STAT 423 Homework 4

1. **Model Selection:** For this problem you will use the file `hw4_problem1.csv`. Using `R`, you will perform 'best subset selection', 'forward step-wise selection', 'backward step-wise selection' and 'bi-direction step-wise selection'.

(a) For the 'best subset selection', use Mallow's C_p, AIC and BIC to determine the best subset (also state which predictors are selected by each criterion for the best subset).

In this sub-part, for the 'best subset selection' method, we will use the Mallow's C_p, AIC and BIC selection criterion to determine the best subset of predictors. In particular, we will state which predictors are selected by each criterion for the best subset. Since the `regsubsets()` function in R cannot compute AIC, we will do this by hand. Below is the output from running the `regsubsets()` function, it will give us information on which variables are selected in each model size.

```
##    (Intercept)   X1    X2    X3    X4    X5    X6    X7    X8    X9   X10
## 1         TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2         TRUE TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 3         TRUE TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 4         TRUE TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
## 5         TRUE TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE
## 6         TRUE TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE
## 7         TRUE TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE
## 8         TRUE TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## 9         TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## 10        TRUE TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

The formatted output of the 'best subset selection' procedure is outputted below for each selection criterion.

```
##    Mallow's Cp      AIC        BIC
## 1  143.7336663 350.0381  -22.04076
## 2   51.4911399 306.3785  -63.28162
## 3   11.3960328 275.4342  -91.80699
## 4   -0.1837246 263.2411 -101.58133
## 5    1.2736277 264.6202  -97.78335
## 6    3.1856803 266.5191  -93.46557
## 7    5.0916048 268.4109  -89.15497
## 8    7.0342058 270.3448  -84.80224
## 9    9.0125586 272.3198  -80.40834
## 10  11.0000000 274.3054  -76.00398
```

As can be seen from the above table, the Mallow's C_p, AIC and BIC selection criterion gives the same result; the model with 4 predictors. Namely, this model includes the intercept, $X1$, $X2$, $X3$, and $X7$.

(b) State the predictors selected by 'forward, backward and bi-direction step-wise selection', respectively.

In this sub-part, we will use the AIC selection criterion, as well as the 'forward, backward and bi-direction step-wise selection' procedures to select the predictors for a model. This is done using the `step()` function from the `MASS` package below. Due to the fact that all of the methods give the exact same result, I will show the code to show that the different methods were in fact used. We will start with using the 'forward selection' procedure.

```
# Fit the full model
fit_full <- lm(Y~., data=hw4_problem1)

# Fit the empty model
fit_empty <- lm(Y~1, data=hw4_problem1)

# Run the forward selection procedure
forward <- step(fit_empty, dir="forward", scope=list(upper=fit_full, lower=fit_empty))
```

```
## (Intercept)          X1          X2          X3          X7
##  -1.2629855   1.1693988   1.2055564   0.7796433  -0.4368278
```

As selected using AIC, the predictors selected are the intercept, $X1$, $X2$, $X3$, and $X7$. We will now use the 'backward selection' procedure. We already did this in part (a) so we expect to get the same results.

```
# Run the backward selection procedure
backward  <- step(fit_full, dir="backward")
```

```
## (Intercept)          X1          X2          X3          X7
##  -1.2629855   1.1693988   1.2055564   0.7796433  -0.4368278
```

As selected using AIC, the predictors selected are the intercept, $X1$, $X2$, $X3$, and $X7$. These are the same as in part (a), as well as the same selected using 'forward selection'. In fact the parameter estimates are even the same. We will now use the 'bi-direction selection' procedure.

```
# Run the hybrid selection procedure
both  <- step(fit_full, dir="both", k=2)
```

```
## (Intercept)          X1          X2          X3          X7
##  -1.2629855   1.1693988   1.2055564   0.7796433  -0.4368278
```

As selected using AIC, the predictors selected are the intercept, $X1$, $X2$, $X3$, and $X7$. These are the same as we found using 'forward and backward selection'. In fact the parameter estimates are even the same. All of the selection procedures so far have selected the intercept, $X1$, $X2$, $X3$, and $X7$ to be included in the "best" model.

(c) What are the flaws of step-wise selection?

In this sub-part, we will point out the flaws of step-wise selection. As mentioned in the notes, the main reason why forward, backward, and forward-backward hybrid procedures are flawed is because all of these procedures aren't guaranteed to select the best possible model, instead they all find a "local" optimum. Another reason why these selection procedures are flawed, is that they can be very time-consuming/computationally expensive for models with a large number of variables (as I learned when using this function on a model made during the project).

(d) Consider the ten models (full to empty). Which of them achieves the best leave-one-out cross-validation score?

In this sub-part, we will consider the eleven models (full to empty) and determine which of them achieves the best leave-one-out cross-validation score. This will be done using the custom `loocv.lm` function given in the specification. Furthermore, we will use the models recommended by `regsubsest` from part (a). Below we will output the *LOOCV* scores for the eleven models (full to empty).

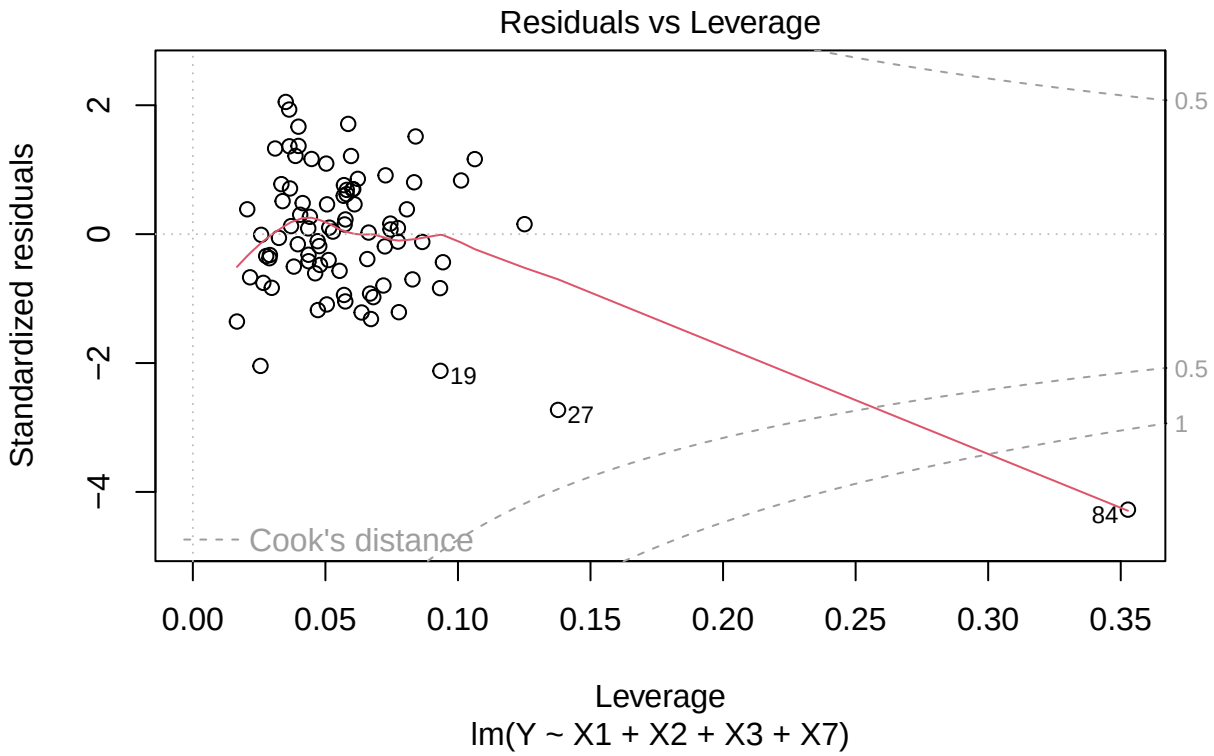
```
##       Number of Predictors LOOCV Scores
##  [1,]                    0     5.626735
##  [2,]                    1     3.969281
##  [3,]                    2     2.361697
##  [4,]                    3     1.616963
##  [5,]                    4     1.415879
##  [6,]                    5     1.432434
##  [7,]                    6     1.467876
##  [8,]                    7     1.511346
##  [9,]                    8     1.552873
## [10,]                    9     1.601774
## [11,]                   10     1.636860
```

Based on the above output, the lowest *LOOCV* score is 1.415879, and corresponds with the model containing 4 predictors (5 parameters). This model in particular corresponds to `Y~X1+X2+X3+X7`, just as we got in all of the previous parts.

(e) Add the following data point to the data set: `c(10, 5, 8, 4, 0, 1, 0.5, 1, 0.8, 0.9, 1.23)`. Use the above selected model on the new data set containing 84 observations (if the methods gave different, choose one of them arbitrarily). Explore the residuals. Is the newly added point a leverage point and/or an outlier?

In this sub-part, we will add the following data point to the data set: `c(10, 5, 8, 4, 0, 1, 0.5, 1, 0.8, 0.9, 1.23)`. Furthermore, we will use the above selected model on the new data set containing 84 observations and explore the residuals. After this is done, we will decide if the newly added point is a leverage point and/or an outlier. The model we will be using is `Y~X1+X2+X3+X7`, which is the model chosen in part (d).

Now that the model has been fit we will analyze the residuals to see if the new data point can be considered a leverage point/outlier. To do this we will use the `plot()` function in R to obtain the 'Residuals vs Leverage' plot. This is done below.



As can be seen above, our new data point, with the index 84, has a high standardized residual, leverage value, and Cook's distance value. Just based on this plot alone, it is safe to say that this new data point is a leverage point and outlier. However, we will do a deeper dive into these values to ensure that they meet the heuristics for a leverage point/outlier.

As computed in R above, the standardized residual of the new data point is -4.2745767, which is considered quite a large residual. Furthermore, the leverage threshold is calculated as $2(p+1)/n$, which for this data set is 0.1904762. The leverage of the new data point was calculated as 0.3527176, which is greater than this threshold (in fact it is the only data point with a leverage value above this threshold). The large standardized residual and leverage is quite concerning. Furthermore, the Cook's distance threshold is calculated as $D_i \geq F_{0.5,p+1,n-p-1}$, which for this data set is 0.9261808. The Cook's distance of the new data point is 1.9913587, which is greater than this threshold.

Therefore, based on the diagnostic plot, as well as the numeric thresholds, it is clear that the new data point can be considered an outlier/leverage point.

2. **Logistic Regression for Binary Data:** A car manufacturer instructed a market research company to analyze which families are going to buy a new car next year using a logistic regression model. The data stems from a random sample of 33 families from an agglomeration area. Assessed variables cover the yearly household income (in 1000 US $) and the age of the oldest car in the family (in years). 12 months later, interviewers assessed which families had bought a new car in the meantime. The data is available in the file `car.RDS` on Canvas.

(a) Perform a logistic regression and report the fitted regression equation.

In this sub-part, we will run a logistic regression on the above data and report the fitted regression equation. This regression model will be fit using `glm()` in R.

```
##
## Call:
## glm(formula = as.factor(purchase) ~ income + age, family = "binomial",
##     data = hw4_problem2)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.73931    2.10195  -2.255   0.0242 *
## income       0.06773    0.02806   2.414   0.0158 *
## age          0.59863    0.39007   1.535   0.1249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 44.987  on 32  degrees of freedom
## Residual deviance: 36.690  on 30  degrees of freedom
## AIC: 42.69
##
## Number of Fisher Scoring iterations: 4
```

As can be seen from the above R output from the `glm()` function, the fitted regression equation, in terms of $P(Y = 1)$, which is the probability of a family buying a new car, is

$$P(Y = 1) = \frac{1}{1 + e^{4.73931 - 0.06773x_{\text{income}} - 0.59863x_{\text{age}}}}$$

In terms of the logit function, which represents the log odds ratio of a family buying a new car, the fitted regression equation is

$$\log\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = -4.73931 + 0.06773x_{\text{income}} + 0.59863x_{\text{age}}$$

Where $P(Y = 1)$ is the probability of a family buying a new car.

(b) Estimate $\exp(\hat{\beta}_{income})$ and $\exp(\hat{\beta}_{age})$ and give an interpretation of these estimates.

In this sub-part, we will estimate $\exp(\hat{\beta}_{income})$ and $\exp(\hat{\beta}_{age})$ and give an interpretation of these estimates.

As found in the previous sub-part, $\hat{\beta}_{income}$ was 0.06773, hence we can see that $\exp(\hat{\beta}_{income})$ is 1.070076. This parameter represents the change in the odds ratio associated with a one-unit increase in the predictor variable. In the context of the problem, this coefficient represents the change in the odds of a family buying a new car with a one unit increase in the yearly household income, holding the age of the oldest car constant. In particular, as the yearly household income increases by 1000 U.S. dollars (a one unit increase), we estimate that the odds of a family buying a new car increases by a factor of 1.070076, holding the age of the oldest car constant.

As found in the previous sub-part, $\hat{\beta}_{age}$ was 0.59863, hence we can see that $\exp(\hat{\beta}_{income})$ is 1.819624. This parameter represents the change in the odds ratio associated with a one-unit increase in the predictor variable. In the context of the problem, this coefficient represents the change in the odds of a family buying a new car with a one unit increase in the age of the oldest care, holding the yearly household income constant. In particular, as the age of the oldest car increases by one year (a one unit increase), we estimate that the odds of a family buying a new car increases by a factor of 1.819624, holding the yearly family income constant.

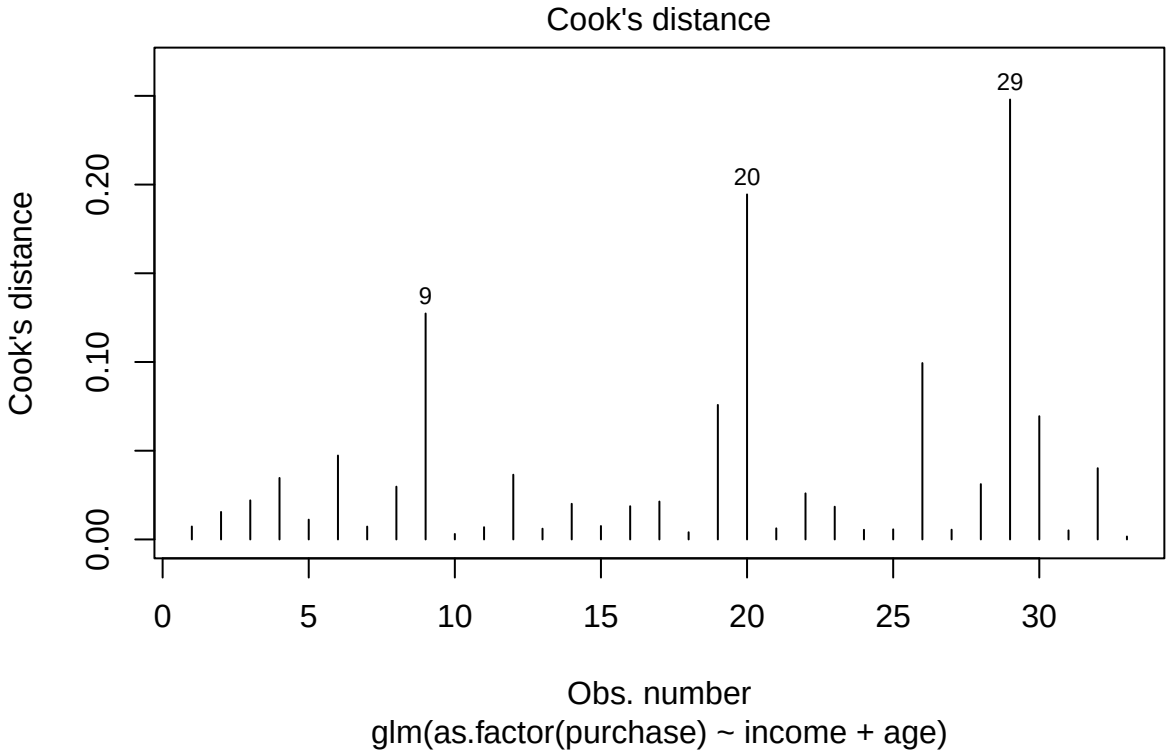
(c) How large is the estimated probability that a family with a yearly household income of 50,000 US $ and whose oldest car is 3 years old will buy a new car?

In this sub-part, we will estimate how large the probability is that a family with a yearly household income of 50,000 U.S. dollars and whose oldest car is 3 years old will buy a new car. This will be done using the `predict()` function in R. This is done below.

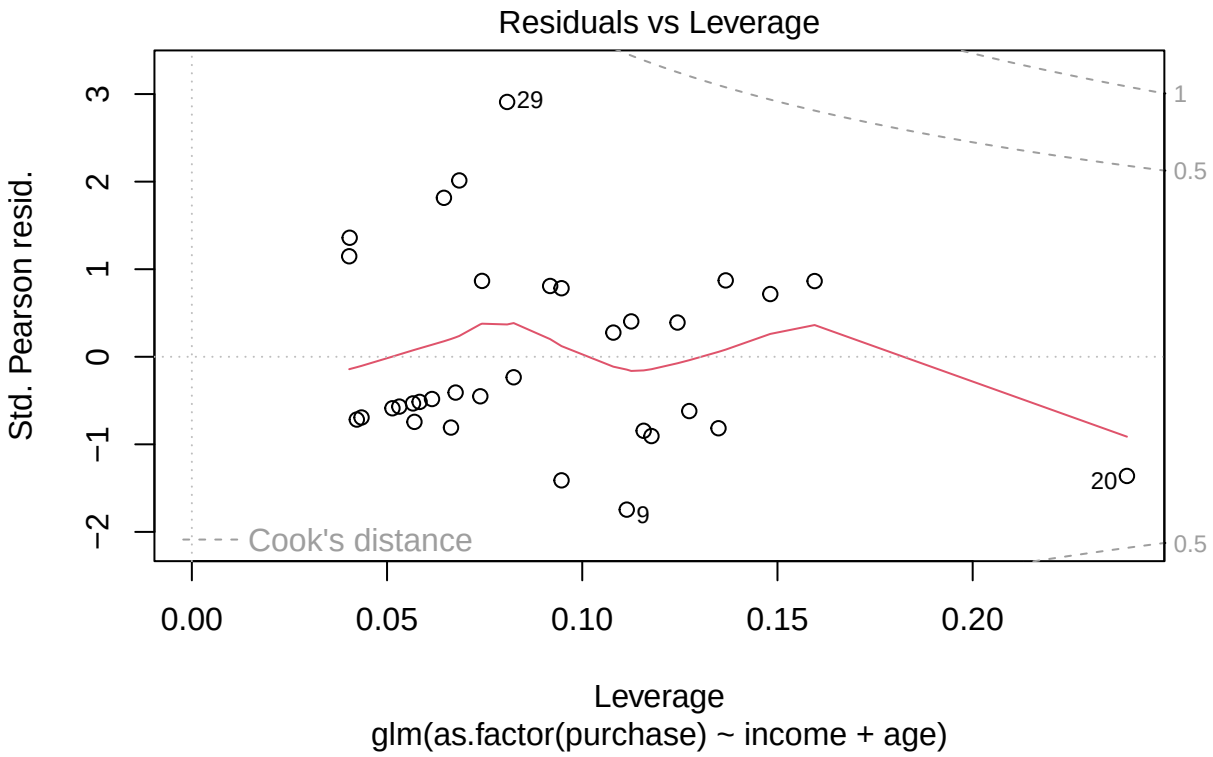
As calculated in R using our logistic regression model, the estimated probability of a family with a yearly household income of 50,000 US $ and whose oldest car is 3 years old buying a new car is 0.6090245.

(d) Check for the presence of points with a large Cook's distance.

In this sub-part, we will check for the presence of points with a large cooks distance. We will do this using the `cooks.distance()` function, as well as the `plot()` function in R. This is done below.



As can be seen by the above plot, the values with large Cook's distances relative to the rest of the data, are observation's with the index 9, 20, and 29. Given that the slides covered no heuristics for what constitutes a "large" Cook's distance for a logistic regression model, we will take these three data points as the one's with "large" Cook's distances. In a deeper analysis, we would look at the aspects of these observations that lead to a high Cook's value, and what further actions we should take based on these findings. To do this, we will now analyze the 'Residuals vs Leverage' plot. This is done below.



Based on the above plot, observation 9 has a moderately large residual and a normal leverage value, it is most likely not an outlier. Observation 20 has a normal residual but has a large leverage, this observation is a leverage point, and could be considered an outlier due to that fact (although the normal residual could prevent this outlier designation). Observation 29 has a normal leverage value and a large residual, this observation is most likely an outlier based on certain heuristics.

(e) Is the predictor `age` significant at the 5% level?

In this problem, we will determine if the predictor `age` is significant at the 5% level. As can be seen from the model fit in sub-part (a), the p-value corresponding to the `age` coefficient was 0.1249. Thus, at the 5% level of significance, we fail to reject the null hypothesis that the coefficient corresponding to the age predictor differs from zero. Thus we have no evidence to say that the age of the oldest car in a family is a significant predictor on if that family will buy a new car within the next year.

(f) Is there a non-negligible interaction between `income` and `age`?

In this sub-part, we will see if there is a non-negligible interaction between `income` and `age`. To do this we will build a new model with an interaction term between `income` and `age`, and check the p-value associated with the coefficient corresponding to the interaction term. This is done using `glm()` in R below.

```
##
## Call:
## glm(formula = as.factor(purchase) ~ income + age + income:age,
##     family = "binomial", data = hw4_problem2)
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.372993   2.862477  -0.829    0.407
## income       0.001326   0.064770   0.020    0.984
## age         -0.303860   0.890512  -0.341    0.733
## income:age   0.028860   0.026493   1.089    0.276
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 44.987  on 32  degrees of freedom
## Residual deviance: 35.404  on 29  degrees of freedom
## AIC: 43.404
##
## Number of Fisher Scoring iterations: 4
```

As can be seen from the above R output from the `glm()` function, the p-value corresponding to the coefficient of the interaction between `income` and `age` was 0.276. Thus, at the 5% level of significance, we fail to reject the null hypothesis that the coefficient corresponding to the interaction between `income` and `age` differs from zero. Thus we have no evidence to say that there a non-negligible interaction between `income` and `age`.