Family (Last) Name _____

Given (First) Name Jerry

Student Number _____

STA 302s13 Quiz 10A

1. (6 points) In a study comparing the effectiveness of different exercise programmes, volunteers were randomly assigned to one of three exercise programmes (A, B, C) or put on a waiting list and told to work out on their own. Aerobic capacity is the body's ability to process oxygen. Aerobic capacity was measured before and after 6 months of participation in the program (or 6 months of being on the waiting list). The response variable was improvement in aerobic capacity. The independent variables were age (a covariate) and treatment group.

   Consider a regression model with an intercept. The model should allow for the possibility of regression lines that are not parallel.

   (a) Write the regression equation. Please use x for age, and make its regression coefficient β_1.

   $$Y_i = \beta_0 + \beta_1 x + \beta_2 e_1 + \beta_3 e_2 + \beta_4 e_3 + \beta_5 x e_1 + \beta_6 x e_2 + \beta_7 x e_3 + \varepsilon_i$$

   (b) Make a table with columns showing how you would set up indicator dummy variables for treatment condition. Waiting List should be the reference category. Make a wider column on the right in which you give $E(Y|\mathbf{X})$ for each treatment condition. Of course Waiting List is one of the treatments.

A	1	0	0	$\beta_0 + \beta_2 + (\beta_1 + \beta_5)x$
B	0	1	0	$\beta_0 + \beta_3 + (\beta_1 + \beta_6)x$
C	0	0	1	$\beta_0 + \beta_4 + (\beta_1 + \beta_7)x$
wait	0	0	0	$\beta_0 + \beta_1 x$

   (c) Suppose you wanted to know whether the slopes of the 4 regression lines were equal. In terms of β values, what null hypothesis would you test?

   $$H_0: \beta_5 = \beta_6 = \beta_7 = 0$$

   (d) Suppose you wanted to know whether the difference in effectiveness between Programme A and the Waiting List depends on the participant's age. In terms of β values, what null hypothesis would you test?

   $$H_0: \beta_5 = 0$$

   (e) Suppose you wanted to estimate the difference in average improvement between programmes A and C for a 27 year old participant. Give your answer in terms of $\hat{\beta}$ values.

   $$\hat{\beta}_2 + 27\hat{\beta}_5 - \hat{\beta}_4 - 27\hat{\beta}_7$$

2. (4 points) In your analysis of the Census Tract data, recall that 1=NE, 2=NC, 3=S, 4=W.

(a) You want to know whether, allowing for the other variables, percent High School graduates is connected with to the crime rate. Write the t or F statistic and the p-value (both numbers from your printout) in the table below.

Test Statistic (F or t)	p-value
$t = 2.484$	$p = 0.01423$

(b) In plain, non-statistical language, what do you conclude from that last test? Be guided by the 0.05 significance level, but don't mention it.

OKAY to say "variable"

Allowing for other characteristics of the census tracts, census tracts with a higher percentage of high school graduates tend to have higher crime rates.

(c) You want to know whether, controlling for the other variables, region is related to the crime rate. Write the t or F statistic and the p-value (both numbers from your printout) in the table below.

Test Statistic (F or t)	p-value
$F = 13.2$	0.0000001368

1.368 e-07 OKAY

(d) Controlling for all other variables, which region has the lowest estimated crime rate? Give a one word answer.

North East

(e) You want to know whether, controlling for other variables, crime rates are the same in the North Central and South regions. If they are not the same, you want to know which one is higher. Write the t or F statistic and the p-value (both numbers from your printout) in the table below.

Test Statistic (F or t)	p-value
$t = -2.71$	0.00761

(f) In plain, non-statistical language, what do you conclude from that last test? Be guided by the 0.05 significance level, but don't mention it. Your answer can start out "All other things being equal, census tracts in the North Central region have . . ."

lower average crime rates than census tracts in the South.

```
R version 2.15.1 (2012-06-22) -- "Roasted Marshmallows"
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-apple-darwin9.8.0/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.52 (6188) i386-apple-darwin9.8.0]

[Workspace restored from /Users/brunner/.RData]
[History restored from /Users/brunner/.Rapp.history]

> # R work for STA302f13 Assignment 10
> rm(list=ls())
> census = read.table("http://www.utstat.toronto.edu/~brunner/302f13/code_n_data/hw/CensusTract.data")
> attach(census)
>
> crimerate = crimes/pop
> region=factor(region,labels=c("NE","NC","S","W" ))
> fullmod = lm(crimerate ~ area + urban + docs + beds + hs + region)
> summary(fullmod)

Call:
lm(formula = crimerate ~ area + urban + docs + beds + hs + region)

Residuals:
    Min      1Q  Median      3Q     Max
-26.715  -8.497  -1.010   7.703  22.749

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.8256300  8.1566926   2.431  0.01642 *
area         0.0006622  0.0003672   1.803  0.07360 .
urban        0.0212986  0.0560259   0.380  0.70444
docs         0.0019495  0.0014347   1.359  0.17653
beds        -0.0003279  0.0005083  -0.645  0.52005
hs           0.3658880  0.1472825   2.484  0.01423 *
regionNC     8.9109681  2.9833246   2.987  0.00336 **
regionS     15.8130903  2.8589149   5.531 1.64e-07 ***
regionW     20.6566143  4.0338907   5.121 1.05e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.75 on 132 degrees of freedom
Multiple R-squared: 0.4773, Adjusted R-squared: 0.4457
F-statistic: 15.07 on 8 and 132 DF,  p-value: 1.513e-15

> justcovs = lm(crimerate ~ area + urban + docs + beds + hs)
> justregion = lm(crimerate ~ region)
> # Other matter controlling for region?
> anova(justregion,fullmod)
```

```
Analysis of Variance Table

Model 1: crimerate ~ region
Model 2: crimerate ~ area + urban + docs + beds + hs + region
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    137 18010
2    132 15244  5    2765.8 4.7898 0.0004685 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # Region controlling for others
> anova(justcovs,fullmod)
Analysis of Variance Table

Model 1: crimerate ~ area + urban + docs + beds + hs
Model 2: crimerate ~ area + urban + docs + beds + hs + region
  Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1    135 19817
2    132 15244  3    4573.3 13.2 1.368e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
>
> # Testing pairwise differences
> betahat = fullmod$coefficients; betahat #$
  (Intercept)          area         urban          docs          beds            hs      regionNC
19.8256299563  0.0006621526  0.0212986117  0.0019494894 -0.0003278516  0.3658879916  8.9109680932
      regionS       regionW
15.8130903085 20.6566143443
> V = vcov(fullmod)
> dfe = fullmod$df.residual #$
> # t-tests
> a23 = rbind(0,0,0,0,0,0,1,-1,0)
> a24 = rbind(0,0,0,0,0,0,1,0,-1)
> a34 = rbind(0,0,0,0,0,0,0,1,-1)
> # NC vs S
> T23 = as.numeric( t(a23)%*%betahat/sqrt(t(a23)%*%V%*%a23) )
> T23; 2*(1-pt(abs(T23),dfe))
[1] -2.710271
[1] 0.00761624
> # NC vs W
> T24 = as.numeric( t(a24)%*%betahat/sqrt(t(a24)%*%V%*%a24) )
> T24; 2*(1-pt(abs(T24),dfe))
[1] -3.534517
[1] 0.0005638497
> # S vs W
> T34 = as.numeric( t(a34)%*%betahat/sqrt(t(a34)%*%V%*%a34) )
> T34; 2*(1-pt(abs(T34),dfe))
[1] -1.40551
[1] 0.1622188
>
>
```