# Data Analysis on Kidney Cancer Gene Expression Dataset

**Introduction**:

Kidney cancer or renal cell carcinoma (RCC) has become one of the 10 most common cancer types in developed countries [1]. According to research conducted in 2005 on examining 122 renal cell carcinoma and normal kidney samples [1], it has been found that 80 genes are capable for researchers to classify renal cell carcinoma into three types: clear cell (ccRCC), papillary cell (pRCC) and chromophobe cell (chRCC) with small error rate [1]. Therenal cell carcinoma and normal kidney samples used in the research are cDNA microarray, which are defined as a grid of DNA segments measured by scientist to find gene expression levels for a large number of genes. The research has found that gene expression signatures are associated with tumor properties and patient variables [1]. For examples, metastasis formation which means metastases were present at the time of surgery is associated with gene expression. In addition, patient survival time after diagnosed of kidney cancer is another important factor associated with gene expression. Based on historic analysis on patients diagnosed with kidney cancer, males are twice likely suffering from kidney cancer than females [1]. Therefore, it is intriguing to conduct further analysis on association tests between subtypes of renal cell carcinoma, gender of patients, age of patients, metastasis status during surgery with patients' survival time since first diagnosis with kidney cancer.

**Data Description**:

The dataset used in this analysis is "Kidpack", which is a R package and can be downloaded in bioinformatics open source "Bioconductor" [2]. Holger Sultmann measured the dataset at the German Cancer Research Centre in 2002. In the experiment, 85 renal cancer cells are hybridized and only 74 good chips are selected based on its good quality. This dataset contains a processed datasets called "phenoData" which includes 5 samples information and it is used in this analysis. "PhenoData" contains five variables of our interest: survival time denoted as "survival", subtypes of renal cancer cell denoted as "subtype", patient's age denoted as "age", and patient's gender denoted as "gender" and metastasis status denoted as "m". In the first step, exploratory data analysis is performed on each of these five variables.

**1. EDA on Subtype**:

Subtype is a categorical variable. Among these 74 renal cancer cells, there are 52 cells are clear cells ("ccRCC"), 9 cells are chromophobe cells ("chRCC"), 13 cells are papillary cell ("pRCC") (Table 1). Therefore, Clear Cells and Papillary Cells are more common subtypes of kidney cancer.

| | Clear Cells | Chromophobe Cells | Papillary Cells |
|---|---|---|---|
| Frequency | 52 | 9 | 13 |
| Proportion | 70.27027% | 12.16216% | 17.56757% |

Table 1: Frequency and Proportion of Three Subtypes of Renal Cancer Cells.

**2. EDA on Survival Time**:
Survival time is a quantitative variable which ranges from 0 to 65 with mean 20.7 and median 16.5. Due to the fact that some patient loss to follow up and censored, survival time contains 14 missing values. The histogram is plotted to reveal its distribution (Figure 1). We can see it doesn't follow Normal distribution since it is asymmetry and there is right skewness. QQplot also shows right skewness since right top points are apart from qqline (Figure 2). Thus, for patients who diagnosed with kidney cancer, 50% of patients have survival time less than 17 years.

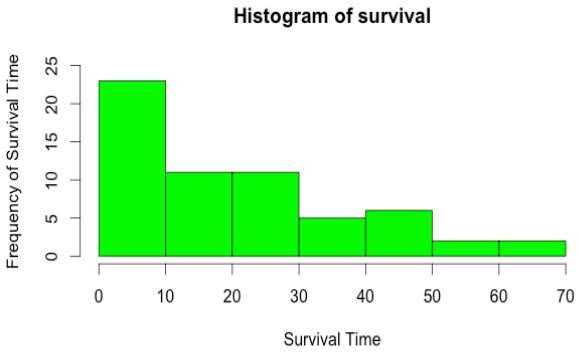


Figure 1: Histogram of Survival Time

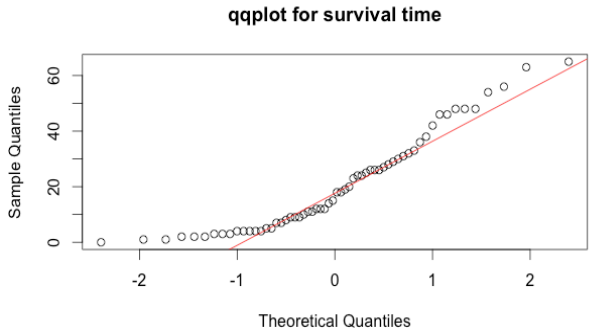


Figure 2: QQplot for Survival Time

**3. EDA on Patient's Age and Gender:**
Patient's age ranges from 26 to 85 with mean 59.77 and median 60. Boxplot shows symmetry and there is one outlier. Patient's gender are consists of 50 males and 24 females (Figure 3). It is consistent with early research on gender for patients with kidney cancer that males are twice likely to suffer from kidney cancer than females.

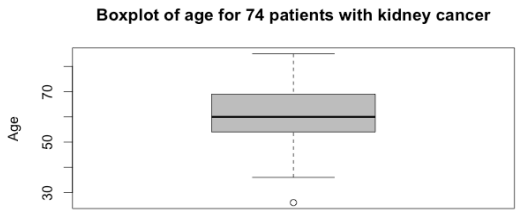


Figure 3: Boxplot for Patient's Age

**4. EDA on Metastasis Status**:
Metastasis is a binary variable. It is "1" if metastasis exists during kidney surgery, and "0" if metastasis does not exist. According to Table 2, there are 24 patients

who are free of metastasis and there are 23 patients who suffer from metastasis. Thus, it is equal likely to suffer from metastasis during surgery.
The relationships between survival time, cancer subtypes and ages are shown in Figure 4. This graph reveals some interesting facts: cancer type "ccRCC" is more common subtype for different age of patients and once diagnosed the survival time varies. Cancer type "pRCC" is more likely to occur for patients between age 45 to 85 and its survival time is either high (above 40 years) or either low (below 25). For cancer type "chRCC", patients with age around 60 are more like to have and once diagnosed their survival time is low (below 25)

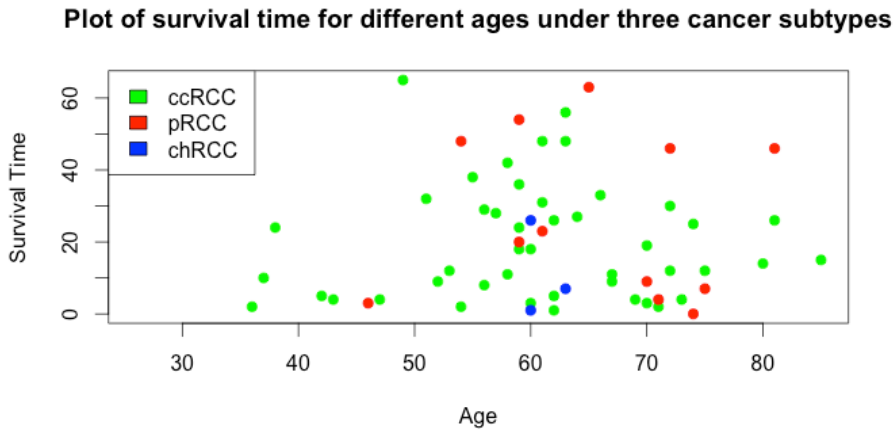


Figure 4: Plot of Survival Time for Different Ages under Three Cancer Types

**Statistical Modeling:**

**1. Multiple Linear Models**

In order to check if there is association between survival time with ages, gender, subtypes and metastasis. A multiple linear regression model is built. The output of significance test is given in figure 5.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.5984    25.2382   1.292   0.2041
age          -0.1792     0.2789  -0.643   0.5242
gender       -1.1924     6.0365  -0.198   0.8444
ccRCC         5.9085    17.6631   0.335   0.7398
pRCC         15.8833    18.6343   0.852   0.3992
m           -14.3082     5.3071  -2.696   0.0103 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5: Significance of Covariates Ages, Gender, Subtypes and Metastasis.

We can see only "metastasis status" is significant in the model (p-value 0.0103). Model can be assessed by residuals and fitted plot (figure 6). It can be seen from figure 6 that there is a pattern between residuals and fitted values. Thus, model should be improved.

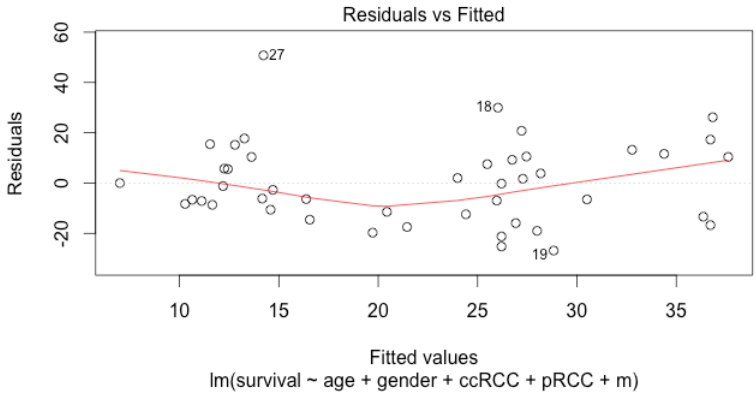


Figure 6: Residual vs Fitted values

**2. Proportional Hazard Model**

Since survival time is always positive and it does not follow Normal distribution, proportional hazard model is fitted [3, 4].

```
   (29 observations deleted due to missingness)

             coef  exp(coef)   se(coef)      z Pr(>|z|)
age     2.152e-02  1.022e+00  3.576e-02  0.602 0.547266
gender -1.614e-01  8.509e-01  5.310e-01 -0.304 0.761139
ccRCC   1.674e+01  1.863e+07  6.502e+03  0.003 0.997946
pRCC    1.685e+01  2.088e+07  6.502e+03  0.003 0.997932
m       3.057e+00  2.127e+01  8.529e-01  3.585 0.000337 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 7: Significance output from PH Model

The result is consistent with Multiple Linear Regression Model. Only “metastasis status” is significant, there is no evidence to reject the null hypothesis that there is no association between age, gender and subtypes with survival time.

**Conclusion and Future Recommendation:**

In this data analysis, even though exploratory data analysis implied possible association between survival time with age, subtypes of cancer cells and gender, statistical model conclude that only metastasis status are associated with survival time. However, further analysis is recommended to check PH assumption and there could be correlation between kidney cancer subtypes with age or gender (Collinearity). In addition, proportion of missing values are relatively high, it will influence the conclusion of association test. Thus, it is

recommended for future studies to analyze a large sample with low missing values proportion.

**References:**

[1]Sultmann et al. (2005) Gene Expression In Kidney Cancer Is Associated with Cytogenetic Abnormalities, Metastasis Formation,and Patient Survival. *Clinical Cancer Research, 11,* 646-655. Retrieved from https://www.ebi.ac.uk/huber/docs/sultmann_2005.pdf.

[2] Wolfgang Huber (2016). kidpack: DKFZ kidney package. R package version 1.14.0. http://www.dkfz.de/mga.

[3] Therneau T (2015). _A Package for Survival Analysis in S_. version 2.38, <URL: http://CRAN.R-project.org/package=survival>.

[4]Terry M. Therneau and Patricia M. Grambsch (2000). _Modeling Survival Data: Extending the Cox Model_. Springer, New York. ISBN 0-387-98784-3.

**Appendix:**

**-R Code:**

```
#Download a dataset from bioconductor
source("http://bioconductor.org/biocLite.R")
biocLite("kidpack") #kidney microarray data
data(eset)
#Subtypes of kidney cancer has three levels
subtype <- as.factor(eset$type) #Change it to be a level variable
summary(subtype) #52 ccRCC, 9 chRCC, 13 pRCC, total 74


#I want to do a mulitple linear regression to check if age, gender and subtypes will affect patient's survival time.

#Step1:EDA for survival time
survival <- eset$survival.time
hist(survival,xlab="Survival Time",ylab="Frequency of Survival Time",ylim=c(0,25),col="green")
boxplot(survival,main="Boxplot of survival time for 74 patients with kidney cancer",ylab="Survival Time",col="grey")
qqnorm(survival,main="qqplot for survival time")
qqline(survival,col="red")

#Step2:EDA for age
age <-phenoData(eset)$age
summary(age) #Ages ranges from 26 to 85 with mean 59.77
```

```
boxplot(age,main="Boxplot of age for 74 patients with kidney
cancer",ylab="Age",col="grey") #Summetric, no skewness

#Step3:EDA for metastasis
m <- phenoData(eset)$m
summary(m) #27 NA's
table(m) #24 without metastasis, 23 with metastasis

#Step4:EDA for gender
gender <-ifelse(phenoData(eset)$sex == 'm',1,0)
table(gender)

#Plot of survival time versus age under three subtypes
plot(age,survival,type='n',xlab="Age",ylab="Survival
Time",xlim=c(0,90),main="Plot of survival time for different ages under three
cancer subtypes",cex.lab=1,cex.axix=1)
points(age[subtype=="ccRCC"],survival[subtype=="ccRCC"],col="green",pch=19,
xlim=c(0,90))
points(age[subtype=="pRCC"],survival[subtype=="pRCC"],col="red",pch=19)
points(age[subtype=="chRCC"],survival[subtype=="chRCC"],col="blue",pch=19)
legend("topleft",c("ccRCC","pRCC","chRCC"),fill=c("green","red",'blue'))

#Statistical modelling: multiple linear regression
ccRCC <- ifelse(subtype =="ccRCC",1,0)
pRCC <- ifelse(subtype =="pRCC",1,0)
chRCC <- ifelse(subtype =="chRCC",1,0)
gender <-ifelse(phenoData(eset)$sex == 'm',1,0)
lmod <- lm(survival~age+gender+ccRCC+pRCC+m)
summary(lmod)
plot(lmod)
lmod2 <- lm(survival~m)
summary(lmod2)
plot(m[which((!is.na(m)) & (!is.na(survival)))],survival[which((!is.na(m)) &
(!is.na(survival)))],ylab="Survival Time",xlab="Metastasis",main="Plot of
Survival Time Versis Metastasis")
points(m[which((!is.na(m)) & (!is.na(survival)))],lmod2$fitted.values,col="red")
#Predicted values are very close to 13 when m=1, 29 when m=0
cor(m[which((!is.na(m)) & (!is.na(survival)))],age[which((!is.na(m)) &
(!is.na(survival)))]) #-0.058 no correlation

#Fitting proportional hazard model for survival time as a function of age, gender,
subtypes and metastasis
library(survival)
death <- phenoData(eset)$died
phmod <-
coxph(Surv(time=survival,event=death)~age+gender+ccRCC+pRCC+m)
summary(phmod)
```