**Final Exam** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Quantitative Methods, Hill, Spring 2014** name

**Question 1.**

Data: Housing Values in Suburbs of Boston

Variables:

crim per capita crime rate by suburb/town.

indus proportion of non-retail business acres per suburb/town.

chas Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

nox nitrogen oxides concentration (parts per 10 million).

rm average number of rooms per dwelling.

tax full-value property-tax rate per \$10,000.

ptratio pupil-teacher ratio by suburb/town.

medv median value of owner-occupied homes in \$1000s.

**Consider the following output from R software:**

Call:

lm(formula = medv ~ crim + indus + chas + nox + rm + tax + ptratio)

Residuals:

Min 1Q Median 3Q Max

-16.926 -3.136 -0.578 1.690 39.915

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.381974 4.519807 1.412 0.158574

crim -0.122889 0.036264 -3.389 0.000758 \*\*\*

indus -0.011725 0.066341 -0.177 0.859786

chas 3.928019 1.016628 3.864 0.000126 \*\*\*

nox -13.324960 3.655497 -3.645 0.000295 \*\*\*

rm 6.770995 0.405641 16.692 < 2e-16 \*\*\*

tax -0.001143 0.002614 -0.437 0.662087

ptratio -0.988400 0.141789 -6.971 0.00000000001 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.654 on 498 degrees of freedom

Multiple R-squared: 0.6273, Adjusted R-squared: 0.622

F-statistic: 119.7 on 7 and 498 DF, p-value: < 2.2e-16

**1.1** What type of hypothesis test is this? *(2 points)*

1. t-test
2. ANOVA
3. Chi Square
4. Simple regression
5. Multiple regression

**1.2** List the dependent variable(s) ***and*** its level(s) of measurement. *(2 points)*

**1.3** List the independent variable(s) ***and*** its level(s) of measurement. *(2 points)*

**1.4** Report the technical details (equation). *(3 points)*

**1.5** Interpret, describe, and comment on the results. Assume the reader will *not* have access to the R output; your answers to this question and to 1.4 should convey the relevant results. *(6 points)*

**1.6** If the pupil-teacher ratio increased by five points in a Boston suburb/town, what is the change in the estimated median home value for that suburb/town? *(5 points)*

**Evaluate the following correlation matrix (Pearson’s R).**

medv crim indus chas nox rm tax ptratio

medv 1.0000000 -0.38830461 -0.48372516 0.17526018 -0.42732077 0.69535995 -0.46853593 -0.5077867

crim -0.3883046 1.00000000 0.40658341 -0.05589158 0.42097171 -0.21924670 0.58276431 0.2899456

indus -0.4837252 0.40658341 1.00000000 0.06293803 0.76365145 -0.39167585 0.72076018 0.3832476

chas 0.1752602 -0.05589158 0.06293803 1.00000000 0.09120281 0.09125123 -0.03558652 -0.1215152

nox -0.4273208 0.42097171 0.76365145 0.09120281 1.00000000 -0.30218819 0.66802320 0.1889327

rm 0.6953599 -0.21924670 -0.39167585 0.09125123 -0.30218819 1.00000000 -0.29204783 -0.3555015

tax -0.4685359 0.58276431 0.72076018 -0.03558652 0.66802320 -0.29204783 1.00000000 0.4608530

ptratio -0.5077867 0.28994558 0.38324756 -0.12151517 0.18893268 -0.35550149 0.46085304 1.0000000

**1.7** Are there any problems with circularity and/or multicollinearity? *(3 points)*

1. There are no problems with either.
2. The assumption of no circularity has been violated, but there is no evidence of multicollinearity.
3. There is no evidence of circularity, but the assumption of no multicollinearity has been violated.
4. Both assumptions have been violated.

**1.8** How do we know when these assumptions have been violated? *(2 points)*

1. The threshold for circularity is 0.70 and the threshold for multicollinearity is 0.70.
2. The threshold for circularity is 0.70 and the threshold for multicollinearity is 0.80.
3. The threshold for circularity is 0.80 and the threshold for multicollinearity is 0.70.
4. The threshold for circularity is 0.80 and the threshold for multicollinearity is 0.80.

**Question 2.**

Data: Presence of Bacteria after Drug Treatments

Description: Tests of the presence of the bacteria H. influenzae in children with otitis media (middle ear infection) in the Northern Territory of Australia.

Variables:

y presence or absence: a factor with levels n and y.

ap active/placebo: a factor with levels a and p (a = Drug Treatment).

**Consider the following code and output from R software:**

> table(y, ap)

ap

y a p

n 31 12

y 93 84

> prop.table(table(y, ap), 2) #column percentages

ap

y a p

n 0.250 0.125

y 0.750 0.875

> chisq.test(table(y, ap), correct=FALSE)

Pearson's Chi-squared test

data: table(y, ap)

X-squared = 5.3764, df = 1, p-value = 0.02041

**2.1** What type of hypothesis test is this? *(2 points)*

1. t-test
2. ANOVA
3. Chi Square
4. Simple regression
5. Multiple regression

**2.2** List the dependent variable(s) ***and*** its level(s) of measurement. *(2 points)*

**2.3** List the independent variable(s) ***and*** its level(s) of measurement. *(2 points)*

**2.4** Report the technical details. *(3 points)*

**2.5** Interpret, describe, and comment on the results. Assume the reader will *not* have access to the R output; your answers to this question and to 2.4 should convey the relevant results. *(6 points)*

**2.6** Which of the following is the appropriate measure of association for this problem? *(2 points)*

1. Phi
2. Cramer’s V
3. Gamma
4. Spearman’s Rho
5. Pearson’s R

**2.**7 The result for the appropriate measure of association in the previous question is: 0.1563276

Interpret this result. *(3 points)*

**Question 3.**

Data: Anatomical Data from Domestic Cats

Description: The heart and body weights of samples of male and female cats. The cats were all adult, over 2 kg body weight.

Variables:

Sex sex: Factor with evels "F" and "M".

Bwt body weight in kg.

**Consider the following code and output from R software:**

> aggregate(Bwt~Sex, cats, mean)

Sex Bwt

1 F 2.359574

2 M 2.900000

> t.test(Bwt~Sex, cats)

Welch Two Sample t-test

data: Bwt by Sex

t = -8.7095, df = 136.838, p-value = 8.831e-15

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.6631268 -0.4177242

sample estimates:

mean in group F mean in group M

2.359574 2.900000

**3.1** What type of hypothesis test is this? *(2 points)*

1. t-test
2. ANOVA
3. Chi Square
4. Simple regression
5. Multiple regression

**3.2** List the dependent variable(s) ***and*** its level(s) of measurement. *(2 points)*

**3.3** List the independent variable(s) ***and*** its level(s) of measurement. *(2 points)*

**3.4** Report the technical details. *(3 points)*

**3.5** Interpret, describe, and comment on the results. Assume the reader will *not* have access to the R output; your answers to this question and to 3.4 should convey the relevant results. *(6 points)*

**Question 4.**

List five ways that you could use statistics in your career. *(5 points)*

1.

2.

3.

4.

5.

**Question 5.**

Data: Old Faithful Geyser Data

Description: A version of the eruptions data from the ‘Old Faithful’ geyser in Yellowstone National Park, Wyoming. This is of continuous measurement from August 1 to August 15, 1985.

Variables:

duration Eruption time in mins

waiting Waiting time for this eruption

**Consider the following output from R software:**

Call:

lm(formula = duration ~ waiting)

Residuals:

Min 1Q Median 3Q Max

-2.21805 -0.72357 -0.01979 0.75071 2.11109

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.313144 0.269935 27.09 <2e-16 \*\*\*

waiting -0.053272 0.003666 -14.53 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.879 on 297 degrees of freedom

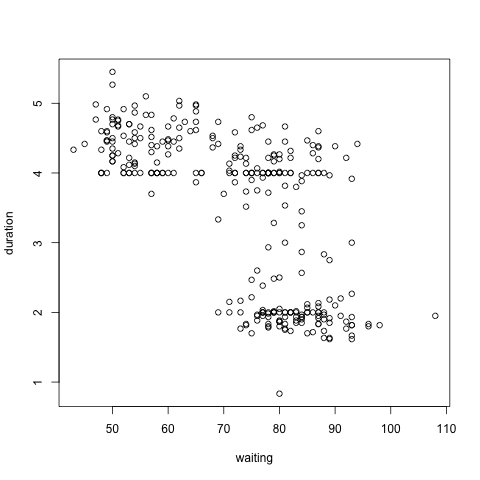
Multiple R-squared: 0.4155, Adjusted R-squared: 0.4136

F-statistic: 211.2 on 1 and 297 DF, p-value: < 2.2e-16

**5.1** Write the linear regression equation. *(2 points)*

**5.2** If you waited 75 minutes to see an eruption, what is your estimate of the duration of the eruption? *(5 points)*

**5.3** Using your answer to 5.1, draw (precisely!) the least-squares regression line on the following scatterplot. *(5 points)*



**5.4** What percentage of variance in duration (eruption time in minutes) can be explained by waiting (waiting time for this eruption)? *(3 points)*

**5.5** Would you use this model to guarantee tourists a precise prediction of eruption duration based on waiting time? Why or why not? *(5 points)*

**Question 6.**

Data: Anorexia Data on Weight Change

Description: Weight change data for young female anorexia patients.

Variables:

Treat Factor of three levels: "Cont" (control), "CBT" (Cognitive Behavioural treatment) and

"FT" (family treatment).

Prewt Weight of patient before study period, in lbs.

Postwt Weight of patient after study period, in lbs.

**Consider the following code and output from R software:**

> aggregate(Postwt~Treat, anorexia, mean)

Treat Postwt

1 CBT 85.69655

2 Cont 81.10769

3 FT 90.49412

> anova.results <- aov(Postwt~Treat, anorexia)

> summary(anova.results)

Df Sum Sq Mean Sq F value Pr(>F)

Treat 2 919 459.5 8.651 0.000444 \*\*\*

Residuals 69 3665 53.1

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

F-statistic: 211.2 on 1 and 297 DF, p-value: < 2.2e-16

**6.1** What type of hypothesis test is this? *(2 points)*

1. t-test
2. ANOVA
3. Chi Square
4. Simple regression
5. Multiple regression

**6.2** List the dependent variable(s) ***and*** its level(s) of measurement. *(2 points)*

**6.3** List the independent variable(s) ***and*** its level(s) of measurement. *(2 points)*

**6.4** Report the technical details. *(3 points)*

**6.5** Interpret, describe, and comment on the results. Assume the reader will *not* have access to the R output; your answers to this question and to 6.4 should convey the relevant results. *(6 points)*

**Question 7.**

Now that you have surmounted all its hurdles, what statistical software will you use forever and ever with great joy and happiness?

**Question 8.**

Data: Nutritional and Marketing Information on US Cereals

Description: The data come from the 1993 ASA Statistical Graphics Exposition, and are taken from the mandatory FDA food label. The data have been normalized here to a portion of one American cup.

Variable:

sugars grams of sugars in one portion.

**Consider the following graph from R software:**



**8.1** Describe the distribution of sugar content in U.S. cereals. *(5 points)*