Problem Set 6: Prediction Modeling

Statistics 139 Teaching Staff

Due: November 03, 2023

This assignment is **due Friday, November 3 at 11:59pm**, handed in on Gradescope (remember, there are two separate submissions, one for your pdf, and another for you rmd file). Show your work and provide clear, explanations when asked. **Incorporate the relevant R output in this R markdown file**. Only the key output should be displayed for each problem and the relevant parts should be **highlighted** in some way. Make sure that you write-up any interpretation of R-code in your own words (don't just provide the output).

Collaboration policy (for this and all future problem sets): You are encouraged to discuss the problems with other students, but you must write up your solutions yourself and in your own words. Copying someone else's solution, or just making trivial changes is not acceptable.

Problem 1.

X_1, X_2, and X_3 are three explanatory variables in a multiple regression with $n = 28$ cases. The following table shows the residual sum of squares and degrees of freedom for all models (note: this table is in the file `ABC.csv` to facilitate using R to do the calculations):

Model Variables	**Residual sum of squares**	**Degrees of freedom**
None	8,100	27
X_1	6,240	26
X_2	5,980	26
X_3	6,760	26
X_1, X_2	5,500	25
X_1, X_3	5,250	25
X_2, X_3	5,750	25
X_1, X_2, X_3	5,160	24

(a) Calculate 3 statistics for each model: the estimate of σ^2, AIC, and BIC.

(b) Summarize which model(s) is/are ranked best for each of the 3 statistics from part (a).

(c) Using the residual sum of squares, find the model indicated by forward selection. Start with the model 'None', and identify the single-variable model that has the smallest residual sum of squares, then perform an extra-sum-of-squares F-test to determine if that variable is significant. If it is, continue with the 2 predictor model. Continue until no more significant predictors can be added. Is this procedure guaranteed to find the "best" model (that is, where a determination of "best" is based on residual sum of squares)?

Problem 2.

What are risk factors for elevated blood pressure in the US (measured by systolic blood pressure, in mm Hg)? Several variables from the National Health and Nutrition Examination Survey (NHANES) are stored in `nhanes.csv`.

Descriptions of the variables are included below.

- `systolic`: systolic blood pressure, measured in mm Hg.

- `workhours`: self-reported number of hours in a typical work week.
- `jobtype`: description of job/work situation. The codes `1` through `5` correspond to an employee of a private company/individual for wages or salary, a federal government employee, a state government employee, a local government employee, or self-employed.
- `smoke`: coded `1` if the participant smokes regularly, `0` otherwise.
- `sleep`: self-reported number of hours study participant usually gets on weeknights or workdays; reported for participants aged 16 years or older.
- `active`: coded `1` if participant does moderate or vigorous intensity sports, fitness, or recreational activities; reported for participants 12 years or older.
- `diabetes`: coded `1` if the participant was told by a health professional that they have diabetes, `0` otherwise.
- `alcohol`: coded `1` if the participant has consumed at least 12 drinks of any type of alcoholic beverage in any one yer; reported for participants aged 18 years or older
- `female`: coded `1` if the participant is female, `0` otherwise.
- `age`: age in years at screening. Subjects 80 years or older were recorded as 80 years of age.
- `poverty`: a ratio of family income to poverty guidelines. Smaller numbers indicate more poverty; i.e., a number below 1 indicates income below the poverty level.
- `married`: marital status of study participant; reported for participants aged 20 or older. The codes `1` through `6` correspond to married, widowed, divorced, separated, never married, or living with partner.
- `education`: highest educational level of study participant, reported for participants aged 20 years or older. The codes `1` through `5` correspond to 8th grade, 9 to 11th grade, high school, some college, or college graduate.
- `race`: reported race of study participant: 1 = Mexican, 2 = Hispanic, 3 = White, 4 = Black, 6 = Asian, or 7 = Other.
- `foreignborn`: coded `1` if participant was not born in the US, `0` otherwise
- `heartrate`: 60 second pulse rate
- `height`: standing height, measured in centimeters.
- `weight`: weight, measured in kilograms.
- `waist`: waist circumference, measured in centimeters.
- `bmi`: body mass index

(a) Explore the data graphically and decide whether the outcome variable (`systolic`) or any predictor variable(s) need to be transformed. Make sure you define any categorical variables as factors in R. If you decide to transform the response, use this transformed version as the response/outcome variable for all future models.

(b) Fit a model with 'main effects' of all available predictors in their transformed states (call this **model1**). Your model should have 2332 degrees of freedom associated with the residuals (unless you used exotic transformations). Identify significant predictors (ignoring multiple comparisons).

(c) Use the **backward** variable selection procedure based on AIC to build a prediction model for `systolic` (transformed appropriately), starting from a model with all main effects. You may find the function `step()` helpful. There is no need to report the intermediate output of the `step()` function in your write-up, just report this model's coefficient estimates (not the full `summary` output), R^2, and AIC. Call the resulting model **model2**.

The function `lm()` has some useful shortcuts for specifying formulas. For example, to include the main effects of all variables in the dataset `MyData`:

```
lm(y ~ ., data = MyData)
```

To include all variables and their pairwise interactions:

```
lm(y ~ .^2, data = MyData)
```

(d) Next, run a **forward** variable selection procedure starting with **model2**, with the upper scope for the final model set to include all the two-way interaction terms for the variables in **model2**. Report this model's coefficient estimates, R^2, and AIC. Call this **model3**.

Note: The predictors from **model2** can be printed in a list in R via the command `model2$terms[[3]]`. This forward variable selection can be performed using the `step()` function as follows, where `interactionModel` is the `lm` fit with all variables from **model2** and their interactions:

```r
step(model2, scope = list(upper = formula(interactionModel)), direction = "forward")
```

(e) Finally, use a combined **stepwise** procedure to perform model selection. Start with a model with all main effects and specify the intercept-only model (**model0**) as a lower limit model and a full model including all two-way interactions of *all* possible predictor variables as the upper-limit as shown below (call this the **fullInteractionModel**). Report this model's coefficient estimates, R^2, and AIC. Call this **model4**. Note, this may take a minute or two… the use of the code chunk option `cache=TRUE` can be helpful so you do not need to wait for this to run every time you want to knit the file into a pdf.

```r
step(model1, scope = list(lower = formula(model0), upper = formula(fullInteractionModel)),
     direction = "both")
```

(f) Select a best final model among 5 models based on their AICs: **model1** through **model4** and the **fullInteractionModel**. Perform a brief model check of assumptions on your selected model.

(g) Use the model chosen in part (f) to interpret the association of systolic blood pressure with the variable `female`. If `female` is not in your chosen model in any fashion, interpret what that means.

Problem 3.

Use cross-validation to compare three different models from the previous problem (as defined below) to predict your (transformed) systolic response variable.

1. **interactionModel** from 2(d)
2. **model3**
3. **model4**

For each of the three models, do the following:

- For 100 iterations, randomly select 2,000 observations on which to train each model, and keep the remaining observations as the validation set in each iteration.
- Predict both systolic blood pressure and log-transformed systolic blood pressure for all observations in the validation sets, and save the residual sums of squares of the validation set in each iteration (for the 2 versions of the response: log and untransformed).

In your solution, summarize your results as to which of the three models performs best in predicting systolic blood pressure and which performs best in predicting log-transformed systolic blood pressure. Be sure to address whether the choice of best model agree with your selection from the problem 2.

Note: if you were to actually then use the best model for future observations, you should always refit the chosen model on all the data sampled, not just one training set (here, of $n = 2000$).

Problem 4.

This problem is intended to investigate the optimal choice of train-validation splitting ratios in a model comparison problem. We will be comparing k-fold cross-validation for $k \in \{2, 10, 25\}$. For $n = 50$ (and eventually also for $n = 500$), create data based on the following data-generating process:

i) Sample $Z_1, Z_2, Z_3, X_1, ..., X_{10}$ all independently from the standard normal distribution.

ii) Sample $(Y|\vec{Z}, \vec{X}) \sim N(1 \cdot Z_1 + 3 \cdot Z_2 + 9 \cdot Z_3, 5^2)$ independently from each other.

*Note, it is more efficient to sample all Z_1, Z_2, Z_3 and $\vec{X}$ for each iteration from one `rnorm` function call, and then reorganize them into the separate variables for data creation and model fitting.

For each of `nsims=200` iterations, perform k-fold cross-validation (for each of the 3 choices of k mentioned above) to determine which of the 4 following models is the best out-of-sample prediction model separately for each choice of k:

1) `lm(Y ~ Z3)`
2) `lm(Y ~ Z2 + Z3)`
3) `lm(Y ~ Z1 + Z2 + Z3)`
4) `lm(Y ~ Z1 + Z2 + Z3 + X1 + X2 + ... + X10)`

In the end for each choice of k, you should determine which of the 4 models above is *best* for 200 iterations.

```
# defining your own simulation function could be useful but not required.
# but much sure your approach is working outside of the function before defining one
# my.sim = function(n = 50, nsims = 200, k = c(2,5,25), seed = NA)

# you simulation code here
```

(a) How often is each model selected as best (as measured by average validation set squared error)? Provide a 4 x 3 table that displays the number of times each model is chosen where the columns represent the choice of k and the rows represent each model mentioned above. Note: the columns should sum up to `nsims=200`.

(b) Interpret the results: is there a clear winner for the choice of k here? How often is each cross-validation choosing the *correct* model? Do they tend to favor the underfit or overfit models? Explain.

(c) Rerun the simulation above for $n = 500$. Provide the analogous 4 x 3 table to part (a) above.

(d) Compare the results for the $n = 50$ and $n = 500$ cases. Is there a clear winner for the choice of k when $n = 500$? How does sample size affect the choice of the *underfit*, *correct*, and *overfit* models (in this context)? Explain.

(e) What is missing/lacking in this small simulation study? How would you modify or extend this simulation to better *investigate the optimal choice of train-validation splitting ratios in a model comparison problem*? Explain your choice(s) in up to 8 sentences (do not implement these changes!!!).