach also confirmed that a linear relationship between depression and duration resulted in the lowest residual

deviance, and also suggested a logarithmic transformation for duration for its relationship with high blood pressure.

```
> anova(
+ glm(hibp ~ dur, data = green, family = binomial),
+ glm(hibp ~ factor(dur), data = green, family = binomial)
+ , test="LRT")
Analysis of Deviance Table

Model 1: hibp ~ dur
Model 2: hibp ~ factor(dur)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1536     1429.0
2      1528     1397.4  8   31.625 0.0001087 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The likelihood ratio test shows that for high blood pressure, using a categorical encoding of the duration variable gives significant improvement in fit. I think that this encoding should be used and may be better than using a log transformation, as it may help with interpretation as well. If duration was log transformed, it would have to be exponentiated twice to interpret the log odds ratio change.

1e. [4 points] Write a short (3-5 sentences) results/conclusion section describing your modeling approach and the parameter estimates (with 95% CI and p-values) of interest. In your conclusion, make specific recommendations about how much time residents should spend in green spaces in order to improve these two health outcomes.

```
Call:
glm(formula = hibp ~ factor(dur), family = binomial, data = green)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.98554    0.16097  -6.123  9.2e-10 ***
factor(dur)7.5    0.10580    0.24938   0.424 0.671374
factor(dur)22.5   0.06186    0.22756   0.272 0.785726
factor(dur)37.5  -1.35740    0.30058  -4.516  6.3e-06 ***
factor(dur)52.5  -0.87192    0.26216  -3.326 0.000881 ***
factor(dur)67.5  -0.83777    0.26567  -3.153 0.001614 **
factor(dur)82.5  -0.91421    0.29397  -3.110 0.001871 **
factor(dur)97.5  -0.79637    0.30763  -2.589 0.009632 **
factor(dur)112.5 -0.87076    0.37621  -2.315 0.020637 *
factor(dur)128.5 -0.68453    0.31650  -2.163 0.030558 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1453.5  on 1537  degrees of freedom
Residual deviance: 1397.4  on 1528  degrees of freedom
AIC: 1417.4

Number of Fisher Scoring iterations: 4
```

Perhaps the optimal time spent in green spaces to reduce high blood pressure is 35.7 (z=-4.5, p<0.001).

Once the assumption of linearity for the relationships between duration of visit in green spaces and high blood pressure and depression were assessed, and the best modeling of the independent variables were determined, the final equations were determined to be: $\hat{Y}_{depression} = -0.65 - 0.0057X_{duration}$ and $\hat{Y}_{hibp} = -1.88 - 0.40\log(X_{duration})$. The beta estimate for $X_{duration}$ in the depression equation, -0.0057, means that a 1 minute increase in time spent in green spaces will decrease the odds of having depression by $e^{-0.0057} = 0.994$ times (CI = (0.993, 0.996), p<0.001). The beta estimate for $X_{duration}$ in the high blood pressure equation, -0.40, means that 1 minute increase in time spent in green spaces is associated with $e^{-0.40} = 0.67$ multiplicative increase in the probability of having high blood pressure (CI = (0.63,0.72), p<0.001), or in other words a 1 minute increase in time spent in green spaces decreases the probability of having high blood pressure by $e^{-0.67} = 1.95$ times. Since duration decreases the odds of both depression and high blood pressure according to both models, residents should aim to spend at least 120 minutes in green spaces to improve these two health outcomes.

Question 2 [11 points]

Read the article by Kalligeros et al. (2020) on Blackboard. The authors performed a logistic regression to determine risk factors for two outcomes: 1) being admitted to the ICU and 2) being intubated within 10 days of hospital admission for COVID-19.

2a. [3 points] Do the authors assess the linearity assumption for BMI? Why do you think they ended up treating BMI as categories instead of as a linear variable?

It was not explicitly stated whether the authors had assessed the linearity assumption for BMI. However, from looking at the odds ratios in the univariate logistic regression using BMI categories against IMV and ICU, it appears that compared to the reference category (BMI < 25), the increase in odds ratios for a one-category increase in BMI are 1.87, 2.80, and 3.02 for ICU, and 2.52, 4.86, and 5.84 for IMV. Converting these into log-odds, they are 0.63, 1.03, and 1.11 for ICU and 0.92, 1.58, 1.76 for IMV. The increase in the logit for an increase in category doesn't appear to be linear, which would be good reason to treat BMI as categories instead of a linear variable.

2b. [3 points] The authors believed that severely obese patients may be at greater risk for ICU admission due to having a higher risk of heart disease. What did the authors do to ensure that the effect of severe obesity on ICU admission was due to COVID and not other comorbidities?

After performing a univariate logistic regression on BMI, the authors adjusted for demographic variables age, race, and gender. Then, the authors further adjusted for chronic heart disease and chronic lung disease and found that severe obesity was still statistically significant after both adjustments.

2c. [3 points] The confidence intervals for some parameter estimates in Tables 2 & 3 are quite wide (e.g., 1.39 – 71.69). Does this concern you? What was the authors' explanation for the size of the confidence intervals?

Yes, the wide confidence intervals for some parameter estimates is a bit of a concern. The authors' explanation for the wide confidence interval is that it is likely due to small sample size.

2d. [2 points] Suppose you are in charge of conducting a new study to replicate these findings. How would you change the study design to prevent these large confidence intervals?

Some improvements that I would make in a re-evaluation of the study is to first get more participants for more statistical power. I noticed that from table 1, only 1.9% of the participants were Non-Hispanic Asian and the same percentage of participants had a transplant, and only 2.9% of participants had cirrhosis. In gathering a larger sample size, I would try to ensure that there are enough participants in each of the categories for better detection of confounding and interactions.