MTH 656 Statistics Method, 2012 Fall

Shuowen Wei’s Project 3

* Introduction

This purpose of this project is to help us understand how to build up multiple regression models for a data set. In this given data set, there are 393 rows of data about a car’s different aspects, like its model, horsepower, acceleration, miles per gallon (MPG), the number of cylinders, weight, etc. I am particularly interested in knowing the influences of other variables on a car’s acceleration, what kind of car has greater acceleration than other cars’. Thus, in the attached excel file ‘Shuowen Wei's Project 3 CARS\_regr\_f12.xls’, I chose the ‘Accelerate’ as my “y” variable, and build up a multiple regression model and through some kinds of analysis to see whether they have big influence or not on the horsepower.

* Results and discussion

First of all, in the original data set, we want to what kinds of variables are highly correlated to each other, such that we may keep only one of these highly correlated variables. So I generate the correlation table of all variables including the acceleration, which is my y-variable, and showed as follows:

The correlation table

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Horsepower* | *Accelerate* | *MPG* | *Cylinders* | *Engine Disp* | *Weight* | *weight\_squared* | *Year* | *Amer* | *Japanese* |
| Horsepower | 1 |  |  |  |  |  |  |  |  |  |
| Accelerate | -0.68919551 | 1 |  |  |  |  |  |  |  |  |
| MPG | -0.77842678 | 0.423328537 | 1 |  |  |  |  |  |  |  |
| Cylinders | 0.842983357 | -0.504683379 | -0.77761751 | 1 |  |  |  |  |  |  |
| Engine Disp | 0.897257002 | -0.543800497 | -0.80512695 | 0.950823 | 1 |  |  |  |  |  |
| Weight | 0.864537738 | -0.416839202 | -0.83224421 | 0.897527 | 0.9329944 | 1 |  |  |  |  |
| weight\_squared | 0.869719798 | -0.426547263 | -0.80668158 | 0.890839 | 0.9287787 | 0.992019 | 1 |  |  |  |
| Year | -0.41636148 | 0.290316113 | 0.580540966 | -0.34565 | -0.3698552 | -0.3091199 | -0.325214222 | 1 |  |  |
| Amer | 0.489624788 | -0.258224128 | -0.56516059 | 0.610494 | 0.65593573 | 0.6009783 | 0.570896341 | -0.1361 | 1 |  |
| Japanese | -0.32193559 | 0.115020399 | 0.45145363 | -0.40421 | -0.4408246 | -0.4479287 | -0.418080838 | 0.19984 | -0.649 | 1 |

Obviously, the ‘weight’ variable and ‘weight\_squared’ variable are highly correlated (0.992019), and ‘Engine displacement’ variable and ‘weight’ are also highly correlated (0.9329944), in order to simply our regression model and get much nicer result, so we will leave both the ‘weight\_squared’ and ‘weight’ variable out of our regression model. Also, as the project required, we will also leave the ‘Accelerate’ variable out regression model, since ‘Accelerate’ variable is also highly correlated to ‘Horsepower’ (but the data in the correlation table doesn’t show that they’re that high ).

After leaving ‘weight\_squared’ and ‘accelerate’ variables out of our regression model, firstly, we build up our first-order multiple regression model depending on the rest variables, as follows:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 12.47186 | 2.665658 | 4.678717 | 4.01E-06 | 7.230744 | 17.71297 | 7.230744 | 17.71297 |
| MPG | 0.068745 | 0.033311 | 2.063744 | 0.039713 | 0.003251 | 0.134239 | 0.003251 | 0.134239 |
| Cylinders | 0.160894 | 0.209822 | 0.766815 | 0.443663 | -0.25165 | 0.573438 | -0.25165 | 0.573438 |
| Engine Disp | -0.03633 | 0.004489 | -8.0935 | 7.73E-15 | -0.04516 | -0.02751 | -0.04516 | -0.02751 |
| Weight | 0.002629 | 0.000433 | 6.075937 | 2.97E-09 | 0.001778 | 0.00348 | 0.001778 | 0.00348 |
| Year | -0.00872 | 0.042534 | -0.20501 | 0.837676 | -0.09235 | 0.074909 | -0.09235 | 0.074909 |
| Amer | 0.916068 | 0.374777 | 2.444303 | 0.014962 | 0.179196 | 1.652939 | 0.179196 | 1.652939 |
| Japanese | -0.4861 | 0.369871 | -1.31423 | 0.189554 | -1.21332 | 0.241132 | -1.21332 | 0.241132 |

In the multiple regression model, the p-value for ‘Year’ is quite high, which is 0.837676 > 0.2, thus we may want to drop this variable. Then let’s look at the scatter plot of this model’s standardized residuals, as follows:

As you can see, there are still about 4 extreme standardized residuals, whose absolute values are greater than 3, thus this model is not good enough. And this model’s adjusted r value is 0.379947, which is not very high (but it turns out in the end that the highest adjusted r value we can get for predicting ‘Acceleration’ is about 0.38).

Now, after dropping the ‘Year’ variable, we repeat what we did above again, we can get out second multiple regression model as follows:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 12.00454 | 1.38012 | 8.698186 | 9.92E-17 | 9.291023 | 14.71805 | 9.291023 | 14.71805 |
| MPG | 0.064472 | 0.025955 | 2.484008 | 0.013416 | 0.013441 | 0.115504 | 0.013441 | 0.115504 |
| Cylinders | 0.158701 | 0.209288 | 0.75829 | 0.448741 | -0.25279 | 0.570191 | -0.25279 | 0.570191 |
| Engine Disp | -0.03611 | 0.004345 | -8.30937 | 1.66E-15 | -0.04465 | -0.02756 | -0.04465 | -0.02756 |
| Weight | 0.002592 | 0.000393 | 6.604179 | 1.33E-10 | 0.00182 | 0.003364 | 0.00182 | 0.003364 |
| Amer | 0.892706 | 0.356588 | 2.503468 | 0.012711 | 0.191603 | 1.59381 | 0.191603 | 1.59381 |
| Japanese | -0.49578 | 0.366386 | -1.35316 | 0.176799 | -1.21615 | 0.22459 | -1.21615 | 0.22459 |

Now, the p-value for ‘Cylinders’ is high, which is 0.448741> 0.2, thus we may want to drop this variable, too. Then the scatter plot of this model’s standardized residuals is as follows:

Which shows that there are still 4 extreme values than the first model we got, and the adjusted r value is 0.38149, which is a little better than the last one. Then we continue to drop the variable whose p-value is greater than 0.2, which is the ‘Cylinders’ variable (p-value = 0.448741).

After dropping ‘Cylinders’ and repeat the same process again, we got the p-value of ‘Japanese’ variable is 0.187965, which is also pretty high, though less than 0.2. Then we drop ‘Japanese’ and come up with almost the best first-order multiple regression model we can get:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 12.1715 | 1.276026 | 9.538597 | 1.66E-19 | 9.66269 | 14.68031 | 9.66269 | 14.68031 |
| MPG | 0.059805 | 0.025779 | 2.319868 | 0.020868 | 0.00912 | 0.11049 | 0.00912 | 0.11049 |
| Engine Disp | -0.03441 | 0.003105 | -11.0828 | 5.61E-25 | -0.04052 | -0.02831 | -0.04052 | -0.02831 |
| Weight | 0.002669 | 0.000389 | 6.85509 | 2.83E-11 | 0.001904 | 0.003435 | 0.001904 | 0.003435 |
| Amer | 1.136657 | 0.301966 | 3.76419 | 0.000193 | 0.542958 | 1.730356 | 0.542958 | 1.730356 |

With every variable’s p-value is far less than 0.2 and the adjusted r value of the this model is 0.38099, which is almost the highest among all the first-order multiple regression models, but there are still some extreme value as you can in the below scatter plot of the standardized residuals below:

Now, to search for the “best” model, we may want to look at some higher order multiple regression models.

And I tried more than 20 kinds of combination of both interaction and self-squared of these variables, finally we find the model below is quite good, which adjusted r value is finally 0.401307.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | 12.70064 | 0.881524 | 14.40759 | 5.58E-38 | 10.96746 | 14.43381 | 10.96746 | 14.43381 |
| MPG | 0.07386 | 0.023977 | 3.08038 | 0.002215 | 0.026717 | 0.121002 | 0.026717 | 0.121002 |
| E W | -9.9E-06 | 8.49E-07 | -11.6379 | 4.78E-27 | -1.2E-05 | -8.2E-06 | -1.2E-05 | -8.2E-06 |
| WW | 7.41E-07 | 8.19E-08 | 9.04119 | 7.55E-18 | 5.8E-07 | 9.02E-07 | 5.8E-07 | 9.02E-07 |
| Amer | 0.871632 | 0.285999 | 3.047674 | 0.002464 | 0.309325 | 1.433938 | 0.309325 | 1.433938 |

Where E W means the interaction of ‘engine displacement’ and ‘weight’, and WW means ‘weight-squared’, and all of their p-value are very small. And there are only 3 outliners at this time, see below:

And among all of the modes I built above, all the variables’ residual plots showed that the errors are mutually independent, which meets the assumptions of our multiple regression model.

* Conclusion:

Based on all the model listed above and showed in the attached excel file, we can conclude that most of the quadratic or higher order models are better than first-order multiple regression models, telling from their adjusted r value, though there’re not much significant improvement. And all of these models’ adjusted r values are less than or equal to 0.4, which is not as high as we expected.

Also, in almost all of those models, all the residual plots of the variables involved show that the errors are mutually independent, and that’s good. But the not that good news is that in almost all the models, there are always 3~6 extreme values out there, considering there are 392 of samples in total, I think 3~6 outliers is tolerable. And there are only 3 outliers in my final model.

The model we came up with in the end indicates that the acceleration of a is highly related to the interaction of the car’s engine displacement and its weight, the square of the car’s weight, which make a lot sense to people, since the the more powerful of a car’s engine and lighter of the car’s weight, the acceleration is much greater, thus meets common sense.