Group P: Project Analysis Component

Sydney Bluestein, Parker Smith, Tate Eppinger, and Juan Perez

2024-04-03

Does "what's under the hood" matter when predicting sports car prices in America? vroom vroom Group P: Sydney Bluestein, Tate Eppinger, Juan Perez, and Parker Smith

Project Aim: We want to explore which factors contribute to sports car prices in America. We have chosen numerous explanatory variables that may influence sports car prices: car origin, car year, 0-60 speed, torque, horsepower, and engine size.

Research Question: Do any of the factors above (or a combination of them) predict sports car prices accurately?

Variables: Car origin (Asian, Domestic, or European) is our categorical variable. Year is a quantitative variable that represents the year the car was made. 0-60 speed is a quantitative variable that represents the time in seconds it takes for the car to go from 0 miles per hour to 60 miles per hour. Torque is a quantitative variable measured in pounds-feet that describes the power of the car's drivetrain. Horsepower is another quantitative variable measured in foot-pounds/minute representing engine power. Lastly, engine size is a quantitative variable in liters that measures how much gas the car can hold.

Horsepower - The plot shows a bimodal distribution with peaks around 450 horsepower and 600 horsepower, indicating two prevalent groups of sports cars with distinct power outputs in this dataset. The IQR is 209.5 lb-ft/minute, demonstrating wide variability of horsepower in this dataset.

Torque - The density plot for the torque (measured in pound-feet) of sports cars in the dataset is mostly normally distributed. The mean torque is around 472.5 lb-ft. The standard deviation of approximately 126.2 lb-feet indicates a moderate variability in the torque figures across the dataset.

Engine Size - The density plot for engine size (in liters) reveals a bimodal distribution with two prominent peaks: one just below 4 liters and another around 6 liters. The IQR of 1.7 liters suggests moderate to high variability of engine sizes among these sports cars.

0 to 60 MPH Time - The distribution of 0-60 mph time is mostly normally distributed. The mean acceleration time is 3.78 seconds (compared to median of 3.6 seconds), suggesting a very slight right skew, though this distribution is mostly normal. The standard deviation is 0.67 seconds.

Car Origin: We also included a bar graph for our categorical variable, car origin. We notice that European cars are most commonly represented in this data set. However, there are still enough Asian and domestic sports cars to make this variable worthwhile.

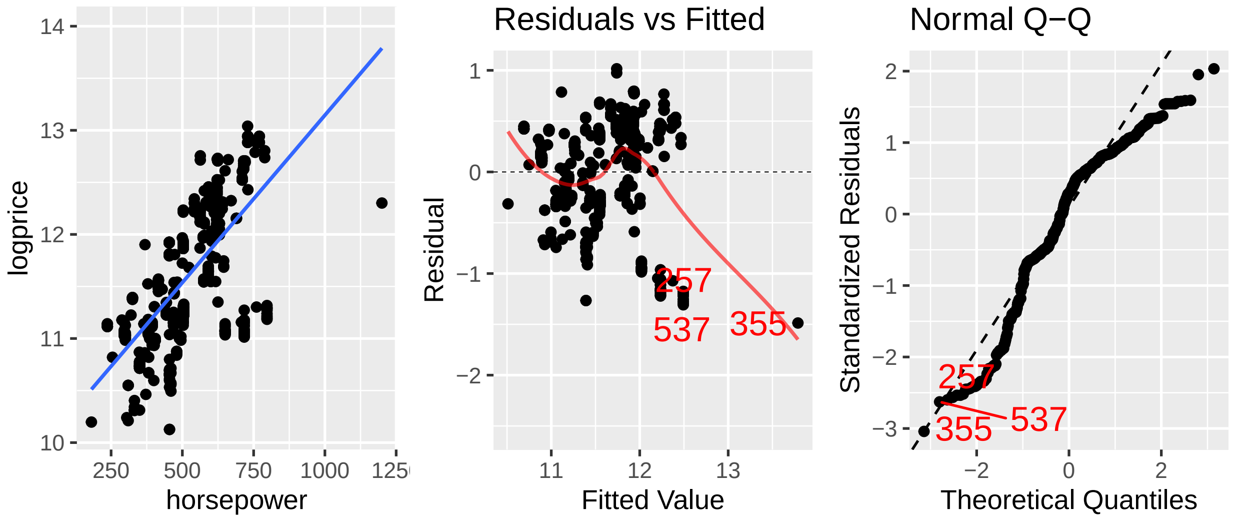
Step 2 – Bivariate Analysis: We will begin with simple linear regression to gauge which explanatory variables help predict sports car prices. Upon the first glimpse, the GGpairs plot demonstrates that horsepower (r=0.648) and 0-60 mph (r=-0.554) may be helpful predictors for sports car prices. These models can be found below.

SLR: Horsepower and 0-60 appeared to be our strongest predictors for car price. However the scatterplot of horsepower vs. car price proved to be pretty curvy. I attempted to transform the data by predicting log (car price) rather than car price, log(horsepower) instead of horsepower, etc. However, conditions still were not met. Similarly for 0-60, conditions were not met for SLR. I removed an outlier, representing a car with extremely high torque and horsepower. I believed this point was pulling the regression line down. After removing this point, I then transformed the data by having -1/sqrt(0-60) predict price. However, conditions for SLR still failed to be met. I then wanted to check if using horsepower and 0-60 together would provide a better fitting model than the predictors on their own. Most of the points in the fitted vs. residuals plot are shifted to the right, demonstrating that there is not perfectly equal variance. Also the QQ plot appears curvy, however it looks less sigmoidal than any of the SLR model QQ plots. Though conditions are still not met perfectly, this model fits better than either of the SLR models. We will consider this bivariate model with caution.

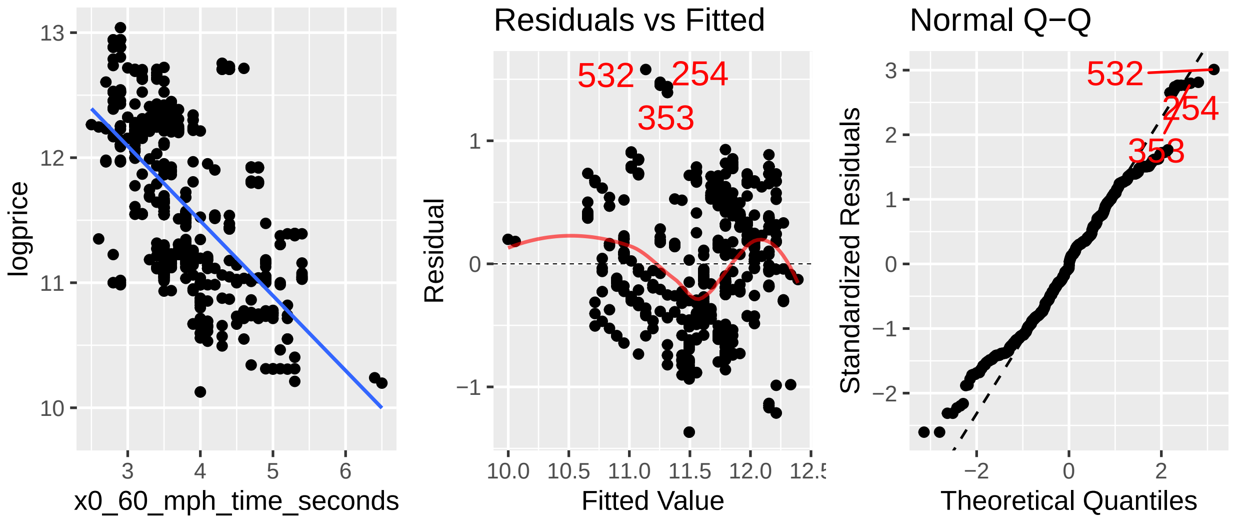
```
SportsCars <- mutate(SportsCars, logprice = log(price_in_usd))
horsepowerSLR2<- lm(logprice ~horsepower, data=SportsCars)
msummary(horsepowerSLR2)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.929184   0.082960  119.69   <2e-16 ***
## horsepower  0.003216   0.000152   21.16   <2e-16 ***
##
## Residual standard error: 0.4999 on 590 degrees of freedom
## Multiple R-squared:  0.4314, Adjusted R-squared:  0.4304
## F-statistic: 447.6 on 1 and 590 DF,  p-value: < 2.2e-16
```

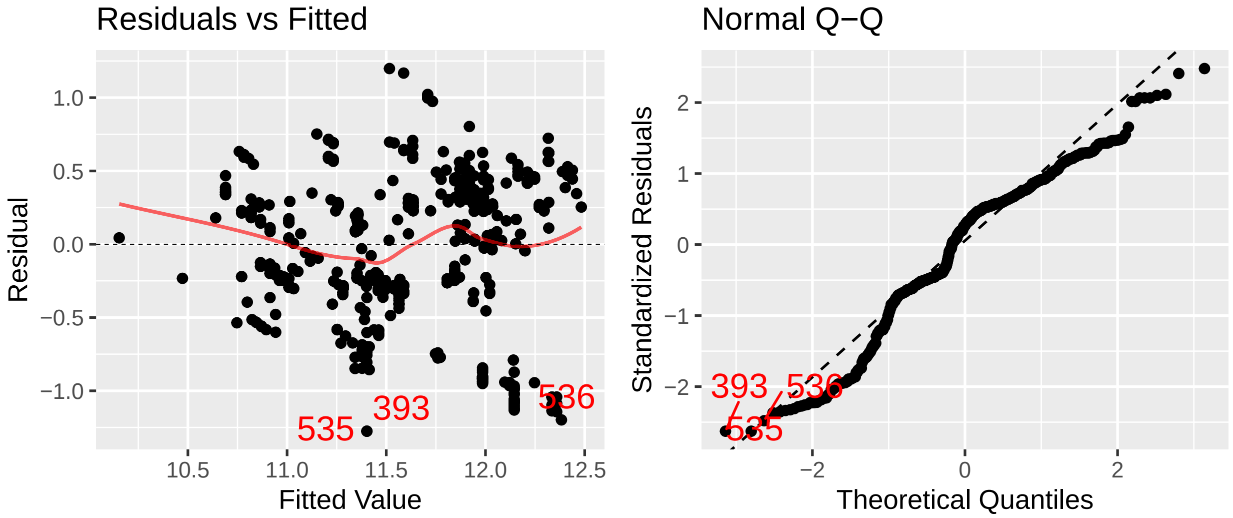
```
grid.arrange(c, a, b, ncol = 3)
```



```
grid.arrange(c, a, b, ncol = 3)
```



```r
grid.arrange(a, b, ncol = 2)
```



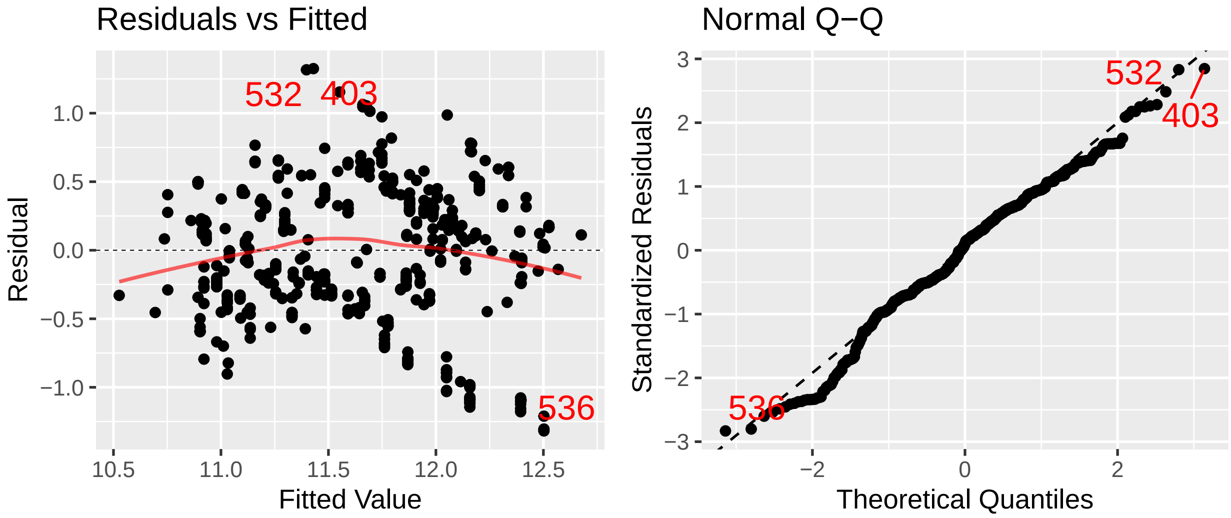
```r
favmod<- lm(logprice ~ year + engine_size_l + horsepower, data=SportsCars)
msummary(favmod)
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.292e+02  3.987e+01   5.748 1.45e-08 ***
## year           -1.084e-01  1.972e-02  -5.498 5.74e-08 ***
## engine_size_l -1.613e-01  2.014e-02  -8.009 6.24e-15 ***
## horsepower      4.307e-03  1.962e-04  21.953  < 2e-16 ***
##
## Residual standard error: 0.4676 on 587 degrees of freedom
## Multiple R-squared:  0.5041, Adjusted R-squared:  0.5015
## F-statistic: 198.9 on 3 and 587 DF,  p-value: < 2.2e-16
```

```
car::vif(favmod) #best model according to stepwise
```

```
##           year engine_size_l    horsepower
##       1.084608      1.938233      1.824585
```

```
grid.arrange(a, b, ncol = 2)
```

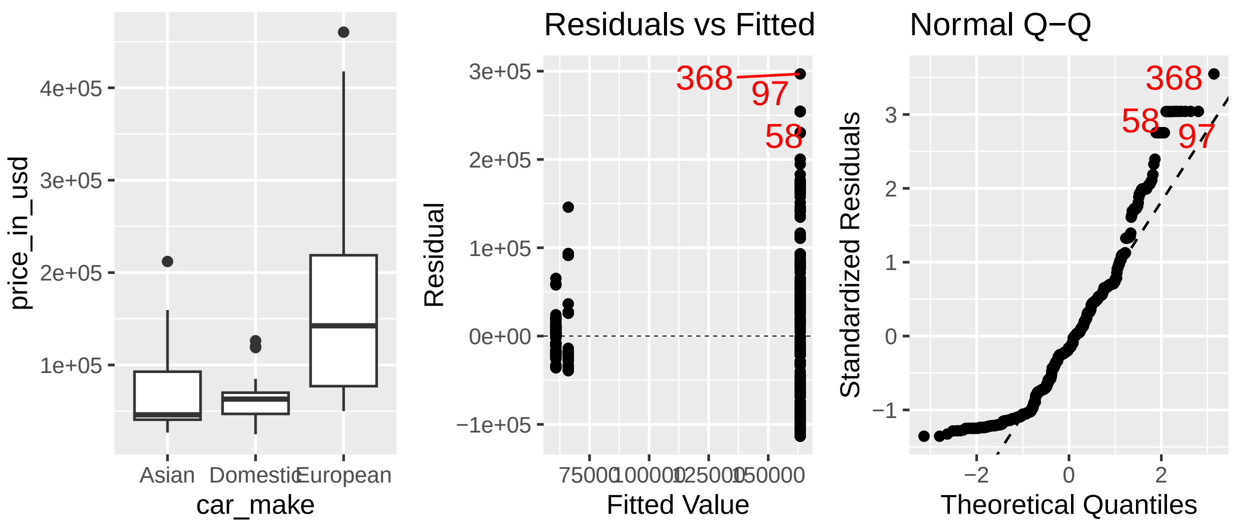


Next, we will attempt to run an ANOVA with the factor car_make, or the car region of origin. The results will allow us to determine if the geographical region of production affects the price enough to add it to a potential multivariate model. This factor has three categories: Asian, European, and Domestic. To get a better sense of this categorical variable, we compared box plots of each category. Immediately, we see that the European cars have a much higher standard deviation than both other categories, but especially domestic cars. Although this could be due to smaller sample sizes of Domestic and Asian car brands, in order to compare the categories at all, we need to check the requisite conditions to run an ANOVA. The condition of equal variance is clearly not passed, due to the largest over smallest standard deviation being well over 2 ($94260.05/18167.74 = 5.188$) and the obvious lack of continuity on the Residuals vs. Fitted plot. Furthermore, the condition of normality is not passed due to the tails of the QQ plot clearly separating from the line.

```
location <- lm(price_in_usd ~ car_make, data = SportsCars)
favstats(price_in_usd ~ car_make, data = SportsCars)
```

```
##   car_make   min      Q1 median       Q3    max      mean       sd   n missing
## 1    Asian 26830 40750.0  46050  92712.5 212000  66062.41 44474.55  54       0
## 2 Domestic 25000 47000.0  63000  69995.0 126190  60880.75 18167.74  87       0
## 3 European 50000 77062.5 142400 218750.0 460247 163421.82 94260.05 450       0
```

```
grid.arrange(c, a, b, ncol = 3)
```

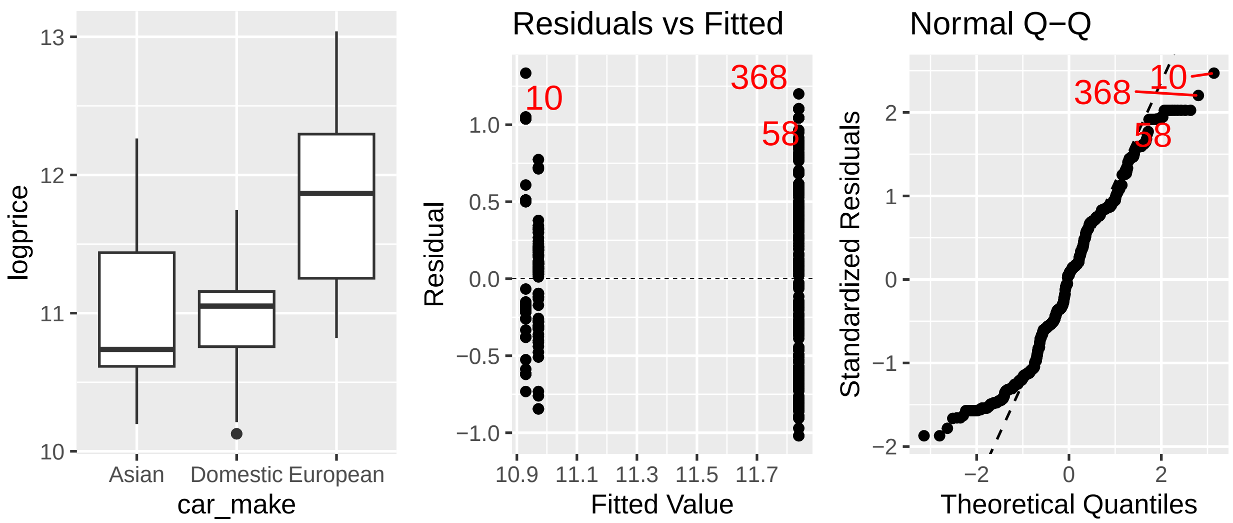


Since we did not pass conditions and still want to run a comparison test on this variable, the next step of action is to attempt a transformation. We attempted to take the natural log of the response variable, price. Although this helped increase our equal variance and pass the standard deviation test $(0.579/0.308 = 1.880)$, we still see a clear lack of normality in the QQ plot.

```
loglocation <- lm(logprice ~ car_make, data = SportsCars)
favstats(logprice ~ car_make, data = SportsCars)
```

```
##   car_make      min       Q1   median       Q3      max     mean        sd   n
## 1    Asian 10.19728 10.61471 10.73748 11.43725 12.26434 10.92960 0.5499697  54
## 2 Domestic 10.12663 10.75700 11.05089 11.15618 11.74554 10.97184 0.3081777  87
## 3 European 10.81978 11.25237 11.86639 12.29568 13.03952 11.83934 0.5793527 450
##   missing
## 1       0
## 2       0
## 3       0
```

```
grid.arrange(c, a, b, ncol = 3)
```



Since we could not find a transformation that allowed us to pass conditions for running an ANOVA, we next moved towards running a randomization F-test. This test can help us compare categories across a response variable even when we do not pass conditions for an ANOVA. First, we randomly shuffled the price 10000 times in order to compare the frequency at which a random shuffle (which should have no main effects) will take on the same F-value as we observe in this data set. Since we consistently found that a transformed response variable log(price) leads to better fits and improved linear model conditions, we only will show the randomization F-test for the average log(price) of each geographical region.

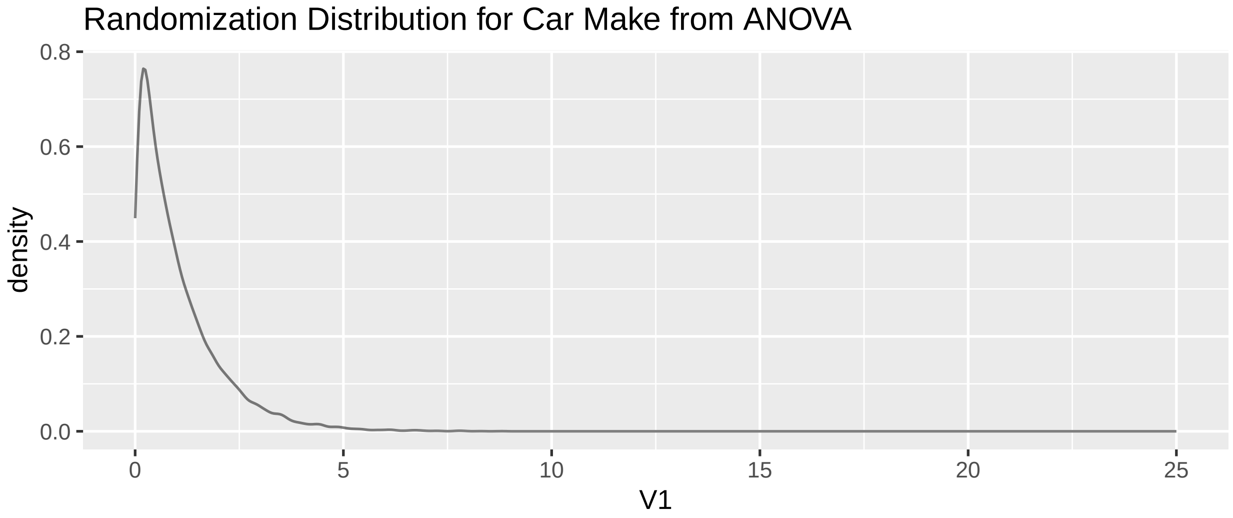
We then calculated the baseline F value, which is 141.02.

```
origres <- anova(loglocation)$"F value"#
origres
```

```
## [1] 141.0201       NA
```

We then will observe how often our randomly sampled sets end up with an F value close to the 141.02 we see in our original data set. We observe that the random samples do not come close to obtaining an F value that high. In fact, we see almost no F-values over 5.

```
gf_dens(~ V1, data = randomizedsamples) %>%
  gf_lims(x = c(0, 25)) %>%
  gf_labs(title = "Randomization Distribution for Car Make from ANOVA")
```



Next, we see the probability of getting an F like ours when the null (sports cars produced everywhere have the same log price) is true. This value is 0, indicating that we should reject the null. We have evidence, using F-randomization, that there is a significant difference in the average log price of a sports car between at least one of these three geographical regions.

```
pdata(~ V1, origres[1], data = randomizedsamples, lower.tail = FALSE)
```

```
## [1] 0
```

Since we got a significant result from the randomization F-test, we then will run a post-Hoc test to determine which categories are significantly different. We ran Tukey's HSD and found evidence that there are significant differences in the average log price of sports cars that are produced in Europe vs. Asia ($p < 0.001$) and significant differences in the average log price of cars produced domestically vs. those produced in Europe ($p < 0.001$). We did not find evidence of a significant difference in the average log price of cars produced domestically vs those produced in Asia ($p = 0.93$).

```
TukeyHSD(loglocation)
```

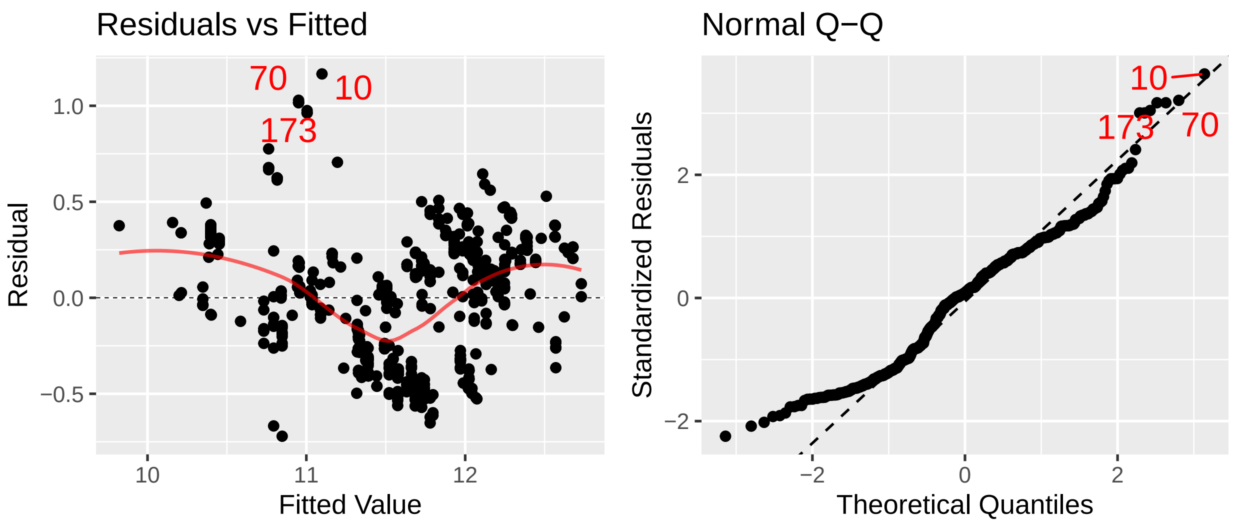
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = x)
##
## $car_make
##                          diff        lwr      upr     p adj
## Domestic-Asian    0.04224292 -0.1797662 0.264252 0.8957139
## European-Asian    0.90973883  0.7251821 1.094296 0.0000000
## European-Domestic 0.86749591  0.7174103 1.017581 0.0000000
```

Since there is evidence that the average log price of European cars is significantly higher than Asian or Domestic makes, we will attempt to add an indicator variable to our best multiple linear regression model to determine if it reduces the error. By indicating all European cars with a "1" and all others with a "0", we can add the variable to our model.

```
MLR6<- lm(logprice ~ year + engine_size_l + horsepower + car_make, data=SportsCars)
msummary(MLR6)
```

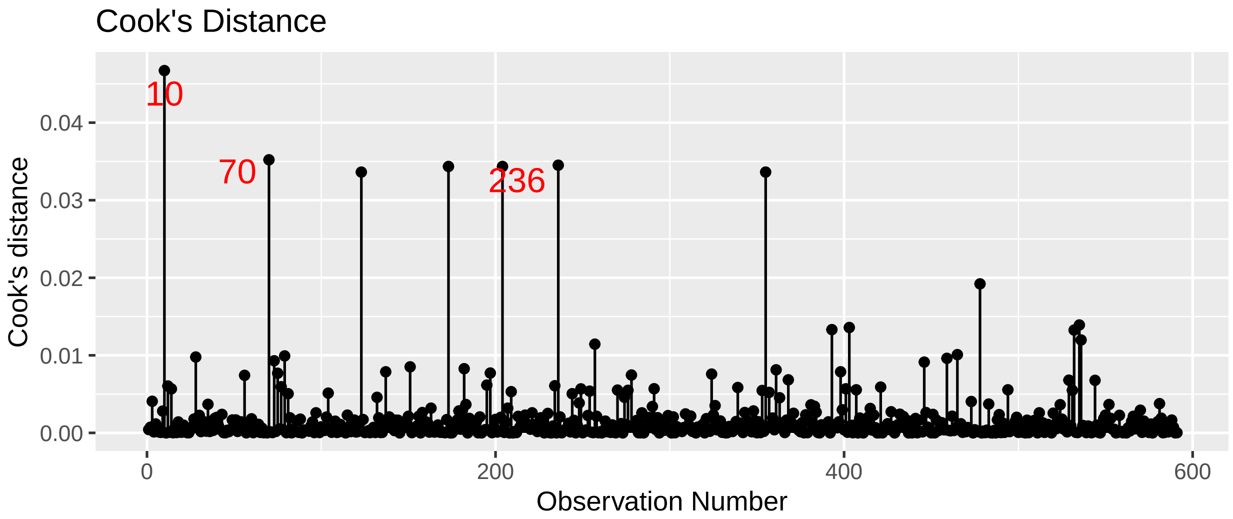
```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.176e+02  2.791e+01   4.212 2.93e-05 ***
## year           -5.362e-02  1.380e-02  -3.885 0.000114 ***
## engine_size_l   6.318e-02  1.650e-02   3.829 0.000143 ***
## horsepower      2.776e-03  1.485e-04  18.692  < 2e-16 ***
## car_make        9.374e-01  3.698e-02  25.346  < 2e-16 ***
##
## Residual standard error: 0.3232 on 586 degrees of freedom
## Multiple R-squared:  0.7634, Adjusted R-squared:  0.7618
## F-statistic: 472.7 on 4 and 586 DF,  p-value: < 2.2e-16
```

```
grid.arrange(a, b, ncol = 2)
```



We found that this model was significantly better than the model without the indicator variable, as the R^2 value increased from 0.50 to 0.76. The residual standard error decreased from 0.47 to 0.32. Although this model is good and passes conditions, we have a couple potential outliers to take a look at. Point #10 in particular looks to be a potential outlier based on the residuals vs. fitted graph. We ran a plot of Cook's distance to determine which, if any, of these points could potentially be removed. However, none of the points even came close to the threshold of a moderately influential point of >0.5. Therefore, we did not remove any additional observations.

```
mplot(MLR6, which = 4)
```

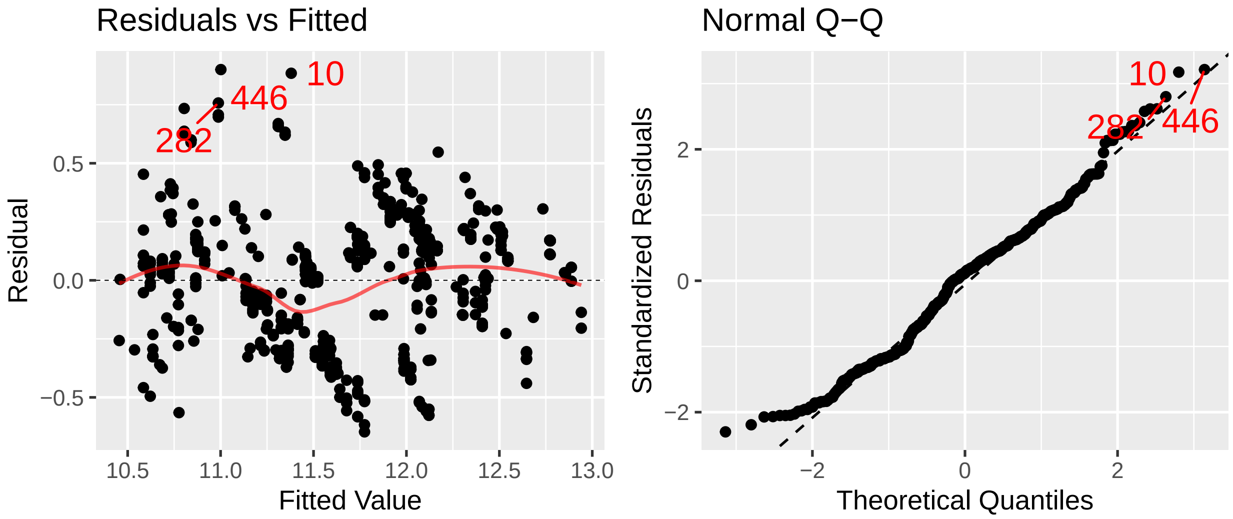


We then checked to see if adding interaction terms to our model improved the fit. Although adding all interaction terms increased the fit to an R^2 of 0.86, this model is not parsimonious. The best model we found added an interaction only between engine size and car region of origin, our indicator variable. This increased our R^2 value to 0.82 and lowered our standard error to 0.25. Adding the interaction term also appears to improve the model's conditions.

```
MLR6<- lm(logprice ~ year + engine_size_l * car_make + horsepower, data=SportsCars)
msummary(MLR6)
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            84.5678788 24.4541769   3.458 0.000583 ***
## year                   -0.0367818  0.0120965  -3.041 0.002466 **
## engine_size_l          -0.1448966  0.0209690  -6.910 1.27e-11 ***
## car_make               -0.4523229  0.1068766  -4.232 2.69e-05 ***
## horsepower              0.0028315  0.0001295  21.861  < 2e-16 ***
## engine_size_l:car_make  0.2876040  0.0210877  13.638  < 2e-16 ***
##
## Residual standard error: 0.2818 on 585 degrees of freedom
## Multiple R-squared:  0.8205, Adjusted R-squared:  0.819
## F-statistic: 534.8 on 5 and 585 DF,  p-value: < 2.2e-16
```

```
grid.arrange(a, b, ncol = 2)
```



We then added the indicator variable (European vs. not European) to our best bivariate model as well. According to the t-test, 0-60 time was not a significant predictor (p = 0.89), so we chose to remove that in future models. We then checked for colinearity using VIF, finding that our variables were not colinear.

```
car::vif(aovbivariate2)
```

```
## x0_60_mph_time_seconds             horsepower
##               2.449456               2.449456
```

```
bivarEuro<-lm(logprice ~x0_60_mph_time_seconds + horsepower + car_make, data=SportsCars)
msummary(bivarEuro)
```

```
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             9.287183   0.210014  44.222   <2e-16 ***
## x0_60_mph_time_seconds -0.004602   0.033262  -0.138     0.89
## horsepower              0.003231   0.000166  19.462   <2e-16 ***
## car_make                0.859114   0.033585  25.580   <2e-16 ***
##
## Residual standard error: 0.3345 on 587 degrees of freedom
## Multiple R-squared:  0.7462, Adjusted R-squared:  0.745
## F-statistic: 575.4 on 3 and 587 DF,  p-value: < 2.2e-16
```

```r
car::vif(bivarEuro)
```

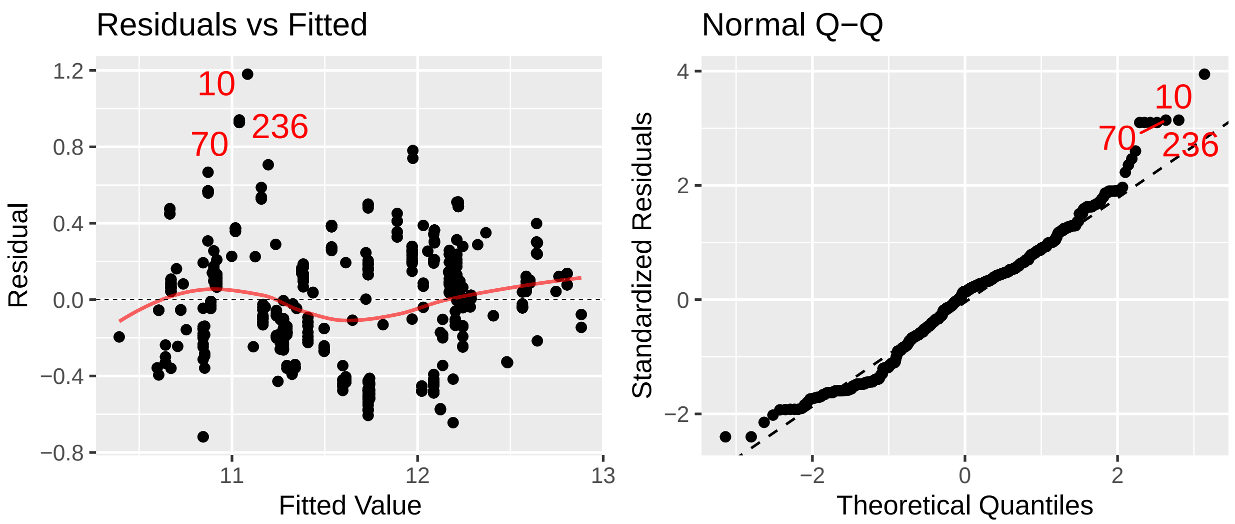
```
## x0_60_mph_time_seconds             horsepower               car_make
##               2.649949               2.552139               1.082428
```

After removing 0-60 time, we then checked to see if an interaction improved the fit of our model. We found that the inclusion of an interaction term between horsepower and car region improved the R^2 from 0.75 to 0.80 and decreased the standard error from 0.33 to 0.30. Furthermore, according to the t-test, the interaction term is significantly contributing the the model (p $<$ 0.001).

```r
bivarcorrected<-lm(logprice ~ horsepower * car_make, data=SportsCars)
msummary(bivarcorrected)
```

```
##                        Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          10.0937434  0.0892414 113.106  < 2e-16 ***
## horsepower            0.0016506  0.0001639  10.071  < 2e-16 ***
## car_make             -0.3801188  0.1085298  -3.502 0.000496 ***
## horsepower:car_make   0.0023651  0.0001994  11.861  < 2e-16 ***
## 
## Residual standard error: 0.3004 on 587 degrees of freedom
## Multiple R-squared:  0.7953, Adjusted R-squared:  0.7943 
## F-statistic: 760.2 on 3 and 587 DF,  p-value: < 2.2e-16
```

```r
grid.arrange(a, b, ncol = 2)
```


Residuals vs Fitted
Residual
Fitted Value
Normal Q−Q
Standardized Residuals
Theoretical Quantiles