The impact of income and culture on prices: Indentifying discrimination in our communities

Monica Buczynski

2/14/2020

a) The summary of the model is shown below.

```
x.t1 <- xtable(summary(discrim_reg),digits = 6)
print(x.t1,comment = F)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.956320	0.018992	50.353788	0.000000
prpblck	0.114988	0.026001	4.422515	0.000013
income	0.000002	0.000000	4.430130	0.000012

b) The estimated regression equation is:

$$\widehat{psoda} = 0.956320 + 0.114988 prpblck + 0.000002 income.$$

Thus, when the proportion of the population that is black is zero in a zipcode and median income is zero, the price of a medium soda is estimated to be \$0.96. Based on the model, for each additional percentage point increase in the black population living in a zipcode, the price of a medium soda is expected to increase by \$0.11, ceteris paribus. Also, when median income increases by one dollar, the price of a medium soda is expected to increase by \$0.000002. In other words, for every additional \$10,000 in median income, the price of soda is predicted to increase by \$0.02, ceteris paribus.

The Standard Error of the Regression (SER) is the measure of the distance that the data points fall from the regression line (the average distance between the observed and predicted values), on average. The SER of *psod* is 0.08611. The SER is also the estimate of the standard deviation of the error term.

The standard error of the intercept is $1.899E^{-2}$. The standard error of the proportion of the population that is black is $2.6E^{-2}$, certeris paribus. The standard error of income is $3.618E{-}7$, ceteris paribus. These represent the estimates of the standard deviation of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$, respectively. In other words, they are a measure of the precision (i.e. amount of variability) in the OLS estimators.

c) To test the significance of β_1, I test the following hypotheses:

The null hypothesis equation for the proportion of the population that is black is:

$$H_0 : \beta_1 = 0$$

The alternative hypothesis equation for the proportion of the population that is black is:

$$H_1 : \beta_1 \neq 0$$

Based on our sample,

$$\hat{\beta}_1 = 0.115$$

and

$$se(\hat{\beta}_1) = 0.026$$

$$t_{\hat{\beta}_1} = \frac{0.115 - 0}{0.026} = 4.423$$

The calculated test statistic and p-value reported in the model summary are $P(t_{398} < -4.423 \text{ or } t_{398} > 4.423) = 1.26 \times 10^{-5}$, respectively.

To test the hypothesis, at the .01 level of significance (99% confidence level), I have enough evidence to reject the null hypothesis (since the p-value is significantly smaller than 0.1) and can safely conclude that *prpblck* is statistically significant (the slope of the *prpblck* is different than zero).

d) The 99% confidence interval estimate for β_1 is calculated as

$$\hat{\beta}_1 \pm t_{\alpha/2} \times se(\hat{\beta}_1).$$

Lower Bound CI: $\hat{\beta}_1$ – c * se$\hat{\beta}_1$ Upper Bound CI: $\hat{\beta}_1$ + c * se$\hat{\beta}_1$

The estimates can be computed in R as shown below.

```
s.dr <-  summary(discrim_reg) # store the summary of the model as an object
bhat1 <- coef(s.dr)[2] #s.dr$coefficients[2,"Estimate"] gives same results
se.bhat1 <- s.dr $coefficients[2,"Std. Error"] # se(\hat{\beta}_1)
alpha <- 0.01 # level of significance
df <- discrim_reg$df[2] # n-k-1
cv <- qt(1-alpha/2,df) # critical value

# Calculate the CI estimate: point estimate +/- cv * se
LB <- bhat1 - cv*se.bhat1 # lower bound
UB <- bhat1 + cv*se.bhat1 # upper bound
c(LB,UB)
```

```
## [1] NA NA
```

In the long run, I have 99% confidence (.01 level of significance), that β_1 is between 0.04769235 and 0.18228404. At the .01 level of significance, I have enough evidence to reject the null hypothesis since the value of the null value (0) does not fall within the range of plausible values (0.0477158 and 0.1823042) and can conclude that *prpblck* is statistically significant.

d) The $R^2 value is 0.06422, meaning that only 6.4% of the variation in price of soda can be explained by the model.

```
s.dr$r.squared # R^2
```

```
## [1] 0.06422039
```

e) The join test of significance tests the following hypotheses;

$$H_0 : \beta_1 = \beta_2 = 0 \quad \text{vs.} \quad H_1 : H_0 \text{ is false.}$$

```
s.dr$fstatistic[1] # F_statistic
```

```
##    value
## 13.65691
```

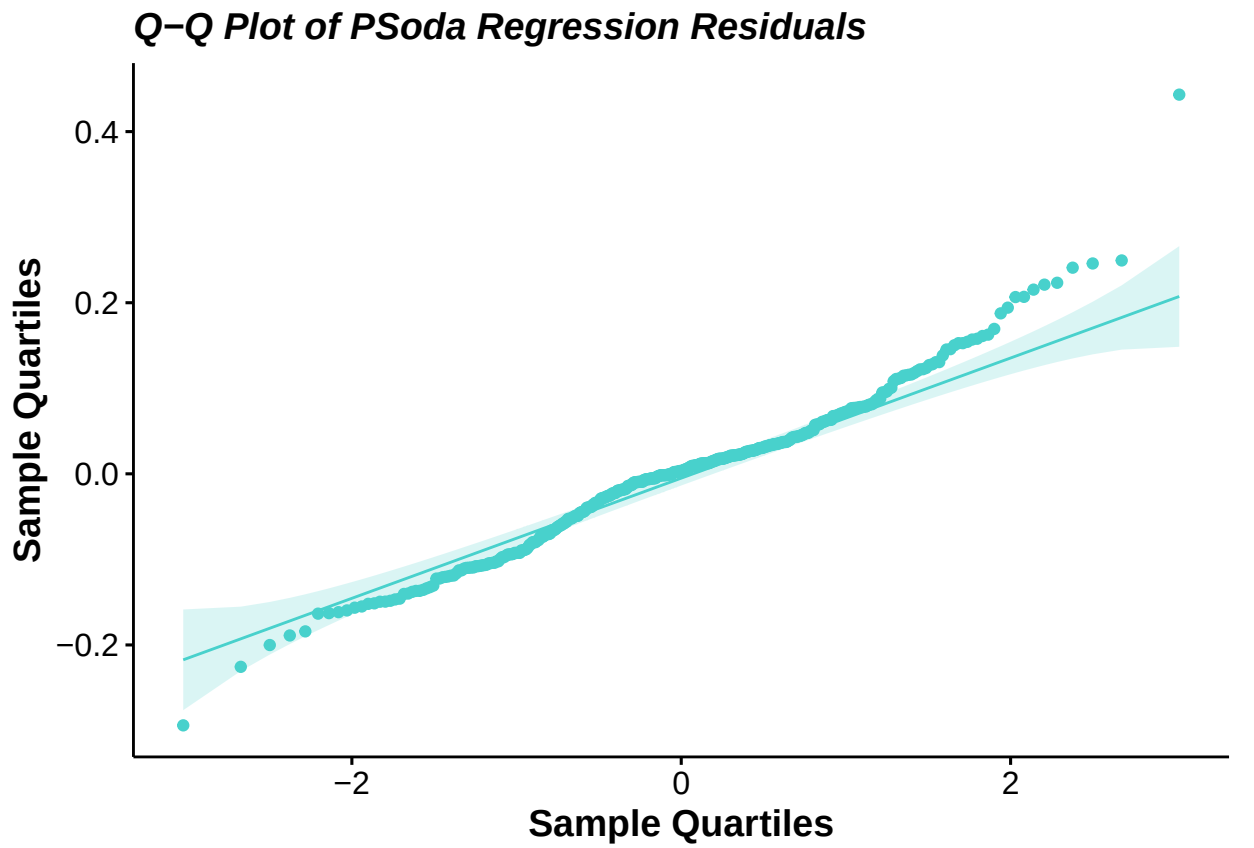
```
pf(s.dr$fstatistic[1], df1 = s.dr$fstatistic[2],df2 = s.dr$fstatistic[3],lower.tail = F) # p-value
```

```
##        value
## 1.834602e-06
```

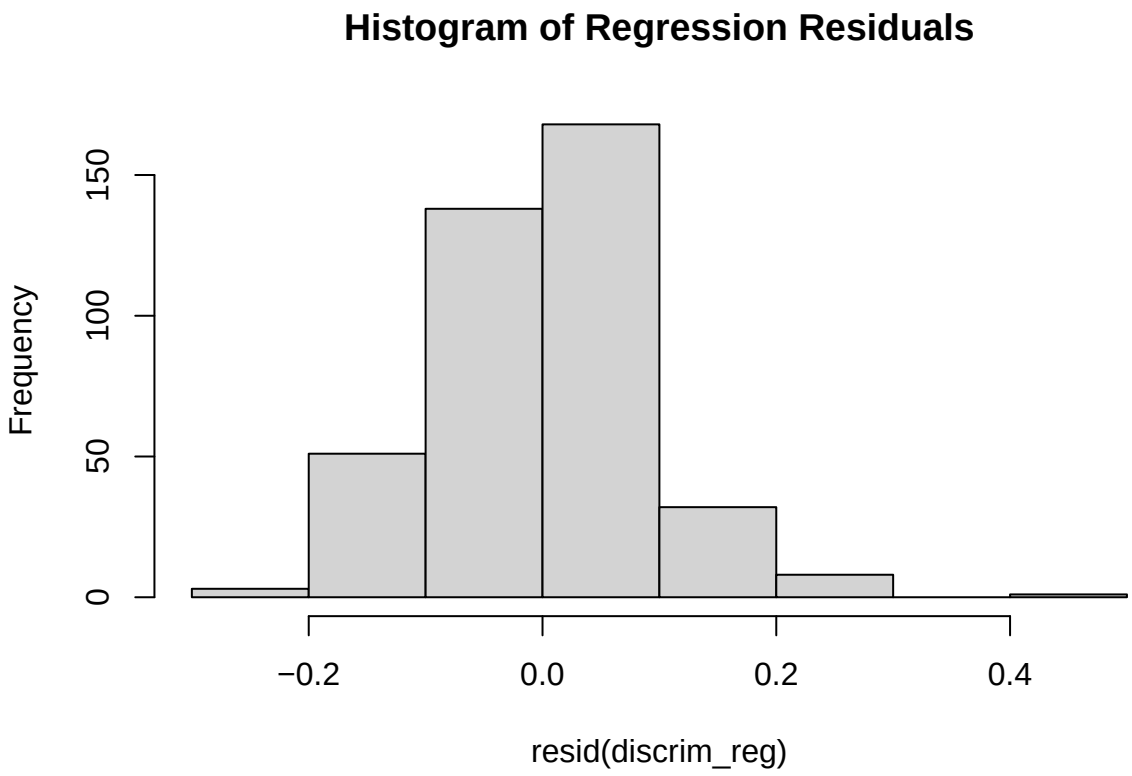
```
#Manual calculation for F statistic: $F = (R^2/k)/[(1-R^2)/(n-k-1)]$
```

f) The Q-Q plot seems to follow a normal distribution, except at the two ends between -3 and -2 and 2 and 3 which would show variation in the tails of the data set. The histogram of the regression residuals should be normally distributed; the histogram is unimodal, but is right skewed. However, since the skew is not extreme and the sample size is reasonably large enough, I can approximate that the shape of the distribution is normal through relying on asymptotics to fulfill the Classical Linear Model assumptions.

```
ggqqplot(resid(discrim_reg),
      color = "mediumturquoise",
      main = "Q-Q Plot of PSoda Regression Residuals",
      xlab = "Sample Quartiles", ylab ="Sample Quartiles",
      font.main = c(14,"bold.italic", "black"),
      font.x = c(14, "bold", "black"),
      font.y = c(14, "bold", "black"))
```




```
hist(resid(discrim_reg), main = "Histogram of Regression Residuals")
```



g) Using all PA and NJ zip-codes, the average proportion of the population that is black is 0.1134864 and the average median income is \$47,053.78. Based on our fitted model, the estimated average price of soda for a zip-code using the average values of *prpblck* and *income* is \$1.04.

```
mean(discrim$prpblck, na.rm = T)
```

```
## [1] 0.1134864
```

```
mean(discrim$income, na.rm = T)
```

```
## [1] 47053.78
```

$$\widehat{psod} = 0.9563 + 0.1150(0.1134864) + 1.6E^{-6}(47053.78) = 1.04curren$$

```
mean(discrim$prpblck,na.rm = T) # mean of prpblck | I use na.rm = T to remove missing values from the c
```

```
## [1] 0.1134864
```

```
mean(discrim$income,na.rm = T) # mean of income
```

```
## [1] 47053.78
```

```
newdf <- data.frame(prpblck = mean(discrim$prpblck,na.rm = T), # create data frame for prediction
                    income = mean(discrim$income,na.rm = T))
predict(discrim_reg,newdata = newdf) # predict psoda for values in the newdf --> 0.9563 + 0.1150*0.1134
```

```
##        1
## 1.044781
```

h) Based on this fitted model, there is evidence higher fast-food restaurants prices are associated with areas with a larger concentration of blacks. However, the low R^2, suggesting the total variation in price is explained by variables which I have not accounted in the model.