ProblemSet4

Aaron Graybill

3/15/2021

Problem 3.1.

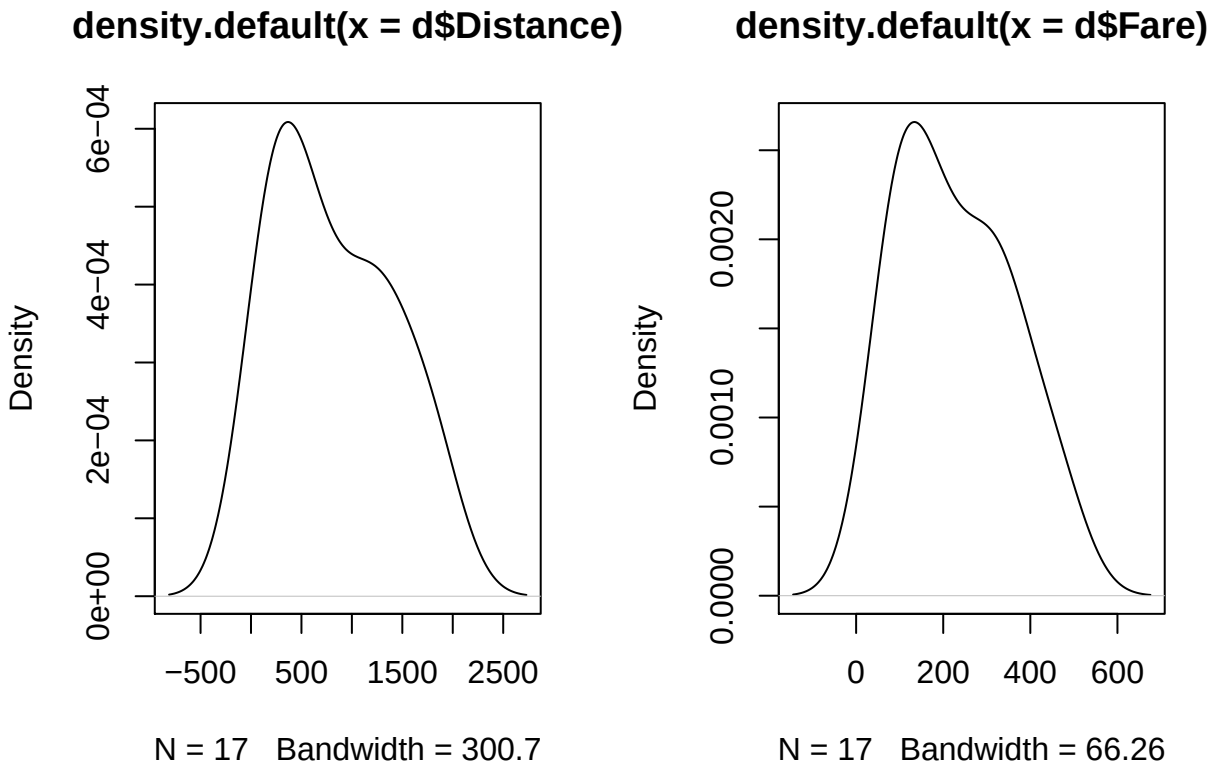
a.

```
## here() starts at /Users/aarongraybill/Documents/Haverford Stuff/Math/Math286
```

The analysis is incomplete. While all of the numbers reported match the regression output, the interpretation should not stated in the way that they was. It is clear from the residuals plot that there is a non-linear relationsihp between the x and y variables because if there was a linear relationship, we would expect there to be no visual relationship between the value of x and the size and direction of residual. However, we see a clear parabolic shape to the residual indicating that the linear model is no suitable for this question. The claim that we can analyze the current effect of distance on fair is not too outlandish because the linear model remains fairly close true values. However, in the future the data might be outside the range currently analyzed which would accentuate the effect of the non-linearity and would make the prediction errors even larger. We cannot forecast for the future values.

b.

All things considered, the model fits the current data quite well, but the can and should be improved through implementing a non-linear model of x and y. The data is not count variables, so a square transformation probably is not the right first step. A log transformation may be appropriate although the plot below shows that the data is not very skewed, so I won't suggest log transformation now.

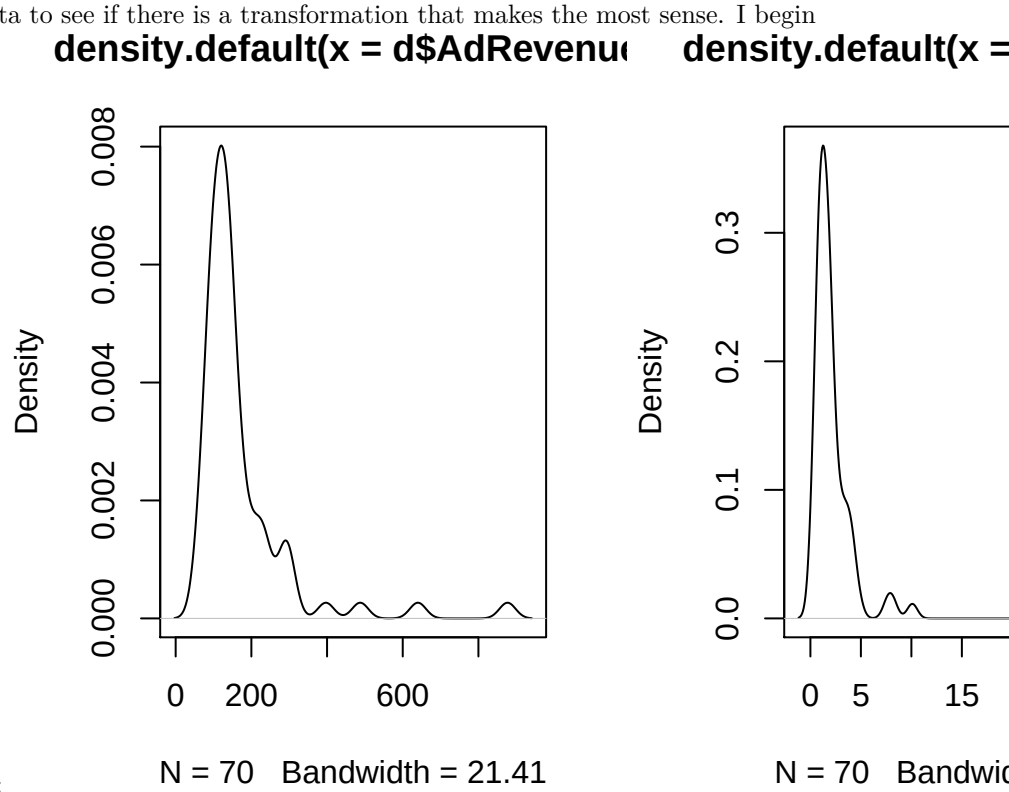


As such the best place to begin might be with an inverse response plot to see if the regression is of the form $Y = g(\beta_0 + \beta_1 x + e)$.

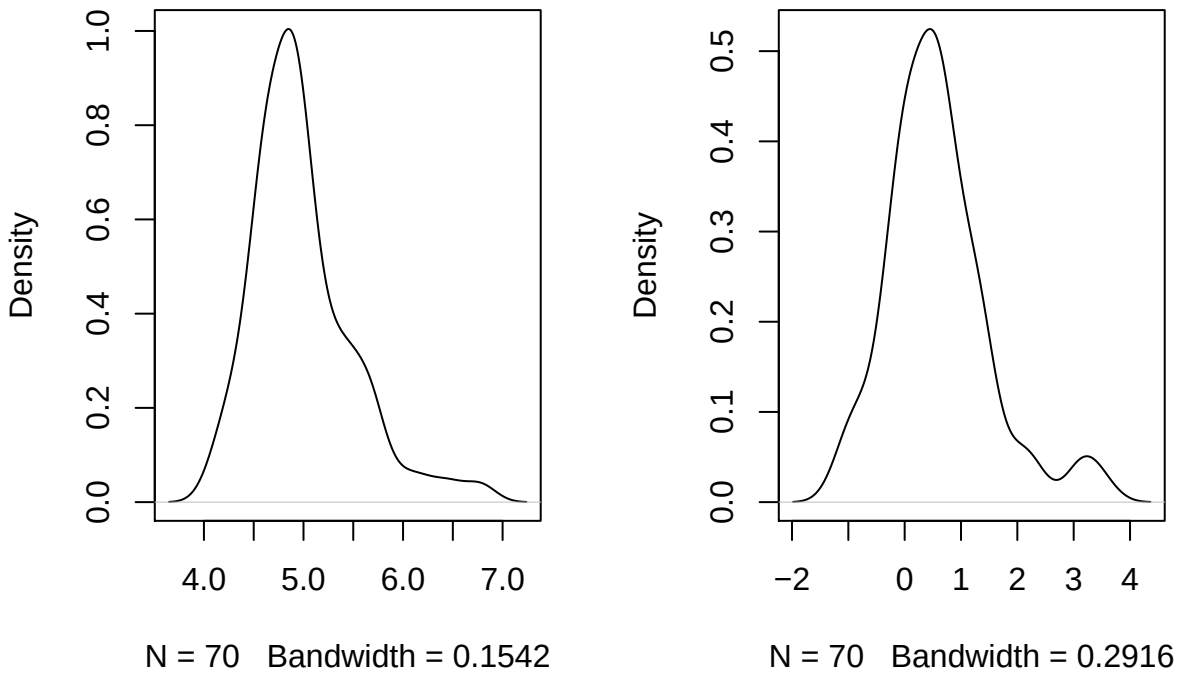
Problem 3.3.A

a.

Before beginning I will explore the data to see if there is a transformation that makes the most sense. I begin with a density plot of the two variables:

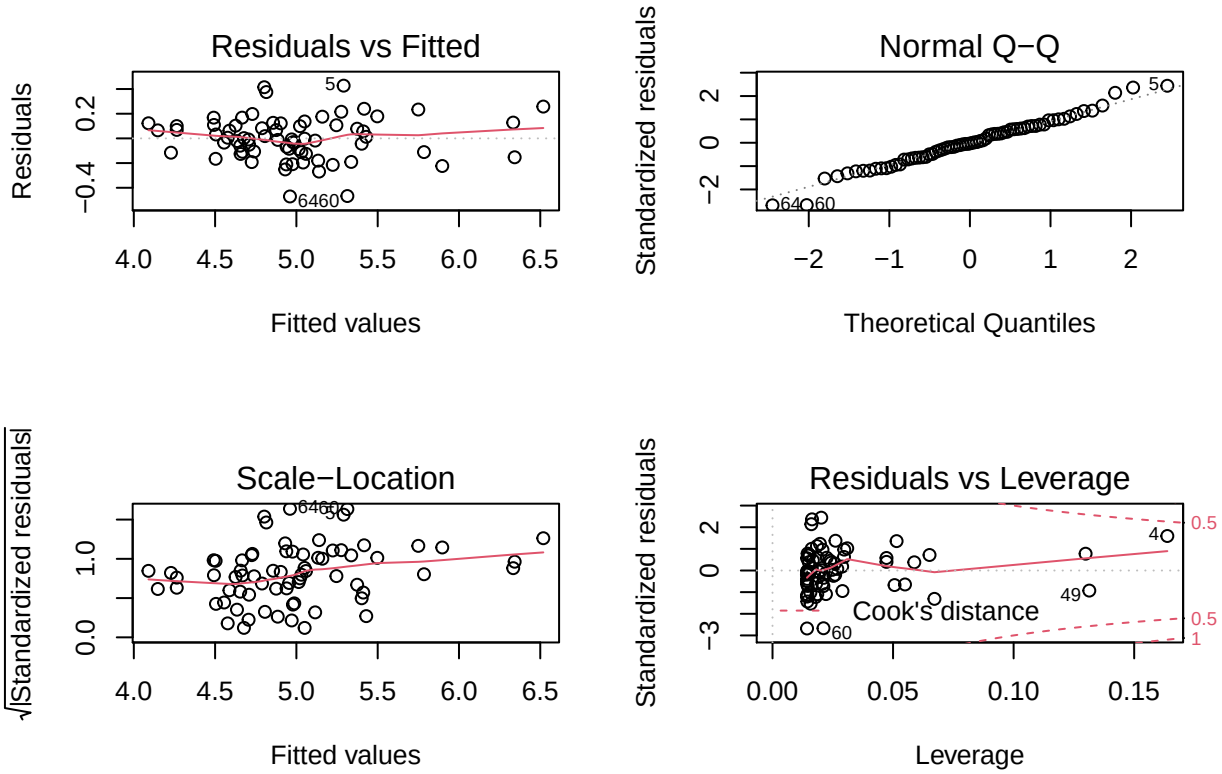


Both variables are quite skew right, so a log transform is in order. Let's apply that and see if it normalized the two:



Okay these variables do look much better, so let's compute the linear model and plot the residuals to ensure they look okay:

```
##
## Call:
## lm(formula = log_AdRevenue ~ log_Circulation, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47022 -0.11142 -0.00532  0.10835  0.42705
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.67473    0.02525  185.16   <2e-16 ***
## log_Circulation  0.52876    0.02356   22.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1768 on 68 degrees of freedom
## Multiple R-squared:  0.881,  Adjusted R-squared:  0.8793
## F-statistic: 503.6 on 1 and 68 DF,  p-value: < 2.2e-16
```



The assumpitions look more or less satistfied with no obvious trends in the plots above.

b.

Since we log transformed the regressor and regressand, we should transform the input and output variables of the prediction interval. By the way this question is asked, it seems like it's not looking for the joint confidence interval, so we can compute in the standard manner.

```
prediction_in <-
  data.frame(log_Circulation=log(c(.5,20)))
prediction_out <-
  exp(predict(out1,prediction_in,interval="prediction",level=.95))
knitr::kable(prediction_out)
```

fit	lwr	upr
74.30864	51.82406	106.5485
522.56626	359.89585	758.7626

Interpreting that table gives that we expect a magazine with half a million readers to attain 74.3 revenue units (thousands of dollars?) with a confidence interval between 51.8 and 106.55. For a magazine with 20 million readers we expect 522.57 units with a lower confidence bound at 359.89 and an upper bound at 758.76.

c.

The biggest weakness in the model here is not really understanding the units and having no theoretical basis for the transformation applied. Some other issues are the slight upward trend in the standardized residual plot indicating we are still not at a perfectly linear relationship.

Problem 3.3.B

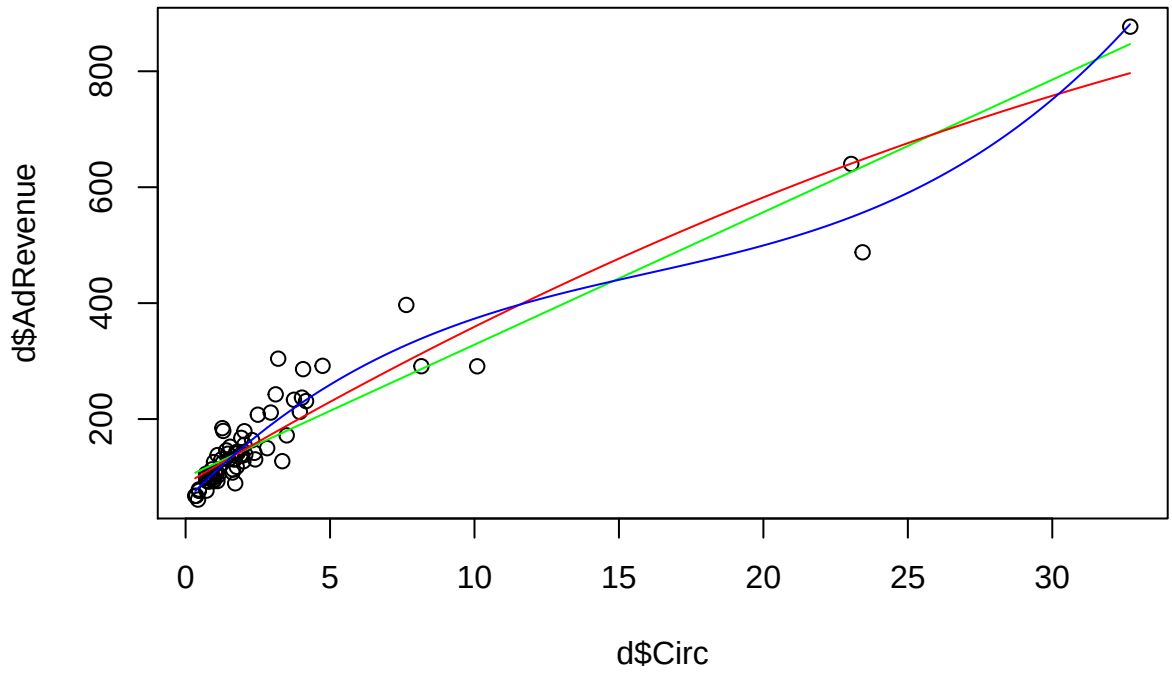
a.

Since all of our polynomial regressions will be using the same, untransformed, dependent variable, we can use the R^2 of the different polynomial regressions to see which is the best predictor of output. I run regressions of orders 1,2,and 3 below:

Table 2:

	Dependent variable:		
	AdRevenue		
	(1)	(2)	(3)
Circ	22.853***	29.501***	51.236***
	(0.952)	(3.299)	(4.711)
Circ2		−0.239**	−2.505***
		(0.114)	(0.411)
Circ3			0.052***
			(0.009)
Constant	99.810***	88.139***	59.170***
	(5.855)	(7.971)	(8.345)
Observations	70	70	70
R^2	0.894	0.901	0.933
Adjusted R^2	0.893	0.898	0.930
Residual Std. Error	42.222 (df = 68)	41.202 (df = 67)	34.065 (df = 66)
F Statistic	576.519*** (df = 1; 68)	304.914*** (df = 2; 67)	308.053*** (df = 3; 66)
Note:			*p<0.1; **p<0.05; ***p<0.01

The R^2 is certainly increasing with every variable, but to be fair the R^2 increases with the addition of any variable. The best we can do is to graphically inspect the three plots to see which matches best. Those graphs



are:
Looking at the three models it seems clear that the 3rd order model is the best, but honestly I'm still not super convinced that this is a viable way of comparing models.

b.

Now i compute the prediction interval:

fit	lwr	upr
84.16846	14.92314	153.4138
499.53342	418.17903	580.8878

c.

I think this is a bad model, it is heavily leveraged on the x outliers, it doesn't even out the data so things are adequately weighted.

Problem 3.3.C

a.

Given the choice, I would choose choose the model from part A because it makes an attempt to normalize the data and make the residuals more random. The outliers in model B are given too much leverage.

b.

For the reasons above, since I was able to normalize both X and Y I would select those prediction intervals for both cirulation .5 and 20. One reason for this is because the log transformations brings these relatively outlier X values closer to the center so they can be better informed by the data around them and not just one point hanging out by itself. That might not be math though.... Oh! one more thing, we learned in class that prediction intervals are senstive to non-normality, so the non-normality in the B model makes it completely inadequate for prediction intervals.

Problem 5.

a.

No prediction intervals cannot be accurately produced from this data because they are senstive to non-normality. The $Q-Q$ plot shows that the theoretical quantiles are not where they should be if the errors were normally distributed. Furthermore, the dealer cost against the standardized residuals is showing two distinct lines which indicate that normality is violated because the errors are not independent of X.

b.

To overcome non-normality, we should first look at the initial scatter plot with the regression line. In this we can see that most of the data is clustered in the bottom left (not just in the bottom or left), this indicates that both the dependent and independent variables need log transformation to normalize them.

c.

This is a huge improvement in terms of not violating the normality assumptions. The standardized residuals seem to be independent of x and the $Q-Q$ is looking much closer to the theoretical line if the data were normal. We cannot really compare the R^2 because the units on the dependent variable have changed.

d.

Every 1% increase in DealerCost is predicted to increase the suggested retail price by 1.01%.

e.

I'm still seeing some deviations from the theoretical quantiles, but overall the model seems pretty good.