



Stat 401 B – Lecture 31



Model Selection

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


1



Explanatory Variables

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

2



"Best" Model

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

Appears to be the "best" model.

3

Stat 401 B – Lecture 31

Prediction Equation

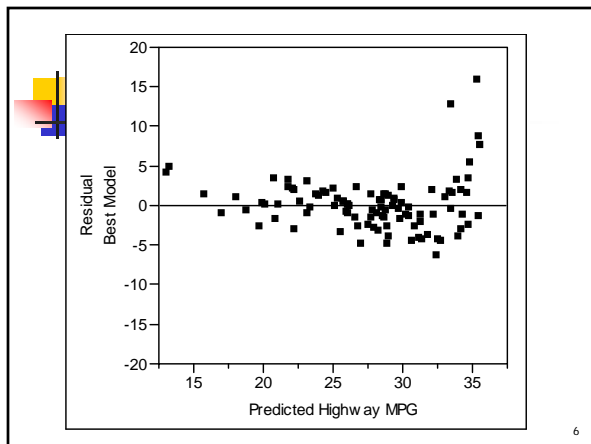
Predicted Highway MPG = $30.74 - 3.15 \cdot \text{SUV} - 3.28 \cdot \text{Minivan} - 2.08 \cdot \text{All Wheel} - 1.65 \cdot \text{Engine} - 0.0226 \cdot \text{Horsepower} - 0.0029 \cdot \text{Weight} + 0.163 \cdot \text{Wheel Base}$

4

Summary

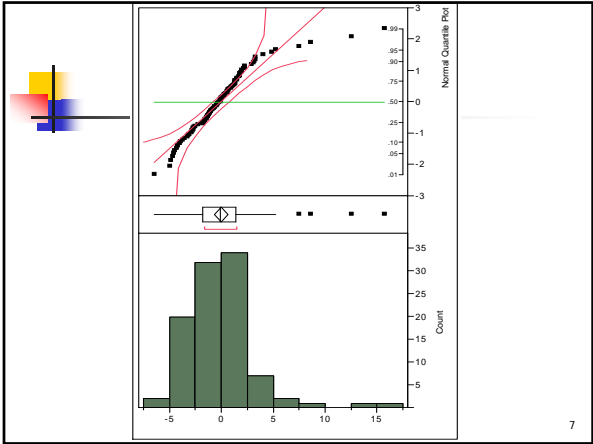
- All variables add significantly.
- $R^2 = 0.705$
- $\text{adj } R^2 = 0.682$
- $\text{RMSE} = 3.430786$
- $C_p = 4.9011$

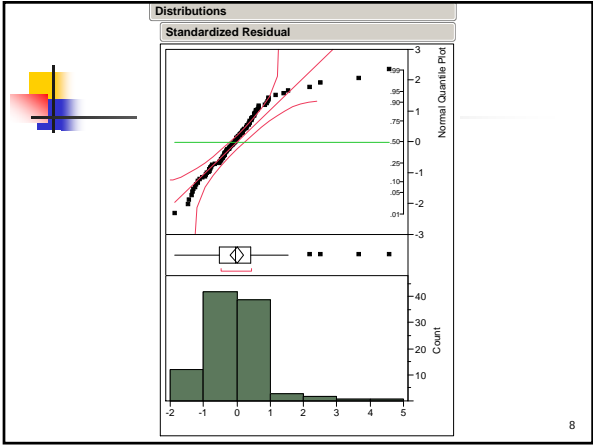
5



6

Stat 401 B – Lecture 31





Box Plot – Potential Outliers

Vehicle Name	Highway MPG	Predicted MPG	Residual	Standardized Residual, z
Honda Civic HX 2dr	44	35.3787	8.6213	2.5129
Toyota Echo 2dr manual	43	35.5299	7.4701	2.1774
Toyota Prius 4dr (gas/electric)	51	35.3093	15.6907	4.5735
Volkswagen Jetta GLS TDI 4dr	46	33.4267	12.5733	3.6648

9

Stat 401 B – Lecture 31

Bonferroni Correction

- Adjust what is a small P-value.

$$\frac{0.05}{\text{\# of residuals}} = \frac{0.05}{100} = 0.0005$$

- If a P-value is less than 0.0005, then the standardized residual is statistically significant.

10

Standardized Residual

Vehicle Name	Standardized Residual, z	Prob > z
Honda Civic HX 2dr	2.5129	0.01197
Toyota Echo 2dr manual	2.1774	0.02945
Toyota Prius 4dr (gas/electric)	4.5735	0.00000
Volkswagen Jetta GLS TDI 4dr	3.6648	0.00025


11

Outliers

- Both the Toyota Prius and the Volkswagen Jetta have standardized residuals so extreme that they are considered statistically significant (P-value < 0.0005).

12


Stat 401 B – Lecture 31



Leverage

- Because we have multiple explanatory variables, there is not an easy formula for leverage, h .
- The leverage, h , value takes into account all of the explanatory variables.


13



Rule of Thumb

- High Leverage Value if
$$h > 2\left(\frac{p+1}{n}\right)$$
- $n = 100, p = 7,$
$$2\left(\frac{p+1}{n}\right) = 2\left(\frac{8}{100}\right) = 0.16$$

14



Leverage

- There are 10 vehicles that have leverage, h , greater than 0.16.
- Of these, 2 have F-statistics large enough to produce P-values smaller than 0.0005.

15

Stat 401 B – Lecture 31

Leverage

	h	F	Prob > F
Chevrolet Corvette convertible 2 dr	0.2532	4.280	0.00039
Porsche 911 GT2 2 dr	0.4084	8.852	0.00000

16

- ## Leverage
- What makes the leverage so high?
 - Have to look for extreme values for the explanatory variables.
- 17

- ## Leverage
- Chevy Corvette
 - Has the 2nd largest engine of all the vehicles – 5.7 liter and the 2nd highest horsepower – 350 horsepower.
 - Porsche 911
 - Has the highest horsepower of all the vehicles – 477 horsepower.
- 18

Stat 401 B – Lecture 31

Influence – Cook's D

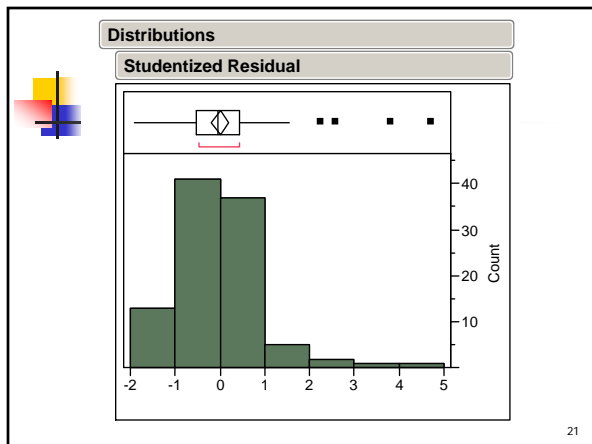
- None of the vehicles has a value of Cook's D that is greater than 1.
- The largest value of Cook's D is 0.16 for the Toyota Prius 4dr (gas/electric). The second largest is 0.12 for the VW Jetta.

19

Influence

- Just because there are no vehicles with Cook's D greater than 1, you should still look at the Studentized residuals.

20



Stat 401 B – Lecture 31

Studentized Residual

Vehicle Name	Studentized Residual, r_s	Prob > $ r_s $
Honda Civic HX 2dr	2.5663	0.01189
Toyota Echo 2dr manual	2.2471	0.02702
Toyota Prius 4dr (gas/electric)	4.5031	0.00001
Volkswagen Jetta GLS TDI 4dr	3.7843	0.00027

22

Outliers

- Both the Toyota Prius and the Volkswagen Jetta have studentized residuals so extreme that they are considered statistically significant (P-value < 0.0005).

23

Summary

- Toyota Prius and Volkswagen Jetta are statistically significant outliers.
- Chevy Corvette and Porsche 911 are statistically significant high leverage values.
- Toyota Prius and Volkswagen Jetta exert statistically significant influence.

24

Stat 401 B – Lecture 31

Response Highway MPG

Summary of Fit

RSquare	0.70486
RSquare Adj	0.682404
Root Mean Square Error	3.430786
Mean of Response	27.7
Observations (or Sum Wgts)	100

Analysis of Variance

Source	DF	Squares	Mean Square	F Ratio
Model	7	2586.1328	369.448	31.3881
Error	92	1082.8672	11.770	Prob > F
C. Total	99	3669.0000		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	30.735611	6.190658	4.96	<.0001*
SUV	-3.147224	1.385628	-2.27	0.0255*
Minivan	-3.283013	1.436711	-2.29	0.0246*
All Wheel	-2.081883	1.01624	-2.05	0.0433*
Engine	-1.654325	0.738966	-2.24	0.0276*
Horsepower	-0.022587	0.008684	-2.60	0.0108*
Weight	-0.002688	0.001221	-2.20	0.0302*
Wheel Base	0.1632806	0.075565	2.16	0.0333*

25

Prediction Equation

- All 100 vehicles

$$\text{Predicted Highway MPG} = 30.74 - 3.15 \cdot \text{SUV} - 3.28 \cdot \text{Minivan} - 2.08 \cdot \text{All Wheel} - 1.65 \cdot \text{Engine} - 0.0226 \cdot \text{Horsepower} - 0.0029 \cdot \text{Weight} + 0.163 \cdot \text{Wheel Base}$$

26

Response Highway MPG

Summary of Fit

RSquare	0.770087
RSquare Adj	0.752205
Root Mean Square Error	2.661825
Mean of Response	27.27551
Observations (or Sum Wgts)	98

Analysis of Variance


Source	DF	Squares	Mean Square	F Ratio
Model	7	2135.8830	305.126	43.0646
Error	90	637.6782	7.085	Prob > F
C. Total	97	2773.5612		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	29.554037	4.84856	6.10	<.0001*
SUV	-2.543539	1.07867	-2.36	0.0205*
Minivan	-2.412591	1.12014	-2.15	0.0339*
All Wheel	-1.640593	0.790945	-2.09	0.0398*
Engine	-1.146582	0.57808	-1.98	0.0504
Horsepower	-0.016019	0.00682	-2.35	0.0210*
Weight	-0.00368	0.000959	-3.84	0.0002*
Wheel Base	0.1740468	0.059234	2.94	0.0042*

27


Stat 401 B – Lecture 31

 **Prediction Equation**

- Excluding Prius and Jetta


Predicted Highway MPG = $29.55 - 2.54 * \text{SUV} - 2.41 * \text{Minivan} - 1.65 * \text{All Wheel} - 1.15 * \text{Engine} - 0.0160 * \text{Horsepower} - 0.0037 * \text{Weight} + 0.174 * \text{Wheel Base}$

28

 **Comment**

- Note that Engine has a P-value of 0.0504 and so is not significant at the 0.05 level.
- This suggests that there is a different “best” model if we exclude Prius and Jetta.


29

 **Comment**

- A similar thing happens if you exclude the Porsche 911 and the Chevy Corvette.

30


Stat 401 B – Lecture 31



Multicollinearity

- High correlation among explanatory variables is called multicollinearity.
- Multicollinearity causes standard errors of estimates to be larger than they should be.

31




Variance Inflation Factor

- A general measure of the effect of multicollinearity is the variance inflation factor, VIF.

$$VIF_i = \frac{1}{1 - R_i^2}$$

32




Multiple R²

- R_i^2 is the value of R^2 among the $k - 1$ explanatory variables excluding explanatory variable i .
- There are k values of R_i^2 .

33


Stat 401 B – Lecture 31



Variance Inflation Factor

- The VIF gives how much the variance of an estimate is inflated by multicollinearity.
- The square root of the VIF gives how much the standard error of an estimate is inflated by multicollinearity.


34



Multiple R²

- SUV excluded: 0.544
- Minivan excluded: 0.304
- All Wheel excluded: 0.356
- Engine excluded: 0.779
- Horsepower excluded: 0.637
- Weight excluded: 0.866
- Wheel Base excluded: 0.673

35




VIF_i

- SUV: 2.19
- Minivan: 1.44
- All Wheel: 1.55
- Engine: 4.52
- Horsepower: 2.75
- Weight: 7.46
- Wheel Base: 3.06

36


Stat 401 B – Lecture 31



Interpretation

- The VIF = 2.19 for SUV.
- This means that the standard error for SUV is 1.48 (the square root of 2.19) times bigger than it would be if SUV were uncorrelated with the other explanatory variables.


37



Interpretation

- The VIF = 7.46 for Weight.
- This means that the standard error for Weight is 2.73 (the square root of 7.46) times bigger than it would be if Weight were uncorrelated with the other explanatory variables.

38



Comment

- If the standard error is 2.73 times what it could be, then the t-statistic is 2.73 times smaller than it could be.
- The corresponding P-value would be less than 0.0001.

39
