SOCIETY OF ACTUARIES

Article from

Predictive Analytics and Futurism Section Newsletter

June 2020

Kinky Business

Nonlinear Relationships in Linear Models

By Nick Hanewinckel

Most of those who build predictive models know the power of the generalized linear model (GLM). These models allow for nonlinearities between regressors and response via their link function. However, the underlying relationship between the regressors and the transformed response is still forced to be linear.

A common technique to overcome this forced linearity is engineering predictors—especially the popular technique of building splines. However, the decision of how to engineer these predictors is largely in the hands of the modeler and can be difficult to optimize.

Fortunately, there are models that use supervised methods to find the best nonlinear shape for the predictors. Two of these are generalized additive models (GAM) and multivariate adaptive regression splines (Earth).

This article will explore the theoretical background of the two and compare their results in a simulated mortality example. Readers are encouraged to research the technical underpinnings of their favorite model, as this article is too short to cover their full depth.

GAM MODELS

A GAM model is one where the predictor depends linearly on unknown "smooth functions," often referred to as just "smooths." Splines are just one particular type of univariate smooth function. This model seeks to estimate the optimal smooth functions in a supervised way, as well as to estimate the coefficients. These smooths can be used in the GAM atop a variety of model structures including, but certainly not limited to, the GLM family.

Given that there are a wide range of theoretical smooth functions, GAM is a very broad model class (as opposed to Earth, which only has one specific smooth function, as we will see). Smooths can be specified as splines (as described above), tensor products (multivariate interactions), and may be further partitioned by factor variables.

R's *mgcv* package and Python's *InterpretML* are popular for their respective languages. However, even these two packages differ in their approaches. With the wide sea of GAM models, users should research available packages and the approach used for their particular underlying model type.

The fitting of the smooth functions themselves take on a variety of techniques. In the packages above, *mgcv* uses a rank-reduced framework and *InterpretML* uses boosting.

In addition, GAM allows extension to a variety of model families outside those available to a GLM. The most notable from an actuarial perspective are likely the Tweedie distribution (P&C—frequency/severity) and Cox Proportional Hazards (Life—Survival).

Due to the complexity and open-ended nature, these models may take a relatively long time to fit. Fitting algorithms may vary due to the model family underlying the GAM (or even the particular GAM package used). Users may want to think about ways to boost efficiency. Besides running in a parallel process, adjusting parameters and simplifying formulae may lead to reduced runtime.

> GAM allows extension to a variety of model families outside those available to a GLM.

EARTH MODELS

You may be wondering why we refer to these models as "Earth" when the acronym for Multivariate Adaptive Regression Splines does not spell the word "Earth." This is because MARS is a trademarked model licensed to Salford Systems. The open-source variety (including R package) is often called *earth*.

Earth's model takes the form:

$$\widehat{f(x)} = \sum_{i=1}^{k} c_i B_i(x)$$

Conceptually, we can think of Earth models as those that find the knots in our prospective splines (Basis Functions—Bi) in a supervised manner. To be able to do this efficiently while also solving for coefficients (ci), Earth uses special splines referred to as "hinge" functions. For simplicity, let's just say that these are hockey stick or ramp shaped. Their products (interactions) can be nonlinear (as opposed to just piecewise-linear).

While we will not go through the minute details of the fitting algorithm, suffice it to say that Earth fits in a vastly different manner to a GAM. The nature of the hinge functions gives Earth the ability to quickly fit the hinges with a least-squares update technique.

A COMPARISON

As mentioned, GAM encompasses a wide range of models making a direct Earth/GAM comparison tricky. However, there are some general ideas that are important, especially in the context of the popular R packages (Earth, mgcv).

- GAM models have a wider variety of smooth functions (including multi-demensional tensor product interactions), hence they may take significantly longer to fit. In the use case below, Earth models took ~2–5 minutes to fit (depending on degree) whereas GAM took over 20.
- Earth models will evaluate interactions up to their "degree" argument; GAM requires interactions to be explicit (by tensor product of continuous variables or using the "by" argument for a factor).
- While GAM has a wider family of models to fit, some (like Tweedie) are available to Earth (often by also using the mgcv package in R). Without the mgcv package, only the GLM family is available to Earth.
- Both functions have a variety of tuning/penalty parameters to explore, which are too numerous to cover here in depth.

USE CASE

To see these model forms in action, I have set up a brief use case in the Life Insurance context. There are a variety of actuarial uses, but mortality often works well with a poisson model, which are simple for both model families here. I have written this in R, though as mentioned, both model forms are available in Python.

An important note—the intent of this use case is to provide a simple example to show functionality and anticipate next steps. This does not represent an ideal model or model-fitting procedure. To say this another way: for now, we care more about how the *smooths* fit vs. how the *model* fits.

Data for this comes from the publicly-available ILEC data set: *https://www.soa.org/resources/research-reports/2019/2009-2015-individual-life-mortality/*

I have made this code available on github: *https://github.com/hanewinckel/KinkyBusiness*

DATA PREPARATION

We read-in the large ILEC dataset using the data.table package for efficiency. To make things even faster, we include only term policies within their level term period that have a "preferred class" designation and a valid exposure. We restrict this to ages 35–75. Next, we engineer a simple "class proportion" to express risk class relative to the number of classes. Then, we make a log-Exposure to fit in the poisson framework. Finally, we hold out the most recent calendar year for a test set.

Users playing along at home may want to tweak my code even further to reduce runtime and/or satisfy memory constraints.

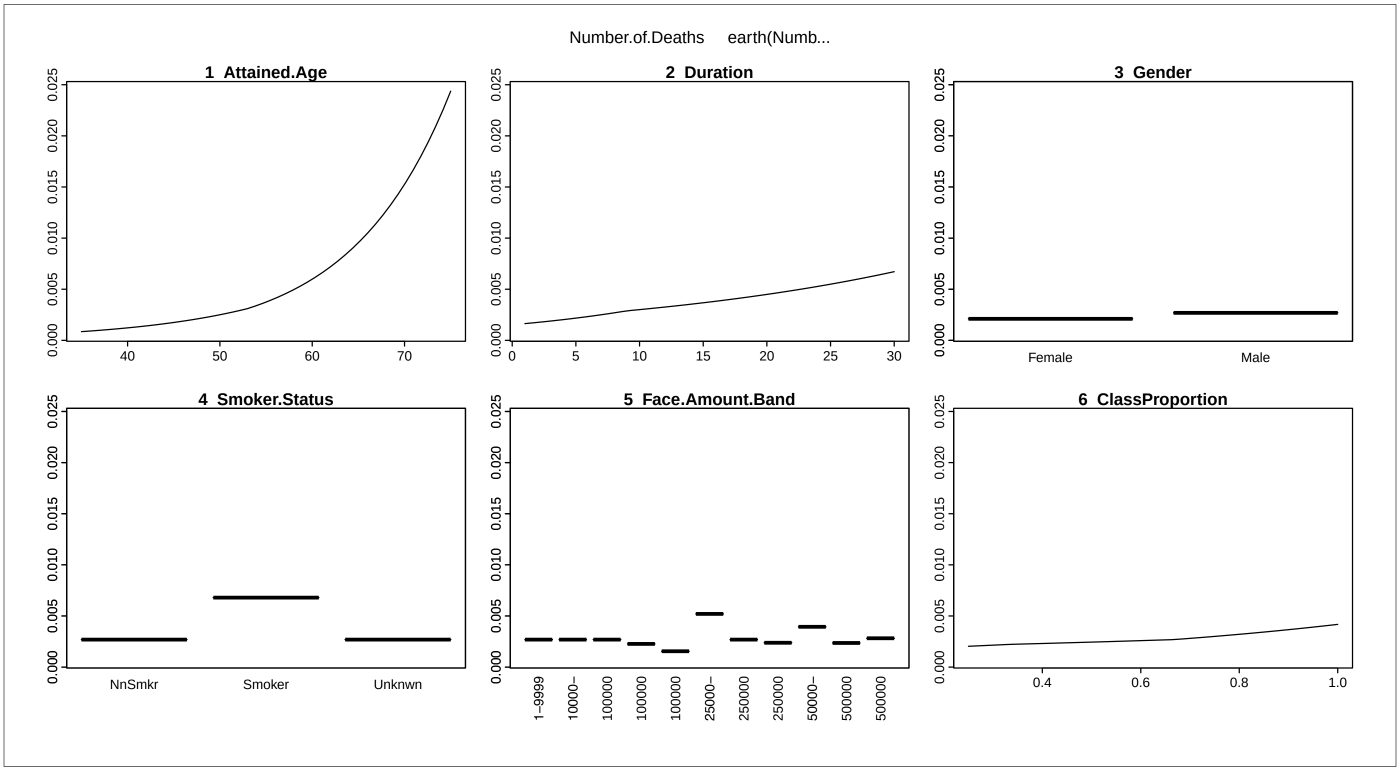
FITTING EARTH MODELS

We fit two relatively simple Earth models. We don't have to specify any smooth functions ourselves (splines) and we leave the model to find statistically meaningful interactions. Our two models differ only in their degree: modelEarth1 is degree 1 (no interactions), and modelEarth2 is degree 2 (considers two-way interactions).

Degree One Model

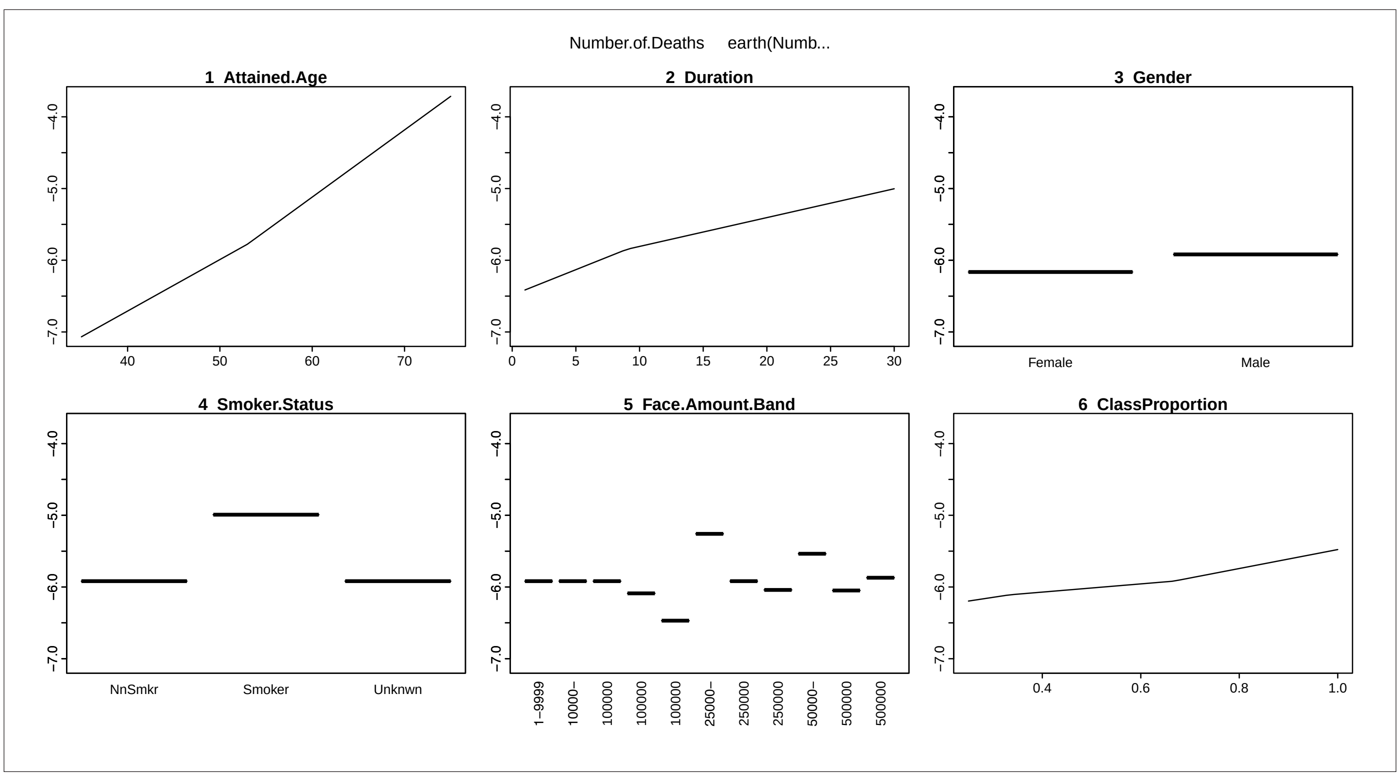
Let's look at the trademark hockey stick kinks that the rigid "hinge" function makes in this simple one-variable case (See Fig. 1).

Figure 1
Degree One Model



Now wait—that doesn't look like any hockey stick I've ever seen! But let's remember: one of the neat tricks of Earth is that the response has had a GLM (poisson-log) transformation. Let's plot these when applying the inverse (See Fig. 2).

Figure 2
GLM (Poisson-log) Transformation

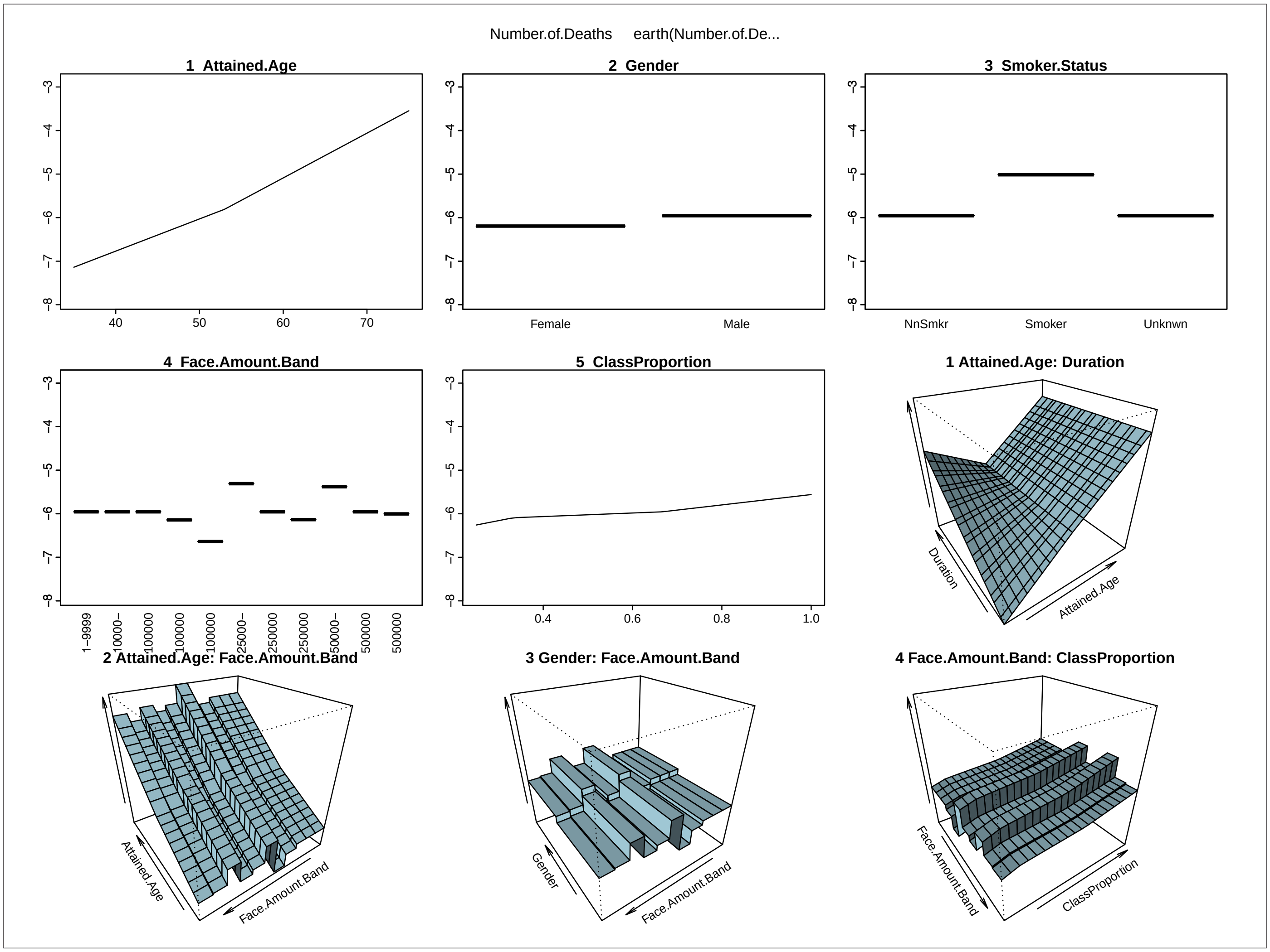


So, now they are more hockey stickish and we also see how subtle this smooth can be (particularly with multiple hinge points).

DEGREE TWO MODEL

Now we change nothing except the argument degree=2. This model fits better (trust me), which I only point out so that we can free our minds to look at the smooths themselves (with the inverse transform already done) (See Fig. 3)

Figure 3
Degree=2



Now we can see how the smooths are formed under interactions (1 Attained.Age:Duration) with factors (4 Face.Amount.Band:-ClassProportion).

FITTING A GAM MODEL

GAM requires a bit more thinking to write the call to the model function. Fortunately, Earth has shown us which interactions make for promising smooths.

I can see, for example, that Attained Age and duration have important interactions. So, I set up the regression equation with `te`(…) which is a full tensor-product, including the original regressors. It is analogous to saying `AttainedAge*Duration` in more familiar function forms. There is also a `ti`() smooth which only contains the interaction that I could have used if AttainedAge and Duration had already been stated in the formula. This is analogous to `AttainedAge:Duration`. I use the "by" argument to calculate these smooths distinctly by gender. This accounts for the different mortality behavior by gender.

Since a more complex model would take (far) longer to run, I trusted that the above multivariate smooth would account for most of the gender effects. For my other numerical predictor (ClassProportion), I simply used `s`() for spline. I declined to calculate this by FaceAmountBand, even though the chart above indicates that is a meaningful interaction. This was mainly done to speed the model and prevent overfit. But you can bet that in a "real" application, we would want to test this.

GAM SMOOTHS

The smooths for GAM are much more open-ended than for Earth. Let's see how the particular smooths and interactions look. We should expect them to look much less "sharp" than an Earth Model (or continuously differentiable, if you like to be technically accurate) (See Fig. 4 and 5).

Figure 4
Smooth for GAM

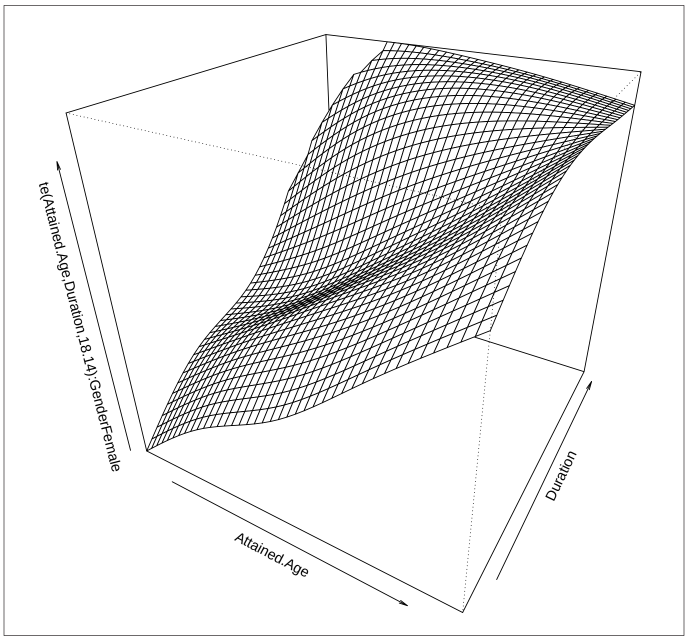


Figure 5
Example of the 2d Age/Duration Smooth Calculated by Gender (Female Shown)

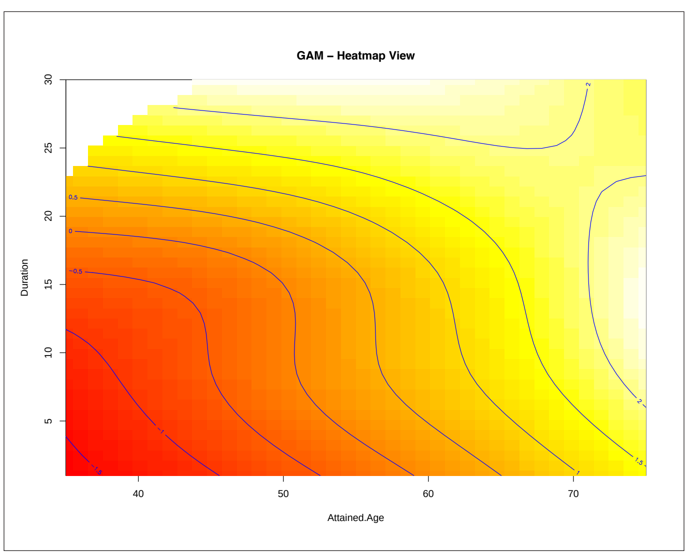
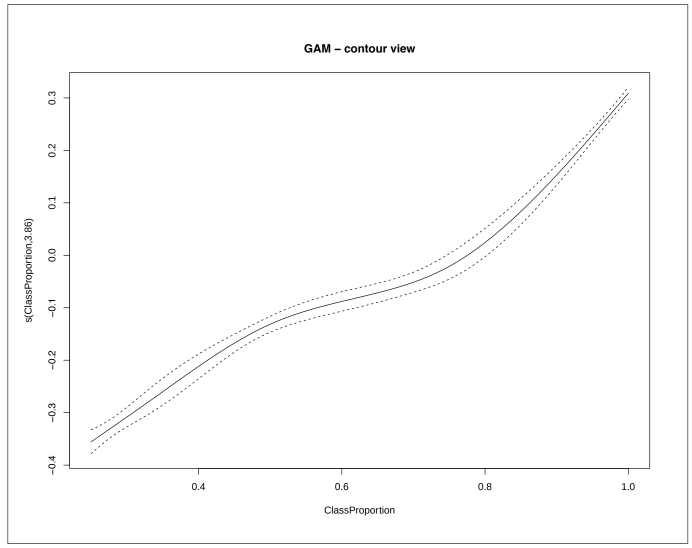


Figure 6
Example of the 1-d Smooth (Spline) for ClassProportion



MODEL COMPARISON

While these models weren't designed to be optimal, it is interesting to note the results given the smooth functions we have just examined (See Fig. 7).

Figure 7
Results of Smooth Functions

testSet	ModelEarth1ExpectedDth	Earth1AE	ModelEarth2ExpectedDth	Earth2AE	ModelGAMEExpectedDth	GAMAE
FALSE	61,566.00	100.00%	61,566.00	100.00%	61,566.00	100.00%
TRUE	22,617.32	96.11%	21,852.65	99.47%	22,437.28	96.88%

From this, we can see that the degree-2 Earth model performed substantially better than a single dimension Earth model. This is to be expected as modeling interactions is one of the benefits of moving to this model type. The 1-d model may suffer from underfit given that its 2-d counterpart was not over-specified.

We should be careful not to draw the conclusion that the GAM did "poorly" just because its holdout A/E was closer to Earth1 than Earth2. There was little to no optimization on this model, and it was partly modified to run quickly. Rather, both models should be experimented with. We'd want to explore most meaningful variables (perhaps taking a look at Earth's ranking of significant smooths), but in addition, we'd want to look at the penalties and other parameters available to determine the smooths. Don't forget—figuring out *which* smooth functions to use is part of GAM's job; Earth always uses hinge functions.

CONCLUSION

Both packages are great tools for modelers to have in their arsenal. With so many potential uses, it would be impossible to universally prefer one over the other. I invite all modelers to play around and see what meets their actuarial needs. Both models are great tools to have, but in the words of LeVar Burton, “you don’t have to take *my* word for it!” ■

Nick Hanewinckel, FSA, CERA, is AVP and actuary for Hannover Life Reassurance Company of America. He can be contacted at *nick.hanewinckel@hlramerica.com*.