Spring 2021 **Ali Berra, Alex Georgiopoulus, Caroline Gear**

University of Central Florida
College of Business

QMB 6911
Capstone Project in Business Analytics

Solutions: Problem Set #10

1 Introduction

We will be estimating a sample-selection model for the Used Truck data set. The sample selection would be useful for this data set because the data includes sales for the trucks in auction and in retail. This would be useful to estimate in order to notice if there is any differences between the two types of sales within the data set and find if there are different significant variables. There could be difference which may be helpful to know when attempting to predict the rate of return for Used Trucks in auction or retail sales.

2 Linear Models

The following table shows the linear models created for both auction and retail respectively. There are some differences shown between the linear model between a few of the variables. The age variable had a slightly higher impact on auction sales than retail sales such that a higher age affects auction slightly more(respectively -0.006 vs -0004). The damage variables lowers the ror more for the retail side versus auction side (respectively -0.1 vs -0.07). Also, some dealers and make models had more of an effect whether they were in retail or auction which can indicate some favoritism from people who go to auctions and retail looking for the specific makes or dealers they visit.

It would be nice to show the coefficients for the full sample (with a type indicator) to compare the coefficeints.

	Model 1	Model 2
(Intercept)	1.799592327^{***}	1.945823822^{***}
	(0.046247226)	(0.055572819)
mileage	0.000000447	−0.000000243
	(0.000000453)	(0.000000745)
mil_sq	-0.000783667^{**}	-0.001013175^{*}
	(0.000281593)	(0.000401824)
age	-0.005838932^{**}	-0.004229296^{*}
	(0.001937950)	(0.001996644)
damage	-0.071722765^{***}	-0.108573010^{***}
	(0.001581872)	(0.001773874)
as.factor(make)2	−0.009834005	0.003728774
	(0.011179170)	(0.009952312)
as.factor(make)3	0.004259219	0.020231647
	(0.011831510)	(0.011289301)
as.factor(make)4	−0.006438908	−0.002394631
	(0.009844842)	(0.009101585)
as.factor(make)5	−0.018918318	0.007172364
	(0.010345810)	(0.009357644)
as.factor(make)6	-0.052526564^{***}	-0.053200228^{***}
	(0.010010891)	(0.009279822)
as.factor(make)7	-0.058748864^{***}	-0.059428784^{***}
	(0.010549409)	(0.009902035)
as.factor(make)8	-0.061683320^{***}	-0.050661770^{***}
	(0.015751650)	(0.012986728)
as.factor(make)9	-0.083534749^{***}	-0.068696799^{***}
	(0.017323728)	(0.013830605)
as.factor(dealer)2	-0.049031618^{***}	-0.080361363^{***}
	(0.013912933)	(0.012980400)
as.factor(dealer)3	0.073761718^{***}	0.066774746^{***}
	(0.010083488)	(0.011106343)
as.factor(dealer)4	-0.069080434^{***}	-0.071685887^{***}
	(0.009801584)	(0.008853971)
as.factor(dealer)5	0.054911458^{***}	0.094591423^{***}
	(0.013004879)	(0.012677209)
as.factor(dealer)6	0.013116486	0.030782389^{***}
	(0.009466773)	(0.008654283)
as.factor(dealer)7	0.033016912^{**}	0.069954520^{***}
	(0.011516114)	(0.010708248)
as.factor(dealer)8	-0.063523587^{***}	-0.080397251^{***}
	(0.009426949)	(0.008344516)
as.factor(dealer)9	-0.097105384^{***}	-0.098017766^{***}
	(0.013340742)	(0.012384392)
R^2	0.392421721	0.643649744
Adj. R^2	0.390624950	0.641317609
Num. obs.	6784	3077

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Tab. 1: Linear Models for Auction and Retail

3 Probit Models for Sample Selection

This would be nice to include to show the rationale for your model specification in the next section.

4 Sample Selection Model

The sample selection model that was best for the variables is show below. The model was created by sampling the auction and retail data within the data set. Different models were played with but to show all the variables on one table, only the final one was shown as it created multiple lines due to dealer being a categorical variable and creating multiple lines. The retail sample model excluded the mileage, square root of mileage and make variables while the auction excluded the mileage and make variables.

The sample model ended up with all the variables being signficant including the different types of dealers between auction and retail. Comparitevly only dealer 6 within the auction data set for the linear model was not signficant while the other dealers were. This could mean that dealer is not as important to some people buying at the auction sales. The age variable coefficient for retail in the sample model was -0.02 while in the linear model it was -0.004. This could show that age has more of an impact in the retail sales than was shown in the linear model. On the other hand, the sample model age variable coefficient was -0.0045 compared to a -0.0058. This is a much smaller increase between the models when compared to the difference in retail age differences. It could be considered that people selling in retail should be more wary of the age of cars sold as it can lower the rate of return for the Used Trucks.

The differences in damage between the linear model and sample model were not as huge. The auction coefficients for the sample was -0.078 versus the linear model at -0.071. The retail coefficients for the sample were -0.098 and for the linear it was -0.108. They both rose almost 0.007 and 0.01 respectively. These is a tiny increase but it could be inferred that perhaps damage has more of an impact on both retail and auction sales than was shown in the linear model.

The dealer variable also experiences some slight differences between the sample model and linear models. For the retail side, all the nine of the dealers coefficients slightly increase from the linear model to the sample model which would indicate that the dealers have a bigger impact on the rate of return. The auction sales also had similar difference in that the sample model increased the coefficient slightly when compared to the linear model.

Comparing the linear models from the split between the auction and retail showed some differences in the r-squared. For auction, the r-squared was lowered from 0.41 to 0.39. However, the retail r-squared increased from 0.41 to 0.64. This increased shows that the model had a better fit for the retail sales than the auction sales. This could indicate that some of the variables within the model may be more significant in retail than in auction and could be looked further to see if reducing the model for auction could create a better linear model for the

auction data.

This would be nice to include the series of model specifications to show the process through which you obtained the final model. For example, it would have followed a logical progression if your first model had a selection equation that matched the outcome of your variable reduction for the probit model. All things considered, this is a good final model.

	Model 1
S: (Intercept)	-1.25322^{***}
	(0.04969)
S: mileage	0.00001^{***}
	(0.00000)
S: damage	0.23856^{***}
	(0.01052)
O: (Intercept) (1)	1.76277^{***}
	(0.01168)
O: age (1)	-0.01999^{***}
	(0.00140)
O: damage (1)	-0.09886^{***}
	(0.00267)
O: as.factor(dealer)2 (1)	-0.07599^{***}
	(0.01356)
O: as.factor(dealer)3 (1)	0.06735^{***}
	(0.01159)
O: as.factor(dealer)4 (1)	-0.07146^{***}
	(0.00927)
O: as.factor(dealer)5 (1)	0.09071^{***}
	(0.01327)
O: as.factor(dealer)6 (1)	0.02808^{**}
	(0.00906)
O: as.factor(dealer)7 (1)	0.06230^{***}
	(0.01118)
O: as.factor(dealer)8 (1)	-0.07888^{***}
	(0.00873)
O: as.factor(dealer)9 (1)	-0.09368^{***}
	(0.01300)
O: (Intercept) (2)	1.85329^{***}
	(0.02105)
O: mil_sq	-0.00070^{***}
	(0.00009)
O: age (2)	-0.00458^{*}
	(0.00194)
O: damage (2)	-0.07813^{***}
	(0.00168)
O: as.factor(dealer)2 (2)	-0.04353^{**}
	(0.01390)
O: as.factor(dealer)3 (2)	0.07621^{***}
	(0.01019)
O: as.factor(dealer)4 (2)	-0.06236^{***}
	(0.00980)
O: as.factor(dealer)5 (2)	0.05792^{***}
	(0.01314)
O: as.factor(dealer)6 (2)	0.01601
	(0.00953)
O: as.factor(dealer)7 (2)	0.03276^{**}
	(0.01164)
O: as.factor(dealer)8 (2)	-0.05614^{***}
	(0.00943)
O: as.factor(dealer)9 (2)	-0.09278^{***}
	(0.01335)
sigma1	0.13803^{***}
	(0.00370)
sigma2	0.20156^{***}
	(0.00262)
rho1	0.44515^{***}
	(0.06530)
rho2	-0.48684^{***}
	(0.03799)
AIC	3628.98886
BIC	3844.87915
Log Likelihood	-1784.49443
Num. obs.	9861
Censored	3077
Observed	6784

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Tab. 2: Selection Model for Used Truck Rate of Return