Weibull Regression with R, Part Two*

```
> rm(list=ls());  options(scipen=999)
> # install.packages("survival",dependencies=TRUE) # Only need to do this once
> library(survival) # Do this every time
> # install.packages("asaur",dependencies=TRUE) # Only need to do this once
> library(asaur)
> # help(pharmacoSmoking)
> head(pharmacoSmoking)
   id ttr relapse         grp age gender     race employment yearsSmoking
1  21 182       0   patchOnly  36   Male    white         ft           26
2 113  14       1   patchOnly  41   Male    white      other           27
3  39   5       1 combination  25 Female    white      other           12
4  80  16       1 combination  54   Male    white         ft           39
5  87   0       1 combination  45   Male    white      other           30
6  29 182       0 combination  43   Male hispanic         ft           30
  levelSmoking ageGroup2 ageGroup4 priorAttempts longestNoSmoke
1        heavy     21-49     35-49             0              0
2        heavy     21-49     35-49             3             90
3        heavy     21-49     21-34             3             21
4        heavy       50+     50-64             0              0
5        heavy     21-49     35-49             0              0
6        heavy     21-49     35-49             2           1825
> summary(pharmacoSmoking)
       id              ttr              relapse                grp    
 Min.   :  1.00   Min.   :  0.00   Min.   :0.000   combination:61  
 1st Qu.: 33.00   1st Qu.:  8.00   1st Qu.:0.000   patchOnly  :64  
 Median : 67.00   Median : 49.00   Median :1.000                   
 Mean   : 66.15   Mean   : 77.44   Mean   :0.712                   
 3rd Qu.: 99.00   3rd Qu.:182.00   3rd Qu.:1.000                   
 Max.   :130.00   Max.   :182.00   Max.   :1.000                   
      age            gender          race      employment  yearsSmoking  
 Min.   :22.00   Female:81   black   :38   ft   :72   Min.   : 9.00  
 1st Qu.:41.00   Male  :44   hispanic: 8   other:39   1st Qu.:22.00  
 Median :49.00               other   : 2   pt   :14   Median :30.00  
 Mean   :48.84               white   :77              Mean   :30.88  
 3rd Qu.:56.00                                        3rd Qu.:39.00  
 Max.   :86.00                                        Max.   :56.00  
 levelSmoking ageGroup2  ageGroup4  priorAttempts     longestNoSmoke  
 heavy:89     21-49:66   21-34:16   Min.   :   0.00   Min.   :   0.0  
 light:36     50+  :59   35-49:50   1st Qu.:   1.00   1st Qu.:   7.0  
                         50-64:48   Median :   2.00   Median :  90.0  
                         65+  :11   Mean   :  12.68   Mean   : 539.7  
                                    3rd Qu.:   5.00   3rd Qu.: 365.0  
                                    Max.   :1000.00   Max.   :6205.0  
> # Make fixed-up data frame called quit
> quit = within(pharmacoSmoking,{
+ DayOfRelapse = Surv(ttr+1,relapse)
+ contrasts(grp) = contr.treatment(2,base=2) # Patch only is reference category
+ colnames(contrasts(grp)) = c('Combo') # Names of dummy vars -- just one
+ # Collapse race categories
+ Race = as.character(race) # Small r race is a factor. This is easier to modify.
+ Race[Race!='white'] = 'blackOther'; Race=factor(Race)
+ }) # Finished making data frame quit
> with(quit, table(race,Race) )
          Race
race       blackOther white
  black            38     0
  hispanic          8     0
  other             2     0
  white             0    77
```

```
> full = survreg(DayOfRelapse ~ grp + age + gender + Race + employment
+         + yearsSmoking + levelSmoking + priorAttempts, dist='weibull', data=quit)
> summary(full)

Call:
survreg(formula = DayOfRelapse ~ grp + age + gender + Race + 
    employment + yearsSmoking + levelSmoking + priorAttempts, 
    data = quit, dist = "weibull")
                      Value Std. Error     z            p
(Intercept)         1.12177    0.97726  1.15       0.2510
grpCombo            1.09225    0.38186  2.86       0.0042
age                 0.08432    0.03411  2.47       0.0134
genderMale          0.03631    0.41417  0.09       0.9301
Racewhite           0.25145    0.39143  0.64       0.5206
employmentother    -1.28799    0.46719 -2.76       0.0058
employmentpt       -1.28482    0.58631 -2.19       0.0284
yearsSmoking       -0.02351    0.03250 -0.72       0.4696
levelSmokinglight -0.07347    0.43151 -0.17       0.8648
priorAttempts      -0.00105    0.00200 -0.52       0.6000
Log(scale)          0.54194    0.08917  6.08 0.0000000012

Scale= 1.72 

Weibull distribution
Loglik(model)= -463.8   Loglik(intercept only)= -476.5
      Chisq= 25.41 on 9 degrees of freedom, p= 0.0025 
Number of Newton-Raphson Iterations: 5 
n= 125 
>
```

I am thinking about dropping Race, yearsSmoking, levelSmoking and priorAttempts. The last 3 variables all represent smoking history and could be correlated highly enough to wash out each other's effects. Test them simultaneously.

```
>
> # Fit the restricted model: Restricted by H0
> rest1 =  survreg(DayOfRelapse ~ grp + age + gender + Race + employment ,
+                    dist='weibull', data=quit)
> anova(rest1,full) # LR test

Terms
1 grp + age + gender + Race + employment
2 grp + age + gender + Race + employment + yearsSmoking + levelSmoking + 
priorAttempts
  Resid. Df    -2*LL Test Df  Deviance  Pr(>Chi)
1       117 928.3771      NA        NA        NA
2       114 927.5513    =  3 0.8258271 0.8432801

> # Is Race significant with those variables dropped?
```

```
> # Is Race significant with those variables dropped?
> summary(rest1)

Call:
survreg(formula = DayOfRelapse ~ grp + age + gender + Race + 
    employment, data = quit, dist = "weibull")
                  Value Std. Error     z            p
(Intercept)      1.3905     0.8684  1.60       0.1093
grpCombo         1.1021     0.3794  2.91       0.0037
age              0.0637     0.0190  3.35       0.0008
genderMale       0.0561     0.4140  0.14       0.8921
Racewhite        0.1880     0.3788  0.50       0.6196
employmentother -1.2821     0.4635 -2.77       0.0057
employmentpt    -1.2251     0.5837 -2.10       0.0358
Log(scale)       0.5444     0.0894  6.09 0.0000000011

Scale= 1.72 

Weibull distribution
Loglik(model)= -464.2   Loglik(intercept only)= -476.5
      Chisq= 24.58 on 6 degrees of freedom, p= 0.00041 
Number of Newton-Raphson Iterations: 5 
n= 125
```

Decision: Drop race and gender.

```
> full2 =  survreg(DayOfRelapse ~ grp + age + employment , dist='weibull', 
data=quit)
> summary(full2)

Call:
survreg(formula = DayOfRelapse ~ grp + age + employment, data = quit, 
    dist = "weibull")
                  Value Std. Error     z           p
(Intercept)      1.4957     0.8414  1.78     0.07545
grpCombo         1.1023     0.3793  2.91     0.00366
age              0.0643     0.0186  3.45     0.00055
employmentother -1.2880     0.4617 -2.79     0.00528
employmentpt    -1.2123     0.5616 -2.16     0.03088
Log(scale)       0.5454     0.0894  6.10 0.000000001

Scale= 1.73 

Weibull distribution
Loglik(model)= -464.3   Loglik(intercept only)= -476.5
      Chisq= 24.31 on 4 degrees of freedom, p= 0.000069 
Number of Newton-Raphson Iterations: 5 
n= 125

> # Test employment status controlling for age and experimental treatment.
> rest2 = survreg(DayOfRelapse ~ grp + age , dist='weibull', data=quit)
> anova(rest2,full2) # LR test
                   Terms Resid. Df    -2*LL Test Df Deviance    Pr(>Chi)
1              grp + age       121 937.9007      NA       NA          NA
2 grp + age + employment       119 928.6554    =  2 9.245333 0.009826558
```

```
> # Test employment status with a Wald test.
> source("http://www.utstat.toronto.edu/~brunner/Rfunctions/Wtest.txt")
> # function(L,Tn,Vn,h=0) # H0: L theta = h
> # Tn is estimated theta, usually a vector.
> # Vn is the estimated asymptotic covariance matrix of Tn.
> # For Wald tests based on numerical MLEs, Tn = theta-hat,
> # and Vn is the inverse of the Hessian of the minus log likelihood.
>
> Vhat = vcov(full2); Vhat
                  (Intercept)      grpCombo           age employmentother
(Intercept)      0.7079360800 -0.0320256900 -0.0147694486     0.111673731
grpCombo        -0.0320256900  0.1438698739 -0.0004703383    -0.013521493
age             -0.0147694486 -0.0004703383  0.0003472409    -0.003893727
employmentother  0.1116737308 -0.0135214927 -0.0038937268     0.213191081
employmentpt    -0.0003554818 -0.0078279548 -0.0013434899     0.077138486
Log(scale)      -0.0098224903  0.0050290739  0.0002048412    -0.003182291
                 employmentpt    Log(scale)
(Intercept)     -0.0003554818 -0.0098224903
grpCombo        -0.0078279548  0.0050290739
age             -0.0013434899  0.0002048412
employmentother  0.0771384860 -0.0031822913
employmentpt     0.3153999894 -0.0035442716
Log(scale)      -0.0035442716  0.0079888732

> thetahat = full2$coefficients; thetahat
    (Intercept)        grpCombo             age employmentother
      1.4957374       1.1023048       0.0643414      -1.2880472
   employmentpt
     -1.2122529
>
```

Note that the asymptotic covariance matrix includes log(sigma), but the "coefficients" vector does not.

```
> sigmahat = full2$scale; sigmahat
[1] 1.725305
> thetahat = c(thetahat,log(sigmahat))
>
> # H0: beta3=beta4=0. Express as H0: L theta = h
> eMat = rbind( c(0,0,0,1,0,0),
+               c(0,0,0,0,1,0)   )

> Wtest(L=eMat, Tn=thetahat, Vn=Vhat)
          W          df     p-value
9.718885315 2.000000000 0.007754805
> anova(rest2,full2) # Repeating LR test for comparison
                 Terms Resid. Df    -2*LL Test Df Deviance    Pr(>Chi)
1 grp + age                  121 937.9007      NA       NA          NA
2 grp + age + employment     119 928.6554    =  2 9.245333 0.009826558

>
> # Test part time versus other
> pto = cbind(0,0,0,1,-1,0); pto
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0    0    0    1   -1    0
> Wtest(L=pto, Tn=thetahat, Vn=Vhat)
         W         df    p-value
0.01534747 1.00000000 0.90140640
>
```

Predict the day of relapse for a 50 year old patient who is employed full time and gets the patch-only treatment.

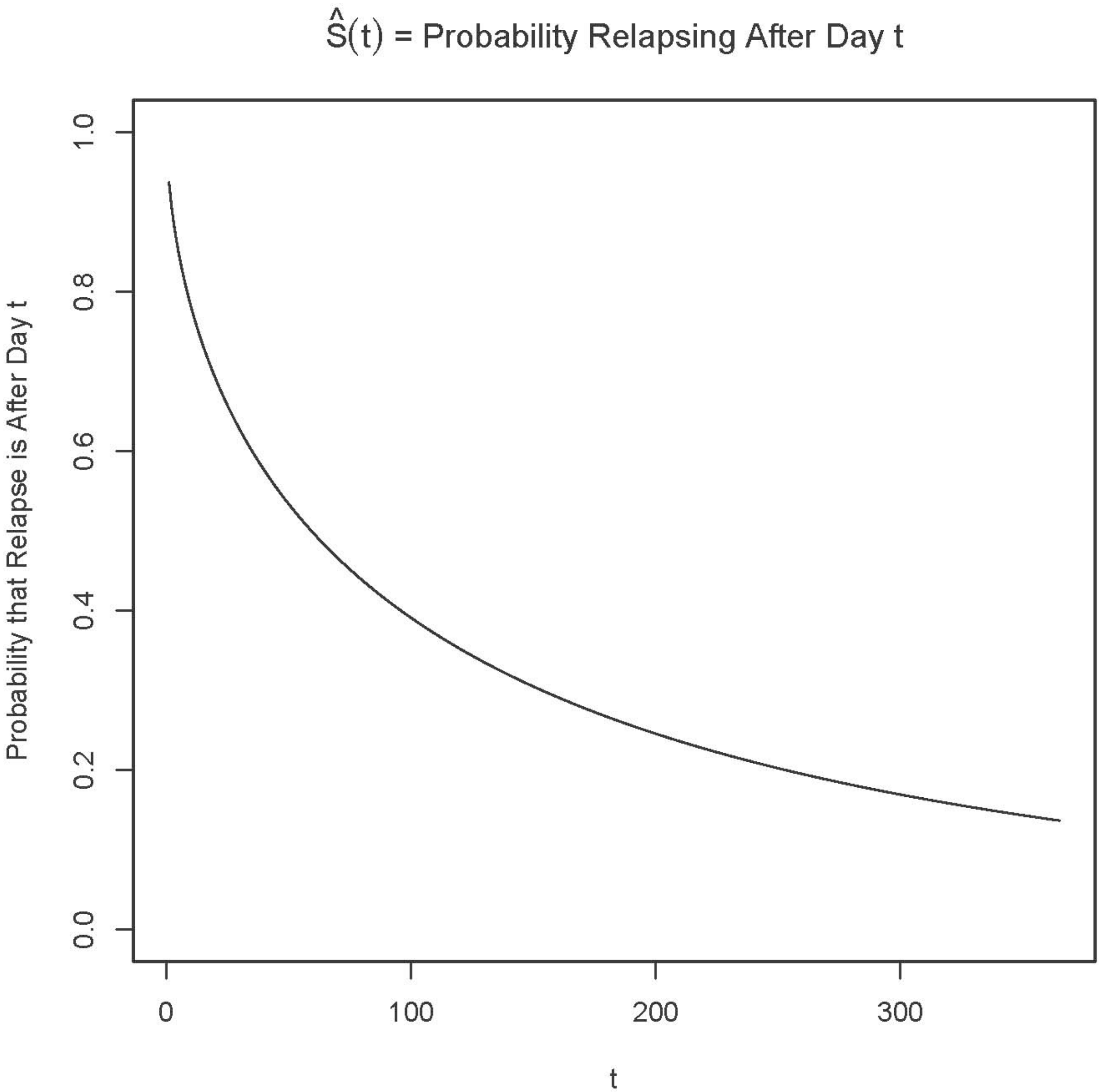
Weibull Regression: $t_i = \exp\{\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1}\} \cdot \epsilon_i^\sigma = e^{\mathbf{x}_i^\top \boldsymbol{\beta}} \epsilon_i^\sigma$, where $\epsilon_1 \sim \exp(1)$.

- $t_i \sim$ Weibull, with $\alpha = 1/\sigma$ and $\lambda = e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}$.
- $E(t_i) = e^{\mathbf{x}_i^\top \boldsymbol{\beta}} \Gamma(\sigma+1)$, $\text{Median}(t_i) = e^{\mathbf{x}_i^\top \boldsymbol{\beta}} \log(2)^\sigma$, $h_i(t) = \frac{1}{\sigma} \exp\{-\frac{1}{\sigma}\mathbf{x}_i^\top \boldsymbol{\beta}\} t^{\frac{1}{\sigma}-1}$.
- $S(t) = \exp\left\{-e^{-\frac{1}{\sigma}\mathbf{x}_i^\top \boldsymbol{\beta}} t^{\frac{1}{\sigma}}\right\}$

```
> thetahat
    (Intercept)         grpCombo             age employmentother
      1.4957374        1.1023048       0.0643414      -1.2880472
   employmentpt
     -1.2122529        0.5454037

> x = c(1,0,50,0,0,0)
> xb = sum(x*thetahat)
>
> # a) The estimated mean
> exp(xb) * gamma(sigmahat+1)
[1] 175.5516
>
> # b) The estimated mean
> exp(xb) * log(2)^sigmahat
[1] 59.17273
>
> # I think the median is preferable to mean because the Weibull distribution
> # is skewed. Also, the predict function for Weibull regression works as expected
> # for medians (but not means).
>
> oldguy = data.frame(grp='patchOnly',age=50,employment='ft')
> predict(full2,newdata=oldguy,type='quantile',p=0.5,se=TRUE)
$fit
       1
59.17273

$se.fit
       1
18.87577

> # The 0.5 quantile is the median. se is from the delta method.
>
> # Estimate and plot S(t) for the old guy.
```

The se of S-hat(t) is straightforward in theory, but messy in practice. G-dot from Mathematica is very ugly. For example, in Wolfram Alpha, try `D[exp( - t^(1/s) exp( -(b0 + 50 b2)/s ),b0]`

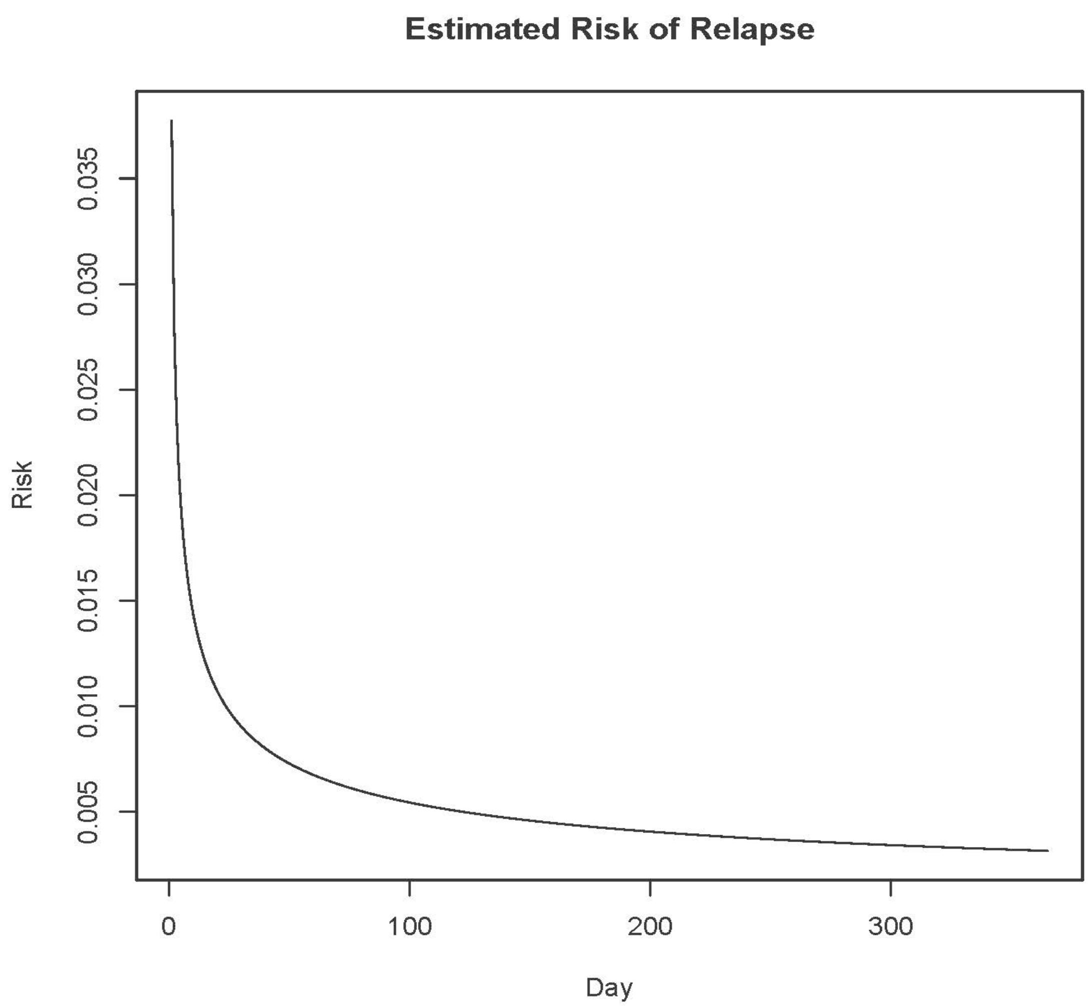
```
> t = 1:365
> Shat = exp( -(exp(-xb/sigmahat)*t^(1/sigmahat)) )
>
> plot(t,Shat,type='l',ylim=c(0,1),xlab='t',ylab='Probability that Relapse is After
Day t')
> tstring = expression(paste(hat(S)(t), " = Probability Relapsing After Day t"))
> title(tstring)
```


$\hat{S}(t)$ = Probability Relapsing After Day t
Probability that Relapse is After Day t
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
t

Plot estimated hazard function for that 50 year old patient who is employed full time and gets the patch-only treatment.

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= \frac{\alpha\lambda(\lambda t)^{\alpha-1}\, e^{-(\lambda t)^{\alpha}}}{e^{-(\lambda t)^{\alpha}}} \\ &= \alpha\lambda^{\alpha} t^{\alpha-1} \\ &= \frac{1}{\sigma} e^{-\frac{1}{\sigma}\mathbf{x}^{\top}\boldsymbol{\beta}}\, t^{\frac{1}{\sigma}-1} \end{aligned}$$

```
> h = 1/sigmahat * exp(-xb/sigmahat) * t^(1/sigmahat - 1)
> plot(t,h,type='l',xlab='Day',ylab='Risk',main='Estimated Risk of Relapse')
```



LaTeX code for the record

```
\noindent
Weibull Regression: $t_i = \exp\{\beta_0+\beta_1x_{i,1} + \ldots + \beta_{p-
1}x_{i,p-1} \} \cdot \epsilon_i^\sigma = e^{\mathbf{x}_i^\top
\boldsymbol{\beta}}\epsilon_i^\sigma$,
where $\epsilon_1 \sim \exp(1)$.
        \begin{itemize}
            \item $t_i \sim$ Weibull, with $\alpha = 1/\sigma$ and  $\lambda =
e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}$.
            \item $E(t_i) = e^{\mathbf{x}_i^\top \boldsymbol{\beta}} \,
\Gamma(\sigma + 1)$,
                    Median($t_i$) = $e^{\mathbf{x}_i^\top \boldsymbol{\beta}} \,
\log(2)^\sigma$,
                    $h_i(t) = \frac{1}{\sigma} \exp\{-\frac{1}
{\sigma}\mathbf{x}_i^\top \boldsymbol{\beta}\}t^{\frac{1}{\sigma}-1}$.
           % \item[] $S(t) = e^{ -\left( e^{-\frac{1}{\sigma}\mathbf{x}_i^\top
\boldsymbol{\beta} } t^{\frac{1}{\sigma}} \right)}$
           \item $S(t) = \exp\left\{ - e^{-\frac{1}{\sigma}\mathbf{x}_i^\top
\boldsymbol{\beta} } t^{\frac{1}{\sigma}} \right\}   $
        \end{itemize}


% Hazard calculation

\begin{eqnarray*}
    h(t) & = & \frac{f(t)}{S(t)} \\
         & = & \frac{\alpha\lambda (\lambda t)^{\alpha-1} \, e^{-(\lambda
t)^\alpha}}
                    {e^{-(\lambda t)^\alpha}} \\
         & = & \alpha\lambda^\alpha t^{\alpha-1} \\
         & = & \frac{1}{\sigma} e^{-\frac{1}
{\sigma}\mathbf{x}^\top\boldsymbol{\beta}}
                      \,         t^{\frac{1}{\sigma} - 1}
\end{eqnarray*}
```