STA 3032 - Practice set 3

1. Ford wants to compare mean assembly times for Explorers at their 3 assembly plants. They observe random samples of 10 cars at each plant, and obtain the following summary statistics on assembly times (in minutes):

Plant	Mean	Std. Dev.
Atlanta	180	12
Chicago	185	10
Detroit	175	9

   (a) Compute the between plant (Treatment) sum of squares and its degrees of freedom.

   Solution: $n_1 = n_2 = n_3 = 10$ so the grand average

$$\bar{y}_{++} = \frac{10(180) + 10(185) + 10(175)}{10 + 10 + 10} = 180$$

   So according to the formula,

$$\text{SST}_{\text{rt}} = 10(180 - 180)^2 + 10(185 - 180)^2 + 10(175 - 180)^2 = 500$$

   (b) Compute the within plant (Error) sum of squares and its degrees of freedom.

   Solution: SSE$= 9(12^2) + 9(10^2) + 9(9^2) = 2925$

   (c) Compute the test statistic and make conclusion.1

   Solution:

$$\text{T.S.} = \frac{\text{SST}_{\text{rt}}/2}{\text{SSE}/27} = 2.307692$$

   The p-value (area to the right of T.S. under a $F_{2,27}$ distribution is 0.1187904, so we fail to reject the null that $\text{H}_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ and conclude that all the Plants have the same mean.

   (d) If applicable, use Tukey's multiple comparison procedure with $\alpha = 0.05$ to determine which plants differ.

   Solution: Not applicable since they are no differences.

2. A study was conducted to relate weight gain in chickens (y) to the amount of the amino acid lysine ingested by the chicken (x). A simple linear regression is fit to the data.

	coefficients	S.E.	t-stat	p-value
Intercept	12.4802	1.2637	9.8762	<0.0000
X	36.8929	7.5640	4.8774	0.0012

Source	df	SS	MS	F	p-value
Regression	1	27.07	27.07	23.79	0.0012
Residual	8	9.10	1.14		
Total	9	36.18			

(a) Give the fitted equation, and the predicted value for $x = 0.20$.

Solution: The equation is

$$\hat{y} = 12.4802 + 36.8929x$$

so, at $x = 0.20, \hat{y} = 19.85878$.

(b) Give a 95% Confidence Interval for the MEAN weight gain of all chickens with $x = 0.20$ (Note: $\bar{x} = 0.16$ and $\sum(x_i - \bar{x})^2 = 0.020$)

Solution: From the notes the CI on the mean response is

$$19.85878 \mp t_{0.975,8}\sqrt{1.14\left(\frac{1}{10} + \frac{(0.20 - 0.16)^2}{0.020}\right)} \rightarrow (18.83405, 20.88351)$$

(c) What proportion of the variation in weight gain is "explained" by lysine intake?

Solution: $R^2 = 27.07/36.18 = 0.7482034$, so approximately 75%.

3. A researcher reports that the correlation between length (inches) and weight (pounds) of a sample of 16 male adults of a species is $r = 0.40$. A colleague from Europe transforms the data from length in inches to centimeters (1 inch=2.54 cm) and weight from pounds to kilograms (1 pound=2.2 kg). What is the colleagues estimate of the correlation?

Solution: The same since correlation is uniteless. Recall that correlation is the standardized covariance.

4. A randomized block design is conducted to compare the output of three weaving looms (treatments) for a sample of 10 operators (blocks), where each operator's output is measured on each loom. The Mean Square Error from the ANOVA is MSE = 500. Compute Bonferroni's B, i.e. margin of error, the minimum significant difference for concluding that two looms' population means differ if their sample means differ by at least B.

   Solution: From the notes

$$B = \underbrace{t_{1-0.008333333,18}}_{t_{1-\frac{0.05}{2(3)},(10-1)(3-1)}} \sqrt{500(2/10)} =$$

5. Late at night you find the following output in your department's computer lab. The data represent numbers of emigrants from Japanese regions, as well as a set of predictor variables from each region.

	coefficients	S.E.	t-stat	p-value
Intercept	407.070	226.341	1.798	0.079
LANDCULT	-1.685	3.5677	-0.472	0.639
AREAFARM	-2.132	1.056	-2.019	0.050
PIONEERS	175.968	61.222	2.874	0.006

R-square=0.274, R-square(adj)=0.222

Source	df	SS	MS	F	p-value
Regression	3	514814.087	171604.696	5.187	0.004
Residual	41	1356447.158	33084.077		
Total	44	1871261.244			

   (a) How many regions are there in the analysis?

   Solution: Since $\text{df}_{total} = n - 1 = 44$ then $n = 45$.

   (b) Give the test statistic and p-value for testing (H_0) that none of the predictors are associated with the response EMGRANTS.

   Solution: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ and the F-test statistic is MSR/MSE=5.187 with p-value 0.004.

   (c) Give the test statistic and p-value for testing whether LANDCULT is associated with EMGRANTS, after controlling for AREAFARM and PIONEERS.

   Solution: Individual test for $H_0 : \beta_1 = 0$ using output from the model with all 3 predictors yields t-test statistic -0.472 with p-value 0.639.

   (d) What proportion of the variation in EMGRANTS is "explained" by the model?

   Solution: $R^2 = 0.274$, so 27.4%.

(e) Give the estimated regression equation.
Solution:

$$\widehat{\text{No. emmigrants}} = 407.070 - 1.685(\text{LANDCULT}) - 2.132(\text{AREAFARM}) + 175.968(\text{PIONEERS})$$

6. A realtor is interested in the determinants of home selling prices in his territory. He takes a random sample of 24 homes that have sold in this area during the past 18 months, observing:

- selling PRICE (y)
- AREA (x_1)
- BEDrooms (x_2)
- BATHrooms (x_3)
- POOL dummy ($x_4 = 1$ if Yes, 0 if No)
- AGE (x_5)

He/She fits the following models (predictor variables to be included in model are given for each model):

Model 1	AREA, BED, BATH, POOL, AGE	SSE1 = 250, SSR1 = 450
Model 2	AREA, BATH, POOL	SSE2 = 325, SSR2 = 375

(a) Test whether neither BED or AGE are associated with PRICE, after adjusting for AREA, BATH, and POOL at the $\alpha = 0.05$ significance level.

Solution: $H_0 : \beta_{\text{BED}} = \beta_{\text{AGE}} = 0$. "Reduced" is model 2 and "full" is model 1, is

$$T.S. = \frac{\frac{325-250}{2}}{\frac{250}{18}} = 2.7$$

with p-value the area to the right of 2.7 under a $F_{2,18}$ is 0.0942996, so fail to reject H_0 and proceed with model 2.

(b) What statement best describes β_4 in Model 1?

i. **Added value (on average) for a POOL, controlling for AREA, BED, BATH, AGE**
ii. Effect of increasing AREA by 1 unit, controlling for other factors
iii. Effect of increasing BED by 1 unit, controlling for other factors
iv. Effect of increasing BATH by 1 unit, controlling for other factors
v. Average price for a house with a POOL

7. Yields of entozoic amoebae in 5 treatment conditions. Conditions:

   1=None

   2=Heat at 70° C for 10min

   3=Addition of 10% Formalin

   4=Heat followed by Formalin

   5=Formalin followed, after 1 hr by heat

Trt 1	Trt 2	Trt3	Trt 4	Trt 5
265	204	191	221	259
292	234	207	205	206
268	197	218	178	179
251	176	201	167	199
245	240	192	224	180
192	190	192	225	146
228	171	214	171	182
291	190	206	214	147
185	222	185	283	182
247	211	163	277	223

Data can also be found at
http://www.stat.ufl.edu/~athienit/STA3032/practice3_entozoic.csv.
Are there differences in the means of the 5 treatments?

Solution: In R we can get the ANOVA table

```
> yield=read.csv("http://www.stat.ufl.edu/~athienit/STA6166/practice3_entozoic.csv",
+                header=TRUE)
>
> library(reshape2)
> yield_long=melt(yield)
No id variables; using all as measure variables
> colnames(yield_long)=c("Treatment","Yield")
>
> model=aov(Yield~Treatment,data=yield_long)
> summary(model)
            Df Sum Sq Mean Sq F value  Pr(>F)
Treatment    4  19666    4916   5.044 0.00191 **
Residuals   45  43858     975
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Assuming that yield is normally distributed the test for $H_0 : \alpha_1 = \cdots \alpha_5 = 0$ is the F-test in the table 5.044 with p-value 0.00191, so conclude that there are differences. Next step is to perform post hoc comparison.

```
> TukeyHSD(model)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Yield ~ Treatment, data = yield_long)

$Treatment
              diff       lwr        upr     p adj
Trt.2-Trt.1 -42.9 -82.57118  -3.228817 0.0281550
Trt3-Trt.1  -49.5 -89.17118  -9.828817 0.0078618
Trt.4-Trt.1 -29.9 -69.57118   9.771183 0.2208842
Trt.5-Trt.1 -56.1 -95.77118 -16.428817 0.0019658
Trt3-Trt.2   -6.6 -46.27118  33.071183 0.9894377
Trt.4-Trt.2  13.0 -26.67118  52.671183 0.8832844
Trt.5-Trt.2 -13.2 -52.87118  26.471183 0.8774716
Trt.4-Trt3   19.6 -20.07118  59.271183 0.6284290
Trt.5-Trt3   -6.6 -46.27118  33.071183 0.9894377
Trt.5-Trt.4 -26.2 -65.87118  13.471183 0.3444453
```

So summarizing which treatments are the same (underlined by the same line) and which are different, we have: $\underline{5\,3}\,\underline{2\,4}\,\underline{1}$

8. Endurance Times (in minutes) for nine well-trained cyclists, on each of 4 doses of caffeine (0,5,9,13 mg)

	Cyclists								
Dose	1	2	3	4	5	6	7	8	9
0	36.05	52.47	56.55	45.20	35.25	66.38	40.57	57.15	28.34
5	42.47	85.15	63.20	52.10	66.20	73.25	44.50	57.17	35.05
9	51.50	65.00	73.10	64.40	57.45	76.49	40.55	66.47	33.17
13	37.55	59.30	79.12	58.33	70.54	69.47	46.48	66.35	36.20

Data can also be found at
http://www.stat.ufl.edu/~athienit/STA3032/practice3_endurance.csv

(a) Which factor is the the treatment and which is the block?
Solution:

- Fixed factor is "Dose"
- Random factor is "Cyclist"

(b) Test whether the factor populations means are all equal.

i. Using parametric technique.

```
> endur=read.csv(
+ "http://www.stat.ufl.edu/~athienit/STA6166/practice3_endurance.csv",
+                header=TRUE)
> library(reshape2)
> endur_long=melt(endur,id="Dose")
> endur_long$Dose=factor(endur_long$Dose)
> colnames(endur_long)[2:3]=c("Cyclist","Time")
>
> model2=aov(Time~Dose+Cyclist,data=endur_long)
> summary(model2)
            Df Sum Sq Mean Sq F value   Pr(>F)
Dose         3    933   311.0   5.917  0.00359 **
Cyclist      8   5558   694.7  13.216 4.17e-07 ***
Residuals   24   1262    52.6
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

As we can see from the p-value of 0.00308, that not all dosage means are equal. Next we should perform multiple comparisons

```
> TukeyHSD(model2,"Dose")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Time ~ Dose + Cyclist, data = endur_long)

$Dose
```

```
           diff       lwr       upr     p adj
5-0  11.2366667  1.808030 20.665303 0.0153292
9-0  12.2411111  2.812474 21.669748 0.0076616
13-0 11.7088889  2.280252 21.137526 0.0110929
9-5   1.0044444 -8.424192 10.433081 0.9909369
13-5  0.4722222 -8.956414  9.900859 0.9990313
13-9 -0.5322222 -9.960859  8.896414 0.9986162
```

0 5 13 9

ii. (optional) Using non-parametric technique.

```
> friedman.test(Time ~ Dose|Cyclist,data=endur_long)

Friedman rank sum test

data:  Time and Dose and Cyclist
Friedman chi-squared = 14.2, df = 3, p-value = 0.002645
```

The multiple comparison yields the same conclusion

```
[1] "95 %  Pairwise CIs"
[1] "0-5"
[1]  0.5496698 29.4503302
[1] "0-13"
[1]  2.54967 31.45033
[1] "0-9"
[1]  3.54967 32.45033
[1] "5-13"
[1] -12.45033  16.45033
[1] "5-9"
[1] -11.45033  17.45033
[1] "13-9"
[1] -13.45033  15.45033
```