19. Survival analysis

survival probability in our dataset here means the probability that the customers will stay

Although this branch of statistics is usually referred to as **Survival Analysis**, the event in question does not need to be related to actual “survival”. The important thing is to understand that we are interested in **the time until something happens**, or whether or not something will happen in a certain time frame.

**Question:** But why is this different? Can’t you just use the techniques you learned so far (e.g., regression models) to predict the time? Take a minute to think about this.

The answer would be yes if you could observe the actual time in all occurrences, but you usually cannot. Frequently, there will be some kind of **censoring** which will not allow you to observe the exact time that the event happened for all units/individuals that are being studied.

Graphical user interface, text, application, email

Description automatically generated

For example in this case, only 1 customer will churn after 2 months, the remaining 4 do not churn even after 50, 2, 29, and 57 months. But if we use linear regression to predict that the first customer will churn after 50 months, then we are wrong.

Solutions:

Approach 1: Only consider the examples where “Churn”=Yes

On average they will be **underestimates** (too small), because we are ignoring the currently subscribed (un-churned) customers. Our dataset is a biased sample of those who churned within the time window of the data collection. Long-time subscribers were more likely to be removed from the dataset! This is a common mistake - see the [Calling Bullshit video](https://www.youtube.com/watch?v=ITWQ5psx9Sw) I posted on the README!

### Approach 2: Assume everyone churns right now[¶](https://ubc-cs.github.io/cpsc330/lectures/19_survival-analysis.html#approach-2-assume-everyone-churns-right-now)

Assume everyone churns right now - in other words, use the original dataset.

It will be an **underestimate** again. For those still subscribed, while we did not remove them, we recorded a total tenure shorter than in reality, because they will keep going for some amount of time.

### Approach 3: Survival analysis[¶](https://ubc-cs.github.io/cpsc330/lectures/19_survival-analysis.html#approach-3-survival-analysis)

## **Kaplan-Meier survival curve**

Graphical user interface, application

Description automatically generated

## **Cox proportional hazards model**

* The Cox proportional hazards model is a commonly used model that allows us to interpret how features influence a censored tenure/duration.
* You can think of it like linear regression for survival analysis: we will get a coefficient for each feature that tells us how it influences survival.
* It makes some strong assumptions (the proportional hazards assumption) that may not be true, but we won’t go into this here.
* The proportional hazard model works multiplicatively, like linear regression with log-transformed targets.
* Looks like month-to-month leads to more churn, two-year contract leads to less churn; this makes sense!!!

Table

Description automatically generated

We can also use it to predict other people’s survival curve

From predict\_survival\_function documentation:

Predict the survival function for individuals, given their covariates. This assumes that the individual just entered the study (that is, we do not condition on how long they have already lived for.)

Graphical user interface, text, application, email

Description automatically generated

T/F questions

 If all customers joined a service at the same time (hypothetically), then censoring would not be an issue.

This is False. Because there is still the problem of having customers who didn't churn in our dataset, how do we interpret their tenure? The issue of censoring is still present, which is why the Cox model should be used.

I think censoring would still be an issue, because the issue isn't just the time somebody joined but also the time that somebody leaves.