Quantifying predictive accuracy in Cox models

Patrick Breheny

November 21

Introduction

- Today's lecture will address the question: Overall, how well can a given model predict survival?
- To illustrate, we'll look at three models for the PBC data:
  - Model 1: `trt` + `albumin`
  - Model 2: `trt` + `stage` + `hepato` + `f(albumin)` + `log(bili)`
  - Model 3: Model 2 + 30 variables of random noise

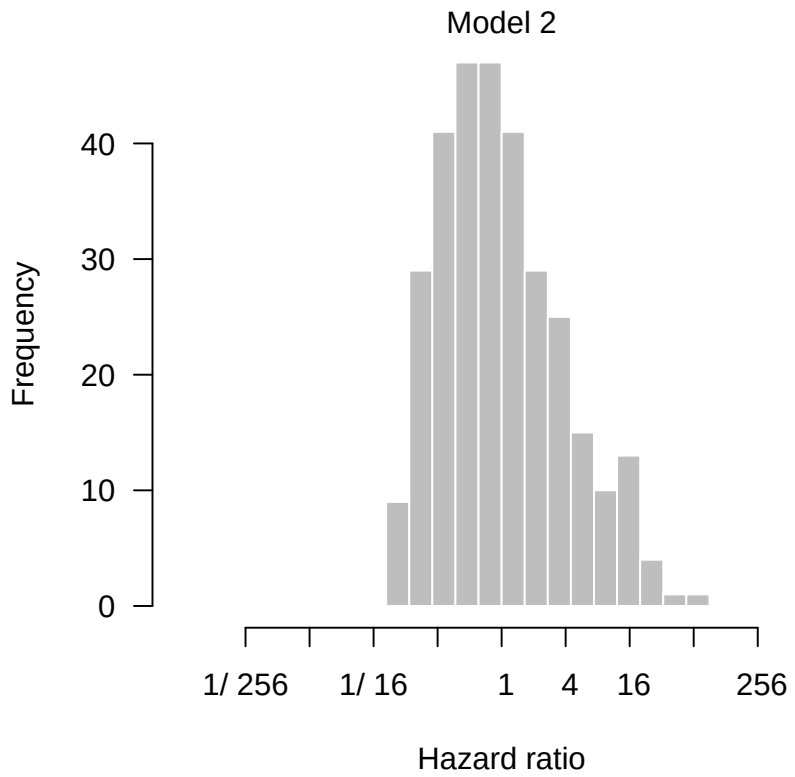
  where `f()` denotes the changepoint function we described in a previous lecture
- The idea here is to see how various metrics compare when applied to a model with OK predictive ability (model 1), a model with good predictive ability (model 2), and a model in which overfitting is present (model 3)

Linear predictors

- One simple approach to describing the amount of signal present in a model is to describe the linear predictors
- Hazard ratios are direct functions of the linear predictors, so by inspecting the distribution of linear predictors, we get a sense of the extent to which our model can identify individuals as high risk and low risk, as opposed to saying that everyone has about the same risk
- For our three models:
  - Model 1: $\mathrm{SD}(\hat{\eta}) = 0.70$
  - Model 2: $\mathrm{SD}(\hat{\eta}) = 1.31$
  - Model 3: $\mathrm{SD}(\hat{\eta}) = 1.75$

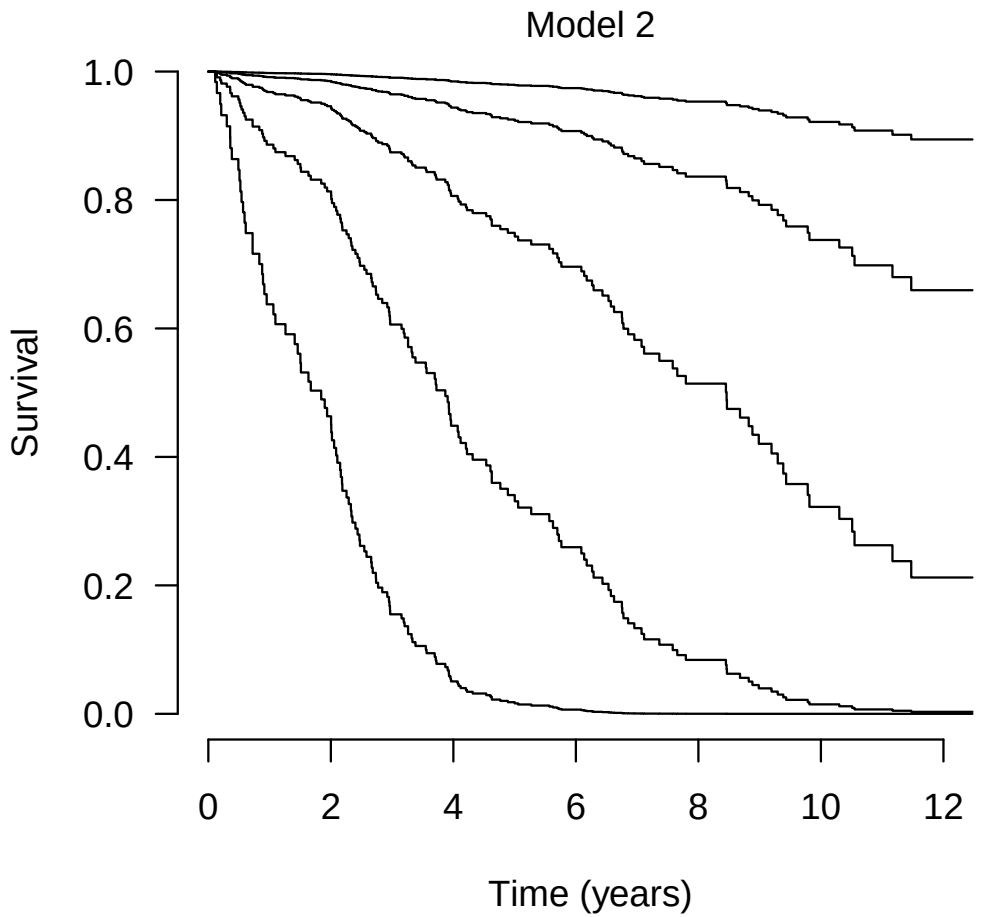
Histograms

Plotting the distribution makes the same point, but also illustrates the distribution of values:



Survival plots

A related idea is to plot the baseline hazard $\pm$ 1 and 2 SDs:



Introduction: R^2

- It is typically desirable to be able to summarize these illustrations into a single number that quantifies a model's accuracy
- For example, in linear regression we have R^2, the proportion of variance in the outcome explained by the model
- There are several ways to construct an R^2-like measure for Cox regression; the motivations typically proceed by analogy

Derivation

- One widely used measure, the *Cox-Snell* R^2, is based on the change in deviance (i.e., the likelihood ratio test statistic):

$$\Delta D = 2(\ell_1 - \ell_0),$$

  where ℓ_1 is the log-likelihood of the fitted model and ℓ_0 is the log-likelihood for the null model
- For linear regression, we have

$$R^2 = 1 - \frac{\mathrm{RSS}_1}{\mathrm{RSS}_0},$$

  where RSS_1 and RSS_0 are the residual sums of squares for the fitted and null models
- For linear regression, we also have

$$\Delta D = n \log \frac{\mathrm{RSS}_0}{\mathrm{RSS}_1}$$

Likelihood ratio R^2

- This suggests

$$R^2 = 1 - \exp(-\Delta D/n)$$

  as a way of calculating an R^2 for Cox models; traditionally the number of observations n is used in the denominator, but the number of events d is probably better
- For our three models:
  - Model 1: $R^2 = 0.18$
  - Model 2: $R^2 = 0.45$
  - Model 3: $R^2 = 0.55$
- This has essentially the same interpretation as R^2 in linear regression, although the analogy isn't perfect
- R^2 is reported by `summary(fit)` in the `survival` package

Concordance: Introduction

- An alternative idea is to quantify a model's accuracy on the basis of concordance
- The idea here is to consider all possible pairs of observations and sort them into concordant and discordant groups based on their outcomes and the model's predictions

Concordant pairs

- For example, suppose we observe a pair with $t_i = 100, d_i = 1, \eta_i = 1$ and $t_j = 150, d_j = 1, \eta_j = 0$
- This is a *concordant* pair, in that the model predicts that subject i will die first, and this coincides with what actually happened
- Note that we can still have concordant pairs in the presence of censoring: $t_i = 100, d_i = 1, \eta_i = 1$ and $t_j = 150, d_j = 0, \eta_j = 0$ also form a concordant pair

Discordant and indeterminate pairs

- Conversely, $t_i = 100, d_i = 1, \eta_i = 0$ and $t_j = 150, d_j = 1, \eta_j = 1$ would be a *discordant pair*: we predict that subject j is higher risk, but they in fact survive longer than subject i
- Not all pairs can be classified as concordant or discordant, however; in the presence of censoring, pairs can also be indeterminate
- For example, suppose we observe $t_i = 100, d_i = 0, \eta_i = 1$ and $t_j = 150, d_j = 1, \eta_j = 0$
- We predict that subject i dies first, but we have no way of knowing whether that actually happened

Scoring

- Finally, we can also have tied pairs, either because the predictors are tied ($\eta_i = \eta_j$) or because the failure times are tied ($t_i = t_j$, with $d_i = d_j = 1$)
- In aggregating the results, the model scores one point for every concordant pair and half a point for every tied pair
- This score is then divided by the total number of non-indeterminate pairs to obtain a *concordance index*
- As a formula,

$$C = \frac{n_c + 0.5 n_t}{n_c + n_d + n_t},$$

where n_c is the number of concordant pairs, n_d is the number of discordant pairs, and n_t is the number of tied pairs

Example: Model 2

- For example, in the pbc data, there are 312 observations, so $\binom{312}{2} = 48,516$ pairs
- For model 2, those pairs fall into the following categories:
  - 23,653 were concordant
  - 5,061 were discordant
  - 17 were tied
  - 19,785 were indeterminate
- This gives $C = 0.82$
- In the pbc data, 14% of the observations are censored, resulting in 41% of the pairs being indeterminate; to contrast, in the VA lung data, only 7% of the observations are censored, and only 5% of the pairs are indeterminate

Concordance results

- By construction, C must be between 0 and 1, with 1 representing perfect agreement between model and observation and 0.5 representing random guesses
- In survival data, $C = 0.6 - 0.8$ is pretty common
- For our three models,
  - Model 1: $C = 0.69$
  - Model 2: $C = 0.82$
  - Model 3: $C = 0.85$
- C is reported by `summary(fit)` along with R^2; you can also obtain a more detailed report from `survConcordance`

Measuring prediction error

- The preceding measures quantify accuracy in some sense, but don't directly quantify prediction error
- So, for some fixed time t_0, let us consider $\hat{S}(t_0|\mathbf{x})$, the model-based probabilistic prediction that an individual will survive past time t_0, along with y, the actual observation of whether this happened (ignoring censoring for the moment)
- Two common ways of quantifying the prediction error are:

$$\text{Brier}(y, \hat{S}(t_0|\mathbf{x})) = \{y - \hat{S}(t_0|\mathbf{x})\}^2$$

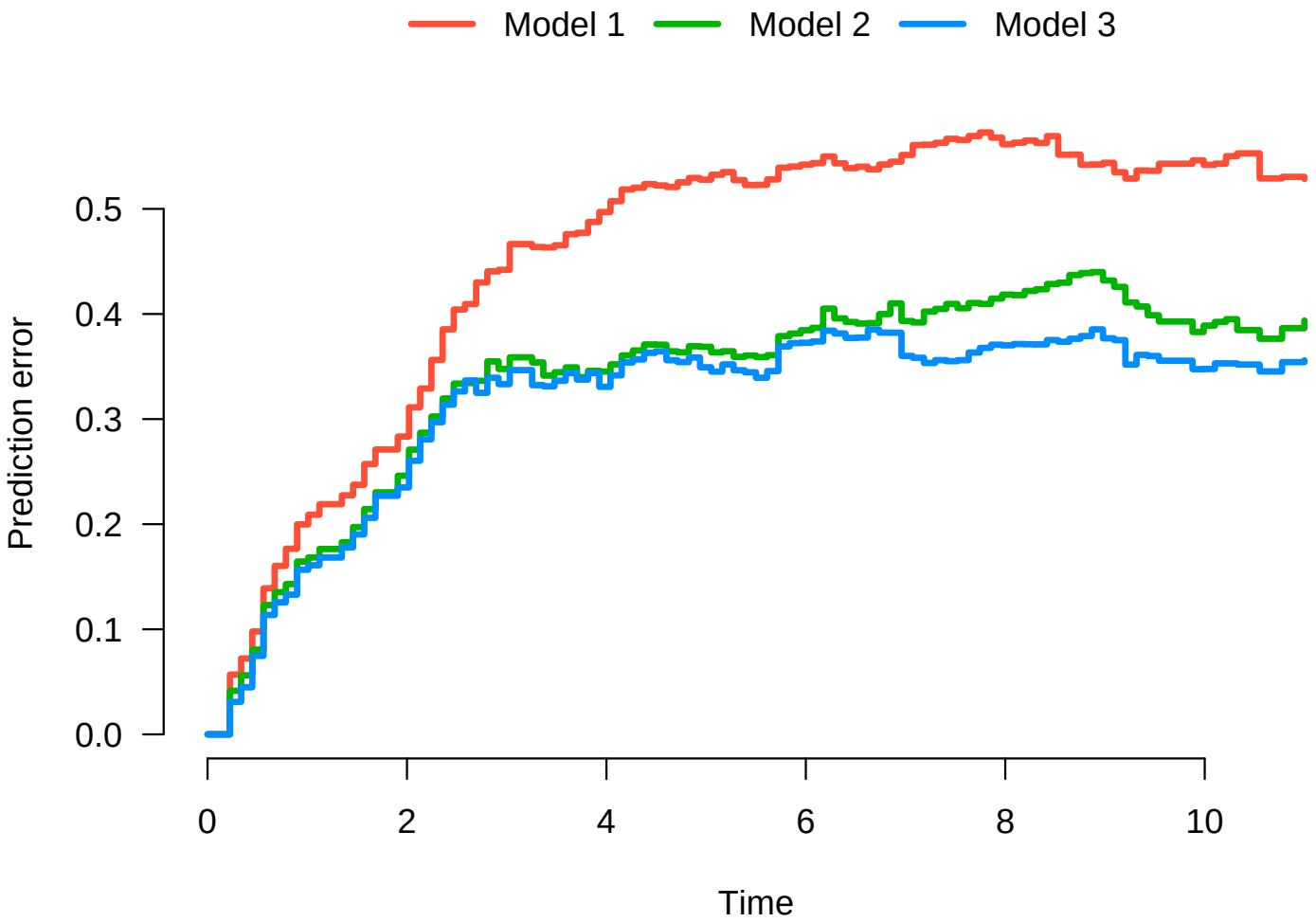
$$\text{KL}(y, \hat{S}(t_0|\mathbf{x})) = -\{y \log \hat{S}(t_0|\mathbf{x}) + (1-y) \log(1 - \hat{S}(t_0|\mathbf{x}))\}$$

- In theory, the Kullback-Liebler score is optimal; in practice, the two are usually pretty similar

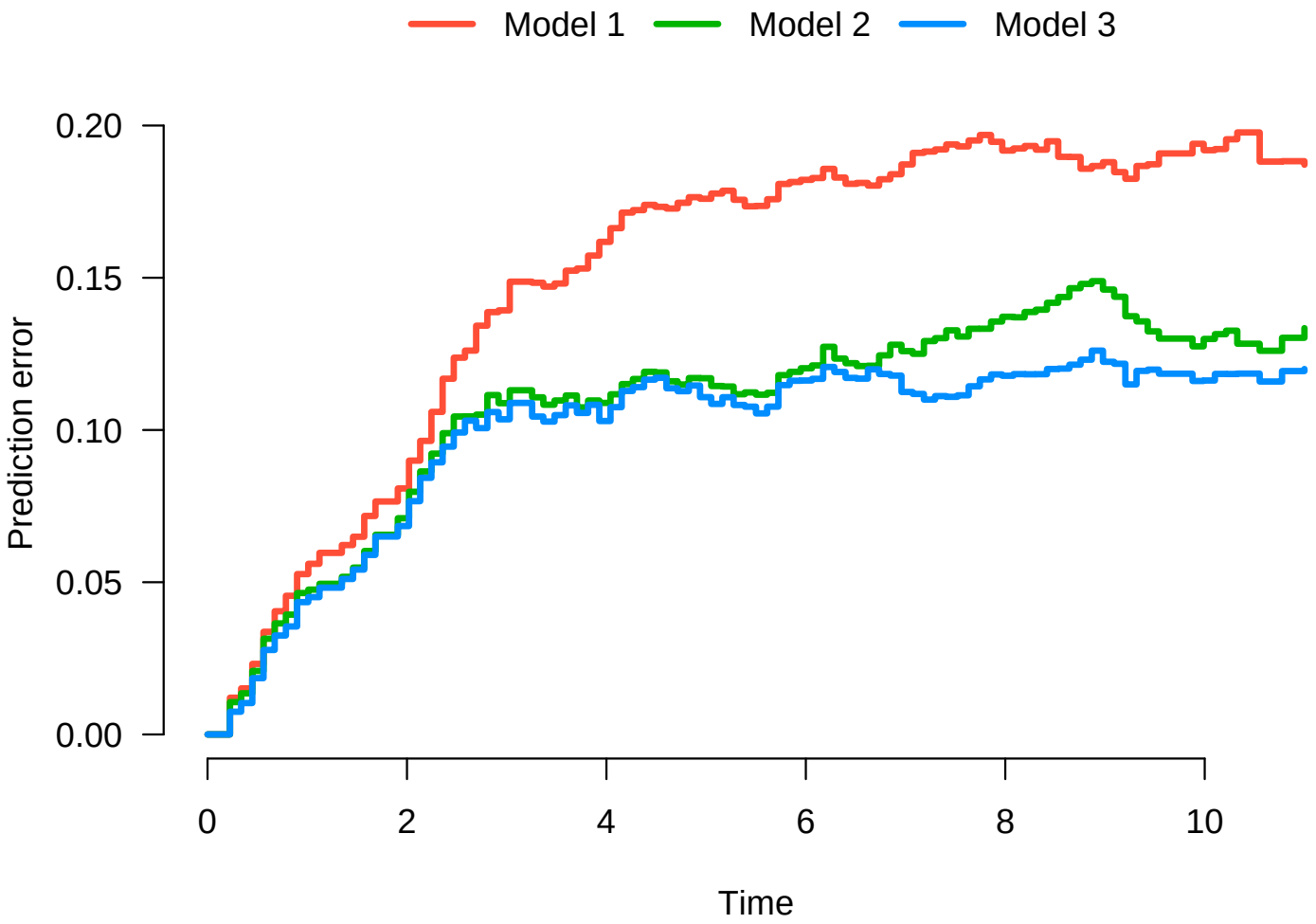
Accounting for censoring

- In the presence of censoring, we will not always know y; this is especially true for large values of t_0
- A common way to deal with this is to fit a Kaplan-Meier curve to the data using *censoring* as the event of interest to obtain $\hat{C}(t_0)$, then up-weight the uncensored observations by $1/\hat{C}(t_0)$
- This scheme, known as inverse probability of censoring weighting (IPCW), produces an unbiased estimate of the true prediction error that we would obtain in the absence of censoring
- In principle, the estimates of $\hat{C}(t_0)$ could depend on covariates as well, although in practice, people often don't spend much time modeling censoring

Kullback-Liebler loss



Brier loss


Model 1
Model 2
Model 3
0.20
0.15
0.10
0.05
0.00
Prediction error
0
2
4
6
8
10
Time

Overfitting

- You have probably noticed that for all of these measures, model 2 is more accurate than model 1 (this is likely genuine) and model 3 is more accurate than model 2 (this is not genuine, as model 3 is just model 2 plus junk)
- This is because none of the methods we have discussed so far address overfitting in any way
- All of these measures describe how well the model agrees with the already observed outcomes; this is not really what we want to know
- What we really want to know is how accurate the model is at *predicting* future observations

Optimism

- Measures of accuracy are almost always better for already observed outcomes than they are for future predictions, because the observed outcomes were used to build the model in the first place
- To be more precise, let M denote a generic measure of accuracy, $\mathbf{y}$ denote the observed outcomes (for survival, this includes t and d), $\mathbf{y}^*$ denote future outcomes, and $f(\mathbf{X})$ denote a model's predictions
- Because of this phenomenon of overfitting, the quantity

$$M\{f(\mathbf{X}), \mathbf{y}\} - M\{f(\mathbf{X}), \mathbf{y}^*\}$$

is almost always positive; this quantity is known as the *optimism* of the model, and it tends to be more severe for complex models than simple models

Shrinkage

- Unfortunately, methods for estimating optimism are underdeveloped in survival analysis, at least with respect to other regression models
- However, one useful approach is the shrinkage heuristic developed by van Houwelingen and le Cessie (1990)
- Those authors developed the estimator for the shrinkage coefficient, γ:

$$\hat{\gamma} = 1 - \frac{\mathrm{df}}{\mathrm{LR}},$$

where df denotes the degrees of freedom of the model and LR is the likelihood ratio test statistic

Calibration

- The idea is that the model's predictions, $\{\hat{\eta}_i\}$, should be shrunken towards zero by γ:

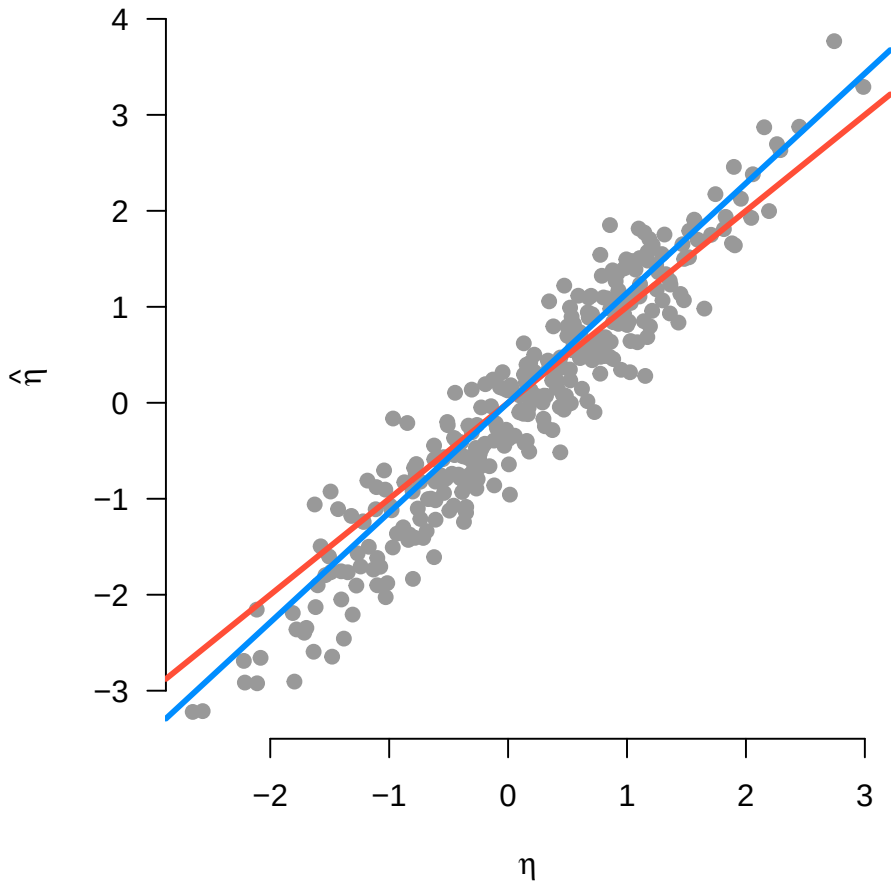
$$\tilde{\eta}_i = \hat{\gamma}\eta_i$$

- This is referred to as *calibration*; the idea is that the model's predictions need to be re-calibrated in order to account for the inevitable optimism that any model possesses
- Remark: This is not the only way to estimate γ; for example, a few authors have proposed estimators based on bootstrapping

Simulation

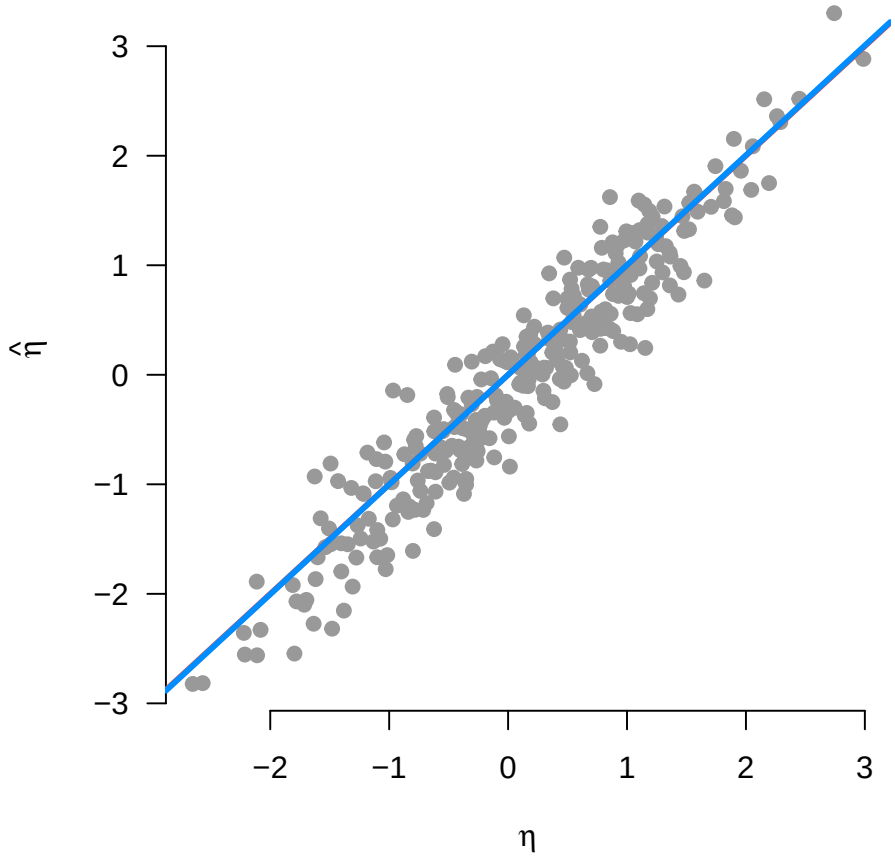
- To illustrate how this works, let's simulate some survival data from an exponential model (for simplicity, all observations are uncensored)
- In the generating model, there are 2 predictors for which a 1 SD change yields a hazard ratio of 2, and 28 predictors that have no effect on hazard
- Since this is simulated data, we can check the agreement between $\{\hat{\eta}_i\}$ and the true $\{\eta_i\}$ values for both the original and shrunken (calibrated) versions (in this example, $\hat{\gamma} = 0.87$)

Original estimates



Red: 1-to-1 line; blue: least squares line

Calibrated estimates



Red: 1-to-1 line; blue: least squares line

Calibration results

- For our models:
  - Model 1: $\hat{\gamma} = 0.97$
  - Model 2: $\hat{\gamma} = 0.97$
  - Model 3: $\hat{\gamma} = 0.86$
- This makes sense: models 1 and 2 are fairly parsimonious, and we shouldn't have to shrink their estimates much, while model 3 deserves some shrinkage
- The calibrated versions of $\mathrm{SD}(\eta)$:
  - Model 1: $\mathrm{SD}(\tilde{\eta}) = 0.68$
  - Model 2: $\mathrm{SD}(\tilde{\eta}) = 1.28$
  - Model 3: $\mathrm{SD}(\tilde{\eta}) = 1.50$

Sample splitting

- An alternative approach to addressing overfitting is to split the sample into two parts, using one for fitting the model and the other for assessing accuracy
- This approach is very common outside of time-to-event analysis, and can be done in a variety of ways: single validation, cross-validation, bootstrapping
- There are, however, some specific challenges that can arise when using these approaches with Cox models, as the partial likelihood-based deviance is entirely based on relative, as opposed to absolute, risk

Leave one out cross validation

- For example, we can fit the data to $\{\mathbf{X}, \mathbf{t}, \mathbf{d}\}_{-i}$ and calculate the linear predictor $\hat{\eta}_i = \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_{-i}$, but that linear predictor quantifies risk relative to the observations in $\{\mathbf{X}, \mathbf{t}, \mathbf{d}\}_{-i}$; upon observing t_i and d_i, how do we evaluate whether this was a good prediction or not?
- We can't use the Cox partial likelihood: with only one observation in the risk set, the likelihood would be 1 regardless of $\hat{\eta}_i$
- A variety of solutions have been proposed (one, for example, would be to just use the KL score instead), but this is still an open research question and papers continue to be published on the topic of quantifying predictive accuracy for Cox models without bias from overfitting

Cross-validated KL loss