498818_SDS496_HW5

2024-10-16

498818 SDS 496 HW 5

Problem 1:

```
library(data.table)
data <- fread("/Users/aidanashrafi/Downloads/pediatric.txt")
head(data)
```

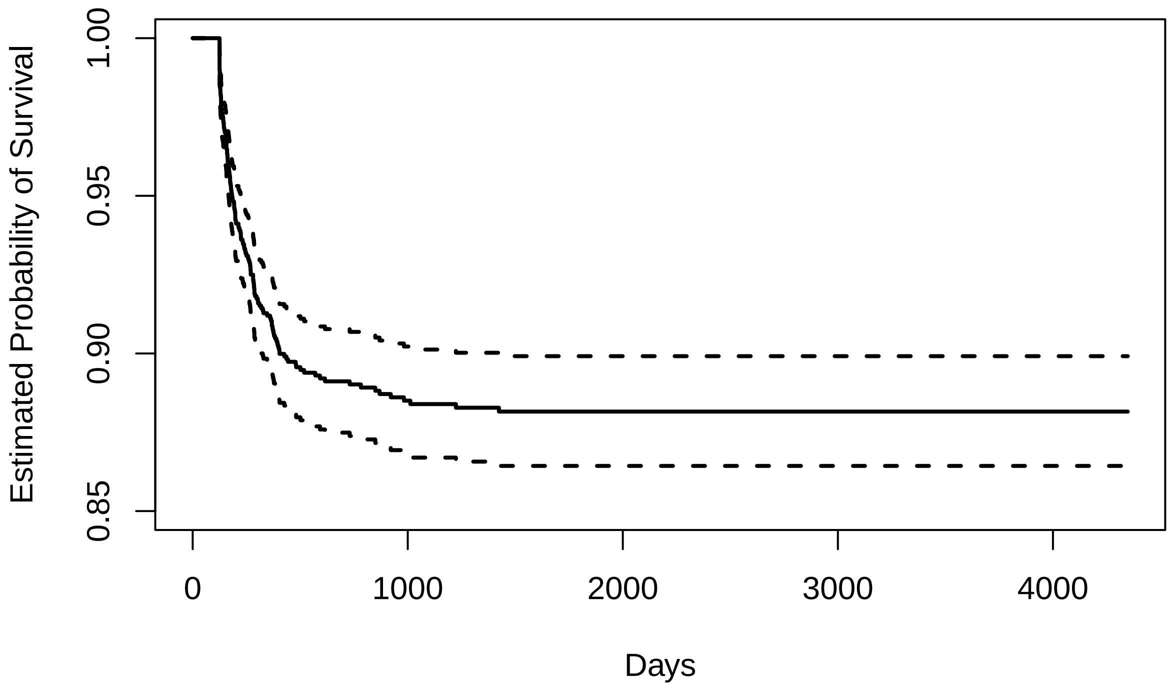
```
##      sex  race   age entry   far  time status
##    <int> <int> <num> <int> <int> <int>  <int>
## 1:     0     0  2.50   710   108   325      0
## 2:     1     0 10.00  1866    38  1451      0
## 3:     1     1 18.17  2531   100   221      0
## 4:     1     0  3.92  2210   100  2158      0
## 5:     0     0 11.83   875    78   760      0
## 6:     1     0 11.17  1419     0   168      0
```

```
names(data)
```

```
## [1] "sex"    "race"   "age"    "entry"  "far"    "time"   "status"
```

i. Plot the Kaplan - Meier estimator of the overall survival curve with 95% confidence bands. What's a rough estimate for the probability that a child survives past the end of the study? You should ideally adjust the limits of the y-axis to make the plot more readable.

```
library(survival)
surv = Surv(data$time, data$status)
fit = survfit(surv~1)
plot(fit, xlab = "Days", ylab = "Estimated Probability of Survival", lwd = 2, ylim= c(0.85,1))
```



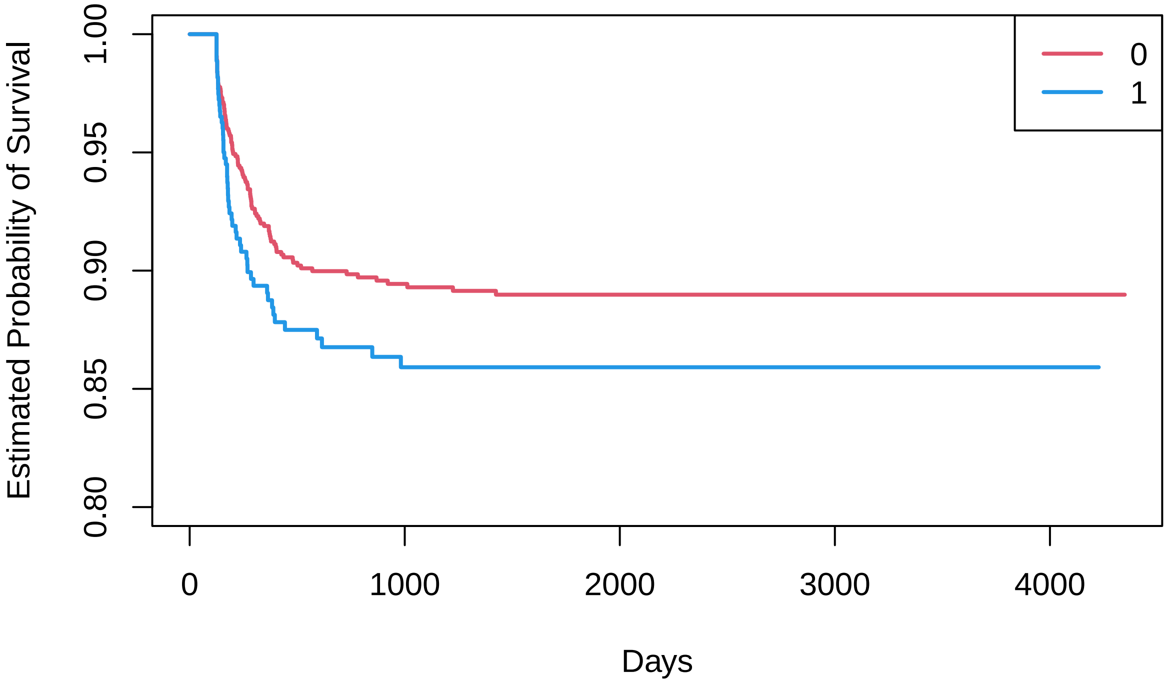
```r
q <- fit$surv
tail(q, n=1)
```

```
## [1] 0.8815554
```

The estimate for the probability that a child survives past the end of the study is **0.8815554.**

ii. Plot the Kaplan - Meier estimators of the stratified survival curves by race. Are there visible differences between the two estimated survival curves? Which group of children appears to have overall higher chances of survival?

```r
new_fit <- survfit(surv ~ race, data)
plot(new_fit, xlab = "Days", col=c(2,4), ylim = c(0.80,1), ylab = "Estimated Probability of Survival", l
legend('topright', legend = unique(data$race), col=c(2,4), lty = 1, lwd = 2)
```



Yes, there are visible differences between the two survival curves. **Group 0 (White Children)** tends to have a higher survival rate than **Group 1 (Other Children)** across all time frames.

iii. Perform a log-rank test to assess whether the survival curves for the two groups of children are significantly different from each other.

```
survdiff(surv~race, data)
```

```
## Call:
## survdiff(formula = surv ~ race, data = data)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## race=0 1179      110    119.7     0.785      3.01
## race=1  441       52     42.3     2.222      3.01
##
##  Chisq= 3  on 1 degrees of freedom, p= 0.08
```

The conclusion of the test at alpha significance level = 0.05 is to fail to reject the null hypothesis because **p = 0.08**, 0.08 > 0.05, therefore we do not have enough evidence to reject the null hypothesis that the survival curves are identical. **Essentially, we cannot conclude the survival curves for the two groups of children are significantly different from each other.**

iiii. Fit Cox's proportional hazards model with only race as a predictor.

```
cox_fit = coxph(surv ~ race, data)
summary(cox_fit)
```

```
## Call:
## coxph(formula = surv ~ race, data = data)
##
##   n= 1620, number of events= 162
##
##        coef exp(coef) se(coef)     z Pr(>|z|)
## race 0.2911    1.3379   0.1683 1.729   0.0838 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## race     1.338     0.7475    0.9619     1.861
##
## Concordance= 0.532  (se = 0.019 )
## Likelihood ratio test= 2.88  on 1 df,   p=0.09
## Wald test            = 2.99  on 1 df,   p=0.08
## Score (logrank) test = 3.01  on 1 df,   p=0.08
```

Interpretation of Cox Proportional Hazards Model Results: The Cox proportional hazards model estimates the effect of race on the hazard function (risk of death) based on the data. The key findings are as follows:

Statistical Significance: Race does not have a statistically significant effect on the hazard function at the 0.05 significance level. This conclusion is supported by the 95% confidence interval for the hazard ratio (exp(coef)) of 1.338, which includes 1, and by the p-values of the likelihood ratio test ($p = 0.09$), Wald test ($p = 0.08$), and score test ($p = 0.08$), all of which exceed 0.05. Therefore, we fail to reject the null hypothesis that race has no effect on the risk of death.

Hazard Ratio (exp(coef)): The hazard ratio for race is 1.338, meaning individuals of other races have a 33.8% higher hazard (risk) of death compared to the baseline group (white individuals). Since the confidence interval for the hazard ratio includes 1 (0.962 to 1.861), this effect is not statistically significant.

Concordance: The concordance index is 0.532, indicating that the model correctly predicts the order of death between two individuals 53.2% of the time. This is only slightly better than random guessing, suggesting limited predictive power.

Overall Model Fit: The likelihood ratio test, Wald test, and score test collectively show that race is not a significant predictor of the hazard function, with p-values greater than 0.05, reinforcing the conclusion that race does not significantly affect the risk of death in this model.

v. Predict the probability of survival past 1000 and 3000 days for a child of white race and other race respectively.

```r
# Fit the Cox proportional hazards model with only race as a predictor
cox_fit_race <- coxph(Surv(time, status) ~ race, data)

# Predict survival probability past 1000 days for a white child (race = 0)
new_individual_white <- data.frame(
  race = 0,  # White race (0 for white, 1 for other)
  status = 0,
  time = 1000
)

predicted_survival_1000_white <- predict(cox_fit_race, newdata = new_individual_white, type = "survival

# Predict survival probability past 3000 days for a child of other race (race = 1)
new_individual_other <- data.frame(
  race = 1,  # Other race (1 for non-white, 0 for white)
  status = 0,
  time = 1000
)
```

```
predicted_survival_1000_other <- predict(cox_fit_race, newdata = new_individual_other, type = "survival

# Print results
print(paste("Predicted survival probability past 1000 days for a white child:", round(predicted_surviva
```

```
## [1] "Predicted survival probability past 1000 days for a white child: 0.8938"
```

```
print(paste("Predicted survival probability past 1000 days for a child of other race:", round(predicted_
```

```
## [1] "Predicted survival probability past 1000 days for a child of other race: 0.8606"
```

```
# Fit the Cox proportional hazards model with only race as a predictor
cox_fit_race <- coxph(Surv(time, status) ~ race, data)

# Predict survival probability past 1000 days for a white child (race = 0)
new_individual_white_2 <- data.frame(
  race = 0,  # White race (0 for white, 1 for other)
  status = 0,
  time = 3000
)

predicted_survival_3000_white <- predict(cox_fit_race, newdata = new_individual_white_2, type = "surviva

# Predict survival probability past 3000 days for a child of other race (race = 1)
new_individual_other_2 <- data.frame(
  race = 1,  # Other race (1 for non-white, 0 for white)
  status = 0,
  time = 3000
)

predicted_survival_3000_other <- predict(cox_fit_race, newdata = new_individual_other_2, type = "surviva

# Print results
print(paste("Predicted survival probability past 3000 days for a white child:", round(predicted_surviva
```

```
## [1] "Predicted survival probability past 3000 days for a white child: 0.8906"
```

```
print(paste("Predicted survival probability past 3000 days for a child of other race:", round(predicted_
```

```
## [1] "Predicted survival probability past 3000 days for a child of other race: 0.8564"
```

Predictions:

Predicted survival probability past **1000 days for a white child: 0.8938**

Predicted survival probability past **1000 days for a child of other race: 0.8606**

Predicted survival probability past **3000 days for a white child: 0.8906**

Predicted survival probability past **3000 days for a child of other race: 0.8564**

vi. Fit Cox's proportional hazards model with all available predictors.

```
cox_fit_full = coxph(surv ~ . , data)
```

```
## Warning in coxph.fit(X, Y, istrat, offset, init, control, weights = weights, :
## Ran out of iterations and did not converge
## Warning in coxph.fit(X, Y, istrat, offset, init, control, weights = weights, :
## one or more coefficients may be infinite
```

```r
summary(cox_fit_full)
```

```
## Call:
## coxph(formula = surv ~ ., data = data)
##
##   n= 1620, number of events= 162
##
##              coef  exp(coef)   se(coef)      z Pr(>|z|)
## sex     6.017e-02  1.062e+00  2.083e-01  0.289    0.773
## race    4.496e-02  1.046e+00  2.301e-01  0.195    0.845
## age     5.986e-03  1.006e+00  2.643e-02  0.226    0.821
## entry   4.231e-05  1.000e+00  1.516e-04  0.279    0.780
## far    -1.886e-04  9.998e-01  2.260e-03 -0.083    0.934
## time   -5.669e-01  5.673e-01  5.738e-02 -9.880   <2e-16 ***
## status  1.876e+01  1.407e+08  6.685e+02  0.028    0.978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## sex    1.062e+00  9.416e-01    0.7060    1.5976
## race   1.046e+00  9.560e-01    0.6663    1.6421
## age    1.006e+00  9.940e-01    0.9552    1.0595
## entry  1.000e+00  1.000e+00    0.9997    1.0003
## far    9.998e-01  1.000e+00    0.9954    1.0043
## time   5.673e-01  1.763e+00    0.5069    0.6348
## status 1.407e+08  7.107e-09    0.0000       Inf
##
## Concordance= 1  (se = 0 )
## Likelihood ratio test= 2077  on 7 df,   p=<2e-16
## Wald test            = 97.79  on 7 df,   p=<2e-16
## Score (logrank) test = 2709  on 7 df,   p=<2e-16
```

Interpretation of Cox Proportional Hazards Model Results: Baseline Hazard Function:
The baseline hazard function in this model reflects the risk when all covariates are set to their reference levels. In this case, most predictors do not have a statistically significant effect on the hazard function at the 0.05 significance level, except for time, which has a strong and highly significant effect (p < 2e-16).

Percentage Change for Sex:
The estimated coefficient for sex is 0.06017, corresponding to a hazard ratio of 1.062. This indicates that being male (or the reference sex) is associated with a 6.2% increase in the hazard (risk of death), though this effect is not statistically significant (p = 0.773).

Statistically Significant Predictors:
At the 0.05 significance level, only time has a statistically significant effect on the hazard function (p < 2e-16). All other predictors, including sex, race, age, entry, far, and status, do not have statistically significant effects, as their p-values exceed 0.05.

Most Significant Predictor:
Time is the most significant predictor, with a hazard ratio of 0.567 and a p-value of less than 2e-16. This suggests that longer time is strongly associated with a lower hazard (risk of death), indicating that the hazard decreases by 43.3% (1 - 0.567) for each unit increase in time.

Concordance Index:
The concordance index for this model is 1, indicating perfect prediction performance. This is a stark

improvement over the model in part iv, which had a concordance of 0.532. However, such a perfect concordance is unusual and might suggest overfitting or data issues that need further exploration.

Statistical Significance of Predictors:
Based on the likelihood ratio test, Wald test, and score test, there is strong evidence (p < 2e-16) that at least one predictor (specifically, time) has a statistically significant effect on the hazard function.

vii. Predict the probability of survival past 2000 days of a female white child who is 15 years old, lives 50 miles away from the treatment center and entered the study on July 1st, 2002.

```r
# Fit Cox model with Surv object in the training data
cox_fit_predict <- coxph(Surv(time, status) ~ race + age + sex + far + entry, data)

# Create new individual with only the predictors used in the model
new_individual <- data.frame(
  race = 0,     # White race (0 for white, 1 for other)
  age = 15,     # Age 15
  sex = 1,      # Assuming sex (0 = female, 1 = male)
  far = 50,     # Distance from treatment center
  status = 0,
  time=2000,
  entry = 365  # Days since July 1st, 2001
)

# Predict survival probability at 2000 days for the new individual
predicted_survival <- predict(cox_fit_predict, newdata = new_individual, type = "survival", time = 2000

# Display the predicted survival probability
print(paste("Predicted probability of survival past 2000 days:", round(predicted_survival, 4)))
```

```
## [1] "Predicted probability of survival past 2000 days: 0.8818"
```

Predictions:

Predicted probability of survival **past 2000 days: 0.8818**

viii. Split the data set into a training set and a test set. Apply the lasso method with 10-fold cross-validation on the training set to select the most important predictors for the hazard function. Fit the optimal model on the test set and report the estimated coefficients.

```r
library(glmnet)
```

```
## Loading required package: Matrix

## Loaded glmnet 4.1-8
```

```r
# Define the number of observations
n <- nrow(data)

# Split data into training and testing sets
set.seed(123)  # For reproducibility
train <- sample(n, 0.8 * n)  # 80% training data
test <- setdiff(1:n, train)  # Remaining 20% as test data

# Create model matrix for glmnet
X_train <- model.matrix(Surv(time, status) ~ ., data = data)[train, ]
X_test <- model.matrix(Surv(time, status) ~ ., data = data)[test, ]
```

```
# Define the survival objects for the training and testing sets
surv_train <- Surv(data$time[train], data$status[train])
surv_test <- Surv(data$time[test], data$status[test])

# Perform cross-validation to find the best lambda (Lasso regularization)
cv.lasso <- cv.glmnet(X_train, surv_train, alpha = 1, family = "cox")

# Fit Lasso model using the optimal lambda found during cross-validation
lasso <- glmnet(X_test, surv_test, alpha = 1, lambda = cv.lasso$lambda.min, family = "cox")

# Print the Lasso model coefficients (beta)
print(lasso$beta)
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept)  .
## sex         -0.2533590111
## race         0.4417509536
## age         -0.0591605790
## entry       -0.0004481667
## far          0.0056846666
```

```
names(data)
```

```
## [1] "sex"    "race"   "age"    "entry"  "far"    "time"   "status"
```

Interpretation:

- **(Intercept):** The intercept term is not displayed as it is typically not penalized in Lasso regression.
- **Sex (female = 1, male = 0):** The coefficient for sex is -0.2534, indicating that being female is associated with a lower hazard (risk of death) compared to being male, after accounting for the other covariates.
- **Race (other = 1, white = 0):** The coefficient for race is 0.4418, suggesting that individuals of other races have a higher hazard (risk of death) compared to white individuals.
- **Age:** The coefficient for age is -0.0592, indicating that as age increases, the hazard (risk of death) decreases slightly.
- **Entry:** The coefficient for entry is -0.0004, indicating a negligible negative effect on the hazard (risk of death).
- **Far:** The coefficient for far is 0.0057, indicating a small positive association with the hazard (risk of death).

These coefficients reflect the relationships between each predictor and the hazard (risk of death) after applying Lasso, which performs variable selection and regularization to shrink less important coefficients toward zero.

Problem 2.

i. Fit a logistic regression model with only race as a predictor.

```
lreg_data = read.table("/Users/aidanashrafi/Downloads/pediatric.txt", header= TRUE)
lreg_fit = glm(status ~ race, lreg_data, family = binomial)
summary(lreg_fit)
```

```
##
## Call:
## glm(formula = status ~ race, family = binomial, data = lreg_data)
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.2740     0.1001 -22.710   <2e-16 ***
## race          0.2617     0.1784   1.467    0.142
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1053.3  on 1619  degrees of freedom
## Residual deviance: 1051.2  on 1618  degrees of freedom
## AIC: 1055.2
##
## Number of Fisher Scoring iterations: 5
```

Interpretation of Logistic Regression Model Results: Intercept Coefficient:

The estimated intercept coefficient is -2.2740. This represents the log-odds of death for the baseline group (white individuals, since race = 0) when all other predictors are held constant. The corresponding odds of death for white individuals can be calculated as exp(-2.2740) is approximately 0.103, meaning the odds of death are about 10.3% for white individuals.

Race Coefficient (other = 1, white = 0):

The estimated coefficient for race is 0.2617. This means that being of another race (as opposed to being white) is associated with an increase in the log-odds of death by 0.2617. The odds ratio is exp(0.2617) is approximately 1.299, indicating that individuals of other races have about 29.9% higher odds of death compared to white individuals.

Statistical Significance:

The p-value for the race coefficient is 0.142, which is greater than the significance level of 0.05. Therefore, we do not have enough evidence to conclude that race has a statistically significant effect on the odds of death in this model.

ii. Split the data set into a training set and a test set. Fit a logistic regression model with all available predictors on the training set.

```r
# Split data into training and testing sets
set.seed(123)  # For reproducibility
n <- nrow(lreg_data)
train <- sample(n, 0.8 * n)  # 80% training data
test <- setdiff(1:n, train)  # Remaining 20% as test data

# Fit logistic regression on the training data with all predictors
logit_model <- glm(status ~ ., data = lreg_data[train, ], family = binomial)

# Summary of the model to extract coefficients and p-values
summary(logit_model)
```

```
##
## Call:
## glm(formula = status ~ ., family = binomial, data = lreg_data[train,
##     ])
##
## Coefficients:
```

```
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.2309906  0.3506982   0.659    0.510
## sex          0.0714616  0.2162583   0.330    0.741
## race         0.0709421  0.2240788   0.317    0.752
## age         -0.0348581  0.0236394  -1.475    0.140
## entry       -0.0005726  0.0001342  -4.266 1.99e-05 ***
## far          0.0053626  0.0021099   2.542    0.011 *
## time        -0.0031022  0.0004402  -7.047 1.83e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 835.56  on 1295  degrees of freedom
## Residual deviance: 583.64  on 1289  degrees of freedom
## AIC: 597.64
##
## Number of Fisher Scoring iterations: 8
```

Interpretation of Logistic Regression Model Results:

Intercept Coefficient:
The estimated intercept coefficient is 0.2310, representing the log-odds of death when all other predictors are held constant at zero. This translates to an odds ratio of exp(0.2310) is approximately 1.260, suggesting a baseline odds of death 26% higher than the reference group (which depends on how other predictors are encoded).

Percentage Change for Sex (female = 1, male = 0):
The coefficient for sex is 0.0715, corresponding to an odds ratio of exp(0.0715) is approximately 1.074. This indicates that being female is associated with a 7.4% increase in the odds of death compared to being male, though this effect is not statistically significant (p = 0.741).

Statistically Significant Predictors:
At the 0.05 significance level, the following predictors have statistically significant effects on the odds of death: - **Entry:** The coefficient is -0.0005726 (p = 1.99e-05), suggesting that as the entry value increases, the odds of death decrease slightly. - **Far:** The coefficient is 0.0054 (p = 0.011), indicating a small positive association between the far predictor and the odds of death. - **Time:** The coefficient is -0.0031 (p = 1.83e-12), suggesting that as time increases, the odds of death decrease significantly.

Most Significant Predictor:
Time is the most significant predictor in this model (p = 1.83e-12), with a coefficient of -0.0031. This indicates that for each unit increase in time, the odds of death decrease by about 0.31% (1 - exp(-0.0031)).

iii. Predict the probability of death for a female white child who is 15 years old, lives 50 miles away from the treatment center and entered the study on July 1st, 2002.

```r
# Create new individual data frame with all variables used in the logistic regression model
new_individual <- data.frame(
  sex = 1,        # Assuming 'sex' is a factor or categorical variable
  race = 0,              # Assuming 'race' is coded numerically (0 for white, 1 for other)
  age = 15,              # Age is numeric
  far = 50,              # Distance from the treatment center
  time = 2000,
  entry = 365
)
```

```
# Predict probability of death for the new individual
predicted_prob_individual <- predict(logit_model, newdata = new_individual, type = "response")
print(paste("Predicted probability of death for the new individual:", round(predicted_prob_individual,
```

```
## [1] "Predicted probability of death for the new individual: 0.0017"
```

Predictions:

Predicted probability of death for the new individual: 0.0017

iv. What's the estimated decision boundary based on this model? Calculate the confusion matrix and the misclassification rate on the test set.

```
# Set response variable from the test data
response <- lreg_data$status[test]

# Make predictions on the test data (logistic regression model)
prediction <- predict(logit_model, lreg_data[test, ], type = "response")

# Convert predictions to binary outcomes (decision boundary = 0.5)
prediction <- as.numeric(prediction > 0.5)

# Confusion matrix
confusion_matrix <- table(response, prediction)
print("Confusion Matrix:")
```

```
## [1] "Confusion Matrix:"
```

```
print(confusion_matrix)
```

```
##         prediction
## response   0   1
##        0 288   2
##        1  32   2
```

```
# Misclassification rate
misclassification_rate <- mean(response != prediction)
print(paste("Misclassification rate:", round(misclassification_rate, 4)))
```

```
## [1] "Misclassification rate: 0.1049"
```

```

# Convert probabilities to binary outcomes (decision boundary = 0.5)
predicted_class <- ifelse(prediction > 0.5, 1, 0)

# Confusion matrix
confusion_matrix <- table(predicted_class, lreg_data$status[test])
print("Confusion Matrix:")
```

Another way to arrive at the same answer

```
## [1] "Confusion Matrix:"
```

```
print(confusion_matrix)
```

```
##
## predicted_class   0   1
##               0 288  32
##                1   2   2
```

```
# Misclassification rate
misclassification_rate <- mean(predicted_class != lreg_data$status[test])
print(paste("Misclassification rate:", round(misclassification_rate, 4)))
```

```
## [1] "Misclassification rate: 0.1049"
```

As a result of having multiple predictors, visualizing a decision boundary becomes challenging. Instead, we assess the performance of the model using a confusion matrix and the misclassification rate. These metrics provide insights into how well the model is able to distinguish between the different classes, without the need for visualizing a decision boundary.

The confusion matrix helps evaluate true positives, false positives, true negatives, and false negatives, while the misclassification rate quantifies the proportion of incorrect predictions made by the model.