Practicum Part II

Josiah Parry

Environment and data set up

The below code chunk loads the requisite libraries for this analysis.

```
library(tidyverse)
library(tidymodels)


cars <- readxl::read_excel("data/kellycarsalesdata.xlsx") %>%
  janitor::clean_names()
```

Using `rsample` I partition my data.

```
init_split <- initial_split(cars)
car_train <- training(init_split)
car_testing <- testing(init_split)
```

3. Outliers

```
skimr::skim(cars) %>%
  select(-c(n_missing, complete_rate))
```

Table 1: Data summary

Name	cars
Number of rows	804
Number of columns	9
Column type frequency:	
character	1
numeric	8
Group variables	None

Variable type: character

skim_variable	min	max	empty	n_unique	whitespace
make	4	9	0	6	0

Variable type: numeric

skim_variable	mean	sd	p0	p25	p50	p75	p100	hist
price	21343.14	9884.85	8638.93	14273.07	18025.0	26717.32	70755.47	
mileage	19831.93	8196.32	266.00	14623.50	20913.5	25213.00	50387.00	
cylinder	5.27	1.39	4.00	4.00	6.0	6.00	8.00	
liter	3.04	1.11	1.60	2.20	2.8	3.80	6.00	
doors	3.53	0.85	2.00	4.00	4.0	4.00	4.00	
cruise	0.75	0.43	0.00	1.00	1.0	1.00	1.00	
sound	0.68	0.47	0.00	0.00	1.0	1.00	1.00	
leather	0.72	0.45	0.00	0.00	1.0	1.00	1.00	

Upon looking at the distributions of the numeric variables, I feel confident in that there are no true outliers. Perhaps we can identify a few in `price`, but dollar values are always heavily right skewed and this is a fact of wealth accumulations and pricing. Perhaps we can find a few values that exceed the 1.5 IQR ranges. To check, I will use the `anomalize` package by Matt Dancho of Business Science University to check both the mileage and the price columns as these are our only continuous variables.

```
# outliers where? in every single column?
# no outliers in mileage
anomalize::anomalize(cars, mileage) %>%
  count(anomaly)
```

```
## # A tibble: 1 x 2
##   anomaly     n
##   <chr>   <int>
## 1 No        804
```

```
# 5 outliers of price using iqr
anomalize::anomalize(cars, price) %>%
  count(anomaly)
```

```
## # A tibble: 2 x 2
##   anomaly     n
##   <chr>   <int>
## 1 No        799
## 2 Yes         5
```

I will create a tibble containing the original data less the 5 observations that are deemed anomalies via IQR method. We could also use the GESD Method, but this is computational intensive and unnecessary at the moment.

```
cars_no_anomaly <- anomalize::anomalize(cars, price) %>%
  filter(anomaly == "No") %>%
  select(-anomaly, -price_l1, -price_l2)
```

4. Distributions

The distributions were visualized with `skimr::skim()` previously. The only variables which are continuous are `mileage` and `price`. These two variables display characteristics of normality with a right skew. We should use a heteroskedastic robust linear regression model to deal with the inevitible heteroskedacity due to a log normal distribution in price. We can transform the variables but the skews are not heavy enough to warrant log or inverse methods. Perhaps a square root transformation would be appropriate. Before making such an adjustment I would prefer to fit the model as is.

5. Correlations

To create the initial correlations, I will use the `corrr` package from tidymodels.

```
corrr::correlate(select_if(cars, is.numeric)) %>%
  corrr::focus(price) %>%
  arrange(-abs(price))
```

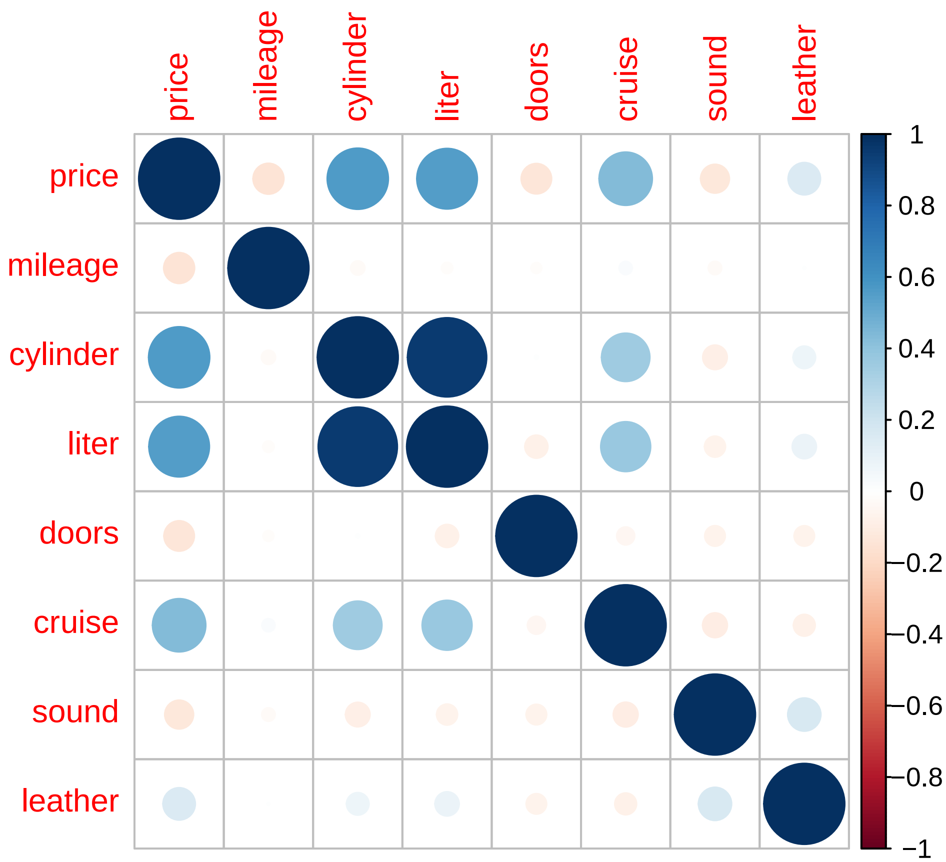
```
##
## Correlation method: 'pearson'
## Missing treated using: 'pairwise.complete.obs'
```

```
## # A tibble: 7 x 2
##   rowname   price
##   <chr>     <dbl>
## 1 cylinder  0.569
## 2 liter     0.558
## 3 cruise    0.431
## 4 leather   0.157
## 5 mileage  -0.143
## 6 doors    -0.139
## 7 sound    -0.124
```

There are strong correlations between price and cylinder, and price and liter.

To visualize the correlation matrix, I will use `corrplot`. First, I select only the numeric variables from the tibble, create a correlation matrix, and then create a correlation plot.

```
select_if(cars, is.numeric) %>%
  cor() %>%
  corrplot::corrplot()
```



Given the above plot there is a high amount of colinearity between liter and cylinder. One of these variables should likely be omitted.

6. Fitting a regression

This question requests principle components but there is no instruction to perform PCA, as such, I will not do that. I will use the `estimatr` package to fit a linear regression.

```
lm_1 <- estimatr::lm_robust(price ~ ., data = car_train)

lm_1 %>%
  broom::tidy() %>%
  knitr::kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
(Intercept)	15042.708	1123.476	13.389	0.000	12836.209	17249.206	590	price
mileage	-0.189	0.022	-8.618	0.000	-0.232	-0.146	590	price
makeCadillac	16042.128	776.420	20.662	0.000	14517.245	17567.012	590	price
makeChevrolet	-2169.519	429.425	-5.052	0.000	-3012.907	-1326.131	590	price
makePontiac	-1995.265	404.336	-4.935	0.000	-2789.378	-1201.152	590	price
makeSAAB	14572.571	463.771	31.422	0.000	13661.727	15483.414	590	price
makeSaturn	-2237.776	443.340	-5.048	0.000	-3108.493	-1367.059	590	price
cylinder	11.309	519.518	0.022	0.983	-1009.020	1031.638	590	price
liter	4523.600	630.169	7.178	0.000	3285.953	5761.247	590	price

term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
doors	-1690.736	300.425	-5.628	0.000	-2280.769	-1100.704	590	price
cruise	-299.176	308.970	-0.968	0.333	-905.991	307.639	590	price
sound	-222.632	341.386	-0.652	0.515	-893.112	447.849	590	price
leather	297.306	262.056	1.135	0.257	-217.370	811.981	590	price

```
broom::glance(lm_1) %>%
  knitr::kable(digits = 3)
```

r.squared	adj.r.squared	statistic	p.value	df.residual	N	se_type
0.88	0.877	466.582	0	590	603	HC2

The above model has created a very strong linear model. We see that this model explains nearly 88% of the variance in our response variable.

It appears that the biggest contributor to a car's price is the make of it. Perhaps, more than anything, we are purchasing names rather than the quality of car. Cadillac's unsurprisingly, add the most value to a car, followed by SAAB. Cadillac's, unlike SAABs, are still being produced today. Given an average car, we can anticipate that the baseline cost will be around $32k.

7 variable selection by p-value

We should not curate our linear models to only include "statistically significant" variables. This ruins the interprative powers of a linear regression. However, I will do so as these are the instructions.

```
lm_2 <- estimatr::lm_robust(price ~ mileage + make + cylinder + liter + doors + cruise + sound, data = c
lm_3 <- estimatr::lm_robust(price ~ mileage + make + liter + doors + cruise + sound, data = car_train)
lm_4 <- estimatr::lm_robust(price ~ mileage + make + liter + doors + sound, data = car_train)
lm_final <- estimatr::lm_robust(price ~ mileage + make + liter + doors, data = car_train)
```

```
broom::tidy(lm_final) %>%
  knitr::kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
(Intercept)	14822.548	1523.720	9.728	0	11830.015	17815.082	594	price
mileage	-0.189	0.022	-8.559	0	-0.232	-0.146	594	price
makeCadillac	16262.490	981.493	16.569	0	14334.871	18190.108	594	price
makeChevrolet	-1998.713	420.632	-4.752	0	-2824.821	-1172.606	594	price
makePontiac	-1867.342	410.421	-4.550	0	-2673.394	-1061.290	594	price
makeSAAB	14628.798	476.107	30.726	0	13693.740	15563.855	594	price
makeSaturn	-2068.268	473.244	-4.370	0	-2997.703	-1138.834	594	price
liter	4502.860	192.530	23.388	0	4124.738	4880.983	594	price
doors	-1678.406	273.486	-6.137	0	-2215.524	-1141.288	594	price

```
broom::glance(lm_final)
```

```
##   r.squared adj.r.squared statistic       p.value df.residual   N se_type
```

```
## 1 0.8793987      0.8777745   640.2201 2.951103e-286         594  603      HC2
```

This final model performs as well as the original one. Now, if we are after performance and not inference, this is a completely fine conclusion.

We find that, like the original model, the make of a car has the biggest impact on the value of a car. Now that we have removed some variables—such as cylinders—other variables will be compensating for the variance that is explained by them. For example we know that cylinders and liters are highly correlated—though not completely—so liters is likely taking on a bit of the explanatory power of cylinders.

8 Leather

According to the initial model, the presence of a leather interior increase value only by 3 dollars. The last model generated does not include the variable at all, thus we can infer that it does not add any value.

9.

The inclusion of year is misleading as this is not included in the dataset.

```r
test_val <- tibble(
  doors = 4,
  make = "SAAB",
  mileage = 61435,
  cruise = 1,
  cylinder = 4,
  liter = 2.3,
  sound = 1,
  leather = 1
)

estimatr:::predict.lm_robust(lm_1, test_val, interval = "confidence")
```

```
## $fit
##            fit     lwr      upr
## [1,] 21458.26 19572.3 23344.21
```

```r
estimatr:::predict.lm_robust(lm_final, test_val, interval = "confidence")
```

```
## $fit
##            fit      lwr      upr
## [1,] 21486.23 19579.03 23393.42
```

The predictions for our example SAAB car are *very* similar. However, due to the inclusion of different variables, we are returned very slightly different prediction intervals.