**QUESTION 1**

Briefly answer the following questions:

1. Although data subsampling is straightforward, it is often not the best method for creating an independent data set for model testing. Why not? **[5 marks]**
2. Traditionally, statisticians have not used resampling techniques when running linear regression models. What is the reasoning behind this? **[5 marks]**

**QUESTION 2**

Import the Excel file “Double\_peak\_21.xlsx” into R and carry out the ORQ (orderNorm) transformation on the variable Length using the package **bestNormalize**.

To show that you have done this correctly, calculate the correlation coefficient (Pearson’s method) between the original variable Length and the new transformed variable.

Write the value of the correlation coefficient below using 4 decimal places. **[10 marks]**

**QUESTION 3**

A student has written the following R code and found that it doesn’t run:

library(carrot)

set.seeds(1234)

trainIndex <- createDataPartition(Vegetables$potatoes, groups = 6, p=0.8, list=FALSE)

veg.train <- Vegetables [ trainIndex,]

veg.test <- Vegetables [-trainIndex]

library(Mass)

meal <- lm(potatoes~sprouts + gravy + beef + meal.time + age, data=Vegetables)

summary(meal)

# inspect the residuals from the model

hist(meal$resid)

ggnorm(meal$resid)

Find five of the mistakes and briefly explain how to correct them. **[2 marks each]**

(A)

(B)

(C)

(D)

(E)

**QUESTION 4**

Examine the output from R below and answer the following questions:

Call:

lm(formula = Mean\_LST ~ NDVI + Build\_ANN12 + Build\_ANN23 + Build\_ANN34 +

Build + dist\_centre + dist\_water + eastness + elevation +

glass + Hard\_ANN12 + Hard\_ANN23 + Hard\_ANN34 + hard + mixed +

Nat\_ANN12 + Nat\_ANN23 + Nat\_ANN34 + nat + northness + slope,

data = Soton\_good)

Residuals:

Min 1Q Median 3Q Max

-8.6470 -0.5420 0.0516 0.5548 6.2217

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.880e+01 6.432e-02 292.349 < 2e-16 \*\*\*

NDVI -1.960e+00 3.298e-02 -59.418 < 2e-16 \*\*\*

Build\_ANN12 2.958e-02 9.313e-04 31.763 < 2e-16 \*\*\*

Build\_ANN23 1.198e-02 1.358e-03 8.820 < 2e-16 \*\*\*

Build\_ANN34 1.208e-03 1.060e-03 1.139 0.254500

Build 8.454e-03 6.553e-04 12.900 < 2e-16 \*\*\*

dist\_centre -4.445e-06 1.798e-06 -2.473 0.013413 \*

dist\_water 3.040e-05 3.063e-06 9.922 < 2e-16 \*\*\*

eastness 5.935e-02 3.464e-03 17.132 < 2e-16 \*\*\*

elevation -1.443e-02 1.779e-04 -81.119 < 2e-16 \*\*\*

glass 5.840e-02 7.158e-03 8.159 3.42e-16 \*\*\*

Hard\_ANN12 3.782e-03 5.901e-04 6.409 1.47e-10 \*\*\*

Hard\_ANN23 2.896e-03 8.554e-04 3.385 0.000711 \*\*\*

Hard\_ANN34 -1.746e-03 6.296e-04 -2.773 0.005558 \*\*

hard 2.841e-03 6.164e-04 4.608 4.07e-06 \*\*\*

mixed -7.875e-03 5.943e-04 -13.251 < 2e-16 \*\*\*

Nat\_ANN12 -6.548e-03 6.147e-04 -10.653 < 2e-16 \*\*\*

Nat\_ANN23 1.512e-03 8.693e-04 1.739 0.082014 .

Nat\_ANN34 -8.679e-03 5.718e-04 -15.178 < 2e-16 \*\*\*

nat -1.896e-02 6.191e-04 -30.632 < 2e-16 \*\*\*

northness -2.670e-01 3.832e-03 -69.679 < 2e-16 \*\*\*

slope -5.522e-02 9.267e-04 -59.585 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8471 on 114351 degrees of freedom

Multiple R-squared: 0.8643, Adjusted R-squared: 0.8643

F-statistic: 3.47e+04 on 21 and 114351 DF, p-value: < 2.2e-16

1. Which variables are not statistically significant at *p* < 0.001? **[2 marks]**
2. Based only on the information supplied here, do the residuals appear to meet the assumptions of the model and why do you believe this? **[2 marks]**
3. As a percentage, how much variance would you expect this model to explain on another dataset? **[2 marks]**
4. What does the minus sign in front of some of the t- values indicate? **[2 marks]**
5. What change would you make to the code to run this model as a GLM? **[2 marks]**