**Regression Modeling of Survival Time Data**

**Why regression models?**

- Groups similar except for the treatment under study – use the nonparametric methods discussed earlier.
- Groups differ in variables (covariates) that may affect the outcome. Compare groups after adjusting for the effects of the covariates.
- Predict the distribution of survival time based on a set of covariates.

**Issue**: Learn how the distribution of T depends on p covariates $x_1, x_2, \ldots, x_p$.

- Fitted values of T must be positive.
- Distribution of T is generally skewed.
- Some of the T observations may be censored.
- Represent a nominal x with m levels through m – 1 design variables (reference cell method).

**Notation**:

- n subjects
- $T_i$ = observation on the i-th subject
- $\delta_i$ = censoring indicator (1 for complete, 0 for censored)
- $\underline{x} = (x_1, x_2, \cdots, x_p)_{p\times 1}$
- $\beta_j$ = coefficient of $x_j$ in the model; $\underline{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)_{p\times 1}$
- $S_0(t)$ = baseline survival function
  - survival function when $\underline{x} = \underline{0}$
  - lies in (0,1)
- $h_0(t)$ = baseline hazard function
  - hazard function when $\underline{x} = \underline{0}$
  - non-negative

**Recall the modeling approaches**:

1. **Accelerated failure time model:** $\ln(T) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \sigma E = \beta_0 + \underline{\beta}' \underline{x} + \sigma E$

- Assume a distribution for E.
- Parametric approach.
- $S(t \mid \underline{x}) = S_0(t \exp\{-\underline{\beta}' \underline{x}\} \mid \underline{x})$

2. **Multiplicative hazard model:** $h(t \mid \underline{x}) = h_0(t) c(\underline{\beta}' \underline{x})$

- Baseline hazard incorporates the effect of time.
- Covariates do not depend on time.
- c(.) is a non-negative function — incorporates the effect of covariates.
- Multiplicative effect of covariates.
- For two settings of covariates $\underline{x}_0$ and $\underline{x}_1$:

- $\frac{h(t \mid \underline{x}_1)}{h(t \mid \underline{x}_0)} = \frac{c(\underline{\beta}' \underline{x}_1)}{c(\underline{\beta}' \underline{x}_0)}$
  - Hazard ratio independent of t.
  - Hazard functions at the two settings are proportional.

- Proportional hazards (PH) model.
- Hazard ratio can be interpreted as *relative risk*.
- Useful for comparing relative survival after adjusting for the covariates.
- $S(t \mid \underline{x}) = [S_0(t)]^{c(\underline{\beta}' \underline{x})}$
- Cox PH model:
  - $c(\underline{\beta}' \underline{x}) = \exp(\underline{\beta}' \underline{x})$
  - Treat baseline hazard as a nuisance parameter and leave it unspecified.
  - Semiparametric approach.
  - Hazard ratio:

**Example**: We want to assess the effectiveness of a treatment of Leukemia using a Cox PH model. Treatment: x = 1 and Placebo: x = 0. Suppose $\beta = -\ln(2)$.

- Hazard ratio:
- This risk of death in treatment group _____ as that of the placebo group.

**3. Additive hazard model:** $h(t \mid \underline{x}) = h_0(t) + \underline{\beta}' \underline{x}$

- Additive effect of covariates.
- Covariate space must be constrained so that the hazard function is positive.

**Our focus in this course**: Cox PH model.

## Fitting a Cox PH Model

n independent subjects.

**Data**: $(T_i, \delta_i, \underline{x}_i)$, i = 1, 2, …, n.

**Model**: $h(t \mid \underline{x}) = h_0(t) \exp(\underline{\beta}' \underline{x}) = h_0(t) \exp\left( \sum_{j=1}^{p} \beta_j x_j \right) = h_0(t) \prod_{j=1}^{p} \exp(\beta_j x_j)$

- Assume non-informative censoring.
- In R, $\underline{x}$ represents $\underline{x} - \bar{\underline{x}}$. So $h_0(t)$ = hazard function for a person with *average* covariate.

## Estimation of $\underline{\beta}$

**Partial Maximum likelihood** (Cox, 1972):

- m ($\leq$ n) distinct times of death.
- Assume no tied death times.
- $t_{(1)} < t_{(2)} < \ldots < t_{(m)}$ = ordered death times.
- For time $t_{(i)}$, define
    - $x_{(i)k}$ = value of k-th covariate (only one individual)
    - $\underline{x}_{(i)} = (x_{(i)1}, \cdots x_{(i)p})$
    - $R(t_{(i)})$ = risk set at time $t_{(i)}$ = set of all individuals who are at risk of death at time $t_{(i)}$

**Partial likelihood function:**

$L_p(\underline{\beta}) =$

- Numerator depends on individual who experiences the event.
- Denominator depends on all individuals who are yet to experience the event.
- Partial likelihood = full censored-data likelihood maximized with respect to $h_o(t)$ keeping $\beta$ fixed = profile likelihood (see handout).
- Treat it as the usual likelihood and maximize it to get partial MLE (PMLE).
- PMLE doesn't exist if a covariate is perfectly correlated with the death times. The maximization algorithms may not always detect this problem.
- Proceed in the usual way for constructing tests (LRT, Wald, Score) and confidence intervals.
- Rao's score = log-rank test when no ties.
- Tests valid only when PH assumption is appropriate. We will learn how to check it later.
- Modify the approach to handle ties:
    - Breslow's method
    - Efron's method – default in R. Tends to be more accurate.
    - See section 8.3 of the book.

**Fitting the Cox model using R**

**Example**: (Brain tumor study) There are 3 covariates in this study.

| | |
|---|---|
| grade | Tumor grade on a scale of 1 to 4.<br>Big is bad |
| age | Patient's age at surgery.<br>Older = worse prognosis. |
| cell | cell = 1 if cell type is astrocytic, 0 if it is oligodendroglial.<br>astrocytic (1) = worse prognosis |

The survival time after the surgery is recorded in days. We want to assess the effect of the covariates on the survival time.

- Baseline hazard rate in R = hazard rate for an individual with covariate value equal to the average covariate value in the sample.

```
# load the survival package #
> library(survival)

# read the data #
> tumor <- read.table(file="brain_tumor.txt",sep="\t", header=TRUE)

# have a look at the data #
> tumor[1:3,]
  cell grade  age status stime
1    1     1  6.3      0  2494
2    1     3 39.9      0  2978
3    1     4 41.8      1   891

# fit the Cox PH model #
> tumor.all <- coxph(Surv(stime, status)~age+grade+cell, data=tumor)
> summary(tumor.all)
Call: coxph(formula = Surv(stime, status)~ age + grade + cell, data = tumor)

  n= 231
       coef exp(coef) se(coef)    z       p
age   0.0166      1.02   0.0057 2.92 3.5e-03
grade 0.9645      2.62   0.1487 6.49 8.8e-11
cell  0.2584      1.29   0.2332 1.11 2.7e-01

      exp(coef) exp(-coef) lower .95 upper .95
age        1.02      0.983      1.01      1.03
grade      2.62      0.381      1.96      3.51
cell       1.29      0.772      0.82      2.05

Rsquare= 0.456   (max possible= 0.996 )

Likelihood ratio test= 141  on 3 df,   p=0
Wald test            = 92.1  on 3 df,   p=0
Score (logrank) test = 127  on 3 df,   p=0
>
```

Conclusion assuming PH assumption is appropriate:

**Testing significance of a subset of coefficients**: [Similar to the logistic regression case]

- Carefully setup the hypotheses.
- Fit the model twice – once the full model (i.e., with the variables that we are testing for significance) and once the reduced model (i.e., without these variables)
- When n is large, 2(logL for full – logL for reduced) ~ $\chi^2$ with degrees of freedom equal to the # of coefficients set to zero in the null hypothesis.

**BT example**: Is the effect of cell significant?

Hypotheses:

```
# get the log-likelihood for the full model #
> tumor.all$loglik
[1] -628.5061 -558.1986
```

(Note that -2*(-628.5061-( -558.1986) )= 140.615 ≈ 141)

```
# get the log-likelihood for the reducted model #
> tumor.age.grade <- coxph(Surv(stime, status)~age+grade, data=tumor)
> tumor.age.grade$loglik
[1] -628.5061 -558.8313
```

So, log-likelihood for the reduced model = -628.5061-(-558.8313) = -69.6748

So, PLRT statistic = 2*(-558.1986-(-558.8313)) = 1.2654
p-value = 1-pchisq(1.2654, df=1) = 0.26

Conclusion:

**Note**: Can also use the `anova()` in `R`.

**Interpretation of the fitted Cox model**

**Model**: $h(t \mid \underline{x}) = h_0(t)\exp(\underline{\beta}' \underline{x}) = h_0(t)\exp(\sum_{j=1}^{p}\beta_j x_j)$

$\underline{x}_0, \underline{x}_1$ = two settings of $\underline{x}$. Generally $\underline{x}_0$ = reference.

$r(\underline{x}_1, \underline{x}_0) =$ hazard ratio $= \dfrac{h(t \mid \underline{x}_1)}{h(t \mid \underline{x}_0)} = \exp\left(\beta'(\underline{x}_1 - \underline{x}_0)\right)$

- Ratio independent of t.
- Interpreted as a *relative risk* — the risk of death with covariate $\underline{x}_1$ relative to covariate $\underline{x}_0$
- $\hat{r}(\underline{x}_1, \underline{x}_0) =$
- Construct CI using the following result:

**Result**: Let $\underline{c}' = (c_1, \cdots, c_p)_{1\times p}$. The large sample Wald CI for $\underline{c}'\underline{\beta}$ is $\underline{c}'\hat{\underline{\beta}} \pm z_{\alpha/2}(c'\hat{V}(\hat{\underline{\beta}})c)^{1/2}$ where $\hat{V}(\hat{\underline{\beta}}) = I^{-1}(\hat{\underline{\beta}})$ is the inverse of the observed information matrix.

- Use `vcov(fit)` in `R` to get the covariance matrix.
- First compute CI for $\underline{c}'\underline{\beta}$ and then exponentiate endpoints to get a CI for exp($\underline{c}'\underline{\beta}$).

```
> vcov(tumor.all)
              [,1]          [,2]          [,3]
[1,]  3.253909e-05 -0.000329754  0.0002559527
[2,] -3.297540e-04  0.022112304 -0.0142072402
[3,]  2.559527e-04 -0.014207240  0.0544042903
>
```

**Interpretation of coefficients**

**Single nominal x with 2 levels 1 and 0**:

**Model**: $h(t \mid x) = h_0(t)\exp(\beta x)$

- r(1, 0) =
- exp(β) = factor by which risk of death increases in group 1 (x = 1) relative to group 0 (x = 0).
- CI easy.

**Single nominal x with K levels**:

- Use K – 1 design variable to represent x
- Reference cell method — one level serves as reference.

**Example**: RACE with 3 levels – White, Black and Hispanic. Reference = White. Create 2 design variables $RACE_B$ and $RACE_H$ as:

| | $RACE_B$ | $RACE_H$ |
|---|---|---|
| White | 0 | 0 |
| Black | 1 | 0 |
| Hispanic | 0 | 1 |

- White = ($RACE_B$, $RACE_H$) = (0, 0).

**Model**: $h(t \mid RACE) = h_0(t)\exp(\beta_B RACE_B + \beta_H RACE_H)$

r(Black, White) =

r(Hispanic, White) =

r(Black, Hispanic) =

- exp(coefficient) = risk of death relative to the reference group.

**Single continuous covariate x**:

**Model**: $h(t \mid x) = h_0(t)\exp(\beta x)$

- $\log h(t \mid x) = \log h_0(t) + \beta x$
- Log-hazard is *linear* in covariate — an important assumption.
- Pick a biologically meaningful c.
- r(x+c, x) =
- exp(β) = factor by which risk of death increases for 1-unit increase in x.

**Multiple covariates model (with no interaction):**

**Model**: $h(t \mid \underline{x}) = h_0(t)\exp(\sum_{j=1}^{p} \beta_j x_j)$

- $r(x_j + c, x_j \mid \text{others}) =$
- $\exp(\beta_j)$ = factor by which risk of death increases for 1 – unit increase in x, after statistically adjusting for other covariates.

**Example**: Consider the Brain tumor study with the full model fitted earlier.

```
> summary(tumor.all)
Call: coxph(formula = Surv(stime, status)~ age + grade + cell, data = tumor)

  n= 231
       coef exp(coef) se(coef)    z       p
age   0.0166      1.02   0.0057 2.92 3.5e-03
grade 0.9645      2.62   0.1487 6.49 8.8e-11
cell  0.2584      1.29   0.2332 1.11 2.7e-01

      exp(coef) exp(-coef) lower .95 upper .95
age        1.02      0.983      1.01      1.03
grade      2.62      0.381      1.96      3.51
cell       1.29      0.772      0.82      2.05
>
```

**Fitted model**:

**Interpretation:**

**Multiple covariates model (with interaction):**

**Example**: Sex = 1 for males and 0 for females.

**Model**: $h(t|age, sex) = h_0(t)\exp(\beta_0\, age + \beta_1\, sex + \beta_2\, age*sex)$

- interaction means $\beta_2 \neq 0$
- r(male, female | age) =
- interaction = relative risk depends on age.
- Practical strategy: Report a table containing the estimate of r and its CI for few key values of age.

**Prediction of survival function based on a Cox PH model**

Consider a patient with $x = \underline{x}_0$. Estimate $S(t \mid \underline{x}_0)$.

Recall:

- $S(t \mid \underline{x}) = [S_0(t)]^{\exp(\beta' \underline{x})}$
- $S_0(t) = \exp(-H_0(t))$
- $t_{(1)} < t_{(2)} < \ldots < t_{(m)}$ = ordered death times
- $d_i$ = # of deaths at time $t_{(i)}$.
- $R(t_{(i)})$ = # individuals at risk of death at time $t_{(i)}$.

Estimator of $H_0(t)$: $\hat{H}_0(t) = \sum_{t_{(i)} \le t} \frac{d_i}{\sum_{j \in R(t_{(i)})} \exp(\hat{\beta}' \underline{x}_j)}$

- Reduces to the Nelson-Aalen estimator of S(t) when there are no covariates.

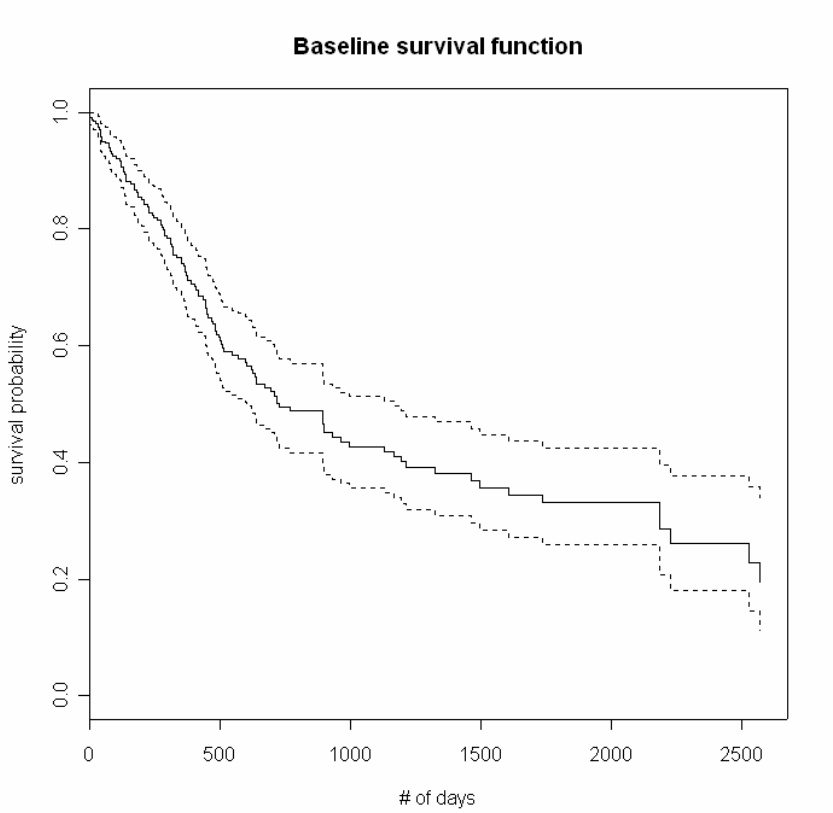
Estimator of $S_0(t)$ =

Estimator of $S(t \mid \underline{x}_0)$ =

- Asymptotically normal with mean $S(t \mid \underline{x}_0)$ and estimated variance given in Section 8.6.
- Use this result to construct CI.
- As usual, log or log-log intervals are better than linear intervals.
- Covariate adjusted survival function.
- Works only when Cox PH model is appropriate.
- In R, baseline = average covariate value.

**Example**: Brain tumor study with just age as the covariate.

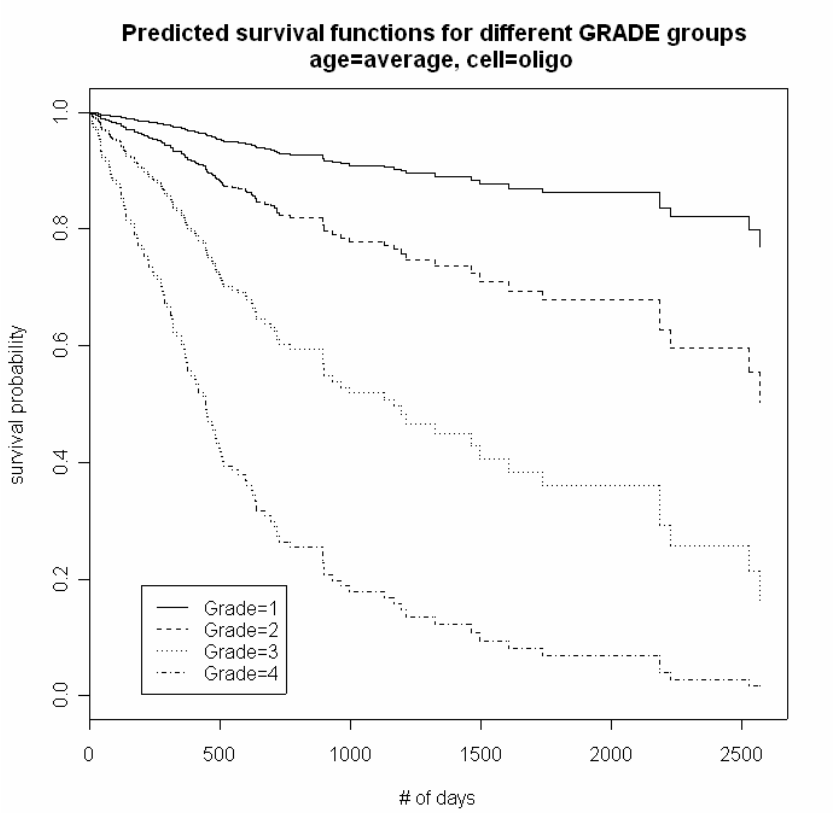
```
> tumor.age <- coxph(Surv(stime, status)~age, data=tumor)

> plot(survfit(tumor.age),main="Baseline survival function", xlab= "# of
days", ylab="survival probability")

> survfit(tumor.age)
Call: survfit.coxph(object = tumor.age)

        n    events     rmean se(rmean)    median   0.95LCL   0.95UCL
    231.0     131.0    1192.2      74.2     724.0     599.0    1164.0
```




- Estimated median survival time for an average aged (mean(tumor$age) = 40.48 years) person =

**Example**: Brain tumor study with all the 3 covariates. Compare the survival functions of the groups formed by their tumor GRADE score among individuals with average age and oligo cell type (cell = 0). GRADE takes the values 1,2,3,4 – an ordered discrete RV. We have treated it as a continuous variable.

```
# attach the dataset in R's memory#
> attach(tumor)

# plot the desired 4 suvival functions #

> plot.survfit(survfit(tumor.all,newdata=list(cell = 0, age = mean(age),
grade=1), conf.type="none"),xlab="# of days",ylab="survival
probability",main="Predicted survival functions for different GRADE groups \n
age=average, cell=oligo")

> lines.survfit(survfit(tumor.all,newdata=list(cell = 0, age = mean(age),
grade=2), conf.type="none"),lty=2)

> lines.survfit(survfit(tumor.all,newdata=list(cell = 0, age = mean(age),
grade=3), conf.type="none"),lty=3)

> lines.survfit(survfit(tumor.all,newdata=list(cell = 0, age = mean(age),
grade=4), conf.type="none"),lty=4)

> legend(locator(1),legend=c("Grade=1","Grade=2","Grade=3","Grade=4"),
lty=1:4)
```

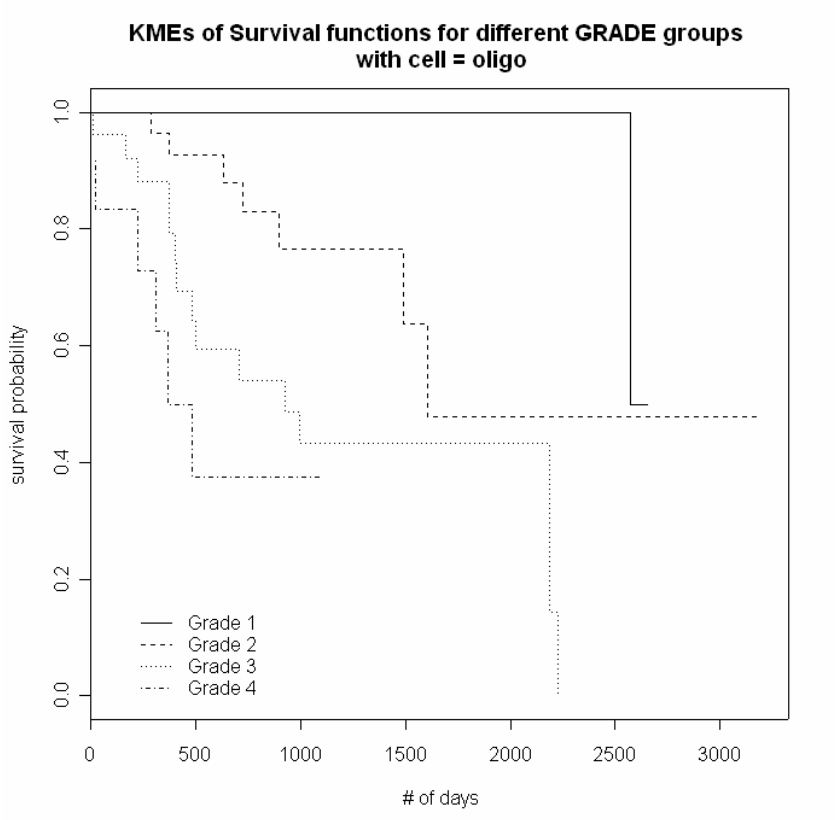



**We observe that**

**How do these estimated (or predicted) survival functions differ from the non-parametric KME?**

```
# get the KME's of the 4 groups with cell = 0#

> plot.survfit(survfit(Surv(stime,status)~grade,subset=cell==0, data=tumor),
mark.time=FALSE,xlab=" # of days", ylab="survival probability",main="KMEs of
Survival functions for different GRADE groups \n with cell = oligo",lty=1:4,
legend.text=c("Grade 1", "Grade 2", "Grade 3", "Grade 4"))
>
```

- Plot A = based on Cox model, Plot B = KME's.
- For Plot A, at every time point r(Grade 2, Grade 1|cell=0, age = avg) = = r(Grade 3, Grade 2|cell=0, age=avg) = r(Grade 4, Grade 3|cell=0, age=avg).
- Plot B doesn't assume any model and it ignores the age.
- Plot A is smoother than Plot B due to the model based extrapolation. Latter B uses only the observed death times in each group, whereas former uses all the observed death times for each group.
- Don't extrapolate outside the observed range of data. The fitted model may not be correct there.




**Stratified Cox PH model**

In the brain tumor example it may happen that the PH hazard assumption is not satisfied for GRADE. (We will learn how to verify it later.) Then we may consider fitting separate Cox PH models for different levels of GRADE.

Suppose the stratifying variable has s levels. The stratified CPH model has the form

$$h_j(t \mid \underline{x}) = h_{0j}(t)\exp(\underline{\beta}'\underline{x}),\ \ j = 1,2,...,\ s$$

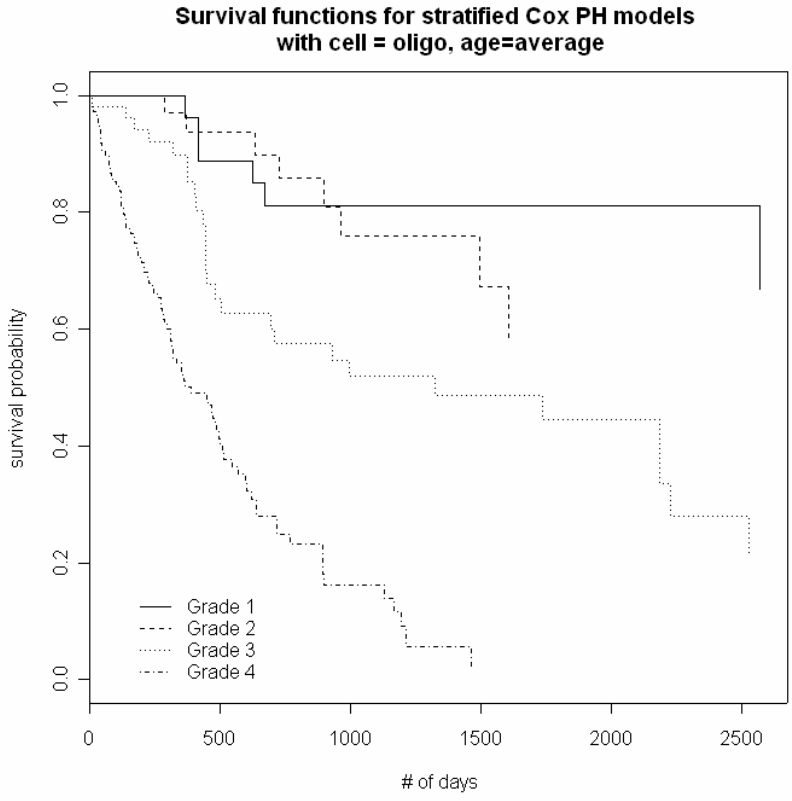
Here:

- Baseline hazards may be different for different strata.
- Regression coefficients are the same in each stratum. Thus the covariates have the same effect in each stratum. See section 9.3 to learn how to verify this assumption.
- The partial log-likelihood function $L(\underline{\beta}) = L_1(\underline{\beta}) + L_2(\underline{\beta}) + \cdots + L_s(\underline{\beta})$, where $L_j(\underline{\beta})$ is the partial log-likelihood function using only the data in the j-th stratum.
- Estimation and testing follows on the same lines as before.
- The large sample stratified tests are appropriate only when the either the sample size within each strata are large or the when the number of strata are large.
- Under a stratified model, the tests of hypotheses on coefficients have good power only if the deviations from the null are the same in each stratum.

Following are the results of fitting a stratified Cox PH model for the brain tumor data:

```
# attach the dataset #
>  attach(tumor)
```

```
# fit a stratified model #
> tumor.str <- coxph(Surv(stime,status)~cell+age+strata(grade), data=tumor)

#look at the summary of fit#
> summary(tumor.str)
Call: coxph(formula = Surv(stime, status) ~ cell + age + strata(grade), data
= tumor)
  n= 231
       coef exp(coef) se(coef)     z      p
cell 0.1107      1.12  0.25229 0.439 0.6600
age  0.0174      1.02  0.00575 3.025 0.0025

     exp(coef) exp(-coef) lower .95 upper .95
cell      1.12      0.895     0.681      1.83
age       1.02      0.983     1.006      1.03

Rsquare= 0.04   (max possible= 0.981 )
Likelihood ratio test= 9.38  on 2 df,   p=0.00918
Wald test            = 9.15  on 2 df,   p=0.0103
Score (logrank) test = 9.11  on 2 df,   p=0.0105

# plot the fitted curves in each GRADE group#
> plot.survfit(survfit(tumor.str, newdata=list(cell=0, age= mean(age) )),
xlab=" # of days", ylab="survival probability",main="Survival functions for
stratified Cox PH models \n with cell = oligo, age=average", lty=1:4,
legend.text=c("Grade 1", "Grade 2", "Grade 3", "Grade 4"))
```




**Observe that**