

ESM 206B  
Hampton  
April 7, 2009

Homework 1 – *SOLUTIONS*

For this homework we'll be using the cars\_1993.xls file from the ESM 206A website. These are real, cross-section data from evaluations of several U.S. and foreign vehicles. Your goal is to determine which variable(s) among the several create the best parsimonious model for predicting fuel efficiency on highways ( 'HighwayMPG').

The following table defines the variables.

<u>Variable</u>	<u>Definition</u>
Manufacturer	Manufacturer name
Model	Car model
Type	Car type
Price	Price in \$000s
CityMPG	MPG in the city
HighwayMPG	MPG on highways
EngineSize	In liters
Horsepower	Horsepower
Fuel Tank	Capacity of tank in gallons
Passengers	Number of passengers car fits
Weight	In pounds
Origin	Whether or not the car is an American brand

1. Run a multivariate regression (Analyze -> Fit Model ->, with HighwayMPG as the response). Include all continuous predictors except CityMPG -- these are marked with blue triangles in JMP).

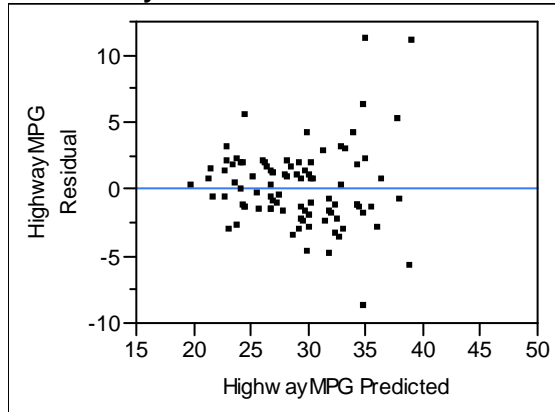
- a. What is  $R^2$  and the p-value for the entire model?

$$R^2 = 0.69$$
$$Adj. R^2 = 0.66$$
$$P < 0.001$$

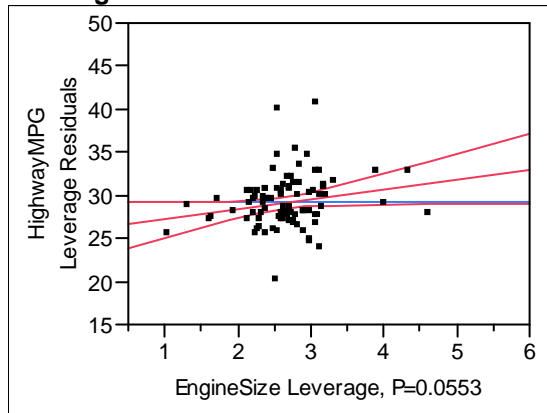
- b. Do the residual and/or leverage plots hint at any problems in the model? If so, briefly explain.

*The residual plot has somewhat of wedge pattern indicating the errors increase with increases in the fuel efficiency of cars. This looks like heteroscedasticity. The 'bunching' patterns in several of the leverage plots hint at potential collinearity (see examples below).*

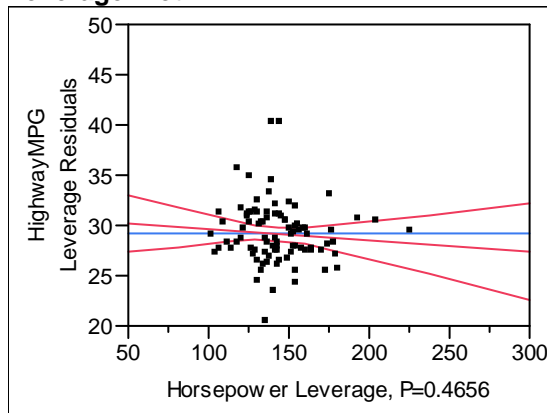
**Residual by Predicted Plot**



**Leverage Plot**



**Leverage Plot**



2. From this model, which coefficient estimates are significantly different from zero at  $\alpha=0.10$ ?

*The significant coefficients are on Weight, Fuel Tank, and EngineSize. Weight and Fuel Tank are negatively related to fuel efficiency as we would*

*predict. EngineSize, however, is positively related to fuel efficiency which is strange.*

<b>Parameter Estimates</b>				
<b>Term</b>	<b>Estimate</b>	<b>Std Error</b>	<b>t Ratio</b>	<b>Prob&gt; t </b>
Intercept	54.397178	2.166259	25.11	<.0001
Price	-0.034119	0.067726	-0.50	0.6157
EngineSize	1.180518	0.607437	1.94	0.0553
Horsepower	-0.011024	0.015039	-0.73	0.4656
FuelTank	-0.453237	0.223157	-2.03	0.0454
Passengers	-0.469113	0.533704	-0.88	0.3819
Weight	-0.005297	0.001829	-2.90	0.0048

3. Compare the partial regression coefficients from this multiple regression to simple linear regressions for the three strongest predictors in this model. (Go to Analyze -> fit Y by X ).

a. How do the regression coefficients from the univariate models compare with the coefficients from the multiple regression above?

*Although Weight and FuelTank are also negatively related to fuel efficiency in the univariate models, their estimated effects are much smaller when all of the variables are included. The sign of the EngineSize coefficient is negative, as we would predict, when it is the only predictor included.*

<b>Parameter Estimates</b>				
<b>Term</b>	<b>Estimate</b>	<b>Std Error</b>	<b>t Ratio</b>	<b>Prob&gt; t </b>
Intercept	51.575677	1.747663	29.51	<.0001
Weight	-0.007316	0.000559	-13.08	<.0001

<b>Parameter Estimates</b>				
<b>Term</b>	<b>Estimate</b>	<b>Std Error</b>	<b>t Ratio</b>	<b>Prob&gt; t </b>
Intercept	50.347208	1.797478	28.01	<.0001
FuelTank	-1.274697	0.105962	-12.03	<.0001

<b>Parameter Estimates</b>				
<b>Term</b>	<b>Estimate</b>	<b>Std Error</b>	<b>t Ratio</b>	<b>Prob&gt; t </b>
Intercept	37.671605	1.205378	31.25	<.0001
EngineSize	-3.208606	0.421992	-7.60	<.0001

b. Why do you think the sign of the effect of EngineSize switches when FuelTank and Weight are also included in the model?

*This is probably a consequence of high collinearity between EngineSize and other predictor variables.*

4. Do a stepwise regression to construct a more parsimonious model. (Analyze -> Fit Model -> Stepwise, run, adjust 'prob to enter' and 'prob to leave' to 0.15.)

a. What can you infer about the univariate x-y relationships from the ordering of the variables in Step History?

*The relationship is strongest for Weight, then FuelTank, then EngineSize.*

**Step History**

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p
1	Weight	Entered	0.0000	1702.427	0.6552	7.0002	2
2	FuelTank	Entered	0.0252	49.31648	0.6742	3.7713	3
3	EngineSize	Entered	0.0661	32.05435	0.6865	2.3727	4

b. What is the AIC from the stepwise regression?

208.65

**Current Estimates**

SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
814.63701	88	9.2572387	0.6865	0.6758	2.3727309	208.6478

5. Run an OLS regression using the variables selected by the stepwise procedure. (Make Model -> Run Model).

a. How do the parameter estimates for Weight, FuelTank, and EngineSize compare with those from the multiple regression that includes all variables?

*The coefficients are pretty similar to those estimated from the full model. Again, there is the troublesome and non-intuitive finding that EngineSize has a positive effect on fuel efficiency once we control for Weight and FuelTank.*

**Parameter Estimates**

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	54.298023	2.030984	26.73	<.0001
EngineSize	1.0647631	0.572203	1.86	0.0661
FuelTank	-0.504588	0.216016	-2.34	0.0218
Weight	-0.00639	0.001464	-4.36	<.0001

b. Do the residual and/or leverage plots look problematic? Explain.

*As before, the residual plot has somewhat of a heteroscedastic, wedge pattern indicating the errors increase with increases in the fuel efficiency of cars. And the 'bunching' pattern in the leverage plots again hint at*

*potential collinearity. These results highlight the fact that stepwise regression does not solve problems caused by heteroscedasticity and/or multi-collinearity.*

6. Examine collinearity among the three predictor variables in the model chosen by the stepwise algorithm. (Analyze -> Multivariate Methods -> Multivariate ->, include Weight, FuelTank, and EngineSize. )
  - a. Are the correlations between the variables ‘high’?

*The correlations are given below and they are very high. These high correlations were suggested earlier by (i) the bunching pattern in the leverage plots; and (ii) the sensitivity of the regression coefficients of Weight, EngineSize, and FuelTank to the inclusion of the other variables in the model.*

**Correlations**

	<b>Weight</b>	<b>EngineSize</b>	<b>FuelTank</b>
Weight	1.0000	0.8447	0.8937
EngineSize	0.8447	1.0000	0.7586
FuelTank	0.8937	0.7586	1.0000

- b. Given the correlations between variables are you inclined to drop any from your regression model? Which ones?

*The correlations are so strong that the stepwise regression model is not reliable. The best we really can do here is run a simple regression using Weight as the only predictor of HighwayMPG.*

7. Drop the variables that you have chosen to drop in 6b and redo the stepwise procedure. (i.e. Analyze ->Fit Model -> Y is HighwayMPG and include Price, Horsepower, Passengers, and Weight -> Stepwise ->Run Model. As before, set ‘prob to enter’ and ‘prob to leave’ at 0.15).
  - a. Which variable(s) does your selected regression include? Is the AIC higher or lower than the stepwise regression in 4b?

*I only include Weight in this regression. The AIC is higher here (213.4 compared to 208.6) but the regression using three explanatory variables is unreliable because of collinearity.*

**Current Estimates**

<b>SSE</b>	<b>DFE</b>	<b>MSE</b>	<b>RSquare</b>	<b>RSquare Adj</b>	<b>Cp</b>	<b>AIC</b>
896.00784	90	9.9556427	0.6552	0.6513	0.6647644	213.4068

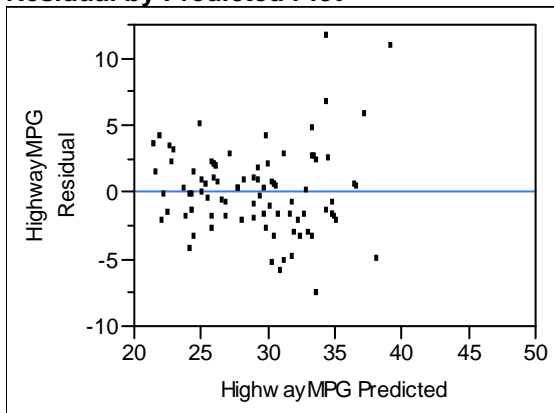
**Step History**

<b>Step</b>	<b>Parameter</b>	<b>Action</b>	<b>"Sig Prob"</b>	<b>Seq SS</b>	<b>RSquare</b>	<b>Cp</b>	<b>p</b>
1	Weight	Entered	0.0000	1702.427	0.6552	0.6648	2

- b. Run the regression that our stepwise procedure chooses in 7a. (Make Model->Run Model). Is there a problem with our residual plot now?

*It doesn't look quite as bad now, but there still appears to be a wedge pattern in the plot.*

**Residual by Predicted Plot**



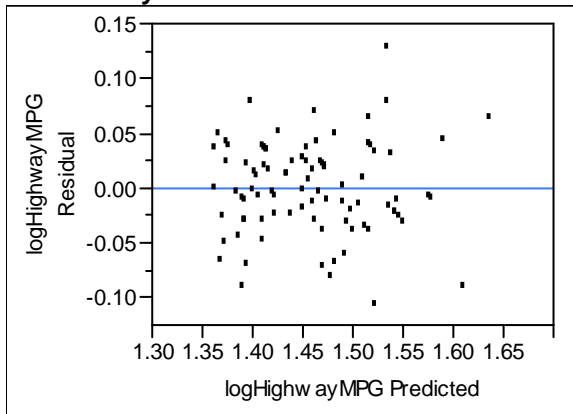
8. Try to deal with the problem identified in 7b by transforming the data. Look at the residual plot of a regression using the transformed data. Does your transformation adequately address the problem?

I transform the data by logging both Weight and HighwayMPG. (Go to the data sheet and right-click to get New Column. Name the new variable 'LogHighwayMPG' -> Column Properties -> Formula->Transcendental ->Log10(HighwayMPG) -> Apply.) Follow the same procedure to log Weight.

Run a regression of  $Y = \text{LogHighwayMPG}$  on  $X = \text{LogWeight}$ . (Analyze -> Fit Model -> Standard Least Squares->Run Model).

The residual plot looks pretty clean now as there is no obvious wedge (or any other) pattern. Now the interpretation of the estimate is as follows: a one percent increase in weight is associated with a -0.71 percent decrease in highway fuel efficiency.

### Residual by Predicted Plot



### Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	3.9434433	0.174183	22.64	<.0001
logWeight	-0.714626	0.05006	-14.28	<.0001