Nikhil Gopal

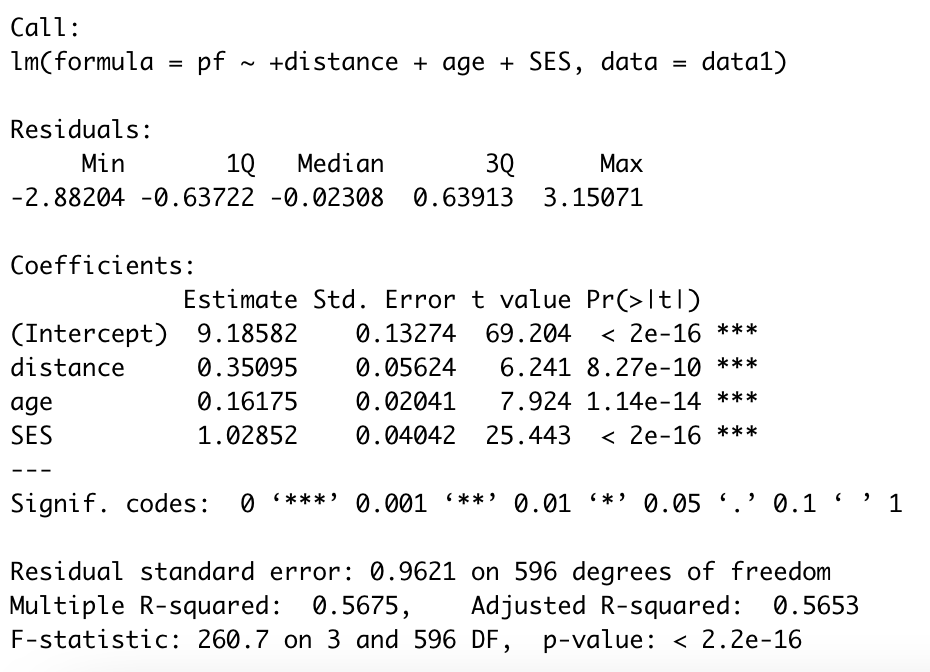
**Question A**

**Methods:**

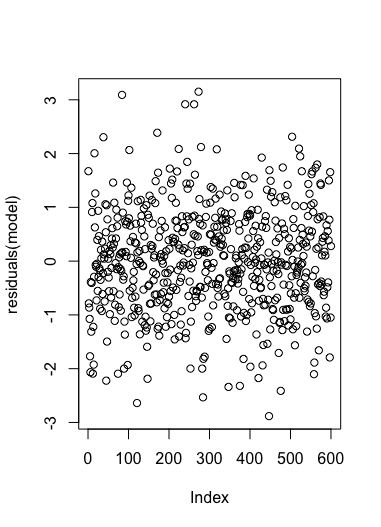
For this question, multivariate regression analysis was used. For each child, 3 measurements were given denoting pulmonary health, as well as the child’s age, socioeconomic status and latitude/longitude coordinates giving that child’s distance from the waste disposal center. Since latitude/longitude coordinates are in a coordinate plane, I used the distance formula (sqrt((x-x)^2+(y-y)^2)) to find the linear distance from the waste disposal center for each child. A linear model was most appropriate, as the distribution of the outcome variable, and most of the predictor variables were approximately normal. Additionally, since the main outcome of interest was not binary, logistic regression was not appropriate. Similarly, since the time to an event was not of interest and we aren’t using censored data, ruling out survival analysis.

I chose a multivariate analysis to minimize the effects of other confounding variables on the predictor. The model ended up having pulmonary health as the outcome variable, with age, socioeconomic status and distance being the predictors. I chose to make age a continuous variable instead of categorical, as any grouping of ages into categories would be arbitrary. I made a residual plot to check for homoscedasticity. Additionally, I made scatter plots to show that variables were independent of each other, and thus to ascertain that this data met the assumptions for linear regression.

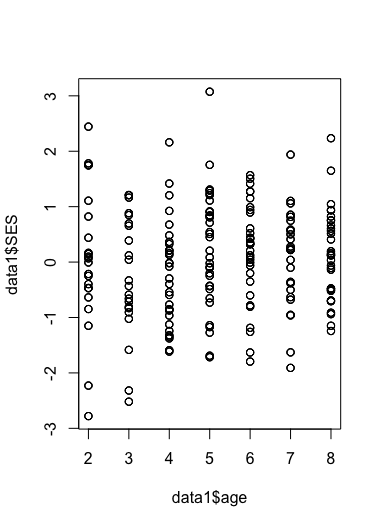
**Results:**

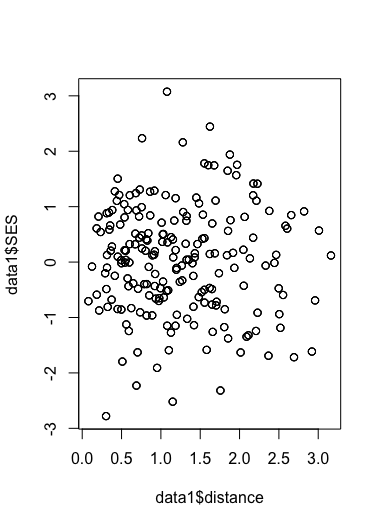


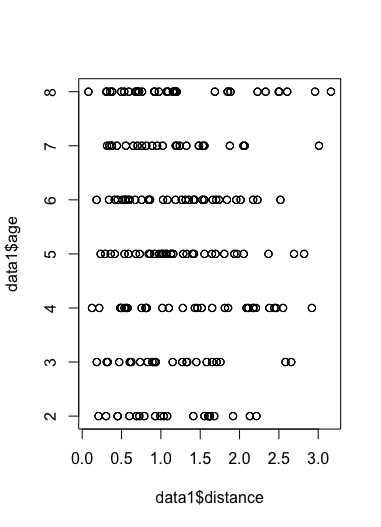
Residuals:



Tests for independence:







Above is the output of my multivariate regression analysis. Socioeconomic status appears to have the largest effect on pulmonary health, with the largest coefficient of 1.02852. Distance and age had coefficients of 0.35095 and 0.16175 respectively. The p values for the coefficients were all less than 0.001, meaning they are all statistically significant.

**Conclusions:**

Three variables are statistically significant, means that if this experiment were repeated, the probability of seeing a similar difference (or greater) would be less than 0.001%. Since the P values were low, it is unlikely that the findings are due to chance alone. There is an extremely high likelihood a relationship exists between distance from the waste disposal plant (and the other variables) and pulmonary health. Additionally, our R^2 value was 0.5653, This demonstrates that a moderate amount of the variability in our data can be explained by our model, implying a strong relationship between the predictors and the outcome.

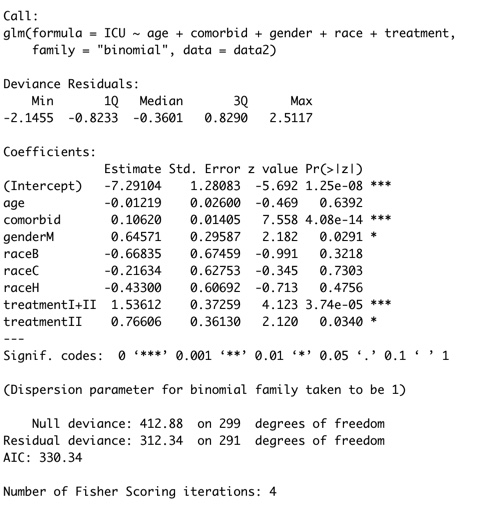
**Question 2**

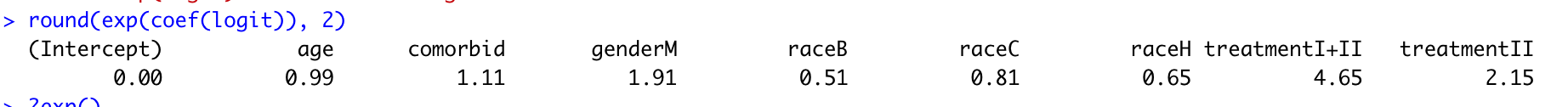
**Methods**

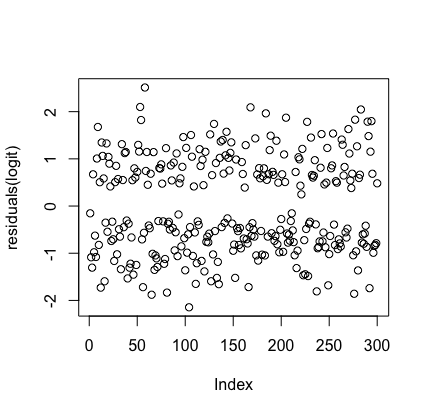
The goal of this research was to examine which treatment option (I, II or I+II) resulted in the lowest rate of ICU admissions. We were given data on patient age, whether or not they were admitted to the ICU, treatment received, gender and race. To investigate this, I decided to use a logistic regression model with ICU admission as the outcome variable, and with treatment, race and gender as predictors. Logistic regression was appropriate since the outcome variable was ICU admissions, which is a binary variable (yes/no). I chose ICU admission as the outcome variable since we wanted to see what the effects of each treatment were on ICU admissions. Since the outcome variable was binary, standard linear regression was not appropriate.

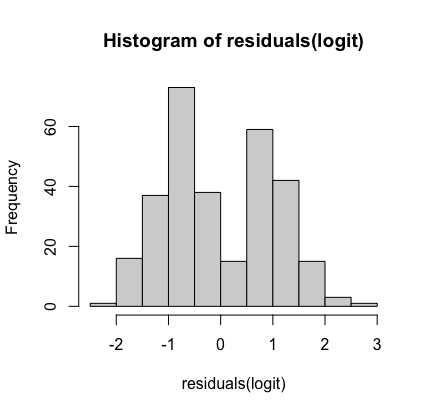
**Results:**

Here is the output of my regression analysis, along with odds ratios and residual plots/histograms:









**Conclusions:**

The logistic regression finds that odds of ICU admission is significantly correlated with comorbidity, gender and treatment. Our AIC was 330, and our residuals were normally distributed, indicating that the model fits the data well. Additionally, log odds ratios the treatments were greater than 1. In this analysis, treatment I was used as the baseline to which treatment II and I+II were compared. Since the odds ratio is greater than 1 for both treatment = II and I+II, it means that the likelihood of ICU admission is greater than the likelihood of not being admitted to the ICU. Treatment I+II had the highest odds, resulting in the most ICU admissions as evidenced by its 4.65 odds ratio. Treatment II had the second highest odds, evidenced by its 2.15 odds ratio, meaning that treatment I is least likely to result in ICU admission. Thus, the data shows that treatment I is the best treatment to reduce the odds of ICU admission. Other variables did not appear to be statistically significant. I also wanted to see if I could find specific treatments that worked well for certain groups. I tested a number of interactions to see if this would be the case, but no interaction gave significant results so I decided to leave them out of this analysis and felt most comfortable with the model above.

**Question 3**

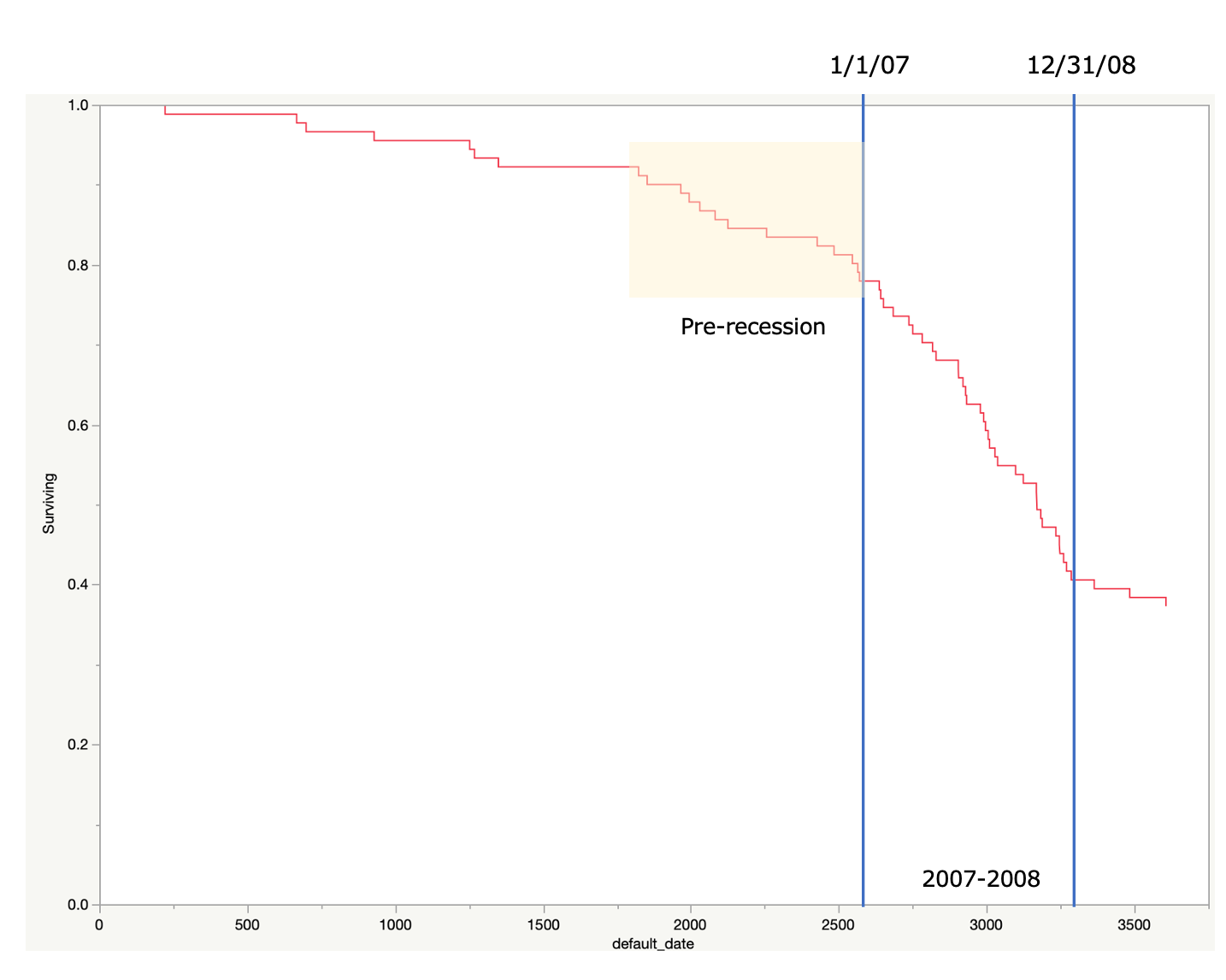
**Methods**

For this question, we are interested in examining if car repossessions increased during the 2007-2008 recession. We were given date of car loan, default date of loan (with no value if debtor didn’t default), debtor’s credit score and the monthly payment amount. I also added columns in the dataset to indicate whether or not the default date was within the desired interval (2007-2008) or if the person defaulted on their debt or not (as evidenced by whether or not there was a value for default date). All of the variables are normally distributed.

Given the censored nature of the data, along with differing periods of follow up, it made most sense to use survival methods to analyze these data. A Kaplan Meir Curve was constructed to measure time to loan default, and a Cox proportional hazard regression was conducted to see if other variables had an effect on survival time. Survival analysis was useful because we were dealing with censored data. An analysis was performed using logistic regression to see if having a loan during the period 2007-2008 had an impact on the outcome Additionally, using the extra 2 columns generated (default and if data was collected within interval), a logistic regression was done with data collected in the interval as the predictor and default as the outcome. Logistic regression was used because the outcome variable was binary, and this analysis would directly tell us if the data falling within the time interval of interest had an effect on car repossessions. Importantly, I am assuming that the bank repossessed every car whose owner defaulted on the loan. All of this analysis was done using SAS JMP software, since I am familiar with that software for fitting Kaplan Meier curves. If there were no defaults recorded after 3655 days, that datapoint was considered to be censored.

**Results:**

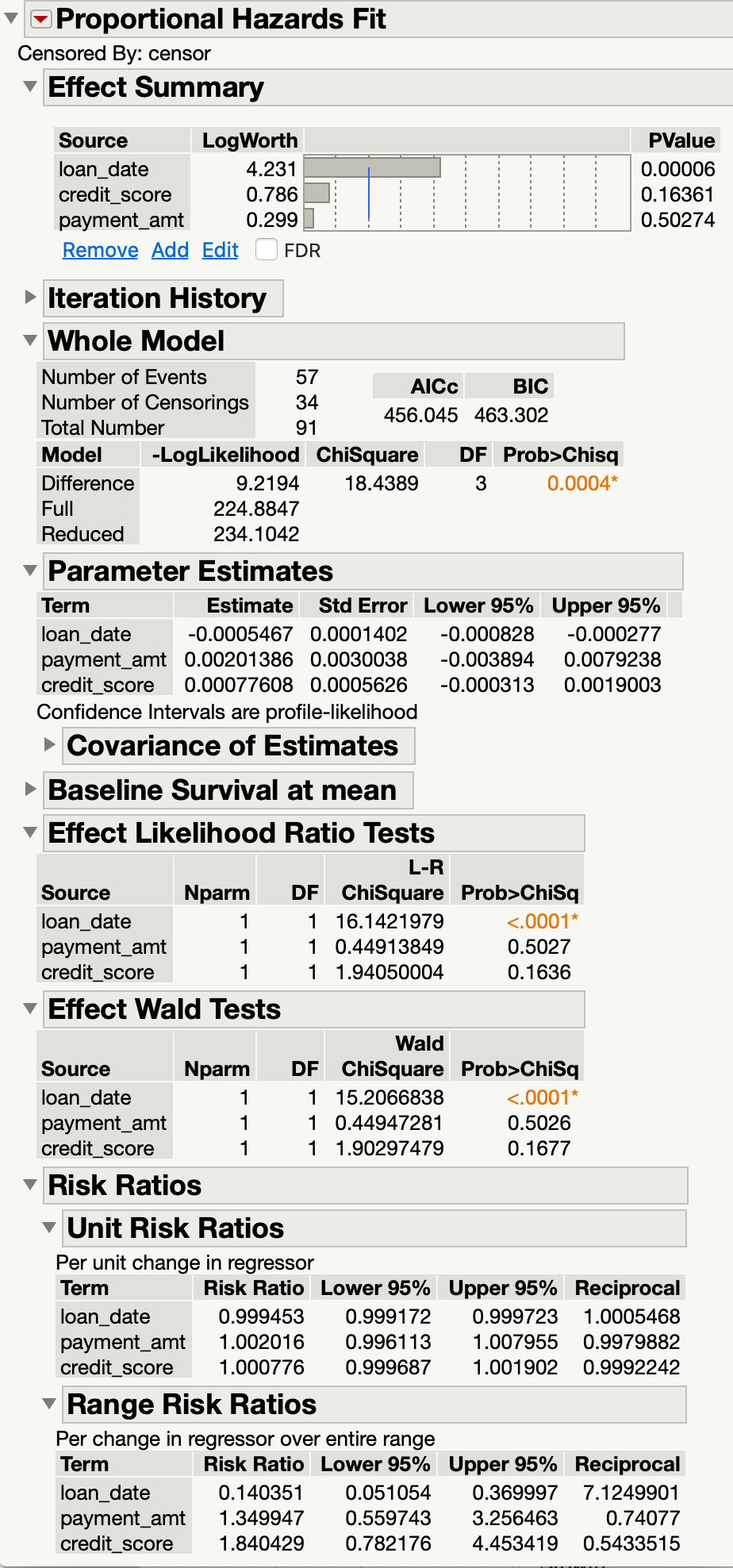
Kaplan Meier Curve (Time to default vs time):



Cox Proportional Hazards Regression:

Outcome variable: time to default (default\_date)

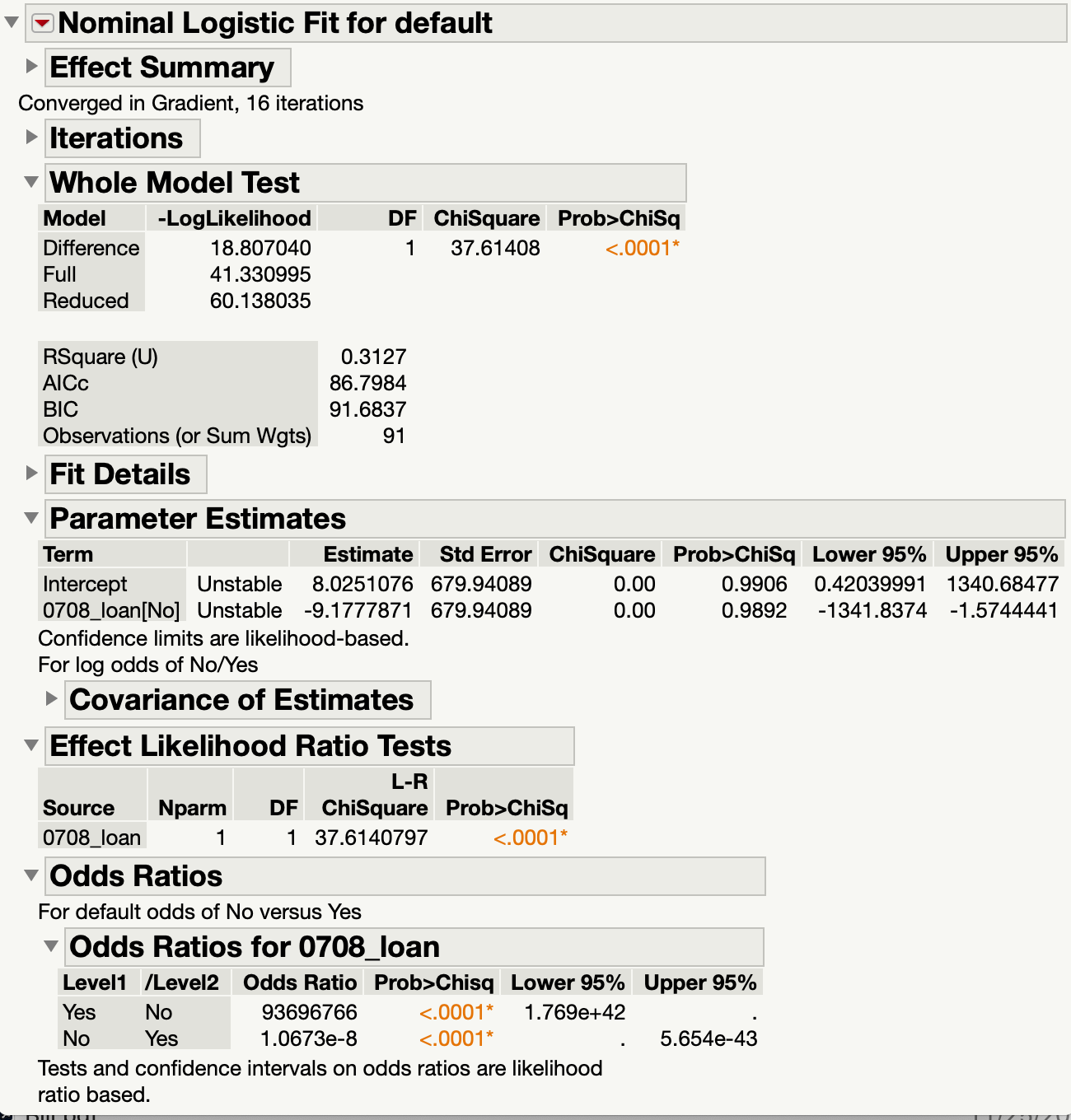
Predictor variables: loan\_date, payment\_amt, credit\_score



Logistic Regression:

Outcome variable: default yes/no (1 = yes; 0 = no)

Predictor: 07088loan (1 = yes loan\_date between 1/1/07 and 12/31/08; 2 = no)



**Conclusions:**

The KM curve on the Y axis shows time to default, and each downward tick in the curve represents another default. The X axis represents time (in days) since Jan 1, 2000. At time zero, there were no defaults, hence the Y axis starts at 1.00. Not everyone in this dataset defaults, and by the end of the time period measured, approximately 40% have not defaulted. So all persons who did not default by XXXX days were censored in the dataset I used for analysis.

Visually, it is unclear if the KM curve shows a constant hazard over time. Specifically, see at the time period up until about Day 1600, the slope of the KM curve is fairly shallow. But in the two-year period prior to the recession, the rate of defaults increases sharply (see yellow shaded area in KM curve). The rate increases even further during the 2007-2008 recession, and then equalizes afterwards. I know that one of the assumptions for survival analysis is that the hazard must be constant, so I’m not sure if this violates those assumptions or not.

But in order to proceed onto the next step (Cox analysis), I assumed that the constant hazard assumption was not violated. For the Cox Proportional hazards model, the outcome variable remained the same (time to default). The following predictors were added to the Cox model: loan\_date, credit\_score and payment\_amt. Of these variables, only loan\_date reached statistical significance (p = 0.00006). The associated hazard ratio (or risk ratio) was 0.999453 (95% CI: 0.999172, 0.999723). The significance of the loan\_date variable deserved further investigation, and hence I performed another analysis using logistic regression.

The outcome variable for the logistic regression was default (yes/no). The only predictor for the model was whether or not the loan\_date was within the 2007-2008 time period. For this a binary predictor variable was computed (yes/no). The results showed that there was a statistically significant relationship between having a loan date in the 2007-2008 time period and odds of default (P < 0.0001), with an odds ratio of 93696766.

So based on three analyses (Kaplan Meier, Cox Proportional Hazards, and Logistic Regression), this seems to suggest that the time to default and rate of defaults during the 2007-2008 time period are different (and higher) than the other time periods.