Description

In this session, the goal is to understand what spine is and we will use spline to explore the relationship of wage and other variables.

Data Information

The data set is called Mid-Atlantic Wage. This data is part of the ISLR package in R, which is cited from James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning with applications in R, www.StatLearning.com, Springer-Verlag, New York. It contains 3000 observations and 11 attributes or variables on workers' wages among other information.

The 11 attributes and a short description of each variable (attribute) is provided below.

The Attributes:

Year: Year that wage information was recorded
Age: Age of worker
Maritl: Marital Status
Race: Race of worker
Education: Education level of worker
Region: Region of the country (mid-atlantic only)
Jobclass: Industrial or Informational type of Job
Health: Health level of worker
Health_ins: Indicates whether worker has health insurance
Logwage Log: of workers wage
Wage: Workers raw wage

HOME MODULE

Select New Project create new project:

Project Name: Mid Atlantic Wage

Select Project Type: Data Analytics

Description: Write the project description

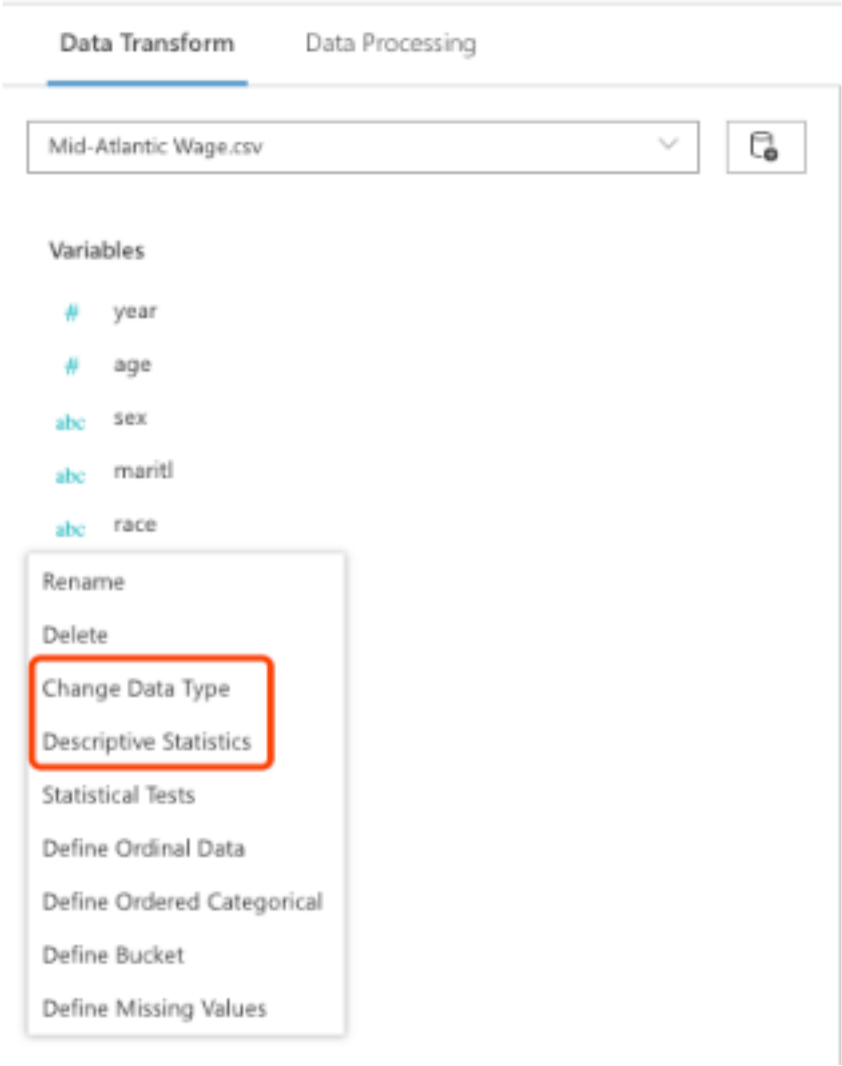
DATA UPLOAD MODEL

Upload Data: Mid-Atlantic Wage.csv

DATA MANAGER MODEL

Raw data and data types need to be checked before processing. The processing can be done in **Data Processing**.

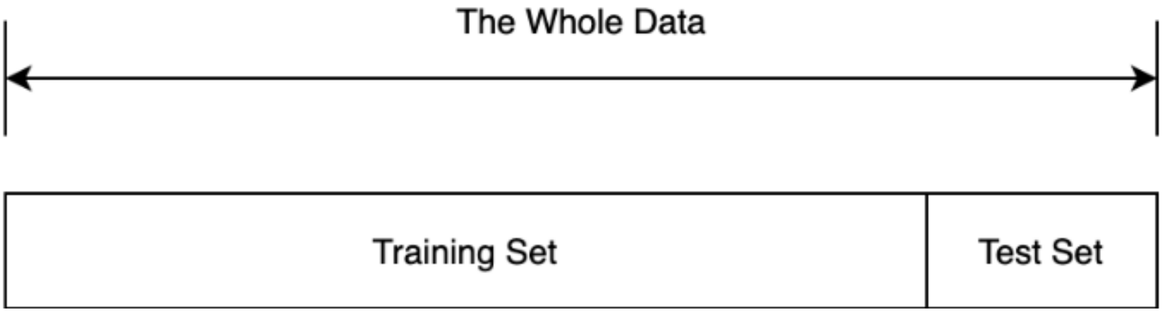
Go to the **data manager** and check the data types by selecting **Change Data Type**, The descriptive statistics will perform some descriptions. You can do this by using the mouse click on the variable of interest within the data manager, providing these options:

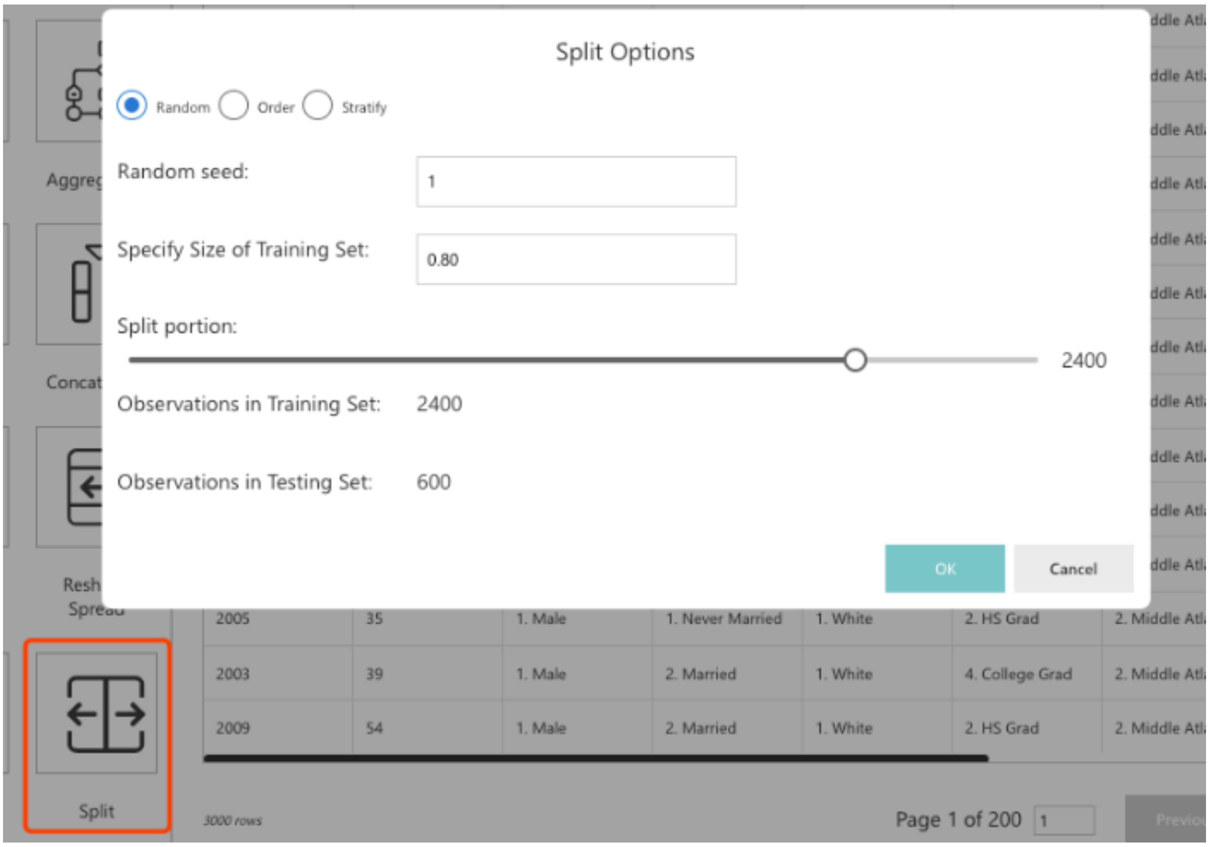


Splitting

Please go to data processing site and select Split.

For Machine Learning, if we want to perform classification or predict statistical models, we can split the data into training and testing sets. Testing set can be used to test if the model will overfit when we use new samples. If a model fit to the training dataset also fits the test dataset well, minimal overfitting has taken place. A better fitting of the training dataset as opposed to the test dataset usually points to overfitting.





Split Options

● Random ○ Order ○ Stratify

Random seed: 1

Specify Size of Training Set: 0.80

Split portion: 2400

Observations in Training Set: 2400

Observations in Testing Set: 600

2005	35	1. Male	1. Never Married	1. White	2. HS Grad	2. Middle Atl
2003	39	1. Male	2. Married	1. White	4. College Grad	2. Middle Atl
2009	54	1. Male	2. Married	1. White	2. HS Grad	2. Middle Atl

3000 rows

Page 1 of 200 1

Change the specific size of the training set from 0.90 to 0.80 which means the whole dataset separated as 80% of training data and 20% of test data. You can change the percentage of the training and testing sets based on your requirements.

Go back to Data Upload module, you can see the split data has been settled.

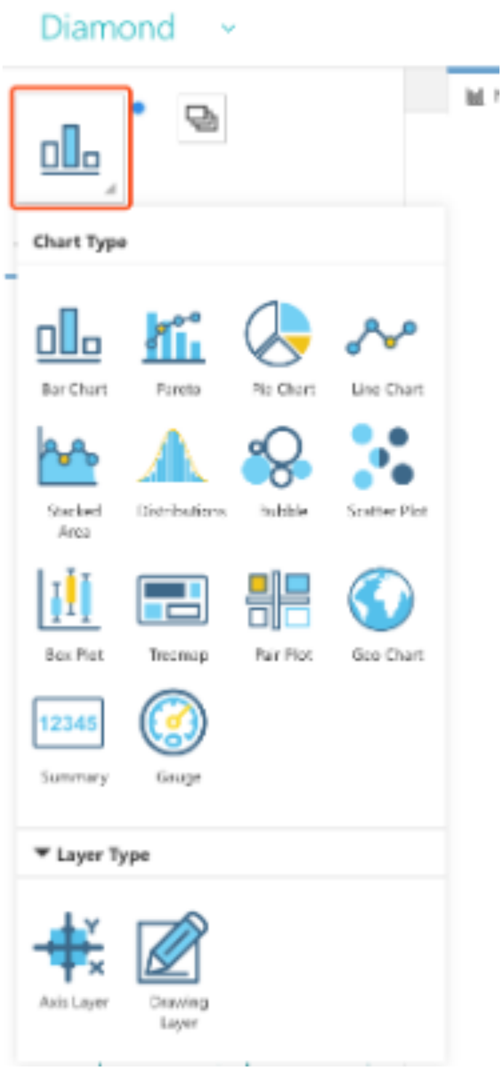
VISUALIZATION MODEL

AutoStat provides an easy drag and drop visualization module.

Before starting modeling data, we should visualize the data variables for a deeper understanding of the relationships between variables.

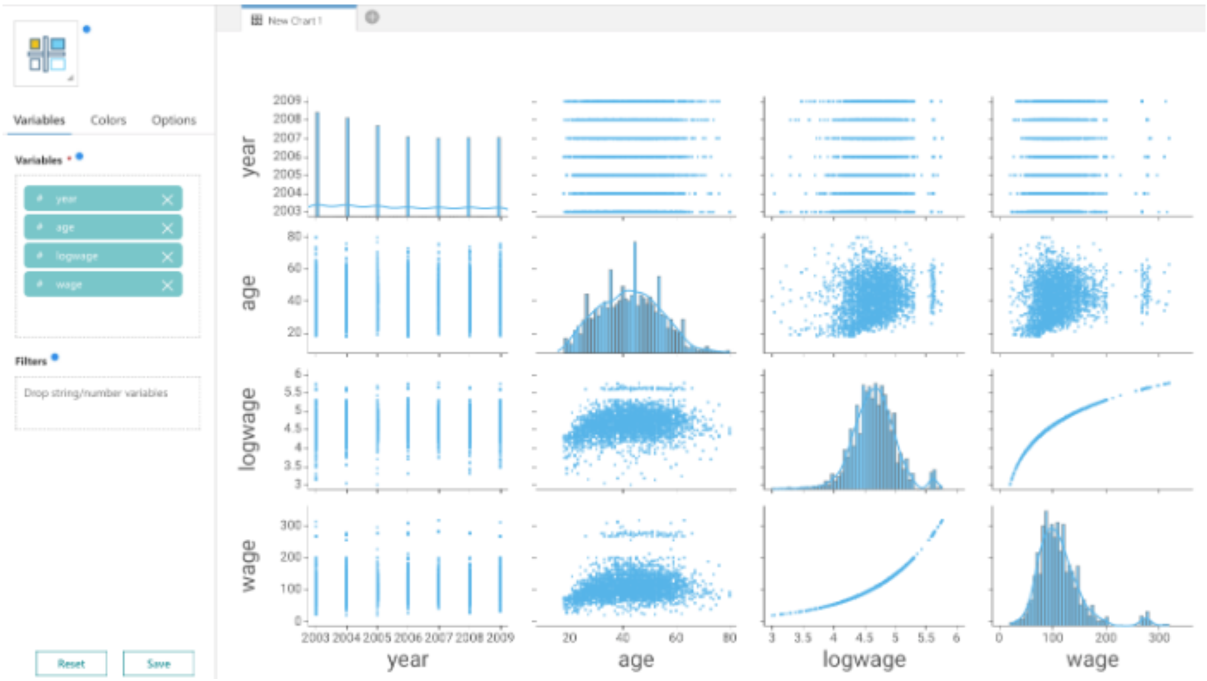
In this case, we are exploring the related variables of heart disease.

There has a variety of Chart Type, click the chart type for start. For each chart start, just click reset, then the chart will empty.

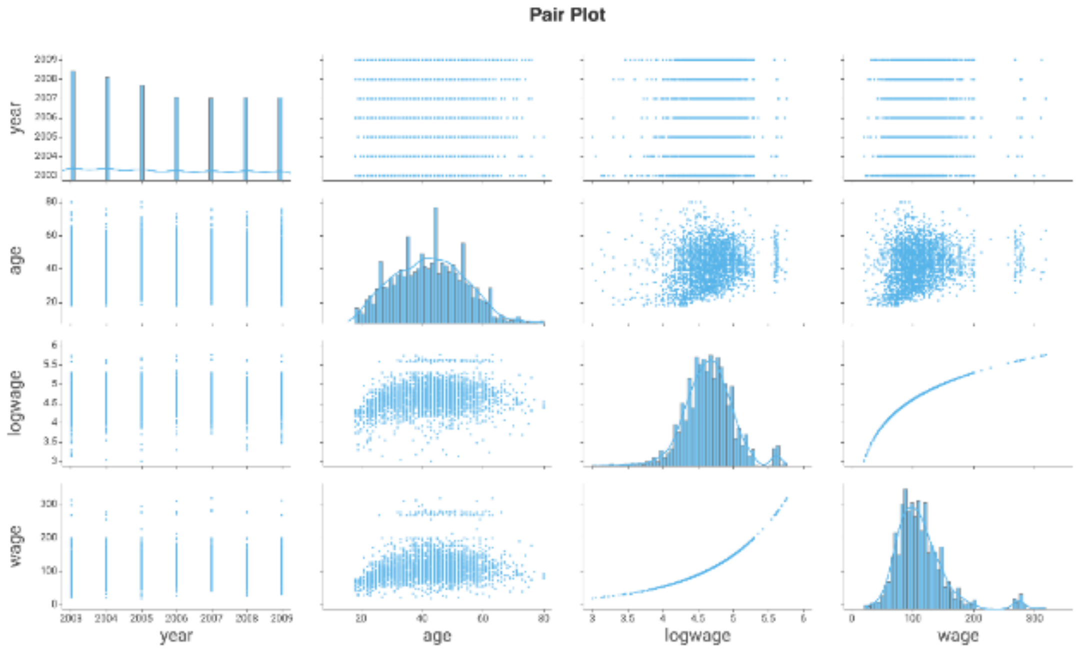


Pairplot:

All numeric variables have been choose to Pair Plot.



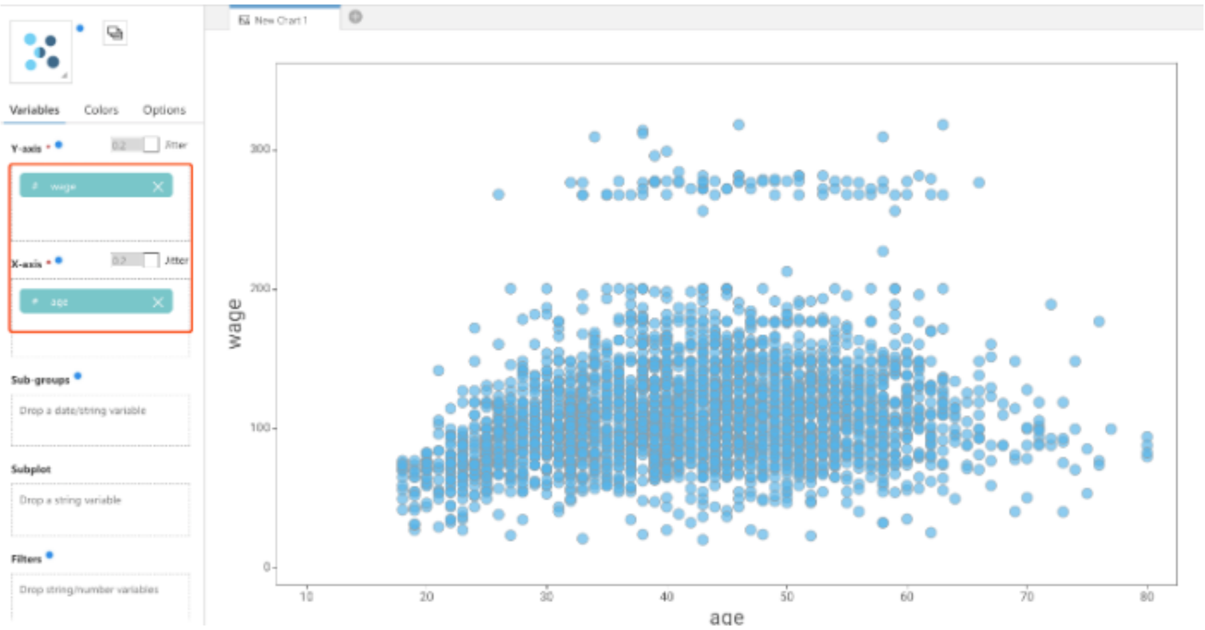
The chart allows us to see Mid Atlantic Wage data distribution of single variables and relationships between two variables, which can be separate with linear relationships or non-linear relationships. It is a great chart type for us to identify the trends for further analysis.



The chart shows there has Non-linear relationship between wage and age and the Non-linear relationship between logwage and age.

Scatter Plot

We are using scatter plot to visualize the relationship between wage and age.



The chart is very clear to show that the relationship between wage and age is Non-linear relationship.

MODEL BUILDER MODULE

Linear Regression

The standard linear regression model equation:

$y_i=\beta_0+\beta_1*b_1(x_i)+\beta_2*b_2(x_i)+\ldots+\beta_{k+3}(x_i)$yi=β0+β1∗b1(xi)+β2∗b2(xi)+...+βk+3(xi)

linear regression is terribly fitted the Non-linear data, and the R-squared and Adjusted R-squared values normally shows very poor.

Polynominal Regression

Polynomial Regression: The simple approach model Non-linear relationship, which adds polynomial terms or quadratic terms to regression.

The polynominal regression model equation:

$y_i=\beta_0+\beta_1 x_i+\beta_2 x_i^2+\ldots+\beta_m x_i^m+\epsilon$yi=β0+β1xi+β2xi2+...+βmxim+ϵ

The behavior of polynomials fitting data tends to be unstable near the boundaries, and extrapolation can be dangerous. Splines exacerbate these problems. The behavior of polynomial fitting beyond the boundary knot is even more crazy than the corresponding global polynomial in that region.

Cubic Splines

Spline Regression: Smooth and flexible function of fitting Non-linear Models and learning the Non-linear interactions from the data. In most of the methods in which we fit Non-linear Models to data and learn Non-linearities is by transforming the data or the variables by applying a Non-linear transformation.

A **linear spline** with knots at $\xi_K \xi K$,*K* = 1, 2 ... $k k$ is a piecewise linear polynomial continuous at each knot. The set of linear splines with fixed knots is a vector space. The number of degrees of freedom is 2(*K*+1) - *K* = *K*+2. We can decompose linear splines on a basis of *K* +2 basis functions

$y={}_{k+2}\sum_{m=1}\beta_m h_m(x)+\epsilon y=\sum m=1k+2\beta mhm(x)+\epsilon$

The basis function can be chosen as$h_1(x)=1h1(x)=1h_2(x)=xh2(x)=xh_{k+2}(x)=(x-\xi_K)_+,k=1,\ldots,khk+2(x)=(x-\xi K)+,k=1,...,k$

$(.)_+(.)+$ denotes the positive part, i.e., $(x-\xi_K)_+=x-\xi_K(x-\xi K)+=x-\xi K$ if x > $\xi_K\xi K$ and $(x-\xi_K)_+=0(x-\xi K)+=0$ otherwise.

A cubic spline with knots at $\xi_K\xi K$, *K* = 1, 2 ... $k k$ is a piece-wise cubic polynomial with continuous derivatives up to order 2 at each knot. Enforcing more than one continuity order will result in a global cubic polynomial. The set of cubic splines with fixed knots is a vector space, and the number of degrees of freedom is $4(K+1)-3K=K+44(K+1)-3K=K+4$. Therefore, we can decompose cubic splines on a basis of *K*+4 basis functions,

$y={}_{k+4}\sum_{m=1}\beta_m h_m(x)+\epsilon y=\sum m=1k+4\beta mhm(x)+\epsilon$

They have continuous first and second derivatives. The continuity order is = ($d d$-1), where $d d$ is the degree of polynomial. Now we can represent the model with the truncated power basis function $b(x)b(x)$.

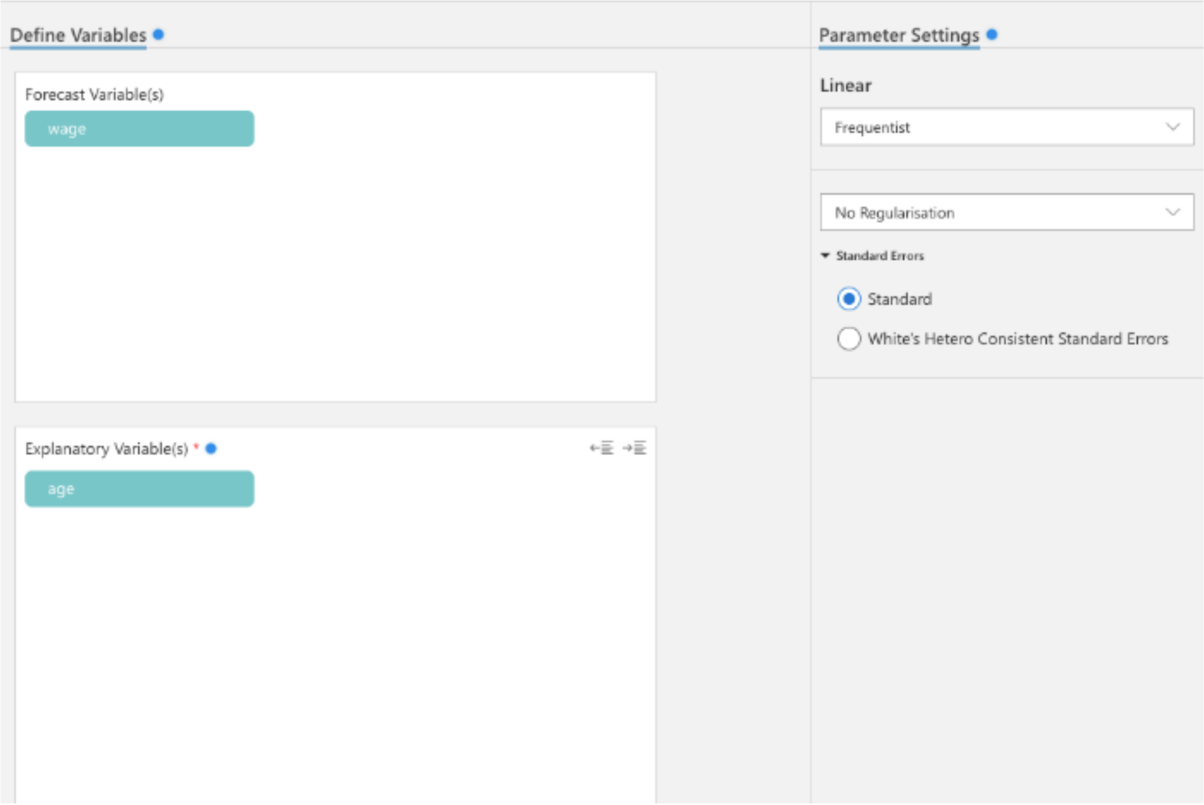
We transform the variables $X_i Xi$ by applying the basic function $b(x)b(x)$, and use these transformed variables to fit the model to add non-linearity to the model and make the splines fit smoother and more flexible non-linearity function. Compared with traditional methods, the cubic splines take into account both smoothness and local effects.

A **natural cubic spline** adds additional constraints, namely that the function is linear outside the boundary knots. This frees up four degrees of freedom (two constraints in each of the two boundary regions), which can be spent more profitably by putting more knots in the interior region. There will be a price paid in bias near the boundary, but assuming the function is linear near the boundaires (where we have less information) is often considered reasonable.

A **natural cubic spline** with *K* knots has *K* degrees of freedom: it can be represented by *K* basis functions. One can start with the basis of a cubic splines, and derive the reduced basis by imposing the boundary constraints. For example, starting from the truncated power series basis, $f(x)=\sum_{j=0}^{3}\beta_jX^j+\sum_{k=1}^{K}\theta_k(X-\xi_K)^3_+$f(x)=∑j=03βjXj+∑k=1kθk(X−ξK)+3The constraints $f''(X)=0$f"(X)=0 and $f^{(3)}(X)=0$f(3)(X)=0 for X < ξ_1ξ1 and X > ξ_kξk lead to the conditions$\beta_2=\beta_3=0,\sum_{k=1}^{k}\theta_k=0,\sum_{k=1}^{k}\xi_k\theta_k=0$β2=β3=0,∑k=1kθk=0,∑k=1kξkθk=0These conditions are automatically satisfied by choosing the following basis, $N_1(X)=1,N_2(X)=X$N1(X)=1,N2(X)=X$N_{k+2}(X)=d_k(X)-d_{k-1}(X),k=1,\ldots,K-2$Nk+2(X)=dk(X)-dk-1(X),k=1,...,K-2with $d_k=\frac{(X-\xi_k)^3_+-(X-\xi_K)^3_+}{\xi_K-\xi_k}$dk=(X-ξk)3-(X-ξK)+3ξK-ξk

Linear Regression

Choosing Linear Regression with the default setting -> Click Analyse



Define Variables

Forecast Variable(s): wage

Explanatory Variable(s) *: age

Parameter Settings

Linear: Frequentist

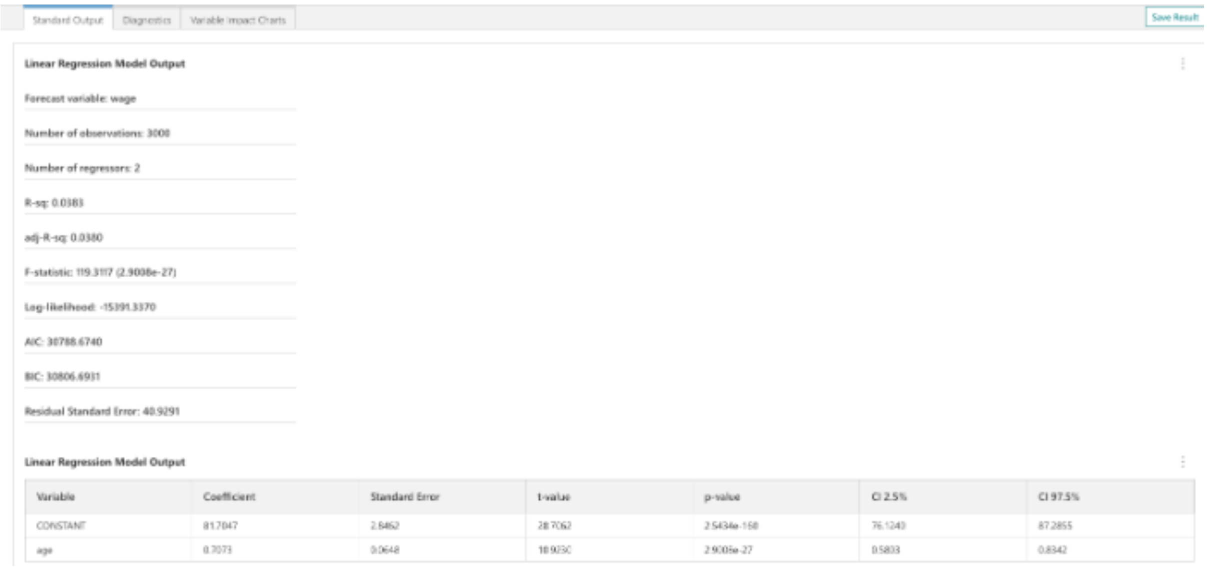
No Regularisation

Standard Errors

● Standard

○ White's Hetero Consistent Standard Errors

The Module Output is shown below:



Linear Regression Model Output

Forecast variable: wage

Number of observations: 3000

Number of regressors: 2

R-sq: 0.0383

adj-R-sq: 0.0380

F-statistic: 119.3117 (2.9008e-27)

Log-likelihood: -15391.3370

AIC: 30788.6740

BIC: 30806.6931

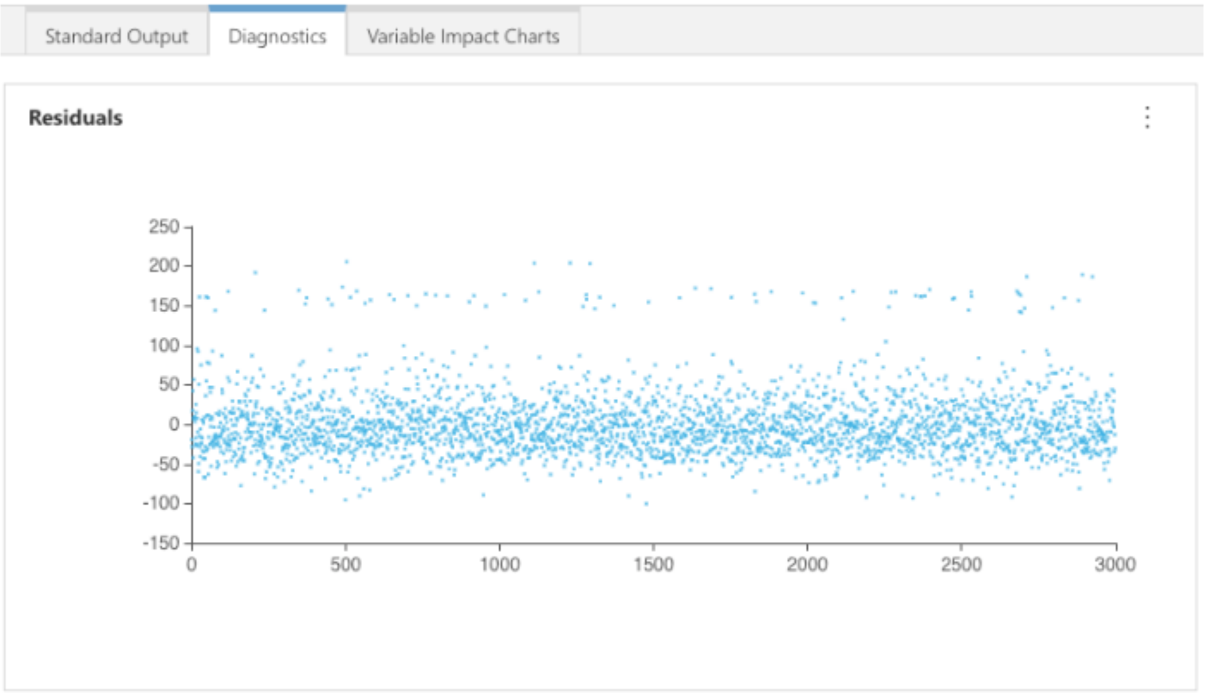
Residual Standard Error: 40.9291

Linear Regression Model Output

Variable	Coefficient	Standard Error	t-value	p-value	CI 2.5%	CI 97.5%
CONSTANT	81.7047	2.8462	28.7062	2.5434e-158	76.1240	87.2855
age	0.7073	0.0648	10.9230	2.9008e-27	0.5803	0.8342

AIC and BIC are the standard to see if the model fits well. The smaller AIC and BIC are, the better. The output of AIC and BIC are quite large in this model.

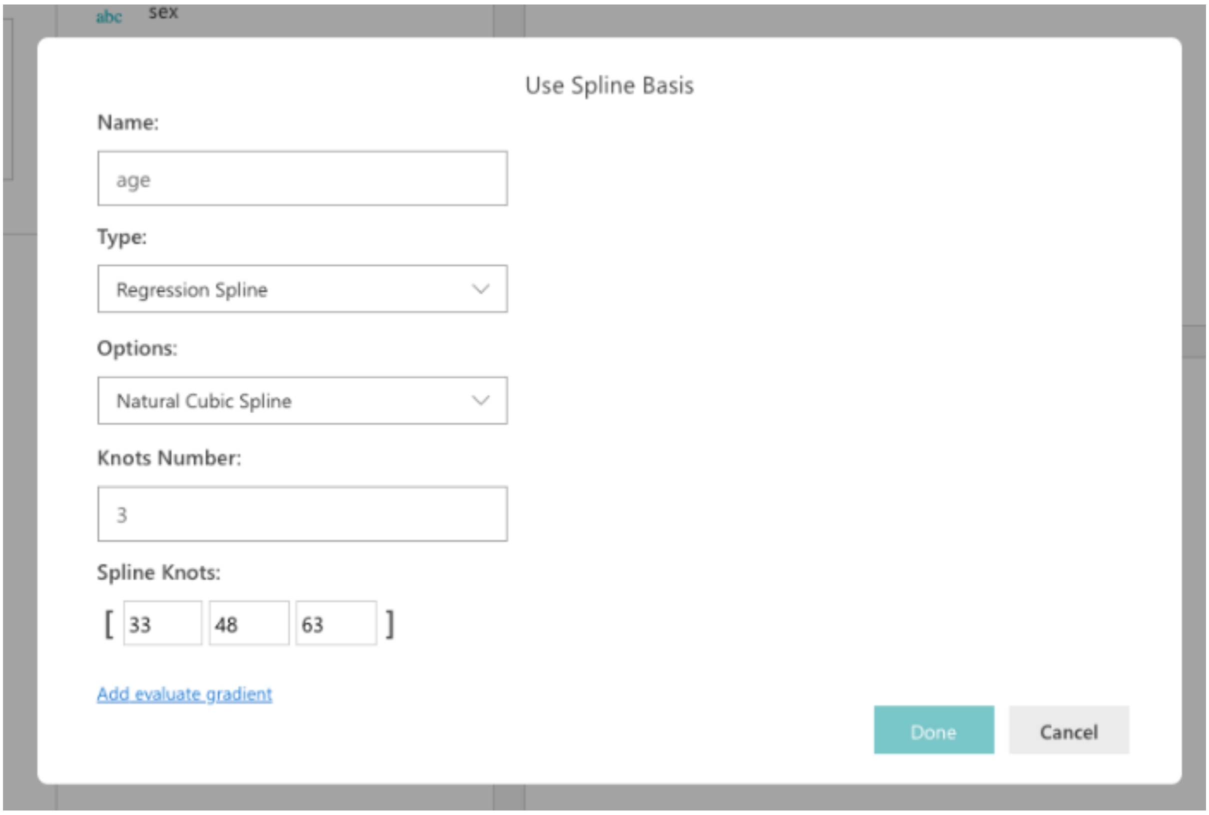
Go to Diagnostics



Spline

For deeper exploring the relationship of them. We need to make age and wage variables spline first.

We choose the knots Number as 3:



Use Spline Basis

Name: age

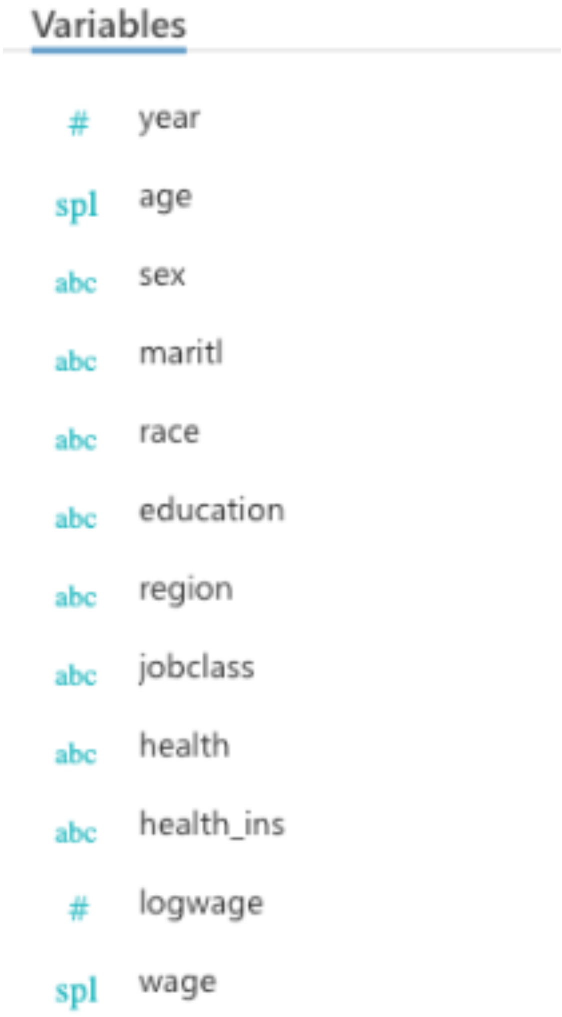
Type: Regression Spline

Options: Natural Cubic Spline

Knots Number: 3

Spline Knots: [33 48 63]

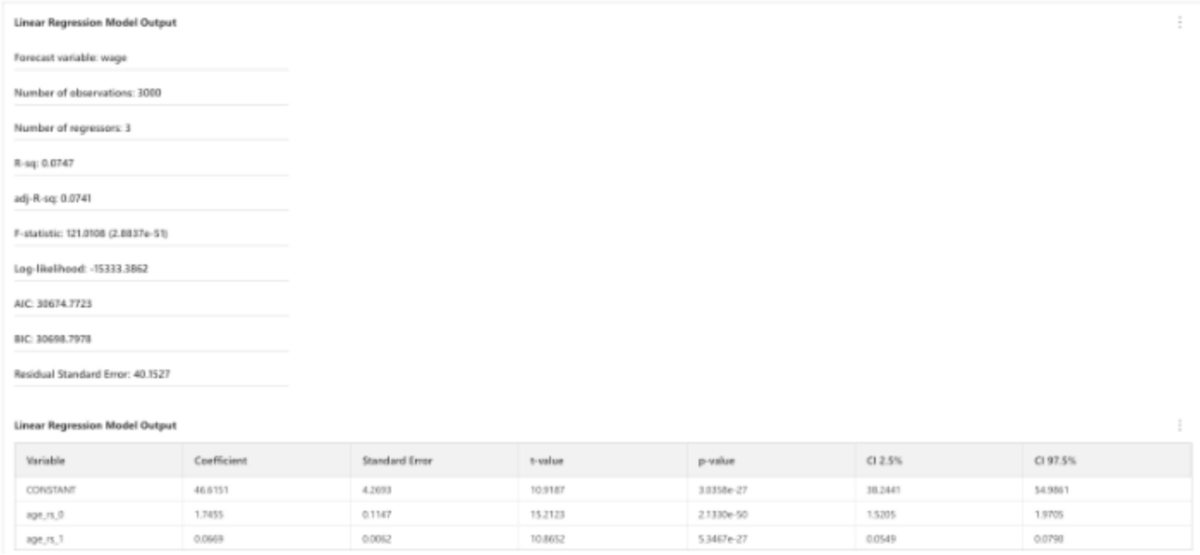
Then the variable will look like the screenshot below



Variables

#	year
spl	age
abc	sex
abc	maritl
abc	race
abc	education
abc	region
abc	jobclass
abc	health
abc	health_ins
#	logwage
spl	wage

The output shows better than before by comparing the AIC and BIC.



Linear Regression Model Output

Forecast variable: wage

Number of observations: 3000

Number of regressors: 3

R-sq: 0.0747

adj-R-sq: 0.0741

F-statistic: 121.0108 (2.8837e-51)

Log-likelihood: -15333.3862

AIC: 30674.7723

BIC: 30698.7978

Residual Standard Error: 40.1527

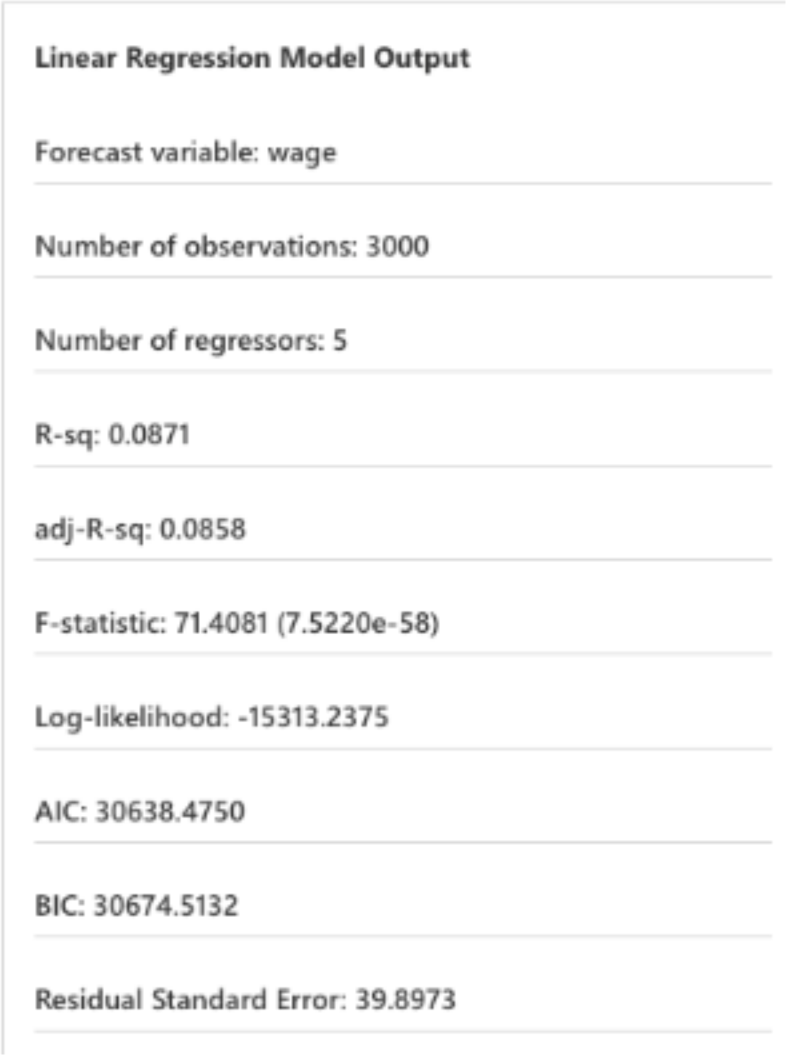
Linear Regression Model Output

Variable	Coefficient	Standard Error	t-value	p-value	CI 2.5%	CI 97.5%
CONSTANT	46.6151	4.2693	10.9187	3.0358e-27	38.2441	54.9861
age_rs_0	1.7455	0.1147	15.2123	2.1330e-50	1.5205	1.9705
age_rs_1	0.0669	0.0062	10.8652	5.3467e-27	0.0549	0.0790

Knot Selection The shape of spline can be controlled by carefully selecting the number of knots and their extraction location to: 1.

Allowing flexibility where the trend changes rapidly 2. Avoiding overfitting where the trend changes are small.

Let change the Knots Number from 3 to 5:



Linear Regression Model Output

Forecast variable: wage

Number of observations: 3000

Number of regressors: 5

R-sq: 0.0871

adj-R-sq: 0.0858

F-statistic: 71.4081 (7.5220e-58)

Log-likelihood: -15313.2375

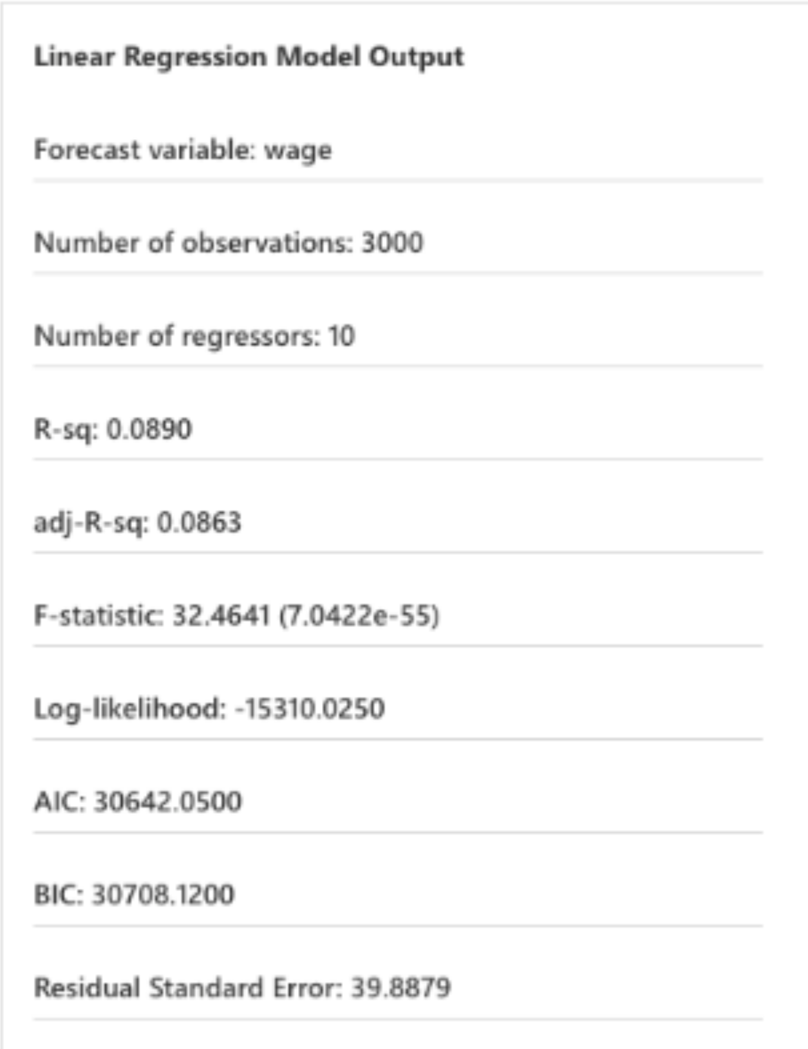
AIC: 30638.4750

BIC: 30674.5132

Residual Standard Error: 39.8973

The performance slightly perform better than before.

How about changing the Knots Number from 5 to 10, is the more knots the better? Let's have a look



Linear Regression Model Output

Forecast variable: wage

Number of observations: 3000

Number of regressors: 10

R-sq: 0.0890

adj-R-sq: 0.0863

F-statistic: 32.4641 (7.0422e-55)

Log-likelihood: -15310.0250

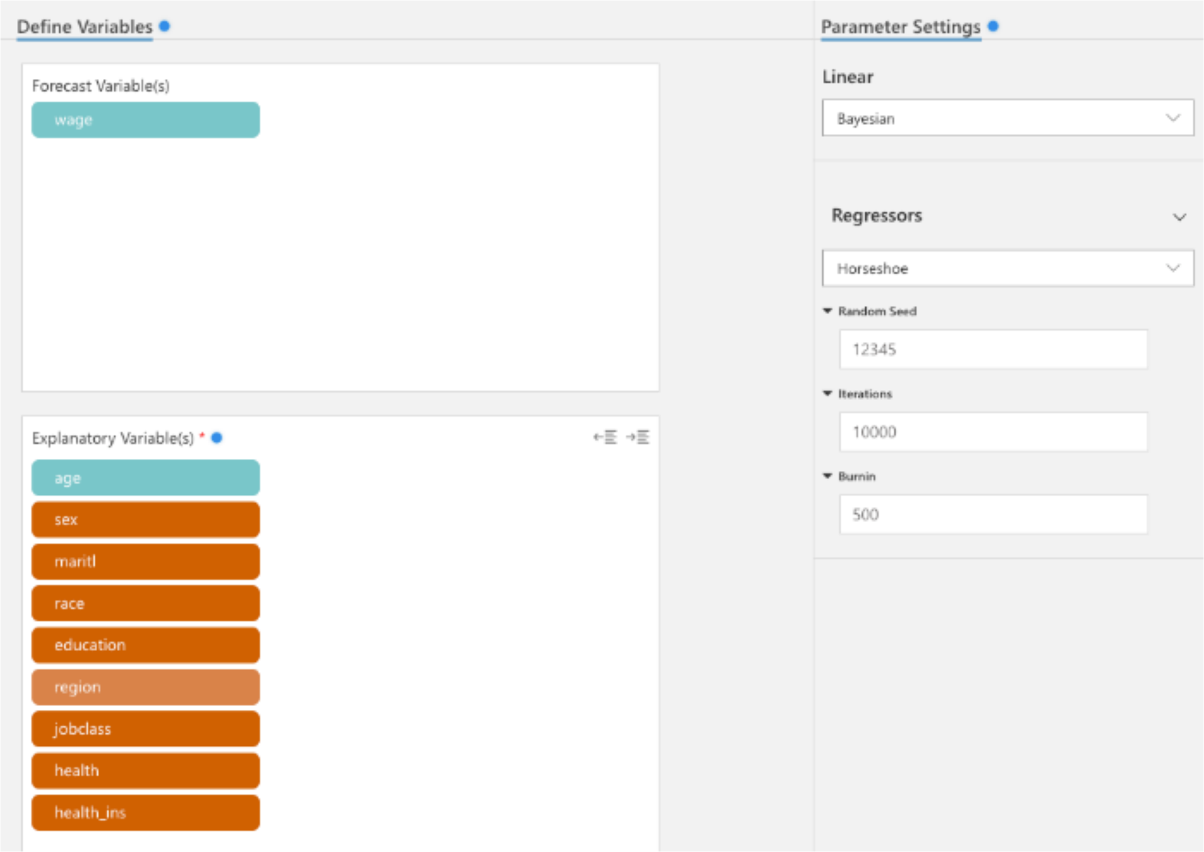
AIC: 30642.0500

BIC: 30708.1200

Residual Standard Error: 39.8879

The output performs worse than before. It seems the Knots Number is overfitting. In this case, we use the Knots Number as 5 for further exploration.

Now we are using Bayesian linear with horseshoe regressors



Forecast Variable(s)

wage

Explanatory Variable(s) *

age

sex

maritl

race

education

region

jobclass

health

health_ins

Linear: Bayesian

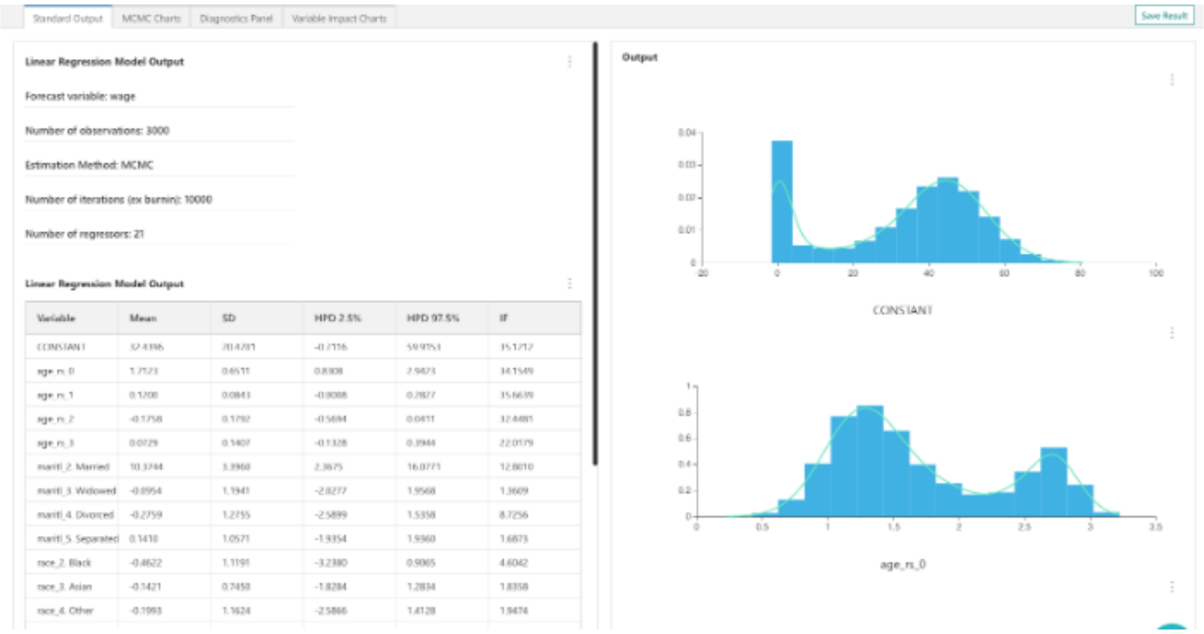
Regressors: Horseshoe

Random Seed: 12345

Iterations: 10000

Burnin: 500

The output shows below



Linear Regression Model Output

Forecast variable: wage

Number of observations: 3000

Estimation Method: MCMC

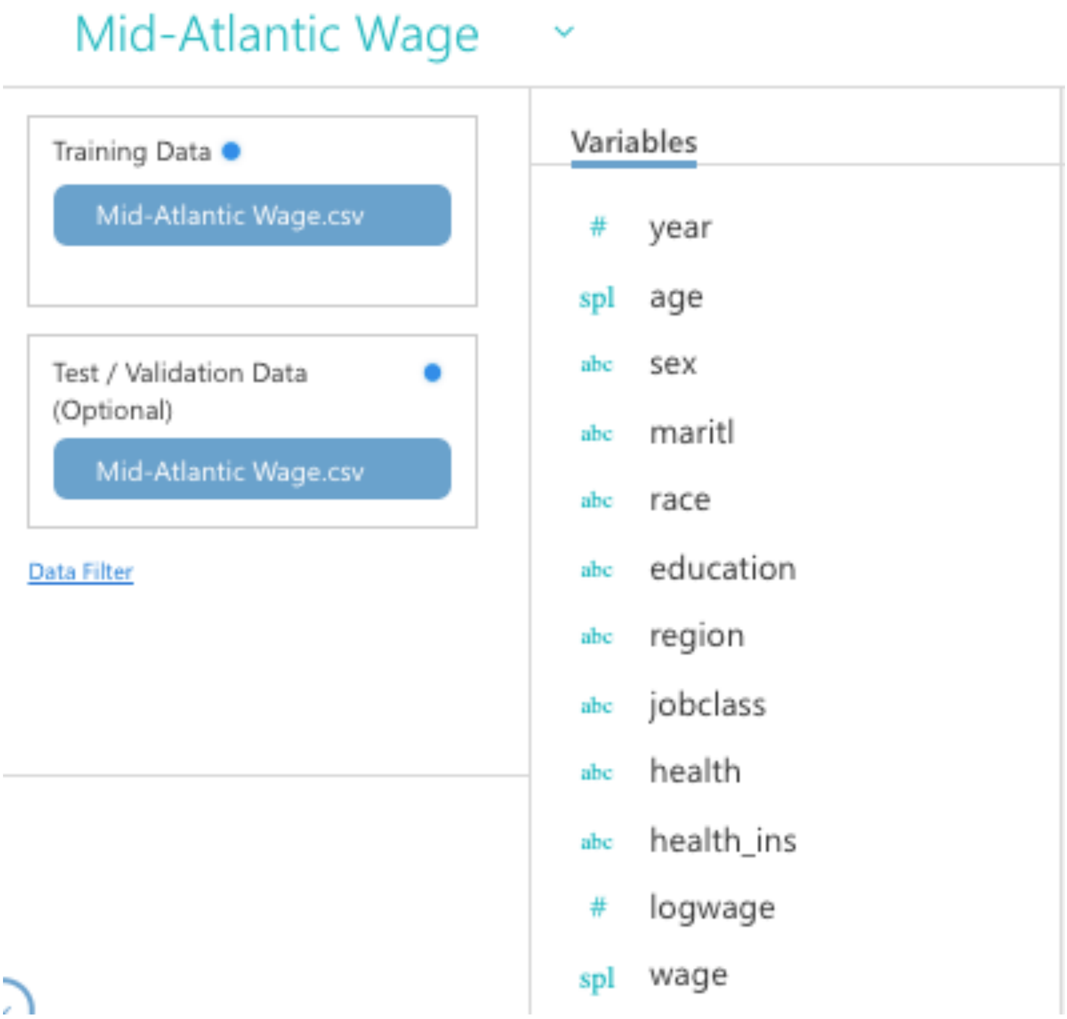
Number of iterations (ex burnin): 10000

Number of regressors: 21

Linear Regression Model Output

Variable	Mean	SD	HPD 2.5%	HPD 97.5%	IF
CONSTANT	32.4396	20.4201	-0.7116	59.9153	35.1212
age_ns_0	1.7123	0.6511	0.8908	2.9423	34.1549
age_ns_1	0.1200	0.0843	-0.0088	0.2827	35.6639
age_ns_2	-0.1758	0.1792	-0.5694	0.0411	32.4481
age_ns_3	0.0729	0.1407	-0.1328	0.3944	22.0179
maritl_2. Married	10.3744	3.3968	2.3675	16.0771	12.8010
maritl_3. Widowed	-0.0954	1.1941	-2.0277	1.9568	1.3609
maritl_4. Divorced	-0.2759	1.2755	-2.5899	1.5358	8.7256
maritl_5. Separated	0.1410	1.0571	-1.9354	1.9960	1.6873
race_2. Black	-0.4622	1.1191	-3.2380	0.9065	4.6042
race_3. Asian	-0.1421	0.7458	-1.8284	1.2834	1.8358
race_4. Other	-0.1993	1.1624	-2.5866	1.4128	1.9474

Let's try to use the same dataset as training data and test data to see the prediction with all variables



Mid-Atlantic Wage

Training Data ●

Mid-Atlantic Wage.csv

Test / Validation Data ●
(Optional)

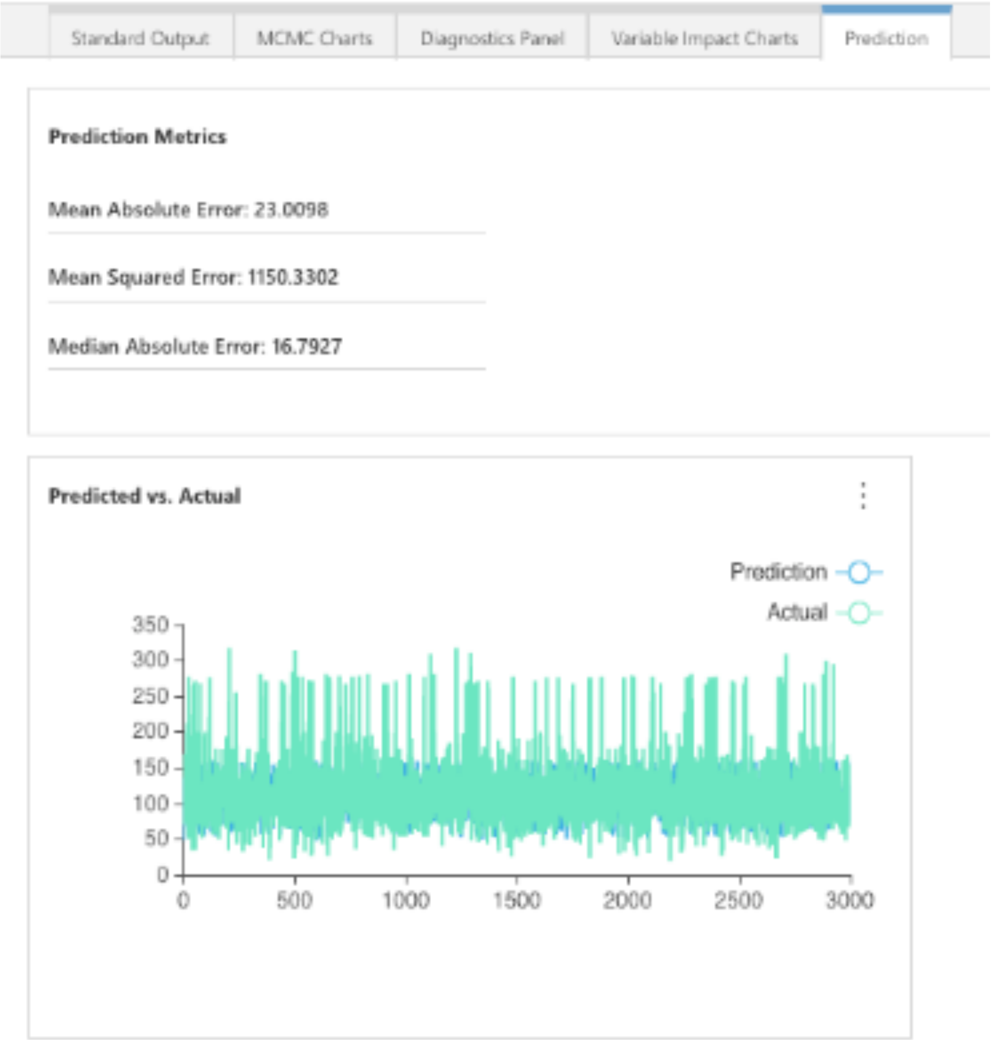
Mid-Atlantic Wage.csv

Data Filter

Variables

#	year
spl	age
abc	sex
abc	maritl
abc	race
abc	education
abc	region
abc	jobclass
abc	health
abc	health_ins
#	logwage
spl	wage

Horseshoe



Prediction Metrics

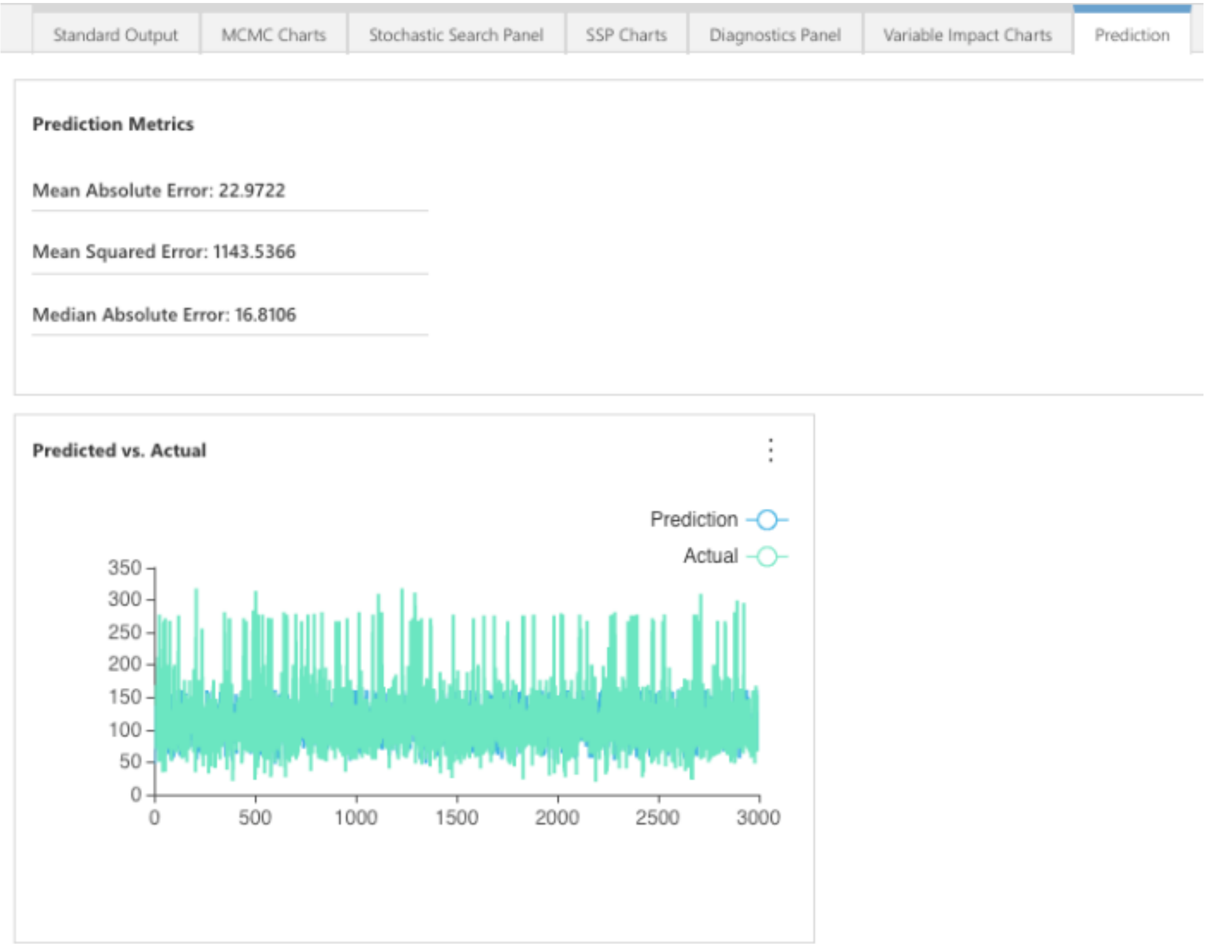
Mean Absolute Error: 23.0098

Mean Squared Error: 1150.3302

Median Absolute Error: 16.7927

Predicted vs. Actual

G-prior Spike Slab



Prediction Metrics

Mean Absolute Error: 22.9722

Mean Squared Error: 1143.5366

Median Absolute Error: 16.8106

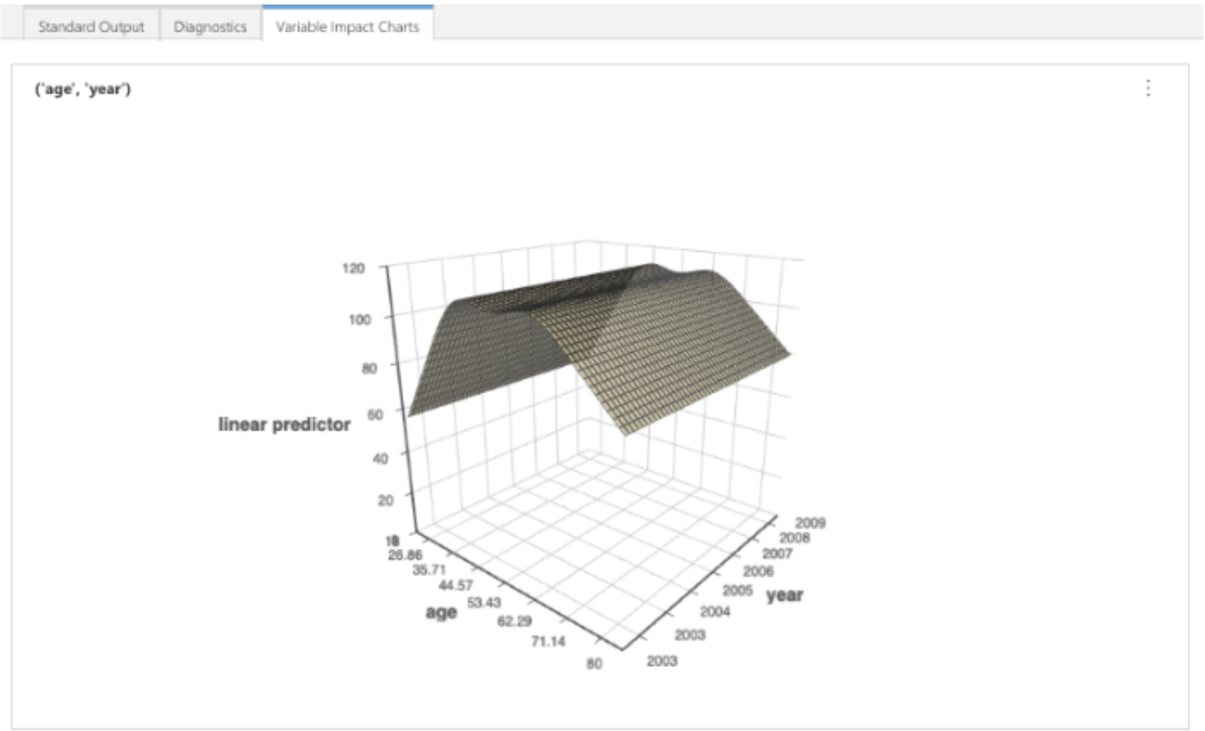
Predicted vs. Actual

The prediction of all variables is quite similar. G-prior spike slab MAE and MSE are slightly lower than horseshoe, but not too much difference between horseshoe and G-prior spike slab in this case.

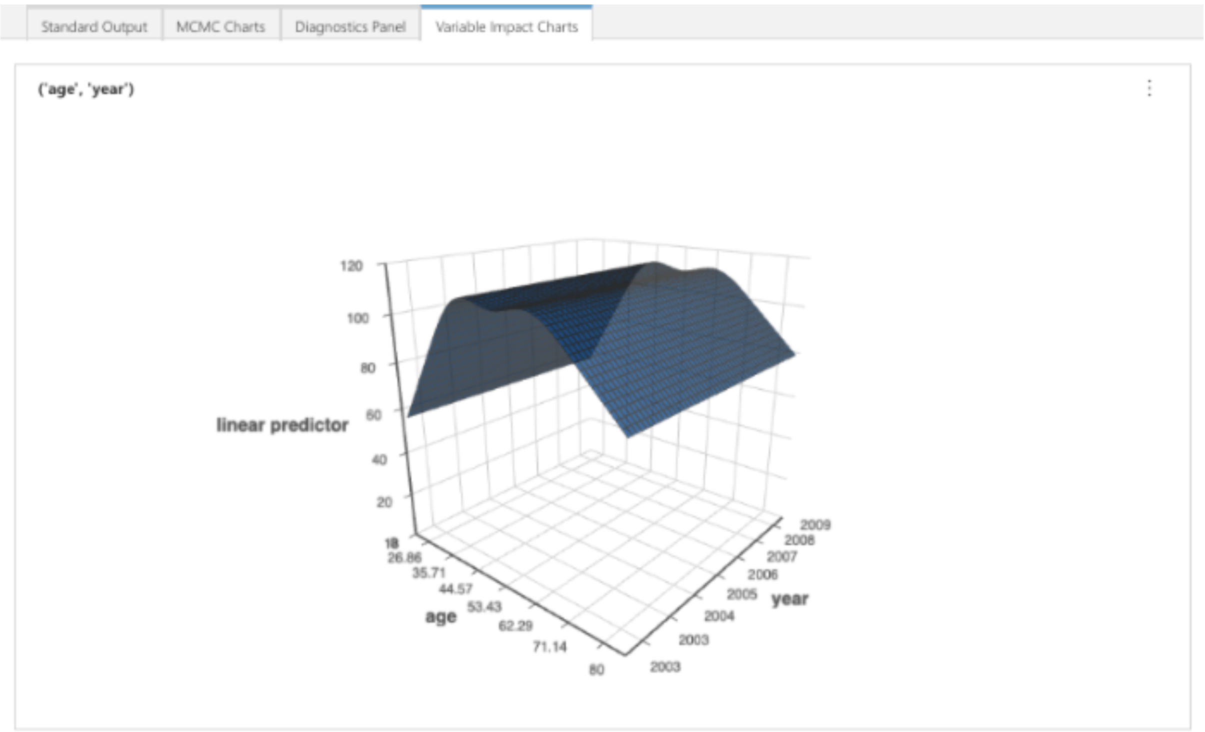
We want to explore the relationship between wage and interaction of age and year.

Frequentist

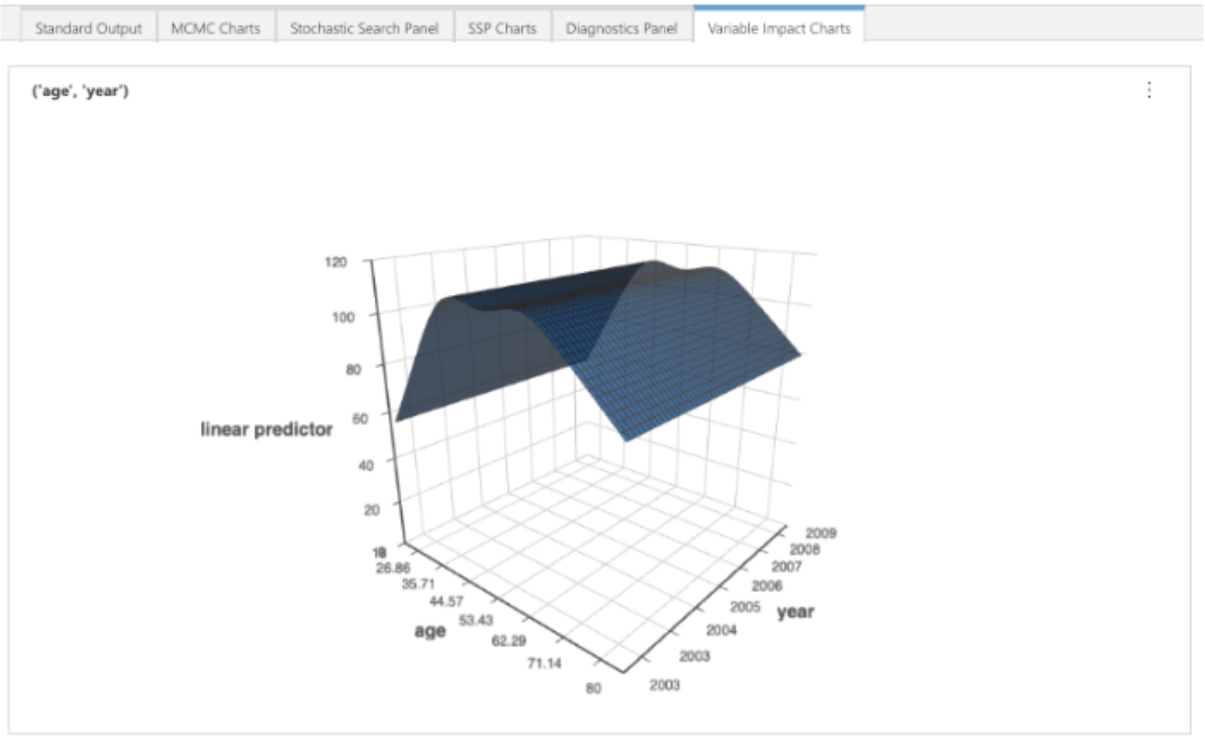
Go to Variable Impact Charts



Horseshoe



G-Prior Spike Slab

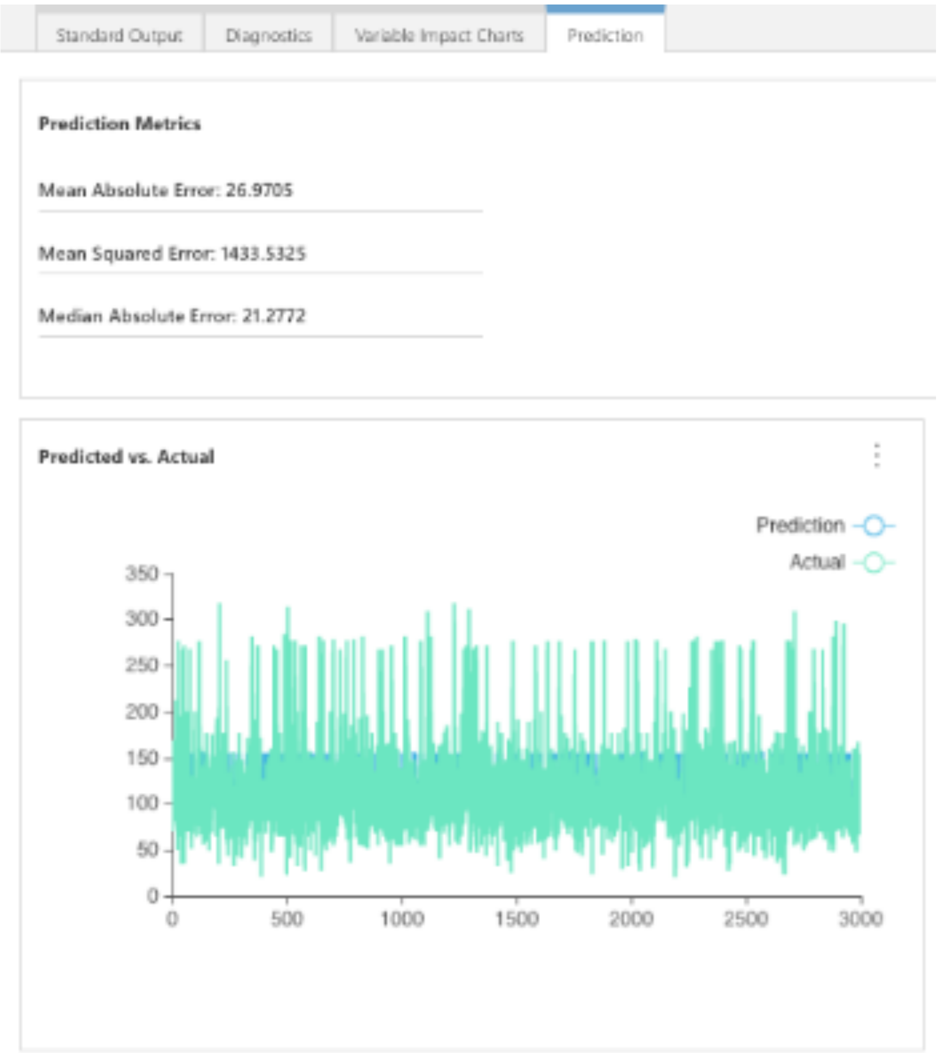


The three charts above seem quite similar. The year have slightly influence to wage while age around 50 has a slightly lower wage.

We will explore the interaction of education and job class related to wage

Frequentist

Go to Prediction



Prediction Metrics

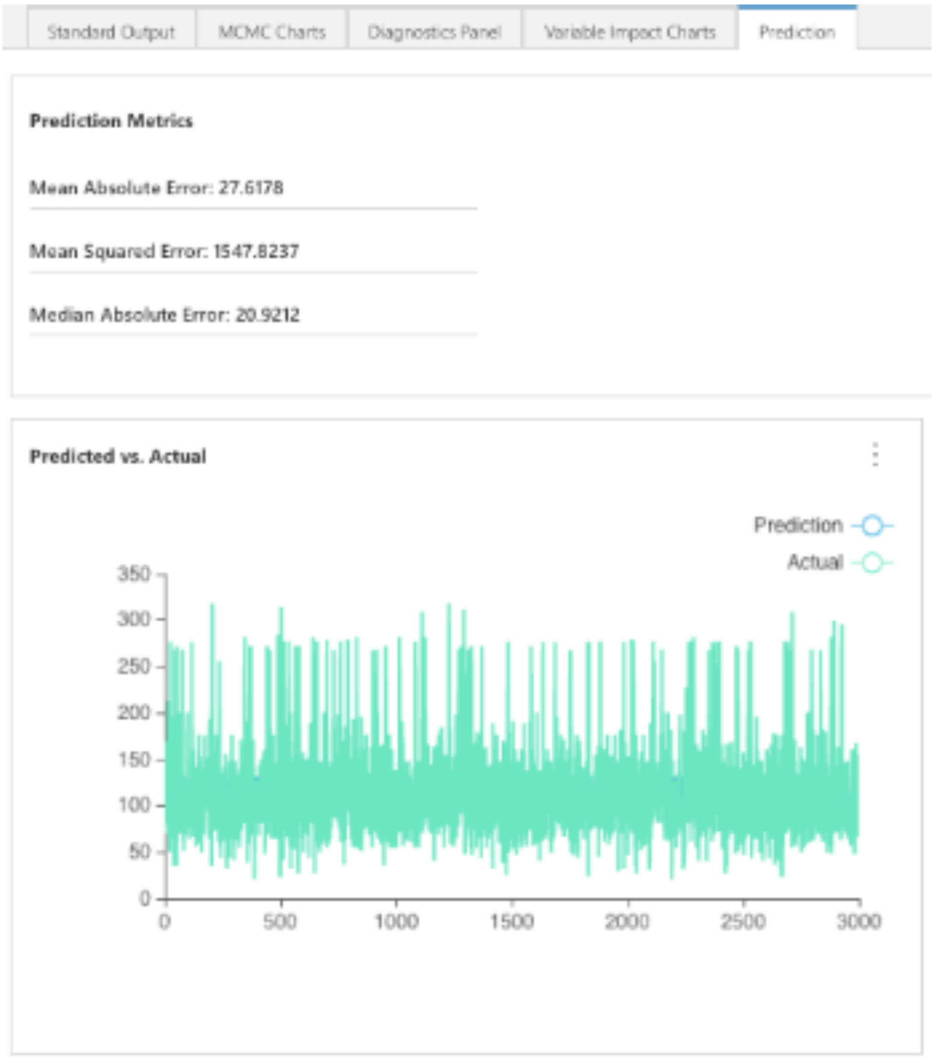
Mean Absolute Error: 26.9705

Mean Squared Error: 1433.5325

Median Absolute Error: 21.2772

Predicted vs. Actual

Horseshoe



Prediction Metrics

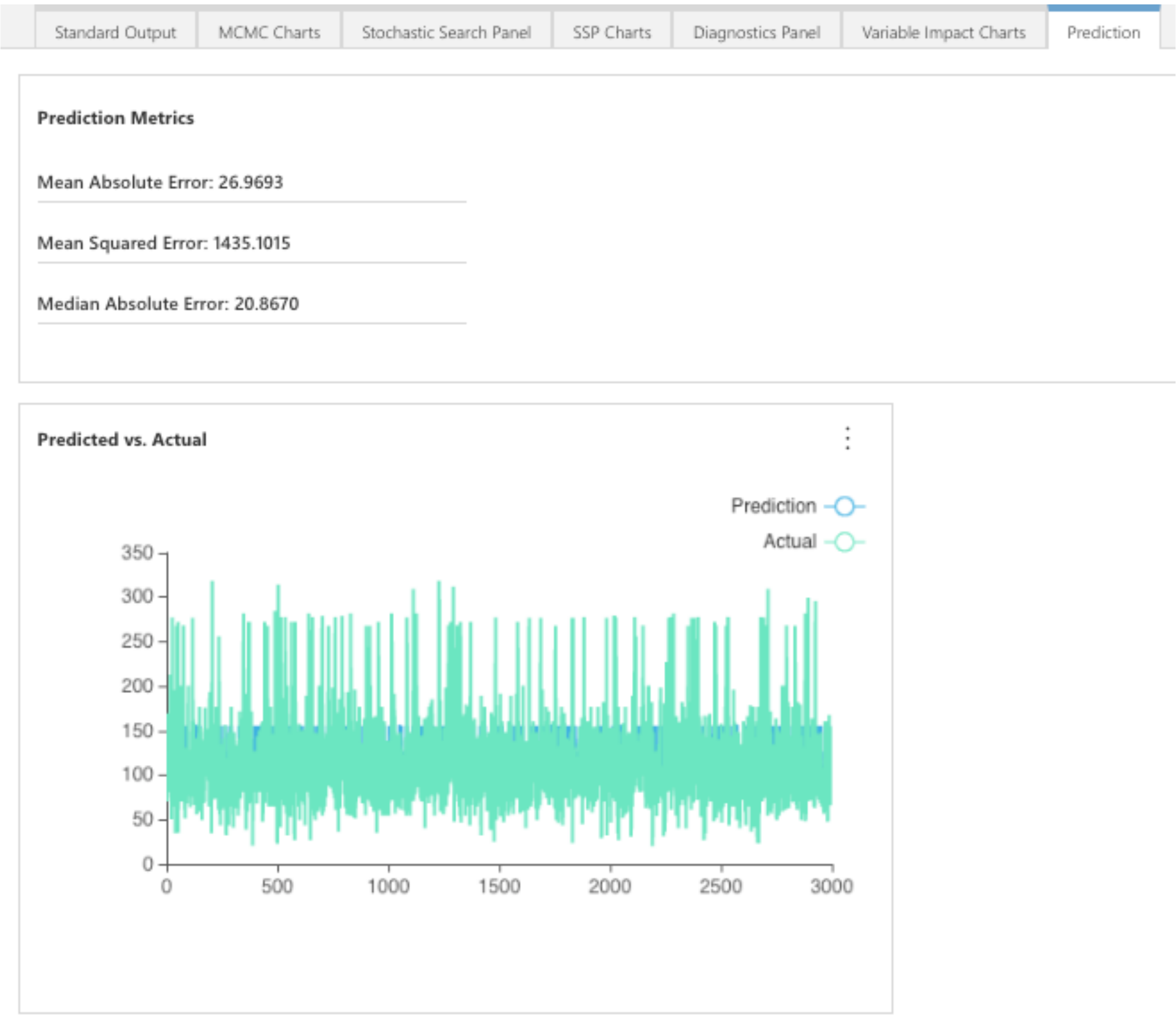
Mean Absolute Error: 27.6178

Mean Squared Error: 1547.8237

Median Absolute Error: 20.9212

Predicted vs. Actual

G-Prior Spike Slab



Prediction Metrics

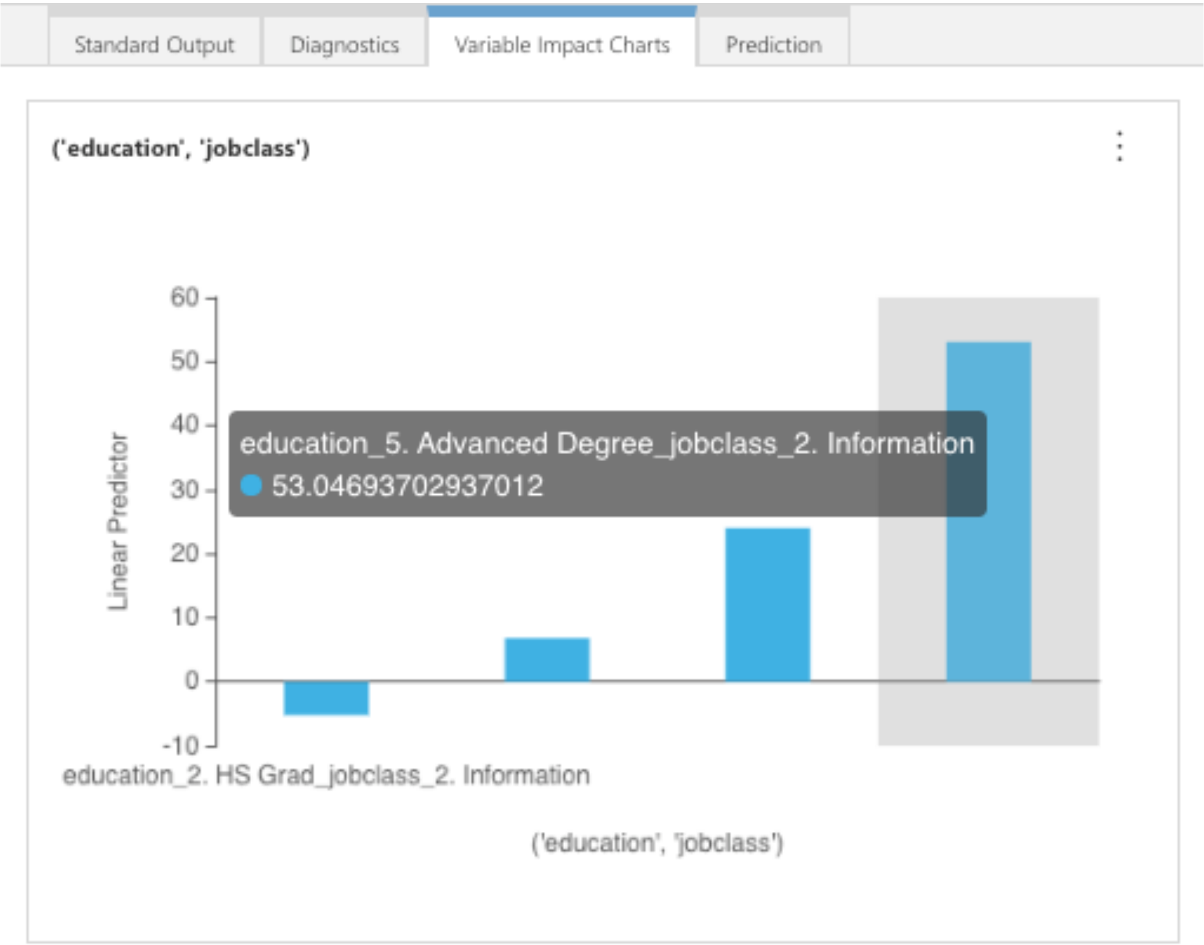
Mean Absolute Error: 26.9693

Mean Squared Error: 1435.1015

Median Absolute Error: 20.8670

Predicted vs. Actual

Go to variable impact charts


Standard Output
Diagnostics
Variable Impact Charts
Prediction
('education', 'jobclass')
Linear Predictor
60
50
40
30
20
10
0
-10
education_5. Advanced Degree_jobclass_2. Information
53.04693702937012
education_2. HS Grad_jobclass_2. Information
('education', 'jobclass')

The chart shows that lower than HS Grad education in information job class has a slightly lower wage than industry job class. With advanced degree education in information job class have much higher wage than in industry job class.

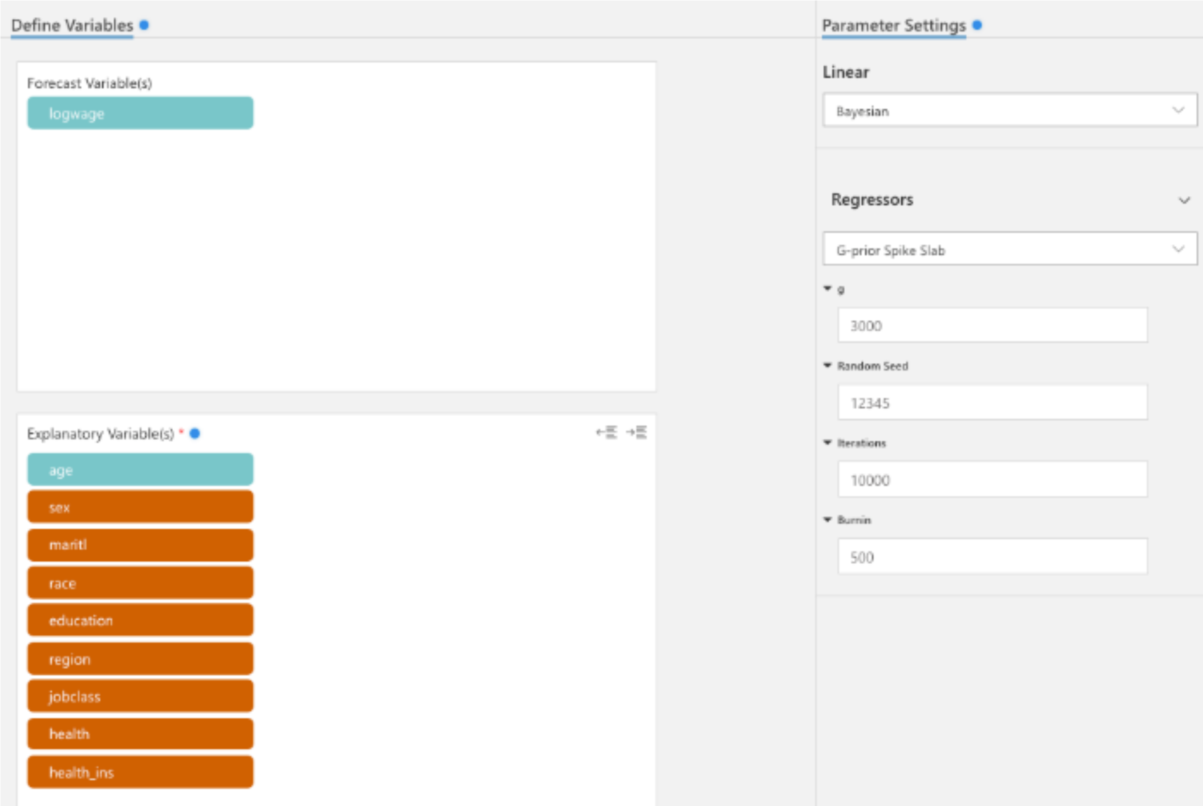
Log transformation

When you have a non-linear relationship, you can also try a logarithm transformation of the forecast variables, especially there have more than one explanatory variables. Using log forecast variable with more than one explanatory variables, each parameter is interpreted as a partial derivative, or the change in the dependent variable for a change in the explanatory variable, holding all other variables constant. For a range of economic variables, substantial forecasting improvements from taking logs are found if the log transformation actually stabilizes the variance of the underlying series.

Let's choose all variables for the comparison of wage and logwage

G-prior Spike Slab

Selecting all variables and setting parameters



Define Variables

Forecast Variable(s)

logwage

Explanatory Variable(s) *

age

sex

maritl

race

education

region

jobclass

health

health_ins

Parameter Settings

Linear

Bayesian

Regressors

G-prior Spike Slab

g

3000

Random Seed

12345

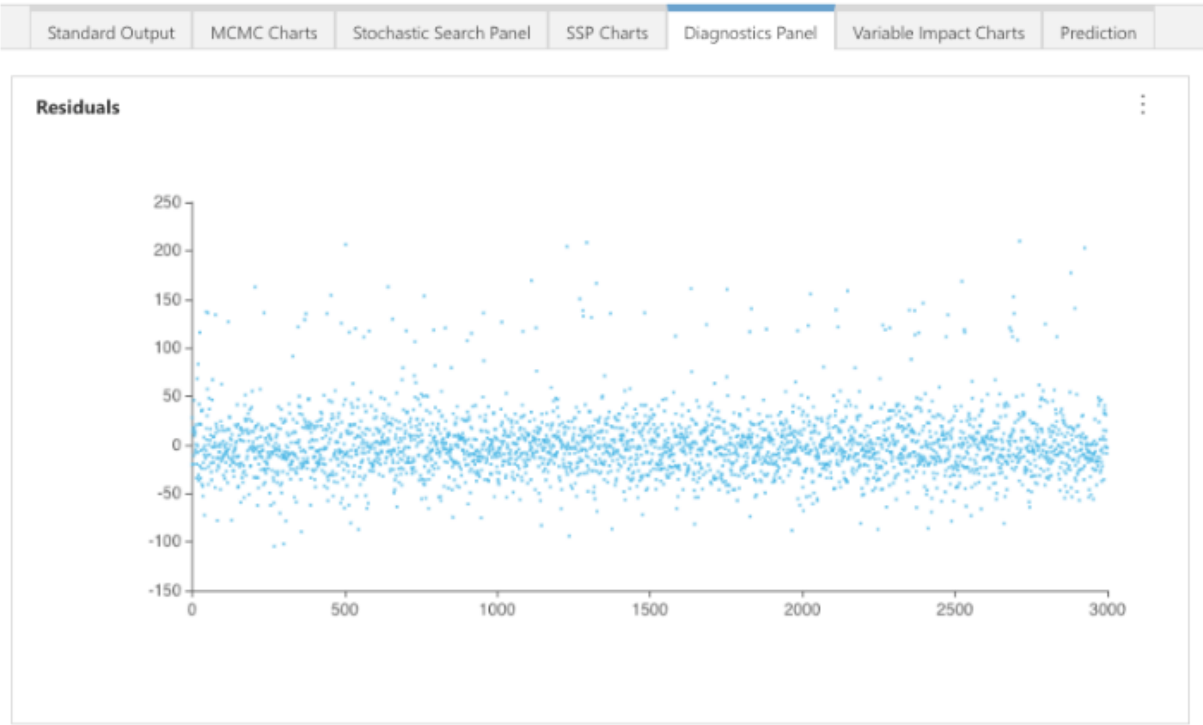
Iterations

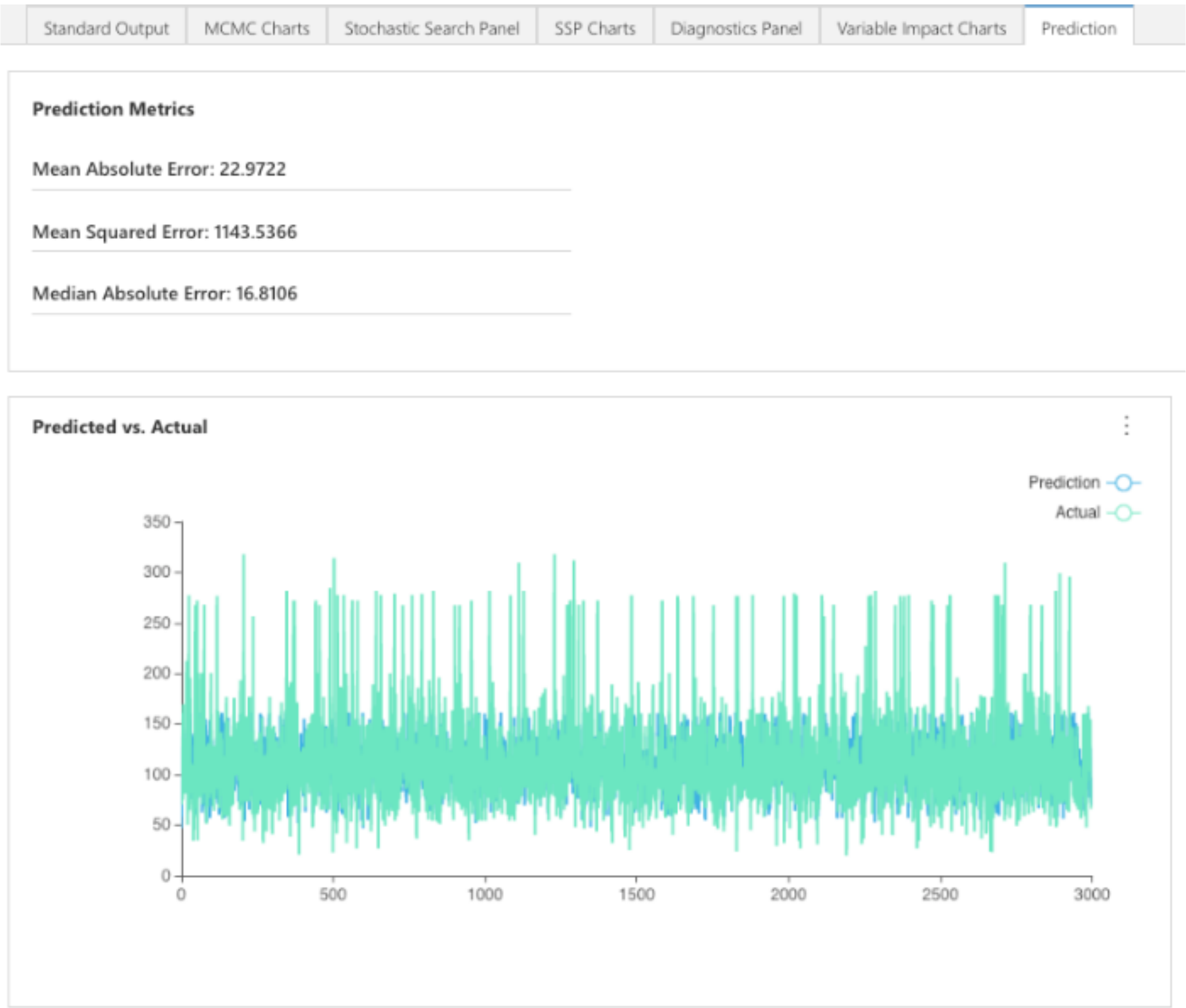
10000

Burnin

500

Wage





Prediction Metrics

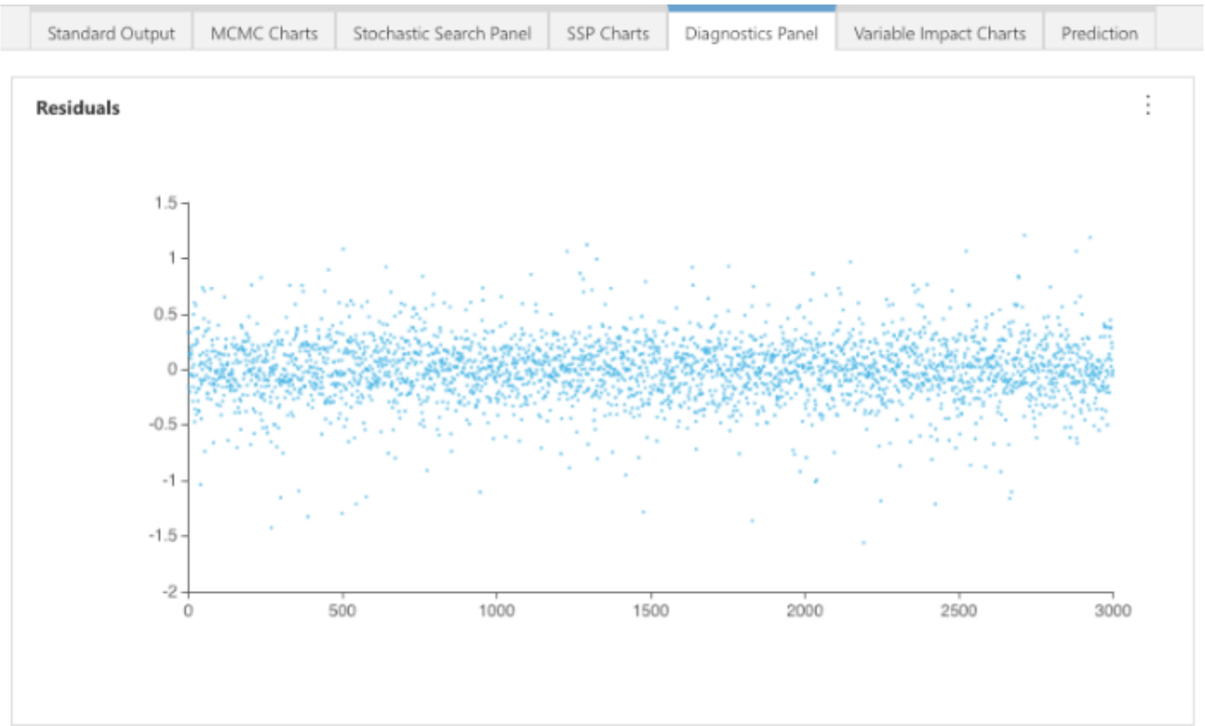
Mean Absolute Error: 22.9722

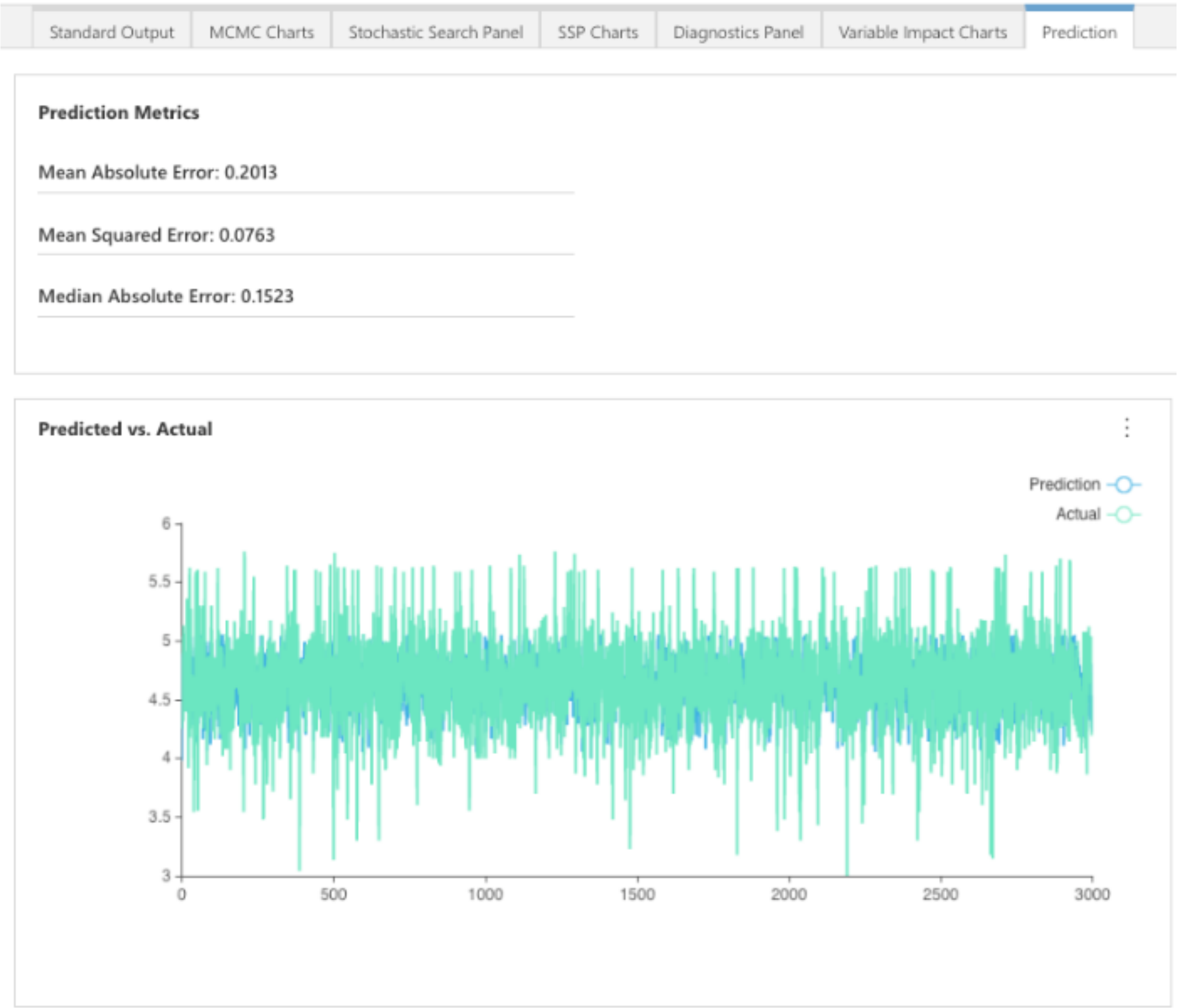
Mean Squared Error: 1143.5366

Median Absolute Error: 16.8106

Predicted vs. Actual

Logwage





Prediction Metrics

Mean Absolute Error: 0.2013

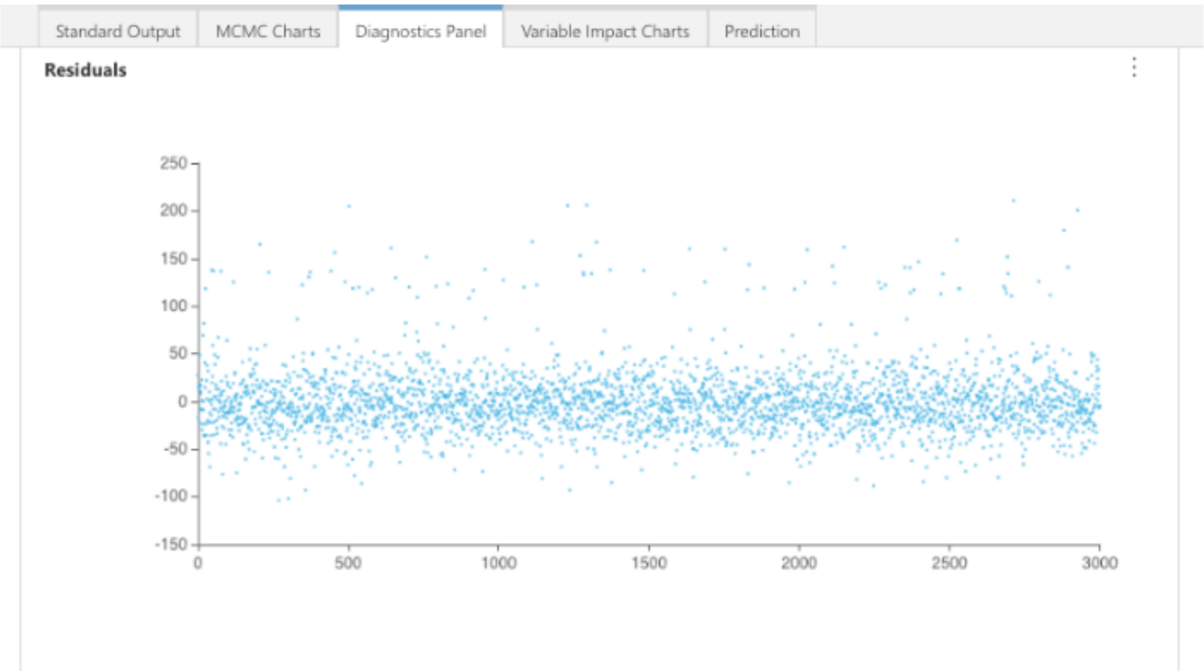
Mean Squared Error: 0.0763

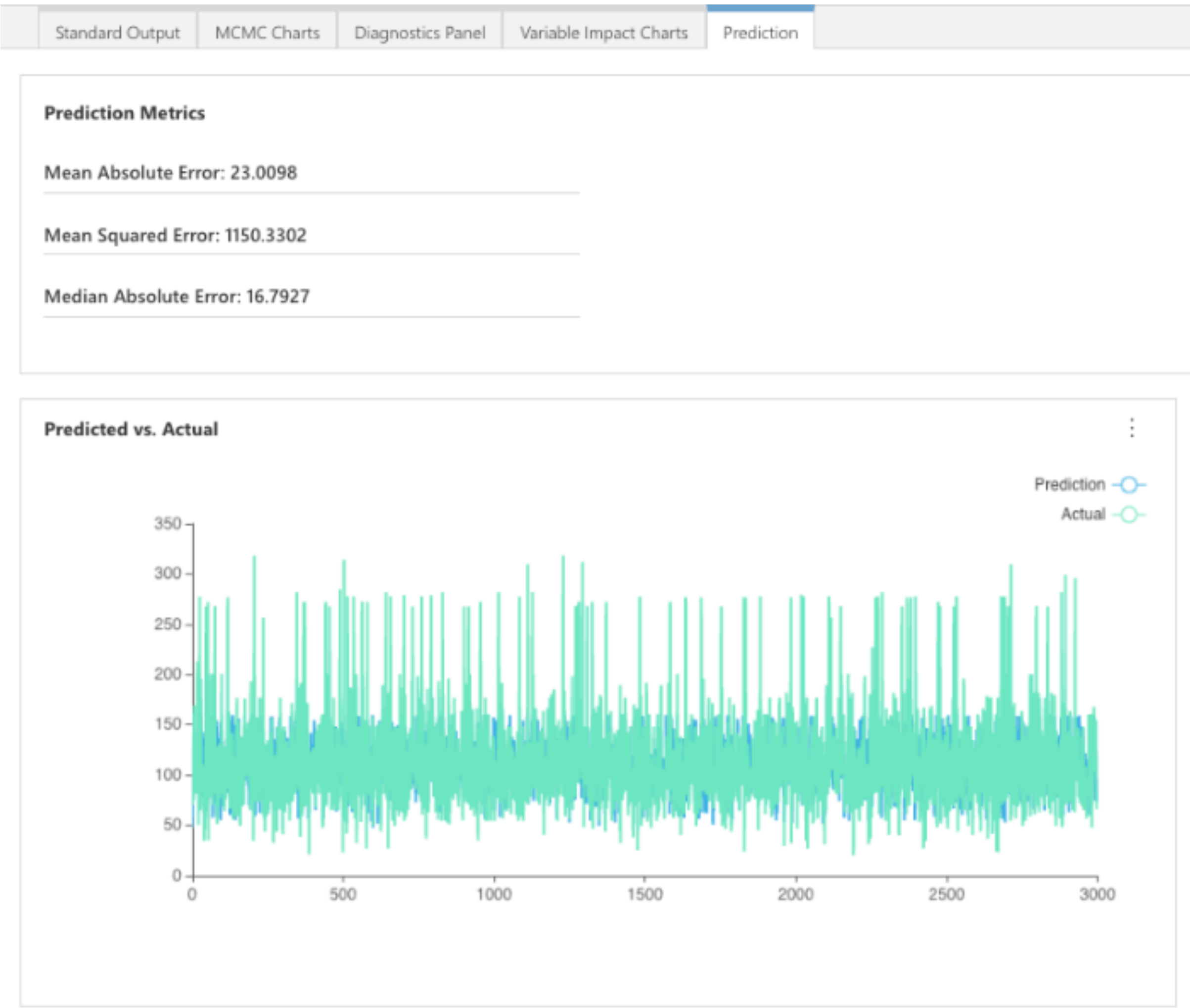
Median Absolute Error: 0.1523

In Diagnostics Panel, the residuals in logwage perform much better than wage in G-prior spike slab. In Prediction, the MAE and MSE in logwage also perform much better than wage. Using log transforming in wage variable for forecasting, the performance has great improvement in this case.

horseshoe

Wage





Prediction Metrics

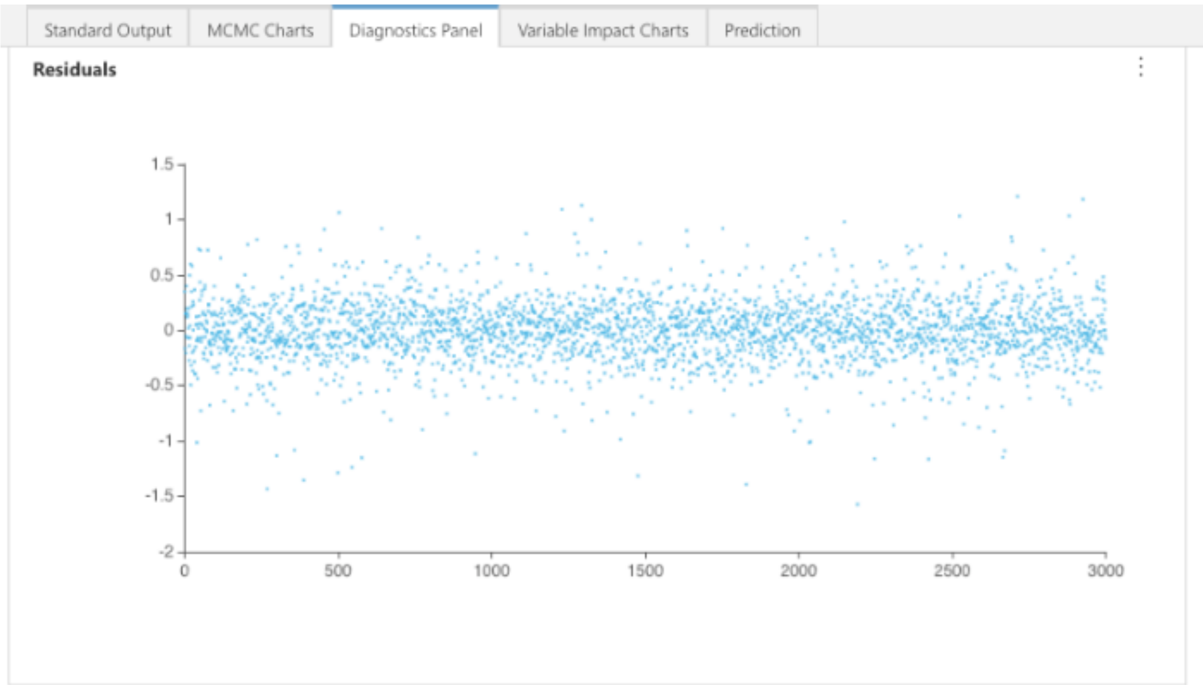
Mean Absolute Error: 23.0098

Mean Squared Error: 1150.3302

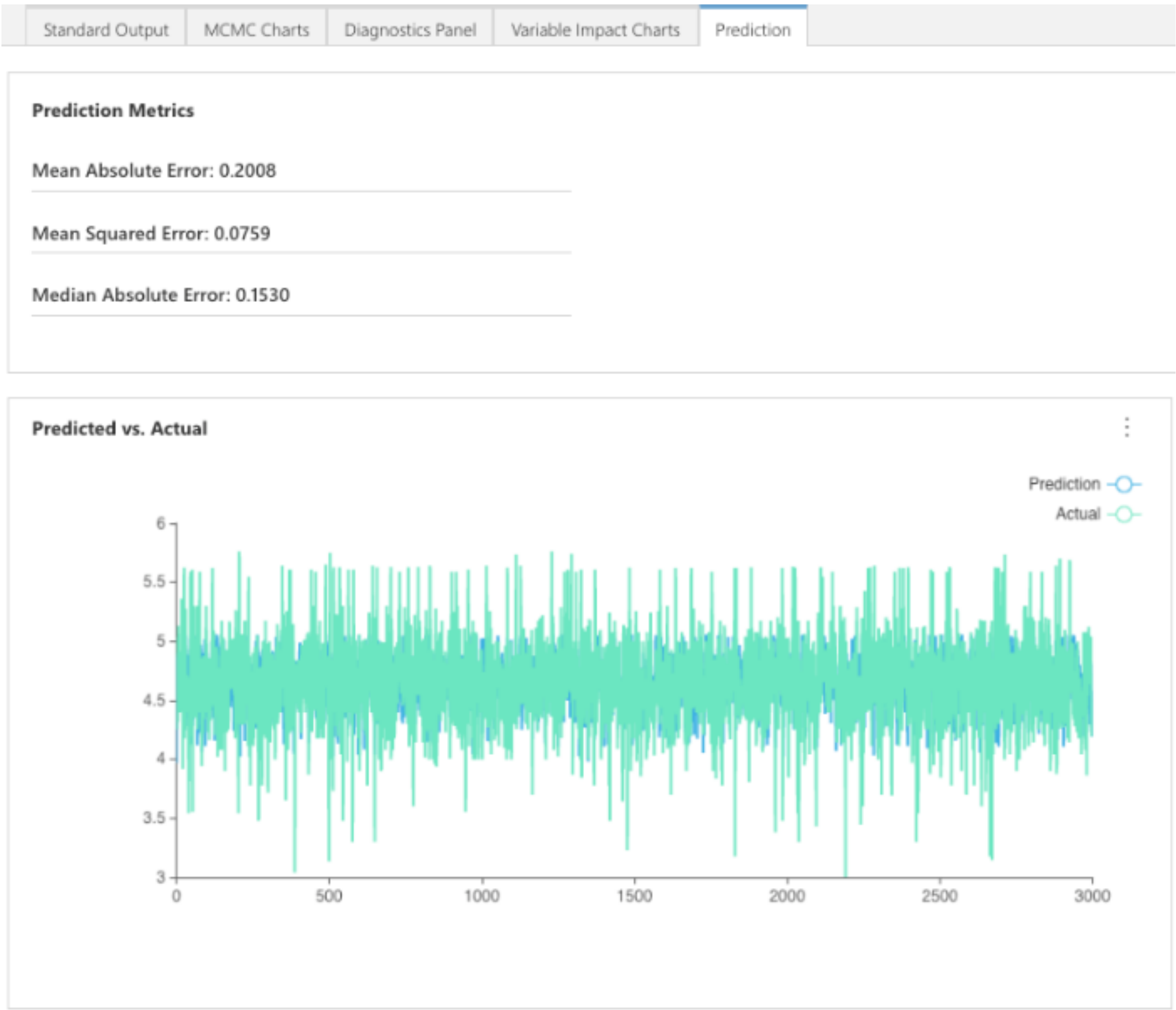
Median Absolute Error: 16.7927

Predicted vs. Actual

Logwage



Residuals



Prediction Metrics

Mean Absolute Error: 0.2008

Mean Squared Error: 0.0759

Median Absolute Error: 0.1530

Predicted vs. Actual

In Diagnostics Panel, the residuals in logwage perform much better than wage in horseshoe. In Prediction, the MAE and MSE in logwage also perform much better than wage. Using log transforming in wage variable for forecasting, the performance has great improvement in this case.