Name - NetID

Homework 13

Homework Instructions

For questions that require code, please create a code chunk directly below the question and type your code there. Your knitted pdf will show both your code and your output. You are encouraged to knit your file as you work to check that your coding and formatting is done so appropriately.

For written responses or multiple choice questions, please bold your (selected) answer.

Grading Details

All questions will be graded full credit (1 point), half credit (0.5 point) or no credit (0 points).

Full credit responses should have the correct response and appropriate code (if applicable). Half credit responses will have a reasonable attempt (typically no more than one small error or oversight), and no credit responses will be either non-attempts or attempts with significant errors.

Exercise 1

Which of the following questions/response variables could be modeled using logistic regression? **Bold all** answers that apply (there may be just one answer, or more than one).

- Modeling time to reach 60mph based on characteristics of a car
- **Estimating the likelihood of developing breast cancer in the next 10 years based on medical information for individuals.**
- Predicting Number of accidents each day in the Chicago city limits based on weather conditions and traffic patterns
- **Predicting presence of heartworms in dogs based on collected samples from a diagnostic test.**
- Predicting which of the four Harry Potter houses someone belongs to based on their answers on a personality test.

Exercise 2

Consider a categorical response Y which takes possible values 0 and 1 as well as three numerical predictors X_1, X_2, and X_3. Recall that

$$p(\mathbf{x}) = P[Y = 1 \mid \mathbf{X} = \mathbf{x}]$$

Consider the model

$$\log\left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

together with parameters

- $\beta_0 = -3$
- $\beta_1 = 1$

- $\beta_2 = 2$
- $\beta_3 = 3$

Calculate $P[Y = 1 \mid X_1 = -1, X_2 = 0.5, X_3 = 0.25]$.

Hint: As stated, the response is the log-odds. You need to un-transform to obtain the desired conditional probability. You can check the book for quick reference of how to complete that inverse operation on the right side of your equation to get an expression for the probability.

```
# solution
eta = -3 + (1 * -1) + (2 * 0.5) + (3 * 0.25)
p = 1 / (1 + exp(-eta))
p
```

```
## [1] 0.09534946
```

Exercise 3

For Exercises 3 - 12, use the built-in `R` dataset `mtcars`. We will use this dataset to attempt to predict whether or not a car has a manual transmission. You may optionally view or learn about the data here.

```
#starter
```

Recall that

$$p(\mathbf{x}) = P[Y = 1 \mid \mathbf{X} = \mathbf{x}]$$

Fit the model

$$\log\left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where

- Y is `am`
- x_1 is `mpg`
- x_2 is `hp`
- x_3 is `qsec`

Report the coefficient for the `qsec` predictor.

```
# solution
fit_full = glm(am ~ mpg + hp + qsec, data = mtcars, family = "binomial")
coef(fit_full)["qsec"]
```

```
##      qsec
## -4.040952
```

Exercise 4

Using the model fit in Exercise 3, estimate the change in log-odds that a car has a manual transmission for an increase in fuel efficiency of one mile per gallon. Report just the value.

Hint: You can report just the value by using double brackets, or using the ***unname*** *function around the extracted coefficient.*

```
# solution
fit_full = glm(am ~ mpg + hp + qsec, data = mtcars, family = "binomial")
coef(fit_full)[["mpg"]]
```

```
## [1] 2.29643
```

Exercise 5

Using the model fit in Exercise 3, estimate the log-odds that a car has a manual transmission with a fuel efficiency of 22 miles per gallon, 123 horsepower, and a quarter mile time of 18 seconds

Hint: remember to set the ***type*** *argument appropriately to get the log odds, rather than the estimated probability.*

```
# solution
fit_full = glm(am ~ mpg + hp + qsec, data = mtcars, family = "binomial")
a_car = data.frame(mpg = 22, hp = 123, qsec = 18)
predict(fit_full, a_car, type = "link")
```

```
##        1
## 2.394592
```

Exercise 6

Using the model fit in Exercise 3, estimate the probability that a car has a manual transmision for a car with a fuel efficiency of 22 miles per gallon, 123 horsepower, and a quarter mile time of 18 seconds

```
# solution
fit_full = glm(am ~ mpg + hp + qsec, data = mtcars, family = "binomial")
a_car = data.frame(mpg = 22, hp = 123, qsec = 18)
predict(fit_full, a_car, type = "response")
```

```
##        1
## 0.916414
```

Exercise 7

Let's now consider more carefully whether it makes most sense to include all of these predictors, or if a simpler model is better.

Using the model from Exercise 3, report the coefficients of this model with 3 predictors.

```r
# solution
fit_full = glm(am ~ mpg + hp + qsec, data = mtcars, family = "binomial")
coef(fit_full)
```

```
##  (Intercept)          mpg           hp         qsec
## 23.712985207  2.296430145  0.007294986 -4.040952199
```

```r
#or alternatively (more helpful for the next Question)
coef(summary(fit_full))
```

```
##                 Estimate  Std. Error    z value  Pr(>|z|)
## (Intercept) 23.712985207 44.06951860  0.5380813 0.5905209
## mpg          2.296430145  1.62323229  1.4147268 0.1571486
## hp           0.007294986  0.04719936  0.1545569 0.8771707
## qsec        -4.040952199  2.89045964 -1.3980310 0.1621038
```

```r
#or just a full model summary is also fine
summary(fit_full)
```

```
##
## Call:
## glm(formula = am ~ mpg + hp + qsec, family = "binomial", data = mtcars)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.25597  -0.04564  -0.00055   0.00211   1.84909
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 23.712985  44.069519   0.538    0.591
## mpg          2.296430   1.623232   1.415    0.157
## hp           0.007295   0.047199   0.155    0.877
## qsec        -4.040952   2.890460  -1.398    0.162
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43.2297  on 31  degrees of freedom
## Residual deviance:  7.4802  on 28  degrees of freedom
## AIC: 15.48
##
## Number of Fisher Scoring iterations: 9
```

Exercise 8

Consider the Wald test to test $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$. What does the result from this Wald test communicate to us? Please **bold** your answer.

- We failed to find evidence that `horsepower` is a significant predictor of automatic vs. manual transmission
- **We failed to find evidence that `horsepower` is a significant predictor of automatic vs. manual transmission when already using `mpg` and `qsec` as predictors.**

- We found evidence that `horsepower` is a significant predictor of automatic vs. manual transmission
- We found evidence that `horsepower` is a significant predictor of automatic vs. manual transmission, even when already using `mpg` and `qsec` as predictors.

Exercise 9

Now compare this result to the result we'd get using the likeliood ratio test (the equivalent to the F-test for looking at a logistic model). Again, test the null hypothesis that

$$H_0 : \beta_2 = 0$$

for the model fit in Exercise 3 using LRT this time. Run the test and report the resulting table.

```
# solution
fit_null = glm(am ~ mpg + qsec, data = mtcars, family = "binomial")
fit_full = glm(am ~ mpg + hp + qsec, data = mtcars, family = "binomial")
anova(fit_null, fit_full, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: am ~ mpg + qsec
## Model 2: am ~ mpg + hp + qsec
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        29     7.5043
## 2        28     7.4802  1  0.02408   0.8767
```

Exercise 10

Now fit a model that only includes `mpg` and compare that to a model that includes `mpg` and `qsec`. Again, complete a Likelihood Ratio Test to compare the two models. Report the summary result.

```
# solution
fit_one = glm(am ~ mpg, data = mtcars, family = "binomial")
fit_two = glm(am ~ mpg + qsec, data = mtcars, family = "binomial")
anova(fit_one, fit_two, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: am ~ mpg
## Model 2: am ~ mpg + qsec
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        30    29.6752
## 2        29     7.5043  1   22.171 2.494e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exercise 11

Based on your exploration in the previous questions, which model appears to be the strongest at predicting `am` without using unnecessary predictors?

- mpg
- **mpg, qsec**
- mpg, qsec, hp

Exercise 12

Consider the simple logistic relationship between `mpg` and `am`. Create a scatterplot that models how well `mpg` serves as a predictor of `am`, and include the "line" of best fit (the estimated logistic model) to represent our estimated probability that a car is automatic based on its reported mpg.

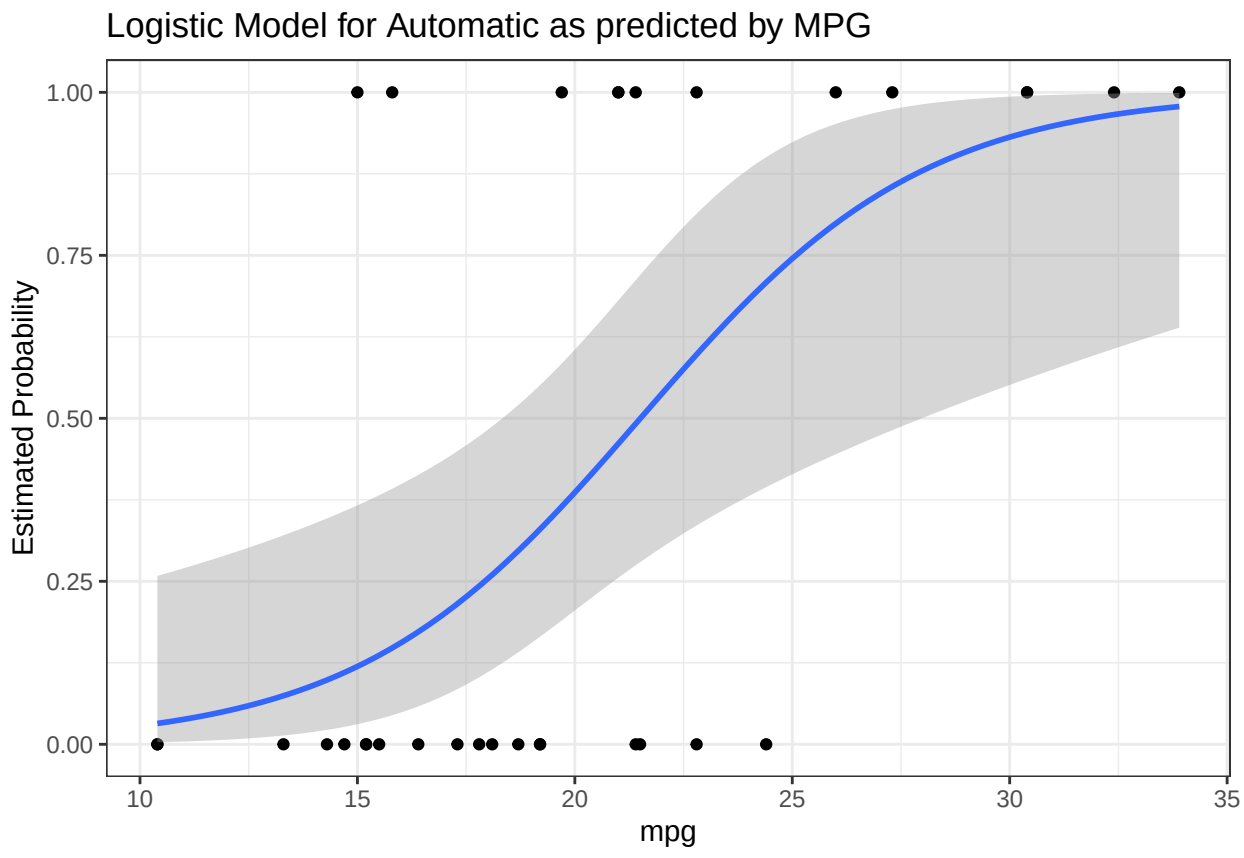
Please title your y axis as "Estimated Probability." All other formatting optional.

You may either use base plot features, or ggplot2. A bit of web searching (especially if using ggplot2) may be of assistance!

```
#solution
simple = glm(am ~ mpg, data = mtcars, family = "binomial")

library(ggplot2)
ggplot(data = mtcars, aes(x = mpg, y = am)) +
  geom_point() +
  geom_smooth(method = "glm", method.args = list(family = "binomial")) +
  labs(y = "Estimated Probability", title = "Logistic Model for Automatic as predicted by MPG") +
  theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```r
#or alternatively

plot(am ~ mpg, data = mtcars,
     pch = 20,
     ylab = "Estimated Probability",
     main = "Logistic Model for Automatic as predicted by MPG")
grid()
curve(predict(simple, data.frame(mpg = x), type = "response"),
      add = TRUE, col = "dodgerblue")
```