HW10

Rhea Toves

11/2/2021

Problem 1

(A) Describe in your own words the difference between linear and logistic regression.

- The difference between linear and logistic regression is that logistic regression estimates the probability of landing in each category; so this process is categorical instead of continuous.

(B) Give an example of a dataset that would be appropriate to analyze with multiple linear regression but not with logistic regression.

- A dataset that would be appropriate to analyze with multiple linear regression but not with logistic regression could focus on temperature, weather, etc.

(C) Give an example of a dataset that would be appropriate to analyze with logistic regression but not with linear regression.

- A dataset that would be appropriate to analyze with logistic regression but not linear regression could focus on race, sex, ages, etc.

(D) Given a model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

what is the interpretation of the coefficient β_2?

- The coefficient β minimize the sum of squared errors for the sample. The coefficient, β_2, represents the change in mean.

(E) Given a model

$$\ln\left(\frac{p}{1-p}\right) = -5 + 3x_1$$

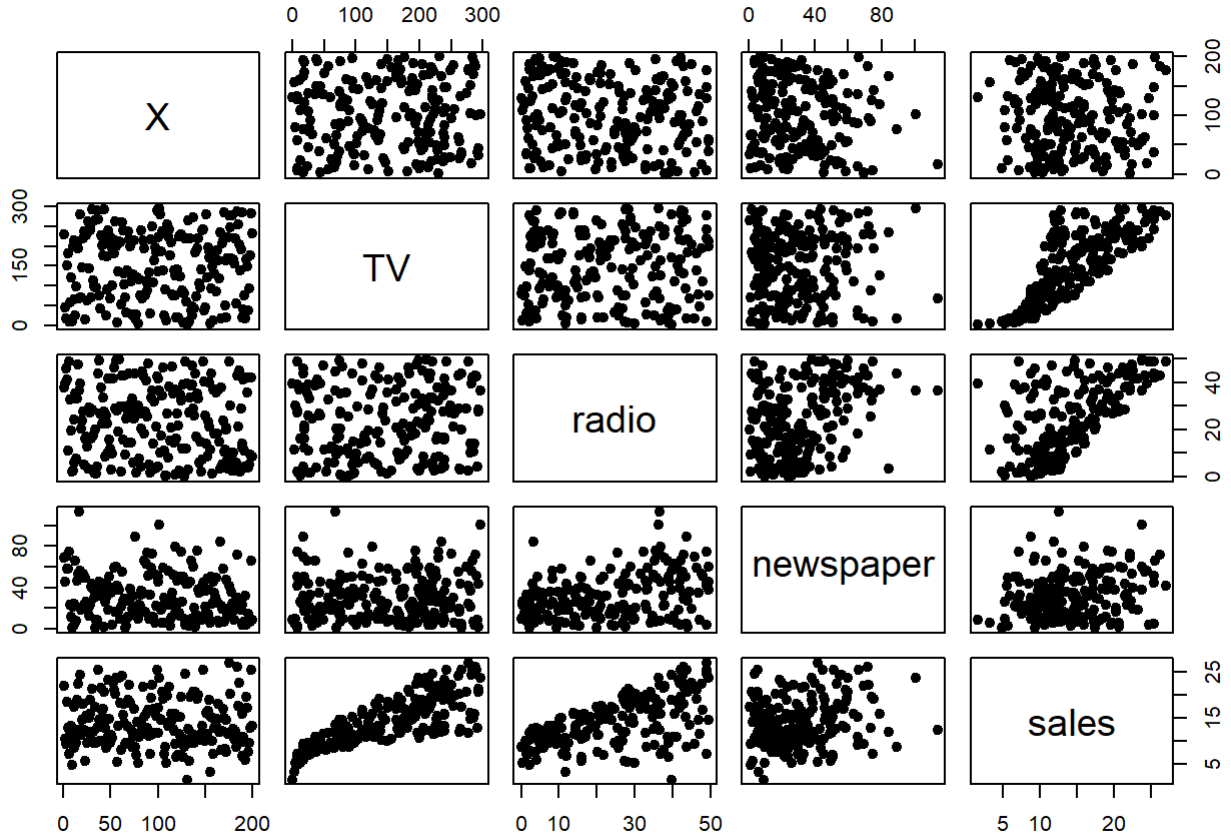
how much do the odds increase if x_1 increases by 1?

- By increasing x_1 by 1, the odds increase by 1 as well.

–

Problem 2 - This problem checks your understanding of multiple linear regression and diagnosis of these fits. Start by loading in the Advertising.csv file as a dataframe.

(A) Make a scatterplot matrix of the columns of the dataframe. Each row of this data set represents a single media market and the TV, Newspaper, and Radio columns contain spending amounts related to each media type while the sales value is the number of units sold (in thousands) in that market. Based on this plot, do you think multiple linear regression is appropriate to attempt?



- I think there are some plots in the scatterplot matrix that would be appropriate to attempt multiple linear regression models on, such as, the columns TV and sales - which have a good structure to perform a model on.

(B) Fit a multiple linear regression model using all three media columns as predictors with the sales column as the dependent variable.

```
##
## Call:
## lm(formula = sales ~ TV + radio + newspaper, data = advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## radio        0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

(C) Write the linear equation estimated by your fit.

$$y_i = 2.938889 + 0.233258 + 1.686$$

(D) Write the coefficient of determination for your fit.

```
## [1] 0.8972106
```

(E) How many sales would your model predict for a market that spent 200 on TV, 50 on radio, and 100 on newspaper?

- Based on my model, I would predict the sales of each electronic device to be: TV = no greater than 30, radio = greater than 10, and newspaper = less than 50.

–

Problem 3 - This problem checks your understanding of implementing logistic regression in R. Start by loading in the default_ISLR.csv file as a dataframe.

(A) Fit a logistic regression model to this data with response variable y being the default column, p being the probability of default, and x being the balance column.

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
##
## Call:
## glm(formula = default ~ balance, family = binomial, data = newISLR)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2697  -0.1465  -0.0589  -0.0221   3.7589
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.065e+01  3.612e-01  -29.49   <2e-16 ***
## balance      5.499e-03  2.204e-04   24.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1596.5  on 9998  degrees of freedom
## AIC: 1600.5
##
## Number of Fisher Scoring iterations: 8
```

(B) Compute the coefficients of the fitted model and write the corresponding equation for the log odds.

- The coefficients of the fitted model computed as significant and the equation for the log odds is,

$$logit(p) = log(p/1 - p)$$

(C) What percentage of the provided data does your model correctly classify?

- I would say, the logistic regression model correctly classifies a good amount of the provided data because both of the coefficients came out as significant.

(D) What is the probability that someone with a balance of 1950 will default, according to your model?

- The null deviance (2920.6) and the residual deviance (1596.5) are not close to the balance of 1950, therefore the probability that someone would have this balance and will be default is low.

–

Problem 4 - Please read the following data analysis project using logistic regression to analyze bias in the COMPAS measure: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm) and the corresponding notebook on github: https://github.com/propublica/compas-analysis/blob/master/Compas%20Analysis.ipynb (https://github.com/propublica/compas-analysis/blob/master/Compas%20Analysis.ipynb) to provide brief responses to the following

(You can skip the section that starts with “To test COMPAS’s overall predictive accuracy, …’ and resume reading down below at:”Finally, we investigated whether certain types of errors…"):

(A) What research question is being addressed in this piece?

- The research question being addressed in this piece focuses on analyzing the COMPAS tool; any biases, errors, patterns, accuracy, etc.

(B) Can you tell where the underlying data came from?

- Yes, the article has a section, “How We Acquired Data”, and this is where the authors discuss what public records they used, how they obtained COMPAS’s data and scores, etc.

(C) The authors give a very thorough description of their data and modeling/cleaning choices. What were the main alterations they had to make to the underlying data set? Do their choices of terminology and metrics seem reasonable to you?

- The authors first focused on the decile (violent and not) scores of black and white defendants to analyze the risk categories for both. After looking at these initial histograms, the authors realized this data, “does not account for other demographic and behavioral factors.” and did some cleaning to the data. The authors then, “created a logistic regression model that considered race, age, criminal history, future recidivism, charge degree, gender and age.” and began to unravel an interesting analysis.

(D) What techniques that we have discussed in our class do the authors use to evaluate the data?

- The authors create logistic regression models to evaluate the data.

(E) What do you think about the visualizations and examples in this piece?

- The visualizations were very nice and helpful to view while reading about the authors’ steps and process of cleaning, sorting, and calculating data.

(F) The authors fit two large logistic regression models to their data. For each model, what is the probability that is being estimated? Which independent variables were found to have the largest impact? Was this surprising to you?

- Risk of general recidivism logistic model: the authors are looking to estimate the probabilities of defendants risk of general recidivism. Every independent variable except for “Asian” and “Native American” were highly significant to this model.

- Risk of violent recidivism logistic model: the authors are looking to estimate the probabilities of defendants risk of violent recidivism. Every independent variable except for “Asian”, “Hispanic”, “Native American”, “Other”, and “Misdemeanor” were highly significant to this model.

(G) At the end of the piece, the authors use contingency tables to compute false positive rates for the COMPAS classification. Do you find this analysis more or less convincing than the logistic model? Why?

- The contingency tables focus on black and white defendants, while the logistic models discuss a lot more data points; therefore I find the logistics regression models to be more convincing.

(H) How strong is the final evidence the authors provide for their conclusions?

- The final evidence the authors provide for their conclusions mainly focus on the misclassified cases for black vs white defendants but every conclusion to each section of the article is strong.