Linear Regression Problem Sheet

1. A linear regression analysis was conducted on the Brain weight (grams) as a function of head size (cm^3). A portion of the R output is given below.

```
Call:
lm(formula = Brain ~ Head, data = Brainsize)

Residuals:
    Min      1Q  Median      3Q     Max
-175.98  -49.76   -1.76   46.60  242.34
Coefficients:
             Estimate Std. Error t-value
(Intercept) 325.5734   47.14085   6.906
Head          0.26343   0.01291  20.409

Residual standard error: 72.43 on 235 degrees of freedom
Multiple R-squared: 0.6393, Adjusted R-squared: 0.6378
F-statistic: 416.5 on 1 and 235 DF, p-value: < 2.2e-16
```

a) Write the formula of the linear regression, explaining the terms used in the expression.

(3 marks)

b) What is the predicted brain weight for a head of size 1280 cm^3.

(3 marks)

c) Calculate the 95% confidence intervals for the coefficients.

(6 marks)

d) How much of the variance in brain weight does the model explain and is it statistically significant.

(4 marks)

2. A linear regression analysis was conducted on the Brain weight (grams) as a function of head size (cm^3). A portion of the R output is given below.

$$\bar{x}_{Head} = 3634 \ \bar{y}_{Brain} = 1283$$

$$SS_{x,x} = 31486154 \ \ Cov(x, y) = 8294377$$

Call:
lm(formula = Brain ~ Head, data = Brainsize)

Residuals:

Min	1Q	Median	3Q	Max
-175.98	-49.76	-1.76	46.60	242.34

Coefficients:

	Estimate	Std. Error	t-value
(Intercept)		47.14085	
Head		0.01291	

Residual standard error: 72.43 on 235 degrees of freedom
Multiple R-squared: 0.6393, Adjusted R-squared: 0.6378
F-statistic: 416.5 on 1 and 235 DF, p-value: < 2.2e-16

a) Calculate the β_0 and β_1 estimates. (6 marks)

b) Write the formula of the linear regression, explaining the terms used in the expression. (3 marks)

c) Calculate the t-value for the β_0 and β_1 estimates. (6 marks)

d) Using a 95% criteria determine if the Coefficients are significant. (6 marks)

3. A linear regression analysis was conducted on the Brain weight (grams) as a function of head size (cm^3), age group (20-46 =0, 46+=1) and sex (male=0, female=1). A portion of the R output is given below.

```
Call:
lm(formula = Brain ~ Head+Age+Sex, data = Brainsize)

Residuals:
    Min      1Q  Median      3Q     Max
-175.98  -49.76   -1.76   46.60  242.34
Coefficients:
             Estimate  Std. Error  t-value  p-value
(Intercept) 464.56281    68.98183    6.735  1.27e-10 ***
Head          0.24421     0.01506   16.212  < 2e-16 ***
Age         -23.96845     9.48065   -2.528  0.0121 *
Sex         -22.54325    11.05789   -2.039  0.0426 *

Residual standard error: 71.36 on 233 degrees of freedom
Multiple R-squared: 0.6528, Adjusted R-squared: 0.6484
F-statistic: 146 on 3 and 233 DF, p-value: < 2.2e-16
```

a) Write the formula of the linear regression, explaining the terms used in the expression. (5 marks)

b) What is the predicted brain weight for a head of size 1280 cm^3, Age group 2, and female. (7 marks)

c) How much of the variance in brain weight does the model explain and is it statistically significant. (3 marks)

d) Explain how the degrees of freedom are calculated for the F statistic. (2 marks)

Formulae

Counting Rules

Permutations

$$^{n}P_{r} = \frac{n!}{(n-r)!}$$

Combinations

$$^{n}C_{r} = \left(\begin{array}{c} n \\ r \end{array} \right) = \frac{n!}{r!(n-r)!}$$

Addition of Probability

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

Conditional Probability

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

Bayes Formula

$$p(H_1|A) = \frac{p(A|H_1)p(H_1)}{p(A|H_1)p(H_1) + p(A|H_2)p(H_2)}$$

Probability Mass Functions

$$E[X] = \Sigma_{i=1} x_i p(x_i)$$

$$Var[X] = \Sigma_{i=1} (x_i - \mu)^2 p(x_i)$$

Geometric Distributions

$$p(k) = q^{(k-1)}p, \quad k = 1, 2, \ldots$$

$$E[k] = \frac{1}{p} \quad Var[k] = \frac{q}{p^2}$$

Binomial Distributions

$$p(k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k}, \quad k = 0, 1, 2, \ldots n$$

$$E[k] = np \quad Var[k] = npq$$

Poisson Distributions

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

$$E[k] = \lambda \quad Var[k] = \lambda$$

Chi squared

$$\chi_{GoF} = \sum \frac{(O-E)^2}{E}\ \chi^2_{k-1},\ \ \chi_{Ind} = \sum \frac{(O-E)^2}{E}\ \chi^2_{(r-1)(c-1)}$$

Linear Regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots$$

$$y_i = \beta_0 + \beta_1 x_{i1}$$

$$\beta_1 = \frac{Cov(x,y)}{SS_{xx}},\ \ \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Logistic Regression

$$p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}},\quad \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots$$