UNIVERSITY OF SOUTH FLORIDA COLLEGE OF PUBLIC HEALTH

Estimating Progression to Plasma Cell Malignancy in Individuals with Monoclonal Gammopathy of Undetermined Significance

Lindsey J. Fiedler, M.Sc.
4/23/2018

Estimating Progression to Plasma Cell Malignancy in Individuals with Monoclonal Gammopathy of Undetermined Significance

Lindsey J. Fiedler, M.Sc.

Abstract

Objective: To model the progression of individuals with monoclonal gammopathy of undetermined significance to a plasma cell malignancy, and to identify the significant predictors of a plasma cell malignancy.

Methods: Patient records collected between 1960 and 1994 for 1384 individuals with monoclonal gammopathy of undetermined significance were analyzed. Kaplan-Meier was used to obtain a crude survivorship; subsequently a Cox proportional hazard model was used to identify the significant predictors of plasma cell malignancy.

Results: Median time before development of a plasma cell malignancy was 31 years. Hgb and monoclonal serum spike were found to be significant predictors of progression to plasma cell malignancy. A hazard ratio of 2.48 per unit increase was estimated for serum monoclonal concentration levels, as well as an 11% reduction in hazard per unit increase of Hemoglobin levels.

Conclusions: High survival probabilities even at longer time points reflect the low prevalence of progression to a plasma cell malignancy. This study is limited by the use of potentially outdated data and by the low number of events, resulting in poor statistical power to detect possible differences in survivorship between genders. Future work should consider the most recent criteria for diagnosing monoclonal gammopathy of undetermined significance as well as the inclusion of race and family history as a covariate.

Introduction

Monoclonal gammopathy of undetermined significance (MGUS) is a condition where there is an abnormal protein in the blood[1]. The monoclonal immunoglobulin (Ig), or M protein, is produced by the plasma cells in the bone marrow. In monoclonal gammopathy of undetermined

significance, the M protein can accumulate to such levels that it inhibits healthy cells and can lead to tissue damage. Although the condition is generally asymptomatic and very seldom problematic, monoclonal gammopathy of undetermined significance can progress to more serious disorders such as blood cancer. In fact, a study done by van de Donk et al. found that MGUS will commonly precede multiple myeloma, a cancer of the plasma cells[2].

The present study aims to model the survivorship of individuals with monoclonal gammopathy of undetermined significance. Since 2003, the International Myeloma Working Group has defined a set of criteria used to diagnose monoclonal gammopathies and multiple myelomas[3]. The criteria are based on the levels of certain proteins and other biological products present in the blood (e.g., creatinine and hemoglobin). In addition to estimating survivorship, the present study will analyze which criteria are useful predictors for progression to plasma cell malignancy.

Methods

Data

The data for this analysis has been donated courtesy of Dr. Robert Kyle of the Mayo Clinic[4]. It contains records for 1384 patients in southeastern Minnesota who were diagnosed with monoclonal gammopathy of undetermined significance between 1960 and 1994. All patient records have been de-identified. The baseline characteristics recorded were age at diagnosis, patient gender and values for M protein, hemoglobin (Hgb) and serum creatinine levels. Inclusion was limited to those with serum monoclonal values of 3g per deciliter or less. Patients were followed for a median of 15.4 years. If a plasma cell malignancy developed, the time at which it was detected was reported as time of event occurrence.

A total of 46 records presented a missing value in at least one measurement. These were distributed as follows: 13 missing hemoglobin, 30 missing serum creatinine, and 11 missing M protein values. Missing values were imputed using Multiple Imputations by Chained Equations with predictive mean matching, a strategy that imputes missing data by estimating the value based on the observations for that record as well as similar records[5]. Unlike other methods, the result of the process is not a single full dataset, but multiple full datasets. As such the results of any statistical analysis should be pooled. To ensure that the imputed data was valid based on the

distribution of the original data points, an evaluation of the created data was performed and can be seen in **Appendix A**.

Statistical Analysis

All analyses were done using R 3.4.4[6] using the mice package[7] for performing imputations and the survival package[8, 9] for the analysis. The event of interest is development of a plasma cell malignancy, and survival time was considered to be the months spent in the study up until detection if the event occurred, or the months up to last contact if the event did not occur.

A crude analysis of survivorship was done using the Kaplan-Meier product limit estimator. Stratification by gender was performed to assess survivorship for each gender and the logrank test was used to evaluate any significant differences. To identify the predictors of progression to a plasma cell malignancy, a Cox proportional hazards model was fitted. An initial model that included all potential predictors was built and then reduced to only include those indicated by a backward stepwise selection. To ensure that the interpretations are valid, the appropriateness of the proportional hazards model was assessed. **Appendix B** presents the validation of the selected model. R code and output for the analysis can be seen in **Appendix C**.

Results

Table 1. Univariate analysis of patient characteristics

	N = 1384 **Median (q1, q3)**
Age	72 (63, 79)
Gender	
Male (%)	54.41%
Hgb	13.5 (12.2, 14.7)
Serum Creatinine	1.1 (0.9, 1.3)
Serum M Spike	1.2 (0.6, 1.5)
PCM event (%)	8.31%
Time to PCM (months)	81 (37, 136.25)

Table 1. shows the characteristics of the study participants as measured at baseline. Median age was 72 years, with the youngest participant being 24 and the oldest 96. There were slightly more

males than females, 54% to 46%, respectively. Monoclonal gammopathy of undetermined significance is known to develop in older adults more than in younger individuals. In adults aged 50 years or older a prevalence of 3.2% has been estimated, and for adults aged 70 years or older it increases to 5.3%[10]. Men also have an increased prevalence compared to women[1], thus, the data is representative of the general population.

The median hemoglobin level for the overall sample was 13.5 g/dL. A normal range for hemoglobin is considered 13.5-17.5 g/dL and 12.0-15.5 g/dL for males and females, respectively[11]. Of the sample, 286 males and 169 females fell below their normal range possibly indicating a progression towards anemia, one of the criteria for diagnosing a plasma cell malignancy. 288 individuals also measured outside the normal range for serum creatinine[11] (0.6-1.3 mg/dL), with the median being reported as 1.1 mg/dL. For serum monoclonal protein levels, there is no accepted range since its presence in the blood is considered abnormal. The median concentration was 1.2 g/dL.

Figure 1. Survival curve for MGUS with no explanatory variables

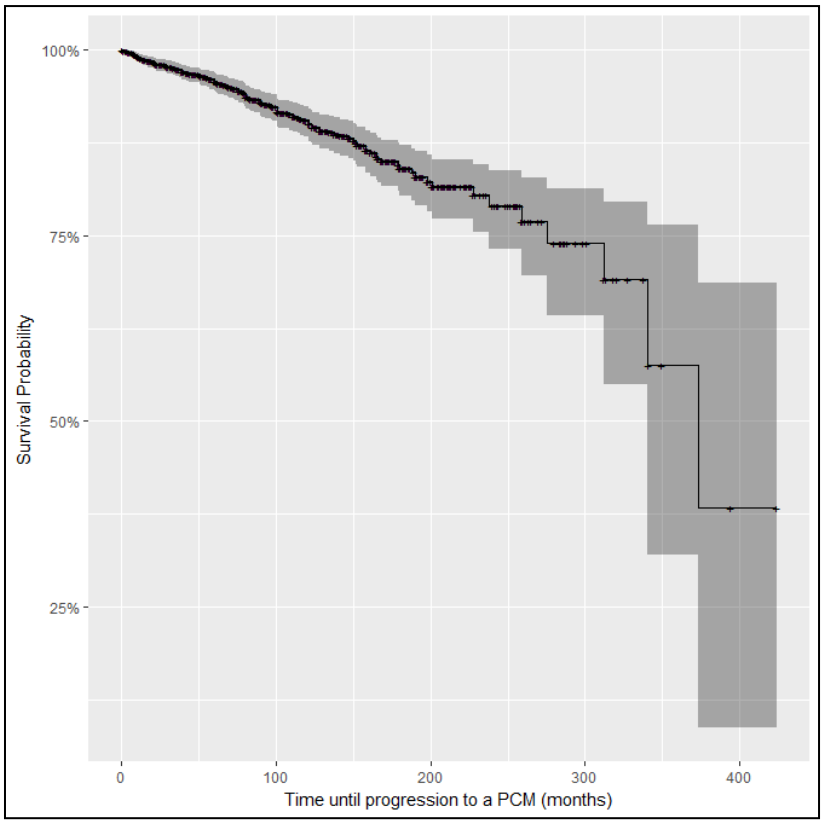
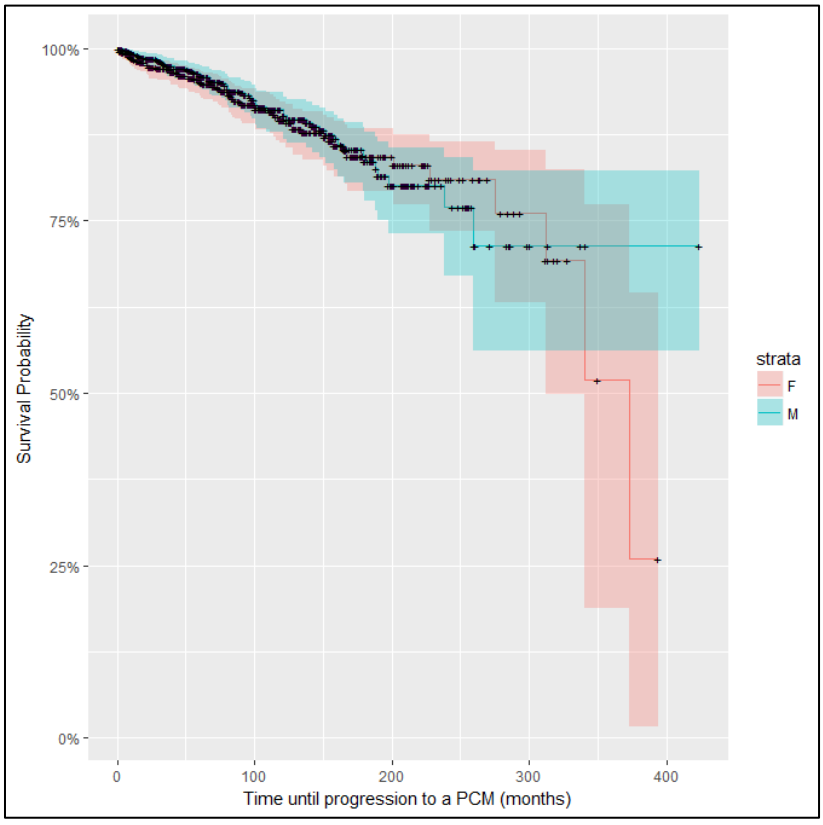


Figure 1. shows the survival curve for a crude analysis using Kaplan-Meier. A total of 115 participants progressed to a plasma cell malignancy (8.31%), with a median time to event of 81 months, or approximately 7 years. At this time point, 93.7% (95% CI: 340, *NA*) had still not experienced the event. Since monoclonal gammopathy of undetermined significance is asymptomatic in most individuals, it is of interest to evaluate survivorship at longer time points where a plasma cell malignancy has had more time to develop. By year 20, survivorship falls to 79% (95% CI: 73.2, 83.8), and by year 35 (the longest measured time point), it has decreased to 38.3% (95% CI: 8.74, 68.6). Median survival for the entire length of study was 31 years.

Because prevalence of monoclonal gammopathy of undetermined significance differs between sexes, stratification by gender was performed to assess if there were any differences in progression to a plasma cell malignancy. **Figure 2.** plots the survival probabilities of men compared to women. A Mantel-Haenszel logrank test revealed no significant difference (*p*-value of 0.751) in the survivorship for men vs. women, indicating that while prevalence of MGUS may be higher in men, it is not a predictor for progression to plasma cell malignancy.

Figure 2. Survival curve for MGUS stratified by gender



The predictors of progression to plasma cell malignancy were found by fitting a Cox proportional hazards model. No covariate was found to be in violation of the proportional hazards assumption, and after reduction of the initial full-model through a backward stepwise selection, the final model included only age, Hgb and monoclonal serum spike as predictors. The final model was re-built using the original dataset with missing values. No significant differences were found between the coefficient estimates for the pooled imputed model and this model. The results presented are therefor from the model built using the original data.

Table 2. Hazard ratios for selected cox proportional hazards model

	HR	95% CI lower	95% CI upper
Age	1.011	0.995	1.028
Hgb	0.889	0.804	0.983
Serum M Spike	2.48	1.793	3.429

Figure 3. Survival curve for reduced Cox proportional hazards model

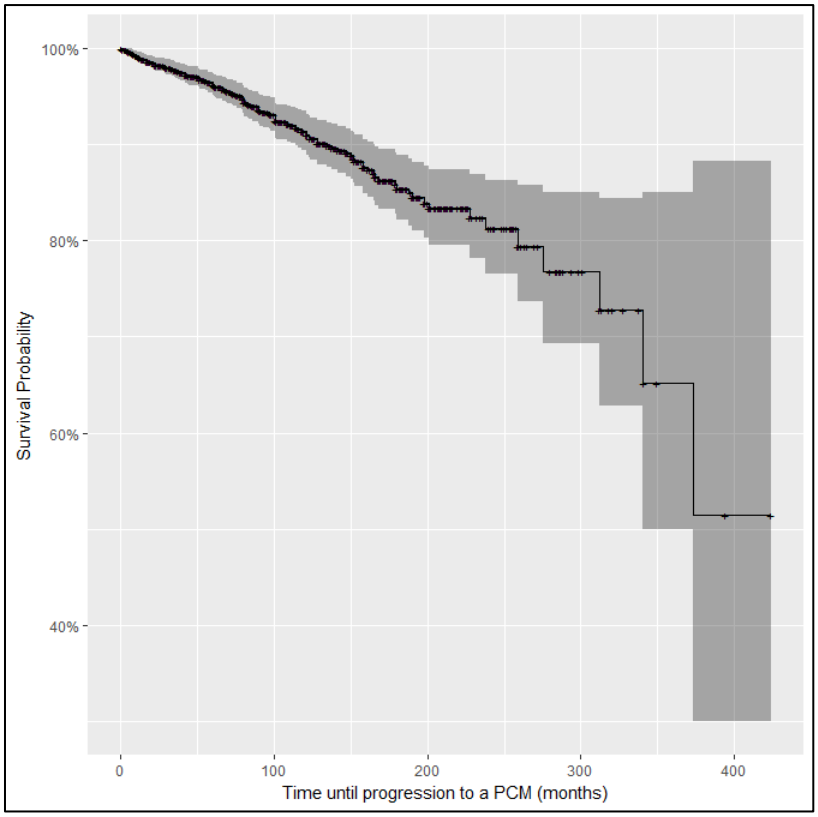


Table 2. presents the selected predictors for progression to plasma cell malignancy and their corresponding hazard ratios. The serum monoclonal spike is the most significant predictor of a plasma cell malignancy with a hazard ratio of 2.48 per unit increase. Hemoglobin levels were found to be protective of plasma cell malignancy, with an 11% reduction in hazard per unit increase. This is consistent with the diagnosis criteria for plasma cell malignancy, since progression to anemia, meaning low Hgb levels, is a characteristic of the disease. Age was not found to be a significant predictor for progression to plasma cell malignancy and its removal does not change the hazard ratios for the other covariates in a meaningful way.

The survival curve for this reduced model is shown in **Figure 3.** Results do not differ substantially from the crude model. Median survival for this adjusted model is again 31 years. At 81 months, the median time to event, survival probabilities are still high at 93.7% (CI: 91.5, 95). For years 20 and 35, survivorship is 79% (CI: 73.2, 83.8) and 38.3% (CI: 8.74, 68.6), respectively. The high survival probabilities even at longer time points reflect the low prevalence of progression to a plasma cell malignancy.

Discussion

Monoclonal gammopathy of undetermined significance has a low prevalence in the general population with even fewer progressing to a plasma cell malignancy. This low prevalence is fortunate since it is mostly common in older adults, and the severity of a plasma cell malignancy can further complicate a potentially already compromised health. Unfortunately, the low number of events results in analyses with low statistical power. The follow-up in the data used for this study spanned 35 years and only 115 events occurred.

A further limitation of this study is the use of potentially outdated data. Collected between 1960 and 1994, the data is perhaps not up to date with the most recent criteria for diagnosing monoclonal gammopathies and plasma cell malignancies. The International Myeloma Working Group has updated the criteria more than once since the original collection, and it is possible that the data now presents misclassifications for event status. Since newer criteria now result in earlier intervention, the misclassification is likely differential with bias towards the null. In other words, using the current guidelines, there are potentially more events than the 115 reported, and the true survival probabilities are likely lower.

Additionally, the newer criteria incorporate the results for other clinical tests for which there are no recorded data and, therefore, cannot be included in the analysis. These are: a value of clonal-bone-marrow-plasma-cells greater than 60%, a serum free-light-chain-ratio of more than 100 and two or more focal-lesions detected on a MRI[3]. The original authors of the study[4] that resulted in the used dataset recently published a new study using the same patient data but extending the follow-up 15 years to December 2015[12]. Their analysis included covariates for the new guidelines for diagnosis. In total 147 events were detected. It is unclear if the 32 additional events all occurred during the 15 year extension or if any were the result of misclassifications. Results from their study found the serum monoclonal spike to be a significant predictor along with an abnormal serum free light-chain ratio.

Since the initial study by Kyle et al. was conducted, other risk factors for monoclonal gammopathy of undetermined significance and plasma cell malignancy have been identified[1]. For example, race and family history of MGUS are known to affect risk of monoclonal gammopathy of undetermined significance. Future work should consider the most recent criteria for diagnosing monoclonal gammopathy of undetermined significance as well as the inclusion of race as a covariate.

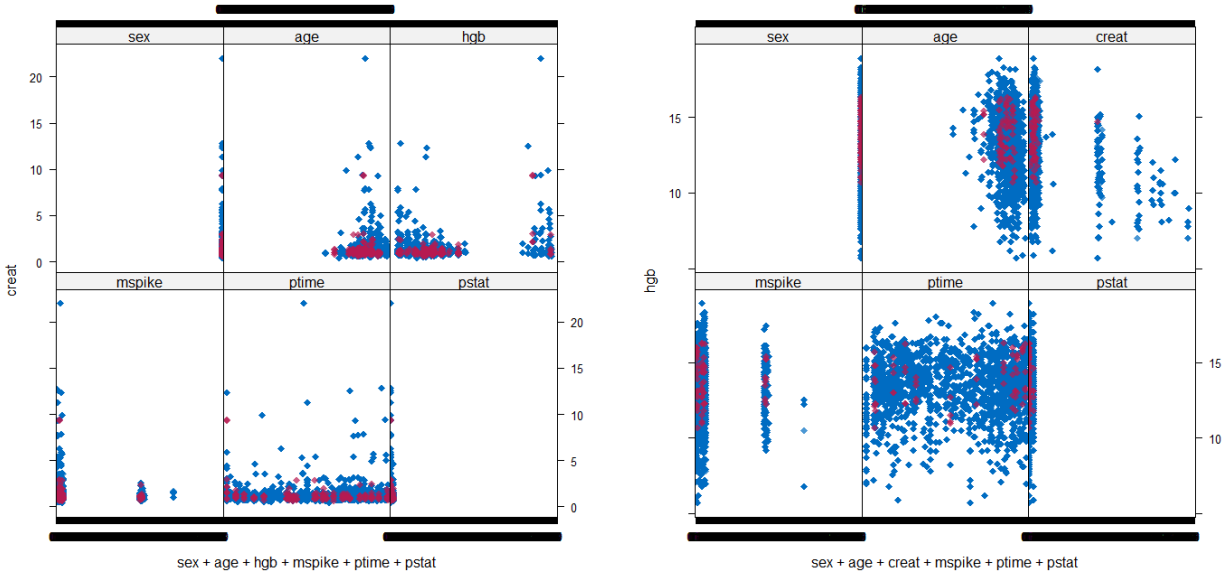
Reference List

1. Monoclonal gammopathy of undetermined significance (MGUS). Mayo Clinic. http://www.mayoclinic.org/diseases-conditions/mgus/symptoms-causes/syc-20352362. Published July 29, 2017. Accessed April 15, 2018.
2. Van de Donk NWCJ, Palumbo A, Johnsen HE, et al. The clinical relevance and management of monoclonal gammopathy of undetermined significance and related disorders: recommendations from the European Myeloma Network. *Haematologica*. 2014;99(6):984-996. doi:10.3324/haematol.2013.100552.
3. International Myeloma Working Group (IMWG) Criteria for the Diagnosis of Multiple Myeloma. International Myeloma Working Group. http://imwg.myeloma.org/international-myeloma-working-group-imwg-criteria-for-the-diagnosis-of-multiple-myeloma/. Published November 9, 2015. Accessed April 15, 2018.
4. Kyle RA, Therneau TM, Rajkumar SV, et al. A Long-Term Study of Prognosis in Monoclonal Gammopathy of Undetermined Significance. *New England Journal of Medicine*. 2002;346(8):564-569. doi:10.1056/nejmoa01133202.
5. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple Imputation by Chained Equations: What is it and how does it work? *International journal of methods in psychiatric research*. 2011;20(1):40-49. doi:10.1002/mpr.329.
6. R: A language and environment for statistical computing [computer program]. Version 3.4.4. Vienna, Austria: R Foundation for Statistical Computing; 2018.
7. Buuren SV, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations inR. *Journal of Statistical Software*. 2011;45(3). doi:10.18637/jss.v045.i03.
8. A Package for Survival Analysis in S [computer program]. Version 2.38. Therneau T; 2015.
9. Borgan Ø. Modeling Survival Data: Extending the Cox Model. Terry M. Therneau and Patricia M. Grambsch, Springer-Verlag, New York, 2000. No. of pages: xiii 350. Price: $69.95. ISBN 0-387-98784-3. *Statistics in Medicine*. 2001;20(13):2053-2054. doi:10.1002/sim.956.abs.
10. Kyle RA, Therneau TM, Rajkumar SV, et al. Prevalence of monoclonal gammopathy of undetermined significance. N Engl J Med. 2006;354(13):1362-9.

11. Understanding your MULTIPLE MYELOMA LAB TESTS. Takeda Oncology. http://www.velcade.com/files/PDFs/Understand_your_Lab_Tests_Resource_(MM). Accessed April 20, 2018.
12. Kyle RA, Larson DR, Therneau TM, et al. Long-Term Follow-up of Monoclonal Gammopathy of Undetermined Significance. N Engl J Med. 2018;378(3):241-249.

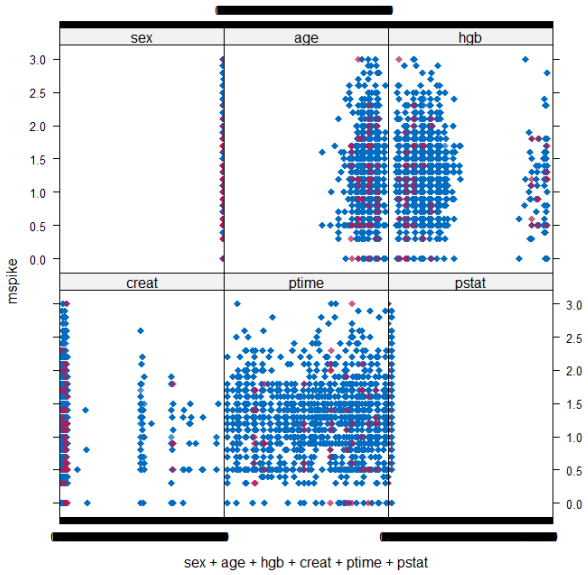
Appendix A: Imputed Data Validation

Figure 4. Scatterplots for variables with imputed values against all other variables



(a) Creatinine

(b) Hgb



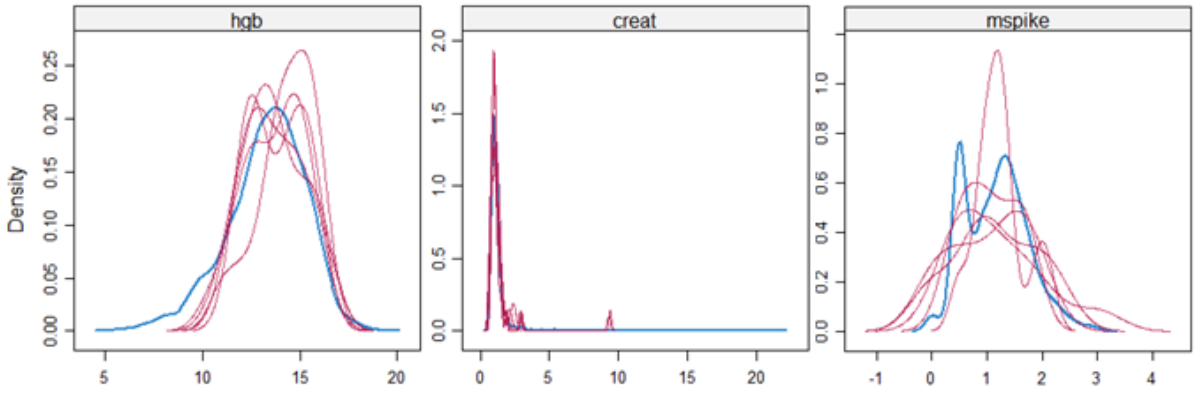
(c) Monoclonal spike

Blue data points correspond to the original data and magenta data points are imputed values

Missing data was imputed using multiple imputations by chained equations, a process which results in more than one dataset being generated. For the purposes of this analysis, 5 datasets were generated. Each dataset was inspected to ensure imputed values were plausible given the distribution of the original data. **Figure 4.** shows a set of scatterplots for one of the generated

datasets. **Figure 4a.** plots imputed data creatinine against all other variables. **Figures 4b.** and **4c.** do the same for Hgb and Monoclonal spike, respectively. Blue points plot the original non-missing data and magenta points correspond to imputed values. In all plots, the magenta points fall along the blue points forming a similar shape. This indicates that it is plausible that both set of points, original and imputed, belong to the same distribution. This is further supported by the density plots shown in **Figure 5.** Again, the blue density curves correspond to the original data and the magenta to imputed values, 1 per generated dataset.

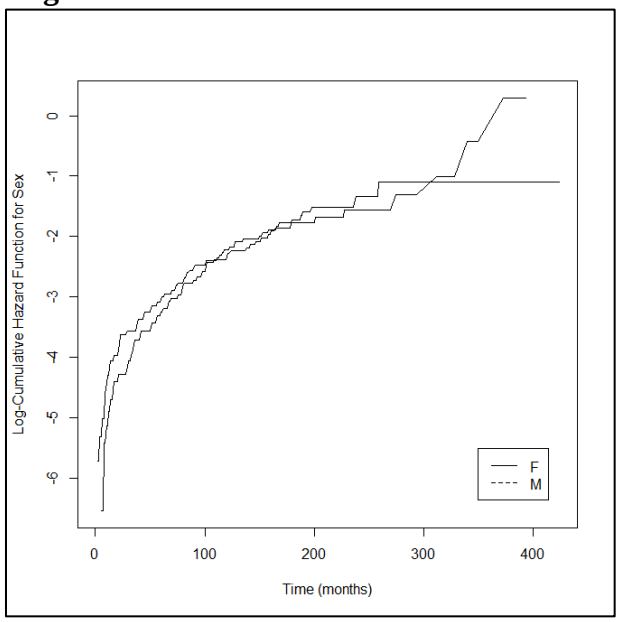
Figure 5. Density curves for variables with imputed values



Appendix B: Model Validation

To ensure that the proportional hazards assumption was met, the hazards curve for the only categorical covariate was plotted and can be seen in **Figure 6.** The curves for male and female are very close together and ascend at a similar rate. Further inspection will be need to evaluate if a correlation with time is present.

Figure 6. Hazard curves for male and females



The initial model included all covariates. **Table 3.** shows the coefficient estimates for each independent predictor. Only M spike and Hgb were found to be statistically significant. **Table 4. g**ives the results of covariate correlation tests with time for one of the imputed data set. The results of the other four are similar and are not shown. From the results, it is revealed that no covariate has a correlation with. The global correlation coefficient is also high.

Table 3. Cox proportional hazards full model estimates coefficient estimates

	Est.	SE	Pr(>|t|)	95% CI lower	95% CI upper
Sex (M)	0.186	0.204	0.360	-0.213	0.585796
Age	0.012	0.008	0.139	-0.004	0.028893
Hgb*	-0.142	0.055	0.009	-0.250	-0.03496
Creatinine	-0.187	0.208	0.368	-0.594	0.22011
Serum M Spike*	0.862	0.164	< 0.001	0.541	1.182819

***** *Coefficient found to be statistically significant*

Table 4. Results of covariate correlation tests with time for a single imputed dataset

	rho	χ^2	p
Sex (M)	0.069	0.52935	0.467
Age	-0.16858	2.27376	0.132
Hgb	-0.08885	0.96138	0.327
Creatinine	-0.02335	0.06438	0.8
Serum M Spike	-0.00469	0.00258	0.959
GLOBAL	NA	3.07841	0.688

Figure 7. plots the results of the covariate time correlation tests. Though *Age* does show a slight downward trend, its correlation coefficient was not found to be significant. Further inspection of the Martingale (**Figure 8a.**) and Schoenfeld residuals (**Figure 9a.**) confirm no correlation.

Figure 7. Plots of covariate time correlation tests for each variable

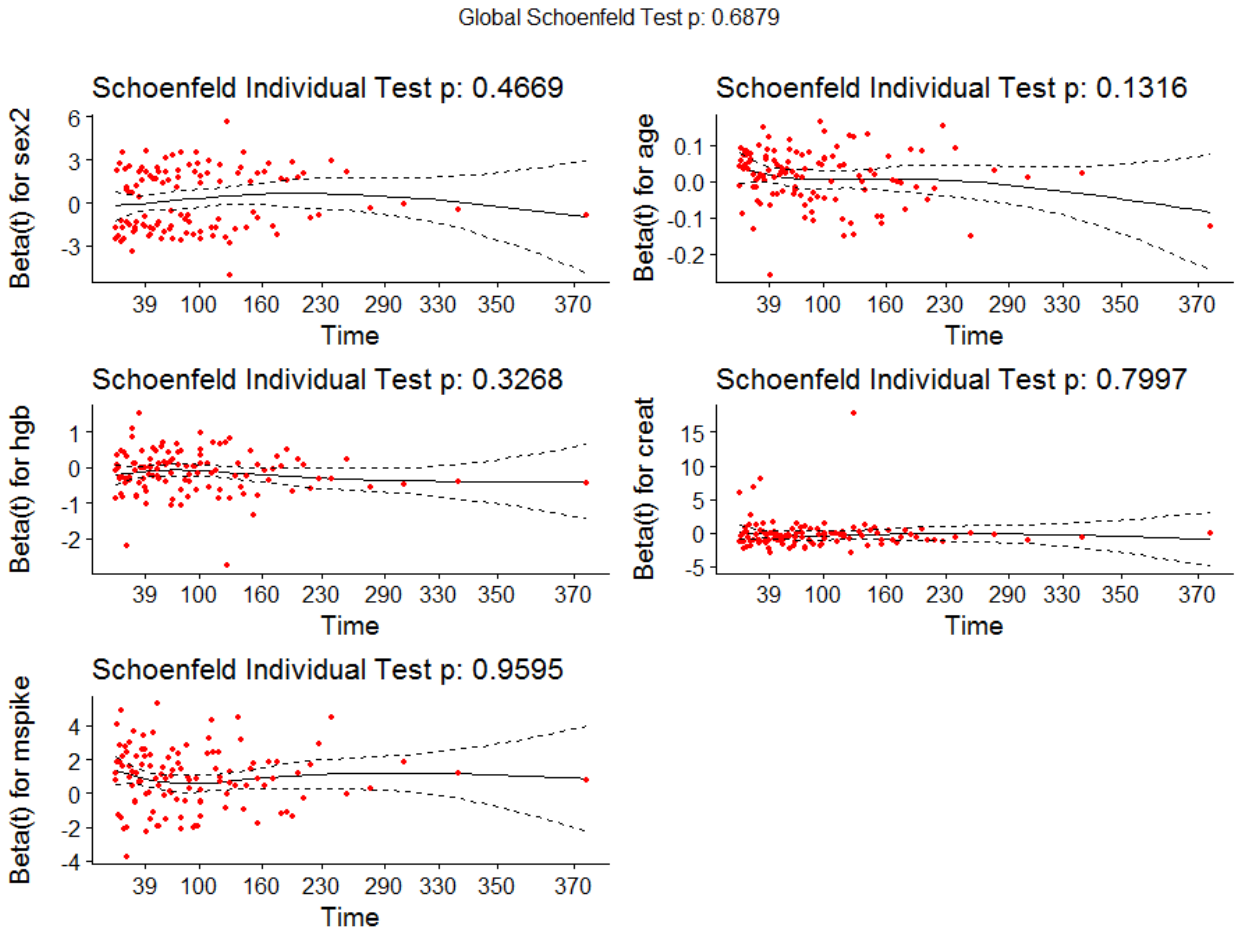
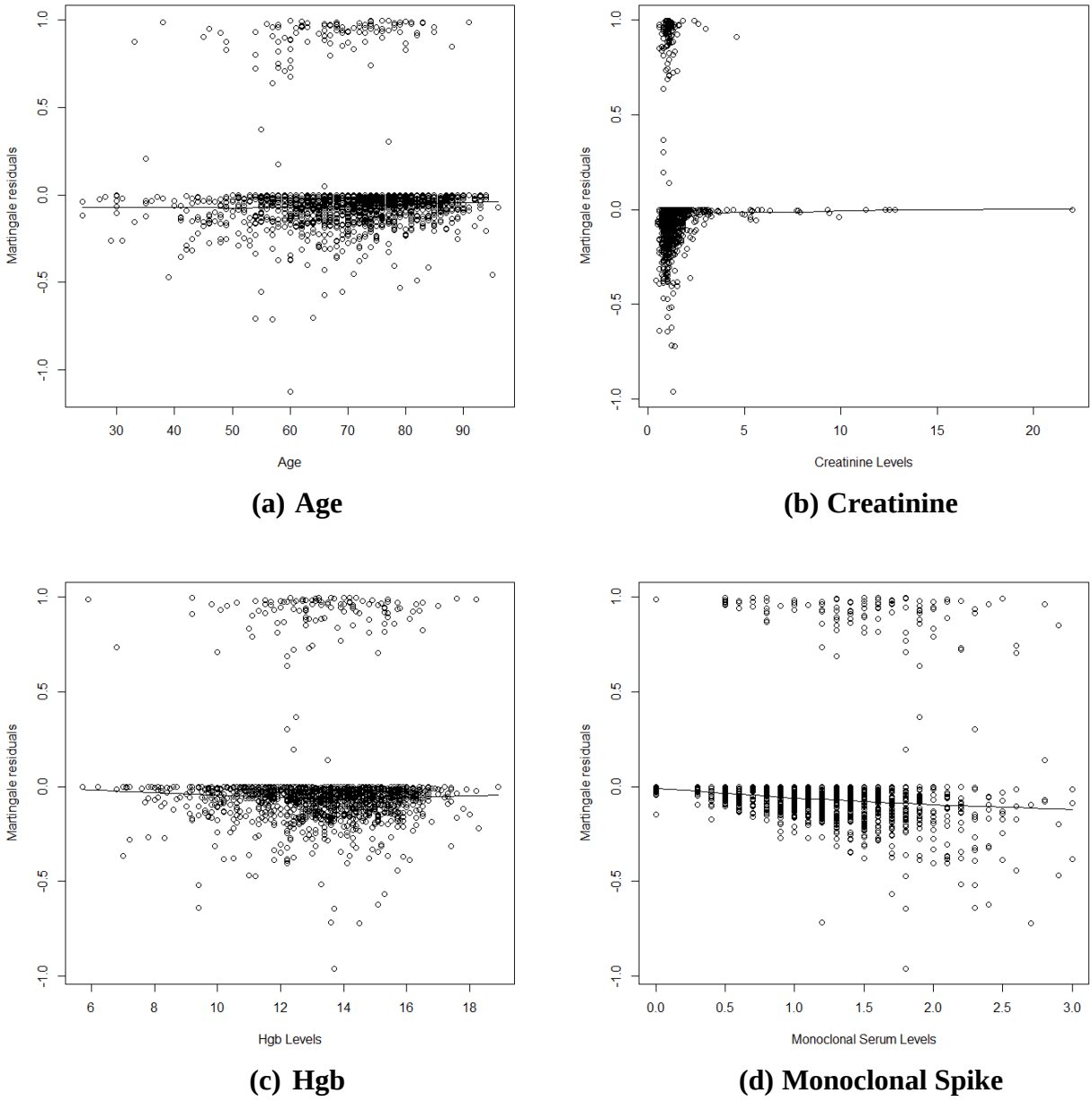


Figure 8. Martingale residual plots for continuous covariates

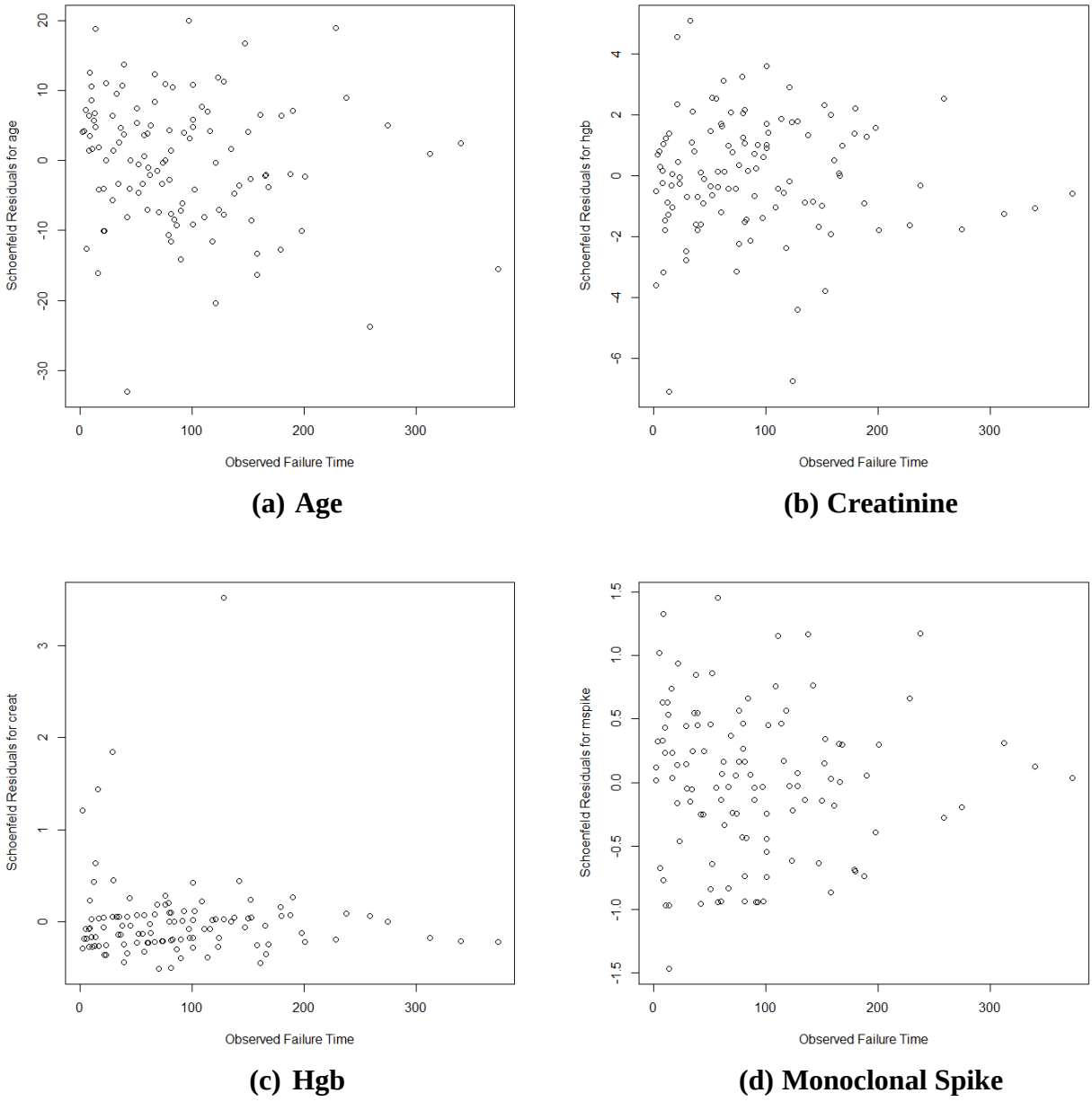


(a) Age **(b) Creatinine**

(c) Hgb **(d) Monoclonal Spike**

Figure 8. plot the Martingale residuals for each of the continuous covariates present in the model. All data points are clustered around zero and no visible trend is detected, though a slight declining trend is seen for *Monoclonal Spike.* Inspection of this variables Schoenfeld plot (**Figure 9d.**) as well as its correlation coefficient with time shows no egregious violation of the proportional hazards assumption.

Figure 9. Schoenfeld residual plots for continuous covariates



(a) Age **(b) Creatinine**

(c) Hgb **(d) Monoclonal Spike**

Figures 9. and **10.** plot the Schoenfeld residuals for continuous covariates and categorical covariates, respectively. No visible trend is seen in any plot and all data points are distributed fairly evenly around zero with convergence occurring in greater values for the *x*-axis.

Finally, the presence of influential outliers was evaluated by plotting the *dfbetas* of every variable for each data point. Large dfbeta values indicate the data point to be very influential in the calculation of the coefficient. **Figure 11.** shows these results. In general, though some larger outliers do appear for *Creatinine,* none seem to be influential.

Figure 10. Schoenfeld residual plots for sex

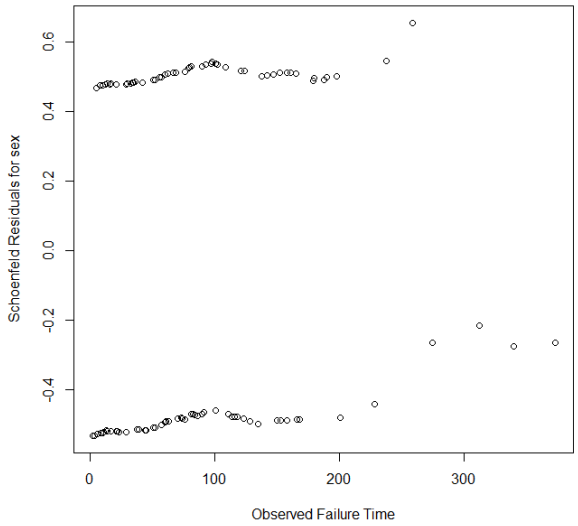
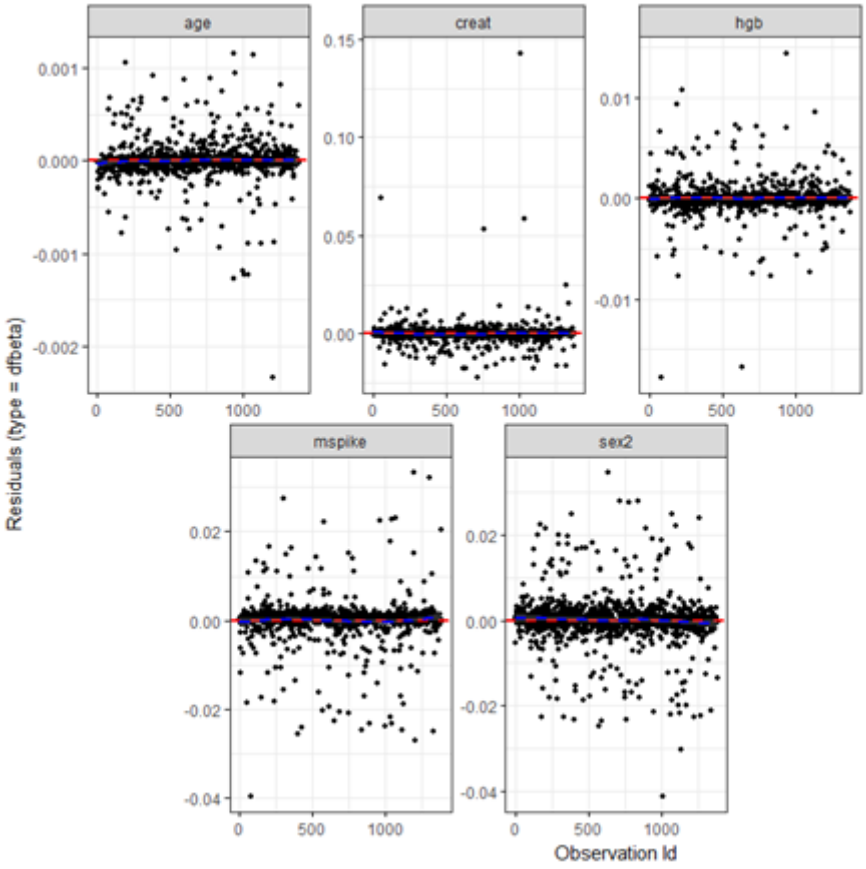


Figure 11. Dfbeta residual plots for each variable



Appendix C: R Code and Output

```
#load the necessary libraries
library(survival)
library(SurvRegCensCov)
library(Hmisc)
library(mice)
library(rms)
library(survminer)
library(ggplot2)
library(ggfortify)

######### Initial data preparation ##############################################

#read data into object
mgus<-read.csv(file.choose(), header=T)

#confirm object is dataframe
is.data.frame(mgus)

#drop columns that will not be used
vars <- names(mgus) %in% c("id", "X", "death", "futime")
mgus <- mgus[!vars]

#get summary statistics
str(mgus)
summary(mgus)

# Impute missing
pMiss <- function(x){sum(is.na(x))/length(x)*100}
apply(mgus,2,pMiss)

mgusImputed <- mice(mgus,m=5,maxit=50,meth='pmm')
summary(mgusImputed)

# Verify imputed data is valid based on distribution of original data points
xyplot(mgusImputed, creat ~ sex+age+hgb+mspike+ptime+pstat,pch=18,cex=1)
xyplot(mgusImputed, hgb ~ sex+age+creat+mspike+ptime+pstat,pch=18,cex=1)
xyplot(mgusImputed, mspike ~ sex+age+hgb+creat+ptime+pstat,pch=18,cex=1)

densityplot(mgusImputed)

stripplot(mgusImputed, pch = 20, cex = 1.2)

######################### Build initial KM models ################################
```

```
#fit survival curve without explanatory variables
mgus.fit<-with(mgusImputed, survfit(Surv(ptime, pstat)~1, error="greenwood", conf.type="log-log", conf.int=0.95))
summary(mgus.fit)
summary(mgus.fit,times=c(81,240,373))

#fit survival curve stratifying by gender
mgus.sex.fit<-with(mgusImputed, survfit(Surv(ptime, pstat)~sex, error="greenwood", conf.type="log-log", conf.int=0.95))
summary(mgus.sex.fit)

#Plot the curves
autoplot(mgus.fit$analyses[1], xlab="Time until progression to a PCM (months)", ylab="Survival Probability")
autoplot(mgus.sex.fit$analyses[1], xlab="Time until progression to a PCM (months)", ylab="Survival Probability")

#Perform the log-rank test with equal weights
mgus.sex.test0<-with(mgusImputed, survdiff(Surv(ptime, pstat)~sex, rho=0))
mgus.sex.test0

########## Visually inspect PH assumptions for categorical variables ###################

#Plot a km model with imputed data for gender
plot(mgus.sex.fit$analyses[[1]]$time, log(-log(mgus.sex.fit$analyses[[1]]$surv)),
xlab="Time (months)", ylab="Log-Cumulative Hazard Function for Sex",type="l",lty=1:2)
legend(350, -5.5, legend=levels(mgus$sex), lty=1:2)


########## Build initial model ########################

#Fit coxph model with imputed data sets and pool the results
mgus.ph.initial <- with(mgusImputed, coxph(Surv(ptime, pstat)~sex+age+hgb+creat+mspike))
mgus.ph.initial.pooled <- pool(mgus.ph.initial)
summary(mgus.ph.initial.pooled)

############### Check for zero correlation with time ###################

#Test each covariate for zero correlation between the time points and the associated sequence of estimates for the regression coefficient
tmp <- with(mgusImputed, cox.zph(coxph(Surv(ptime, pstat)~sex+age+hgb+creat+mspike)))
tmp

# Plot Schoenfeld test for each covariate
ggcoxzph(tmp$analyses[[1]])
```

```
#Check martingale residuals for each covariate
m.resid<-with(mgusImputed, resid(coxph(Surv(ptime, pstat)~sex+age+hgb+creat+mspike), "mart"))
par(mfrow=c(1,1))
plot(m.resid$analyses[[1]]~complete(mgusImputed,1)$sex, xlab="Sex", ylab="Martingale residuals")
plot(m.resid$analyses[[1]]~complete(mgusImputed,1)$age, xlab="Age", ylab="Martingale residuals")
lines(lowess(complete(mgusImputed,1)$age, m.resid$analyses[[1]]))
plot(m.resid$analyses[[1]]~complete(mgusImputed,1)$creat, xlab="Creatinine Levels", ylab="Martingale residuals")
lines(lowess(complete(mgusImputed,1)$creat, m.resid$analyses[[1]]))
plot(m.resid$analyses[[1]]~complete(mgusImputed,1)$hgb, xlab="Hgb Levels", ylab="Martingale residuals")
lines(lowess(complete(mgusImputed,1)$hgb, m.resid$analyses[[1]]))
plot(m.resid$analyses[[1]]~complete(mgusImputed,1)$mspike, xlab="Monoclonal Serum Levels", ylab="Martingale residuals")
lines(lowess(complete(mgusImputed,1)$mspike, m.resid$analyses[[1]]))

# Check Schoenfeld residuals for each covariate
s.resid<-with(mgusImputed, resid(coxph(Surv(ptime, pstat)~sex+age+hgb+creat+mspike), "scho"))
par(mfrow=c(1,1))
plot(as.numeric(dimnames(s.resid$analyses[[1]])[[1]]), s.resid$analyses[[1]][,1], xlab="Observed Failure Time", ylab="Schoenfeld Residuals for sex")
plot(as.numeric(dimnames(s.resid$analyses[[1]])[[1]]), s.resid$analyses[[1]][,2], xlab="Observed Failure Time", ylab="Schoenfeld Residuals for age")
plot(as.numeric(dimnames(s.resid$analyses[[1]])[[1]]), s.resid$analyses[[1]][,3], xlab="Observed Failure Time", ylab="Schoenfeld Residuals for hgb")
plot(as.numeric(dimnames(s.resid$analyses[[1]])[[1]]), s.resid$analyses[[1]][,4], xlab="Observed Failure Time", ylab="Schoenfeld Residuals for creat")
plot(as.numeric(dimnames(s.resid$analyses[[1]])[[1]]), s.resid$analyses[[1]][,5], xlab="Observed Failure Time", ylab="Schoenfeld Residuals for mspike")

# Check for influential outliers
ggcoxdiagnostics(mgus.ph.initial$analyses[[1]], type = "dfbeta", linear.predictions = FALSE, ggtheme = theme_bw())
ggcoxdiagnostics(mgus.ph.initial$analyses[[1]], type = "deviance", linear.predictions = FALSE, ggtheme = theme_bw())

###################### Preform variable selection ############################
# Backwards stepwise variable selection
mgus.ph.var <- with(mgusImputed, step(coxph(Surv(ptime, pstat)~sex+age+hgb+creat+mspike)))
mgus.ph.var.pooled <- pool(mgus.ph.var)
summary(mgus.ph.var.pooled)
```

```
################## Check Assumptions again for selected variables ###################
#Test each covariate for zero correlation between the time points and the associated sequence of estimates for the regression coefficient
tmp <- with(mgusImputed, cox.zph(coxph(Surv(ptime, pstat)~age+hgb+mspike)))
tmp

# Plot Schoenfeld test for each covariate
ggcoxzph(tmp$analyses[[1]])

#Check martingale residuals for each covariate
m.resid<-with(mgusImputed, resid(coxph(Surv(ptime, pstat)~age+hgb+mspike), "mart"))
par(mfrow=c(1,1))
plot(m.resid$analyses[[1]]~complete(mgusImputed,1)$age, xlab="Age", ylab="Martingale residuals")
lines(lowess(complete(mgusImputed,1)$age, m.resid$analyses[[1]]))
plot(m.resid$analyses[[1]]~complete(mgusImputed,1)$hgb, xlab="Hgb Levels", ylab="Martingale residuals")
lines(lowess(complete(mgusImputed,1)$hgb, m.resid$analyses[[1]]))
plot(m.resid$analyses[[1]]~complete(mgusImputed,1)$mspike, xlab="Monoclonal Serum Levels", ylab="Martingale residuals")
lines(lowess(complete(mgusImputed,1)$mspike, m.resid$analyses[[1]]))

# Check Schoenfeld residuals for each covariate
s.resid<-with(mgusImputed, resid(coxph(Surv(ptime, pstat)~sex+age+hgb+creat+mspike), "scho"))
par(mfrow=c(1,1))
plot(as.numeric(dimnames(s.resid$analyses[[1]])[[1]]), s.resid$analyses[[1]][,1], xlab="Observed Failure Time", ylab="Schoenfeld Residuals for age")
plot(as.numeric(dimnames(s.resid$analyses[[1]])[[1]]), s.resid$analyses[[1]][,2], xlab="Observed Failure Time", ylab="Schoenfeld Residuals for hgb")
plot(as.numeric(dimnames(s.resid$analyses[[1]])[[1]]), s.resid$analyses[[1]][,3], xlab="Observed Failure Time", ylab="Schoenfeld Residuals for mspike")


################### Build final model ###################################
#Build final model with imputations
mgus.ph.final <- with(mgusImputed, coxph(Surv(ptime, pstat)~age+hgb+mspike))
mgus.ph.final.pooled <- pool(mgus.ph.final)
summary(mgus.ph.final.pooled)
summary(survfit(mgus.ph.final$analyses[[1]]),times=c(81,240,373))
autoplot(survfit(mgus.ph.final$analyses[[1]]), xlab="Time until progression to a PCM (months)", ylab="Survival Probability")

#Build model without imputations
no_imputations <- coxph(Surv(ptime, pstat)~age+hgb+mspike, data=mgus)
```

```
no_imputations.fit <- survfit(no_imputations, error="greenwood", conf.type="log-log", conf.int=0.95)
summary(no_imputations.fit)
summary(no_imputations.fit, times=c(81,240,373))
summary(no_imputations)

autoplot(no_imputations.fit, xlab="Time until progression to a PCM (months)", ylab="Survival Probability")
```

Output

```
> #load the necessary libraries
> library(survival)
> library(SurvRegCensCov)
> library(Hmisc)
> library(mice)
> library(rms)
Loading required package: SparseM

Attaching package: ‘SparseM’

The following object is masked from ‘package:base’:

    backsolve

> library(survminer)
> library(ggplot2)
> library(ggfortify)
>
> ############### Initial data preparation ####################################
>
> #read data into object
> mgus<-read.csv(file.choose(), header=T)
>
> #confirm object is dataframe
> is.data.frame(mgus)
[1] TRUE
>
> #drop columns that will not be used
> vars <- names(mgus) %in% c("id", "X", "death", "futime")
> mgus <- mgus[!vars]
>
> #get summary statistics
> str(mgus)
'data.frame':   1384 obs. of  7 variables:
 $ age   : int  88 78 94 68 90 90 89 87 86 79 ...
```

```
 $ sex   : Factor w/ 2 levels "F","M": 1 1 2 2 1 2 1 1 1 1 ...
 $ hgb   : num  13.1 11.5 10.5 15.2 10.7 12.9 10.5 12.3 14.5 9.4 ...
 $ creat : num  1.3 1.2 1.5 1.2 0.8 1 0.9 1.2 0.9 1.1 ...
 $ mspike: num  0.5 2 2.6 1.2 1 0.5 1.3 1.6 2.4 2.3 ...
 $ ptime : int  30 25 46 92 8 4 151 2 57 136 ...
 $ pstat : int  0 0 0 0 0 0 0 0 0 0 ...
> summary(mgus)
      age         sex          hgb            creat            mspike
 Min.   :24.00   F:631   Min.   : 5.7   Min.   : 0.400   Min.   :0.000
 1st Qu.:63.00   M:753   1st Qu.:12.2   1st Qu.: 0.900   1st Qu.:0.600
 Median :72.00           Median :13.5   Median : 1.100   Median :1.200
 Mean   :70.42           Mean   :13.3   Mean   : 1.292   Mean   :1.164
 3rd Qu.:79.00           3rd Qu.:14.7   3rd Qu.: 1.300   3rd Qu.:1.500
 Max.   :96.00           Max.   :18.9   Max.   :22.000   Max.   :3.000
                         NA's   :13     NA's   :30       NA's   :11
    ptime            pstat
 Min.   :  1.00   Min.   :0.00000
 1st Qu.: 37.00   1st Qu.:0.00000
 Median : 81.00   Median :0.00000
 Mean   : 93.54   Mean   :0.08309
 3rd Qu.:136.25   3rd Qu.:0.00000
 Max.   :424.00   Max.   :1.00000

>
> # Impute missing
> pMiss <- function(x){sum(is.na(x))/length(x)*100}
> apply(mgus,2,pMiss)
      age       sex       hgb     creat    mspike     ptime     pstat
0.0000000 0.0000000 0.9393064 2.1676301 0.7947977 0.0000000 0.0000000
>
> mgusImputed <- mice(mgus,m=5,maxit=50,meth='pmm')

 iter imp variable
  1   1  hgb  creat  mspike
  1   2  hgb  creat  mspike
  1   3  hgb  creat  mspike
  1   4  hgb  creat  mspike
  1   5  hgb  creat  mspike
  .
  .
  .
  50   1  hgb  creat  mspike
  50   2  hgb  creat  mspike
  50   3  hgb  creat  mspike
  50   4  hgb  creat  mspike
  50   5  hgb  creat  mspike
```

```
> summary(mgusImputed)
Multiply imputed data set
Call:
mice(data = mgus, m = 5, method = "pmm", maxit = 50)
Number of multiple imputations:  5
Missing cells per column:
  age    sex    hgb  creat mspike  ptime  pstat
    0      0     13     30     11      0      0
Imputation methods:
  age    sex    hgb  creat mspike  ptime  pstat
 "pmm"  "pmm"  "pmm"  "pmm"  "pmm"  "pmm"  "pmm"
VisitSequence:
  hgb  creat mspike
    3      4      5
PredictorMatrix:
       age sex hgb creat mspike ptime pstat
age      0   0   0     0      0     0     0
sex      0   0   0     0      0     0     0
hgb      1   1   0     1      1     1     1
creat    1   1   1     0      1     1     1
mspike   1   1   1     1      0     1     1
ptime    0   0   0     0      0     0     0
pstat    0   0   0     0      0     0     0
Random generator seed value:  NA
>
> # Verify imputed data is valid based on distribution of original data points
> xyplot(mgusImputed, creat ~ sex+age+hgb+mspike+ptime+pstat,pch=18,cex=1)
> xyplot(mgusImputed, hgb ~ sex+age+creat+mspike+ptime+pstat,pch=18,cex=1)
> xyplot(mgusImputed, mspike ~ sex+age+hgb+creat+ptime+pstat,pch=18,cex=1)
>
> densityplot(mgusImputed)
>
> stripplot(mgusImputed, pch = 20, cex = 1.2)
>
> ###################### Build initial KM models ###################################
>
> #fit survival curve without explanatory variables
> mgus.fit<-with(mgusImputed, survfit(Surv(ptime, pstat)~1, error="greenwood",
conf.type="log-log", conf.int=0.95))
> summary(mgus.fit,times=c(81,240,373))

 ## summary of imputation 1 :
Call: survfit(formula = Surv(ptime, pstat) ~ 1, error = "greenwood",
    conf.type = "log-log", conf.int = 0.95)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
```

```
  81    697      64    0.937 0.00781       0.9195       0.950
 240     57      46    0.790 0.02679       0.7320       0.838
 373      3       5    0.383 0.17496       0.0874       0.686

 ## summary of imputation 2 :
Call: survfit(formula = Surv(ptime, pstat) ~ 1, error = "greenwood",
    conf.type = "log-log", conf.int = 0.95)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  81    697      64    0.937 0.00781       0.9195       0.950
 240     57      46    0.790 0.02679       0.7320       0.838
 373      3       5    0.383 0.17496       0.0874       0.686

 ## summary of imputation 3 :
Call: survfit(formula = Surv(ptime, pstat) ~ 1, error = "greenwood",
    conf.type = "log-log", conf.int = 0.95)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  81    697      64    0.937 0.00781       0.9195       0.950
 240     57      46    0.790 0.02679       0.7320       0.838
 373      3       5    0.383 0.17496       0.0874       0.686

 ## summary of imputation 4 :
Call: survfit(formula = Surv(ptime, pstat) ~ 1, error = "greenwood",
    conf.type = "log-log", conf.int = 0.95)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  81    697      64    0.937 0.00781       0.9195       0.950
 240     57      46    0.790 0.02679       0.7320       0.838
 373      3       5    0.383 0.17496       0.0874       0.686

 ## summary of imputation 5 :
Call: survfit(formula = Surv(ptime, pstat) ~ 1, error = "greenwood",
    conf.type = "log-log", conf.int = 0.95)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  81    697      64    0.937 0.00781       0.9195       0.950
 240     57      46    0.790 0.02679       0.7320       0.838
 373      3       5    0.383 0.17496       0.0874       0.686
>
> #fit survival curve stratifying by gender
> mgus.sex.fit<-with(mgusImputed, survfit(Surv(ptime, pstat)~sex, error="greenwood",
conf.type="log-log", conf.int=0.95))

>
> #Plot the curves
```

```
> autoplot(mgus.fit$analyses[1], xlab="Time until progression to a PCM (months)",
ylab="Survival Probability")
> autoplot(mgus.sex.fit$analyses[1], xlab="Time until progression to a PCM (months)",
ylab="Survival Probability")
>
> #Perform the log-rank test with equal weights
> mgus.sex.test0<-with(mgusImputed, survdiff(Surv(ptime, pstat)~sex, rho=0))
> mgus.sex.test0
call :
with.mids(data = mgusImputed, expr = survdiff(Surv(ptime, pstat) ~
    sex, rho = 0))

call1 :
mice(data = mgus, m = 5, method = "pmm", maxit = 50)

nmis :
   age    sex    hgb  creat mspike  ptime  pstat
     0      0     13     30     11      0      0

analyses :
[[1]]
Call:
survdiff(formula = Surv(ptime, pstat) ~ sex, rho = 0)

        N Observed Expected (O-E)^2/E (O-E)^2/V
sex=F 631       59     57.3    0.0501     0.101
sex=M 753       56     57.7    0.0498     0.101

 Chisq= 0.1  on 1 degrees of freedom, p= 0.751

[[2]]
Call:
survdiff(formula = Surv(ptime, pstat) ~ sex, rho = 0)

        N Observed Expected (O-E)^2/E (O-E)^2/V
sex=F 631       59     57.3    0.0501     0.101
sex=M 753       56     57.7    0.0498     0.101

 Chisq= 0.1  on 1 degrees of freedom, p= 0.751

[[3]]
Call:
survdiff(formula = Surv(ptime, pstat) ~ sex, rho = 0)

        N Observed Expected (O-E)^2/E (O-E)^2/V
sex=F 631       59     57.3    0.0501     0.101
```

```
sex=M 753       56     57.7    0.0498     0.101

 Chisq= 0.1  on 1 degrees of freedom, p= 0.751

[[4]]
Call:
survdiff(formula = Surv(ptime, pstat) ~ sex, rho = 0)

       N Observed Expected (O-E)^2/E (O-E)^2/V
sex=F 631       59     57.3    0.0501     0.101
sex=M 753       56     57.7    0.0498     0.101

 Chisq= 0.1  on 1 degrees of freedom, p= 0.751

[[5]]
Call:
survdiff(formula = Surv(ptime, pstat) ~ sex, rho = 0)

       N Observed Expected (O-E)^2/E (O-E)^2/V
sex=F 631       59     57.3    0.0501     0.101
sex=M 753       56     57.7    0.0498     0.101

 Chisq= 0.1  on 1 degrees of freedom, p= 0.751

> ########## Visually inspect PH assumptions for categorical variables ##################
>
> #Plot a km model with imputed data for gender
> plot(mgus.sex.fit$analyses[[1]]$time, log(-log(mgus.sex.fit$analyses[[1]]$surv)),
+ xlab="Time (months)", ylab="Log-Cumulative Hazard Function for Sex",type="l",lty=1:2)
> legend(350, -5.5, legend=levels(mgus$sex), lty=1:2)
>
>
> ################### Build initial model #####################################
>
> #Fit coxph model with imputed data sets and pool the results
> mgus.ph.initial <- with(mgusImputed, coxph(Surv(ptime, pstat)~sex+age+hgb+creat+mspike))
> mgus.ph.initial.pooled <- pool(mgus.ph.initial)
Warning message:
In mice.df(m, lambda, dfcom, method) : Large sample assumed.
> summary(mgus.ph.initial.pooled)
              est         se          t       df     Pr(>|t|)       lo 95
sex2    0.18699521 0.203543739  0.9186979 778730.84 3.582539e-01 -0.21194381
age     0.01221687 0.008397618  1.4548019 942740.97 1.457245e-01 -0.00424218
hgb    -0.14442317 0.054723873 -2.6391255  73016.41 8.313786e-03 -0.25168177
creat  -0.17799800 0.200273455 -0.8887748 503601.76 3.741245e-01 -0.57052771
mspike  0.86187307 0.163884672  5.2590219 973164.39 1.448541e-07  0.54066462
```

```
            hi 95 nmis          fmi       lambda
sex2    0.58593422   NA 0.0010666567 0.0010640912
age     0.02867592    0 0.0004930079 0.0004908875
hgb    -0.03716457   13 0.0071513433 0.0071241485
creat   0.21453170   30 0.0019876020 0.0019836385
mspike  1.18308153   11 0.0003321599 0.0003301055
>
> ######################################## Check for zero correlation with time
############################################################
>
> #Test each covariate for zero correlation between the time points and the associated sequence
of estimates for the regression coefficient
> tmp <- with(mgusImputed, cox.zph(coxph(Surv(ptime, pstat)~sex+age+hgb+creat+mspike)))
Warning messages:
1: contrasts dropped from factor sex
2: contrasts dropped from factor sex
3: contrasts dropped from factor sex
4: contrasts dropped from factor sex
5: contrasts dropped from factor sex
> tmp
call :
with.mids(data = mgusImputed, expr = cox.zph(coxph(Surv(ptime,
    pstat) ~ sex + age + hgb + creat + mspike)))

call1 :
mice(data = mgus, m = 5, method = "pmm", maxit = 50)

nmis :
   age    sex    hgb  creat mspike  ptime  pstat
     0      0     13     30     11      0      0

analyses :
[[1]]
           rho   chisq     p
sex2    0.06714 0.50179 0.479
age    -0.16737 2.24029 0.134
hgb    -0.08754 0.93951 0.332
creat  -0.01796 0.03635 0.849
mspike -0.00839 0.00843 0.927
GLOBAL       NA 2.99564 0.701

[[2]]
           rho   chisq     p
sex2    0.0756 0.62976 0.427
age    -0.1687 2.27112 0.132
hgb    -0.0961 1.11397 0.291
```

```
creat  -0.0532 0.31102 0.577
mspike -0.0041 0.00201 0.964
GLOBAL      NA 3.42065 0.635

[[3]]
          rho   chisq     p
sex2    0.07012 0.53775 0.463
age    -0.16720 2.22429 0.136
hgb    -0.08757 0.95199 0.329
creat  -0.02686 0.06921 0.792
mspike -0.00749 0.00668 0.935
GLOBAL       NA 3.02583 0.696

[[4]]
          rho   chisq     p
sex2    0.06699 0.49572 0.481
age    -0.17022 2.31217 0.128
hgb    -0.09029 0.98930 0.320
creat  -0.01416 0.01989 0.888
mspike -0.00593 0.00421 0.948
GLOBAL       NA 3.07257 0.689

[[5]]
          rho  chisq     p
sex2    0.06885 0.5255 0.469
age    -0.16025 2.0413 0.153
hgb    -0.08075 0.8039 0.370
creat  -0.02778 0.0849 0.771
mspike -0.00912 0.0099 0.921
GLOBAL       NA 2.7990 0.731

>
> # Plot Schoenfeld test for each covariate
> ggcoxzph(tmp$analyses[[1]])
>
> #Check martingale residuals for each covariate
> m.resid<-with(mgusImputed, resid(coxph(Surv(ptime, pstat)~sex+age+hgb+creat+mspike),
"mart"))
> par(mfrow=c(1,1))
> plot(m.resid$analyses[[1]]~complete(mgusImputed,1)$sex, xlab="Sex", ylab="Martingale
residuals")
> plot(m.resid$analyses[[1]]~complete(mgusImputed,1)$age, xlab="Age", ylab="Martingale
residuals")
> lines(lowess(complete(mgusImputed,1)$age, m.resid$analyses[[1]]))
> plot(m.resid$analyses[[1]]~complete(mgusImputed,1)$creat, xlab="Creatinine Levels",
ylab="Martingale residuals")
```

```
> lines(lowess(complete(mgusImputed,1)$creat, m.resid$analyses[[1]]))
> plot(m.resid$analyses[[1]]~complete(mgusImputed,1)$hgb, xlab="Hgb Levels",
ylab="Martingale residuals")
> lines(lowess(complete(mgusImputed,1)$hgb, m.resid$analyses[[1]]))
> plot(m.resid$analyses[[1]]~complete(mgusImputed,1)$mspike, xlab="Monoclonal Serum
Levels", ylab="Martingale residuals")
> lines(lowess(complete(mgusImputed,1)$mspike, m.resid$analyses[[1]]))
>
> # Check Schoenfeld residuals for each covariate
> s.resid<-with(mgusImputed, resid(coxph(Surv(ptime, pstat)~sex+age+hgb+creat+mspike),
"scho"))
Warning messages:
1: contrasts dropped from factor sex
2: contrasts dropped from factor sex
3: contrasts dropped from factor sex
4: contrasts dropped from factor sex
5: contrasts dropped from factor sex
> par(mfrow=c(1,1))
> plot(as.numeric(dimnames(s.resid$analyses[[1]])[[1]]), s.resid$analyses[[1]][,1],
xlab="Observed Failure Time", ylab="Schoenfeld Residuals for sex")
> plot(as.numeric(dimnames(s.resid$analyses[[1]])[[1]]), s.resid$analyses[[1]][,2],
xlab="Observed Failure Time", ylab="Schoenfeld Residuals for age")
> plot(as.numeric(dimnames(s.resid$analyses[[1]])[[1]]), s.resid$analyses[[1]][,3],
xlab="Observed Failure Time", ylab="Schoenfeld Residuals for hgb")
> plot(as.numeric(dimnames(s.resid$analyses[[1]])[[1]]), s.resid$analyses[[1]][,4],
xlab="Observed Failure Time", ylab="Schoenfeld Residuals for creat")
> plot(as.numeric(dimnames(s.resid$analyses[[1]])[[1]]), s.resid$analyses[[1]][,5],
xlab="Observed Failure Time", ylab="Schoenfeld Residuals for mspike")
>
> # Check for influential outliers
> ggcoxdiagnostics(mgus.ph.initial$analyses[[1]], type = "dfbeta", linear.predictions = FALSE,
ggtheme = theme_bw())
Warning message:
contrasts dropped from factor sex
> ggcoxdiagnostics(mgus.ph.initial$analyses[[1]], type = "deviance", linear.predictions =
FALSE, ggtheme = theme_bw())
>
> ######################## Preform variable selection ###############################
> # Backwards stepwise variable selection
> mgus.ph.var <- with(mgusImputed, step(coxph(Surv(ptime,
pstat)~sex+age+hgb+creat+mspike)))
Start:  AIC=1416.83
Surv(ptime, pstat) ~ sex + age + hgb + creat + mspike

       Df    AIC
- sex    1 1415.7
```

```
- creat   1 1416.4
<none>      1416.8
- age     1 1417.0
- hgb     1 1421.2
- mspike  1 1441.5

Step:  AIC=1415.66
Surv(ptime, pstat) ~ age + hgb + creat + mspike

        Df    AIC
- creat   1 1414.8
<none>      1415.7
- age     1 1415.8
- hgb     1 1419.2
- mspike  1 1440.2

Step:  AIC=1414.82
Surv(ptime, pstat) ~ age + hgb + mspike

        Df    AIC
<none>      1414.8
- age     1 1415.0
- hgb     1 1417.7
- mspike  1 1439.9
Start:  AIC=1417.13
Surv(ptime, pstat) ~ sex + age + hgb + creat + mspike

        Df    AIC
- sex     1 1415.9
- creat   1 1416.6
<none>      1417.1
- age     1 1417.4
- hgb     1 1421.4
- mspike  1 1441.7

Step:  AIC=1415.89
Surv(ptime, pstat) ~ age + hgb + creat + mspike

        Df    AIC
- creat   1 1415.0
<none>      1415.9
- age     1 1416.1
- hgb     1 1419.4
- mspike  1 1440.3

Step:  AIC=1414.99
```

Surv(ptime, pstat) ~ age + hgb + mspike

	Df	AIC
<none>		1415.0
- age	1	1415.2
- hgb	1	1417.9
- mspike	1	1439.8

Start: AIC=1415.92
Surv(ptime, pstat) ~ sex + age + hgb + creat + mspike

	Df	AIC
- sex	1	1414.8
- creat	1	1415.3
<none>		1415.9
- age	1	1416.1
- hgb	1	1421.0
- mspike	1	1441.0

Step: AIC=1414.78
Surv(ptime, pstat) ~ age + hgb + creat + mspike

	Df	AIC
- creat	1	1413.8
<none>		1414.8
- age	1	1414.9
- hgb	1	1419.0
- mspike	1	1439.7

Step: AIC=1413.82
Surv(ptime, pstat) ~ age + hgb + mspike

	Df	AIC
<none>		1413.8
- age	1	1414.0
- hgb	1	1417.4
- mspike	1	1439.1

Start: AIC=1416.29
Surv(ptime, pstat) ~ sex + age + hgb + creat + mspike

	Df	AIC
- sex	1	1415.2
- creat	1	1415.7
<none>		1416.3
- age	1	1416.4
- hgb	1	1421.3
- mspike	1	1440.9

```
Step:  AIC=1415.19
Surv(ptime, pstat) ~ age + hgb + creat + mspike

        Df    AIC
- creat   1 1414.3
<none>      1415.2
- age     1 1415.3
- hgb     1 1419.3
- mspike  1 1439.7

Step:  AIC=1414.26
Surv(ptime, pstat) ~ age + hgb + mspike

        Df    AIC
<none>      1414.3
- age     1 1414.3
- hgb     1 1417.8
- mspike  1 1439.2
Start:  AIC=1416.51
Surv(ptime, pstat) ~ sex + age + hgb + creat + mspike

        Df    AIC
- sex     1 1415.4
- creat   1 1416.0
<none>      1416.5
- age     1 1416.7
- hgb     1 1421.1
- mspike  1 1441.2

Step:  AIC=1415.39
Surv(ptime, pstat) ~ age + hgb + creat + mspike

        Df    AIC
- creat   1 1414.5
<none>      1415.4
- age     1 1415.5
- hgb     1 1419.1
- mspike  1 1440.0

Step:  AIC=1414.5
Surv(ptime, pstat) ~ age + hgb + mspike

        Df    AIC
<none>      1414.5
- age     1 1414.7
```

```
- hgb     1 1417.6
- mspike  1 1439.5
> mgus.ph.var.pooled <- pool(mgus.ph.var)
Warning message:
In mice.df(m, lambda, dfcom, method) : Large sample assumed.
> summary(mgus.ph.var.pooled)
              est          se         t       df     Pr(>|t|)        lo 95
age     0.01205842 0.008356942  1.442923 967779.0 1.490425e-01 -0.004320902
hgb    -0.11866683 0.051249060 -2.315493 103701.5 2.058792e-02 -0.219114314
mspike  0.86836563 0.164281136  5.285851 963827.8 1.251487e-07  0.546380117
            hi 95 nmis          fmi       lambda
age     0.02843775    0 0.0003649855 0.0003629197
hgb    -0.01821935   13 0.0058969782 0.0058778060
mspike  1.19035114   11 0.0003875149 0.0003854406
>
> ############################################## Check Assumptions again for
selected variables ####################################
> #Test each covariate for zero correlation between the time points and the associated sequence
of estimates for the regression coefficient
> tmp <- with(mgusImputed, cox.zph(coxph(Surv(ptime, pstat)~age+hgb+mspike)))
> tmp
call :
with.mids(data = mgusImputed, expr = cox.zph(coxph(Surv(ptime,
    pstat) ~ age + hgb + mspike)))

call1 :
mice(data = mgus, m = 5, method = "pmm", maxit = 50)

nmis :
   age    sex    hgb  creat mspike  ptime  pstat
     0      0     13     30     11      0      0

analyses :
[[1]]
           rho  chisq     p
age    -0.1730 2.3166 0.128
hgb    -0.0621 0.4975 0.481
mspike -0.0111 0.0147 0.903
GLOBAL      NA 2.4613 0.482

[[2]]
            rho   chisq     p
age    -0.17938 2.49932 0.114
hgb    -0.06926 0.60927 0.435
mspike -0.00776 0.00722 0.932
GLOBAL      NA 2.69629 0.441
```

```
[[3]]
          rho  chisq     p
age    -0.1729 2.3226 0.128
hgb    -0.0612 0.4882 0.485
mspike -0.0107 0.0137 0.907
GLOBAL      NA 2.4681 0.481

[[4]]
           rho   chisq     p
age    -0.17546 2.39053 0.122
hgb    -0.06623 0.55961 0.454
mspike -0.00851 0.00871 0.926
GLOBAL      NA 2.56205 0.464

[[5]]
          rho  chisq     p
age    -0.1681 2.1771 0.140
hgb    -0.0543 0.3808 0.537
mspike -0.0127 0.0195 0.889
GLOBAL      NA 2.2769 0.517

> # Plot Schoenfeld test for each covariate
> ggcoxzph(tmp$analyses[[1]])
>
> #Check martingale residuals for each covariate
> m.resid<-with(mgusImputed, resid(coxph(Surv(ptime, pstat)~age+hgb+mspike), "mart"))
> par(mfrow=c(1,1))
> plot(m.resid$analyses[[1]]~complete(mgusImputed,1)$age, xlab="Age", ylab="Martingale
residuals")
> lines(lowess(complete(mgusImputed,1)$age, m.resid$analyses[[1]]))
> plot(m.resid$analyses[[1]]~complete(mgusImputed,1)$hgb, xlab="Hgb Levels",
ylab="Martingale residuals")
> lines(lowess(complete(mgusImputed,1)$hgb, m.resid$analyses[[1]]))
> plot(m.resid$analyses[[1]]~complete(mgusImputed,1)$mspike, xlab="Monoclonal Serum
Levels", ylab="Martingale residuals")
> lines(lowess(complete(mgusImputed,1)$mspike, m.resid$analyses[[1]]))
>
> # Check Schoenfeld residuals for each covariate
> s.resid<-with(mgusImputed, resid(coxph(Surv(ptime, pstat)~sex+age+hgb+creat+mspike),
"scho"))
Warning messages:
1: contrasts dropped from factor sex
2: contrasts dropped from factor sex
3: contrasts dropped from factor sex
4: contrasts dropped from factor sex
```

```
5: contrasts dropped from factor sex
> par(mfrow=c(1,1))
> plot(as.numeric(dimnames(s.resid$analyses[[1]])[[1]]), s.resid$analyses[[1]][,1],
xlab="Observed Failure Time", ylab="Schoenfeld Residuals for age")
> plot(as.numeric(dimnames(s.resid$analyses[[1]])[[1]]), s.resid$analyses[[1]][,2],
xlab="Observed Failure Time", ylab="Schoenfeld Residuals for hgb")
> plot(as.numeric(dimnames(s.resid$analyses[[1]])[[1]]), s.resid$analyses[[1]][,3],
xlab="Observed Failure Time", ylab="Schoenfeld Residuals for mspike")
>
>
> ################# Build final model ######################################
> #Build final model with imputations
> mgus.ph.final <- with(mgusImputed, coxph(Surv(ptime, pstat)~age+hgb+mspike))
> mgus.ph.final.pooled <- pool(mgus.ph.final)
Warning message:
In mice.df(m, lambda, dfcom, method) : Large sample assumed.
> summary(mgus.ph.final.pooled)
              est         se         t       df     Pr(>|t|)       lo 95
age     0.01205842 0.008356942  1.442923 967779.0 1.490425e-01 -0.004320902
hgb    -0.11866683 0.051249060 -2.315493 103701.5 2.058792e-02 -0.219114314
mspike  0.86836563 0.164281136  5.285851 963827.8 1.251487e-07  0.546380117
            hi 95 nmis          fmi       lambda
age     0.02843775    0 0.0003649855 0.0003629197
hgb    -0.01821935   13 0.0058969782 0.0058778060
mspike  1.19035114   11 0.0003875149 0.0003854406
> summary(survfit(mgus.ph.final$analyses[[1]]),times=c(81,240,373))
Call: survfit(formula = mgus.ph.final$analyses[[1]])

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   81    697      64    0.943  0.0076        0.928        0.958
  240     57      46    0.799  0.0289        0.744        0.858
  373      3       5    0.453  0.1621        0.225        0.914
> autoplot(survfit(mgus.ph.final$analyses[[1]]), xlab="Time until progression to a PCM
(months)", ylab="Survival Probability")
>
> #Build model without imputations
> no_imputations <- coxph(Surv(ptime, pstat)~age+hgb+mspike, data=mgus)
> no_imputations.fit <- survfit(no_imputations, error="greenwood", conf.type="log-log",
conf.int=0.95)
> summary(no_imputations.fit)
Call: survfit(formula = no_imputations, conf.int = 0.95, conf.type = "log-log",
    error = "greenwood")

 time n.risk n.event survival  std.err lower 95% CI upper 95% CI
    2   1317       2    0.999 0.000892        0.995        1.000
    4   1273       1    0.998 0.001110        0.994        0.999
```

5	1257	1	0.997	0.001298	0.993	0.999
6	1246	1	0.997	0.001466	0.992	0.999
8	1230	2	0.995	0.001766	0.990	0.998
9	1216	2	0.994	0.002031	0.988	0.997
10	1205	2	0.993	0.002272	0.986	0.996
11	1192	1	0.992	0.002386	0.986	0.995
12	1185	1	0.991	0.002497	0.985	0.995
13	1178	1	0.990	0.002605	0.984	0.994
14	1172	2	0.989	0.002812	0.982	0.993
16	1164	1	0.988	0.002912	0.981	0.993
17	1159	2	0.987	0.003106	0.979	0.992
21	1131	2	0.985	0.003300	0.977	0.991
22	1127	1	0.985	0.003394	0.976	0.990
23	1115	2	0.983	0.003579	0.974	0.989
29	1080	2	0.981	0.003770	0.972	0.988
30	1072	1	0.981	0.003864	0.971	0.987
33	1048	1	0.980	0.003960	0.970	0.986
34	1043	1	0.979	0.004055	0.969	0.986
35	1036	1	0.978	0.004148	0.968	0.985
36	1028	1	0.977	0.004241	0.967	0.984
38	1018	1	0.976	0.004334	0.966	0.984
39	1011	2	0.975	0.004518	0.964	0.982
42	996	2	0.973	0.004700	0.962	0.981
44	979	1	0.972	0.004791	0.961	0.980
45	972	1	0.971	0.004882	0.960	0.979
51	923	2	0.969	0.005081	0.958	0.978
52	920	2	0.967	0.005272	0.955	0.976
56	890	1	0.966	0.005372	0.954	0.976
57	884	2	0.964	0.005569	0.952	0.974
60	857	2	0.962	0.005773	0.949	0.972
61	848	1	0.961	0.005875	0.948	0.971
62	838	1	0.960	0.005976	0.947	0.971
63	832	1	0.959	0.006077	0.946	0.970
67	799	2	0.957	0.006292	0.943	0.968
69	784	1	0.956	0.006401	0.942	0.967
70	777	1	0.955	0.006510	0.940	0.966
73	753	1	0.954	0.006623	0.939	0.965
74	743	1	0.953	0.006738	0.937	0.964
76	732	2	0.950	0.006970	0.935	0.962
79	703	1	0.949	0.007094	0.933	0.961
80	697	2	0.946	0.007338	0.930	0.959
81	684	2	0.944	0.007581	0.927	0.957
83	667	1	0.943	0.007705	0.925	0.956
84	662	1	0.941	0.007829	0.924	0.955
86	647	1	0.940	0.007957	0.922	0.954
90	625	2	0.937	0.008225	0.919	0.951

```
 91    613       1    0.936 0.008361        0.917        0.950
 93    601       1    0.934 0.008499        0.915        0.949
 97    570       1    0.933 0.008650        0.913        0.948
 98    559       1    0.931 0.008805        0.912        0.946
101    532       4    0.924 0.009458        0.904        0.941
102    522       1    0.923 0.009618        0.902        0.940
109    478       1    0.921 0.009808        0.899        0.938
111    467       1    0.919 0.010004        0.897        0.937
114    451       1    0.917 0.010211        0.895        0.935
116    438       1    0.915 0.010427        0.892        0.933
118    424       1    0.913 0.010652        0.890        0.932
121    416       2    0.909 0.011110        0.884        0.928
123    402       1    0.906 0.011349        0.881        0.926
124    401       1    0.904 0.011584        0.879        0.924
128    385       2    0.899 0.012082        0.873        0.921
135    357       1    0.897 0.012364        0.870        0.919
138    338       1    0.894 0.012675        0.866        0.917
142    319       1    0.891 0.013014        0.863        0.914
147    298       1    0.888 0.013389        0.859        0.912
150    285       1    0.885 0.013783        0.855        0.909
152    278       1    0.882 0.014182        0.851        0.907
153    274       1    0.879 0.014579        0.847        0.904
158    258       2    0.872 0.015427        0.838        0.899
161    241       1    0.868 0.015875        0.834        0.896
165    231       1    0.865 0.016343        0.829        0.893
166    227       1    0.861 0.016810        0.824        0.891
168    217       1    0.857 0.017308        0.819        0.888
179    182       1    0.852 0.018019        0.813        0.884
180    177       1    0.848 0.018728        0.807        0.880
188    156       1    0.842 0.019568        0.799        0.877
190    153       1    0.837 0.020390        0.792        0.873
198    127       1    0.831 0.021476        0.784        0.868
201    120       1    0.824 0.022572        0.775        0.864
228     67       1    0.813 0.025468        0.757        0.857
238     57       1    0.800 0.028956        0.736        0.850
259     36       1    0.778 0.036282        0.697        0.840
275     25       1    0.746 0.048484        0.636        0.827
312     14       1    0.695 0.068551        0.538        0.807
340      6       1    0.611 0.102006        0.384        0.776
373      3       1    0.456 0.159259        0.153        0.720
> summary(no_imputations.fit, times=c(81,240,373))
Call: survfit(formula = no_imputations, conf.int = 0.95, conf.type = "log-log",
    error = "greenwood")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   81    684      63    0.944 0.00758        0.927        0.957
```

```
  240     56      46    0.800 0.02896        0.736        0.850
  373      3       5    0.456 0.15926        0.153        0.720
> summary(no_imputations)
Call:
coxph(formula = Surv(ptime, pstat) ~ age + hgb + mspike, data = mgus)

  n= 1360, number of events= 114
   (24 observations deleted due to missingness)

            coef exp(coef)  se(coef)      z Pr(>|z|)
age     0.011411  1.011476  0.008434  1.353   0.1761
hgb    -0.117460  0.889176  0.051317 -2.289   0.0221 *
mspike  0.908083  2.479565  0.165367  5.491 3.99e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

       exp(coef) exp(-coef) lower .95 upper .95
age       1.0115     0.9887    0.9949    1.0283
hgb       0.8892     1.1246    0.8041    0.9833
mspike    2.4796     0.4033    1.7931    3.4288

Concordance= 0.668  (se = 0.031 )
Rsquare= 0.027   (max possible= 0.65 )
Likelihood ratio test= 37.27  on 3 df,   p=4.041e-08
Wald test            = 39.31  on 3 df,   p=1.495e-08
Score (logrank) test = 39.75  on 3 df,   p=1.206e-08

>
> autoplot(no_imputations.fit, xlab="Time until progression to a PCM (months)",
ylab="Survival Probability")
>
```