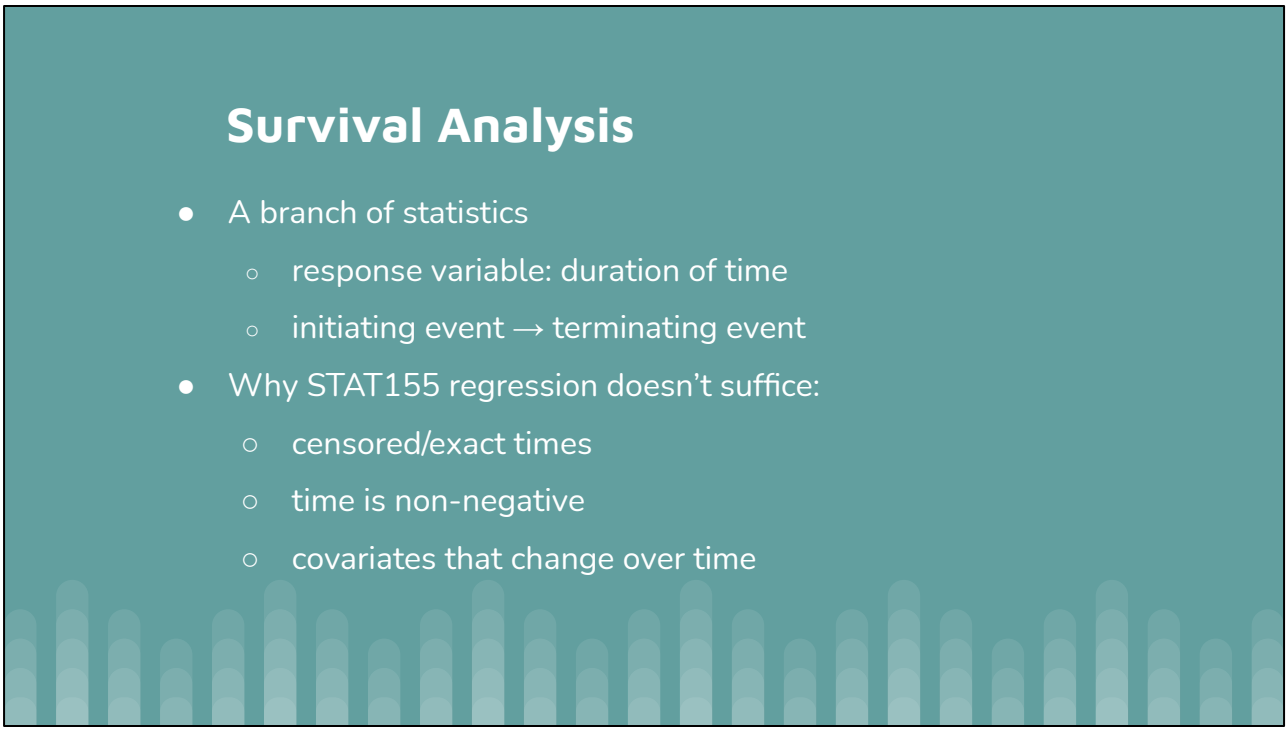
How long after founding does a company go public?

Prof. Vittorio Addona - STAT 453
Charlotte Giang and Yunyang Zhong

Hi everyone and welcome to our capstone presentation in STAT 453: Survival Analysis. We are Yunyang Zhong and Charlotte Giang, guided by our professor Vittorio Addona. Our research question is “How long after founding does a company go public?”

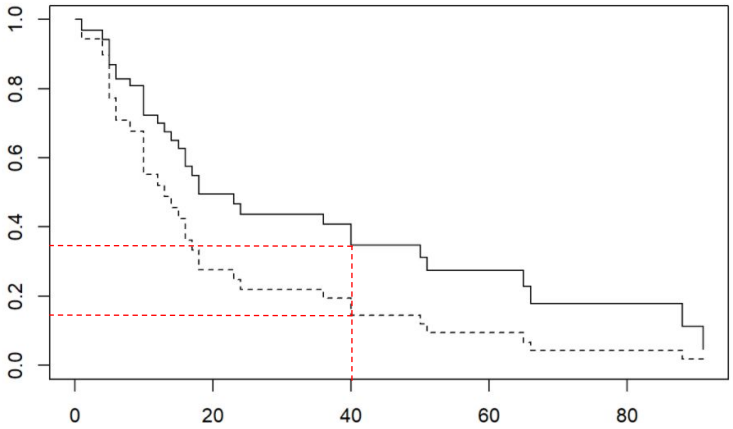
Some background on this topic. A company can go from private to public through a process called Initial Public Offering, or IPO, which allows the company's stocks to be listed on a stock exchange and purchased by public investors. A company may seek IPO to raise capital by selling its shares to fuel its **operations** and **expansion**. With that in mind, we seek to model the time from a company's founding date to its IPO date, with covariates concerning the company's profitability, size, sector and leadership. To model this time duration, we use the techniques we've learned from this very class, Survival Analysis.



What is survival analysis? Survival analysis is a branch of statistics where the response variable is a duration of time from some initiating event to some terminating event.

Why can't we just use regression models in STAT155 to model this duration of time? Because it's a bit more complicated than that. Sometimes it is impossible to observe the terminating event, so survival models may involve censored times for observations with no exact initiating or terminating time. Durations of time also cannot be negative, making certain distributional assumptions more appropriate than others. In addition, some covariates can change over time.

Survival Curves



When one thinks of survival analysis, this image can immediately pop into mind. This is an example of what survival curves look like. The survival function is the complement of the cumulative distribution function, which is the integral of the density function of a random variable.

This graph plots the survival function of two groups. Let's consider time 40. The solid-line group has a 35% chance of surviving past time 40, while the dashed-line group has a 15% chance.

Our Data

- Source: Stocks IPO information & results | Kaggle
  - Selection bias - no censored data
- Data wrangling:
  - Response variable:
    - DaysToIPO: approximated
    - YearsToIPO: exact
  - Selected covariates “happened” before IPO
  - Dimensions: 2,821 observations x 13 covariates

Our data is from a Kaggle dataset named Stocks IPO information & results. All the stocks were publicly traded on 1/1/2018, which means there is no censored data.

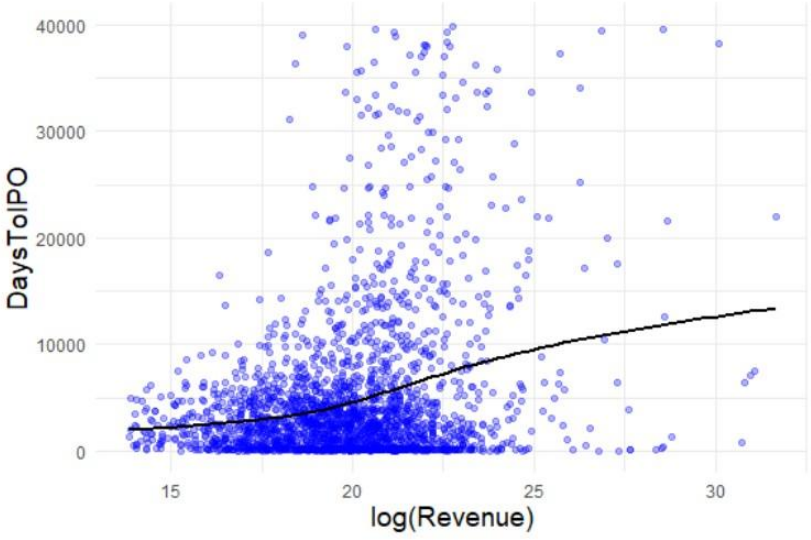
Our response variables are DaysToIPO and YearsToIPO, both of which measure the duration of time from the founding date to the IPO date of a company.

In the original dataset, some values in founding dates were missing. If all the date, the month and the year were missing, we removed those observations. For founding dates with only years recorded, the founding date in year was assumed to be June 30^{th}. For founding dates with only months and years recorded, the founding day of month was assumed to be the 15^{th}. Thus, in our cleaned dataset, DaysToIPO is estimated and YearsToIPO is exact.

Regarding covariates, we selected those that happened before IPO in time, concerning a company’s profitability, size, sector and leadership.

After data wrangling, our dataset has 2,821 observations and 13 covariates, ready for next steps.

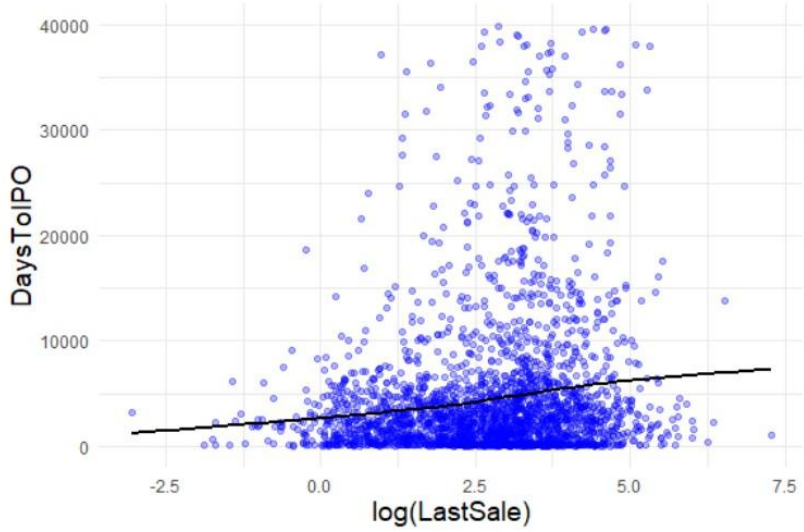
Exploratory Data Analysis



We made several graphs to explore the relationships between covariates prior to modeling.

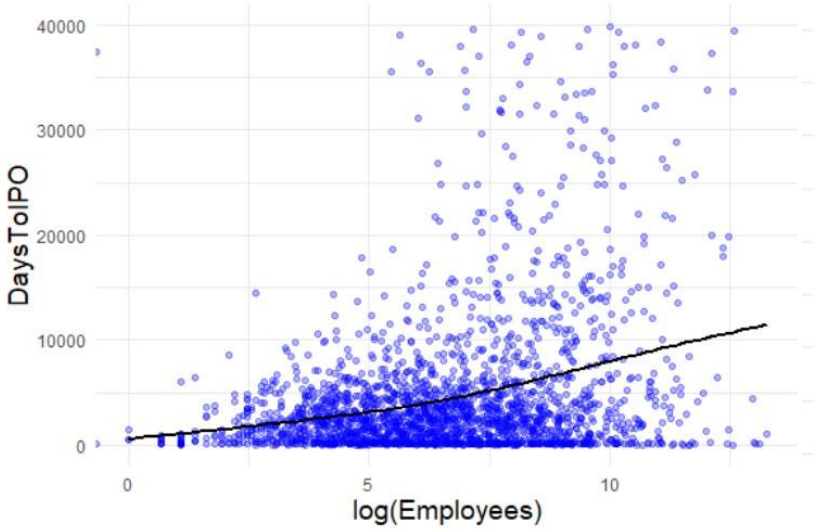
This graph shows that log(Revenue) is positively correlated to DaysToIPO. It indicates that companies with higher revenues tended to take longer before going public. This may be because those companies were less in need of capital so they continued to stay private.

Exploratory Data Analysis



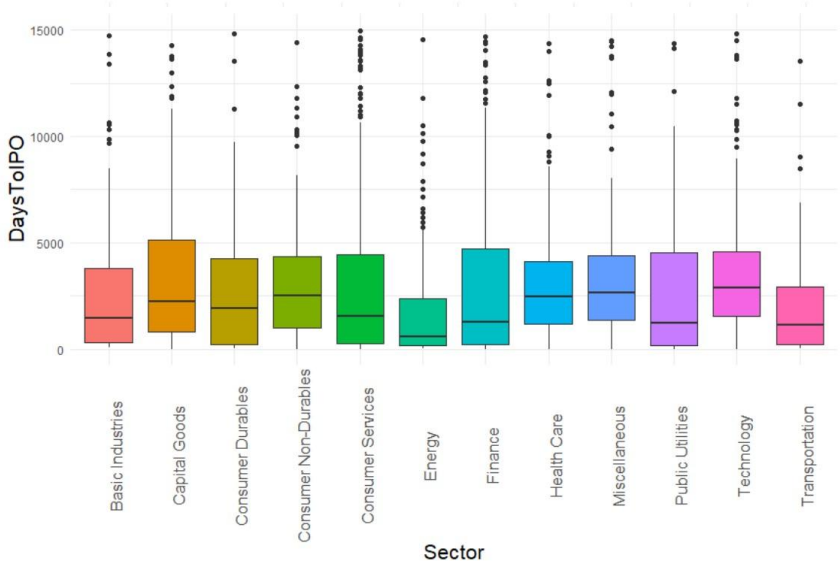
In this graph, the positive relationship between log(LastSale) and DaysToIPO indicates that companies with higher sales in the year before IPO tended to take longer to go public. This may mean those companies were more profitable and less in need of capital, or that companies that went public later simply had higher previous year's sales due to inflation.

Exploratory Data Analysis



In this graph, log(Employees) is positively correlated to DaysToIPO. This suggests that companies with more employees took longer to go public, possibly because they were less in need of expansion.

Exploratory Data Analysis



In this graph, we can see that companies in different sectors had different median DaysToIPO's and different interquantile ranges too. For example, companies in Energy had relatively shorter DaysToIPO compared to companies in Transportation.

Parametric Survival Modeling

Weibull model:

DaysToIPO ~ LastSale + Employees + CEOAge + Sector

After exploratory analysis, we considered fitting survival parametric models to predict DaysToIPO. Here, we fitted a Weibull model to predict DaysToIPO using previous years' sales, number of employees, CEO's age and sector.

Parametric Survival Modeling

Weibull model:

DaysToIPO ~ LastSale + Employees + CEOAge + Sector

	Coefficient	p-value
LastSale	**0.0018**	0.0165
Employees	**8.11×10^{-6}**	7.8×10^{-7}
CEOAge	**0.0169**	6.6×10^{-5}
SectorEnergy	**-0.0715**	0.7089
SectorTransportation	**0.0781**	0.7549

Here's the output coefficients for our covariates. The lines "SectorEnergy" and "SectorTransportation" are comparing companies in those sectors to those in the baseline sector, "Basic Industries". There are 12 sectors in total -- we're only displaying 2 of them for illustration purposes. The coefficients of LastSale, Employees and CEOAge are significant at a 5% level.

What do the coefficients mean, though? It turns out that, unlike in STAT155, we cannot directly interpret the coefficients as an increase in DaysToIPO per unit increase in each covariate.

Parametric Survival Modeling

Weibull model: Accelerated Failure Time Model

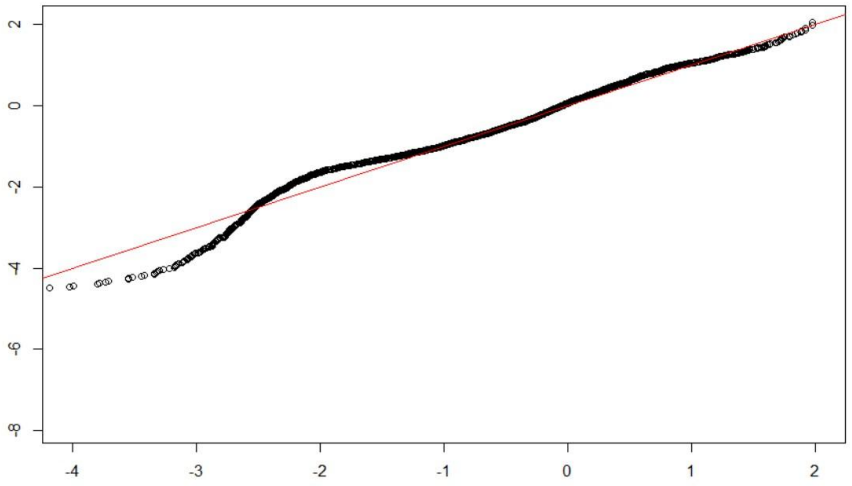
DaysToIPO ~ LastSale + Employees + CEOAge + Sector

	$e^{\text{Coefficient}}$	= Time Ratio	p-value
LastSale	$e^{0.0018}$	= **1.002**	0.0165
Employees	$e^{8.11x10\text{^}(-6)}$	≈ **1**	$7.8x10^{-7}$
CEOAge	$e^{0.0169}$	= **1.017**	$6.6x10^{-5}$
SectorEnergy	$e^{-0.0715}$	= **0.9310**	0.7089
SectorTransportation	$e^{0.0781}$	= **1.0812**	0.7549

Rather, the exponents of those covariates are multiplicative changes in DaysToIPO per unit increase in each covariate. For example, the coefficient of LastSale is 0.0018, which means for one unit increase in the previous years' sales, the time from founding to IPO of an average company is increased by 0.02%, keeping other covariates constant. For levels of the Sector covariate, the times to IPO of companies in each sector is compared to those in the baseline sector, "Basic Industries". For example, the time to IPO of a company in Energy is 93% that of a company in Basic Industries. The time to IPO of a company in Transportation, on the other hand, are 8% longer than that of a company in Basic Industries.

The property where the time ratio is the exponents of coefficients is quite special, and parametric models with this property are called Accelerated Failure Time models. Aside from the Weibull model, the Exponential and the Lognormal models are also Accelerated Failure Time models.

Parametric Model Appropriateness



Why did we choose to fit a Weibull model instead of other parametric models for our data?

We made this decision after performing model comparison. In survival analysis, we can use Cox-Snell residuals to check the adequacy of **any** parametric models. Using the Probability Integral Transform, we can prove that for a random variable T of **any** continuous distribution, -log(S(T)) is Exponential with lambda = 1. Thus, if a parametric model is appropriate for a given dataset, the plot of the Cox-Snell residuals will be a straight line with slope=1. In our case, we plotted the graph for 4 parametric models: normal, exponential, weibull and lognormal. Based on the 4 Cox-Snell residuals plots, we chose the Weibull model because it was the most appropriate one for our data. This plot is the Cox-Snell residuals plot for the Weibull model. We can see that the dots follow the red line with slope 1 quite closely, indicating that the Weibull model is appropriate.

Parametric Model Comparison -> Weibull

Likelihood Ratio Test (LRT)

If p-value < α, then more complex model is worthwhile.

Akaike's Information Criterion (AIC)

AIC = 2p - 2L

Lower AIC is better

Besides using c-log-log graphs to check model adequacy, we also tried AIC and LRT to compare models. On the left is an example of an LRT, which can only be applied to nested models, such as weibull and exponential (because exponential is a special case of weibull). This test focuses on how much the likelihood increases. If p-value < a significance level, then the more complex model has provided a sufficient gain in likelihood to warrant the added complexity. In our case, we get a p-value = 0 so Weibull is chosen.

Utilizing AIC, we need to find which model has the lowest AIC metric, which indicates a better fit. AIC can be applied to both nested and non-nested models and thus we can compare weibull and lognormal. p = number of parameters and L = log-likelihood. Similarly, Weibull has the lowest AIC and it is the better fit. This agrees with our conclusion from the c-log-log graphs that Weibull is the best. Fortunately, all results agree with each other and our conclusion is robust.

After considering what model to choose, we need to think about what covariates to include as well. Same as above, LRT and AIC will tell us whether to add a covariate or not. For example, both of the tests suggest that including the covariate Sector is worthwhile.

Time vs Hazard

- Transition from modeling time to modeling hazard
- Lower hazards are associated with longer times
- Hazard function: the chance of failure at time *t* given that failure has not yet occurred up to *t*

Another way to model survival is to find the hazard instead of time, which are negatively associated with each other: lower hazard means longer time from the initiating event to the terminating event. A hazard function represents the chance of failure at time t, given that failure has not yet occurred up to t, and thus the area under the hazard curve equals a conditional probability. It provides an alternative way of describing survival which can be useful for illustrating the “instantaneous” risk of failure.

Cox Proportional Hazards (PH) Models

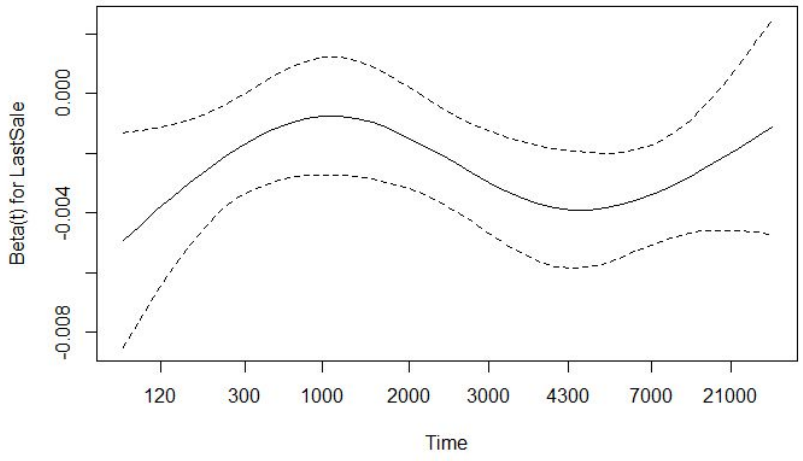
- The Cox PH model assumes that, for any 2 values of a covariate, the **hazard ratio is constant over time** (i.e., multiplicative impact of a covariate on h(*t*) is identical at all time *t*).

- How to check the PH assumption for a covariate?

The Cox PH model assumes that, for any 2 values of a covariate, the hazard ratio is constant over time, which means the multiplicative impact of a covariate on h(t) is identical at all time t, thus the name “Proportional Hazards”. Therefore, only the covariates that satisfy the PH assumption should be added to a PH model.

Then, the question is, how to check the PH assumption for a covariate? Based on the definition, one way to do so is to calculate the hazard ratio at different times and check if they are the same. Another way is to graphically check the PH assumption using Schoenfeld residuals.

Checking PHs Assumption



In this plot of the Schoenfeld residuals, the y axis is “Beta(t) for LastSale”, which is the coefficient of netIncome in a CoxPH model, and also equivalent to the log hazard ratio when LastSale is increased by one unit. If the PH assumption is valid for LastSale then the Schoenfeld residuals will show no trend over time. Based on this plot, the general direction satisfies the criterion.

However, there are too many data points close to each other and it is hard to tell exactly if the PH assumption is appropriate for this covariate. Therefore, we also used a formal test check the assumption. Here, the null hypothesis refers to a valid PH while the alternative hypothesis refers to an invalid PH. As the p-value is large, we do not have enough evidence against the null hypothesis. Thus, the PH assumption is valid for LastSale and it can be added to a CoxPH model.

Cox Proportional Hazards Modeling

DaysToIPO ~ LastSale + netIncome + strata(Sector) + strata(CEOGender)

	$e^{Coefficient}$	= Hazard Ratio	p-value
LastSale	$e^{-1.98*10^{(-3)}}$	= 0.998	0.0003
netIncome	$e^{-2.03*10^{(-13)}}$	= 1.000	0.1607

For covariates satisfying the PH assumption, including netIncome, our next step was to test if its relationship with DaysToIPO is statistically significant. In this output, the small p-value for LastSale indicates its significance. Although netIncome satisfies the PH assumption, it is not significant in this model.

Since it is a CoxPH model, the exponent of a coefficient is the hazard ratio when the corresponding covariate is increased by one unit. For example, when LastSale is increased by one unit, the hazard is decreased by 0.2%, thus a longer time for a company to go public. This result agrees with that from the Weibull model, where a higher value of previous year's sales is associated with a longer time to IPO.

The hazard ratio when netIncome increases by one unit is equal to 1.0, so its value does not affect the hazard of going public.

Key Findings

- Weibull and CoxPH models agree
- LastSale, CEOAge and Employees are positively correlated to DaysToIPO
  - LastSale is the most statistically significant
- Effect of Sector

Through graphical and formal model comparison methods, the Weibull model was the most appropriate parametric model for our data. The results obtained from our Weibull model were consistent with those from our CoxPH model.

In summary, we've found that previous year's sales, CEO's age, and the number of employees each has a positive relationship with time from founding to IPO, among which previous year's sales is the most statistically significant covariate. Companies in different sectors also had different times to IPO. With "Basic Industries" being the baseline sector, holding other covariates constant, companies in Capital Goods, Consumer Non-Durables, Public Utilities and Transportation had longer times, while Consumer Durables, Consumer Services, Energy, Finance, Health Care, Technology and Miscellaneous sectors had shorter times from founding to IPO.

Last but not least, we're gonna briefly talk about some of the limitations of our analysis as well as suggestions for future work.

First, we have no censored data since all the companies in our dataset had already gone public by Jan 1st, 2018. This can be a selection bias towards a shorter time to IPO: companies that are not included might take longer to go public. Therefore, our models could underestimate a company's time from founding to IPO. Second, companies with no founding date available were filtered out, which could influence the accuracy of our result. The third limitation is that our models were built using DaysToIPO for more convenient interpretation, although DaysToIPO are approximated where the exact founding dates were missing. However, when we built the same models using YearsToIPO, the result was very much similar to when we used DaysToIPO. Thus, we believe that the general direction of the relationships should be the same. Our final note is that there are many other covariates in the original dataset, so if given more time, we would be able to find other meaningful associations.

Thank you for your time listening to our presentation!