Duration Models #1 v.4

Paul Johnson <pauljohn@ku.edu>

August 8, 2011

1 Let's get some math out of the way

All duration–or "hazard"–models are based on the idea that, as time passes, the object under study can undergo an "event". The event might be death, quitting, getting fired, etc. At a given time, the probability of an event is defined as $f(t)$.
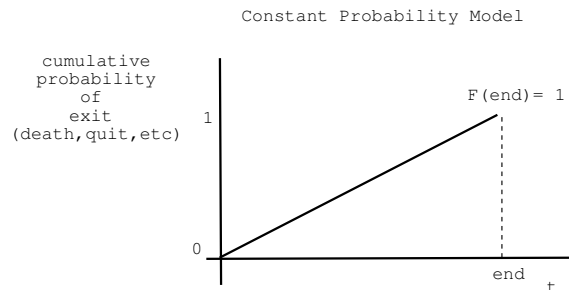
1.1 Event probability

If the probability of an event is constant, then the model would look like this:

Constant Probability Model

1

probability
of
exit
(death,quit,etc)

f(t)= k

0

end

t
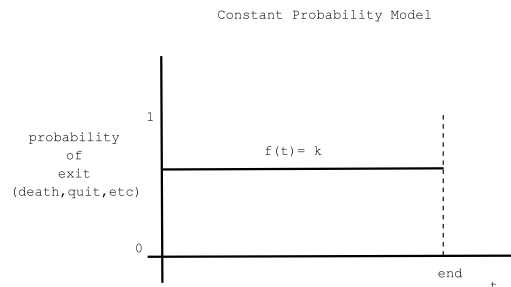
That is a "probability density function," it gives probability over an interval from 0 to end.

The "cumulative distribution function", more commonly called just the **distribution function**, $F(t)$, tells the probability that an observation will have an event time "T" less than a given critical number "t:

Constant Probability Model

cumulative
probability
of
exit
(death,quit,etc)

F(end)= 1

1

0

end

t

$$F(t) = Prob(T < t)$$

1.2 Survival function

The probability that an individual observation will not have an event by time "t" is thus $1 - F(t)$. That quantity is sometimes called the "stayer" function, or the "survival" function, and because both of those words start with "s", it is typically referred to as $S(t)$.

$$S(t) = Prob(T > t) = 1 - F(t) \tag{1}$$

Note that by definition,

$$\frac{\partial S(t)}{\partial t} = -\frac{\partial F(t)}{\partial t} = -f(t) \tag{2}$$

And, because $dln(f)/dx = \frac{1}{f(x)}\frac{df}{dx}$, it is also the case that

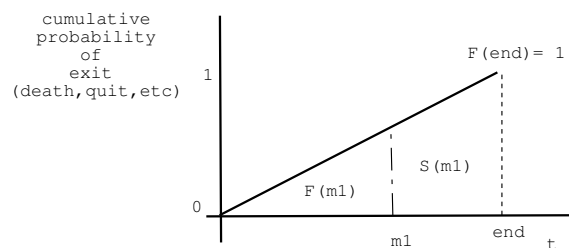$$\frac{\partial lnS(t)}{\partial t} = \frac{1}{S(t)}\frac{\partial S(t)}{\partial t} \tag{3}$$

and if you connect the dots between these, you see

$$\frac{\partial lnS(t)}{\partial t} = -\frac{f(t)}{S(t)} \tag{4}$$

The change in the survivor function is the negative of the hazard function, which is defined next.

1.3 Hazard function

Now, suppose I told you that an individual had survived until time m_1 (m is for milestone). Take a look at the next picture and tell me what you can infer from it:

If I told you somebody lived until m_1, then you could make a more accurate statement about their probability of exit at m_1 than you could before I told you that. If I did not give you that information, you could only say their probability of exit is $f(m_1)$. But now, you know more! You know the probability they will exit at that time has to use $S(m_1)$ as a denominator, since that is the total amount of "risk" left after reaching time m_1.

Given a unit lasts until t, then, the probability that unit will "exit" at that time is

$$h(t) = \frac{f(t)}{S(t)} \tag{5}$$

2

Harrell uses $\lambda(t)$ instead of $h(t)$ in his notation. Hazard is sometimes given exciting names, like the *instantaneous event* or *instantaneous failure rate*.

In light of equation 4, it is also true that:

$$h(t) = -\frac{\partial \ln S(t)}{\partial t} \qquad (6)$$

I tend to think of this as a pure probability exercise. The total probability of an exit at or after t is $S(t)$ and the unconditional chance of exiting at t is $f(t)$, so the conditional chance of exiting is $h(t)$.

If you want a formal proof of equation (5), consult Harrell, p. 395. Hazard is the probability of an "event" that occurs at time T will happen in a really small unit of time between t and $t + \Delta t$, divided by the probability that the event happens after t:

$$h(t) = \frac{lim_{\Delta t \to 0} \frac{Prob(t < T < t + \Delta t)}{\Delta t}}{Prob(t < T)} \qquad (7)$$

Harrel p. 395 shows all the steps.

It is really just a lightly veiled application of Bayes's theorem, come to think of it.

Anyway, the function $h(t)$ is the so-called hazard function, or hazard rate. It has an interpretation that goes like "given a unit survives until t, what is the probability that it will die exactly at t."

Before you go further, stop and note the importance of this:

$$f(t) = h(t)S(t) \qquad (8)$$

1.4   Its intertwined! Hazard function, cumulative hazard, survivor function...

The cumulative hazard function, often denoted as $\Lambda(t)$, is the "accumulated risk" up to time t. Cumulative risk never decreases: it is always either steady or increasing in t.

$$\Lambda(t) = \int_0^t h(u)du \qquad (9)$$

Note, given equation 6, that is:

$$\Lambda(t) = \int_0^t -\frac{\partial \ln S(u)}{\partial u}du \qquad (10)$$

$$\Lambda(t) = -\ln S(t) \qquad (11)$$

$$S(t) = exp(-\Lambda(t)) \qquad (12)$$

The survival curve is thus tightly tied to the accumulated hazard rate.

3

2   No Covariates. Why bother?

2.1   Constant hazard=exponential survival function

I don't know why anybody would want this, but suppose the hazard rate were a constant:

$$h(t) = \lambda$$

That is, no matter how far along one survives, the chance of exiting is fixed. If you think backward from there, then you can use your math skills to assert that

$$f(t) = \lambda exp(-\lambda t) \tag{13}$$

and

$$S(t) = exp(-\lambda t) \tag{14}$$

and

$$\Lambda(t) = \lambda t \tag{15}$$

and

$$S(t) = exp(-\Lambda(t)) \tag{16}$$

Maybe you don't know enough calculus to believe it, but you could learn it some day and see this is right.

In other words, if the hazard is a constant, then the Survivor function is an exponential function.

2.2   Weibull hazard

Suppose you wanted to tell a story in which the hazard rate changes over time. Here would be a good starting point:

$$h(t) = \lambda \alpha t^{\alpha-1} \tag{17}$$

You can see the time-dependence of the hazard depends on α, right?

If α is bigger than 1, then the hazard rate gets bigger, there is "positive duration dependence."

If α is smaller than 1, then the exponent is negative, and the hazard gets smaller, and there is "negative duration dependence".

Because α determines the "shape" it is often called a shape parameter.

The coefficient λ is the "scale" parameter. Not that adjusting λ changes the hazard up and down.

Please note that the exponential model is a "special case" of the Weibull. If in the Weibull we found $\alpha = 1$, then this would degenerate to an exponential model.

If you get on your math hat, you can figure out the f and F and S which imply the previous hazard rate. My notes have

$$S(t) = exp(-\lambda t^{\alpha}) \tag{18}$$

and I know from browsing the internet that

$$f(t) = \alpha\lambda(t\lambda)^{\alpha-1}exp(-(t\lambda)^{\alpha}) \tag{19}$$

And it is as obvious and plain as the nose on your face that

$$F(t) = 1 - \exp\left(-(t\lambda)^{\alpha}\right) \tag{20}$$

Sometimes I see people who use different letters for these, and sometimes they write $\frac{1}{\lambda} = \beta$ and then they replace the λ in 19:

$$f(t) = \frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1}\exp\left(-\left(\frac{t}{\beta}\right)^{\alpha}\right) \tag{21}$$

Watch out if you go digging around for Weibull formulae, because it appears to me that it is against the law for any two authors to us the same letters and symbols.

2.3 Gamma hazard.

Condgon (p. 427) mentions the possibility that the hazard rate might be given by a model that depends on a Gamma variate. Suppose a Gamma distribution $G(\gamma, \lambda)$ with the density

$$f(t) = \frac{\lambda}{\Gamma(\gamma)}(\lambda t)^{\gamma-1}\exp(-\lambda t) \tag{22}$$

The coefficients γ and λ are the shape and scale parameters, respectively.

Recall that the gamma function is just some number you can look up in a book or computer.

The cumulative distribution that goes along with this is somewhat magical as far as I can see. Writing down the formula for F(t) requires the "incomplete gamma function" $\Gamma_x(\gamma)$ and as far as I can see, for our purposes, it is just more distraction.

Notice, if $\gamma = 1$ then this is the same as the exponential distribution.

The gamma model has a hazard function that is monotonic in t. It appears to me after considerable searching that there is no general closed form for the hazard function for all values of the parameters. Nevertheless, supposing $\lambda = 1$:

$$h(t) = \frac{t^{\gamma-1}e^{-t}}{\Gamma(\gamma) - \Gamma_x(\gamma)} \tag{23}$$

We know, in general, that if $\gamma \in (0,1)$ then hazard decreases monotonically, while if $\gamma > 1$ then the hazard increases from 0 to λ as time goes from 0 to ∞.

2.4 Nonparametric survival curves: K-M curves

I'll write something here someday.

5

3 Proportional Hazards: finally some input variables

Donald Cox made the proportional hazards model famous. He pioneered many areas in modern statistics. The "Cox Proportional Hazards" model is a specific strategy for estimating regression coefficients.

Caution: Not all proportional hazards models are CPH models.

3.1 General Proportional Hazard

The proportional hazards model assumes that the "time dependent" hazard is multiplied by the part that depends on the input variables:

$$h(t) = \lambda(t|X_i) = \lambda_0(t) \cdot f(X_i, b)$$

The hazard is a function of time which reflects a "baseline hazard" $\lambda_0(t)$ multiplied against a function of the input variables.

Please note the significance.

> **Hazard separates into two parts**, an individual dependent part and a part that depends only on time.

I like this notation $\lambda(t|X)$ as a way of remembering that hazard is separate from the function $\lambda_0(t)$. It is very common, but not absolutely necessary, to assume

$$f(x, b) = exp(Xb)$$

So the general definition of a proportional hazards model is

$$h(t) = \lambda_0(t) \cdot exp(Xb) \tag{24}$$

Fiddle that around in various ways:

$$h(t) = e^{ln\lambda_0(t) + Xb}$$

So you can think of the hazard rate as being the exponentiated sum of the logged time-related element and the input variables. In other words, if you did have the hazard value as an observed quantity, and you wanted to use it in a regression model, you would need logged time as an input.

To refer to a case i, we might write

$$h_i(t) = \lambda_0(t) \cdot exp(X_i b) \tag{25}$$

Note the premise here is that all cases at time t share a certain amount of hazard, $\lambda_0(t)$ and the there is case-specific customization with $exp(X_i b)$.

4 The Cox Proportional Hazards model: nonParametric approach.

There is a division of opinion on this. Box-Steffensmeier and Jones cite authorities who discourage the use of the CPH model, whereas Harrell seems to be more enthusiastic. Take your pick, apply your diagnostics.

Some people call this a nonparametric approach because we don't end up estimating the baseline hazard at all. In fact, we don't even end up using the precise times to event data. We only use the ranking of the event times. But I prefer to say it is semi parametric, because we do estimate the *b*'s.

4.1 About conditional probability

Suppose the following events might happen: {A,B,C,D,E,F}. Suppose these are independent events, and the probability of each one is given by P(A), P(B), ..., P(F), and, of course, these sum up to 1.0.

Suppose you are told that either A, B, or C happened. What is the probability that the thing which happened was A?

$$P(A|A\,or\,B\,or\,C) = \frac{P(A)}{P(A) + P(B) + P(C)} \tag{26}$$

4.2 Cox's model depends on ordering and conditional probability

Cox's argument was that we should not worry about the function $\lambda_0(t)$. Its not our main focus. We want regression coefficients!

So, how can we make $\lambda_0(t)$ disappear? The event time for a case i is T_i and the *risk Set* at time t is the set of observations ($j \in risk\,Set$) such that the event did not occur, meaning $T_j > t$. Let's figure out the probability that observation 1 will be the first one to have an event. Observe (following the logic in 26):

$$\frac{h_1(t)}{\sum_{i \in risk\,Set} h_i(t)} = \frac{\lambda_0(t) \cdot exp(X_1 b)}{\sum_{i \in risk\,Set} \lambda_0(t) \cdot exp(X_i b)} = \frac{exp(X_1 b)}{\sum_{t < T_i} exp(X_i b)} \tag{27}$$

Whew! $\lambda_0(t)$ disappeared! What we are left with is the probability that case 1 will be first.

Now find a way to justify that: If only one case can fail at time t (a big assumption that haunts CPR applications), then the one which fails will be case 1 is given by the hazard experienced by case 1 divided by the hazard experienced by all of the cases.

Note this is somewhat similar to the probability that $y = 1$ in a logistic regression model, but it is not exactly the same thing. Its not remotely similar, really, because the numerator concerns case i and the denominator concerns cases other than i which have longer survival times.

So we take a data set and sort the observations in order of event time, so $T_1 < T_2 < T_3... < T_N$ and then we figure the conditional probability that each case which does fail will be the one that actually does, as given for case 1 in expression 27.

Each observation in the dataset has a conditional calculation which indicates the probability that it will be next. It will be "next" in the sense that we know which others "already

happened" and we know which others did not happen yet. Now, we know from ordered data where all of the events fall into line, so you can see where this is heading. The model is predicting the ordering of events across the sample.

It's not maximum likelihood because we aren't maximizing the full probability of observing each observation. Instead, we are looking at the probability that the observed order will be replicated by each case considered one at a time.

This approach is called "partial likelihood". Cox showed that maximizing partial likelihood is going to give parameter estimates similar to full maximum likelihood, but partial likelihood is easier to calculate and makes some impossible problems "do-able."

Cox Proportional Hazards (CPH) models are sometimes written with the stipulation that all event times are observed. None of the observations can be censored:

$$partialL(b) = \prod_{i=1}^{N} \frac{exp(X_i b)}{\sum_{t<T_j} exp(X_j b)} \tag{28}$$

$$ln\,partialL(b) = \sum_{i=1}^{N} \left\{ ln\left[exp(X_i b)\right] - ln\left[\sum_{t<T_j} exp(X_j b) \right] \right\} \tag{29}$$

$$= \sum_{i=1}^{N} \left\{ X_i b - ln\left[\sum_{t<T_j} exp(X_j b) \right] \right\} \tag{30}$$

Note the following:

1. There's no constant in a CPH model. If there is a constant in the variable matrix X_i, then the division described in 39 will result in its elimination from the model (division of $exp(b_o)/exp(b_0)$) makes the constant disappear.

2. We CAN deal with ties. As Harrell explains, the "one failure at a time" restriction that justifies this approach has been a source of trouble, but workarounds have been proposed. Breslow's method seems to be the most widely used, but it is perhaps not so accurate (see Harrell, p. 467).

3. Stratified CPH models are possible. If you think the assumption that underlies 39 is too strong, perhaps the problem is that your data includes "separate strata" of observations. If you have 2 collections of observations, and you think they have different baseline hazards, you can specify a model in which the 2 collections have the same b coefficients but the baseline hazards are treated separately.

4. Diagnostics exist. There are ways to find out if the hazards are not truly proportional. That means the "input variables" in $exp(X_i b)$ include some component that depends on t, such as $exp(X_i b + log(t))$ or so forth. If this is the case, then the "explanatory part" of the model is no longer truly proportional to the time-based part.

5 Parametric Hazards models

The term "hazard model" became synonymous with the Cox proportional hazards model, but is is not always necessarily so. The alternative approach is a maximum likelihood approach that

1. requires us to specify a formula for $\lambda_0(t)$

2. allows us to incorporate "censored" observations in our analysis.

The likelihood of observing a particular exit at time t is given by the hazard rate for that value of t times the Survivorfunction. If we observe all cases to completion, then the likelihood is simply found by multiplying the hazard values.

Maximize the log likelihood:

$$lnL = \sum_{i=1}^{N} ln(f(t_i)) \tag{31}$$

The probability that an event will be observed at time t is the hazard rate at time t multiplied by the surivival rate for t: $h(t) \cdot S(t) = \lambda_0(t) \cdot exp(X\beta) \cdot S(t)$. If the event times are not all observed when the study comes to an end, then we have "censored" data. If we have, say, an observation that has not exited, that means it is still a survivor, and the exit time is unknown. If the study ends at t_{end}, the probability of such an observation is $S(t_{end})$. To put the two possibilities into one expression, introduce the dummy variable δ_i which has the value 0 if an observation is "censored" and 1 if it is an observed failure.

$$likelihood\, of\, case\, i: \quad h(t_i)^{\delta_i} S(t_i) \tag{32}$$

If the case is censored, then the h() part gets thrown away because it is raised to the power of 0, and then only the $S()$ part is taken into account.

6 Time-Varying Covariates

Wading through the literature, it seems to me there is not a complete consensus on the incorporation of time-varying covariates. Here is my best understanding of the situation.

6.1 Extend the CPH model

Recall the CPH partial likelihood is:

$$partialL(b) = \prod_{i=1}^{N} \frac{exp(X_i b)}{\sum_{t < T_j} exp(X_j b)} \tag{33}$$

What if the variable X_j is not fixed in time, but varies.

$$partialL(b) = \prod_{i=1}^{N} \frac{exp(X_{it} b)}{\sum_{t < T_j} exp(X_{jt} b)} \tag{34}$$

The numerator is easy to calculate. Instead of a "constant" X_i, we just put in the observed value of X_{it} at time t. This assumes that the hazard depends only on the current setting of the variable X, not past values.

The denominator is superficially more complicated, but really it is just as simple. The hazard depends only on the value of X at the current time, so customize each denominator.

6.2 Parametric Survival: Petersen's approach

Trond Peterson. 1986. Fitting Parametric Survival Models with Time-Dependent Covariates. Applied Statistics. 35(3): 281-288

Begin with all the apparatus of the parametric survival model.

Suppose you take a full (as opposed to partial) likelihood approach. Then each term in the likelihood has to be the probability of having an event at time t_i. But we don't have a way to calculate the Survivor function that takes into account the fact that the input is varying in time. Peterson considers 2 kinds of covariates:

X_{it} varies in time in discrete "jumps"

Z_{it} varies in time continuously

First, just "assume" hazard depends on time-varying variables

$$h(t) = \lambda(t|X_{it}, Z_i(t)) = lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t | T \geq t, X_{it}, Z_i(t)]}{\Delta t} \tag{35}$$

Since X_{it} is observed to change at certain intervals, it means we can "break up" time so that X_{it} is constant within the intervals. Divide time into sections marked of by

$$t_0 = 0 \quad t_0 < t_1 < t_2 ... < t_k \tag{36}$$

Create a sequence of hazard functions for case i at t_1, t_2,..., t_k. These are separate functions which apply in the intervals:

$$\lambda_{i1} = \lambda_1(t|X_{i1}), \lambda_{i2} = \lambda_2(t|X_{i2})..., \lambda_{ik} = \lambda_k(t|X_{it_k}) \tag{37}$$

On each little time slice, we can figure the hazard rate. So, if there were no continuous variables like $Z_i(t)$ involved, we could think of the Survivor function as the probability of surviving a string of hazards.

$$S(t_k|X_{it}) = exp[-\int_0^{t_1} \lambda_1(s|X_{i0})ds] \times exp[-\int_{t_1}^{t_2} \lambda_2(s|X_{it_2})ds]... \times exp[-\int_{t_{k-1}}^{t_k} \lambda_k(s|X_{it_k})] \tag{38}$$

The elements in the likelihood function are exactly like the usual parametric hazard model, with the noncensored cases contributing observed event times contributing the probability of an event at the observed time times the Survival function. The probability of event "at the observed time" is given by one of the λ_{ij} above, and the Survival curve

$$P(t = t_i) = \lambda_{it_i} \cdot S(t_i|X_{it})$$

And a censored case contributes the $S(t_{end\,point})$ value.

Peterson shows how, if you have continuously varying inputs like $Z_i(s)$, then the survivor function 38 is easily written to take those variables into account. Each term is just

$$\lambda(s|X_{it}, Z_i(s))$$

6.3 Counting Process Model

The counting process model is the most general, and most recent, development. I believe that it can eventually replace all of the thing we think of as survival/duration analysis. However, the conceptualization is not quite so easy to understand as proportional hazard.

Counting Process models allow multiple events per unit and they more-or-less automatically incorporate time-varying covariates.

It is difficult to explain this without throwing a lot of notation around. The Appendix 2 in Hosmer & Lemeshow explains some details and refers us to other readings. Personally, I think volume 2 of the S-Plus statistical manual is a good place to start.

Revise the risk set

Let's attack the problem by starting with a generalization of the risk set as used in the CPH model. Recall the risk set was stated as the observations for which the survival time exceeded the current time, $t < T_i$. We can generalize that to allow for the possibility that subjects go in and out of the risk set by defining an indicator variable Y_{it}, which is coded 1 if i is in the risk set at time t, 0 otherwise. If you adopt that idea, the CPH model can be stated:

$$partialL(b) = \prod_{i=1}^{N} \frac{exp(X_i b)}{\sum Y_{it} \cdot exp(X_j b)} \tag{39}$$

If the dependent phenomenon is something that can happen over and over again, then one might have a column of observations Y_{it} that switches back and forth.

Reconceptualize censored events

It is implicit in the previous that we are treating the censored observations–the ones for which the observation period ends before the event is observed–just like observations for which we do have event observations later in the data. Censored cases are no longer thought of as incomplete data (S-plus II, p. 358).

Reconceptualize event

The "counting process" idea begins with the premise that we want to understand the accumulated number of events that a case i experiences up to time t. That is called N_{it}.

If you have a project in which the cases "die", then N_{it} can only take on the values 0 and 1. But you can imagine a model of heart attacks in which there are repeats that don't kill the patient.

The S-plus manual (p. 358) says you should think of an individual has having a set of observations, one for each "interval" on which data is gathered. Each interval observed for a case contributes one row, and in that row we have this information:

s_{ij} "start" time for the interval of risk represented by this row

t_{ij} "stop" time for the interval of risk represented by this row

δ_{ij} "event" indicator variable, 1 if event occurred for observation of case i in time j

x_{ij} covariates that apply in the interval represented by the row

k_{ij} the stratum (cluster) that includes this observation

CPH is easy

Re-organize the data so that there is one row for each unique combination of input values. Make sure each line has a "start" and "stop" time variable.

The Survival object in R is created with a command like
$mySurv < -Surv(start = s, stop = t, event = \delta)$
and that is used on the left hand side of a regression model,
$coxph(mySurv \sim x + y, data = dat, method = "breslow", robust = T)$

7 Frailty (heterogeneity)

Remember when we worked on "count" models with heterogeneity, and we put in a rather simple error term to account for "overdispersion" of observed data?

If you don't remember, I'll wait until you look it up.

Hmm.

Hmm.

Now suppose the duration model is based on the Weibull distribution:

$$h(t_i) = \alpha t_i^{\alpha-1} \exp(Xb + \theta_i)$$

The beauty of the exponential functional form is that sums inside the brackets are easily brought out of the brackets, as in:

$$h(t_i) = \alpha t_i^{\alpha-1} \exp(Xb) \exp(\theta_i)$$

And if we want to make θ_i appear "all by itself" as a multiplier, we just use the sleight of hand of redefining the original $\theta = ln(\delta)$ and so this hazard rate is just the original one multiplied by the random variable δ.

So we want to insert a variable δ that has an expected value of 1, which means it "doesn't matter" on average, but some cases are higher and some are lower. One way to achieve that is to use a Gamma distribution for δ because Gamma(1,1) has a mean of 1 and a variance of 1.

8 It ain't necessarily so....

8.1 Nonproportional hazards are possible

Not all hazard models incorporate input variables with hazards in a proportional way. Congdon (p. 427) refers to the logistic hazard model:

$$h(t) = \frac{1}{\sigma \left[\frac{e^{t-Xb}}{1+e^{t-Xb}} \right]}$$

8.2   Nonmonotonic hazards are possible.

The log-logistic model has

$$h(t) = \alpha(\frac{t}{\theta})^{\alpha-1} / \left[\theta + t^{\alpha}\theta^{1-\alpha}\right]$$