# Model Choice & Checking

- Confirmatory Goals.
- General Empirical Model Development.
- Residuals.
  - Martingale residuals
- Influence
- Predictive Model Development & Assessment.
  - Bias versus variance
- Accuracy Ideas for Survival Data.
  - Extension of $R^2$?
  - ROC curves? C index?

# Regression Analysis

**Q:** What are the goals of regression analysis?

**A:** Estimation, Testing, & Prediction

- Estimation of the "effect" of one variable (exposure), called the *predictor of interest* (POI), after "adjusting", or controlling for other measured variables.
  - ▷ Remove confounding effects.
  - ▷ Remove bias.
- Testing whether variables are associated with the response.
- Prediction of a response variable given a collection of covariates.

# Regression Analysis

- First step is to identify the scientific question. Apriori consideration of the goals of analysis is crucial!

- $\boxed{\text{Classification of Variables:}}$
  - ▷ **Response variable**
    - ∗ Dependent variable.
    - ∗ Outcome variable.
  - ▷ **Predictor of interest**
    - ∗ Exposure variable.
    - ∗ Treatment assignment.

- $\boxed{\text{Classification of Variables:}}$ continued...

  - **Confounding variables**
    - Associated with response and POI.
    - Not intermediate.
  - **Precision variables**
    - Associated with response and not POI.
    - Reduces response uncertainty.

# Model Building Strategies

Kleinbaum, *Logistic Regression* Chapter 6:

"Most epidemiologic research studies in the literature... provide a minimum of information about modeling methods used in the data analysis."

"Without meaningful information about the modeling strategy used, it is difficult to assess the validity of the results provided. Thus, there is need for guidelines regarding modeling strategy to help researchers know what information to provide."

"In practice, most modeling strategies are *ad hoc*; in other words, researchers often make up a strategy as they go along in their analysis. The general guidelines that we recommend here encourage more consistency in the strategy used by different researchers."

Information often <u>not</u> provided:

1. how variables chosen / selected
2. how effect modifiers assessed
3. how confounders assessed

# Model Goals

$\boxed{\text{Main Model Goals:}}$

- Valid "effect" estimation (exposure $\rightarrow$ disease)
- Good prediction of the outcome
- Parsimonious description of correlations / associations

# Model Goals

Classification of Analysis:

- **Confimatory Data Analysis (CDA)**
  - Formal hypothesis testing
  - Protocol contains analysis plans

- **Exploratory Data Analysis (EDA)**
  - Hypothesis generating
  - Write / create analysis plans
  - Confirmatory studies to follow

# Guidelines for CDA

Carefully decide the scientific question that you want answered.

Outcome: **T** = survival time

Exposure of Interest: **E**

$\boxed{\text{Variable Specification}}$

- Restrict attention to clinically or biologically meaningful variables.
  - Study goals
  - Literature review
  - Theoretical basis
  - Define these variables as $C_1, C_2, \ldots, C_p$

# Guidelines for CDA

- Decide how you will use the variables
  - Will you include interactions among these covariates?
  - Call the combinations $V_j$ (ie. this may be just be the $C$'s or may include some $C_j C_k$ terms)
- Decide what interactions between **E** and the $V_j$ are to be considered.
  - In most cases the number of *apriori* interactions is limited.

# Guidelines for CDA

Kleinbaum (1994):
"In general, regardless of the number of $V$'s in one's model, the method for assessing confounding when there is no interaction is to monitor changes in the effect measure corresponding to different subsets of potential confounders in the model."

"To evaluate how much of a change is a **meaningful** change when considering the collection of coefficients... is quite **subjective**."

Recommendations:

- 10% change in the measure of interest
  - Mickey & Greenland (1989) AJE
  - Maldonado & Greenland (1993) AJE

# Regression Models

"It will rarely be necessary to include a large number of variables in the analysis, because only a few exposures are of genuine scientific interest in any one study, and there are usually very few variables of sufficient *a priori* importance for their potential confounding effect to be controlled for. Most scientists are aware of the dangers of analyses which search a long list of potentially relevant exposures. These are known as *data dredging* or *blind fishing* and carry considerable danger of false positive findings. Such analyses are as likely to impede scientific progress as to advance it.

There are similar dangers if a long list of potential confounders is searched, either with a view to explaining the observed relationship between disease and exposure or to enhancing it – findings will inevitably be biased. Confounders should be chosen *a priori* and not on the basis of statistical significance."

Clayton and Hills (1993)
Statistical Methods in Epidemiology
page 273

# Multiple Comparisons

Example:

Hilsenbeck, Clark, and McGuire

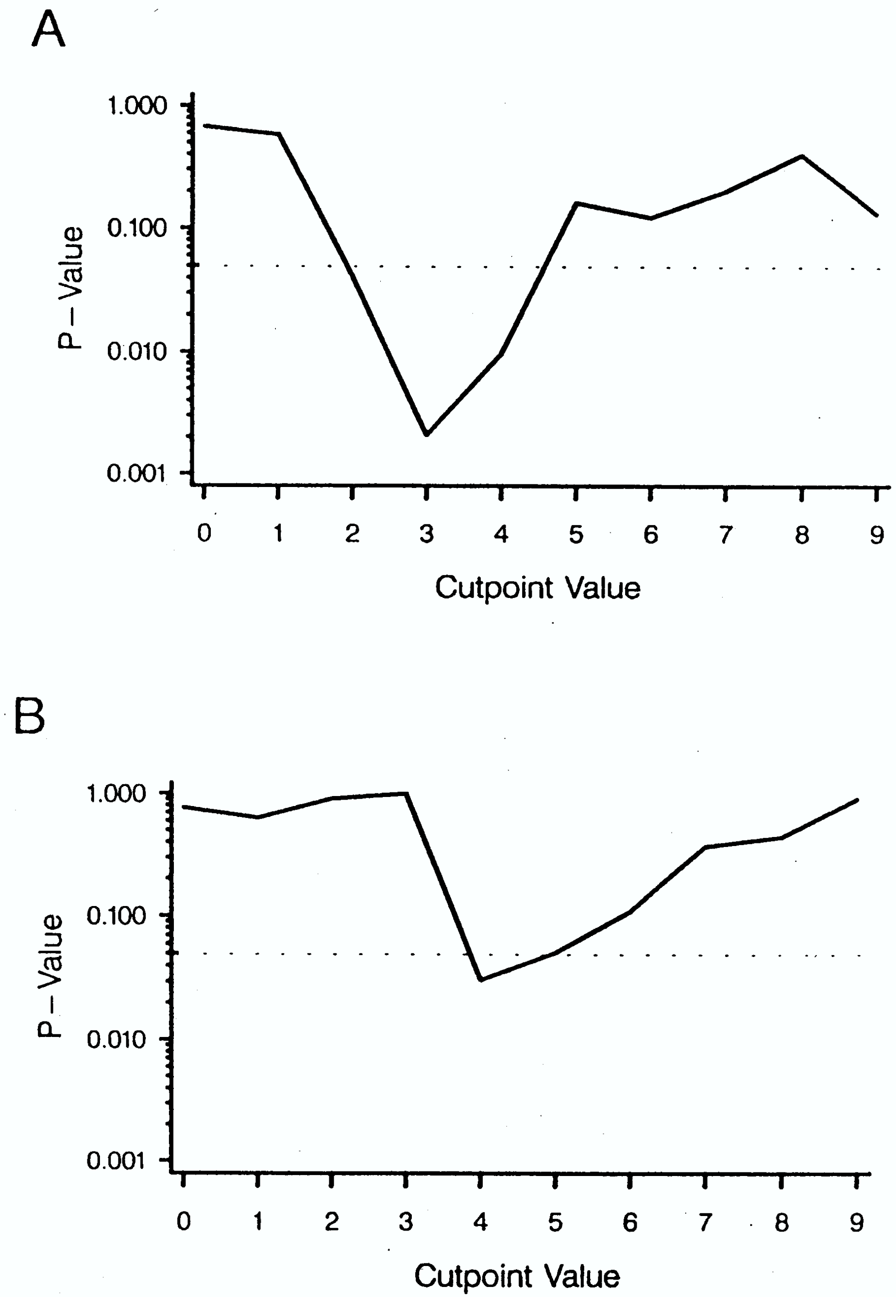
"Why do so many prognostic factors fail to pan out?"

*Breast Cancer Research and Treatment*

(1992) **22**: 197-206

- **Figure 1**: Cutpoint analysis curves for typical simulated datasets ($n = 250$) with: A) a true 10% difference in 5-year relapse-free survival; and B) no difference in 5-year RFS.

- **Figure 3**: Type I errors rates for training and validation datasets with no difference in prognosis. Square = sample size of 250, Diamond = sample size of 125, dased line is observed rates, solid line is fitted non-linear regression.


A
1.000
0.100
0.010
0.001
P – Value
0
1
2
3
4
5
6
7
8
9
Cutpoint Value
B
1.000
0.100
0.010
0.001
P – Value
0
1
2
3
4
5
6
7
8
9
Cutpoint Value

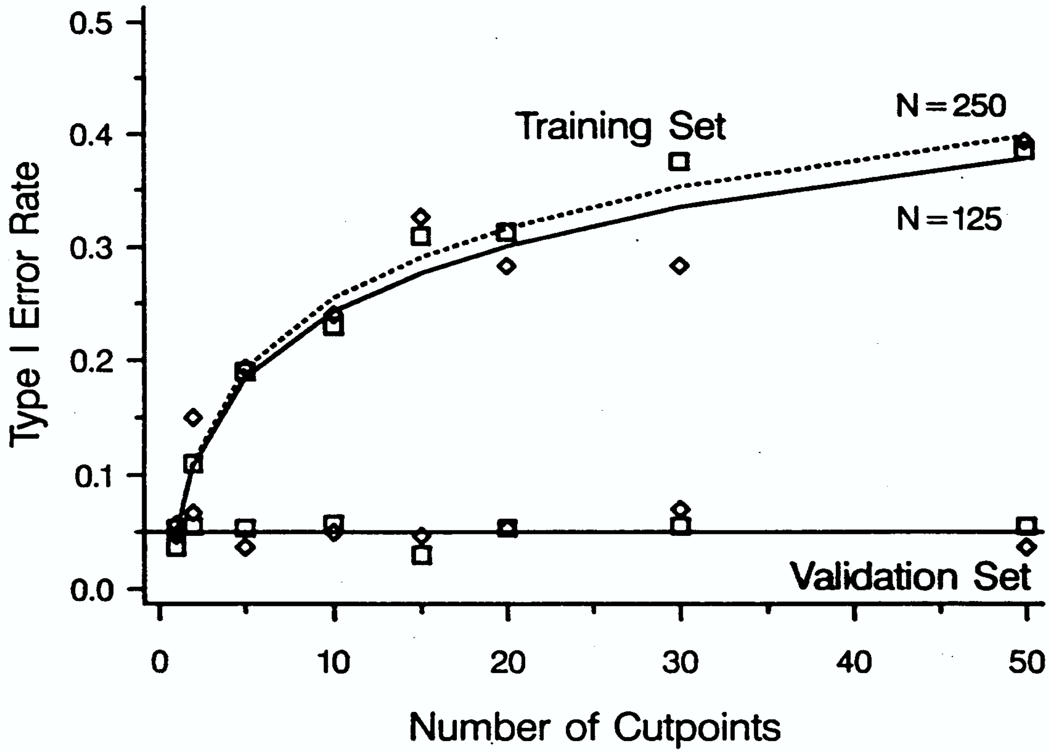



*Figure 3.* Type I error rates for training and validation datasets with no difference in prognosis. Observed rates for samples of 250 (□) and 125 (◇), and the corresponding fitted nonlinear regression lines (--- and —, respectively).

# Empirical Model Building

$\boxed{\text{A Strategy}}$ Cox and Wermuth (1996) section 7.2

1. Establish the main scientific research question.
2. Check Data Quality
   - Look for:
     - Possible errors in coding.
     - Outliers.
     - Missing values.
   - Produce univariate summaries.

3. Classification of variables based on substantive grounds.

- Document pairwise associations:
  - correlations
  - mean differences with s.e.'s
  - log odds ratios with s.e.'s
- Additional stratified analyses for key variables.

4. Regression models developed:
   - Main effects regression.
   - Forward selection for strong interactive effects.
   - Backward selection for stable effects.
   - Summary of standardized regression coefficient(s) for each omitted variable if added individually in order to reassure that no important further effects have been overlooked.
5. Presentation of model(s):
   - Coefficients and s.e.'s
   - Graphical display

# Empirical Model Building

$\boxed{\text{Model Development}}$ Cox and Wermuth (1996) section 7.3

1. Specify required explanatory variables:
   - Predictor of interest.
   - Known important variables (a priori).
2. Regress response on all explanatory variables considered. Include nonlinear terms (ie. age$^2$) that are *a priori* considered important.
3. Eliminate (individually) variables that have small standardized regression coefficients. Identify one or more simple, well-fitting models.

4. Examine, one term at a time, the impact of adding squared terms and cross-product terms into the model.

   Any significant terms uncovered in this way need detailed interpretation before we decide how to incorporate them into the "final" model.

5. List the contribution of omitted primary explanatory variables if added back to the model one at a time (individually) to check that no important effect has been overlooked.

# Statistical Thinking

Breslow (1999)

"As a medical statistician, I am appalled by the large number of irreproducible results published in the medical literature. There is a general, and likely correct, perception that this problem is associated more with statistical, as opposed to laboratory, research. I am convinced, however, that results of clinical and epidemiological investigations could become more reproducible if only the investigators would apply more rigorous statistical thinking and adhere more closely to well established principles of the scientific method. While I agree that the investigative cycle is an iterative process, I believe that is works best when it is hypothesis driven."

# Statistical Thinking

Breslow (1999)

"The epidemiology literature is replete with irreproducible results stemming from the failure to clearly distinguish between analyses that were specified in the protocol and that test the *a priori* hypotheses whose specification was needed to secure funding, and those that were performed *post-hoc* as part of a serendipitous process of data exploration."

## Some Recommendations

- For CDA fit three regression models:
  - ▷ Unadjusted
  - ▷ Adjusted for known confounders
  - ▷ Adjusted for known confounders, and candidate confounders
- **Analysis Plan**:
  - ▷ Scientific Aims
  - ▷ Classification of variables (by group)
  - ▷ List of Basic Tables
  - ▷ Details for Primary Regression Analyses by Aim
  - ▷ Details for Secondary Regression Analyses by Aim

## Model Selection / Building Summary

- ⋆ Determine the scientific question.
- ⋆ Classify the measured variables.
- ⋆ Thorough univariate and bivariate summaries.
- ⋆ Choose an analysis plan.
  - Confirmatory regression.
    formal analysis plan *a priori*
    limited number of models (three!)
  - Exploratory / Empirical regression.
    more flexible
    have a strategy! (ie. reproducible analysis)

# Model Checking Methods

- We have seen that we can check the PH assumption using:
  - ▷ Global and individual tests (`stphtest`)
  - ▷ Evaluate $\beta(t)$ using scaled Schoenfeld residuals.
- **Q**: How can assess whether $X_j$ is modeled using an appropriate functional form?
  - ▷ Use splines to create a flexible relationship, and plot the fitted values.
  - ▷ Use **Martingale residuals** to evaluate non-linearity.
- **Q**: How can we assess whether certain individuals have a large influence on the fitted model?
  - ▷ Calculate **delta-betas** to estimate impact of dropping each subject.

# Martingale Residuals

- $\boxed{\text{Recall:}}$ used to check PH
  - ▷ Schoenfeld residual = "observed covariate" - "expected covariate"
  - ▷ Only for observed failure times.
- $\boxed{\text{Martingale Residual:}}$

$$\widehat{M}_i = \delta_i - \widehat{\Lambda}_0(t_i) \exp(\widehat{\beta}_1 X_{1i} + \ldots + \widehat{\beta}_k X_{ki})$$

  - ▷ $\delta_i$ : event indicator for subject $i$.
  - ▷ $\widehat{\Lambda}_0(t_i)$ : Estimated cumulative hazard at final follow-up time for subject $i$.
  - ▷ $\exp(\widehat{\beta}_1 X_{1i} + \ldots)$ : estimated coefficients applied to observed covariate for subject $i$.

## Martingale Residuals

- $\boxed{\text{Use:}}$
  - ▷ These residuals can be plotted against covariates, $X_j$, that are either included in the model, or excluded, to see if appropriately modeled.
  - ▷ With time-dependent covariates it is more difficult to generate these since we use:

$$\Lambda(t) = \int_0^t \lambda(s) ds \ = \ \int_0^t \lambda_0(s) \exp[\beta X(s)] ds$$

- **Q**: Justification?
  - ▷ These are justified by the "counting process" theory that is used for non- and semi-parametric analysis of censored survival data.

# Counting Process Idea (* = extra)

- $\boxed{\text{Define:}}$ $N(t) = 1(T^* \leq t, \delta = 1)$
- Consider breaking the time axis into very small intervals $t, t + \Delta$.

$$\begin{aligned} E[N(t+\Delta) - N(t) \mid T \geq t] &= P[T \in (t, t+\Delta) \mid T \geq t] \\ &\approx \Delta \cdot \lambda(t) \end{aligned}$$

- This comes from the basic definition of the hazard.
- Define $dN(t) = N(t + \Delta) - N(t)$. Let $t_j = j \cdot \Delta$. Then we have

$$N(t) = \sum_{t_j < t} dN(t_j)$$

# Counting Process Idea (* = extra)

- From this we can show

$$
\begin{aligned}
E[N(t)] &= \sum_{t_j < t} \Delta \cdot \lambda(t_j) \\
&\approx \int_0^t \lambda(s) ds = \Lambda(t) \\
0 &= E[N(t) - \Lambda(t)]
\end{aligned}
$$

- And furthermore, if we provide information about survival (and perhaps covariates) throught time $s < t$ then it can be shown that:

$$
E[N(t) - \Lambda(t) \mid T \geq s] = N(s) - \Lambda(s)
$$

- This particular property is known as the "martingale property".
- See Hosmer & Lemeshow (1999) Appendix 2.

## Deviance Residuals

- The Martingale residuals, $\widehat{M}_i$, have a skewed distribution:
  - ▷ maximum possible value for $\widehat{M}_i$: 1
  - ▷ mininum possible value for $\widehat{M}_i$: $-\infty$
- Transformations to achieve a more symmetric distribution are helpful.
- One such transformation is motivated by **deviance** residuals used for logistic and poisson regression.
- $\boxed{\text{Define:}}$

$$d_i = \mathsf{sign}(\widehat{M}_i)\sqrt{2}\sqrt{-\widehat{M}_i - \delta_i \log(\delta_i - \widehat{M}_i)}$$

# Deviance Residuals

- Note that $d_i = 0$ only when $\widehat{M}_i = 0$.
- The square root shrinks the large negative martingale residuals, while the logarithm transformation expands those residuals that are close to zero.
- $\boxed{\text{Usage:}}$
  - ▷ Again, these residuals can be plotted against covariates, $X_j$, that are either included in the model, or excluded, to see if appropriately modeled.
  - ▷ Plot versus observation number as an indication of the discrepancy between the fitted model and the observed data for each observation.

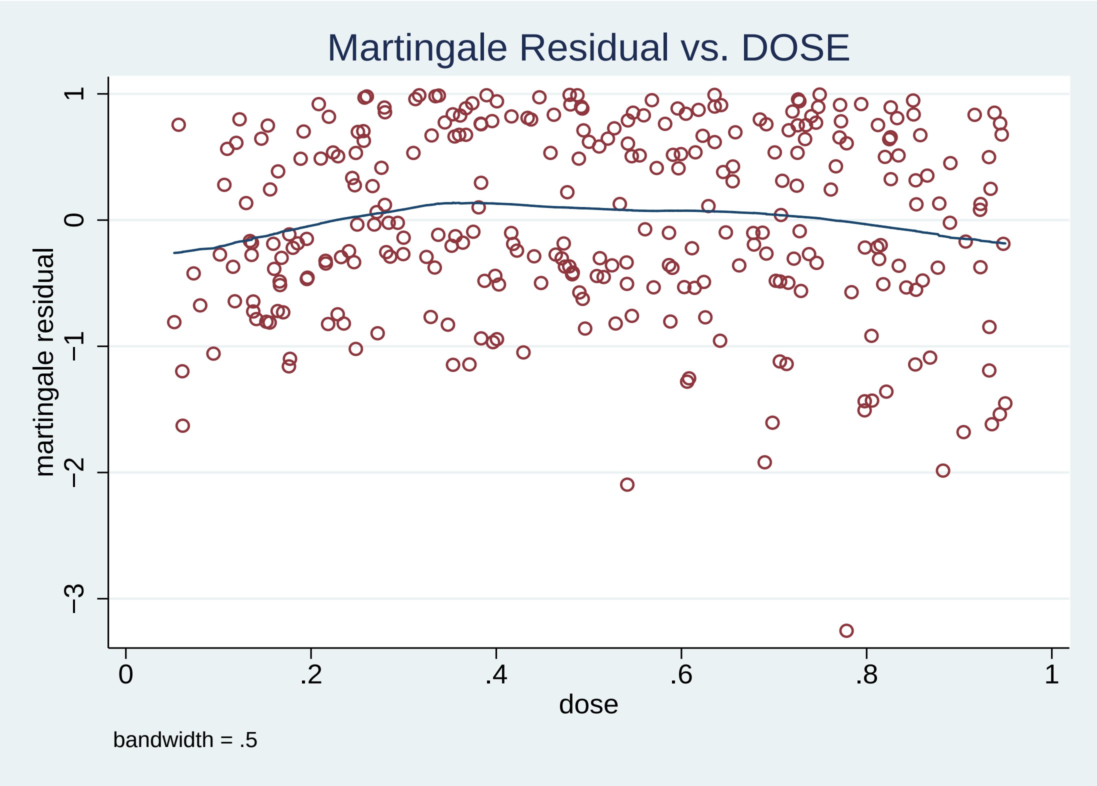
## Illustration with Simulated Data

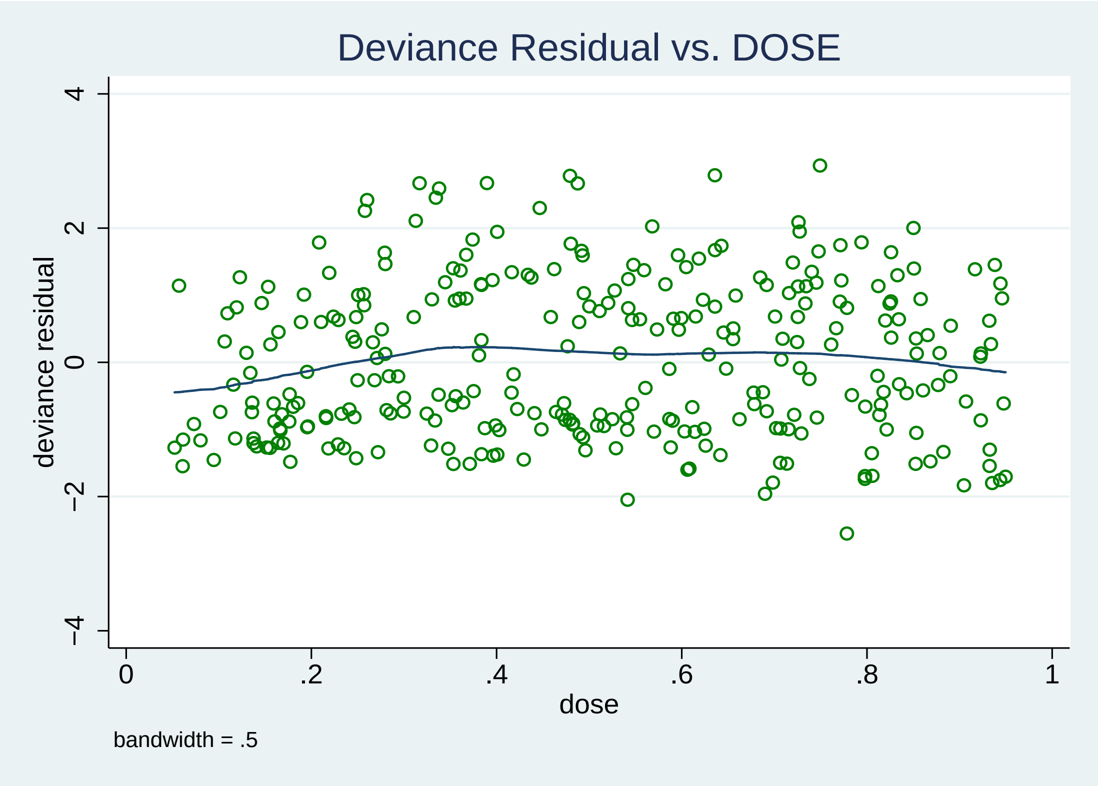
- To learn about the behavior of these residuals we explore simulated data where we know what the "right" functional form is:

$$\lambda(t \mid \texttt{Dose}) = \lambda_0(t) \exp[\beta \cdot \underbrace{\texttt{log}(\texttt{Dose})}_{X}]$$

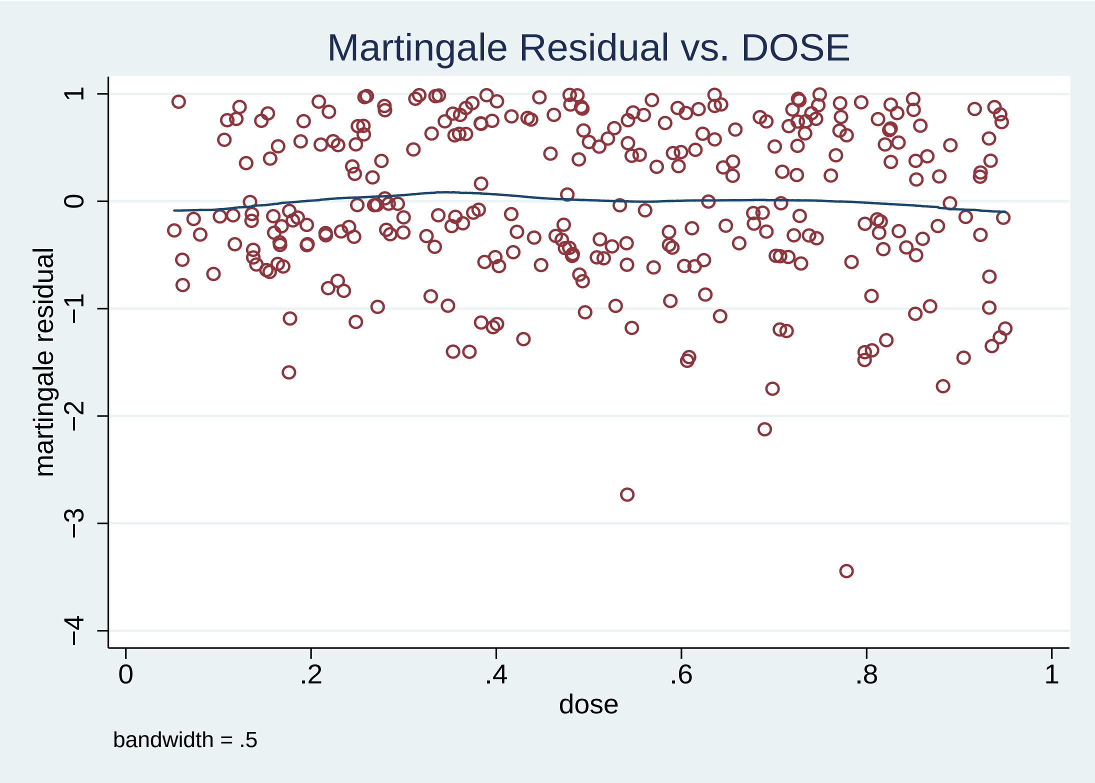
- We consider the residuals from different fits:
  - ▷ Model that assumes log hazard is linear in **Dose**
  - ▷ Model that assumes log hazard is linear in **log-Dose**
  - ▷ Model that assumes log hazard follows a **linear spline in Dose**
- $N = 300$ observations; 40% censored.

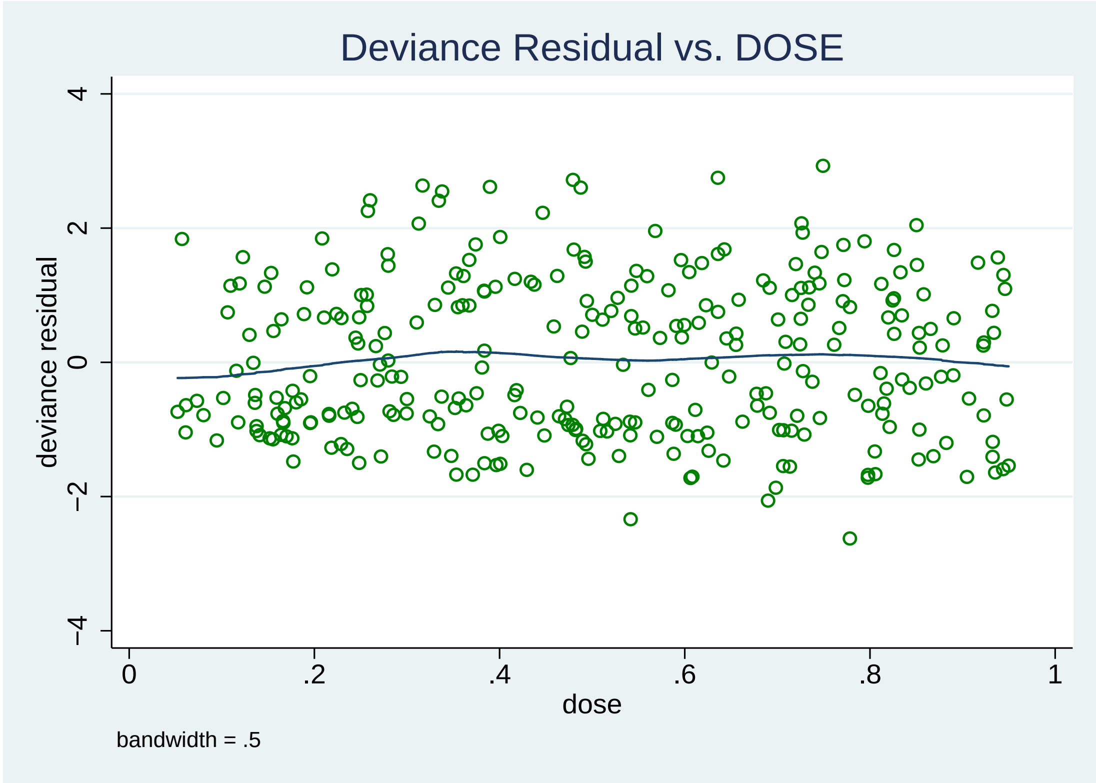
# log hazard linear in DOSE



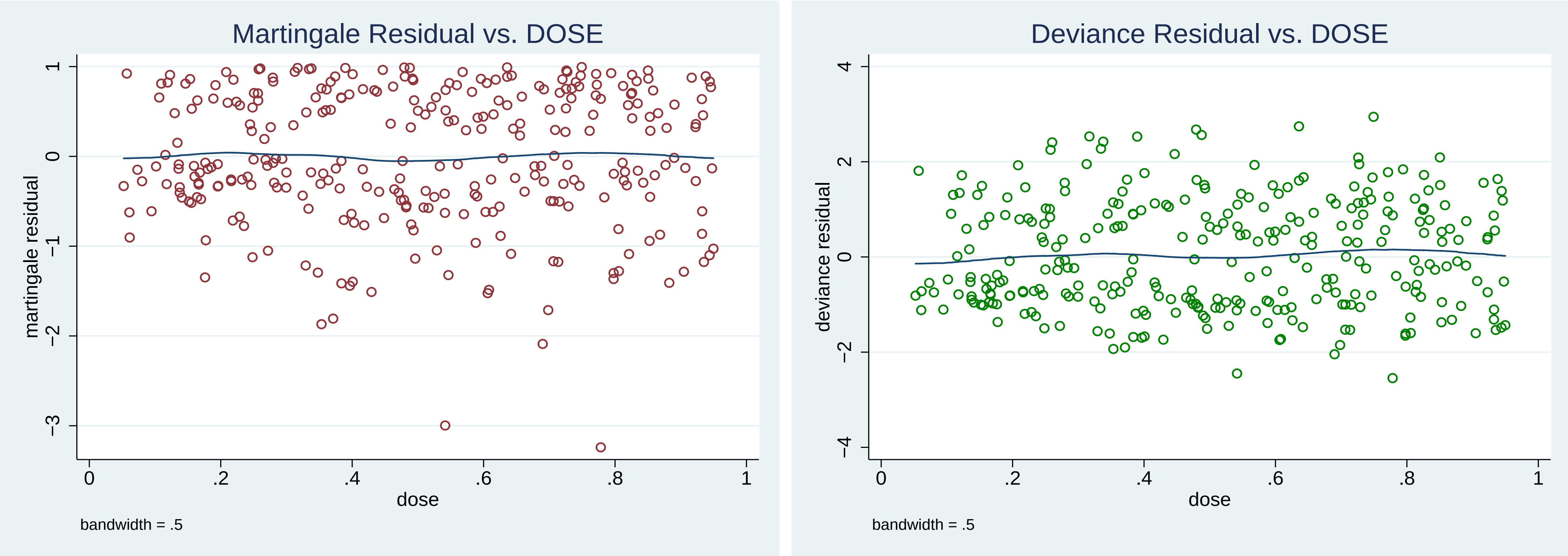



# log hazard linear in logDOSE

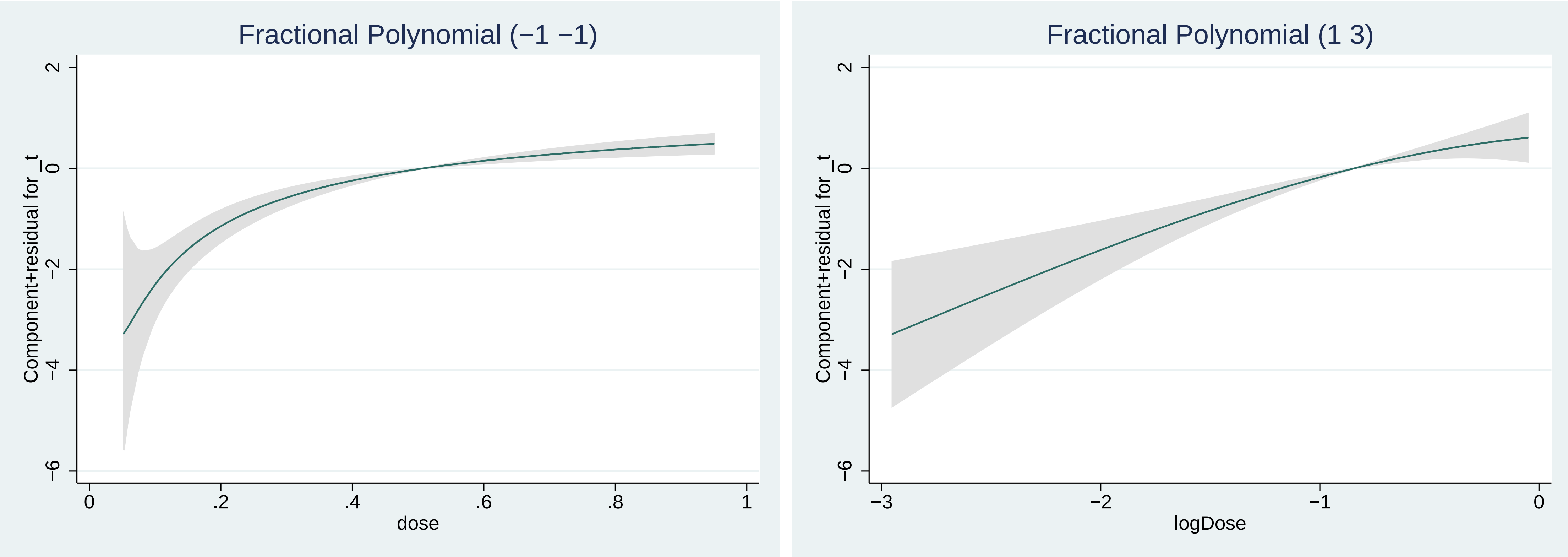


# log hazard linear in DOSE SPLINES



# log hazard curve using `fracpoly`



## Summary: Simulation Illustration

- We can use the martingale and/or deviance residuals to look for **systematic** patterns in the residuals which suggest a lack of fit (non-linearity).

- We can estimate the "dose response" curve (log hazard) using either standard spline methods and/or fractional polynomial methods.

- The spline approach allows inference regarding violation of "linear" assumptions for continuous covariates.

# Example: Mayo PBC Data

- To obtain **martingale** residuals with STATA we modify the Cox regression call so that these are created and saved:
  - ▷ `stcox x1 x2 x3, mgale(`*varname*`)`

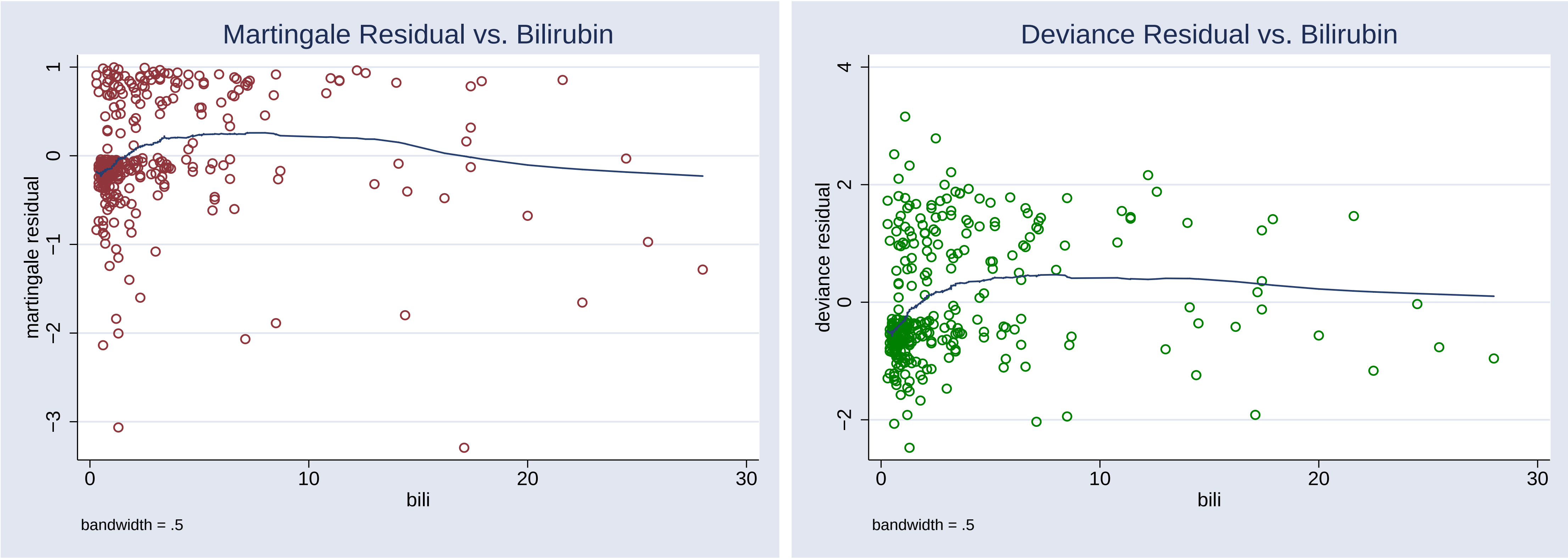
- To obtain **deviance** residuals with STATA we use the "predict" command following a Cox regression fit:
  - ▷ `predict` *varname*, `deviance`

# Example: Mayo PBC Data

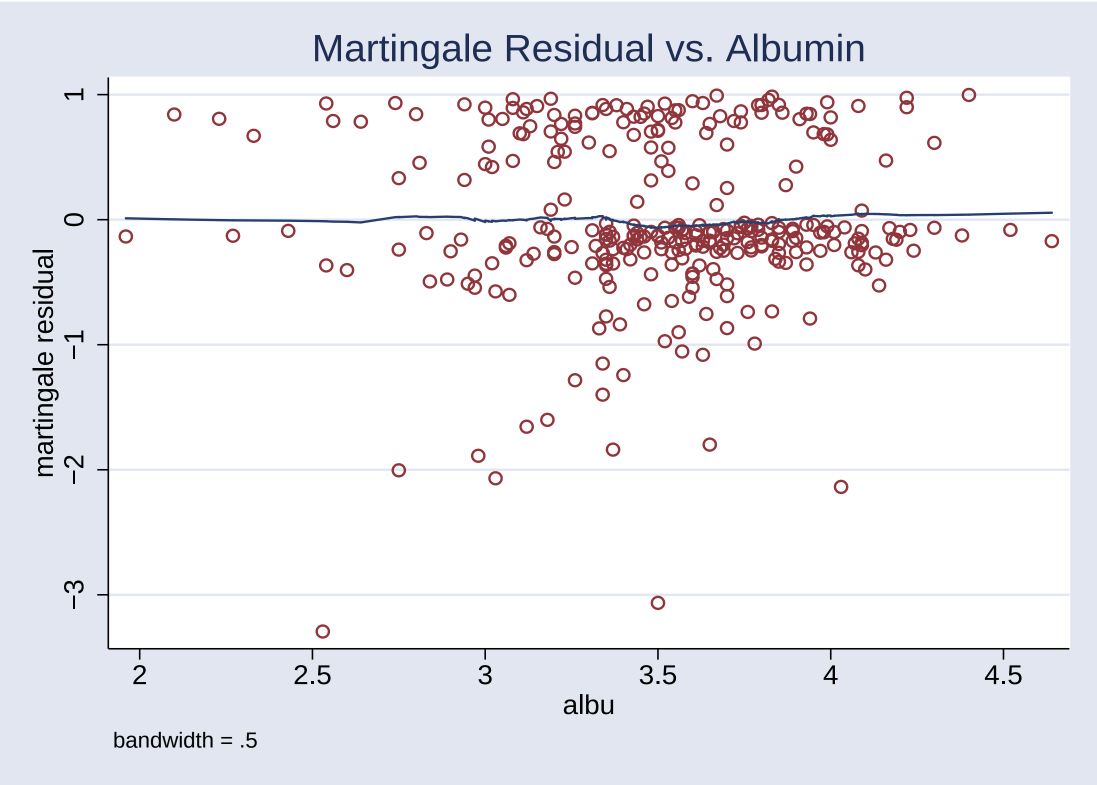
```
************************************************************
***
*** Cox regression -- using LINEAR versions of PREDICTORS
***
stcox bili albu age proth edema, nohr ///
      scaledsch(resid0*) esr(esr*) mgale(mres)


*** generate deviance residuals

predict dres, deviance

label variable mres "martingale residual"
label variable dres "deviance residual"
```

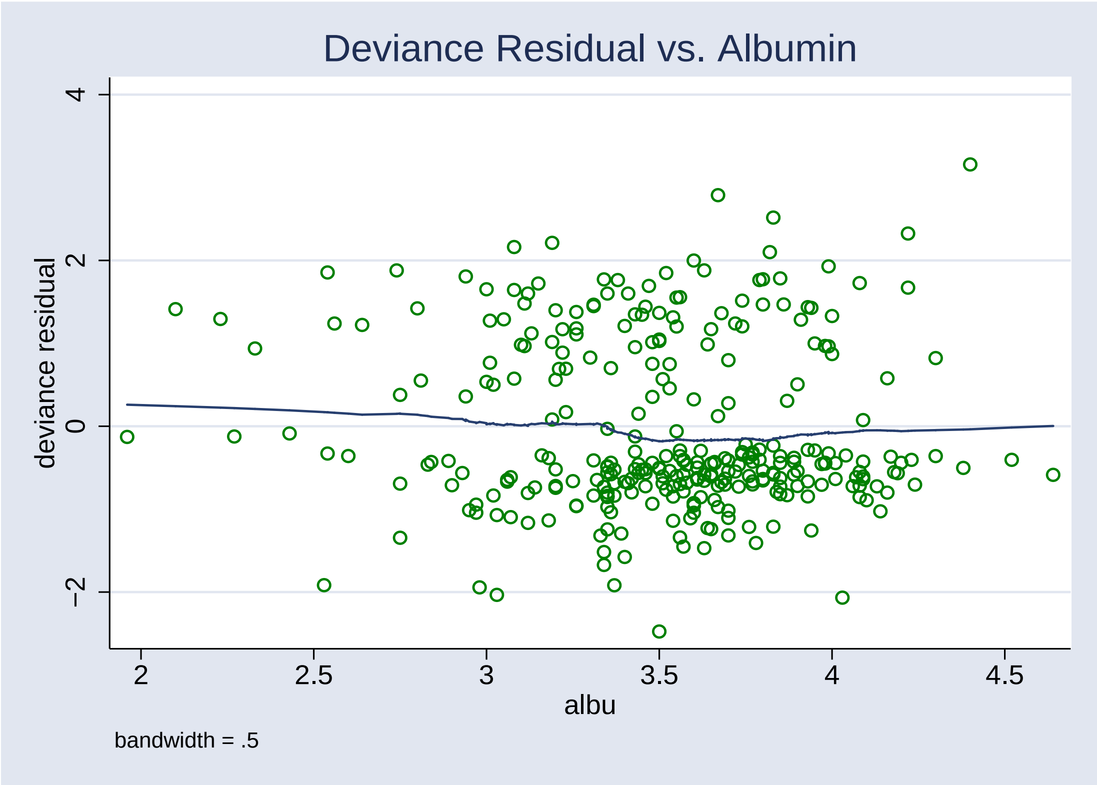
```
***
*** plot residuals vs predictor(s)
***

lowess mres bili, mean bwidth(5) ///
      title("Martingale Residual vs. Bilirubin") ///
      mcolor(maroon) msymbol(Oh)
graph export c:/COURSES/SURVIVAL/CoxRegn/bili-mres.ps, as(eps) replace

lowess dres bili, mean bwidth(5) ///
      title("Deviance Residual vs. Bilirubin") ///
      mcolor(green) msymbol(Oh)
graph export c:/COURSES/SURVIVAL/CoxRegn/bili-dres.ps, as(eps) replace
```

# Example: Mayo PBC Data

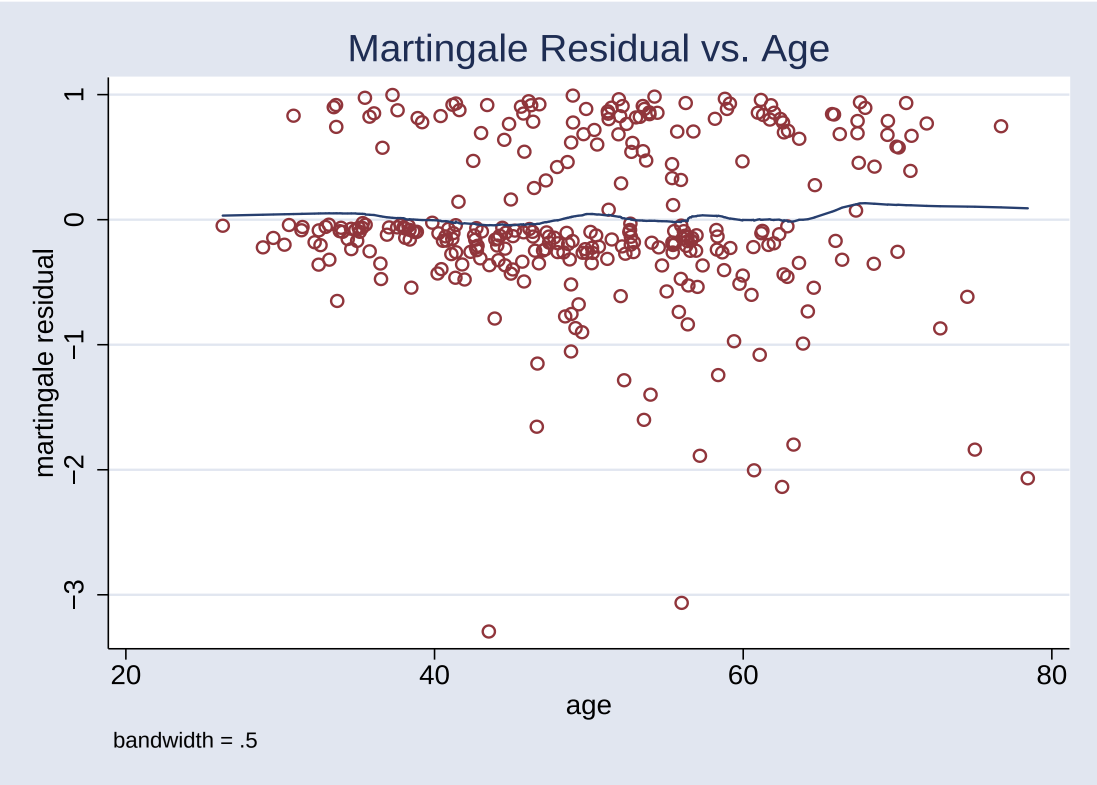


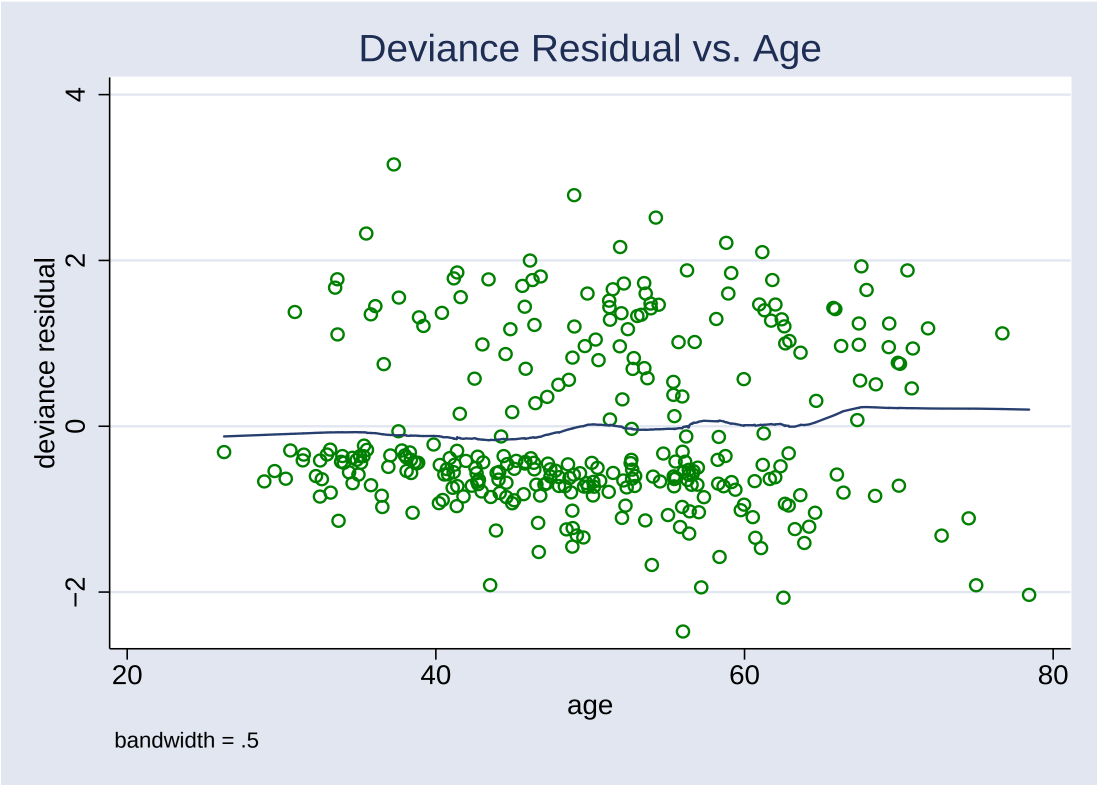
# Example: Mayo PBC Data



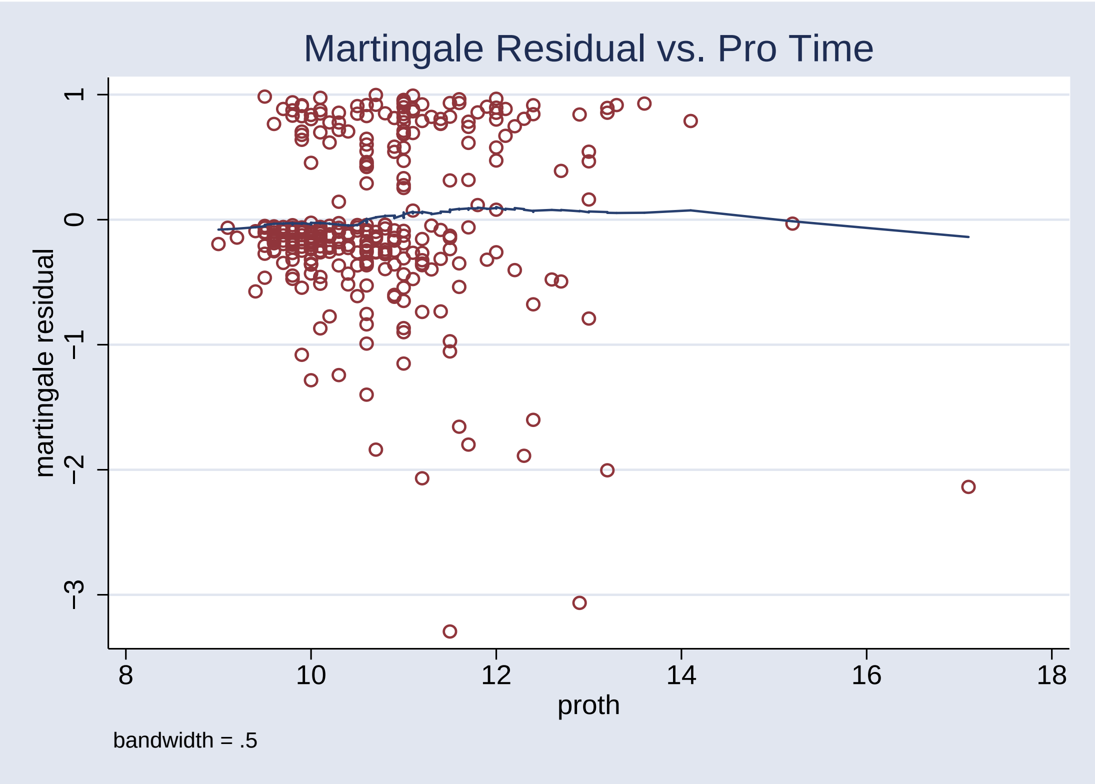



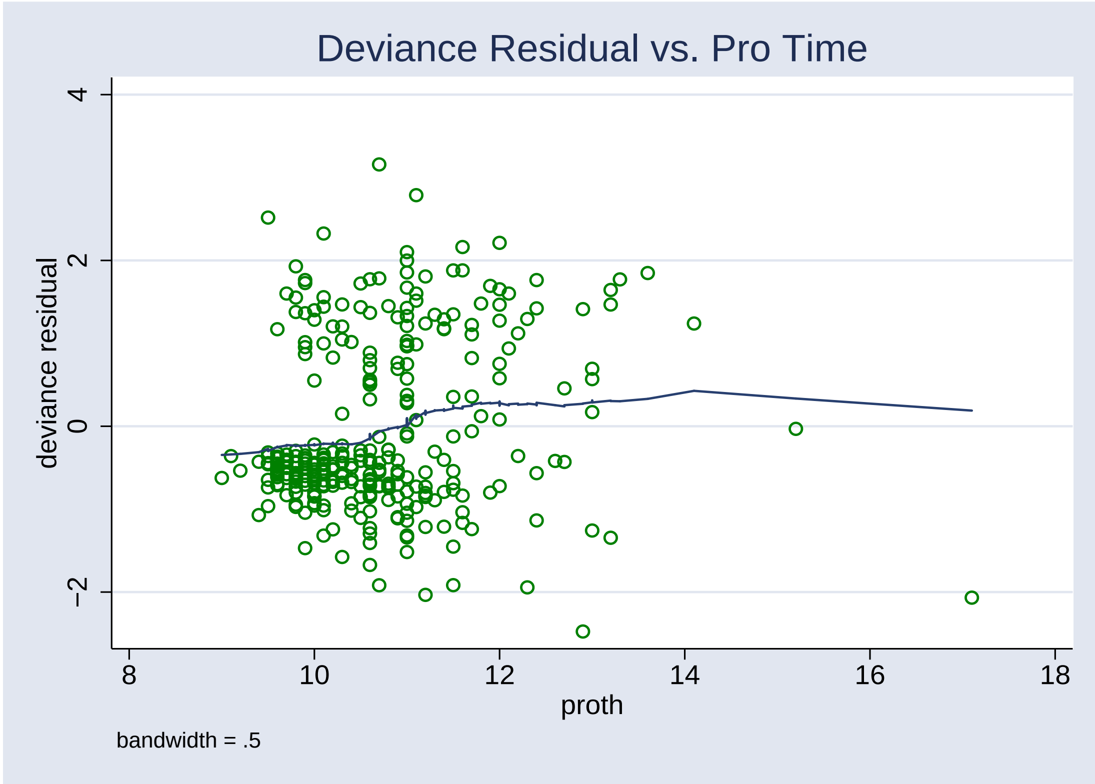
# Example: Mayo PBC Data



# Example: Mayo PBC Data



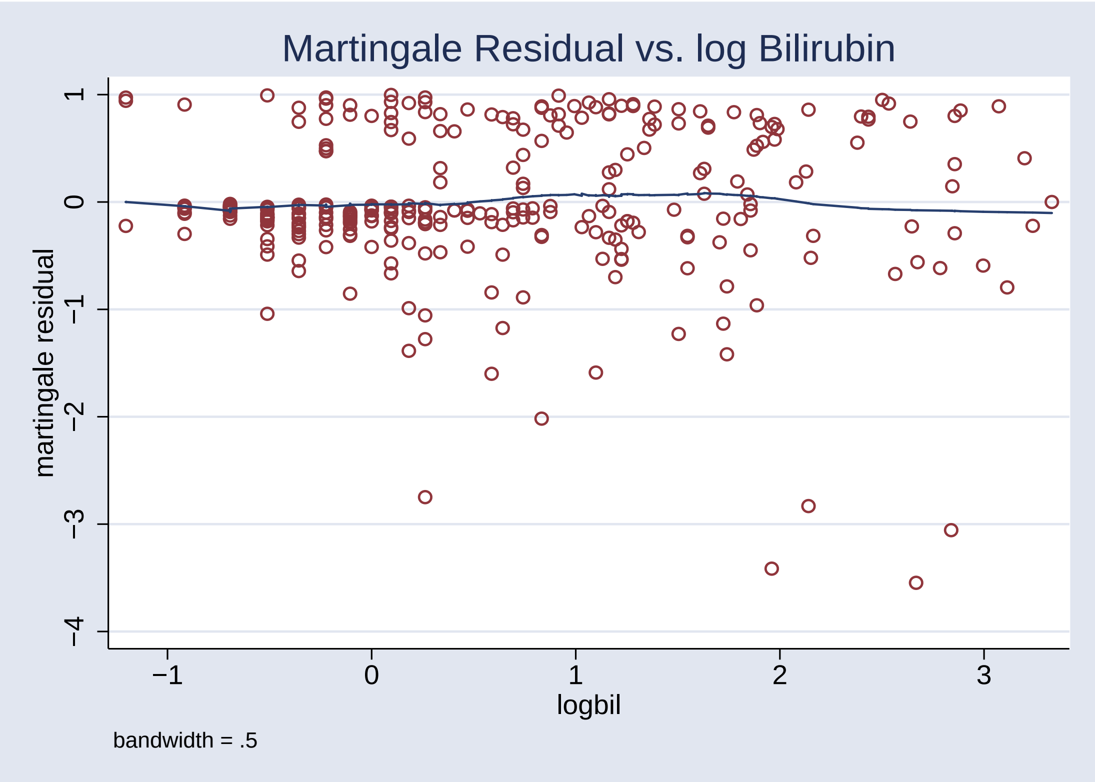
# Example: Mayo PBC Data

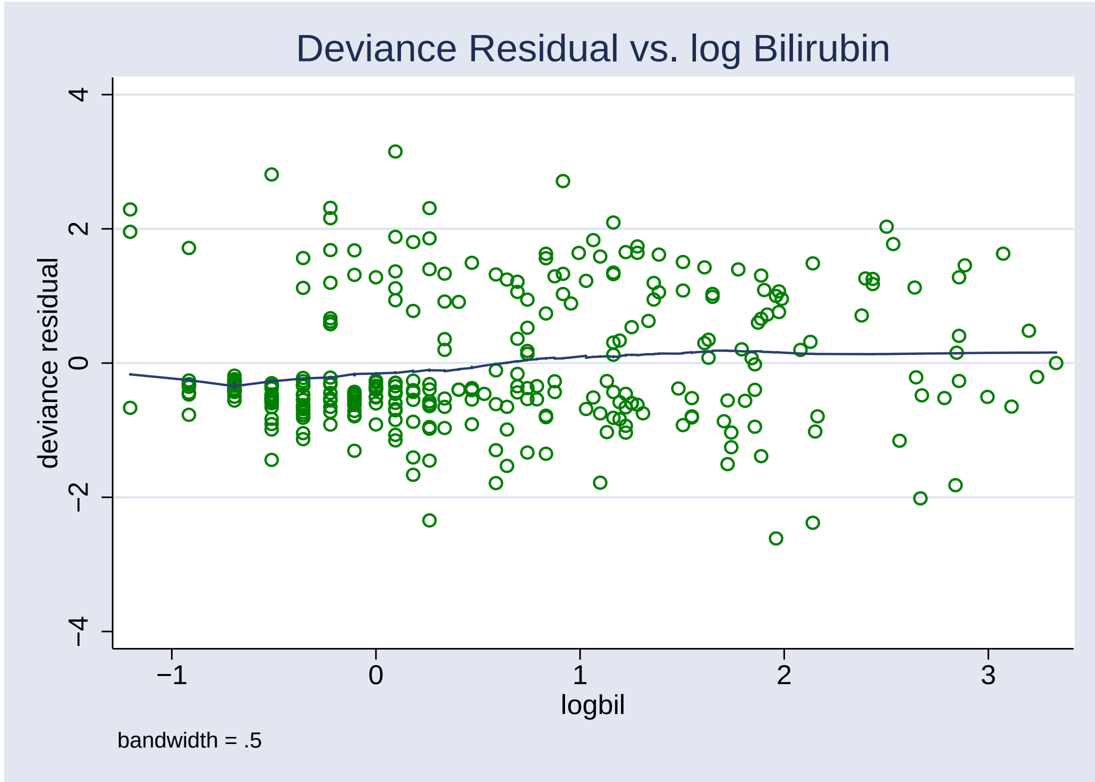
```
*****************************************************************
***
*** Cox regression -- using TRANSFORMED versions of PREDICTORS
***

drop mres dres resid0* esr*

stcox logbil logalb age logpro edema, nohr ///
      scaledsch(resid0*) esr(esr*) mgale(mres)

*** generate deviance residuals
predict dres, deviance
label variable mres "martingale residual"
label variable dres "deviance residual"
```

# Example: Mayo PBC Data

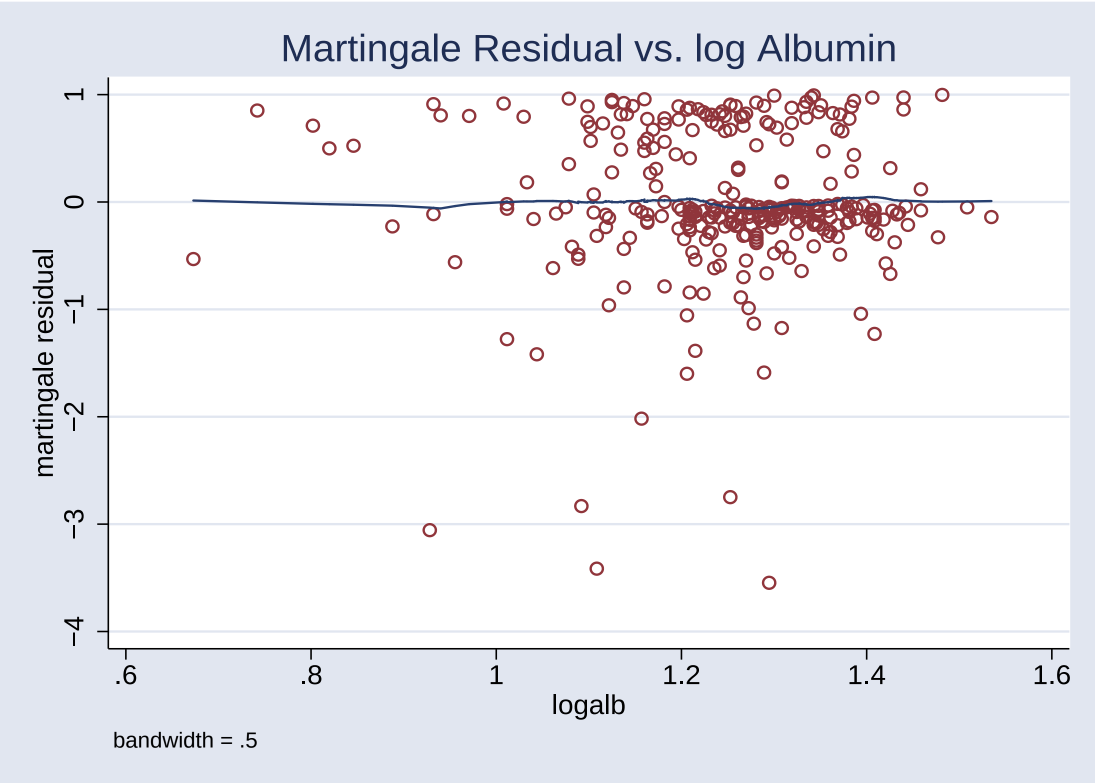
```
Cox regression -- Breslow method for ties
No. of subjects =          312      No. of failures =          125

Log likelihood   =   -541.70071
---------------------------------------------------------------------
      _t |   Coef.    Std. Err.     z     P>|z|     [95% Conf. Interval]
---------+-----------------------------------------------------------
  logbil |   0.9017   0.0982      9.18   0.000      0.7091    1.0944
  logalb |  -3.0973   0.7228     -4.28   0.000     -4.5142   -1.6805
     age |   0.0327   0.0085      3.82   0.000      0.0159    0.0495
  logpro |   3.1844   1.0072      3.16   0.002      1.2102    5.1587
   edema |   0.4839   0.2373      2.04   0.041      0.0187    0.9491
---------------------------------------------------------------------
```

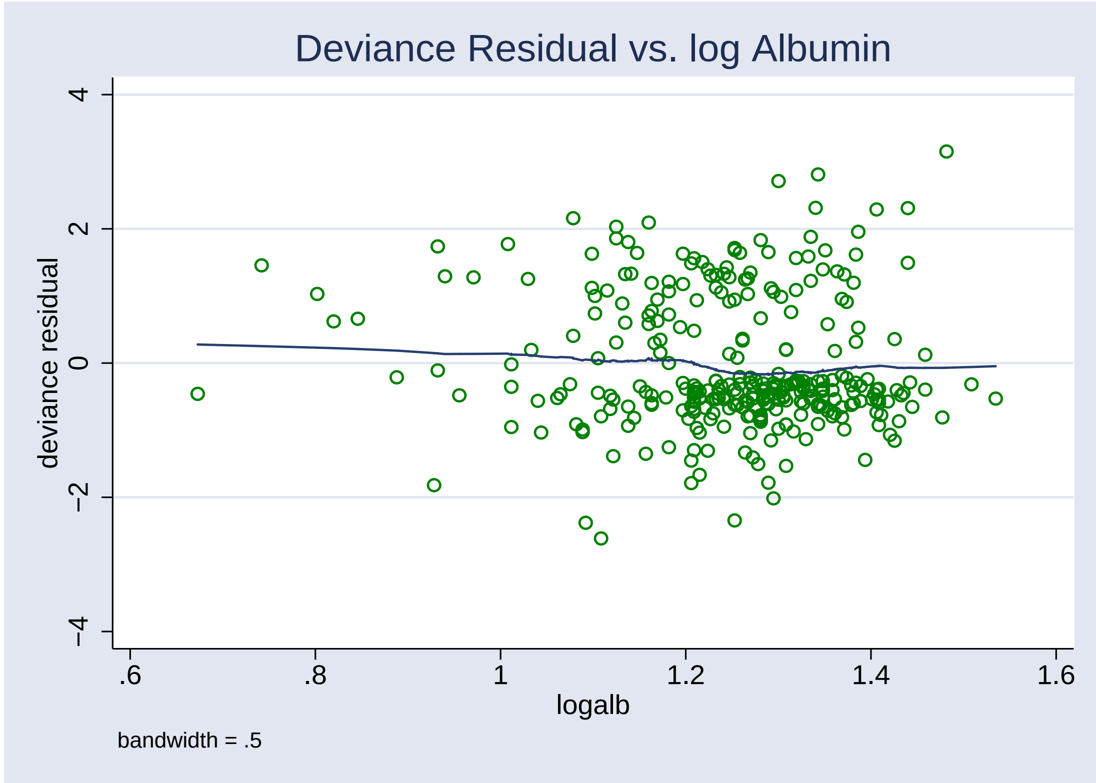
# Example: Mayo PBC Data



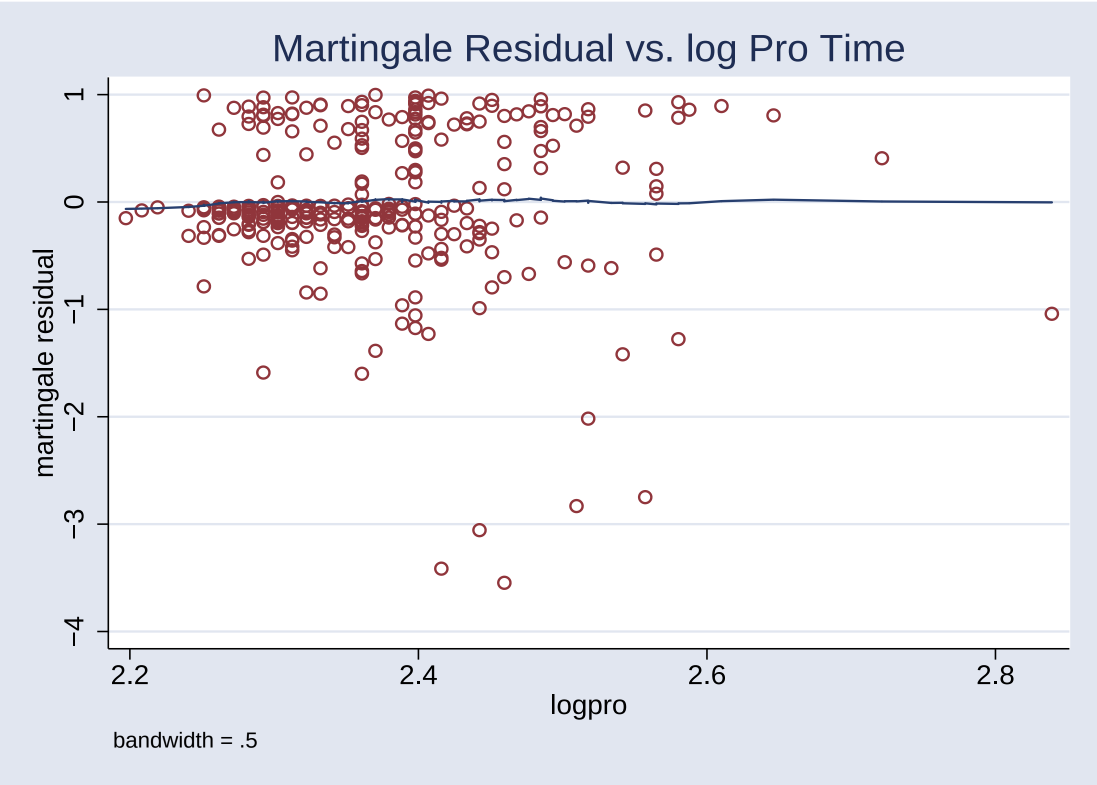



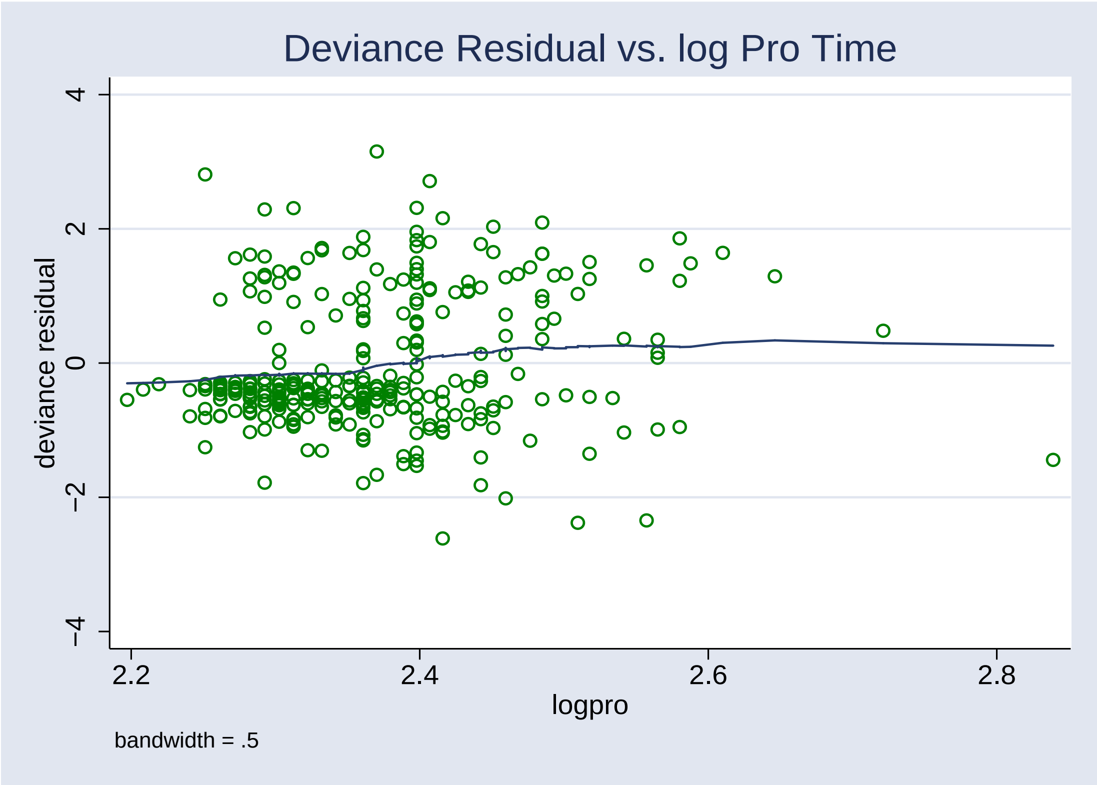
# Example: Mayo PBC Data



# Example: Mayo PBC Data



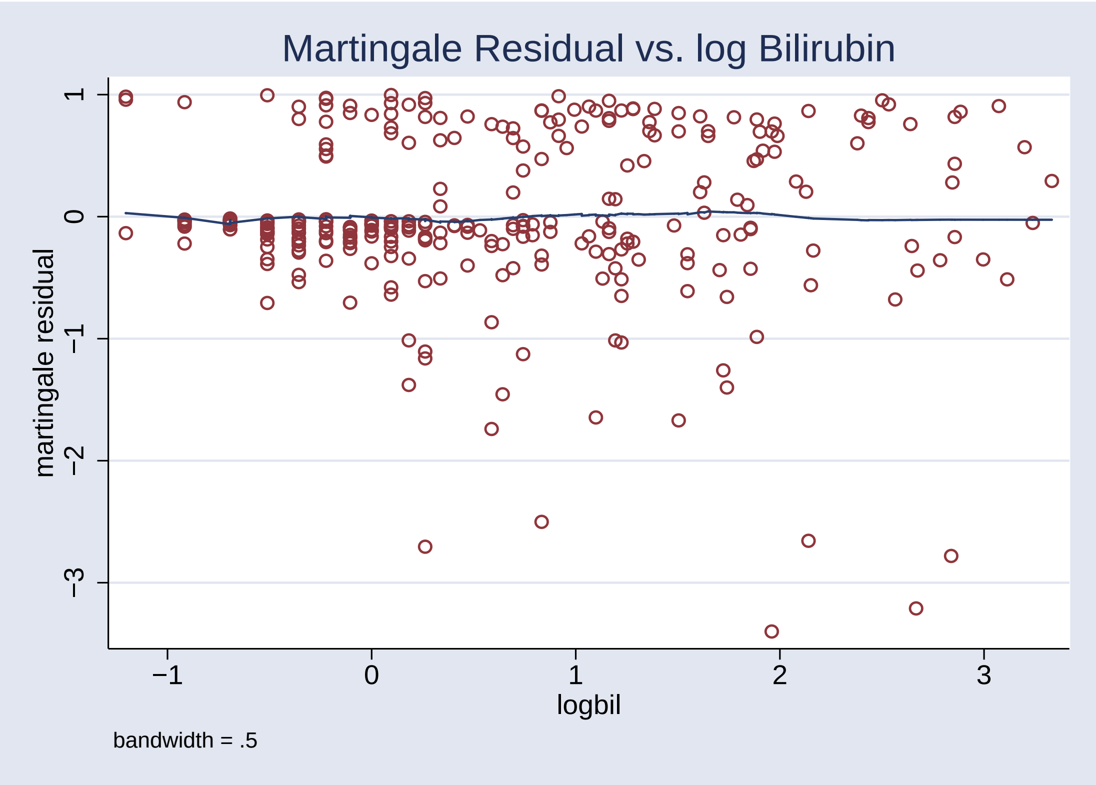
# Example: Mayo PBC Data

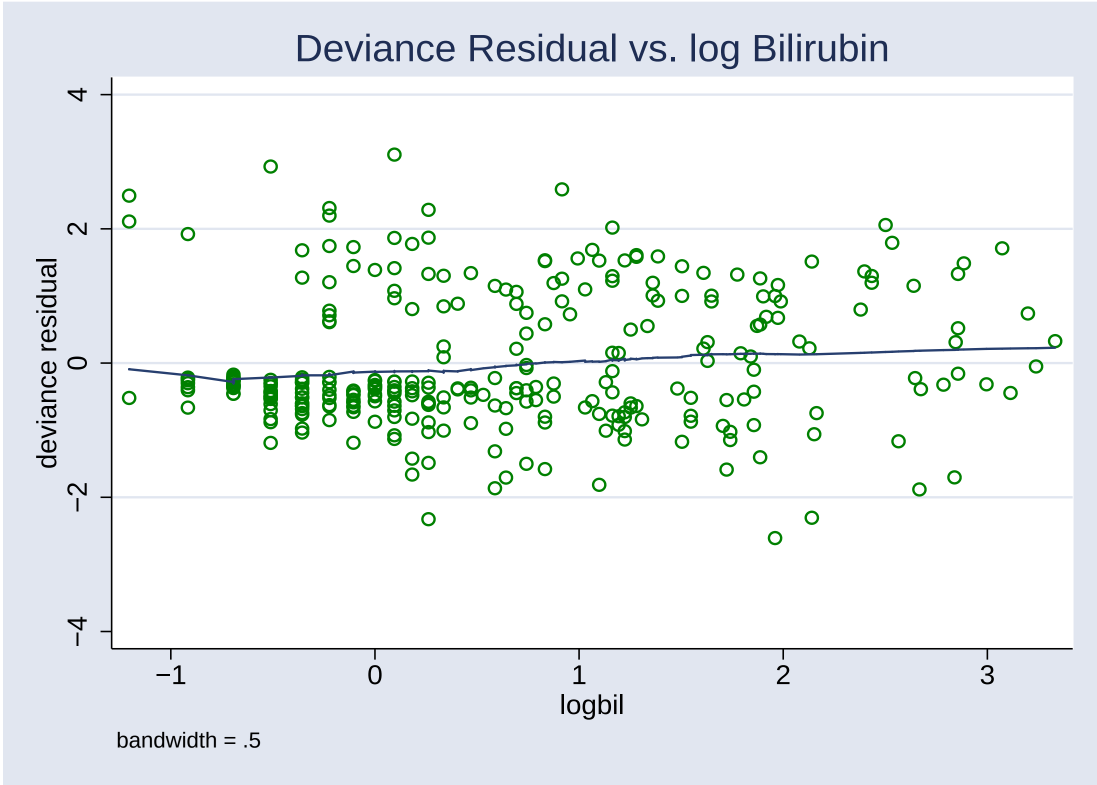
```
*****************************************************************
***
*** Cox regression -- using TRANSFORMED with splines
***
mkspline logbil1 0.8 logbil2 = logbil, marginal
mkspline logalb1 1.1 logalb2 = logalb, marginal
mkspline logpro1 2.4 logpro2 = logpro, marginal

stcox logbil1 logbil2 logalb1 logalb2 age logpro1 logpro2 edema, nohr ///
      scaledsch(resid0*) esr(esr*) mgale(mres)

*** generate deviance residuals
predict dres, deviance
label variable mres "martingale residual"
label variable dres "deviance residual"
```

# Example: Mayo PBC Data

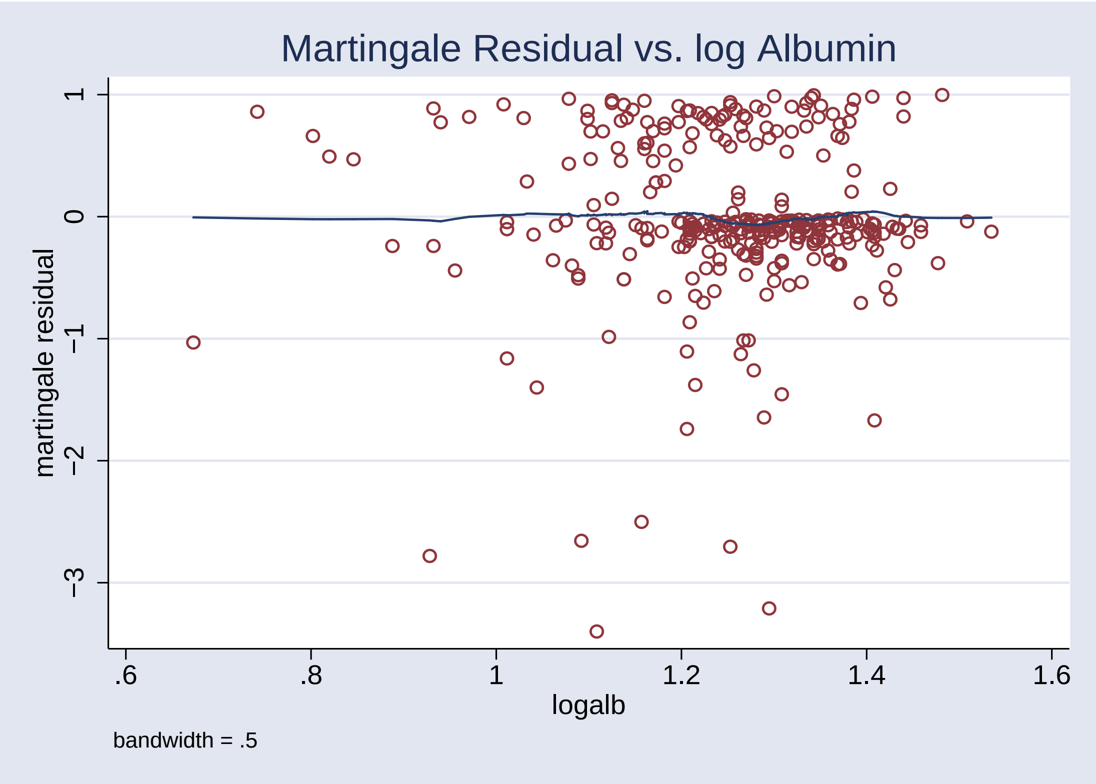
```
Cox regression -- Breslow method for ties
No. of subjects =          312     No. of failures =         125
Log likelihood  =   -540.34442
------------------------------------------------------------------
      _t |   Coef.    Std. Err.     z     P>|z|    [95% Conf. Interval]
---------+--------------------------------------------------------
 logbil1 |   1.2337   0.2662      4.63   0.000     0.7118    1.7556
 logbil2 |  -0.5111   0.3596     -1.42   0.155    -1.2159    0.1936
 logalb1 |  -3.5134   1.3852     -2.54   0.011    -6.2285   -0.7984
 logalb2 |   0.8529   2.0487      0.42   0.677    -3.1624    4.8684
     age |   0.0321   0.0085      3.78   0.000     0.0154    0.0488
 logpro1 |   4.9560   2.5182      1.97   0.049     0.0204    9.8916
 logpro2 |  -2.3762   3.4911     -0.68   0.496    -9.2186    4.4661
   edema |   0.4920   0.2369      2.08   0.038     0.0276    0.9565
------------------------------------------------------------------
```

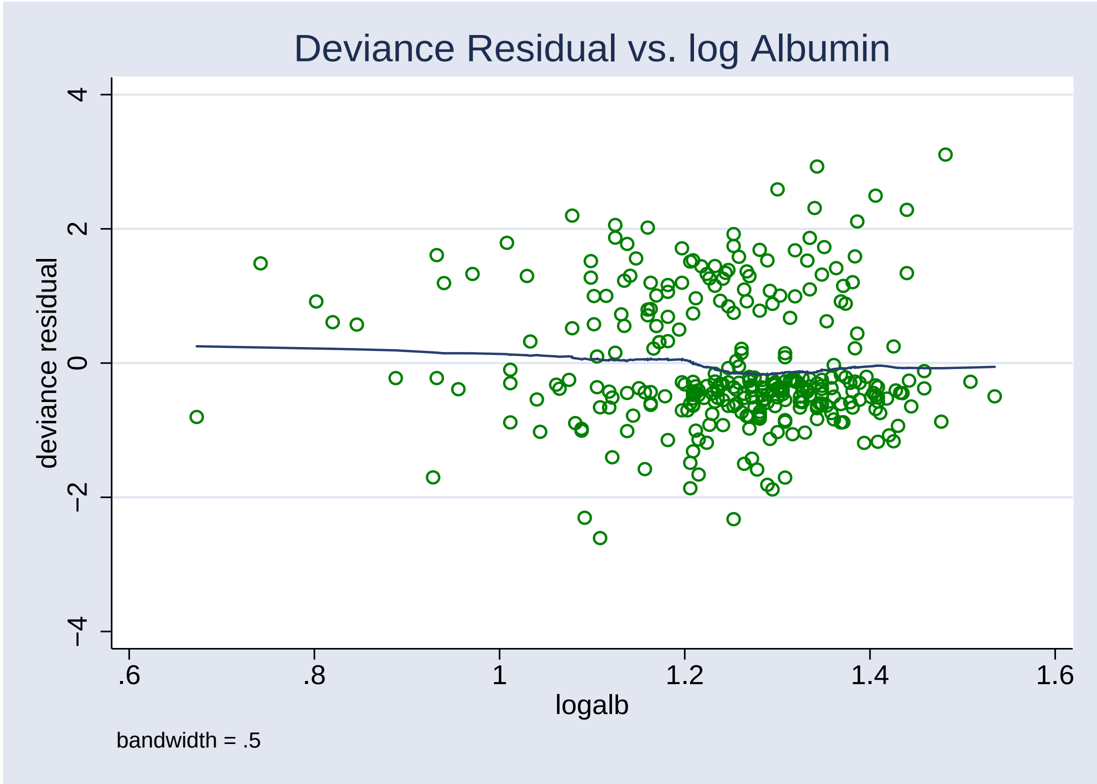
# Example: Mayo PBC Data



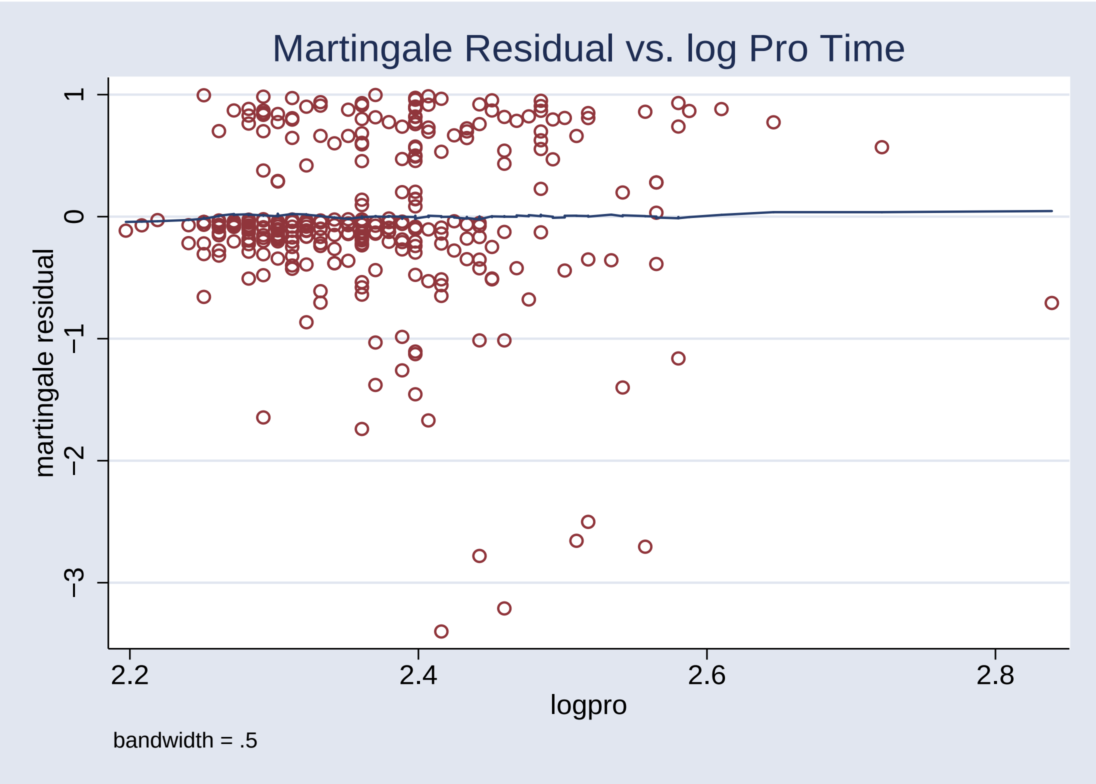



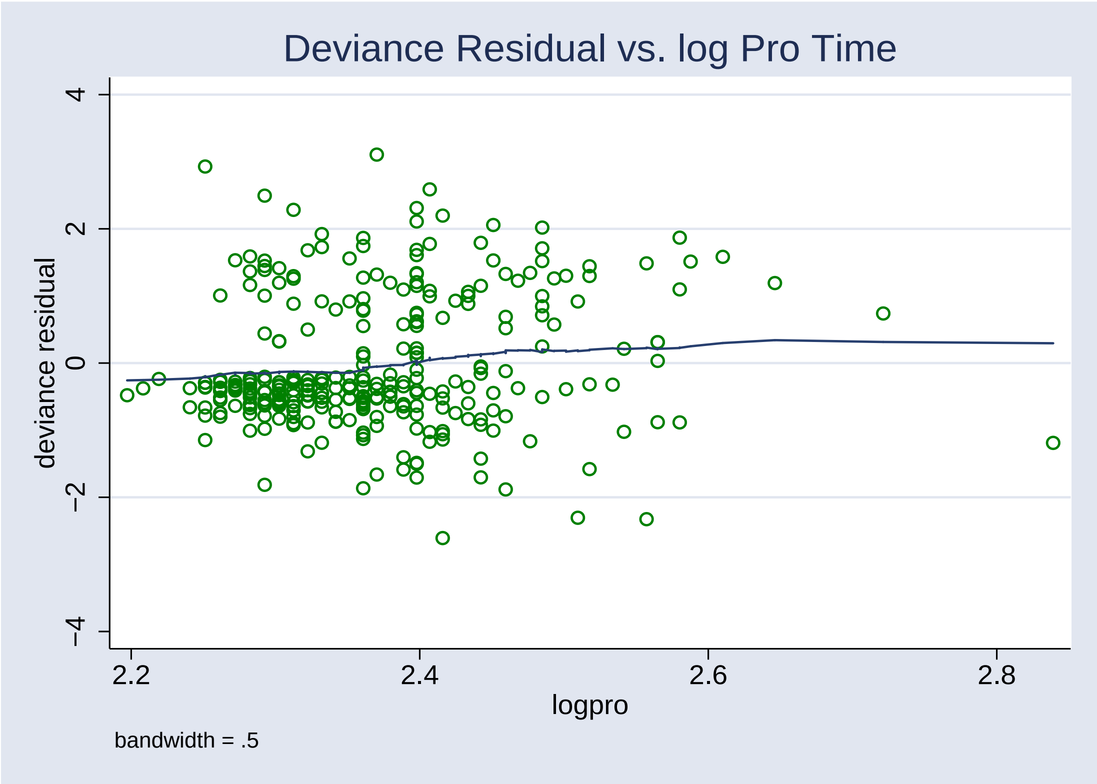
# Example: Mayo PBC Data



# Example: Mayo PBC Data



# Detecting Influential Observations

- $\boxed{\text{Delta-Betas}}$
- **Idea**: Just like with linear regression and logistic regression, we measure how coefficient estimates would change if each observation was individually omitted.
- Influence of the $i$th observation on $\widehat{\beta}$:

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

  ▷ $\widehat{\beta}_{(i)}$ = the estimated coefficient without subject $i$.
- Cox Regression: exact calculation of delta-beta would require refitting the model – so as many regression fits as subjects = computing time (but still possible).
- An alternative is to approximate $\Delta\beta_{(i)}$ with a "one-step" estimator that is a single iteration away from the entire data estimate, $\widehat{\beta}$.

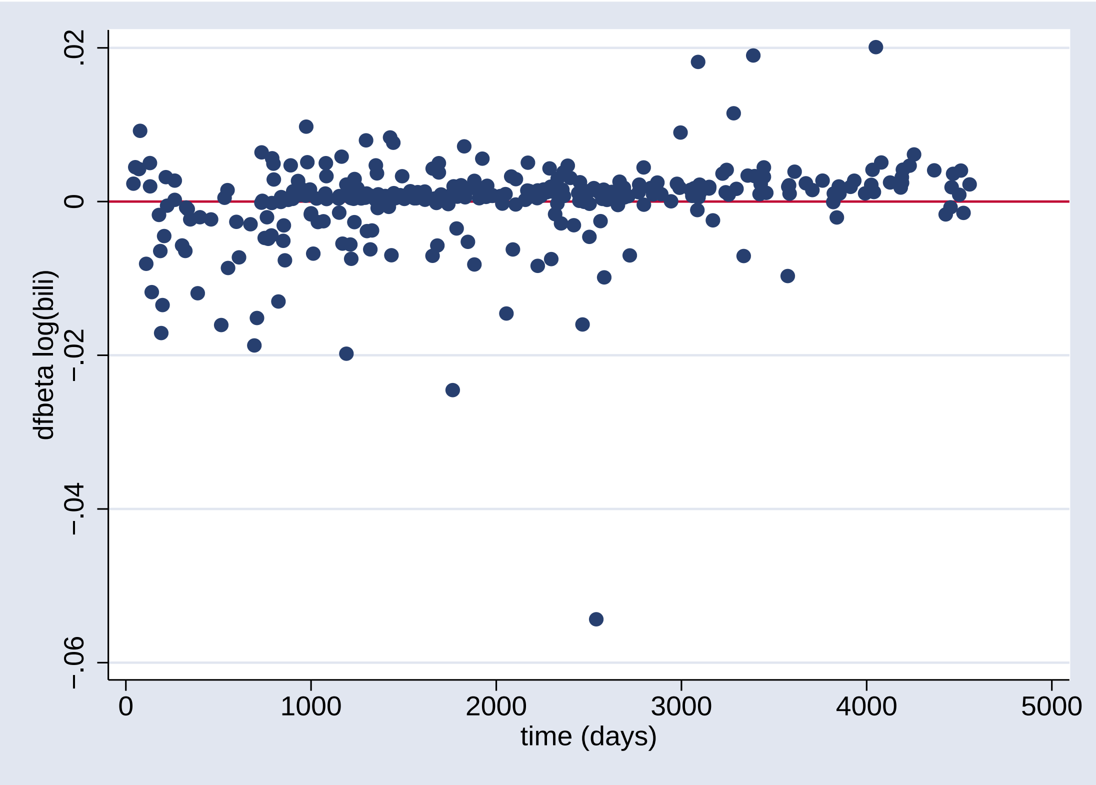
## Detecting Influential Observations

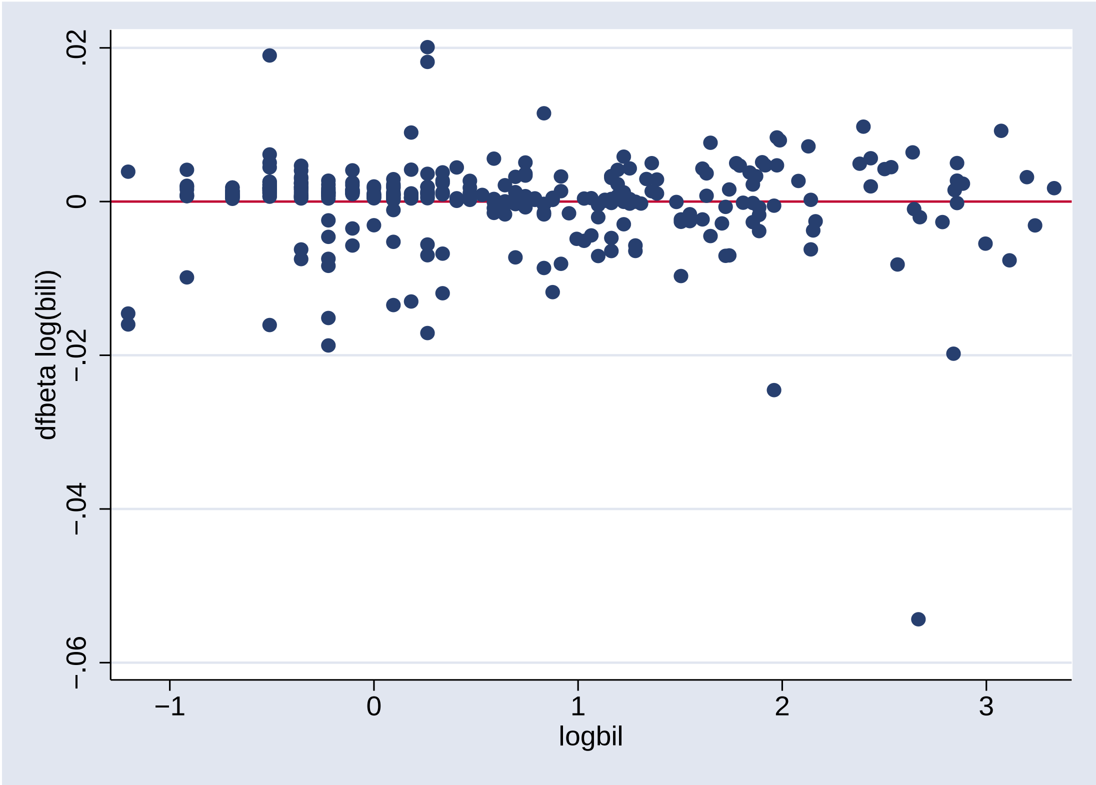
- By looking at the delta-betas we can see whether the value of the estimated relative hazard is influenced by one observation.
- We may also consider alternative impacts such as the impact on the test statistic $Z = \widehat{\beta}/\mathtt{s.e.}$, or on the p-value associated with this test.
- Impact of omitting observation $i$:
  - ▷ If a censored observation then dropping alters the contribution to every risk set in which it appears.
  - ▷ If an observed failure then dropping alters the contribution to every risk set in which it appears, and removes one risk set (if no ties).

# Example: Mayo PBC Data

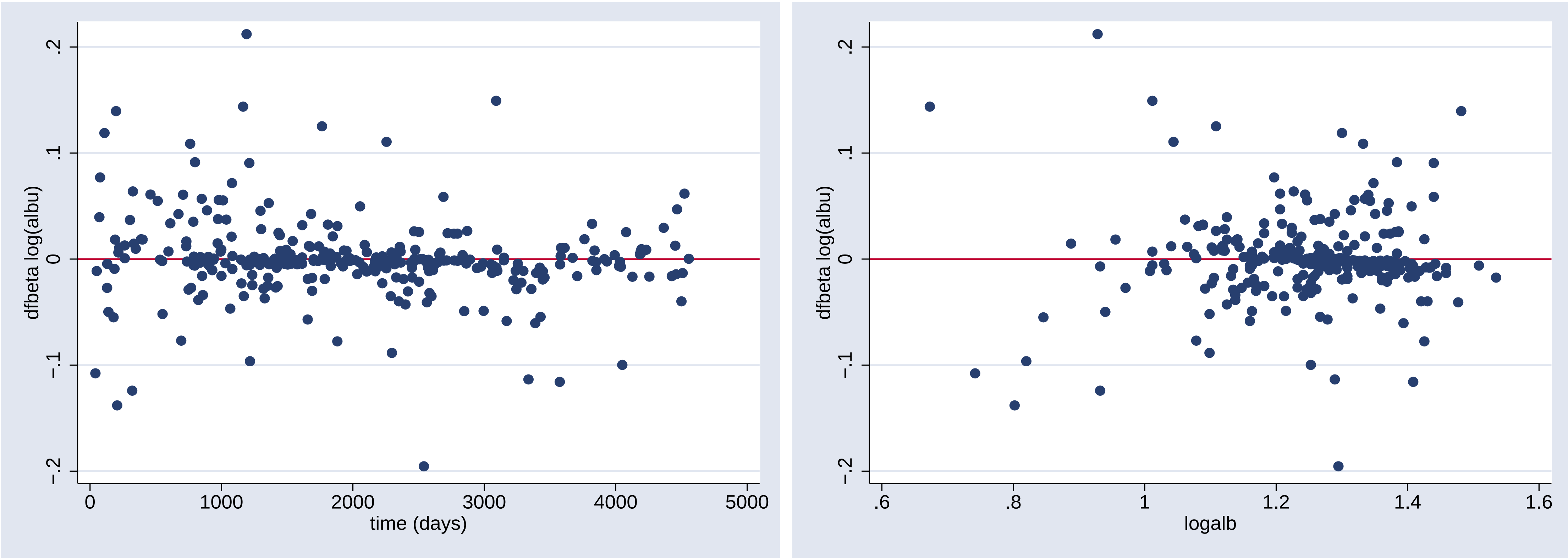
```
***
*** compute delta-betas
***
set matsize 400
mkmat esr1 esr2 esr3 esr4 esr5, matrix(esr)
mat V = e(V)
mat Inf = esr*V
svmat Inf, names(dfb)
label var dfb1 "dfbeta log(bili)"
label var dfb2 "dfbeta log(albu)"
label var dfb3 "dfbeta Age"
label var dfb4 "dfbeta log(prot)"
label var dfb5 "dfbeta Edema"
```

# Example: Mayo PBC Data

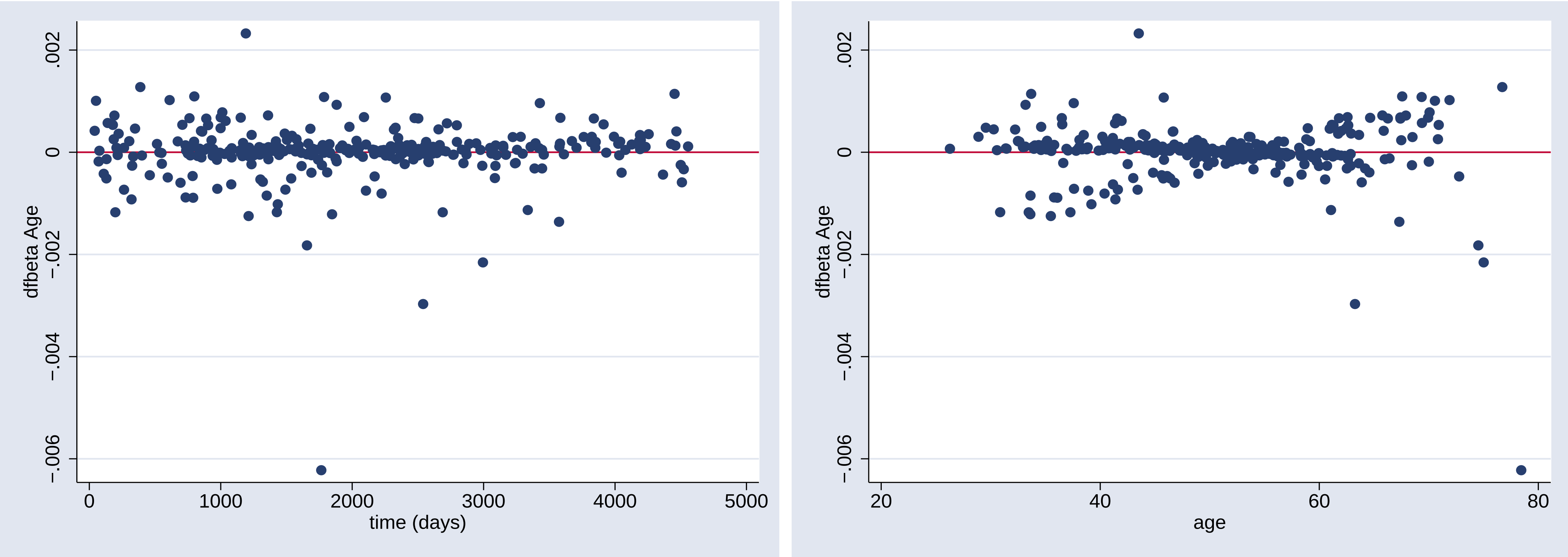


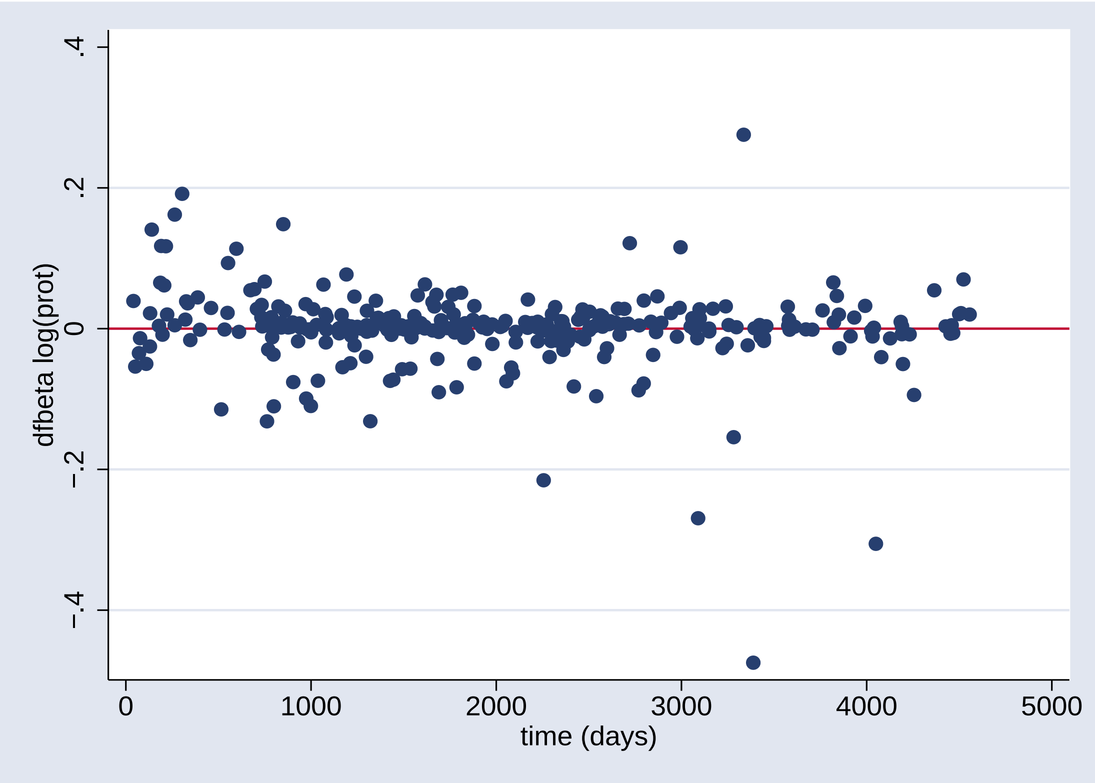
# Example: Mayo PBC Data

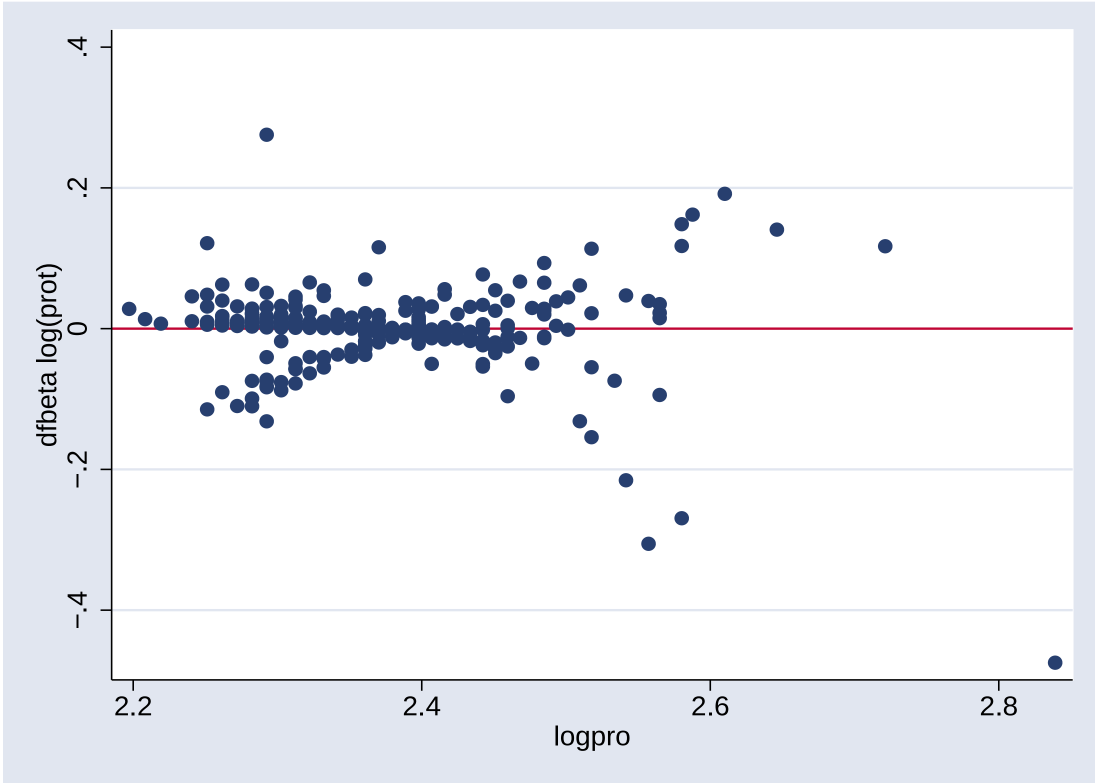


# Example: Mayo PBC Data

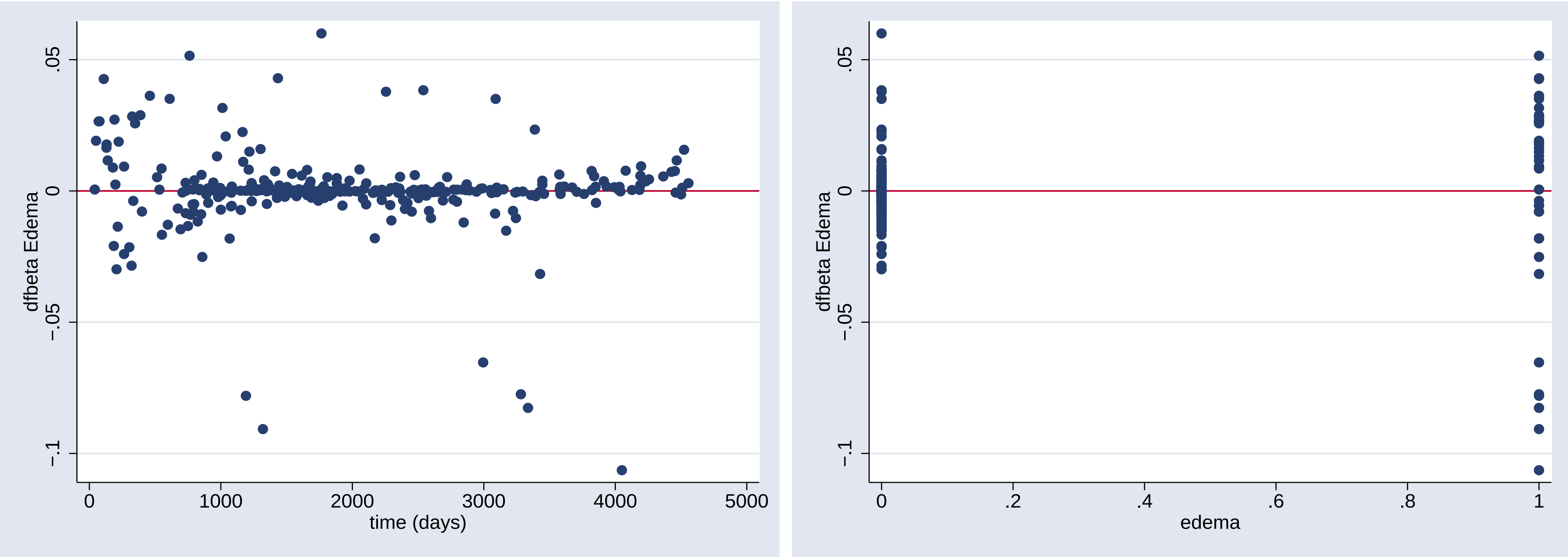


# Example: Mayo PBC Data



# Example: Mayo PBC Data



# Summary

- Formulate an analysis plan.
- Model checking:
  - ▷ PH assumption
  - ▷ Functional form (martingale residuals)
  - ▷ Influential observations (delta-betas)
- **Q**: How to choose a good model for prediction?