**Gillian Patrick, Anil Poonai, Cher Ee Chong, Michael Moxey, Margaret May Binshtok**

**Dataset:** <https://archive.ics.uci.edu/ml/datasets/Auto+MPG>

**Part 1**

1. **Provide the justification of determining the response (y) and predictor (x) variables.**

Miles per gallon (mpg) was chosen as the response variable and vehicle weight (lbs) was chosen as the predictor variable because we think that they have an inverse relationship with vehicles because the heavier the vehicle is, the less mileage the vehicle is going to cover and vice versa. Also, the weight of the vehicle acts independently in order to make mpg respond.

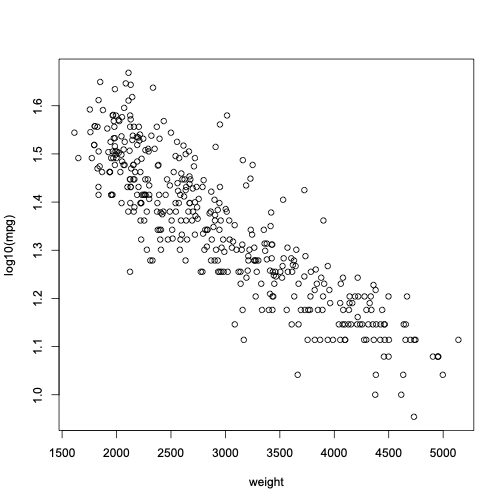
1. **Construct a scatterplot of against. Comment on the association between the two variables.**

From the scatterplot of mpg against weight, there is visual evidence to suggest that there is a curvilinear relationship between mpg and weight. The correlation is negative and strong based on the correlation coefficient of -0.8317409. A close up of a mans face

Description automatically generated

1. **Apply transformations on either y or x , or on both, if any, to make a linear model more appropriate for the transformed data. If so, then work on the transformed data. Re-construct a scatterplot of (transformed) y against (transformed) x.**

We applied a transformation on y (mpg) of log base 10 in order to attempt to linearize the curvilinear relation.



1. **Regress the (transformed) y on the (transformed) x. State the fitted regression line and superimpose it on the scatterplot. Interpret the regression coefficients in the context of the problem.**

A screenshot of a cell phone

Description automatically generated

A close up of a map

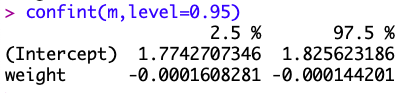
Description automatically generated

After the transformation, the fitted regression line is log10(mpg)=1.800-0.0001525145(weight) . The slope of -0.0001525145 implies that by holding all other factors constant, an increase in vehicle weight by 1 pound results in a decrease in the miles per gallon by 0.0001525145. The intercept of 1.800 implies that when a vehicle weighs 0 lbs, the mpg would be 1.800. This is not realistically possible because a vehicle cannot weigh 0 lbs.

1. **Report the values of se and R2. Interpret them, respectively, in the context of the problem.**

With an R-squared value of 0.7666, approximately 77% of the variation in mpg is accounted for by the regression model by using weight as the predictor. This is an indication that the model is appropriate. The residual standard error of 0.07135 means the average distance between the sample of mpg values and predicted mpg values from the fitted regression line is about 7.1%.

1. **Obtain a 95% confidence interval for and interpret it in the context of the problem.**



We are 95% confident that the interval from -0.0001608281 to -0.000144201 captures the true rate at which mpg is decreasing per pound of vehicle weight.

1. **Test whether B1 is different from zero or not at the 0.05 level of significance. State the hypotheses, the value of t-test statistic, p-value, and your conclusion.**

H0: β1=0

HA: β10

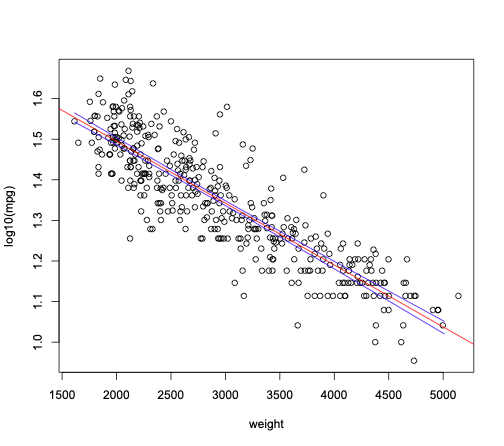
α=0.05

t-test statistic = -36.07

p-value = <2e-16

**Conclusion:** Since t does not equal 0 and the p-value is less than .0001, we reject the null hypothesis. There is significant and sufficient statistical evidence to support the claim that there is a linear association between miles per gallon and vehicle weight.

1. **Superimpose the (pointwise) 90% confidence band on the scatterplot.**



1. **Construct a residual plot against the predicted values and a normal probability plot of residuals and perform model diagnostics using the two plots.**

Based on the residual plot, the model passes the equal variance assumption because the spread of the residuals around the line is about the same for all the predicted values and the scatter appears to be random with no visible patterns. Based on the normal probability plot, the model passes

normality assumption because the plot is reasonably straight and not curved.A close up of a mans face

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**Part 2**

**Model #1: a regression model including three numerical predictor variables that you think are closely associated with the response variable.**

**(i)                  State the fitted regression equation.**

lm(formula = (mpg) ~ weight + acceleration + displacement, data = auto\_mpg)

 Residuals:

 Min   1Q   Median     3Q    Max

-11.7382    -2.8112      -0.3607     2.5231  16.1845

 Coefficients:

            Estimate  Std. Error t value Pr(>|t|)

(Intercept)  41.2990756   1.8614975  22.186  < 2e-16 \*\*\*

weight   -0.0061889   0.0007396    -8.368 1.03e-15 \*\*\*

acceleration  0.1738507   0.0975107    1.783  0.0754 .

displacement -0.0108953   0.0065036    -1.675    0.0947 .

---

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

Residual standard error: 4.294 on 394 degrees of freedom

Multiple R-squared:  0.7004, Adjusted R-squared:  0.6981

F-statistic:   307 on 3 and 394 DF,  p-value: < 2.2e-16

The regression equation is (mpg)=  41.29907559 - 0.00618888(weight) + 0.17385071(acceleration) - 0.01089526(displacement)

**(ii)                 Interpret the regression coefficients.**

(Intercept)   weight acceleration displacement

 41.29907559  -0.00618888   0.17385071  -0.01089526

 The intercept means that when weight, acceleration, and displacement are all 0, the mpg is 41.29907559. The weight coefficient means that with the allowance of the effects of displacement and acceleration an increase in vehicle weight by 1 lb will result in a decrease in miles per gallon by 0.00618888. The acceleration coefficient means that with the allowance of the effects of weight and displacement an increase in the time to accelerate from 0 to 60mph by 1 second will result in an increase in miles per gallon by 0.17385071. The displacement coefficient means that with the allowance of the effects of weight and acceleration an increase in engine displacement by 1 cubic inch will result in a decrease in miles per gallon by 0.01089526.

**(iii)  Produce an ANOVA table. Report SSR, SST, and SSE, and their corresponding degrees of freedom.**

Response: (mpg)

        Df  Sum Sq Mean Sq        F value Pr(>F)

weight         1      16777.8    16777.8       909.7485  <2.2e-16 \*\*\*

acceleration   1       156.8   156.8          8.5039    0.003747 \*\*

displacement   1   51.8       51.8           2.8065  0.094676 .

Residuals     394     7266.2   18.4

---

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

SST = 24252.5755 df = 397, SSE = 16986.3495 df= 3, SSR = 7266.2260 df= 394

**(iv)   Perform an F test of overall linear relationship. State the hypotheses, the value of F test statistic, p-value, and your conclusion.**

**H0:**  βweight=βacceleration=βdisplacement=0

**HA:** At least one βj0

**F test statistic** = 307

**p-value** = <2.2e-16

**Conclusion:**  Since the p-value is less than α=0.05, we reject the null hypothesis. There is significant and sufficient evidence that at least one βj does not equal 0.

**(v)  Test whether x1 is helpful, given that x2 and x3  are in the model. State the hypotheses, the value of the T-test statistic, p-value, and your conclusion.**

**X1 = weight**

**X2 = acceleration**

**X3 = displacement**

**H0:** βweight = 0

**HA:** βweight 0

**t-test Statistic: -**8.368

**p-value** = 1.03e-15

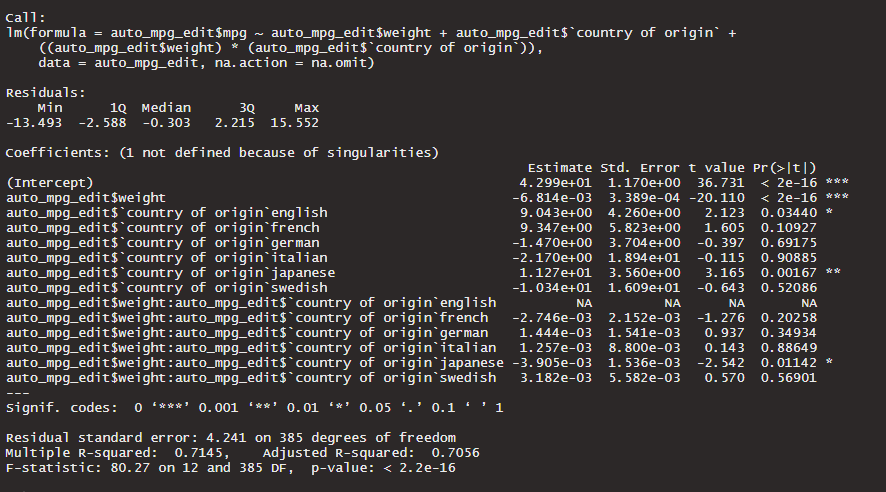
**Conclusion:**  Since the p-value is less than α=0.05, we reject the null hypothesis. There is significant and sufficient evidence that βweight does not equal 0 and that weight is helpful given that acceleration and displacement are in the model.

**Model #2**

**vi) Incorporate the categorical variable into the model by defining indicator variable(s).**

We defined the indicator variables as the country or origins such as England, France, Germany, Italy, Japan and Sweden. As well as each of their own interaction with the weight variable.

**vii)State the fitted regression equation**

****

The regression equation is :

Mpg = 4.299e+01 + -6.814e-03(weight) + 9.043e+00(english) + 9.347e+00(french) + -1.470e+00(german) + -2.170e+00(italian) + 1.127e+01(japanese) + -1.034e+01(swedish) + -2.746e-03(weight\*french) + 1.444e-03(weight\*german) + 1.257e-03(weight\*italian) + -3.905e-03(weight\*japanese) + 3.182e-03(weight\*swedish)

**viii) Interpret the coefficient of the categorical variable and that of the interaction term, respectively**

The coefficient of the categorical variables is the amount of increase or decrease in the miles per gallon of the car due to the car’s origin(country) going up or down by one unit. The interaction term is the result of one of the predictor variables acting differently depending on the levels of the other predictor variable.

**ix)Should you drop the interaction term? Explain**

We shouldn't drop the interaction term because the residual standard error has gone down and the R-squared value has gone up. The p values aren’t significant but we believe the coefficient of determination is more valuable as it actually explains the variance in the graph even if it is only slightly more.

**Model #3**

**x) Use AIC as the criterion to obtain the “best” model using the “backward elimination” procedure. Report the subset of predictor variables to be included in the model and the corresponding AIC value.**

We created a linear regression model with all of the predictor variables then relied on a step function to do a backward selection elimination as that would check all of the possibilities of removing certain predictor variables and get the lowest AIC value. We ended up getting an AIC value of 656.32 and the subset of predictor variables are now cylinders, horsepower, weight, acceleration, model year, and car name.