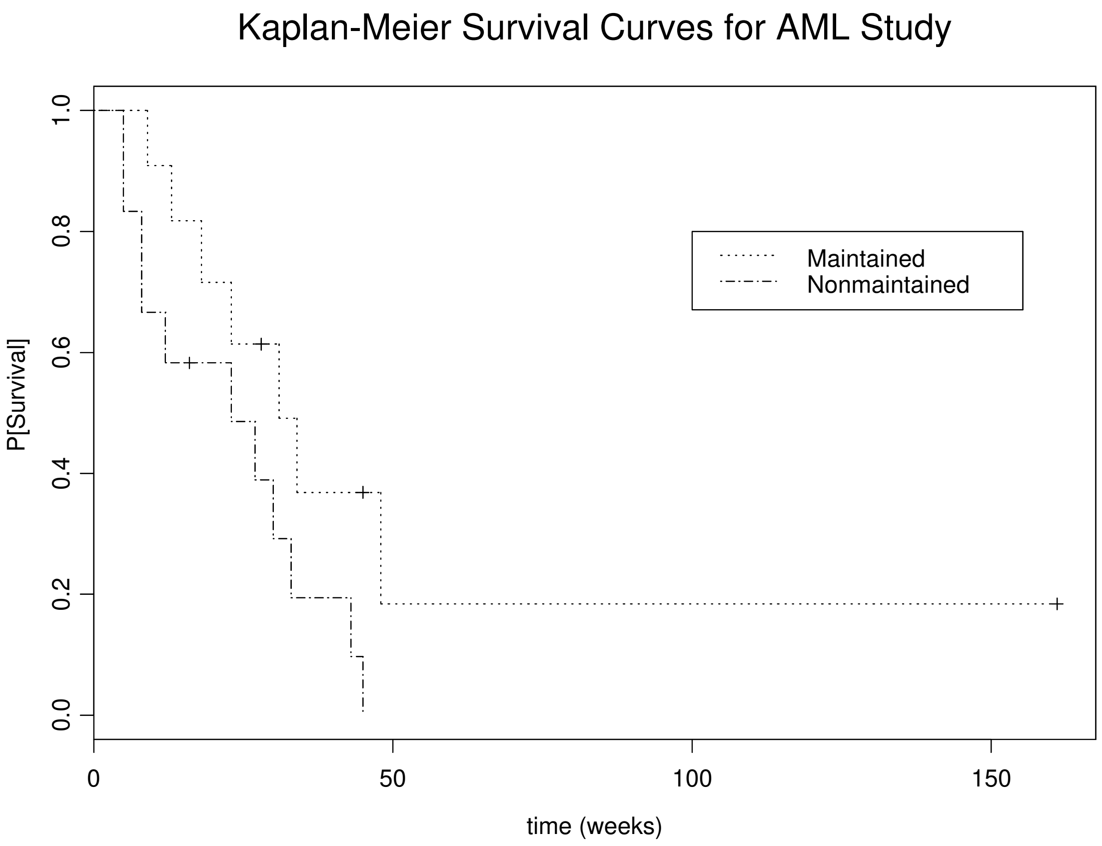
Notes for Thursday April 2, 1998

How does one compare two survival curves?

In the last example from last time we computed and graphed Kaplan-Meier survival curves for leukemia patients that were on a special maintenance program and those that were not.



Did maintenance help? How different are these two curves?

That is a very good question. One answer is to say, just use the Splus function `survdiff`, which we will do shortly. But, to really understand what this is all about, we need a digression.

Digression on some Nonparametric Techniques

Wilcoxon or Mann-Whitney Test

Suppose we have samples from two distributions F and G and we wish to compare the distributions. Let $X_1, \ldots X_m$ and $Y_1, \ldots Y_n$ denote the random samples from the two distributions.

One test of the difference between the distributions proceeds as

1. Rank all the data—i.e., mix the two samples.

2. Let $R(X_i)$ denote the ranks of the F sample in the combined data.
3. Form $S = \sum_{i=1}^{m} R(X_i)$, the rank sum.

Now note that S will be large if the F-distribution has larger values than the G-distribution and vice-versa. It is easy to show that

$$\mathrm{E}(S) = \frac{m(n+m+1)}{2}$$

and

$$\mathrm{Var}(S) = \frac{m(n+m+1)n}{2}$$

(This comes from assuming that $F = G$ in which case the ranks are just sampled form a discrete uniform distribution.

For small values of n and m there are tables of exact tail probabilities for S (or some linear transformation of it).

For large samples, just take a normal approximation

$$z = \frac{S - \mathrm{E}(S)}{\sqrt{\mathrm{Var}(S)}}$$

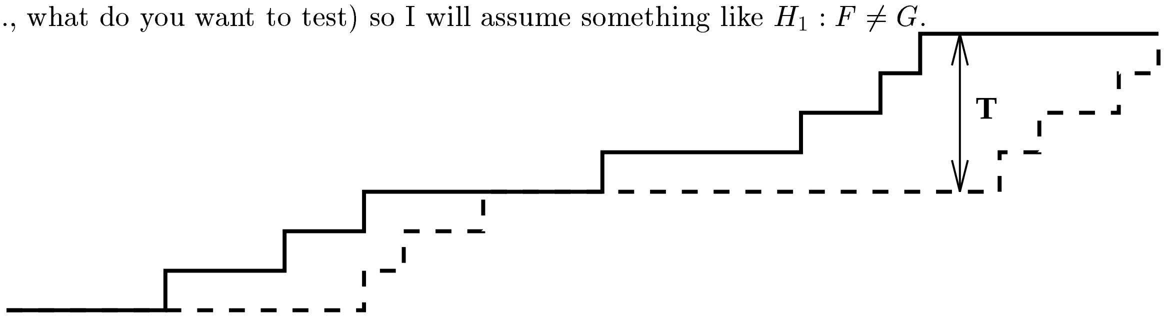
The test makes minimal assumptions and works quite well in certain situations? However, what is the test sensitive to? When normality can be assumed, the 2-sample t-test is a competitor and like most specific purpose tools does better than the Wilcoxon test.

You can find this test as `wilcox.test` in Splus.

Kolmogorov-Smirnov test

The previous test is not very sensitive to differences in the distribution shape, as opposed to solely location differences. A test which is sometimes useful here is the (Kolmogorov-)Smirnov test.

Let $\hat{F}$ and $\hat{G}$ be estimates of two distribution functions. (For censored data one might think about using the Kaplan-Meier estimates). The exact form of the test depends on the alternate hypothesis (i.e., what do you want to test) so I will assume something like $H_1 : F \neq G$.



Now form $T = \max|F - G|$, and look up the results in a table, OR just use the Splus function `ks.gof`.

The test is sensitive to all sorts of departures, particularly changes in tails and different variances.

End Digression

So how does one compare two Kaplan-Meier estimates? One of the things that `survdiff` does is (a version) of Gehan's test, which itself is just a modification of the Wilcoxon test.

Another way of looking at the ranksum S in the Wilcoxon test is to consider

$$U(X_i, Y_j) = \begin{cases} +1 & \text{if} \quad X_i > Y_j \\ 0 & \text{if} \quad X_i = Y_j \\ -1 & \text{if} \quad X_i < Y_j \end{cases}$$

Let $U = \sum_{i=1}^{m} \sum_{j=1}^{n} U_{i,j}$, which can be read as: the number of times that $X > Y$, minus the number of times that $X < Y$. (Would $\sum_{i=1}^{m} \sum_{j=1}^{n} |U_{i,j}|$ make sense?). It isn't hard to see that

$$\begin{aligned} S &= \{\text{sum of the ranks of the } X_i\text{'s when } X\text{'s and } Y\text{'s are ranked together}\} \\ &= \sum_{i=1}^{m} R(X_i) \\ &= \sum_{i=1}^{m} \{1 + (\text{number of } X's \text{ and } Y's \text{ less than } X_i)\} \\ &= \sum_{i=1}^{m} \{n + m - (\text{number of } X's \text{ and } Y's \text{ greater than } X_i)\} \\ &= \frac{1}{2}(S + S) \;=\; \frac{1}{2}\left[m(m+n+1) + U\right], \qquad \text{(after some algebra).} \end{aligned}$$

If X and Y are independent samples from the same distribution, it follows from the expressions derived earlier for $E(S)$ and $\text{Var}(S)$ that $E(U) = 0$, and $\text{Var}(U) = m(m+n+1)n/8$.

For censored data, the Gehan test suggests defining $U_{i,j}$ in a different form.

Our basic data is (X_i, δ_i) and (Y_i, ϵ_i) where δ and ϵ are the indicators of censoring.

The null hypothesis is $F_1 = F_2$ (distribution) and $G_1 = G_2$ (censoring), with alternate interest in $F_1 \neq F_2$. (I am not sure what happens when the censoring is different).

Now define

$$U_{i,j} = \begin{cases} +1 & \text{if} & \begin{array}{l} \text{we know that } t_i > u_j \\ \text{i.e., } (x_i > y_j, \epsilon_j = 1) \text{ or } (x_i = y_j, \delta_i = 0, \epsilon_i = 1) \end{array} \\ 0 & \text{otherwise} & \\ -1 & \text{if} & \begin{array}{l} \text{we know that } t_i < u_j \\ \text{i.e., } (x_i < y_j, \delta_j = 1) \text{ or } (x_i = y_j, \delta_i = 1, \epsilon_i = 0) \end{array} \end{cases}$$

and $U = \sum_{i=1}^{m} \sum_{j=1}^{n} U_{i,j}$

It is possible to show (via some permutation theory) that $\text{E}(U) = 0$ and $\text{Var}(U) =$ (a big mess), but there is a computational formula, and packages.

Digression on the Mantel-Haenszel Test

Suppose there are two populations and each individual in each population can have one of two characteristics. For example, two populations of cancer patients with different treatments who can either survive beyond a year or die within the year. We could write out the table of data as,

	Dead	Alive	
Pop1	a	b	n_1
Pop2	c	d	n_2
	m_1	m_2	n

Let

$$\begin{aligned} p_1 &= P(\text{Dead}|\text{Pop1}) \\ p_2 &= P(\text{Dead}|\text{Pop2}) \end{aligned}$$

We could test for $p_1 = p_2$ by investigating the interaction in the log-linear model `Count ~ Pop + Status + Pop:Status` (i.e., test for independence or monotone homogeneity). Alternatively we could look directly at the statistic

$$\chi^2 = \left[\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}\right]^2 = \frac{n(ad-bc)^2}{n_1 n_2 m_1 m_2}$$

where $\hat{p}_1 = \frac{a}{n_1}$, $\hat{p}_2 = \frac{c}{n_2}$, and $\hat{p} = \frac{m_1}{n}$. The χ^2 is approximately distributed as a χ_1^2 random variable. The intuition for this test comes (again) from the 2-sample t-test, but the test itself is derived from a hypergeometric distribution for the contingency table.

$$P(A = a) = \frac{\binom{n_1}{a}\binom{n_2}{m_1 - a}}{\binom{n}{m_1}}$$

This test assumes that the n and m margins are BOTH fixed in advance.

Under the hypergeometric distribution and the null hypothesis that $p_1 = p_2$ (i.e. independence),

$$\mathrm{E}(A) = \frac{n_1 m_1}{n} \quad \text{and} \quad \mathrm{Var}(A) = \frac{n_1 n_2 m_1 m_2}{n^2(n-1)}$$

Some algebra shows that

$$ad - bc = n(a - \mathrm{E}(A))$$

and thus

$$\chi^2 = \frac{n}{n-1}\left[\frac{a - \mathrm{E}(A)}{\sqrt{\mathrm{Var}(A)}}\right]^2$$

For the example above, the hypergeometric assumption is suspect. If I collected the data by finding m_1 deaths and m_2 survivors and then recorded their populations, then it is probably OK. On the other hand, if I just start out with two populations and watch for deaths a binomial (or product binomial) seem more reasonable. But the particular form of the χ^2 test is (I think) just the same as the binomial one would have been, modulo a fudge about n or $n-1$.

Continue Digression, a sequence of 2×2 tables

The same setup as before, but now suppose we look at the two patient populations in each of K hospitals. Because there may be differences between hospitals we don't wish to combine all the 2×2 tables. Rather we would like to test

$$p_1^1 = p_2^1, p_1^2 = p_2^2, \ldots, p_1^k = p_2^k,$$

i.e., equality in each population, but not overall equality.

The data are now

	Dead	Alive	
Pop1	a_1	b_1	n_{11}
Pop2	c_1	d_1	n_{12}
	m_{11}	m_{12}	n_1

$\ldots$

	Dead	Alive	
Pop1	a_k	b_k	n_{k1}
Pop2	c_k	d_k	n_{k2}
	m_{k1}	m_{k2}	n_k

The Mantel-Haenszel statistic for this problem is

$$\text{MH} = \frac{\sum_{i=1}^{k}(a_i - \text{E}(A_i))}{\sqrt{\sum_{i=1}^{k}(\text{Var}(A_i))}}$$

If the tables are independent then MH tends towards a $N(0,1)$ when either k is fixed and $n_i \to \infty$, or $k \to \infty$ and the tables are iid.

When used in survival analysis, neither of these conditions for asymptotic normality holds—but the result still does.

> *Digression in a digression:* Think for a moment about modeling the probability of death with a logistic regression model like `cbind(dead,alive) ~ Hosp + Pop + Hosp:Pop`, where `Hosp` is a factor with K levels and `Pop` is 0 for Pop1 and 1 for Pop2. As we have seen before, if the interaction term `Hosp:Pop` is absent, the estimated coefficient for `Pop` in this model is the log-odds-ratio
>
> $$\hat{\beta}_{\text{Pop}} = \log\left[\frac{P[\text{Dead}|\text{Pop2}]/(1 - P[\text{Dead}|\text{Pop2})}{P[\text{Dead}|\text{Pop1}]/(1 - P[\text{Dead}|\text{Pop1})}\right]$$
>
> This suggests a way of using logistic regression to get at the same thing the MH test is getting at: test to drop the interaction `Hosp:Pop` to see whether the assumption of a common odds ratio across all tables is valid, and if you can drop it then test to see if the coefficient $\beta_{\text{Pop}} = 0$ to see if the survival is independent of population, across hospitals.
>
> Of course, you end up with slightly different tests using the logistic regression approach than you do with the above MH calculations, but the differences are small for large data sets and the spirit is the same.

End Digression

Mantel-Haenszel in Survival Analysis

In survival analysis this test is used in the following manner. Let

$$X_1, \ldots, X_n$$
$$\delta_1, \ldots, \delta_n$$

be one population and

$$Y_1, \ldots, Y_n$$
$$\epsilon_1, \ldots, \epsilon_n$$

be the other. Consider $Z_{(1)}, \ldots, Z_{(n+m)}$ to be the combined ordered populations. For each $Z_{(i)}$ which corresponds to a death (in either population) construct a 2×2 table and compute the MH test for this sequence of tables. (E.G., at time (i) we have

	Dead	Alive	
Pop1	a	b	n_1
Pop2	c	d	n_2

and at time (i+1) we have

	Dead	Alive	
Pop1	$a+1$	$b-1$	n_1
Pop2	c	d	n_2

etc.,)

In other words, test $F_1 = F_2$ (which is the same as $S_1 = S_2$) by using the MH test. (Independence of tables is violated here, but asymptotic normality of the test can still be pushed through).

Example Continued: The AML Study

In the Leukemia survival study we considered previously, we wanted to know whether the two survival curves, for patients on maintenance, and not on maintenance, were distinct.

The `survdiff` function in SPLUS implements a variety of weighted Mantel-Haenszel style tests for looking at the difference between such estimated survival curves.

The overall weighted difference is $\sum_1^K w_i(O_i - E_i)$, with variance $\sum_i w_i^2 V_i$. It can be shown that when $w_i = 1$ this is the Mantel-Haenszel test (a.k.a. the *log-rank* test in the survival literature), for $w_i = n_i$ it is the Gehan/Wilcoxon test, and for $w_i = \hat{S}(t_i)$ it is the Peto-Peto modification of the Gehan/Wilcoxon test. A single parameter $\rho \in [0, 1]$ in `survdiff` controls the weights given to different survival times; $\rho = 0$ (the default) yields the log-rank test and $\rho = 1$ yields the Peto-Peto modification of Gehan/Wilcoxon. Other values give a test that is intermediate to these two.

```
402 > # The Mantel-Haenszel (log-rank) test:

402 > survdiff(Surv(time, status) ~ group, data = leukemia,rho=0)
Call:
```

```
survdiff(formula = Surv(time, status) ~ group, data = leukemia, rho = 0)

                     N Observed Expected (O-E)^2/E (O-E)^2/V
   group=Maintained 11        7    10.69      1.27       3.4
group=Nonmaintained 12       11     7.31      1.86       3.4

 Chisq= 3.4  on 1 degrees of freedom, p= 0.0653

402 > # The Peto/Gehan/Wilcoxon test

402 > survdiff(Surv(time, status) ~ group, data = leukemia,rho=1)
Call:
survdiff(formula = Surv(time, status) ~ group, data = leukemia, rho = 1)

                     N Observed Expected (O-E)^2/E (O-E)^2/V
   group=Maintained 11     3.85     6.14     0.859      2.78
group=Nonmaintained 12     7.18     4.88     1.081      2.78

 Chisq= 2.8  on 1 degrees of freedom, p= 0.0955
```

The Mantel-Haenszel test can be show to be most powerful against a "proportional hazards" model; in this case there is mild evidence of a difference in survival between the leukemia patientes on maintenance and those that were not on maintenance.

The Cox Proportional Hazards Model

The Cox proportional hazards model is the most commonly used regression model for survival data. It allows not only the fitting of different survival curves to different groups, like the Kaplan-Meier product-limit estimate, but also allows "regression" of the survival curves onto other covariates, that also may be changing with time.

Let $Z_i(t)$ be a vector of covariates (predictors in a regression) for the i^{th} individual at time t. The proportional hazards model is of the form

$$\lambda(t; Z_i) = \lambda_0(t) r_i(t)$$

where

$$r_i(t) = e^{\beta' Z_i(t)}$$

is referred to as the *risk score* for the i^{th} subject, β is a vector of regression coefficients, and $\lambda_0(t)$ is an arbitrary and unspecified baseline hazard function. Thus the fitted survival curves are of the form

$$S(t; Z_i) = \exp\left[-\int_0^t \lambda(s; Z_i) ds\right]$$

The regression coefficients β do not include an intercept term, which is absorbed into $\lambda_0(t)$.

Now, assume that a death occurs as time t^*. Conditional on this death occurring, it can be shown that the likelihood that it would be subject i rather than some other subject is

$$L_i(\beta) = \frac{\lambda_0(t^*) r_i(t^*)}{\sum_j Y_j(t^*)\lambda_0(t^*) r_j(t^*)} = \frac{r_i(t^*)}{\sum_j Y_j(t^*) r_j(t^*)}$$

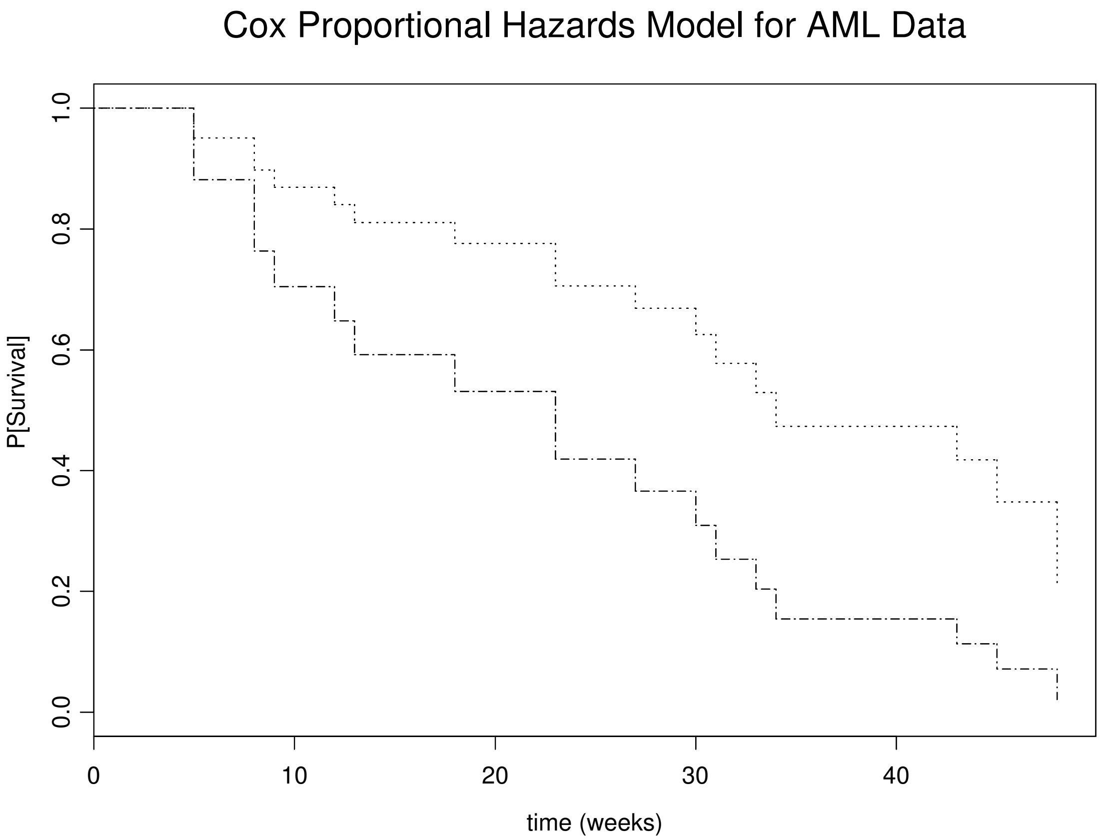
where $Y_j(t) = 1$ if subject j is still at risk and hasn't died yet at time t, and $= 0$ otherwise. The product of $L_i(\beta)$ over all death times, $L(\beta) = \prod_i L_i(\beta)$, was called a *partial likelihood* by Cox (1992); maximizing $L(\beta)$ gives an estimate of β without the need to estimate the nusiance parameter $\lambda_0(t)$ *(when was the last time you saw a whole* **function** *being called a parameter, and just a "nuisance parameter" at that??).* $\lambda_0(t)$ is estimated separately using a method similar to the likelihood version of the Kaplan-Meier estimator. The Cox proportional hazards model is thus a *semiparametric* model: $\lambda_0(t)$ is estimated *non-parametrically* using Kaplan-Meier, and β is estimated *parametrically.*

Example continued: The AML study

SPLUS has a function, `coxph`, for fitting proportional hazards models.

```
402 > leukemia.ph _ coxph(Surv(time,status) ~ group,data=leukemia)
402 > summary(survfit(leukemia.ph,newdata=
+ data.frame(group=unique(leukemia$group))))
Call: survfit.coxph(object = leukemia.ph, newdata = data.frame(group = unique(leuke
mia$
        group)))

 time n.risk n.event survival1 survival2
    5     23       2     0.951    0.8817
    8     21       2     0.898    0.7634
    9     19       1     0.869    0.7043
   12     18       1     0.841    0.6480
   13     17       1     0.811    0.5917
   18     14       1     0.776    0.5308
   23     13       2     0.706    0.4188
   27     11       1     0.669    0.3658
   30      9       1     0.625    0.3097
   31      8       1     0.577    0.2536
   33      7       1     0.529    0.2040
   34      6       1     0.474    0.1546
   43      5       1     0.418    0.1131
   45      4       1     0.348    0.0718
   48      2       1     0.211    0.0206
402 > plot(survfit(leukemia.ph,newdata=
+ data.frame(group=unique(leukemia$group))),lty=2:3)
402 > title("Cox Proportional Hazards Model for AML Data",xlab="time (weeks)",
+ ylab="P[Survival]")
```



Fitting a Cox model in SPLUS also automatically produces three tests of hypothesis, which are asymptotically equivalent but not always in practice. Let β_0 be the initial value of the coefficients and $\hat{\beta}$ the solution after fitting the model.

The *likelihood ratio test* is defined as

$$-2\{\log L(\beta_0) - \log L(\hat{\beta})\}$$

and is the most reliable. It is produced by simply printing out a `coxph` object.

```
402 > leukemia.ph
Call:
coxph(formula = Surv(time, status) ~ group, data = leukemia)


       coef exp(coef) se(coef)    z     p
group 0.458      1.58    0.256 1.79 0.074

Likelihood ratio test=3.38  on 1 df, p=0.0658  n= 23
```

The *Wald statistic*, $(\hat{\beta} - \beta_0)'\hat{\Sigma}_{\hat{\beta}}^{-1}(\hat{\beta} - \beta_0)$, where $\Sigma_{\hat{\beta}}^{-1}$ is the estimated variance-covariance matrix, is perhaps the most natural because it provides a per variable test rather than an overall measure of significance. The *score test* is defined as $U'IU$ where the vector U is the so-called score function

$\partial/\partial\beta \log L(\beta)$, and I is the information matrix for the model, both evaluated at β_0. The score test does not require iteration and, consequently, is more computationally efficient if a large number of models are to be tested. These are produced by asking for a summary of a `coxph` object.

```
402 > summary(leukemia.ph)
Call:
coxph(formula = Surv(time, status) ~ group, data = leukemia)

  n= 23

       coef exp(coef) se(coef)    z     p
group 0.458      1.58    0.256 1.79 0.074

      exp(coef) exp(-coef) lower .95 upper .95
group      1.58      0.633     0.957      2.61

Rsquare= 0.137   (max possible= 0.976 )
Likelihood ratio test= 3.38  on 1 df,   p=0.0658
Wald test            = 3.2  on 1 df,   p=0.0737
Efficient score test = 3.42  on 1 df,   p=0.0645
```

To compare different models you can compute the likelihood ratio statistic by hand. For example,

```
402 > leukemia0.ph _ coxph(Surv(time,status) ~ 1,data=leukemia)
402 > names(leukemia.ph)
 [1] "coefficients"      "var"               "loglik"
 [4] "score"             "iter"              "linear.predictors"
 [7] "residuals"         "means"             "method"
[10] "n"                 "terms"             "assign"
[13] "y"                 "formula"           "call"
```

The `loglik` component of the fit is a vector of the log likelihoods for two fits. The null model (intercept only) is the first value, and the current model is the second value.

```
402 > -2*(leukemia0.ph$loglik[2] - leukemia.ph$loglik[2])
[1]  3.384447
402 > 1 - pchisq(3.38,1)
[1] 0.06599206
```

In this case, since the smaller model was the "null model" with $\beta_0 = 0$ anyway, the explicit calculation `-2*(leukemia0.ph$loglik[2] - leukemia.ph$loglik[2])` gives the same likelihood ratio result as before. When comparing more complicated nested models, one has to resort to explicit calculation since the summary printout generally does not do enough.

Example: Ovarian Cancer

This example uses data from a study of ovarian cancer. The variables are:

`futime`... The number of days from enrollment until death or censoring, whichever comes first.
`fustat`... An indicator of death (1) or censoring (0). age The patient age in years (actually, the age in days divided by 365.25)
`residual.dz`... An indicator of the extent of residual disease (1 or 2).
`rx`... An indicator of the treatment given (1 or 2).
`ecog.ps`... A measure of performance score or functional status, using the Eastern Cooperative Oncology Group's scale. It ranges from 0 (fully functional) to 4 (completely disabled). Level 4 subjects are usually considered too ill to enter a randomized trial such as this. In this study only 1's and 2's were recorded.

The data frame `ovarian` comes with SPLUS. A summary produces the following:

```
402 > summary(ovarian)
     futime           fustat             age          residual.dz
 Min.   :  59.0   Min.   :0.0000   Min.   :38.89   Min.   :1.000
 1st Qu.: 368.0   1st Qu.:0.0000   1st Qu.:50.17   1st Qu.:1.000
 Median : 476.0   Median :0.0000   Median :56.85   Median :2.000
 Mean   : 599.5   Mean   :0.4615   Mean   :56.17   Mean   :1.577
 3rd Qu.: 794.8   3rd Qu.:1.0000   3rd Qu.:62.38   3rd Qu.:2.000
 Max.   :1227.0   Max.   :1.0000   Max.   :74.50   Max.   :2.000
       rx         ecog.ps
 Min.   :1.0   Min.   :1.000
 1st Qu.:1.0   1st Qu.:1.000
 Median :1.5   Median :1.000
 Mean   :1.5   Mean   :1.462
 3rd Qu.:2.0   3rd Qu.:2.000
 Max.   :2.0   Max.   :2.000
402 > sapply(ovarian,unique)
$futime:
 [1]   59  115  156  421  431  448  464  475  477  563  638  744  769  770  803
[16]  855 1040 1106 1129 1206 1227  268  329  353  365  377

$fustat:
[1] 1 0

$age:
 [1] 72.3315 74.4932 66.4658 53.3644 50.3397 56.4301 56.9370 59.8548 64.1753
[10] 55.1781 56.7562 50.1096 59.6301 57.0521 39.2712 43.1233 38.8932 44.6000
[19] 53.9068 44.2055 59.5890 74.5041 43.1370 63.2192 64.4247 58.3096

$residual.dz:
```

```
[1] 2 1

$rx:
[1] 1 2

$ecog.ps:
[1] 1 2
```

We begin by modeling survival as a function of age only.

```
402 > ov.fit1 <- coxph(Surv(futime, fustat) ~ age, ovarian)
402 > summary(ov.fit1)
Call:
coxph(formula = Surv(futime, fustat) ~ age, data = ovarian)

  n= 26

     coef exp(coef) se(coef)    z      p
age 0.162      1.18   0.0497 3.25 0.0012

    exp(coef) exp(-coef) lower .95 upper .95
age      1.18      0.851      1.07       1.3

Rsquare= 0.423   (max possible= 0.932 )
Likelihood ratio test= 14.3  on 1 df,   p=0.000156
Wald test            = 10.6  on 1 df,   p=0.00116
Efficient score test = 12.3  on 1 df,   p=0.000463
```

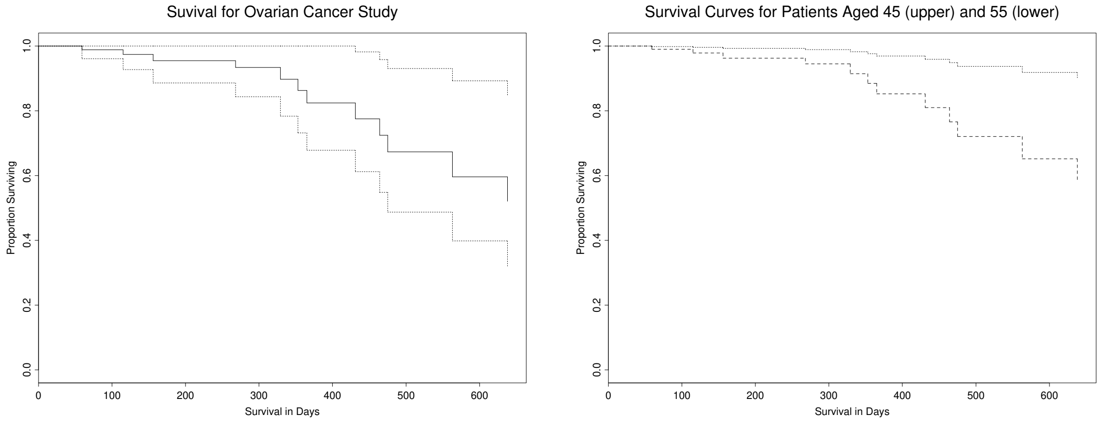
Printing the resulting fit produces the estimated coefficient $\hat{\beta}$, the estimated relative risk for a one unit change in the variable $e^{\hat{\beta}}$, the standard error of the estimated coefficient $se(\hat{\beta})$, a z-test $\hat{\beta}/se(\hat{\beta})$ along with its p-value for the significance of the estimated coefficient, and the three tests for goodness of fit. The z-test is sometimes referred to as Wald's test. An estimate of the relative risk of dying of ovarian cancer for two patients in the study differing in age by one year is 1.18, which is significantly larger than one ($p = 0.000156$). The older patient has an estimated 1.18 times higher risk of dying of ovarian cancer than the younger patient.

If we plot the survival curve with no `newdata` argument, we get an estimate of survival for a patient with age equal to the average, 56.17.

```
402 > plot(survfit(ov.fit1), xlab = "Survival in Days",
+ ylab = "Proportion Surviving")
402 > title("Suvival for Ovarian Cancer Study")
```

The default when printing one curve is to also include 95% confidence bands. If we request specific ages with the `newdata` argument, we get just those survival curves, with no confidence bands.

```
402 > plot(survfit(ov.fit1,newdata=data.frame(age=c(45,55))),
+ xlab="Survival in Days",ylab="Proportion Surviving",lty=2:3)
402 > title("Survival Curves for Patients Aged 45 (upper) and 55 (lower)")
```



It is clear from summary(ov.fit1) that age is an important predictor of survival. Let's add the other predictors to the model.

```
402 > ov.fit2 <- coxph(Surv(futime, fustat) ~ age +
+ residual.dz + rx + ecog.ps, ovarian)
402 > summary(ov.fit2)
Call:
coxph(formula = Surv(futime, fustat) ~ age + residual.dz + rx + ecog.ps, data
        = ovarian)
  n= 26
              coef exp(coef) se(coef)      z      p
        age  0.125     1.133   0.0469  2.662 0.0078
residual.dz  0.826     2.285   0.7896  1.046 0.3000
         rx -0.914     0.401   0.6533 -1.400 0.1600
    ecog.ps  0.336     1.400   0.6439  0.522 0.6000

            exp(coef) exp(-coef) lower .95 upper .95
        age     1.133      0.883     1.033      1.24
residual.dz     2.285      0.438     0.486     10.74
         rx     0.401      2.496     0.111      1.44
    ecog.ps     1.400      0.714     0.396      4.94

Rsquare= 0.481   (max possible= 0.932 )
Likelihood ratio test= 17  on 4 df,   p=0.0019
Wald test            = 14.2  on 4 df,   p=0.00654
Efficient score test = 20.8  on 4 df,   p=0.000345
```

To check for an overall improved fit over the age only model compute the likelihood ratio test between the models "by hand", as we did before. Note that this likelihood ratio test is not printed out by the summaries.

```
402 > -2*(ov.fit1$loglik[2] - ov.fit2$loglik[2])
[1] 2.749708
```

Noting that there is a difference of three degrees of freedom between the models, the p-value for the likelihood ratio test is computed as follows:

```
402 > 1-pchisq(2.75, df = 3)
[1] 0.4317971
```

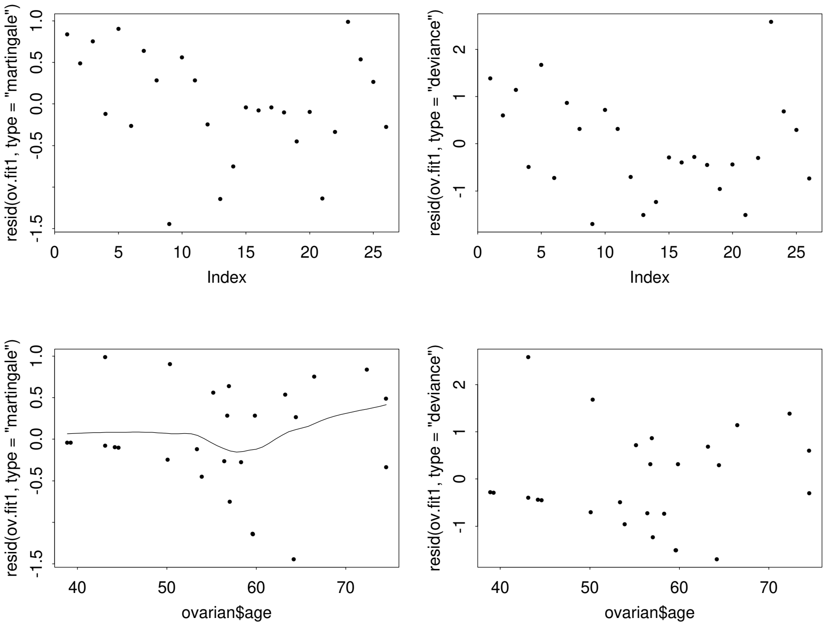
There is no significant difference between the two models indicating that `residual.dz`, `rx`, and `ecog.ps` don't improve the fit.

It is also possible to make residual plots from Cox regression models.

- The *martingale* residuals are obtained from viewing the deviation between data and model in a survival curve problem as a martingale[1].
- The *deviance* residuals are a transformation of the martingale residuals to something that should have approximately a normal distribution.

and there are other kinds of residuals possible too. We briefly illustrate plots using deviance and martingale residuals.

```
402 > par(mfrow=c(2,2))
402 > plot(resid(ov.fit1,type="martingale"))
402 > plot(resid(ov.fit1,type="deviance"))
402 > scatter.smooth(ovarian$age,resid(ov.fit1,type="martingale"))
402 > plot(ovarian$age,resid(ov.fit1,type="deviance"))
```



[1]A "martingale" is a stochastic process, $X_1, X_2, X_3, \ldots$, characterized by the property $E[X_n|X_{n-1}, X_{n-2}, \ldots, X_1] = X_{n-1}$. Martingales are usually the type of stochastic process you learn about after you know all about Markov processes. Martingales are also important in the theory of finance.

A trend in the martingale residuals vs a covariate often suggests what the functional form of the regression on that covariate should be. In this case, there appears to be some slight quadratic trend, if anything.

One can also assess influential points by asking for `resids(ov.fit1,type="dfbetas")` as well.

Additional references:

- Fleming, T. and Harrington, D. (1991). *Counting Processes and Survival Analysis.* Wiley, New York.

  Update of Kalbfleisch and Prentice, discusses the partial likelihood approach and the more modern "counting processes" perspective on survival analysis in detail.

- Therneau, T. M. and Grambsch, P. M. and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika,* 77, 147–160.

  A modern treatment of residuals in survival analysis.