Stat 401 B – Lecture 27

Model Selection

- Response: Highway MPG
- Explanatory: 13 explanatory variables
  - Indicator variables for types of car – Sports Car, SUV, Wagon, Minivan

1

Indicator Variables

- The indicator variable takes on the value 1 if it is that kind of vehicle and 0 otherwise.
- If all four indicator variables are 0, then the vehicle is a sedan.

2

Explanatory Variables

- Indicator variables for All Wheel and Rear Wheel drive.
- If both indicator variables are 0, then the vehicle has Front Wheel drive.

3

Stat 401 B – Lecture 27

Explanatory Variables

- Engine size (liters)
- Cylinders (number)
- Horsepower
- Weight
- Wheel Base
- Length
- Width

4

Forward Selection

- Fit Model – Personality: Stepwise
- Y, Response – Highway MPG
- Put all 13 variables into the Construct Model Effects box.
- Click on Run Model

5

Stepwise Fit

- Prob to Enter = 0.050
- Prob to Leave = 0.050
- Direction = Forward
- Click on Go

6

Stat 401 B – Lecture 27

Stepwise Fit

Response: Highway MPG

Stepwise Regression Control

Prob to Enter 0.050
Prob to Leave 0.050
Direction Forward

Current Estimates

SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
1268.269	96	13.211136	0.6543	0.6435	12.122214	262.0238

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	33.0251054	1	0	0.000	1.0000
☐	☐	Sports Car	0	1	6.928014	0.522	0.4719
☐	☐	SUV	0	1	48.78157	3.800	0.0542
☐	☐	Wagon	0	1	1.952058	0.146	0.7028
☐	☐	Minivan	0	1	9.296884	0.702	0.4044
☐	☐	All Wheel	0	1	45.27064	3.517	0.0638
☐	☐	Rear Wheel	0	1	2.570807	0.193	0.6615
☐	☐	Engine	0	1	38.31665	2.960	0.0886
☐	☐	Cylinders	0	1	31.96371	2.456	0.1204
☐	☑	Horsepower	-0.0257556	1	181.8159	13.762	0.0003
☐	☑	Weight	-0.0062597	1	703.0322	53.215	0.0000
☐	☑	Wheel Base	0.20569376	1	106.6696	8.074	0.0055
☐	☐	Length	0	1	3.045635	0.229	0.6336
☐	☐	Width	0	1	0.141521	0.011	0.9182

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p
1	Weight	Entered	0.0000	2065.965	0.5631	35.606	2
2	Horsepower	Entered	0.0001	228.0966	0.6253	18.88	3
3	Wheel Base	Entered	0.0055	106.6696	0.6543	12.122	4

Forward Selection

- Three variables are added
  - Weight
  - Horsepower
  - Wheel Base
- All variables added are still statistically significant.

Forward Selection

- Model with Weight, Horsepower and Wheel Base.
  - $R^2 = 0.6543$
  - adj $R^2 = 0.6435$
  - RMSE = 3.635
  - $C_p = 12.1222$

Stat 401 B – Lecture 27

Stepwise Fit

- Enter All
- Prob to Enter = 0.050
- Prob to Leave = 0.050
- Direction = Backward
- Click on Go

Stepwise Fit

Response: Highway MPG

Stepwise Regression Control

Prob to Enter 0.050

Prob to Leave 0.050

Direction Backward

Current Estimates

SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
1146.8151	94	12.200161	0.6874	0.6708	6.151101	255.9574

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
✓	✓	Intercept	43.5501018	1	0	0.000	1.0000
		Sports Car	0	1	22.677	1.876	0.1741
	✓	SUV	-4.9350964	1	230.5484	18.897	0.0000
		Wagon	0	1	3.239053	0.263	0.6090
	✓	Minivan	-4.2562699	1	141.4604	11.595	0.0010
	✓	All Wheel	-2.674427	1	95.75119	7.848	0.0062
		Rear Wheel	0	1	10.81537	0.885	0.3492
		Engine	0	1	13.31194	1.092	0.2987
	✓	Cylinders	-1.3933688	1	127.8662	10.481	0.0017
	✓	Horsepower	-0.0297786	1	170.5613	13.980	0.0003
		Weight	0	1	20.39484	1.684	0.1976
		Wheel Base	0	1	4.623746	0.376	0.5410
		Length	0	1	0.159931	0.013	0.9096
		Width	0	1	2.552708	0.207	0.6498

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p
1	Length	Removed	0.6466	2.579335	0.7138	12.212	13
2	Rear Wheel	Removed	0.6784	2.088733	0.7132	10.383	12
3	Wagon	Removed	0.4941	5.636558	0.7117	8.846	11
4	Width	Removed	0.4902	5.704726	0.7101	7.3143	10
5	Engine	Removed	0.4246	7.603433	0.7081	5.9386	9
6	Wheel Base	Removed	0.3576	10.06508	0.7053	4.7649	8
7	Weight	Removed	0.0591	42.93028	0.6936	6.2894	7
8	Sports Car	Removed	0.1741	22.677	0.6874	6.1511	6

Backward Selection

- Eight variables are removed
  - Length, Rear Wheel, Wagon, Width, Engine, Wheel Base, Weight, Sports Car.
- All variables left are statistically significant.

Stat 401 B – Lecture 27

Backward Selection

- Model with SUV, Minivan, All Wheel, Cylinders and Horsepower.
  - R^2 = 0.6874
  - adj R^2 = 0.6708
  - RMSE = 3.493
  - C_p = 6.1511

13

Backward Selection

- The final model from Backward selection is better than the final model from Forward selection. It has a higher R^2 value, higher adj R^2 value, lower RMSE and lower C_p value.

14

Stepwise Fit

- Prob to Enter = 0.050
- Prob to Leave = 0.050
- Direction = Mixed
- Click on Go

15

Stat 401 B – Lecture 27

Stepwise Fit

Response: Highway MPG

Stepwise Regression Control

Prob to Enter 0.050
Prob to Leave 0.050
Direction Mixed

Current Estimates

SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
1268.269	96	13.211136	0.6543	0.6435	12.122214	262.0238

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	33.0251054	1	0	0.000	1.0000
☐	☐	Sports Car	0	1	6.928014	0.522	0.4719
☐	☐	SUV	0	1	48.78157	3.800	0.0542
☐	☐	Wagon	0	1	1.952058	0.146	0.7028
☐	☐	Minivan	0	1	9.296884	0.702	0.4044
☐	☐	All Wheel	0	1	45.27064	3.517	0.0638
☐	☐	Rear Wheel	0	1	2.570807	0.193	0.6615
☐	☐	Engine	0	1	38.31665	2.960	0.0886
☐	☐	Cylinders	0	1	31.96371	2.456	0.1204
☐	☑	Horsepower	-0.0257556	1	181.8159	13.762	0.0003
☐	☑	Weight	-0.0062597	1	703.0322	53.215	0.0000
☐	☑	Wheel Base	0.20569376	1	106.6696	8.074	0.0055
☐	☐	Length	0	1	3.045635	0.229	0.6336
☐	☐	Width	0	1	0.141521	0.011	0.9182

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p
1	Weight	Entered	0.0000	2065.965	0.5631	35.606	2
2	Horsepower	Entered	0.0001	228.0966	0.6253	18.88	3
3	Wheel Base	Entered	0.0055	106.6696	0.6543	12.122	4

Mixed Selection

- Three variables are added
  - Weight
  - Horsepower
  - Wheel Base
- No variables are removed.
- This is the same as with Forward Selection.

All Possible Models

- Lists all 8191 models.
- 13-variable (full) model first.
- 12-variable models – listed in order of the R^2 value.
- etc.
- 1-variable models – listed in order of the R^2 value.

Stat 401 B – Lecture 27

All Possible Models

- Model with all 13 variables has the highest R^2 value.
- R^2 = 0.7145
- Is this the "best" model?
- No, several variables are not statistically significant.

19

All Possible Models

- Model with 7 variables has the lowest RMSE value.
  - Sports Car, SUV, Minivan, All Wheel, Cylinders, Horsepower, Weight
- RMSE = 3.4282

20

Model with lowest RMSE

- Is this the "best" model?
- No, several variables are not statistically significant.
  - Weight: F=3.653, P-value=0.0591
  - Horsepower: F=3.761, P-value=0.0555
  - Sports Car: F=3.847, P-value=0.0529

21

Stat 401 B – Lecture 27

All Possible Models

- Model with 7 variables has the lowest C_p value.
  - Sports Car, SUV, Minivan, All Wheel, Cylinders, Horsepower, Weight
- C_p = 4.76488
- This is the same model as the one with the lowest RMSE.

22

Strategies

- Start with the "best" 1-variable model.
- Find a 2-variable model that beats it.
- Find a 3-variable model that beats the "best" 2-variable model.
- Etc.

23

Strategies

- Start with the full (13-variable) model. Is it "best"?
- Go to the 12-variable models. Are any of these "best"?
- Etc.

24

Stat 401 B – Lecture 27

"Best" Model

- The 7-variable model with
  - SUV, Minivan, All Wheel, Engine, Horsepower, Weight and Wheel Base

  Appears to be the "best" model.

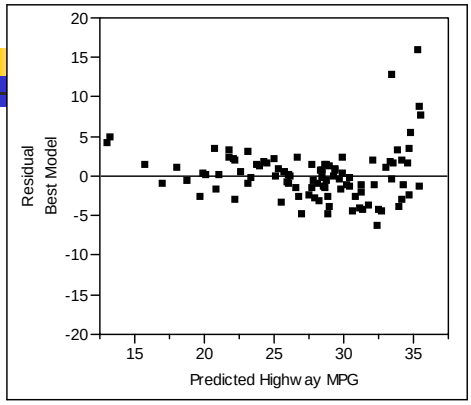
Prediction Equation

Predicted Highway MPG = 30.74 – 3.15*SUV – 3.28*Minivan – 2.08*All Wheel – 1.65*Engine – 0.0226*Horsepower – 0.0029*Weight + 0.163*Wheel Base

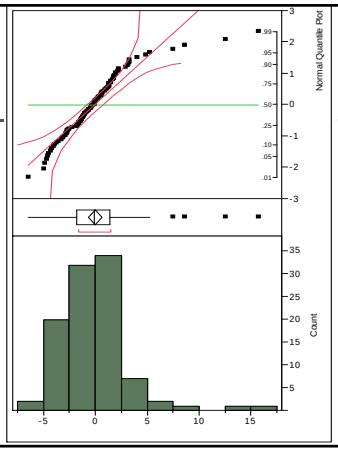
Summary

- All variables add significantly.
- R^2 = 0.705
- adj R^2 = 0.682
- RMSE = 3.431
- C_p = 4.9011

Stat 401 B – Lecture 27



28



29

Box Plot – Potential Outliers

Vehicle Name	Highway MPG	Predicted MPG	Residual
Honda Civic HX 2dr	44	35.4	8.6
Toyota Echo 2dr manual	43	35.5	7.5
Toyota Prius 4dr (gas/electric)	51	35.3	15.7
Volkswagen Jetta GLS TDI 4dr	46	33.4	12.6

30

Stat 401 B – Lecture 27

Outlier

- How do we determine if an potential outlier identified on the box plot is statistically significant?

31

Unusual Points in Regression

- Outlier – a point with an unusually large residual.
- High leverage point – a point with an extreme value for one, or more, of the explanatory variables

32

Influential Points

- Does a point influence where the regression line goes?
- An outlier can.
- A high leverage point can.
- Are they statistically significant?

33