Stat 5021, Second exam, Fall '05 Name ____________________

Thirty five people took this test. The average score was 70.4 with a standard deviation of 21.7.

1. Suppose $X_1, \ldots, X_{10}$ are iid each Normal$(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_{15}$ are iid each Normal$(\mu_2, \sigma^2)$ where $\mu_1$, $\mu_2$ and $\sigma^2$ are unknown and the two samples are independent. Suppose we observed $\bar{x} = 21.9$, $\sum(x_i - 21.9)^2 = 216.5$, $\bar{y} = 26.3$ and $\sum(y_i - 26.3)^2 = 249.4$.

i) Find a 95% confidence interval for $\mu_1 - \mu_2$.

ii) For testing H:$\mu_1 - \mu_2 = 0$ against K:$\mu_1 - \mu_2 < 0$ find the approximate *p-value* or *level of significance* for these data.

***Solution:***

i)

$$(21.9 - 26.3) \quad \mp \quad t_{.025,23}\sqrt{\frac{216.5 + 249.4}{9 + 14}}\sqrt{\frac{1}{10} + \frac{1}{15}}$$

where $t_{.025,23} = 2.069$.

ii)

$$\frac{(21.9 - 26.3)}{\sqrt{\frac{216.5+249.4}{9+14}}\sqrt{\frac{1}{10} + \frac{1}{15}}} = \frac{-4.4}{1.837} = -2.395$$

But $P(T_{23} < -2.395) = .0123$ which is the *p-value*

∎

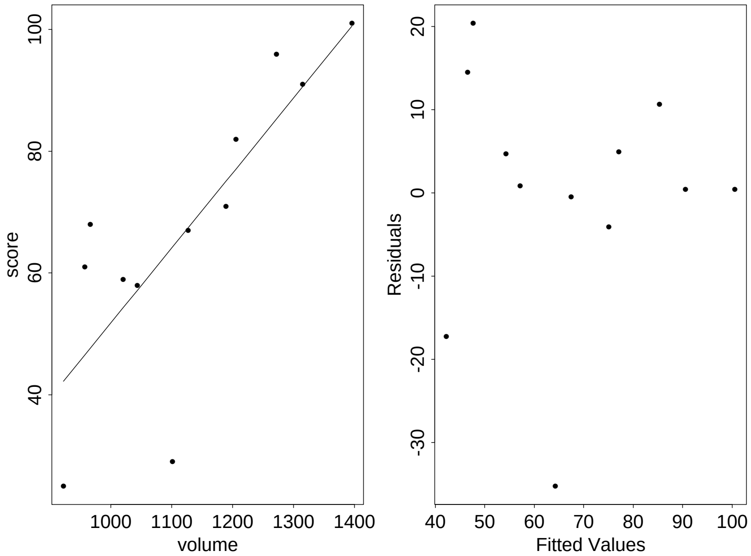
2. Neurologists believe that a part of the brain called the hippocampus plays an important role in short term memory. For 12 subjects the volume of the right hippocampus was measured and each was given a memory test and their scores where recorded. (The better their results on the test the bigger their score.) A simple linear regression model

```
out.lm <- lm(score ~ volume)
```

was fit to these data. Use the following computer output, a plot of the data along with the least squares line and a plot of the residuals against the fitted values to answer the following questions.

```
>summary(out.lm)
Call: lm(formula = score ~ volume)
Residuals:
    Min     1Q Median    3Q   Max
 -35.25 -1.354 0.6659 6.389 20.37

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept) -71.3016  34.4377    -2.0705   0.0652
     volume   0.1231   0.0303     4.0592   0.0023

Residual standard error: 15.31 on 10 degrees of freedom
Multiple R-Squared: 0.6223
F-statistic: 16.48 on 1 and 10 degrees of freedom, the p-value is 0.002289

>predict(prb3.lm,data.frame(volume=c(950,1050,1150)),se.fit=T)
$fit
        1        2        3
```

```
 45.65527 57.96652 70.27777

$se.fit
        1        2        3
 6.931261 4.984649 4.477516
```




i) What does the regression equation predict for the test score for someone whose right hippocampus has a volume of 885.

ii) At level $\alpha = .05$ test H: $\beta_1 = 0.1$ against K: $\beta_1 \neq 0.1$.

iii) Find a 95% confidence interval for the expected score of someone with a volume of 1050.

iv) Sometimes to study how influential a given data point is in the regression equation statisticians do the regression again with that data point removed. Select the point whose removal would result, you believe, in the largest change in the estimate of $\beta_1$. Briefly justify your answer.

***Solution:***

i) $-71.30 + 0.1231 \times 885 = 37.6$

ii)

$$\frac{0.1231 - 0.1}{0.0303} = 0.762$$

But $|0.762| < 2.228 = t_{.025,10}$ so we accept H.

iii) $57.97 \quad \mp \quad 2.228 \times 4.985$

iv) The sixth point has the largest residual in absolute value so you might think that it would be the most influential. However removing the first point makes a bigger change in the estimate of $\beta_1$. This happens because removing the first point allows the least squares line to flatten out more than when we remove a point from the middle of the data. In fact when the first point is removed we have $\hat{\beta}_1 = 0.1048$ while if we remove the sixth point we have $\hat{\beta}_1 = 0.1193$.

∎

3. A study was carried out to see if one could predict the sale prices of a certain manufacturer's antique vases. For twenty recent sales the following data were collected.

- $P$: Sale price in dollars.
- $A$: Age in years.

- $C$: Condition of the vase. This took on four values: 1, 2, 3 and 4. The larger the number the better the condition of the vase.

The model

$$P = \beta_0 + \beta_a A + \beta_c C + \beta_{ac} A \times C + Z$$

was fit to the data and the usual output is given below. In addition two anova tables were computed.

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 10287.07 | 14066.22 | 0.731 | 0.475 |
| Age | 32.33 | 83.62 | 0.387 | 0.704 |
| Condition | - 2006.60 | 4366.65 | -0.460 | 0.652 |
| Age:Condition | 28.61 | 26.02 | 1.100 | 0.288 |

```
Residual standard error: 2089 on 16 degrees of freedom
Multiple R-Squared: 0.7563,
F-statistic: 16.55 on 3 and 16 DF,  p-value: 3.672e-05
```

Response: P

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| Age | 1 | 95716926 | 95716926 | 21.9292 | 0.0002495 *** |
| Condition | 1 | 115775516 | 115775516 | 26.5247 | 9.679e-05 *** |
| Age:Condition | 1 | 5280233 | 5280233 | 1.2097 | 0.2876604 |
| Residuals | 16 | 69837142 | 4364821 | | |

Response: P1

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| Condition | 1 | 116173237 | 116173237 | 26.6158 | 9.51e-05 *** |
| Age | 1 | 95319206 | 95319206 | 21.8381 | 0.0002545 *** |
| Age:condition | 1 | 5280233 | 5280233 | 1.2097 | 0.2876604 |
| Residuals | 16 | 69837142 | 4364821 | | |

i) In the full model test $H : \beta_a = 0$ against $K : \beta_a \neq 0$ at level $\alpha = 0.05$.
ii) In the model $P = \beta_0 + \beta_a A + Z$ test $H : \beta_a = 0$ against $K : \beta_a \neq 0$ at level $\alpha = 0.05$.
iii) What is the explanation for these "contradictory" results?
iv) In the full model test $H : \beta_c = \beta_{ac} = 0$ against $K :$ At least one is not zero at level $\alpha = 0.05$.
***Solution:***
i) We see from the computed t value = 0.387 we fail to reject $H$.
ii) We see from the first anova table that for this reduced model

$$SSR(\beta_a|\beta_0) = 95716926$$
$$RSS = 115775516 + 5280233 + 69837142 = 190892891$$
$$\frac{95716926/1}{190892891/18} = 95716926/9468699 = 10.11 > 4.41 = f_{1,18;.05}$$

so we reject $H$.

iii) The test in part i) is measuring how important Age is in explaining Price when the other two variables are in the model while the test in part ii) is measuring how important Age is by itself.

iv) From the first anova table we see that

$$\frac{(SSR(\beta_c|\beta_0,\beta_a)+SSR(\beta_{ac}|\beta_0,\beta_a,\beta_c))/2}{ResSS/16}=\frac{(115775516+5280233)/2}{4364821}=\frac{60527874}{4364821}=13.87$$

Since this is greater than $3.634 = f_{2,16;.05}$ we reject $H$.

∎