Lecture 9: Multiple Regression

BMI 713
November 16, 2017
Peter J Park

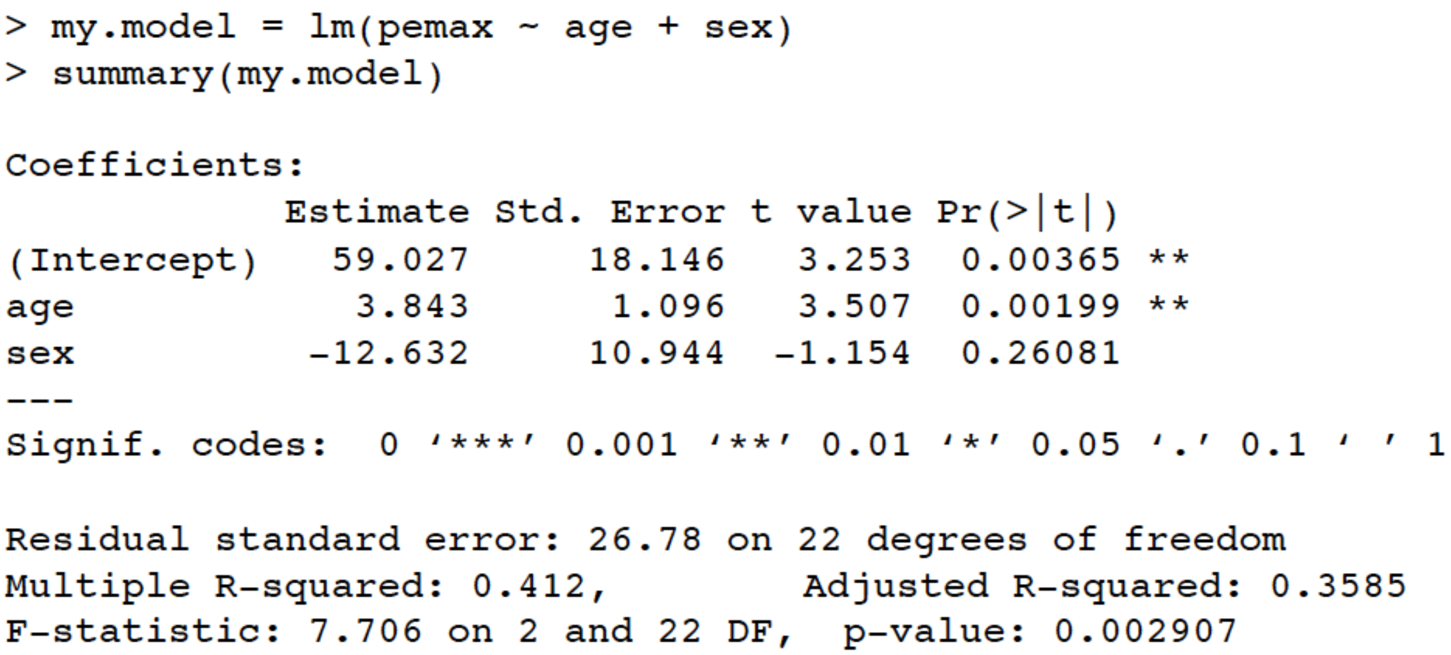
Categorical Variables

- Since “sex” is a categorical variable with two categories, we can represent a patient’s category by creating a variable that takes values “0” and “1” for men and women, respectively.
- This is not a continuous variable, but that is not a problem

$$\boxed{y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e}$$

- A binary variable created to represent a categorical variable with two categories is called a **dummy variable**, since its value (1 vs. 0) is arbitrarily chosen as numeric quantity.

Categorical Variables



```
> my.model = lm(pemax ~ age + sex)
> summary(my.model)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   59.027     18.146   3.253  0.00365 **
age            3.843      1.096   3.507  0.00199 **
sex          -12.632     10.944  -1.154  0.26081
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.78 on 22 degrees of freedom
Multiple R-squared: 0.412,      Adjusted R-squared: 0.3585
F-statistic: 7.706 on 2 and 22 DF,  p-value: 0.002907
```

- PEmax = 59 + 3.84 * age - 12.6 * sex
- For men, PEmax = 59 + 3.84 * age
- For women, PEmax = 46.4 + 3.84 * age
- So b_2 is the difference between the predicted values of a man

Categorical Variables

- What if there are multiple categories?
- Examples:
  - Geographic location: “Northeast”, “South”, “Midwest”, etc.
  - Race: “White”, “Black”, “Asian”, “American Indian”

- How about X = 0 “White”, X = 1 “Black”, X = 2 “Asian”, X = 3 “American Indian”?
- We need to create multiple dummy variables
- X_1 = “Black”, X_2 = “Asian”, X_3 = “American Indian”
- One category, by default, is the reference category
- So, for a multiple-category variable, the number of dummy variables needed is one fewer than the number of categories

Interaction Terms

- In the regression models above, each explanatory variable is related to the outcome independently of all others.
- Sometimes the effect of an explanatory variable depends on the level of another explanatory variable.
- If this occurs, it is called an **interaction effect**
- **If we believe interaction effects are present, we include them in the model.**

$$\boxed{y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e}$$

Interaction Terms

- For instance, we can see whether the effect of age is different for men and women

$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

men: $\hat{y} = a + b_1 x_1$

women: $\hat{y} = (a + b_2) + b_1 x_1$

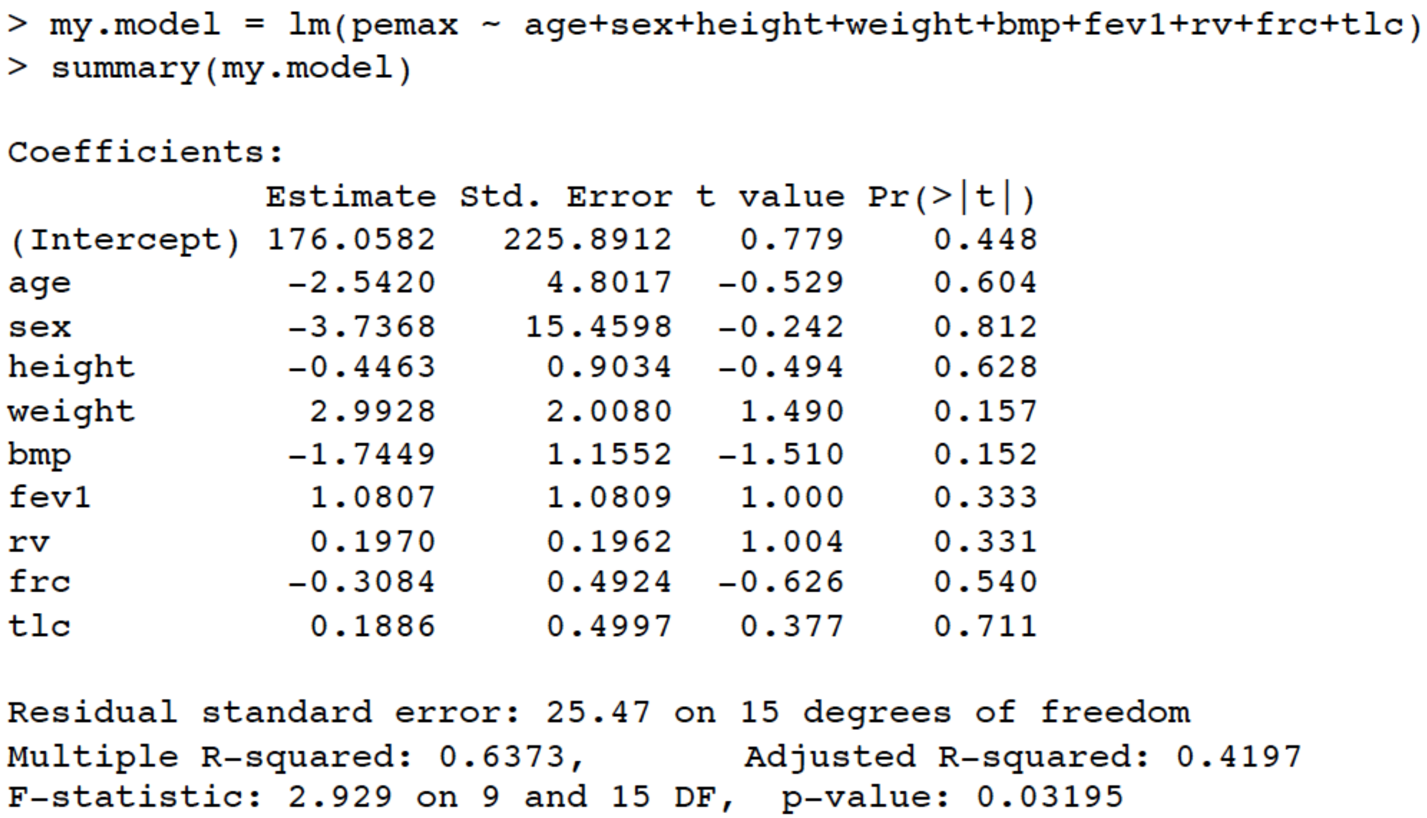
$$\hat{y} = a + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2$$

men: $\hat{y} = a + b_1 x_1$

women: $\hat{y} = (a + b_2) + (b_1 + b_3) x_1$

A Full Model

- We can do inference about all the model parameters together (F-test), or we can do inference about them separately (t-tests).



```
> my.model = lm(pemax ~ age+sex+height+weight+bmp+fev1+rv+frc+tlc)
> summary(my.model)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 176.0582   225.8912   0.779    0.448
age          -2.5420     4.8017  -0.529    0.604
sex          -3.7368    15.4598  -0.242    0.812
height       -0.4463     0.9034  -0.494    0.628
weight        2.9928     2.0080   1.490    0.157
bmp          -1.7449     1.1552  -1.510    0.152
fev1          1.0807     1.0809   1.000    0.333
rv            0.1970     0.1962   1.004    0.331
frc          -0.3084     0.4924  -0.626    0.540
tlc           0.1886     0.4997   0.377    0.711

Residual standard error: 25.47 on 15 degrees of freedom
Multiple R-squared: 0.6373,      Adjusted R-squared: 0.4197
F-statistic: 2.929 on 9 and 15 DF,  p-value: 0.03195
```

CF Example

- Age was a statistically significant variable before
- Note that none of the variables are significant now!

- But the joint F-test is still significant; there must be some effect
- From the p-values in the full model, you cannot tell whether a variable would be significant in a reduced model
- A predictor may become non-significant when there is a highly correlated predictor

Model Selection

- How do we select the ‘best’ model?
- As a general rule, we prefer to include only those explanatory variables that help us to predict the response y, the coefficients of which can be accurately estimated
- To study the full effect of each explanatory variable on the response it would be necessary to perform a separate regression analysis for each combination of the variables
- While thorough, the all possible models approach is usually extremely time-consuming
- More frequently, we use a stepwise approach to choose a “best-fitting” model

Forward Selection

- Two commonly used procedures are **forward selection** and **backward elimination**
- Forward selection beings with no variables in the model and introduces variables one at a time
- The model is evaluated at each step
- For example, we might begin by including the single variable that yields the largest R^2
- We next add the variable that increases R^2 the most (the increase must be statistically significant)
- We continue this procedure until none of the remaining variables explains a significant amount of the additional variability in y

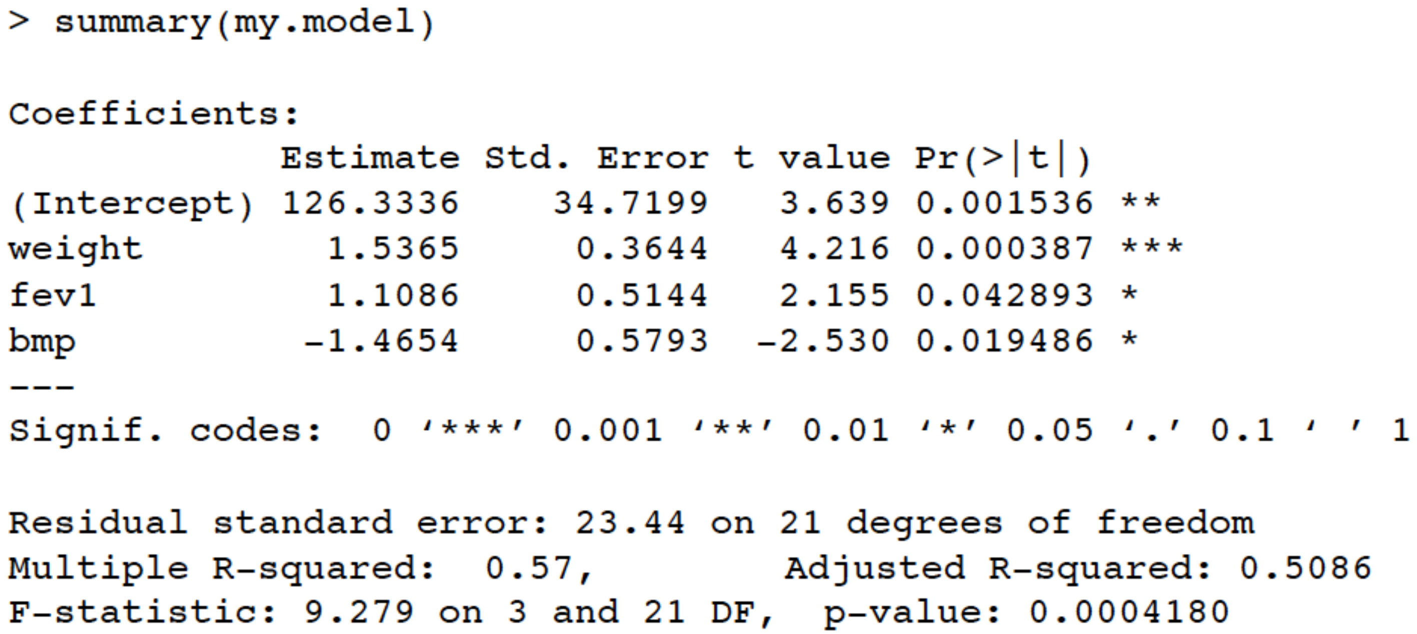
Backward Selection

- Backward elimination begins by including all explanatory variables in the model
- We drop the variable that contributes the least to the overall R^2
- The process continues until each remaining variable explains a significant portion of the variability in y

- Your software probably uses a more sophisticated penalty
- Example: AIC (Alkaike information criterion): 2k + n * log (RSS/n)
- This finds the model that best explains the data with a minimum of free parameters

CF Example

- Model from the original paper



```
> summary(my.model)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 126.3336    34.7199   3.639 0.001536 **
weight        1.5365     0.3644   4.216 0.000387 ***
fev1          1.1086     0.5144   2.155 0.042893 *
bmp          -1.4654     0.5793  -2.530 0.019486 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.44 on 21 degrees of freedom
Multiple R-squared:  0.57,        Adjusted R-squared: 0.5086
F-statistic: 9.279 on 3 and 21 DF,  p-value: 0.0004180
```

- It is somewhat arbitrary that weight ended up in the model
- You may not get the same result, depending on your selection criteria
- PEmax is probably connected to the patient’s physical size

Stepwise Model

- A stepwise procedure allows variables that have been dropped from the model to re-enter at a later time
- Usually the p-value criteria are relaxed in these procedures
- in a forward stepwise method, one may have
  - p=0.1 to include or remove a variable
  - p=0.1 to include and p=0.2 to remove a variable
- It is possible to end up with different final models, depending on which strategy is used
- The decision is usually made based on a combination of statistical and non statistical considerations
- **Key aspects are simplicity of the model and whether the model can be easily interpreted**

Logistic Regression

- In linear regression, the response variable *Y* is continuous

$$\boxed{\begin{gathered} y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e \\ e \sim N\left(0, \sigma^2\right) \end{gathered}}$$

- We are interested in identifying explanatory variables that help us to predict the mean value of the response by explaining the observed variation in the outcomes
- However, we often have a response variable *Y* that is **dichotomous** rather than continuous

Logistic Regression

- We cannot simply use linear regression with 0/1 as outcome
- We define the ‘logit’ transformation for the probability of the outcome variable being a “1” (vs “0”)

$$\boxed{\mathrm{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k}$$

- logit(p) can take any value and we can perform regression
- How do we interpret coefficients?
- If X is a binary variable, β is the log of the odds ratio
- We estimate parameters by the maximum likelihood method

Introduction to Survival Analysis

- “Time-to-event” data: in some studies, the response variable of interest is the length of time between an initial observation and the occurrence of a subsequent event.
- Despite the name, “survival” analysis isn’t only for analyzing time until death. It deals with any situation where the quantity of interest is amount of time until study subject experiences some relevant endpoint.
- This subsequent event is often called a “failure”; the terms “failure” and “event” are used interchangeably for the endpoint of interest.

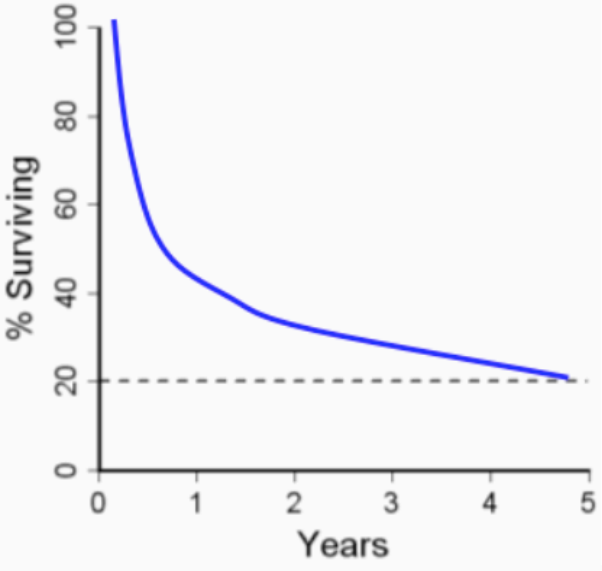
Survival Analysis

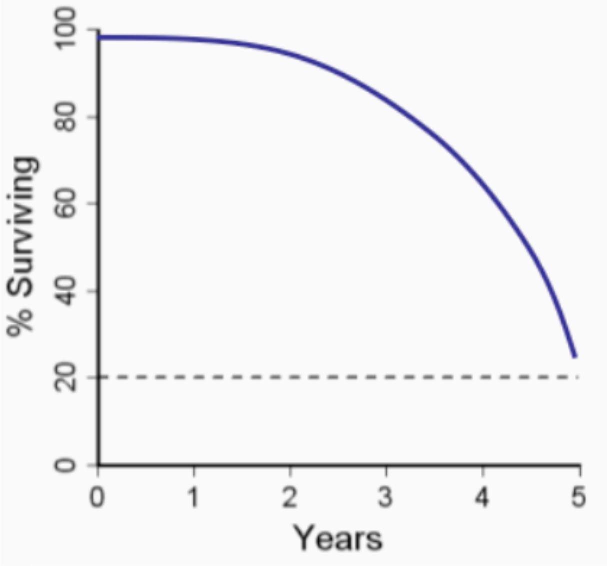
- Examples:
  - Time from birth until death
  - Time from start of treatment until remission of disease
  - Injection of a lentivirus with a growth factor into mice till the development of tumor
- The time from the initial event until failure is called the survival time
- Time is a continuous measurement that cannot assume negative values - it is rarely normally distributed
- We will study estimation, one-sample/two-sample inference, and regression in this context

How to Measure Survival

- **Idea:** Report the mean survival time?
- **Problem:** Not robust to outliers
- **Idea:** 5-year survival rate? “long” vs “short” survival
- **Problem:** how to choose the cutoff time?

- These two have the same 5-yr survival rates:





Survival Function

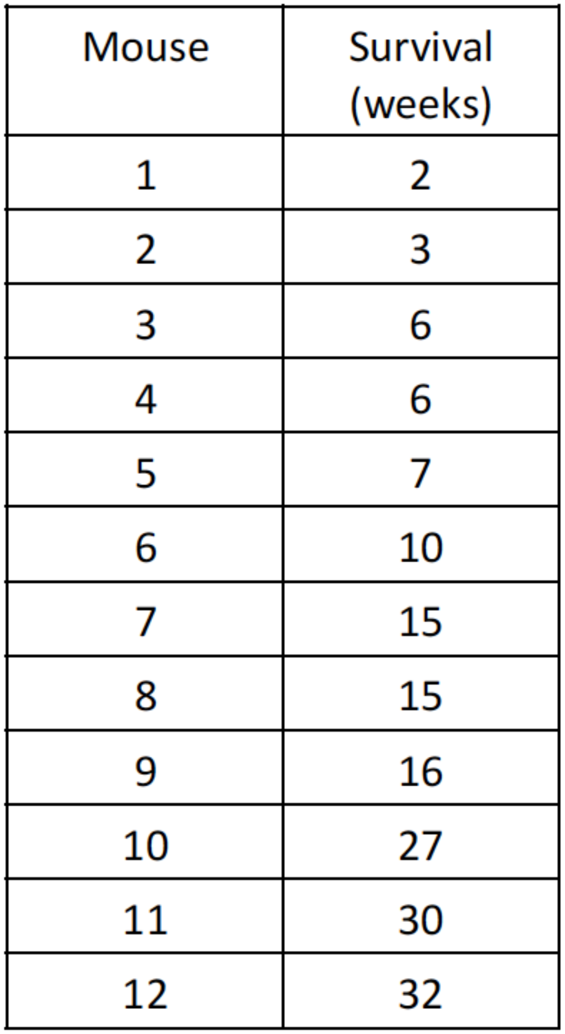
- A distribution of survival times can be characterized by a survival function ***S(t)***
- *S(t)* is the probability that an individual survives beyond time t, or the proportion of subjects who have not yet failed
- If *T* is a continuous random variable representing survival time, then $S(t) = P(T>t)$
- The graph of the survival function *S(t)* versus time *t* is called a survival curve

The Kaplan-Meier Survival Curve

- The Kaplan-Meier method, also known as the product-limit method, can be used to estimate a survival curve
- It is a **nonparametric** technique that does not take any assumptions about the underlying distribution of survival times
- **Example:** A total of 12 mice with a particular genotype were exposed to radiation and were followed until death

The Kaplan-Meier Curve

- Based on this sample, what can we infer about the survival?
- We begin by ordering the survival times associated with each of the 12 mice
- Again, ‘survival’ is a general term
- Another example: the mice with tumor were treated with a small molecule and were followed to remission



Mouse	Survival (weeks)
1	2
2	3
3	6
4	6
5	7
6	10
7	15
8	15
9	16
10	27
11	30
12	32

Survival Function

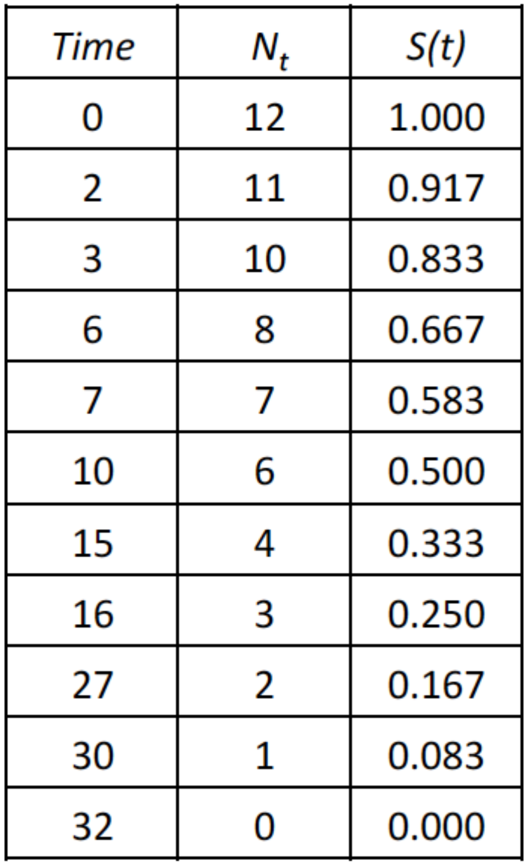
- Note that no one fails at 0 weeks or at 1 week following exposure, one mouse fails at 2 weeks, one at 3 weeks, none at 4 weeks, and so on
- N_t is defined as the number of mice who have not yet failed at time t:
  - N_0 = 12, N_1 = 12, N_2 = 11, …
- If everyone in the sample fails, then the survival function is estimated by $S(t) = N_t / N_0$

Survival Function

- Therefore,
  S(0) = 12/12 = 1.000
  S(2) = 11/12 = 0.917
  S(3) = 10/12 = 0.833
  S(6) = 8/12 = 0.667

  ….
  S(32) = 0/12 = 0.000

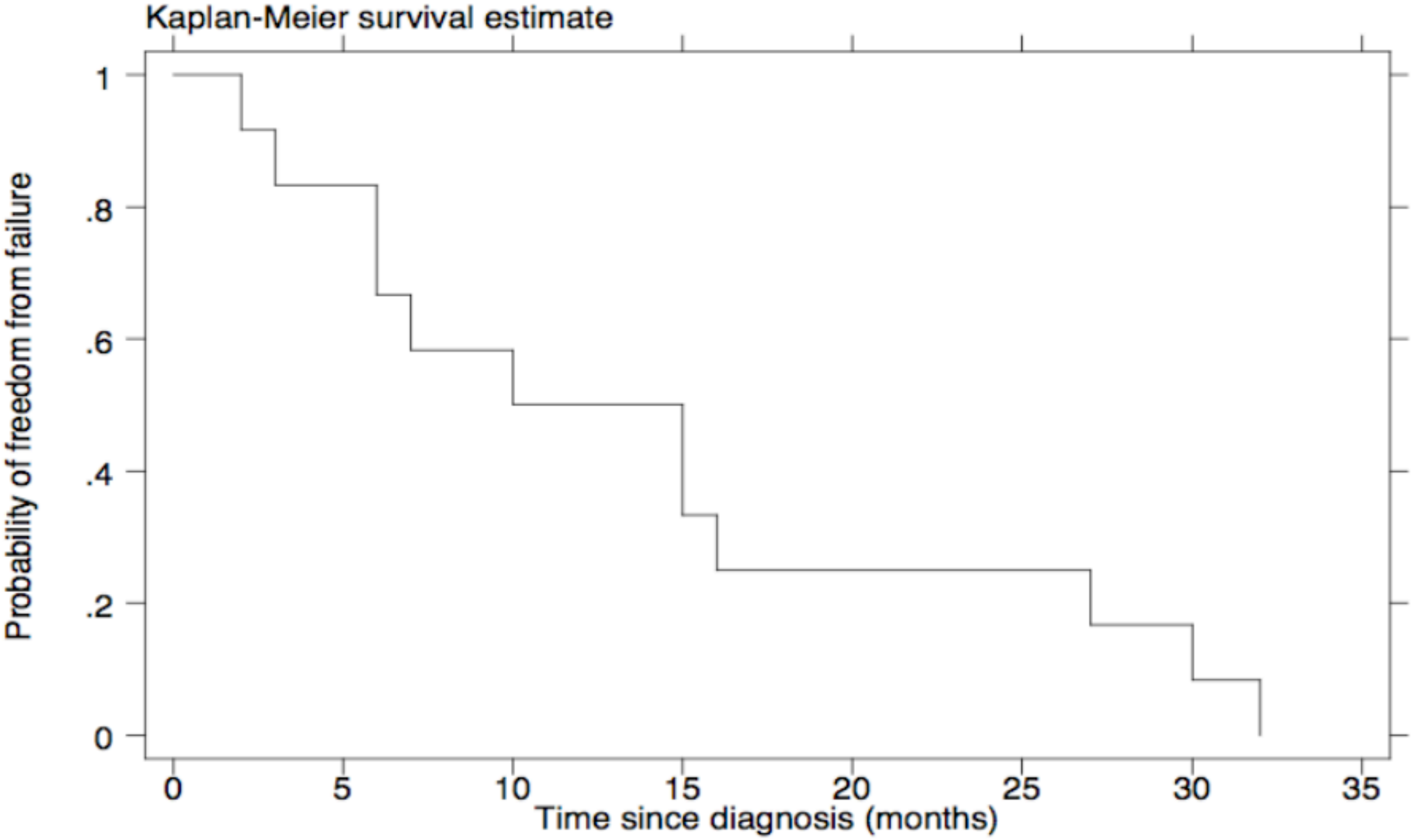
- The Kaplan-Meier estimate of the survival curve is:



Time	N_t	*S(t)*
0	12	1.000
2	11	0.917
3	10	0.833
6	8	0.667
7	7	0.583
10	6	0.500
15	4	0.333
16	3	0.250
27	2	0.167
30	1	0.083
32	0	0.000

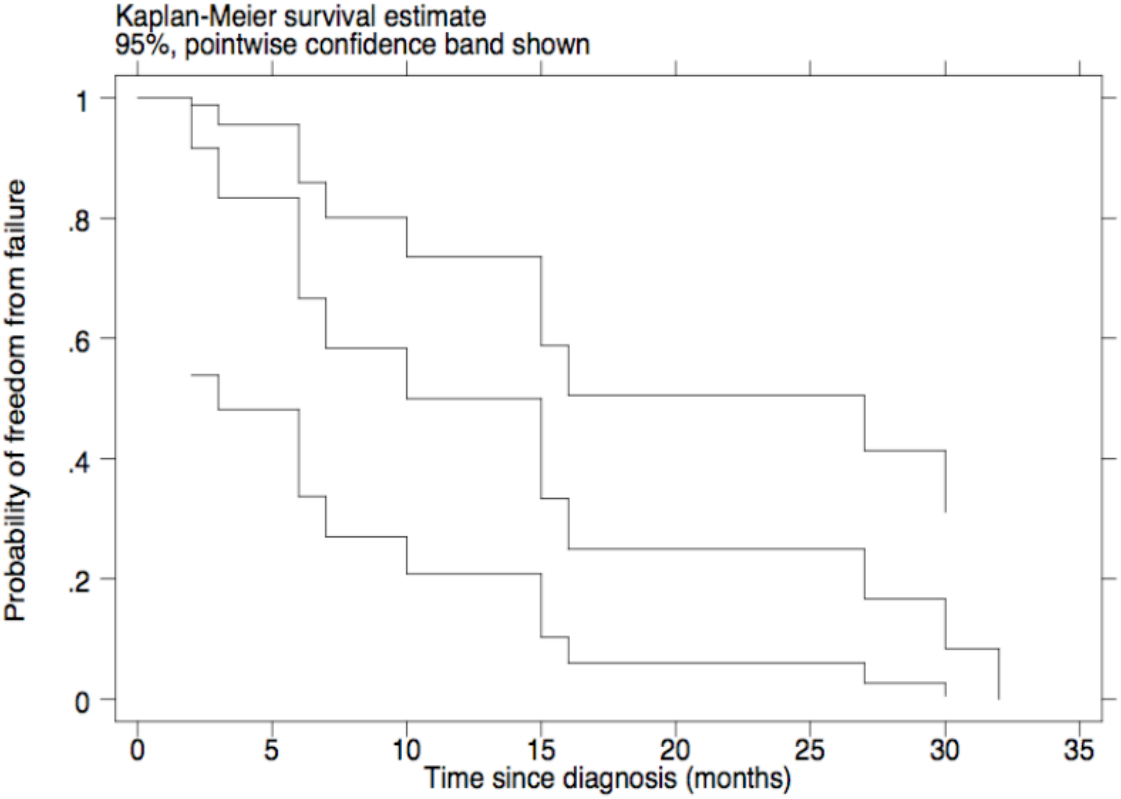
Survival Function

- *S(t)* is an estimate of the true population survival function calculated using the information in a sample of observations



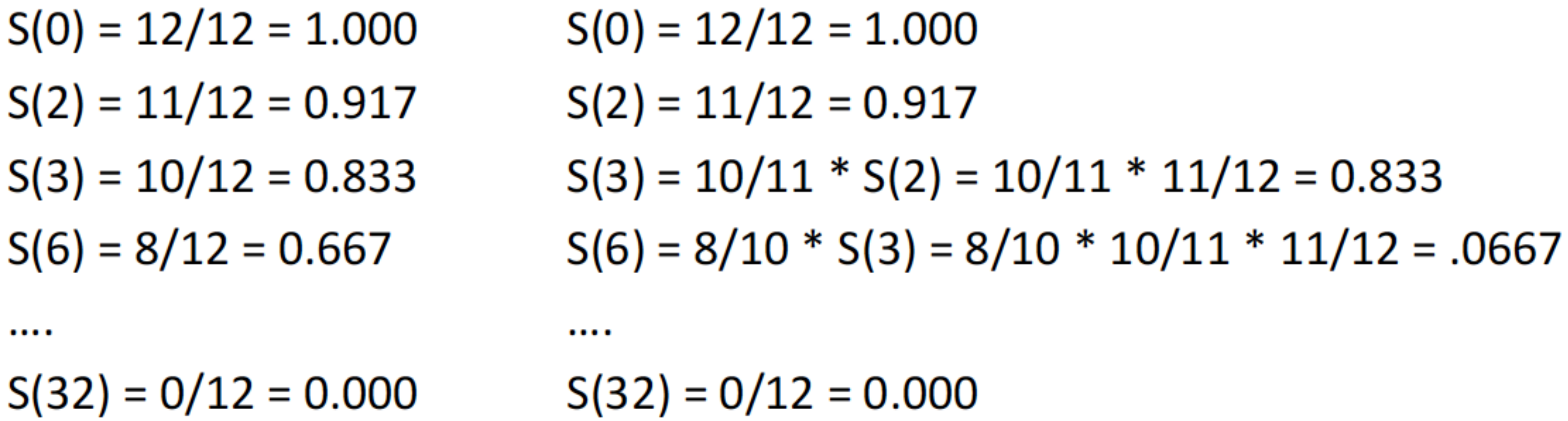
Confidence Bands

- To quantify the amount of sampling variability involved, we can calculate the standard error of *S(t)* and use it to construct confidence bands



Survival Function

- Another way: $S(t_i)$ = P(alive at t_i | alive at t_{i-1}) x $S(t_{i-1})$
- Probability you survive until time t_i = probability you survive until t_{i-1} and then survive until t_i given you made it to t_{i-1}
- P(alive at t_i | alive at t_{i-1}) = (# alive at t_i) / (# alive at t_{i-1})



S(0) = 12/12 = 1.000

S(2) = 11/12 = 0.917

S(3) = 10/12 = 0.833

S(6) = 8/12 = 0.667

….

S(32) = 0/12 = 0.000

S(0) = 12/12 = 1.000

S(2) = 11/12 = 0.917

S(3) = 10/11 * S(2) = 10/11 * 11/12 = 0.833

S(6) = 8/10 * S(3) = 8/10 * 10/11 * 11/12 = .0667

….

S(32) = 0/12 = 0.000

Missing Data: Censoring

- Time-to-event data often have data missing in particular way: **individuals may be lost to follow-up or may drop out of the study before they experience the event of interest**
- This incomplete observation of time of failure is known as **censoring**
- Censored data provide partial information: you do not know how long a patient lived, but you know that she/he lived at least as long as the time before being lost to follow-up.
- Why would a person be lost to follow-up? The person could have, e.g., moved to another city, withdrawn from the study, or died of a different cause.
- The data might be analyzed before the event of interest has occurred in all subjects

Censoring

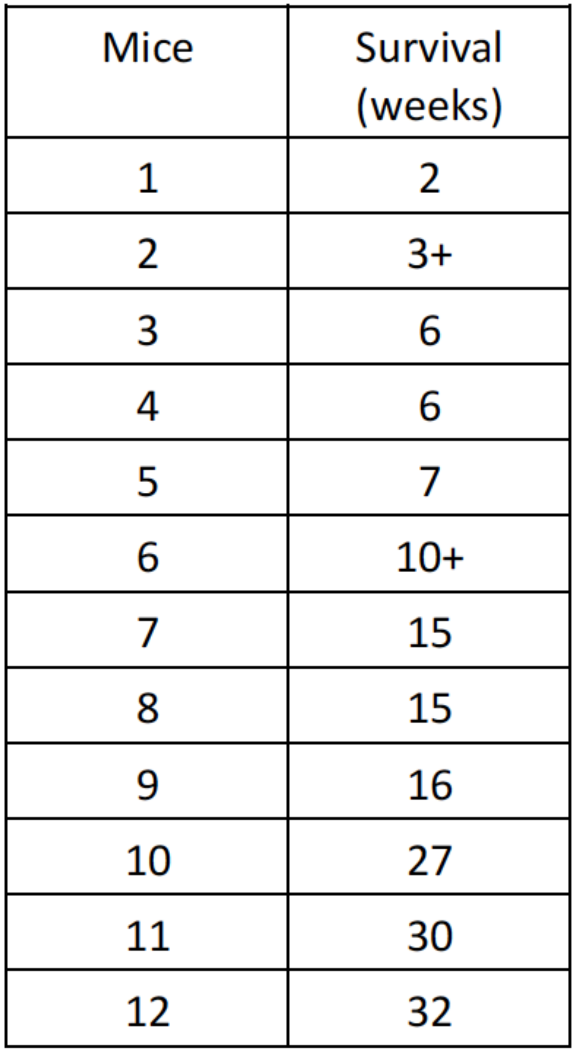
- We would like to be able to take advantage of the partial information contained in censored observations
- To do inference in the setting of missing data, we must be willing to make a big assumption that **censoring is non-informative**
- In other words, assume that being lost to follow-up is unrelated to prognosis
- If this assumption can't be made, inference becomes more complicated if not impossible
- If the reason a person is lost to follow-up is related to prognosis, then our data is **biased**

Informative Censoring

- Example: Researchers administer a new chemotherapy drug to 10 cancer patients to estimate survival time while on the drug
- 5 patients can’t tolerate the side effects and drop out of the study
- If non-informative censoring were assumed, the drug would probably appear falsely impressive.
- Those who dropped out were probably more ill; hence shorter survival times were disproportionately removed from the sample.

Kaplan-Meier Estimator

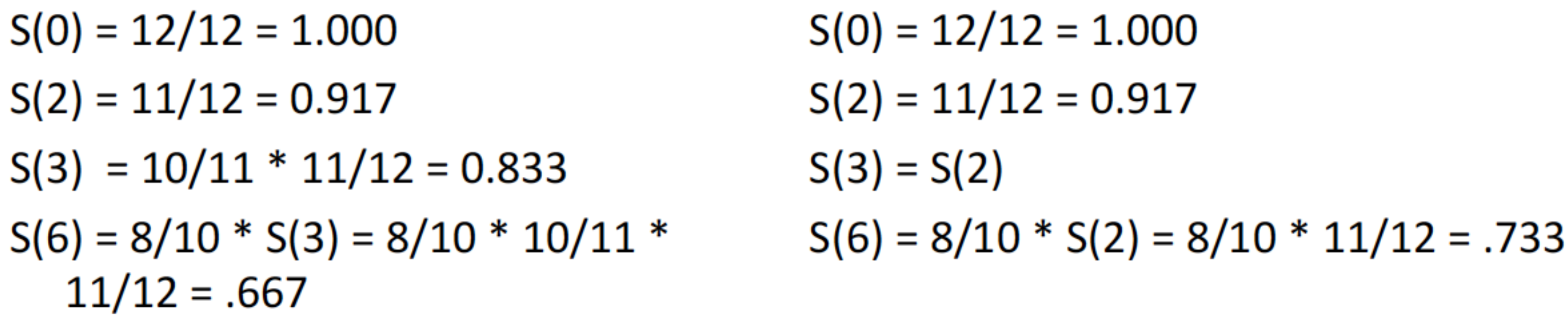
- The Kaplan-Meier method can be modified to account for the partial information about survival times that is available from censored observations
- Example: suppose that, in the sample of 12 mice, mice 2 and 6 have not yet died
- Instead, they are alive after 3 and 10 months of follow-up, respectively, but are lost to follow-up



Mice	Survival (weeks)
1	2
2	3+
3	6
4	6
5	7
6	10+
7	15
8	15
9	16
10	27
11	30
12	32

Kaplan-Meier Estimator

- $S(t_i)$ = P(alive at t_i | alive at t_{i-1}) x $S(t_{i-1})$
- If there were no censoring, P(alive at t_i | alive at t_{i-1}) = (# alive at t_i)/(# alive at t_{i-1})
- However, a patient who is alive but censored at time t_{i-1} never really had a chance to make it to t_i. That patients was not eligible to die during the interval from t_{i-1} to t_i and therefore should not be counted for computing survival rate in this interval.



S(0) = 12/12 = 1.000

S(2) = 11/12 = 0.917

S(3) = 10/11 * 11/12 = 0.833

S(6) = 8/10 * S(3) = 8/10 * 10/11 * 11/12 = .667

S(0) = 12/12 = 1.000

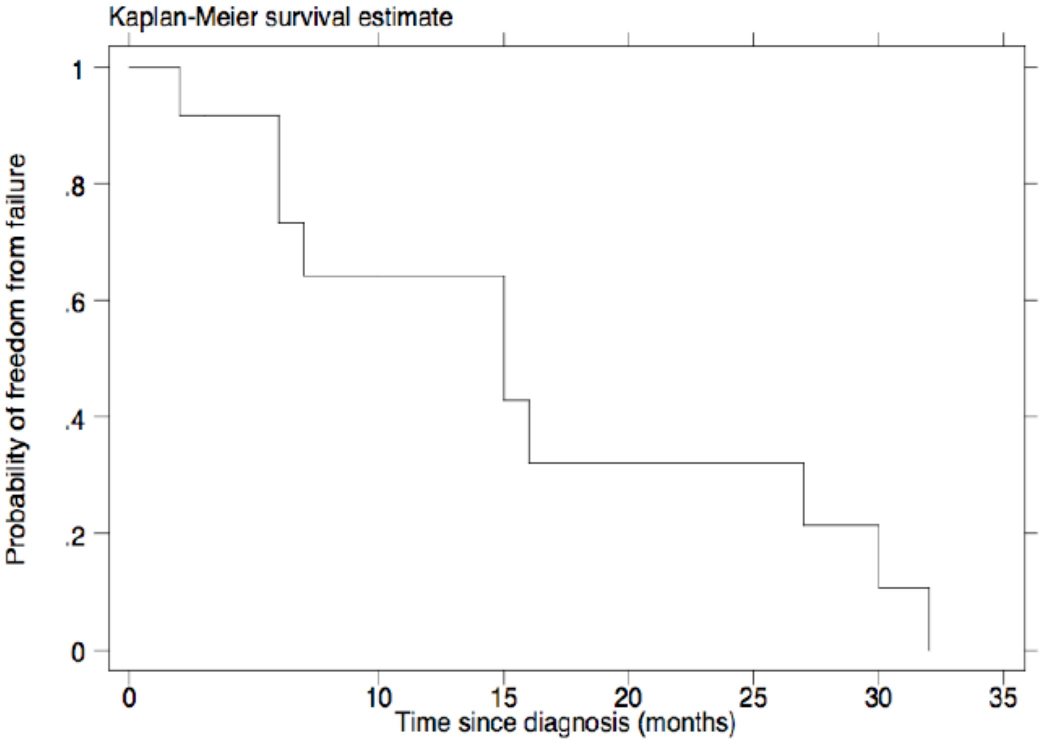
S(2) = 11/12 = 0.917

S(3) = S(2)

S(6) = 8/10 * S(2) = 8/10 * 11/12 = .733

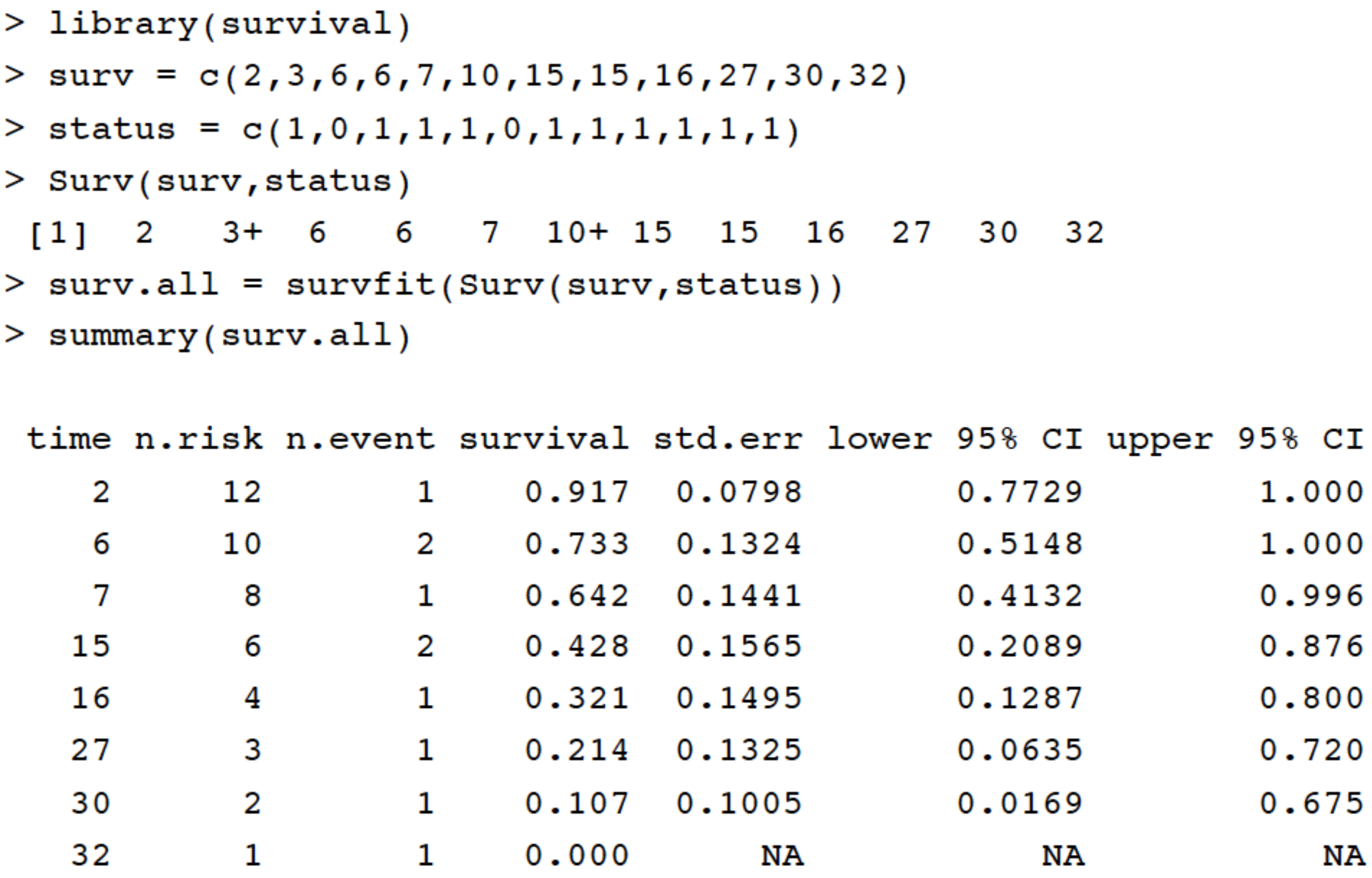
Kaplan-Meier Estimator

- In this case, *S(t)* does not change from its previous value if the observation at the t is censored *S(3)=S(2)=0.917*
- However, the observation is not used to calculate the probability of failure at any subsequent time - it is removed from the denominator



KM estimator in R

- Data: 2, 3+, 6, 6, 7, 10+, 15, 15, 16, 27, 30, 32

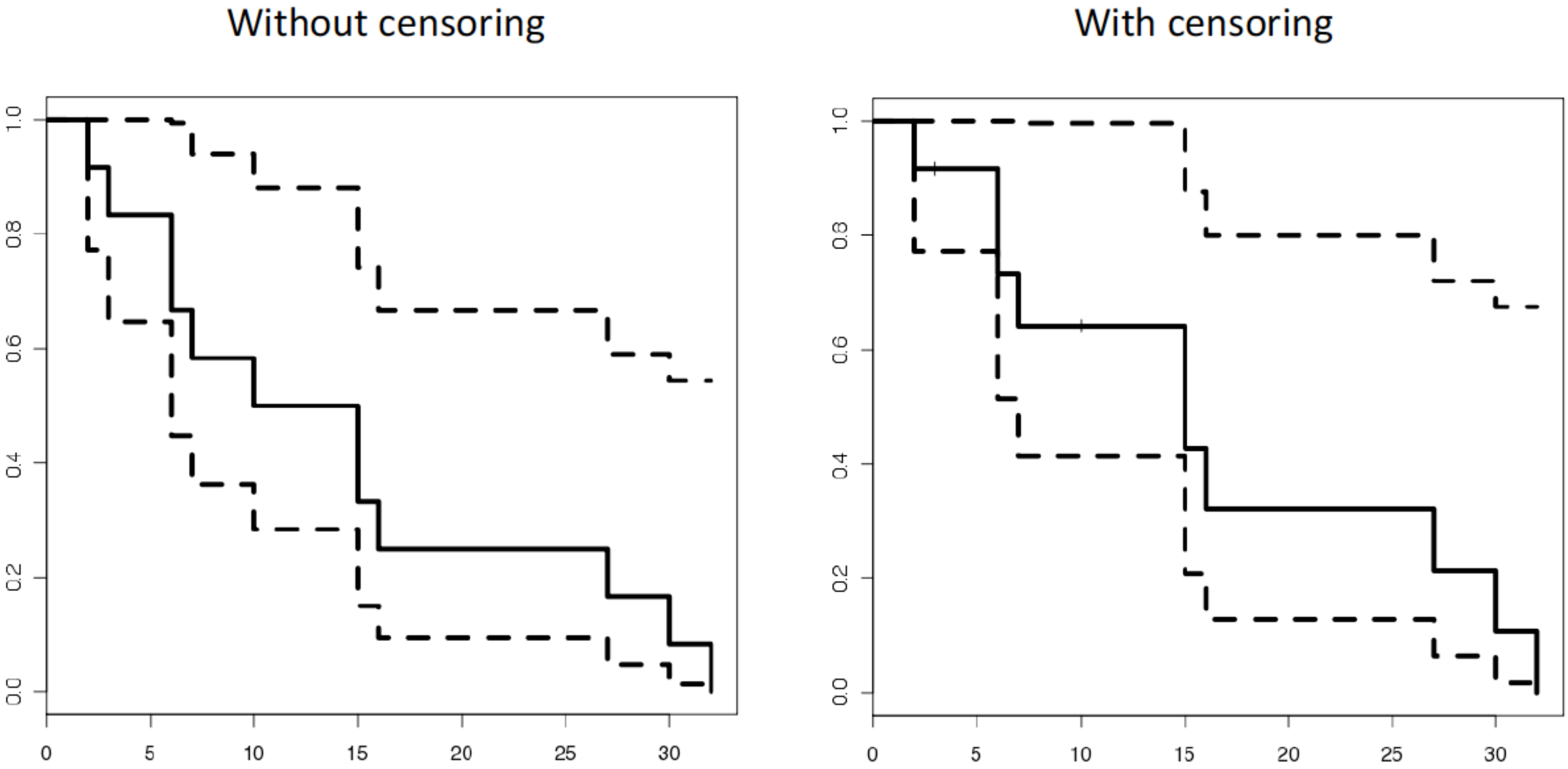


```
> library(survival)
> surv = c(2,3,6,6,7,10,15,15,16,27,30,32)
> status = c(1,0,1,1,1,0,1,1,1,1,1,1)
> Surv(surv,status)
 [1]  2   3+  6   6   7  10+ 15  15  16  27  30  32
> surv.all = survfit(Surv(surv,status))
> summary(surv.all)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2     12       1    0.917  0.0798       0.7729        1.000
    6     10       2    0.733  0.1324       0.5148        1.000
    7      8       1    0.642  0.1441       0.4132        0.996
   15      6       2    0.428  0.1565       0.2089        0.876
   16      4       1    0.321  0.1495       0.1287        0.800
   27      3       1    0.214  0.1325       0.0635        0.720
   30      2       1    0.107  0.1005       0.0169        0.675
   32      1       1    0.000      NA           NA           NA
```

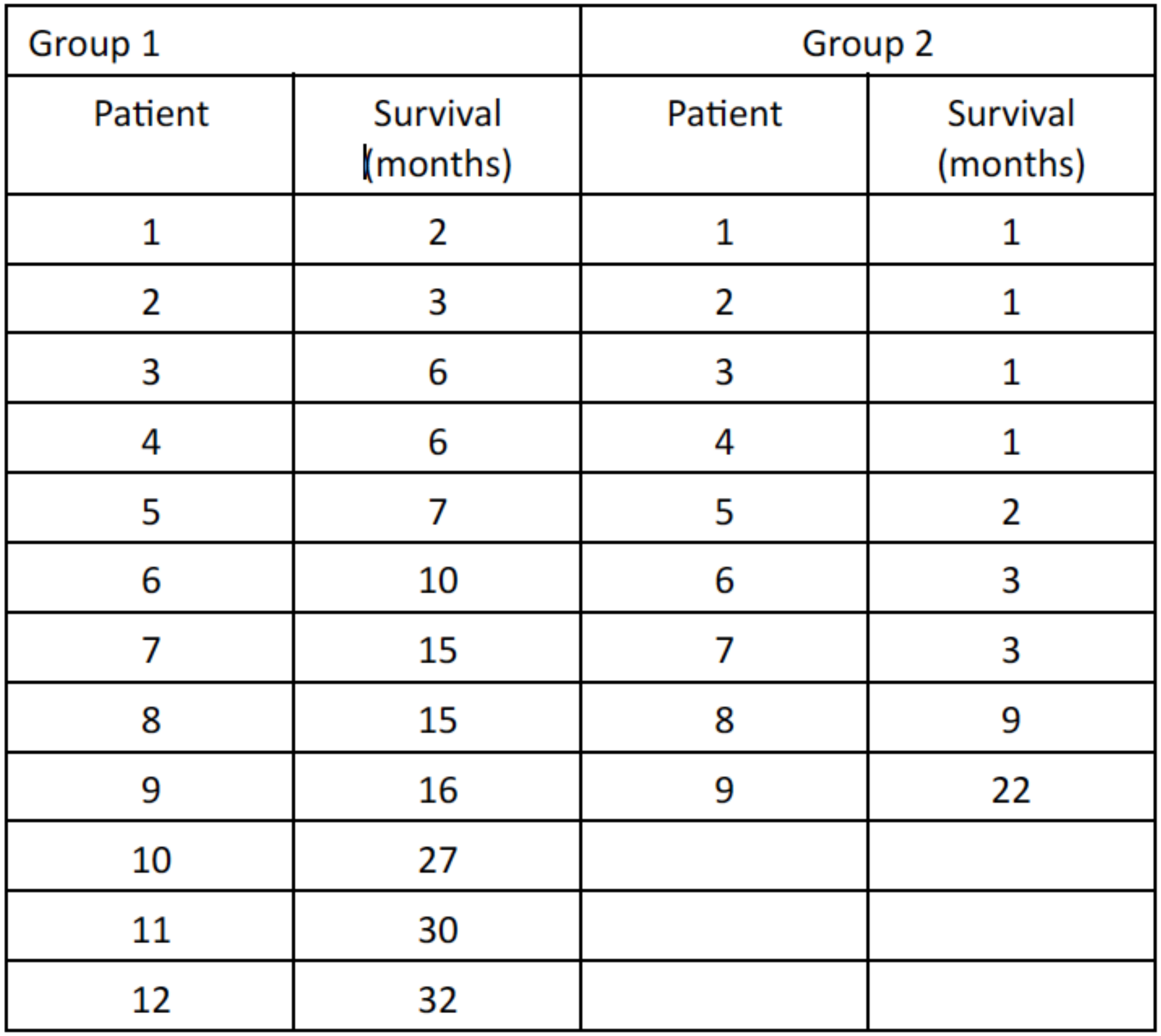
KM estimator in R

```
> plot(surv.all)
```



Log-Rank Test

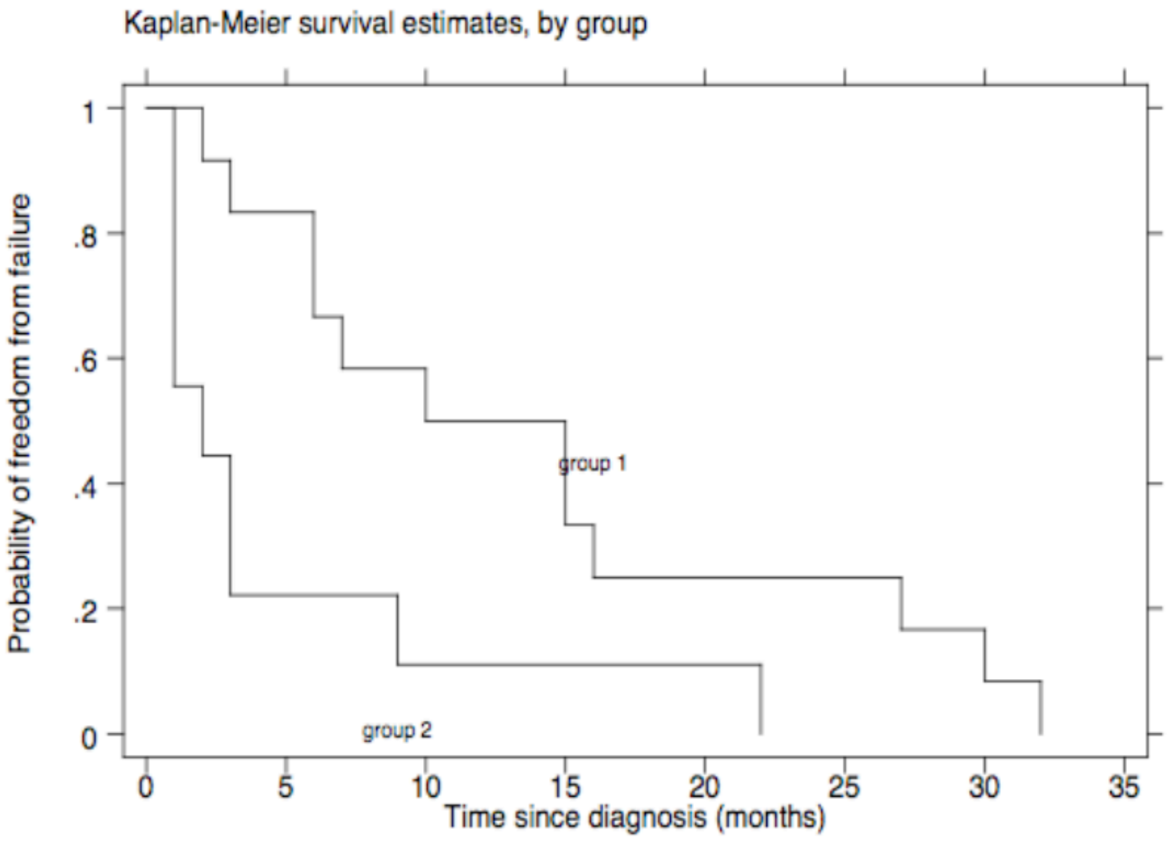
- We often want to compare the distributions of survival times in two (or more) different populations to determine whether survival differs between the groups



Group 1		Group 2	
Patient	Survival (months)	Patient	Survival (months)
1	2	1	1
2	3	2	1
3	6	3	1
4	6	4	1
5	7	5	2
6	10	6	3
7	15	7	3
8	15	8	9
9	16	9	22
10	27		
11	30		
12	32		

Comparison of Two Groups

- It appears that mice in group 1 survive longer than those in group 2



Log-Rank Test

- We can test the null hypothesis that the two (or more) distributions of survival times are identical using a technique called the log-rank test

$$H_0 : S_1(t) = S_2(t)$$

- This test compares the observed number of failures at time t to the expected number of failures (assuming that the two curves are identical), and then accumulates the information over all times
- Under the null hypothesis, the test statistic has a chi-square distribution with 1df
- Since p = 0.013, we reject H_0 and conclude that the two distributions of survival times are not identical

Log-Rank Test

- How does the log-rank test work?
- A direct application of the **Mantel-Haenszel test**, which combines data from a series of 2x2 tables
- Example: exposure/no exposure vs death/no death divided by another variable: sex, genotype, etc.
- The study period is subdivided into *k* intervals. For each interval, a 2x2 table is created. The test statistic is calculated from the *k* tables just as in Mantel-Haenszel
- The only extra thing to worry about is to remove the censored cases in between intervals.

Cox PH Model

- We are often interested in the relationship between survival time and a continuous risk factor, or to evaluate the simultaneous effects of more than one risk factor
- Log-rank test is only for one dichotomous variable
- Multivariable analysis can be performed using the **Cox proportional hazards model**
- Multiple linear regression analysis cannot be used because survival time is rarely normally distributed, and because it cannot account for censored observations
- The Cox model is an example of a **semiparametric** model

Cox PH Model

- We need a new function called the **hazard function, *h(t)***
- This is the probability that you will die in the very instant after time *t*, given that you have survived until time *t*
- The proportional-hazards model assumes that the hazard rate for any individual can be modeled as a function of covariates X_1, …, X_k as follows:

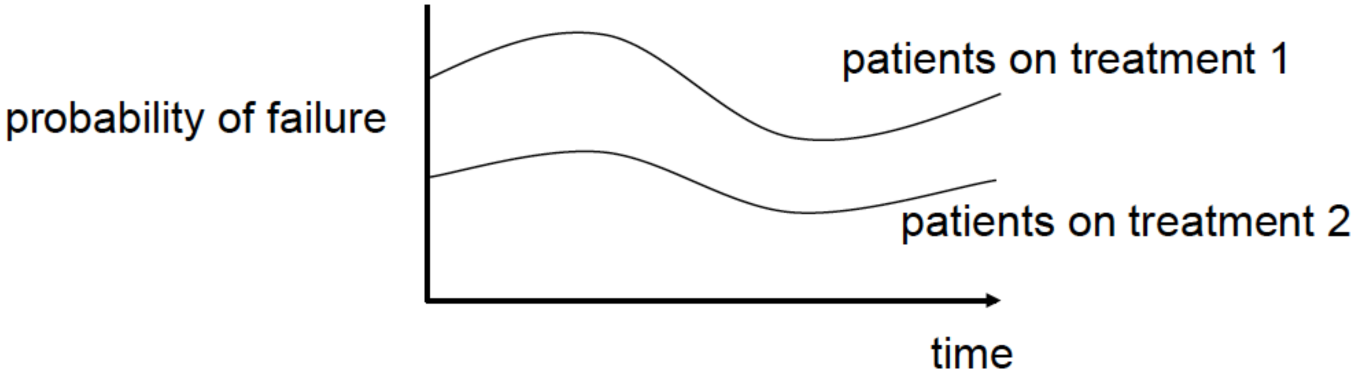
$$h(t) = h_0(t)e^{\beta_1 x_1 + \cdots + \beta_k x_k}$$

$$\ln\left(\frac{h(t)}{h_0(t)}\right) = \beta_1 x_1 + \cdots + \beta_k x_k$$

Cox PH Model

$$\boxed{h(t) = h_0(t)e^{\beta_1 x_1 + \cdots + \beta_k x_k}}$$

- h0(t) is called the “baseline hazard rate”
- We make no assumptions about its shape
- This is why the model is called semi parametric. We don’t completely specify the distribution of survival times; we only specify that changes in covariates will change the hazard rate proportionally to whatever it was.



Interpretation of the Coefficients

- Interpreting the parameters of the model is a bit difficult. The easiest case to understand is when a variable is dichotomous.
- Example: Suppose we are analyzing survival times using a Cox proportional hazards model with covariates X_1 = gender (1=F), X_2 = drug dosage. What is the ratio of hazards between a man and a woman on the same dose of the drug?

$$\frac{h_{woman}(t)}{h_{man}(t)} = \frac{h_0(t)e^{\beta_1(1)+\beta_2 x_2}}{h_0(t)e^{\beta_1(0)+\beta_2 x_2}} = e^{\beta_1}$$

- β_1 is the logarithm of the **"hazard ratio"**, which can be thought of as the instantaneous relative risk of death per unit time of a woman vs. of a man, given that both have survived until time t and with all other covariates held constant

Summary

- **Survival analysis** to handle survival data which usually have censored data points and are non-normally distributed
- **Kaplan-Meier estimator** for estimation & one-sample inference
- **Log-Rank test** for Two-sample comparisons
- **Cox Proportional Hazards model** for regression modeling

Regression Models - Summary

- Binary (disease vs. normal) ➔ Logistic regression (and many others!)
- Discrete
  - Non-ordered (multiple subclasses) ➔ Polytomous regression
  - Ordered (number of recurrences) ➔ Poisson regression
- Continuous (gene expression) ➔ Linear regression
- Censored (patient survival time) ➔ Cox model