MISSISSIPPI STATE UNIVERSITY

Data Analysis One

Emily Hubbard

October 1, 2025

HW 1

SECTION ONE

Executive Summary:

A. Introduction:
It is important to understand the factors and influences that go into the survival time of leukemia patients. Questions such as: Does a certain amount of sleep increase length of life? Does going through chemotherapy during a certain timezone of illness positively impact survival time? Does white blood count indicate how long a patient will live? These questions are crucial in understanding the illness and being able to accurately predict the lifespan of someone with leukemia. Those are the kinds of questions we will address in this report. Specifically, we will be assessing whether or not white blood count(wbc) or presenc of antigens(ag) indicates longer or shorter survival times in leukemia patients. We will use logistic linear regression to create a model that best predicts whether or not a person will survive for 24 weeks or more given certain values of wbc and ag. We want to know if these two variables, wbc and ag, have a meaningful relationship with survival time under a logistic model.

A logistic linear model is based on the following formula:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \alpha + \beta_i X_i + \epsilon$$

where $\log\left(\frac{p(X)}{1-p(X)}\right)$ is the log-odds or logit.

Using R to create a logistic linear regression model will give us coefficients for each of the covariates. These coefficients gives us the "log-odds" of the effect that covariate has on the response. For example, if the coefficient were 4.35, then we could say that, holding other variables constant, a 1-unit increase in that covariate multiplies the odds of being in category 1 by $e^{4.35}$.

B. Data Collection

Table 7.6: `leuk` data (package **MASS**). Survival times of patients suffering from leukemia.

wbc	ag	time	wbc	ag	time
2300	present	65	4400	absent	56
750	present	156	3000	absent	65
4300	present	100	4000	absent	17
2600	present	134	1500	absent	7
6000	present	16	9000	absent	16
10500	present	108	5300	absent	22
10000	present	121	10000	absent	3
17000	present	4	19000	absent	4
5400	present	39	27000	absent	2
7000	present	143	28000	absent	3
9400	present	56	31000	absent	8
32000	present	26	26000	absent	4
35000	present	22	21000	absent	3
100000	present	1	79000	absent	30
100000	present	1	100000	absent	4
52000	present	5	100000	absent	43
100000	present	65			

Figure 0.1: Leuk Data

The data in Figure 0.1, Table 7.6 shows the survival times of patients diagnosed with leukemia and the values of two explanatory variables, the white blood cell count (wbc) and the presence or absence of antigens within the white blood cells (ag).

The following portion of R code describes the variable types of the original data variables as well as log transformed or factored variables that were added to aid in model creation.

```
> str(leuk)
'data.frame':   33 obs. of  6 variables:
 $ wbc               : int  2300 750 4300 2600 6000 10500 10000 17000 5400 7000 ...
 $ ag                : Factor w/ 2 levels "absent","present": 2 2 2 2 2 2 2 2 2 2 ...
 $ time              : int  65 156 100 134 16 108 121 4 39 143 ...
 $ time.binary       : num  1 1 1 1 0 1 1 0 1 1 ...
 $ time.binary.factor: Factor w/ 2 levels "< 24 weeks","≥ 24 weeks": 2 2 2 2 1 2 2 1 2 2 ...
 $ wbc.log           : num  7.74 6.62 8.37 7.86 8.7 ...
```

Figure 0.2: Leuk Data Variable Types

- **"wbc"** represents the white blood cell count for each of the 33 leukemia patients. It is a quantitative, continuous variable.
- **"ag"** represents the presence or absence of certain antigens in the cells. This variable is qualitative, categorical, and has two levels ("present" and "absent").
- **"time"** represents the survival time in weeks of each patient. This variable is quantitative and continuous.
- **"time.binary"** represents the variable time transformed into a binary outcome where 0 means the patient survived less than 24 weeks and 1 means the patient survived at least 24 weeks. This is a binary, quantitative variable.
- **"time.binary.factor"** represents the same binary outcome recoded as a factor variable with two levels ("> 24 weeks" and "≤ 24 weeks"). To build a logistic regression model, the response variable must be categorical, aka have different factor levels. This redefined variable allows for R to run a logistic regression model on the data. This variable is binary and categorical.
- **"wbc.log"** represents the log transformation of the wbc variable. This will be utilized to create the most efficient model to predict survival time because it helps reduce the effect of extreme outliers. Just like wbc, it remains a quantitative, continuous variable.

To be clear, the response variable will be "time" (in factor form) and the covariates will be "wbc" (log transformed to reduce skew) and "ag".

C. Summary Information

```
> summary(leuk)
      wbc               ag          time
 Min.   :   750   absent :16   Min.   :  1.00
 1st Qu.:  5300   present:17   1st Qu.:  4.00
 Median : 10500                Median : 22.00
 Mean   : 29165                Mean   : 40.88
 3rd Qu.: 32000                3rd Qu.: 65.00
 Max.   :100000                Max.   :156.00
```

Figure 0.3: Basic Summary of Leuk Data

Given this summary of the data, there are a few observations we can make:

1. There is a significant difference in the mean and median of both wbc and time. This indicates that wbc and time columns contain outliers. Some transformations will need to be made on the data to create a more stable and reliable logistic model.

2. The presence of antigens is almost evenly split among the observed patients.

3. "Time" is currently a numerical value and to run a logistic model, we need a categorical response variable. We will fix this issue by transforming time into a binary outcome variable and assigning it to two factor levels, "< 24 weeks" and "≥ 24 weeks".

Let's take a closer look at the response variable and associated covariates with some visuals.

Response Variable: "Time"

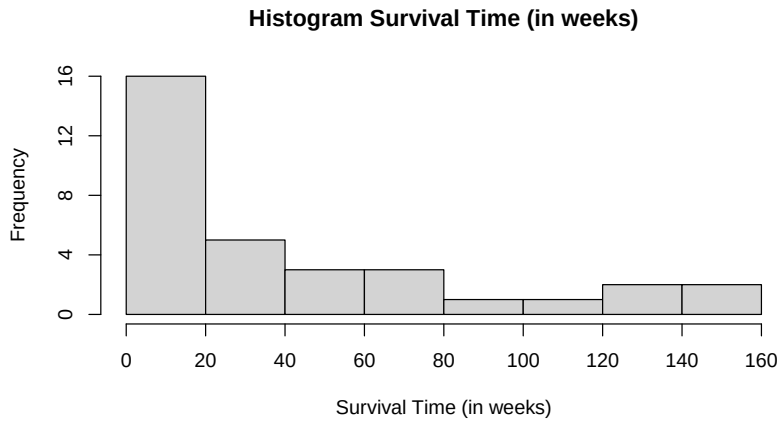

Figure 0.4: Histogram of Survival Time (in weeks)

Observations: It looks as though from this chart, the more freguent outcome of patients is that they live less than 24 weeks. There is also a very wide spread of survival times and it is right skewed. Categorizing them into < 24 weeks or ≥ 24 weeks will remove the effect of the outliers because magnitude of the value will not be taken into account.

Covariate 1: White Blood Count (wbc)

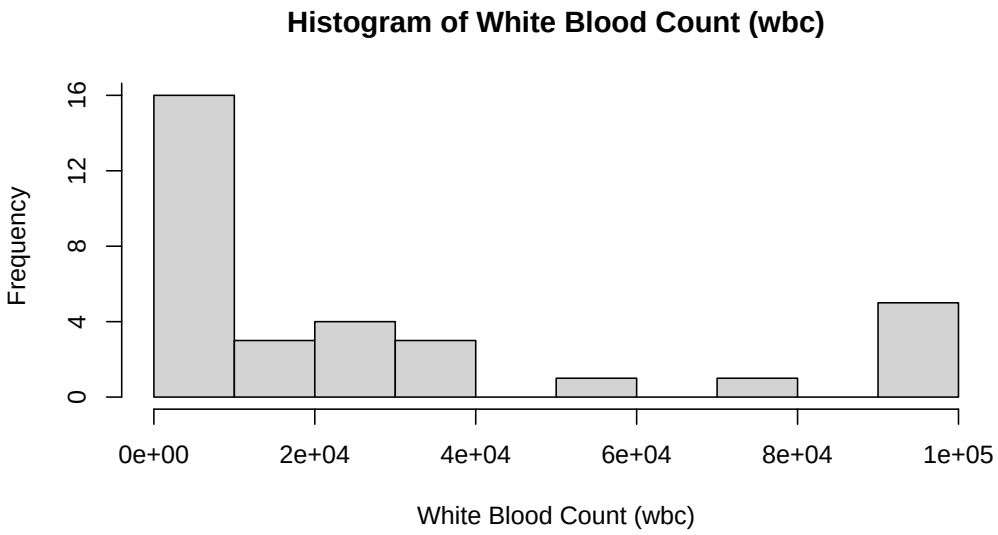

Figure 0.5: Histogram of White Blood Count (wbc)

Observations: This figure allows us to visualize outliers within the white blood count observations. The effect of these outliers will be greatly reduced when we take the log transformation and use that in its place when creating the model.

Covariate 2: Antigen Presence (ag)

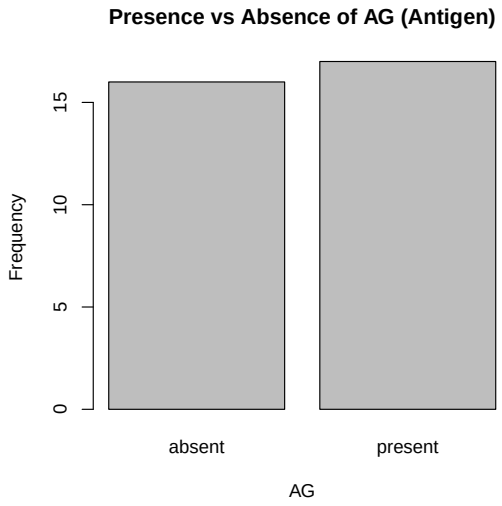

Figure 0.6: Antigen Presence (ag)

Observations: Just as we observed in the summary output from R, figure 0.6 also shows that the presence of antigens is almost evenly split among the observed patients.

Now, we will look at each covariate individually and assess its effect on the response variable.

Visualization of "ag" and "time" together:

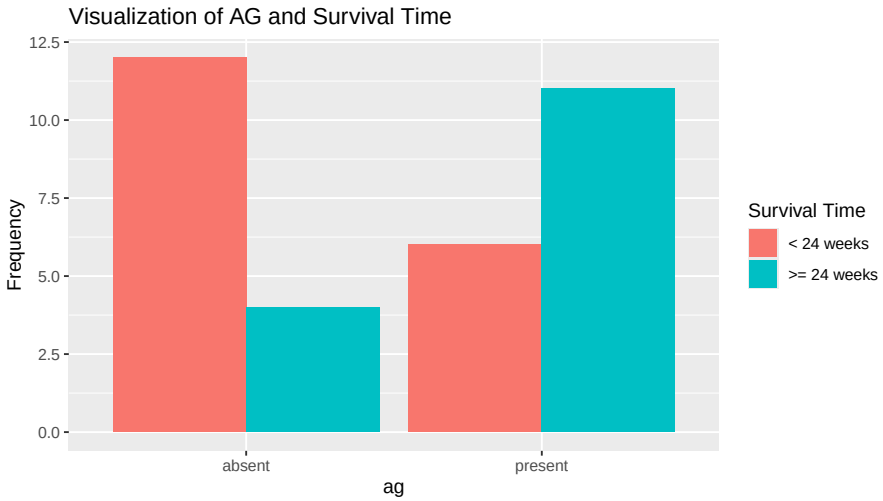

Figure 0.7: ag and Survival Time

Observations: This group bar chart indicates a possible relationship with these two variables. It seems that shorter survival times are more likely to occur when ag is absent.

Visualization of "wbc" and "time" together:

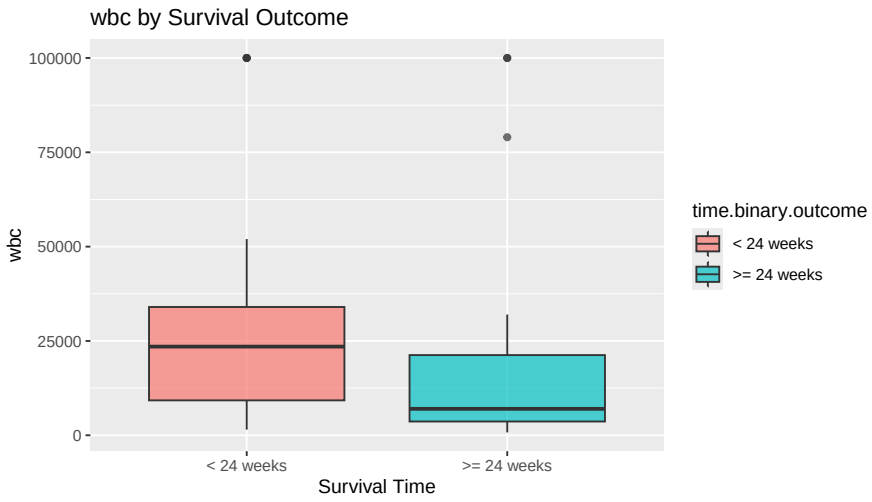

Figure 0.8: wbc and Survival Time

Observations: This box-plot visual tells us a little bit about the relationship between survival time and wbc. It seems as though a lower wbc is more common in those who survived more than 24 weeks. Conversely, a higher wbc seems to be more likely in shorter survival times. This is noticed by looking at the medians (black bars) and noticing their positions in relation to the y-axis.

Overall Observations Given the Summary Data and Visuals: After looking at the response variable and each covariate individually, I wonder if there will be a good logistic regression model to predict the response because there does seem to be a positive association between lower wbc and survival times as well as a positive association between ag presence and survival time.

The question remains: Can we explain the relationship between wbc and ag on time with a logistic regression model? Let's find out.

1. First, we will define a binary outcome variable according to whether or not patients lived for at least 24 weeks after diagnosis.

$$Y = \begin{cases} 0 & \text{if} < 24 \text{ weeks} \\ 1 & \text{if} \geq 24 \text{ weeks} \end{cases}$$

   The response variable Y, survival time, will be split into two categories that will be sorted into 1 or 0. The logistic regression model will help us predict the likelihood of the response variable falling into category 1, which in this case is ≥ 24 weeks.

2. Now, we will fit a logistic regression model to the data.

Summary of the Model:

```
Call:
glm(formula = leuk$time ~ log.wbc + leuk.ag, family = binomial)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.4556     2.9821   1.159   0.2466
log.wbc        -0.4822     0.3149  -1.531   0.1257
leuk.agpresent  1.7621     0.8093   2.177   0.0295 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 45.475  on 32  degrees of freedom
Residual deviance: 37.498  on 30  degrees of freedom
AIC: 43.498

Number of Fisher Scoring iterations: 3
```

Figure 0.9: R Summary Output

Regression Line Equation:

$$\text{Intercept}(\alpha) = 3.456$$
$$\text{Coefficient } 1(\beta_1) = -0.482$$
$$\text{Coefficient } 2(\beta_2) = 1.762$$
$$\log\left(\frac{p(X)}{1-p(X)}\right) = 3.456 - 0.482(\text{log.wbc}) + 1.762(\text{ag}) + \epsilon$$

Summary Statistics:

$$\frac{\text{Residual Deviance}}{\text{DF}} = \frac{37.498}{30} = 1.25$$
$$\text{AIC} = 43.498$$

Interpretation of the Summary:
The presence of antigens are shown to have a significant impact on the response variable survival time. This is indicated by the p-value being less than a standard 5 percent

α(0.03). People with the presence of antigens have a 5.8 times higher chance of surviving at least 24 weeks. This is calculated by taking e and raising it to the coefficient corresponding to AG. e^1.76 = 5.8. This logistic regression model suggests that white blood count does not have a significant impact on survival time. This is also shown by the p-value being much greater than 5 percent(0.13). To check whether or not this model is overfit, we divide the residual deviance by the degrees of freedom. In this case, we get 1.25 as stated above which is below the threshold of 1.5, indicating this model is not overfit. This is good. This means that our data, if significant, can be applied to other data sets and at least somewhat help in predicting the response variable outcome.

3. Finally, we will construct some graphics useful in the interpretation of the fitted model.

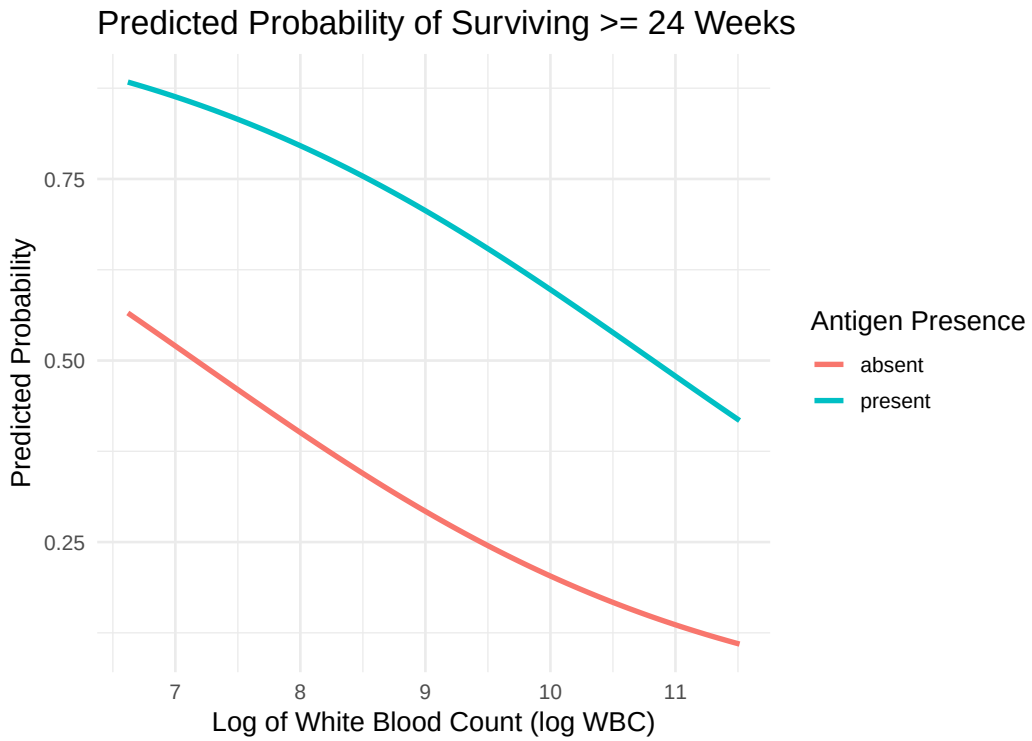

Figure 0.10: Predicted Probabilities of Survival time ≥ 24 weeks

Explanation:
This plot shows us the probabilities of the response variable falling into category 1 (≥ 24 weeks) given antigen presence and white blood count. The blue line represents patients with antigens present in their white blood cells. These people overall have a greater chance of surviving at least 24 weeks, compared to their absent antigen counterparts. It is also important to note that when antigens are present or absent, survival

times decrease as the white blood count increases. So, best case scenario for survival is to have a lower white blood count with antigens present.

SECTION TWO

Executive Summary:

A. Introduction:
The stock market is an important part of the modern financial landscape. It is also a way to make a lot of money. Many people would kill to know just what the stock market will do on any given day, so many a statistical model has been built in an attempt to predict the direction and magnitude in which the stock market will change. We will attempt to predict the direction of the market in a given week with the information of 9 covariates and a logistic regression model. Will we be able to fit a model that can make meaningful predictions about the stock market?

As previously stated, a logistic linear model is based on the following formula:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \alpha + \beta_i X_i + \epsilon$$

$$\text{where } \log\left(\frac{p(X)}{1-p(X)}\right) \text{ is the log-odds or logit.}$$

B. Data

The data that we are working with is from the ISLR R package and consists of the weekly percentage returns for the S and P 500 stock index between 1990 and 2010.

The following portion of R code describes the variable types of the original data variables.

```
> str(Weekly)
'data.frame':   1089 obs. of  9 variables:
 $ Year     : num  1990 1990 1990 1990 1990 1990 1990 1990 1990 1990 ...
 $ Lag1     : num  0.816 -0.27 -2.576 3.514 0.712 ...
 $ Lag2     : num  1.572 0.816 -0.27 -2.576 3.514 ...
 $ Lag3     : num  -3.936 1.572 0.816 -0.27 -2.576 ...
 $ Lag4     : num  -0.229 -3.936 1.572 0.816 -0.27 ...
 $ Lag5     : num  -3.484 -0.229 -3.936 1.572 0.816 ...
 $ Volume   : num  0.155 0.149 0.16 0.162 0.154 ...
 $ Today    : num  -0.27 -2.576 3.514 0.712 1.178 ...
 $ Direction: Factor w/ 2 levels "Down","Up": 1 1 2 2 2 1 2 2 2 1 ...
```

Figure 0.11: Weekly Data Variable Types

- **"Year"** The year that the observation was recorded. This is a discrrete variable.
- **"Lag1 - Lag5"** Percentage return for n weeks previous (n= 1,2,3,4,5). Each Lag is a continuous, quantitative variable.
- **"Volume"** Volume of shares traded (average number of daily shares traded in billions). This is a continuous, quantitative variable.
- **"Today"** Percentage return for this week. This is also a continuous, quantitative variable.
- **"Direction"** levels Down and Up indicating whether the market had a positive or negative return on a given week. This variable is a two-level factor with binary outcome 0 referring to "Down" and 1 meaning "Up".

To be clear, the response variable will be "direction" and the covariates will be "Lag 1 - Lag5" (log transformed to reduce skew) and "Volume". We are not going to include the "Year" or the "Today" column in our model.

C. Summary Information

Numerical Summary

```
> summary(Weekly)
      Year           Lag1               Lag2               Lag3               Lag4         
 Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950  
 1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580   1st Qu.: -1.1580  
 Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410   Median :  0.2380  
 Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472   Mean   :  0.1458  
 3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090   3rd Qu.:  1.4090  
 Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260  
      Lag5             Volume            Today          Direction  
 Min.   :-18.1950   Min.   :0.08747   Min.   :-18.1950   Down:484  
 1st Qu.: -1.1660   1st Qu.:0.33202   1st Qu.: -1.1540   Up  :605  
 Median :  0.2340   Median :1.00268   Median :  0.2410             
 Mean   :  0.1399   Mean   :1.57462   Mean   :  0.1499             
 3rd Qu.:  1.4050   3rd Qu.:2.05373   3rd Qu.:  1.4050             
 Max.   : 12.0260   Max.   :9.32821   Max.   : 12.0260             
```

Figure 0.12: Basic Summary of Weekly Data

Given this summary of the data, there are a few observations we can make:

a) The medians and means of all of the Lag variables have similar values of difference. Each of them having around a 0.10 difference between median and mean.

b) There are more "Up" observations than "Down".

c) There will be no need to transform any existing variables because they are already in perfect form to run a logistic regression model.

Graphical Summary

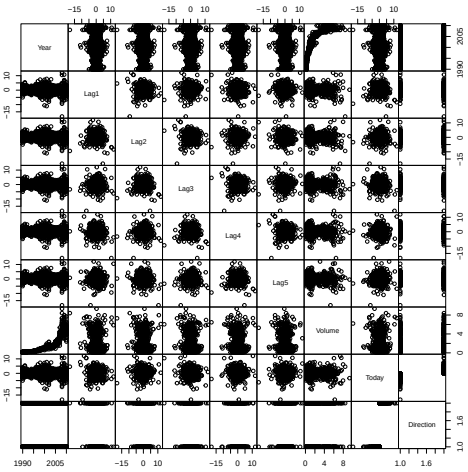

Figure 0.13: Correlation Chart (Weekly Data Set)

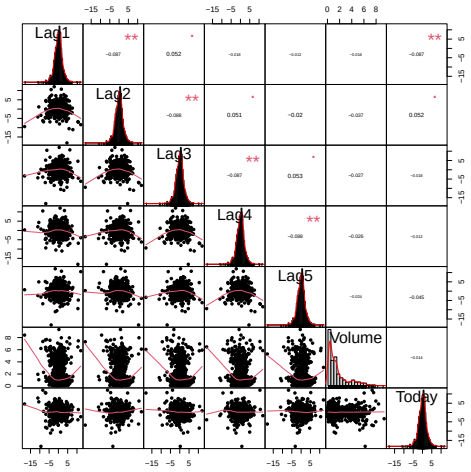

Figure 0.14: Spearman Correlation Chart (Weekly Data Set)

Interpretation:
Looking at both the normal correlation chart (Figure 0.13) as well as the spearman correlation chart (Figure 0.14), there seems to be very little, if any, correlation between lags. There are some significant correlation values shown the spearman chart, however, they are so small that with such a large data set, it does not make a meaningful difference. Based on the first correlation chart, we can see a positive exponential relationship between time and volume, which makes sense. Since the development of the stock market, increasingly more people use it and the network affect occurs. As more people use it, it becomes a better market tool, meaning even more people want to participate. This is an easy explanation for the pattern.

Now, we will create our logistic regression model to answer the question: Will we be able to fit a model that can make meaningful predictions about the stock market?

a) We will use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Then we wil use the

summary function to print the results. We will then discuss if any of the predictors appear to be statistically significant and if so, which ones.

$$Y = \begin{cases} 0 & \text{if Down} \\ 1 & \text{if Up} \end{cases}$$

Summary of the Model:

```
Call:
glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
    Volume, family = binomial, data = Weekly)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.26686    0.08593   3.106   0.0019 **
Lag1        -0.04127    0.02641  -1.563   0.1181
Lag2         0.05844    0.02686   2.175   0.0296 *
Lag3        -0.01606    0.02666  -0.602   0.5469
Lag4        -0.02779    0.02646  -1.050   0.2937
Lag5        -0.01447    0.02638  -0.549   0.5833
Volume      -0.02274    0.03690  -0.616   0.5377
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 0.15: R Summary Output

Regression Line Equation:

$$\text{Intercept}(\alpha) = 0.267$$
$$\text{Coefficient } 1(\beta_1) = -0.041$$
$$\text{Coefficient } 2(\beta_2) = 0.058$$
$$\text{Coefficient } 3(\beta_3) = -0.016$$
$$\text{Coefficient } 4(\beta_4) = -0.028$$
$$\text{Coefficient } 5(\beta_5) = -0.014$$
$$\text{Coefficient } 6(\beta_6) = -0.023$$
$$\log\left(\frac{p(X)}{1-p(X)}\right) = 0.267 - 0.041(\text{Lag1}) + 0.058(\text{Lag2}) - 0.016(\text{Lag3})$$
$$- 0.028(\text{Lag4}) - 0.014(\text{Lag5}) - 0.023(\text{Volume}) + \epsilon$$

Summary Statistics:

$$\frac{\text{Residual Deviance}}{\text{DF}} = \frac{1496}{1082} = 1.38$$
$$\text{AIC} = 1500$$

Interpretation of the Summary:
Using the summary function in R, it can be seen that only one covariate is calculated to be significant. Based on its p-value less than 0.05, Lag 2 is significant in predicting the response variable direction. No other covariate has a significant p-value. When dividing Residual deviance by degrees of freedom to check the fit of the model, the result indicates it is close to being overfit. Greater than 1.5 means a model is overfit and this model is 1.38. It also has a 1500 AIC score. This will become more meaningful in comparison to another model.

b) We will now compute the confusion matrix and explain what it is telling us about the mistakes made by logistic regression. We will then calculate what the misclassification rate is. This number will tell us the percentage of times the model makes the wrong prediction about the response variable's category.

Confusion Matrix:

```
> conf.mat
          Actual
Predicted Down  Up
     Down   54  48
     Up    430 557
```

Figure 0.16: Confusion Matrix for Full Model

Misclassification Rate: 0.4389 or 43.89 percent

Interpretation:
This misclassification stat tells us that 44 percent of the time, this model will predict the wrong category for the response variable to go in. This is supported by the confusion matrix. 430 observations were misclassified as "Up" when in reality they were "Down". Another 48 observations were categorized as "Down" when they were actually "Up". This indicates just how faulty this model is. It is not surprising since only one covariate was calculated to be significant.

c) Now, what if we were to use only Lag 2 (the significant covariate) in a logistic regression model to predict Direction. Would that help in increasing the accuracy of the model? That is what we will see. In this next model, we will also use a training set and a test set so that we can more officially test if this model gives us the right predictions.

Summary of the Model (Only Lag2):

```
Call:
glm(formula = Direction ~ Lag2, family = binomial, data = Weekly.train)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.20326    0.06428   3.162  0.00157 **
Lag2         0.05810    0.02870   2.024  0.04298 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1354.7  on 984  degrees of freedom
Residual deviance: 1350.5  on 983  degrees of freedom
AIC: 1354.5
```

Figure 0.17: R Summary Output

Regression Line Equation:

$$\text{Intercept}(\alpha) = 0.203$$
$$\text{Coefficient } 1(\beta_1) = 0.058$$
$$\log\left(\frac{p(X)}{1-p(X)}\right) = 0.203 + 0.058(\text{Lag2}) + \epsilon$$

Summary Statistics:

$$\frac{\text{Residual Deviance}}{\text{DF}} = \frac{1350.5}{983} = 1.37$$
$$\text{AIC} = 1354.5$$

Interpretation of the Summary:
This model has a slightly lower AIC score, but that is the only number in this summary output that can give us any indication of how it compares to the other test.

A confusion matrix for this model followed by a misclassification rate can more clearly indicate which model is superior if any.

Confusion Matrix:

```
> conf.mat.2
         Direction.test
glm.pred Down  Up
    Down   32  25
    Up    451 580
```

Figure 0.18: Confusion Matrix for Lag2 Only Model

Misclassification Rate: 0.4375 or 43.75 percent

Interpretation:
Compared to the full model, this model that only takes into account Lag 2 has about the same amount of predictive power. This can be supported by a few things. The misclassification rate is a mere 0.0012 lower than with the full model. There were 476 misclassified observations out of 1088. The only other thing that points to this model being superior is the slight decrease in the AIC score from 1500 (full model) to 1354 (current model). However, this model still does not do a great job of predicting the likelihood of the response variable falling into the "Up" Direction category.

d) We are now going to repeat this process using the Linear Discriminant Analysis Model (LDA) and eventually compare which if either of the models are superior.

Confusion Matrix:

```
> conf.mat.3
      Direction.test
       Down  Up
  Down   31  25
  Up    452 580
```

Figure 0.19: Confusion Matrix for LDA Model

Misclassification Rate: 0.4384 or 43.84 percent

Interpretation:

e) We are now going to repeat this process using the Quadratic Discriminant Analysis Model (QDA) and compare if this model is better than any of the previous ones.

Confusion Matrix:

```
> conf.mat.4
      Direction.test
       Down  Up
Down    0   0
Up    483 605
```

Figure 0.20: Confustion Matrix for QDA Model

Misclassification Rate: 0.4439 or 44.39 percent

Interpretation:
This QDA model does not increase the accuracy. In fact, it decreases it by a small amount. This again points to Lag2 not having enough significane for the test to differentiate the correct category for the response variable. Especially for this test. It predicted only "Up" classifications.

f) Considering each of these three models, which of these methods appears to provide the best results on this data?

All three of these models, Logistic, LDA, and QDA, come out with very similar predictions and misclassifications rates. All of them being approximately 56 percent accurate and 44 percent wrong. This is barely better than flipping a coin to choose a category. There confusion matrices, though slightly different, hold pretty much the same amount of misclassified information. The Logistic Regression Model is the best of the 3 based on smallest misclassification rate though the difference is practically negligible. Overall, I would say that this model, nor any of the others in this report, can accurately predict the response variable direction.

APPENDIX

R code:

```
  #### HW 3 ####
#### Section One ####
library(MASS)

##Visualization and Understanding of Original Data:
?leuk
data("leuk")
View(leuk)
names(leuk)
# Y: "time" (in weeks) X's: "wbc" (white blood count) "ag" (antigen presence or abs

dim(leuk)
# 3x33

str(leuk)

summary(leuk)
#looking at "time", it is noticeable how big a difference there is between the medi
#the mean. The median, 22, is a better representation of this data set of 33 people
#the max observation skews the rest of the sample.

##Visualization of Response Variable "Time":
hist(leuk$time,xlab = "Survival␣Time␣(in␣weeks)", main = "Histogram␣Survival␣Time␣(
axis(1, at = seq(0, 160, by = 20))
axis(2, at = seq(0, 16, by = 4))

##Visualization of Response Variable "wbc":
hist(leuk$wbc,xlab = "White␣Blood␣Count␣(wbc)", main = "Histogram␣of␣White␣Blood␣Co
axis(2, at = seq(0, 20, by = 4))

##Visualization of Response Variable "ag":
str(leuk$ag)
table(leuk$ag)
contrasts(leuk$ag)

barplot(table(leuk$ag),main = "Presence␣vs␣Absence␣of␣AG␣(Antigen)", xlab = "AG",yla
#What I notice from this barplot is that it is pretty 50/50 on whether or not the a

##Visualize "AG" and "Time" together:
library(ggplot2)
```

```
ggplot(leuk, aes(x =  ag, fill = time.binary.factor)) +
  geom_bar(position = "dodge") +
  labs(title = "Visualization of AG and Survival Time", x = "ag", y = "Frequency",
#This group bar chart indicates a possible relationship with these two variables. I
#shorter survival times are more likely to occur when ag is absent.

##Visualize ""WBC" and "Time" together:
ggplot(leuk, aes(x = time.binary.factor, y = wbc, fill = time.binary.factor)) +
  geom_boxplot(alpha = 0.7) +
  labs(title = "wbc by Survival Outcome", x = "Survival Time", y = "wbc")
#This box-plot visual tells us a little bit about the relationship between survival
#It seems as though a lower wbc is more common in those who survived more than 24 u
#This is noticed by looking at the medians (black bars) and noticing their position

##After looking at the response variable and each covariate individually, I wonder
#to predict the response because there does seem to be a positive association betwe
#positive association between ag presence and survival time. So maybe, it doesn't m
#the antigen is present withing them. So, can we explain this relationship with a l


#### 1 ####
leuk$time.binary <- ifelse(leuk$time < 24, 0, 1) #creating a binary outcome
leuk$time.binary.factor <- factor(leuk$time.binary, levels = c(0, 1), labels = c("<

##Data Checking:
str(leuk$time.binary.factor) #checking that it was properly converted into a factor
table(leuk$time.binary.factor)  #counts within factor levels
contrasts(leuk$time.binary.factor) #shows us the number assigned to each level...
#In our logistic regression, we will be calculating the probability that the respon

#### 2 ####
leuk$wbc.log <- log(leuk$wbc)  #natural log transformation for wbc
str(leuk$wbc.log)

##Visualize the log transformation of wbc
hist(leuk$wbc.log,xlab = "White Blood Count (wbc)", main = "Histogram of Log Transfo

model.1 <- glm(time.binary ~ wbc.log + ag, data = leuk, family = binomial)
model.1

summary(model.1)
exp(model.1$coefficients) #This indicates what odds a patient has of surviving at l
#Logistic Equation: logit = 3.4556 - 0.4822(log.wbc) + 1.7621 (leuk.agpresent)
```

```
#Interpretation: The presence of antigens have shown a significant impact on the re
#This is indicated by the p-value being less than a standard 5% alpha. People with
#higher chance of surviving at least 24 weeks. This is calculated by exponentiating
#blood count does not have a significant impact on survival time. This is also show
#To check whether or not this model is overfit, we divide the residual deviance by
# get 1.25 which is below the threshhold of 1.5 which would indicate an overfit mod
#to other data sets and at least somewhat help in predicting the response variable


model.2 <- glm(leuk$time.binary ~ leuk$wbc + leuk$ag, family = binomial)
model.2
#Logistic Equation: logit = -8.706e-01 - 8.436e-06(wbc) + 1.733 (leuk.agpresent)
#This is a test of what the model would look like in if we did not take the natural
#this model underperforms model 1 slightly. Model 1 is also slightly easier for int


#### 3 ####
##Graphic for model interpretation:
# 2) Build prediction grid
newdata <- with(leuk, expand.grid(
  wbc.log = seq(min(wbc.log), max(wbc.log), length.out = 200),
  ag      = levels(ag)
))

# 3) Predict from the model
newdata$predicted_prob <- predict(model.1, newdata = newdata, type = "response")

# Plot
library(ggplot2)
ggplot(newdata, aes(x = wbc.log, y = predicted_prob, color = ag)) +
  geom_line(size = 1.2) +
  labs(title = "Predicted␣Probability␣of␣Surviving␣  ␣24␣Weeks",
       x = "Log␣of␣White␣Blood␣Count␣(log␣WBC)",
       y = "Predicted␣Probability", color = "Antigen␣Presence") +
  theme_minimal(base_size = 13)



#### Section Two ####
library(ISLR)
##Visualization and Understanding of Original Data:
?Weekly
data("Weekly")
```

```
View(Weekly)
names(Weekly)
# "Year"       "Lag1"       "Lag2"       "Lag3"       "Lag4"       "Lag5"
"Volume"     "Today"      "Direction"
dim(Weekly)
# 1089 x 9
str(Weekly)
#All are continuous except for direction, which is a factor.

summary(Weekly)

contrasts(Weekly$Direction) #Up is 1 which means the odds that we calculate will in

#### 1 ####
##Visualization of Original Data
par(mex=0.5)     #test for independent covariates
pairs(Weekly, gap=0, cex.labels=0.9)
cor(Weekly)

library(PerformanceAnalytics)

chart.Correlation(Weekly[,-c(1,9)],method="spearman",histogram=TRUE,pch=16)

##Looking at both the normal correlation chart as well as the spearman correlation
#There are some significant correlation values, however, they are so small that wit
#Based on the first correlation chart, we can see an positive exponential relationsh
#stock market, increasingly more people use it the net work affect occurs. As more
#want to participate. This is an easy explanation for that pattern.

#### 2 ####
model.3 <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume, data = Weekly
model.3

summary(model.3)
#Using the summary function in R, it can be seen that only one covariate is calcula
#Based on its p-value less than 0.05, Lag 2 is significant in predicting the respon
#No other covariate has a significant p-value. When dividing Residual deviance by d
#the fit of the model, indicates it is close to being overfit. Greater than 1.5 mea
#model is 1.37. It also has a 1500 AIC score. This will become more meaningful in c

#### 3 ####
prob <- predict(model.3, type = "response")
predicted.classes <- ifelse(prob > 0.5, "Up", "Down")
conf.mat <- table(Predicted = predicted.classes, Actual = Weekly$Direction)
```

```
conf.mat

accuracy <- mean(predicted.classes == Weekly$Direction)
misclassification_rate <- 1 - accuracy
accuracy #=0.5611
misclassification_rate #=0.4389
##This misclassification stat tells us that 44 % of the time, this model will predi
#This is supported by the confusion matrix. 430 observations were misclassified as
#were categorized as "Down" when they were actually "Up". This indicates just how f
#calculated to be significant.

#Now, what if we were to use only Lag 2 (the significant covariate) in a logistic r
#Would that help in increasing the accuracy of the model? That is what we will see.
#a training set and a test set so that we can more officially test if this model gi

#### 4 ####
##Training data:
train <- (Weekly$Year < 2009)
Weekly.train <- Weekly[train, ]
Weekly.test <- Weekly[-train, ]

model.4 <- glm(Direction ~ Lag2, data = Weekly.train, family = binomial)
model.4

summary(model.4)

##Predication using test data
glm.probs <- predict(model.4, newdata = Weekly.test, type = "response")
glm.pred <- rep("Down", nrow(Weekly.test))
glm.pred[glm.probs > 0.5] <- "Up"

# Confusion matrix
Direction.test <- Weekly.test$Direction
conf.mat.2 <- table(glm.pred, Direction.test)
conf.mat.2
accuracy.2 <- mean(glm.pred == Direction.test)
misclassification_rate.2 <- 1 - accuracy.2
accuracy.2 #=0.5625
misclassification_rate.2 #=0.4375
#Compared to the full model, this model that only takes into account Lag 2 has abou
#This can be supported by a few things. The misclassification rate is a mere 0.0012
#The only other thing that points to this model being superior is the slight decrea
#However, this model still does not do a good job of predicting the likelihood of t
```

```
#### 5 ####
##LDA Model Fitting
model.4.lda <- lda(Direction ~ Lag2, data = Weekly.train)
model.4.lda

lda.pred <- predict(model.4.lda, Weekly.test)

Direction.test <- Weekly.test$Direction
conf.mat.3 <- table(lda.pred$class, Direction.test)
conf.mat.3
accuracy.lda <- mean(lda.pred$class == Direction.test)
misclassification_rate.lda <- 1 - accuracy.lda
accuracy.lda #=0.5616
misclassification_rate.lda #=0.4384

#### 6 ####
##QDA Model Fitting
model.4.qda <- qda(Direction ~ Lag2, data = Weekly.train)
model.4.qda

qda.pred <- predict(model.4.qda, Weekly.test)
Direction.test <- Weekly.test$Direction
conf.mat.4 <- table(qda.pred$class, Direction.test)
conf.mat.4
accuracy.qda <- mean(qda.pred$class == Direction.test)
misclassification_rate.qda <- 1 - accuracy.qda
accuracy.qda #= 0.5560
misclassification_rate.qda #0.4439
##This QDA model does not increase the accuracy. In fact, it decreases it by a smal
#significane for the test to differentiate the correct category for the response va
#"Up" classifications.


#### 7 ####
#All three of these models, Logistic, LDA, and QDA, come out with very similar pred
#This is barely better than flipping a coin to choose a category. There confusion m
#The Logistic Regression Model is the best of the 3 though the difference is practi
#Overall, I would say that this model model, nor any of the others in this report,
```