STAT5291: Advanced Data Analysis
Final Project

Survival Analysis :
Evidence from Lung Cancer

Rui Cao, Xiaoyi Li, Yi Chen, Yunfan Li

Columbia University

Summary

The goal of this project is to study the effects of 7 potential prognostic factors on the survival of advanced lung cancer patients. They are age, sex, Karnofsky performance score rated by physician, Karnofsky performance score rated by patient, ECOG performance score, calories consumed at meals and weight loss in last six months.

The analysis shows that sex and ECOG performance score are two significant factors that affect the survival of advanced lung cancer patients. Specifically, if we hold ECOG performance score, male patients with advanced lung cancer are about 1.67 times more likely to die than female patients with advanced lung cancer. If we hold sex factor fixed, then increment ECOG performance score by one, the patient becomes 1.62 times more likely to die than before.

1. Introduction

As one of the common types of cancer, lung cancer leads to millions of deaths every year. According to the table of age-standardized net survival for patients diagnosed with lung cancer during 2010-2011 in England and Wales prepared by Cancer Research UK, 32% of adults survive lung cancer for at least one year and it then falls to 9.5% for five years or more and 5% for 10 years or more[1].

Table 1: Lung Cancer Survival Probability Summary

		1-Year Survival (%)	5-Year Survival (%)	10-Year Survival (%)
Men	Net Survival	30.4%	8.4%	4.0%
	95% LCL	30.1%	7.5%	2.8%
	95% UCL	30.7%	9.3%	5.5%
Women	Net Survival	35.1%	11.6%	6.5%
	95% LCL	34.8%	10.5%	4.9%
	95% UCL	35.3%	12.6%	8.4%
Adults	Net Survival	32.1%	9.5%	4.9%
	95% LCL	31.9%	8.8%	3.9%
	95% UCL	32.3%	10.2%	6.1%

Data Source: UK Cancer Research

Based on the definition of lung cancer staging provided by Cancer Research UK, stage four A and B with symptoms like cancer in both lungs and cancer spread to one or more organs is normally called advanced lung cancer[2]. Several factors could affect the survival of advanced lung cancer patients and our project uses a advanced lung cancer data set to analyze which ones are potentially significant and to what extent they would affect the survival of patients.

The data set used for this project comes from a study conducted by the North Central Cancer Treatment Group. The study was originally presented and analyzed in Lprinzi et al.(1994). The data were obtained from patients completed questionnaires and there were over a thousand of advanced lung cancer patients involved in this study[3]. The data set records the survival status of patients with advanced lung cancer and several performance features of patients measured by either physicians and by the patients themselves that could provide prognostic information, including age, sex, ECOG performance score, Karnofsky performance score rated by physician, Karnofsky performance score rated by patient, calories consumed at meals and weight loss in last six months.

2. Data Description

This dataset has 167 observations. Each one represents a patient with advanced lung cancer from the North Central Cancer Treatment Group. There are 10 independent variables (including survival time and condition of censor). We put more attention on ”ph.ecog”, ” ph.karno” and ”pat.karno” because they are hard to understand for non-medical people.

- **Data Introduction**

The basic decription about the data we can be seen as follow table:

Table2: Data Description

Variable name	Understanding	Type and Range
ph.ecog	a index of performance status rated by physician	Categorical variable “0”(very good) to “5”(dead)
ph.karno	Karnofsky performance score rated by physician	Categorical variable “0”(death) to “100”(very good)
pat.karno	Karnofsky performance score rated by patient	Categorical variable “0” (dead) to “100”(very good)
sex	Women/Men	Categorical variable Male:”1” Female: “2”
meal.cal	Calories consumed at each meals	Continuous variable Min:96 to Max:2600
wt.loss	Weight loss in last six months	Continuous variable Min:-24 to Max:68
Age	Age in years	Contiunous Variable Min:39 to Max:82

Data Source: North Central Cancer Treatment Group

Here are the seven variables which may have the influence on the survival time for a lung cancer patient. We will discuss the effect of these variables later in the section analysis.

1. “ph.ecog” stands for ECOG performance score. It is a index of performance status and can be used to evaluate the effectiveness of therapies. It has 6 levels from 0 to 5. “0” represents fully active, able to carry on all performance without restriction. “1” represents restricted in physically strenuous activities but able to carry out work of a light nature such as light house work. “2” represent ambulatory and capable of all selfcare but unable to carry out any work activities. “3” is even worse, capable of only limited selfcare and confined to bed or chair. “4” means completely disabled. “5” is dead. So the degree of severity is deeper from “0” to “5”.
2. “ph.karno” stands for Karnofsky performance score rated by physician. It is another widely used method to assess the functional status of a patient. Every level increases by 10 numbers, so it has 10 levels from “0” to “100”. Higher scores of “ph.karno” mean higher healthy status. They are rated by physician. ”pat.karno” is totally as same as “ph.karno”.The only difference is it is rated by patient themselves.
3. “sex” is a categorical variable in the dataset. It has two levels. “1” represents male and “2” represents female. The other variables are “age”(ages in years), “meal.cal”(calories consumed at meals), “wt.loss”(weight loss in last six months), “time”(survival times in days) and “status” (1=censored, 2=dead).

- **Explory Data Analysis**

In this part, we explored the data through descriptive statistics analysis.The pie chart shows that 63% of observation are male and 37% of observation are female.

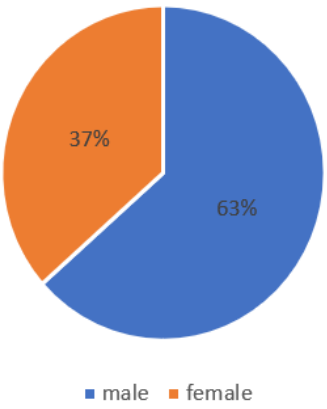


Fig 1: Ratio Of Sex

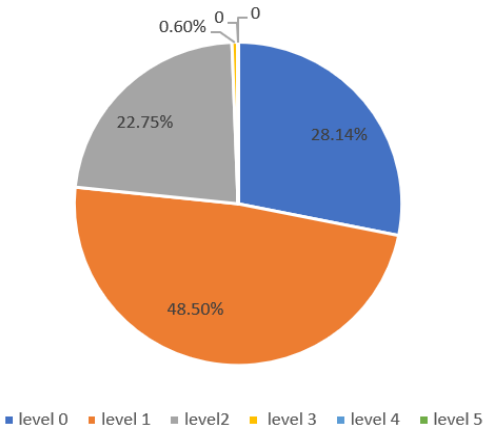


Fig 2: histogram of ph.ecog

We also analyzed the variable “ph.ecog”. 28.14% of patients are in Level 0, 48.5% of patients are in Level 1, 22.75% of patients are in Level 2, 0.6% of patients are in Level 3 and nobody in Level 4 or 5.

Clearly, among all the patients, the outcome is acceptable. Higher levels represent bad conditions.So we will put more attention on survival patient with good conditions. Then from the histogram over half of the patient can carry out work of light nature. And no one is completely disabled or died after the treatment.

The following histogram shows the number of death influenced by “ph.ecog”. The blue represents the number of death. Although the number of patients in each level is different we can see the rate of death has an increasing trends with the level of “ph.ecog”. Of those who were in Level 0 of ECOG, 42.55% of patient was censored and 57.44% of patient was dead. Of those who were in Level 1, 27.16% of patient was censored and 72.84% of patient was dead. Of those who were in Level 2, 13.16% of patient was censored and 86.84% of patient was dead. Because only 1 patient in level 3. So the death rate is 100%. It shows that patients in high level of ECOG has a higher death rate.

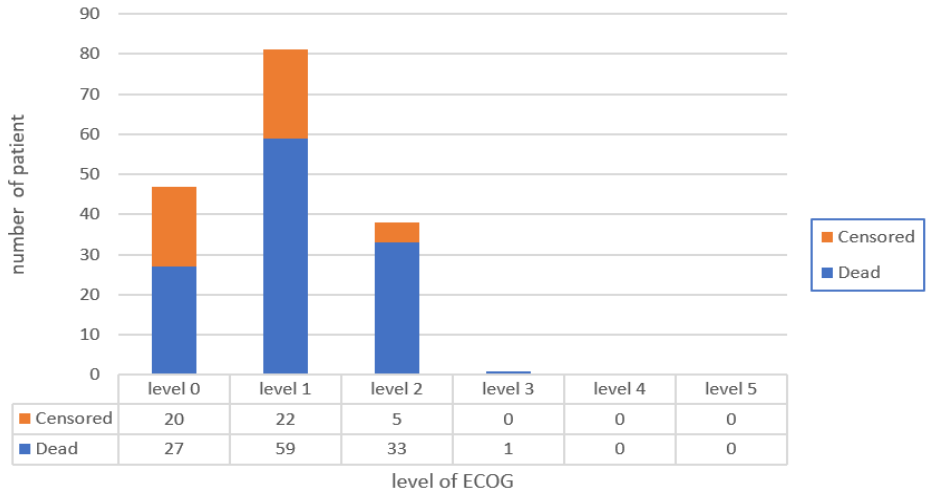


Fig 3: Number of death influenced by ph.ecog

This graph shows the histogram of “physician.Karno” and “Patient.Karno”.The variance for patient karno are more bigger. It shows that patients are more pessimistic when they get cancer. They usually rated themselves more serious than doctor think. But the general trends are the same. Most of patients are in the level 80 and level 90.

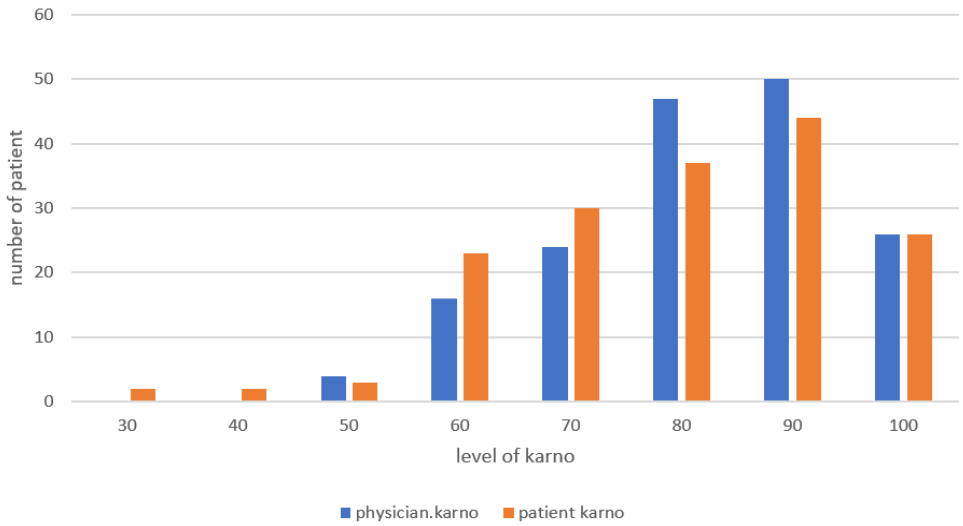


Fig 4: Number of patients in each level of Karno

The correlation matrix represents the correlation between each independent variables and dependent variable. there are three interesting features. First,among 7 independent variables,we judged that maybe only “pat.karno”,”ph.karno”,”ph.ecog” and “sex” are significant. We can see only this four number in the red circle are near 0.1 or greater than 0.1. The other variables such as “wt.loss” and “meal.col” are so smaller. So they are weak correlate to the dependent variable. Second, “Ph.ecog” and “Ph.karno” are highly negative correlated, which is reasonable since they are both the performance scores for the patient and the only difference is the range of level and direction. Last, The correlation coefficient between “pat.karno” and“ph.karno” are 0.54 bigger than 0.5. Also highly correlate with each other. So we can make a hypothesis that we can only keep one of them at the end of the model because “pat.karno”,“ph.karno”and “Ph.ecog” are highly correlate with each other. We will use the model to prove this later.

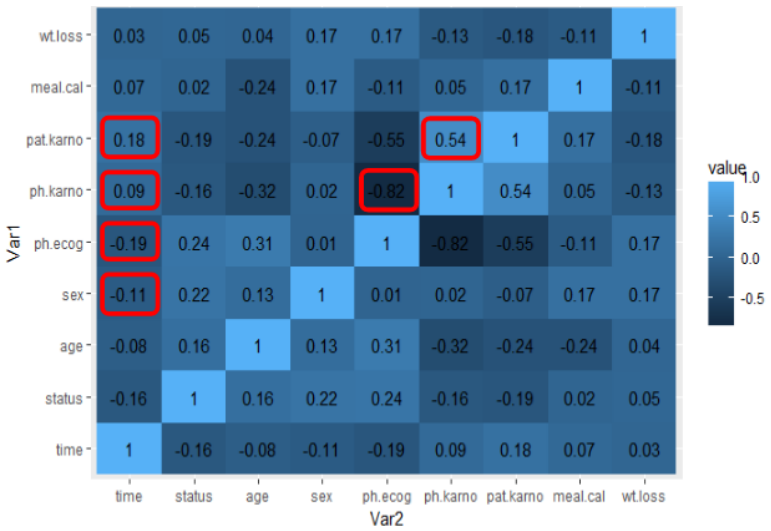
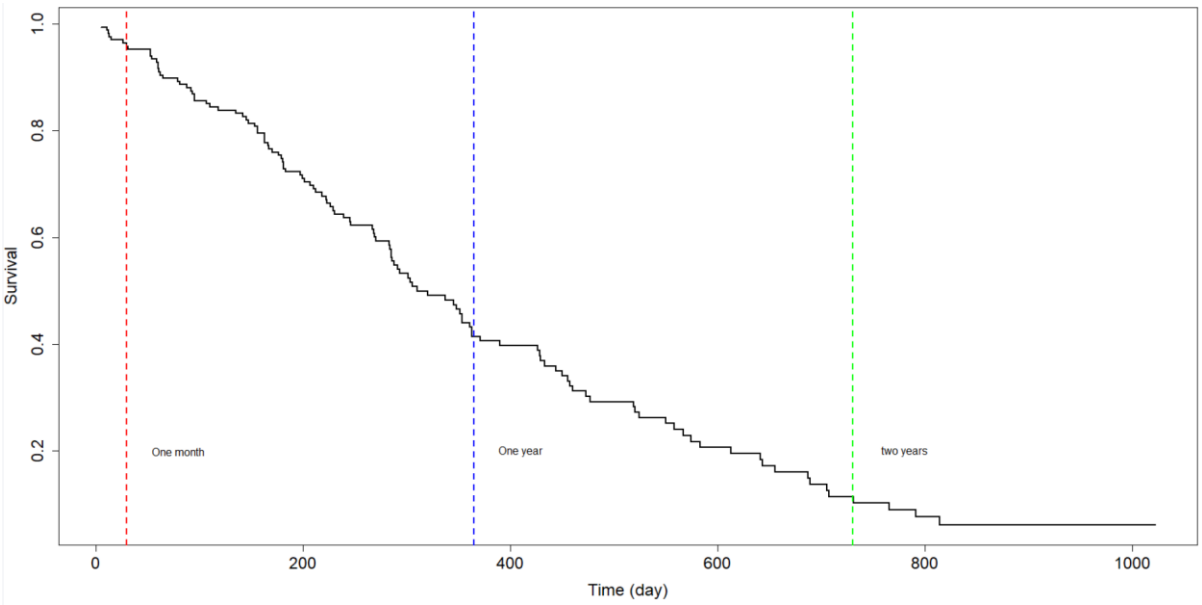


Fig 5: Correlation Materix

3. Analysis

In this research, we use the Cox Proportional Hazard Model (PML). PML is based on the regression approach to make the analysis on the survival time. This model has the strength in inference. A total of 165 patients were entered onto this study. As we described before, we explored the effect of 7 potential factors which may influence the survival time of the lung cancer patient. Finally, we use the backward model selection method to get the final model. (More detail information can be seen in the appendix).

Based on our data, we draw the survival function plot of the lung patient. It is a decreasing function. Over 95% of patients can live longer or equal to one month. But this value decrease dramatically to about 40% for one year. And less than 5% of the patient can live longer or equal to two years. This just once again prove how dangerous the lung cancer it is.



Data Source: North Central Cancer Treatment Group

Fig 6: Survival Function Plot

The analysis showed that only sex and ph.ecog have the significant effect on the survival time of a lung cancer patient. While all the other factors, including ph.karno, kat.karno, age, meal.cal and wt.loss don't have significant effect on the survival time.

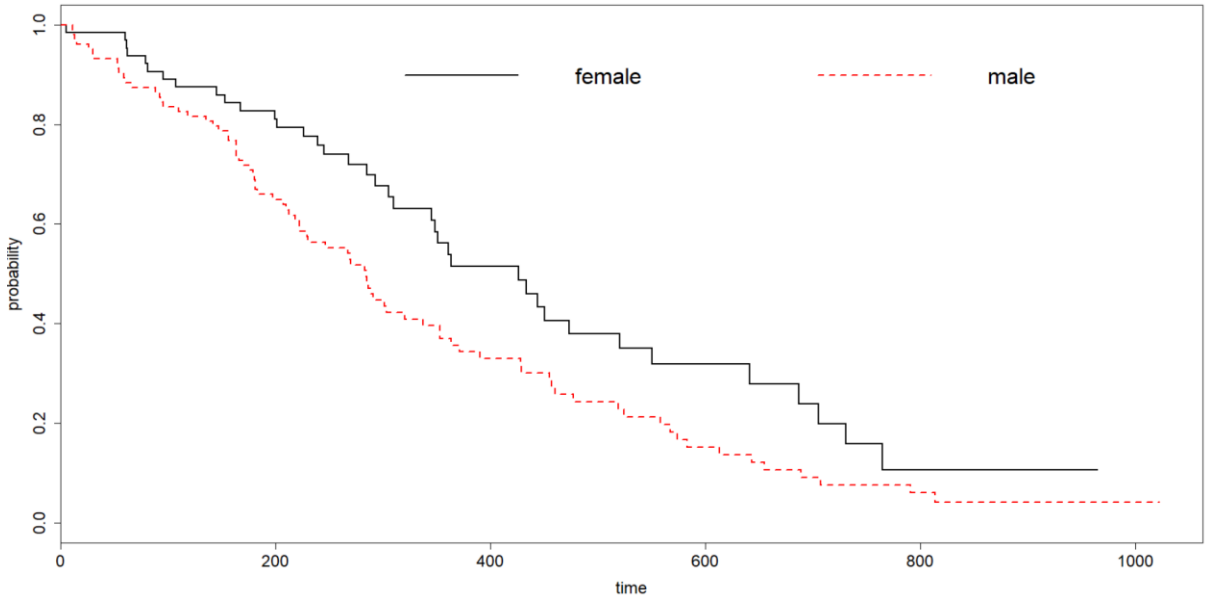
Specifically, the effect of sex and ph.ecog can be summarized as follows:

1. If we fixed all the other variables (i.e. ph.ecog) to a constant, at any instant in time, a male patient would be approximately 1.665 times more likely to die than a female patient.
2. If we fixed all the other variables (i.e. sex) to be the same, if the ECOG performance increase 1 unit, at any instant in time, the patient would be approximately 1.62 times more likely to die.

In addition, we can be 95% confidence to conclude that:

1. If we fixed all the other variables (i.e. ph.ecog) to a constant, at any instant in time, a male patient would be approximately between 1.132 to 2.450 times more likely to die than a female patient.
2. If we fixed all the other variables (i.e. sex) to the same, if the ECOG performance increase 1 unit, at any instant in time, the patient would be approximately between 1.250 to 2.10 times more likely to die.

As we have already know that the sex has the significant effect on the survival time. Here we provide the survival function plot on different sex:



Data Source: North Central Cancer Treatment Group

Fig 6: Survival Function Plot Based on Different Sex

As we can see from the plot above. For a given time, the female have a higher probability to survive. Or in other words, for a given survive probability, the female tends to live longer than a male.

According to the Dr Robot from Harvard Medical Center, there are three possible reasons why sex can be influential to survival time[3]:

1. Males face more risks. For example, males have more unhealthy habit, including smoke (harmful to lung) and drinking.
2. Males have less social connection and are less willing to share their ideas and feeling with their family and friends.
3. Males are more likely to avoid the help from the doctors. And when they have physical problem they rely more on themselves which may make them miss the best treatment time.

Here we also give the reasons why ph.ecog can be influential to survival time but other physical condition assessment indexes (i.e. ph.karno and kat.karno) not: Even though, ph.ecog, ph.karno and kat.karno each alone has clear inference on the survival time, all of them are the assessment of the physical condition of the patient and they are correlated with each other. Ph.ecog is the one have the highest effect on the survival time. Given that ph.ecog has already in the model, the effect from ph.karno and kat.karno will not be significant anymore. If we include all these variables in our model, this will lead to the problem of multiple-collinearity.

4. Conclusion

In the project we studied the effect of different potential prognostic variables to the survival time of advanced lung cancer patients. Based on a dataset of 167 observations, we use the survival model to analyze the effects of 7 different prognostic variables.

First, we introduce the background of the lung cancer and data set that we used in this research. Utilized the descriptive statistical analysis, we summarize the basic characteristics of the data. And we use the correlation matrix to get an overall idea of the association among the variables. We find that "ph.ecog" and "ph.karno" are negatively and strongly correlated. Then, we use the Kaplan-Meier estimator and Cox proportional hazard model to select the variables of final model and quantify the relationship between the predictors and response variable.

After applying the method above, we choose "ph.ecog" and "sex" as our predictors. Holding "ph.ecog" fixed, the hazard rate of a male patient with advanced lung cancer dies is about 1.67 times the hazard rate of a female patient with advanced lung cancer dies at any instant time. Holding "sex" fixed, a patient with advanced lung cancer will be 1.62 times more likely to die for every incremental unit of Karnofsky performance score rated by patients.

The advantage of this model is that only two prognostic variables are included, which makes it easier and more straightforward to interpret. However, the test results might be biased due to the small sample size in the dataset. To improve the accuracy of this model, we may collect more data from different institutions. Since our sample size is too small and all the data are from the same institute, our final model can be biased. Besides, more information can be collected on each individual. Other prognostic variables such as race, region might also have a significant effect on the survival time. For example, people in the Africa and America have a higher rate of lung cancer than people from Asian area. Including more variables in our dataset will definitely improve the accuracy of final model.

Reference

[1] Cancer Research UK, Accessed [04] [2018]

www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/survival#heading-Zero.

[2]Cancer Research UK, Accessed [04] [2018]

http://www.cancerresearchuk.org/about-cancer/lung-cancer/stages-types-grades/stage-4

[3] Lprinzi et al.(1994). Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology, 12(3), 601-607.

[4] Robert H. Shmerling (2016): Why Men Often Die Earlier Than Women, Harvard Health Publishing, www.health.harvard.edu/blog/why-men-often-die-earlier-than-women-201602199137

Appendix

1. final model and results

In this research, we use the Cox Proportional Hazard Model (phm) to find the influential factors for the survival time of a lung cancer patient. We used the backward model selection method to get our final model. Here is the our result:

$$\lambda(t, \mathbf{x}) = \lambda_0(t) e^{0.5101\, Sex + 0.4825\, Ph.ecog}$$

Se: (0.1969) (0.1323)

P value: (0.009) (0.000)

Likelihood ratio test = 19.48 on 2 df, p = 5.882e − 05

Wald test = 19.35 on 2 df, p = 6.297e − 05

Score (logrank)test = 19.62 on 2 df, p = 5.493e − 05

2. Statistical Testing

- **Z Test:**

As we can see from the table below: Sex and ph.ecog have a significant P value for the Z test, while other variables not. Thus, we can reject the hypothesis that the parameter of these two variables are equals to zero. In other words, these two parameters are influential while others not. Even though pat.karno also have relatively high significance level but as we mentioned before we will not keep it in the final model in avoid of the problem of multiple-collinearity.

Table 3: Result of Z Test

	Coef	Se	Z	Pr(>|Z|)
Sex	0.5536	0.2016	2.746	0.0060**
ph.ecog	0.7395	0.2250	3.287	0.0010**
ph.karno	0.2244	0.0116	0.931	0.3516
pat.karno	-0.1207	0.0112	1.998	0.0457*
age	0.0108	0.0081	-1.488	0.1368
meal.cal	0.0002	0.0002	0.109	0.9129
wt.loss	-0.0014	0.0076	-1.828	0.0674

- **Likelihood Ratio Test/ Wald Test and Score Test:**

As we can see from the result of the model, all of these three test have the similar target that is to test whether the parameter of the other parameter are all equal to zero at the same time. The corresponding p value of these three statistics are all less than 0.05 which means that the given sex and ph.ecog in the model, the other variables indeed are not influential to the survival time.

3. Code

```
# read and clean the data
library(survival)
data <- survival::cancer
data <- na.omit(data)
status <- as.factor(data$status)
time <- data$time
```

```
data$sex[data$sex == 2] = 0
sex <- as.factor(data$sex)

km <- survfit(Surv(time,status)~1,data=data,type="kaplan-meier")
# survival funcion plot
plot(km$time,km$surv, type="s",xlab="Time (day)",ylab="Survival",main="survival function")
text(80,0.2,"One month")
text(410,0.2,"One year")
text(780,0.2,"two years")
abline(v=30, lty=2, lwd=2, col="red")
abline(v=365, lty=2, lwd=2, col="blue")
abline(v=730, lty=2, lwd=2, col="green")
# survival function plot based on differenct sex
fit <- survfit(Surv(time,status)~sex,data=data)
plot(fit,lty=1:2,col=1:2,main = "Survival function plot for male patient and female patient",xlab =
"time",ylab = "probability",lwd=2,,cex.axis=1.5,cex.lab=1.5)
legend("topright",legend=
c("female","male"),lwd=2,lty=1:2,col=1:2,bty="n",ncol=2,cex=2,pt.cex=0.7)
# fit the model and test the results
fitphm <-
coxph(Surv(time,status)~sex+ph.ecog+age+ph.karno+pat.karno+meal.cal+wt.loss,data=data)
fitreduced <- coxph(Surv(time,status)~sex+ph.ecog,data = data)
predict(fitreduced,sex=0,ph.ecog=0)
summary(fitphm)
summary(fitreduced)
anova(fitphm,fitreduced)
```