Problem 1) Match the terms to their definitions. Every term is used exactly once. (In the real midterm, there are fewer terms).

1. Bayesian Information Criterion_____

2. Cross-Validation _____

3. Robust _____

4. Imputation _____

5. Quantile-Quantile _____

6. Principle of Heirarchy / Heredity* _____

7. R-squared **G**

8. Main effect _____

9. Sensitive _____

10. Leave-One-Out _____

11. K-Fold _____

12. Variance Inflation Factor _____

13. Baseline category _____

14. Influence / Leverage _____

A) Plot used to explore any deviations from normality (or other specied distribution) in a collection of values.

B) A regression term made of only one variable, without transformations.

C) Describes a statistic, test, or method that can be greatly affected by small changes.

D) Category designated as the basis of comparison for all the dummy variables of that category.

E) A measure to compare statistical models. Stricter against complexity when N is large.

F) A special case of K-fold cross validation using training sets of N-1 observations.

G) A measure of the proportion of variance explained by a model. Also, a measure to compare regression models that only considers model fit.

H) A method of doing multiple cross-validations where every observation is in a test set once, and a training set every other time.

I) A term for the amount a single observation changes a model. Measured by Cook's distance. Observations with a lot of this harm the predictive ability of a model.

J) Method of checking the predictive ability of a model. Typically uses most of the data to make the model, and the rest to check it.

K) If an interaction is included in a model, both main effects that made it should also be included.

L) A measure of co-linearity of a regression term.

M) Describes a statistic, test, or method that is not greatly affected by changes.

N) General term for methods of replacing missing data.

Problem 2) In data set A, daily measurements were taken from a creek (a very small river) for 40 days in the summer 2014.

In data set B, daily measurements were taken from the same creek, for the same days, but in 2015.

The researcher is interested in predicting flow as a function of the previous day's temperature, precipitation (rain), and humidity.

Consider the results below for each data set.

```
Call:
lm(formula = Flow ~ Temp + Rain + Humid, data = creek2014)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 110.0699    29.6366   3.714 0.000688 ***
Temp         -2.8174     1.2400  -2.272 0.029151 *
Rain          2.6187     1.8068   1.449 0.155896
Humid         0.5151     0.4879   1.056 0.298112

> vif(mod2014)
    Temp     Rain    Humid
1.021788 8.594291 8.548616
```

```
Call:
lm(formula = Flow ~ Temp + Rain + Humid, data = creek2015)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.7910    26.8140   3.013  0.00471 **
Temp         -0.4108     1.0337  -0.397  0.69343
Rain          6.3268     1.3966   4.530 6.25e-05 ***
Humid        -0.5732     0.4390  -1.306  0.19994

> vif(mod2015)
    Temp     Rain    Humid
1.025110 9.005564 8.981679
```

a) Give the regression equation for both models.

b) What does do the VIF values indicate about the data?

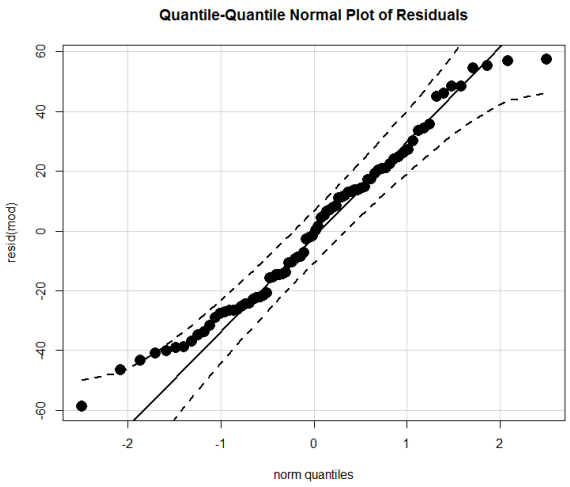
c) Why might the coefficients for Rain and Humid be so different between the two years?

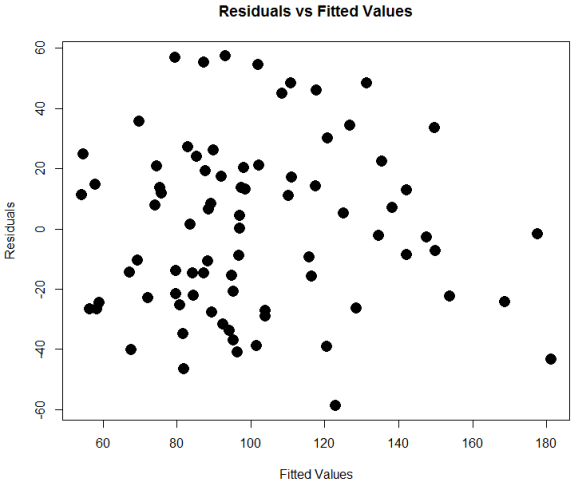
d) If either model was used (or the data combined and a new model made), are the differences between the years going to interfere with flow predictions? Why or why not?

e) Describe a term that could be added to the models to possibly improve them.

f) Consider the following diagnostic plots

(Q-Q plot of residuals), (Residuals vs Fitted).





Are there any potential problems that are showing in these plots?

g) Consider this ANOVA table (from both years):

```
Response: Flow
          Df Sum Sq Mean Sq F value    Pr(>F)
Temp       1   1333    1333  1.6294    0.2057
Rain       1  63064   63064 77.0840 3.526e-13 ***
Humid      1     53      53  0.0645    0.8002
Residuals 76  62177     818
```

What proportion of variance is explained by each of the explanatory terms?

Problem 3) Data set with 2x2 categorical

A glass factory is trying a number of different variables to make glass with the most hardness. They can control two factors: the type of sand they use (fine or coarse), and the type of cooling method used (natural or fast).

The results of their model are below

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)               84.14      10.90   7.719 1.31e-08 ***
SandFine                 113.36      14.21   7.975 6.68e-09 ***
CoolingB-Fast             96.86      15.42   6.283 6.35e-07 ***
SandFine:CoolingB-Fast    82.54      20.10   4.106 0.000285 ***
```

a) What does the intercept mean in this model?

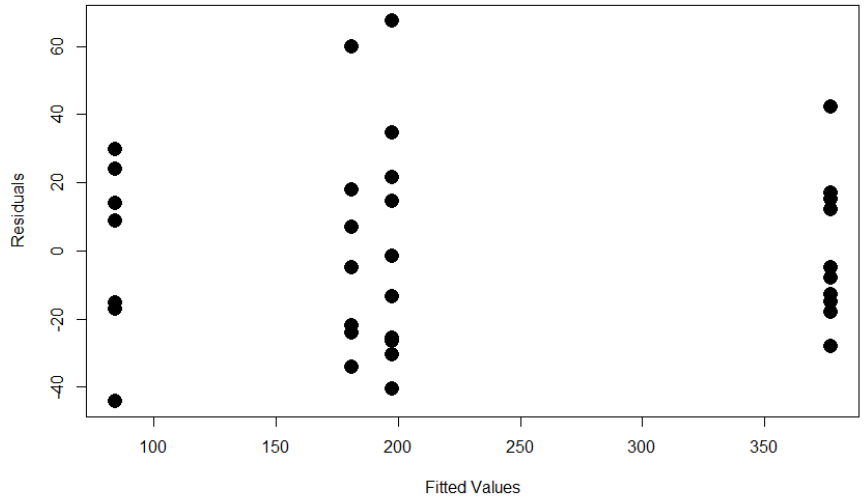
b) What does the 'sandfine' dummy variable main effect mean?

c) Predict the average hardness for...

i)) Using fine sand and fast cooling

ii)) Using fine sand and natural cooling.

d) Consider this residual vs fitted diagnostic plot. Explain why we see this vertical bar pattern in the residuals.



e) The glass shatters for some of the coarse sand cases before it can be tested for hardness, we have some missing data. Can we ignore this missingness? Why or why not?

Problem 4) A global health group is interested in understanding the individual factors behind child mortality in different areas.

They have a dataset of dozens of areas and dozens of possible variables to use, including average temperature, crime level, adult life expectancy, poverty rates, average family size, corruption index, food security, and average income.

a) If this group uses a model with every variable and interaction they have, what problems could they run into? (there are at least three answers, mention two)

b) If some of the variables they select in their model are co-linear, could that interfere with the results the group is interested in?

c) If the group is worried that their model is overfitted, how could they check against that?

d) The group uses the stepwise method and the AIC to select a model. How could they modify this approach to obtain a simpler model? Is it guaranteed to be simpler?

e) If the group were to use the stepwise method and the R-squared to select a model. What sort of model would you expect. Why?

Problem 5) A hospital is interested in whether a new anti-swelling drug works better than a placebo or the old drug after surgery.

They have the treatment ('A-Placebo' or 'DrugOld', or 'DrugNew'),

'Invasive', a continuous rating of 0-10 (0 being very minor, 10 being open heart) for the invasiveness of the surgery, and

'GenHealth', a continuous rating 0-10 of the patient's general health (0 being nearly dead, 10 being very healthy).

"placebo" is marked "A-Placebo" to force it to be the baseline.

First, consider this model with just main effects.

```
Call:
lm(formula = Swelling ~ Treatment + Invasive + GenHealth)
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        23.4913     3.1175   7.535 1.31e-10 ***
TreatmentDrugNew -18.0205     2.8586  -6.304 2.26e-08 ***
TreatmentDrugOld  -1.6528     2.8298  -0.584  0.56106
Invasive           1.2233     0.4228   2.893  0.00508 **
GenHealth         -1.2831     0.4282  -2.997  0.00377 **
```

a) What does the intercept mean in this particular case.

b) What does the dummy variable 'DrugNew' mean in this case?

c) Explain why the dummy variables in this model are more useful than in a simpler model that only looked at the treatment, and not 'Invasive' or 'GenHealth'

d) If you were to find the difference in responses between 'DrugNew' and 'DrugOld' (say, through a TukeyHSD analysis). What would be the value?

e) One of the surgeons spills coffee on their notes and loses the 'GenHealth' rating for some of the patients. What kind of missingness is this?

Problem 6) The same hospital from the last question has decided to add an interaction term to their model and do further analysis.

```
Call:
lm(formula = Swelling ~ Treatment + Invasive + GenHealth + Treatment:Invasive)

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                20.9155     4.3344   4.825 8.24e-06 ***
TreatmentDrugNew           -7.4318     5.7051  -1.303  0.19708
TreatmentDrugOld           -4.6663     5.2595  -0.887  0.37809
Invasive                    1.7083     0.7308   2.338  0.02236 *
GenHealth                  -1.1995     0.4189  -2.864  0.00557 **
TreatmentDrugNew:Invasive  -1.9533     0.9910  -1.971  0.05278 .
TreatmentDrugOld:Invasive   0.6293     1.0047   0.626  0.53317
```

a) Consider the ANOVA table and/or the regression summary table. Is there evidence that at least one of the interactions is significant? How do you know?

```
Response: Swelling
                  Df Sum Sq Mean Sq F value    Pr(>F)
Treatment          2 4083.3 2041.65 22.5426 3.087e-08 ***
Invasive           1  475.9  475.90  5.2546   0.02499 *
GenHealth          1  887.7  887.69  9.8013   0.00257 **
Treatment:Invasive 2  761.2  380.62  4.2026   0.01902 *
Residuals         68 6158.7   90.57
```

b) What does the interaction term DrugOld:Invasive mean?

c) What does the interaction term DrugNew:Invasive mean?

d) Which drug would you recommend to reduce swelling in a patient coming out of a highly-invasive surgery? Why?

e) Would you be surprised if there was a large difference in your patient's swelling level, and what the model predicts? Why or why not?

To be covered:

5-3: Two-way ANOVA

7-1: ANOVA, Regression, and R-squared

7-2: Co-linearity and Perturbations

7-3: Polynomial fits (and VIF)

8-1: AIC and BIC

8-2: Dummy Variables

8-3: Interactions, Stepwise method

9-1,2: No new material

9-3: Q-Q Plots

10-1: Shapiro-Wilks,

10-2: Cross-validation (general, K-fold), Missing Data

10-3: Imputation

*for problem 1:

Heredity is saying 'A and B should be included in a model if the interaction A:B is included'.

Heirarchy is saying 'all main effects should be included in a model if any interactions are'.

The difference between these is too subtle for this class, so we will consider the terms interchangeable.

*for Problem 2: average temperature and humidity over the last 24 hours, total precipitation last 24 hours.