# Report 2: Logistic Regression for Modeling Leukemia Treatment

Adapted from Stat2 Textbook and Resources

March 13, 2025

## 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 helps to fight back at the leukemia.

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

# Getting Started

The data was saved in the csv file provided. Once you download it, save it in the same location on your computer as your rmd file. 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)
```

# 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.

$$\log\left(\frac{1-P(\text{Resp}=1)}{P(\text{Resp}=1)}\right) = \beta_0 + \beta_1 \cdot \text{Age}$$

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"`.)

```
# Fit a logistic regression model
age_model <- glm(Resp ~ Age, data = leukemia_df, family="binomial")
summary(age_model)
```

```
##
## Call:
## glm(formula = Resp ~ Age, family = "binomial", 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
```

```r
# coefficients
beta_0 <- coef(age_model)[1]
beta_1 <- coef(age_model)[2]

# odds ratio
odds_ratio <- exp(beta_1)
odds_ratio
```

```
##       Age
## 0.9543163
```

Looking at the summary of the age model we can see that the p-value for age is 0.0166 which is less than 0.05, so we reject the null hypothesis that age has no effect of response. Age is a statistically significant predictor of response.

The odds ratio tells us that for each additional year of age multiplies the odds of response by 0.9543. This is less than 1 so the odds or responding to treatment decrease with age. We can understand this as older patients are less likely to respond to treatment compared to those younger.

If we don't use family = "binomial", it will be assumes that the model is a linear regression model instead of a logistic regression model.

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
# compute a confidence interval for slope
confint.default(age_model)
```

```
##                   2.5 %       97.5 %
## (Intercept)  0.22607822  4.167477056
## Age         -0.08502798 -0.008492264
```

```r
# compute a confidence interval for odds ratio
exp(confint.default(age_model))
```

```
##                 2.5 %     97.5 %
## (Intercept) 1.2536737 64.5523845
## Age         0.9184866  0.9915437
```

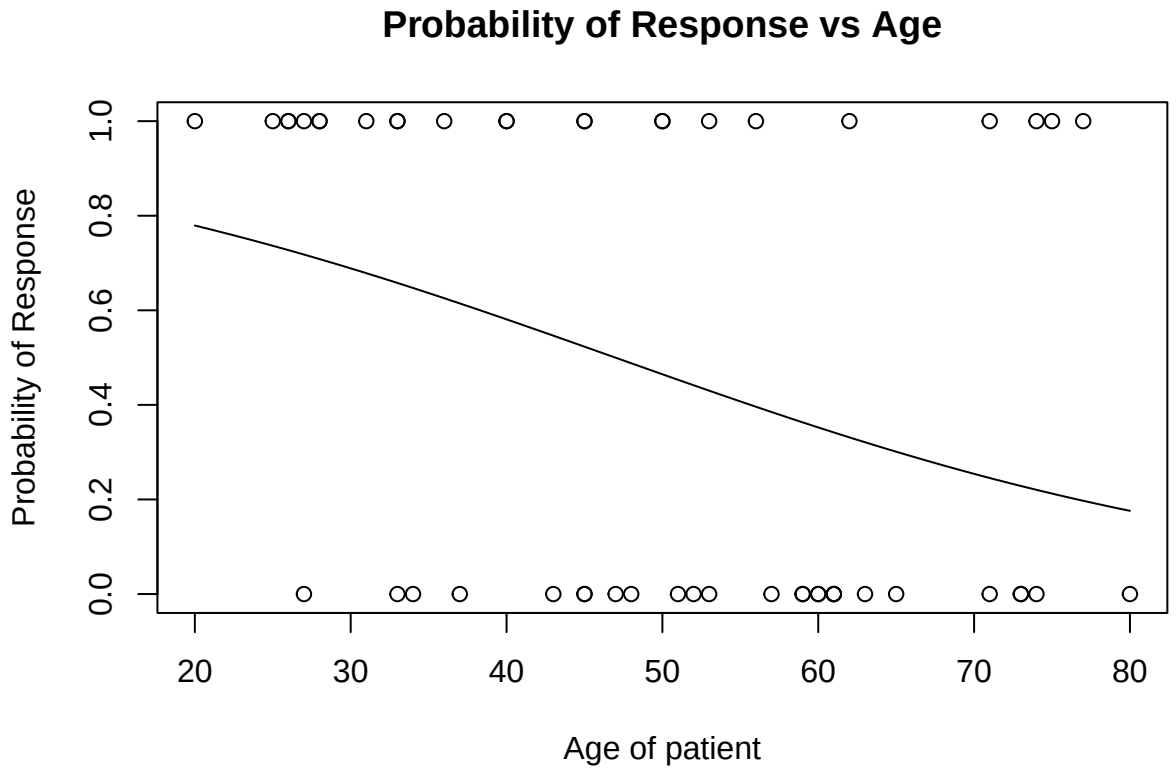
The confidence interval for slope (the first part) is (-0.0850, -0.0085). Zero is not in this interval meaning Age is statistically significant of Resp. Since both numbers are negative the log-odds of Resp will decrease as Age increases.

The confidence interval for the odds ratio is (0.918, 0.992). We can see that 1 is not in this interval meaning Age is significant with Resp. As Age has a one unit increase the odds of Resp decrease by a factor between 0.918 and 0.992.

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
# plot the logistic fit
plot(Resp ~ Age, data=leukemia_df,
     xlab="Age of patient", ylab="Probability of Response",
     main = "Probability of Response vs Age")

# Draw the curve onto the plot
curve(predict(age_model, data.frame(Age = x), type="response"), add=TRUE)
```




Looking at this plot, we can see the logistic regression curve that represents the predicted probability of response across the ages. This curve is showing a decrease meaning that the probability of response being 1 decreases as age increases. The curve is showing a negative relationship between age and response as age gets older, but the curve may not perfectly predict response probability on age as it seems like a gradual decrease. The model overall shows the relationship but other predictors could help make it better.

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"`.)

$$\log\left(\frac{1-P(\text{Resp}=1)}{P(\text{Resp}=1)}\right) = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{SurvivalTime}$$

```r
beta_2 <- 0.5

age <- 50
```

```r
survivaltime <- 5

# log(odds)
log_odds <- beta_0 + (beta_1 * age) + (beta_2 * survivaltime)
log_odds
```

```
## (Intercept)
##    2.358772
```

```r
# log(odds) to probability
prob <- 1/(1 + exp(-log_odds))
prob
```

```
## (Intercept)
##   0.9136289
```

The log(odds) is 2.358772 and then the log(odds) to probability is 0.9136289. We can see that the log-odds is bigger than 0 which means a high probability of response which we see with 0.9136289. Since 0.9136289 is greater than 0.5, the person is likely to respond to treatment.

```r
# log(odds) without survivaltime
log_odds_model <- predict(age_model, data.frame(Age=50))

#adjust by adding survivaltime
log_odds_adjusted <- log_odds_model + (beta_2 * survivaltime)
log_odds_adjusted
```

```
##        1
## 2.358772
```

```r
# probability
prob_adjusted <- 1 / (1 + exp(-log_odds_adjusted))
prob_adjusted
```

```
##         1
## 0.9136289
```

Here we add survival time (5 months) in and have age be 50. We get a log-odds with survival time of 2.358772. The predicted probability of response is 91.36% and since 0.9136 is greater than 0.5 we predict that this person will 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"`.)

```r
null_model <- glm(Resp ~ 1, data=leukemia_df, family = "binomial")
anova(null_model, age_model, test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Resp ~ 1
## Model 2: Resp ~ Age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        50     70.524
## 2        49     64.004  1   6.5207  0.01066 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, we are doing the likelihood ratio test (LRT) which compares two models (null model (no predictors) and age model). We use this to see if having age in the model improves the model compared to having no predictors at all.

We can see that the G-statistic (deviance) is 6.5207 which is the difference in deviance between the models. The p-value is 0.01066 which is less than 0.05, so we reject the null hypothesis. Thus, having age in the model significantly improves the model and is a useful predictor of response.

## 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.

$$\log\left(\frac{1-P(\text{Resp}=1)}{P(\text{Resp}=1)}\right) = \beta_0 + \beta_1 \cdot \text{Time}$$

```
# logistic regression model
time_model <- glm(Resp ~ Time, data=leukemia_df, family="binomial")
summary(time_model)
```

```
##
## Call:
## glm(formula = Resp ~ Time, family = "binomial", 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
```

Here we can see that the p-value is very small showing that time is a strong predictor of response. We can also see a big difference in the deviance and that since the deviance is lower with time, this shows a significant amount of variation. We also have a very small AIC.

```r
# coefficients
time_beta_0 <- coef(time_model)[1]
time_beta_1 <- coef(time_model)[2]

# get odds ratio
time_odds_ratio <- exp(time_beta_1)
time_odds_ratio
```

```
##     Time
## 1.714877
```

Here we get the odds ratio for time and it tells is that for each additional month the odds of responding to treatment increase by 71%. The odds ratio is greater than 1 so it has a positve effect on response probability.

```r
# confidence interval for slope
confint.default(time_model)
```

```
##                   2.5 %     97.5 %
## (Intercept) -7.4371028 -1.8954009
## Time         0.2163111  0.8623719
```

```r
# CI for odds ratio
exp(confint.default(time_model))
```

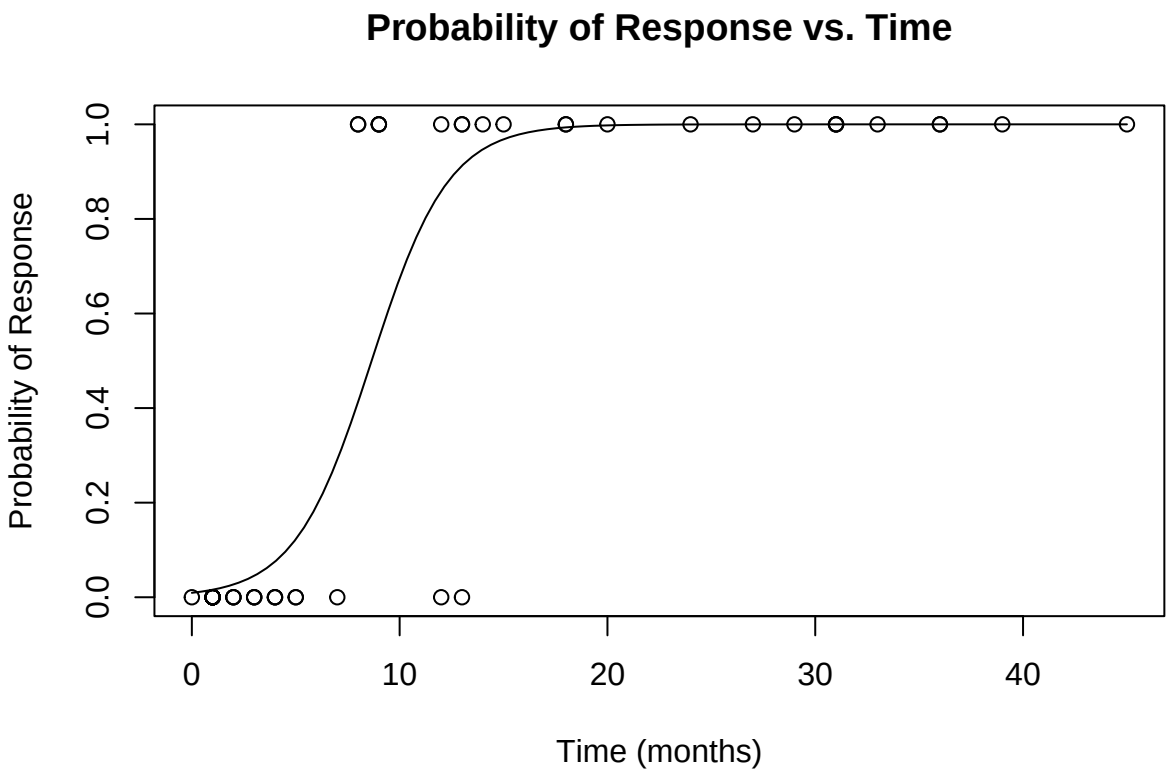
```
##                     2.5 %    97.5 %
## (Intercept) 0.0005889892 0.1502581
## Time        1.2414885276 2.3687726
```

The CI for slope is (0.2163111, 0.8623719). This interval does not include 0, showing time is statistically significant. The CI for odds ratio is (1.2414885276, 2.3687726). This shows that the odds of responding to treatment increase by a factor between 1.2414885276 and 2.3687726 for each additional month.

Looking at the intercept for the CI for slope we see that they are negative numbers meaning the response probability is very low showing most patients do not respond.

```r
# plot the logistic fit
plot(Resp ~ Time, data=leukemia_df,
     xlab="Time (months)", ylab="Probability of Response",
     main="Probability of Response vs. Time")

# Draw logistic curve onto plot
curve(predict(time_model, data.frame(Time = x), type="response"), add=TRUE)
```




This chart shows us that over time there is a probability of response. We see a steep increase at the 5 to 10 month range and that after 15 months the probability is 1 showing a response. This chart shows how longer treatment time is associated with a higher probability of response.

```
predicted_prob_time <- predict(time_model, data.frame(Time = 5), type="response")
predicted_prob_time
```

```
##         1
## 0.1224378
```

The predicted probability of response at 5 months is 0.1224378 (12.24%) This is pretty low meaning that a longer treatment duration is needed for a higher response.

```
# LRT
anova(null_model, time_model)
```

```
## Analysis of Deviance Table
##
## Model 1: Resp ~ 1
## Model 2: Resp ~ Time
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        50     70.524
## 2        49     19.346  1   51.178 8.436e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we are running the likelihood ratio test and using the anova model to compare the null model with the time model. We can see that time significantly improves the prediction of response. We see this with the large decrease in deviance from 70.524 to 19.346 and the very small p-value. Thus, showing that time is a significant predictor of response. However, looking to the next question we can see why using time can cause an issue.

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).

Looking at the summary of the time model we can see that time as a predictor of response is statistically significant with a p-value of 0.001066. However, there is an issue with using time as a predictor here. Time doesn't have a meaningful relationship to the response as it is elapsed time and doesn't have a direct relationship to the response of treatment. So, since time seems significant it is an unreliable predictor. This examples shows why it is impornatnt to understand the data.

# 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.

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

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.

```
# Fit the multiple logistic regression model
model_3 <- glm(Resp ~ Age + Smear + Infil + Index + Blasts + Temp,
               data = leukemia_df, family = "binomial")

# View a summary of the model
summary(model_3)
```

```
##
## Call:
## glm(formula = Resp ~ Age + Smear + Infil + Index + Blasts + Temp,
##     family = "binomial", 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
```

Looking at the summary of model_3 we can see that Smear has the highest p-value of 0.90677. This means that this predictor is not statistically significant to predicting the response. Since, Smear has the high p-value we would fail to reject the null hypothesis in that Smear has no significant effect on the response.

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()`.)

```r
# Correlation matrix for predictors
cor_matrix <- cor(leukemia_df[, c("Age", "Smear", "Infil", "Index", "Blasts", "Temp")])
print(cor_matrix)
```

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

Here we are using the correlation matrix to see the relationships between the different predictors in the multiple logistic model. Looking at the matrix Smear has a strong correlation with Infil meaning they are highly correlated to each other and could explain why Smear has a high p-value. Having both of these in the same model as predictors could lead to multicollinearity. It might be beneficial to remove Smear from the model to reduce this and see if the model improves.

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?

Looking at the summary of model_3, the variables that seemed most significant were Age (p-value: 0.02327), Index(p-value: 0.00489), and Temp(p-value: 0.00884). These variables seemt to add the most to the power of the predictive model based on their significance.

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.)

From the summary of model_3 we can see that the coefficient for Age is -0.06231. So, for each 1 year increase the log-odds of a response decrease by 0.06231. Below we get the odds ratio for age:

```
# coefficients
coef_model_3 <- summary(model_3)$coefficients

# coefficient for Age
coef_age <- coef_model_3["Age", "Estimate"]

# odds ratio
odds_ratio_age <- exp(coef_age)
odds_ratio_age
```

```
## [1] 0.9395896
```

```
1-odds_ratio_age
```

```
## [1] 0.06041043
```

Based on the odds ratio for age we can see that for each 1 year increase in Age, the odds or responding are 0.9395896 times the odds before the increase in Age. The odds of responding decreases by 6.1% as Age increases.

Looking back at the summary for model_3 but for Temp this time we see the coefficient is -0.11162. This means that for each 1 increase in Temp the log-odds of a response decrease by 0.11162. The odds ratio is below:

```
# coefficient for Temp
coef_temp <- coef_model_3["Temp", "Estimate"]

# odds ratio
odds_ratio_temp <- exp(coef_temp)
odds_ratio_temp
```

```
## [1] 0.8943819
```

```
1-odds_ratio_temp
```

```
## [1] 0.1056181
```

Based on the odds ratio for Temp we can see that for each 1 increase in Temp, the odds or responding are 0.8943819 times the odds before the increase in Temp. The odds of responding decreases by 10.6% as Temp increases.

## 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.

$$\log\left(\frac{P(\text{Resp}=1)}{1-P(\text{Resp}=1)}\right) = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{Index}) + \beta_3(\text{Temp})$$

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.

```
model_4 <- glm(Resp ~ Age + Index + Temp,
               data = leukemia_df, family = "binomial")

# Likelihood Ratio Test
anova(model_4, model_3, test = "LRT")
```

```
## 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
```

Looking at the likelihood ratio test we can see that the p-value is 0.2625. This tells us that since we have such a high p-value there is not significant enough evidence to say that model_3 is better than model_4. Thus meaning we fail to reject the null hypothesis.

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!)

```
# Predict probability for given values
predict(model_4, newdata = data.frame(Age = 46, Index = 15, Temp = 998),
        type = "response")
```

```
##         1
## 0.8423224
```

The predicted probability (0.8423224) is greater than 0.5, so we can predict that a patient with these conditions will likely respond to treatment.

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.

Since, model_4 looks at how a patient responds to treatment based on Age, Index, and Temperature this can help in seeing who is most likely to respond to treatment, help with decision making of treatment, and seeing what factors impact response. Uisng the example above, we can see a 84.2% predicted probability of that patient having a response to treatment. Having insights just like this can help in making decisions based on factors that are impact for the patient.

In terms of model_4, we are only looking at Age, Index, and Temp, but there are other factors that could be seen as important to the response. The model may look different and have a different response if we were using other predictors. This is true in the case if there is one predictor that is highly important as the model may not fully show the true predictors for treatment.

There is also the practical side of the predictors in that two could be correlated with a better response together, but that does not mean they are causing the response. When we use the logistic regression model we are showing relationships between variables not a direct cause or effect of relationship.

Based on these findings from this model, we should take into consideration all the findings and limitations. As model_4 should not be used alone to determining things, but more as an supporting piece to go with more information. In the future, we could look at other predictors that could help with making the model better with treatment predicitons.

# 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 c. 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?
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.
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?

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."

Reflection on Facebook Experiment on Users using Statistics

This Facebook Experiment got their data from 689,003 (unknowingly) Facebook users in January of 2012. Facebook wanted to see if emotional contagion could occur through social networks. They did this by influencing the users content that they viewed and then looked into what the emotional tone was of the user's post's. There were two experiments that took place. There was one that had positive emotional content from their friends reduced and the other had negative emotional content from their friends reduced. The posts were determined positive or negative based on if the had at least one positive or negative word. The researchers were working with real time data and they created controlled variables to what the users were exposed to. They then used computer based analysis tools to see the positive and negative effects on users posts. The statistical techniques that the researchers used were significance testing, regression analysis and comparative analysis. The significance testing was used determine whether a drop in positive word usage when positive content was reduced was significant rather than randomness. The regression model was fitting the models to predict the likelihood of these emotions in the experimental condition and then as well in the controlled. The comparative analysis to shows before and afters of emotions being shown. Looking at there ways of statistical methods I think they are very similar in terms of fitting models to predict what predictors will work best. This relates to predicitng likelihood especially with the LRT (anova). The Facebook experiment works with a controlled manipulation which isn't something we directly worked with but doing this showed complexity of the data. Overall the point of this research was looking into whether emotions from others played a role into our own emotions in an online setting. Taking what the experiment showed us we can say that yes, emotional contagion is real and does happen in the social network setting. Users who were exposed to fewer positive posts, their posts followed in line to that and same for negative.

It can be said that this experiment can be seen as controversial because of the lack of consent that was informed to the users. These users were not completely aware of this experiment that was being taken and that their emotional states were apart of it, leading to users feeling manipulated and having a lack of privacy. If the users were informed they would've had the opportunity to deny or approve their participation in the experiment. Since, this study was intentionally messing with users emotions could lead to some psychological effects that would've been prevented if the users had knowledge to the experiment. As Christians we are called to love your neighbor as yourself (Mark 12:31), so having this in mind is important when running experiments like this. Letting all parties and knowledge be known before starting is how we can love one another and avoiding harmful situations. Manipulation and having a lack of privacy go completely against this, so its important to remember to love your neighbor as yourself.

The revisions I would make to the experiment would be by starting with an option for people to decide if they want to be in the experiment or not. This allows for people to volunteer knowingly with clear communication from the beginning. I think this would allow more people to willingly want to do the experiment if they knew more about what was happening. I also think that rephrasing the question into a way that is more positive could also be beneficial. We could focus an experiment more onto if having more positive content boots the users mood, posts, and more. Overall, keeping the users in mind is the best way to approach an experiment like this, so changing a few things could help this research.