FACTORS CONTRIBUTING TO HEALTH CARE COST AMONG THE ELDERLY

1 RESEARCH QUESTION

What factors contributing to health care expenditures will help identify what sort of programs to implement to reduce future expenditures for the elderly

2 UNDERSTANDING THE DATA

2.1 DATA SOURCE

The data comes from the 2005 Medical Expenditures Panel Survey. We explore a total of 13 variables that impact the total expanse of health care among the elderly

2.1.1 Exploring the Response Variable: Total Expense

The total expense is right-skewed, as shown in Fig 1. Using a log transformation on the variable provided a more normal distribution.

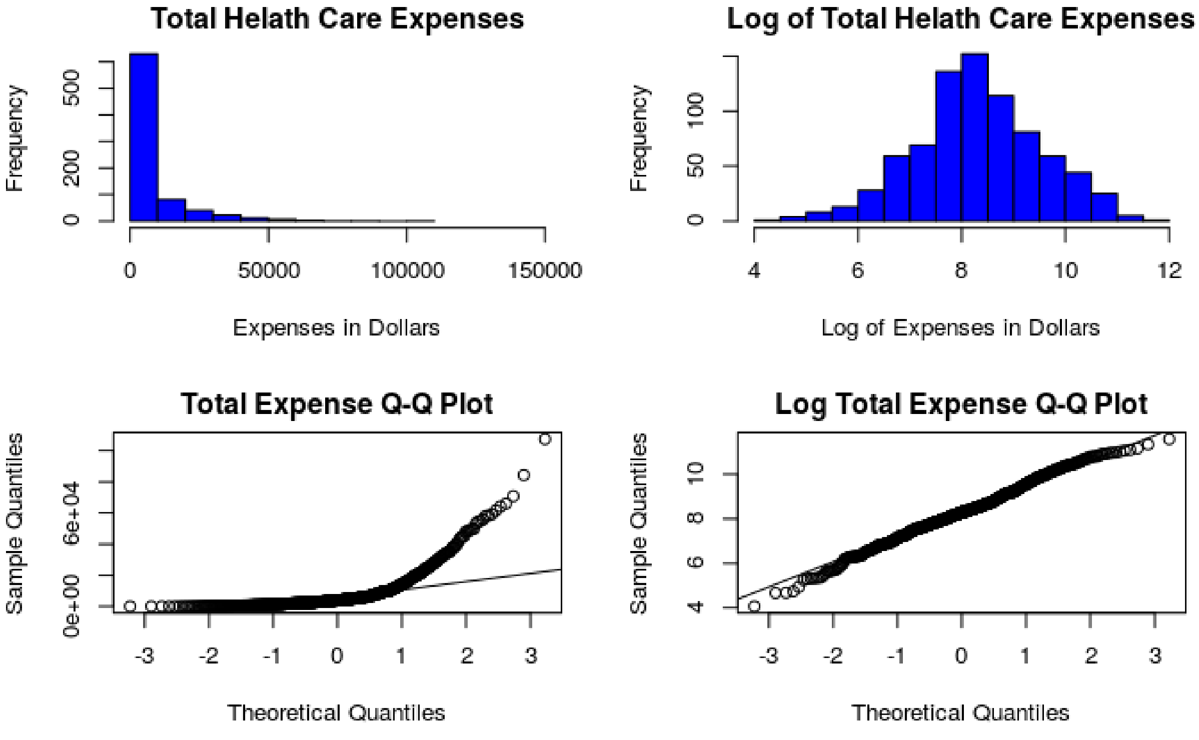


Fig. 1 – Response Variable – Total Expense

2.1.2 Exploring the predictor variables

A quick review of the correlation of the data indicates that multicollinearity may not be an issue, as all correlations are less than 0.5, which we will validate through the variance inflation factor (VIF).

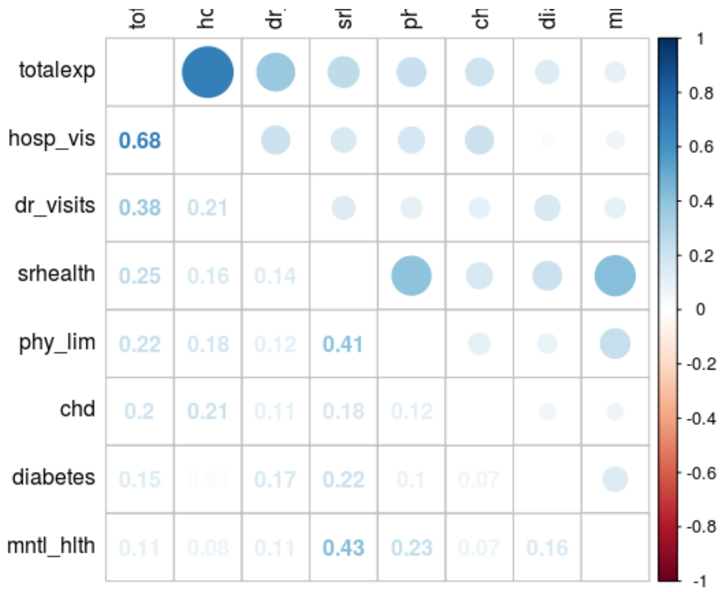


Fig. 2 – Correlation

2.2 TEST & TRAIN DATASETS

We split the data into a training set and a test set, by randomly allocating 80% to the train set and 20% to the test set. We used the train set to fit various regression models and predicted using the test data.

3 METHODOLOGY

We started with standard linear regression model. We checked for multicollinearity utilizing variance inflation factor (VIF). We noticed that the VIF (Table 1) for all the variables was less than 2, indicating that there is no correlation between the predictor variables.

Varible	VIF	Varible	VIF	Varible	VIF	Varible	VIF
age	1.217	mntl_hlth	1.327	high_chol	1.093	race_grp	1.263
marital	1.122	phy_lim	1.324	diabetes	1.156	smoker	1.086
educ	1.513	bmi	1.179	dr_visits	1.144	male	1.132
income	1.276	chd	1.169	msa	1.041	high_bp	1.137
srhealth	1.639					hosp_vis	1.138

Table 1 – VIF for predicator variables

Next we tested the model for heteroscedasticity utilizing the non-constant variance test (ncvTest) and also plotting the standardized residuals to see if they displayed a statistically normal distribution, (Fig. 3), indicating that the errors have the same but unknown variance. The ncvTest indicated that the residuals showed heteroscedasticity.

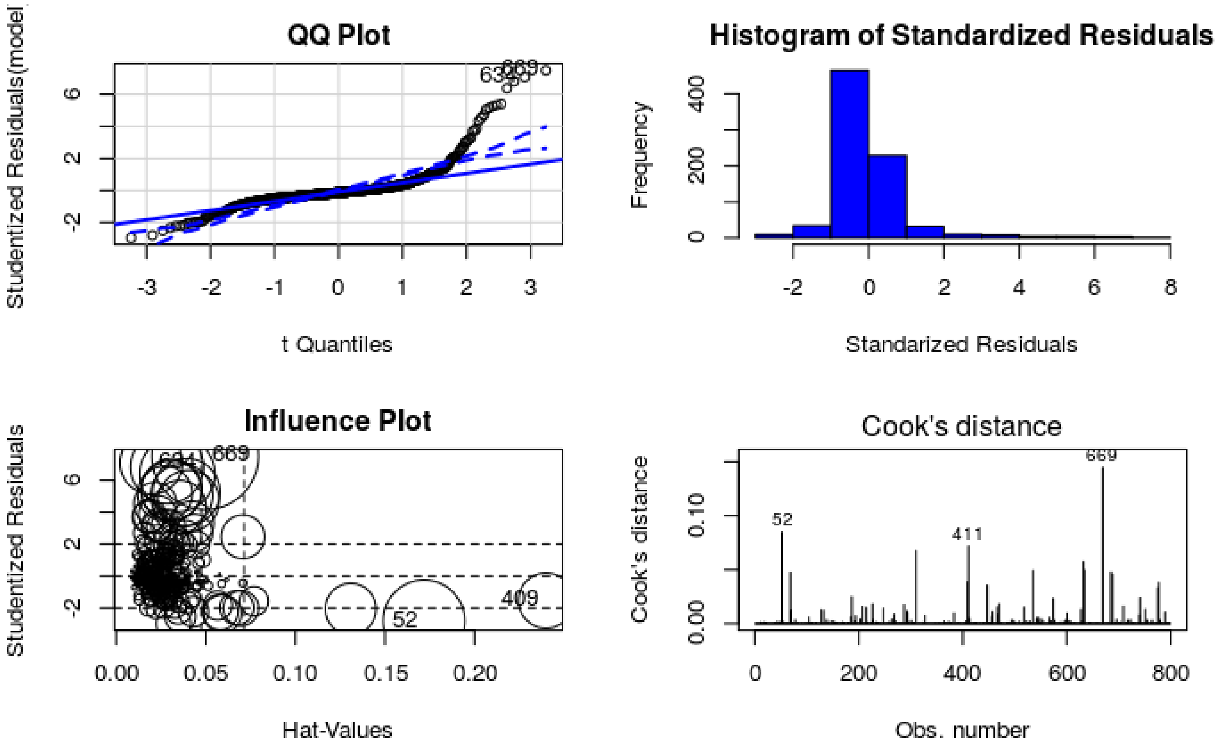
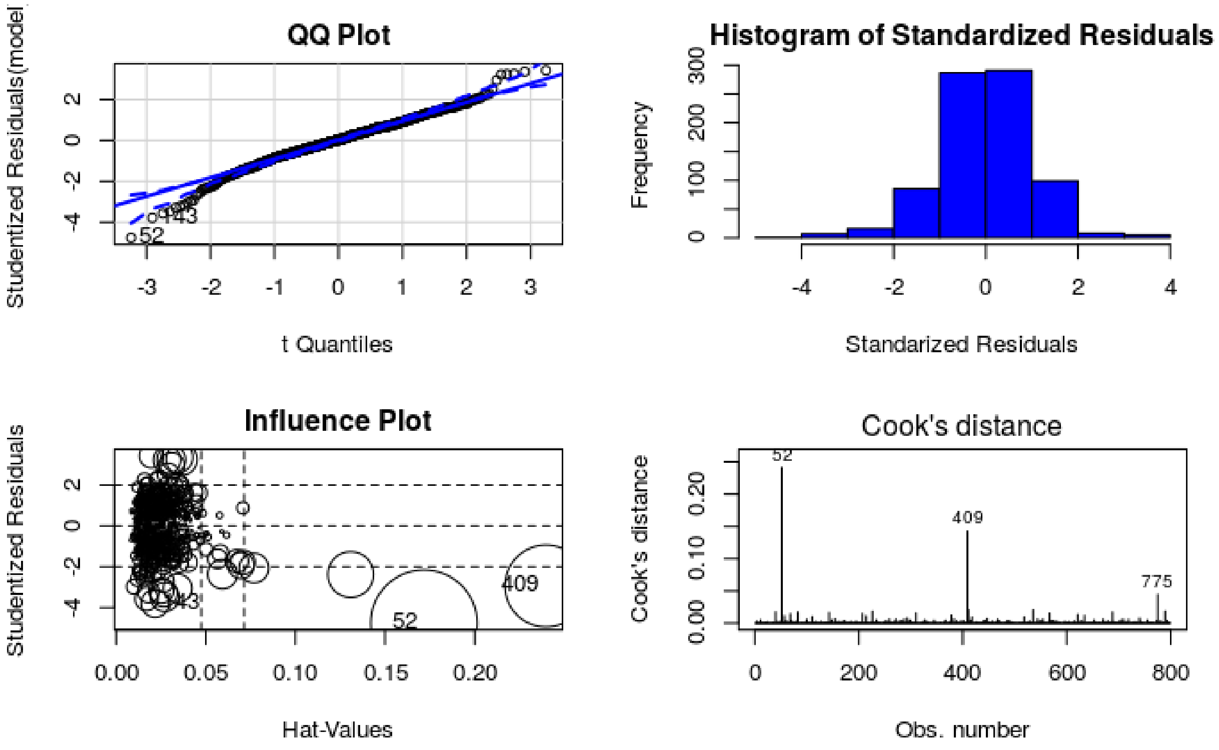


Fig. 3 – Initial regression model

To address the heteroscedasticity nature of the model, we ran a new model by applying a logarithmic transformation of the response variable ‘totalexpense’ (Fig. 4).



. Fig. 4 – Log transformation

We then used boxcox() function to find the best λ (Fig. 5) and applied the transformation to the response variable ‘totalexpanse’ (Fig. 6) .

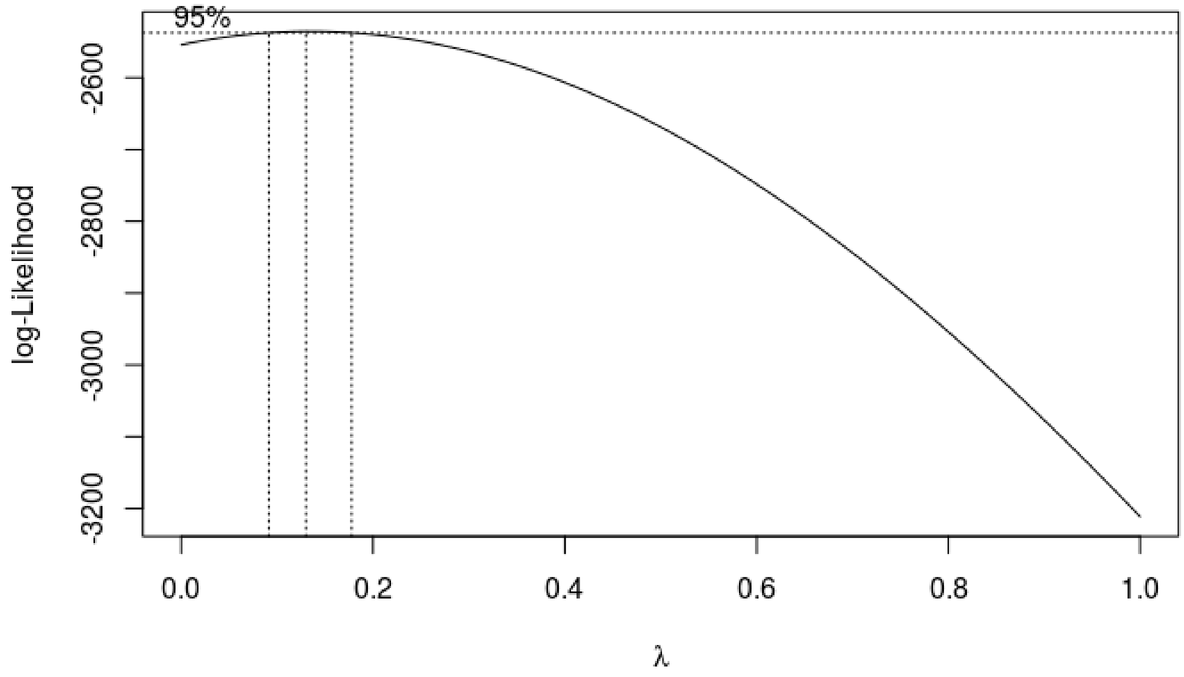


Fig. 5 - log-Likelihood as a function of possible λ values

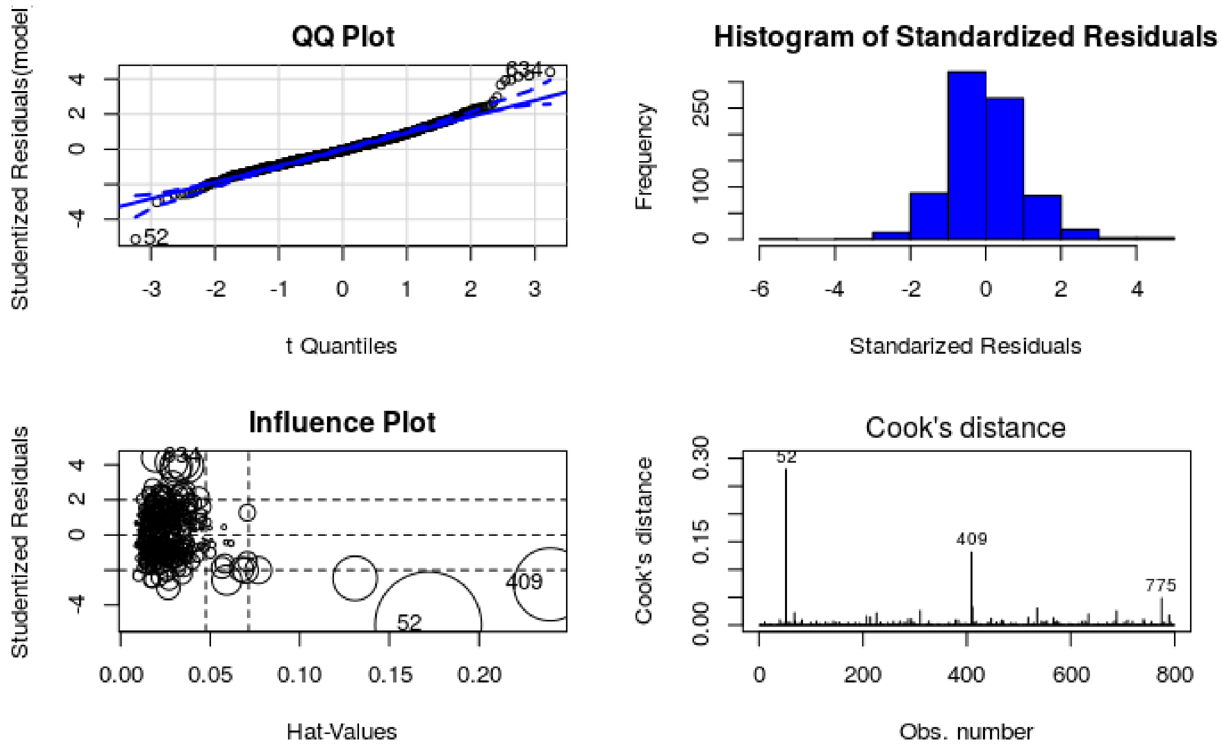


Fig. 6 – Box Cox Transformation Model

After that, we used Cook’s distance (Fig. 7) to prune any outliers and/or influential points from our data set. For this we used a cutoff of 4/(k-n-1), where n is the number of observations in our data set and k the number of parameters in the model. Any data points with Cook's distance greater than this threshold were deemed to be of high influence, and are removed from the data set.

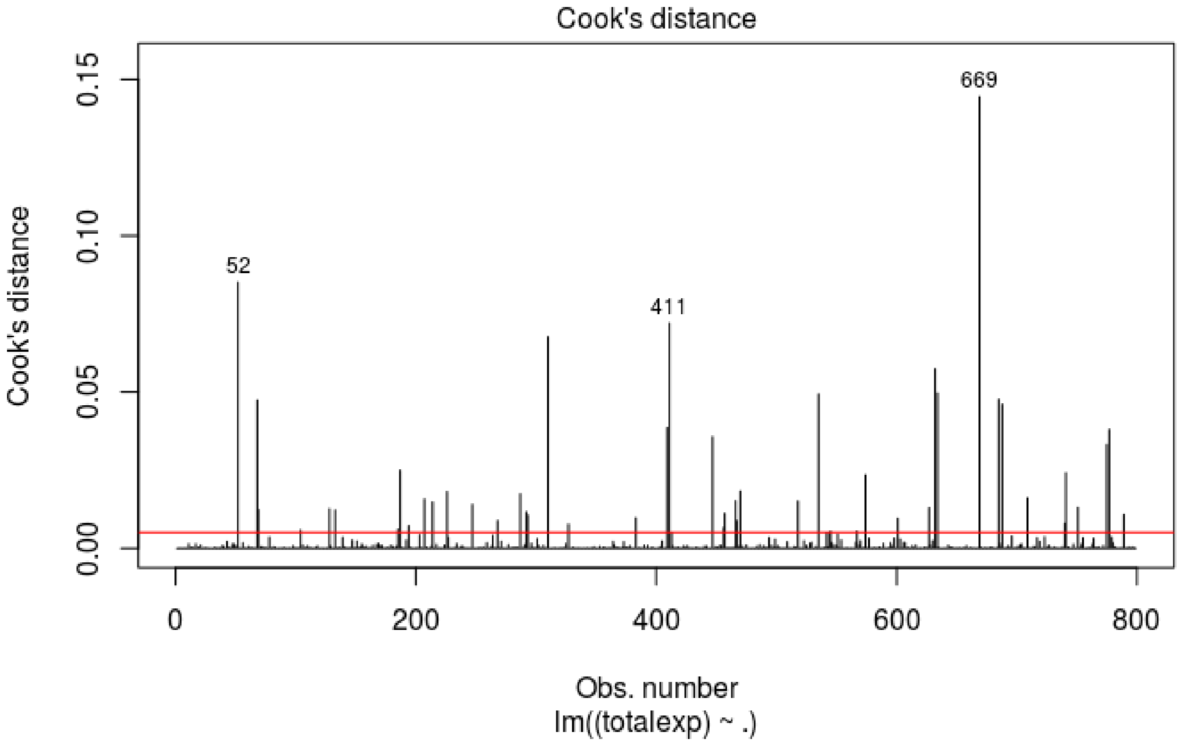


Fig.7 – Cook's Distance
The red line is the cutoff at which a point is considered influential: 4/(n-k1)

We then performed cross validation, with n-folds = 10, with the full regression model, significant P-value model (where we iteratively ran the full model till we found variables with significant p-values), stepwise model, ridge model, lasso model and elastic model. We performed similar analysis with the log and λ transformation. We repeated the tests without transformation with purged data set (purged the original data set from all the outliers)

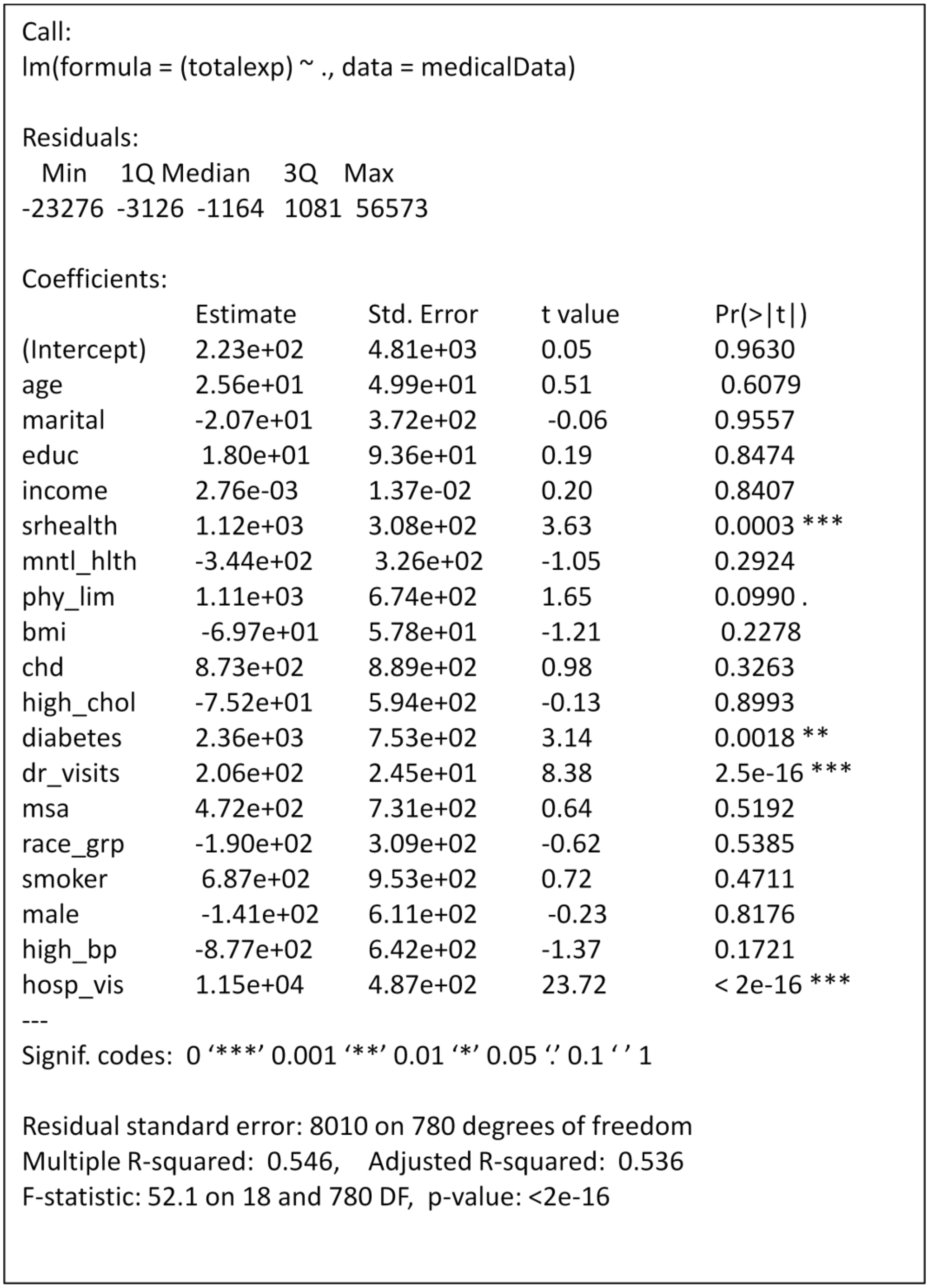
We ranked the models using three measures – Root Mean Squared Errors, R-Square and accuracy (ratio of correct predictions)

4 RESULTS

4.1 Basic Regression

We generated a baseline by running the basic regression model with all the predictor variables. We used the results to spot any problems that need to be addressed. The summary of the regression model are shown below (Fig. 8)

The initial model indicates that only 4 out of the 18 predictor variables are significant. The adjusted R-square for the model is 0.536.

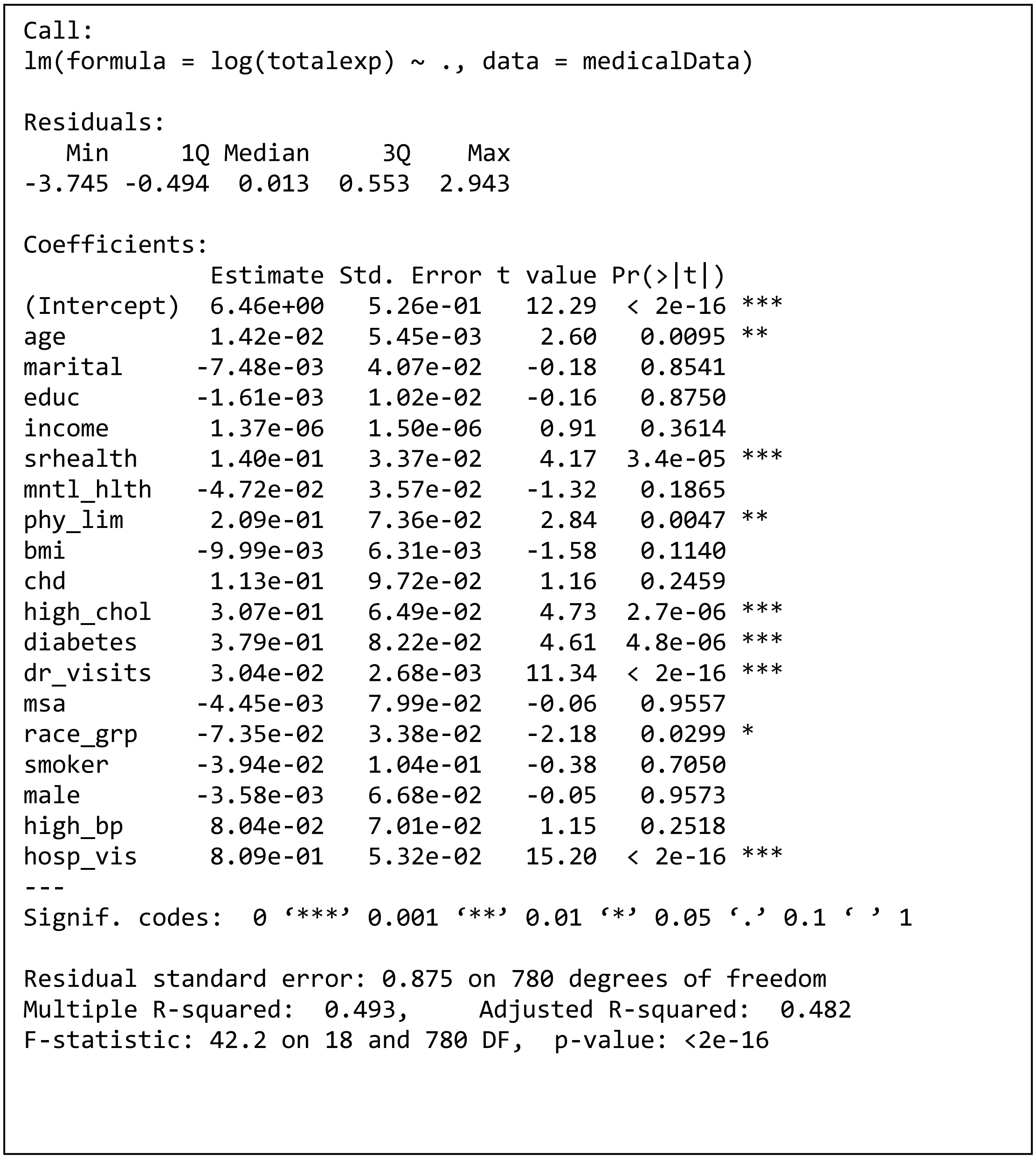


```
Call:
lm(formula = (totalexp) ~ ., data = medicalData)

Residuals:
   Min     1Q Median     3Q    Max
-23276  -3126  -1164   1081  56573

Coefficients:
                Estimate      Std. Error     t value       Pr(>|t|)
(Intercept)     2.23e+02      4.81e+03       0.05          0.9630
age             2.56e+01      4.99e+01       0.51           0.6079
marital         -2.07e+01     3.72e+02        -0.06        0.9557
educ             1.80e+01     9.36e+01       0.19          0.8474
income          2.76e-03      1.37e-02       0.20          0.8407
srhealth        1.12e+03      3.08e+02       3.63          0.0003 ***
mntl_hlth       -3.44e+02      3.26e+02      -1.05         0.2924
phy_lim         1.11e+03      6.74e+02       1.65          0.0990 .
bmi              -6.97e+01    5.78e+01       -1.21          0.2278
chd             8.73e+02      8.89e+02       0.98          0.3263
high_chol       -7.52e+01     5.94e+02       -0.13         0.8993
diabetes        2.36e+03      7.53e+02       3.14          0.0018 **
dr_visits       2.06e+02      2.45e+01       8.38          2.5e-16 ***
msa             4.72e+02      7.31e+02       0.64          0.5192
race_grp        -1.90e+02     3.09e+02       -0.62         0.5385
smoker           6.87e+02     9.53e+02       0.72          0.4711
male             -1.41e+02    6.11e+02        -0.23        0.8176
high_bp         -8.77e+02     6.42e+02       -1.37         0.1721
hosp_vis        1.15e+04      4.87e+02       23.72         < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8010 on 780 degrees of freedom
Multiple R-squared:  0.546,    Adjusted R-squared:  0.536
F-statistic: 52.1 on 18 and 780 DF,  p-value: <2e-16
```

Fig. 8 – Initial Regression Model Summary Result

4.2 Regression with Log transformation

To answer the heteroscedasticity nature of the basic regression model, we applied a log transformation on the response variable. The summary of the log transformation model are shown in Fig. 9



```
Call:
lm(formula = log(totalexp) ~ ., data = medicalData)

Residuals:
   Min     1Q Median     3Q    Max 
-3.745 -0.494  0.013  0.553  2.943 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.46e+00   5.26e-01   12.29  < 2e-16 ***
age          1.42e-02   5.45e-03    2.60   0.0095 ** 
marital     -7.48e-03   4.07e-02   -0.18   0.8541    
educ        -1.61e-03   1.02e-02   -0.16   0.8750    
income       1.37e-06   1.50e-06    0.91   0.3614    
srhealth     1.40e-01   3.37e-02    4.17  3.4e-05 ***
mntl_hlth   -4.72e-02   3.57e-02   -1.32   0.1865    
phy_lim      2.09e-01   7.36e-02    2.84   0.0047 ** 
bmi         -9.99e-03   6.31e-03   -1.58   0.1140    
chd          1.13e-01   9.72e-02    1.16   0.2459    
high_chol    3.07e-01   6.49e-02    4.73  2.7e-06 ***
diabetes     3.79e-01   8.22e-02    4.61  4.8e-06 ***
dr_visits    3.04e-02   2.68e-03   11.34  < 2e-16 ***
msa         -4.45e-03   7.99e-02   -0.06   0.9557    
race_grp    -7.35e-02   3.38e-02   -2.18   0.0299 *  
smoker      -3.94e-02   1.04e-01   -0.38   0.7050    
male        -3.58e-03   6.68e-02   -0.05   0.9573    
high_bp      8.04e-02   7.01e-02    1.15   0.2518    
hosp_vis     8.09e-01   5.32e-02   15.20  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.875 on 780 degrees of freedom
Multiple R-squared:  0.493,	Adjusted R-squared:  0.482 
F-statistic: 42.2 on 18 and 780 DF,  p-value: <2e-16
```

Fig. 9 – Log Transform Model Summary Result

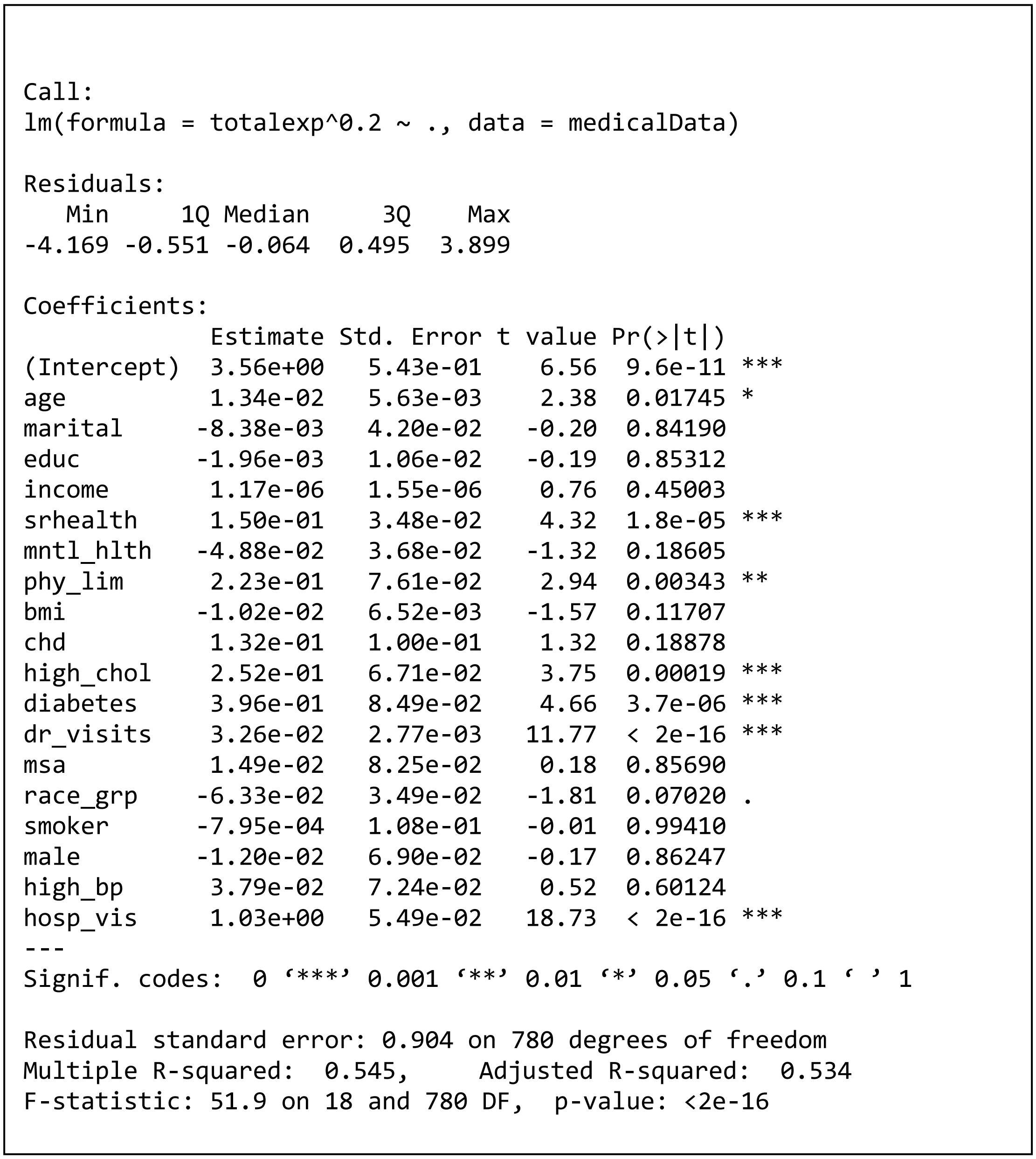
This model indicates that 8 out of the 18 predictor variables are significant. The adjusted R-square for the model is 0.482.

The histogram of the residuals indicates a close to normal distribution (Fig. 4). But the Anderson-Darling normality test returned a p-value of 2e^{-5} suggesting that we reject the null hypothesis that the residuals are normally distributed.

4.3 Regression with Box Cox Transformation

We also decided to model by transforming the response variable by λ. To determine the best λ for transforming we used the boxcox() function on the base model. (Fig. 5).

The model indicates that 8 out of the 18 predictor variables are significant. The adjusted R-square for the model is 0.534.



```
Call:
lm(formula = totalexp^0.2 ~ ., data = medicalData)

Residuals:
   Min     1Q Median     3Q    Max 
-4.169 -0.551 -0.064  0.495  3.899 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.56e+00   5.43e-01    6.56  9.6e-11 ***
age          1.34e-02   5.63e-03    2.38  0.01745 *  
marital     -8.38e-03   4.20e-02   -0.20  0.84190    
educ        -1.96e-03   1.06e-02   -0.19  0.85312    
income       1.17e-06   1.55e-06    0.76  0.45003    
srhealth     1.50e-01   3.48e-02    4.32  1.8e-05 ***
mntl_hlth   -4.88e-02   3.68e-02   -1.32  0.18605    
phy_lim      2.23e-01   7.61e-02    2.94  0.00343 ** 
bmi         -1.02e-02   6.52e-03   -1.57  0.11707    
chd          1.32e-01   1.00e-01    1.32  0.18878    
high_chol    2.52e-01   6.71e-02    3.75  0.00019 ***
diabetes     3.96e-01   8.49e-02    4.66  3.7e-06 ***
dr_visits    3.26e-02   2.77e-03   11.77  < 2e-16 ***
msa          1.49e-02   8.25e-02    0.18  0.85690    
race_grp    -6.33e-02   3.49e-02   -1.81  0.07020 .  
smoker      -7.95e-04   1.08e-01   -0.01  0.99410    
male        -1.20e-02   6.90e-02   -0.17  0.86247    
high_bp      3.79e-02   7.24e-02    0.52  0.60124    
hosp_vis     1.03e+00   5.49e-02   18.73  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.904 on 780 degrees of freedom
Multiple R-squared:  0.545,	Adjusted R-squared:  0.534 
F-statistic: 51.9 on 18 and 780 DF,  p-value: <2e-16
```

Fig. 10 – λ Transform Model Summary Result

The histogram of the residuals indicates a close to normal distribution (Fig. 6). But the non constant variance test returned a p-value of `<2e-16` suggesting that the residuals are heteroscedastic

4.4 Regression using variables with significant P-Values

We ran the three models to iteratively select predictor variables that are significant (p-value < 0.1). Code in the appendix.

The iteration resulted in the following three models:

Base Model	totalexp ~ srhealth + diabetes + dr_visits + hosp_vis
Log Transform Model	log(totalexp) ~ age + srhealth + phy_lim + high_chol + diabetes + dr_visit + race+grp + hosp_vis
Box Cox Transform Model	totalexp^0.2 ~ age + srhealth + phy_lim + high_chol + diabetes + dr_visit + race+grp + hosp_vis

4.5 Step Regression

The step regression resulted in the following best model using :

Base Model	totalexp ~ srhealth + phy_lim + bmi + chd + diabetes + dr_visits + hosp_vis
Log Transform Model	log(totalexp) ~ age + income + srhealth + phy_lim + bmi + high_chol + diabetes + dr_visits + race_grp + high_bp + hosp_vis
Box Cox Transform Model	totalexp^0.2 ~ age + income + srhealth + phy_lim + bmi + high_chol + diabetes + dr_visits + race_grp + hosp_vis

4.6 Ridge, Lasso and Elastic Regression

We used the glmnet package to run ridge, lasso and elastic regression.

Ridge regression: We find the coefficients (the b's) that minimizes, the below equation for a given λ.

$$\sum e_i^2 + \lambda \sum b_i^2 \quad e_i = Yi - \hat{Y}i$$

Lasso regression: We find the coefficients (the b's) that minimizes, the below equation for a given λ.

$$\sum e_i^2 + \lambda \sum |b_i| \qquad e_i = Yi - \hat{Y}i$$

Elastic Net regression: We find the coefficients (the b's) that minimizes, the below equation for a given λ and α.

$$\sum e_i^2 + \lambda\, [\alpha\, (\sum |b_i|) + (1- \alpha)(\, \sum b_i^2\,)] \qquad e_i = Yi - \hat{Y}i$$

5 Ranking and Selecting the model

We used cross-validation with n-folds equal to 10, to see how well the various models predict and ranked them by measuring the Root Mean Squared Errors, the R-Square and the Accuracy of the models.

5.2 Base Model

The elastic net regression model had the best RMSE of 8025. The ridge regression had the best accuracy score ofo0.02

Name	RMSE	Accuracy	R2
Full	8046	0.015625	0.5598
Significat P-Value	8001	0.008125	0.5655
Step	14629	0.0075	0.5047
Step Pair	8740	0.017219	0.4951
Ridge	**8050**	**0.02**	**0.56**
Lasso	7979	0.014375	0.5692
Elastic	8025	0.0175	0.5646

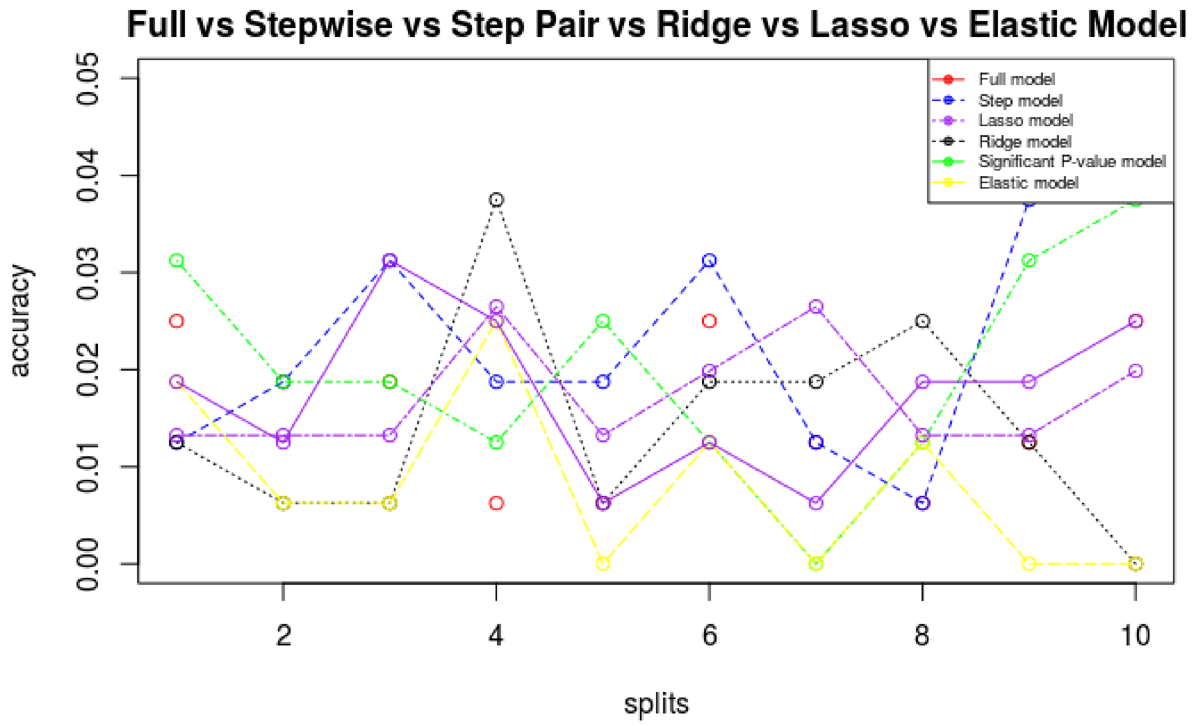


Fig. 11 – Accuracy Measure for the Base Model

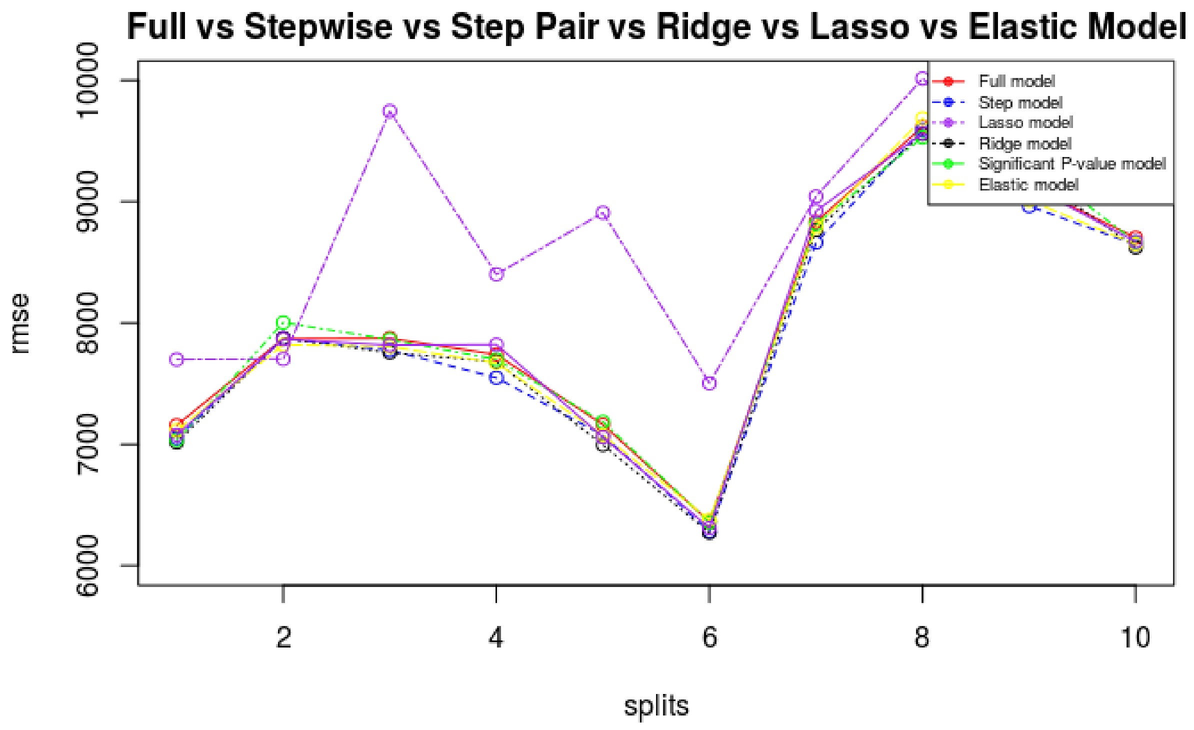


Fig.12 – RMSE Measure for Base Model

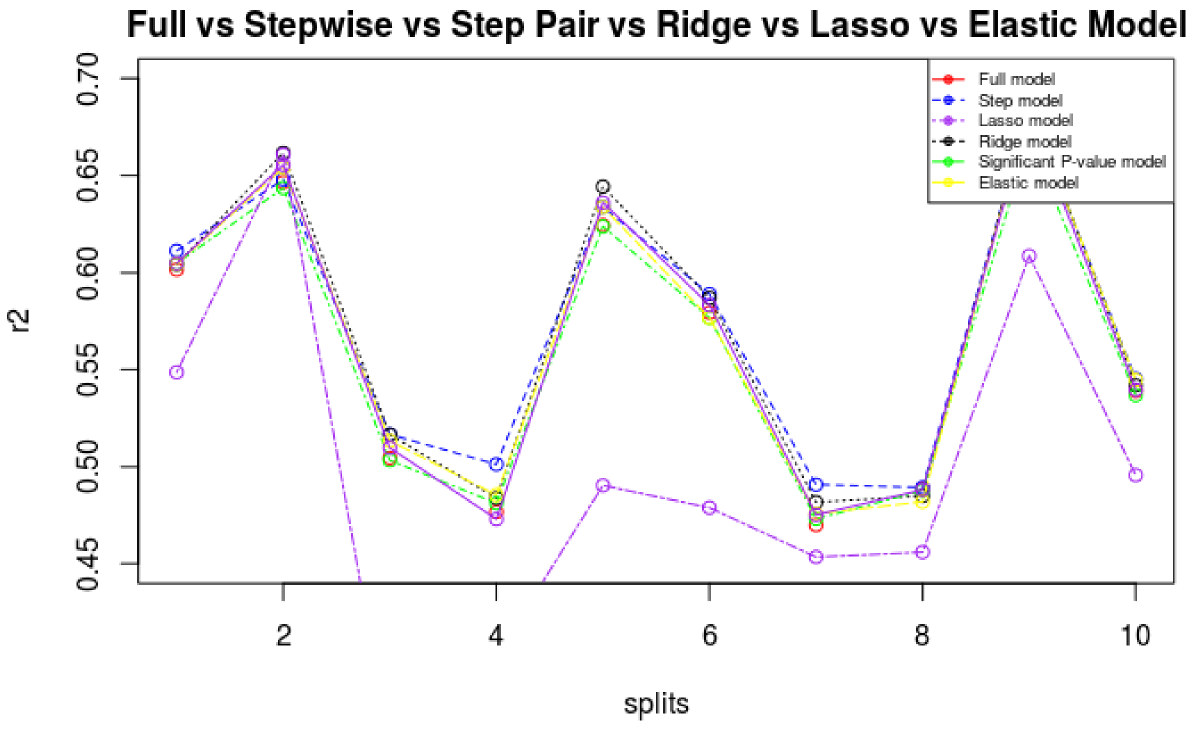


Fig. 13 – R2 Measure for Base Model

5.3 Log Transform Model

All the models had the same RMSE score. So we used the accuracy score to select the best model. The model based on the accuracy score was Elastic Model.

Name	**RMSE**	**Accuracy**	**R2**
Log	14627	0.008125	0.4223
Significant P-Value LOG	14627	0.008125	0.4224
Step LOG	14627	0.008125	0.4221
Ridge LOG	14627	0.006875	0.4283
Lasso LOG	14627	0.006875	0.4427
Elastic	**14627**	**0.018125**	**0.4928**

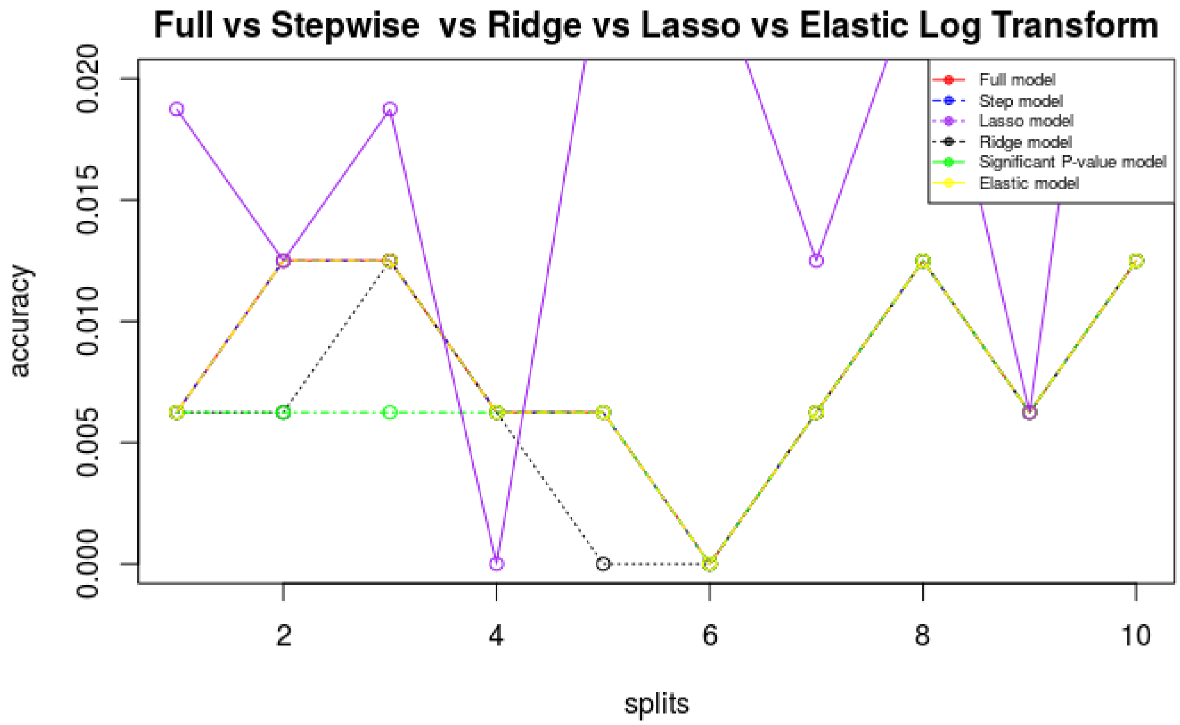


Fig. 14 – Accuracy Measure for Log Transform Model

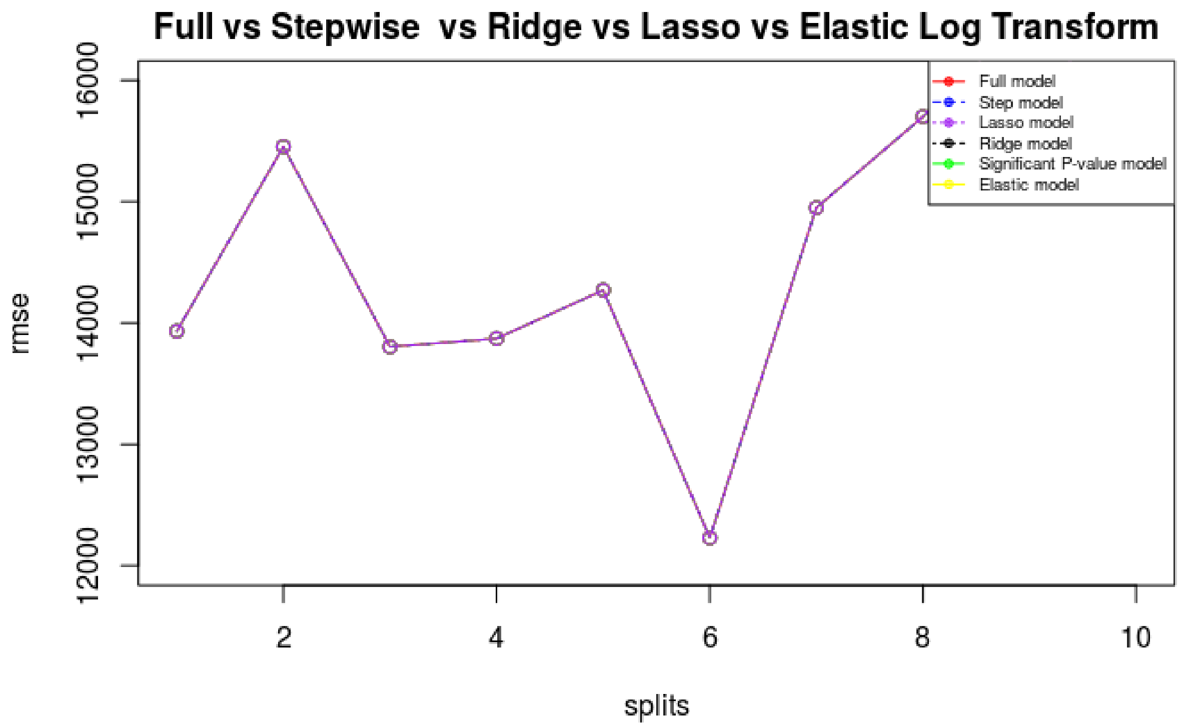


Fig. 15 – Accuracy Measure for Log Transform Model

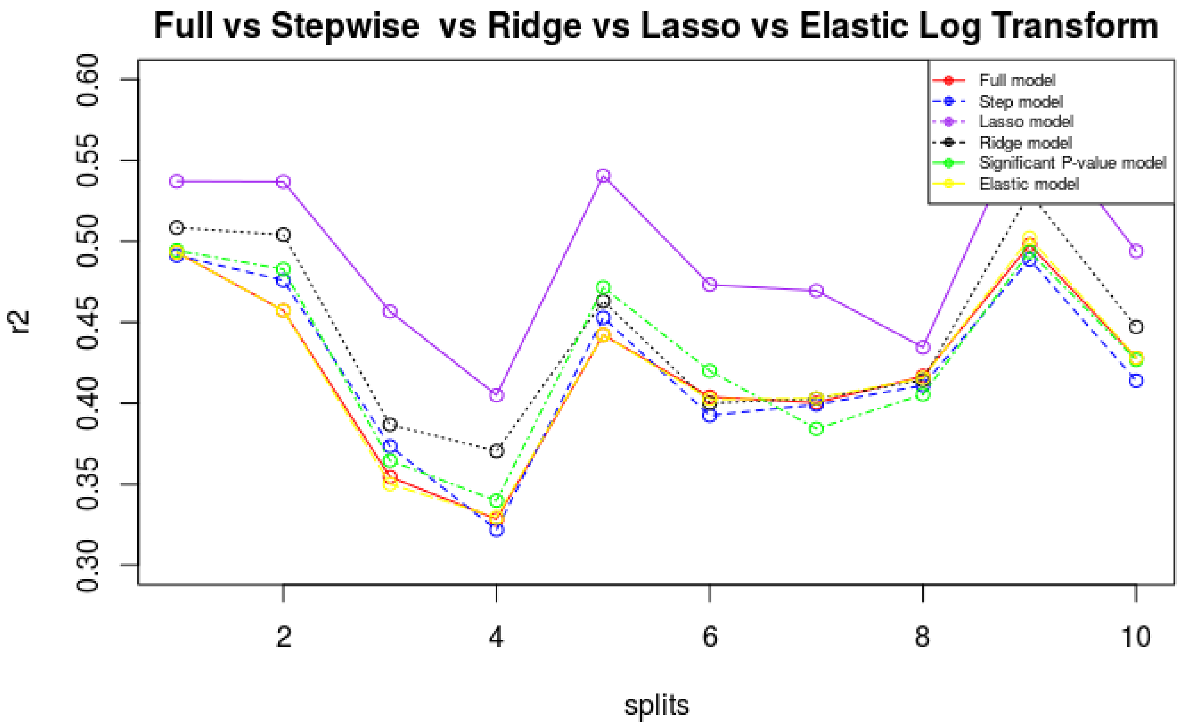


Fig. 16 – R2 Measure for Log Transform Model

5.4 Box Cox Transform Model

All the models had the same RMSE score. So we used the accuracy score to select the best model. The model based on the accuracy score was Elastic Model.

.Name	RMSE	Accuracy	R2
Box Cox Transform	14629	0.006875	0.4988
Significat P-Value BCT	14629	0.006875	0.4928
Step BCT	14629	0.0075	0.5047
Ridge BCT	14629	0.00625	0.4767
Lasso BCT	14629	0.005625	0.4973
Elastic	**14629**	**0.0175**	**0.5266**

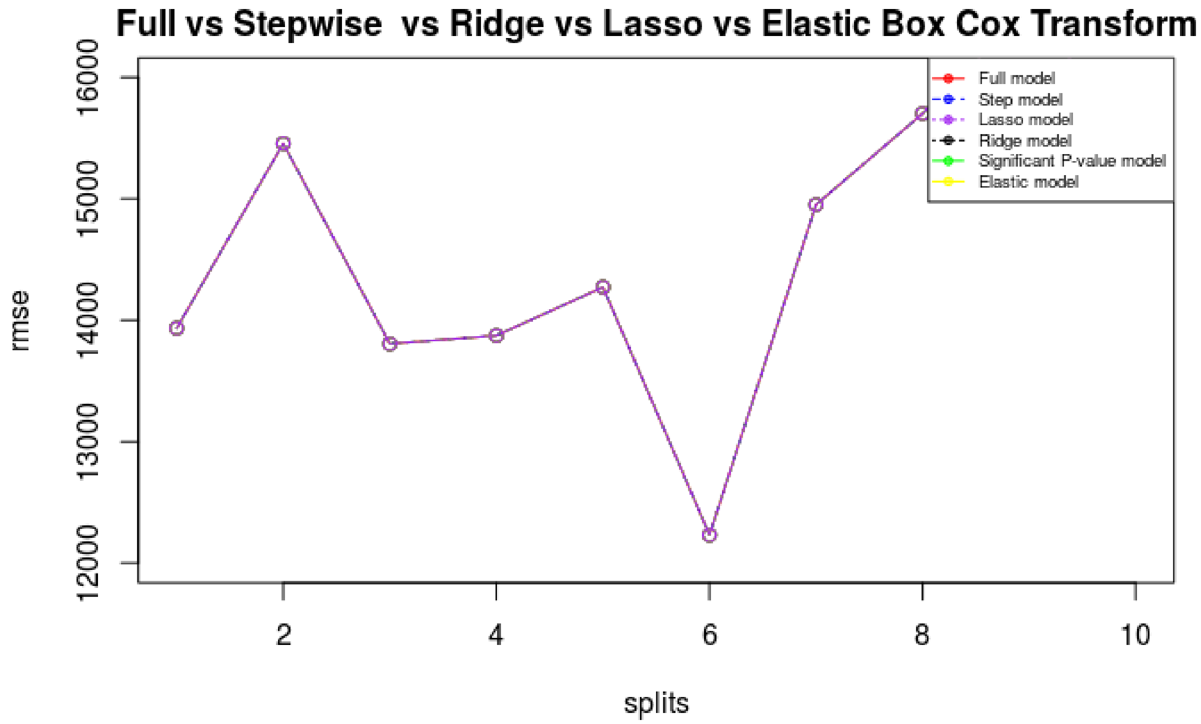


Fig. 17 – RMSE Measure for Box Cox Transform Model

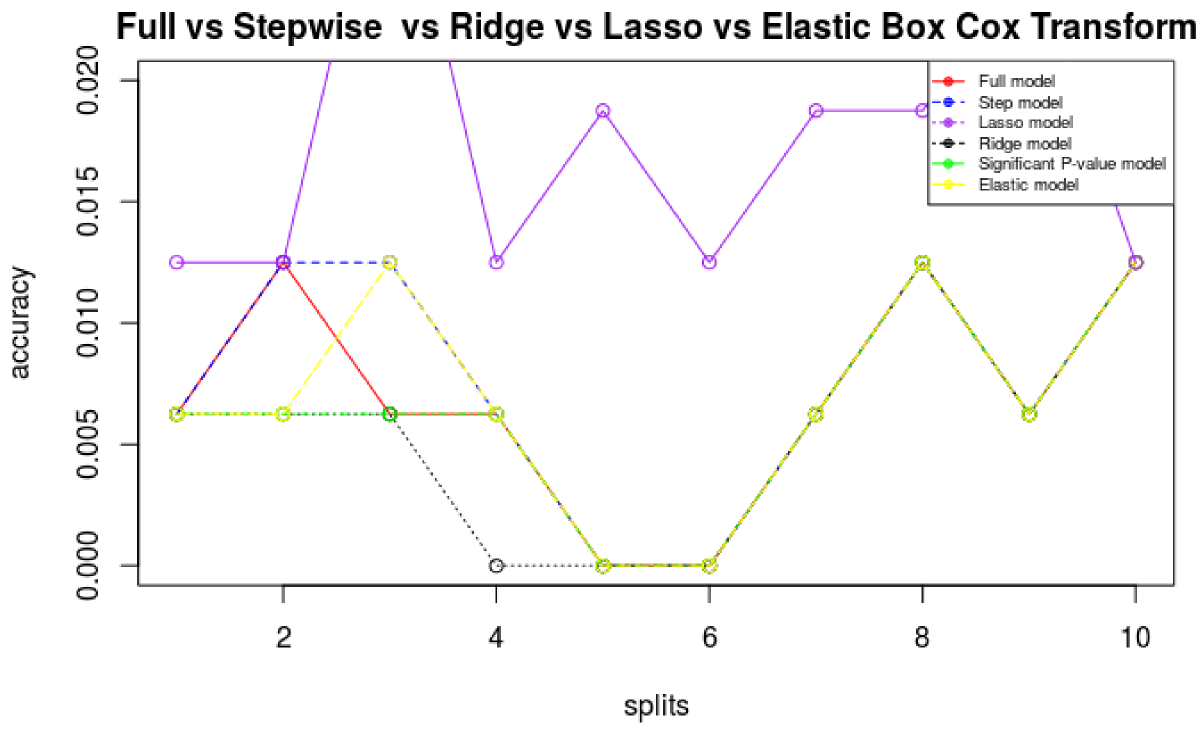


Fig. 18 – Accuracy Measure for Box Cox Transform Model

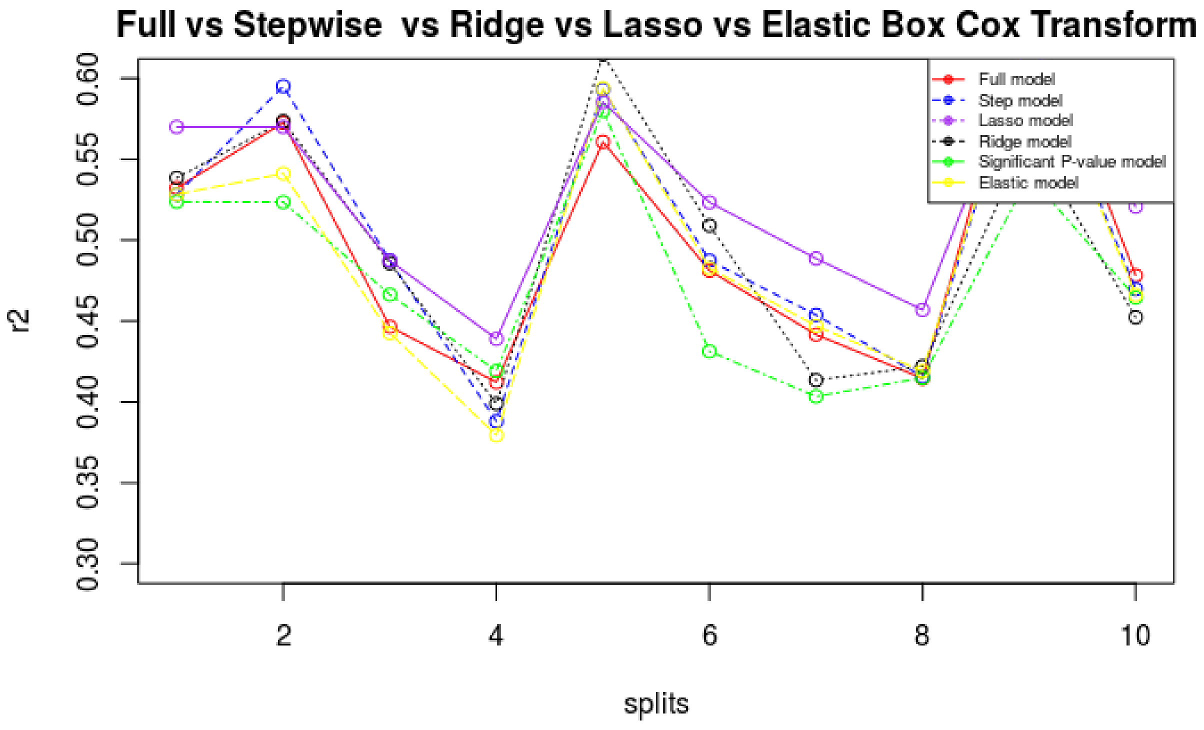


Fig. 19 – R2 Measure for Box Cox Transform Model

The ridge regression with no transformation was the best model for prediction with an accuracy score of 0.02.

6 Conclusion

All the models displayed heteroscedastic nature of residuals. Heteroscedasticity is a problem because ordinary least squares (OLS) regression assumes that all residuals are drawn from a population that has a constant variance (homoscedasticity). To satisfy the regression assumptions and be able to trust the results, the residuals should have a constant variance. This could be due to the large range between the largest and the smallest observed values. The response variable (totalexp) has smallest value of 58 and largest value of 107355. One of the predictor variable (income) has smallest value of 125 and largest value of 176839.

We think other predictor variables should be explored to better understand the cost of health care amongst the elders.