Report 2: Logistic Regression for Modeling Leukemia Treatment

Nathan Klimt

March 11, 2024

Context

A study involved 51 untreated adult patients with acute myeloblastic leukemia who were given a course of treatment, after which they were assessed as to their response. To respond to treatment means that the treatment begins to fight back at the leukemia. The data was saved in the csv file provided. You can load the data into your Rproject environment by running the following:

```
# setting header to TRUE reads the first row of the csv as titles for the columns
leukemia_df <- read.csv("leukemia_data.csv", header=TRUE)
```

The data set contains the following 9 variables for 51 observations.

Variable	Description
`Age`	Age at diagnosis (in years)
`Smear`	Differential percentage of blasts
`Infil`	Percentage of absolute marrow leukemia infiltrate
`Index`	Percentage labeling index of the bone marrow leukemia cells
`Blasts`	Absolute number of blasts, in thousands
`Temp`	Highest temperature of the patient prior to treatment (in tenths of °F)
`Resp`	1=responded to treatment or 0=failed to respond
`Time`	Survival time from diagnosis (in months)
`Status`	0=dead or 1=alive

Instructions

Your report must be made using R markdown. Your submission must include both the Rmd file and the outputted PDF. Make sure your code, graphs, and results are displayed on the PDF clearly. When performing calculations in this report, you do not need to typeset them (but typesetting with LaTeX is welcome!)

You must submit your own individual report, but you are free to consult with up to two other classmates and please identify them clearly at the beginning of your report.

Each question on the assignment will be graded on a 0-3 scale with

- 0: No attempt - problem has not been attempted faithfully
- 1: Major revisions needed - problem needs to be redone due to incompleteness or many errors
- 2: Minor revisions needed - problem is almost complete or there are a couple of minor errors
- 3: Full credit - problem is completed fully and there are no errors

Part A

This part of the report can be completed as Chapter 9 is covered.

Model 1: Logistic Regression Using `Age`

1. Write down the equation for a binary logistic regression model using `Age` as the predictor variable to predict the probability of `Resp`. Use the Logit form of the logistic regression model.

$logit(p) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 \times Age$

2. Fit a logistic model using `Resp` as the response variable and `Age` as the predictor variable. Interpret the slope coefficient (in terms of an odds ratio) and interpret the test for the slope. Be sure to do these in the context of your data situation. (Hint: Use the `glm()` function with the parameter `family="binomial"`. For your own knowledge, see what happens if you don't set `family="binomial"`.)

```
logistic_model <- glm(Resp ~ Age, data = leukemia_df, family = binomial(link = "logit"))

summary(logistic_model)
```

```
##
## Call:
## glm(formula = Resp ~ Age, family = binomial(link = "logit"),
##     data = leukemia_df)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.19678    1.00548   2.185   0.0289 *
## Age         -0.04676    0.01952  -2.395   0.0166 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.524  on 50  degrees of freedom
## Residual deviance: 64.004  on 49  degrees of freedom
## AIC: 68.004
##
## Number of Fisher Scoring iterations: 4
```

Given the above information, there is an odds ratio of approximately 0.954, which means that for every one-year increase in Age, the odds of responding to treatment decrease by approximately 4.6%.

Given the above information, Age has a statistically significant effect on the probability of responding to treatment for Leukemia.

3. Compute a 95% confidence interval for your slope and use it to find a confidence interval for the odds ratio. Does your interval around the odds ratio include the value 1? Why does that matter?

```r
coef_age <- coef(logistic_model)["Age"]
std_err_age <- summary(logistic_model)$coefficients["Age", "Std. Error"]

conf_int_slope <- coef_age + c(-1, 1) * qnorm(0.975) * std_err_age
conf_int_slope
```

```
## [1] -0.085027981 -0.008492264
```

Given the above information, the confidence interval for the slope coefficient (Age) is approximately (-0.0850, -0.0085). This means that we are 95% confident that the true coefficient for Age falls within this interval.

Now we will calculate the confidence interval for the odds ratio.

```r
conf_int_odds_ratio <- exp(conf_int_slope)
conf_int_odds_ratio
```

```
## [1] 0.9184866 0.9915437
```

Given the above information, the confidence interval for the odds ratio associated with a one-unit increase in Time is approximately (0.918, 0.992). More specifically, for every one-unit increase in Time, the odds of responding to treatment decrease by approximately 8.2% to 0.8%.

Now we will check whether the interval includes the value 1.

```r
conf_int_odds_ratio[1] <= 1 & conf_int_odds_ratio[2] >= 1
```

```
## [1] FALSE
```

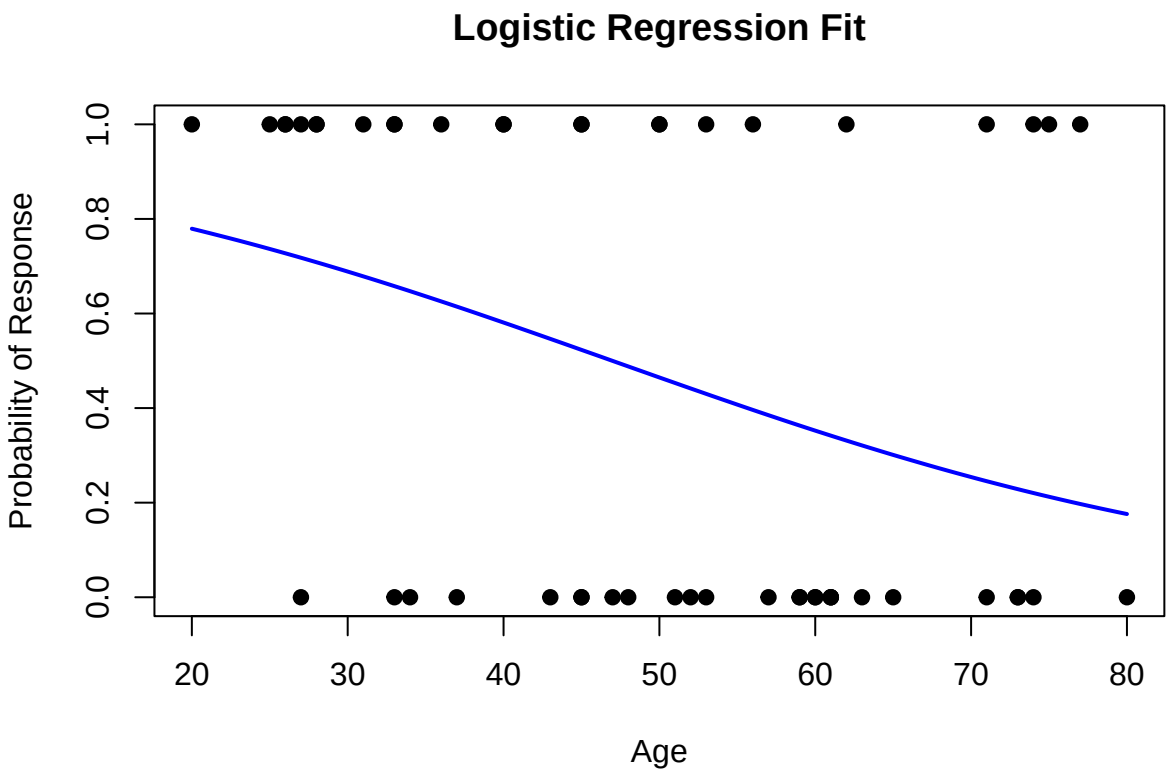
Since the confidence interval for the odds ratio does not include the value 1, it suggests that there is a statistically significant association between Time and the probability of responding to treatment for Leukemia.

4. Assess the model by generating a plot of the logistic fit. Comment on what you learn from the plot. How good is this logistic model at modeling `Resp`? (Hint: You may want to use the `curve()` function as used in the R manual).

```r
age_seq <- seq(min(leukemia_df$Age), max(leukemia_df$Age), length.out = 100)

predicted_probs <- predict(logistic_model, newdata = data.frame(Age = age_seq), type = "response")

plot(leukemia_df$Age, leukemia_df$Resp, pch = 19, xlab = "Age", ylab = "Probability of Response", main =
lines(age_seq, predicted_probs, col = "blue", lwd = 2)
```



This plot shows a negative correlation between the Probability of Response and Age. This model is not very accurate at modeling Resp. Observing the chart, we can see many outliers where the probability is 0 at a young age and 1 at an old age. We can also see that the Probability of Response goes down as Time increases.

5. Show (by hand) how to use the fitted model for predicting the probability of response to treatment for someone who is 50 years old and has a survival time from diagnosis of 5 months. Using a 0.5 threshold, does your calculation suggest this person will respond to the treatment or fail to respond? (Hint: Use the `predict()` function and set the parameter `type="response"`. For your own knowledge, investigate what happens if you don't set `type="response"`.)

Given the following logistic regression equation: $logit(p) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 \times Age$, and the coefficients obtained from the fitted model: $\beta_0 = 2.19678$ and $\beta_1 = -0.04676$, we can predict the response to treatment for a 50-year-old. Substituting the above values into the equation, we get $logit(p) = 2.19678 - 0.04676 * 50$. By simplifying this equation, we get $logit(p) = -0.14122$.

Next, we will convert the log odds to probabilities using the logistic function. Given the equation $p = \frac{1}{1+e^{-\text{logit}(p)}}$, we can plug in the above numbers into it, giving us $p = \frac{1}{1+e^{0.14122}}$, which can be simplified to $p = 0.465$. Given this information, the predicted probability of response to treatment for a 50-year-old individual is approximately 0.465. Based on the 0.5 threshold, since the predicted probability is less than 0.5, this calculation suggests that the person is predicted to fail to respond to the treatment.

```
coef_age <- coef(logistic_model)["Age"]
coef_intercept <- coef(logistic_model)["(Intercept)"]

age <- 50
```

```r
survival_time <- 5

log_odds <- coef_intercept + coef_age * age

probability <- exp(log_odds) / (1 + exp(log_odds))
probability
```

```
## (Intercept)
##   0.4647514
```

Using the above formulas, we get 0.4647514, which gives us the same calculation, suggesting that the person is predicted to fail to respond to the treatment.

6. Use the likelihood ratio test (LRT) to assess the utility of this simple logistic regression model. Set up the hypotheses, report the G-statistic, the associated p-value, and the interpretation in context of the scenario.(Hint: You will need to save another model that does not use any predictors and then compare it to the model using `Age` that you have developed. Use `anova()` with `test="LRT"`.)

H_0: The model with Age as a predictor is not significantly better than the null model.

H_1: The model with Age as a predictor is significantly better than the null model.

```r
null_model <- glm(Resp ~ 1, data = leukemia_df, family = binomial)

logistic_model_age <- glm(Resp ~ Age, data = leukemia_df, family = binomial)

lrt_age <- anova(null_model, logistic_model_age, test = "LRT")

G_statistic_age <- lrt_age$Deviance[2]
G_statistic_age
```

```
## [1] 6.520702
```

```r
p_value_age <- lrt_age$"Pr(>Chi)"[2]
p_value_age
```

```
## [1] 0.01066259
```

Given the above information, the G-statistic is 6.5207018 and the p-value is 0.0106626. Since the p-value is less than the significance level, we can reject the null hypothesis. Therefore, we have evidence to suggest that the simple logistic regression model with Age as a predictor is significantly better than the null model, indicating that age has a significant effect on the probability of responding to treatment for Leukemia.

Model 2: Logistic Regression Using `Time`

Repeat exercises 1-6 from the previous section with `Resp` as the response variable and `Time` as the single predictor variable.

1.

$logit(p) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 * Time$

2.

```
logistic_model_time <- glm(Resp ~ Time, data = leukemia_df, family = binomial(link = "logit"))

summary(logistic_model_time)
```

```
##
## Call:
## glm(formula = Resp ~ Time, family = binomial(link = "logit"),
##     data = leukemia_df)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -4.6663     1.4137  -3.301 0.000965 ***
## Time          0.5393     0.1648   3.272 0.001066 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.524  on 50  degrees of freedom
## Residual deviance: 19.346  on 49  degrees of freedom
## AIC: 23.346
##
## Number of Fisher Scoring iterations: 7
```

Given the above information, there is an odds ratio of approximately 1.714, which means that for every one-unit increase in Time, the odds of responding to treatment decrease by approximately 71.4%.

Given the above information, Time has a statistically significant effect on the probability of responding to treatment for Leukemia.

3.

```
coef_time <- coef(logistic_model_time)["Time"]
std_err_time <- summary(logistic_model_time)$coefficients["Time", "Std. Error"]

conf_int_slope_time <- coef_time + c(-1, 1) * qnorm(0.975) * std_err_time
conf_int_slope
```

```
## [1] -0.085027981 -0.008492264
```

Given the above information, the confidence interval for the slope coefficient (Time) is approximately (-0.0850, -0.0085). This means that we are 95% confident that the true coefficient for Time falls within this interval.

Now we will calculate the confidence interval for the odds ratio.

```
conf_int_odds_ratio_time <- exp(conf_int_slope_time)
conf_int_odds_ratio_time
```

```
## [1] 1.241489 2.368773
```

Given the above information, the confidence interval for the odds ratio associated with a one-unit increase in Time is approximately (1.241, 2.369). More specifically, for every one-unit increase in Time, the odds of responding to treatment decrease by approximately 24.1% to 136.9%.

Now we will check whether the interval includes the value 1.

```
conf_int_odds_ratio_time[1] <= 1 & conf_int_odds_ratio_time[2] >= 1
```

```
## [1] FALSE
```

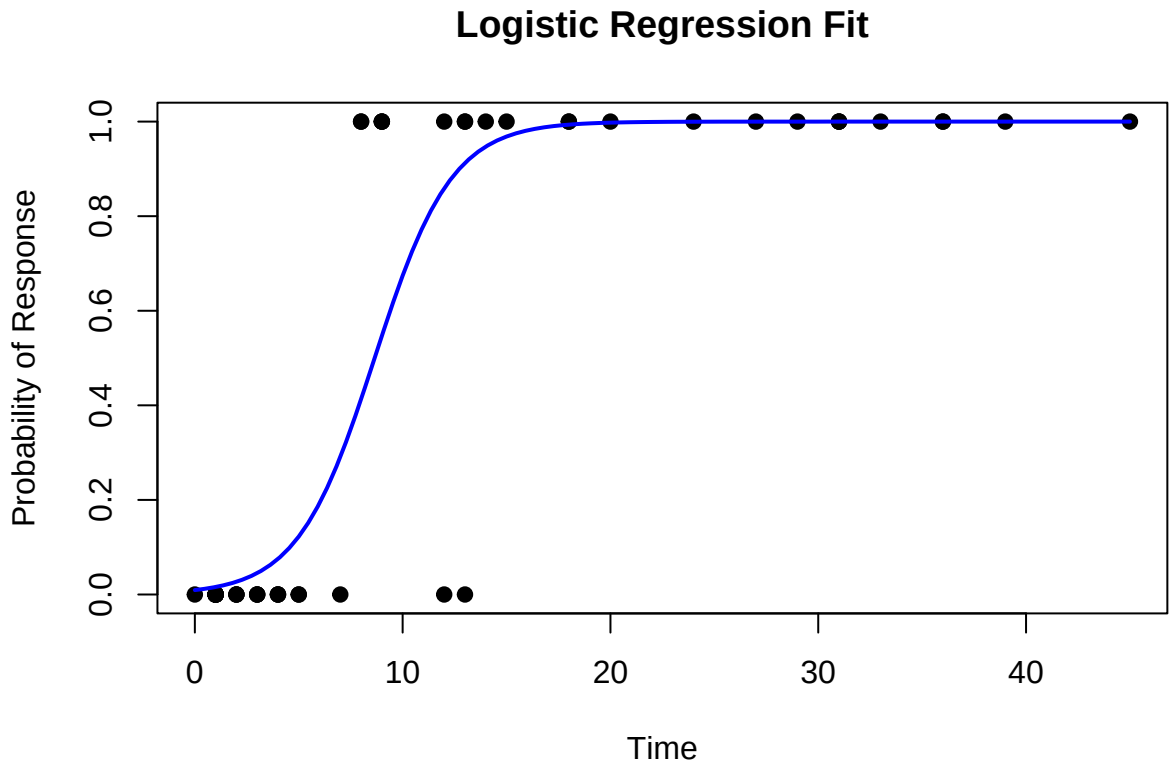
Since the confidence interval for the odds ratio does not include the value 1, it suggests that there is a statistically significant association between Time and the probability of responding to treatment for Leukemia.

4.

```
time_seq <- seq(min(leukemia_df$Time), max(leukemia_df$Time), length.out = 100)

predicted_probs <- predict(logistic_model_time, newdata = data.frame(Time = time_seq), type = "response

plot(leukemia_df$Time, leukemia_df$Resp, pch = 19, xlab = "Time", ylab = "Probability of Response", mai
lines(time_seq, predicted_probs, col = "blue", lwd = 2)
```



This plot shows a positive correlation between the Probability of Response and Time. This model is very accurate at modeling Resp. Observing the chart, we can see very few outliers. We can also see that the Probability of Response clearly goes up as Time increases.

5.

Given the following logistic regression equation: $logit(p) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 * Time$, and the coefficients obtained from the fitted model: $\beta_0 = -4.6663$ and $\beta_1 = 0.5393$, we can predict the response to treatment for a patient with a survival time from diagnosis of 5 months. Substituting the above values into the equation, we get $logit(p) = -4.6663 + 0.5393 * 5$. By simplifying this equation, we get $logit(p) = -1.9698$.

Next, we will convert the log odds to probabilities using the logistic function. Given the equation $p = \frac{1}{1+e^{-\text{logit}(p)}}$, we can plug in the above numbers into it, giving us $p = \frac{1}{1+e^{1.9698}}$, which can be simplified to $p = 0.122$. Given this information, the predicted probability of response to treatment for a patient with a survival time from diagnosis of 5 months is approximately 0.122. Based on the 0.5 threshold, since the predicted probability is less than 0.5, this calculation suggests that the person is predicted to fail to respond to the treatment.

```
coef_time <- coef(logistic_model_time)["Time"]
coef_intercept <- coef(logistic_model_time)["(Intercept)"]

age <- 50
survival_time <- 5

log_odds <- coef_intercept + coef_time * survival_time

probability <- exp(log_odds) / (1 + exp(log_odds))
probability
```

```
## (Intercept)
##   0.1224378
```

Using the above formulas, we get 0.1224378, which gives us the same calculation, suggesting that the person is predicted to fail to respond to the treatment.

6.

H_0: The model with Age as a predictor is not significantly better than the null model.

H_1: The model with Age as a predictor is significantly better than the null model.

```
null_model <- glm(Resp ~ 1, data = leukemia_df, family = binomial)

logistic_model_time <- glm(Resp ~ Time, data = leukemia_df, family = binomial)

lrt_time <- anova(null_model, logistic_model_time, test = "LRT")

G_statistic_time <- lrt_time$Deviance[2]
G_statistic_time
```

```
## [1] 51.17803
```

```
p_value_time <- lrt_time$"Pr(>Chi)"[2]
p_value_time
```

```
## [1] 8.435746e-13
```

Given the above information, the G-statistic is 51.1780306 and the p-value is $8.4357464 \times 10^{-13}$. Since the p-value is less than the significance level, we can reject the null hypothesis. Therefore, we have evidence to suggest that the simple logistic regression model with Time as a predictor is significantly better than the null model, indicating that Time has a significant effect on the Probability of Response to treatment for Leukemia.

7. Take a step back and consider the limitations of Model 2. There is at least one major issue with using `Time` to predict `Resp`. Consider what these variables mean and explain why this model would not be practically useful. (This illustrates the importance of understanding the context of your data before jumping into doing statistical modeling).

Using only the variable Time to predict Resp might not work well. While Time tells us how long a person has been dealing with Leukemia, it doesn't tell us everything we need to know. We also need to consider things such as the kind of treatment they're getting, how bad the Leukemia is when they're diagnosed, how healthy they are overall, and if there are any problems during treatment.

Even if Time seems to be connected to Resp, it's not enough to make accurate predictions on its own. Using only Time could potentially produce inaccurate or unreliable results. In order to accurately predict how likely someone is to respond to treatment, we need to look at other variables.

Part B

This part of the report can be completed as Chapter 10 is covered.

Model 3: Multiple Logistic Regression (Full Model)

1. Write down the equation for a multiple logistic regression model using all six variables (exclude `Time` and `Status`) to predict the probability of `Resp`. Use the Logit form of the logistic regression model.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times \text{Age} + \beta_2 \times \text{Smear} + \beta_3 \times \text{Infil} + \beta_4 \times \text{Index} + \beta_5 \times \text{Blasts} + \beta_6 \times \text{Temp}$$

2. Fit a multiple logistic regression model to model 3. State the predictor with the highest p-value in the summary output. (You should see that it's p-value is greater than 0.9.) Interpret what this means in context.

```
logistic_model_3 <- glm(Resp ~ Age + Smear + Infil + Index + Blasts + Temp, family = binomial(link = "l

summary(logistic_model_3)
```

```
##
## Call:
## glm(formula = Resp ~ Age + Smear + Infil + Index + Blasts + Temp,
##     family = binomial(link = "logit"), data = leukemia_df)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 108.33115   41.84379   2.589  0.00963 **
## Age          -0.06231    0.02746  -2.269  0.02327 *
## Smear        -0.00469    0.04005  -0.117  0.90677
## Infil         0.03104    0.03789   0.819  0.41264
```

```
## Index         0.37281    0.13247   2.814  0.00489 **
## Blasts        0.03267    0.04605   0.710  0.47801
## Temp         -0.11162    0.04263  -2.618  0.00884 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.524  on 50  degrees of freedom
## Residual deviance: 39.275  on 44  degrees of freedom
## AIC: 53.275
##
## Number of Fisher Scoring iterations: 6
```

Given the above information, the predictor with the highest p-value is Smear, with a value of 0.90677. Therefore, the variable Smear does not have a significant effect on the probability of response to treatment for Leukemia when considering the other variables in the model.

3. Investigate why the predictor you stated in the previous question has such a high p-value. One reason might be that it is highly correlated with another predictor or combination of predictors. What other predictor(s) is it highly correlated with? (Hint: It may be helpful to analyze a matrix scatterplot or the matrix of correlations using `cor()`.)

```
correlation_matrix <- cor(leukemia_df[, c("Smear", "Age", "Infil", "Index", "Blasts", "Temp")])
correlation_matrix
```

```
##              Smear         Age        Infil       Index     Blasts         Temp
## Smear   1.00000000 -0.20378215  0.847132591  0.10269246 0.32598642 -0.028249230
## Age    -0.20378215  1.00000000 -0.136998888 -0.12425459 0.04710552  0.084589073
## Infil   0.84713259 -0.13699889  1.000000000  0.14437713 0.34015968 -0.006709947
## Index   0.10269246 -0.12425459  0.144377132  1.00000000 0.37802894  0.070529145
## Blasts  0.32598642  0.04710552  0.340159684  0.37802894 1.00000000  0.360247536
## Temp   -0.02824923  0.08458907 -0.006709947  0.07052914 0.36024754  1.000000000
```

Given the above information, we can see that Smear has a high correlation with Infil (0.847) and a moderate negative correlation with Age (-0.204). This high correlation with Infil suggests that Smear might be redundant when Infil is already included in the model. Therefore, the high p-value for Smear could be due to its redundancy with Infil or other predictors in the model.

4. Based on values from a summary of your model 3, which of the six pretreatment variables appear to add to the predictive power of the model (i.e. which predictors appear to be significant), given that other variables are in the model?

Given the logistic model above, the predictors that appear to add to the predictive power of the model are Age, Index, and Temp. These three predictors are the only ones that contribute significantly to predicting the probability of response to treatment when other variables are in the model.

5. Specifically, interpret the relationship (if any) between `Age` and `Resp` and also between `Temp` and `Resp` indicated in the fitted model 3. Use the coefficients to help you make statements about the probability of response and the odds ratio. (i.e. State what happens to the odds of responding and the probability of responding for a 1 unit increase in each of those predictors separately.)

According to the summary output above, the coefficient estimate for Age is -0.06231, signifying that for every additional year in age, the log odds of responding to treatment diminish by approximately 0.06231 units, holding all other predictors constant. Correspondingly, the odds ratio associated with age is 0.939, indicating that with each one-year increase in age, the odds of responding to treatment decline by around 6.1%. Consequently, as a patient's age advances, the likelihood of responding to treatment tends to decrease.

According to the summary output above, the coefficient estimate for Temp is -0.11162, suggesting that for every additional tenth of a degree Fahrenheit in temperature, the log odds of responding to treatment decrease by approximately 0.11162 units, when other predictors are held constant. The odds ratio for temperature is 0.894, indicating that with each one-tenth of a degree Fahrenheit increase in temperature, the odds of responding to treatment drop by approximately 10.6%. Consequently, higher temperatures before treatment are associated with reduced probabilities of responding positively to the treatment regimen.

Model 4: Multiple Logistic Regression (Reduced Model)

6. Write down the equation for a multiple logistic regression model using just `Temp`, `Age`, `Index` to predict the probability of `Resp`. Use the Logit form of the logistic regression model.

$$\text{logit}(p) = \beta_0 + \beta_1 \times \text{Temp} + \beta_2 \times \text{Age} + \beta_3 \times \text{Index}$$

7. Use a nested likelihood ratio (drop-in-deviance) test to see if the model that excludes precisely the non-significant variables seen in model 3 is a reasonable choice for a final model (model 4). Set up the hypotheses, report the relevant test statistic, report the conclusion, and interpret your results in context.

H_0: The reduced model is not significantly worse than the full model.

H_1: The reduced model is significantly worse than the full model.

```
model3 <- glm(Resp ~ Age + Smear + Infil + Index + Blasts + Temp,
              family = binomial(link = "logit"), data = leukemia_df)

model4 <- glm(Resp ~ Age + Index + Temp,
              family = binomial(link = "logit"), data = leukemia_df)

nested_test <- anova(model4, model3, test = "LRT")
nested_test
```

```
## Analysis of Deviance Table
##
## Model 1: Resp ~ Age + Index + Temp
## Model 2: Resp ~ Age + Smear + Infil + Index + Blasts + Temp
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        47     43.265
## 2        44     39.275  3   3.9902   0.2625
```

Given the above information, the p-value is 0.2625, meaning it is greater than the significance level. Therefore, we fail to reject the null hypothesis. This suggests that the reduced model is not significantly worse than the full model. Therefore, based on the nested likelihood ratio test, the reduced model, which includes only the predictors Age, Index, and Temp, is a reasonable choice compared to the full model.

8. Using model 4, predict the probability of a successful response to treatment for an individual who is 46 years old with `Index=15` and had their highest body temperature at 99.8°F before the treatment. (Hint: take a look at the `Temp` column of the dataset carefully; you cannot just plug in `Temp=99.8` into your model!)

```
model4 <- glm(Resp ~ Age + Index + Temp, family = binomial(link = "logit"), data = leukemia_df)
summary(model4)
```

```
##
## Call:
## glm(formula = Resp ~ Age + Index + Temp, family = binomial(link = "logit"),
##     data = leukemia_df)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 87.38804   35.45816   2.465  0.01372 *
## Age         -0.05850    0.02558  -2.287  0.02218 *
## Index        0.38493    0.12152   3.168  0.00154 **
## Temp        -0.08897    0.03607  -2.467  0.01363 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.524  on 50  degrees of freedom
## Residual deviance: 43.265  on 47  degrees of freedom
## AIC: 51.265
##
## Number of Fisher Scoring iterations: 6
```

Given the above information, we can plug in the intercept, coef_age, coef_index, and coef_temp into our model.

```
intercept <- 87.38804
coef_age <- -0.05850
coef_index <- 0.38493
coef_temp <- -0.08897

age <- 46
index <- 15
temp <- 998

log_odds <- intercept + coef_age * age + coef_index * index + coef_temp * temp

probability <- exp(log_odds) / (1 + exp(log_odds))
probability
```

```
## [1] 0.8427628
```

According to the output provided, the calculated probability of success is approximately 0.843 for an individual who is 46 years old with an Index value of 15 and had their highest body temperature at 99.8°F before the treatment. In other words, the model predicts with somewhat confidence that the individual will have a successful response to treatment, as the predicted probability is relatively close to 1.

9. Comment on how your model may be used in the context of treating leukemia. What shortcomings might your modeling technique have statistically and practically? You need to identify at least 2 meaningful limitations.

The logistic regression model could be used to predict the probability of treatment response, offering valuable insights. However, it has a couple of limitations.

The logistic regression model assumes that there's a straight-line relationship between the things we're measuring and how likely they are to respond well to treatment. However, real-life medical data is often more complicated than that. The model might miss important details about how different patient factors, types of Leukemia, and treatments interact, which could make its predictions less accurate.

Logistic regression is good at sorting things into two groups, such as "responds well to treatment" or "doesn't respond well." However, when it comes to something complex such as Leukemia, where there are lots of different factors at play, it might not be able to make accurate predictions. The model only looks at certain factors that we think are important, but it might miss other important things such as genetics or how strong someone's immune system is. This can make its predictions less reliable and not cover all the different situations that patients might face.

Part C

Reflection on Facebook Experiment on Users using Statistics

Now that you have seen and constructed a variety of regression models, you will read about an application of regression to a real research experiment conducted by Facebook in 2012. The research was published in 2014 in PNAS (Proceedings of the National Academy of Sciences of the United States of America) which is considered a prestigious and influential scientific journal.

Read the research publication article at https://www.pnas.org/doi/full/10.1073/pnas.1320040111. It is titled, "Experimental evidence of massive-scale emotional contagion through social networks." Respond to the following prompts with at least a one page reflection total. This reflection should prepare you for a class discussion on this article.

1. Data and Statistics: How did the researchers obtain the data they worked with? What sorts of statistical methods and techniques did they use? Research into what the methods are and describe them more in detail. Compare and contrast their statistical methods what you have learned in class. What was the question they set out to address and what did they conclude?

The researchers randomly selected 689,003 Facebook users. The only caveat was that they needed to be English speakers, or at least use the English version of the app.

The researchers used many statistical methods in this experiment. First, they used descriptive statistics to summarize the characteristics of the data set, including the number of participants, posts analyzed, and word counts. Descriptive statistics are "a set of brief descriptive coefficients that summarize a given data set representative of an entire or sample population." They also weighted linear regression to examine the effects of the experimental conditions on participants' emotional expression. Linear regression is "a statistical model which estimates the linear relationship between a scalar response and one or more explanatory variables." They also used Poisson regression to analyze the overall posting rate, with the percent of posts excluded serving as a regression weight. Poisson regression is "a generalized linear model form of regression analysis used to model count data and contingency tables." They also used interaction effects to assess the strength of the experimental manipulations. Interaction effects occur "when the relationship between at least two variables is modified by at least one other variable." They also used post hoc tests to compare effect sizes and assess the significance of observed results. Post hoc tests are "a statistical analysis that is performed after a study has concluded and the data has been collected." Out of the above methods, the only one that

we have learned in class so far is linear regression. However, I had heard about post hoc tests and Poisson regression before reading this article.

Within this experiment, the researchers set out to address the question of whether emotional states can be transferred to others through "emotional contagion" in the context of online social networks, specifically focusing on Facebook. In the end, they were able to conclude that emotional states can be transferred to others through "emotional contagion," as exposure to emotional content in the News Feed influenced the participants' emotional expression. They also came to realize that "emotional contagion" occurs even in the absence of direct interaction between individuals, showing the impact of online social networks on one's emotions.

2. Ethics: Why was this experiment so controversial? What is your opinion on the methods of this research? Discuss a biblical basis for your position on conducting this experiment on user news feeds.

There are a couple of reasons as to why this experiment was so controversial. First, the subjects of the experiment did not know that they were being tested. They did not give explicit consent to being tested. Second, the potential harms of emotional manipulation. People thought that emotional tests such as these could have unintended consequences on the subjects.

Although I am aware that all users are required to agree with the terms and conditions when signing up for Facebook, I do not think that a test such as this is permissible under these guidelines. I do not think that people should be tested without their knowing. Although I do not necessarily believe that this test was harmful to its subjects, I firmly believe that people should be informed that they are being tested and have the option to not be. A verse that comes to mind when discussing such matters is Philippians 2:4, which states, "Let each of you look not only to his own interests, but also to the interests of others." I believe that this verse accurately depicts what it means to be a Christian, but also fits the idea of being tested without consent. I do not think that we, as Christians, should be only considering our interests when making a decision. We should consider the interests of others as well and think about how our actions may affect them, either positively or negatively.

3. Critique: How would you revise this experiment or change the research question so the methods of obtaining data, performing statistics on it, and publishing the results become less concerning?

There are a few different ways to make this experiment more ethical. First, I would ask all subjects to provide written consent. We live in a day and age where consent is everything, so asking subjects to formally agree to it would go a long way. This does not mean creating a new terms and conditions page that will be ignored, but rather creating a page that stands out and has to be read before continuing, making it very clear what they are asking. Second, I would ensure transparency throughout the whole process. This means providing clear explanations of how their data will be used and assuring them of their privacy and confidentiality. Being open about how data is being used should go a long way with the subjects and help them feel protected. Third, I would make sure that data from the experiment is completely protected. This means adhering to data protection regulations and ensuring the best practice is being done in handling sensitive data. Ensuring data security is the most important task here. Although not everyone understands just how serious a security attack is, it is still important to protect them and make sure that nothing will happen. These three methods of critiquing the experiment would help not only the subjects feel safe, but also the researchers as it would ensure that there would be no legalities down the road.

To give you more context, here is a direct quote from an article written by Gregory S. McNeal at Forbes titled "Facebook Manipulated User News Feeds To Create Emotional Responses":

> Facebook conducted a massive psychological experiment on 689,003 users, manipulating their news feeds to assess the effects on their emotions. The details of the experiment were published in an article entitled "Experimental Evidence Of Massive-Scale Emotional Contagion Through Social Networks" published in the journal Proceedings of the National Academy of Sciences of the United States of America.

The short version is, Facebook has the ability to make you feel good or bad, just by tweaking what shows up in your news feed.

The experiment tested whether emotional contagion occurs between individuals on Facebook, a question the authors (a Facebook scientist and two academics) tested by using an automated system to reduce the amount of emotional content in Facebook news feeds. The authors found that when they manipulated user timelines to reduce positive expressions displayed by others "people produced fewer positive posts and more negative posts; when negative expressions were reduced, the opposite pattern occurred."

The results suggest that "emotions expressed by others on Facebook influence our own emotions, constituting experimental evidence for massive-scale contagion via social networks." For a long time research on emotional contagion was premised on the need for in-person and nonverbal cues, this experiment suggests "in-person interaction and nonverbal cues are not strictly necessary for emotional contagion, and that the observation of others' positive experiences constitutes a positive experience for people."