ECOM20001 Econometrics 1 Assignment 1

Name	Student ID
Yixin (Chelsea) Wu	1178379
Lang (Ron) Chen	1181506

1.

The typical observation for beer consumption is 1.399 Gallons/Capita, 0.469 USD/Gallon for beer tax and 35.068 USD/Pack for cigarette tax. This means the average beer consumption for beer across 47 US states from 1981 to 2007 is 1.399 Gallons/Capita, and so forth for beer tax and cigarette tax (i.e. 0.469 USD/Gallon and 35.068 USD/Pack).

The minimum value of beer consumption is 0.738 Gallons/Capita and maximum is 2.359 Gallons/Capita, whilst 0.237 USD/Gallon and 1.164 USD/Gallon respectively for beer tax, and 2.00 USD/Pack and 246 USD/Pack for cigarette tax. It is noted that the maximum value of cigarette tax is significantly higher than its mean (7 times), indicating potential for existence of outliers. The range for cigarette tax is also significantly large as the maximum is 123 times the minimum value.

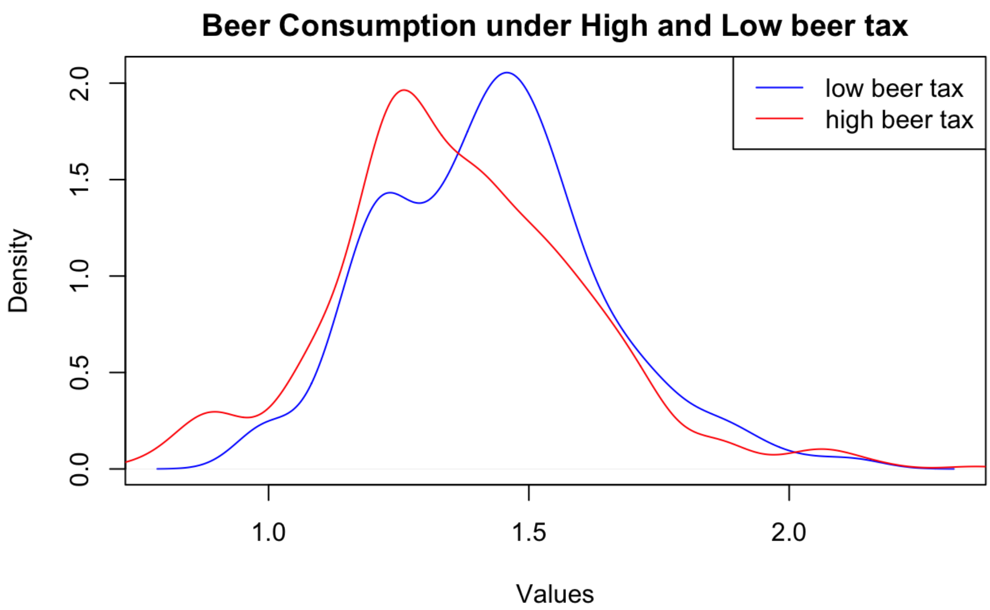
2.

The 95% CI for mean of beer consumption is (1.385784, 1.412432)

The 95% CI for mean of beer tax is (0.460936, 0.4760864)

The 95% CI for mean of cigarette tax is (33.04162, 37.09414)

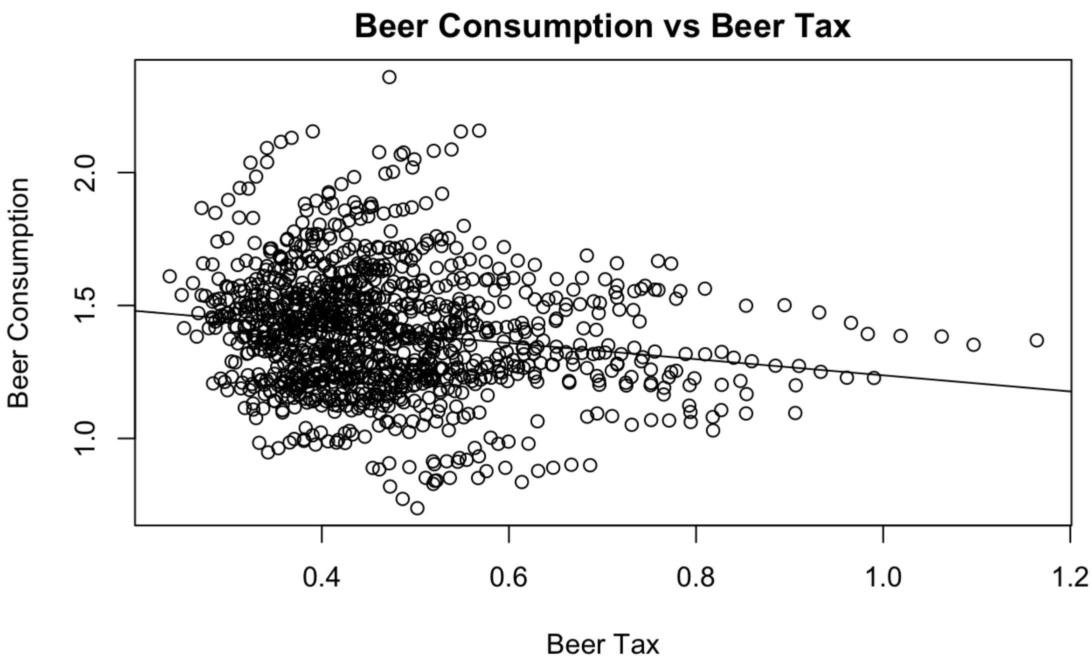
3.



The mode of the high beertax density sits to the left of that of the low beertax density, and the density line for low beer tax is higher than the high beer tax density line at almost all values above approximately beercons = 1.4; hence, the mean value of beer consumption for the low beer tax is almost certainly higher than that of high beertax. An intuitive explanation for both difference in density and mean using the supply demand model is that higher tax increases the price, which shifts the equilibrium market clearing quantity to be lower, and hence the mean consumption is lower and also more people will consume less alcohol (hence the shift in density).

4.
The difference between mean beer consumption at low tax and at high tax is -0.06098, with 95% confidence interval = (-0.08740573 -0.03456050) and the t-test (assuming non-constant variance in two samples as used welch test) returned a p value of $6.579*10^{-6}$. As this p-value is smaller than 0.05, there is enough evidence to reject the H0 of mean(beercons if hightax) = mean(beercons if lowtax).
This result suggests there is statistically significant evidence that having high beer tax affects the mean of beer consumption (i.e. the difference is by more than random chance).
By percentage, having high beer tax reduces the beer consumption by 4.27% compared to low beer tax.

5.



The correlation coefficient between beer consumption and beer tax is -0.1727, meaning a weak negative relationship. The linear regression line (for beer consumption against beer tax) has a negative gradient of -0.30387 (implying a 0.3087 gallon decrease in per capita consumption per USD increase in beer tax per gallon) – meaning increasing beer tax has a negative effect on beer consumption. This matches with the result from Question 4 which shows it is statistically significant that high beer tax affects the beer consumption level. As the negative correlation coefficient suggests that the higher the beer tax, the lower the beer consumption; this also matches the result from Question 3 that the mean of beer consumptions when beer tax is high is statistically significantly lower than mean of beer consumption when beer tax is low, hence negative relationship between the two. It can be also argued that the small percentage change based on the level of tax found in Question 4 matches with the low magnitude of the correlation coefficient in Question 5.

6.
Assuming homoscedasticity, the estimates are

	Model1	**Model2**
Beta0	1.542	1.443
Beta1	-0.304	-0.00125
SE	0.225	0.225

1 standard deviation of beertax = 0.1299822
1 standard deviation of cigtax = 34.77676

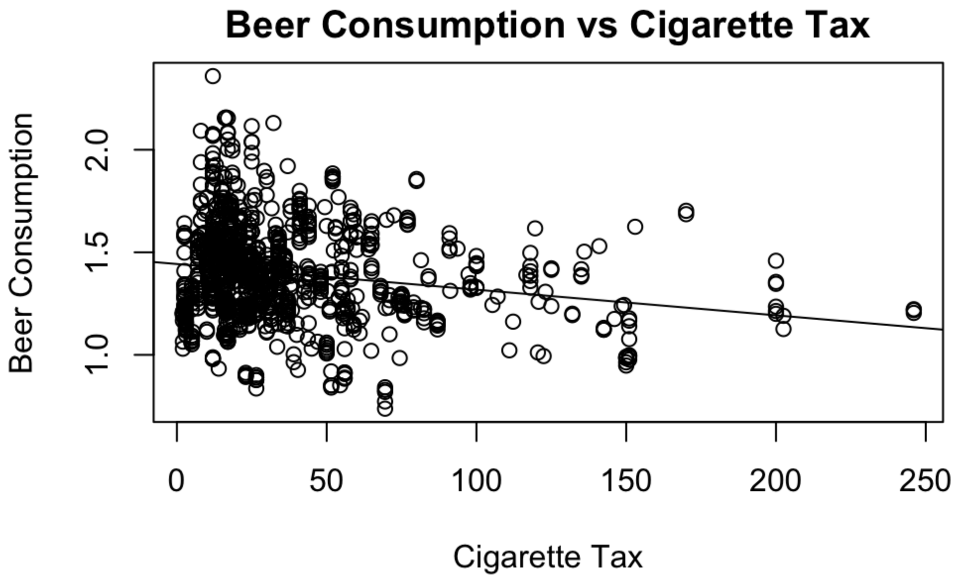
The magnitude of the predicted change in beer consumption corresponding to one standard deviation of beer tax is a 0.03949 decrease (-0.3038 * 0.12998 = - 0.03949); the magnitude of the predicted change in beer consumption corresponding to one standard deviation of cigarette tax is a 0.04351 decrease (-0.001251 * 34.77676 = - 0.04351).

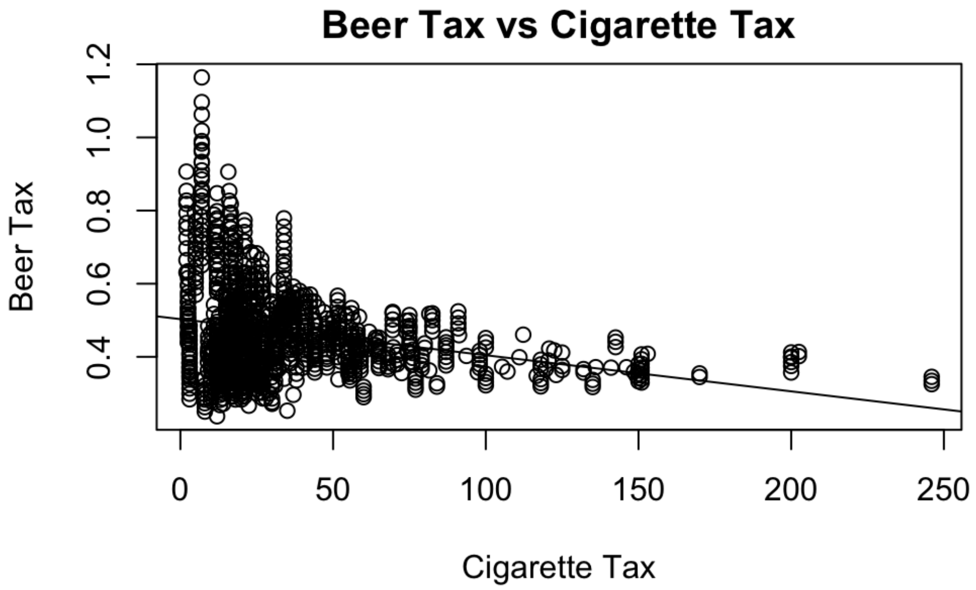
Conducting hypothesis tests using the Confidence interval (i.e. reject model 1 H0: beta1_model1 = 0 if CI does not include 0):

> *model 1 H0: beta1_model1 = 0; H1: beta1_model1 != 0*
> *model 2 H0: beta1_model2 = 0; H1: beta1_model2 != 0*

The confidence intervals for beta1_model1 is (-0.4048277, -0.2029123), for beta1_model2 is (-0.0016267171, -0.000874)
The confidence interval for beta1_model1 is 0.1299822*(-0.4048277, -0.2029123) = (-0.05262040, -0.02637498)
for beta1_model2 is 34.77676*(-0.0016267171, -0.000874) = (-0.05657195, -0.03041168).
As neither contains 0, can reject the two null hypothesis at 5% significance and conclude that the predicted changes are statistically significant different from 0. (Also, the p-values for the beta1's test is 3.651734e-09 and beta2's 7.175807e-11, both < 0.05 so can reject H0 that they are not significant)

7.





Q6 regression1's beta1 = -0.30387, and the regression undertaken in Q7's beta1 = -0.4217305.

Considering cigarette tax as a previously omitted variable in q6 regression1; as shown in the two graphs above, cigarette tax has a negative correlation with beer tax and also negative correlation with beer consumption, so with negative * negative = positive sign:

$\hat{\beta} = \beta + ((+)\rho * \frac{\sigma_u}{\sigma_x})$, and as the second term is positive as ρ is positive, and σ being standard deviation is always positive

$$\Rightarrow \hat{\beta} - \left(\rho * \frac{\sigma_u}{\sigma_x}\right) = \beta$$
$$\Rightarrow \hat{\beta} \geq \beta \Rightarrow \beta \leq \hat{\beta}$$

And this matches the fact that Q7's beta1 (more likely to be the 'true' beta1 as we have included the omitted variable of cigarette) is lower than Q6 regression 1's beta1.

The magnitude of beta1 of Q7 is 0.1178 larger than beta1 of Q6. This means previously the beta1 has accounted for the influence of cigtax on beer consumption as well. Now by removing the omitted variable bias, with cigtax also in the regression as a variable explaining for itself, a better representation of the magnitude for the effect of change in beer consumption given change in beer tax and holding the level of cigarette tax fixed is estimated to be – 0.422.

However, the magnitude of the change of 0.1178 does not fit with the formula for the omitted variable bias $\left(\rho * \frac{\sigma_u}{\sigma_x}\right)$, which calculates to be 0.00155. Given the significant difference, based on intuition it is highly unlikely that this gap is due to randomness.

8.
Q6 regression1 (all data) beta1 = -0.30387
Q9 regression1 (<= 1994) beta1 = -0.43945
Q9 regression2 (> 1994) beta1 = -0.06844

```
Call:
lm(formula = beercons ~ beertax, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.6513 -0.1544 -0.0063  0.1322  0.9610 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.54148    0.02504   61.56  < 2e-16 ***
beertax     -0.30387    0.05151   -5.90  4.8e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2253 on 1132 degrees of freedom
Multiple R-squared:  0.02983,	Adjusted R-squared:  0.02897 
F-statistic: 34.81 on 1 and 1132 DF,  p-value: 4.804e-09
```

```
Call:
lm(formula = data$beercons[data$year <= 1994] ~ data$beertax[data$year <= 
    1994])

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5407 -0.1422 -0.0205  0.1112  0.9166 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      1.64990    0.03122  52.840  < 2e-16 ***
data$beertax[data$year <= 1994] -0.43945    0.06201  -7.086 3.97e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2293 on 586 degrees of freedom
Multiple R-squared:  0.07893,	Adjusted R-squared:  0.07736 
F-statistic: 50.22 on 1 and 586 DF,  p-value: 3.97e-12
```

```
Call:
lm(formula = data$beercons[data$year > 1994] ~ data$beertax[data$year > 
    1994])

Residuals:
     Min       1Q   Median       3Q      Max 
-0.61540 -0.13087 -0.02054  0.15960  0.56921 

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     1.38735    0.04238  32.734   <2e-16 ***
data$beertax[data$year > 1994] -0.06844    0.09079  -0.754    0.451    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2093 on 544 degrees of freedom
Multiple R-squared:  0.001044,	Adjusted R-squared:  -0.0007927 
F-statistic: 0.5683 on 1 and 544 DF,  p-value: 0.4513
```

The value of the beertax's regression coefficient decreases (but increases in absolute value/magnitude of coefficient) for the regression taken pre and during 1994 compared to regression on full data, whilst increases (decreases in absolute value/magnitude of coefficient) for post 1994. As p value for the hypothesis that coefficient = 0 is $3.97*10^{-12} < 0.05$, hence we reject the null, and conclude that the coefficient for the pre and during 1994 beercons~beertax regression is statistically significant at 0.05 significance level. Whilst post 1994 has p value for the hypothesis that

coefficient = 0 is 0.451 > 0.05 so the coefficient is statistically insignificant. (The beertax coefficient for the full data regression is also significant as the p value for the hypothesis that coefficient = 0 is $4.8*10^{-9}$)
Plainly speaking, there is a statistically significant relationship between beer consumption and beertax before and during 1994, but this relationship is indistinguishable from noise/randomness after 1994 (i.e. the relationship has weakened to the point of statistical insignificance).

ECOM20001 Asmt 1 R Code

Setup

```
setwd('/Users/tg.chenny/Desktop/1. University/1. Undergraduate/23.
Econometrics 1/Asmt')
data = read.csv('./as1_beer.csv')
library(stargazer)

##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary
Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

q1

```
stargazer(data,
          summary.stat = c("n", "mean", "sd", "median", "min", "max"),
          type="text", title="Descriptive Statistics",
          out="sumstats1.txt")

##
## Descriptive Statistics
## =========================================================
## Statistic   N     Mean    St. Dev. Median  Min    Max
## ---------------------------------------------------------
## state     1,134  21.500    12.126   21.5    1      42
## year      1,134 1,994.000  7.792   1,994  1,981  2,007
## beercons  1,134   1.399    0.229    1.394  0.738  2.359
## beertax   1,134   0.469    0.130    0.440  0.237  1.164
## cigtax    1,134  35.068    34.777  23.692 2.000 246.000
## ---------------------------------------------------------
```

q2

```
# use t test, which generates the CI, to get the required output
(CI_beercons = t.test(data$beercons, conf.level=0.95))

##
##  One Sample t-test
##
## data:  data$beercons
## t = 206.02, df = 1133, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1.385784 1.412432
## sample estimates:
```

```
## mean of x
##  1.399108
```

```
(CI_beertax = t.test(data$beertax, conf.level=0.95))
```

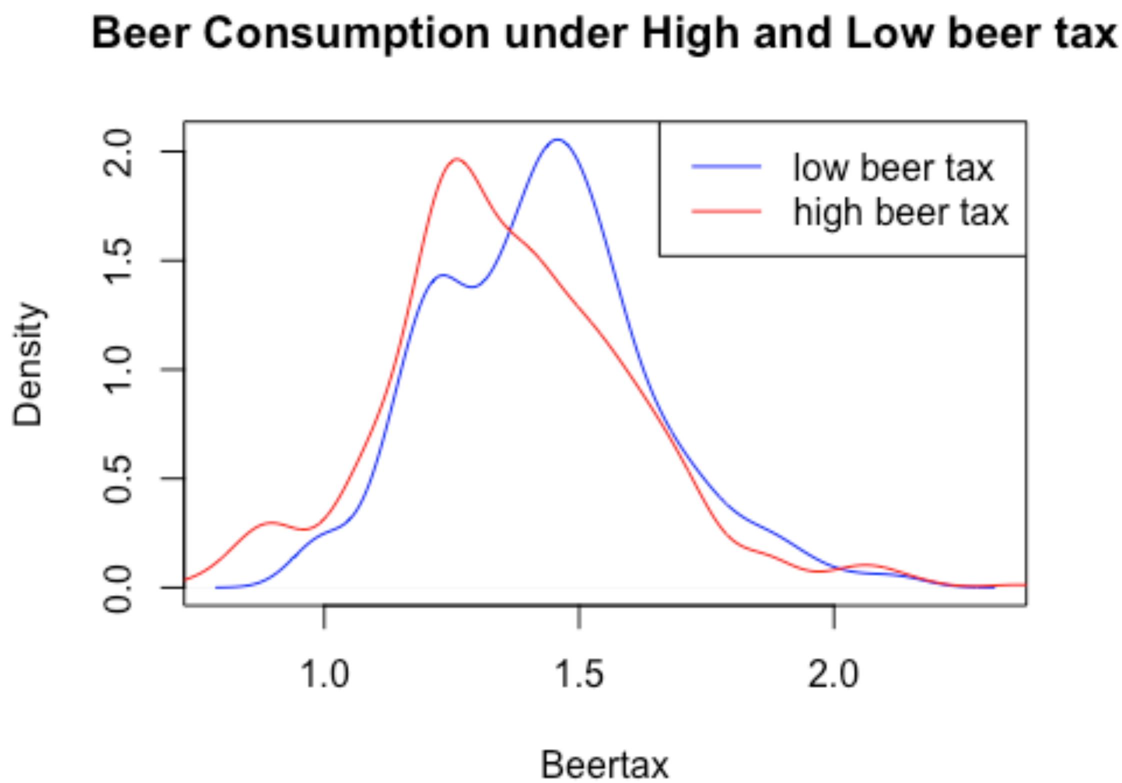
```
##
##  One Sample t-test
##
## data:  data$beertax
## t = 121.38, df = 1133, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.4609396 0.4760864
## sample estimates:
## mean of x
##  0.468513
```

```
(CI_cigtax = t.test(data$cigtax, conf.level=0.95))
```

```
##
##  One Sample t-test
##
## data:  data$cigtax
## t = 33.957, df = 1133, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  33.04162 37.09414
## sample estimates:
## mean of x
##  35.06788
```

q3

```
# create new dummy variable
data$hightax = as.integer(data$beertax >= median(data$beertax))

# create the density plot
plot(density(data$beercons[data$hightax==0]), col="blue", main='Beer
Consumption under High and Low beer tax', xlab = 'Beertax', ylab =
'Density',pch = 16)

lines(density(data$beercons[data$hightax==1]), col="red")

legend('topright', legend=c('low beer tax', 'high beer tax'), col = c('blue',
'red'), lty=1:1)
```




```
# get  mean of beer consumption when beertax is low
mean(data$beercons[data$hightax==0])

## [1] 1.429599

# get  mean of beer consumption when beertax is high
mean(data$beercons[data$hightax==1])

## [1] 1.368616
```

q4

```
# use T-test to conduct hypothesis testing for H0: means are equal; H1: means
are not equal
t.test(data$beercons[data$hightax == 1], data$beercons[data$hightax == 0])

##
##  Welch Two Sample t-test
##
## data:  data$beercons[data$hightax == 1] and data$beercons[data$hightax ==
0]
## t = -4.5285, df = 1115.7, p-value = 6.579e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.08740573 -0.03456050
## sample estimates:
## mean of x mean of y
##  1.368616  1.429599
```

```r
# compute %diff when going from lowtax to hightax
(meandiff = mean(data$beercons[data$hightax == 1]) -
mean(data$beercons[data$hightax == 0]))
```

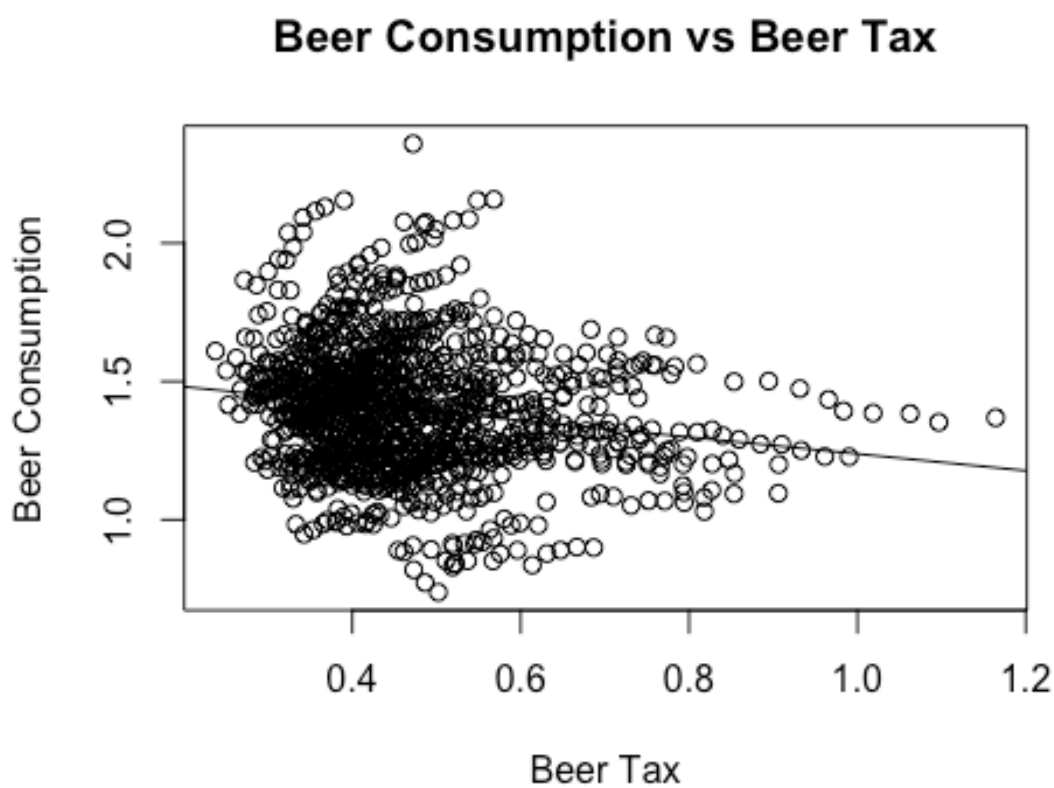
```
## [1] -0.06098312
```

```r
(diff_percentage = meandiff/mean(data$beercons[data$hightax == 0]))
```

```
## [1] -0.04265748
```

q5

```r
# create the scatter plot
plot(data$beertax, data$beercons, main='Beer Consumption vs Beer Tax', ylab =
'Beer Consumption', xlab = 'Beer Tax')
# create the linear regression and plot the line
model1 = lm(beercons ~ beertax, data = data)
abline(model1)
```



```r
# get the correlation between the two variables
cor(data$beertax, data$beercons)

## [1] -0.1727163
```

q6

```r
# fit linear model for beer tax and beer consumption
model1 = lm(beercons ~ beertax, data = data)
# fit linear model for cigarette tax and beer consumption
model2 = lm(beercons ~ cigtax, data = data)

summary(model1)

##
## Call:
## lm(formula = beercons ~ beertax, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6513 -0.1544 -0.0063  0.1322  0.9610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.54148    0.02504   61.56  < 2e-16 ***
## beertax     -0.30387    0.05151   -5.90  4.8e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2253 on 1132 degrees of freedom
## Multiple R-squared:  0.02983,    Adjusted R-squared:  0.02897
## F-statistic: 34.81 on 1 and 1132 DF,  p-value: 4.804e-09

summary(model2)

##
## Call:
## lm(formula = beercons ~ cigtax, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6185 -0.1583 -0.0014  0.1292  0.9311
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.4429630  0.0094745 152.300  < 2e-16 ***
## cigtax      -0.0012506  0.0001919  -6.518 1.07e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2246 on 1132 degrees of freedom
```

```
## Multiple R-squared:  0.03617,    Adjusted R-squared:  0.03532
## F-statistic: 42.48 on 1 and 1132 DF,  p-value: 1.074e-10
```

```
# get the one standard deviation effect for both models
SE_beertax = sd(data$beertax) # standard deviation of original dataset
(model1$coefficients[2] * SE_beertax)
```

```
##     beertax
## -0.03949782
```

```
SE_cigtax = sd(data$cigtax)
(model2$coefficients[2] * SE_cigtax)
```

```
##      cigtax
## -0.04349101
```

```
# get confidence intervals

beta1_model1 = -0.30387 # get beta
beta1SE_model1 = 0.05151 # get beta's SE

beta1_model2 = -0.0012506
beta1SE_model2 = 0.0001919

# create CI
(CI_beta1_model1 = beta1_model1 + c(-1, 1) * beta1SE_model1 * qnorm(0.975, 0,
1))
```

```
## [1] -0.4048277 -0.2029123
```

```
(CI_beta1_model2 = beta1_model2 + c(-1, 1) * beta1SE_model2 * qnorm(0.975, 0,
1))
```

```
## [1] -0.0016267171 -0.0008744829
```

```
# SE_data * CI
SE_beertax*CI_beta1_model1
```

```
## [1] -0.05262040 -0.02637498
```

```
SE_cigtax*CI_beta1_model2
```

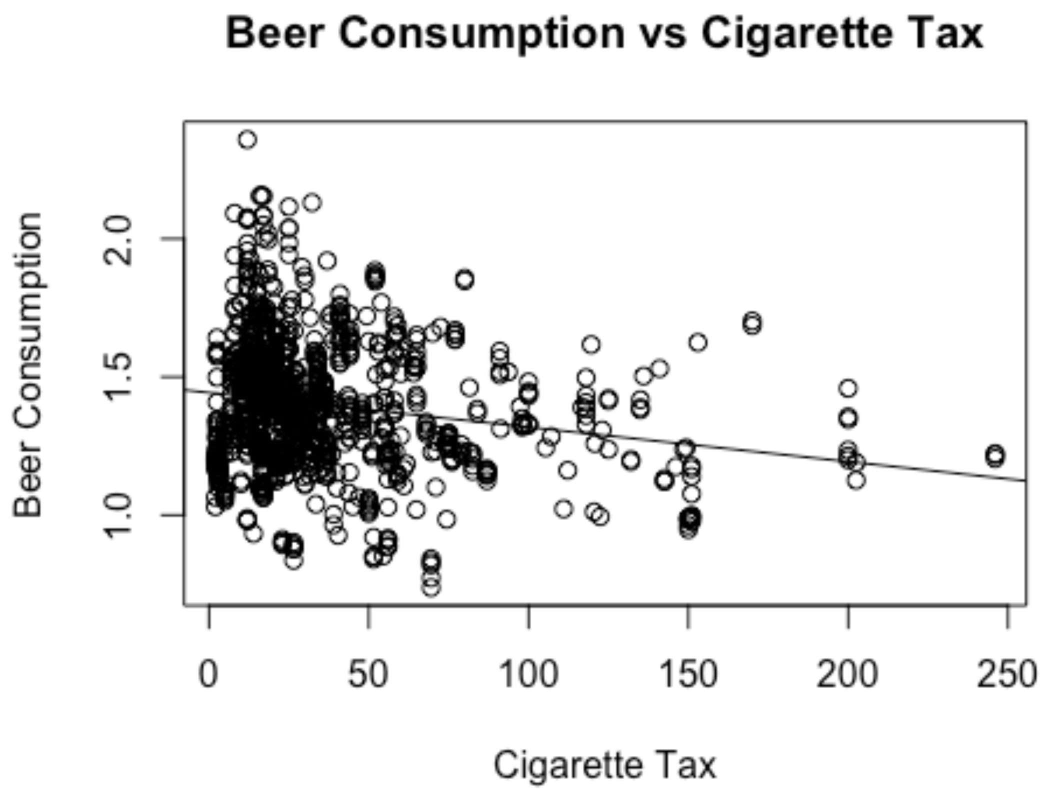
```
## [1] -0.05657195 -0.03041168
```

```
# conduct hypothesis test

# get t-values
t_model1 = abs(beta1_model1/beta1SE_model1)
t_model2 = abs(beta1_model2/beta1SE_model2)

# get p-values
(pvalue_1 = 2*pnorm(t_model1, 0, 1, lower.tail = FALSE))
```

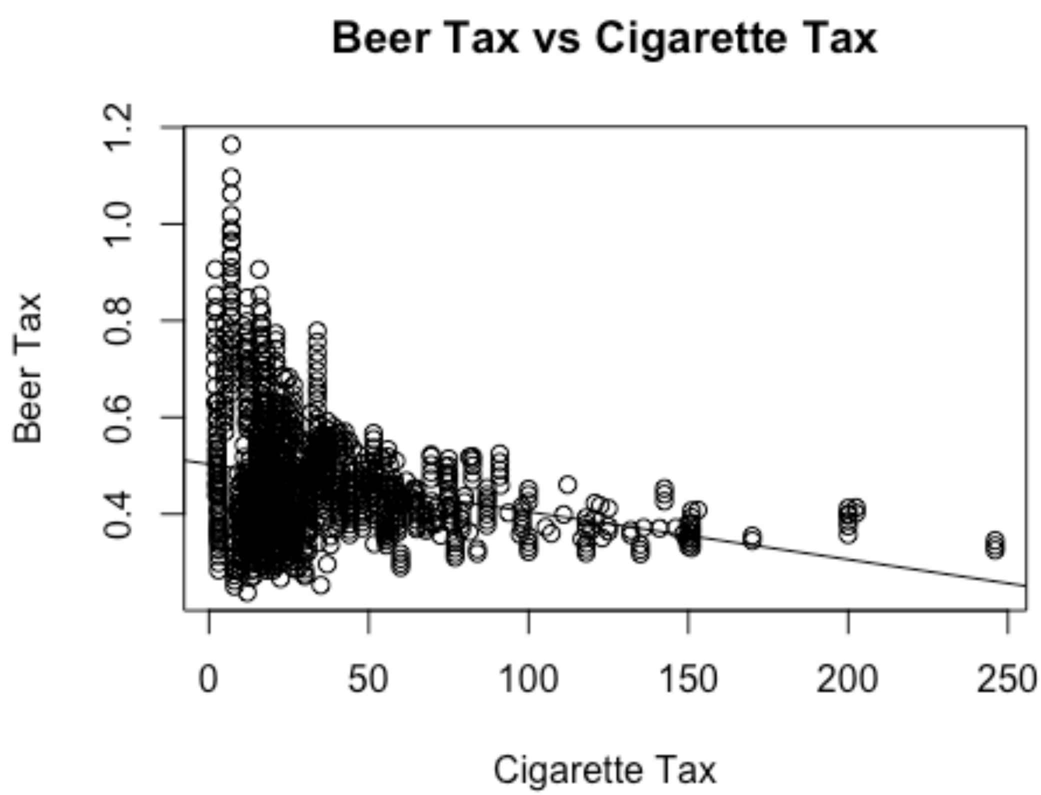
```
## [1] 3.651734e-09

(pvalue_2 = 2*pnorm(t_model2, 0, 1, lower.tail = FALSE))

## [1] 7.175807e-11
```

q7

```
# plot multiple linear regression
model3 = lm(beercons ~ beertax + cigtax, data=data)
summary(model3)

##
## Call:
## lm(formula = beercons ~ beertax + cigtax, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58996 -0.13875 -0.01115  0.12621  0.92302
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.6551555  0.0276219  59.922  < 2e-16 ***
## beertax     -0.4217305  0.0517553  -8.149 9.68e-16 ***
## cigtax      -0.0016671  0.0001934  -8.618  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2184 on 1131 degrees of freedom
## Multiple R-squared:  0.08961,    Adjusted R-squared:  0.088
## F-statistic: 55.67 on 2 and 1131 DF,  p-value: < 2.2e-16

# plot beer consumption and cigarette tax
plot(data$cigtax, data$beercons, main = 'Beer Consumption vs Cigarette Tax',
ylab = 'Beer Consumption', xlab = 'Cigarette Tax')
abline(model2)
```




```r
# plot beer tax and cigarette tax
tmp_model = lm(beertax ~ cigtax, data=data)
plot(data$cigtax, data$beertax, main = 'Beer Tax vs Cigarette Tax', xlab =
'Cigarette Tax', ylab = 'Beer Tax')
abline(tmp_model)
```




```
# calculating magnitude of omitted variable bias
(corr_ux = cor(model1$residuals, data$cigtax))

## [1] -0.2394145

(res_sd = sd(model1$residuals))

## [1] 0.2252493

(cigtax_sd = sd(data$cigtax))

## [1] 34.77676

(magnitude_omitted_var_bias = corr_ux * res_sd/cigtax_sd)

## [1] -0.001550689
```

q8

```
# do more linear models after splitting data
model4 = lm(data$beercons[data$year <= 1994] ~ data$beertax[data$year <=
1994])
model5 = lm(data$beercons[data$year > 1994] ~ data$beertax[data$year > 1994])

summary(model4)
```

```
##
## Call:
## lm(formula = data$beercons[data$year <= 1994] ~ data$beertax[data$year <=
##     1994])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5407 -0.1422 -0.0205  0.1112  0.9166
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        1.64990    0.03122  52.840  < 2e-16 ***
## data$beertax[data$year <= 1994] -0.43945    0.06201  -7.086 3.97e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2293 on 586 degrees of freedom
## Multiple R-squared:  0.07893,    Adjusted R-squared:  0.07736
## F-statistic: 50.22 on 1 and 586 DF,  p-value: 3.97e-12
```

```
summary(model5)
```

```
##
## Call:
## lm(formula = data$beercons[data$year > 1994] ~ data$beertax[data$year >
##     1994])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61540 -0.13087 -0.02054  0.15960  0.56921
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       1.38735    0.04238  32.734   <2e-16 ***
## data$beertax[data$year > 1994] -0.06844    0.09079  -0.754    0.451
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2093 on 544 degrees of freedom
## Multiple R-squared:  0.001044,   Adjusted R-squared:  -0.0007927
## F-statistic: 0.5683 on 1 and 544 DF,  p-value: 0.4513
```