STA 371H Midterm, Spring 2017

Instructions

1) Do not turn this page and begin the exam until instructed to do so.

2) The time limit for this exam is 75 minutes. If you are concise, you should have no trouble with the time limit. (If you ramble, you might.)

3) Write in ink, in a blue book. I will not grade anything written on the exam sheet itself, so you may use this as scratch paper. If you need more scratch paper, come ask me.

4) Turn in this exam paper along with your written exam.

5) This is a closed-book exam. You are allowed pens and a blue book on your desk, and nothing more.

6) Switch off all cell phones, mobile communication devices, etc. Do not merely turn them to silent or vibrate mode. If I catch you using your phone, I will collect your exam and give you a zero.

7) The exam has three parts, with point values labeled. You may take these point values as roughly proportional to appropriate length, but I strongly suggest that you spend at least 35-40 minutes on the data-analysis question (Question 3).

8) Good luck, and be safe over Spring Break.

This page left intentionally blank.

Question 1: Short answers (30 points)

A) Define the term "p-value."

B) Concisely describe the stepwise-selection algorithm. (It's OK to just list the steps.)

C) Under what general circumstances should we expect to need an interaction term in a statistical model?

Question 2: Essay (30 points)

We have met the concept of statistical adjustment in at least two different contexts. First, we met it in ordinary linear regression with a single predictor variable. Here, we wished to look at the y (response) variable after "controlling for" or "adjusting for" the x (predictor) variable. Second, we met this concept in multiple regression models, with more than one numerical predictor.

Briefly describe what is meant by "statistical adjustment" in each context, making sure to address/explain the claim we encountered in class that "statistical adjustment is just subtraction." Give one example of statistical adjustment, either from class/reading or your own imagination. Though it's not necessary to draw pictures, feel free to do so if you find that it helps you convey an idea.

Question 3: Interpreting data analysis (40 points)

In this question, you will look at data on traffic fatalities and seatbelt usage collected from all 50 US States (plus DC) from 1983–1997, a period in which many states were implementing laws requiring that all passengers in a car wear seatbelts.

The variables in the data set are:
- state: a categorical variable for the state in which the observation was taken.
- year: the year of the observation.
- YearsSince1983: same as year, but centered so that 1983 starts at 0, 1984 is 1, 1985 is 2, etc.
- fatalities: number of traffic fatalities per billion miles logged by drivers in that state in that year.
- seatbelt: seat-belt usage rate by drivers in that state in that year, as self-reported by respondents to a survey in the state. This is expressed as percent out of 100, so 100 means everyone uses their seatbelts.
- speed65: whether there is a 65 MPH (or lower) speed limit on state highways. States with "no" for this variable had highway speed limits of higher than 65 MPH.
- drinkage: whether there was a minimum drinking age of 21 in the state during that year.

Over the next several pages, the results of several statistical plots and analyses are shown. Use these results to decide whether the following statements are true, false, or undecidable/ambiguous in light of the evidence provided. If true, cite supporting evidence. If false, propose a correction and cite supporting evidence. If undecidable/ambiguous, *make your best guess in light of what you do know* and explain what evidence you'd like to see in order to decide the question to your satisfaction. (Note: all quoted numbers are rounded off a bit; I'm not trying to trick you here by making subtle rounding errors that invalidate an otherwise true statement.)

The statements: true or false?

A) Holding other relevant factors equal, increasing a state's seatbelt-usage rate by 1% seems to decrease traffic fatalities by about 0.11 deaths per billion miles driven (on average across all states), with a 95% confidence interval of about (-0.14, -0.09) for the partial slope on seatbelt usage rate.

B) Holding other relevant factors equal, the average fatality rate across all states seemed to change by about -0.65 fatalities per billion miles per year from 1983-1997, with a 95% confidence interval of roughly (-0.8, -0.5).

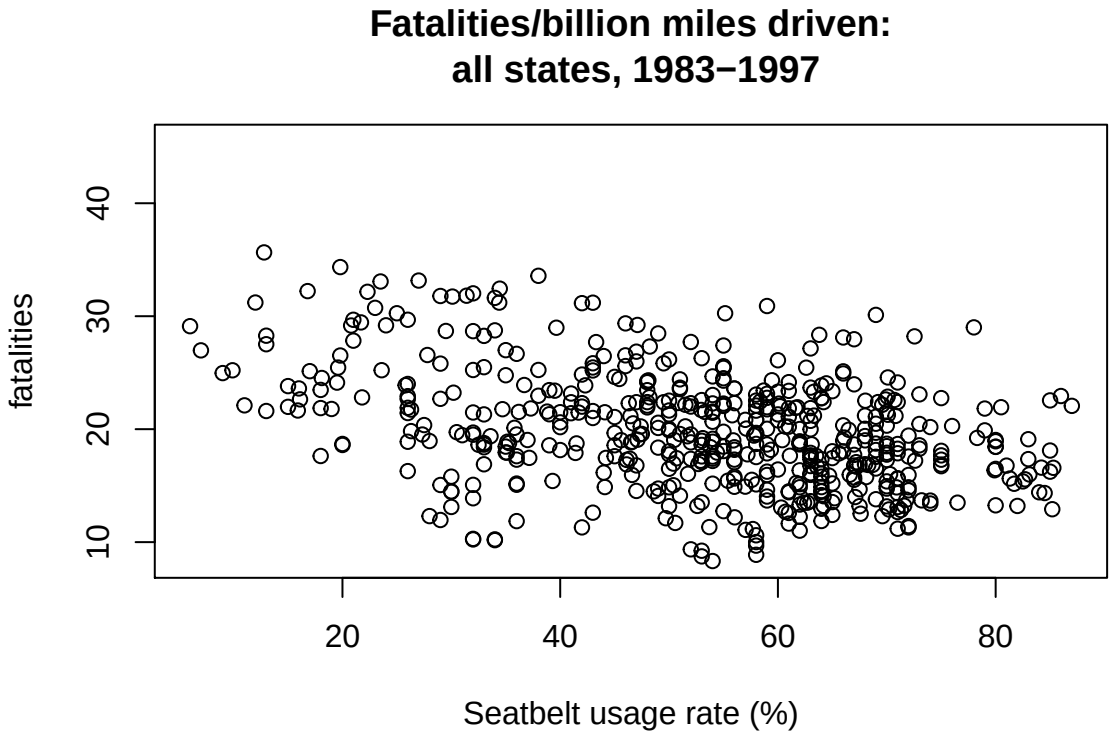
C) Even once we adjust for other relevant confounders, there are statistically significant differences among the states in their overall fatality rates.

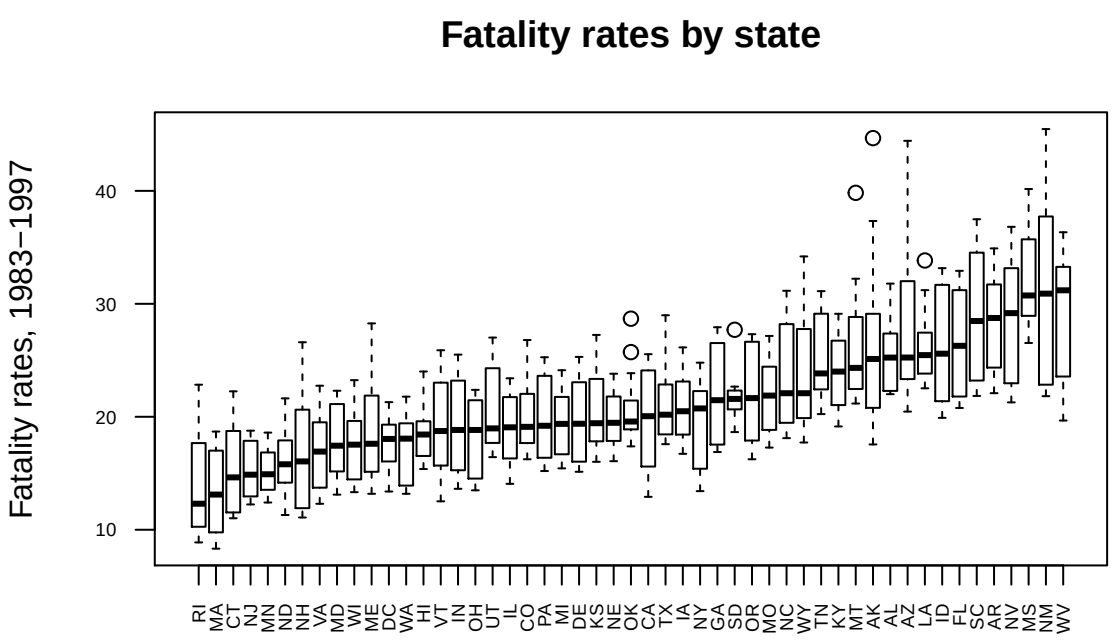
D) The relationship between seatbelt usage and fatality rates differs from state to state.

E) States with higher drinking ages (21) had lower traffic fatality rates, holding other relevant factors equal.

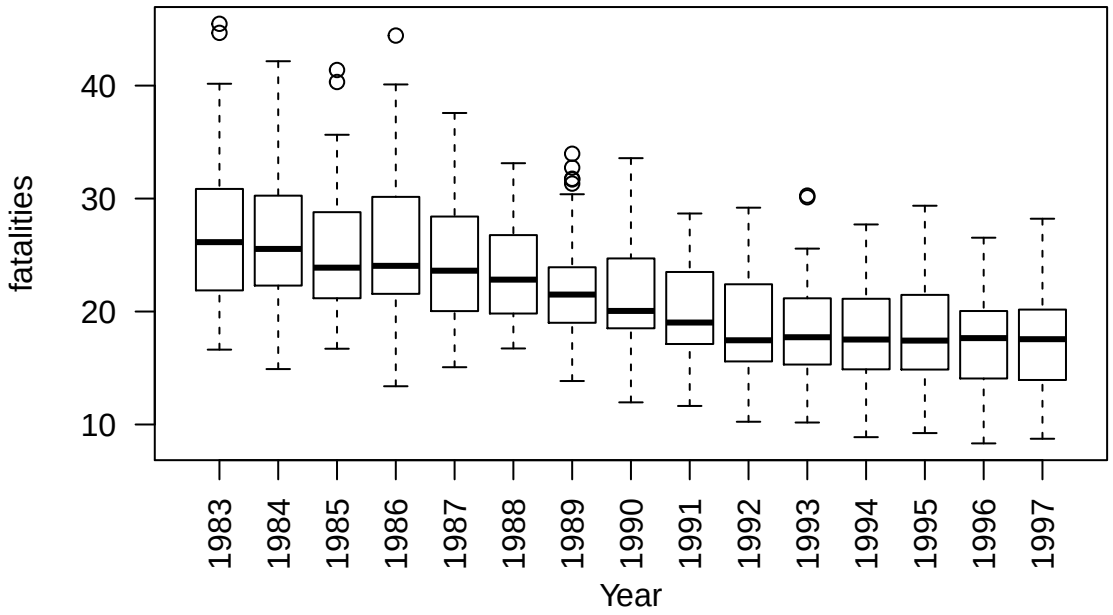
Plots

First, several exploratory plots of fatalities versus other variables were created.

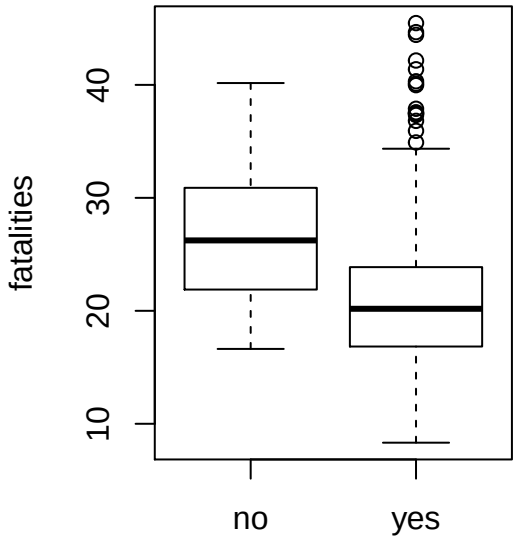




Fatality rate over time

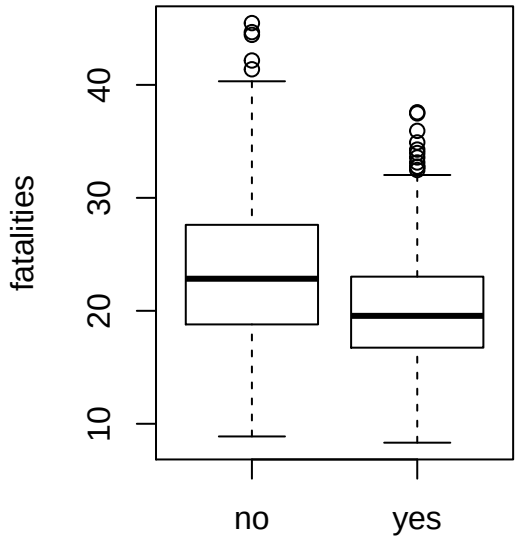


Fatality rates versus drinking−age laws



Does the state have a minimum drinking age of 21?

Fatality rates versus speed limits



Does the state have a highway speed limit of 65 MPH or lower?

Models and model outputs

Several statistical models were fit to the data.

Model 1

Model 1 has fatalities versus seatbelt usage rate, drinking-age laws, and speed-limit laws. This model was bootstrapped, and 95% confidence intervals were created for all model parameters.

```
lm1 = lm(fatalities ~ seatbelt + drinkage + speed65, data=seatbelts)
boot1 = do(5000)*{
    lm(fatalities ~ seatbelt + drinkage + speed65, data=resample(seatbelts))
}
confint(boot1)[1:4,]
```

```
##          name      lower       upper level     method   estimate
## 1   Intercept 25.9928788 30.57360941  0.95 percentile 28.2014495
## 2    seatbelt -0.1381498 -0.08833426  0.95 percentile -0.1138482
## 3 drinkageyes -6.1395614 -0.98344886  0.95 percentile -3.5015586
## 4  speed65yes  0.1573488  2.27715417  0.95 percentile  1.2096013
```

Model 2

Model 2 has all the variables that Model 1 does, but adds a variable for YearsSince1983. This model was bootstrapped to get 95% confidence intervals.

```
lm2 = lm(fatalities ~ seatbelt + drinkage + speed65 + YearsSince1983, data=seatbelts)
boot2 = do(5000)*{
    lm(fatalities ~ seatbelt + drinkage + speed65 + YearsSince1983,
       data=resample(seatbelts))
}
confint(boot2)[1:5,]
```

```
##             name       lower       upper level     method    estimate
## 1      Intercept 25.46371919 30.10920297  0.95 percentile 27.72869965
## 2       seatbelt -0.06721328 -0.00489516  0.95 percentile -0.03606921
## 3    drinkageyes -5.39324902 -0.24250553  0.95 percentile -2.76540069
## 4     speed65yes  1.85456908  3.84917340  0.95 percentile  2.89037126
## 5 YearsSince1983 -0.80243443 -0.49735204  0.95 percentile -0.65010482
```

Model 3

Model 3 has all the variables that Model 2 does, but adds the "state" variable. As before, this model was bootstrapped to get 95% confidence intervals. (Most of the state-level coefficients are omitted from the output below, to save space.)

```
lm3 = lm(fatalities ~ seatbelt + drinkage + speed65 + YearsSince1983 + state, data=seatbelts)
boot3 = do(5000)*{
    lm(fatalities ~ seatbelt + drinkage + speed65 + YearsSince1983 +
    state, data=resample(seatbelts))
}
confint(boot3)
```

```
##             name      lower      upper level     method    estimate
## 1      Intercept 16.1288183 19.7529450  0.95 percentile 17.96685327
## 2       seatbelt -0.0803017 -0.0266463  0.95 percentile -0.05328746
## 3    drinkageyes -0.6832013  1.7810261  0.95 percentile  0.52917718
## 4     speed65yes -1.3680711  0.1348818  0.95 percentile -0.63736206
## 5 YearsSince1983 -0.6171308 -0.4160902  0.95 percentile -0.51509799
## 6        stateMA -0.8864815  2.1495200  0.95 percentile  0.61349086
## 7        stateCT  0.8398869  4.0417942  0.95 percentile  2.50454070
```

In addition, an analysis of variance was run on Model 3:

```
simple_anova(lm3)
```

```
##                 Df      R2 R2_improve     sd sd_improve
## Intercept        1 0.00000            5.0295
## seatbelt         1 0.16218    0.16218 4.6078    0.42172
## drinkage         1 0.17130    0.00912 4.5868    0.02100
## speed65          1 0.17906    0.00776 4.5694    0.01740
## YearsSince1983   1 0.27823    0.09917 4.2884    0.28098
## state           50 0.88343    0.60520 1.8074    2.48106
## Residuals      501
```

Model 4

Finally, Model 4 has all the same variables that Model 3 does, but adds an interaction between the state and seatbelt variables. Again, this model was bootstrapped to get 95% confidence intervals. (Most of the state-level coefficients and interaction terms are omitted from the output below, to save space.)

```
lm4 = lm(fatalities ~ seatbelt + drinkage + speed65 + YearsSince1983
+ state + state:seatbelt, data=seatbelts)
boot4 = do(5000)*{
    lm(fatalities ~ seatbelt + drinkage + speed65 + YearsSince1983 +
    state + state:seatbelt, data=resample(seatbelts))
}
confint(boot4)
```

```
##             name       lower       upper level     method    estimate
## 1      Intercept 11.63160379 17.90555572  0.95 percentile 14.71028959
## 2       seatbelt -0.04483169  0.09533092  0.95 percentile  0.03102808
## 3    drinkageyes -1.37239983  1.71518753  0.95 percentile  0.13971964
## 4     speed65yes -1.03691366  0.67220862  0.95 percentile -0.24577886
## 5 YearsSince1983 -0.64485382 -0.41463363  0.95 percentile -0.53616899
## 6        stateMA  1.08405695  9.03642634  0.95 percentile  5.63425326
## 7        stateCT -9.84364135 19.65405318  0.95 percentile  5.97836308
```

ANOVA and permutation test

An analysis of variance of Model 4 was conducted:

```
simple_anova(lm4)
```

```
##                  Df      R2 R2_improve     sd sd_improve
## Intercept         1 0.00000            5.0295
## seatbelt          1 0.16218    0.16218 4.6078    0.42172
## drinkage          1 0.17130    0.00912 4.5868    0.02100
## speed65           1 0.17906    0.00776 4.5694    0.01740
## YearsSince1983    1 0.27823    0.09917 4.2884    0.28098
## state            50 0.88343    0.60520 1.8074    2.48106
## seatbelt:state   50 0.91319    0.02976 1.6438    0.16354
## Residuals       451
```

Finally, a permutation test of the state-seatbelt interaction term was conducted, using R^2 as a test statistic:

```
perm4 = do(5000)*{
    lm(fatalities ~ seatbelt + drinkage + speed65 + YearsSince1983 +
    state + shuffle(state):shuffle(seatbelt), data=seatbelts)
}
# Histogram of R-squared
hist(perm4$r.squared)
```