Home assignment 4

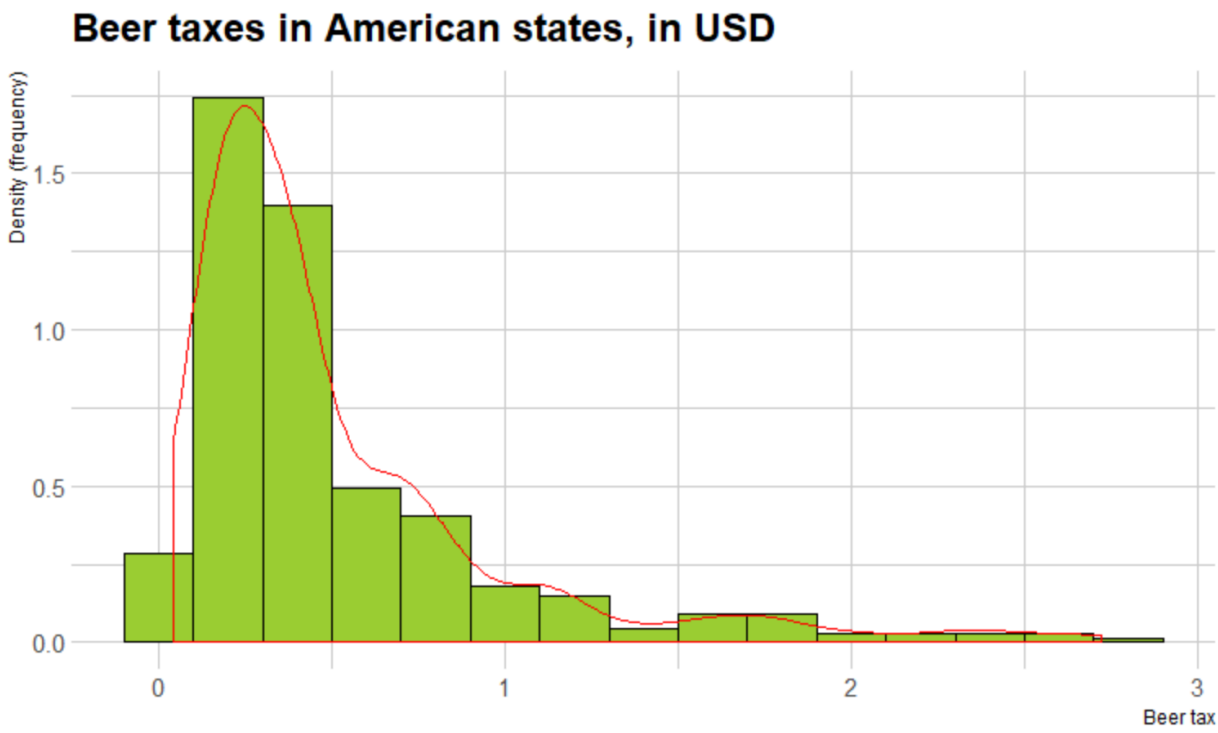
Description: Many governments use tax on alcoholic beverages as a policy instrument to improve some social indicators, such as vehicle fatality rate. The data contain observations from 1982 to 1988 on 48 American states: vehicle fatality rates per capita (mrall), state beer tax in 1988 USD (beertax), spirits consumption in gallons per capita (spircons), and other variables.

Problem 1:

Data contains 336 observations.

Problem 2:

Nice histogram is presented below; theme_ipsum from hrbrthemes package is used for all graphs.



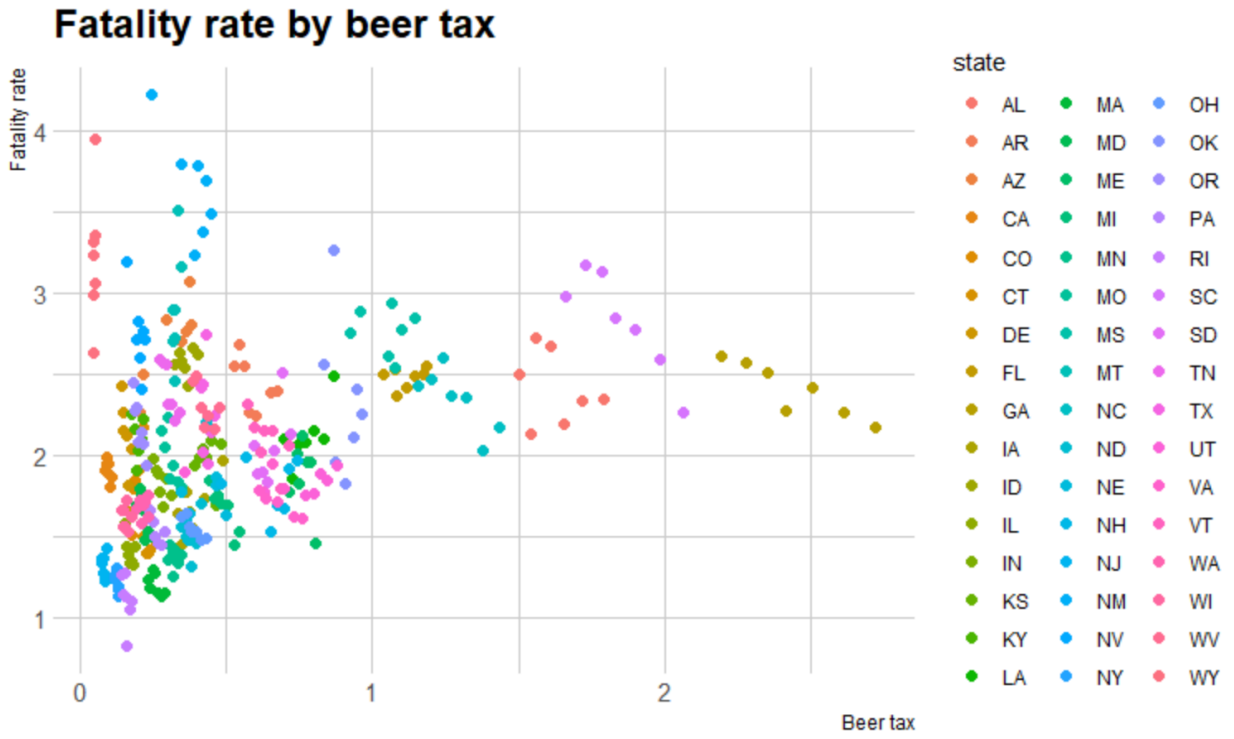
Problem 3:

Summary for fatalityrate; we can see that 2.04 is an average value

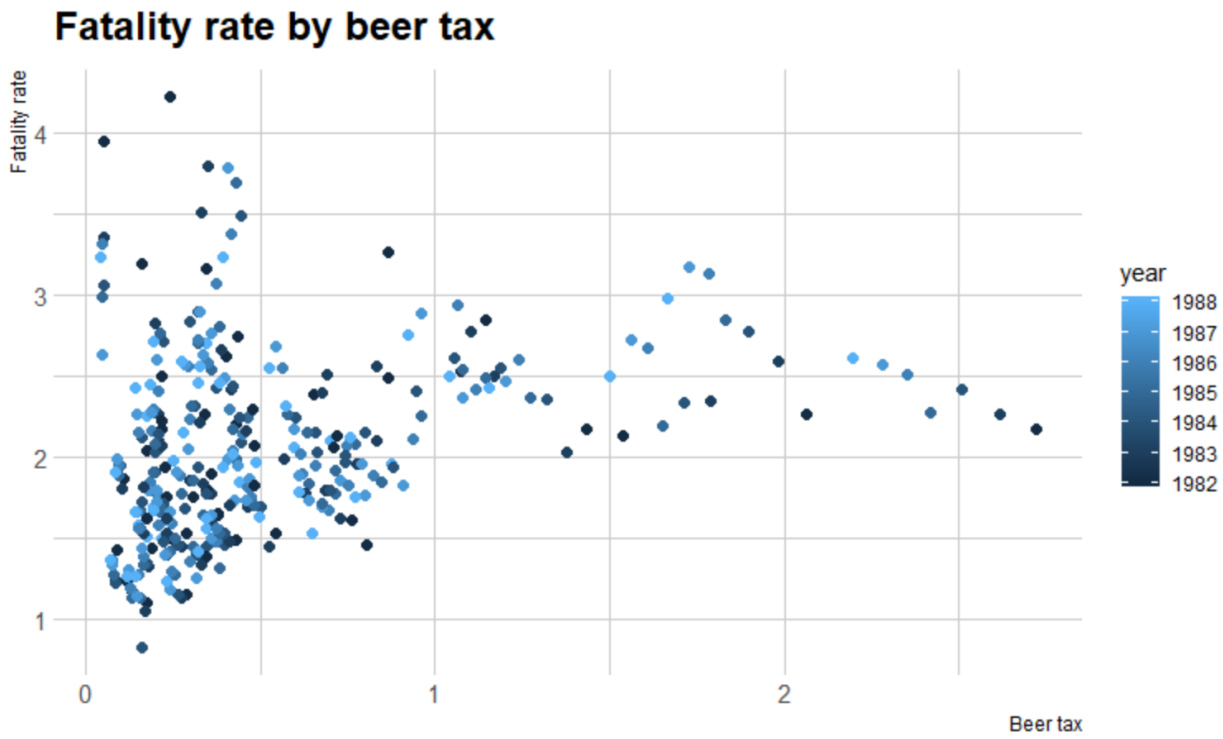
```
> summary(fatrat$fatalityrate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.8212  1.6237  1.9560  2.0404  2.4179  4.2178
```

Problem 4:

Plot with same states having the same color:



Plot with same color for same years; darker colors correspond to earlier years



Problem 5:

Using test for coefficients with heteroskedasticity-robust standard errors:

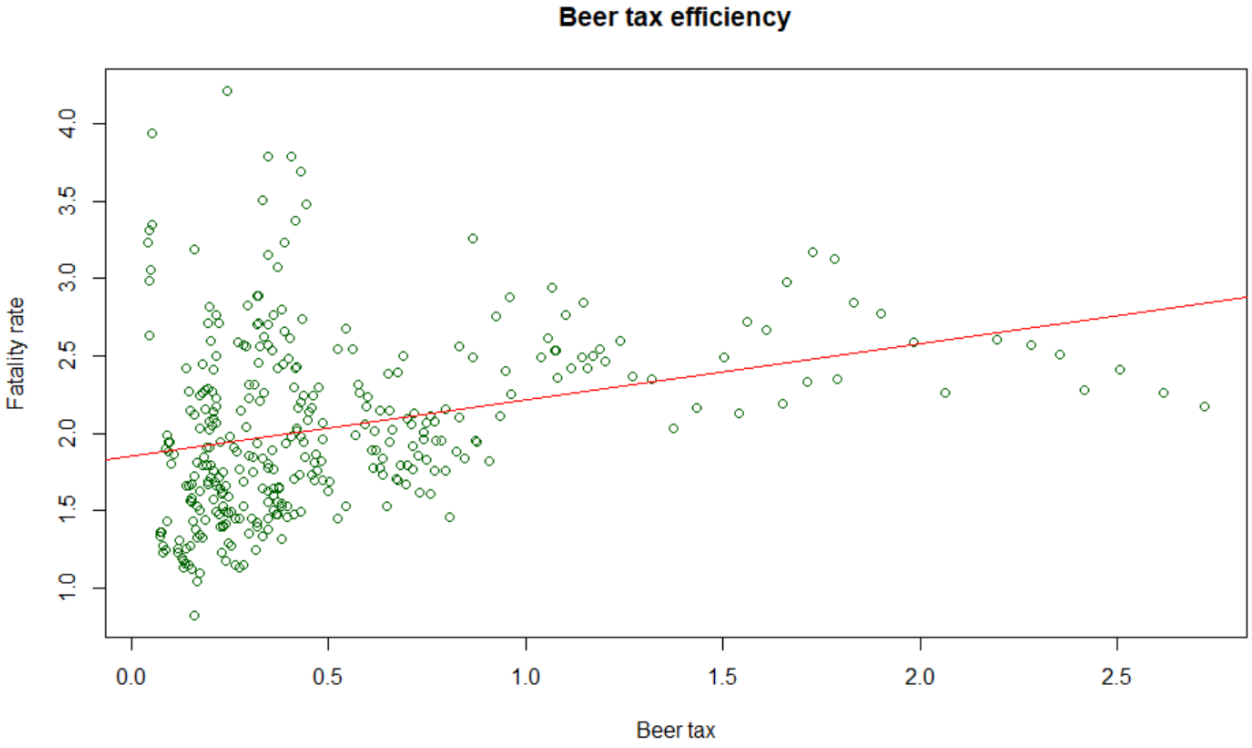
```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 1.853308   0.047402 39.0974 < 2.2e-16 ***
beertax     0.364605   0.054104  6.7389 7.008e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We get coefficients 1.853 for constant and 0.365 for beer tax (what means that linear regression is given by y=1.853+0.365x), standard errors for coefficients are 0.047 and 0.054.

This means that as beer tax grows, fatality rate grows. Therefore, it is not an effective measure. Coefficients are significant.

```
> summary(model, vcov. = vcovHC)

Call:
lm(formula = fatalityrate ~ beertax, data = fatrat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.09060 -0.37768 -0.09436  0.28548  2.27643

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.85331    0.04357  42.539  < 2e-16 ***
beertax      0.36461    0.06217   5.865 1.08e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5437 on 334 degrees of freedom
Multiple R-squared:  0.09336,   Adjusted R-squared:  0.09065
F-statistic: 34.39 on 1 and 334 DF,  p-value: 1.082e-08
```

Plot with regression line:



Problem 6:

As we can see from the plot above, our observations can vary a lot from regression line, so there could be some other important factors which can affect fatality rate. Probably beer tax increasing works different in different places on in different time of observation. So, we are going to use state and year variables.

Also, it could depend from spirits consumption in gallons per capita (spircons variable)

Problems 7:

Using test for coefficients with heteroskedasticity-robust standard errors:

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  1.910587   0.098174 19.4613 < 2.2e-16 ***
beertax      0.360576   0.053165  6.7822 5.412e-11 ***
spircons    -0.031483   0.052842 -0.5958    0.5517
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We get the following coefficients: 1.910 for constant term, 0.361 for beer tax term, and -0.031 for spirits consumption per capita term, standard errors are 0.098, 0.053, and 0.053, respectively. We can see that coefficient for beertax is not really different from the coefficient in previous model. As we can see, spircons is insignificant, and beertax is significant on any reasonable significance level.

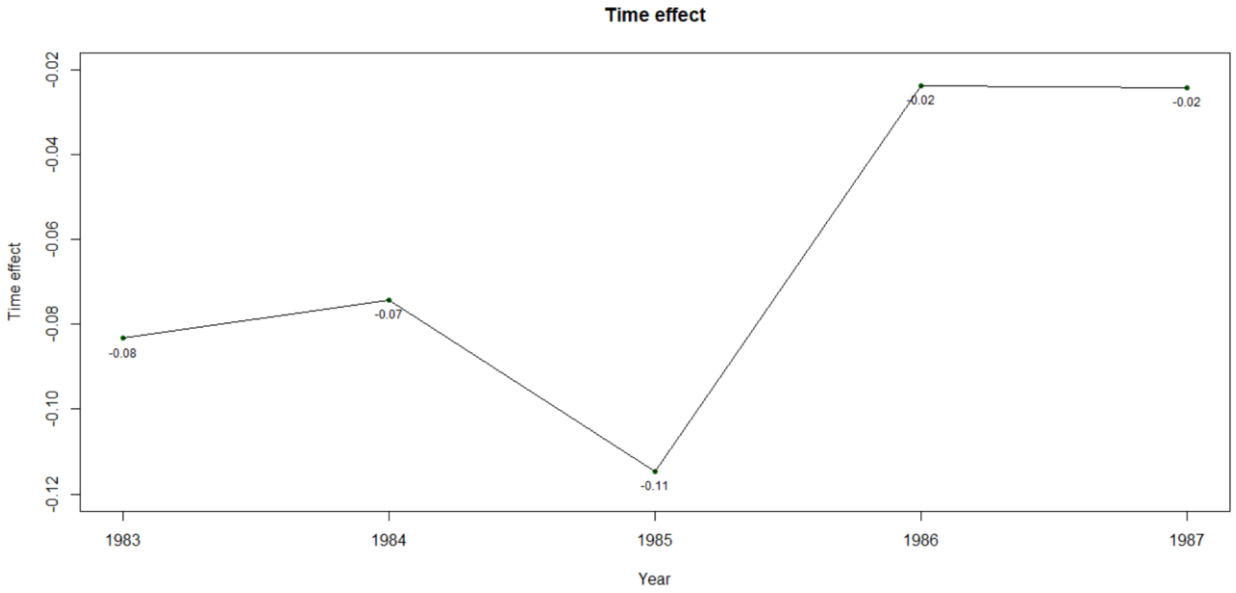
```
> summary(model2, vcov. = vcovHC)

Call:
lm(formula = fatalityrate ~ beertax + spircons, data = fatrat)

Residuals:
     Min       1Q   Median       3Q      Max
-1.08237 -0.37571 -0.09345  0.27568  2.26986

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.91059    0.09062  21.083  < 2e-16 ***
beertax      0.36058    0.06247   5.772 1.79e-08 ***
spircons    -0.03148    0.04367  -0.721    0.471
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5441 on 333 degrees of freedom
Multiple R-squared:  0.09478,   Adjusted R-squared:  0.08934
F-statistic: 17.43 on 2 and 333 DF,  p-value: 6.308e-08

> linearHypothesis(model2, c("beertax = 0", "spircons=0"), test = "F", vcov. = v
covHC)
Linear hypothesis test

Hypothesis:
beertax = 0
spircons = 0

Model 1: restricted model
Model 2: fatalityrate ~ beertax + spircons

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1    335
2    333  2 23.041 4.247e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As value of F-statistic corresponds to p-value ~ $4*10^{-10}$, this is very small p-value, what means that null hypothesis is rejected and jointly these two coefficients are significant.

This is not really counter-intuitive, as our first model explained data with only one variable. Now it still predicts data, but the second variable, spircons, doesn’t give as a lot of new information. Therefore, it is insignificant, as we can see, and it is better to try something else to make better model…

Problem 8:

Plot is presented below:



We can see, in 1985th year fatality rate changed, and also it changed in 1986th-1987th. So overall, fatality rate is changing over time if all other variables are unchanged.

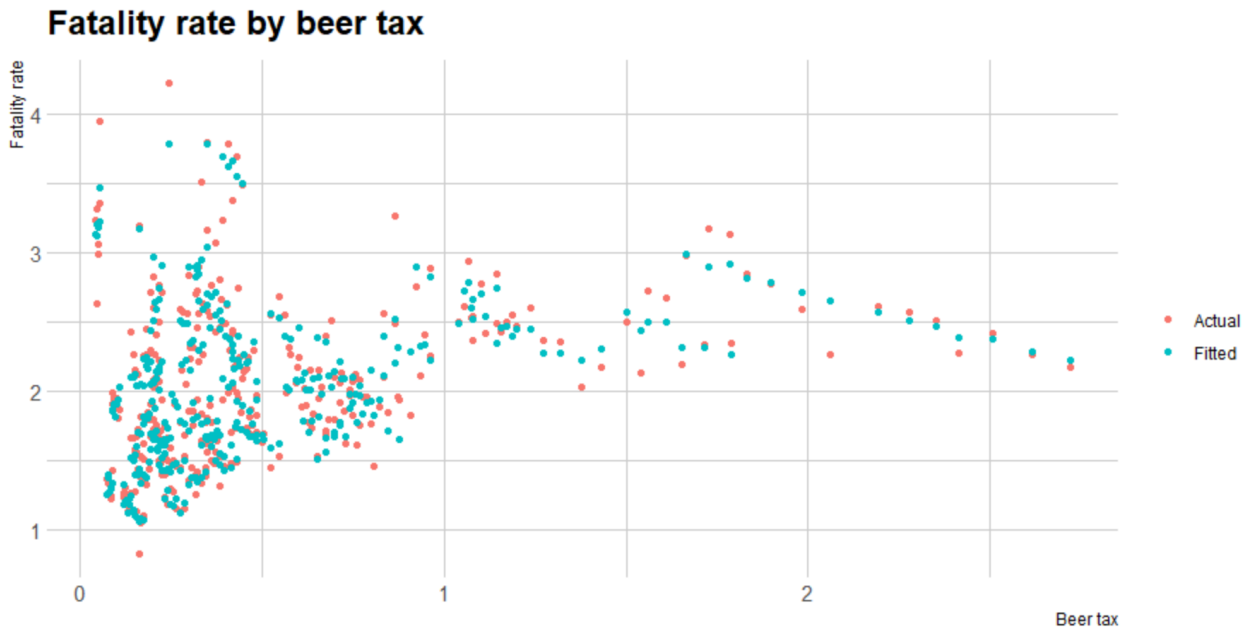
Problem 9:

Using the model, which depends from state; standard errors for all variables can be seen in the third column (shown below)

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  1.784603   0.465915  3.8303 0.0001580 ***
beertax     -0.492419   0.232975 -2.1136 0.0354328 *
spircons     1.027293   0.155602  6.6021 2.029e-10 ***
fyear1983   -0.033773   0.046605 -0.7247 0.4692608
fyear1984    0.017568   0.041128  0.4272 0.6695882
fyear1985    0.027379   0.044725  0.6122 0.5409301
fyear1986    0.241195   0.053824  4.4812 1.083e-05 ***
fyear1987    0.270177   0.056527  4.7796 2.840e-06 ***
fyear1988    0.317549   0.063537  4.9979 1.023e-06 ***
fstateAR    -0.301155   0.270225 -1.1145 0.2660380
fstateAZ    -0.960722   0.344882 -2.7857 0.0057067 **
fstateCA    -1.970099   0.389375 -5.0596 7.614e-07 ***
fstateCO    -1.987073   0.375247 -5.2954 2.405e-07 ***
fstateCT    -2.735360   0.381727 -7.1657 6.873e-12 ***
fstateDE    -2.320853   0.414569 -5.5982 5.163e-08 ***
fstateFL    -1.237308   0.221590 -5.5838 5.565e-08 ***
fstateGA    -0.219398   0.216955 -1.0113 0.3127642
fstateIA    -1.167215   0.325098 -3.5903 0.0003898 ***
fstateID    -0.387810   0.317732 -1.2206 0.2232806
fstateIL    -2.302070   0.368810 -6.2419 1.596e-09 ***
fstateIN    -1.322166   0.334808 -3.9490 9.935e-05 ***
fstateKS    -0.931905   0.304408 -3.0614 0.0024173 **
fstateKY    -0.979366   0.353984 -2.7667 0.0060397 **
fstateLA    -1.041859   0.237877 -4.3798 1.680e-05 ***
fstateMA    -2.978187   0.373679 -7.9699 4.045e-14 ***
fstateMD    -2.462567   0.379536 -6.4884 3.924e-10 ***
fstateME    -1.583404   0.256938 -6.1626 2.485e-09 ***
fstateMI    -1.823033   0.297229 -6.1334 2.922e-09 ***
fstateMN    -2.381218   0.339621 -7.0114 1.769e-11 ***
fstateMO    -1.131562   0.328804 -3.4415 0.0006669 ***
fstateMS    -0.018453   0.171755 -0.1074 0.9145198
fstateMT    -0.549006   0.349896 -1.5691 0.1177642
fstateNC    -0.456604   0.134978 -3.3828 0.0008198 ***
fstateND    -1.975748   0.331136 -5.9666 7.301e-09 ***
fstateNE    -1.450270   0.309908 -4.6797 4.480e-06 ***
fstateNH    -4.167868   0.517439 -8.0548 2.307e-14 ***
fstateNJ    -2.706109   0.398187 -6.7961 6.463e-11 ***
fstateNM     0.494270   0.319133  1.5488 0.1225614
fstateNV    -3.568907   0.610435 -5.8465 1.395e-08 ***
fstateNY    -2.618579   0.380393 -6.8839 3.823e-11 ***
fstateOH    -1.370728   0.311376 -4.4022 1.526e-05 ***
fstateOK    -0.446419   0.241546 -1.8482 0.0656321 .
fstateOR    -1.185086   0.358021 -3.3101 0.0010549 **
fstatePA    -1.545508   0.338688 -4.5632 7.546e-06 ***
fstateRI    -2.776574   0.379154 -7.3231 2.586e-12 ***
fstateSC    -0.047690   0.151428 -0.3149 0.7530473
fstateSD    -1.219823   0.275765 -4.4234 1.392e-05 ***
fstateTN    -0.671554   0.329385 -2.0388 0.0424078 *
fstateTX    -0.783076   0.311918 -2.5105 0.0126192 *
fstateUT    -0.597898   0.268115 -2.2300 0.0265398 *
fstateVA    -1.382593   0.249774 -5.5354 7.143e-08 ***
fstateVT    -1.676680   0.296053 -5.6634 3.676e-08 ***
fstateWA    -1.847476   0.353291 -5.2293 3.335e-07 ***
fstateWI    -2.208974   0.377642 -5.8494 1.374e-08 ***
fstateWV    -0.278524   0.309508 -0.8999 0.3689520
fstateWY    -0.543310   0.412732 -1.3164 0.1891248
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see now coefficient at beertax term is -0.492, negative! Before, in models 1-3 it was positive and significant, and in this model it is insignificant! This happens because fatality rates are well modelled if we know the states, and global effect of increasing beer tax can have different effect in different states, therefore it lacks predictive power.

Plot is shown below; we can see, now it is a lot better than a straight line in the model 1:



This can be overfitting (adjusted R-squared is approximately 0.91, quite high), but we are not going to think about this problem…

```
Residual standard error: 0.1688 on 280 degrees of freedom
Multiple R-squared:  0.9268,    Adjusted R-squared:  0.9124
F-statistic: 64.43 on 55 and 280 DF,  p-value: < 2.2e-16
```

Problem 10:

Resulting table with all four models is shown below:

Regression Results

	Dependent variable:			
	fatalityrate			
	Model 1	Model 2	Model 3	Model 4
Constant	1.853***	1.911***	1.950***	1.785***
	(0.047)	(0.098)	(0.141)	(0.466)
beertax	0.365***	0.361***	0.363***	-0.492**
	(0.054)	(0.053)	(0.054)	(0.233)
spircons		-0.031	-0.028	1.027***
		(0.053)	(0.054)	(0.156)
Time effects	No	No	Yes	Yes
State fixed effects	No	No	No	Yes
Observations	336	336	336	336
R^2	0.093	0.095	0.100	0.927
Adjusted R^2	0.091	0.089	0.078	0.912
Residual Std. Error	0.544 (df = 334)	0.544 (df = 333)	0.548 (df = 327)	0.169 (df = 280)
F Statistic	34.394*** (df = 1; 334)	17.432*** (df = 2; 333)	4.526*** (df = 8; 327)	64.427*** (df = 55; 280)
Note:				*p<0.1; **p<0.05; ***p<0.01

Problem 11:

Conclusion.

We investigated step-by-step several models, starting from easy ones, model 1 and model 2, which are linear and lacks generalizational power (R-squared is approximately 0.09, what means that 91% of our data variance still was unexplained).

In order to overcome this, we found more factors, that could have effect on fatality rates: year and state, where the event happened.

The addition of year didn't help a lot; we still haven't explained plenty of variance in our data. So, we also took into account the state and managed to build the model, which fits to the data quite well. Main coefficients (free term, beertax, and spircons) for that model are significant on 0.05 significance level.

We obtained different signs of coefficients for beertax term for different models; this happened because of the fact that first models are not describing our data well; for policymakers it is logical to believe model №4 (as it is the closest one to the real data), and think that true coefficient for beertax term is -0.49.

And in order to decrease fatality rates in all states, it is better to increase beer taxes.