Chapter 11 Regression Diagnostics

PubH 7450

©Wei Pan

Email: weip@biostat.umn.edu

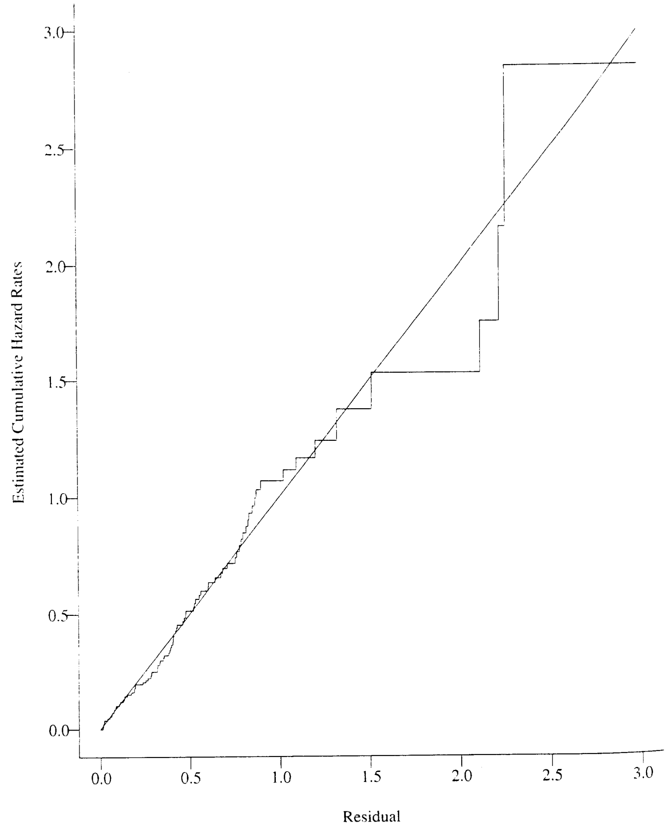
Http: www.biostat.umn.edu/∼weip

§11.2 Cox-Snell residuals

- Goal/use: a graphical assessment of the overall fit of a model.
- Basic idea:
  1. $X \sim F$ (cdf) $\Longrightarrow F(X) \sim U(0,1)$;
  A rough proof: for any $0 \leq y \leq 1$,
  $Pr[F(X) \leq y] = Pr[X \leq F^{-1}(y)] = F[F^{-1}(y)] = y$.
  2. $H(X) = -\log S(X) = -\log[1 - F(X)] \sim Exp(1)$
  $\Longrightarrow h(t) = 1, H(t) = t$.
- Given: 1) data (T_j, δ_j, Z_j), $j = 1, ..., n$;
  2) a Cox PHM: $h(t|Z) = h_0(t)\exp(Z'\beta)$.
- How?
  1) fit the model $\longrightarrow \hat{H}_0(t)$, $\hat{\beta}$;
  2) $r_j = H_j(T_j) = \hat{H}_0(T_j)\exp(Z_j'\hat{\beta})$.
  —-*Cox-Snell residuals*
  Q: what is the distribution of r_j?

3) (r_j, δ_j)'s: a ...... sample from ...
4) plot ...... based on (r_j, δ_j)'s; compare with ...... to see whether there is a strong discrepancy between the two; if yes, the model is inadequate!

- Example 11.1: Fig. 11.1-3.



11.1 *Cox–Snell residual plot treating MTX as a fixed time covariate*

3-1

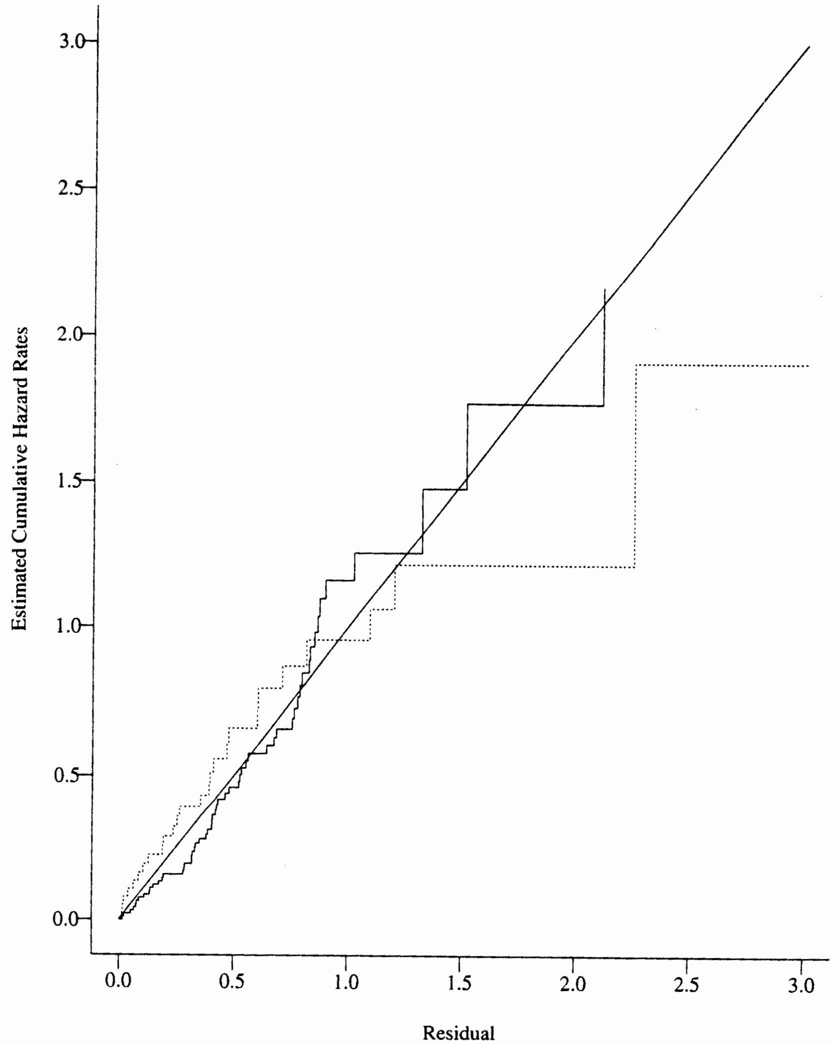


Figure 11.2 *Cox–Snell residual plots for MTX and no MTX patients separately treating MTX as a fixed covariate in the model. MTX patients (------) No MTX patients (———)*

3-2

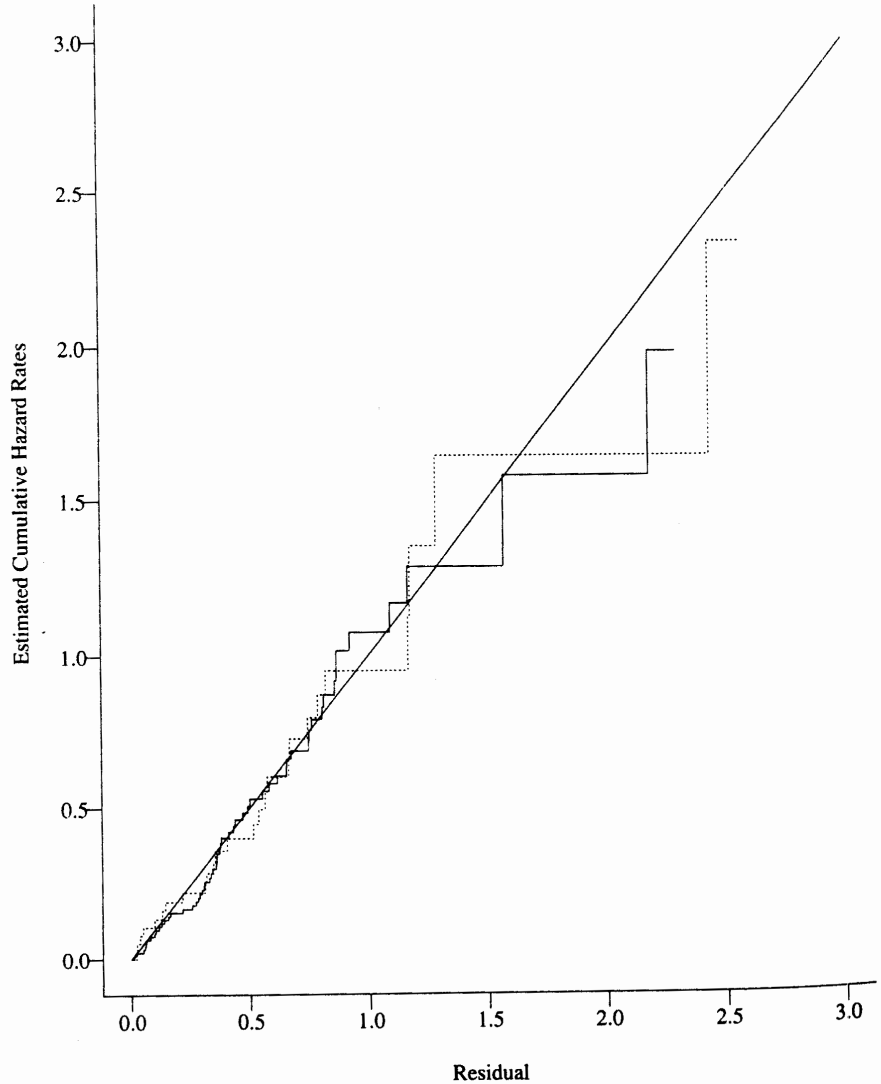


Figure 11.3 *Cox–Snell residual plots for MTX and no MTX patients based a model stratified on MTX usage. MTX patients (------) No MTX patients (—*

§11.3 Martingale residuals

- Goal: to determine the functional form of a covariate.
  similar to (partial) residual plot?
- Given: 1) data (T_j, δ_j, Z_j), $j = 1, ..., n$;
  2) a Cox PHM: $h(t|Z) = h_0(t)\exp(Z'\beta)$.
  martingale residuals:
  $\hat{M}_j = \delta_j - \hat{H}_0(T_j)\exp(Z_j'\hat{\beta})$ =#obs'ed events - #exp'ed events,
  $j = 1, ..., n$.
- Given: $Z = (Z_1, Z_2')'$ and we know functional form of Z_2.
- Q: find functional form of Z_1.
- How?
  1) fit a PHM w/o Z_1: $h(t|Z_2) = h_0(t)\exp(Z_2'\beta) \Longrightarrow \hat{M}_j$,
  $j = 1, ..., n$.
  2) plot $\hat{M}_j$ vs Z_1: the trend tells the functional form of Z_1.

- Example 11.2: Fig 11.4.
- Q: why not just do H.T.? why do graphics?

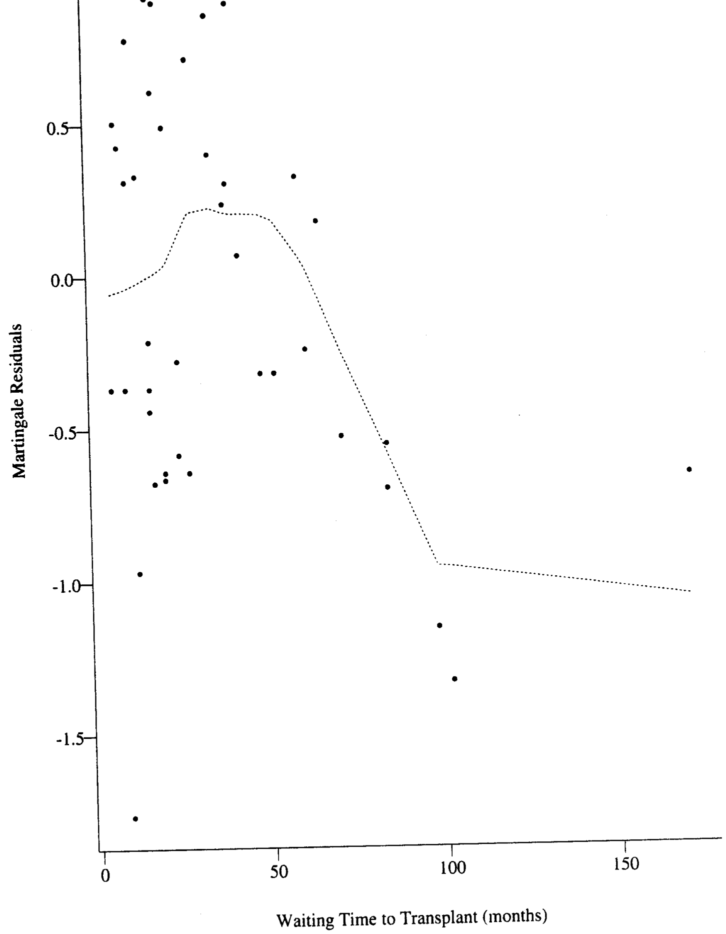


Figure 11.4 *Plot of martingale residual verus waiting time to transplant and LOWESS smooth*

§11.5 Deviance residuals

- Goal: to identify possible outliers (and assess overall model fiting).
- Motivation:
  Martingale residuals: highly skewed!
  $\min(\hat{M}_j) = -\infty$, $\max(\hat{M}_j) = 1$.
- *Deviance residuals*: transform $\hat{M}_j$ so that it is more symmetric (like a Normal variate),
  $D_j = sgn(\hat{M}_j)\{(-2)[\hat{M}_j + \delta_j \log(\delta_j - \hat{M}_j)]\}^{1/2}$.
- Some properties:
  $\hat{M}_j = 0 \Longrightarrow D_j = 0$.
  D_j increases as $\hat{M}_j \to 1$.
  D_j shrinks a large negative $\hat{M}_j$.
- Goal 1: to identify outliers, use index plot:
  plot D_j vs j,...

- Goal 2: for general model checking,
  plot D_j vs $Z_j'\hat{\beta}$ (linear rpedictor or risk score); if any trend, ...
- Example 11.2: Fig 11.20-21.

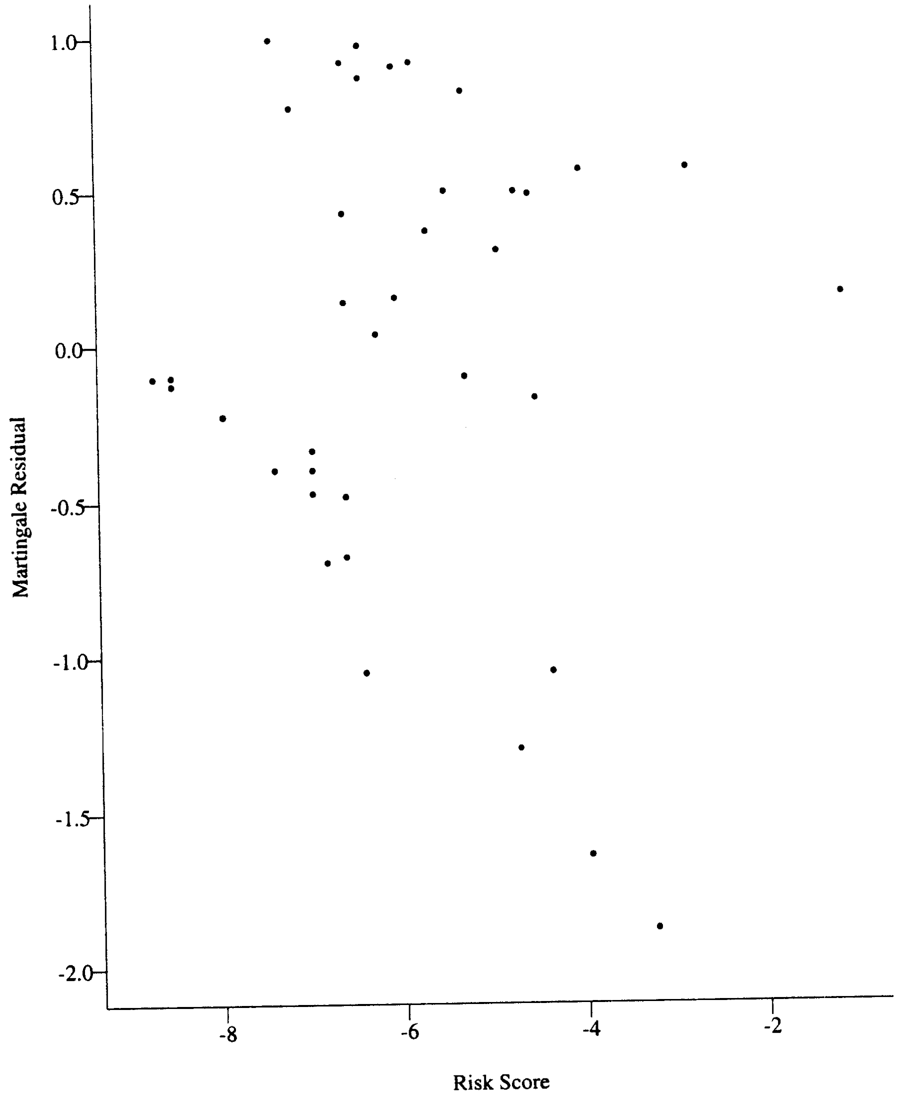


Figure 11.20 *Plot of the martingale residuals versus risk scores for the bone marrow transplant example*

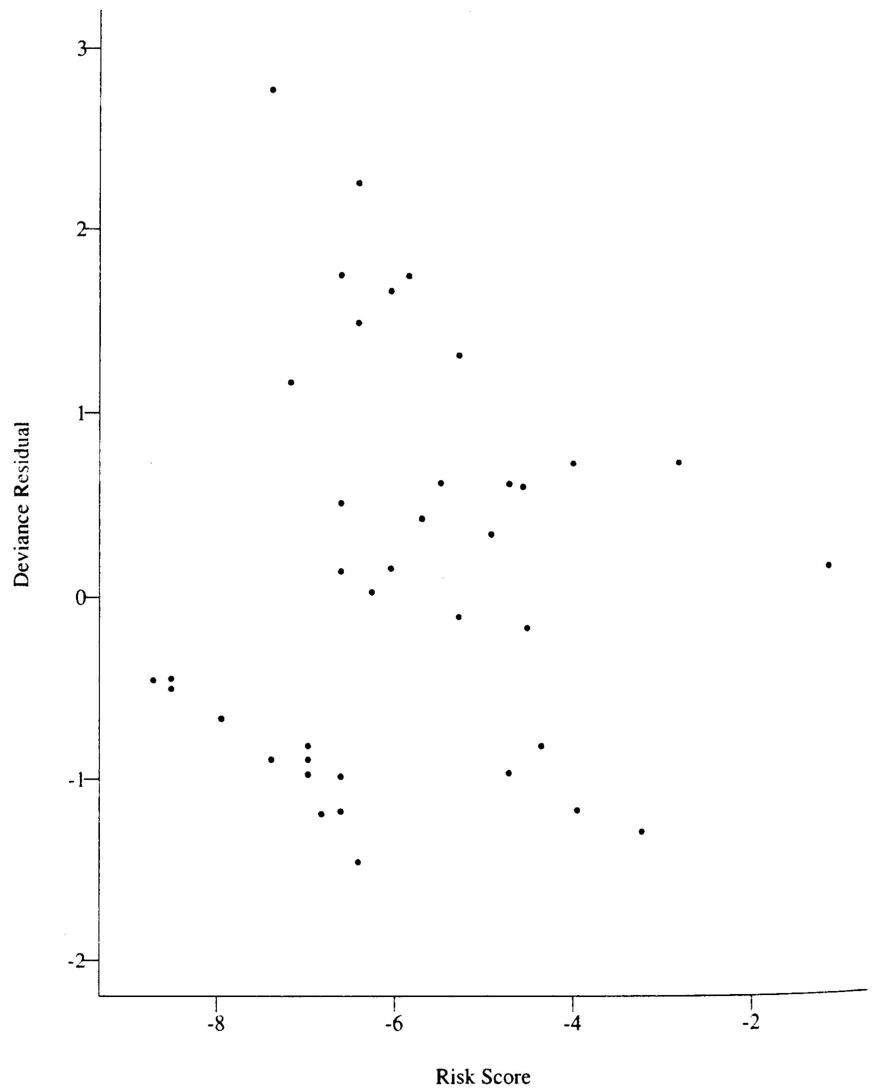


Figure 11.21 *Plot of the deviance residuals versus risk scores for the bone marrow transplant example*

§New. Other residuals

- References: p.376 in §11.4; Therneau and Grambsch, §6.2.
- *Schoenfeld (1982) residuals*: Assume no ties, no time-dependent covariate, at each time point t_i, $U_i = U(t_i) = Z_{(i)} - \bar{Z}(t_i)$, where $\bar{Z}(t_i) = \frac{\sum_{j \in R(t_i)} Z_j \exp(Z_j' \hat{\beta})}{\sum_{j \in R(t_i)} \exp(Z_j' \hat{\beta})}$.
- $U = \sum_{i=1}^{D} U_i$ is the score eq.
- U_i is a vector, as $Z_{(i)}$ and $\bar{Z}(t_i)$.
- With tied event times, then give multiple U_i at t_i, one for each observation with the tied event time.
- will be used later for model checking.
  time-varying coefficient models and GOF tests.

- *Score residuals*: with time-dependent covariates,
  $U = \sum_{j=1}^{n} S_j$
  $S_j = \int_0^\infty [Z_j(u) - \bar{Z}(u)] d\hat{M}_j(u)$, score residual.
  used to simplify delete-1 stat's for influence analysis.

§New. Time-dependent coefficient model

- Reference: Therneau and Grambsch, §6.2.
- Goal: a generalized version of a standard PHM; can be used to check the standard PHM.
- Standard PHM:
  $h(t|Z) = h_0(t)\exp(Z'\beta)$.
  Note: β are constants, do not change over t.
- Time-dependent **coefficient** PHM:
  $h(t|Z) = h_0(t)\exp(Z'\beta(t))$.
  Note: $\beta(t)$ is in general a function of t.
- Model checking:
  If $\beta(t) = const$, say β, then the standard PHM holds;
  otherwise, it gives evidence against the standard PHM.
- Basic idea:

Use scaled (or weighted, as called in SAS) Schoenfeld residuals s^*_{ij}; i for time point t_i, and j for component j of the covariate/coefficient vector.

- Theory: by Grambsch and Therneau (1995),
  $E(s^*_{ij}) + \hat{\beta}_j \approx \beta_j(t_i)$,
  where $\hat{\beta}_j$ is obtained from the standard PHM.
  $\Longrightarrow$ (nonparametrically) smooth $s^*_{ij} + \hat{\beta}_j$ over t to obtain $\hat{\beta}_j(t)$!
  And
- A formal check on each covariate:
  Plot $s^*_{ij} + \hat{\beta}_j$ against t_i or $g(t_i)$ (e.g. $\log(t_i)$);
  Fit a line;
  Test whether the slope $\theta_j = 0$.
  If yes, then $\hat{\beta}_j(t)$ is not constant, and thus ...
  Note: applies to each covariate j or β_j.
- A global check:

H_0: $\beta_1(t) = \beta_1$, $\beta_2(t) = \beta_2$,...
H_0': $\theta_1 = \theta_2 = ... = 0$

- Choice of $g(t)$:
  different $g(t)$ leads to different test;
  $g(t) = \log(t)$ leads to the score test of the zero-coefficient for $Z_j \log(t)$!
- Example: R

§11.6 Influence analysis

- Goal: to find influential observations.
  Outliers may or may not be influential.
  Influence on what?
- General model-fitting:
  General model-fitting measured by ...
  How to measure influence? Likelihood change/displacement with and without an observation.
- β
  $\Delta\beta_j = \hat{\beta}_j - \hat{\beta}_{j(-i)}$, DFBETA for each j.
  before and after deleting obs i.
  Overall?
  $\Delta\beta = (\hat{\beta} - \hat{\beta}_{(-i)})' V^{-1} (\hat{\beta} - \hat{\beta}_{(-i)})$, DFBETAS
- Brute force: requires fitting the model $n+1$ times; some tricks appply so that only fitting with the full data is needed.

Use score residuals: $\hat{\beta} - \hat{\beta}_{(-i)} \approx I(\hat{\beta})^{-1} S_i$
See eq (11.6.1) on p.385 for an expression of S_i.

- Example: SAS