Survival Analysis with Accelerated Failure Time Model in XGBoost

Avinash Barnwal, Hyunsu Cho, Toby Hocking

March 2020

1 What is survival analysis?

Survival analysis models **time to an event of interest**. Survival analysis is a special kind of regression and differs from the conventional regression task as follows:

- The label is always positive, since you cannot wait a negative amount of time until the event occurs.
- The label may not be fully known, or **censored**, because "it takes time to measure time."

The second bullet point is crucial and we should dwell on it more. As you may have guessed from the name, one of the earliest applications of survival analysis is to model mortality of a given population. Let's take NCCTG Lung Cancer Dataset as an example. The first 8 columns represent features and the last column, Time to death, represents the label.

Table 1: Example of survival data: NCCTG Lung Cancer Dataset

Inst	Age	Sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss	**Time to death (days)**
3	74	1	1	90	100	1175	N/A	306
3	68	1	0	90	90	1225	15	455
3	56	1	0	90	90	N/A	15	$[1010, +\infty)$
5	57	1	1	90	60	1150	11	210
1	60	1	0	100	90	N/A	0	883
12	74	1	1	50	80	513	0	$[1022, +\infty)$
7	68	2	2	70	60	384	10	310

Take a close look at the label for the third patient. **His label is a range, not a single number.** The third patient's label is said to be **censored**, because for some reason the experimenters could not get a complete measurement for that label. One possible scenario: the patient survived the first 1010 days and walked out of the clinic on the 1011th day, so his death was not directly observed. Another possibility: The experiment was cut short (since you cannot run it forever) before his death could be observed. In any case, his label is $[1010, +\infty)$, meaning his time to death can be any number that's higher than 1010, e.g. 2000, 3000, or 10000.

There are four kinds of censoring:

- **Uncensored**: the label is not censored and given as a single number.
- **Right-censored**: the label is of form $[a, +\infty)$, where a is the lower bound.
- **Left-censored**: the label is of form $(-\infty, b]$, where b is the upper bound.
- **Interval-censored**: the label is of form $[a, b]$, where a and b are the lower and upper bounds, respectively.

Right-censoring is the most commonly used.

2 Accelerated Failure Time model

Accelerated Failure Time (AFT) model is one of the most commonly used models in survival analysis. The model is of the following form:

$$\ln Y = \langle \mathbf{w}, \mathbf{x} \rangle + \sigma Z \tag{1}$$

where

- $\mathbf{x}$ is a vector in $\mathbb{R}^d$ representing the features.
- $\mathbf{w}$ is a vector consisting of d coefficients, each corresponding to a feature.
- $\langle \cdot, \cdot \rangle$ is the usual dot product in $\mathbb{R}^d$.
- $\ln(\cdot)$ is the natural logarithm.
- Y and Z are random variables.
  - Y is the output label.
  - Z is a random variable of a known probability distribution. Common choices are the normal distribution, the logistic distribution, and the extreme distribution. Intuitively, Z represents the "noise" that pulls the prediction $\langle \mathbf{w}, \mathbf{x} \rangle$ away from the true log label $\ln Y$.
- σ is a parameter that scales the size of Z.

Note that this model is a generalized form of a linear regression model $Y = \langle \mathbf{w}, \mathbf{x} \rangle$. In order to make AFT work with gradient boosting, we revise the model as follows:

$$\ln Y = \mathcal{T}(\mathbf{x}) + \sigma Z \tag{1'}$$

where $\mathcal{T}(\mathbf{x})$ represents the output from a decision tree ensemble, given input $\mathbf{x}$.

3 Derivation of AFT loss function

XGBoost requires a twice-differentiable loss function $\ell(y_i, \hat{y}_i)$ to perform gradient boosting. Common choices are

- Regression: $\ell(y, \hat{y}) = (1/2)(y - \hat{y})^2$
- Binary classification[1]: $\ell(y, \hat{y}) = -y \ln(\text{sigmoid}(\hat{y})) + (y-1)\ln(1 - \text{sigmoid}(\hat{y}))$.

The notation y represents the true label, whereas $\hat{y}$ represents the predicted label.

We will now define a suitable loss function ℓ_{AFT} to represent the AFT model. The loss function should represent the fitness of the model, i.e. how well the model fits the training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$.

Since Y in (1') is a random variable, it is possible to compute the probability density for the i-th data point:

$$\mathbb{P}[Y_i = y_i] = f_Y(y_i) \tag{2}$$

where

- Y_i is the random variable representing the i-th "drawing" from the distribution of Y in (1'). $Y_1, Y_2, \ldots, Y_n$ are i.i.d. This construction is justified as long as we assume that the data points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are independently determined from one another.
- f_Y indicates the probability density function (PDF) of the random variable Y. $f_Y(y_i)$ is the value of the PDF evaluated at $Y = y_i$.

[1] In this context, $y \in \{0, 1\}$ is a binary label, $\hat{y} \in \mathbb{R}$ is a margin score, and $\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$.

The likelihood function for the whole training data $\mathcal{D}$ is the product of probability densities for individual data points:

$$L(\mathcal{D}) = \mathbb{P}[Y_1 = y_1, \ldots, Y_n = y_n] = \mathbb{P}[Y_1 = y_1] \cdots \mathbb{P}[Y_n = y_n] = \prod_{i=1}^{n} \mathbb{P}[Y_i = y_i] = \prod_{i=1}^{n} f_Y(y_i) \tag{3}$$

To improve the fitness of our model, we should aim to maximize the likelihood function (3). This practice is known as **maximum likelihood estimation**. Usually, we elect to maximize the log likelihood instead[2], in order to substitute product with sum:

$$\ln L(\mathcal{D}) = \sum_{i=1}^{n} \ln \mathbb{P}[Y_i = y_i] = \sum_{i=1}^{n} \ln f_Y(y_i) \tag{4}$$

Unfortunately, we do not know y_i for some data points, due to label censoring. So we need to revise (4) to take account of these data points whose label is censored:

$$\begin{aligned}\ln L(\mathcal{D}) &= \underbrace{\sum \ln \mathbb{P}[Y_i = y_i]}_{\text{uncensored label}} + \underbrace{\sum \ln \mathbb{P}[Y_i \geq y_i^l]}_{\substack{\text{right-censored label} \\ y_i \in [y_i^l, +\infty)}} + \underbrace{\sum \ln \mathbb{P}[Y_i \leq y_i^u]}_{\substack{\text{left-censored label} \\ y_i \in (-\infty, y_i^u]}} + \underbrace{\sum \ln \mathbb{P}[y_i^l \leq Y_i \leq y_i^u]}_{\substack{\text{interval-censored label} \\ y_i \in [y_i^l, y_i^u]}} \\ &= \underbrace{\sum \ln f_Y(y_i)}_{\text{uncensored label}} + \underbrace{\sum \ln\left(1 - F_Y(y_i^l)\right)}_{\text{right-censored label}} + \underbrace{\sum \ln F_Y(y_i^u)}_{\text{left-censored label}} + \underbrace{\sum \ln\left(F_Y(y_i^u) - F_Y(y_i^l)\right)}_{\text{interval-censored label}}\end{aligned} \tag{4'}$$

where

- y_i^l and y_i^u are lower and upper bounds for the label y_i, respectively.
- F_Y is the cumulative distribution function (CDF), defined as $F_Y(y) = \mathbb{P}[Y \leq y] = \int_{-\infty}^{y} f_Y(t)dt$.

Notice that we are still applying maximum likelihood estimation: Each time we are given the range of the label in the form $[y_i^l, y_i^u]$, we maximize the likelihood for that range, $\mathbb{P}[y_i^l \leq Y_i \leq y_i^u]$.

The likelihood function in (4') can be expressed more precisely using index sets:

$$\ln L(\mathcal{D}) = \sum_{i \in U} \ln f_Y(y_i) + \sum_{i \in R} \ln\left(1 - F_Y(y_i^l)\right) + \sum_{i \in L} \ln F_Y(y_i^u) + \sum_{i \in V} \ln\left(F_Y(y_i^u) - F_Y(y_i^l)\right) \tag{4''}$$

where

$$\begin{aligned} U &= \{i : y_i \text{ is uncensored}\} & R &= \{i : y_i \text{ is right-censored}\} \\ L &= \{i : y_i \text{ is left-censored}\} & V &= \{i : y_i \text{ is interval-censored}\} \end{aligned} \tag{5}$$

We are now ready to define the loss function ℓ_{AFT}.

Definition 1 (Loss function for AFT survival regression)**.**

$$\ell_{\mathsf{AFT}}(y, \hat{y}) = \begin{cases} -\ln f_Y(y) & \textit{if } y \textit{ is not censored} \\ -\ln\left(1 - F_Y(y^l)\right) & \textit{if } y \textit{ is right-censored with } y \in [y^l, +\infty) \\ -\ln F_Y(y^u) & \textit{if } y \textit{ is left-censored with } y \in (-\infty, y^u] \\ -\ln\left(F_Y(y^u) - F_Y(y^l)\right) & \textit{if } y \textit{ is interval-censored with } y \in [y^l, y^u] \end{cases} \tag{6}$$

Under this definition, the sum of losses $\sum_{i=1}^{n} \ell(y_i, \hat{y}_i)$ over the training data is identical to $-\ln L(\mathcal{D})$. The minus sign is particularly useful because the gradient boosting will minimize any given loss function. Setting the loss to be negative of the likelihood ensures that the likelihood will be maximized as the loss is minimized. This definition is not yet complete, however, since we don't know yet how to actually compute (6), so let's unpack the terms $f_Y(y)$ and $F_Y(y)$ in terms of $\hat{y}$. We will need the following lemma:

[2] This step is justified because the logarithm is a monotone increasing function, i.e. $x < y$ if and only if $\ln x < \ln y$.

Lemma 1. *Let Y and Z be random variables. If $Y = g(Z)$ with monotone increasing function $g(\cdot)$, then the PDF and CDF of Y can be expressed in terms of the PDF and CDF of Z, respectively, as follows:*

$$f_Y(y) = f_Z(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y) \tag{7}$$

$$F_Y(y) = F_Z(g^{-1}(y)) \tag{8}$$

Proof. (8) follows from $\mathbb{P}[Y \le y] = \mathbb{P}[Z \le g^{-1}(y)]$. Differentiate both sides of (8) with respect to y (use the Chain Rule) to obtain (7). ∎

From (1'), we have

$$Y = \exp(\hat{y} + \sigma Z), \tag{9}$$

where we replaced $\mathcal{T}(\mathbf{x})$ with $\hat{y}$. Let's apply Lemma 1 with $g(Z) = \exp(\hat{y} + \sigma Z)$:

$$\begin{aligned} f_Y(y) &= f_Z\left(\frac{\ln y - \hat{y}}{\sigma}\right) \cdot \frac{1}{\sigma y} = f_Z(s(y)) \cdot \frac{1}{\sigma y} \\ F_Y(y) &= F_Z\left(\frac{\ln y - \hat{y}}{\sigma}\right) = F_Z(s(y)) \end{aligned} \tag{10}$$

where $s(y)$ is a shorthand for the expression $(\ln y - \hat{y})/\sigma$.

Now we have a working formula for ℓ_{AFT}, as we know how to compute the PDF and CDF of Z.

Definition 2 (Loss function for AFT survival regression, in terms of known PDF and CDF)**.**

$$\ell_{\mathsf{AFT}}(y, \hat{y}) = \begin{cases} -\ln\left[f_Z(s(y)) \cdot \dfrac{1}{\sigma y}\right] & \textit{if } y \textit{ is not censored} \\ -\ln\left[1 - F_Z(s(y^l))\right] & \textit{if } y \textit{ is right-censored with } y \in [y^l, +\infty) \\ -\ln F_Z(s(y^u)) & \textit{if } y \textit{ is left-censored with } y \in (-\infty, y^u] \\ -\ln\left[F_Z(s(y^u)) - F_Z(s(y^l))\right] & \textit{if } y \textit{ is interval-censored with } y \in [y^l, y^u] \end{cases} \tag{6'}$$

where f_Z and F_Z are given by Table 2 and $s(y) = (\ln y - \hat{y})/\sigma$ is defined the same way as in (10).

Table 2: Probability distributions for Z

Distribution	PDF ($f_Z(z)$)	CDF ($F_Z(z)$)
Normal	$\dfrac{\exp(-z^2/2)}{\sqrt{2\pi}}$	$\dfrac{1}{2}\left(1 + \operatorname{erf}\left(\dfrac{z}{\sqrt{2}}\right)\right)$
Logistic	$\dfrac{e^z}{(1+e^z)^2}$	$\dfrac{e^z}{1+e^z}$
Extreme[3]	$e^z e^{-\exp z}$	$1 - e^{-\exp z}$

4 Gradient and hessian of the AFT loss

The gradient boosting algorithm in XGBoost uses the gradient and hessian of the loss function, which are first and second partial derivatives of ℓ with respect to $\hat{y}$:

$$\text{Gradient} = \frac{\partial \ell}{\partial \hat{y}} \qquad\qquad \text{Hessian} = \frac{\partial^2 \ell}{\partial \hat{y}^2} \tag{11}$$

We first state the full formula for the gradient and hessian of the AFT loss function.

[3] We follow the formulation found in *A Package for Survival Analysis in S* (2012) by Terry M. Therneau, Mayo Foundation.

Definition 3 (Gradient and hessian of AFT loss)**.**

$$\left.\frac{\partial \ell_{\mathrm{AFT}}}{\partial \hat{y}}\right|_{y,\hat{y}} = \begin{cases} \dfrac{f_Z'(s(y))}{\sigma f_Z(s(y))} & \textit{if } y \textit{ is not censored} \\ \dfrac{-f_Z(s(y^l))}{\sigma[1-F_Z(s(y^l))]} & \textit{if } y \textit{ is right-censored with } y \in [y^l, +\infty) \\ \dfrac{f_Z(s(y^u))}{\sigma F_Z(s(y^u))} & \textit{if } y \textit{ is left-censored with } y \in (-\infty, y^u] \\ \dfrac{f_Z(s(y^u)) - f_Z(s(y^l))}{\sigma[F_Z(s(y^u)) - F_Z(s(y^l))]} & \textit{if } y \textit{ is interval-censored with } y \in [y^l, y^u] \end{cases} \tag{12}$$

$$\left.\frac{\partial^2 \ell_{\mathrm{AFT}}}{\partial \hat{y}^2}\right|_{y,\hat{y}} = \begin{cases} -\dfrac{f_Z(s(y))f_Z''(s(y)) - f_Z'(s(y))^2}{\sigma^2 f_Z(s(y))^2} & \textit{if } y \textit{ is not censored} \\ \dfrac{[1-F_Z(s(y^l))]f_Z'(s(y^l)) + f_Z(s(y^l))^2}{\sigma^2[1-F_Z(s(y^l))]^2} & \textit{if } y \textit{ is right-censored} \\ -\dfrac{F_Z(s(y^u))f_Z'(s(y^u)) - f_Z(s(y^u))^2}{\sigma^2 F_Z(s(y^u))^2} & \textit{if } y \textit{ is left-censored} \\ -\dfrac{[F_Z(s(y^u)) - F_Z(s(y^l))][f_Z'(s(y^u)) - f_Z'(s(y^l))] - [f_Z(s(y^u)) - f_Z(s(y^l))]^2}{\sigma^2[F_Z(s(y^u)) - F_Z(s(y^l))]^2} & \textit{if } y \textit{ is interval-censored} \end{cases} \tag{13}$$

where f_Z' and f_Z'' are the first and second derivatives of the PDF f_Z, respectively, and $s(y) = (\ln y - \hat{y})/\sigma$ is defined the same way as in Definition 2.

Table 3: First and second derivatives of PDF

Distribution	$f_Z'(z)$	$f_Z''(z)$
Normal	$-zf_Z(z)$	$(z^2-1)f_Z(z)$
Logistic	$\dfrac{f_Z(z)(1-e^z)}{1+e^z}$	$\dfrac{f_Z(z)(e^{2z}-4e^z+1)}{(1+e^z)^2}$
Extreme	$(1-e^z)f_Z(z)$	$(e^{2z}-3e^z+1)f_Z(z)$

4.1 Derivation of gradient and hessian

There's a lot of arithmetic happening in this section. You may choose to skip this section.

4.1.1 Uncensored data

$$\ell(y,\hat{y}) = -\ln\left[f_Z(s(y)) \cdot \frac{1}{\sigma y}\right] \tag{14}$$

$$\frac{\partial \ell}{\partial \hat{y}} = -\frac{\partial}{\partial \hat{y}} \ln\left[f_Z(s(y)) \cdot \frac{1}{\sigma y}\right] \tag{15}$$

$$= -\frac{1}{f_Z(s(y)) \cdot \dfrac{1}{\sigma y}} \cdot \frac{\partial}{\partial \hat{y}}\left[f_Z(s(y)) \cdot \frac{1}{\sigma y}\right] \qquad \text{Chain Rule} \tag{16}$$

$$= -\frac{\sigma y}{f_Z(s(y))} \cdot \left[f_Z'(s(y)) \cdot \frac{\partial}{\partial \hat{y}} \frac{\ln y - \hat{y}}{\sigma} \cdot \frac{1}{\sigma y}\right] \qquad \text{Chain Rule} \tag{17}$$

$$= -\frac{\sigma y}{f_Z(s(y))} \cdot \left[f_Z'(s(y)) \cdot -\frac{1}{\sigma} \cdot \frac{1}{\sigma y} \right] \tag{18}$$

$$= \frac{f_Z'(s(y))}{\sigma f_Z(s(y))} \tag{19}$$

$$\frac{\partial^2 \ell}{\partial \hat{y}^2} = \frac{\partial}{\partial \hat{y}} \frac{\partial \ell}{\partial \hat{y}} \tag{20}$$

$$= \frac{\partial}{\partial \hat{y}} \frac{f_Z'(s(y))}{\sigma f_Z(s(y))} \tag{21}$$

$$= \frac{\partial/\partial\hat{y}[f_Z'(s(y))] \cdot \sigma f_Z(s(y)) - f_Z'(s(y)) \cdot \sigma \partial/\partial\hat{y}[f_Z(s(y))]}{\sigma^2 f_Z(s(y))^2} \quad \text{Quotient Rule} \tag{22}$$

$$= \frac{f_Z''(s(y)) \cdot (-1/\sigma) \cdot \sigma f_Z(s(y)) - f_Z'(s(y)) \cdot \sigma f_Z'(s(y)) \cdot (-1/\sigma)}{\sigma^2 f_Z(s(y))^2} \quad \text{Chain Rule} \tag{23}$$

$$= -\frac{f_Z''(s(y)) f_Z(s(y)) - f_Z'(s(y))^2}{\sigma^2 f_Z(s(y))^2} \tag{24}$$

4.1.2 Right-censored data

$$\ell(y, \hat{y}) = -\ln\left[1 - F_Z(s(y^l))\right] \tag{25}$$

$$\frac{\partial \ell}{\partial \hat{y}} = -\frac{\partial}{\partial \hat{y}} \ln\left[1 - F_Z(s(y^l))\right] \tag{26}$$

$$= -\frac{1}{1 - F_Z(s(y^l))} \cdot \frac{\partial}{\partial \hat{y}} \left[1 - F_Z(s(y^l))\right] \quad \text{Chain Rule} \tag{27}$$

$$= -\frac{1}{1 - F_Z(s(y^l))} \cdot -f_Z(s(y^l)) \cdot -\frac{1}{\sigma} \quad \text{Chain Rule; } f_Z = F_Z' \tag{28}$$

$$= \frac{-f_Z(s(y^l))}{\sigma[1 - F_Z(s(y^l))]} \tag{29}$$

$$\frac{\partial^2 \ell}{\partial \hat{y}^2} = \frac{\partial}{\partial \hat{y}} \frac{\partial \ell}{\partial \hat{y}} \tag{30}$$

$$= \frac{\partial}{\partial \hat{y}} \frac{-f_Z(s(y^l))}{\sigma[1 - F_Z(s(y^l))]} \tag{31}$$

$$= \frac{-\partial/\partial\hat{y}[f_Z(s(y^l))] \cdot \sigma[1 - F_Z(s(y^l))] - [-f_Z(s(y^l))]\sigma \partial/\partial\hat{y}[1 - F_Z(s(y^l))]}{\sigma^2[1 - F_Z(s(y^l))]^2} \quad \text{Quotient Rule} \tag{32}$$

$$= \frac{-f_Z'(s(y^l)) \cdot (-1/\sigma) \cdot \sigma[1 - F_Z(s(y^l))] + f_Z(s(y^l))\sigma(-f_Z(s(y^l))) \cdot (-1/\sigma)}{\sigma^2[1 - F_Z(s(y^l))]^2} \quad \text{Chain Rule; } f_Z = F_Z' \tag{33}$$

$$= \frac{[1 - F_Z(s(y^l))] f_Z'(s(y^l)) + f_Z(s(y^l))^2}{\sigma^2[1 - F_Z(s(y^l))]^2} \tag{34}$$

4.1.3 Left-censored data

$$\ell(y, \hat{y}) = -\ln F_Z(s(y^u)) \tag{35}$$

$$\frac{\partial \ell}{\partial \hat{y}} = -\frac{\partial}{\partial \hat{y}} \ln F_Z(s(y^u)) \tag{36}$$

$$= -\frac{1}{F_Z(s(y^u))} \cdot \frac{\partial}{\partial \hat{y}} F_Z(s(y^u)) \quad \text{Chain Rule} \tag{37}$$

$$= -\frac{1}{F_Z(s(y^u))} \cdot f_Z(s(y^u)) \cdot -\frac{1}{\sigma} \quad \text{Chain Rule; } f_Z = F_Z' \tag{38}$$

$$= \frac{f_Z(s(y^u))}{\sigma F_Z(s(y^u))} \tag{39}$$

$$
\begin{aligned}
\frac{\partial^2 \ell}{\partial \hat{y}^2} &= \frac{\partial}{\partial \hat{y}} \frac{\partial \ell}{\partial \hat{y}} && (40)\\
&= \frac{\partial}{\partial \hat{y}} \frac{f_Z(s(y^u))}{\sigma F_Z(s(y^u))} && (41)\\
&= \frac{\partial/\partial \hat{y}[f_Z(s(y^u))] \cdot \sigma F_Z(s(y^u)) - f_Z(s(y^u)) \cdot \sigma \partial/\partial \hat{y}[F_Z(s(y^u))]}{\sigma^2 F_Z(s(y^u))^2} && \text{Quotient Rule} \quad (42)\\
&= \frac{f'_Z(s(y^u)) \cdot (-1/\sigma) \cdot \sigma F_Z(s(y^u)) - f_Z(s(y^u)) \cdot \sigma f_Z(s(y^u)) \cdot (-1/\sigma)}{\sigma^2 F_Z(s(y^u))^2} && \text{Chain Rule; } f_Z = F'_Z \quad (43)\\
&= \frac{-f'_Z(s(y^u)) F_Z(s(y^u)) + f_Z(s(y^u))^2}{\sigma^2 F_Z(s(y^u))^2} && (44)\\
&= -\frac{F_Z(s(y^u)) f'_Z(s(y^u)) - f_Z(s(y^u))^2}{\sigma^2 F_Z(s(y^u))^2} && (45)
\end{aligned}
$$

4.1.4 Interval-censored data

$$
\begin{aligned}
\ell(y, \hat{y}) &= -\ln\left[F_Z(s(y^u)) - F_Z(s(y^l))\right] && (46)\\
\frac{\partial \ell}{\partial \hat{y}} &= -\frac{\partial}{\partial \hat{y}} \ln\left[F_Z(s(y^u)) - F_Z(s(y^l))\right] && (47)\\
&= -\frac{1}{F_Z(s(y^u)) - F_Z(s(y^l))} \cdot \frac{\partial}{\partial \hat{y}}\left[F_Z(s(y^u)) - F_Z(s(y^l))\right] && \text{Chain Rule} \quad (48)\\
&= -\frac{1}{F_Z(s(y^u)) - F_Z(s(y^l))} \cdot \left[f_Z(s(y^u)) \cdot -\frac{1}{\sigma} - f_Z(s(y^l)) \cdot -\frac{1}{\sigma}\right] && \text{Chain Rule; } f_Z = F'_Z \quad (49)\\
&= \frac{f_Z(s(y^u)) - f_Z(s(y^l))}{\sigma[F_Z(s(y^u)) - F_Z(s(y^l))]} && (50)\\
\frac{\partial^2 \ell}{\partial \hat{y}^2} &= \frac{\partial}{\partial \hat{y}} \frac{\partial \ell}{\partial \hat{y}} && (51)\\
&= \frac{\partial}{\partial \hat{y}} \frac{f_Z(s(y^u)) - f_Z(s(y^l))}{\sigma[F_Z(s(y^u)) - F_Z(s(y^l))]} && (52)\\
&= \frac{\partial/\partial \hat{y}[f_Z(s(y^u)) - f_Z(s(y^l))] \cdot \sigma[F_Z(s(y^u)) - F_Z(s(y^l))]}{\sigma^2[F_Z(s(y^u)) - F_Z(s(y^l))]^2}\\
&\quad - \frac{[f_Z(s(y^u)) - f_Z(s(y^l))] \cdot \sigma \partial/\partial \hat{y}[F_Z(s(y^u)) - F_Z(s(y^l))]}{\sigma^2[F_Z(s(y^u)) - F_Z(s(y^l))]^2} && \text{Quotient Rule} \quad (53)\\
&= \frac{[f'_Z(s(y^u)) - f'_Z(s(y^l))] \cdot (-1/\sigma) \cdot \sigma[F_Z(s(y^u)) - F_Z(s(y^l))]}{\sigma^2[F_Z(s(y^u)) - F_Z(s(y^l))]^2}\\
&\quad - \frac{[f_Z(s(y^u)) - f_Z(s(y^l))] \cdot \sigma[f_Z(s(y^u)) - f_Z(s(y^l))] \cdot (-1/\sigma)}{\sigma^2[F_Z(s(y^u)) - F_Z(s(y^l))]^2} && \text{Chain Rule; } f_Z = F'_Z \quad (54)\\
&= -\frac{[F_Z(s(y^u)) - F_Z(s(y^l))][f'_Z(s(y^u)) - f'_Z(s(y^l))] - [f_Z(s(y^u)) - f_Z(s(y^l))]^2}{\sigma^2[F_Z(s(y^u)) - F_Z(s(y^l))]^2} && (55)
\end{aligned}
$$