Stat536 HW3 - Cars Data

618518

February 21, 2014

Introduction:

When car dealers buy a used car, they stand the risk of not being able to resell the used car for a profit. To increase the probability of selling their used cars form a profit, car dealers would like to predict the price they can sell the used cars for. Since it is impossible to know in advanced the willingness of buyers to pay for each (unique) car, we will model car selling price with respect to variables (features of cars) provided in the cars data set.

Methods / Models Used:

The response variable in the cars data set is Price (quantitative). Miles and Weight are also quantitative variables. All other variables will be considered as categorical variables. Hence, we have multiple variables, and we need to use multiple regression. However, the a plot of Price against Miles reveals that the two variables are not linearly correlated (See Figure 1). So, we cannot use multiple linear regression. A proper model to use in this case would then be a nonlinear model, specifically, a General Additive Model (GAM):

$$y_i = \beta_0 + \sum_{p=1}^{P} f_p(x_{ip}) + \epsilon_i,$$

where $f_p(x_{ip})$ is some function for the p^{th} variable. In this case, y, our response variable, is Price. Our covariates make up x. And $\epsilon \sim\ N(0, \sigma^2)$.

Model Justification:

The advantage of using a GAM for this problem is that it can model non-linear relationships that a linear regression will miss, and potentially give better predictions while maintaining interpretability. Note that while the GAM is an restricted to be additive, like the linear model, it is much more flexible than the linear model because of its ability to model nonlinear relationships.

A spline would suitably model the nonlinear relationship between Miles and Price. I chose to use a smoothing spline over b-splines or a natural spline as (1) I can get a smooth curve which has good tail behavior, and (2) I don't need to determine the number of knots to create the spline. A natural approach to getting a smoothing spline is to find the function g that minimizes:

$$\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt \ldots (eq.1)$$

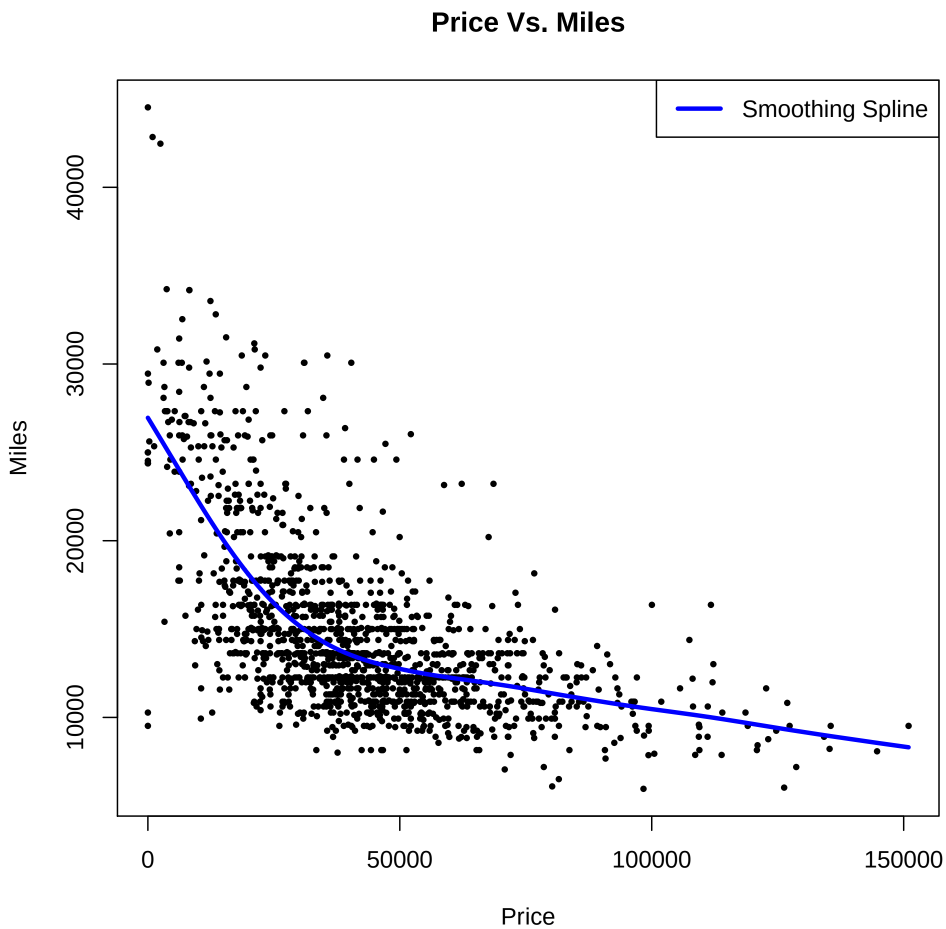

Figure 1: Price vs. Weight Plot

where λ is a nonnegative tuning parameter. The function g that minimizes (eq.1) is known as a smoothing spline. The first term in (eq.1) is a loss function, which encourages g to fit the data well. The second term is a penalty term that penalizes variability in g. $g''(t)^2$ is a convenient measure of the roughness of g. In essence, if g is very smooth, then $g'(t)$ will be close to constant and $\int g''(t)^2 dt$ will be small. Conversely, if g is jumpy, then $g'(t)$ varies greatly, and $\int g''(t)^2 dt$ will be large. So the larger λ is, the smoother g will be.

Results:

A smoothing spline was applied to Miles. No other functions were applied to other variables. After some variable selection, the GAM obtained contained:

- s(Miles)
- Manufacturing Year
- Fuel Type
- Horse Power
- Automatic
- Cylinder Capacity
- Manufacturing Guarantee
- Weight
- Automatic Air-conditioning
- Powered Windows

From Figure 2, we can see that the residuals are approximately evenly spread out and centered about 0. A Q-Q plot (Figure 3) helps us see that the residuals are not strictly normally distributed. There may be some outliers. However, those points were not identified.

To measure the accuracy of predictions under the GAM model, I computed the coverage, which was 95.6%. The average prediction interval width was \$5589.635. Given that the range of price of cars was (5959.5, 44525.0) in this data set, I believe that a prediction interval width of of \$5589 is acceptable.

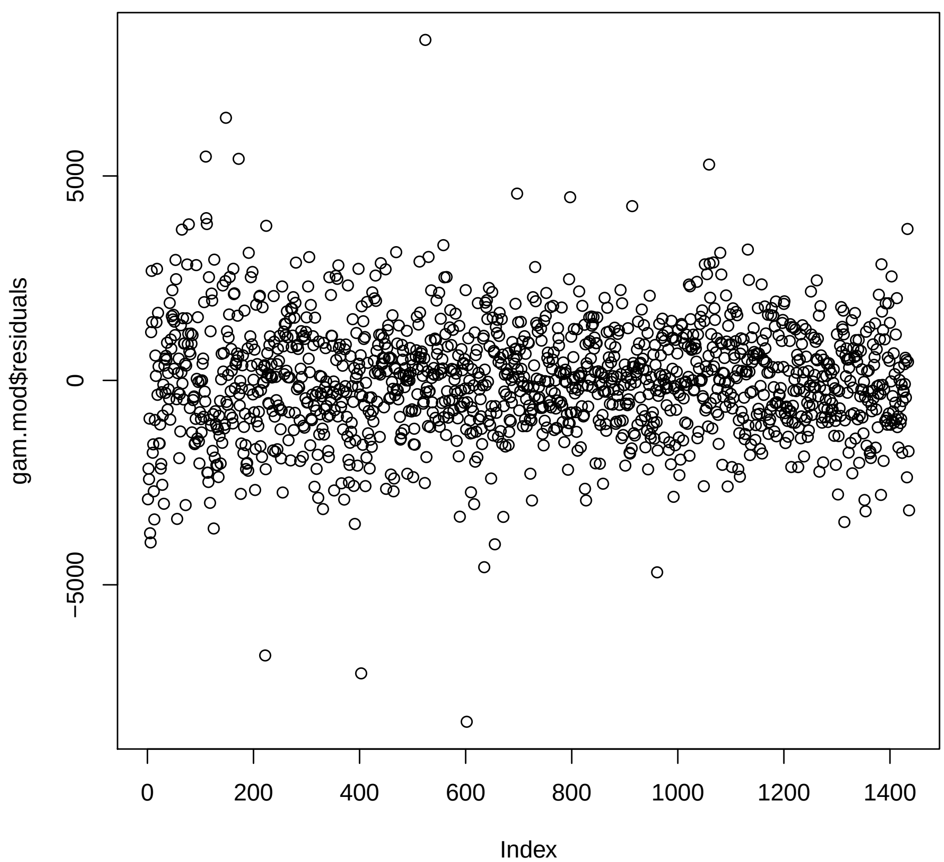

Figure 2: Residuals Plot

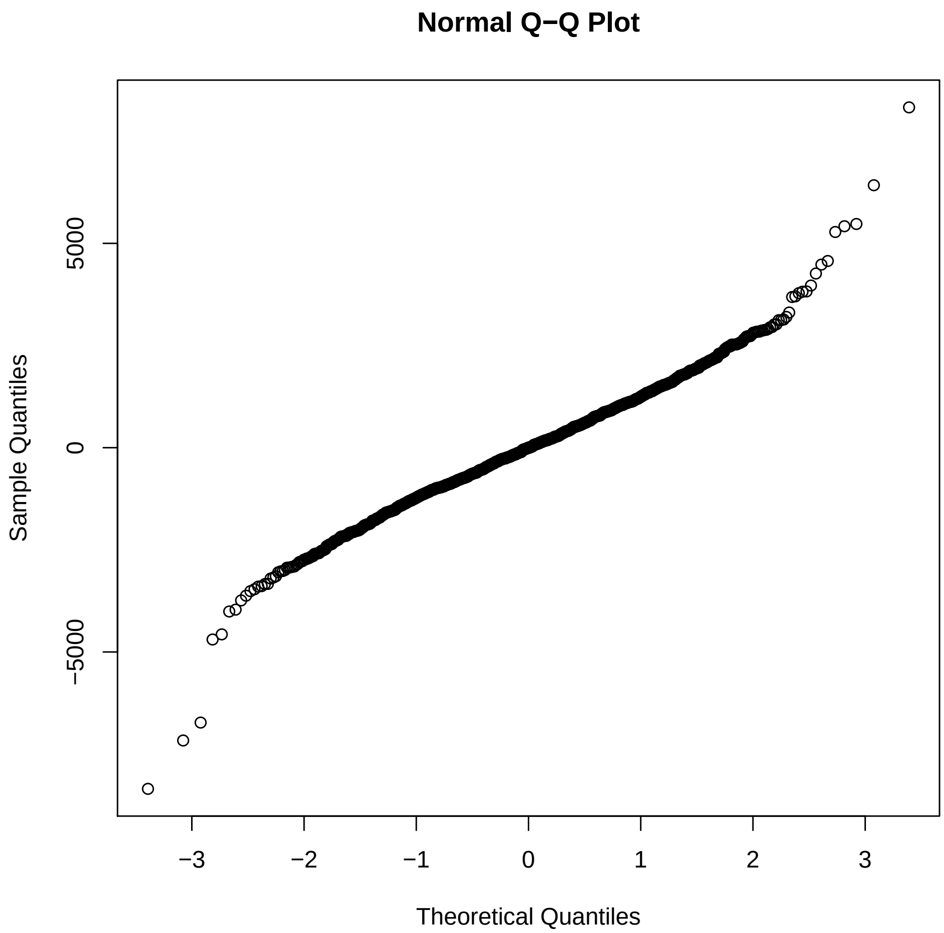

Figure 3: Q-Q norm plot

Parameter Estimations:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3181.97	1592.90	-2.00	0.05
Mfg_Year1999	1493.35	101.51	14.71	0.00
Mfg_Year2000	3486.25	138.08	25.25	0.00
Mfg_Year2001	5109.46	152.38	33.53	0.00
Mfg_Year2002	8747.39	241.65	36.20	0.00
Mfg_Year2003	11243.38	255.52	44.00	0.00
Mfg_Year2004	13994.57	407.58	34.34	0.00
Fuel_TypeDiesel	-3500.44	632.32	-5.54	0.00
Fuel_TypePetrol	651.81	361.30	1.80	0.07
HP71	3449.20	1733.94	1.99	0.05
HP72	3064.45	1426.58	2.15	0.03
HP73	4096.80	1998.38	2.05	0.04
HP86	751.98	823.12	0.91	0.36
HP90	2898.34	1436.61	2.02	0.04
HP97	1957.02	1184.40	1.65	0.10
HP98	1271.42	1486.20	0.86	0.39
HP107	6389.77	1638.67	3.90	0.00
HP110	6523.96	1597.28	4.08	0.00
HP116	7762.08	1519.53	5.11	0.00
HP192	9697.50	871.19	11.13	0.00
Automatic1	556.42	195.26	2.85	0.00
cc1332	-453.35	1001.64	-0.45	0.65
cc1398	-2944.88	1730.91	-1.70	0.09
cc1400	-1997.59	1417.08	-1.41	0.16
cc1587	-6422.05	1271.54	-5.05	0.00
cc1598	-4634.06	1262.04	-3.67	0.00
cc1600	-5752.55	1103.57	-5.21	0.00
cc1800	-3741.64	1030.06	-3.63	0.00
cc1900	3897.71	894.36	4.36	0.00
cc1975	2848.69	1409.56	2.02	0.04
cc1995	1383.86	1158.56	1.19	0.23
cc2000	1808.44	759.80	2.38	0.02
Weight	11.81	1.41	8.41	0.00
Mfr_Guarantee1	472.97	79.36	5.96	0.00
Airco1	488.42	97.05	5.03	0.00
Automatic_airco1	1982.19	216.35	9.16	0.00
Powered_Windows1	441.52	93.20	4.74	0.00

Smoothing Spline Function:

	edf	Ref.df	F	p-value
s(cars$Miles)	2.65	3.39	67.28	0.00

Conclusion:

The GAM models additive effects, so interaction terms can be missed. It is possible, however, to manually add interaction terms, as with linear regression. This may be done in a future investigation. Further investigation could also include identifying outliers.