

# QMWS - Survival Analysis

## Regression models for survival analysis

Instructor: Kevin McGregor

York University  
Department of Mathematics and Statistics

We've seen how to get multiple survival curves and how to test for differences (log-rank test).

However, we may want to do more sophisticated modelling, with multiple continuous and/or categorical variables affecting survival probabilities.

We have many techniques for **regression analysis** in the context of survival data.

- To begin, we will do a **very short** review of linear regression.
- It is assumed you have seen multiple linear regression before.

# Review of simple linear regression

In **simple linear regression**, we wish to study the relationship between a **predictor** variable  $x_i$  and an **outcome** variable  $y_i$ , where  $i = 1, \dots, n$ .

The predictor  $x_i$  is assumed to be **fixed** and the outcome  $y_i$  is assumed to be **random**. **Examples:**

- $y_i$  is blood pressure,  $x_i$  is age.
- $y_i$  is a measure of user satisfaction after using a web portal,  $x_i \in \{0, 1\}$  is one of two portal designs.

# A basic probabilistic model for linear regression

The linear regression model takes the form:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where  $\epsilon_i$  is a random error term, e.g. a random variable with zero mean and finite variance ( $E[\epsilon_i] = 0$ ,  $\text{var}[\epsilon_i] = \sigma^2$ ); it represents the error present in the measurement of  $y_i$ .

Terminology:

- $\beta_0$  - **Intercept** parameter
- $\beta_1$  - **Slope** parameter

# A basic probabilistic model for linear regression

- $\beta_1 > 0$ : increasing  $y_i$  with increasing  $x_i$
- $\beta_1 < 0$ : decreasing  $y_i$  with increasing  $x_i$
- $\beta_1 = 0$ : no *linear* relationship between  $x_i$  and  $y_i$

We sometimes write  $Y_i$  when we think of the outcome as a **random variable**. We have that:

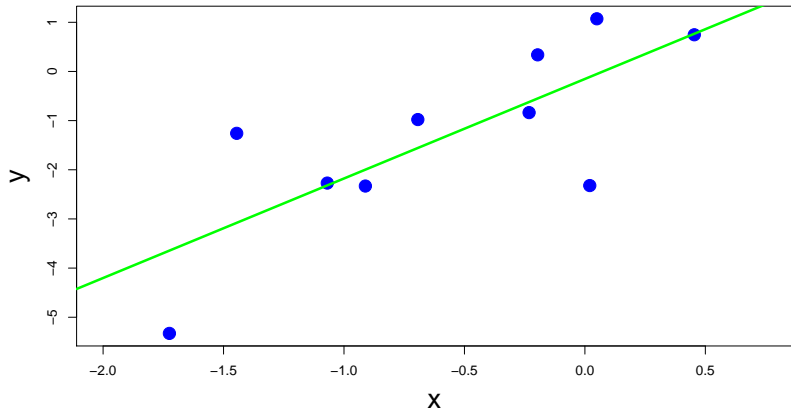
$$E[Y_i|x] = \beta_0 + \beta_1 x_i$$

where  $E[Y_i|x_i]$  is the expected (mean) value of  $Y_i$  for fixed value of  $x_i$ .

$$E[Y_i|x_i] = \beta_0 + \beta_1 x_i$$

- $\beta_1$  is the expected difference in response  $Y_i$  between two groups of individuals, where one group has covariate value  $x_i$  one unit greater than the other group.
- $\beta_0$  is the expected value of  $Y_i$  in a group of individuals with covariate value  $x_i = 0$ .

# Linear regression



We also assume a **normal distribution** for the error term  $\epsilon_i$ :

$$\epsilon_i \sim \text{Normal}(0, \sigma^2)$$

This is equivalent to saying that:

$$Y_i | x_i \sim \text{Normal}(\beta_0 + \beta_1 x_i, \sigma^2)$$



## Multiple linear regression

In **multiple linear regression** we have  $p$  predictor variables  $x_{i1}, \dots, x_{ip}$  and outcome  $y_i$  for each  $i = 1, \dots, n$  and assume that:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i,$$

where  $\epsilon_i \sim \text{Normal}(0, \sigma^2)$  and the observations are independent.

Alternatively we have:

$$y_i | x_i \sim \text{Normal}(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}, \sigma^2)$$

We have that:

$$E[Y_i | x_{i1}, \dots, x_{ip}] = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

- $\beta_j$  is the expected difference in response  $Y_i$  between two individuals, where one individual has predictor value  $x_j$  one unit greater than the other individual, **assuming all other predictors are held constant**.
- $\beta_0$  is the expected value of  $Y_i$  in an individual with **all** covariate values set to  $x_{ij} = 0$ .

How do we perform regression analysis in survival data?

In linear regression, we modelled the **mean** of the outcome given the predictors:

$$E[Y_i | x_{i1}, \dots, x_{ip}] = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

For regression in survival analysis, we typically model the **hazard function** as a function of covariates.

- Interpretability: the hazard function describes the chance of having an event given you've survived up to time  $t$ :

$$\lambda(t, x_{i1}, \dots, x_{ip}) = h(t, x_i^\top \beta)$$

for some non-negative function  $h(\cdot, \cdot)$ .

In **exponential regression**, we assume that the underlying time-to-event data is  $z_i \sim \text{Exp}(\theta_i)$ , for  $i = 1, \dots, n$ .

For now, we'll assume **right-censored** data. Our **observed** time variable  $t_i$  is defined like usual:

- If individual  $i$  is censored,  $t_i$  is the censoring time.
- If we observe the event, then  $t_i$  is the event time  $z_i$ .

Like before  $\delta_i = 0$  if right censored,  $\delta_i = 1$  if the event is observed.

Now, assume we have  $p$  predictors measured for each individual  $x_{i1}, \dots, x_{ip}$ .

- We want to study the effect of each predictor on the time to an event.

Recall that the exponential distribution has **constant hazard**. The hazard function for the exponential distribution is:

$$\lambda(t) = \theta_i$$

To study the effects of  $x_{i1}, \dots, x_{ip}$  on  $\theta_i$  (the hazard), we need to choose some function  $h$  so that:

$$\theta_i = h(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip})$$

What if we choose  $h$  as the **identity function**? That is:

$$h(x) = x$$

Therefore, our model would be:

$$\theta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

**Problem:** We need  $\theta_i > 0$ . This means we would need to somehow constrain  $\beta_0, \beta_1, \dots, \beta_p$  to make sure  $\theta_i$  is in the proper range.

- This is quite difficult.

Alternative choice:

$$h(x) = \exp\{x\}$$

In this case, our model would be:

$$\theta_i = \exp\{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\}$$

By doing this, we **guarantee** that  $\theta_i > 0$ , regardless of the values of the  $\beta$  parameters or the predictors.

- **Advantage:** We do not have to constrain the  $\beta$  parameters, each  $\theta_j$  could be any real number.

## Interpretation of exponential regression parameters:

$$\lambda(x_i) = \theta_i = \exp\{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\}$$

What happens if we have two individuals, such that  $x_{1j} = a$  and  $x_{2j} = a + 1$ , and all other predictors the same between the two? Consider the **ratio** of their hazard functions:

$$\frac{\theta_2}{\theta_1} =$$

Thus, when  $x_{ij}$  increases by one unit, the hazard function is multiplied by a factor of  $\exp\{\beta_j\}$ .



## Interpretation of exponential regression parameters:

$$\lambda(x_i) = \theta_i = \exp\{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\}$$

Similarly,  $\exp\{\beta_0\}$  is the hazard when all  $x_j = 0$  for  $j = 1, \dots, p$ .

If  $x_j = 0$  is possible for all  $j = 1, \dots, p$  in the data, then we call:

$$\exp\{\beta_0\}$$

the **baseline hazard**.

In **exponential regression**, we modelled the **hazard function** of the exponential distribution with rate parameter  $\theta_i$  as:

$$\lambda(x_i) = \exp \{ \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \}$$

Remember, the exponential distribution assumes **constant hazard**; this means the chance of an event **does not change over time**.

In many cases, this assumption does not hold.

- Ex: All cause mortality, with  $t = 0$  being birth. The chance of death is higher in infancy and in old age.
- Exponential regression is **not appropriate** in cases like this.

One alternative is **Weibull regression**.

One parameterization of the Weibull distribution is:

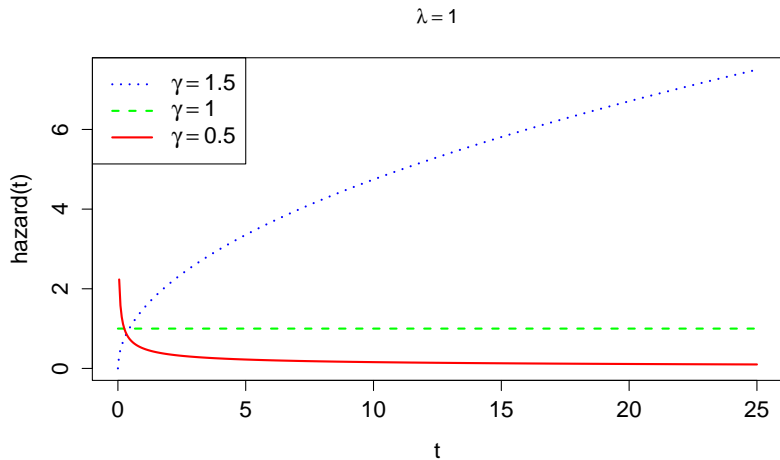
$$f(t) = \lambda \gamma (\lambda t)^{\gamma-1} e^{-(\lambda t)^\gamma} \quad (t > 0)$$

where  $\lambda > 0$  and  $\gamma > 0$ .

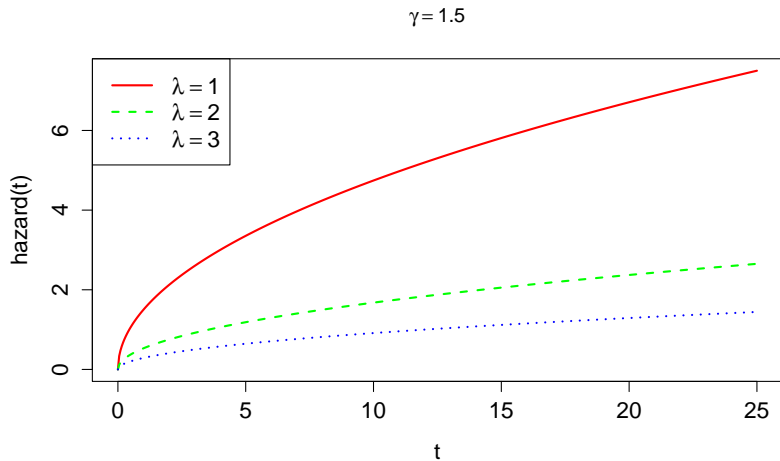
The **hazard function** for the Weibull distribution is given by:

$$\lambda(t) = \lambda \gamma (\lambda t)^{\gamma-1}$$

# Weibull hazard



# Weibull hazard



In **Weibull regression**, we posit the following model for the hazard function, as a function of a vector of covariates  $x_i = (x_{i1}, \dots, x_{ip})$ ,

$$\lambda(t, x_i) = \gamma(\lambda t)^{\gamma-1} \exp \{ \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} \},$$

where  $\beta = (\beta_1, \dots, \beta_p)^\top$ .

Note: **there is no intercept**. Why not?

Interpretation of parameters: Once again, assume that we have two individuals whose covariate  $j$  differs by exactly one unit:

$x_{1j} = a$  and  $x_{2j} = a + 1$ .

Assume that  $x_{1k} = x_{2k}$  for all other covariates  $k \neq j$ .

$$\lambda(t, x_i) = \gamma(\lambda t)^{\gamma-1} \exp \left\{ x_i^\top \beta \right\},$$

The **ratio** of the hazards between individual 2 and 1 at time  $t$  is:

**Example:** Veteran lung cancer dataset. Outcome is death. Two covariates we'll use in the model are:

- age: Ranges from 34-81 years.
- karno: Karnofsky score. Higher value corresponds to higher ability for a patient to care for themselves. Very low score refers to hospitalization.

We'll fit a Weibull regression model with age and karno as predictors.



## Weibull hazard: veterans example

**Result:** The maximum likelihood estimates of  $\gamma$  and  $\lambda$  are:

$$\hat{\gamma} = 14.26 \quad \hat{\lambda} = 1.02$$

The coefficients for age and karno (respectively) are:

$$\beta_{\text{age}} = 0.00018 \quad \beta_{\text{karno}} = -0.03419$$

Interpretation of  $\beta_{\text{karno}}$ :

# Weibull hazard: veterans example

