Smoothing Spline ANOVA Models II. Variable Selection and Model Building via Likelihood Basis Pursuit

Hao Helen Zhang (first author), with
Grace Wahba, Yi Lin, Meta Voelker, Michael Ferris,
Ronald Klein, and Barbara Klein

http://www4.stat.ncsu.edu/~hzhang2
http://www.stat.wisc.edu/~wahba
http://www.stat.wisc.edu/~yilin
http://www.cs.wisc.edu/~ferris
http://www.ssc.wisc.edu/aging/kleinr.htm
http://www.ssc.wisc.edu/aging/kleinb.htm

Joint Statistical Meetings
San Francisco, CA
August 6, 2003

Abstract

We describe Likelihood Basis Pursuit, a nonparametric method for variable selection and model building, based on merging ideas from Lasso and Basis Pursuit works and from smoothing spline ANOVA models. An application to nonparametric variable selection for risk factor modeling in the Wisconsin Epidemiological Study of Diabetic Retinopathy is described.

Although there are many approaches to variable and model selection, we believe that this one has some novel and useful aspects.

References

[ZWL02]H. Zhang, G. Wahba, Y. Lin, M. Voelker, M. Ferris, R. Klein, and B. Klein. Variable selection and model building via likelihood basis pursuit. Technical Report 1059, UW-Madison Statistics Department, 2002, under review.[on web]

[CDS98]S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20:33–61, 1998.

[T96]R. J. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, B*, 58:267–288, 1996.

[F98]W. J. Fu. Penalized regression: the bridge versus the LASSO. *Journal of Computational and Graphical Statistics*, 7:397–416, 1998.

References (cont.)

[WWGKK95]G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Ann. Statist.*, 23:1865–1895, 1995.

[W90] G. Wahba. Spline models for observational data, SIAM, 1990.

OUTLINE

1. Motivation: The WESDR data.

2. Likelihood Basis Pursuit for the WESDR data.

3. What is Likelihood Basis Pursuit?

4. Why do l_1 penalties give sparser solutions?

5. Building an overcomplete set of basis functions for the Likelihood Basis Pursuit, from a smoothing spline ANOVA model.

6. Computations.

7. The importance measure for the model terms.

8. Choosing the importance threshold.

9. Back to the WESDR results.

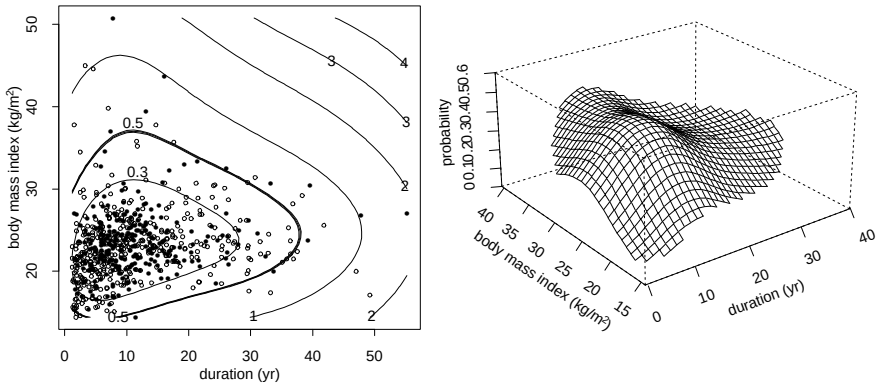
10. Closing remarks.

♣♣ 1. Motivation: WESDR (Wisconsin Epidemiological Study of Diabetic Retinopathy).

WESDR is an ongoing epidemiological study of a cohort of patients receiving their medical care in an 11-county area in southern Wisconsin. Baseline exam, 1980, with four, ten, fourteen and twenty year followups. (refs in [ZWL02]). [WWGKK95] built a smoothing spline ANOVA model for four year risk of progression of diabetic retinopathy from baseline, as a function of three risk factors. We started out with twenty possible risk factors, narrowed it down to four variables by very laborious means, mostly repeated applications of parametric and nonparametric regression analyses of small groups of variables at a time, and existing analyses by others. Finally, three variables were selected as apparently the most important, by ad hoc means. It was evident that it would be highly desireable to have a model selection procedure that could simultaneously select important variables/components of a spline ANOVA model. Such a procedure has been obtained in [Z02][ZWL02], and is today's topic.

♣♣ 1. Motivation: WESDR (Wisconsin Epidemiological Study of Diabetic Retinopathy)(cont).

[WWGKK95] looked for the four year risk of progression of diabetic retinopathy from baseline at a cohort of (selected) $n = 669$ younger onset subjects. Let $p(t)$ be the probability of progression for a subject with risk factor vector t and $f = log[p/(1-p)]$. The model fitted was $f(t) = f(\texttt{dur}, \texttt{gly}, \texttt{bmi}) = \mu + f_1(\texttt{dur}) + a_2 \cdot \texttt{gly} + f_3(\texttt{bmi}) + f_{13}(\texttt{dur}, \texttt{bmi})$ where `dur` = duration of diabetes, `gly` = glycosylated hemoglobin, and `bmi` = body mass index.



(Right: `probability` plotted against `bmi` and `dur`; `gly` at median.) (Note: model is not monotonic in `dur`) From [WWGKK95]

Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR)

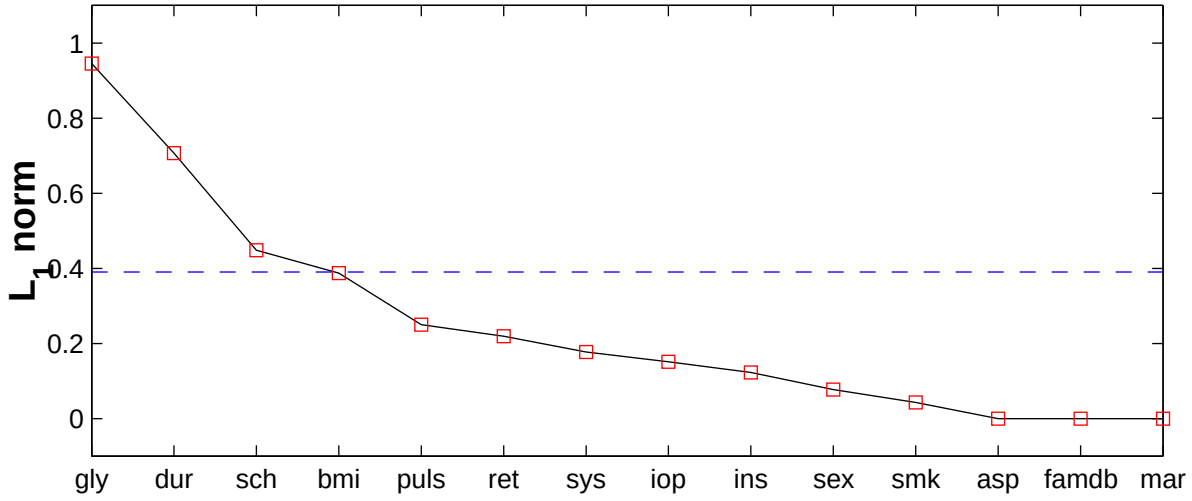
- Continuous covariates:

X_1:	(dur)	duration of diabetes at the time of baseline examination, years
X_2:	(gly)	glycosylated hemoglobin, a measure of hyperglycemia, %
X_3:	(bmi)	body mass index, kg/m^2
X_4:	(sys)	systolic blood pressure, $mmHg$
X_5:	(ret)	retinopathy level
X_6:	$(pulse)$	pulse rate, count for 30 seconds
X_7:	(ins)	insulin dose, kg/day
X_8:	(sch)	years of school completed
X_9:	(iop)	intraocular pressure, $mmHg$

- Categorical covariates:

Z_1:	(smk)	smoking status	(0 = no, 1 = any)
Z_2:	(sex)	gender	(0 = female, 1 = male)
Z_3:	(asp)	use of at least one aspirin for at least three months while diabetic	(0 = no, 1 = yes)
Z_4:	$(famdb)$	family history of diabetes	(0 = none, 1 = yes)
Z_5:	(mar)	marital status	(0 = no, 1 = yes/ever)

♣♣ 2. WESDR: The Likelihood Basis Pursuit result for the WESDR data.



L_1 norm scores for the WESDR main effects model, from a Likelihood Basis Pursuit analysis. The method selected `gly`, `dur`, `sch` and `bmi`, in that order, as important variables in a smoothing spline ANOVA main effects model (smoothing spline additive model).

Note: `sch` is *years of schooling completed*. It came up in some of the analyses in [WWGKK95] but was not considered a direct cause of disease and was not further considered.

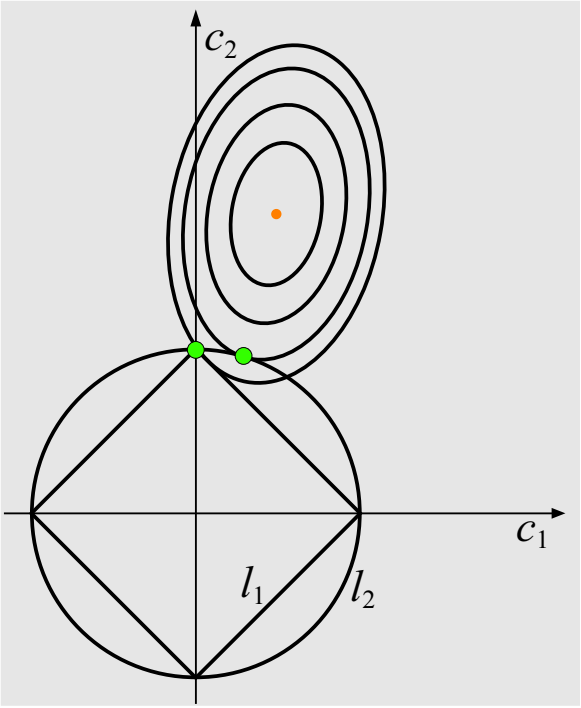
Next: How it was done.

♣♣ 3. What is Likelihood Basis Pursuit?

Likelihood Basis Pursuit combines ideas from the LASSO [T96][F98], basis pursuit [CDS98], and smoothing spline ANOVA models to generate an overcomplete set of basis functions, which are then used in a penalized likelihood variational problem with l_1 penalties. Basis pursuit uses l_1 penalties, instead of quadratic penalties, to obtain solutions which are relatively sparse in the number of basis functions with non-0 coefficients.

Why do l_1 penalties result in sparser solutions than quatratic penalties?

♣♣ 4. Why do l_1 penalties give sparser solutions?



Circle: $\sum_{j=1}^{N} c_j^2 \leq M$, diamond: $\sum_{j=1}^{N} |c_j| \leq M$.
Ellipses: contours of constant $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{N} x_{ij}c_j)^2$.
Find c to minimize:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{N} x_{ij}c_j)^2$$

subject to $\sum_{j=1}^{N} |c_j|^p \leq M$. Green dots: minimizers for (l. to r.) $p = 1$ and $p = 2$. Note that $c_1 = 0$ for $p = 1$.

♣♣ 4. Why do l_1 penalties give sparser solutions ? (cont.)

The problem: Find c to minimize:

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{N} x_{ij}c_j)^2$$

subject to $\sum_{j=1}^{N} |c_j|^p \leq M$ is generally equivalent to the problem: Find c to min

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{N} x_{ij}c_j)^2 + \lambda \sum_{j=1}^{N} |c_j|^p$$

for some $\lambda = \lambda(M)$.

♣♣ 5. Building an overcomplete set of basis functions for the Likelihood Basis Pursuit, from a smoothing spline ANOVA model.

1. Main effects model, continuous variables.

First: the usual penalized likelihood: Let $l!k_l(u)$ be the lth Bernoulli polynomial, and let $K(u,v) = k_2(u)k_2(v) - k_4(|u-v|), u, v \in [0,1]$ (spline kernel [W90]). Let $x = (x^1, \cdots, x^d)$, and the observations be $\{y_i, x_i\}$ where $x_i = (x_i^1, \cdots, x_i^d), i = 1, \cdots, n$.

The solution to the problem: Find f (in an appropriate space) of the form $f(x) = \mu + \sum_{\alpha=1}^{d} f_\alpha(x^\alpha)$ to min

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{C}(y_i, f(x_i)) + \sum_{\alpha=1}^{d} \theta_\alpha^{-1} \int (f_\alpha{}'')^2$$

has a representation

$$f(x) = \mu + \sum_{\alpha=1}^{d} b_\alpha k_1(x_\alpha) + \sum_{i=1}^{n} c_i \left(\sum_{\alpha=1}^{d} \theta_\alpha K(x^\alpha, x_i^\alpha) \right)$$

♣♣♣ 5. Building an overcomplete set of basis functions for the Likelihood Basis Pursuit, from smoothing spline ANOVA model (cont.).

1. Main effects model, continuous variables (cont.).

Since generally an excellent approximation to the solution to the variational problem can be obtained with fewer basis functions, a selected subset, $x_{i_1}, \cdots, x_{i_N}$, of the x_i can be used to generate the solution. Thus

$$\sum_{i=1}^{n} c_i \left(\sum_{\alpha=1}^{d} \theta_\alpha K(x^\alpha, x_i^\alpha) \right)$$

is replaced by

$$\sum_{j*=1}^{N} c_{j*} \left(\sum_{\alpha=1}^{d} \theta_\alpha K(x^\alpha, x_{j*}^\alpha) \right)$$

where $\{x_{j*} = (x_{j*}^1, \cdots, x_{j*}^d),\ j* = 1, \cdots, N\}$ is the selected subset of the x_i.

♣♣ 5. Building an overcomplete set of basis functions...(cont.).

1. Main effects model, continuous variables (cont.). This suggests the overcomplete set of $1 + d + dN$ basis functions:

$$\{1,\ b^{\alpha}(x) \equiv k_1(x^{\alpha}),\ B^{\alpha}_{j*}(x) \equiv K(x^{\alpha}, x^{\alpha}_{j*})\}.$$

for $\alpha = 1, \cdots, d, j* = 1, \cdots, N$

The basis pursuit variational problem is to find $f(x)$,

$$f(x) = \mu + \sum_{\alpha=1}^{d} b_{\alpha} b^{\alpha}(x) + \sum_{\alpha=1}^{d} \sum_{j*=1}^{N} c_{\alpha j*} B^{\alpha}_{j*}(x)$$

to minimize

$$\frac{1}{n} \sum_{i=1}^{n} \mathcal{C}(y_i, f(x_i)) + \lambda_{\pi} (\sum_{\alpha=1}^{d} |b_{\alpha}|) + \lambda_s (\sum_{\alpha=1}^{d} \sum_{j*=1}^{N} |c_{\alpha j*}|)$$

♣♣ 5. Building an overcomplete set of basis functions...(cont.).

2. Two factor interaction model, continuous variables.

Add tensor products of the previous basis functions: (Similar rationale as before-from the two-factor interaction smoothing spline ANOVA model.)

$$\{span\ \ b^{\alpha}b^{\beta}, \alpha < \beta; b^{\alpha}B^{\beta}_{j*}, \alpha \neq \beta, j* = 1, \cdots, N;$$

$$B^{\alpha}_{j*}B^{\beta}_{j*}, \alpha < \beta, j* = 1, \cdots, N\}$$

to the previous set of basis functions.

♣♣ 5. Building an overcomplete set of basis functions...(cont.).

2. Two factor interaction model (more) ..

Recalling the definitions

$$\{b^{\alpha}(x) \equiv k_1(x^{\alpha}),\ B^{\alpha}_{j*}(x) \equiv K(x^{\alpha}, x^{\alpha}_{j*})\}.\}$$

f is now of the form:

$$f(x) = \mu + \sum_{\alpha=1}^{d} b_{\alpha} b^{\alpha}(x) + \sum_{\alpha=1}^{d} \sum_{j*=1}^{N} c_{\alpha j*} B^{\alpha}_{j*}(x)$$

$$+\lambda_{\pi\pi} \sum_{\alpha<\beta} b_{\alpha\beta} b^{\alpha}(x) b^{\beta}(x)$$

$$+\lambda_{\pi s} \sum_{\alpha\neq\beta} \sum_{j*=1}^{N} c^{\pi s}_{\alpha\beta j*} b^{\alpha}(x) B^{\beta}(x)$$

$$+\lambda_{ss} \sum_{\alpha<\beta} \sum_{j*=1}^{N} c^{ss}_{\alpha\beta j*} B^{\alpha}(x) B^{\beta}(x).$$

♣♣ 5. Building an overcomplete set of basis functions...(cont.).

2. Two factor interaction model (more)...

...and now the variational problem becomes: Find f to minimize;

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{C}(y_i, f(x_i)) + \lambda_\pi(\sum_{\alpha=1}^{d}|b_\alpha|) + \lambda_s(\sum_{\alpha=1}^{d}\sum_{j*=1}^{N}|c_{\alpha j*}|)$$

$$+\lambda_{\pi\pi}(\sum_{\alpha<\beta}|b_{\alpha\beta}|) + \lambda_{\pi s}(\sum_{\alpha\neq\beta}\sum_{j*=1}^{N}|c^{\pi s}_{\alpha\beta j*}|)$$

$$+\lambda_{ss}(\sum_{\alpha<\beta}\sum_{j*=1}^{N}|c^{ss}_{\alpha\beta j*}|).$$

♣♣ 5. Building an overcomplete set of basis functions...(cont.).

2. Two factor interaction model, continuous variables. There are now five smoothing parameters in this setup:

$$
\begin{array}{cc}
\lambda_\pi & parametric\ main\ effects \\
\lambda_s & smooth\ main\ effects \\
\lambda_{\pi\pi} & parametric - parametric\ interactions \\
\lambda_{\pi s} & parametric - smooth\ interactions \\
\lambda_{ss} & smooth - smooth\ interactions
\end{array}
$$

Other variants of smoothing parameters are possible, as well as other reproducing kernels.

Categorical variables are added to the model in a straight-forward way, but now there will be more combinations of possible interaction terms.

♣♣ 6. Computations

The model is fitted via a tailored mathematical programming algorithm called slice modeling, see [ZWL02], originally due to Michael Ferris. MINOS was the underlying optimization code.

Given Bernoulli data, the smoothing parameters are fitted via GACV (Generalized Approximate Cross Validation) for basis pursuit models. The GACV looks essentially the same as the GACV for Bernoulli data with the usual RKHS penalty functional [XW96], and the randomized trace technique is used to compute it.

♣♣ 7. The importance measure for the model terms.

We have adopted the empirical L_1 norm to assess the relative importance of the various terms. (Since the smoothing spline ANOVA basis functions all average to 0 this is makes sense). The empirical L_1 norms of the main effects f_α and the two-factor interactions $f_{\alpha\beta}$ are defined as

$$\begin{aligned} L_1(f_\alpha) &= \tfrac{1}{n}\textstyle\sum_{i=1}^n |f_\alpha(x_i^\alpha)| \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^n |b_\alpha k_1(x_i^\alpha) + \sum_{j=1}^N c_{\alpha j*} K_1(x_i^\alpha, x_{j*}^\alpha)| \end{aligned}$$

and

$$\begin{aligned} L_1(f_{\alpha\beta}) &= \tfrac{1}{n}\textstyle\sum_{i=1}^n |f_{\alpha\beta}(x_i^\alpha, x_i^\beta)| \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^n |b_{\alpha\beta} k_1(x_i^\alpha) k_1(x_i^\beta) \\ &+ \textstyle\sum_{j=1}^N c^{\pi s}_{\alpha\beta j*} K_1(x_i^\alpha, x_{j*}^\alpha) k_1(x_i^\beta) k_1(x_{j*}^\beta) \\ &+ \textstyle\sum_{j=1}^N c^{\pi s}_{\beta\alpha j*} K_1(x_i^\beta, x_{j*}^\beta) k_1(x_i^\alpha) k_1(x_{j*}^\alpha) \\ &+ \textstyle\sum_{j=1}^N c^{ss}_{\alpha\beta j*} K_1(x^\alpha, x_{j*}^\alpha) K_1(x^\beta, x_{j*}^\beta)|. \end{aligned}$$

The empirical L_2 norm gives essentially the same results.

♣♣ 8. Choosing the importance threshold.

A Monte Carlo bootstrap test is used to sequentially add terms to the model, beginning in the order of the L_1 norms, until the next term which is a candidate for inclusion fails the importance test.

For $\eta = 0$ until end, do

1. Posit a null model which includes the first η terms in rank order of their L_1 norms. Fit this model, call it f^η. Here we want to test whether the $(\eta + 1)$st term is important.

2. Generate T independent Monte Carlo bootstrap data sets $\{y_i^{\eta t}, x_i\}, t = 1, \cdots, T$ from f^η, using the original design $x_1, \cdots, x_n$.

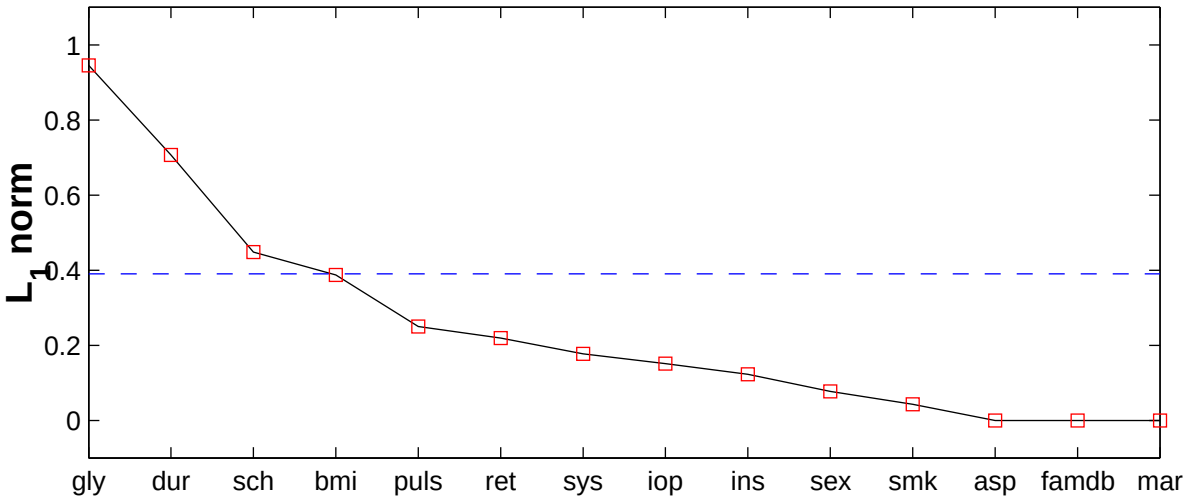
♣♣ 8. Choosing the importance threshold (cont.).

3 Using the Monte Carlo bootstrap data generated above, fit the full model to each bootstrap sample, and compute the L_1 scores $L_1^t(\eta+1), t = 1, 2, \cdots, T$ for the next term. This results in a Monte Carlo bootstrap sample of T observations of $L_1(\eta+1)$ under the null model hypothesis.

4 Compare $L_1(\eta+1)$ to the bootstrap samples $L_1^t(\eta+1), t = 1, 2, \cdots, T$. If $L_1(\eta+1)$ is between the rth and the $r+1$st largest $L_1^t(\eta+1)$, then the Monte Carlo p-value is $p_\eta = \frac{r+1}{T+1}$.

end. For an α level test, increment η until $p_\eta \geq \alpha$ (or the full model is reached), and declare that the importance threshold is $L_1(\eta)$ if $\eta > 1$ and any number larger than $L_1(1)$ if $\eta = 0$.

♣♣ 9. Back to the WESDR results.

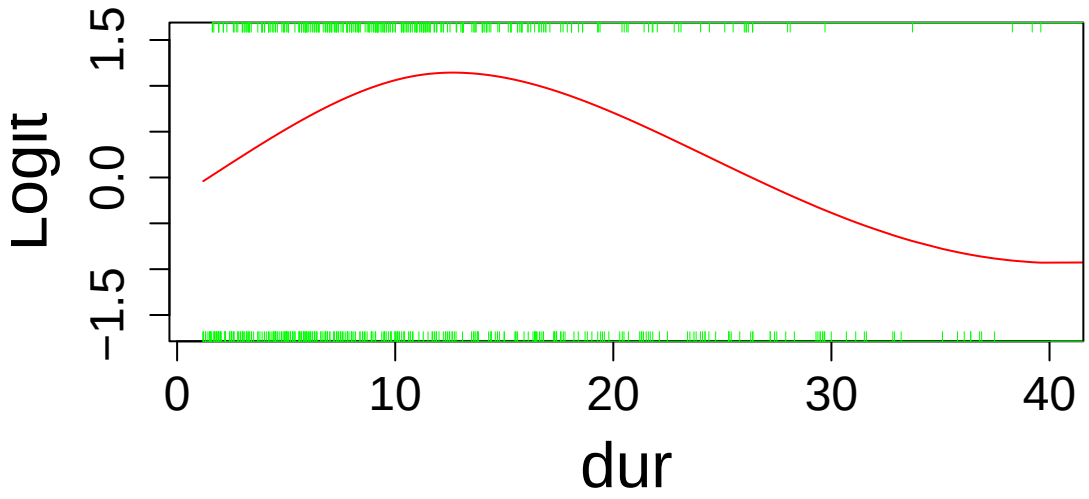


L_1 norm scores for the WESDR main effects model.

The importance threshhold is .39 (blue line). The p-values for all four of the selected variables `gly`, `dur`, `sch`, `bmi`, were .02. The solution was very sparse, with about 90% of the coefficients 0.

We were happy with the results - the method returned important variables, that had previously been selected by much more tedious methods. Simulation results in [ZWL02] have also demonstrated the efficacy of the approach in data sets where the 'truth' is known.

♣♣ 9. Back to the WESDR results(cont.).



Estimated logit component for dur

For comparison, the linear logistic regression model using the function *glm* in *R* was fit, and the linear coefficient for *dur* was not significant at level $\alpha = 0.05$. From the plot: A linear fit would not be good for dur.

Refitting the linear logistic model by including dur^2, the hypothesis test for dur^2 was significant with p-value 0.02. But with 14 variables, this is not easily discovered using parametric logistic regression.

♣♣ 10. Closing remarks.

We claim that that Likelihood Basis Pursuit can be a useful screening tool in data sets with many potential predictor variables, and can provide helpful insurance against loss of information due to ill-fitting parametric models.