Problem 1: A study examines the vital capacity measurements of 60 adult males classified by 4 different types of occupations and three age groups. Data stored in the file `test2_1.csv` list the values of vital capacity (VC) vs age group (AGE) and occupation (OCC).

i) Apply the Anova test to answer the following questions: (a) does the vital capacity differs among individuals with different occupations, (b) does the vital capacity differences among individuals with different age groups, and (c) is the interaction between age and occupation? Let $\alpha = 0.01$ for all tests.

ii) Use the Tukey’s HSD procedure to test for significant differences among individual pairs of means for age group and occupation, if appropriate (you can ignore the interaction term in the Tukey’s HSD procedure). Justify your conclusion.

(i) We use the Anova to test the null hypothesis that there is no difference among the means of (a) individuals with different occupations, (b) different age and (b) that there is no interaction between age and occupation.

```r
> data21 <- read.csv("C:/Users/dlabate/Desktop/Teaching/ma4310/test2_1.csv")
> data21$AGE = factor(data21$AGE, levels=unique(data21$AGE))
> data21$OCC = factor(data21$OCC, levels=unique(data21$OCC))
> data21.model = aov(VC~AGE+OCC+AGE:OCC, data = data21)
> anova(data21.model)
```

Analysis of Variance Table

Response: VC

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGEGROUP</td>
<td>2</td>
<td>12.3088</td>
<td>6.1544</td>
<td>29.3817</td>
<td>4.652e-09 ***</td>
</tr>
<tr>
<td>OCC</td>
<td>3</td>
<td>19.7785</td>
<td>6.5928</td>
<td>31.4750</td>
<td>2.129e-11 ***</td>
</tr>
<tr>
<td>AGEGROUP:OCC</td>
<td>6</td>
<td>8.9489</td>
<td>1.4915</td>
<td>7.1205</td>
<td>1.825e-05 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>48</td>
<td>10.0542</td>
<td>0.2095</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The p-value in the last row shows that for each of the 3 cases p-value < 0.01. Thus we reject the null hypothesis and we conclude that (a) vital capacity differs among individuals with different occupations, (b) vital capacity differs among individuals with different age groups, and (c) there is an interaction between age and occupation.

(ii) We run the Tukey’s HSD procedure:

```r
> TukeyHSD(data21.model)
```

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = VC ~ AGE + OCC + AGE:OCC, data = data21)

$AGE

<table>
<thead>
<tr>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>-0.7395</td>
<td>-1.08952404</td>
<td>0.389476</td>
</tr>
<tr>
<td>3-1</td>
<td>0.3465</td>
<td>-0.00352404</td>
<td>0.696524</td>
</tr>
<tr>
<td>3-2</td>
<td>1.0860</td>
<td>0.73597596</td>
<td>1.436024</td>
</tr>
</tbody>
</table>

$OCC

<table>
<thead>
<tr>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>b-a</td>
<td>0.2073333</td>
<td>-0.23743004</td>
<td>0.6520967</td>
</tr>
<tr>
<td>c-a</td>
<td>0.4613333</td>
<td>0.01656996</td>
<td>0.9060967</td>
</tr>
<tr>
<td>d-a</td>
<td>1.4940000</td>
<td>1.04923663</td>
<td>1.9387634</td>
</tr>
<tr>
<td>c-b</td>
<td>0.2540000</td>
<td>-0.19076337</td>
<td>0.6987634</td>
</tr>
<tr>
<td>d-b</td>
<td>1.2866667</td>
<td>0.84190330</td>
<td>1.7314300</td>
</tr>
<tr>
<td>d-c</td>
<td>1.0326667</td>
<td>0.58790330</td>
<td>1.4774300</td>
</tr>
</tbody>
</table>
For the age factor, we observe \textit{p-value} < 0.01 only for the comparison 2-1 and 3-2. For the occupation factor, we observe \textit{p-value} < 0.01 only for the comparison d-a, d-b and d-c. All the other comparisons are not statistically significant at level $\alpha = 0.01$.

**Problem 2:** Pulmonary blood flow (PBF) and pulmonary blood volume (PBV) values were recorded for 16 infants and children with congenital heart disease, see file \texttt{test2_2.csv}

1. Write the equation of the linear regression equation of the PBF (the response variable) as a function of the PBV (the explanatory variable). (round to 3 decimal digits)
2. Test the hypothesis $H_0: \beta_1=0$ vs $H_1: \beta_1 \neq 0$ with significance level $\alpha = 0.05$ and $\alpha = 0.005$
3. Compute the approximate 95\% confidence interval of $\beta_1$ (round to 3 decimal digits)
4. Would we obtain the same value of $\beta_1$ (regression coefficient) and $r^2$ (coefficient of determination) if we interchange x and y in the R formulas? Explain why we obtain or we do not obtain the same quantity.

```r
> data22 <- read.csv("C:/Users/dlabate/Desktop/Teaching/ma4310/test2_2.csv")
> x <- data22$PBV
> y <- data22$PBF
> plot(x, y, main="Scatterplot", xlab="PBV ", ylab="PBF ", pch=19)
```

![Scatterplot](scatterplot.png)

```r
> relation <- lm(y~x)
> print(summary(relation))
Call:
  lm(formula = y ~ x)
Residuals:
     Min      1Q  Median      3Q     Max
-6.4389  -3.5963  0.1949  3.3508  6.7782
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.028332   3.267897  -0.009  0.99320
x            0.025119   0.008331   3.015  0.00927 **
---
```

\texttt{relation} is an object of class \texttt{lm}. The summary function is used to print the results of a linear model. The \texttt{summary} function returns a list containing the coefficients, standard errors, t-values, and p-values. The coefficients are estimates of the regression parameters, and the standard errors are used to calculate the t-values. The p-values are used to test the null hypothesis that the coefficient is zero. A small p-value (less than 0.05) indicates that the null hypothesis can be rejected, and the coefficient is significantly different from zero.
Residual standard error: 4.262 on 14 degrees of freedom  
Multiple R-squared: 0.3937, Adjusted R-squared: 0.3504  
F-statistic: 9.091 on 1 and 14 DF, p-value: 0.009269

i) We write the equation of the regression line

\[ y = -0.028 + 0.025 \, x \]

ii) Test the hypothesis about \( \beta_1 \) at significance levels \( \alpha = 0.05 \) and \( \alpha = 0.005 \).

We use the p-value = 0.00927 from the table and conclude that, at level \( \alpha = 0.05 \) we do reject H0 since p-value < 0.05 but at level \( \alpha = 0.005 \) we do not reject H0 since p-value > 0.005.

iii) We compute the approximate 95\% confidence interval of \( \beta_1 \). Since \( n=16 \), the number of degrees of freedom is \( r=16-2=14 \). Hence \( t(\alpha/2, r=14) = qt(1-0.05/2, 14) = 2.145 \)

\[ C.I. = 0.025 \pm 2.145 \times 0.008 = (0.008, 0.042) \]

iv) Would we obtain the same value of \( \beta_1 \) and \( r^2 \) if we interchange \( x \) and \( y \)?

If we interchange \( x \) and \( y \), the new value of \( \beta_1 \) would change since the model would become \( x = \beta_0 + \beta_1 y \) and in this case \( \beta_1 \) would measure the slope of a different line (new slope is the reciprocal of the prior one).

The value of the correlation coefficient would not change since the model associated with the bivariate normal distribution is symmetric in \( x \) and \( y \). The formula of the correlation coefficient is symmetric with respect to \( x \) and \( y \). In fact, correlation measures the strength of linear relationship between two variables and this is independent of the order in which variables are taken.

\[ \text{cor}(y, x) \]

[1] 0.6274564

\[ \text{cor}(x, y) \]

[1] 0.6274564
Problem 3: A study aims to analyze the somatosensory evoked potentials and their interrelations following stimulation of digit I of the hand. Measurements were collected from 114 healthy subjects to see how well you can predict the peak spinal latency (CVDIGI) for digit I when age in years (AGE) and arm length in cm (ARMLENGTH) are the predictor variables. Data are stored in the file test2_3.csv

i) Write the expression of the regression equation. What is the geometrical region defined by the regression equation?

\[ y = -1.947 + 0.036 x_1 + 0.196 x_2 \]

This equation describes a plane

ii) Test the hypothesis \( H_0: \beta_1=0 \) vs \( H_1: \beta_1 \neq 0 \) and \( H_0: \beta_2=0 \) vs \( H_1: \beta_2 \neq 0 \) at significance level \( \alpha = 0.001 \)

According to the table, in both cases the p-value of the test is inferior to 0.001, hence we reject \( H_0 \)

iii) Based on the regression model, can you predict that there is a direct or inverse linear relationship between CVDIGI and AGE? that there is a direct or inverse linear relationship between CVDIGI and ARMLENGTH?

Based on the regression model and the result of hypothesis testing, we conclude that data support the assumption that there is a linear model. In addition, since the coefficients \( \beta_1 \) and \( \beta_2 \) are positive, we conclude that there is a direct linear relationship between CVDIGI and AGE and that there is a direct linear relationship between CVDIGI and ARMLENGTH.

```r
> data23 <- read.csv("C:/Users/dlabate/Desktop/Teaching/ma4310/test2_3.csv")
> AGE <- data23$AGE
> ARMLENGTH <- data23$ARMLENGTH
> CVDIGI <- data23$CVDIGI
> relation <- lm(CVDIGI ~ AGE + ARMLENGTH, data = data23)
> print(summary(relation))
Call:
  lm(formula = CVDIGI ~ AGE + ARMLENGTH, data = hwR1006)
Residuals:
   Min     1Q Median     3Q    Max
-1.64884 -0.42040 -0.00034  0.41672  1.75042
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.94651 0.786771    -2.474  0.0149 *
   AGE         0.03631 0.005606     6.477 2.65e-09 ***
   ARMLENGTH   0.19627 0.008311   23.617  < 2e-16 ***
---
Residual standard error: 0.6519 on 111 degrees of freedom
Multiple R-squared: 0.8403, Adjusted R-squared: 0.8374
F-statistic: 292 on 2 and 111 DF, p-value: < 2.2e-16
```