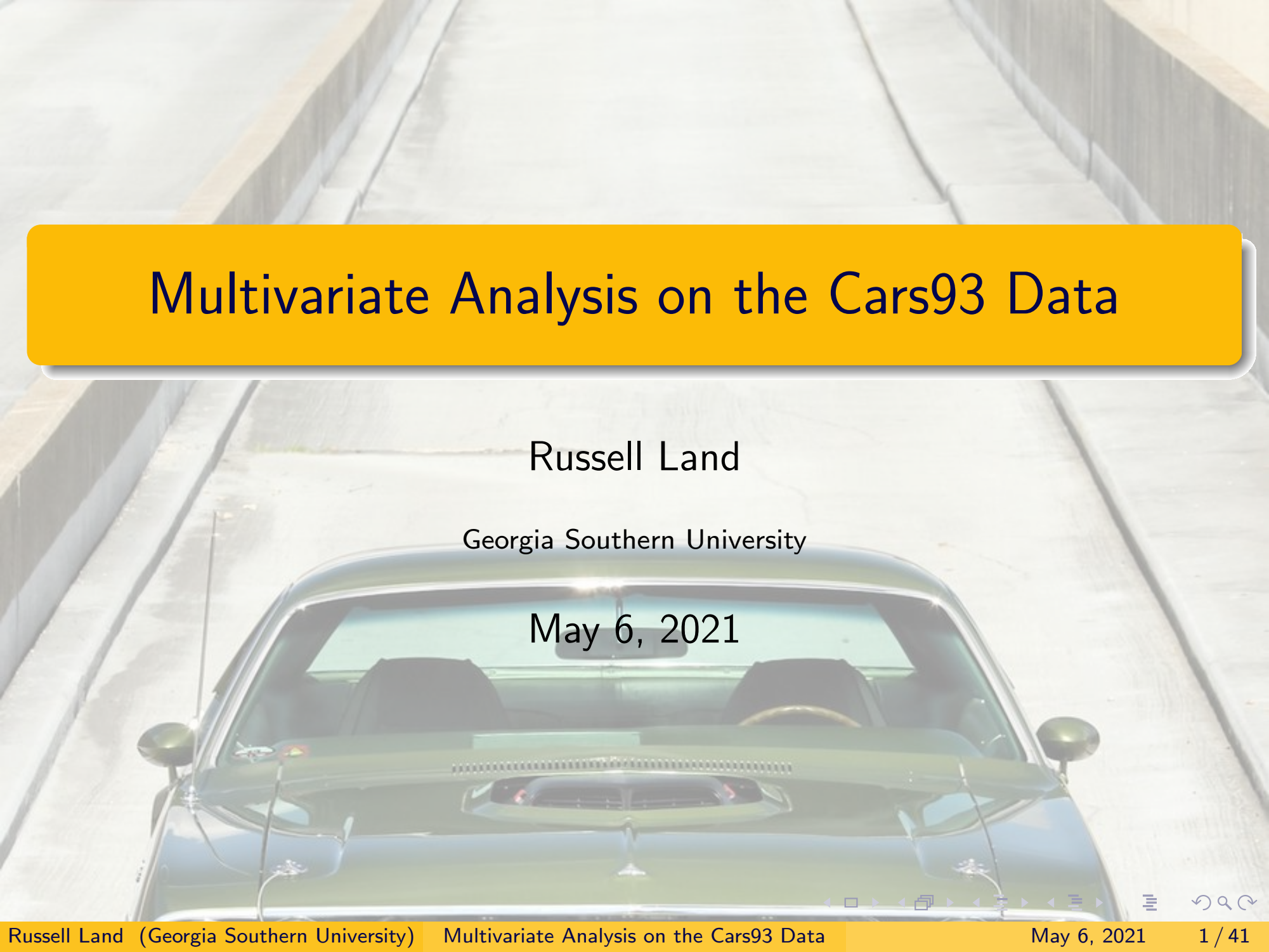
Multivariate Analysis on the Cars93 Data

Russell Land

Georgia Southern University

May 6, 2021



Introduction

Most of you have probably heard of the Cars93 data set. It is a data set in the "MASS" package for R.

Here are my goals for this project:

- Clean the data set.
- Create a suitable model that can make predictions for response variables of my choosing.
- Use multivariate linear regression to make predictions with principal component analysis.
- Use multivariate logistic regression to make predictions with principal component analysis.
- Run hypothesis test to see how close the car data is to my car.
- Run a cluster analysis and compare clusters.

The Data

From the R Documentation:

> Cars were selected at random from among 1993 passenger car models that were listed in both the Consumer Reports issue and the PACE Buying Guide. Pickup trucks and Sport/Utility vehicles were eliminated due to incomplete information in the Consumer Reports source. Duplicate models (e.g., Dodge Shadow and Plymouth Sundance) were listed at most once.

The Data

Here are the variables and the class of each:

No.	Variables	Class
1	Manufacturer	factor
2	Model	factor
3	Type	factor
4	Minimum Price	numeric
5	Price	numeric
6	Maximum Price	numeric
7	City MPG	integer
8	Highway MPG	integer
9	Airbags	factor

The Data

No.	Variables	Class
10	Drive Train	factor
11	Cylinders	factor
12	Engine Size	numeric
13	Horsepower	integer
14	RPM	integer
15	Revolutions per Mile	integer
16	Manual Transmission (Y/N)	factor
17	Fuel Tank Capacity	numeric
18	Passengers	integer

The Data

No.	Variables	Class
19	Length	integer
20	Wheelbase	integer
21	Width	integer
22	Turn Circle	integer
23	Rear Seat Room	numeric
24	Luggage Room	integer
25	Weight	integer
26	Origin (USA/non-USA)	factor
27	Make	factor

Unique Car

The Mazda RX-7 is the only car in the data set with a different type of engine; a rotary engine. We will see this car again...

Cleaning the Data

The first thing I want to do is change the factor variables into dummy variables. I can do this with the "`fastDummies`" package. I will do this for the variables:

- Type (small, midsize, compact, large, sporty, van)
- Airbags (none, driver only, driver and passenger)
- DriveTrain (FWD, RWD, 4WD)
- Man.trans.avail (yes, no)
- Origin (USA, non-USA)

"Man.trans.avail" will be become binary (0,1) with 1 being manual transmission available. "Origin" will be the same with 1 begin origin in the USA.

Cleaning the Data

I will remove the following variables from the data set:

- Manufacturer
- Model
- Min.Price
- Max.Price
- Rev.per.mile
- Make

I removed Manufacturer, Model, and Make because each observation was a unique car. If I want to identify a particular car later on I can simply call the index the car appears.

Model Building

I want to predict gas mileage (both highway and city). Therefore I want to eliminate variables that do not provide much information for this prediction. I will use **stepwise variable selection** to accomplish this.

I will use the built-in `stepAIC()` function to perform both forward and backward stepwise selection. After running, I remained with 13 variables out of the original 30.

Principal Component Analysis

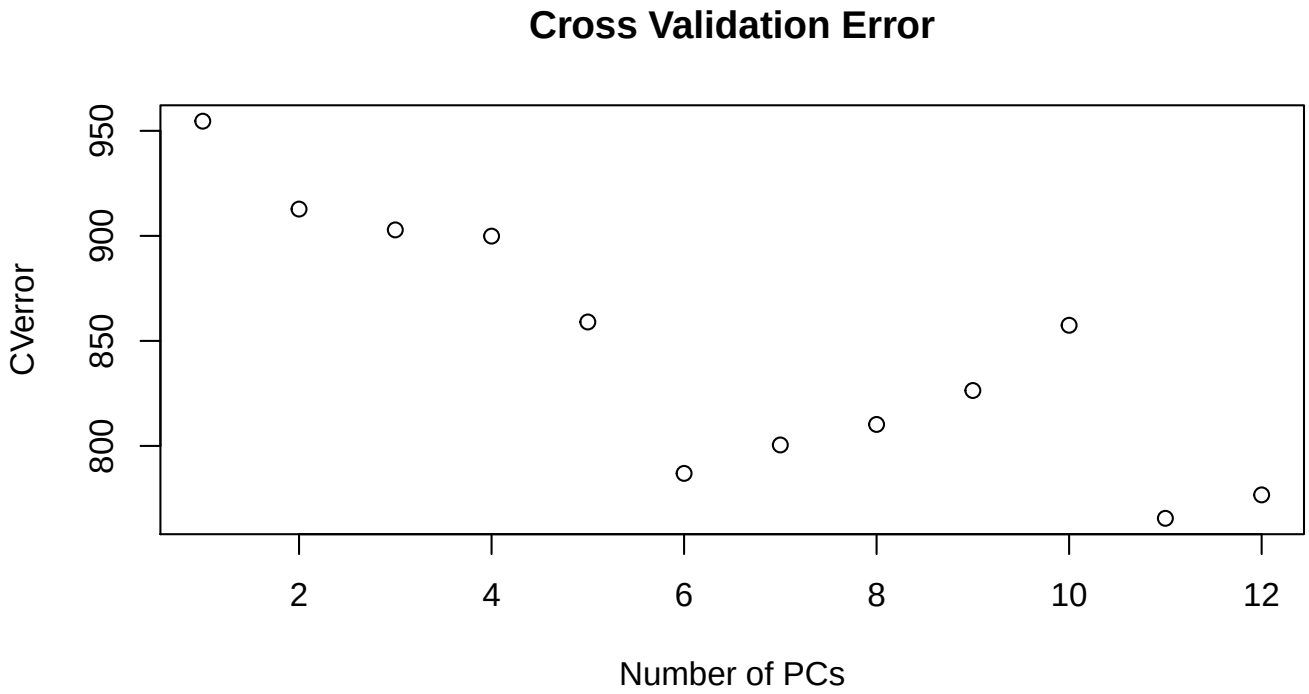
I want to use principal component analysis to reduce dimensions of the model. I will use **cross validation** to choose the number of principal components that minimize the prediction error. Here is psudeocode for the process:

```
for i in length(variables)-1:
      for j in variables:
            remove an observation
            train model and make prediction on removed observation
      repeat for each principal component
      record the sum-of-squared error

choose number of principal components that minimized the error
```

Starting with 13 variables, the number of principal components that minimized the error after cross validation is 10. Although 10 was the minimum, let's plot the cross validation error.

Cross Validation Plot



This looks like if we choose 11, we are barely improving over choosing 6. For our final model, let's choose 6 principal components.

Predictions

Let's begin by making predictions on each car, for both city and highway MPG.

It performs fine.

More Predictions

Now let's make predictions on my car. I have recorded the specifications for my model, a 2014 Volkswagen Passat.

Almost!

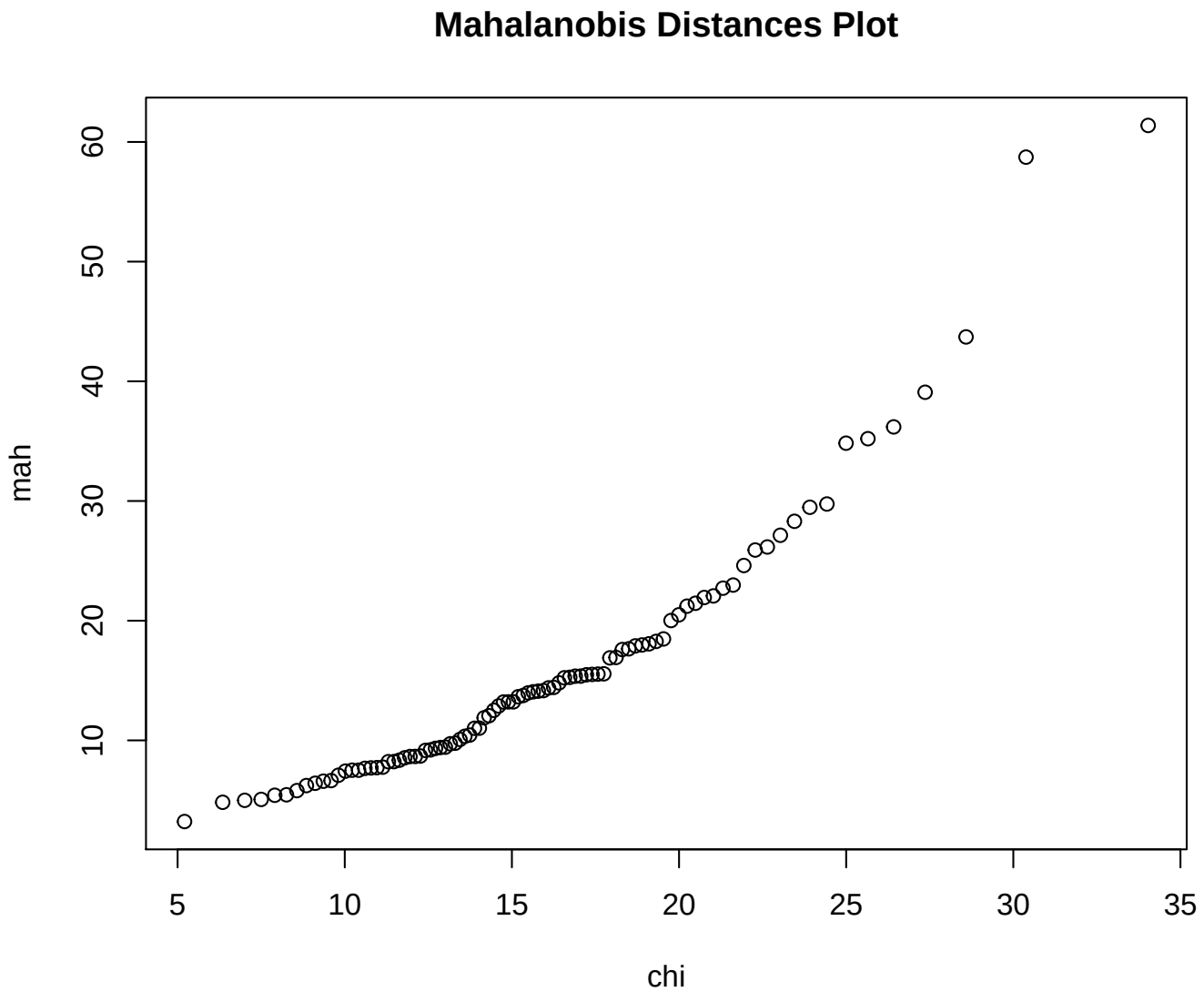
Multivariate Linear Regression (MvLR)

Now I will perform multivariate linear regression where there are two response variables, city and highway MPG. I will also be using six principal components just like I did before.

Logistic Regression

Now to consider different variables, let's take a look at "Origin". Let's use logistic regression to make predictions whether a car is made domestically or internationally. For this I will continue to use principal components when building the model and making predictions.

Mahalanobis Distances



Hypothesis Test

I want to test the following claim:

> The mean of the Cars93 data set is the same, or similar, to the specifications of my car.

Equivalently we can form the hypotheses,

$$H_0 : \vec{\mu} = \mu_{\text{passat}} \qquad H_1 : \vec{\mu} \neq \mu_{\text{passat}}$$

I will only consider the first 16 variables from the model, because the means of binary variables do not have an interpretation for a test for means.

Before we run this test, I predict that the data is very different from my cars specifications. This data is from 1993, and cars today are bigger, better, and more efficient.

Hypothesis Test

I will complete this test in R at a 95% significance level, and here are the expression that I will calculate for both the test statistic and critical value.

$$n\left(\vec{x}-\mu_0\right)' S^{-1}\left(\vec{x}-\mu_0\right) \sim \frac{(n-1)\,p}{n-p} F_{p,n-p}$$

The test statistic is 8,135.094 and the critical value is 33.963. As I anticipated, we reject the null hypothesis and conclude that the mean from the Cars93 data set is much different than the specifications of my car. I predict that this test would be similar for most modern vehicles.

Confidence Intervals for Means

For the same variables used in the hypothesis test, let's make simultaneous confidence intervals for the means at 95% level of significance. I will make both shadows of confidence regions and Bonferroni simultaneous confidence intervals. To do so, I will use the following expressions:

Shadows of Confidence Regions:

$$\left(\bar{x}_i - \sqrt{\frac{(n-1)\,p}{n-p} F_{p,n-p}(\alpha)} \left(\frac{s_{ii}}{\sqrt{n}}\right),\ \bar{x}_i + \sqrt{\frac{(n-1)\,p}{n-p} F_{p,n-p}(\alpha)} \left(\frac{s_{ii}}{\sqrt{n}}\right)\right)$$

Bonferroni Simultaneous confidence intervals:

$$\left(\bar{x}_i - t_{n-1}(\alpha/2p) \left(\frac{s_{ii}}{\sqrt{n}}\right),\ \bar{x}_i + t_{n-1}(\alpha/2p) \left(\frac{s_{ii}}{\sqrt{n}}\right)\right)$$

Shadows of Confidence Regions

	LB	UB	My Car
Price	13.672321	25.347034	20.8
MPG.city	18.969444	25.761738	24.0
MPG.highway	25.863969	32.308074	34.0
Cylinders	4.070303	5.757654	4.0
EngineSize	2.040846	3.294638	1.8
Horsepower	112.177217	175.478697	170.0
RPM	4920.030121	5641.260202	4800.0
Fuel.tank.capacity	14.682737	18.646295	18.5
Passengers	4.458149	5.713894	5.0
Length	174.379835	192.028767	191.6
Wheelbase	99.824993	108.067480	110.4
Width	67.092639	71.660049	72.2
Turn.circle	37.009116	40.904862	36.4
Rear.seat.room	24.196634	30.265732	39.1
Luggage.room	9.035131	15.459493	15.9
Weight	2716.418800	3429.387652	3166.0

Bonferroni Simultaneous Confidence Intervals

	LB	UB	My Car
Price	16.469443	22.549912	20.8
MPG.city	20.596797	24.134386	24.0
MPG.highway	27.407900	30.764143	34.0
Cylinders	4.474572	5.353385	4.0
EngineSize	2.341240	2.994244	1.8
Horsepower	127.343497	160.312417	170.0
RPM	5092.828253	5468.462069	4800.0
Fuel.tank.capacity	15.632358	17.696674	18.5
Passengers	4.759011	5.413032	5.0
Length	178.608309	187.800293	191.6
Wheelbase	101.799794	106.092679	110.4
Width	68.186936	70.565752	72.2
Turn.circle	37.942490	39.971488	36.4
Rear.seat.room	25.650717	28.811648	39.1
Luggage.room	10.574331	13.920292	15.9
Weight	2887.237640	3258.568811	3166.0

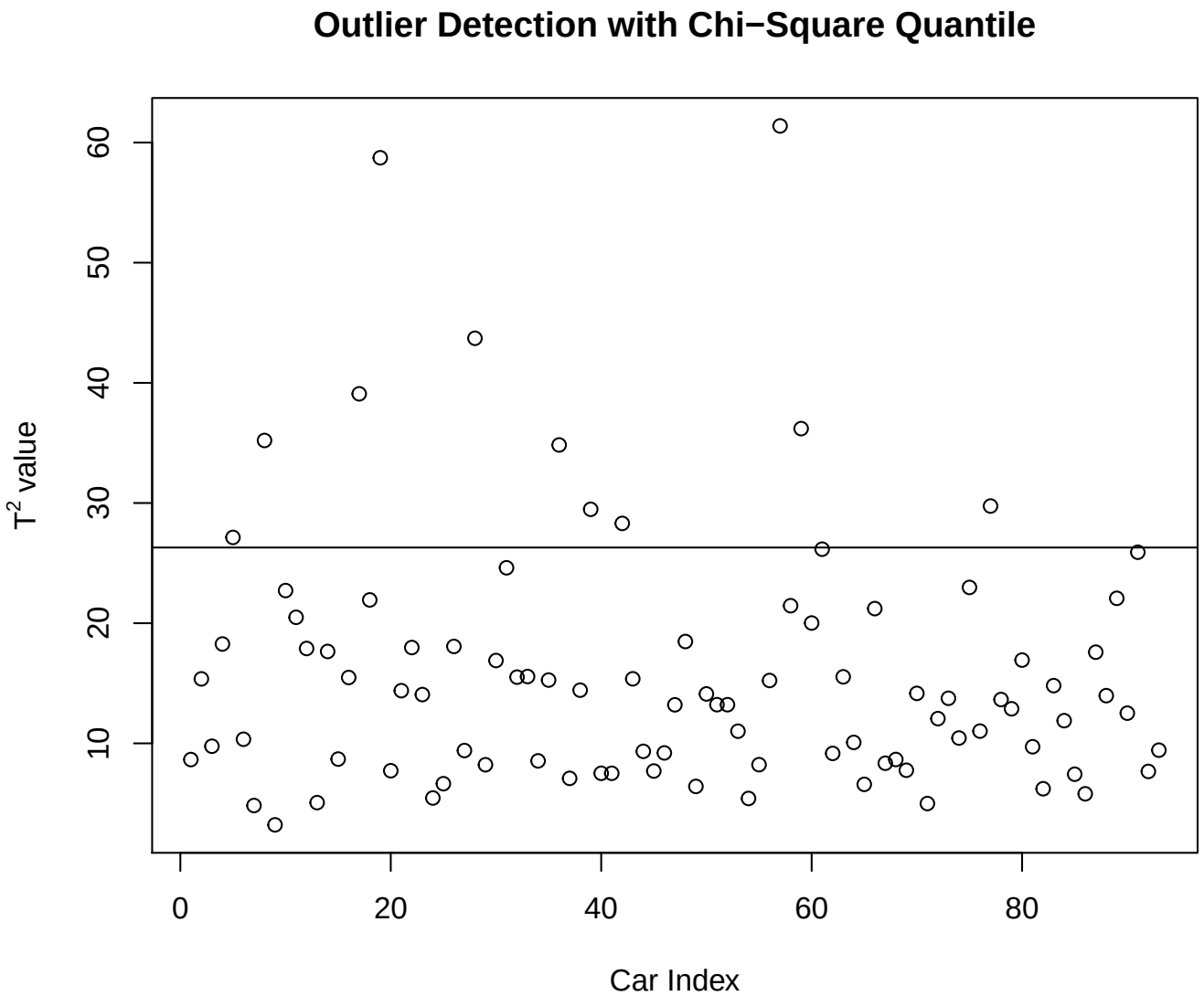
Outlier Detection

To find outliers we can use a chi-square quantile with

$$T_i^2 = (x_i - \hat{x})' S^{-1} (x_i - \hat{x})$$

Of course this data must meet the assumptions of normality. Even though this assumption is not met, let's still take a look at the outliers it gives.

Outlier Detection



Outliers

(a) BMW 535i

(b) Chevrolet Corvette

(c) Mercedes-Benz 300E

(d) Mazda RX-7

Outliers

(a) Buick Roadmaster

(b) Dodge Stealth

(c) Chevrolet Astro

(d) Ford Aerostar

Outliers

(a) Pontiac Bonneville

(b) Geo Metro

(c) Honda Civic

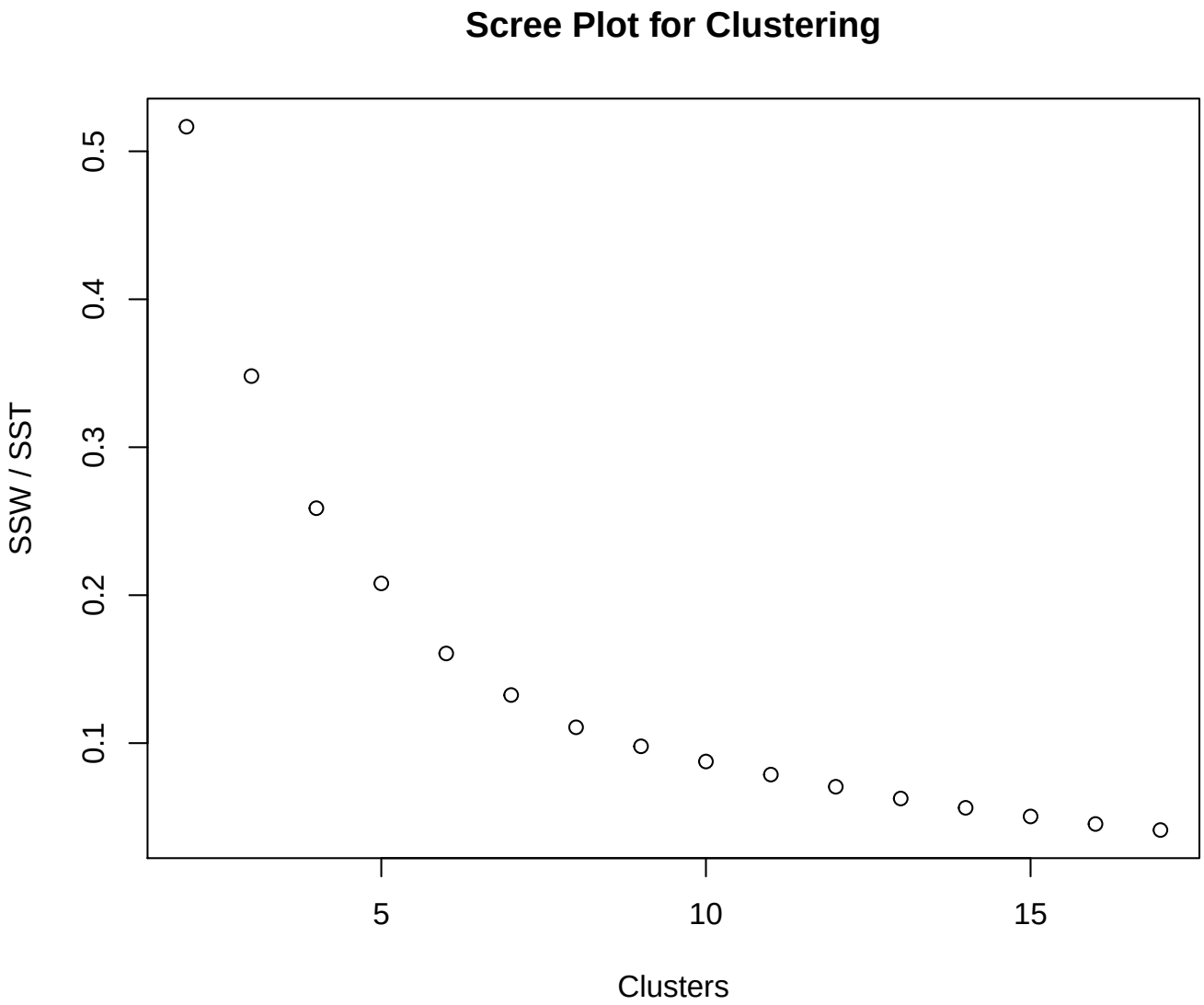
Cluster Analysis

I will perform two different clustering approaches. Maybe we will see some similarities.

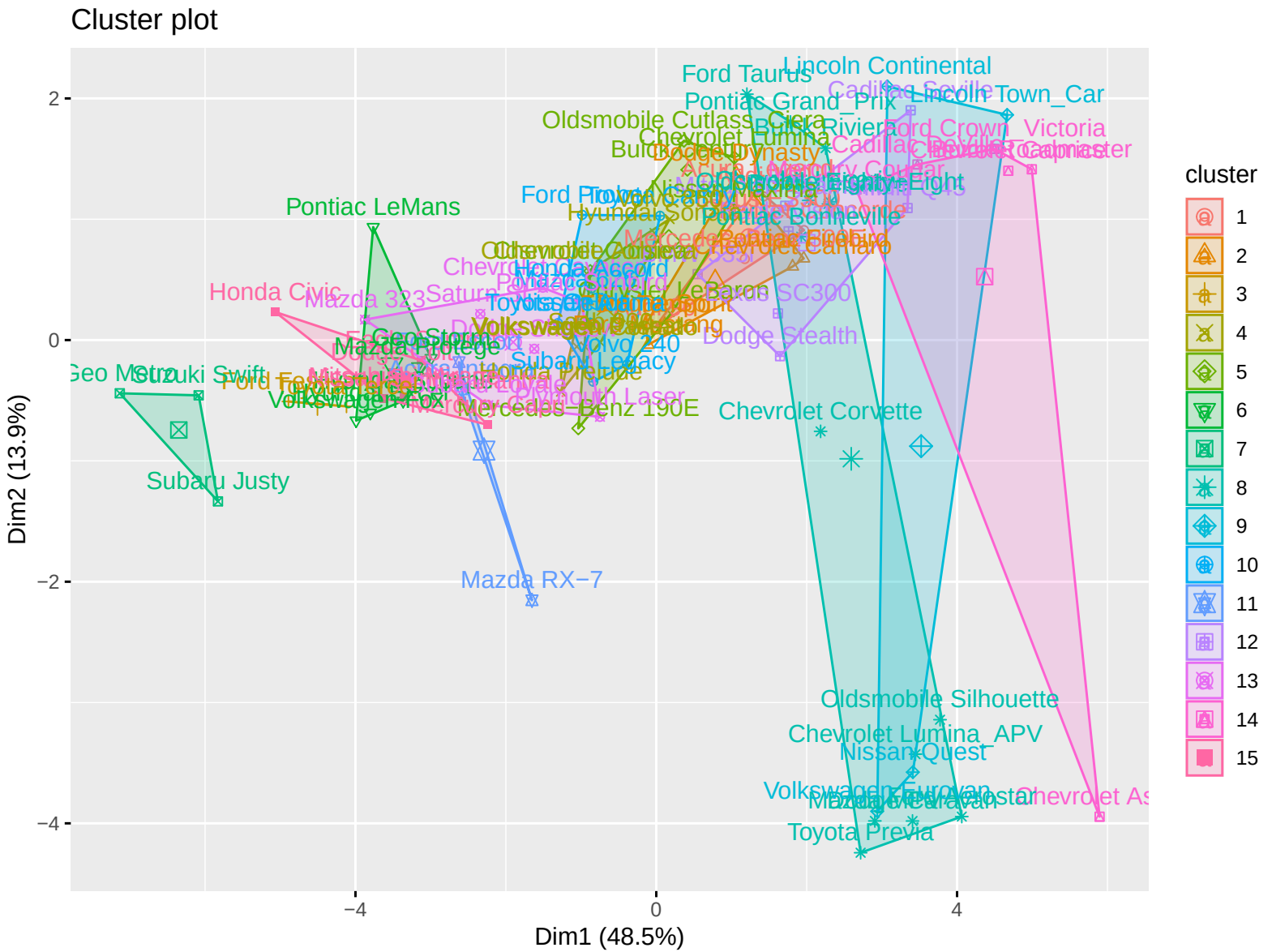
I am going to perform k-means clustering to form clusters that contain similar cars. I will use the built-in `kmeans()` function in the "`factoextra`" package. I need to pick a suitable k by looking at a scree plot.

Afterwards, I will perform complete linkage hierarchical clustering using the "`cluster`" package in R. To create a distance matrix I will use the euclidean distances between each observation.

k-means Cluster Analysis



k-means Cluster Analysis



Cluster 3

(a) Ford Festiva

(b) Toyota Tercel

Cluster 10

(a) Ford Probe

(b) Honda Accord

(c) Mazda 626

(d) Nissan Altima

Cluster 10

(a) Subaru Legacy

(b) Toyota Celica

(c) Toyota Camry

(d) Volvo 240

Cluster 12

(a) BMW 535i

(b) Cadillac Seville

(c) Dodge Stealth

(d) Infiniti Q45

Cluster 12

(a) Lexus SC300

(b) Mitsubishi Diamante

(c) Eagle Vision

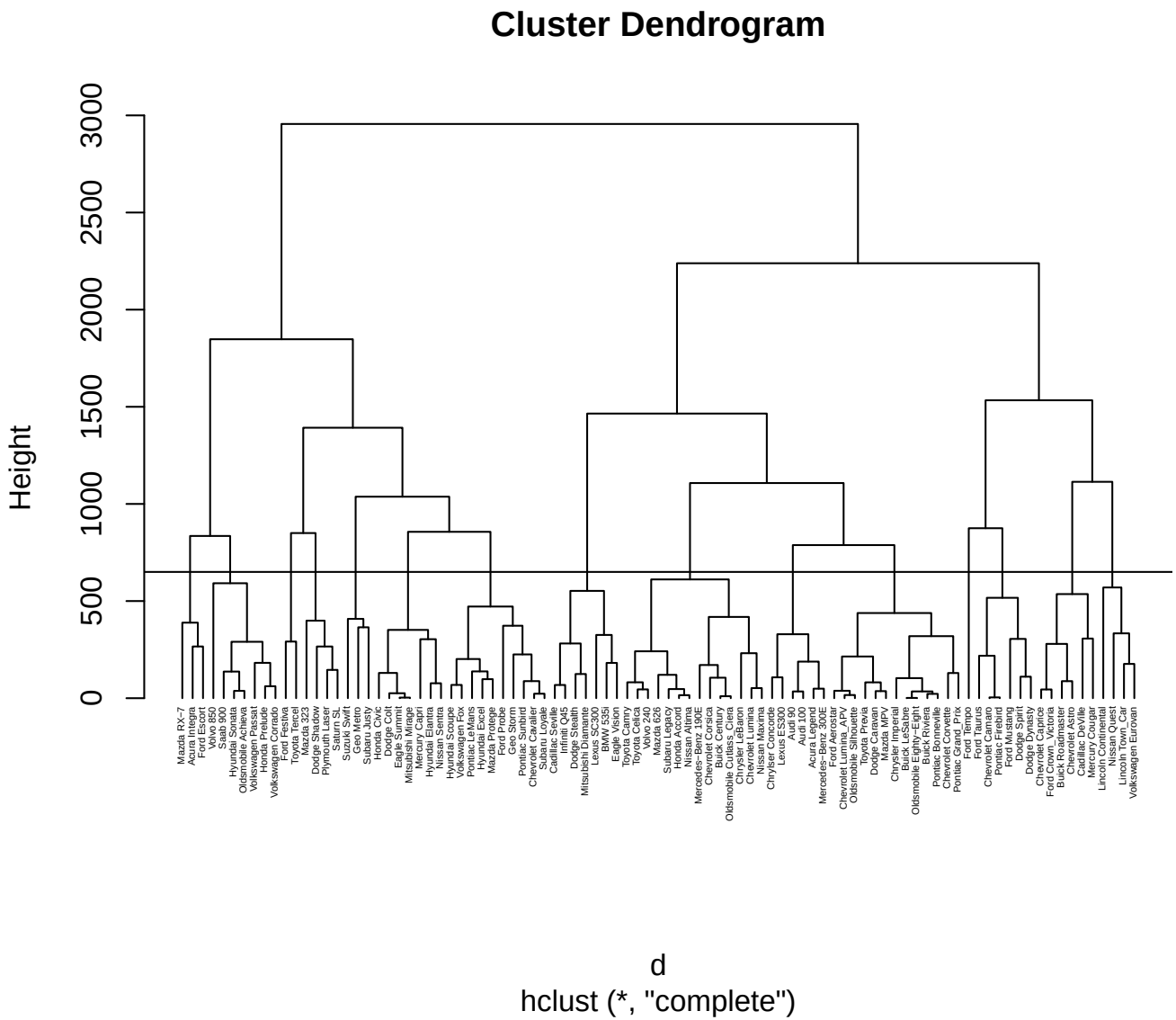
Complete Linkage Cluster Analysis

To perform this, we need a distance matrix. I will use the euclidean distances between each observation.

```
distances <- matrix(0, nrow = nrow(cars), ncol = nrow(cars))
for (i in 1:nrow(distances)) {
  for (j in 1:ncol(distances)) {
    if (i > j) {
      distances[i,j] <- sqrt(sum((cars[i,] - cars[j,])^2))
    }
    else {
      distances[i,j] <- 0
    }
  }
}
```

The result is a 93×93 lower-triangle distance matrix. Next, I will use the `hclust()` function to perform the complete linkage cluster analysis. At a height of 650, there are 15 clusters formed.

Complete Linkage Cluster Analysis



Cluster 3

(a) BMW 535i

(b) Cadillac Seville

(c) Dodge Stealth

(d) Eagle Vision

Cluster 3

(a) Infiniti Q45

(b) Lexus SC300

(c) Mitsubishi Diamante

Conclusion

- Every aspect of this project worked out better than I thought it would.
- The regression models were making a lot of correct predictions, besides the outliers.
- The cluster analysis was able to group similar cars together.
- I was able to identify outliers, which were clearly outliers compared to the majority of other cars.
- I learned more about cars in general, especially cars from the 90's!

Thank You