M8S2 - Regression In Practice

Professor Jarad Niemi

STAT 226 - Iowa State University

December 4, 2018

Outline

Regression assumptions

Regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

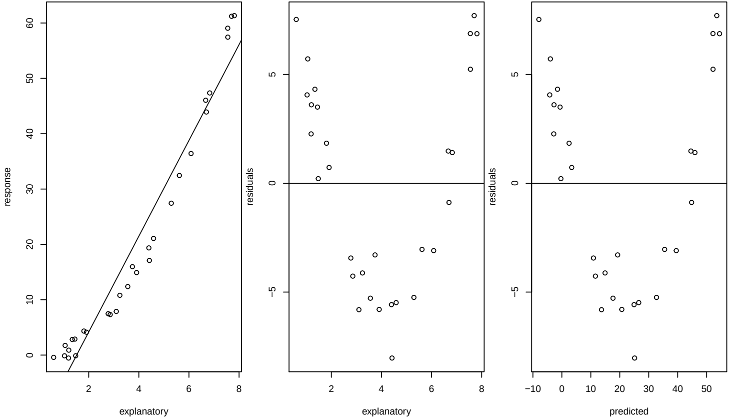
Regression assumptions are

- Errors are independent
- Errors are normally distributed
- Errors are identically distributed with a mean of 0 and constant variance of σ^2
- Linear relationship between explanatory variable and mean of the response

Assessing linearity assumption

Look for non-linearity in

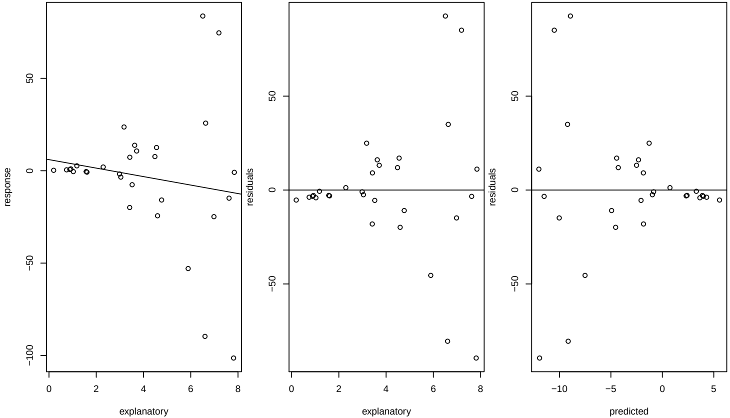
- response vs explanatory plot
- residuals vs explanatory plot
- residuals vs predicted value plot



Assessing constant variance assumption

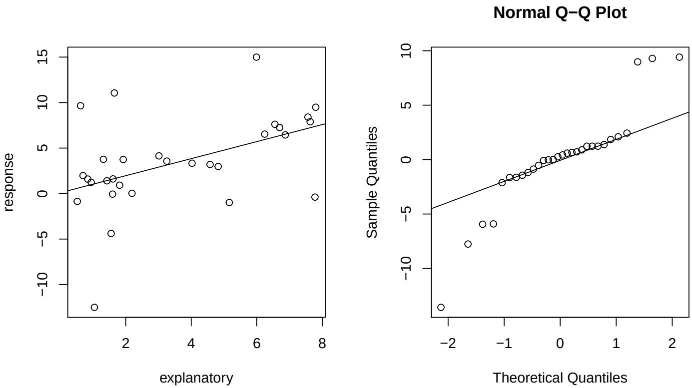
Look for a bugle horn pattern

- residuals vs explanatory plot
- residuals vs predicted value plot



Assessing normality assumption

Deviations from a straight line in a normal quantile plot (qq-plot)



Assessing the independence assumption

The main ways that the independence assumption is violated are

- temporal effects
- spatial effects
- clustering effects

Each of these requires a relatively sophisticated plot or analysis and thus, for this course, we will assess the independence assumption using the context of the problem. If one of the above effects are present in the problem, then there may be a violation of the independence assumption.

Influential individuals

In addition to violation of model assumptions, we should be on the lookout for individuals who are influential.

Recall

- if the explanatory variable value is far from the other explanatory variable values, then the individual has high leverage, and
- if removing an observation changes the intercept or slope a lot, then the individual has high influence.

Regression analysis procedure

1. Determine hypotheses, i.e. why are you collecting data
2. Collect data (at least two variables per individual)
3. Identify explanatory and response variables
4. Plot the data
5. Run regression
6. Assess regression assumptions
7. Interpret regression output

Gas mileage

To understand changes in our 2011 Toyota Sienna, we record the miles driven and amount of fuel consumed since our last fill-up. From this we can calculate the miles per gallon (mpg) since out last fill-up. Understanding changes in mpg through time may give us an indication of problems with our car.

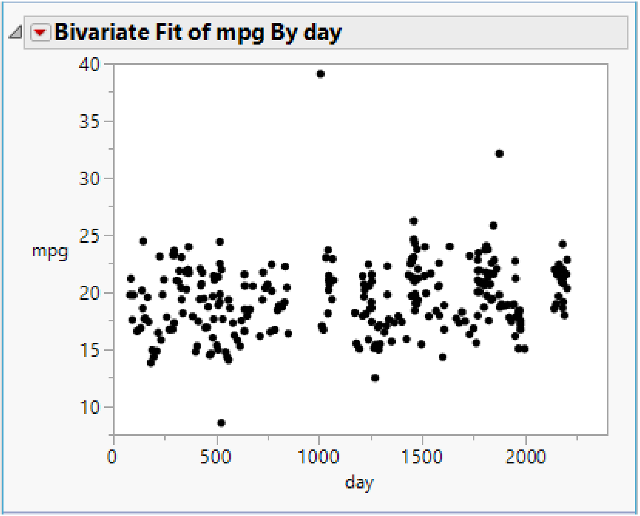
In the following analysis, we use

- miles per gallon as our response variable
- days since purchase as our explanatory variable

Example data sheet

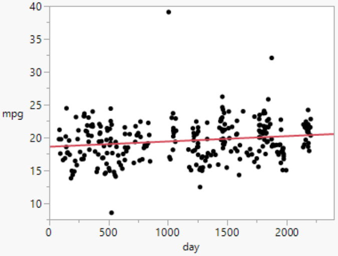
date	cost	fuel	mileage	octane	ethanol	notes
6/8	44.52	15.357	284.2	87	10%	Phillips66
6/10	15.561	$45.11	341.6	87	0%	Loves
6/18	$44.20	16.877	319.2	87	10%	Sams
6/21	$38.47	14.264	307.6	87	10%	Kum&Go
6/26	$34.00	13.234	284.3	87	10%	Sams
6/29	$28.13	10.197	200.1	87	10%	Phillips66
7/1	$34.10	12.451	278.9	87	0%	Pilot
7/2	$35.59	13.185	291.0	87	0%	Holiday
7/5	$35.66	14.865	326.4	87	0%	Costco
7/11	$49.10	17.542	370.9	87	0%	Holiday
7/13	$47.40	17.563	366.1	87	10%	Casey's
7/19	$33.90	12.895	239.5	87	10%	Swift Stop
7/19	$18.12	6.664	146.6	87	0%	Holiday
7/19	$22.10	7.894	190.8	87	0%	Holiday
7/22	$27.86	10.322	197.3	87	10%	Cenex
7/22	$18.24	6.859	145.5	87	0%	Holiday
7/22	18.43	6.778	147.7	87	0%	Holiday
7/23	16.99	7.449	154.3	87	10%	Sams
7/28	24.09	8.762	157.2	87	10%	Phillips66
8/7	33.23	12.043	259.4	87	10%	SuperAmerica
8/10	31.08	11.388	231.0	87	10%	Swift Stop
8/10	17.42	6.455	147.1	87	0	Holiday

Plot


Bivariate Fit of mpg By day
40
35
30
25
20
15
10
mpg
0
500
1000
1500
2000
day

Regression

Bivariate Fit of mpg By day



Linear Fit

mpg = 18.567468 + 0.0008083*day

Summary of Fit

RSquare	0.029453
RSquare Adj	0.025923
Root Mean Square Error	3.022426
Mean of Response	19.51018
Observations (or Sum Wgts)	277

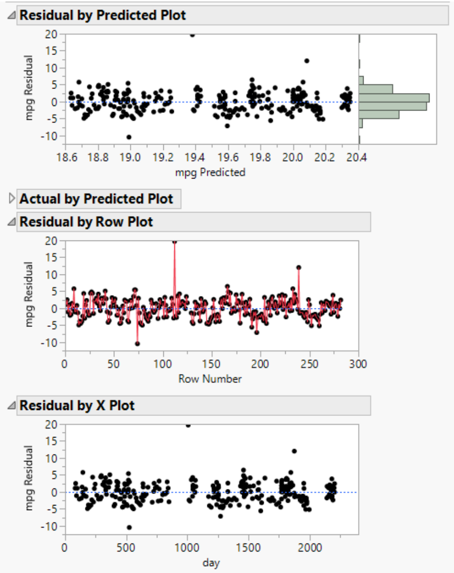
Lack Of Fit

Analysis of Variance

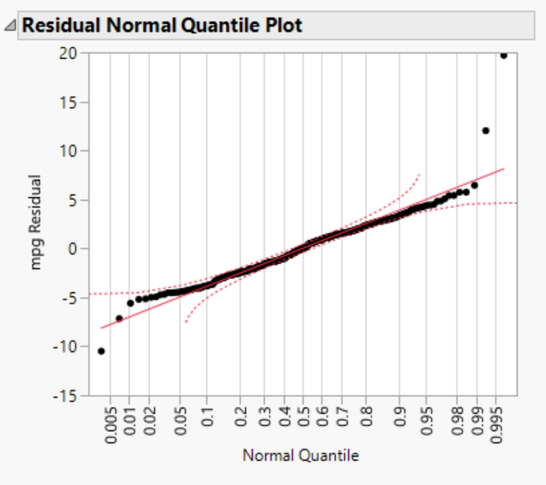
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	18.567468	0.373457	49.72	<.0001*
day	0.0008083	0.00028	2.89	0.0042*

Residuals


Residual by Predicted Plot
20
15
10
5
0
-5
-10
mpg Residual
18.6 18.8 19.0 19.2 19.4 19.6 19.8 20.0 20.2 20.4
mpg Predicted
Actual by Predicted Plot
Residual by Row Plot
20
15
10
5
0
-5
-10
mpg Residual
0 50 100 150 200 250 300
Row Number
Residual by X Plot
20
15
10
5
0
-5
-10
mpg Residual
0 500 1000 1500 2000
day

Normal quantile plot



Regression

Linear Fit

mpg = 18.567468 + 0.0008083*day

Summary of Fit

RSquare	0.029453
RSquare Adj	0.025923
Root Mean Square Error	3.022426
Mean of Response	19.51018
Observations (or Sum Wgts)	277

Lack Of Fit

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	18.567468	0.373457	49.72	<.0001*
day	0.0008083	0.00028	2.89	0.0042*

Interpretation

- When the car was purchased (day 0), the predicted miles per gallons was 18.6 mpg.
- Each additional day that passes, the miles per gallons increases by 0.0008 mpg on average. Over the course of a year, this is an increase of 0.29 mpg on average.
- Only 2.9% of the variability in miles per gallon is explained by day.

Confidence intervals

To construct a $100(1-\alpha)\%$ confidence interval, we use the generic formula

$$\text{estimate } \pm t_{n-2,\alpha/2} \text{ SE(estimate)}$$

Suppose we are interested in 90% confidence intervals for the intercept and slope. We have

$$t_{275,0.05} < t_{100,0.05} = 1.66.$$

Thus, a 90% confidence interval for the intercept is

$$18.567468 \pm 1.66 \times 0.373457 = (17.9, 19.2)$$

and a 90% confidence interval for the slope is

$$0.0008083 \pm 1.66 \times 0.00028 = (0.0003, 0.0013).$$

Hypothesis tests

JMP reports two p-values:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	18.567468	0.373457	49.72	<.0001*
day	0.0008083	0.00028	2.89	0.0042*

These correspond to the hypothesis tests

$$\begin{array}{llcl} \text{Intercept} & H_0 : \beta_0 = 0 & \text{vs} & H_a : \beta_0 \neq 0 \\ \text{day} & H_0 : \beta_1 = 0 & \text{vs} & H_a : \beta_1 \neq 0 \end{array}$$

To obtain the one-sided p-values, you need to divided the p-value in half and, if the alternative is not consistent with the estimate, subtract from 1. So the

$$\begin{array}{lcll} \text{Hypotheses} & & & p\text{-value} \\ H_0 : \beta_0 = 0 & \text{vs} & H_a : \beta_0 > 0 & < 0.0001 \\ H_0 : \beta_1 = 0 & \text{vs} & H_a : \beta_1 < 0 & 0.9979 \end{array}$$