Important Facts to keep in mind

Survival versus Failure

S(t) defines the probability of surviving beyond time t (i.e. $P(T > t)$)

F(t) is the failure probability, or the probability of failing (having event occur) prior to time t ($P(T \leq t)$)

When we predict outcomes from a survival regression, we are predicting when we think the event will occur (in other words, when will it “fail”)

When you use the “predict” command in R, you will be predicting the mean for each observation (is that reasonable for survival data?)

Checking assumption on distribution

One of the **BIGGEST** assumptions for the AFT model is that we correctly specified the distribution of the error

Since we cannot rely on p-values until AFTER we have the correct assumption made, recommend using all variables to decide which distribution is best (well, remove any multicollinearities first!!)

Once you decide on distribution, now go through and choose which variables are important in the model

Using graphical procedure: pro…can see if distribution is good for data; con…no statistical test

Using statistical procedure: pro…can perform a statistical test; con…can ONLY compare distributions (see which one is best….if they both are bad, this test will NOT tell you that!!)

Making predictions

We can predict the mean (that is what we do in other regressions):

```
head(predict(recid.aft.w))
```

```
[1] 128.26394  58.86229  43.55317 156.35349  87.52751
[6] 119.17415
```

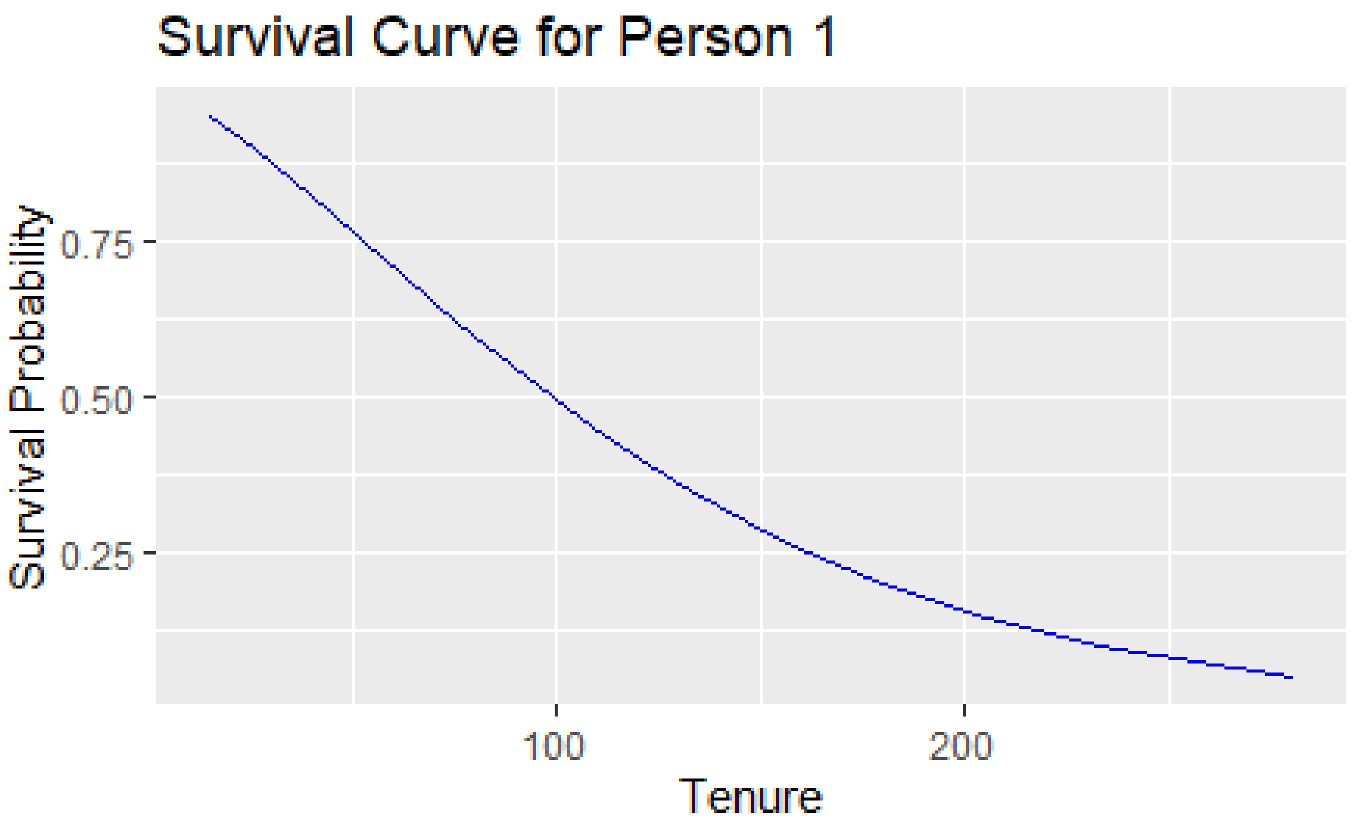
Predict quantiles..

Or we can predict quantiles (0.25, 0.50, 0.75)

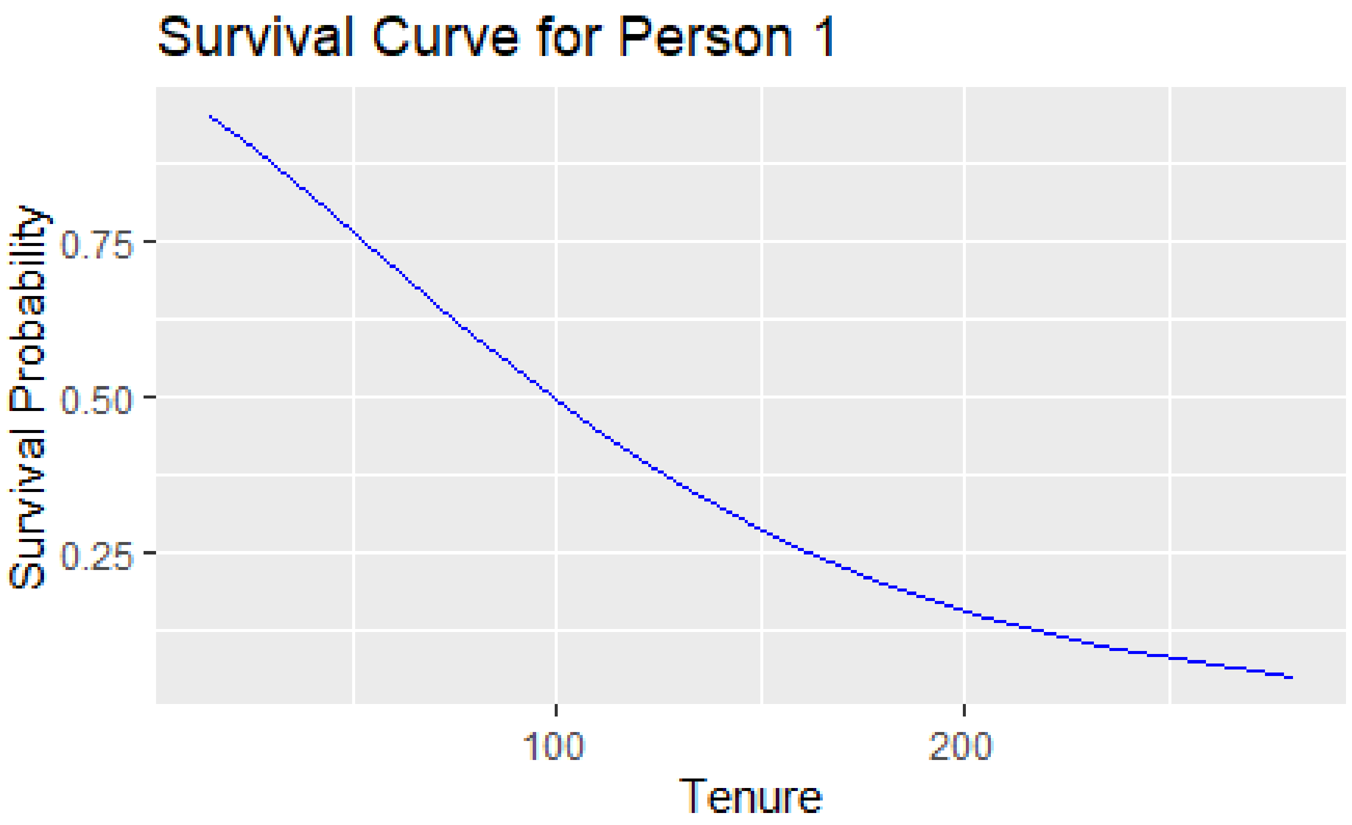
```
survprob.75.50.25 = predict(recid.aft.w, type = "quantile", se.fit = TRUE,p = c(0.25, 0.5, 0.75))
head(survprob.75.50.25$fit)
```

```
         [,1]      [,2]      [,3]
[1,] 52.68849  98.72758 161.95827
[2,] 24.17956  45.30760  74.32514
[3,] 17.89085  33.52383  54.99438
[4,] 64.22717 120.34873 197.42682
[5,] 35.95471  67.37185 110.52057
[6,] 48.95457  91.73097 150.48064
```

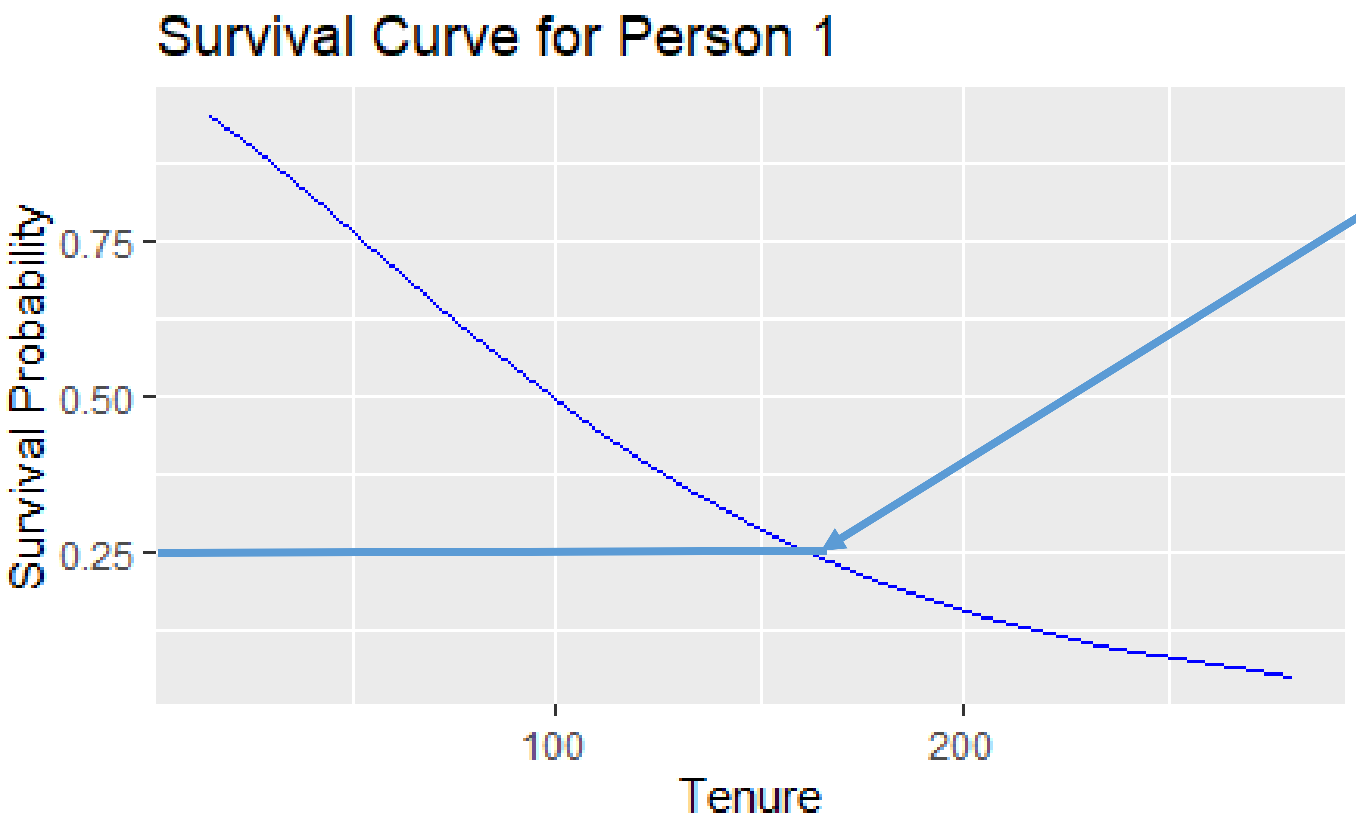
```
quant.prob=seq(0.05,0.95,by=0.05)
survprob = predict(recid.aft.w, type = "quantile", se.fit = TRUE,p = quant.prob)
surv.prob=rev(quant.prob)
graph.dat=data.frame(cbind(survprob$fit[1,],surv.prob))
colnames(graph.dat)=c("Tenure","SurvivalProb")
ggplot(graph.dat,aes(x=Tenure,y=SurvivalProb))+geom_line(color="blue")+labs(title="
Survival Curve for Person 1",x="Tenure",y="Survival Probability")
```


Survival Curve for Person 1
Survival Probability
0.75
0.50
0.25
100
200
Tenure

	0.25	0.5	0.75
[1,]	52.68849	98.72758	161.95827

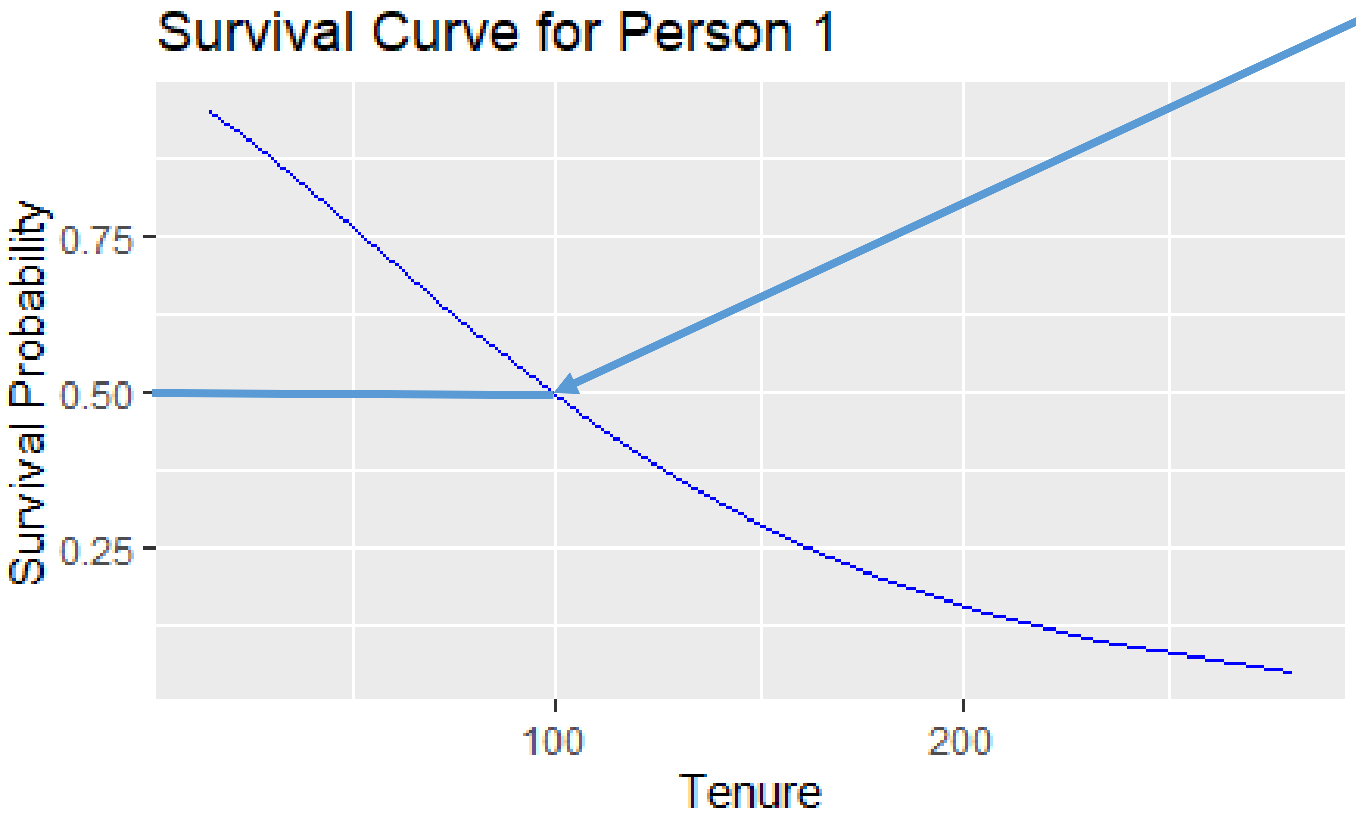


	0.25	0.5	0.75
[1,]	52.68849	98.72758	$\boxed{161.95827}$



Probability of event occurring on or before 161.95 is 0.75 (probability of "surviving" beyond 161.95 is 0.25!)

	0.25	0.5	0.75
[1,]	52.68849	$\boxed{98.72758}$	161.95827



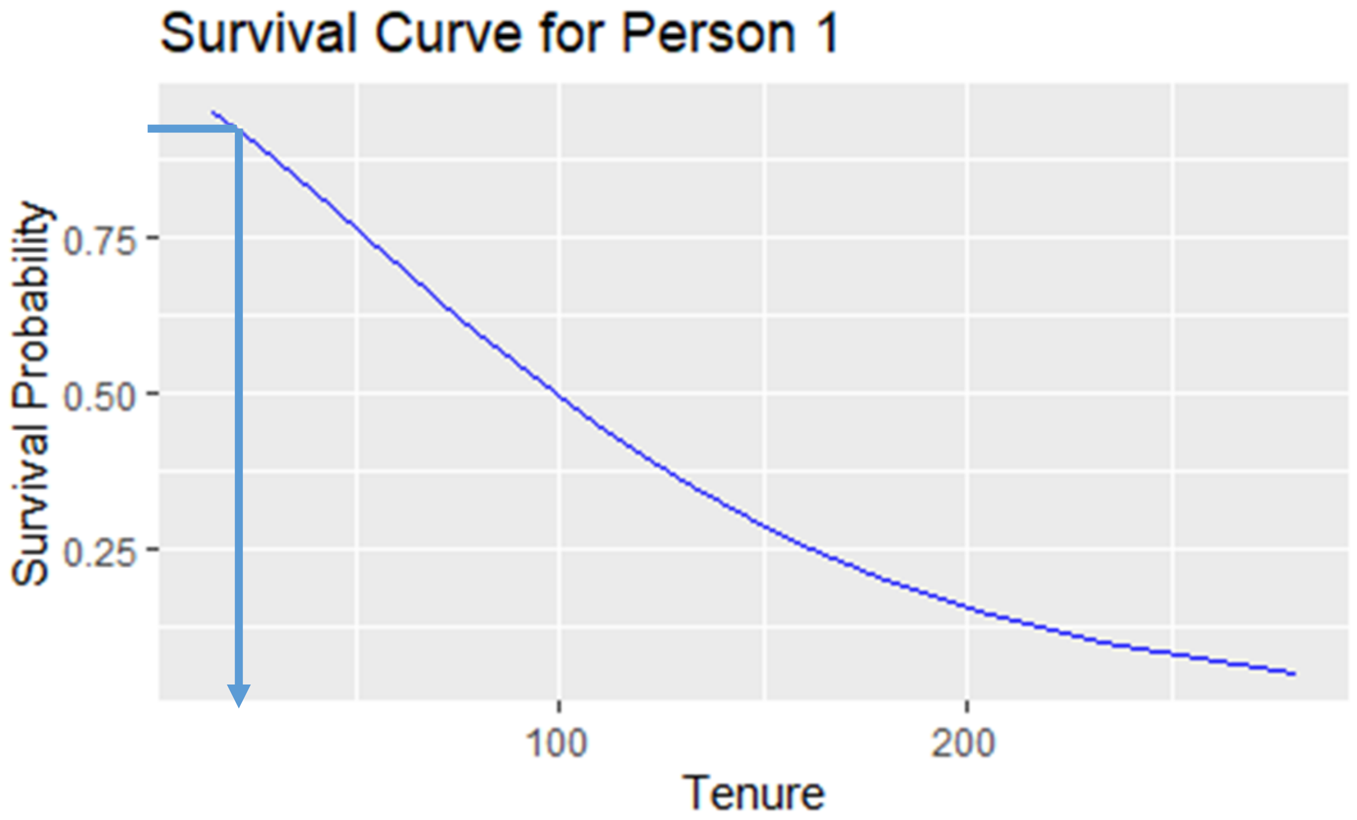
Probability of event occurring on or before 98.73 is 0.5 (probability of "surviving" beyond 98.73 is 0.5!)

Finding survival probabilities

We can go the “opposite” direction..find probabilities instead of quantiles!! We will find the survival probability for each of the observed values….

Actual arrest weeks..

```
> head(recid$week)
[1] 20    17    25     52    52     52
```


Survival Curve for Person 1
Survival Probability
0.75
0.50
0.25
100
200
Tenure

```
survprob.actual = 1 - psurvreg(recid$week,
      mean = predict(recid.aft.w, type = "lp"),
      scale = recid.aft.w$scale, distribution = recid.aft.w$dist)
head(survprob.actual, n = 10)
```

```
> head(recid$week)
[1] 20    17    25     52    52     52
```

```
[1] 0.9285822 0.8389085 0.6315234 0.8073231
[5] 0.6173609 0.7312118 0.9260438 0.7203354
[9] 0.5891529 0.7143008
```

We can also do this for a point in time…10 weeks…

```
survprob.10wk = 1 - psurvreg(10,
    mean = predict(recid.aft.w, type = "lp"),
    scale = recid.aft.w$scale,
    distribution = recid.aft.w$dist)
head(survprob.10wk)
```

```
[1] 0.9723202 0.9198457 0.8803901 0.9789527 0.9531961
[6] 0.9693657
```

Impact of a variable

We can use this information to help us find the impact of changing a variable

For example: let’s take a look at what would be the impact to those individuals who did NOT have financial aid if they would have had it…

```
new_time = qsurvreg(1 - survprob.actual,
  mean = predict(recid.aft.w, type = "lp") +
  coef(recid.aft.w)['fin'],
  scale = recid.aft.w$scale,
  distribution = recid.aft.w$dist)
```

Impact of a variable

We can use this information to help us find the impact of changing a variable

For example: let’s take a look at what would be the impact to those individuals who did NOT have financial aid if they would have had it…

```
new_time = qsurvreg(1 - survprob.actual,
  mean = predict(recid.aft.w, type = "lp") +
  coef(recid.aft.w)['fin'],
  scale = recid.aft.w$scale,
  distribution = recid.aft.w$dist)
```

Finding a quantile..

Impact of a variable

We can use this information to help us find the impact of changing a variable

For example: let's take a look at what would be the impact to those individuals who did NOT have financial aid if they would have had it…

```
new_time = qsurvreg(1 - survprob.actual,
  mean = predict(recid.aft.w, type = "lp") +
  coef(recid.aft.w)['fin'],
  scale = recid.aft.w$scale,
  distribution = recid.aft.w$dist)
```

Keeping the same location on the curve as the original data..

Impact of a variable

We can use this information to help us find the impact of changing a variable

For example: let’s take a look at what would be the impact to those individuals who did NOT have financial aid if they would have had it…

```
new_time = qsurvreg(1 - survprob.actual,
  mean = predict(recid.aft.w, type = "lp") +
  coef(recid.aft.w)['fin'],
  scale = recid.aft.w$scale,
  distribution = recid.aft.w$dist)
```

For the linear predictor, add in the coefficient for financial aid

```
recid$new_time = new_time
recid$diff = recid$new_time - recid$week


impact.fin=data.frame(recid$week, recid$new_time,
recid$diff,recid$arrest,recid$fin)
colnames(impact.fin)=c("O.Week","N.Week","Diff","Arrest"
,"Fin")
head(impact.fin2)
```

	O.Week	N.Week	Diff	Arrest	Fin
1	20	25.66776	5.667764	1	0
2	17	21.81760	4.817600	1	0
3	25	32.08471	7.084706	1	0
7	23	29.51793	6.517929	1	0
13	37	47.48536	10.485364	1	0
15	25	32.08471	7.084706	1	0

6 rows