# Prediction of the longitudinal *Fragile Families* survey via regularized regression

**Tom Bertalan**
Chemical and Biological Engineering
bertalan@princeton.edu

## Abstract

A large longitudinal dataset of feature values describing about four thousand children from mostly unmarried parents is treated as a black-box regression, in which a small set of response variables from the most recent wave of the survey is estimated using a subset of the more than twelve thousand predictor variables estimated over the course the participants' childhoods.

Data is pre-processed by dropping columns with an excessive count of missing values, imputing remaining missing values with per-column modes, and converting categorical variables to a one-hot encoding.

Methods used include radial basis function support vector machine regression, as well as ridge, LASSO, and elastic net regression, with best results being obtained from the last of these. Directions for future work are discussed, including performing linear and nonlinear dimension-reduction on the predictor features, as well as using matrix- and manifold-completion methods in place of simple modal imputation.

## 1 Introduction

The Fragile Families and Child Wellbeing Study is a longitudinal study of about 5,000 American children and their immediate environments. [7] A large fraction of the children were born to unmarried parents, families which are defined for these purposes as "fragile". The study contains records to survey questions such as "Who does child usually live with?" (questions `m3a3a` and `f3a3a` – with categorical responses) and "About how many days did you see child in past 30 days?" (with numeric responses).

In total, there are about 12,000 features in the dataset, from surveys conducted when children were ages one, three, five, and nine. The latest wave of the survey, conducted at age 15, is currently underway, but data available for this project include six outcome features–eviction, job training, layoff, GPA, "grit", and material hardship. "Grit" is a sociological/psychological construction reflecting a child's "perseverance and passion for long-term goals." [6] The first three of these are binary-valued, and the last three are numeric. For the purposes of this project, the year-15 data are considered to be response features, and the years 1-9 data are considered to be predictor features.

About 9,800 of the predictor features are free of missing values. Some of the remaining $\sim$2,200 features have a small fraction of missing values, while some are completely missing.

The data as provided describes a 4,242 children. However, only half of the cases have the six year-15 outcomes provided. The task is to train a predictor on the 2,121 given predictor/response pairs, then upload predictions for the missing 2,121 rows in the six-column response array (in addition to approximate predictions for the training data). For the purposes of this paper, I chose to regress only the continuous outcomes (GPA, grit, and material hardship). For predicting the remaining categorical outcomes, a classification rather than regression approach would be more appropriate. Submissions

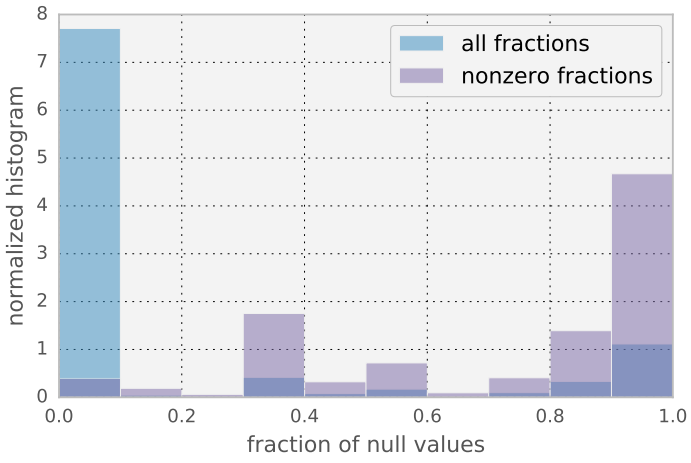


Figure 1: **Histogram of counts of missing values across the original set of more than 12,000 features.** Most features (blue histogram) were complete (had no missing values). However, drawing a histogram for only those features which which were missing some nonzero fraction of their entries shows some variation, with some features nearly completely missing. Histograms are normalized s.t. the sum of column areas is one.

are uploaded at `http://codalab.fragilefamilieschallenge.org/`, where they are automatically evaluated against the 2,121 held-out true values for six year-15 response variables. (This report and corresponding results were uploaded under the username `bertalan`.)

## 2 Methods

### 2.1 Pruning and imputation

Data was loaded with Pandas [8], and codebook files were parsed to determine whether each feature was numeric or categorical. Categorical variables were expanded using a one-hot encoding, which would have caused the number of predictor features to balloon to over 100,000. However, before this, some features (columns) and samples (rows) were pruned or imputed, substantially reducing the number of predictor features.

Due to missed questions, or survey skip patterns (in which conditional on a previous question, a subsequent question might not be applicable in one survey instance), The data contains many missing values. Some are coded with reasons such as "not in wave", "skip", "missing", "don't know", or "refuse", and some are missing without a numeric reason code. In future work (see below), I plan to apply more sophisticated matrix-completion techniques for filling in a subset of these values, such as those not coded as "skip". Additionally, there is reason to believe that some "missing" data are actually informative–or, in other words For pruning, predictor features whose fraction of missing values was greater than 0.5 were dropped entirely. For imputation, remaining missing predictor feature values were replaced with the mode of the feature across all samples.

Response variables were separately imputed using Multiple Imputation by Chained Equations (MICE) [1]. This is as alternative to simply dropping rows with any missing response values, which performed worse on the internal test set (§2.2).

### 2.2 Train/test split

In addition to the enforced 2,121-train/2,121-test split imposed by the form of the provided data (see §1), I further split off $n_{\text{test}} = 121$ of the provided training rows to create an internal validation set. This proved useful e.g. for diagnosing overfitting before upload of predictions to the challenge

website (see §3). Therefore, below I will use the labels "train", "test", and "challenge" to refer to the 2,121-$n_{\text{test}}$ training rows, $n_{\text{test}}$ internal validation rows, and 2,121 online challenge rows; both predictors $\boldsymbol{x}$ and responses $\boldsymbol{y}$.

### 2.3 Regression

Four regression techniques were used on this data: ridge regression, LASSO, elastic net, and radial-basis-function support vector machine.

#### 2.3.1 Regularized regression

Ridge regression, LASSO, and the Elastic Net [9] are all forms of penalized linear regression. Elastic net [12] can be seen as a generalization of LASSO (least absolute shrinkage and selection operator) [11] and ridge regression.

In standard ordinary least squares (OLS) regression, fitting parameters $w_0$, $\boldsymbol{w}$ are found to minimize an objective function [9]

$$J_{\text{OLS}}(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - (w_0 + \boldsymbol{w}^T\boldsymbol{x}_i))^2, \tag{1}$$

where $y_i$ is the response to the $i_th$ feature vector $\boldsymbol{x}_i$. For models with multiple outputs, as in this case, this can be generalized by writing several independent regressions, or, equivalently, one with multiple columns of coefficients $\boldsymbol{w}$ and responses $\boldsymbol{y}$.

In ridge regression, also known as $l_2$ regularization, a Gaussian prior is imposed on the parameters $\boldsymbol{w}$. This manifests as the addition of the 2-norm of the weights as a new term in objective function

$$J_{\text{ridge}}(\boldsymbol{w}) = J_{\text{OLS}}(\boldsymbol{w}) + \lambda_{\text{ridge}}||\boldsymbol{w}||_2^2 \tag{2}$$

This has benefits especially in situations where some features are highly collinear, for which the corresponding coefficients might become very large and of opposing signs. $l_2$ regularization generates a maximum a-posteriori (MAP) estimate for the parameters $\boldsymbol{w}$ rather than the maximum likelihood (ML) estimate–loosely speaking, the estimate is shifted towards the origin, and more so along directions of greater uncertainty. [9] This shifting towards zero is described as "shrinkage".

With the LASSO [9, 11], an $l_1$ penalty is used instead.

$$J_{\text{LASSO}}(\boldsymbol{w}) = J_{\text{OLS}}(\boldsymbol{w}) + \lambda_{\text{LASSO}}||\boldsymbol{w}||_1. \tag{3}$$

Here, $||\boldsymbol{z}||_1 = \sum z_k^1$ is the $l_1$ norm of the parameters. Eqn. 3 is the Lagrangian form of an alternative program

$$\boldsymbol{w}^* = \text{argmin}_{\boldsymbol{w}} J_{\text{OLS}}(\boldsymbol{w}) \quad \text{subject to} \quad ||\boldsymbol{w}||_1 \leq t. \tag{4}$$

As before, imposing this regularization is equivalent to finding a MAP estimate for the parameters rather than the MLE of OLS. Now however, the prior is a Laplace distribution, strongly peaked at zero with a discontinuous first derivative. For this reason, the LASSO naturally performs features selection, by preferring fits in which some coefficients are exactly zero. [1]

Elastic net combines the $l_1$ and $l_2$ regularizations, providing the $l_2$ benefits in cases with strongly correlated features, and the feature-selection of $l_1$.

The Python package `sklearn.linear_model` [10] provides a multi-task elastic net implementation for performing several regression tasks jointly. For six columns of responses $\mathbf{Y}$ with one predictor feature per column of $\mathbf{X}$, this package finds weights in the corresponding six columns of $\mathbf{W}$ by minimizing

$$\frac{1}{2n_{\text{samples}}}||\mathbf{XW} - \mathbf{Y}||_{\text{Fro}}^2] + \alpha\rho||\mathbf{W}||_{2,1} + \frac{\alpha(1-\rho)}{2}||\mathbf{W}||_{\text{Fro}}^2, \tag{5}$$

where $||\mathbf{Z}||_{\text{Fro}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|z_{ij}||^2}$ is the Frobenius norm, and $||Z||_{2,1} = \sum_{j=1}^{n}\left(\sum_{i=1}^{m}|z_{ij}|^2\right)^{1/2}$, (the sum of Euclidean norms of columns) is the $l_{2,1}$ norm. Default hyperparameter values are $\alpha = 1$ and $\rho = 1/2$. In an effort to encourage sparsity, $\alpha$ was doubled from its default value.

[1] $l_1$ regularization can also be seen as a convex relaxation of the more direct alternative of $l_0$ regularization

### 2.3.2 Support vector machine regression

As a briefly-considered alternative to regularized linear regression, I also tried support vector machine (SVM) regression. Briefly, SVM regression implicitly views predictors $\boldsymbol{x}_i$ not in their native feature space, but in a space transformed in a manner governed by a chosen kernel function. If, as in this case, the radial basis kernel function is chosen, this is equivalent to viewing each data points as samples of a random walk along a submanifold of the native features space. The implicitly transformed coordinates correspond to natural coordinates within this (possibly curved) submanifold. For SVM regression (or classification), the actual transformation need not actually be computed–using the so-called kernel trick, any methods reliant on inner products between predictors $\boldsymbol{x}$ (as in ridge-regression) can be "kernelized" by replacing these inner products with calls to the chosen kernel function.

## 3 Results

FIX

Ridge regression clearly overfit the data, as visible in Table 1. LASSO presented a large improvement on ridge regression,

On the Fragile Families Challenge website, the mean-squared error of these methods' predictions for chosen (continuous) responses on the the held-out challenge data were 0.44424 for GPA, 0.30139 for grit, and 0.02968 for material hardship.

| | $\hat{r^2}$ **(train)** | $\hat{r^2}$ **(test)** |
|---|---|---|
| **ridge** | 0.999999 | -0.771324 |
| **lasso** | 0.243894 | -0.066203 |
| **elasticNet($\alpha$=2.0)** | 0.265761 | -0.063879 |
| **svm(3)** | 0.804006 | -0.106628 |

Table 1: Average coefficients of determination $\bar{r^2}$ for several fitted models. For this multiple-output regression (i.e., regression for the GPA, grit, and material hardship response variables), we take the mean over the three response variables of the per-variable coefficient of determination $r^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$, where $SS_{\text{res}} = \sum_i (y_i - f_i)^2$, $SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$, $y_i$ are true response values and $f_i$ are model predictions.

## 4 Discussion and conclusion

The poor predictive ability of the ridge-regression (see Table 1) is a clear indication of over-fitting. Since the dataset had several times the number of features as training samples, and ridge-regression does not perform any feature selection, this result is unsurprising–it's possible to encode nearly perfectly the training responses in such a large number of trained weights. LASSO and Elastic Net, in contrast, were able to eliminate a large number of features, greatly improving the model's robustness to overfitting, and producing similar test and train coefficients of determination.

SVM performed fairly well on the training data, but less well on the internal test data, and so was not chosen for submission. The motivation for using an SVM was its implied ability to perform nonlinear dimensionality reduction on the predictors. In future work, I plan to explore this further, combining the dimensionality reduction of manifold learning techniques such as simple principal component analysis, or nonlinear equivalents such as diffusion maps [3, 5, 2, 4]. An adaptation of the second of these is currently being examined for use in both pre-regression dimensionality reduction and for imputing missing values (a task called "manifold completion" or "matrix completion").

More immediately, I plan to use cross-validation to tune the hyperparameters of the LASSO fit beyond intuitive tweaks to the default values.

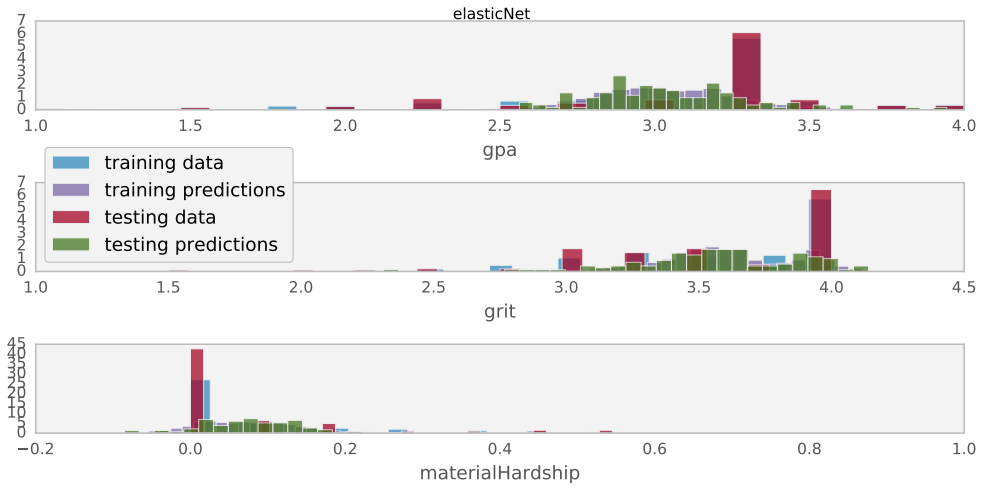


Figure 2: **Histograms of response variables in training and testing data, as well as predictions of both.** The regression captures the primary mode of all three variables, as well as the secondary mode of the grit variable.

## References


[1] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–9, mar 2011.

[2] Mikhail Belkin and Partha Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15(6):1373–1396, 2003.

[3] Ronald R Coifman and Stéphane Lafon. Geometric harmonics: A novel tool for multiscale out-of-sample extension of empirical functions. *Applied and Computational Harmonic Analysis*, 21(1):31–52, 2006.

[4] Ronald R Coifman, Stéphane Lafon, Ann B Lee, Mauro Maggioni, Boaz Nadler, F Warner, and S W Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *PNAS*, 102(21), may 2005.

[5] Carmeline J Dsilva, Ronen Talmon, Ronald R Coifman, and Ioannis G Kevrekidis. Parsimonious representation of nonlinear dynamical systems through manifold learning: A chemotaxis case study. *Applied and Computational Harmonic Analysis*, 1:1–15, 2015.

[6] Angela L. Duckworth, Christopher Peterson, Michael D. Matthews, and Dennis R. Kelly. Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, 92(6):1087–1101, jun 2007.

[7] Ian D. Lundberg, Matthew J. Salganik, Sara McLanahan, Irwin Garfinkel, Janet Currie, Dan Notterman, Jeanne Brooks-Gunn, Ron Mincy, and Jane Waldfogel. About the Fragile Families and Child Wellbeing Study — Fragile Families and Child Wellbeing Study, 2017.

[8] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51–56, 2010.

[9] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. 2012.

[10] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[11] Robert Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, jun 2011.

[12] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, apr 2005.