Spring 2021 **Firstname M. Lastname**

University of Central Florida
College of Business

QMB 6911
Capstone Project in Business Analytics

Solutions: Problem Set #10

1 Data Description

This analysis follows the script `Tractor_Reg_Model.R` to produce a more accurate model for used tractor prices with the data from `TRACTOR7.csv` in the `Data` folder. The dataset includes the following variables.

$saleprice_i$ = the price paid for tractor i in dollars
$horsepower_i$ = the horsepower of tractor i
age_i = the number of years since tractor i was manufactured
$enghours_i$ = the number of hours of use recorded for tractor i
$diesel_i$ = an indicator of whether tractor i runs on diesel fuel
fwd_i = an indicator of whether tractor i has four-wheel drive
$manual_i$ = an indicator of whether tractor i has a manual transmission
$johndeere_i$ = an indicator of whether tractor i is manufactured by John Deere
cab_i = an indicator of whether tractor i has an enclosed cab
$spring_i$ = an indicator of whether tractor i was sold in April or May
$summer_i$ = an indicator of whether tractor i was sold between June and September
$winter_i$ = an indicator of whether tractor i was sold between December and March

I will revisit the recommended linear model from Problem Set #7, which was supported in Problem Sets #8 and #9 by considering other nonlinear specifications within a Generalized Additive Model.

Then I will further investigate this nonlinear relationship by considering the issue of sample selection: John Deere may produce tractors of specific qualities based on their perceived value to typical John Deere customers, in ways that are not represented by the variables in the dataset.

	Model 1	Model 2
(Intercept)	8.72792^{***}	8.72792^{***}
	(0.10602)	(0.05228)
horsepower	0.01112^{***}	0.01112^{***}
	(0.00107)	(0.00053)
squared_horsepower	-0.00001^{***}	-0.00001^{***}
	(0.00000)	(0.00000)
age	-0.03233^{***}	-0.03233^{***}
	(0.00358)	(0.00177)
enghours	-0.00004^{***}	-0.00004^{***}
	(0.00001)	(0.00000)
diesel	0.20350^{*}	0.20350^{***}
	(0.09805)	(0.04835)
fwd	0.26539^{***}	0.26539^{***}
	(0.05820)	(0.02870)
manual	-0.15015^{*}	-0.15015^{***}
	(0.06189)	(0.03052)
johndeere	0.31872^{***}	0.31872^{***}
	(0.07186)	(0.03543)
cab	0.48345^{***}	0.48345^{***}
	(0.07003)	(0.03453)
R^2	0.80591	0.80591
Adj. R^2	0.79935	0.80432
Num. obs.	276	1104

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

Tab. 1: Quadratic Model for Tractor Prices

2 Linear Regression Model

A natural staring point is the recommended linear model from Problem Set #7.

2.1 Quadratic Specification for Horsepower

In the demo for Problem Set #7, we considered the advice of a used tractor dealer who reported that overpowered used tractors are hard to sell, since they consume more fuel. This implies that tractor prices often increase with horsepower, up to a point, but beyond that they decrease. To incorporate this advice, I created and included a variable for squared horsepower. A decreasing relationship for high values of horsepower is characterized by a positive coefficient on the horsepower variable and a negative coefficient on the squared horsepower variable.

The results of this regression specification are shown in Table 1. The squared horsepower variable has a coefficient of $-2.081e-05$, which is nearly ten times as large as the standard error of $2.199e-06$, which is very strong evidence against the null hypothesis of a positive or zero coefficient. I conclude that the log of the sale price does decline for large values of horsepower.

With the squared horsepower variable, the $\bar{R}^2$ is 0.764, indicating that it is a much stronger model than the others we considered. The F-statistic is large, indicating that it is a better candidate than the simple average log sale price. The new squared horsepower variable is statistically significant and the theory behind it is sound, since above a certain point, added horsepower may not improve performance but will cost more to operate. This new model is much improved over the previous models with a linear specification for horsepower. Next, I will attempt to improve on this specification, using Tobit models for sample selection.

2.1.1 Separate Models by Brand

To test for many possible differences in models by brand of tractor, Table 2 shows the estimates for two separate models by brand of tractor. Model 1 shows the estimates for the full sample, Model 2 shows the estimates from the full model for John Deere tractors and Model 4 represents all other brands. Models 3 and 5 show the estimates from a reduced version of each model, in which all coefficients are statistically significant. The coefficients appear similar across the two subsamples. Notable differences include the statistical significance for the indicators for four-wheel drive, manual transmission and an enclosed cab. These features seem to change the value of other tractors, but perhaps these coefficients are not measured accurately for the small sample of 39 John Deere tractors.

We can also test for all of the differences at the same time by using an F-test. In this case, the full, unrestricted model has $K = 2 \times 9 = 18$ parameters, one for each variable in two models. The test that all of the coefficients are the same has $M = 9 - 1 = 8$ restrictions. The one restriction fewer accounts for the John Deere indicator in the full model, which allows for two separate intercepts. The F-statistic has a value of

$$\frac{(RSS_M - RSS)/M}{RSS/(N-K-1)} = \frac{(42.15882 - 41.1432)/3}{41.1432/263} = 0.7929991.$$

This is also a very low value for the F-statistic. There is no evidence to reject the null that all coefficients are equal across both samples and conclude that the John Deere indicator should be the only brand difference left in the model.

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	8.72792^{***}	8.86706^{***}	9.03796^{***}	8.77320^{***}	8.90792^{***}
	(0.10602)	(0.22409)	(0.16430)	(0.12450)	(0.08769)
horsepower	0.01112^{***}	0.01502^{***}	0.01580^{***}	0.01032^{***}	0.01057^{***}
	(0.00107)	(0.00250)	(0.00223)	(0.00119)	(0.00119)
squared_horsepower	-0.00001^{***}	-0.00002^{***}	-0.00002^{***}	-0.00001^{***}	-0.00001^{***}
	(0.00000)	(0.00000)	(0.00000)	(0.00000)	(0.00000)
age	-0.03233^{***}	-0.03038^{**}	-0.03295^{***}	-0.03164^{***}	-0.03283^{***}
	(0.00358)	(0.00914)	(0.00738)	(0.00399)	(0.00392)
enghours	-0.00004^{***}	-0.00006^{*}	-0.00006^{**}	-0.00004^{***}	-0.00004^{***}
	(0.00001)	(0.00002)	(0.00002)	(0.00001)	(0.00001)
diesel	0.20350^{*}	0.08485		0.18218	
	(0.09805)	(0.18242)		(0.11984)	
fwd	0.26539^{***}	0.12882		0.29072^{***}	0.30003^{***}
	(0.05820)	(0.15529)		(0.06308)	(0.06296)
manual	-0.15015^{*}	0.06749		-0.17919^{**}	-0.14668^{*}
	(0.06189)	(0.17288)		(0.06743)	(0.06413)
johndeere	0.31872^{***}				
	(0.07186)				
cab	0.48345^{***}	0.32344	0.38517^{*}	0.51732^{***}	0.52756^{***}
	(0.07003)	(0.17555)	(0.16365)	(0.07696)	(0.07688)
R^2	0.80591	0.91993	0.91606	0.77992	0.77769
Adj. R^2	0.79935	0.89858	0.90334	0.77220	0.77090
Num. obs.	276	39	39	237	237

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

Tab. 2: Separate Models by Brand

	Model 1	Model 2
(Intercept)	-0.77781^{*}	-0.51461
	(0.38590)	(0.26756)
horsepower	0.00019	
	(0.00434)	
squared_horsepower	0.00000	
	(0.00001)	
age	0.00852	
	(0.01433)	
enghours	0.00003	
	(0.00004)	
diesel	-0.97286^{**}	-1.01766^{**}
	(0.35873)	(0.34206)
fwd	0.16677	
	(0.24458)	
manual	0.63033^{*}	0.73176^{**}
	(0.29427)	(0.27747)
cab	-0.67541^{*}	-0.44114^{*}
	(0.30337)	(0.21490)
AIC	220.18598	214.08158
BIC	252.76958	228.56319
Log Likelihood	-101.09299	-103.04079
Deviance	202.18598	206.08158
Num. obs.	276	276

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

Tab. 3: Probit Models for Brand Selection of Tractors

3 Sample Selection

3.1 Predicting the Selection into Samples

The specification in Table 1 assumes a quadratic functional form for the relationship between price and horsepower, without selecting into samples by brand. To investigate this relationship further, consider the set of variables that are related to whether or not John Deere makes a tractor with the characteristics observed in the dataset.

Table 3 shows the estimates for a probit model to predict the selection into samples by brand name. Model 1 in Table 3 shows a preliminary probit model to predict the selection indicator, with all the other explanatory variables in the model. John Deere tractors are more likely to be gasoline-powered, have manual transmissions, and less likely to have an enclosed cab. Model 2 shows the result of a variable-reduction exercise to eliminate variables that are not statistically significant. These estimates provide a concise but useful model to indicate the tractor designs that would be favored by John Deere engineers and customers. This model is used to specify the selection equation of the sample selection estimates investigated next.

3.2 Estimating a Sample Selection Model

In this section, I will make a valiant attempt to fit a sample selection model to the tractor sales data. This exercise is useful because it illustrates a level of difficulty that is often encountered when conducting maximum likelihood estimation, or any for of estimation that involves numeric optimization with a complex objective function.

In contrast, in the sections above, I applied linear regression to separate samples by brand of tractor. I also fit a probit model to predict the brand of a tractor based on its characteristics. In each of these cases, the estimation run much more smoothly. With linear regression, the parameter estimates are obtained from calculation following a closed-form analytical solution obtained by calculus. This calculation is well-defined, except in very well understood cases that can be identified beforehand. Similarly, in the case of the probit model, the likelihood function is well-behaved, with a convex shape like a smooth hill.

When one fits a much more complex model, such as the Tobit type 5 switching model, the situation is different. Not only are we estimating parameters for three equations at once—one selection equation into two brand groups and two observation equations for those groups—the likelihood function can have multiple optima or, as we will see, will often encounter regions of the parameter space in which the required calculations are undefined.

3.2.1 Sample Selection Model 1: Full Model

For a first model, I use the entire set of variables for both observation equations, for John Deere and other brands. I specify the selection equation with the variables above from the probit model.

```
tobit_5_sel_1 <-
  selection(selection = johndeere ~
              diesel + manual + cab,
            outcome = list(log_price_other ~
                             horsepower +
                             squared_horsepower +
                             age +
                             enghours +
                             diesel +
                             fwd +
                             manual +
                             cab,
                           log_price_JD ~
                             horsepower +
                             squared_horsepower +
                             age +
                             enghours +
                             diesel +
                             fwd +
                             manual +
                             cab),
            iterlim = 20,
            # method = '2step',
            data = tractor_sales)
```

Note that this printed some warning messages:

```
Warning messages:
1: In heckit5fit(selection, as.formula(formula1), as.formula(formula2), :
 Inverse Mills Ratio is virtually multicollinear to the rest of explanatory
 variables in the outcome equation 1
2: In heckit5fit(selection, as.formula(formula1), as.formula(formula2), :
 Inverse Mills Ratio is virtually multicollinear to the rest of explanatory
 variables in the outcome equation 2
```

This suggests much overlap between the two models. This is a risk of starting with too rich a model. Inspecting the model summary gives the following result.

```
R> summary(tobit_5_sel_1)
--------------------------------------------
Tobit 5 model (switching regression model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 12 iterations
Return code 3: Last step could not find a value above the current.
Boundary of parameter space?
Consider switching to a more robust optimisation method temporarily.
Log-Likelihood: -268.2932
276 observations: 237 selection 1 (0) and 39 selection 2 (1)
26 free parameters (df = 250)
Probit selection equation:
         Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.4064  Inf     0      1
diesel    -0.9974    Inf     0      1
manual     0.5731    Inf     0      1
cab       -0.5105    Inf     0      1
Outcome equation 1:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.373e+00    Inf     0      1
horsepower     8.717e-03    Inf     0      1
squared_horsepower -1.024e-05 Inf  0      1
age           -2.403e-02    Inf     0      1
enghours      -2.961e-05    Inf     0      1
diesel         4.411e-01    Inf     0      1
fwd            3.966e-01    Inf     0      1
manual        -3.577e-01    Inf     0      1
cab            6.171e-01    Inf     0      1
Outcome equation 2:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.245e+00    Inf     0      1
horsepower     1.838e-02    Inf     0      1
squared_horsepower -2.557e-05 Inf  0      1
age           -1.895e-02    Inf     0      1
enghours      -7.279e-05    Inf     0      1
diesel         6.111e-01    Inf     0      1
fwd            3.582e-01    Inf     0      1
manual        -1.247e-01    Inf     0      1
```

```
cab            4.143e-01    Inf     0      1
  Error terms:
     Estimate Std. Error t value Pr(>|t|)
sigma1 0.5428    Inf     0      1
sigma2 0.7614    Inf     0      1
rho1  -0.9598    Inf     0      1
rho2  -0.9999    Inf     0      1
--------------------------------------------
```

The messages surrounding the optimization suggest a problematic search for the parameter values. Also, the standard errors could not be calculated. Furthermore, when I toggled the `method = ’2step’` commented line, the situation did not improve, aside from returning the results from separate models as in the analysis above. Again, problems such as these are often encountered when starting with too rich a model. Perhaps a less ambitious model is a better starting point.

3.2.2 Sample Selection Model 2: Reduced Model from Separate Estimation

Instead of starting with a “big-to-small” approach, I reconsider the best models for each tractor brand group that were recommended above. I specify the selection equation with the variables above from the probit model. For this refined model, I use the set of variables for separate observation equations, to match the reduced models from separate linear regressions for John Deere and other tractor brands.

```
tobit_5_sel_2 <-
  selection(selection = johndeere ~
            diesel + manual + cab,
          outcome = list(log_price_other ~
                           horsepower +
                           squared_horsepower +
                           age +
                           enghours +
                           # diesel +
                           fwd +
                           manual +
                           cab,
                         log_price_JD ~
                           horsepower +
                           squared_horsepower +
                           age +
                           enghours +
                           # diesel +
                           # fwd +
                           # manual +
                           cab),
          iterlim = 20,
          # method = ’2step’,
          data = tractor_sales)
```

Although there were no error messages this time, the results were similarly disappointing.

Ideally, this estimation method produces standard errors in the above summary but I can inspect the results from the separate models for each sample. For instance, from the linear model for other brands, after using the `method = '2step'` argument, I obtain the following.

```
> summary(tobit_5_sel_2$lm1)

Call:
lm(formula = YO1 ~ -1 + XO1)

Residuals:
    Min      1Q  Median     3Q     Max
-1.65134 -0.20946 0.04333 0.27656 0.70671

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
XO1(Intercept)       8.973e+00 1.142e-01 78.588 < 2e-16 ***
XO1horsepower        1.046e-02 1.193e-03 8.764 4.38e-16 ***
XO1squared_horsepower -1.311e-05 2.559e-06 -5.122 6.44e-07 ***
XO1age              -3.210e-02 4.004e-03 -8.016 5.66e-14 ***
XO1enghours         -3.861e-05 1.077e-05 -3.586 0.000411 ***
XO1fwd               2.965e-01 6.311e-02 4.699 4.53e-06 ***
XO1manual           -1.107e-01 7.570e-02 -1.462 0.145173
XO1cab               4.843e-01 9.078e-02 5.335 2.30e-07 ***
XO1invMillsRatio -2.694e-01 3.005e-01 -0.897 0.370771
---
Signif. codes: 0 *** 0.001 **   0.01  *   0.05  .   0.1       1

Residual standard error: 0.406 on 228 degrees of freedom
Multiple R-squared: 0.9983, Adjusted R-squared: 0.9982
F-statistic: 1.447e+04 on 9 and 228 DF, p-value: < 2.2e-16
```

One should be cautious in interpreting these results, since the standard errors are underestimated, since this linear model does not account for the uncertainty in the switching component of the sample selection model. Still, this suggests that I can remove the `manual` indicator. A similar investigation of the `tobit_5_sel_2$lm2` model object suggests that the `cab` variable can be removed as well.

3.2.3 Sample Selection Model 3: Reduced Model

I estimate a model that is the same as the last in all ways, except that the `manual` variable was removed from the first observation equation (other tractors) and the `cab` variable from the second (John Deere) equation.

```
tobit_5_sel_3 <-
  selection(selection = johndeere ~
            diesel + manual + cab,
          outcome = list(log_price_other ~
                      horsepower +
                      squared_horsepower +
                      age +
                      enghours +
                      # diesel +
                      fwd +
                      # manual +
                      cab,
                    log_price_JD ~
                      horsepower +
                      squared_horsepower +
                      age # +
                      # enghours # +
                      # diesel +
                      # fwd +
                      # manual +
                      # cab
                    ),
          iterlim = 20,
          method = '2step',
          data = tractor_sales)
```

As above, I retrieve the summary from each model.

```
> summary(tobit_5_sel_3$lm1)

Call:
lm(formula = YO1 ~ -1 + XO1)

Residuals:
    Min    1Q  Median    3Q    Max
-1.68195 -0.22481 0.05553 0.26875 0.68681

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
XO1(Intercept)    9.012e+00 1.115e-01 80.855 < 2e-16 ***
XO1horsepower     1.035e-02 1.194e-03 8.667 8.16e-16 ***
XO1squared_horsepower -1.293e-05 2.563e-06 -5.044 9.28e-07 ***
XO1age           -3.283e-02 3.983e-03 -8.243 1.31e-14 ***
XO1enghours      -4.037e-05 1.073e-05 -3.763 0.000213 ***
XO1fwd            2.861e-01 6.286e-02 4.551 8.67e-06 ***
XO1cab            4.367e-01 8.495e-02 5.141 5.86e-07 ***
XO1invMillsRatio -5.025e-01 2.553e-01 -1.968 0.050239 .
---
Signif. codes: 0 *** 0.001 **  0.01 *  0.05 .  0.1     1

Residual standard error: 0.407 on 229 degrees of freedom
Multiple R-squared: 0.9982, Adjusted R-squared: 0.9982
F-statistic: 1.619e+04 on 8 and 229 DF, p-value: < 2.2e-16

> summary(tobit_5_sel_3$lm2)

Call:
lm(formula = YO2 ~ -1 + XO2)
```

```
Residuals:
    Min     1Q  Median    3Q     Max
-0.84032 -0.25558 -0.01936 0.27142 0.87561

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
XO2(Intercept)    8.626e+00 3.177e-01 27.154 < 2e-16 ***
XO2horsepower     1.536e-02 2.044e-03 7.516 1.00e-08 ***
XO2squared_horsepower -2.018e-05 4.536e-06 -4.449 8.78e-05 ***
XO2age           -4.599e-02 6.331e-03 -7.264 2.07e-08 ***
XO2invMillsRatio  3.800e-01 2.097e-01 1.812 0.0788 .
---
Signif. codes: 0 *** 0.001 **  0.01 *  0.05 .  0.1     1

Residual standard error: 0.3892 on 34 degrees of freedom
Multiple R-squared: 0.9985, Adjusted R-squared: 0.9983
F-statistic: 4670 on 5 and 34 DF, p-value: < 2.2e-16
```

This time, all of the explanatory variables appear significant, although, again, I caution the reader that the standard errors are underestimated. The results do not suggest a clear path to a further reduced model. Another model-building strategy is in order.

3.2.4 Sample Selection Model 4: Simple Model

The variable reduction strategy performed above has lead to a dead end. As another alternative, I follow a "small-to-big" strategy by estimating a simple model with only a few variables in each equation. On example is the following model with one variable in the selection equation and only two continuous variables in the observation equations.

```
tobit_5_sel_4 <-
  selection(selection = johndeere ~
            # diesel +
            manual # +
            # cab
          ,
          outcome = list(log_price_other ~
                       horsepower +
                       # squared_horsepower +
                       age # +
                       # enghours +
                       # diesel +
                       # fwd +
                       # manual +
                       # cab
                       ,
                     log_price_JD ~
                       horsepower +
                       # squared_horsepower +
                       age # +
                     # enghours # +
                     # diesel +
                     # fwd +
                     # manual +
                     # cab
          ),
          iterlim = 20,
          method = '2step',
          data = tractor_sales)
```

Even for such a simple model, however, the estimation was problematic and no standard errors could be calculated. This suggests a fundamental problem with the likelihood function. The likelihood function is very flat over this region of the parameter space, enough that the curvature could not be used to calculate a nonsingular Hessian matrix. Again, I investigate the fit of the individual models.

```
> summary(tobit_5_sel_4$lm1)

Call:
lm(formula = YO1 ~ -1 + XO1)

Residuals:
    Min      1Q  Median     3Q     Max
-1.58645 -0.34390 0.01728 0.40160 1.16196

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
XO1(Intercept) 9.5619331 0.2889648 33.090 <2e-16 ***
XO1horsepower 0.0068993 0.0004473 15.425 <2e-16 ***
XO1age       -0.0421246 0.0040024 -10.525 <2e-16 ***
XO1invMillsRatio -0.4595151 1.1962781 -0.384 0.701
---
Signif. codes: 0 *** 0.001 **  0.01 *  0.05 .  0.1     1

Residual standard error: 0.543 on 233 degrees of freedom
Multiple R-squared: 0.9968, Adjusted R-squared: 0.9967
F-statistic: 1.817e+04 on 4 and 233 DF, p-value: < 2.2e-16

> summary(tobit_5_sel_4$lm2)

Call:
lm(formula = YO2 ~ -1 + XO2)

Residuals:
    Min      1Q  Median     3Q     Max
-0.76425 -0.41769 -0.01199 0.30249 1.05331

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
XO2(Intercept) 13.7697066 1.8495934 7.445 1.03e-08 ***
XO2horsepower 0.0068541 0.0008473 8.089 1.59e-09 ***
```

```
XO2age       -0.0444887 0.0078166 -5.692 1.97e-06 ***
XO2invMillsRatio -2.6490903 1.1416835 -2.320 0.0263 *
---
Signif. codes: 0 *** 0.001 **   0.01  *   0.05  .   0.1      1

Residual standard error: 0.4832 on 35 degrees of freedom
Multiple R-squared: 0.9977, Adjusted R-squared: 0.9974
F-statistic: 3784 on 4 and 35 DF, p-value: < 2.2e-16
```

Even with inflated standard errors, the above results do not suggest an obvious flaw with the model, however simple. At this point, I concede that the sample selection model has limited potential with this dataset.

3.3 Discussion

Although this outcome appears negative to the researcher who has invested valuable time in the estimation, it still conveys some understanding of the data. If a model is not suited to the data, the model will not fit well and the numerical optimization technique may present may difficulties. In these circumstances, one should be careful to make many careful attempts to fit the model, since even a well-specified model often presents numerical challenges. For example, choose different starting values or different optimization methods.

In the end, I did not succeed in finding an optimum in which the estimates were well-defined. Looking back at the model, however, this outcome makes sense, given the nature of the data: in our preliminary analysis, we rejected the possibility of separate models by brand name. We found that the only interaction between brand name and other variables is with the intercept term: John Deere tractors pull in a higher resale value than other tractors. In order for a switching model to make sense, the value of one set of characteristics should be higher for one model in some circumstances and lower in others. In this case, however, the recommended model predicts a uniformly higher value for John Deere tractors.

In order for the switching model to make sense here, the engineers at John Deere would be choosing designs among all conceivable tractor designs and producing those that are most valuable to John Deere customers. According to the investigation from separate models by sample, all tractor designs are more valuable for John Deere tractors and the notion of selection has little to support it. Perhaps the premium for John Deere tractors is derived from some actions not related to the characteristics of the tractors that were recorded in the dataset, such as their quality control practices, their marketing strategy, or simply the reputation from a long history in the business. Such an analysis is outside the scope of the current study.