

ML Problem Set 2

I worked with Wasil Engel as a partner on this PSet. (UCID: 12231558)

All working code can be located via my [GitHub](#).

Chapter 4 Problems

6.

6 (a)

$$\hat{y} = \Pr[y=1 | x] = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$
$$Y = -6 + (0.05)(40) + (1)(3.5)$$
$$= -0.5$$

132 Equation 4.2. $P(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{e^{-0.5}}{1 + e^{-0.5}} = .606530$

Logit model. Prob(A in class) = 0.37754066879814546.

(B) How many hours would the student need to study to have a 50% chance of getting an A in the class?

$$\Pr(Y=1 | x = ? \text{ hours}) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2}}$$
$$= \frac{e^{-2.5 + (0.05)x_1}}{1 + e^{-2.5 + (0.05)x_1}} = .50$$
$$= e^{-2.5 + 0.05x_1} = .50(1 + e^{-2.5 + 0.05x_1})$$
$$= 1.5 + 1.5 e^{-2.5 + 0.05x_1}$$
$$\frac{1}{2}(e^{-2.5 + 0.05x_1}) = \ln(\frac{1}{2})$$
$$\ln(\frac{1}{2})/e^{-2.5 + 0.05x_1} = -0.301$$
$$-0.301 - 2.5 + 0.05x_1 = -0.301$$
$$-0.301 - 2.5 + 0.05x_1 = -0.301$$
$$0.05x_1 = 2.5$$
$$\frac{0.05}{0.05} x_1 = \frac{2.5}{0.05}$$
$$x_1 = 50 \text{ hrs}$$

7.

?: Predict the probability that
a company will issue a
dividend this year given that
its percentage profit was $X=4$
last year

Bayes Theorem $\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$

$\bar{X}_1 = 10$, $\bar{X}_2 = 0$, $\hat{\sigma}^2 = 36$, 80% of companies issued dividends

$\pi_{\text{res}} = 0.20$, $\pi_{\text{no}} = 0.80$

$\Pr(Y=K | X=x) = \frac{\pi_K f_K(x)}{\sum_{k=1}^K \pi_k f_k(x)}$ = Probability that company $x=4$ will issue a dividend given $y=k$

where $f_k(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_k}{\hat{\sigma}}\right)^2\right)$.

$\bar{X}_1 = 10$, $\pi_{\text{div}} = 0.80$, $f_{\text{div}}(x) = \frac{1}{\sqrt{2\pi \cdot 36}} \exp\left(-\frac{1}{2}\left(\frac{x-10}{6}\right)^2\right)$

$f_{\text{div}}(4) = \frac{1}{\sqrt{2\pi \cdot 36}} \exp\left(-\frac{1}{2}\left(\frac{4-10}{6}\right)^2\right)$

$\pi_{\text{div}} \cdot f_{\text{div}}(4) = 0.8 \cdot \frac{e^{-\frac{(4-10)^2}{72}}}{\sqrt{2\pi \cdot 6}}$

$\pi_{\text{div}} \cdot f_{\text{div}}(4) = 0.8 \cdot \frac{e^{-\frac{36}{72}}}{15.0397} = \underline{\underline{0.0397}}$

$\bar{X}_2 = 0$, $\pi_{\text{nd}} = 0.20$, $f_{\text{nd}}(x) = \frac{1}{\sqrt{2\pi \cdot 6}} \exp\left(-\frac{1}{2}\left(\frac{x-0}{6}\right)^2\right)$

$\pi_{\text{nd}} = 0$, $f_{\text{nd}}(0) = \frac{1}{\sqrt{2\pi \cdot 6}} \exp\left(-\frac{1}{2}\left(\frac{0-0}{6}\right)^2\right)$

$\pi_{\text{nd}} \cdot f_{\text{nd}}(0) = 0.2 \cdot \frac{1}{15.0397} = \underline{\underline{0.0132}}$

$\therefore \Pr(Y=\text{div} | X=4) = \frac{\pi_{\text{div}} \cdot f_{\text{div}}(4)}{\pi_{\text{div}} \cdot f_{\text{div}}(4) + \pi_{\text{nd}} \cdot f_{\text{nd}}(0)}$

$= 0.8 \cdot \frac{\underline{\underline{0.0397}}}{\underline{\underline{0.0397}} + 0.2 \cdot \underline{\underline{0.0132}}} = \underline{\underline{0.751634}}$

$\Pr(Y=\text{div} | X=4) = \underline{\underline{0.751634}} = 75.16\%$

The probability that dividends are given when $X = 4$ is 75.16%

9.

Ch 4, #9

(a) Odds : $\frac{\Pr(Y=1|x)}{1 - \Pr(Y=1|x)}$

$$\frac{\Pr(Y=1|x)}{1 - \Pr(Y=1|x)} = .37$$

$$\Pr(Y=1|x) = .37 - .37(\Pr(Y=1|x))$$

$$+ .37\Pr(Y=1|x)$$

$$\frac{1.37\Pr(Y=1|x)}{1.37} = \frac{.37}{1.37}$$

$$\Pr(Y=1|x) = \frac{.37}{1.37} = .27$$

27% of people with odds of .37 of defaulting will in fact default.

b) 16% chance of defaulting credit card payment. what are the odds that they will default?

$$\Pr(Y=1|x) = 16\%$$

$$\therefore \frac{.16}{1 - .16} = \frac{.16}{.84} = 0.19.$$

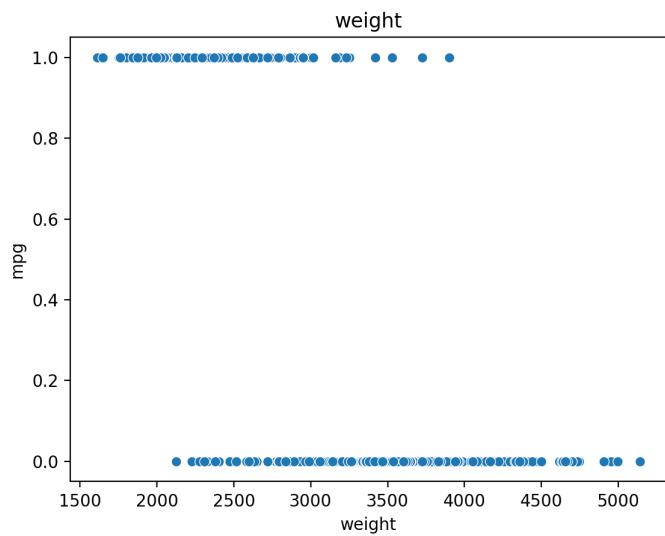
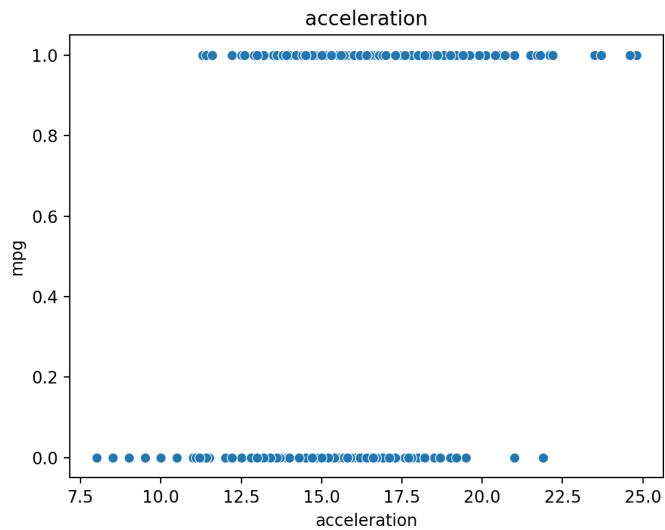
The odds of defaulting are 19%.

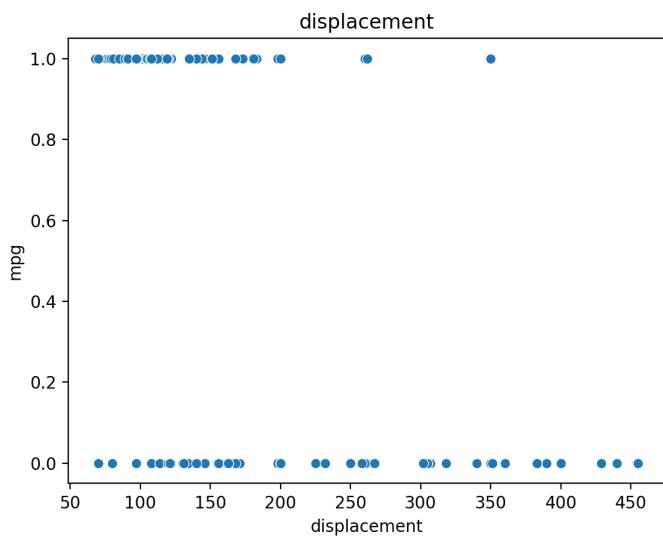
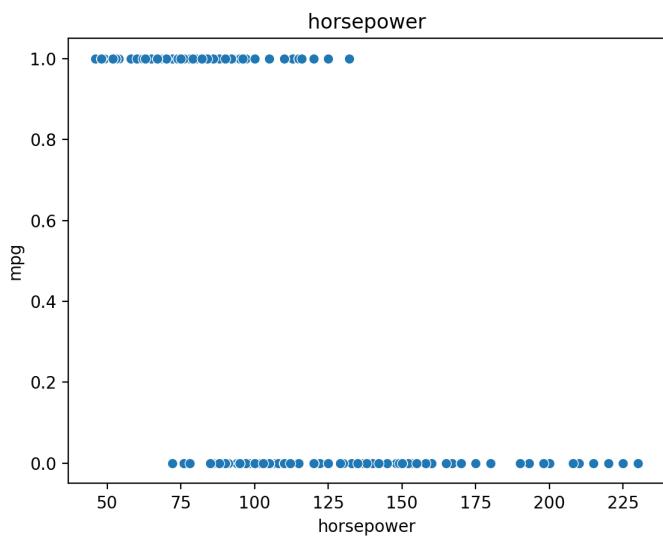
11.

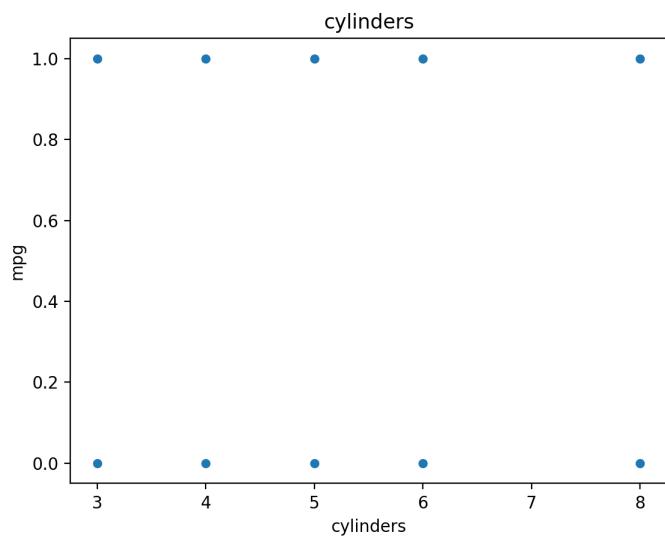
A) Median value is 22.75

B) Checking correlations, as well as a scatterplot, I feel as if cylinders, displacement, horsepower, weight, and acceleration are good predictors for mpg01. They each have high correlations and intuitively make sense in describing miles per gallon. Even if some

variables are negative, I believe these predictors are still valuable to add into our fitted models. I excluded MPG because this would have quickly overfit our model.







C) Check code.

D, E, F, G) For the models proceeding forward, I decided to go with cylinders, displacement, horsepower, weight, and acceleration.

```
Accuracy score for LDA model
0.8571428571428571
Test Error for LDA model
0.1428571428571429

Accuracy score for QDA model
0.8877551020408163
Test Error for QDA model
0.11224489795918369

Accuracy for Logistic Regression model
0.8877551020408163
Test Error for a Logistic Regression model
0.11224489795918369

K-nearest Neighbors
KNN model accuracy
0.8571428571428571
Test error of KNN
0.1428571428571429
```

Floating between 4 and 5, I decided to go with $K = 5$. This gave me the lowest test error, and the highest accuracy score; in reflection, this is the same result we see with our LDA model.

Chapter 5:

5)

Chapter 5

```
part A
Accuracy score
0.9737
Validation set error
0.0262999999999999

part B
part ii: Multiple logistic regression
Accuracy score
0.968
Validation Set Error
0.03200000000000003
part iv
Validation Set Error
0.03200000000000003

part C
Validation set error with set 1  is: 0.027000000000000024
Validation set error with set 2  is: 0.026000000000000023
Validation set error with set 3  is: 0.03433333333333333

part D
Accuracy score Using Balance, Income, and Student status
0.967
Test Error by using Validation Set
0.03300000000000003
```

C) In this case, I observe a validation set error that is different 100% of the time from the prior question. Using different seed values will produce a different subset of possible training and validation sets. Because we are given different answers each time, this is a strong approach, especially when we converge the results from each split so we can arrive at the model's mean accuracy and test error. With what might have appeared to be a high validation set error in part B could actually be much lower when we subset and rerun our model.

D) When we add the dummy variable for student, it appears as if the test error rate increases. There is probable cause that our model could be interpreting more information whenever we add in a dummy variable, therefore changing our prediction. Which a larger test error, we cannot argue that it is insignificant, however, we can argue that the model incorrectly predicts default status by adding in the dummy variable.

