STA 371H Midterm, Spring 2016

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, either in a blue book or on separate sheets of paper. If you are not using a blue book, write your name and UT EID on each page, and staple all pages together. I will not grade anything written on the exam sheet itself, so you may use this as scratch paper.

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

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

6) Switch off all cell phones, mobile communication devices, iPods, and so forth. 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.

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

This page left intentionally blank.

Question 1: Short answers (20 points)

A) We have studied the "frequentist coverage property" in our discussion of confidence intervals. Briefly explain this property.

B) Suppose that we have data on a response y_i and a predictor x_i and that we want to estimate a power law of the form $\hat{y}_i = Kx_i^{\beta}$ for some constant K and exponent β. How we would use linear least squares to fit this nonlinear model? How do the parameters of the linear model you fit correspond to the parameters (K and β) of the power law?

Question 2: Essay (40 points)

Suppose you looked at data from a clinical trial in which 100 cancer patients were randomly assigned to a treatment group who received a new chemotherapy drug, and another 100 patients were assigned to a control group who received the best chemotherapy drug currently on the market. Each patient was assessed one year later to see whether his or her cancer had recurred. The trial gave the following results:

	Recurred	Did not recur
Treatment	40	60
Control	50	50

Thus it appears that patients in the treatment group experienced a lower rate of recurrence. Briefly describe what statistic/summary you might use to quantify this apparent relationship. (You don't have to actually calculate any numbers here; just explain how you would set up the calculation.)

Then explain in detail how you would use a permutation test to ascertain whether this apparent relationship between treatment and cancer recurrence could plausibly have been explained by chance. In your answer, make sure to include both the how (i.e. what steps we take to conduct the test) and the why (the role that each step plays in answering the overall question of interest).

Question 3: Interpreting data analysis (40 points)

In November 2015, the outbreak of Zika virus in Latin America led the Pan American Health Organization (PAHO) to issue an epidemiological alert regarding the spread of the virus and its putative links to fetal anomalies – specifically, microcephaly, a condition in which a baby's head is smaller than expected. Shortly afterwards, several countries in the region issued public-health advisories regarding the risk of Zika to women who were pregnant or who might become pregnant. These advisories ranged from cautions from the medical community about Zika-related microcephaly, to declared states of national emergency, to unprecedented public-health warnings urging women not to get pregnant while Zika remained a threat.

Brazil is considered to be the epicenter of the Zika outbreak. Below, you are shown data on patient visits to four family-planning clinics in Rio de Janeiro, Brazil. These are clinics where women go to obtain access to contraception and family-planning counseling, and the managers of the clinics want to understand the impact of Zika-related fears on women in the area. Thus the fundamental issue here is whether demand for access to these clinics increased in the wake of the PAHO alert in November 2015 due to fears over Zika—and if so, by how much.

The data begin on February 25th, 2015 and end one year later on February 24th, 2016. The variables in the data set are:
- *date*: the date of the observation.
- *clinic*: a categorical variable indicating which clinic the observation is from. The four clinics are labeled by the neighborhood of Rio de Janeiro in which they are located (Botafogo, Lagoa, Laranjeiras, and Santa Teresa).
- *visits*: the number of patient visits on the day in question (e.g. 25 means that 25 patients visited the clinic on that day
- *paho*: an indicator or dummy variable that marks which days came after the pregnancy alert issued by the Pan American Health Organization (PAHO) on November 17th, 2015. This variable is equal to 1 for all days since November 17th, and 0 for all days before then.

Over the next several pages, the results of several statistical analyses are shown. Use these results to decide whether the following statements are true, false, or undecidable in light of the evidence provided. If true, cite supporting evidence. If false, propose a correction and cite supporting evidence. If undecidable, *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.)

A) The available evidence suggests that the rate of patient visits to these four clinics after the PAHO alert on November 17th was the same, on average, as the rate of visits had been before November 17th.

B) We can say with 95% confidence that, after the PAHO pregnancy alert was issued on November 17th, patient visits increased somewhere between 5.8 and 7.6 visits per day, and that this increase was uniform across all four clinics.

C) There were statistically significant differences among the clinics in how their daily rate of patient visits changed after the PAHO alert.

D) After the PAHO alert was issued on November 17th, patient visits at the Laranjeiras clinic actually went down compared to what they had been before. The magnitude of change at the Laranjeiras clinic was -1.96 visits per day after the PAHO alert, with a 95% confidence interval of about -3.8 to -0.1.

E) Rates of patient visits at all four clinics seemed to go up after November 17th, and the clinic with the largest increase saw about 10 new patients per day after the alert.

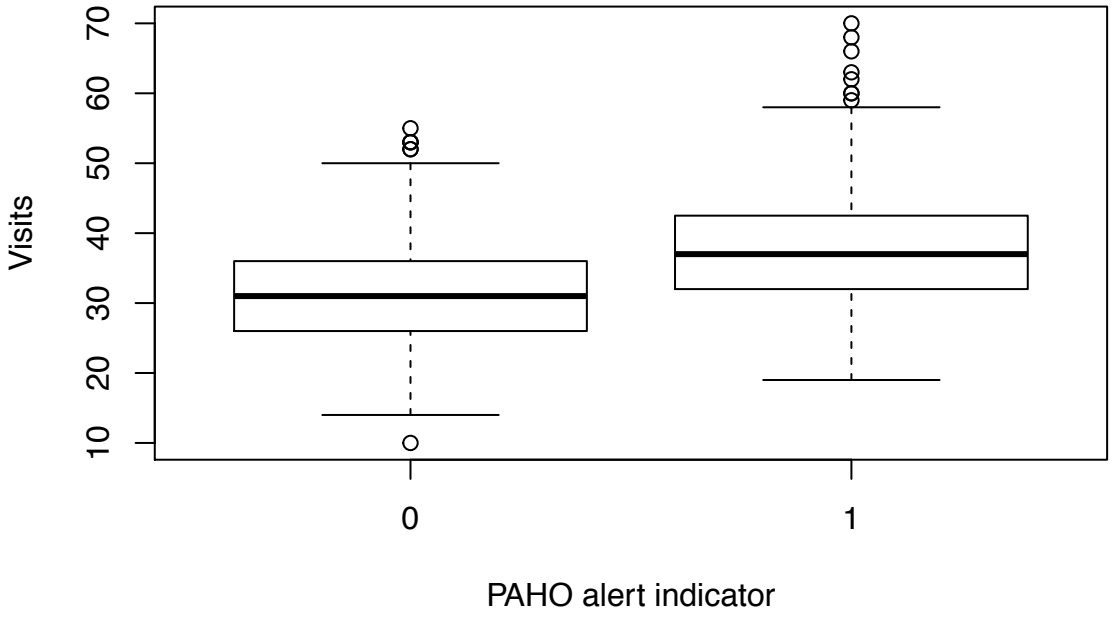
Description and plots

First, the data were loaded in and a boxplot of visits versus the PAHO indicator variable was made.

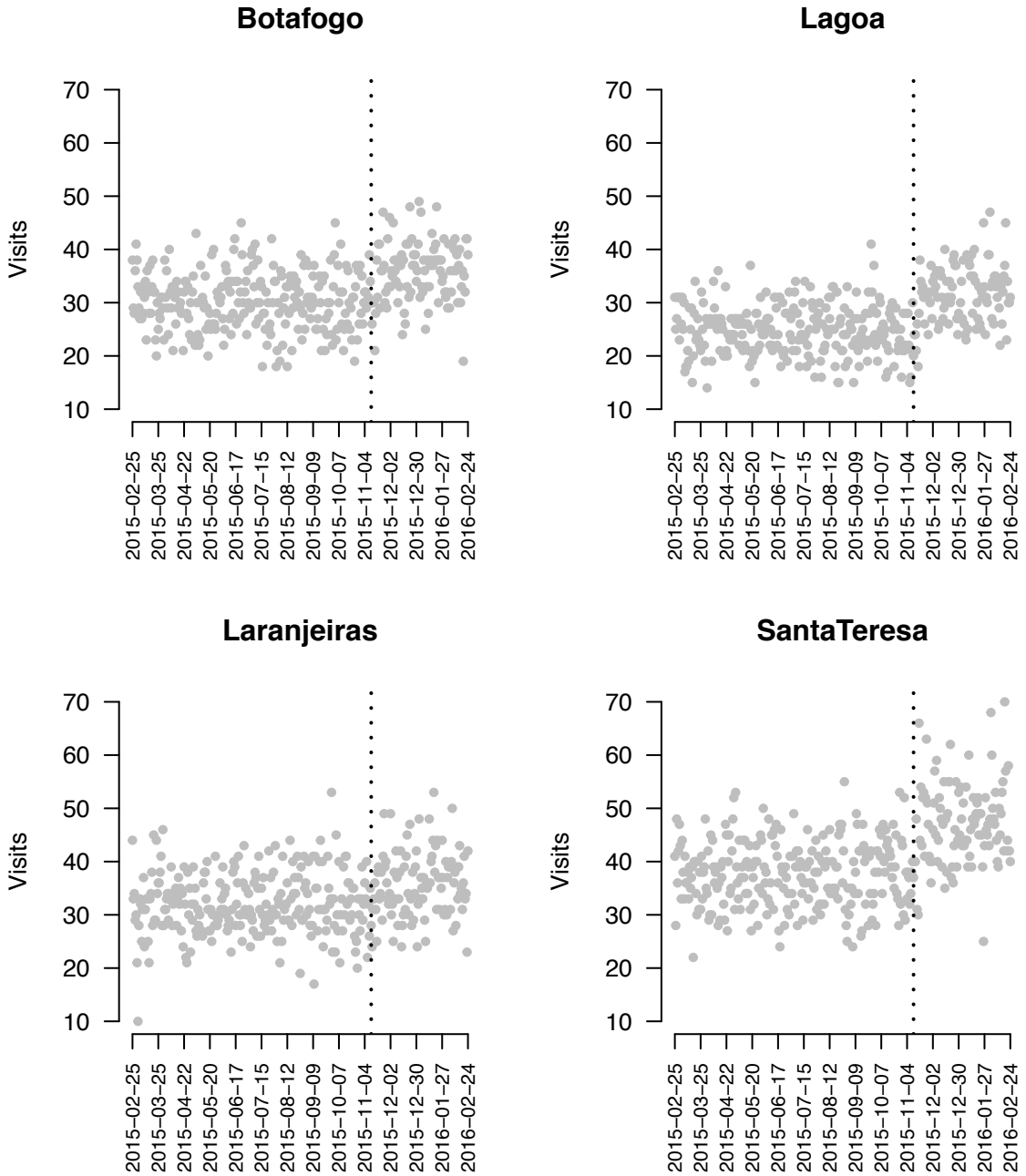
```
library(mosaic)
famplan = read.csv('famplan.csv')
head(famplan, 10)  # First 10 lines
```

```
##    visits      clinic       date paho
## 1      44 Laranjeiras 2015-02-25    0
## 2      41 SantaTeresa 2015-02-25    0
## 3      38    Botafogo 2015-02-25    0
## 4      31       Lagoa 2015-02-25    0
## 5      33 Laranjeiras 2015-02-26    0
## 6      28 SantaTeresa 2015-02-26    0
## 7      29    Botafogo 2015-02-26    0
## 8      25       Lagoa 2015-02-26    0
## 9      34 Laranjeiras 2015-02-27    0
## 10     48 SantaTeresa 2015-02-27    0
```

```
boxplot(visits ~ paho, data=famplan,
        xlab="PAHO alert indicator", ylab="Visits")
```



Here is a lattice plot showing the number of visits over time, stratified by clinic. In each panel, the vertical dotted line is on November 17th, 2015, the date of the pregnancy alert from the Pan American Health Organization. For all earlier dates, the "paho" variable is 0; for all dates thereafter, it is 1.



Model 1

In all of the regression models below, the outcome variable is the number of patient visits each day.

First, we have a model incorporating PAHO alert indicator as a predictor. The coefficients and the model's R-squared were found.

```
lm1 = lm(visits ~ paho, data=famplan)
coef(lm1)
```

```
## (Intercept)        paho
##   31.192453    6.697547
```

```
rsquared(lm1)
```

```
## [1] 0.1355188
```

Then the model was bootstrapped, and 95% confidence intervals were constructed for the model parameters.

```
boot1 = do(10000)*{
    lm(visits ~ paho, data=resample(famplan))
}
```

```
confint(boot1)
```

```
##        name      lower      upper level   estimate
## 1 Intercept 30.7643555 31.6169137  0.95 31.1924528
## 2      paho  5.7826214  7.6293006  0.95  6.6975472
## 3     sigma  7.2513495  7.8416003  0.95  7.5495573
## 4 r.squared  0.1044744  0.1686566  0.95  0.1355188
```

Note: these five columns represent the name of the variable; the upper and lower bounds for the confidence interval; the level (95%); and the estimate from the least-squares fit to the original data set. "Sigma" refers to the residual standard deviation of the model.

Model 2

Second, we have a model with main effects for the PAHO alert indicator and the clinic. The coefficients and the model's R-squared were found.

```
lm2 = lm(visits ~ paho + clinic, data=famplan)
coef(lm2)
```

```
##       (Intercept)              paho       clinicLagoa clinicLaranjeiras
##         30.000672          6.697547         -5.079452          1.265753
## clinicSantaTeresa
##          8.580822
```

```
rsquared(lm2)
```

```
## [1] 0.4975769
```

Then the model was bootstrapped, and 95% confidence intervals were constructed for the model parameters.

```
boot2 = do(10000)*{
    lm(visits ~ paho + clinic, data=resample(famplan))
}
```

```
confint(boot2)
```

```
##                 name      lower      upper level   estimate
## 1          Intercept 29.4155269 30.5969143  0.95 30.0006720
## 2               paho  5.9852976  7.3966393  0.95  6.6975472
## 3       clinicLagoa -5.8342446 -4.3293866  0.95 -5.0794521
## 4 clinicLaranjeiras  0.4255535  2.1075740  0.95  1.2657534
## 5 clinicSantaTeresa  7.6898268  9.4711340  0.95  8.5808219
## 6             sigma  5.5210677  5.9884081  0.95  5.7613705
## 7         r.squared  0.4608665  0.5345156  0.95  0.4975769
```

Note: the clinicLagoa, clinicLaranjeiras, and clinicSantaTeresa terms are the coefficients on the corresponding dummy variables for clinics.

Model 3

Third, we have a model with main effects for the PAHO alert indicator and the clinic, and an interaction between these two variables. The coefficients and the model's R-squared were found.

```
lm3 = lm(visits ~ paho + clinic + paho:clinic, data=famplan)
coef(lm3)
```

```
##            (Intercept)                    paho            clinicLagoa
##              30.215094                5.914906              -5.384906
##      clinicLaranjeiras      clinicSantaTeresa       paho:clinicLagoa
##               1.803774                7.490566                1.114906
## paho:clinicLaranjeiras paho:clinicSantaTeresa
##              -1.963774                3.979434
```

```
rsquared(lm3)
```

```
## [1] 0.5115384
```

Then the model was bootstrapped, and 95% confidence intervals were constructed for the model parameters.

```
boot3 = do(10000)*{
    lm(visits ~ paho + clinic + paho:clinic, data=resample(famplan))
}
```

```
confint(boot3)
```

```
##                      name      lower      upper level   estimate
## 1               Intercept 29.5785249 30.8634628  0.95 30.2150943
## 2                    paho  4.6473697  7.1384111  0.95  5.9149057
## 3             clinicLagoa -6.2275566 -4.5142763  0.95 -5.3849057
## 4       clinicLaranjeiras  0.8703592  2.7472384  0.95  1.8037736
## 5       clinicSantaTeresa  6.4826471  8.4651645  0.95  7.4905660
## 6        paho.clinicLagoa -0.5616584  2.8561177  0.95  1.1149057
## 7  paho.clinicLaranjeiras -3.8041333 -0.1107044  0.95 -1.9637736
## 8  paho.clinicSantaTeresa  1.9412579  6.1019851  0.95  3.9794340
## 9                  sigma  5.4480601  5.8987514  0.95  5.6866225
## 10             r.squared  0.4752604  0.5497544  0.95  0.5115384
```

Note: the coefficient labeled "paho.clinicLagoa" is the coefficient on the interaction between the PAHO indicator/dummy variable and the Lagao indicator/dummy variable. Similarly, "paho.clinicLaranjeiras" and "paho.clinicSantaTeresa" are the coefficients on their corresponding interaction terms.

ANOVA and permutation test

An analysis of variance was run on Model 3.

```
simple_anova(lm3)
```

```
##                Df      R2 R2_improve     sd sd_improve
## Intercept       1 0.00000            8.1170
## paho            1 0.13552    0.13552 7.5496    0.56743
## clinic          3 0.49758    0.36206 5.7614    1.78819
## paho:clinic     3 0.51154    0.01396 5.6866    0.07475
## Residuals    1452
```

These columns refer to the variable being added; the number of new parameters being added as a result (Df, which stands for degrees of freedom); and the improvement in model fit upon adding the variable, as measured by R-squared and the residual standard deviation (sd).

Finally, a permutation test was conducted to assess the statistical significance of the interaction between the PAHO alert and clinic, by comparing Model 3 to Model 2 using R-squared as a test statistic.

```
perm_test = do(10000)*{
    lm(visits ~ paho + clinic + shuffle(paho):shuffle(clinic), data=famplan)
}
```

The approximate sampling distribution of R-squared under the null hypothesis is shown below, together with the 95% quantile of this distribution (vertical grey line).

```
hist(perm_test$r.squared, 30)
alpha_level = 0.05
critical_value = qdata(perm_test$r.squared, 1-alpha_level)
abline(v = critical_value, col='grey', lwd=3)
```