ElectronMobility

Charissa Martin

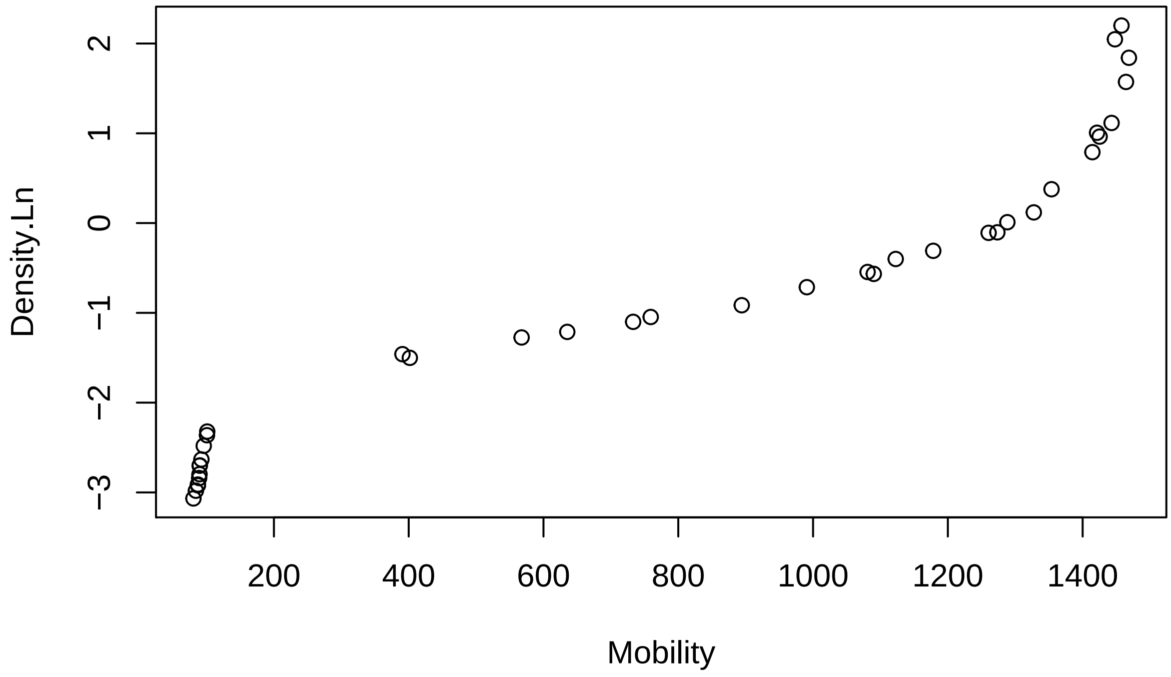
9/9/2019

```
# read the dataset
dataset = read.csv("ElectronMobility.csv")
```

1. Create a scatter plot of the two variables. Describe whether the nature of relationship can be considered linear

Answer: No, it's not linear. It looks like it is logistic.

```
# plot the data
plot(dataset)
```



2. Nexy t, split the dataset into two parts; the first part has all the rows where Mobility $<$ 1000, and the second part has the remainder of the rows.

```
rlt1000 <- dataset[dataset$Mobility<1000,]
rgt1000 <- dataset[dataset$Mobility>1000,]
```

3. Nexy t fit a linear regression model on the first part, where Mobility is the predictor and Density.Ln is the outcome. What is the adjusted R^2?

Answer: Adjusted R-squared: 0.8032

```
linreg <- lm(Density.Ln~Mobility, data = rgt1000)
summary(linreg)
```

```
##
## Call:
## lm(formula = Density.Ln ~ Mobility, data = rgt1000)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4877 -0.3502 -0.1764  0.2841  0.7952
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.5381770  1.0031091  -7.515 1.84e-06 ***
## Mobility     0.0061353  0.0007535   8.142 6.92e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4129 on 15 degrees of freedom
## Multiple R-squared:  0.8155, Adjusted R-squared:  0.8032
## F-statistic:  66.3 on 1 and 15 DF,  p-value: 6.925e-07
```

4. Then, using the predict() function, use this model to predict the value of Density.Ln in the second part of the dataset. Store the predicted values of the outcome in a vector, called predvals.

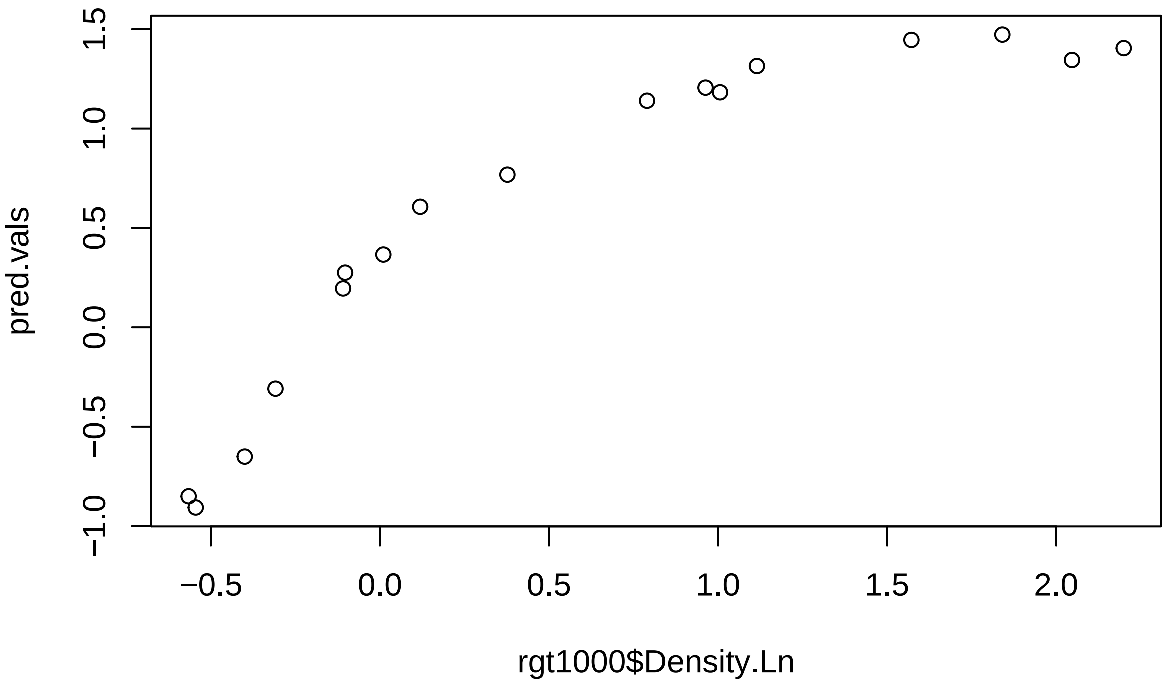
```
pred.vals <- predict.lm(linreg, rgt1000)
```

5. Next, compute the errors of prediction, using the observed values (from Density.Ln in the second half of the dataset) and the predicted values of the outcome

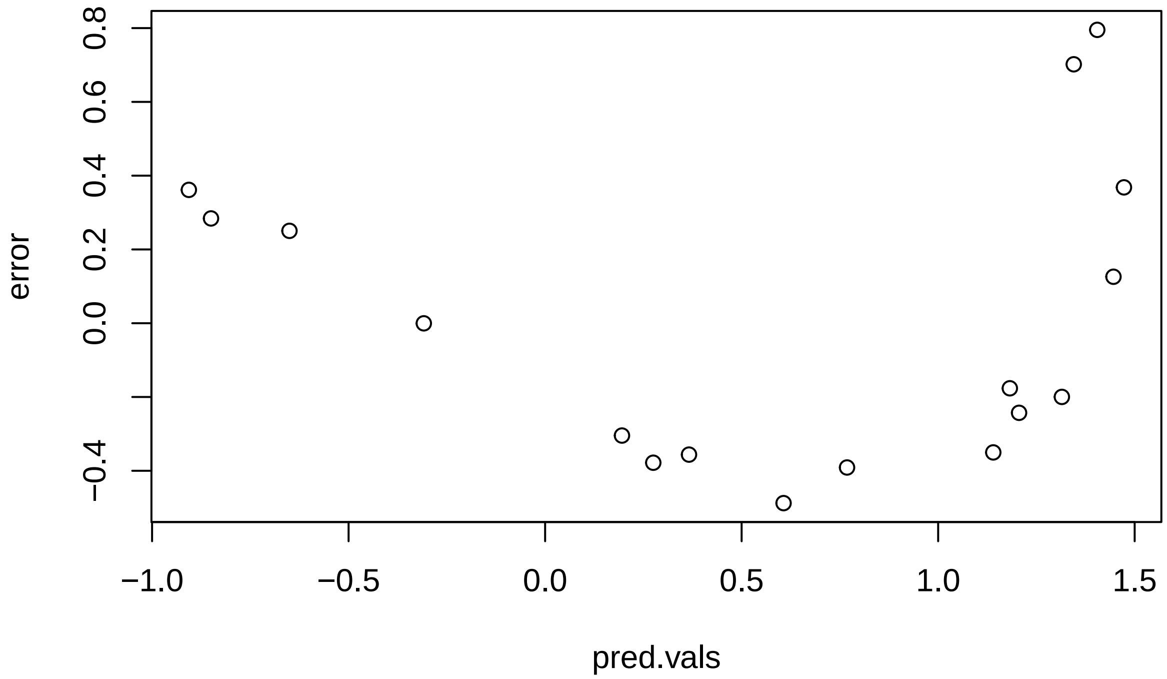
```
error <- rgt1000$Density.Ln - pred.vals
```

6. Create two bivariate plots: 1) between observed and predicted values (from the second half of the dataset), 2) between predicted values and the errors.

```
plot(rgt1000$Density.Ln, pred.vals)
```



```
plot(pred.vals, error)
```



7. Comment on the plots created in part 6, describing the relationships between observed and predicted values, and between predicted and outcome values.

Answer: For predicted and observed values plot: Since the plot shows a pattern curving in in a way that demonstrates higher predicted values than observed values, it might suggest a bias for the predicted values. If the predicted and observed were the same, then we would exy pect the points to ordinate linearlly. For difference between predicted and observed errors plot: A good error distribution would have values as close to zero as possible. That is not the case here, which suggests there is a bias.

8. Next, take into account the plot created for addressing part 1, and combine it with the plots created in part 6. Comment on a), whether the pattern observed between the predictor and the outcome is the same throughout the entire range of predictor values; (b) how, using a model fit on the first half has produced a certain pattern of errors of prediction made when using the model fit on the first part of the dataset for making predictions on the second part of the dataset.

Answer: a) No, it’s not the same throughout. The pattern for the early part of the dataset is different from the rest of it. b) I’m not sure, but I think the model fit based on the data split around 1000 contributted to a pattern of errors, because the data itself does not have differentiated pattens around 1000. I think having it split around ~400 might have fit better. However, that also cause overfitting.