Applied Statistics - MATH 7343

Sai Nikhil

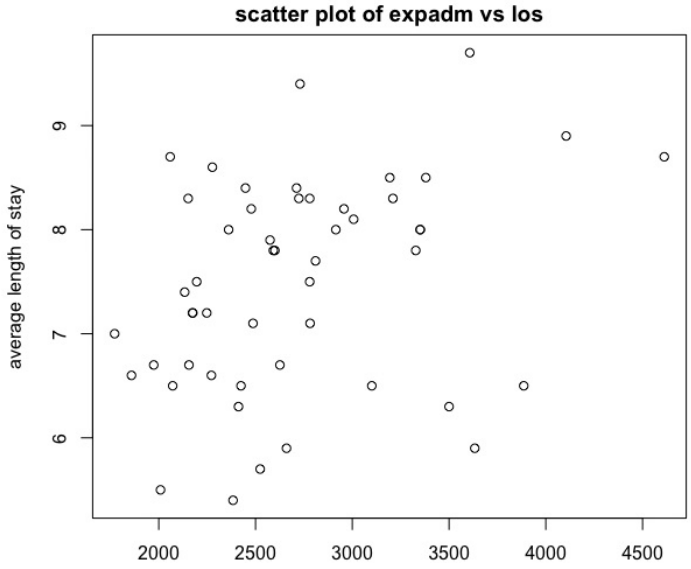
NUID: 002584864

Sai Nikhil

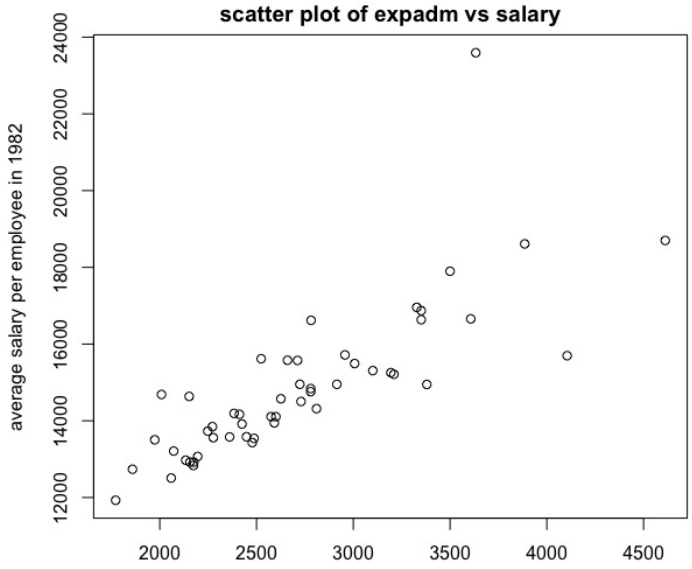
1

(a)

```
> data <- read.table(file="hospital.txt", header = TRUE)
>
> plot(data$expadm, data$los, xlab = "average expense per admission into a community hospital",
+      ylab = "average length of stay",
+      main = "scatter plot of expadm vs los")
>
```




```
> plot(data$expadm, data$salary, xlab = "average expense per admission into a community hospital",
+      ylab = "average salary per employee in 1982",
+      main = "scatter plot of expadm vs salary")
>
```



(b)

```
> line <- lm(data$expadm ~ data$los)
> summary(line)

Call:
lm(formula = data$expadm ~ data$los)

Residuals:
   Min     1Q Median     3Q    Max
-889.6 -428.1 -102.1  265.8 1663.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1281.96     608.10   2.108   0.0402 *
data$los      191.56      80.47   2.381   0.0212 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 577.6 on 49 degrees of freedom
Multiple R-squared:  0.1037, Adjusted R-squared:  0.08538
F-statistic: 5.668 on 1 and 49 DF,  p-value: 0.02121
```

From above output,

$$y = 1281.96 + 191.56x$$

(c)

```
> predict(line, data.frame(los = 6), interval = "confidence")
        fit      lwr      upr
6  2852.776 2653.803 3051.749
```

∴ 95% prediction interval $= (2653.803, 3051.749)$

(d)

$H_0: \beta = 0$

$H_A: \beta \neq 0$

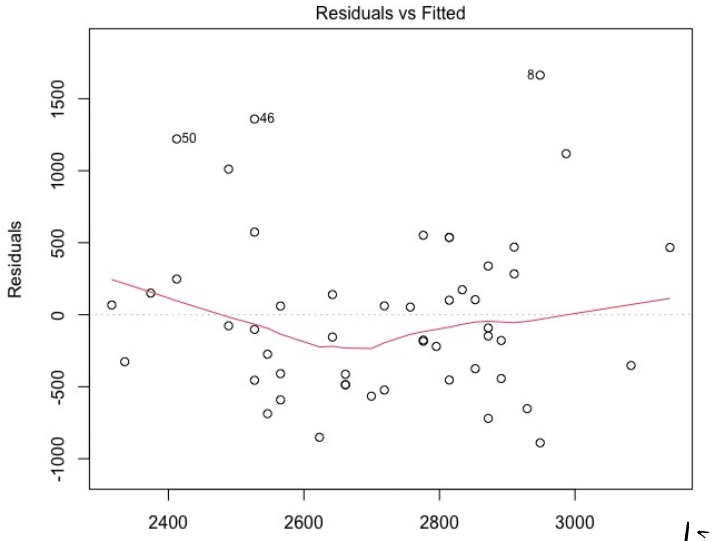
p-value $= 0.212 < \alpha = 0.05$

$\Rightarrow$ reject H_0

∴ we have significant linear relationship between expense per admission and length of stay.

(e)

```
> plot(line, which = 1)
```

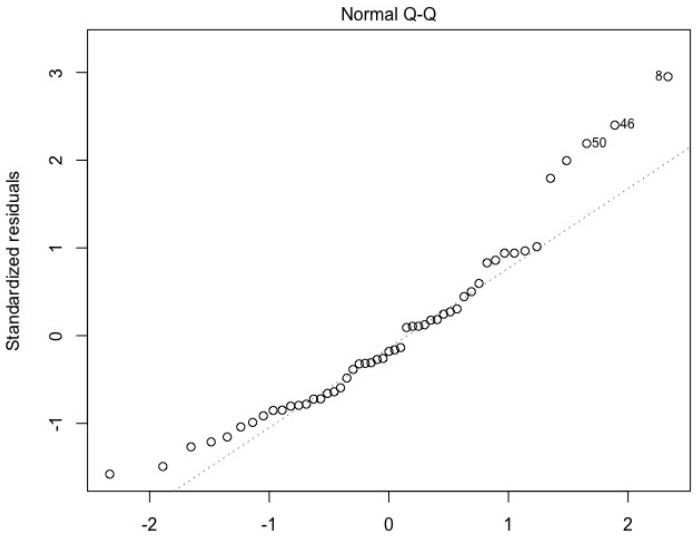


These plots say does not look like a great linear fit to the data. Also, there are several outliers (8, 46, 50)

Also, $R^2 = 0.203$

=> poor fit and no evidence of homoscedasticity has been violated

```
> plot(line, which = 2)
```



Most of them are normally distributed ignoring 8th, 46th, 50th.

(f)

```
> line2 <- lm(expadm ~ los + salary, data = data)
> summary(line2)

Call:
lm(formula = expadm ~ los + salary, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-920.43 -134.15    0.57  133.77  876.74

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2582.7364   464.7700  -5.557 1.18e-06 ***
los           213.7967    42.2077   5.065 6.45e-06 ***
salary          0.2490     0.0218  11.422 2.73e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 302.6 on 48 degrees of freedom
Multiple R-squared:  0.7589, Adjusted R-squared:  0.7489
F-statistic: 75.55 on 2 and 48 DF,  p-value: 1.485e-15
```

$\therefore \quad y = -2582.7364 + 213.7967 x_1 + 0.2490 x_2$

$\therefore \quad \beta_1 = 213.7967 \Rightarrow$ when average salary is constant, 1 day increase in length of stay, increases average expenses by 213.7967

$\beta_2 = 0.2490 \Rightarrow$ when length of stay is constant, 1 unit increase in salary increases average expenses by 0.2490

(g) When average salary is added to model, the estimated coefficient of length of stay increases from 191.56 to 213.79.

(b)

```
> line3 <- lm(expadm ~ los, data = data)
> anova(line2, line3)
Analysis of Variance Table

Model 1: expadm ~ los + salary
Model 2: expadm ~ los
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1     48  4396616
2     49 16346819 -1 -11950203 130.47 2.731e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output says,

R^2 increases from 0.1037 to 0.7589

$\Rightarrow$ inclusion of salary in addition to average length of stay, will improve ability to predict mean expense per admission.

Also, p-value $= 2.731e-15 < \alpha = 0.05$

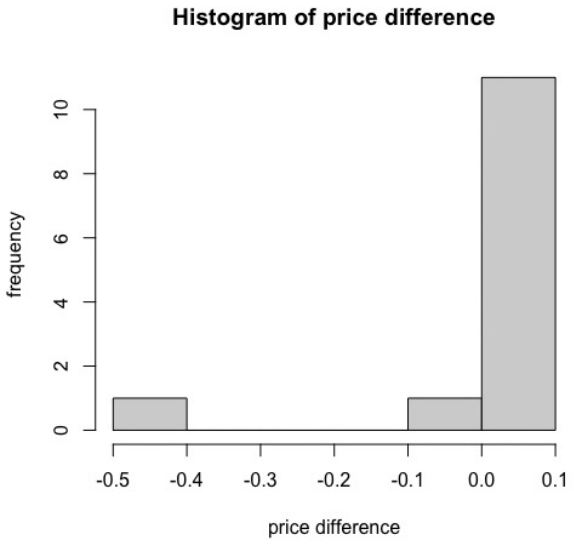
This conclusion also supports to include salary.

(2)

We want to see if there is a price difference for items between old account and new account.

Data Exploration:

```
> data <- read.table(file="shopping.txt", header = TRUE)
> data$difference <- (data$OldAccount - data$NewAccount)
> hist(data$difference, xlab = "price difference", ylab = "frequency",
+      main = "Histogram of price difference")
>
```



Item	OldAccount	NewAccount	difference
1	27.61	27.55	0.06
2	10.29	10.19	0.10
3	52.20	52.19	0.01
4	33.89	33.87	0.02
5	340.49	340.99	-0.50
6	65.34	65.29	0.05
7	54.69	54.64	0.05
8	15.23	15.21	0.02
9	102.19	102.09	0.10
10	4.88	4.89	-0.01
11	161.39	161.35	0.04
12	12.51	12.49	0.02
13	96.99	96.96	0.03

The above histogram clearly says that the distribution is not normal. Hence, it makes sense to use a non-parametric test such as Wilcoxon's test.

H_0: median difference ≤ 0

H_1: median difference > 0

```
> wilcox.test(data$OldAccount, data$NewAccount, paired = TRUE, alternative = "greater")

   Wilcoxon signed rank test with continuity correction

data:  data$OldAccount and data$NewAccount
V = 77, p-value = 0.01509
alternative hypothesis: true location shift is greater than 0
```

p-value $= 0.1509 < \alpha = 0.05 \Rightarrow$ Reject H_0

$\therefore$ There is significant evidence that student's suspicion on the e-commerce company's pricing practice is true.

(3)

(a) H_0: proportion of years in which Phil sees his shadow and March temperatures are below average is equal to proportion of yes when phil does not see his shadow and March temperatures are below average.

H_A: proportion of years in which Phil sees his shadow and March temperatures are below average is not equal to proportion of years when Phil does not see his shadow and March temperatures are below average.

(b) It is true that χ^2 is a two-sided test. But in this case, large differences between observed and expected can be considered as outliers and this can still be thought effectively as one-sided test. Another main issue here is number of observation = 1, while χ^2 is applicable when number of observations ≥ 5.

(c)

Observed

March Temperature	Phil sees its shadow: Yes	No
Above Normal	16	8
Below Normal	5	1

Expected

March Temperature	Phil sees its shadow: Yes	No
Above Normal	16.8	7.2
Below Normal	4.2	1.8

$$\chi^2 = \sum_{\text{cells}} \frac{(O-E)^2}{E}$$

$$= \frac{(16-16.8)^2}{16.8} + \frac{(8-7.2)^2}{7.2} + \frac{(5-4.2)^2}{4.2} + \frac{(1-1.8)^2}{1.8}$$

$$\approx 0.0893$$

$$df = (2-1)(2-1) = 1$$

From table, $p\text{-value} \approx 0.76 > \alpha = 0.05$

$\therefore$ We accept H_0.

$\therefore$ Therefore we can say that there is no association between phil seeing his shadow and March temperatures.

(d) p-value > 0.05

$\Rightarrow$ Groundhog Day's information cannot be effectively used to predict the March temperature.

---

(4)

(a) In order to find the effect of nematodes on the plant growth, we can use an ANOVA test, where we measure the mean growth of plants for different variations of nematodes.

$\therefore$ H_0 : $\mu_0 = \mu_{1000} = \mu_{5000} = \mu_{10000}$

H_A : Atleast two of the means are different

```
> data <- data.frame(Nematodes = factor(c(0, 0, 0, 0,
+                                         1000, 1000, 1000, 1000,
+                                         5000, 5000, 5000, 5000,
+                                         10000, 10000, 10000, 10000)),
+                    SeedlingGrowth = c(10.8, 9.1, 13.5, 9.2,
+                                       11.1, 11.1, 8.2, 11.3,
+                                       5.4, 4.6, 7.4, 5.0,
+                                       5.8, 5.3, 3.2, 7.5))
> summary(data)
 Nematodes SeedlingGrowth
 0    :4   Min.   : 3.200
 1000 :4   1st Qu.: 5.375
 5000 :4   Median : 7.850
 10000:4   Mean   : 8.031
           3rd Qu.:10.875
           Max.   :13.500
> aov.fit <- aov(SeedlingGrowth~Nematodes, data=data)
> summary(aov.fit)
            Df Sum Sq Mean Sq F value   Pr(>F)
Nematodes    3 100.65   33.55   12.08 0.000616 ***
Residuals   12  33.33    2.78
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

$P = 0.0006 < \alpha = 0.05 \Rightarrow$ reject H_0

$\therefore$ nematodes probably have effect on growth of plants.

(b) H_0 : $\mu_0 > \mu_{1000}$ or μ_{5000} or μ_{10000}

H_A : $\mu_0 \leq \mu_{1000}$ and μ_{5000} and μ_{1000}

For this pairwise analysis, we can use Tukey HSD test.

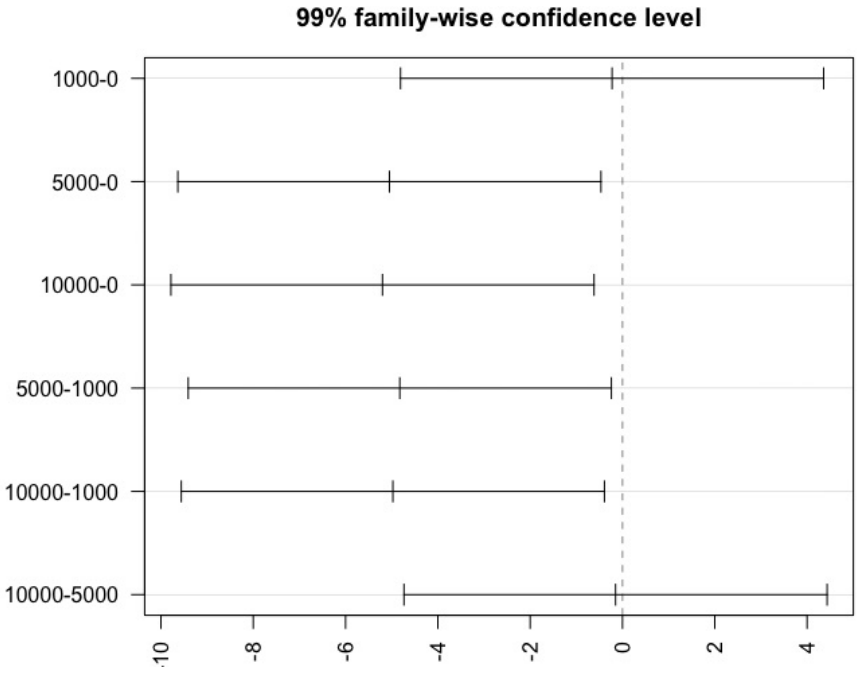
```
> TukeyHSD(aov.fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = SeedlingGrowth ~ Nematodes, data = data)

$Nematodes
               diff       lwr       upr     p adj
1000-0      -0.225 -3.723577  3.273577 0.9973921
5000-0      -5.050 -8.548577 -1.551423 0.0050470
10000-0     -5.200 -8.698577 -1.701423 0.0040599
5000-1000   -4.825 -8.323577 -1.326423 0.0070131
10000-1000  -4.975 -8.473577 -1.476423 0.0056301
10000-5000  -0.150 -3.648577  3.348577 0.9992199
```

} Accept H_0

} Reject H_0

} Accept H_0

```
> par(mar=c(2, 7, 2, 2))
> plot(TukeyHSD(aov.fit, conf.level=0.99), las=2)
```



∴ For 1000-0 and 10000-5000, we accept H_0

else, reject H_0

⇒ nematodes significantly reduce the growth of plants.

(c) In order to find if 1000 nematodes have different growth then 5000 nematodes, we carry out pairwise test for 1000-5000 nematodes.

From part (b) code,

we get,

$$p\text{-value} = 0.007 < \alpha = 0.05$$

$$\Rightarrow \text{reject } H_0$$

∴ There is no significant difference in growth from 1000 to 5000.

However, this conclusion is not inline with the data as I could see from the table.

---

(5) (a) We can use logistic analysis as the output is discrete here.

H_0: $\beta = 0$ (no correlation)

H_A: $\beta \neq 0$

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta x$$

```
> data <- read.table(file="programmer.txt", header = TRUE)
> lr.fit<-glm(formula = success ~ Experience, family = binomial(link = "logit"),data = data)
> summary(lr.fit)

Call:
glm(formula = success ~ Experience, family = binomial(link = "logit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8924  -0.7591  -0.4030   0.7715   2.0147

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.2206     1.2770  -2.522   0.0117 *
Experience    0.1665     0.0659   2.527   0.0115 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35.426  on 25  degrees of freedom
Residual deviance: 26.140  on 24  degrees of freedom
AIC: 30.14

Number of Fisher Scoring iterations: 4
```

$\therefore \quad \alpha = -3.2206$

$\beta = 0.1665 > 0$

p-value $= 0.0115 < \alpha = 0.05$

$\Rightarrow$ reject H_0

$\therefore$ Correlation is statistically significant

$\Rightarrow$ work experience improves programmer's ability.

(b) $(0.1665 \times 12 \pm 1.96 \times 0.0659 \times 12)$

95% CI $= e$

$\approx (1.565229, 34.742649)$

(c) $\ln\left(\frac{p}{1-p}\right) = -3.2206 + 0.1665(24)$

$\Rightarrow p = 0.684$

$\therefore$ Expected Salary per year $= \$90,000 \times 0.684$

$= \$61,560$

(d) $(X - 10K) \Rightarrow \$51,560$ ← Programmer 1

$\ln\left(\frac{p}{1-p}\right) = -3.2206 + 0.1665(18)$

$\Rightarrow p \approx 0.444$

$\therefore$ Expected Salary per year $= \$90000 \times 0.444$

$= \$39,960$ ← Programmer 2

Second programmer would be a better if company wants to save money but would be less efficient. Hence, not a long term benefit.

6

(a)

```
> library("psych")
> data <- read.table(file = "cyto.txt", header = TRUE)
> summary(data)
      time           censor           group
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000
 1st Qu.: 35.0   1st Qu.:1.0000   1st Qu.:1.000
 Median :101.0   Median :1.0000   Median :1.000
 Mean   :135.8   Mean   :0.8966   Mean   :1.379
 3rd Qu.:204.0   3rd Qu.:1.0000   3rd Qu.:2.000
 Max.   :400.0   Max.   :1.0000   Max.   :2.000
>
> #part a
> describeBy(data$censor,data$group)

 Descriptive statistics by group
group: 1
   vars  n mean   sd median trimmed mad min max range  skew kurtosis   se
X1    1 18 0.83 0.38      1    0.88   0   0   1     1 -1.64     0.75 0.09
------------------------------------------------------------------------------
group: 2
   vars  n mean sd median trimmed mad min max range skew kurtosis se
X1    1 11    1  0      1       1   0   1   1     0  NaN      NaN  0
```

From above statistics,

number of deaths in group 1 (n_1) = 18

number of deaths in group2 (n_2) = 11

(b)

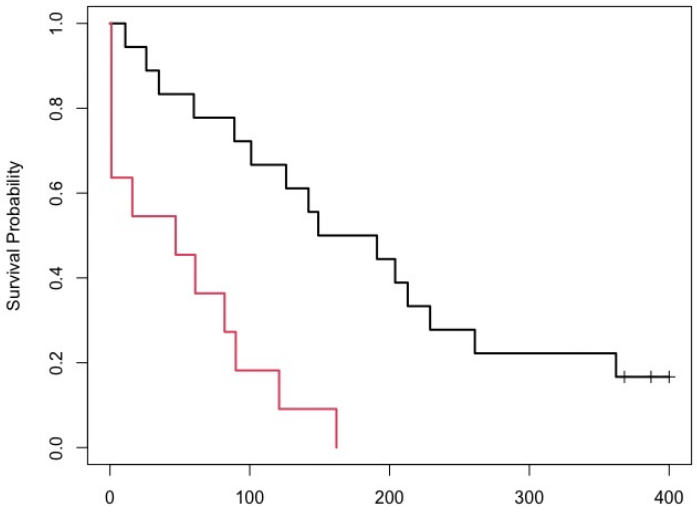
```
> library("survival")
> a.fit <- survfit(Surv(time, censor) ~ group,data=data)
> summary(a.fit)
Call: survfit(formula = Surv(time, censor) ~ group, data = data)

                group=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   11     18       1    0.944  0.0540        0.8443        1.000
   26     17       1    0.889  0.0741        0.7549        1.000
   35     16       1    0.833  0.0878        0.6778        1.000
   60     15       1    0.778  0.0980        0.6076        0.996
   89     14       1    0.722  0.1056        0.5423        0.962
  101     13       1    0.667  0.1111        0.4809        0.924
  126     12       1    0.611  0.1149        0.4227        0.883
  142     11       1    0.556  0.1171        0.3675        0.840
  149     10       1    0.500  0.1179        0.3150        0.794
  191      9       1    0.444  0.1171        0.2652        0.745
  204      8       1    0.389  0.1149        0.2179        0.694
  213      7       1    0.333  0.1111        0.1734        0.641
  229      6       1    0.278  0.1056        0.1319        0.585
  261      5       1    0.222  0.0980        0.0936        0.527
  362      4       1    0.167  0.0878        0.0593        0.468

                group=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     11       4   0.6364  0.1450        0.4071        0.995
   16      7       1   0.5455  0.1501        0.3180        0.936
   47      6       1   0.4545  0.1501        0.2379        0.868
   61      5       1   0.3636  0.1450        0.1664        0.795
   82      4       1   0.2727  0.1343        0.1039        0.716
   90      3       1   0.1818  0.1163        0.0519        0.637
  121      2       1   0.0909  0.0867        0.0140        0.589
  162      1       1   0.0000     NaN           NA           NA
```

Survival function for group 2

Survival function for group 2

(c)

```
> ggsurvplot(a.fit, data = data)
> plot(a.fit, col=1:2, xscale=1, lwd=2, mark.time=TRUE, xlab="time", ylab="Survival Probability")
>
```



(d) It appears that group 1 survived longer than group 2

(e)

```
> survdiff(Surv(time, censor) ~ group, data = data)
Call:
survdiff(formula = Surv(time, censor) ~ group, data = data)

         N Observed Expected (O-E)^2/E (O-E)^2/V
group=1 18       15    21.37      1.90      12.4
group=2 11       11     4.63      8.74      12.4

 Chisq= 12.4  on 1 degrees of freedom, p= 4e-04
```

p-value = 4e-04 $< \alpha = 0.05 \Rightarrow$ reject H_0

$\therefore$ Distribution of survival times are not identical in the two groups.

(7)

$H_0 : \rho = 0$

$H_A : \rho \neq 0$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Here, $n = 40$, $r^2 = 0.36$

$\Rightarrow r = 0.6$

$$\therefore t = 0.6\sqrt{\frac{40-2}{1-0.36}} \approx 4.62$$

$\therefore df = 38, t = 4.62$

$\Rightarrow p = 4.27e{-5} < \alpha = 0.05$

$\Rightarrow$ reject H_0

$\therefore$ X and Y are associated with each other.