Human Perception of Important Variables in Linear Regression Model

Qinmei Wu, Xiaojun Wang, Jiaoyan Chen and Qing Xu

Abstract— Multiple linear regression have settled on a model which contains several predictor variables that are statistically significant. At this point, it's common to ask that which variable is most important. In this paper, we are going to measure the human perception of important features in linear regression model by given scatterplot. Compared to the standardized regression coefficients and selected features in LASSO, we found that the performance of human perception of important features is quite good when we take that features selected by Lasso as important ones.

Index Terms—Perception, Visualization, Evaluation.

1 Introduction

Machine learning is an important subject in data analyzing such as prediction, classification and clustering. Corporation are trying to seek common rules of operation behind the daily data they stored. Predictive analytics is frequently mentioned which used to analyze current data and make predictions about unknown future events. It uses many techniques from data mining, statistics and machine learning to capture relationship among features. For the dataset which contains multiple variables, it is important to do feature selection before modeling [1]. If all variables taken into consideration for modeling, overfitting is often the case [2].

In this research, we are going to identify the importance of features in regression models. The simple and direct way people can think about is plotting relationship between the observation and prediction variable and selecting the more important ones. In the experiment, we make several scatter plots to visualize the relationship of each feature and the dependent variable which is the value we are trying to predicating by model. And then we make a web survey which contains the scatter plots and asks people to rank the importance of features based on these plots. In this experiment, we collect people's opinions and compared with standardized regression coefficients in multiple linear model [3] and features selected by lasso regression models. We are trying to find out whether the features selection can be observed by human perception.

We will review basic terminology in linear regression first and introduce two different criterions we used to identify the importance of a variable. Then, we will explain our experiment on comparing human perception of important variables to criterions mentioned above. The discussion and conclusion is focus on if scatter plot is a good way for feature selection.

2 Background

In statistics, linear regression is an approach for modeling the relationship between a scalar dependent prediction y and one or more independent observations which are denoted as x [4]. The case of studying only one observation is called simple linear regression. For more than one independent observations, the process is called multiple linear regression [5]. What need to notice is this term is different from multivariate linear regression. In multivariate linear regression, multiple correlated dependent observations are predicted, rather than a single scalar variable [6].

Basically, linear regression (such as OLS) makes a prediction based on an inner product between trained coefficients β and feature variables x. The coefficient value represents the mean change in the response given a one-unit increase in the predictor. Consequently, it's easy to think that variables with larger coefficients are more important because they represent a larger change in the response. However, the scales vary between different variables, which makes it impossible to compare them directly. For example, the meaning of a one-unit change is very different if you're talking about population(which is commonly up to millions) and local income(which is commonly scaled in several thousands).

What's more, the problem is further complicated by the fact that there are different units within each type of measurement. For example, weight can be measured in kilogram, gram, ounce. If you fit models for the same data set using grams in one model and kilograms in another, the coefficient for weight changes by a factor of a thousand even the model remains unchanged. However, there are still some statistical method that can help people determine which predictor variables are relatively more important in regression models. Here we use two different ways to evaluate the importance of variables. One is compare the coefficients after normalization, another is using regularization models.

Data normalization is a method which standardize the range of independent variables of data [7]. The point of normalization is to make variables comparable to each other. The reason this is a problem is that measurements made using such scales of measurement as nominal, ordinal, interval and ratio are not unique. The range of input data could be spanned widely and might cause misleading in some machine learning algorithms. To eliminate misleading brought by data, the range of data will be scaled into [0,1], and each feature will contain a same distribution in this new range.

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{1}$$

In more complicated cases, normalization may refer to more sophisticated adjustments where the intention is to bring the entire probability distributions of adjusted values into alignment.

Overfitting Due to the noise of data, a model fits a given data with a unbelievable high accuracy might face a problem of overfitting [8]. An overfit model can cause the regression coefficients, p-values, and R-squared to be misleading [9]. Overfitting will be detected will apply the test data into training model. If the error in test increased dramatically, it means the model face the problems of overfitting. In order to avoid overfitting, it is common to add a penalty to existing linear model.

Regularization refers to the method of preventing overfitting, by explicitly controlling the model complexity. It leads to smoothen of the regression line [10] and thus prevents overfitting. It does so by

- *Qinmei Wu - qwu4@gmail.com*
- *Jiaoyan Chen - jch11@wpi.edu*
- *Xiaojun Wang - xwang18@wpi.edu*
- *Qing Xu - qxu@wpi.edu*

Manuscript received 31 March 2014; accepted 1 August 2014; posted online 13 October 2014; mailed on 4 October 2014.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

penalizing the bent of the regression line that tries to closely match the noisy data points.

A popular method based on the knowledge of regularization is ridge regression [11]. One motivation of ridge regression is that the dataset could be high dimensional. In some situation, the number of input variables even could greatly exceed the number of observations. With a number of predictors, the covariates of predicting attributes could be collinear, which means that certain attributes are highly linearly related. Therefore, these internal relationships among attributes will impact the estimates of the regression parameters which could cause a large errors if we don't detect relationships among attributes. Ridge regression will put constraints on parameters which make those internal correlated parameters become quite low.

Also, fitting the whole model without penalization or regularize the coefficients will cause large prediction intervals. Therefore, Ridge regression will imposed constraints on parameters β while calculating penalized sum of squares.

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

This penalty term λ is a constant which penalize the function while β takes a large number. If we set a large λ, the parameters will be constrained severely and the degrees of freedom will be lower.

A more popular method for regularization is Lasso Regression [12], which is a relatively recent alternative to Ridge Regression. Lasso regression will shrink the coefficients of predictors that are less statistically important to zero by by adding a penalty.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3}$$

It was introduced by Robert Tibshirani in 1996 based on Leo Breiman's Nonnegative Garrote [13]. Lasso was originally formulated for least squares models. It also reveals that for standard linear regression, the coefficient estimates need not be unique if covariates are collinear. In this paper, we use Lasso model to select important features.

3 Experiment

In order to measure human perception of important features by given scatter plots [14] [15], we compare the important features selected by our participants to significant features in standardized linear regression and regularization linear regression. We selected four independent datasets: Auto(392 observations with 7 variables), Bike sharing(10887 observations with 8 variables), Cigarette Consumption(1380 observations with 7 variables) and Child Seat Sales(400 observations with 10 variables) from R standard library. These datasets have multidimensional data frame. Among the variables, only one attribute was prediction variable, it could be selected as the dependent variable. All of the other attributes were selected as possible observation variables, which are independent variables. The data frame and method of data recording satisfied the requirement of regression model. To visualize the correlation between the dependent variable and independent variables, scatterplot is selected for human perception.

In the beginning of our experiment, 20 participants were asked to estimate how these independent variables influence the dependant variable by observing scatter plots. participants also need to select all features and sort them from the most influential one to the least influential one. Next we ran the standard linear regression and lasso to select important features. For standard linear regression, we computed the coefficient for each feature. Feature with larger coefficient was considered as more important one. For LASSO, we did not care about the value of coefficient. Instead, we thought all features selected by lasso model were important.

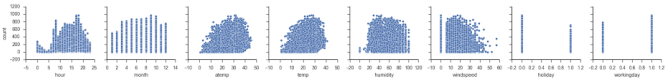

Fig. 1: Scatter plot for bike sharing

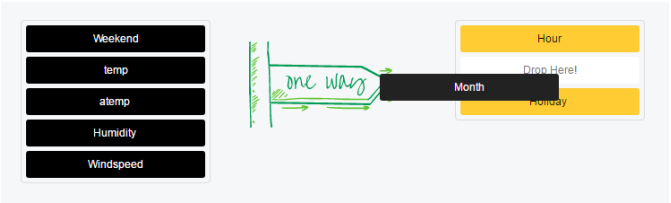

Fig. 2: Ranking process

4 Methodology

In experiment, participants need answer 4 questions for each dataset through the questionnaire. Every question was presented to participant with data descriptions and 1 scatter plot. For instance, in question 2, the Bike sharing dataset have one dependent variable called count ,the number of total rentals in one hour, which is vertical coordinate variable. The other independent variables are hour, month, holiday, weekend, temperature,'feel like' temp, humidity and wind speed. They are vertical coordinate variables in each subplot. As the Fig. 1 showed, each subplot represent the correlation of mpg and other independent variables. Participant will observe the scatter plot and give an importance feature ranking within the independent variables(Fig. 2).

After finished collect the experiment results, we calculate the coefficients of Multiple linear regression model and lasso regression model through Seaborn which is a statistics python library. In the next, we compare the user results with two model's coefficients respectively. For Lasso regression, we ordered the important coefficient descendingly. Then we count how many important coefficients will orderly matched with the lasso coefficient ranking list. For example, in bike sharing part, there are 8 variables in all, but only 6 variables selected by lasso model. We need to count how many variables in high ranking selected by each participant could hit in theses 6 variables and calculate the percentage of experiment group. For multiple linear regression, some coefficients have a huge difference, to ensure accuracy, we used the size of difference to categorize the coefficients in significant level and secondary level. For instance, in bike sharing part, the coefficient of hour and humidity are 52 and -44, the coefficient of atemp and month is 17 and 25. There are huge gap existed. So the hour and humidity are significant variables, the atemp and month are secondary variables. As the lasso model part, we calculate how the situation that participants hit the two categories variables distributed. Finally, through the comparison results, we evaluate how is the human perception of important features by given scatter plots.

5 Results

First, we compared the human perception with Lasso feature selection. For Auto data set,as Fig. 3 showed, 5.3% of participants exactly selected all the important features which were also selected by lasso model. 36.84% of participants selected 3 important features overlapped by lasso and 57.89% of participants selected 2. All participants selected at least half of the important features. For Bike Sharing data set,as Fig. 4 showed, 21.05% selected all important features, 10.53% selected 5 and 68.4% selected 4. For Cigar Consumption data set,as Fig. 5 showed, 15.79% selected 3 important features and 84.21% selected 2. For Car seat sales data set, as Fig. 6 showed, 57.89% selected the most important features.

Then we compared the human perception with multiple linear regression. If we just consider the significant impact, for Auto data set (Fig. 7)) 57.89% of participants could select the most important features which had the largest coefficient in multiple linear regression. For Bike Sharing data set(Fig. 8) , 57.89% of participants selected one

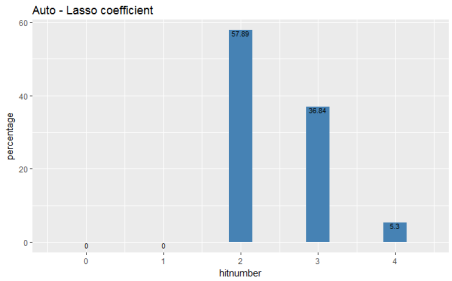

Fig. 3: Auto Dataset - Percentage of important features selected by human perception hitted in the Lasso coefficients

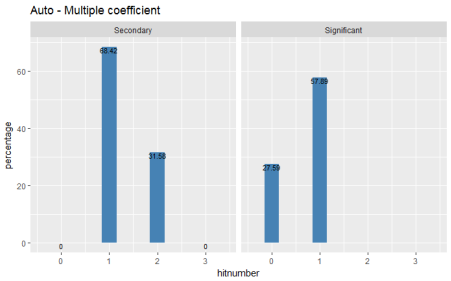

Fig. 7: Auto Dataset - Percentage of important features selected by human perception hitted in the MLR coefficients

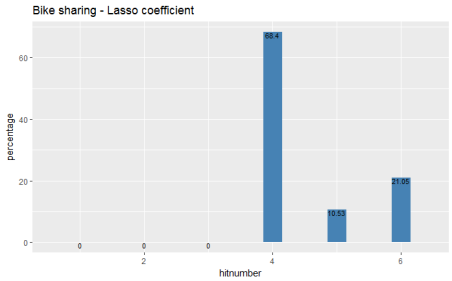

Fig. 4: Bike Sharing Dataset - Percentage of important features selected by human perception hitted in the Lasso coefficients

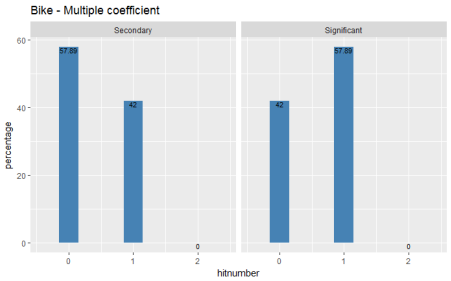

Fig. 8: Bike Sharing Dataset - Percentage of important features selected by human perception hitted in the MLR coefficients

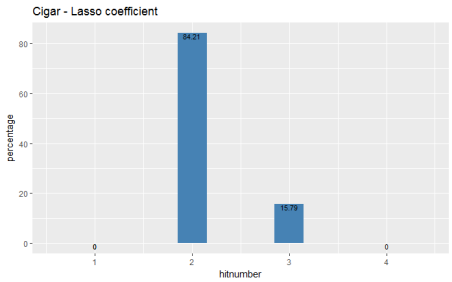

Fig. 5: Cigarette Consumption Dataset - Percentage of important features selected by human perception hitted in the Lasso coefficients

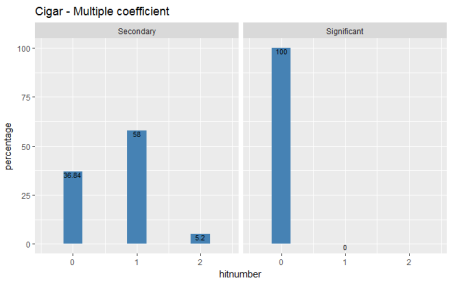

Fig. 9: Cigarette Consumption Dataset - Percentage of important features selected by human perception hitted in the MLR coefficients

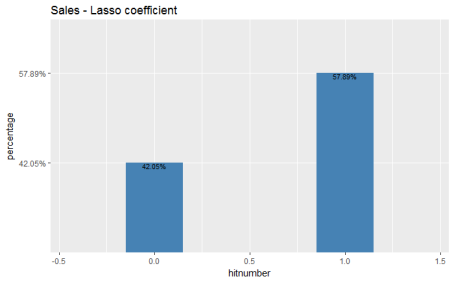

Fig. 6: Sales of Child Seats Dataset - Percentage of important features selected by human perception hitted in the Lasso coefficients

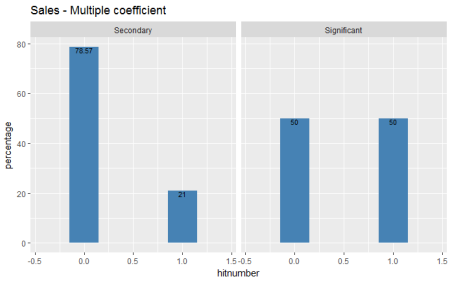

Fig. 10: Sales of Child Seats Dataset - Percentage of important features selected by human perception hitted in the MLR coefficients

of the two important features. For Cigar Consumption data set((Fig. 9)), no one selected out the most important features. For Car seat sales data set(Fig. 10) , 50% selected the most important features.

6 Discussion

6.1 Human Perception of Important Features VS Feature selection by Lasso

Participants can capture over half of the significant features by their observation which are selected by Lasso model. We can find that the situation where participants cannot select any key feature rarely happens. For the Child Seat Sales, though nine features given into consideration when predicting sales, only price can actually influence the final result according to Lasso. In this experiment, it has more strict requirement on participants that they must find out the most important feature by the given scatter plot. However, half of the participant can still make it.

6.2 Human Perception of Important Features VS Standardized Regression Coefficients of Multiple Linear

RegressionWhen use the value of standardized regression coefficients of multiple linear regression as the criterion for feature importance, the performance of these participants is hard to judge. According to the result, it seems human perception of important features quite depends on the data set. For Auto and Bike Sharing dataset, participants did a relatively better job than Child Seat Sales and Cigar Consumption data set. Especially, none of participants capture the most important feature 'year' in Cigar data. What's more, we found it is hard for human perception of the importance of a categorical feature.

7 Conclusion

The performance of human perception of significant features in linear model depends on which criterion we use to define the actually significant features. In this paper, we design an experiment to ask participants select relatively more important features by given scatter plots. However, there is nouniversal rules to identify what are the important features in linear regression model. Here we considered two different criterion: one is considering the features selected by lasso. All features with coefficients not shrink to zero are taken as significant features. Another method is evaluating the importance by the value of coefficient. If we use the former one as our evaluation criterion, human perception is good for selecting important features, which can help avoid overfitting by leaving out trivial features in multiple linear regression. Our feature work is exploring a more reasonable criterion to identify the important features, so we can set a comparison for human perception.

References

[1] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[2] John Loughrey and Pádraig Cunningham. Using early-stopping to avoid overfitting in wrapper-based feature selection employing stochastic search. In *Proceedings of the Twenty-fifth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, 2005.

[3] Richard Taylor. Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, 6(1):35–39, 1990.

[4] David F Andrews. A robust method for multiple linear regression. *Technometrics*, 16(4):523–531, 1974.

[5] David A Freedman. *Statistical models: theory and practice*. cambridge university press, 2009.

[6] Alvin C Rencher. *Methods of multivariate analysis*, volume 492. John Wiley & Sons, 2003.

[7] David A Freedman. *Statistical models: theory and practice*. cambridge university press, 2009.

[8] Wallace E Larimore and Raman K Mehra. Problem of overfitting data. *Byte*, 10(10):167–178, 1985.

[9] Frank E Harrell Jr, Kerry L Lee, David B Matchar, and Thomas A Reichert. Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer treatment reports*, 69(10):1071–1077, 1985.

[10] Yuichiro Anzai. *Pattern recognition and machine learning*. Elsevier, 2012.

[11] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[12] Leo Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

[13] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[14] Michael E Doherty, Richard B Anderson, Andrea M Angott, and Dale S Klopfer. The perception of scatterplots. *Attention, Perception, & Psychophysics*, 69(7):1261–1272, 2007.

[15] William S Cleveland, Persi Diaconis, and Robert McGill. Variables on scatterplots look more highly correlated when the scales are increased. Technical report, DTIC Document, 1982.