Lecture 10

Proportional Hazards Regression Diagnostics

Statistics 255 - Survival Analysis

Presented March 1, 2016

Dan Gillen
Department of Statistics
University of California, Irvine

UCIrvine
University of California, Irvine

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Proportional Hazards Regression Diagnostics

Questions to address

- Are model assumptions correct?
  - Is the proportional hazards assumption correct?
  - Should covariates be left as is, or should they be transformed?
- Are there observations that are not well-captured by the model? Outliers?
- Are there observations with unduly-strong influence on the fitted model?

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Proportional Hazards Regression Diagnostics

Questions to address

- Some of these questions can be addressed with hypothesis tests
- In addition, these questions can be addressed graphically with residual plots:
  - Martingale residuals
  - Deviance residuals
  - Score residuals and delta-beta residuals
  - Schoenfeld residuals
- Here, consider only time-fixed (baseline) covariates; extensions exist

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Model Fit and Function Form

Martingale residuals

- Recall:
  - data for each subject is (y_i, δ_i, x_i)
  - δ_i "counts" the number of events for the ith subject (0 or 1)
  - $\widehat{\Lambda}_0(t)$ is an estimate of the *baseline* cumulative hazard function
  - Therefore,

$$\widehat{\Lambda}_i(t \mid x_i) = \widehat{\Lambda}_0(t) \exp(\widehat{\beta}^T x_i)$$

- Taking the ith subjects **total observation time** y_i:

$$\widehat{\Lambda}_i(y_i \mid x_i) = \widehat{\Lambda}_0(y_i) \exp(\widehat{\beta}^T x_i)$$

is an **estimate** of the expected value of δ_i

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Model Fit and Function Form

Martingale residuals

- Martingale residuals compare "observed" to "expected":

$$r_{Mi} = \delta_i - \widehat{\Lambda}_i(y_i \mid x_i)$$

- They are motivated by the fact that, *for large samples*, the quantity

$$\delta_i - \Lambda_i(Y_i \mid x_i)$$

  would be a *martingale* evaluated at the time Y_i

- In particular, under correct model specification they:
  - have mean zero
  - are uncorrelated with one another across subjects

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Model Fit and Function Form

Martingale residuals

- Interpretation:
  - δ_i is the *observed* number of events for the ith person (either 1 or 0)
  - $E_{Y_i}\{\Lambda_i(Y_i \mid x_i)\}$ is the *expected* number of events for the ith person, *accounting for censoring*
  - So r_{Mi} is like the "excess" number of events for the ith subject
  - It is like **observed** $-$ **expected**
  - In fact, these residuals sum to zero
- The residuals r_{Mi} can be used to examine overall model fit and whether transformation is needed in covariates, after other covariates have already been entered in the model

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Model Fit and Function Form

Martingale residuals

- Martingale residuals are very similar to residuals in linear regression
  - In particular, the functional form of covariate x_k is *very close* to the regression of r_{Mi} on x_{ik} (or, the residual of x_{ik} after regression onto the other x_{il}'s)
  - We can use martingale residuals to examine graphically whether certain covariates are important and what their functional form might be

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Model Fit and Function Form

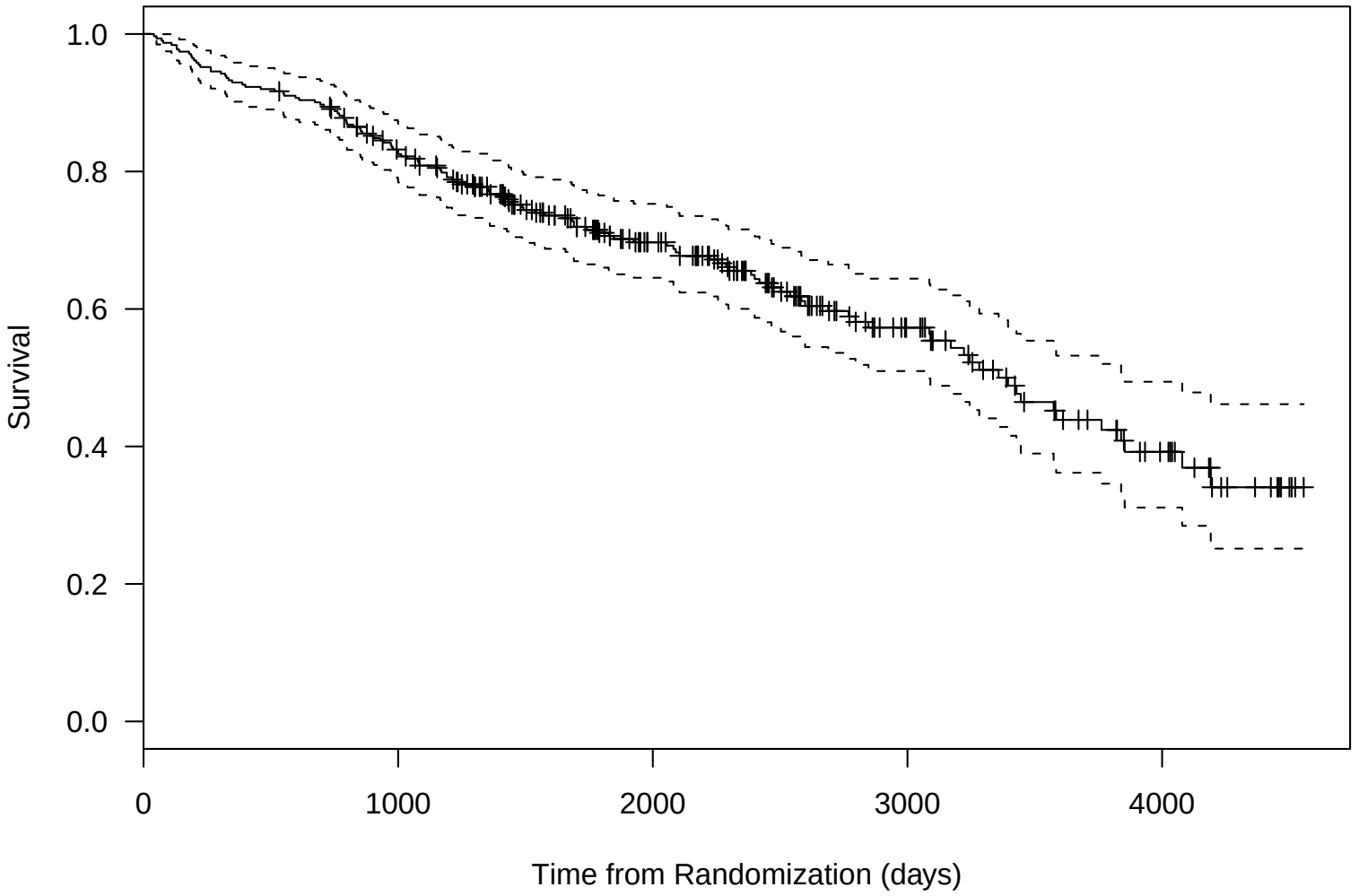
Ex: PBC Data (Fleming and Harrington, 1991)

- There were 424 patients referred to the Mayo Clinic with primary biliary cirrhosis (PBC) between January 1974 and May 1984.
- 312 of these were randomized to to treatment with D-penicillamine (DPCA).
- Clinical, biochemical, serologic and histologic measures were taken at intake.
- Subjects were followed up for mortality through July 1986. Censoring events were the end of study, LTFU or liver transplantation. 11 deaths are not attributable to PBC, but are apparently included as failures.
- We use the data here to develop a natural history model, ignoring treatment, to describe how survival depends on baseline status.

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Model Fit and Function Form

Ex: PBC Data (Fleming and Harrington, 1991)



Proportional Hazards Regression Diagnostics

Questions to address

Model Fit and Functional Form

Martingale residuals

Ex: PBC Data

Identification of Outliers

Deviance residuals

Assessment of Influence

Score residuals

Delta-beta values

Ex: PBC Data

The Proportional Hazards Assumption

Schoenfeld residuals

Summary

Model Fit and Function Form

Ex: PBC Data (Fleming and Harrington, 1991)

- The covariates of interest are
  - Albumin in g/dl
  - Serum bilirubin in mg/dl
  - Prothrombin time, in sec
  - Presence of edema

```
> summary(pbc[,c("age", "album", "protime", "bilir", "edema" )])
      age            album          protime
 Min.   :26.3   Min.   :1.96   Min.   : 9.0
 1st Qu.:42.2   1st Qu.:3.31   1st Qu.:10.0
 Median :49.8   Median :3.55   Median :10.6
 Mean   :49.9   Mean   :3.52   Mean   :10.7
 3rd Qu.:56.6   3rd Qu.:3.80   3rd Qu.:11.1
 Max.   :76.7   Max.   :4.64   Max.   :15.2
     bilir            edema
 Min.   : 0.30   Min.   :0.000
 1st Qu.: 0.80   1st Qu.:0.000
 Median : 1.35   Median :0.000
 Mean   : 3.26   Mean   :0.119
 3rd Qu.: 3.42   3rd Qu.:0.000
 Max.   :28.00   Max.   :1.000
```


Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Model Fit and Function Form

Ex: PBC Data (Fleming and Harrington, 1991)

- We will consider albumin, and prothrombin time on the log scale and consider doing so for bilirubin
- The starting model is one **without** bilirubin

```
> ##
> #####         Fit model without bilirubin
> ##
> fit <- coxph( Surv(time,death) ~ age + log(album) + log(protime)
                  + edema, data=pbc )
> summary(fit)
Call:
                    coef exp(coef)  se(coef)      z Pr(>|z|)
age              0.02764   1.02802   0.00961   2.88    0.004 **
log(album)      -4.02771   0.01782   0.65717 -6.13  8.8e-10 ***
log(protime)     5.99803 402.63670   1.04634   5.73  9.9e-09 ***
edema            0.56680   1.76262   0.23396   2.42    0.015 *

                exp(coef) exp(-coef) lower .95 upper .95
age                1.0280    0.97274   1.00885  1.05e+00
log(album)         0.0178   56.13238   0.00491  6.46e-02
log(protime)     402.6367    0.00248  51.79220  3.13e+03
edema              1.7626    0.56734   1.11432  2.79e+00
```


Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Model Fit and Function Form

Ex: PBC Data (Fleming and Harrington, 1991)

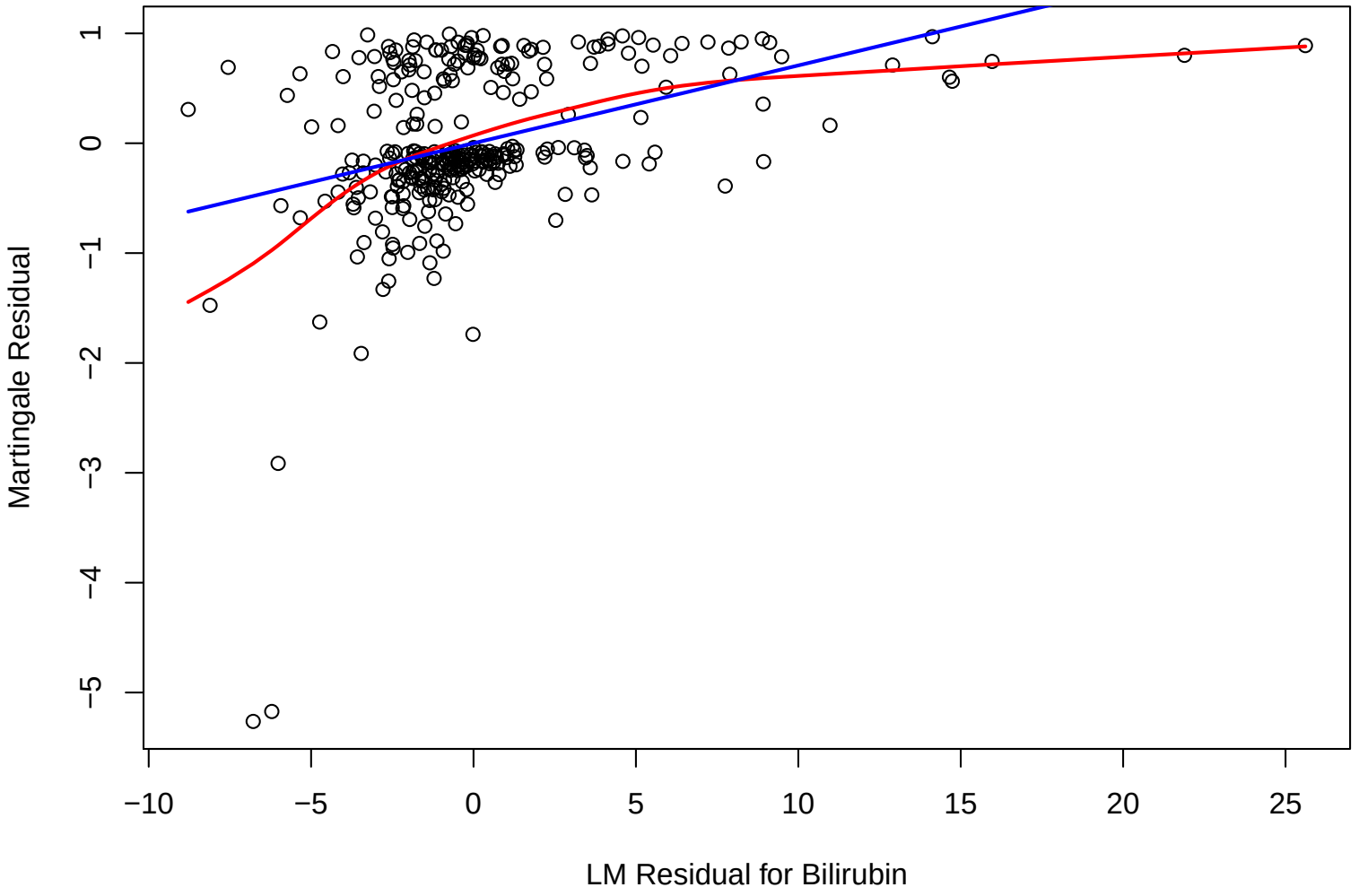
- Now, what is the correct functional form for bilirubin in the context of this model (that is, for predicting mortality risk, adjusting for the other covariates)?
- Martingale residual plot for bilirubin:
  - need to adjust for other covariates
  - use a smoother
  - include regression line

```
> mresids <- residuals( fit, type="martingale" )
> lmfit <- lm( bilir ~ age + log(album) + log(protime) + edema,
               data=pbc )
> rbili <- lmfit$resid
> ord <- order( rbili )
> mresids <- mresids[ ord ]
> rbili <- rbili[ ord ]
> plot( rbili, mresids )
> lines( smooth.spline( rbili, mresids, df=6  ), col="red", lwd=2 )
> lines( rbili, fitted(lm( mresids ~ rbili )), col="blue", lwd=2 )
```


Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Model Fit and Function Form

Ex: PBC Data (Fleming and Harrington, 1991)



Proportional Hazards Regression Diagnostics

Questions to address

Model Fit and Functional Form

Martingale residuals

Ex: PBC Data

Identification of Outliers

Deviance residuals

Assessment of Influence

Score residuals

Delta-beta values

Ex: PBC Data

The Proportional Hazards Assumption

Schoenfeld residuals

Summary

Model Fit and Function Form

Ex: PBC Data (Fleming and Harrington, 1991)

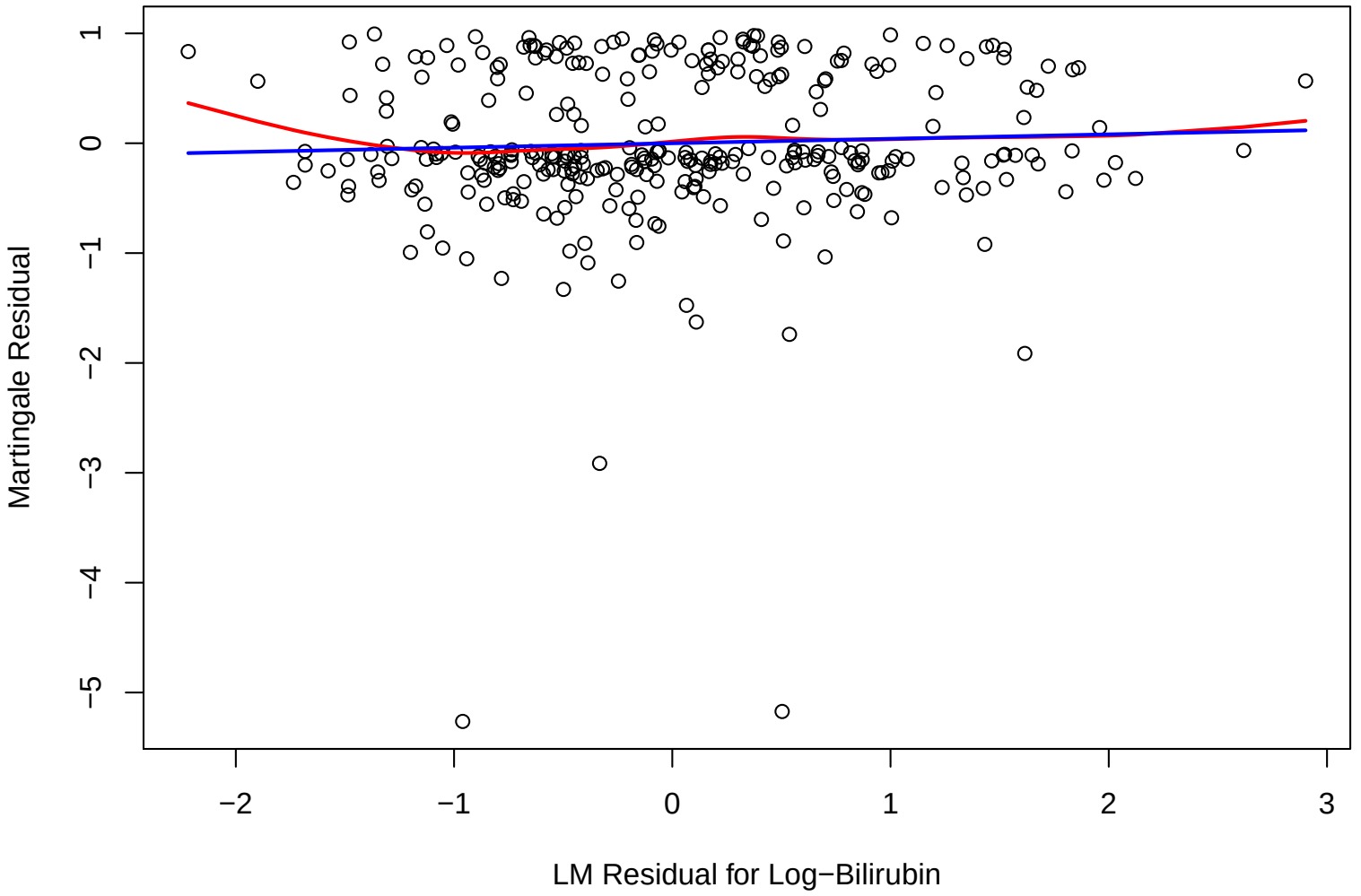
- Now, let's consider a log-transformation for biliruibin

```
> lmfit <- lm( log(bilir) ~ age + log(album) + log(protime) + edema,
              data=pbc )
> rlogbili <- lmfit$resid
> ord <- order( rlogbili )
> mresids <- mresids[ ord ]
> rlogbili <- rlogbili[ ord ]
> plot( rlogbili, mresids )
> lines( smooth.spline( rlogbili, mresids, df=6  ), col="red", lwd=2 )
> lines(rlogbili, fitted(lm( mresids ~ rlogbili )), col="blue", lwd=2)
```

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Model Fit and Function Form

Ex: PBC Data (Fleming and Harrington, 1991)



Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Model Fit and Function Form

Ex: PBC Data (Fleming and Harrington, 1991)

- Conclusion: In the context of this model, with the other 4 covariates, the effect of log(bilirubin) on the log mortality hazard is approximately linear

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Identification of Outliers

Deviance residuals

- **Outlier ($Y|X$-space)**: an unusual failure-time observation (y_i, δ_i), given the covariate value, x_i:

  – large (positive or negative) martingale or large deviance residual

- The martingale residual r_{Mi} is a measure of the degree to which the ith subject is an outlier, after adjusting for the effect of x_i

- **But note:** While martingale residuals are uncorrelated and have mean zero, their disadvantage is that:

  1. their maximum is $+1$, but their minimum is $-\infty$
  2. their distribution is quite skewed (left)

  The heavily skewed distribution of martingale residuals makes them hard to use to identify outliers

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Identification of Outliers

Deviance residuals

- For this, we have **deviance residuals**:

$$r_{Di} = \text{sign}(r_{Mi}) \left[-2 \left\{ r_{Mi} + \delta_i \log(\delta_i - r_{Mi}) \right\} \right]^{1/2}$$

- Why are they called deviance residuals?
  - From GLMs, the deviance of a model is defined as

$$\text{dev}(\text{model}) = 2[\log L(\text{saturated model}) - \log L(\text{model})]$$

    where a "saturated model" is one that *perfectly reproduces the data*

  - Deviance residuals are created in the same spirit. Namely,

$$\text{dev}(\text{model}) = \sum_i r_{Di}^2$$

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Identification of Outliers

Behavior of deviance residuals

- r_{Di} has the *same sign* as r_{Mi}:
  - The quantity inside the [ ]'s is positive (so we can take the square root), while sign$(\cdot)$ assures that the deviance residual has the same sign as the martingale residual
- What happens when . . .
  - $r_{Mi} \approx 0$?
  - r_{Mi} is close to 1
  - r_{Mi} is large and negative

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Identification of Outliers

Behavior of deviance residuals

- Compared to r_{Mi}, r_{Di} has a *shorter left* and a *longer right* tail
  $\longrightarrow$ r_{Di} is more symmetrical around zero
- The distribution of the deviance residual is *better approximated by a Gaussian distribution* than is the distribution of the martingale residuals
- Because they are approximately normally distributed, you can think of outliers as values outside of the range of $(-3, +3)$ or even $(-2.5, 2.5)$

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Identification of Outliers

Ex: PBC Data (Fleming and Harrington, 1991)

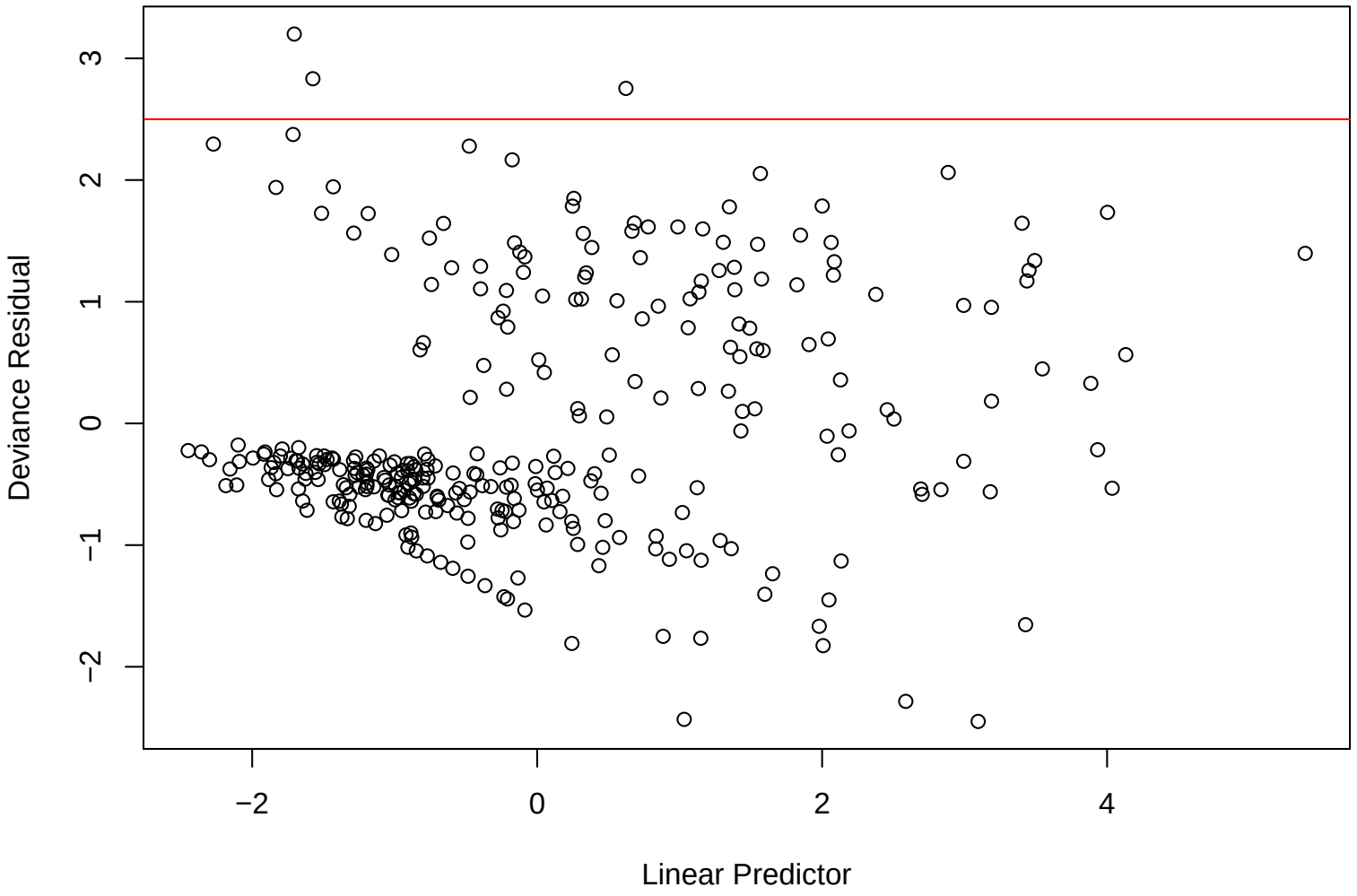
- Goal: Determine if there are any outliers in the model with all covariates, plus log(bilirubin)
- Approach: Plot residuals versus the "risk scores" $\widehat{\beta}^T x_i$:
- Start by fitting the model with log(bilirubin), then obtain linear predictor estimates and the deviance residuals...

```
> ##
> #####         Consider outliers (in the X-space)
> fit <- coxph( Surv(time,death) ~ age + log(album) + log(protime) +
                        edema + log(bilir), data=pbc )
> summary(fit)

              exp(coef) exp(-coef) lower .95 upper .95
age              1.0415     0.9602    1.0226     1.061
log(album)       0.0441    22.6812    0.0109     0.178
log(protime)    42.5586     0.0235    4.7547   380.937
edema            1.5055     0.6643    0.9450     2.398
log(bilir)       2.4623     0.4061    2.0257     2.993

> dresids <- residuals( fit, type="deviance" )
> lp <- predict( fit, type="lp" )
> plot(lp, dresids, xlab="Linear Predictor", ylab="Deviance Residual")
```


Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Identification of Outliers

Ex: PBC Data (Fleming and Harrington, 1991)



Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Identification of Outliers

Ex: PBC Data (Fleming and Harrington, 1991)

- Let's investigate the three outliers...

```
> summary(pbc[,c("age", "album", "protime", "bilir", "edema" )])
      age            album          protime          bilir
 Min.   :26.3   Min.   :1.96   Min.   : 9.0   Min.   : 0.30
 1st Qu.:42.2   1st Qu.:3.31   1st Qu.:10.0   1st Qu.: 0.80
 Median :49.8   Median :3.55   Median :10.6   Median : 1.35
 Mean   :49.9   Mean   :3.52   Mean   :10.7   Mean   : 3.26
 3rd Qu.:56.6   3rd Qu.:3.80   3rd Qu.:11.1   3rd Qu.: 3.42
 Max.   :76.7   Max.   :4.64   Max.   :15.2   Max.   :28.00
     edema
 Min.   :0.000
 1st Qu.:0.000
 Median :0.000
 Mean   :0.119
 3rd Qu.:0.000
 Max.   :1.000
```

```
> cbind( dresids,pbc[,c("time", "death", "age", "album", "protime",
                  "bilir", "edema" )] )[ abs(dresids) >= 2.5, ]
    dresids time death    age album protime bilir edema
87   3.2003  198     1 37.279  4.40    10.7   1.1     0
103  2.7535  110     1 48.964  3.67    11.1   2.5     1
119  2.8327  515     1 54.256  3.83     9.5   0.6     0
```


Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Assessment of Influence

Influence

- Consider only time-fixed (baseline) covariates
- **Outlier**: unusual (extreme) failure-time observation (y_i, δ_i), *given* the covariate value, x_i:

  $\longrightarrow$ large martingale or large deviance residual
- High **leverage** observation: an unusual observation with respect to the covariate (vector) x_i

  $\longrightarrow$ an "outlier in X-space"
- High **influence** observation:

  $\longrightarrow$ An observation for which the combination of the degree to which it is an outlier *and* its leverage means that it strongly influences estimates of β

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Assessment of Influence

How influence is operationalized

- Recall that the *martingale residual* is ...

$$r_{Mi} = \delta_i - \widehat{\Lambda}_i(Y_i \mid x_i)$$

- The martingale residual r_{Mi} is a measure of the degree to which the ith subject is an outlier, *after adjusting* for the effect of x_i ... and note that the martingale residual could be rewritten as

$$\begin{aligned} r_{Mi} &= \sum_{t_{(k)} \le Y_i} \left\{ \delta_i(t_{(k)}) - e^{\widehat{\beta}^T x_i} [\widehat{\Lambda}_0(t_{(k)}) - \widehat{\Lambda}_0(t_{(k-1)})] \right\} \\ r_{Mi} &= \sum_{t_{(k)} \le Y_i} \left\{ \delta_i(t_{(k)}) - e^{\widehat{\beta}^T x_i} d\widehat{\Lambda}_0(t_{(k)}) \right\} \\ r_{Mi} &= \sum_{t_{(k)} \le Y_i} r_{Mik} \end{aligned}$$

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Assessment of Influence

How influence is operationalized

- Here, $\delta_i(t) = 0$ for $t < Y_i$ and $\delta_i(Y_i) = \delta_i$, since Y_i is the "exit time"
- $\delta_i(t)$ "counts" the number of failure events for the ith subject, up to time t
- Also,

$$d\widehat{\Lambda}_0(t_{(k)}) = \widehat{\Lambda}_0(t_{(k)}) - \widehat{\Lambda}_0(t_{(k-1)})$$

  is the "jump" in the baseline CHF at time $t_{(k)}$
- The piece of martingale residual r_{Mik} is a measure of the degree to which the ith subject is an outlier *at time* $t_{(k)}$, after adjusting for the effect of x_i

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Assessment of Influence

How influence is operationalized

- **Leverage defined:**
  - The "weighted average" of covariate x_l at the observation time $t_{(k)}$ can be written

$$\bar{x}_l(t_{(k)}) = \frac{\sum_{i \in R_{(k)}} x_{il} \exp(\widehat{\beta}^T x_i)}{\sum_{i \in R_{(k)}} \exp(\widehat{\beta}^T x_i)}$$

  - Then the **leverage** of the ith subject for the lth covariate at time $t_{(k)}$ is

$$x_{il} - \bar{x}_l(t_{(k)})$$

  - This is the distance between x_{il} and the average x_l at $t_{(k)}$
  - This quantity is a measure of the degree to which the ith subject differs from the others in the risk set, *with respect to covariate* x_l, at time $t_{(k)}$

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Assessment of Influence

How influence is operationalized – Score residuals

- **Influence** is then operationalized as the integral of leverage times the martingale residuals:

$$r_{Sli} = \sum_{t_{(k)} \le Y_i} \underbrace{(x_{il} - \bar{x}_l(t_{(k)}))}_{\text{leverage}} \times \underbrace{\left\{ \delta_i(t_{(k)}) - e^{\widehat{\beta}^T x_i} d\widehat{\Lambda}_0(t_{(k)}) \right\}}_{\text{Martingale residual}}$$

  - Qualitatively, **influence** is the product of leverage and outlying tendency
  - The quantities r_{Sli} are called **score residuals**
  - There is one set of score residuals for *each covariate* x_{il} in the model, $l = 1, \ldots, p$
  - Large values of r_{Sli} imply large influence of the ith subject on the estimate of β_l, the coefficient for x_l
  - Obtain in R with `residuals(fit, type="score")`

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Assessment of Influence

How influence is operationalized – Delta-beta values

- **Delta-beta** values:
  - suppose $\widehat{\beta}_l$ is the estimate of β_l from the whole data set
  - and, suppose $\widehat{\beta}_{l(i)}$ is the estimate of β_l from the data set *with the* ith *subject removed*
  - the quantity (called a **delta-beta**):

$$\Delta\beta_{li} = \widehat{\beta}_l - \widehat{\beta}_{l(i)}$$

    is a measure of the *influence* of the ith subject on the estimate of β_l

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Assessment of Influence

How influence is operationalized – Delta-beta values

- As it turns out, $\Delta\beta_{li}$ can be approximated by:

$$\Delta\beta_{li} = \widehat{\beta}_l - \widehat{\beta}_{l(i)} \approx \widehat{V}_l \cdot r_{Si}$$

where r_{Si} is the vector

$$r_{Si} = (r_{S1i}, \ldots, r_{Spi})$$

of score residuals for the ith subject (across all covariates)and $\widehat{V}_l$ is the lth row of the estimated variance-covariance matrix of $\widehat{\beta}$

- Each subject i has one $\Delta\beta$ value for each covariate in the model

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Assessment of Influence

Ex: PBC Data (Fleming and Harrington, 1991

- Goal: Investigate the influence of observations on the coefficients of in the 5-variable model which we have been investigating

```
> ##
> #####        A look at delta-betas for influential points
> ##
> dfbeta <- residuals( fit, type="dfbeta" )
> colnames( dfbeta ) <- names(fit$coef)
> summary( dfbeta )
      age                log(album)          log(protime)
 Min.   :-4.49e-03   Min.   :-1.96e-01   Min.   :-4.43e-01
 1st Qu.:-6.50e-05   1st Qu.:-1.07e-02   1st Qu.:-1.15e-02
 Median : 4.75e-05   Median :-1.67e-03   Median : 3.25e-03
 Mean   : 1.72e-18   Mean   :-4.58e-17   Mean   : 1.10e-16
 3rd Qu.: 1.86e-04   3rd Qu.: 6.30e-03   3rd Qu.: 1.82e-02
 Max.   : 2.28e-03   Max.   : 1.95e-01   Max.   : 3.09e-01
     edema              log(bilir)
 Min.   :-1.08e-01   Min.   :-6.55e-02
 1st Qu.:-7.98e-04   1st Qu.:-4.07e-04
 Median : 4.52e-05   Median : 8.29e-04
 Mean   :-7.09e-16   Mean   : 1.65e-17
 3rd Qu.: 1.72e-03   3rd Qu.: 2.10e-03
 Max.   : 5.48e-02   Max.   : 2.58e-02
```


Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Assessment of Influence

Ex: PBC Data (Fleming and Harrington, 1991

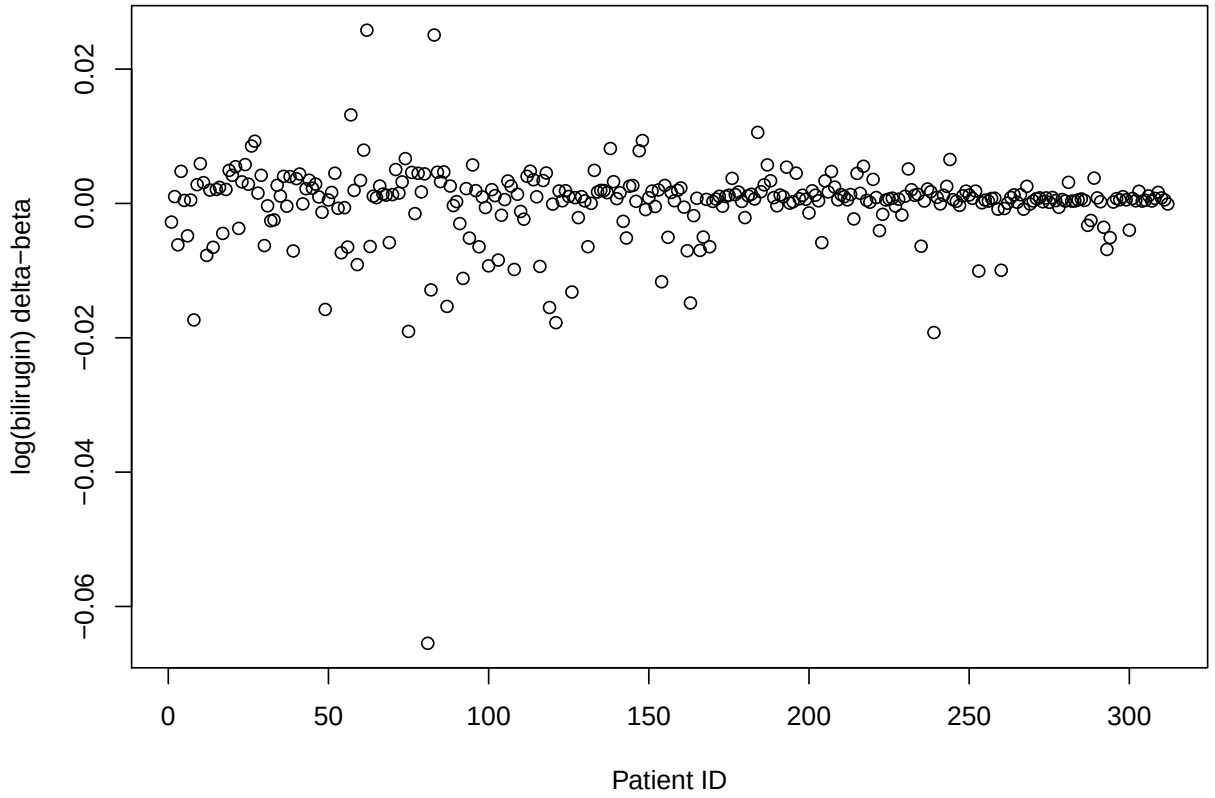
- Conclusion: For log(albumin), log(protime) and edema, no single very influential observations. For age, one observation has a large negative influence. For log(bilirubin), one has a large negative influence.
- Let's plot and print out the influential observation

```
> plot( pbc$id, dfbeta[,5], xlab="Patient ID",
          ylab="log(bilirugin) delta-beta" )
```


Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Assessment of Influence

Ex: PBC Data (Fleming and Harrington, 1991




```
> pbc[ dfbeta[,5] < -.04, ]
      age album alkph ascites bilir chol edema edematx hepat
81 63.264  3.65  1218       0  14.4  448     0       0     1
   time plate protime sex  sgot spiders stage death treat trigl
81 2540   385    11.7   1 60.45       1     4     1     1   318
   ucopp rand id
81    34    1 81
```


Proportional Hazards Regression Diagnostics

Questions to address

Model Fit and Functional Form

Martingale residuals

Ex: PBC Data

Identification of Outliers

Deviance residuals

Assessment of Influence

Score residuals

Delta-beta values

Ex: PBC Data

The Proportional Hazards Assumption

Schoenfeld residuals

Summary

Assessment of Influence

Ex: PBC Data (Fleming and Harrington, 1991

- **Conclusion:** Subject 81 is older and has a high serum bilirubin (2 sd above mean on log scale). Bilirubin is an important predictor of high risk, yet subject is in the 40th (or so) percentile of survival times
- **Recommendation:** If interest is on assessing the effect of bilirubin, might do a sensitivity analysis (ie. present results with this case and without)
- **Important:** Unless it is *very, very* clear that there is some sort of data entry error causing a problem, it is generally never a good idea to permanently remove an observation from your data!

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

The Proportional Hazards Assumption

Schoenfeld residuals

- **Recall:** If we have proportional hazards, then

$$\lambda_1(t) = \phi\lambda_0(t)$$

for all t, so that

$$\log \Lambda_1(t) = \log(\phi) + \log \Lambda_0(t)$$

- Thus, the log cumulative hazards should be parallel if the proportional hazards assumption holds.
- We looked at unadjusted and adjusted versions of these plots for categorical variables earlier (See Lectures 4 and 8)

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

The Proportional Hazards Assumption

Schoenfeld residuals

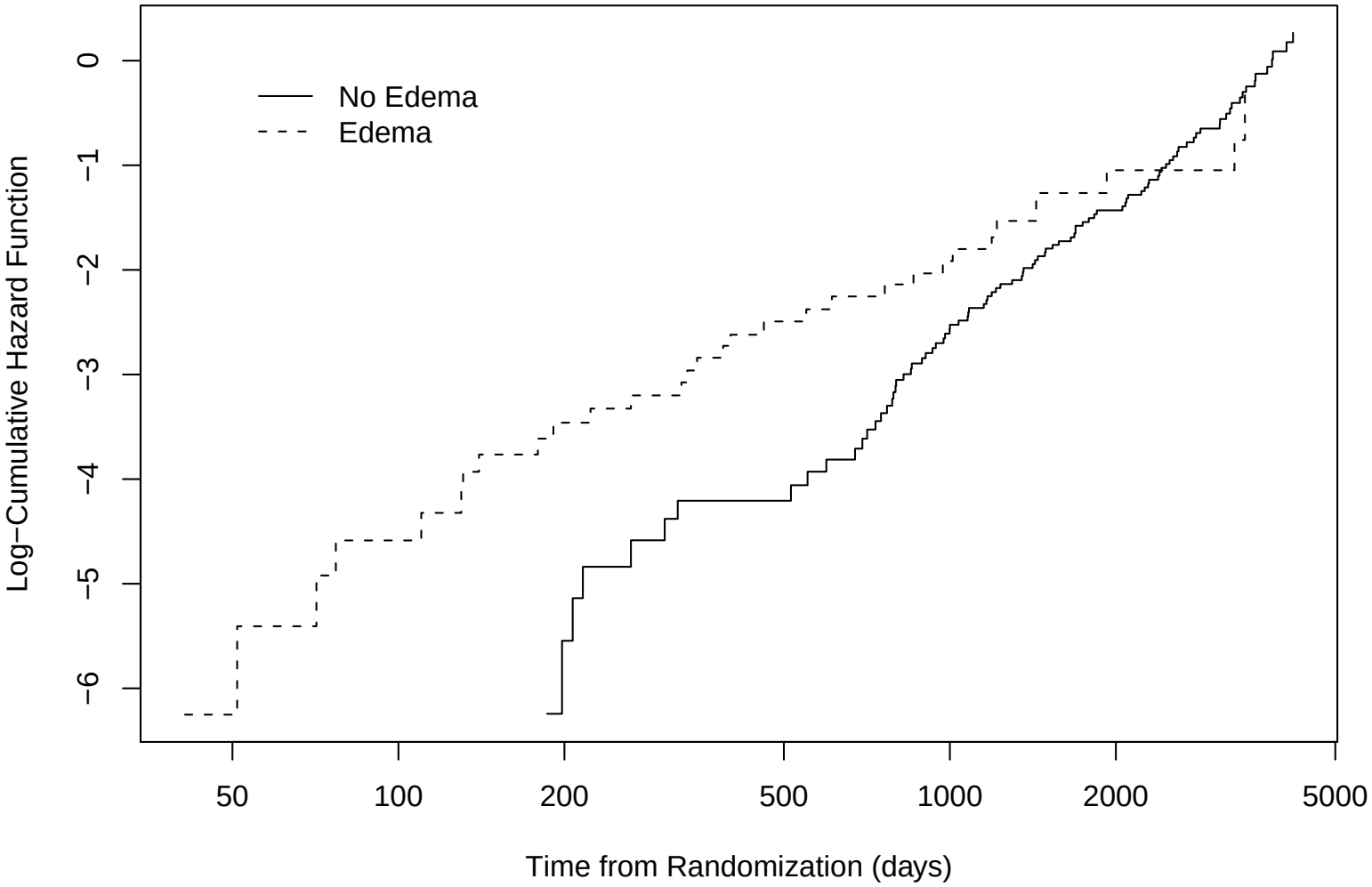
- Let's consider whether `edema` exhibits a non-proportional hazards effect

```
> fit <- coxph( Surv(time,death) ~ age + log(album) +
              log(protime) + log(bilir) + strata(edema), data=pbc )
> plot( survfit(fit), fun="cloglog", lty=1:2,
+         xlab="Time from Randomization (days)",
+         ylab="Log-Cumulative Hazard Function" )
> legend( 50, 0, lty=1:2, legend=c("No Edema", "Edema"), bty="n" )
```


Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

The Proportional Hazards Assumption

Schoenfeld residuals



Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

The Proportional Hazards Assumption

Schoenfeld residuals

- Clearly the log cumulative hazards are not parallel
- This suggests that the proportional hazards assumption may be violated, ie. The hazard ratio associated with edema may be changing with respect to time.
- We have looked at one test of this assumption using time dependent covariates. Another relies upon the *Schoenfeld* residuals...

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

The Proportional Hazards Assumption

Schoenfeld residuals

- **Recall:** Under the Cox model the probability that any particular member j of $\mathcal{R}(t_k)$ fails at t_k, given that one does, is

$$w_j(\beta, t_k) = \frac{e^{\beta^T x_j}}{\sum_{l \in \mathcal{R}(t_k)} e^{\beta^T x_l}}$$

- The (weighted) average of the covariate values for members of $\mathcal{R}(t_k)$, with weights proportional to $w_j(\beta, t_k)$, is

$$\bar{x}(\beta, t_k) = \sum_{j \in \mathcal{R}(t_k)} x_j w_j(\beta, t_k)$$

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

The Proportional Hazards Assumption

Schoenfeld residuals

- The **Schoenfeld residual** for any subject $i \in \mathcal{D}(t_k)$ (the set of d_k failures at time t_k) is the *difference* between the covariate for that subject and the weighted average of covariates in the risk set, namely

$$x_i - \bar{x}(\beta, t_k)$$

- The sum of the Schoenfeld residuals over all d_k subjects who fail at t_k, also known as the **Schoenfeld residual corresponding to** t_k, is

$$r_{S,k} = r_{S,k}(\beta) = \sum_{i \in \mathcal{R}(t_k)} \delta_{ik}[x_i - \bar{x}(\beta, t_k)]$$

where δ_{ik} equals one if subject i fails at t_k and zero otherwise.

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

The Proportional Hazards Assumption

Schoenfeld residuals

- Provided the PH model holds and β is the true regression coefficient, the $r_{S,k}(\beta)$ are *uncorrelated* and have *mean zero*.

- In practice the Schoenfeld residuals are calculated as

$$\hat{r}_{S,k} = r_{S,k}(\widehat{\beta})$$

  where $\hat{\beta}$ is the partial likelihood estimate of the regression coefficients.

- Schoenfeld residuals are also known as *partial score residuals*, because their total equals the partial likelihood *score*, or estimating equation, whose solution is $\hat{\beta}$:

$$\sum_k r_{S,k}(\hat{\beta}) = 0$$

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

The Proportional Hazards Assumption

Schoenfeld residuals

- **Scaled Schoefeld residuals** are residuals after multiplication by the inverse weighted covariance matrix of $\hat{\beta}$:

$$r^*_{S,k} = r^*_{S,k}(\beta) = V^{-1}(\beta) r_{S,k}(\beta)$$

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

The Proportional Hazards Assumption

Schoenfeld residuals

- **Key Point:** When the scaled Schoenfeld residuals, $\hat{r}^*_{S,k}$ are plotted against any transformation $g(t_k)$ of time t_k, for example $\log(t_k)$ or t_k itself, the smooth curve through the plotted points approximates the manner in which the associated coefficients depend on time.
  - If a specific covariate has a time-varying coffecient (effect):

$$\beta(t) = \beta + \gamma g(t)$$

where $g(t)$ is a specified function of time t, such as $g(t) = t$ or $g(t) = \log(t)$, then the approximate expectation of the scaled Schoenfeld residual at time t_k is

$$E[\hat{r}_{S,k}] \approx \gamma g(t)$$

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

The Proportional Hazards Assumption

Schoenfeld residuals

- This suggests:
  - Plotting $\hat{r}^*_{S,k}$ against $g(t_k)$ and examining trends
  - Slope of linear regression gives numerator of the score statistic, $\hat{\gamma}$ for testing $H_0 : \gamma = 0$ (proportionality)
  - This test is implemented in R via the `cox.zph()` command

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

The Proportional Hazards Assumption

Schoenfeld residuals

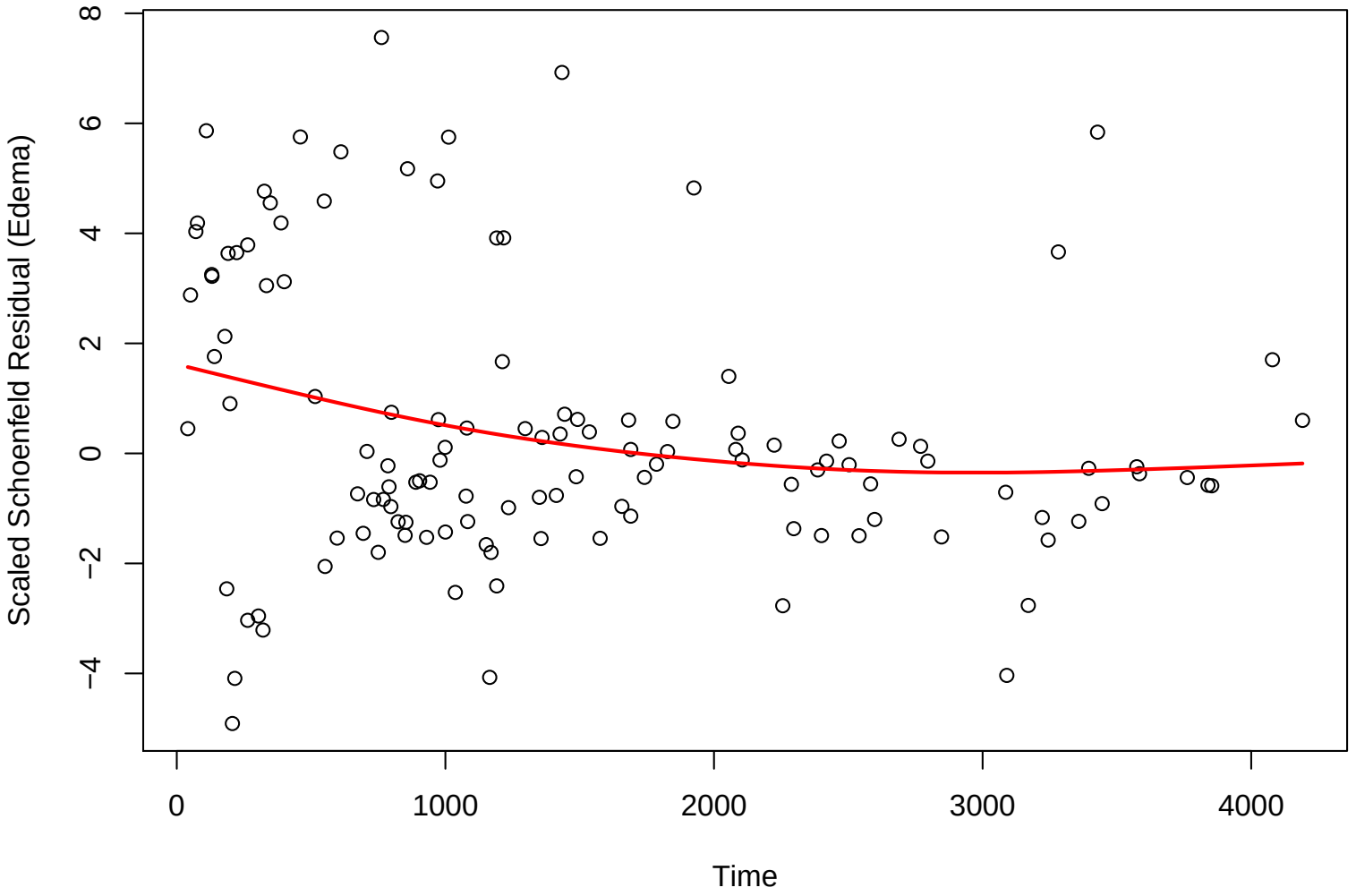
- Goal: Test each covariate in the PBC data to determine if any significantly violate the PH assumption
- This can be done using the function `cox.zph()`
- First let's plot the scaled Schoenfeld residuals for edema and prothrombin vs. time

```
> fit <- coxph( Surv(time,death) ~ age + log(album) +
            log(protime) + log(bilir) + edema, data=pbc )
> sresids <- residuals( fit, type="scaledsch" )
> colnames( sresids ) <- names( fit$coef )
> time <- as.numeric( rownames( sresids ) )

> plot( time, sresids[,5], xlab="Time",
      ylab="Scaled Schoenfeld Residual (Edema)" )
> lines( smooth.spline( time, sresids[,5] ), col="red", lwd=2 )

> plot( time, sresids[,3], xlab="Time",
       ylab="Scaled Schoenfeld Residual (Log-Prothrombin Time)" )
> lines( smooth.spline( time, sresids[,3] ), col="red", lwd=2 )
```


Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

The Proportional Hazards Assumption

Schoenfeld residuals



Proportional Hazards Regression Diagnostics

Questions to address

Model Fit and Functional Form

Martingale residuals

Ex: PBC Data

Identification of Outliers

Deviance residuals

Assessment of Influence

Score residuals

Delta-beta values

Ex: PBC Data

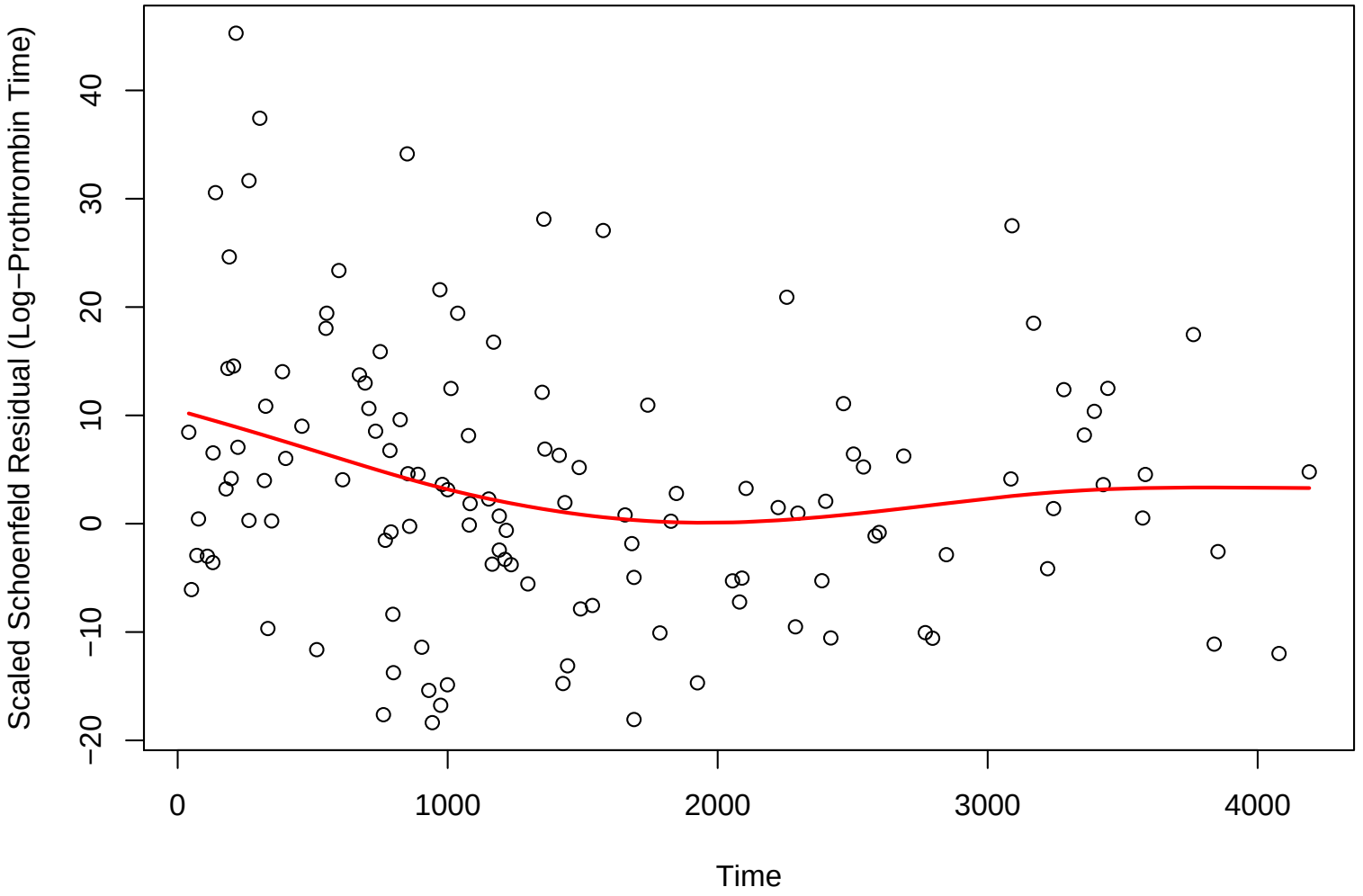
The Proportional Hazards Assumption

Schoenfeld residuals

Summary

The Proportional Hazards Assumption

Schoenfeld residuals



Proportional Hazards Regression Diagnostics

Questions to address

Model Fit and Functional Form

Martingale residuals

Ex: PBC Data

Identification of Outliers

Deviance residuals

Assessment of Influence

Score residuals

Delta-beta values

Ex: PBC Data

The Proportional Hazards Assumption

Schoenfeld residuals

Summary

The Proportional Hazards Assumption

Schoenfeld residuals

- Now, let's test the slopes using `cox.zph()`

```
> cox.zph( fit, transform="identity" )
                 rho  chisq      p
age          -0.0610  0.461 0.4971
log(album)   -0.0431  0.237 0.6262
log(protime) -0.1570  2.967 0.0850
log(bilir)    0.1154  1.563 0.2112
edema        -0.2195  5.407 0.0201
GLOBAL           NA 12.197 0.0322
```


Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

The Proportional Hazards Assumption

Schoenfeld residuals

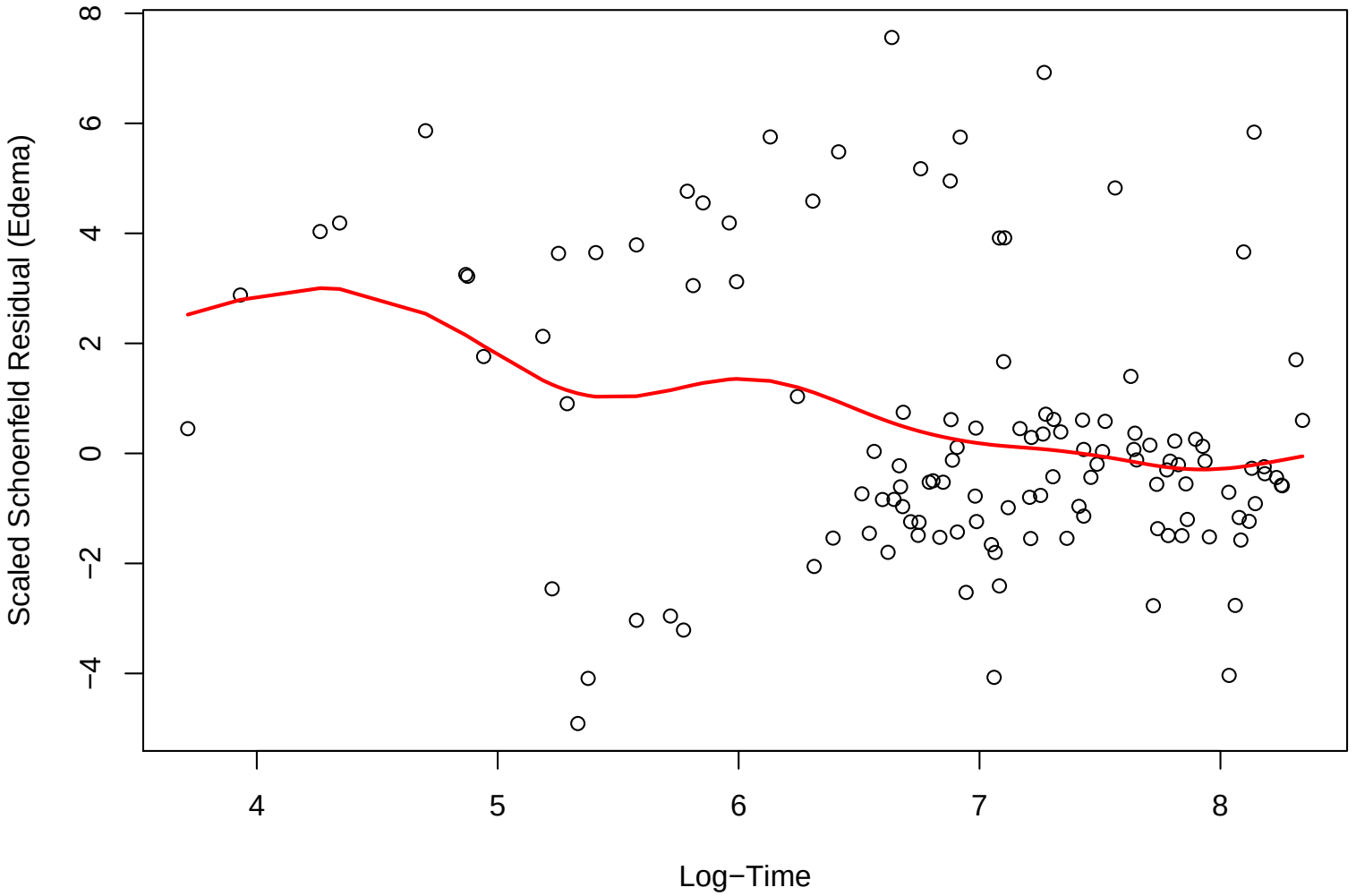
- In each case smoother certainly appears to have a pattern and we reject the proportionality assumption with our hypothesis test. But does the effect really change linearly in time? The smoother is not linear in either case.
- Let's explore the relationship between edema (and log(prothrombin)) and log(time)?

```
> plot( log(time), sresids[,5], xlab="Log-Time",
        ylab="Scaled Schoenfeld Residual (Edema)" )
> lines( smooth.spline( log(time), sresids[,5], df=6 ), col="red", lwd=2 )

> plot( log(time), sresids[,3], xlab="Log-Time",
       ylab="Scaled Schoenfeld Residual (Log-Prothrombin Time)" )
> lines( smooth.spline( log(time), sresids[,3], df=6 ), col="red", lwd=2 )
```

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

The Proportional Hazards Assumption

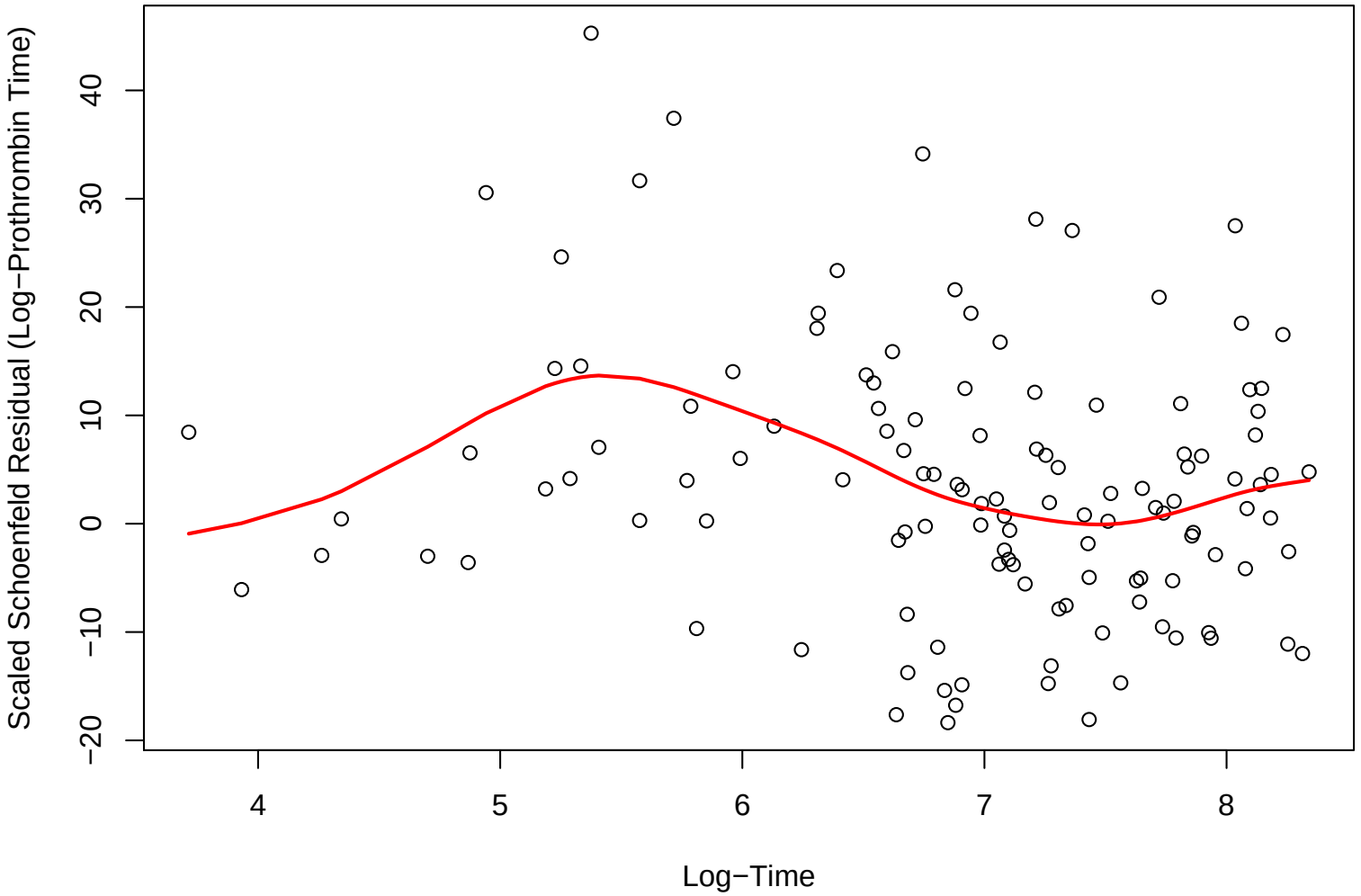
Schoenfeld residuals



Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

The Proportional Hazards Assumption

Schoenfeld residuals



Proportional Hazards Regression Diagnostics

Questions to address

Model Fit and Functional Form

Martingale residuals

Ex: PBC Data

Identification of Outliers

Deviance residuals

Assessment of Influence

Score residuals

Delta-beta values

Ex: PBC Data

The Proportional Hazards Assumption

Schoenfeld residuals

Summary

The Proportional Hazards Assumption

Schoenfeld residuals

- Again, let's test the slopes using `cox.zph()`

```
> cox.zph( fit, transform=log )
                 rho  chisq       p
age          -0.0878  0.955 0.32857
log(album)   -0.0313  0.125 0.72334
log(protime) -0.2005  4.844 0.02774
log(bilir)    0.1068  1.338 0.24747
edema        -0.2809  8.853 0.00293
GLOBAL           NA 20.221 0.00114
```


Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

The Proportional Hazards Assumption

Schoenfeld residuals

- Conclusion: We reject the null hypothesis of proportional hazards for both log(prothombin)and edema, and conclude that their effect varies as a function of log(time).
- Compare the results for edema with what we found looking at time-varying covariates!
- Note: We are attempting to disprove the proportional hazards assumption. Just because we fail to reject the null hypothesis does not guarantee proportional hazards, our test may just be underpowered.

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Model Diagnostics: Summary

Summary

- Model:
  - Proportional hazards assumption:
  - Functional form of covariates (log, square-root, etc.):

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary

Model Diagnostics: Summary

Summary

- Observations:
  - Observations not well-described by the model (outliers):
  - Observations with undue influence on results:

    (here, "results" refers to one β at a time)

Proportional Hazards Regression Diagnostics
Questions to address
Model Fit and Functional Form
Martingale residuals
Ex: PBC Data
Identification of Outliers
Deviance residuals
Assessment of Influence
Score residuals
Delta-beta values
Ex: PBC Data
The Proportional Hazards Assumption
Schoenfeld residuals
Summary