Spring 2023 **Firstname M. Lastname**

University of Central Florida
College of Business

QMB 6911
Capstone Project in Business Analytics

Solutions: Problem Set #8

1 Data Description

By engaging an industry consultant to gather relevant and appropriate information, your firm has been able to put together data concerning 248 different fly-fishing reels, over one-half of which are produced in the United States, with the remainder being produced in Asia—either in China or Korea. These data are contained in the file `FlyReels.csv`, which is available in the `Data` folder. Each fly-fishing reel in the data set is a row, while the columns correspond to the variables whose names and definitions are the following:

Variable	Definition
`Name`	product name (a string)
`Brand`	brand name (a string)
`Weight`	weight of reel in ounces (a real number)
`Diameter`	diameter of reel in inches (a real number)
`Width`	width of reel in inches (a real number)
`Price`	price of reel in dollars (a real number)
`Sealed`	whether the reel is sealed; `"Yes"` versus `"No"` (a string)
`Country`	country of manufacture, (a string)
`Machined`	whether the reel is machined versus cast; machined=`"Yes"`, while cast=`"No"` (a string)

I will revisit the recommended linear model from Problem Set #7, which included

I will investigate any nonlinear relationships by incorporating a nonparametric specification for the value of the dimensions of the reels: the width, diameter, and density, which constitute the continuous variables in the dataset. The nonparametric analysis will be performed to investigate whether the value of these characteristics are best described with nonlinear forms.

	Model 1
(Intercept)	2.00999***
	(0.26125)
Width	0.33575*
	(0.15622)
Diameter	0.39567***
	(0.05076)
Density	1.21296***
	(0.21948)
SealedYes	0.62731***
	(0.08622)
MachinedYes	0.64934***
	(0.08320)
made_in_USATRUE	0.74633***
	(0.09247)
SealedYes:made_in_USATRUE	−0.29519**
	(0.10092)
R^2	0.74893
Adj. R^2	0.74160
Num. obs.	248

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

Tab. 1: Linear Model for Fly Reel Prices

2 Linear Regression Model

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

2.1 Linear Model with `Sealed*Made_in_USA` Interaction

Last week I investigated whether the functional form should include different specifications by country of manufacture. The model included the continuous variables width, diameter, and density, as well as categorical variables for country of manufacture, and whether or not the reels were sealed or machined. In addition to the indicator for the country of manufacture, the model included an indicator for an interaction between the the country of manufacture indicator and the indicator for whether the reels were sealed or unsealed. The dependent variable was chosen as the logarithm of the fly reel price, since the results were similar to those from the model with the optimal Box-Cox transformation, without the added complexity. The results of this regression specification are shown in Table 1.

Next, I will attempt to improve on this specification by investigating the potential for nonlinear functional forms.

3 Nonlinear Specifications

3.1 Nonparametric Specification for Width

As above, I first conduct FWL regressions to reduce the problem to two dimensions. The results are not shown here, since the comparison only verifies the conclusion of the FWL theorem.

To illustrate the fit of the linear model, Figure 1 shows a scatter plot of the residual log prices on the width of fly reels. The observations are shown in blue and the fitted values are shown in red. The variation in the fitted values results from the fact that it is not plotted against the transformed excess width variable used in the regressions. Still, the linear pattern is apparent and appears to match the data.

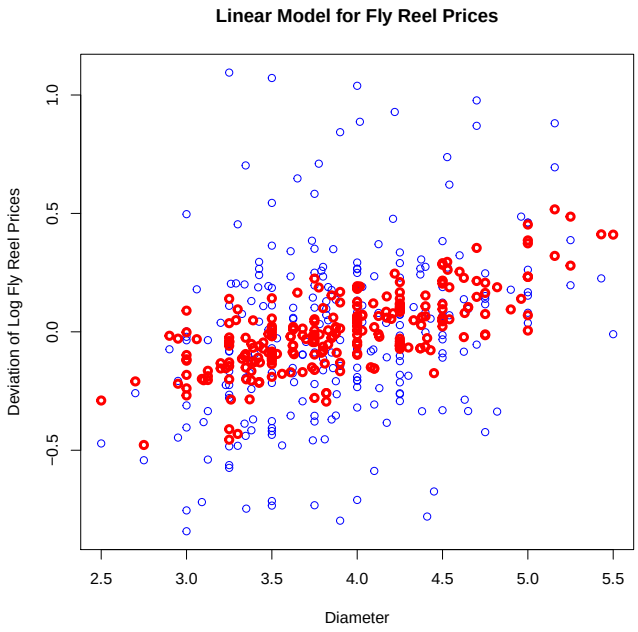

Fig. 1: Linear Model for Fly Reel Prices vs. Width

As a comparison, Figure 2 augments the above by showing the plot against the residuals from the regression for width: the "excess width" of a fly reel compared to what would be expected given the other characteristics of the fly reel. The fit follows a straight line, as specified in the model. I move directly to the nonparametric specification for the relationship between prices and width. Figure 2 overlays the nonparametric estimate, shown in green. The pattern has more variation in slope but closely follows the prediction from the linear model. Although the nonparametric estimate varies around the linear estimate, it appears that the linear form is also a close enough approximation.

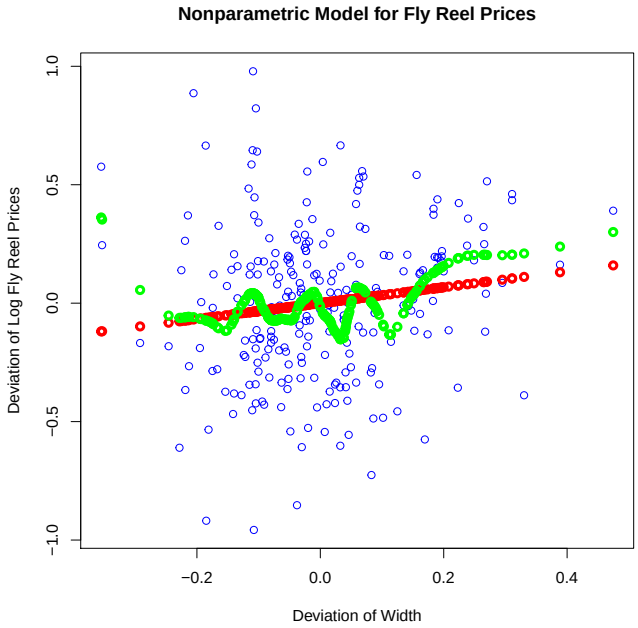

Fig. 2: Nonparametric Model for Fly Reel Prices: Excess Width

3.2 Nonparametric Specification for Diameter

To illustrate the fit of the linear model, Figure 3 shows a scatter plot of the residual log prices on the diameter of fly reels. The observations are shown in blue and the fitted values are shown in red. The variation in the fitted values results from the fact that it is not plotted against the transformed excess diameter variable used in the regressions. Still, the linear pattern is apparent and appears to match the data.

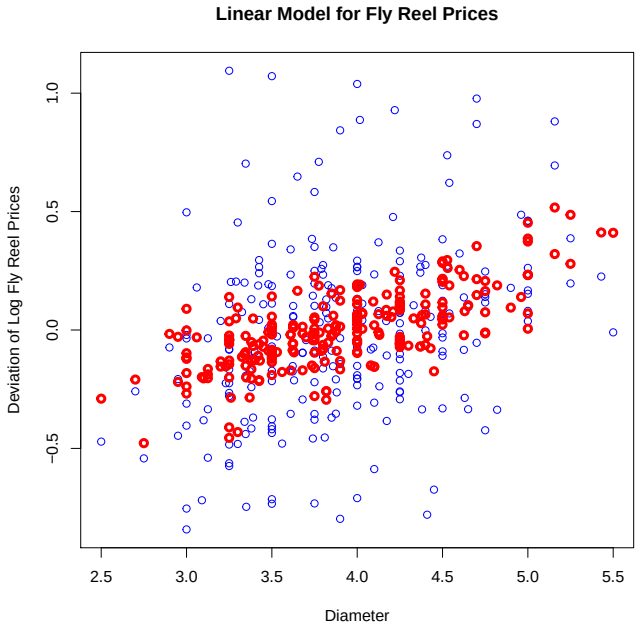

Fig. 3: Linear Model for Fly Reel Prices vs. Diameter

As a comparison, Figure 4 augments the above by showing the plot against the residuals from the regression for diameter: the "excess diameter" of a fly reel compared to what would be expected given the other characteristics of the fly reel. The fit follows a straight line, as specified in the model. I move directly to the nonparametric specification for the relationship between prices and diameter. Figure 4 overlays the nonparametric estimate, shown in green. The pattern has more variation in slope but closely follows the prediction from the linear model. Although the nonparametric estimate varies around the linear estimate, it appears that the linear form is also a close enough approximation.

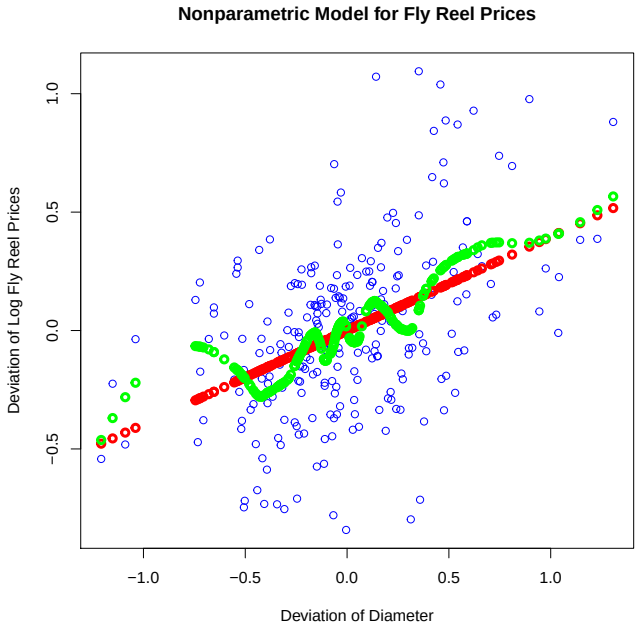

Fig. 4: Nonparametric Model for Fly Reel Prices: Excess Diameter

3.3 Nonparametric Specification for Density

To illustrate the fit of the linear model, Figure 5 shows a scatter plot of the residual log prices on the density of fly reels. The observations are shown in blue and the fitted values are shown in red. The variation in the fitted values results from the fact that it is not plotted against the transformed excess density variable used in the regressions. Still, the linear pattern is apparent and appears to match the data.

Fig. 5: Linear Model for Fly Reel Prices vs. Density

As a comparison, Figure 6 augments the above by showing the plot against the residuals from the regression for density: the "excess density" of a fly reel compared to what would be expected given the other characteristics of the fly reel. The fit follows a straight line, as specified in the model. I move directly to the nonparametric specification for the relationship between prices and density. Figure 6 overlays the nonparametric estimate, shown in green. The pattern has more variation in slope but closely follows the prediction from the linear model. Although the nonparametric estimate varies around the linear estimate, it appears that the linear form is also a close enough approximation.

Fig. 6: Nonparametric Model for Fly Reel Prices: Excess Density

4 Semiparametric Estimates

As I was building the above nonparametric models, I stored the predictions and will now use them as variables in linear models. Table 2 shows the estimates from a set of models. Model 1 is the benchmark linear model in Table 1. Model 2 is a semi-parametric model with a nonparametric fit on width substituted in for the width variable. Models 3 and 4 are semi-parametric models with nonparametric fits on diameter and density, respectively. Model 5 is a maximally semiparametric model, with nonparametric fits for all continuous variables. For each of the single-variable semiparametric models, the coefficients are near one and the fits are similar to the linear model. Even with maximal flexibility, the fit of Model 5 is slightly better than the benchmark linear model. Across all models, the adjusted $\bar{R}^2$ values are all hovering around 0.75, with the full parametric model up to 0.80. All things considered, these are excellent models and the linear model is sufficient but you might recommend the full semiparametric model if you can justify the additional complexity.

One factor to keep in mind, however, is that the above semiparametric models essentially take the nonparametric functions as known, and do not account for the additional variability of the nonarametric parts of the model. The next specification estimates both linear and nonlinear parts jointly.

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	2.00999***	2.34926***	2.98091***	3.23438***	4.49531***
	(0.26125)	(0.22596)	(0.22704)	(0.15261)	(0.05897)
Width	0.33575*		0.92791***	0.03044	
	(0.15622)		(0.12784)	(0.14082)	
Diameter	0.39567***	0.43144***		0.33658***	
	(0.05076)	(0.04150)		(0.04671)	
Density	1.21296***	1.07566***	0.81613***		
	(0.21948)	(0.20093)	(0.20645)		
SealedYes	0.62731***	0.61960***	0.70050***	0.56858***	0.69858***
	(0.08622)	(0.08246)	(0.08226)	(0.08047)	(0.07509)
MachinedYes	0.64934***	0.58954***	0.71659***	0.65070***	0.61103***
	(0.08320)	(0.07933)	(0.07938)	(0.07819)	(0.07386)
made_in_USATRUE	0.74633***	0.77354***	0.79615***	0.70473***	0.79326***
	(0.09247)	(0.08855)	(0.08879)	(0.08692)	(0.08296)
SealedYes:made_in_USATRUE	−0.29519**	−0.29826**	−0.33376***	−0.27253**	−0.31356***
	(0.10092)	(0.09642)	(0.09694)	(0.09500)	(0.09038)
width_np		1.11995***			1.35512***
		(0.21565)			(0.20456)
diameter_np			1.00650***		1.00083***
			(0.10926)		(0.10411)
density_np				1.03923***	0.76790***
				(0.12864)	(0.12582)
R^2	0.74893	0.76995	0.76755	0.77748	0.79771
Adj. R^2	0.74160	0.76324	0.76077	0.77099	0.79181
Num. obs.	248	248	248	248	248

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

Tab. 2: Semiparametric Models for Fly Reel Prices

5 Generalized Additive Model

5.1 Linear Model

As an example of the output from the GAM specification, I first estimated the model with no nonlinear terms, which is essentially a linear regression.

```
Family: gaussian
Link function: identity

Formula:
log_Price ~ Width + Diameter + Density + Sealed + Machined +
    made_in_USA + made_in_USA * Sealed

Parametric coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 2.00999    0.26125   7.694 3.69e-13 ***
Width                       0.33575    0.15622   2.149  0.03262 *
Diameter                    0.39567    0.05076   7.795 1.95e-13 ***
Density                     1.21296    0.21948   5.527 8.49e-08 ***
SealedYes                   0.62731    0.08622   7.275 4.88e-12 ***
MachinedYes                 0.64934    0.08320   7.805 1.84e-13 ***
made_in_USATRUE             0.74633    0.09247   8.071 3.35e-14 ***
SealedYes:made_in_USATRUE -0.29519    0.10092  -2.925  0.00378 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


R-sq.(adj) =  0.742   Deviance explained = 74.9%
GCV = 0.10913  Scale est. = 0.10561   n = 248
```

5.2 Semiparametric Model

Since the results of the full semiparametric specification, in Model 5 of Table 2, were so promising, I estimated the model with all three continuous variables specified as nonparametric functions. The result was that all the variables—both linear and nonlinear—were statistically significant. On the other hand, the adjusted R-squared has not increased very much, from 0.742 to 0.769 under this specification, which may not justify the added complexity of the model. Perhaps more importantly, the coefficients on the linear terms are very similar across models, indicating that the models support similar conclusions relating to any business decision involving the "Made in USA" premium. With this second model, we have even more support for those conclusions and are certain that the conclusions are not coincidental results of the functional form decisions for previous models.

```
Family: gaussian
Link function: identity

Formula:
log_Price ~ s(Width) + s(Diameter) + s(Density) + Sealed + Machined +
    made_in_USA + made_in_USA * Sealed

Parametric coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                4.52027    0.06539  69.123  < 2e-16 ***
SealedYes                  0.64536    0.08257   7.816 1.95e-13 ***
MachinedYes                0.62750    0.07980   7.863 1.45e-13 ***
made_in_USATRUE            0.78661    0.08961   8.778 3.85e-16 ***
SealedYes:made_in_USATRUE -0.31047    0.09742  -3.187  0.00164 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
              edf Ref.df      F  p-value
s(Width)    7.052  8.082  3.686 0.000411 ***
s(Diameter) 3.490  4.409 16.859  < 2e-16 ***
s(Density)  2.663  3.403  9.610 3.02e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.769   Deviance explained = 78.5%
GCV = 0.10198  Scale est. = 0.094497  n = 248
```