# Cox proportional hazards regression

Professor Dr. Syed Hatim Noor
Unit of Biostatistics & Research Methodology
School of Medical Sciences
Universiti Sains Malaysia



## Cox proportional hazards regression

- To assess effect of multiple covariates on survival
- The most commonly used multivariable survival method
- Can be used to analyze data that contain censored observations



## Cox proportional hazards model

- Works with hazard model
- Can handle both continuous and categorical predictor variables
- **Proportional hazards assumption**:
  - Hazards ratio should be constant across time
- The proportionality of hazards should not vary over time



## Limitations of Cox PH model

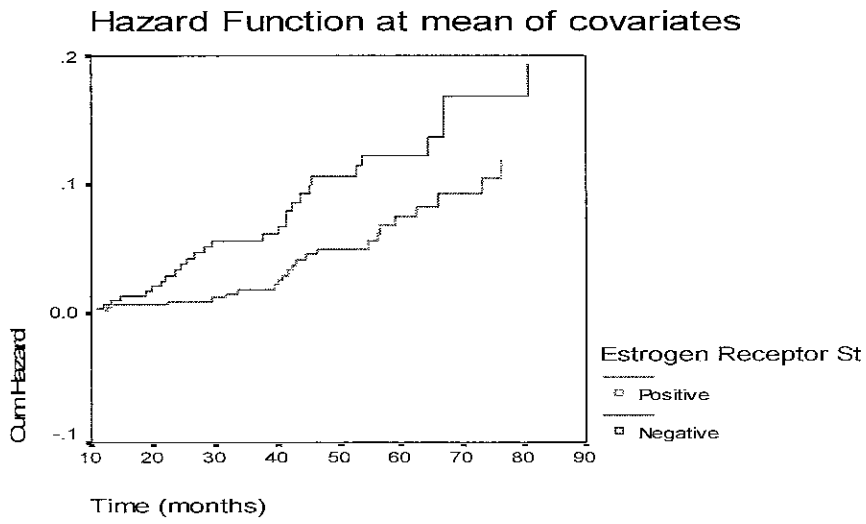
- Covariates normally do not vary over time **(time independent covariates)**
  - True with respect to gender, ethnicity or congenital condition
- If proportional hazards assumption does not hold (hazard ratios change across time: value of one or more covariates are different at different time points)
  - Extended Cox regression model **(time-dependent or time-varying covariates)** needs to be applied

# Hazard function {h(t)}

- Cox regression uses hazard function to estimate the hazards ratio
- A high hazard function indicates a high rate of mortality



# Hazard functions plot




# Mathematical expression (Simple Cox)

- $h(t)=[h_0 (t)]\ e^{bx}$
- h(t)=hazard function (expected risk of death for a case with condition x)
- $h_0$ (t)=baseline hazard function (when x is set to zero) (expected risk without treatment or condition)
- b= regression coefficient
- x= independent prognostic variable
- e=base of the natural logarithm

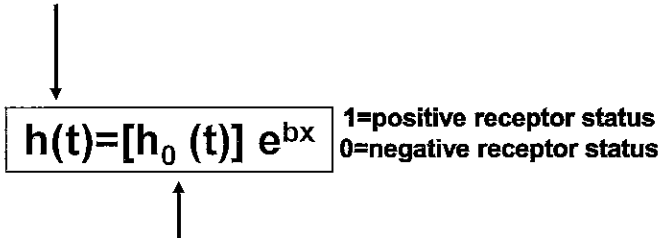


# Explanation

- Dichotomous covariate x
- 0=control / no condition
- 1=treatment / condition
- $h(t)=[h_0 (t)]\ e^{bx}$

**Expected risk of death with condition X**

**eg. Expected risk of death from breast cancer in a patient who has a positive estrogen receptor**




**Expected risk of death with condition X**

**eg. Expected risk of death from breast cancer in a patient who has a negative estrogen receptor**



# Hazard ratio

- Indicates the increase (or decrease) in risk incurred by applying the treatment or condition
- Also called as relative hazard
- $h(t)/[h_0 (t)] = e^{bx}$
- Significance of the hazard ratios (HRs)
  - Confidence intervals
  - Statistical test



# Steps in Cox proportional hazards regression model

Step 1: Data exploration and cleaning
Step 2: Simple Cox regression / Kaplan-Meier survival analysis (log-rank test)
Step 3: Variable selection (**prototyping**)
Step 4: Checking multicollinearity and interactions (**fine modeling**)
Step 5: Checking model assumptions
Step 6: Interpretation, conclusion and presentation

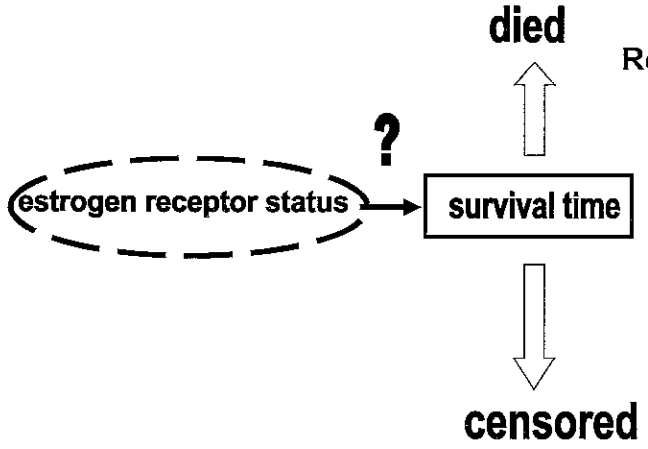


# Step 1:
# Data exploration and cleaning

# Step 2:
# Simple Cox Proportional Hazards Regression



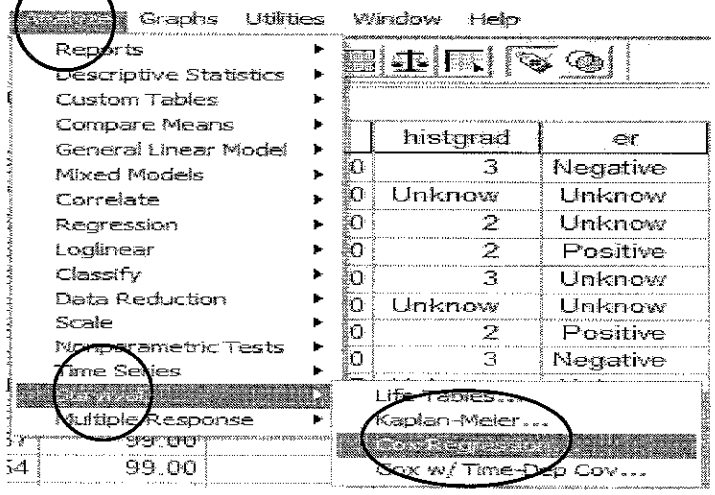
# Example: breast cancer survival study (Simple Cox)




**Research question:**
**Is the estrogen receptor (ER) status the prognostic factor for survival time amongst patients with breast cancer ?**

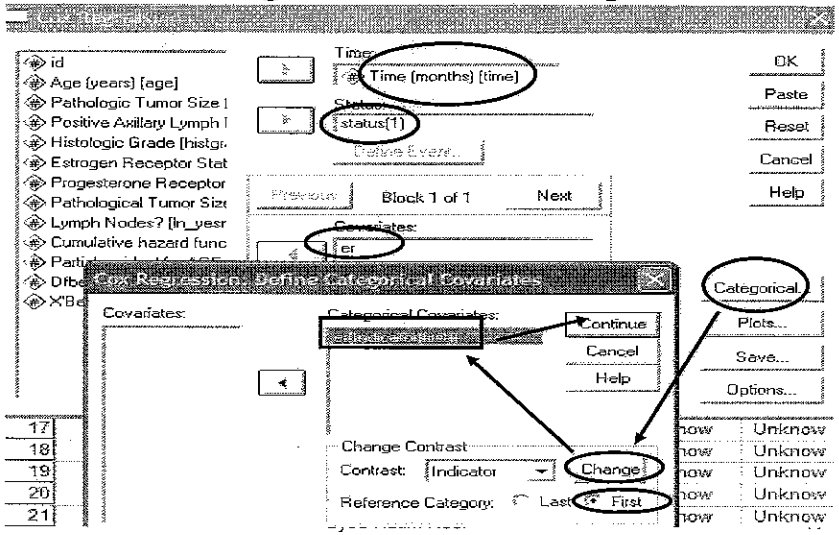


# Steps in Simple Cox regression analysis

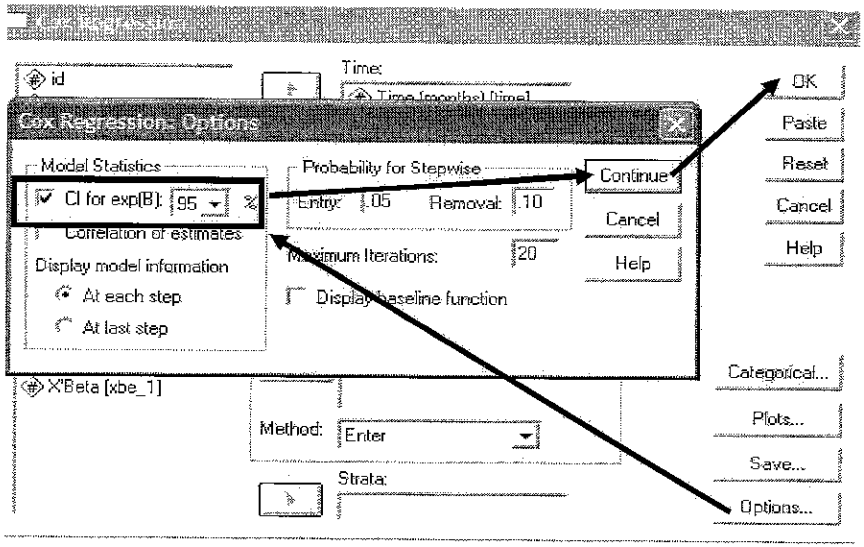





# Steps in Simple Cox regression analysis



# Steps in Simple Cox regression analysis




**Case Processing Summary**

| | | N | Percent |
|---|---|---|---|
| Cases availa in analysis | Event | 51 | 4.2% |
| | Censored | 726 | 60.1% |
| | Total | 777 | 64.4% |
| Cases dropp | Cases with missing value | 338 | 28.0% |
| | Cases with non-positive t | 0 | .0% |
| | Censored cas before the ea event in a stra | 92 | 7.6% |
| | Total | 430 | 35.6% |
| Total | | 1207 | 100.0% |

a. Dependent Variable: Time (months)

**Categorical Variable Codings**[a,b]

| | | Freqency | (1) |
|---|---|---|---|
| ER | 0=Negative | 338 | .000 |
| | 1=Positive | 531 | 1.000 |

a. Indicator Parameter Coding

b. Category variable: ER (Estrogen Receptor Sta



# Hypothesis testing

- $H_o$: $B_1 = B_2 = B_3 = 0$
- Three main likelihood methods
  - Likelihood ratio test (LR test)
  - Wald test
  - Score test (global or overall chi-square test)



**Omnibus Tests of Model Coefficients**[a,b]

| -2 Log likelihood | Overall (score) | | | Change From Previous Step | | | Change From Previous Block | | |
|---|---|---|---|---|---|---|---|---|---|
| | Chi-square | df | Sig. | Chi-square | df | Sig. | Chi-square | df | Sig. |
| 607.723 | 5.537 | 1 | .019 | 5.365 | 1 | .021 | 5.365 | 1 | .021 |

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 613.088

b. Beginning Block Number 1. Method = Enter

**613.088 - 607.723 = 5.365**

**global chi-square (or) overall chi-square**

Omnibus Tests of Model Coefficients[a,b]

| -2 Log Likelihood | Overall (score) | | | Change From Previous Step | | | Change From Previous Block | | |
|---|---|---|---|---|---|---|---|---|---|
| | Chi-square | df | Sig. | Chi-square | df | Sig. | Chi-square | df | Sig. |
| 607.723 | 5.537 | 1 | .019 | 5.365 | 1 | .021 | 5.365 | 1 | .021 |

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 613.088

b. Beginning Block Number 1. Method = Enter



**Likelihood ratio test compared to -2LL of previous step in stepwise regression procdure**

Omnibus Tests of Model Coefficients[a,b]

| -2 Log Likelihood | Overall (score) | | | Change From Previous Step | | | Change From Previous Block | | |
|---|---|---|---|---|---|---|---|---|---|
| | Chi-square | df | Sig. | Chi-square | df | Sig. | Chi-square | df | Sig. |
| 607.723 | 5.537 | 1 | .019 | 5.365 | 1 | .021 | 5.365 | 1 | .021 |

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 613.088

b. Beginning Block Number 1. Method = Enter

**613.088 - 607.723 = 5.365**



**Likelihood ratio test based on the difference between -2LL of two blocks**

Omnibus Tests of Model Coefficients[a,b]

| -2 Log Likelihood | Overall (score) | | | Change From Previous Step | | | Change From Previous Block | | |
|---|---|---|---|---|---|---|---|---|---|
| | Chi-square | df | Sig. | Chi-square | df | Sig. | Chi-square | df | Sig. |
| 607.723 | 5.537 | 1 | .019 | 5.365 | 1 | .021 | 5.365 | 1 | .021 |

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 613.088

b. Beginning Block Number 1. Method = Enter

**613.088 - 607.723 = 5.365**

**Significance is based on the chi-square test. Both score and -2LL are distributed as chi-sqaure.**



**prognostic factor**

**Estimated regression coefficient**

**Standard error of estimated regression coefficient**

**Degree of freedom for the Wald statistic**

**significance level for the Wald statistic**

Variables in the Equation

| | B | SE | Wald | df | Sig. | Exp(B) | 95.0% CI for Exp(B) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower | Upper |
| ER | -.651 | .281 | 5.346 | 1 | .021 | .522 | .301 | .906 |

**Wald statistic**

**Hazard ratio and 95% CI of HR**

- The negative B means that as the value of ER increases, the hazards decrease. coefficient is away from zero
- As the coding is 0=negative and 1=positive, hazards of dying is expected to be less when estrogen receptor status is positive.
- 95% confidence interval of hazards ratio does not include one and corresponding p-value is significant, estrogen receptor status appears to be the effective predictor variable when no other variables are considered.



- The hazard function for estrogen receptor status can be written as

$$h(t) = [h_0(t)]e^{(-0.6507 * 1)}$$

- Since HR is less than one, it indicates there is a decreased hazards when estrogen receptor status is positive.



# Hands-on : Simple Cox

- Do Simple Cox regression analysis for all other independent variables ( age, pathologic tumor size, number of positive lymph nodes, histologic grade, progesterone receptor status)
- Make a table for Simple Cox analysis results



# Step 3: Multiple Cox Proportional hazards Regression Variables selection (**prototyping**)

# Mathematical expression (Multiple Cox)

- $h(t)=[h_0(t)]\, e^{\{B_1x_1+ B_2x_2+ B_3x_3+ \ldots..B_nx_n\}}$
- h(t)=hazard function
- $h_0(t)$=baseline hazard function
- $B_1, B_2, B_3, \ldots..$ regression coefficients
- $x_1, x_2, x_3, \ldots..$ prognostic factors (categorical / continuous)

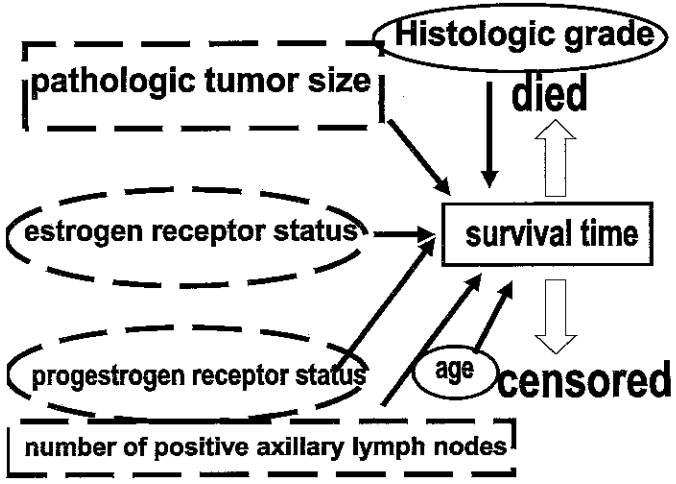


# Methods for selecting variables

- Forced entry (enter method) – all of the variables have been forced into the model in one step
- Forward, Backward and stepwise selections – variables are entered and deleted according to specified criteria



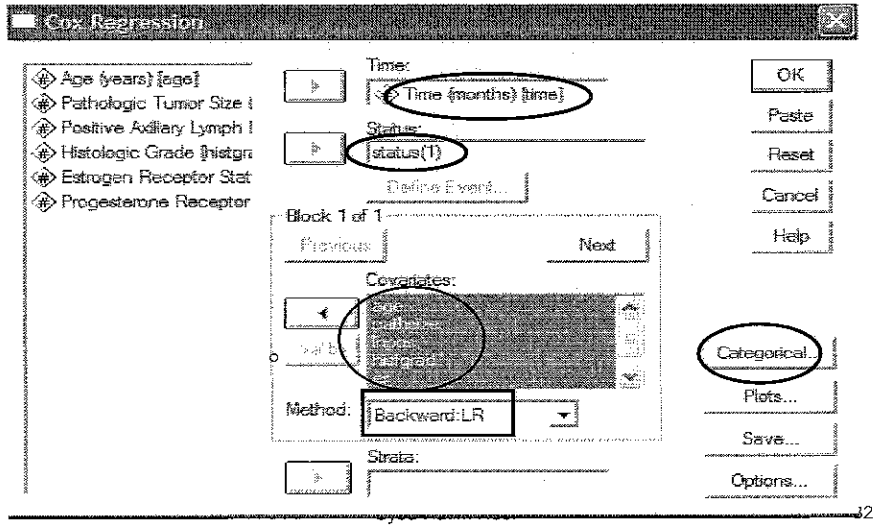
# Breast cancer survival study (Multiple Cox)

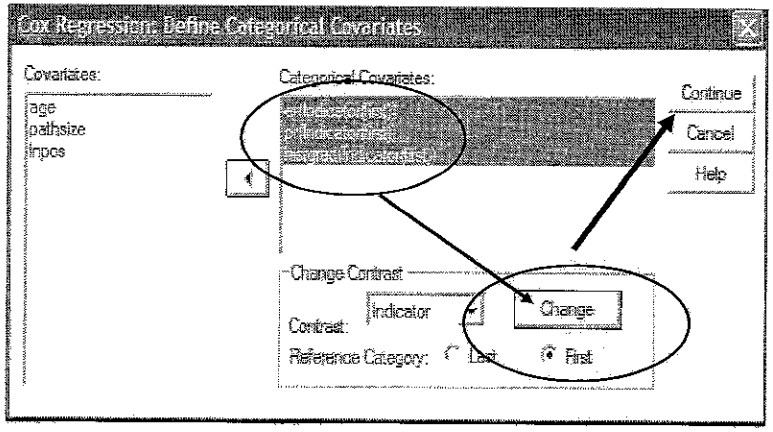



**Research question:**
**Are estrogen receptor status, progesterone receptor status, number of positive axillary lymph nodes, pathologic tumor size, age and histologic grade the prognostic factors for survival time amongst patients with breast cancer ?**

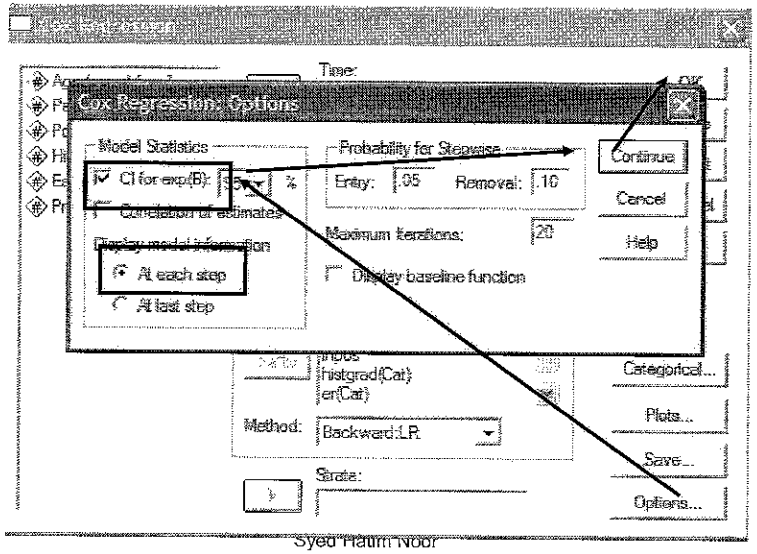


# Steps in Multiple Cox Regression










**Case Processing Summary**

| | | N | Percent |
|---|---|---|---|
| Cases availab | Event | 40 | 3.3% |
| in analysis | Censored | 550 | 45.6% |
| | Total | 590 | 48.9% |
| Cases droppe | Cases with missing values | 547 | 45.3% |
| | Cases with negative | 0 | .0% |
| | Censored cases bef the earliest event in stratum | 70 | 5.8% |
| | Total | 617 | 51.1% |
| Total | | 1207 | 100.0% |

a. Dependent Variable: Time (months)

**Categorical Variable Codings**[b,c,d]

| | | Frequency | (1) | (2) |
|---|---|---|---|---|
| histgrad | 1=1 | 56 | 0 | 0 |
| | 2=2 | 352 | 1 | 0 |
| | 3=3 | 252 | 0 | 1 |
| er[a] | 0=Negative | 262 | 0 | |
| | 1=Positive | 398 | 1 | |
| pr[a] | 0=Negative | 299 | 0 | |
| | 1=Positive | 361 | 1 | |

a. Indicator Parameter Coding

b. Category variable: histgrad (Histologic Grade)

c. Category variable: er (Estrogen Receptor Status

d. Category variable: pr (Progesterone Receptor S



**Variables in the Equation**

| | | B | SE | Wald | df | Sig. | Exp(B) | 95.0% CI for Exp(B) Lower | Upper |
|---|---|---|---|---|---|---|---|---|---|
| Step 1 | age | -.020 | .014 | 2.007 | 1 | .157 | .980 | .954 | 1.008 |
| | pathsize | .392 | .140 | 7.894 | 1 | .005 | 1.480 | 1.126 | 1.946 |
| | lnpos | .133 | .049 | 7.342 | 1 | .007 | 1.143 | 1.038 | 1.258 |
| | histgrad | | | .836 | 2 | .658 | | | |
| | histgrad(1) | .675 | 1.033 | .427 | 1 | .513 | 1.965 | .259 | 14.890 |
| | histgrad(2) | .866 | 1.049 | .682 | 1 | .409 | 2.378 | .304 | 18.596 |
| | er | .078 | .431 | .033 | 1 | .857 | 1.081 | .464 | 2.518 |
| | pr | -.542 | .425 | 1.629 | 1 | .202 | .581 | .253 | 1.337 |
| Step 2 | age | -.020 | .014 | 1.980 | 1 | .159 | .981 | .954 | 1.008 |
| | pathsize | .389 | .138 | 7.904 | 1 | .005 | 1.475 | 1.125 | 1.934 |
| | lnpos | .134 | .049 | 7.382 | 1 | .007 | 1.143 | 1.038 | 1.259 |
| | histgrad | | | .805 | 2 | .669 | | | |
| | histgrad(1) | .674 | 1.033 | .426 | 1 | .514 | 1.963 | .259 | 14.876 |
| | histgrad(2) | .852 | 1.046 | .664 | 1 | .415 | 2.345 | .302 | 18.239 |
| | pr | -.499 | .353 | 2.002 | 1 | .157 | .607 | .304 | 1.212 |
| Step 3 | age | -.022 | .014 | 2.525 | 1 | .112 | .979 | .953 | 1.005 |
| | pathsize | .395 | .138 | 8.217 | 1 | .004 | 1.484 | 1.133 | 1.944 |
| | lnpos | .138 | .049 | 7.965 | 1 | .005 | 1.148 | 1.043 | 1.264 |
| | pr | -.561 | .344 | 2.650 | 1 | .104 | .571 | .291 | 1.121 |
| Step 4 | pathsize | .449 | .132 | 11.565 | 1 | .001 | 1.566 | 1.209 | 2.028 |
| | lnpos | .137 | .047 | 8.369 | 1 | .004 | 1.147 | 1.045 | 1.259 |
| | pr | -.672 | .336 | 4.001 | 1 | .045 | .511 | .265 | .987 |

Model if Term Removed

| Term Removed | | Loss Chi-square | df | Sig. |
|---|---|---|---|---|
| Step 1 | age | 2.054 | 1 | .152 |
| | pathsize | 7.183 | 1 | .007 |
| | lnpos | 6.044 | 1 | .014 |
| | histgrad | .966 | 2 | .617 |
| | er | .033 | 1 | .857 |
| | pr | 1.636 | 1 | .201 |
| Step 2 | age | 2.022 | 1 | .155 |
| | pathsize | 7.165 | 1 | .007 |
| | lnpos | 6.080 | 1 | .014 |
| | histgrad | .934 | 2 | .627 |
| | pr | 2.050 | 1 | .152 |
| Step 3 | age | 2.588 | 1 | .108 |
| | pathsize | 7.448 | 1 | .006 |
| | lnpos | 6.528 | 1 | .011 |
| | pr | 2.727 | 1 | .099 |
| Step 4 | pathsize | 10.251 | 1 | .001 |
| | lnpos | 6.778 | 1 | .009 |
| | pr | 4.158 | 1 | .041 |

Variables not in the Equation[b,c]

| | | Score | df | Sig. |
|---|---|---|---|---|
| Step 2 | er | .033 | 1 | .857 |
| Step 3 | histgrad | .831 | 2 | .660 |
| | histgrad(1) | .092 | 1 | .761 |
| | histgrad(2) | .416 | 1 | .519 |
| | er | .001 | 1 | .971 |
| Step 4 | age | 2.547 | 1 | .111 |
| | histgrad | 1.343 | 2 | .511 |
| | histgrad(1) | .254 | 1 | .614 |
| | histgrad(2) | .847 | 1 | .357 |
| | er | .063 | 1 | .801 |

a. Residual Chi Square = .033 with 1 df Sig. = .857
b. Residual Chi Square = .864 with 3 df Sig. = .834
c. Residual Chi Square = 3.334 with 4 df Sig. = .50



# Results and Interpretation

- **Each unit increase in pathologic tumor size is expected to increase the hazards of dying by 1.6 times (b=0.45, HR=1.57, 95% CI 1.21,2.03, p=0.001)**
- **Each unit increase in number of positive lymph nodes is expected to increase the hazards of dying by 1.2 times (b=0.14, HR=1.15, 95% CI 1.05,1.26, p=0.004)**
- **Hazards of dying is decreased by half when progesterone receptor is positive (b=-0.67, HR=0.51, 95% CI 0.27, 0.99, p=0.045)**

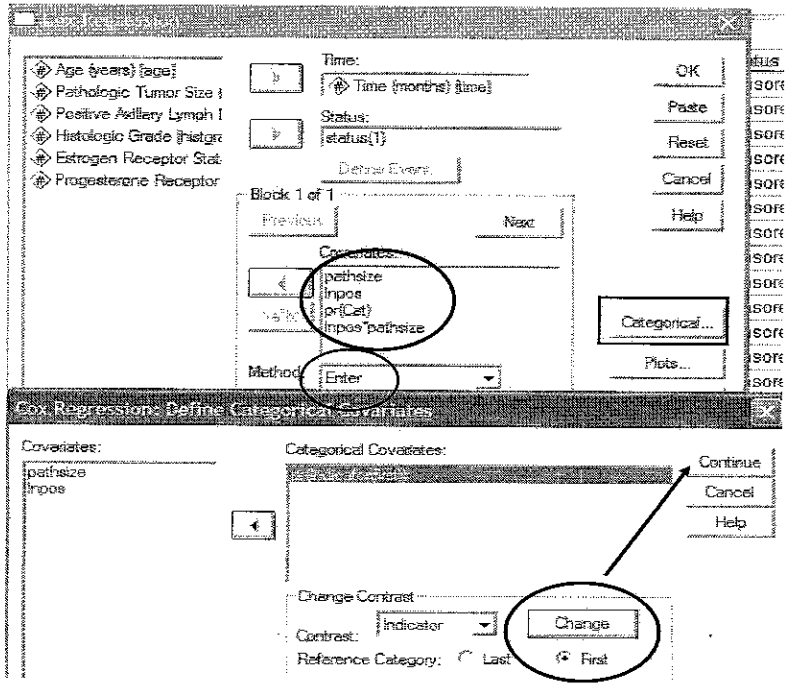
$$h(t)=[h_0(t)]\, e^{(0.45 \times pathsize)+ (0.14 \times lnpos)+ (-0.67 \times pr)}$$

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Step 4 | pathsize | .449 | .122 | 11.565 | 1 | .001 | 1.566 | 1.269 | 2.028 |
| | lnpos | .137 | .047 | 8.369 | 1 | .004 | 1.147 | 1.045 | 1.259 |
| | pr | -.672 | .336 | 4.001 | 1 | .045 | .511 | .265 | .987 |



# Step 4:
# Checking interactions
# (**fine modeling**)





**Variables in the Equation**

| | B | SE | Wald | df | Sig. | Exp(B) | 95.0% CI for Exp(B) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower | Upper |
| pathsize | .562 | .141 | 15.891 | 1 | .000 | 1.754 | 1.331 | 2.312 |
| lnpos | .263 | .074 | 12.798 | 1 | .000 | 1.301 | 1.127 | 1.504 |
| pr | -.742 | .307 | 5.865 | 1 | .015 | .476 | .261 | .868 |
| lnpos*pathsize | -.044 | .027 | 2.682 | 1 | .101 | .957 | .908 | 1.009 |

**Variables in the Equation**

| | B | SE | Wald | df | Sig. | Exp(B) | 95.0% CI for Exp(B) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower | Upper |
| pathsize | .387 | .132 | 8.585 | 1 | .003 | 1.473 | 1.137 | 1.908 |
| lnpos | .149 | .037 | 16.541 | 1 | .000 | 1.161 | 1.080 | 1.247 |
| pr | -.818 | .660 | 1.535 | 1 | .215 | .441 | .121 | 1.610 |
| pathsize*pr | .070 | .252 | .078 | 1 | .780 | 1.073 | .655 | 1.758 |

**Variables in the Equation**

| | B | SE | Wald | df | Sig. | Exp(B) | 95.0% CI for Exp(B) | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower | Upper |
| pathsize | .446 | .125 | 12.694 | 1 | .000 | 1.562 | 1.222 | 1.996 |
| lnpos | .105 | .065 | 2.574 | 1 | .109 | 1.111 | .977 | 1.262 |
| pr | -.788 | .343 | 5.261 | 1 | .022 | .455 | .232 | .892 |
| lnpos*pr | .062 | .076 | .667 | 1 | .414 | 1.064 | .917 | 1.234 |



# Step 5:
# Checking model assumptions



- Proportional assumption: covariates are independent with respect to time and their hazards ratio are constant over time

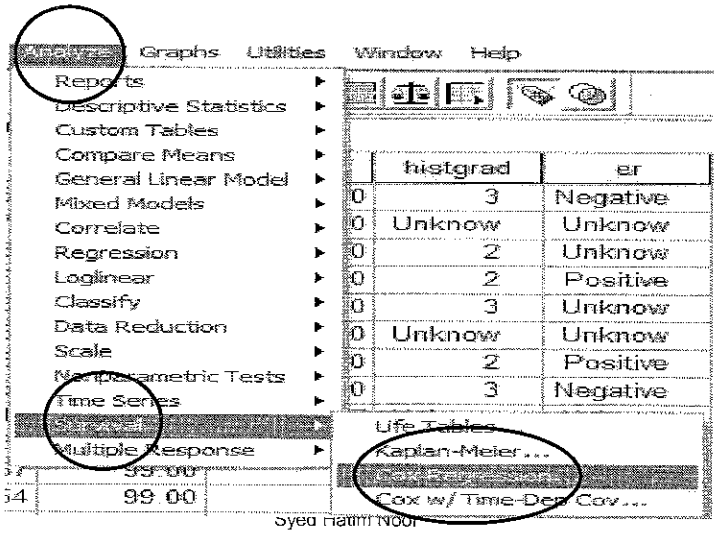
- General ways to examine model adequacy
  - Graphically
  - Mathematically



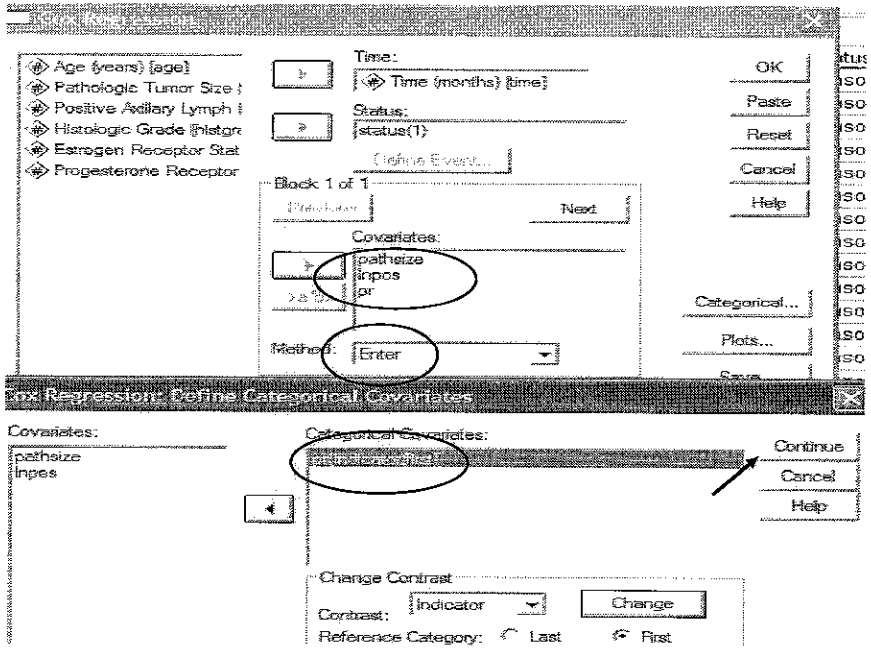
# Graphical approaches

- Several graphical approaches
  - Check proportionality assumption with **hazard functions plot and Log-minus-log plots** of the baseline survival functions
  - Check residuals
    - **Partial residual (Schoenfeld residuals)**
    - **Martingale residuals**
    - **Cox-Snell residuals (cumulative hazard function)**
    - **Deviance residuals**
    - **X'Beta**
  - **Influential statistics**
    - **DfBeta**

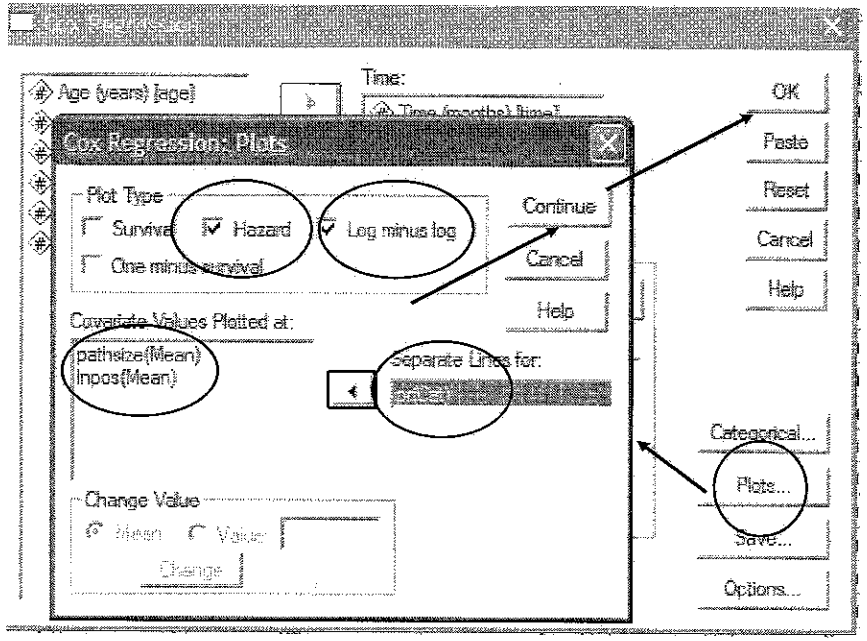
# Steps in getting a hazard function plot and a LML plot



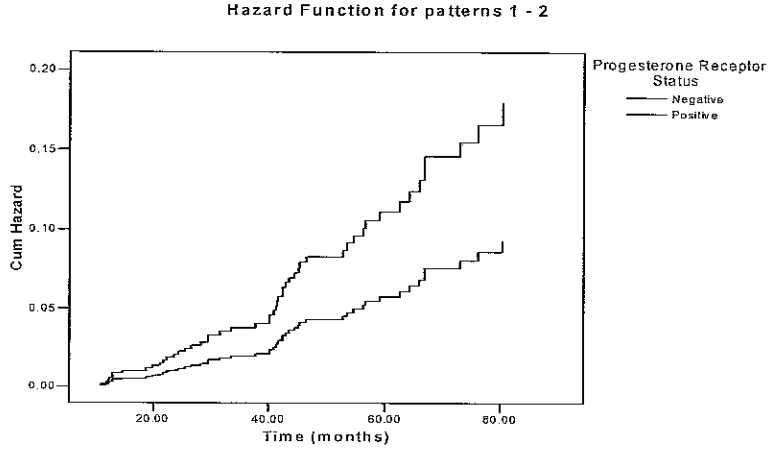




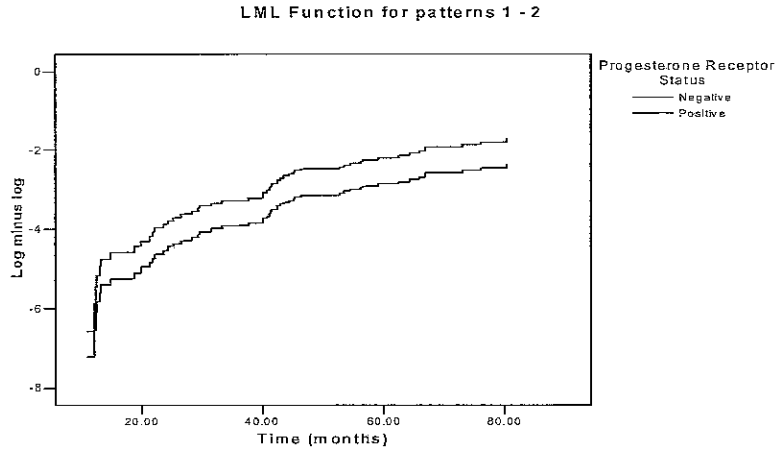




# Hazard functions plot




**Interpretation :The difference between the two curves remains proportional over time.**
**The proportional hazard assumption is met.**
**Hazards ratio is constant over time**

# Log-minus-log (LML) plot




**Interpretation : The curves seem sufficiently parallel. The proportional hazard assumption is met.**



# Checking proportional hazards assumption

- Can be checked by examining the (LML) log-minus-log survival plot
- Should display parallel lines
- Can be checked only for categorical variables



# Checking residuals

⇒ Partial residual (Schoenfeld residuals)

Martingale residuals

Cox-Snell residuals (cumulative hazard function)

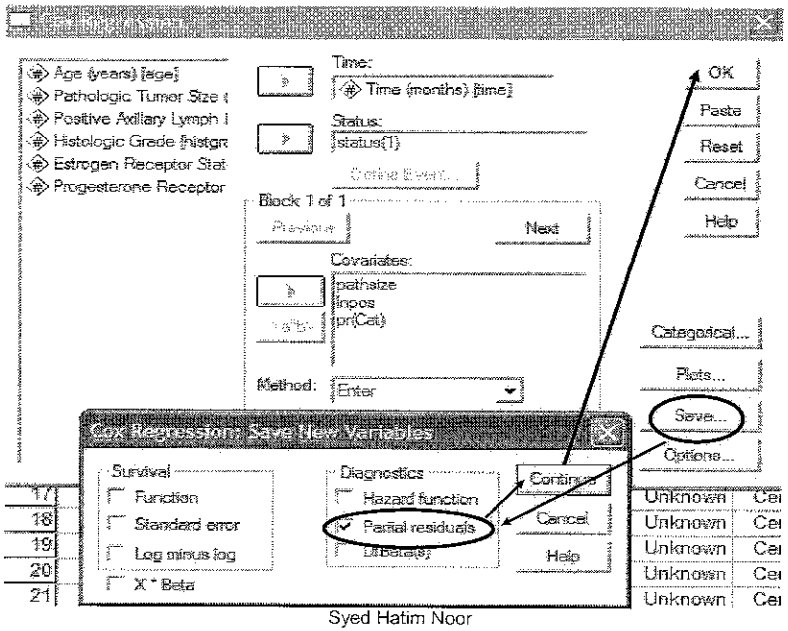
Deviance residuals

X'Beta

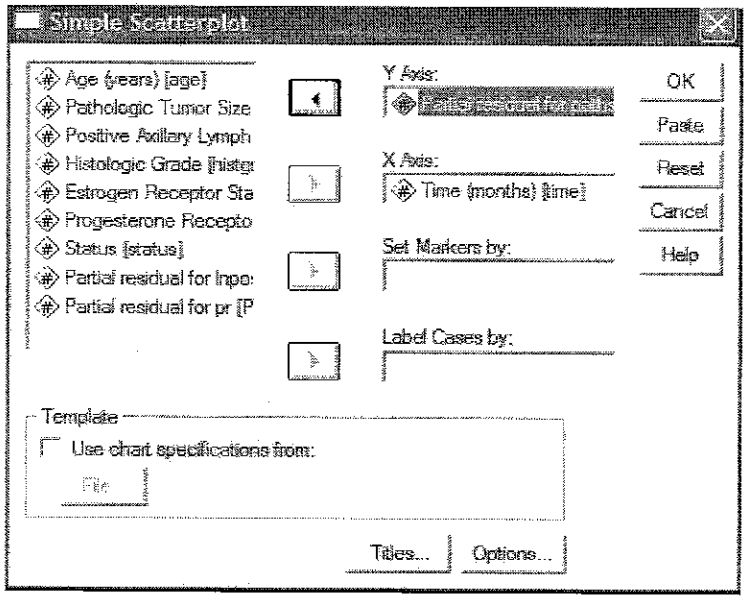


# Partial residuals (Schoenfeld residuals)

- Can be plotted against time to test for violations of the proportional hazards assumption
- Partial residual for each case is the difference between the observed value of the case and its expected value
- Difference should be approximately zero if proportionality assumption holds



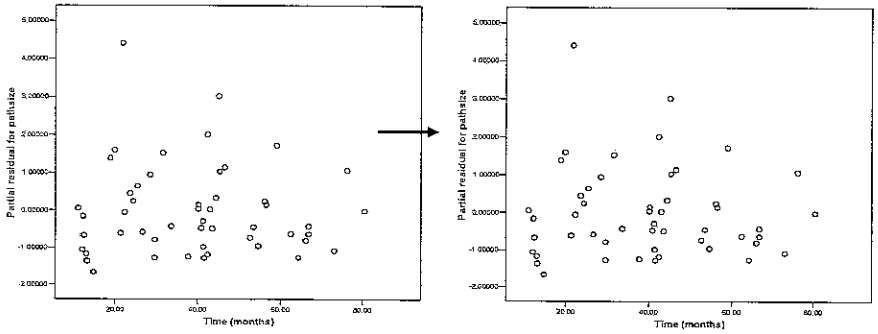


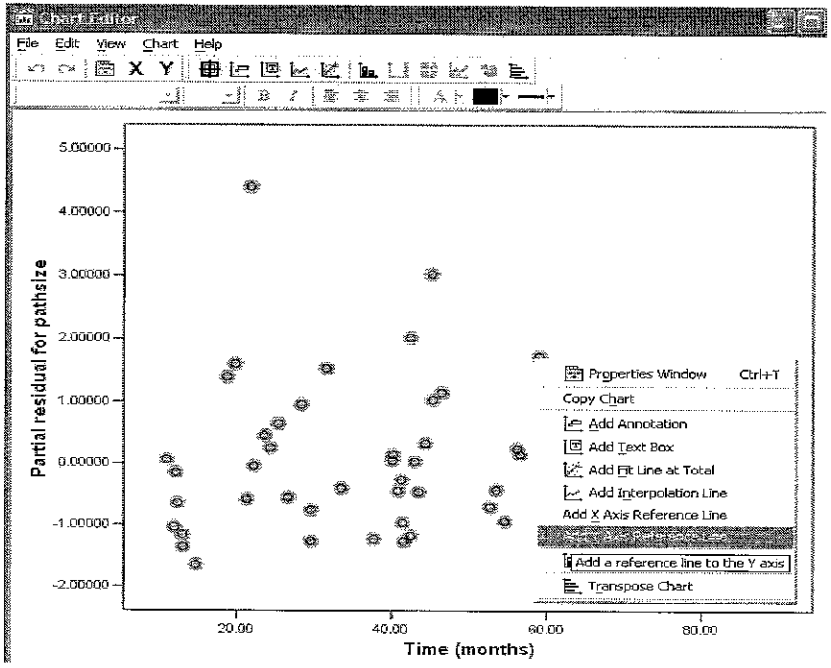


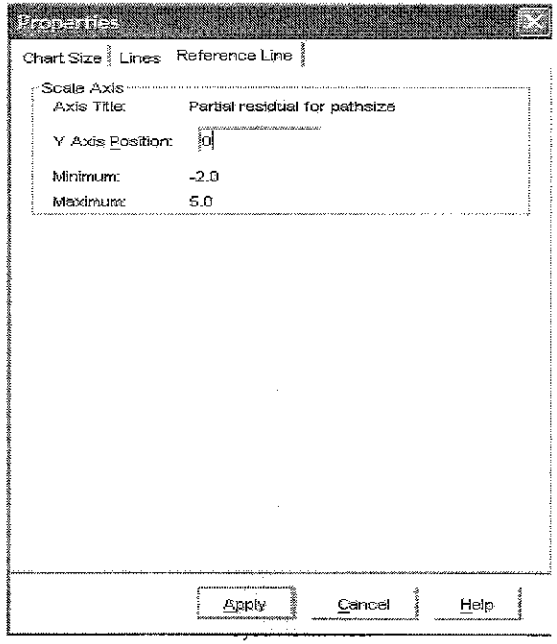

# Steps in getting a partial residuals plot

**double click on the graph**

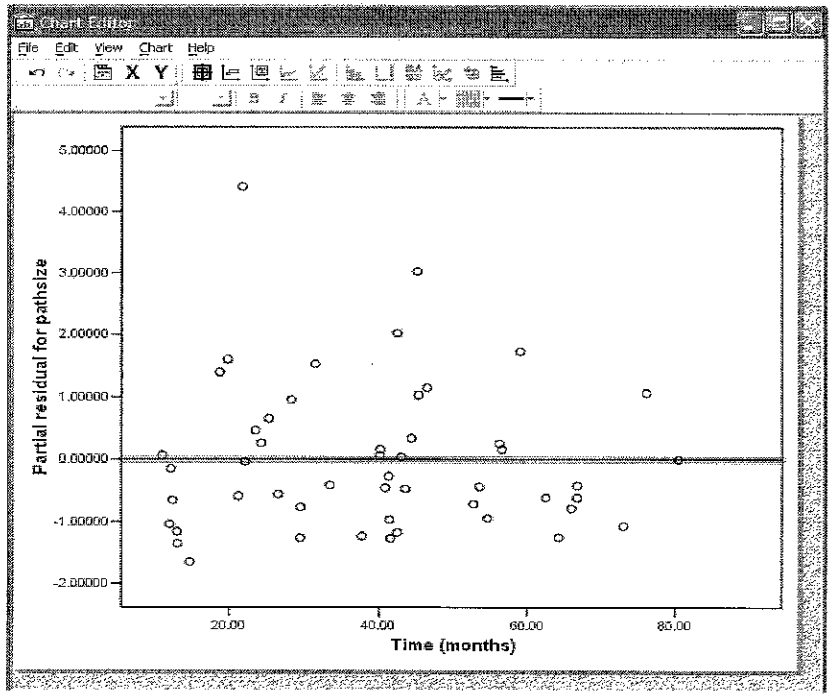


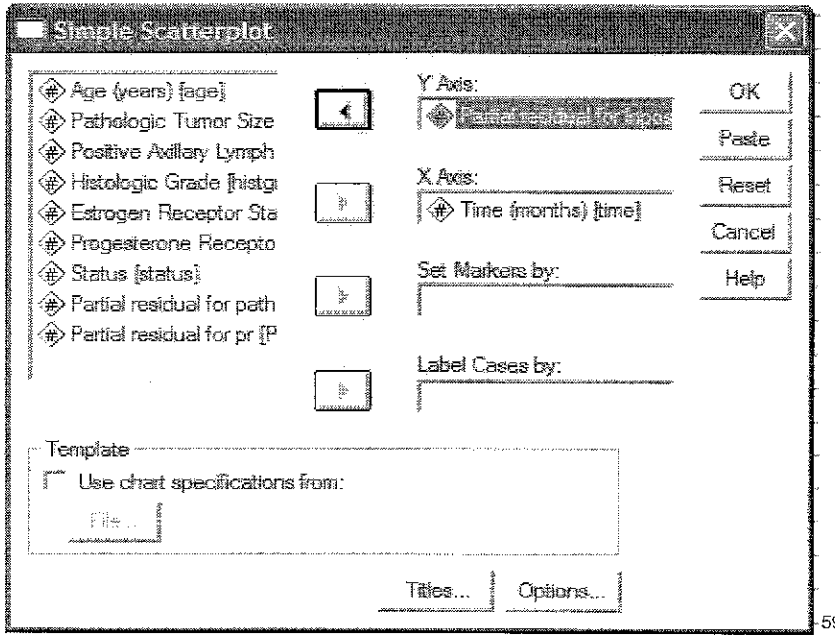






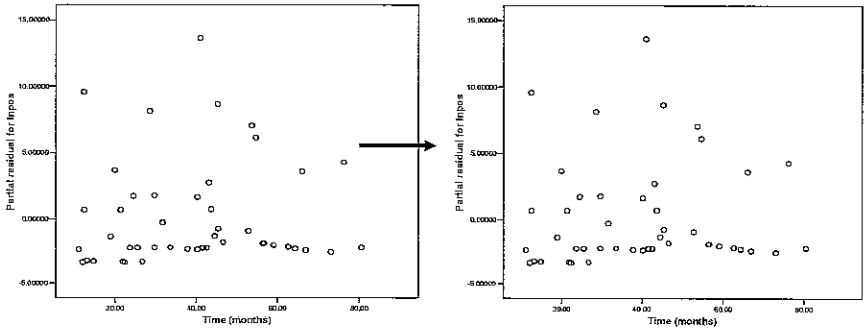



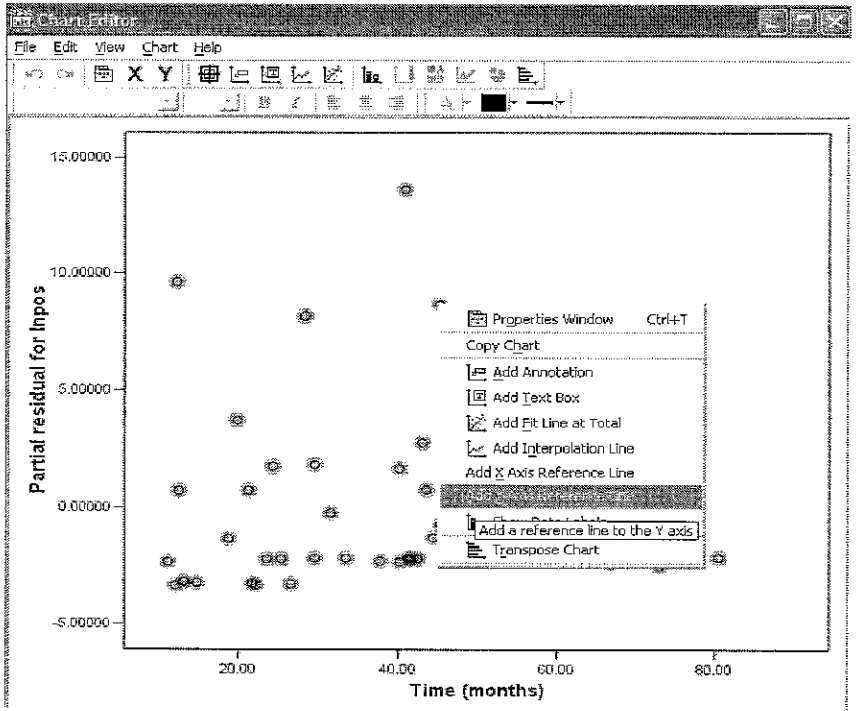





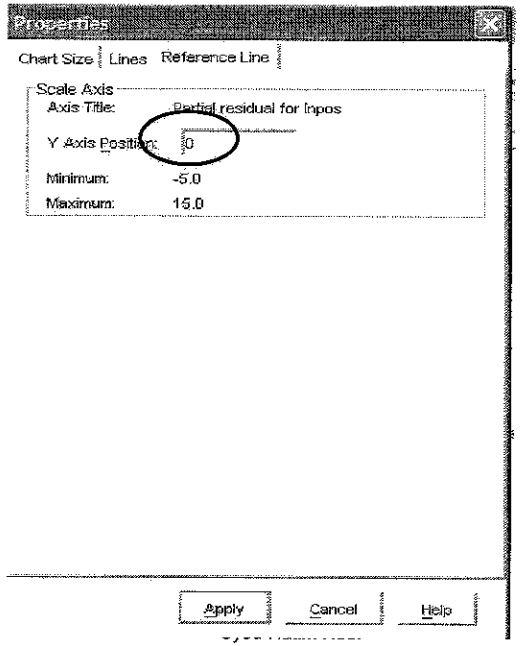

# Steps in getting a partial residuals plot

**double click on the graph**

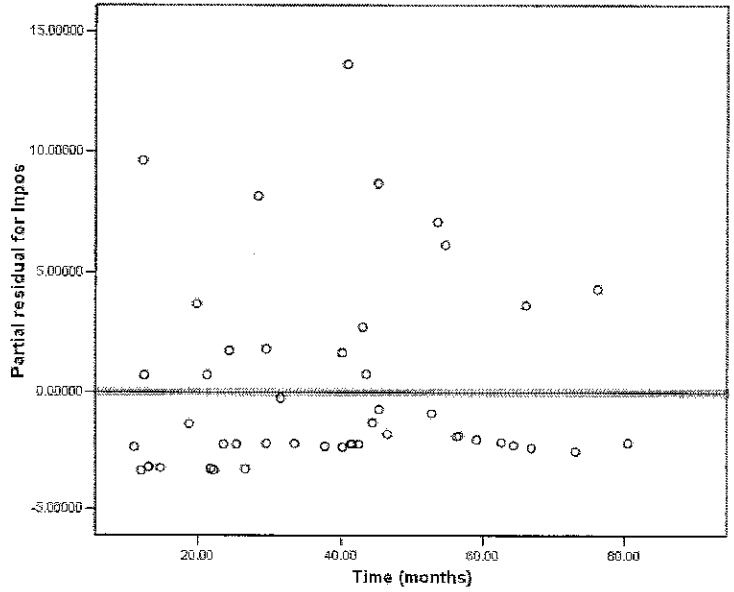






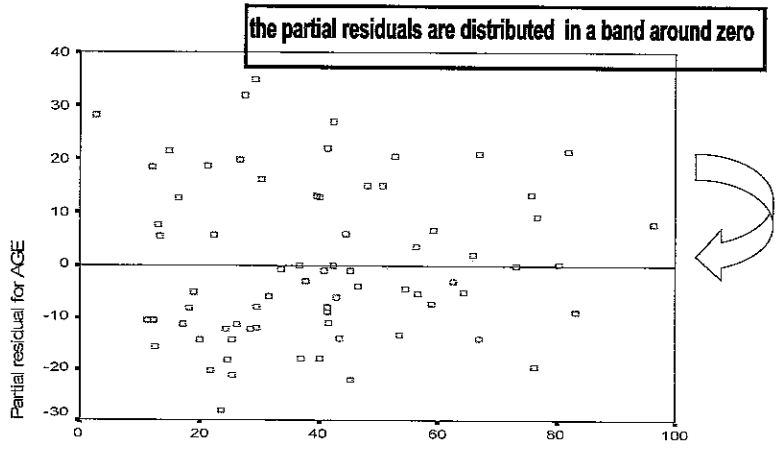








# Other example




**Assumption of proportional hazards is met**

# Step 6:
# Interpretation, presentation and conclusion



# Hands-on Exercise

By using the data set (breast cancer), perform the following data analyses:

(1) Use the Simple Cox regression model to determine whether the variables (estrogen receptor status, progestrone receptor status, age, pathological tumor size, number of positive lymph nodes and histologic grade ) are potential prognostic factors of survival time of patients with carcinoma of the breast. Comment on the findings.



# Hands-on exercise

(2) Perform multiple Cox regression to determine the prognostic factors of breast cancer by applying different methods for selecting models (forced entry, forward stepwise and backward stepwise). Compare the results and make comments.

(3) Check possible clinically sound interactions between covariates by including two-way interactions terms.



# Hands-on exercise

(4) Check whether the model holds the proportional hazards assumption by applying hazard function plot and Log-minus-log plot. Check each covariate from the final model with regard to partial residuals (Schoenfeld residuals) to check whether proportional hazards assumption is met.

## Hands-on exercise

(5) Interpret the results and make the conclusion.

(6) Present the results in tables.



## Summary

- Survival analyses quantifies **time to a dichotomous event**
- Handles **censored data** well
- **Kaplan-Meier survival analysis** can be compared statistically and graphically
- **Cox proportional hazards models** help distinguish individual contributions of covariates on survival, provided certain assumptions are met

Table (1) Prognostic factors of cancer breast by Simple Cox proportional hazards model

| Variable | Regression coefficient (b) | Crude Hazards ratio (95% CI) | Wald statistic | p-value |
|---|---|---|---|---|
| Age | -0.02 | 0.98(0.96,0.99) | 4.98 | 0.026 |
| Pathologic tumor size | 0.60 | 1.82(1.53,2.17) | 45.28 | <0.001 |
| Number of positive lymph nodes | 0.10 | 1.11(1.07,1.15) | 29.47 | <0.001 |
| Histologic grade | | | | |
| Grade 1 | 0.00 | 1.00 | - | - |
| Grade 2 | 0.71 | 2.03(0.48,8.55) | 0.94 | 0.333 |
| Grade 3 | 1.35 | 3.86(0.92,16.25) | 3.40 | 0.065 |
| Estrogen receptor status | | | | |
| Negative | 0.00 | 1.00 | - | - |
| positive | -0.65 | 0.52(0.30,0.91) | 5.35 | 0.021 |
| Progesterone receptor status | | | | |
| Negative | 0.00 | 1.00 | - | - |
| positive | -0.64 | 0.53(0.30,0.93) | 4.97 | 0.026 |

Table (2) Prognostic factors of cancer breast by multiple Cox proportional hazards model

| Variable | Regression coefficient (b) | Adjusted Hazards ratio (95% CI) | Wald statistic | p-value |
|---|---|---|---|---|
| Pathologic tumor size | 0.45 | 1.57(1.21, 2.03) | 11.57 | 0.001 |
| Number of positive lymph nodes | 0.14 | 1.15(1.05, 1.26) | 8.37 | 0.004 |
| Progesterone receptor status | | | | |
| Negative | 0.00 | 1.00 | - | - |
| positive | -0.67 | 0.51(0.27, 0.99) | 4.00 | 0.045 |

Backward stepwise Cox proportional hazards regression model applied
Log-minus-log plot, hazard function plot and partial residuals were applied to check the model assumption and found fulfilled

Table (3) Prognostic factors of cancer breast by univariable and multivariable Cox proportional hazards models

| Variable | Simple Cox regression | | | Multiple Cox regression | | |
|---|---|---|---|---|---|---|
| | b | Crude HR(95%CI) | p | b | Adjusted HR(95% CI) | p |
| Number of positive lymph nodes | 0.10 | 1.11(1.07,1.15) | <0.001 | 0.14 | 1.15(1.05,1.26) | 0.004 |
| Progesterone receptor status | | | | | | |
| Negative | 0.00 | 1.00 | - | 0.00 | 1.00 | - |
| positive | -0.64 | 0.53(0.30,0.93) | 0.026 | -0.67 | 0.51(0.27, 0.99) | 0.045 |
| Pathologic tumor size | 0.60 | 1.82(1.53,2.17) | <0.001 | 0.45 | 1.57(1.21, 2.03) | 0.001 |

Backward stepwise Cox proportional hazards regression model applied
Log-minus-log plot, hazard function plot and partial residuals were applied to check the model assumption and found fulfilled