Overall

An automobile manufacturer, AutosRUs, is testing its newest prototype, the MechaCar, but is encountering production issues. There are two areas of concern: one is the potential impact of specific design characteristics on the fuel efficiency of the prototype; and the second involves possible manufacturing problems with the suspension coils installed on the vehicles. This project is divided into two main analyses to address the two issues. The first deals with the issue of fuel efficiency as measured against five design variables. This analysis uses the MechaCar data set. The second analyzes differences in the weight capacity of three manufacturing lots of suspension coils. Data for this analysis is contained in the Suspension Coils data set. We begin with the MechaCar analysis of fuel consumption (miles per gallon, or mpg).

Part I - MechaCar MPG

Purpose

The purpose of this part of the analysis is to determine if there is a statistical relationship between five design attributes of the prototype and its fuel efficiency. These attributes include length, weight, spoiler angle, ground clearance, and drive type. Using a linear model to regress the five independent variables against mpg, this analysis did prove useful in predicting mpg variability based on three of the five attributes: ground clearance, vehicle length, and to a lesser degree, vehicle weight.

Data Review

The MechaCar data set consists of six variables and 50 observations in csv format read in as an R dataframe. The first step was to establish if the data came from a normal distribution. Smaller data sets, usually defined as less than 50 observations, may have non-normal distributions which can affect the linear model. For that reason I employed two major tests for normality, the Shapiro-Wilk test, and histograms with density plotting.

Tests of Normality

Shapiro-Wilk

Description

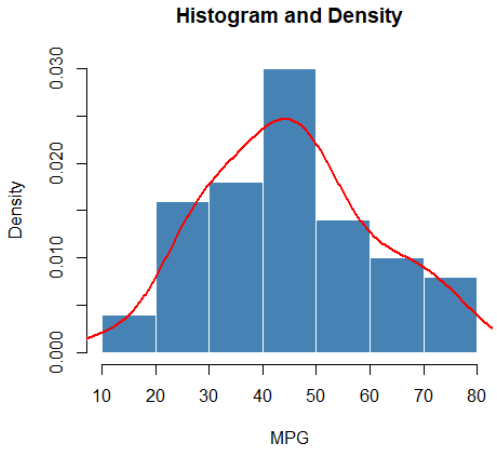
Shapiro-Wilk identifies the probability that the data being analyzed is from a normal distribution. It generates two statistics, the w and the p. The w statistic in conjunction with the degrees of freedom calculated in the test provides the p-value. The p value is the statistic by which we measure the validity of the null hypothesis based on the level of confidence set at the outset of the analysis.

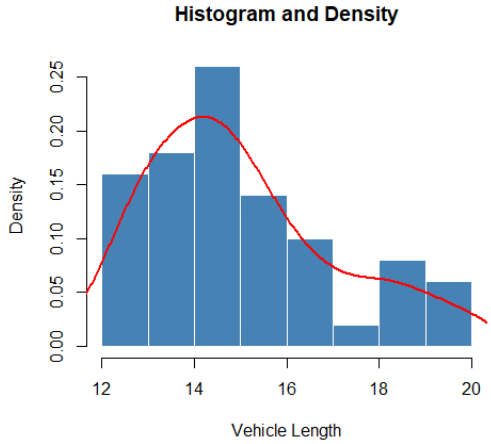
The null hypothesis (Ho) for Shapiro-Wilk is of a normal distribution. The alternate hypothesis (Ha) is non-normal distribution. Assuming an alpha of .05 (95% confidence level) the following is the guide:

P>.05: fail to reject Ho: not sufficient evidence to say the sample does not come from a normal distribution.

p<.05: reject the null hypothesis: sufficient evidence to say that sample does not come from a normal distribution.

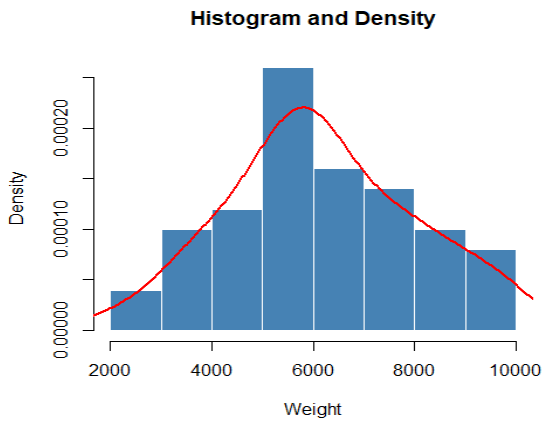
Results of Shapiro-Wilk with Histogram and Density Plots
All but one feature displayed normal distribution. That feature was vehicle length; however non-normality was slight.

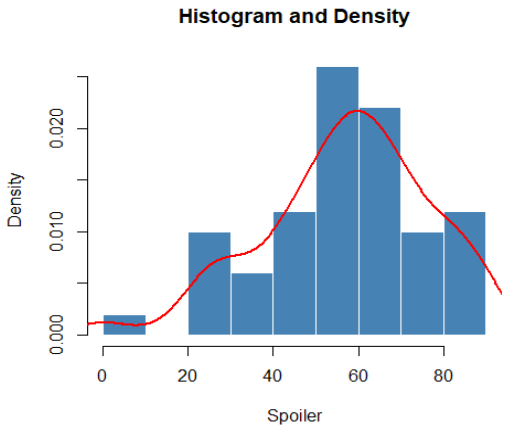




MPG
data: Mecha_Car$mpg
W = 0.98536, p-value = 0.7869
p-value is >.05; fail to reject Ho: sample is from a normal distribution

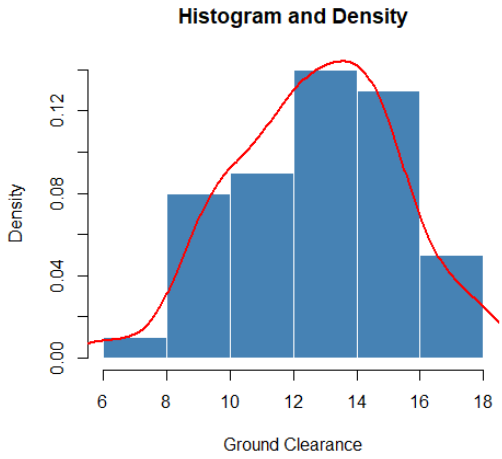
Length
data: Mecha_Car$vehicle_length
W = 0.93421, p-value = 0.008003
p-value is <.05; reject Ho:
sample is not from a normal distribution; but, w is very close to 1, and by this metric indicates normal distribution





Weight
data: Mecha_Car$vehicle_weight
W = 0.98626, p-value = 0.8242
p-value is >.05; fail to reject Ho: sample is from a normal distribution

Spoiler Angle
data: Mecha_Car$spoiler_angle
W = 0.97201, p-value = 0.2788
p-value is >.05; fail to reject Ho: sample is from a normal distribution

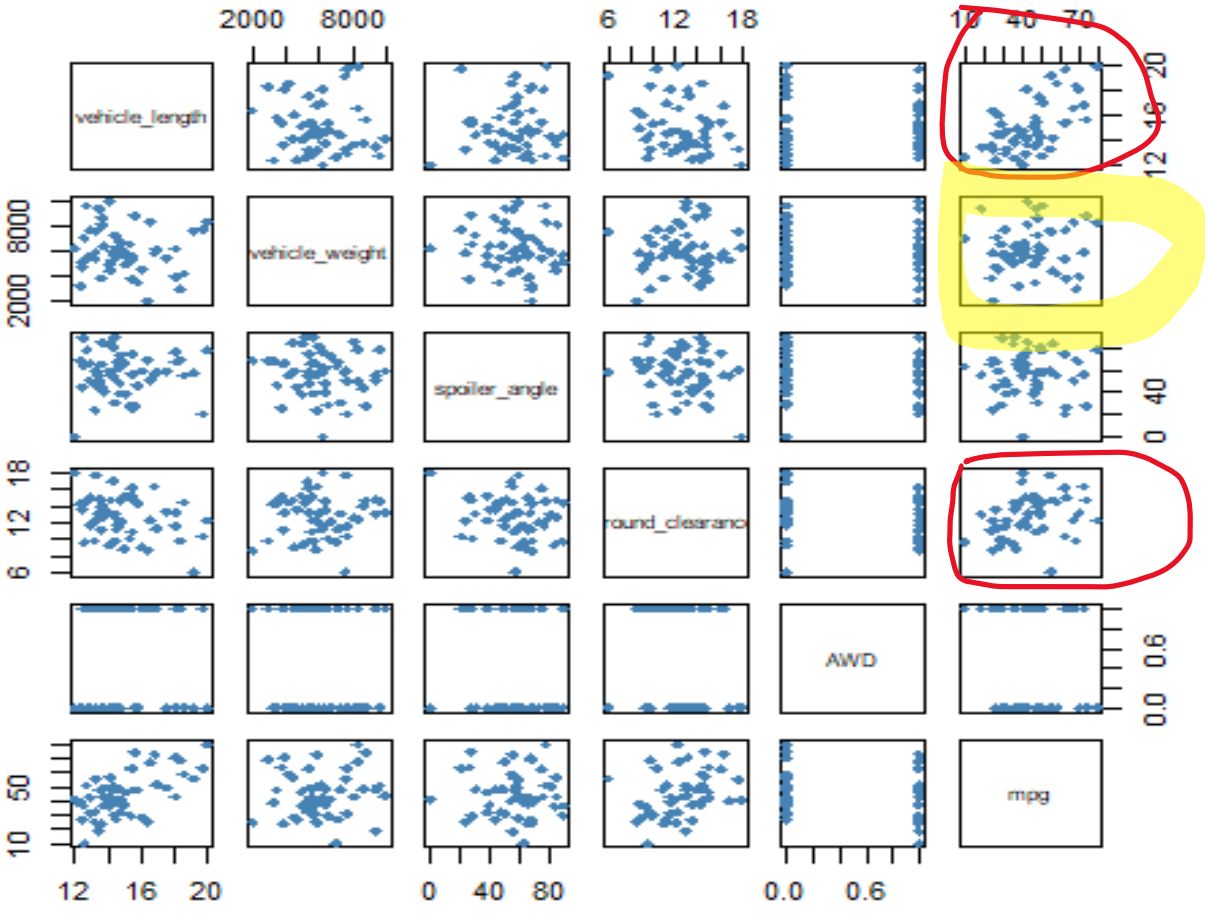


Ground Clearance
data: Mecha_Car$ground_clearance
W = 0.98678, p-value = 0.8446
p-value is >.05; fail to reject Ho: sample is from a normal distribution

Tests of Linear Relationship

After establishing data normality, the analysis turns to tests of linear relationship to establish if there is a possible relationship between the five independent and the one dependent variables. These tests are regression and its summary, and include a residuals analysis to confirm the integrity of the regression model. Tests of linear relationship look for correlations between or among variables to develop a predictive equation. The equation predicts the value of the dependent variable in relation to changes in the values of the independent attributes. Regression's null hypothesis assumes no relationship between the variables; specifically, the correlation or regression coefficient is zero, and that as one variable increases, there is no corresponding increase or decrease for the other variable. The alternate hypothesis is that there is a demonstrable relationship.
Ho: no relationship
Ha: relationship

To begin, displaying a quick scatter plot of pairs helps visualize the total landscape of the data. It shows modest relationship in the clustering of vehicle length and ground clearance relative to mpg:



Additionally, there appears to be a less pronounced relationship between vehicle weight and mpg. In the regression analysis which follows, the analysis will examine these interactions in finer detail.

Regression
The regression model begins to quantify these relationships. The model regresses the dependent variable, termed Y, on the chosen independent variables:
regressing
y (dependent): mpg
on:
x (independent): vehicle_length
vehicle_weight
spoiler_angle
ground_clearance
AWD

The model in RScript is:
lm(mpg ~vehicle_length + vehicle_weight +spoiler_angle+ground_clearance+AWD,data=Mecha_Car)

Results
Coefficients:

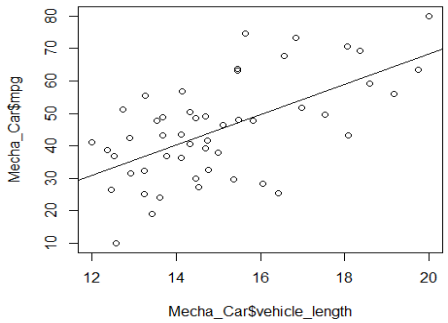
(Intercept)	vehicle_length	vehicle_weight	spoiler_angle	ground_clearance	AWD
-104	6.276	.00125	.066877	3.546	3.411

The coefficients are the intercept and the slope for each variable. These become the equation for the value predictions for mpg given the value of the independent variable.

The formula for the plot of the regression line is y = a+bx where a is the intercept, b is the slope, and x is the value of the independent variable for that observation. Thus the formula for vehicle length regressed against mpg is:

length: Y Coefficients:
(Intercept) vehicle_length
-25.062 4.673
Y = -25.06+4.67*vehicle_length

MPG	Length
33.51036	12.53421
41.85461	14.31984
47.15445	15.45398



As vehicle length increases, so does mpg.

The formula for the model is:
mpg^= -104 + (6.27*length)+(0.00125*weight) +(.069*spoiler)+(3.55*clearance)+(-3.41*AWD)

Thus, the slope of the linear model is not zero.

Summary
The R summary function provides more statistical detail to the regression function. Specifically, it provides the coefficients and residuals as well as the slope for each variable and the Y intercept. It also displays the p, R squared, and the F statistic, and identifies those variables having significance at different levels of alpha.

The summary function is called through the following syntax:

```
summary(lm(mpg ~vehicle_length +vehicle_weight+spoiler_angle+ground_clearance+AWD,data=Mecha_Car))
```

Results

The results tabulate the residuals
Residuals:

Min	1Q	Median	3Q	Max
-19.4701	-4.4994	-0.0692	5.4433	18.5849

and the coefficients.

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-104	15.85	-6.559	0.0000000508 ***
vehicle_length	6.267	0.6553	9.563	0.00000000000260 ***
vehicle_weight	0.001245	0.000689	1.807	0.0776000000 .
spoiler_angle	0.06877	0.06653	1.034	0.3069
ground_clearance	3.546	0.5412	6.551	0.0000000521 ***
AWD	-3.411	2.535	-1.346	0.1852

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.774 on 44 degrees of freedom
Multiple R-squared: 0.7149, Adjusted R-squared: 0.6825
F-statistic: 22.07 on 5 and 44 DF, p-value: 5.35e-11 (.0000000000535)

Findings
Overall, based on the F-statistic of 22.07 and a p value well below zero, the model is useful in establishing a statistical relationship between the dependent variable and the independent variables.

Vehicle_length and vehicle_weight are statistically significant at .001 level, and vehicle_weight at .1.

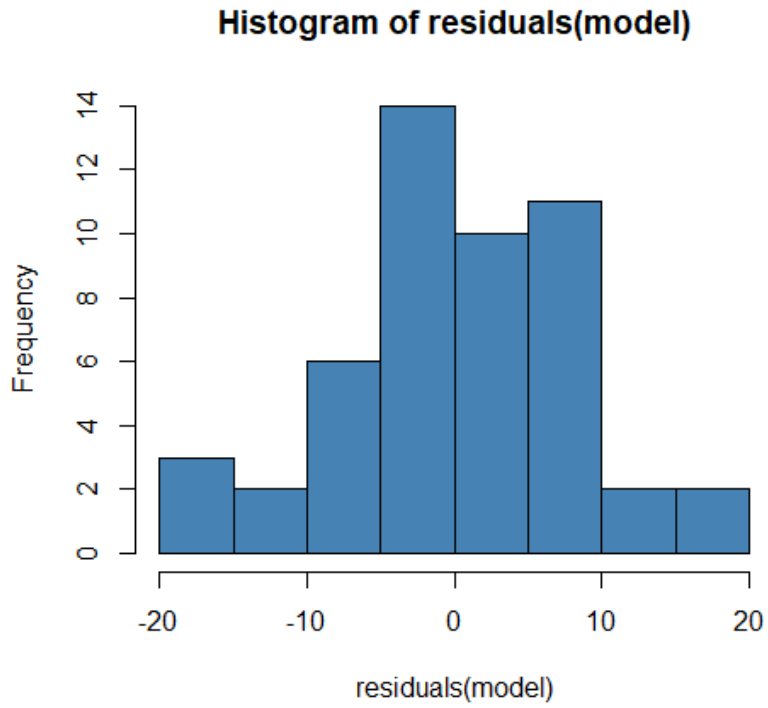
Residuals

Residuals refers to the difference between the actual response and the predicted response of the model. For every observation, there is an actual value and a predicted value. The variance of the residuals should be consistent for all observations.

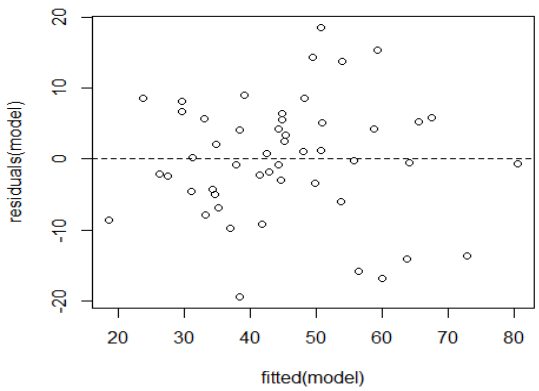
To verify that the residuals are normally distributed, the analysis shows the plotted values in a histogram:

hist(residuals(model), col = "steelblue") #LM_plotted_residual_Mecha_Car.png

Based on the resulting graph, the data appear as a normal distribution:



And a fitted value vs. residual plot shows a fairly symmetrical distribution:



R-squared

R-squared shows how closely the data fits the model. R-squared value always lies within a range of -1 to 1. The closer the R-squared value is to 1, the weaker the relationship. Conversely, values closer to 1 or -1 display strong correlations. In sum, -1 is perfect negative correlation; 1 is perfect positive correlation.

Based on adjusted R2 of .68, the model demonstrates a statistically significant relationship between the dependent variable, mpg, and the independent variables.

Coefficient – Standard Error: The standard error is the estimation of error between the response variable's actual and predicted value. The smaller the standard error, the more precise the model.

Coefficient – t value: This value provides the measure to reject/fail to reject the null hypothesis. The greater the magnitude in absolute value, the greater the reliability of rejecting the null hypothesis and accepting the alternate hypothesis of a relationship between the dependent and independent variables. By this metric, there exists a relationship between vehicle length and mpg, and ground clearance and mpg.

Coefficient – Pr(>t): this is the p-value. The closer it is to zero, the greater the guidance to reject the null hypothesis. Both ground clearance and vehicle length are all but zero, indicating rejection of the Ho; there is a relationship between those two attributes and mpg.

Conclusion to Part I

The MechaCar data displays properties of normal distribution. Regressing the target variable on the five attributes gave statistically significant results for three of the five, notably vehicle length and ground clearance, and to a slighter degree, vehicle weight. The model yielded a useful and accurate prediction equation which can help guide further design efforts for the manufacturer. It would be interesting to look at the relationship between vehicle length and ground clearance, if indeed one exists, or if there are further confounding factors to consider as they relate to fuel efficiency.

Part II – Suspension Coils Weight Capacity

Purpose

The objective is to determine if the manufacturing process is consistent across different manufacturing lots for one automobile manufacturer, AutosRUs, which is testing its newest prototype, the MechaCar. This part of the analysis compares the relative strength of three different manufacturing lots of suspension coils by measuring weight capacity. The analysis gathers key metrics across the groups for comparison: mean, median, variance, and standard deviation and develops a design specification requirement.

Data set
The dataset consists of three variables, vehicle ID (VehicleID), manufacturing lot (Manufacturing_Lot), and PSI, and 150 records. Within the manufacturing lot column are three (3) manufacturing lots: Lot1, Lot2, and Lot3. This analysis briefly compares the three lots based on measures of central tendency. In the Part III following, the analysis conducts t-tests to compare differences in means across each of the lots.

Descriptive Statistics: PSI

Summary by group using tapply
Call:
tapply(Coils$PSI, Coils$Manufacturing_Lot, summary)

Lot	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std.dev	Var
1	1498	1499	1500	1500	1501	1502	0.99	0.98
2	1494	1499	1500	1500	1502	1506	2.73	7.47
3	1452	1494	1498	1496	1501	1542	13	170
All Observations	1452	1498	1500	1499	1501	1542	7.89	62.3

Call:
```
Coils %>%  group_by(Manufacturing_Lot) %>%
  summarise(mean = mean(PSI), median=median(PSI),var=var(PSI),sd = sd(PSI))
```

Manufacturing_Lot	mean	median	var	sd
Lot1	1500	1500	0.98	0.99
Lot2	1500	1500	7.47	2.73
Lot3	1496	1498	170	13
All Observations	1499	1500	62.3	7.89

Design Specification

- Suspension coils should meet an optimal strength in weight capacity between a minimum of 1498 psi and a maximum of 1542
- Vehicle level targets should be identified and quantified as part of the design process. These would include roll and steer
- Analyze the impact of the design on vehicle dynamics and performance

Part III – Suspension Coils t-test

The requirement for this part is to determine if all manufacturing lots and each lot individually are statistically different from the population mean by conducting t-tests. The tests are executed on the entire population sample considered as a whole, and on each individual manufacturing lot. Using the specified population mean of 1,500 pounds per square inch, the results are given below.

One Sample t-test

All Data (no grouping)
call
t.test(Coils$PSI,mu=1500) #compare sample versus population means

Results
data: Coils$PSI
t = -1.8931, df = 149, p-value = 0.06028
Ho: true mean is equal to 1500.
alternative hypothesis: true mean is not equal to 1500

95 percent confidence interval (.05):
1497.507 1500.053
sample estimates:
mean of x
1498.78

p-value of .06 >.05. In this instance , fail to reject Ho: the true mean of the sample is 1500; the means of the population and the sample are statistically similar

One Sample t-test: Each lot against population mean (1500)

Lot1
data: subset(Coils$PSI, Coils$Manufacturing_Lot == "Lot1")
t = 0, df = 49, p-value = 1
alternative hypothesis: true mean is not equal to 1500
95 percent confidence interval:
1499.719 1500.281
sample estimates:
mean of x
1500

p-value of 1.0 >.05. Fail to reject Ho: the true mean of the sample is 1500; the means of the population and the sample are statistically similar

Lot2
data: subset(Coils$PSI, Coils$Manufacturing_Lot == "Lot2")
t = 0.51745, df = 49, p-value = 0.6072
alternative hypothesis: true mean is not equal to 1500
95 percent confidence interval:

1499.423 1500.977
sample estimates:
mean of x
1500.2

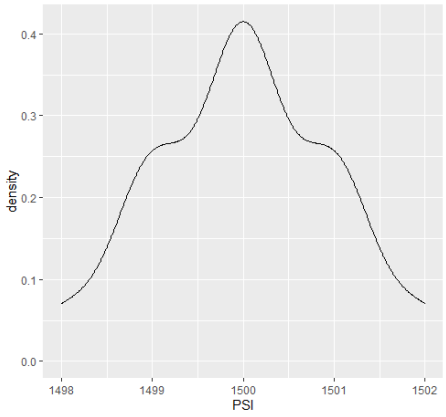
p-value of .06 >.05. Fail to reject Ho: the true mean of the sample is 1500; the means of the population and the sample are statistically similar

Lot3
data: subset(Coils$PSI, Coils$Manufacturing_Lot == "Lot3")
t = -2.0916, df = 49, p-value = 0.04168
alternative hypothesis: true mean is not equal to 1500
95 percent confidence interval:
1492.431 1499.849
sample estimates:
mean of x
1496.14

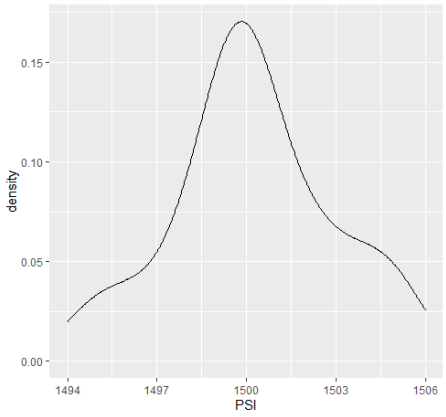
p-value of .04 <.05. Fail to reject Ho: the true mean of the sample is not equal to 1500; the means of the population and the sample statistically differ

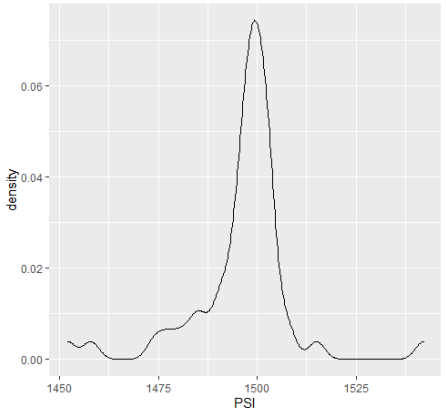
Distributions

Lot 1



Lot 2





Lot 3

Part IV

Statistical Study Design

AutosRUs is developing a new model, the MechaCar. Having encountered production issues which it is addressing, further refinement of the car's features and developing a robust marketing plan would be a critical step in moving forward. To assist in those activities, I recommend creating and implementing strategic studies measuring the target market's opinion of the MechaCar and its competitors. In a broad range, steps include:

- Identifying MechaCar's target market, including:
    - Demographics
    - Lifestage
    - Lifestyle
    - Geography: urban, suburban, rural
- Identifying particular preferences in automobile features
    - Safety
    - Power
    - Convenience
    - Family-friendly
    - Design
    - Ergonomics
    - Aesthetics
    - Fuel Economy
- Identifying competitors' similar vehicles in class
    - Gathering metrics on technical specifications
- Designing studies to elicit consumers' reactions to the MechaCar and to the competitors' models
    - Focus groups, surveys

Choosing safety as the initial focus, and in particular seat-belt harness safety, is the subject of the proposed study. Because automobiles are designed to accommodate a range in body size, seats and seat belts fit different groups differently. Body mass, height, musculature, and posture when sitting vary greatly. Seat belts, if not properly fitted can cause serious injuries, particularly for women and children.

This study would consider the impacts of seat belt design on safety (effective restraint); flexible fit to provide effective restraint while considering size, stature, and weight; ability to adjust easily to fit a range of body types. The study would call for three designs to be tested.

Data should include

- A variable array of body dimensions for 3 major body types (height, weight, build);
- A variable array capturing how each model fits each body type; that is where it is positioned when worn (neck, shoulder, chest, solar plexus, abdomen, hip);
- The strength of the harness
- How the belt responds to forces measured by direction and magnitude
- Adjustability
- How effective specific adjustments are in terms of restraint and conformity with body size and type
- Surveys measuring consumer satisfaction with the models.

Metrics

- Sample size of 50
- For adjustability, use an ordinal scale: 1 – 5; 1 not adjustable, 5 very adjustable. While more difficult to extract quantitative results, the median should provide some insight
- For consumer preference, Likert scale corresponding with the ordinal scale for adjustability

The point analysis is seat belt or seat harness safety and adjustability, using t-tests and ANOVA across all groups. The null hypothesis is that there are differences between groups.