Cox regression in SAS version 9

Paul W. Dickman
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

paul.dickman@mep.ki.se

May 27, 2005

Slides, data, and SAS code available at
http://www.pauldickman.com/teaching/sas/phreg/

Upgrading to SAS v.9 at MEB

- SAS v.9 is available via the remote installation tool, which theoretically means that you just need to send an e-mail to IT support and it should be available for remote installation within several hours.
  intra.meb.ki.se/ > IT Support > FAQ > How do I install new programs?
- But...if a previous SAS version is installed then IT support must manually uninstall the old version before version 9 can be installed via the remote installation tool.
- See the IT support FAQ
  https://intra.meb.ki.se/ > IT Support > FAQ > Questions regarding software/programs > "All questions regarding SAS".
- Full documentation is available online
  http://support.sas.com/onlinedoc/913/docMainpage.jsp

The Cox proportional hazards model

- The most commonly applied model in medical time-to-event studies is the Cox proportional hazards model [1].
- The Cox proportional hazards model does not make any assumption about the shape of the underlying hazards, but makes the assumption that the hazards for patient subgroups are proportional over follow-up time.
- We are usually more interested in studying how survival varies as a function of explanatory variables rather than the shape of the underlying hazard function.
- In most statistical models in epidemiology (e.g. linear regression, logistic regression, Poisson regression) the outcome variable (or a transformation of the outcome variable) is equated to the 'linear predictor', $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$.
- $X_1, \ldots, X_k$ are explanatory variables and $\beta_0, \ldots, \beta_k$ are regression coefficients (parameters) to be estimated.

- The Xs can be continuous (age, blood pressure, etc.) or if we have categorical predictor variables we can create a series of indicator variables (Xs with values 1 or 0) to represent each category.
- We are interested in modelling the hazard function, $\lambda(t; \boldsymbol{X})$, for an individual with covariate vector $\boldsymbol{X}$, where $\boldsymbol{X}$ represents $X_1, \ldots, X_k$.
- The hazard function should be non-negative for all $t > 0$; thus, using

$$\lambda(t; \boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

  may be inappropriate since we cannot guarantee that the linear predictor is always non-negative for all choices of $X_1, \ldots, X_k$ and $\beta_0, \ldots, \beta_k$.

- However, $\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$ is always positive so another option would be

$$\log \lambda(t; \boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

- In this formulation, both the left and right hand side of the equation can assume any value, positive or negative.
- This formulation is identical to the Poisson regression model. That is,

$$\log \frac{\text{no. events}}{\text{person-time}} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

- The one flaw in this potential model is that $\lambda(t; \boldsymbol{X})$ is a function of t, whereas the right hand side will have a constant value once the values of the βs and Xs are known.
- This does not cause any mathematical problems, although experience has shown that a constant hazard rate is unrealistic in most practical situations.

- The remedy is to replace β_0, the 'intercept' in the linear predictor, by an arbitrary function of time — say $\log \lambda_0(t)$; thus, the resulting model equation is

$$\log \lambda(t; \boldsymbol{X}) = \log \lambda_0(t) + \beta_1 X_1 + \cdots + \beta_k X_k.$$

- The arbitrary function, $\lambda_0(t)$, is evidently equal to the hazard rate, $\lambda(t; \boldsymbol{X})$, when the value of $\boldsymbol{X}$ is zero, i.e., when $X_1 = \cdots = X_k = 0$.
- The model is often written as

$$\lambda(t; \boldsymbol{X}) = \lambda_0(t) \exp(\boldsymbol{X}\beta).$$

- It is not important that an individual having all values of the explanatory variables equal to zero be realistic; rather, $\lambda_0(t)$ represents a reference point that depends on time, just as β_0 denotes an arbitrary reference point in other types of regression models.

- This regression model for the hazard rate was first introduced by Cox [1], and is frequently referred to as the Cox regression model, the Cox proportional hazards model, or simply the Cox model.
- Estimates of $\beta_1, \ldots, \beta_k$ are obtained using the method of maximum partial likelihood.
- As in all other regression models, if a particular regression coefficient, say β_j, is zero, then the corresponding explanatory variable, X_j, is not associated with the hazard rate of the response of interest; in that case, we may wish to omit X_j from any final model for the observed data.
- As with logistic regression and Poisson regression, the statistical significance of explanatory variables is assessed using Wald tests or, preferably, likelihood ratio tests.
- The Wald test is an approximation to the likelihood ratio test. The likelihood is approximated by a quadratic function, an approximation which is generally quite good when the model fits.

- In most situations, the test statistics will be similar. Differences between the test statistics are indicative of possible problems with the fit of the model.
- The assumption of proportional hazards is a strong assumption, and should be tested (see slide 39).
- Because of the inter-relationship between the hazard function, $\lambda(t)$, and the survivor function, $S(t)$, we can show that the PH regression model is equivalent to specifying that

$$S(t; \boldsymbol{X}) = \{S_0(t)\}^{\exp(\beta_1 X_1 + \cdots + \beta_k X_k)}, \quad (1)$$

  where $S(t; \boldsymbol{X})$ denotes the survivor function for a subject with explanatory variables $\boldsymbol{X}$, and $S_0(t)$ is the corresponding survivor function for an individual with all covariate values equal to zero.
- Most software packages, will provide estimates of $S(t)$ based on the fitted proportional hazards model for any specified values of explanatory variables (e.g., the BASELINE statement in PROC PHREG).

Interpreting the estimated regression coefficients

- Recall that the basic PH regression model specifies

$$\lambda(t; \boldsymbol{X}) = \lambda_0(t) \exp(\beta_1 X_1 + \cdots + \beta_k X_k) \; ;$$

equivalently,

$$\log \lambda(t; \boldsymbol{X}) = \log \lambda_0(t) + \beta_1 X_1 + \cdots + \beta_k X_k.$$

- Note the similarity to the basic equation for multiple linear regression, i.e.,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k.$$

- In ordinary regression we derive estimates of all the regression coefficients, i.e., $\beta_1, \ldots, \beta_k$ and β_0.

- In PH regression, the baseline hazard component, $\lambda_0(t)$, vanishes from the partial likelihood; we only obtain estimates of the regression coefficients associated with the explanatory variates $X_1, \ldots, X_k$.

- Consider the simplest possible setup, one involving only a single binary variable, X; then the PH regression model is

$$\log \lambda(t; X) = \log \lambda_0(t) + \beta X \; ,$$

or equivalently,

$$\begin{aligned} \beta X &= \log \lambda(t; X) - \log \lambda_0(t) \\ &= \log \{\lambda(t; X)/\lambda_0(t)\} \; . \end{aligned}$$

- Since $\lambda_0(t)$ corresponds to the value $X = 0$,

$$\beta = \log \{\lambda(t; X = 1)/\lambda_0(t)\} \; .$$

- That is, β is the logarithm of the ratio of the hazard rate for subjects belonging to the group denoted by $X = 1$ to the hazard function for subjects belonging to the group indicated by $X = 0$.

- The parameter β is a log relative risk and $\exp(\beta)$ is a relative risk of response; PH regression is sometimes called "relative risk regression".

- If we conclude that the data provide reasonable evidence to contradict the hypothesis that X is unrelated to response, $\exp(\hat{\beta})$ is a point estimate of the rate at which response occurs in the group denoted by $X = 1$ relative to the rate at which response occurs at the same time in the group denoted by $X = 0$.

- A confidence interval for β, given by $\hat{\beta} \pm 1.96$SE, represents a range of plausible values for the log relative risk associated with the corresponding explanatory variable.

- Corresponding confidence intervals for the relative risk associated with the same covariate are obtained by transforming the confidence interval for β, i.e.,

$$(\beta_\ell, \beta_u) \Rightarrow \left(e^{\beta_\ell}, e^{\beta_u}\right).$$

- When more than one covariate is involved, the principle is the same; $\exp(\hat{\beta}_j)$ is the estimated relative risk of failure for subjects that differ only with respect to the covariate X_j.

- If X_j is binary, $\exp(\hat{\beta}_j)$ estimates the increased/reduced risk of response for subjects corresponding to $X_j = 1$ versus those denoted by $X_j = 0$.

- When X_j is a numerical measurement then $\exp(\hat{\beta}_j)$ represents the estimated change in relative risk associated with a unit change in X_j.

- Since the estimates $\hat{\beta}_1, \ldots, \hat{\beta}_k$ are obtained simultaneously, these estimated relative risks adjust for the effect of all the remaining covariates included in the fitted model.

Example: Localised colon carcinoma 1975–1994

- The data file (`colon.sas7bdat`) contains individual-level data for 15,564 patients diagnosed with colon carcinoma in Finland 1975-1994 with follow-up to the end of 1995.

- We will primarily study mortality among the 6,274 patients diagnosed with localised tumours (stage=1).

The patient data file (`colon.sas7bdat`)

Variable	Type	Format	Label
AGE	Num		Age at diagnosis
DX	Num	DATE.	Date of diagnosis
EXIT	Num	DATE.	Date of exit
MMDX	Num		Month of diagnosis
SEX	Num	SEX.	Sex
STAGE	Num	STAGE.	Clinical stage at diagnosis
STATUS	Num	STATUS.	Vital status at last date of contact
SUBSITE	Num	COLONSUB.	Anatomical subsite of tumour
SURV_MM	Num		Survival time in completed months
SURV_YY	Num		Survival time in completed years
YEAR8594	Num		Indicator for year of dx 1985-94
YYDX	Num		Year of diagnosis

Coding of vital status (for localised stage)

STATUS	Frequency	Cumulative Frequency
0, Alive	2979	2979
1, Dead: colon cancer	1734	4713
2, Dead: other	1557	6270
4, Lost to follow-up	4	6274

Now let's fit a Cox model (where stage=1)

```
proc phreg data=rsmodel.colon(where=(stage=1));
model surv_mm*status(0,2,4) = sex yydx / risklimits;
run;
```

- The syntax of the model statement is

```
MODEL time < *censor ( list ) > = effects < /options > ;
```

- That is, our time scale is time since diagnosis (measured in completed months) and patients with STATUS=0, 2, or 4 are considered censored.

- Patients with any other value of STATUS are assumed to have experienced the event of interest.

Output

Model Information

Data Set	RSMODEL.COLON	
Dependent Variable	SURV_MM	Survival time in completed months
Censoring Variable	STATUS	Vital status at last date of contact
Censoring Value(s)	0 2 4	
Ties Handling	BRESLOW	

Number of Observations Read	6274
Number of Observations Used	6274

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
6274	1734	4540	72.36

Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	28895.004	28859.884

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	35.1199	2	<.0001
Score	35.4870	2	<.0001
Wald	35.3436	2	<.0001

- This output is not especially interesting.
- -2 log likelihood (used for performing likelihood ratio tests) is 28859.884.

- Now for the most interesting part of the output.

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
SEX	-0.00589	0.04891	0.0145	0.9041	0.994	0.903	1.094
YYDX	-0.02749	0.00462	35.3425	<.0001	0.973	0.964	0.982

- There is no evidence that mortality depends on gender (while adjusting only for year of diagnosis).
- Strong association between mortality and year of diagnosis. On assuming a linear association we estimate that mortality is 2.7% lower for each one year increase in year of diagnosis.
- The estimated HR for a 10-year difference would be $0.973^{10} = 0.761$.

Let's categorise year of diagnosis into two periods

- I created a variable, year8594, which takes the value 1 for patients diagnosed 1985-94 and 0 otherwise. That is, we assume a step function.

```
proc phreg data=rsmodel.colon(where=(stage=1));
model surv_mm*status(0,2,4) = sex year8594 / risklimits;
run;
```

Variable	Parameter Estimate	Standard Error	Hazard Ratio	95% Hazard Ratio Confidence Limits	
SEX	-0.00212	0.04889	0.998	0.907	1.098
YEAR8594	-0.23210	0.04920	0.793	0.720	0.873

- We estimate that mortality is 21% lower during the more recent period.
- This code will work in versions 6, 7, and 8.

- A large annoyance with PROC PHREG in versions 8 and earlier was that there was no CLASS statement; if we wanted to model categorical variables we needed to create dummy variables.
- SAS version 9 includes PROC TPHREG (officially an experimental procedure) which contains a CLASS statement.
- Variables listed in the CLASS statement are modelled as categorical variables.
- The syntax is similar to the CLASS statement introduced to PROC LOGISTIC in version 8. That is, one can specify the reference categories using the CLASS statement.

Let's categorise year into two periods using a format

```
proc format;
value yydx
75-84='1975-84'
85-94='1985-94'
;
run;

proc tphreg data=rsmodel.colon(where=(stage=1));
class yydx / ref=first;
model surv_mm*status(0,2,4) = sex yydx / risklimits;
format yydx yydx.;
run;
```

Parameter		DF	Parameter Estimate	Standard Error	Hazard Ratio	95% Hazard Ratio Confidence Limits	
SEX		1	-0.00212	0.04889	0.998	0.907	1.098
YYDX	1985-94	1	-0.23210	0.04920	0.793	0.720	0.873

- Results are the same as when we used a dummy variable to categorise period.

Let's include age at diagnosis as an explanatory variable

```
proc tphreg data=rsmodel.colon(where=(stage=1));
class yydx / ref=first;
model surv_mm*status(0,2,4) = sex yydx age / risklimits;
format yydx yydx.;
run;
```

Parameter		DF	Parameter Estimate	Standard Error	Hazard Ratio	95% Hazard Ratio Confidence Limits	
SEX		1	-0.10208	0.04936	0.903	0.820	0.995
YYDX	1985-94	1	-0.28920	0.04934	0.749	0.680	0.825
AGE		1	0.03342	0.00234	1.034	1.029	1.039

- AGE is not listed in the CLASS statement so it is being modelled as a metric variable in the analysis above.

Modelling age as a categorical variable

```
proc format;
value age
0-44='0-44'
45-59='45-59'
60-74='60-74'
75-high='75+'
;
run;

proc tphreg data=rsmodel.colon(where=(stage=1));
class yydx age / ref=first;
model surv_mm*status(0,2,4) = sex yydx age / risklimits;
format yydx yydx. age age.;
run;
```

Parameter		Parameter Estimate	Standard Error	Chi-Sq	P	Hazard Ratio	95% Hazard Ratio Confidence Limits	
SEX		-0.08871	0.04937	3.2291	0.0723	0.915	0.831	1.008
YYDX	1985-94	-0.28121	0.04937	32.4467	<.0001	0.755	0.685	0.832
AGE	45-59	-0.05153	0.13847	0.1385	0.7098	0.950	0.724	1.246
AGE	60-74	0.29240	0.12576	5.4055	0.0201	1.340	1.047	1.714
AGE	75+	0.81053	0.12611	41.3108	<.0001	2.249	1.757	2.880

Interpreting the estimated hazard ratios

- The variable sex is coded as 1 for males and 2 for females. Since each parameter represents the effect of a one unit increase in the corresponding variable, the estimated hazard ratio for sex represents the ratio of the hazards for females compared to males.
- That is, the estimated hazard ratio is 0.92 indicating that females have an estimated 8% lower colon cancer mortality than males. There is some evidence that the difference is statistically significant ($P = 0.07$).
- The model assumes that the estimated hazard ratio of 0.92 is the same at each and every point during follow-up and for all combinations of the other covariates.
- That is, the hazard ratio is the same for females diagnosed in 1975–1984 aged 0–44 (compared to males diagnosed in 1975–1984 aged 0–44) as it is for females diagnosed in 1985–1994 aged 75+ (compared to males diagnosed in 1985–1994 aged 75+).

- The estimated hazard ratio for YYDX is 0.755. We estimate that, after controlling for age and sex, patients diagnosed 1985–1994 have a 25% lower mortality than patients diagnosed during 1975–1984. The difference is statistically significant ($P < 0.0001$).
- We chose to group age at diagnosis into four categories; 0–44, 45–59, 60–74, and 75+ years.
- It is estimated that individuals aged 75+ at diagnosis experience 2.25 times higher risk of death due to colon carcinoma than individuals aged 0–44 at diagnosis, a difference which is statistically significant ($P < 0.0001$).
- Similarly, individuals aged 60–74 at diagnosis have an estimated 34% higher risk of death due to colon carcinoma than individuals aged 0–44 at diagnosis, a difference which is statistically significant ($P < 0.02$).
- As yet, we have not performed a global test for the effect of age (see slide 29).

Selecting another reference category for age

```
proc format;
value age
0-44='0-44'
45-59='45-59'
60-74='60-74'
75-high='75+'
;
run;
```

```
proc tphreg data=rsmodel.colon(where=(stage=1));
class yydx age(ref='45-59') / ref=first;
model surv_mm*status(0,2,4) = sex yydx age / risklimits;
format yydx yydx. age age.;
run;
```

Parameter		DF	Parameter Estimate	Standard Error	Hazard Ratio	95% Hazard Ratio Confidence Limits	
SEX		1	-0.08871	0.04937	0.915	0.831	1.008
YYDX	1985-94	1	-0.28121	0.04937	0.755	0.685	0.832
AGE	0-44	1	0.05153	0.13847	1.053	0.803	1.381
AGE	60-74	1	0.34392	0.07942	1.410	1.207	1.648
AGE	75+	1	0.86206	0.07950	2.368	2.026	2.767

- The `ref=first` option specifies that, by default, the first category (of the formatted values) is to be used as the reference category.
- We have, however, specified a specific reference category for age which overrides the global option.
- We could also create a variable, called for example AGEGRP, rather than using a format to categorise age.
- I feel, however, that using a format is more efficient. One can, for example, use a different categorisation without having to remake the data set.

Some options for the CLASS statement

- As with PROC LOGISTIC, there is also a PARAM=keyword option to the CLASS statement which can be used to specify the parameterisation method for categorical variables.
- Unlike PROC LOGISTIC, however, the default in PROC PHREG is PARAM=REF (reference cell parameterisation) which is the method we generally want.
- The MISSING option allows missing value (for example,'.' for a numeric variable and blanks for a character variable) as a valid value for the CLASS variable.
- `ORDER=DATA | FORMATTED | FREQ | INTERNAL` specifies the sort criteria.
- `REF=FIRST | LAST`.

Testing the significance of categorical variables (TPHREG)

- New in version 9: if the model contains an effect involving a CLASS variable, a 'Type 3 Tests' table is displayed, which gives the Wald chi-square statistic, the degrees of freedom, and the p-value for each effect in the model (including those effects not listed in the CLASS statement).

Type 3 Tests

Effect	DF	Wald Chi-Square	Pr > ChiSq
SEX	1	3.2291	0.0723
YYDX	1	32.4467	<.0001
AGE	3	173.9180	<.0001

- The Wald test statistic for YYDX is (estimate/SE)$^2 = (-0.28121/0.04937)^2 = 32.4467$ and is displayed by default in the table of parameter estimates (see slide 23; I have removed these columns from some tables to save space).

- In PROC PHREG we would have to create dummy variables and use the TEST statement.

```
Age: Test age_gr2=age_gr3=age_gr4=0;
```

- These are Wald tests; to get LR tests we have to fit models with and without AGE and calculate the test statistic 'by hand'.
- -2 Log L for the model with SEX and YYDX is 28872.77
- -2 Log L for the model with SEX, YYDX, and AGE is 28697.77
- The LR test statistic is $28872.77 - 28697.77 = 175.0$ (close to the Wald test statistic as expected).

Including stage and subsite in the model

```
proc tphreg data=rsmodel.colon;
class yydx age(ref='45-59') stage(ref='Localised') subsite / ref=first;
model surv_mm*status(0,2,4) = sex yydx age stage subsite / risklimits;
format yydx yydx. age age.;
run;
```

Parameter		Parameter Estimate	Standard Error	Hazard Ratio	95% Hazard Ratio Confidence Limits	
SEX		-0.03465	0.02269	0.966	0.924	1.010
YYDX	1985-94	-0.16625	0.02222	0.847	0.811	0.885
AGE	0-44	-0.12404	0.06171	0.883	0.783	0.997
AGE	60-74	0.17420	0.03442	1.190	1.113	1.273
AGE	75+	0.60308	0.03487	1.828	1.707	1.957
STAGE	Distant	2.04294	0.02926	7.713	7.283	8.169
STAGE	Regional	0.82354	0.04113	2.279	2.102	2.470
STAGE	Unknown	0.88945	0.03802	2.434	2.259	2.622
SUBSITE	Descending	-0.04949	0.02547	0.952	0.905	1.000
SUBSITE	Other	0.06913	0.04758	1.072	0.976	1.176
SUBSITE	Transverse	0.10187	0.03125	1.107	1.041	1.177

Type 3 Tests

Effect	DF	Wald Chi-Square	Pr > ChiSq
SEX	1	2.3323	0.1267
YYDX	1	55.9921	<.0001
AGE	3	506.3685	<.0001
STAGE	3	5342.6076	<.0001
SUBSITE	3	26.7450	<.0001

Estimating interaction effects

- Let's study whether the effect of calendar period is modified by stage. We'll fit the interaction term and test if it is statistically significant.

```
proc tphreg data=rsmodel.colon;
class yydx age(ref='45-59') stage(ref='Localised') subsite / ref=first;
model surv_mm*status(0,2,4) = sex yydx age stage subsite yydx*stage / risklimits;
format yydx yydx. age age.;
run;
```

Type 3 Tests

Effect	DF	Wald Chi-Square	Pr > ChiSq
SEX	1	2.3135	0.1283
YYDX	1	51.3153	<.0001
AGE	3	510.3487	<.0001
STAGE	3	2317.8063	<.0001
SUBSITE	3	26.9741	<.0001
YYDX*STAGE	3	29.8496	<.0001

- We see that the interaction effect is statistically significant.

Analysis of Maximum Likelihood Estimates

Parameter			Parameter Estimate	Standard Error	P	Hazard Ratio
SEX			-0.03451	0.02269	0.1283	0.966
YYDX	1985-94		-0.34676	0.04841	<.0001	.
AGE	0-44		-0.12665	0.06172	0.0401	0.881
AGE	60-74		0.17701	0.03443	<.0001	1.194
AGE	75+		0.60633	0.03488	<.0001	1.834
STAGE	Distant		1.89757	0.04087	<.0001	.
STAGE	Regional		0.81554	0.06040	<.0001	.
STAGE	Unknown		0.80801	0.05290	<.0001	.
SUBSITE	Descending		-0.04900	0.02547	0.0544	0.952
SUBSITE	Other		0.07481	0.04761	0.1161	1.078
SUBSITE	Transverse		0.10188	0.03125	0.0011	1.107
YYDX*STAGE	1985-94	Distant	0.28067	0.05672	<.0001	.
YYDX*STAGE	1985-94	Regional	0.03826	0.08240	0.6425	.
YYDX*STAGE	1985-94	Unknown	0.16069	0.07547	0.0332	.

- It seems that SAS will not present the estimated hazard ratios for variables that figure in interaction terms. PROC LOGISTIC also behaves this way.
- The HR for YYDX from the main effects model was 0.85.
- The HR for YYDX at the reference level of stage (localised) is $\exp(-0.34676) = 0.71$
- The HR for YYDX for distant stage is $\exp(-0.34676 + 0.28067) = 0.94$
- The HR for YYDX for regional stage is $\exp(-0.34676 + 0.03826) = 0.73$

- A trick to estimate the effect of an exposure for each level of a modifier, that works for many SAS procedures, is to 'leave out' the main effect of the exposure.

```
proc tphreg data=rsmodel.colon;
class yydx age(ref='45-59') stage(ref='Localised') subsite / ref=first;
model surv_mm*status(0,2,4) = sex age stage subsite yydx*stage / risklimits;
format yydx yydx. age age.;
run;
```

- This doesn't appear to work with TPHREG; SAS estimates a model with one less parameter rather than the same model with a different parameterisation.

Estimating interaction effects using the CONTRAST statement

- It's possible to estimate the effect of period for each level of stage using the CONTRAST statement. Thanks to Mats Talbäck for this suggestion.

```
proc tphreg data=rsmodel.colon;
class yydx age(ref='45-59') stage(ref='Localised') subsite / ref=first;
model surv_mm*status(0,2,4) = sex yydx age stage subsite yydx*stage / risklimits;
format yydx yydx. age age.;
contrast 'Effect of period for localised' YYDX 1 / estimate=exp;
contrast 'Effect of period for distant' YYDX 1 YYDX*STAGE 1 0 0 / estimate=exp;
contrast 'Effect of period for regional' YYDX 1 YYDX*STAGE 0 1 0 / estimate=exp;
contrast 'Effect of period for unknown' YYDX 1 YYDX*STAGE 0 0 1 / estimate=exp;
run;
```


Contrast Rows Estimation and Testing Results

Contrast	Estimate	Confidence Limits	
Effect of period for localised	0.7070	0.6430	0.7773
Effect of period for distant	0.9360	0.8827	0.9926
Effect of period for regional	0.7345	0.6441	0.8377
Effect of period for unknown	0.8302	0.7405	0.9308

- These are exactly the hazard ratios we estimated on slide 35.

Assessing the appropriateness of the proportional hazards assumption

- The proportional hazards assumption is a strong assumption and its appropriateness should always be assessed.
- The model assumes that the *ratio* of the hazard functions for any two patient subgroups (i.e. two groups with different values of the explanatory variable $\boldsymbol{X}$) is constant over follow-up time.
- Note that it is the hazard ratio which is assumed to be constant. The hazard can vary freely with time.
- When comparing an aggressive therapy vs a conservative therapy, for example, it is not unusual that the patients receiving the aggressive therapy do worse earlier, but then have a lower hazard (i.e. better survival) than those receiving the conservative therapy.

The ASSESS statement

- An experimental statement in version 9 of PHREG (not TPHREG).

```
ASSESS < VAR=(list) > < PH > < /options > ;
```

- The ASSESS statement performs the graphical and numerical methods of Lin, Wei, and Ying (1993) [2] for checking the adequacy of the Cox regression model.
- Can assess the functional form of a covariate or check the proportional hazards assumption for each covariate in the Cox model.
- PROC PHREG uses the experimental ODS graphics for the graphical displays.
- `VAR=(list)` specifies the list of explanatory variables for which their functional forms are assessed. For each variable on the list, the observed cumulative martingale residuals are plotted against the values of the explanatory variable along with 20 simulated residual patterns.

- `PH` requests the checking of the proportional hazards assumption. For each explanatory variable in the model, the observed score process component is plotted against the follow-up time along with 20 simulated patterns.
- The following code should work:

```
ods html;
ods graphics on;

proc phreg data=rsmodel.colon(where=(stage=1));
assess var=(age) ph;
model surv_mm*status(0,2,4) = sex yydx age / risklimits;
format yydx yydx. age age.;
run;
quit;

ods graphics off;
ods html close;
```


Using time-varying covariates to assess the PH assumption

- If the effect of an exposure is modified by time then this can be modelled using what is often called a time-varying covariate.
- This is nothing more than an interaction between the exposure and the effect modifier, except the situation is slightly complicated when the effect modifier is time.
- Using a time-varying covariate for an explanatory variable implies that we have removed the assumption that the hazard ratio for that variable is constant with time.
- We can make use of time-varying covariates to test whether the hazard ratio for a fixed covariate is constant over time.

- Consider again a proportional hazards model with one single binary variable, X_1, which takes the value 1 if an exposure is present and 0 if it is absent

$$\lambda(t; \boldsymbol{X}) = \lambda_0(t) \exp(\beta_1 X_1).$$

- The hazard ratio for exposed to unexposed is given by $\exp(\beta_1)$.
- We now construct a second variable, $X_2 = X_1 t$ and include this in the model, in addition to X_1. The variable X_2 takes the value t if the exposure is present and 0 if it is absent

$$\lambda(t; \boldsymbol{X}) = \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 X_1 t).$$

- Based on this model, the hazard ratio for exposed to unexposed is given by $\exp(\beta_1 + \beta_2 t)$.
- An estimate for β_2 significantly different from 0 indicates that the hazard ratio is non-constant over time. $\beta_2 > 0$ indicates that the hazard ratio increases with time and $\beta_2 < 0$ indicates it decreases with time.

- This is not a general test of the proportional hazards assumption. It tests against the alternative that the hazard ratio changes monotonically with time.
- Another alternative might be that the hazard ratio is constant for an initial time period, say $t = 2$ years, but takes on a different (constant) value for the remainder of follow-up.
- To test against this alternative, we construct a variable X_2 which takes the value 1 if the exposure is present and $t > 2$ years, and 0 otherwise.
- In the resulting model containing the variables X_1 and X_2, the hazard ratio for exposed to unexposed for the period $t \leq 1$ year is given by $\exp(\beta_1)$ and for $t > 2$ years it is given by $\exp(\beta_1 + \beta_2)$.
- An estimate for β_2 significantly different from 0 indicates that the hazard ratio is different between the two time periods.

- We will now extend the model for the colon carcinoma data by including a term which allows different hazard ratios for calendar period before and after 2 years (24 months).

```
proc tphreg data=rsmodel.colon(where=(stage=1));
class age / ref=first;
model surv_mm*status(0,2,4) = sex age year8594 t_yr8594 / risklimits;
if surv_mm ge 24 then t_yr8594=year8594;
else t_yr8594=0;
format age age.; run;
```

- We have used SAS programming statements to construct the time varying covariate, `t_yr8594`, which corresponds to the variable X_2 (see Table 1).

Table 1: Values of the time varying covariate

period	$t <$ 24mths	$t \geq$ 24mths
1975–84	0	0
1985–94	0	1

- The coefficient for this variable represents the additional hazard experienced by patients diagnosed in 1985–94 during the period beyond 24 months after diagnosis.

Table 2: Estimated hazard ratios

Variable	$\hat{\beta}$	P-value	Hazard Ratio	95% CI
SEX	-0.0893	0.070	0.915	0.83–1.01
AGE 45–59	-0.0519	0.708	0.949	0.72–1.25
AGE 60–74	0.2904	0.021	1.337	1.05–1.71
AGE 75+	0.8110	0.000	2.250	1.76–2.88
YEAR8594	-0.4207	0.000	0.657	0.58–0.75
T_YR8594	0.3212	0.001	1.379	1.14–1.67

- The time varying covariate was statistically significant in the model ($P = 0.001$).
- That is, the PH assumption was not appropriate for calendar period.

- The estimated hazard ratio, based on the above model, for patients diagnosed 1985–94 compared to 1975–84 is $\exp(-0.4207) = 0.657$ for the period up to 2 years of follow-up and $\exp(-0.4207 + 0.3212) = 0.905$ for the period after 2 years of follow-up.
- The estimated hazard ratio and CI reported by SAS for the variable `YEAR8594` refer to the period prior to 2 years of follow-up.
- The estimated hazard ratio for the period after two years of follow-up can be obtained by multiplying the two hazard ratios, $0.657 \times 1.379 = 0.905$.
- The cutoff at 24 months was chosen arbitrarily. For the first 6 months of follow-up the estimated hazard ratio was 0.724, for the first year it was 0.676, and for the first two years it was 0.657.
- Choosing the cutpoint after inspection of the data will invalidate statistical inference (i.e. reported P-values will be too low).

- We have described two possible alternatives to proportional hazards. In practice, it is possible to fit any model of the form

$$\lambda(t; \boldsymbol{X}) = \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 X_1 f(t)),$$

where $f(t)$ is a function of time.

- To test for non-proportional hazards by age, we must construct three time varying covariates and test them as a group.

```
proc phreg data=survival.colon(where=(stage=1));
model surv_mm*status(0,2,4) = sex age_gr2-age_gr4
 t_age2-t_age4 year8594 t_yr8594 / risklimits;
 t_yr8594=0; t_age2=0; t_age3=0; t_age4=0;
    if surv_mm ge 24 then do;
      t_yr8594=year8594; t_age2=age_gr2;
      t_age3=age_gr3; t_age4=age_gr4;
    end;
Age: Test age_gr2=age_gr3=age_gr4=0;
t_by_age: Test t_age2=t_age3=t_age4=0;
run;
```


Stratified Cox model

- The Cox model assumes that the baseline hazard (mortality rate in the reference group) is an arbitrary function of time.
- The hazard functions for each of the other groups are assumed to be proportional to the baseline.
- It is possible to relax this assumption to allow separate baseline hazards for each level of, for example, age at diagnosis.
- This is known as a stratified proportional hazards model and is a useful method for modelling data where non-proportional hazards are suspected for a factor that is not of primary interest.
- Use the STRATA statement in PROC PHREG.

```
STRATA variable < ( list ) > < ... variable < ( list ) >> < /option > ;
```



```
proc tphreg data=rsmodel.colon(where=(stage=1));
class yydx / ref=first;
model surv_mm*status(0,2,4) = sex yydx / risklimits;
strata age (45,60,75);
format yydx yydx.;
run;
```

Summary of the Number of Event and Censored Values

Stratum	AGE	Total	Event	Censored	Percent Censored
1	<45	297	70	227	76.43
2	52.5	993	206	787	79.25
3	67.5	2716	698	2018	74.30
4	>=75	2268	760	1508	66.49
Total		6274	1734	4540	72.36

Analysis of Maximum Likelihood Estimates

Parameter		Parameter Estimate	Hazard Ratio	95% Hazard Ratio Confidence Limits	
SEX		-0.08871	0.915	0.831	1.008
YYDX	1985-94	-0.28056	0.755	0.686	0.832

- We have allowed a separate baseline hazard within each age group but the effects of sex and period are assumed to be constant across age groups.
- That is, the baseline hazard is the instantaneous mortality rate for males diagnosed in the early period and varies in an unspecified manner as a function of time since diagnosis.
- The instantaneous mortality rate for females diagnosed in the early period is assumed to be 8% lower than the rate for males (which is allowed to be different for each age group).

Time-varying exposures vs time-varying effect of exposure

- We have seen how 'time-varying covariates' can be used in order to allow the effect of exposure to depend on time.
- We may also encounter the situation where the exposure varies with time (effect of the exposure may or may not depend on time), for example, CD4 count, blood pressure, or cumulative exposure to cigarettes or HRT.
- A distinction is made between internal variables (which relate to an individual and can only be measured while a patient is alive) and external variables (which do not necessarily require the survival of the patient for their existence).
- Care should be taken when modelling time-dependent covariates, particularly with internal variables [3, 4].

Late entry / choosing a different time scale

- We used time since diagnosis as the time scale; a sensible choice since mortality depends heavily on time since diagnosis.
- If we wanted to instead use calendar time as the timescale we could use:

```
proc tphreg data=rsmodel.colon(where=(stage=1));
class age(ref='45-59') / ref=first;
model exit*status(0,2,4) = sex age / risklimits entry=dx;
format age age.;
run;
```

- This is not an appropriate model for these data since we have not adjusted for time since diagnosis.

- If we had variables containing age at diagnosis and age at exit we could use attained age as the timescale.

```
model ageexit*status(0,2,4) = sex yydx / risklimits entry=agedx;
```

References

[1] Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society Series B* 1972;**34**:187–220.

[2] Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* 1993;**80**:557–572.

[3] Fisher LD, Lin DY. Time-dependent covariates in the cox proportional-hazards regression model. *Annu Rev Public Health* 1999; **20**:145–57.

[4] Wolfe RA, Strawderman RL. Logical and statistical fallacies in the use of cox regression models. *Am J Kidney Dis* 1996;**27**:124–9.