Stat 6021: Homework Set 9

Tom Lever

11/05/22

1. You will continue to use the `birthwt` data set from the `MASS` package for this question. The data were collected at Baystate Medical Center, Springfield, MA in 1986. The data contain information regarding weights of newborn babies as well as potential predictors. Before proceeding, be sure to read the documentation about the data set by typing `?birthwt`. The birthweight of newborns may be related to characteristics of their mothers during pregnancy.

   (a) Which of these variables is categorical? Ensure that R is viewing the categorical variables correctly. If needed, use the `factor` function to force R to treat the necessary variables as categorical.

   Considering discrete variables, Jeffrey Woo suggests to "ask if arithmetic operations can be performed on the variable. If yes, treat as quantitative, if no, treat as categorical. We will use this for the purpose of the HW question."

   The following predictors are discrete and categorical:

   - *low* (0 indicates newborn birthweight is less than 2.5 kg, 1 indicates newborn birthweight is greater than or equal to 2.5 kg),
   - *race* (1 indicates white, 2 indicates black, 3 indicates other),
   - *smoke* (0 indicates non-smoking, 1 indicates smoking),
   - *ht* (0 indicates no history of hypertension, 1 indicates history of hypertension),
   - *ui* (0 indicates no presence of uterine irritability, 1 indicates presence of uterine irritability), and

   The following predictors are discrete and quantitative:

   - *ptl* (value represents number of previous premature labors in $\{0, 1, 2, 3\}$),
   - ftv (value represents number of physician visits during the first trimester in $\{0, 1, 2, 3, 4, 6\}$)

   On loading the `MASS` package and the `birthwt` data frame, R interprets the columns corresponding to these variables as vectors of integers.

```
library(MASS)
library(TomLeversRPackage)
birthwt$low <-
    convert_to_categorical_vector(birthwt$low, c("N", "Y"))
birthwt$race <-
    convert_to_categorical_vector(birthwt$race, c("white", "black", "other"))
birthwt$smoke <-
    convert_to_categorical_vector(birthwt$smoke, c("N", "Y"))
birthwt$ht <-
    convert_to_categorical_vector(birthwt$ht, c("N", "Y"))
birthwt$ui <-
    convert_to_categorical_vector(birthwt$ui, c("N", "Y"))
head(birthwt, n = 3)
```

```
##    low age lwt  race smoke ptl ht ui ftv  bwt
## 85   N  19 182 black     N   0  N  Y   0 2523
```

```
## 86   N  33 155 other     N   0  N  N   3 2551
## 87   N  20 105 white     Y   0  N  N   1 2557
```

(b) A classmate makes the following suggestion: "We should remove the variable *low* as a predictor for the birth weight of babies. Do you agree with your classmate? Briefly explain. Hint: You do not need to do any statistical analysis to answer this question.

I agree. The predictor *low* is dependent on the response / birth weight *bwt*.

```
library(dplyr)
birthwt <- birthwt %>% select(-low)
head(birthwt, n = 3)
```

```
##    age lwt  race smoke ptl ht ui ftv  bwt
## 85  19 182 black     N   0  N  Y   0 2523
## 86  33 155 other     N   0  N  N   3 2551
## 87  20 105 white     Y   0  N  N   1 2557
```

(c) Based on your answer to part 1b, perform all possible regressions using the `regsubsets` function from the `leaps` package. Write down the predictors that lead to a first-order model having the best

i. adjusted R^2,

```
library(leaps)
subset_selection_object <- regsubsets(
    bwt ~ .,
    data = birthwt,
    nbest = 2,
    really.big = TRUE
)
summary_for_subset_selection_object <- summary(subset_selection_object)
adjusted_R2 <- summary_for_subset_selection_object$adjr2
index_of_model_with_maximum_adjusted_R2 <- which.max(adjusted_R2)
coefficients <- coef(
    subset_selection_object, index_of_model_with_maximum_adjusted_R2
)
predictors <- names(coefficients[2:length(coefficients)])
predictors
```

```
## [1] "lwt"       "raceblack" "raceother" "smokeY"    "htY"       "uiY"
```

ii. Mallow's C_p, and

```
Cp <- summary_for_subset_selection_object$cp
index_of_model_with_minimum_Cp <- which.min(Cp)
coefficients <- coef(subset_selection_object, index_of_model_with_minimum_Cp)
predictors <- names(coefficients[2:length(coefficients)])
predictors
```

```
## [1] "lwt"       "raceblack" "raceother" "smokeY"    "htY"       "uiY"
```

iii. Schwartz Bayesian Information Criterion ($BIC_{Schwartz}$)

```
BICSchwartz <- summary_for_subset_selection_object$bic
index_of_model_with_minimum_BICSchwartz <- which.min(BICSchwartz)
coefficients <- coef(
    subset_selection_object, index_of_model_with_minimum_BICSchwartz
)
```

```
predictors <- names(coefficients[2:length(coefficients)])
predictors
```

```
## [1] "lwt"       "raceblack" "raceother" "smokeY"    "htY"       "uiY"
```

(d) Based on your answer to part 1b, use backward selection using the Akaike Information Criterion (AIC) to find the best model. Start with the first-order model with all predictors. What is the regression equation selected?

```
intercept_only_model <- lm(bwt ~ 1, data = birthwt)
full_model <- lm(bwt ~ ., data = birthwt)
step(
    full_model,
    scope = list(lower = intercept_only_model, upper = full_model),
    direction = "backward"
)
```

```
## Start:  AIC=2458.21
## bwt ~ age + lwt + race + smoke + ptl + ht + ui + ftv
##
##         Df Sum of Sq      RSS    AIC
## - ftv    1     38708 75741025 2456.3
## - age    1     58238 75760555 2456.3
## - ptl    1     95285 75797602 2456.4
## <none>               75702317 2458.2
## - lwt    1   2661604 78363921 2462.7
## - ht     1   3631032 79333349 2465.1
## - smoke  1   4623219 80325536 2467.4
## - race   2   6578597 82280914 2470.0
## - ui     1   5839544 81541861 2470.2
##
## Step:  AIC=2456.3
## bwt ~ age + lwt + race + smoke + ptl + ht + ui
##
##         Df Sum of Sq      RSS    AIC
## - age    1     79115 75820139 2454.5
## - ptl    1     91560 75832585 2454.5
## <none>               75741025 2456.3
## - lwt    1   2623988 78365013 2460.7
## - ht     1   3592430 79333455 2463.1
## - smoke  1   4606425 80347449 2465.5
## - race   2   6552496 82293521 2468.0
## - ui     1   5817995 81559020 2468.3
##
## Step:  AIC=2454.5
## bwt ~ lwt + race + smoke + ptl + ht + ui
##
##         Df Sum of Sq      RSS    AIC
## - ptl    1    117366 75937505 2452.8
## <none>               75820139 2454.5
## - lwt    1   2545892 78366031 2458.7
## - ht     1   3546591 79366731 2461.1
## - smoke  1   4530009 80350149 2463.5
## - race   2   6571668 82391807 2466.2
## - ui     1   5751122 81571261 2466.3
```

```
##
## Step:  AIC=2452.79
## bwt ~ lwt + race + smoke + ht + ui
##
##         Df Sum of Sq      RSS    AIC
## <none>               75937505 2452.8
## - lwt    1   2674229 78611734 2457.3
## - ht     1   3584838 79522343 2459.5
## - smoke  1   4950633 80888138 2462.7
## - race   2   6630123 82567628 2464.6
## - ui     1   6353218 82290723 2466.0
```

```
##
## Call:
## lm(formula = bwt ~ lwt + race + smoke + ht + ui, data = birthwt)
##
## Coefficients:
## (Intercept)          lwt    raceblack    raceother       smokeY          htY
##    2837.264        4.242     -475.058     -348.150     -356.321     -585.193
##         uiY
##    -525.524
```

```
best_model <- lm(bwt ~ lwt + race + smoke + ht + ui, data = birthwt)
summarize_linear_model(best_model)
```

```
##
## Call:
## lm(formula = bwt ~ lwt + race + smoke + ht + ui, data = birthwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1842.14  -433.19    67.09   459.21  1631.03
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2837.264    243.676  11.644  < 2e-16 ***
## lwt            4.242      1.675   2.532 0.012198 *
## raceblack   -475.058    145.603  -3.263 0.001318 **
## raceother   -348.150    112.361  -3.099 0.002254 **
## smokeY      -356.321    103.444  -3.445 0.000710 ***
## htY         -585.193    199.644  -2.931 0.003810 **
## uiY         -525.524    134.675  -3.902 0.000134 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 645.9 on 182 degrees of freedom
## Multiple R-squared:  0.2404, Adjusted R-squared:  0.2154
## F-statistic:   9.6 on 6 and 182 DF,  p-value: 3.601e-09
##
## E(y | x) =
##     B_0 +
##     B_lwt * lwt +
##     B_raceblack * raceblack +
##     B_raceother * raceother +
##     B_smokeY * smokeY +
```

```
##     B_htY * htY +
##     B_uiY * uiY
## E(y | x) =
##     2837.26392020575 +
##     4.24154999761198 * lwt +
##     -475.057604364483 * raceblack +
##     -348.150381131656 * raceother +
##     -356.32094981266 * smokeY +
##     -585.193120939356 * htY +
##     -525.523897271075 * uiY
## Number of observations: 189
## Estimated variance of errors: 417239.03742102
## Prediction R2: 0.177290334497079
## Multiple R:  0.490300372976123   Adjusted R:  0.464060895487847
## Critical value t(alpha/2 = 0.05/2, DFRes = 182): 1.9730840773359
## Critical value F(alpha = 0.05, DFR = 6, DFRes = 182): 2.14868632747573
```

Let $\boldsymbol{\beta}_{predictor}$ be a column vector of the coefficients of the non-reference indicator variables associated with predictor *predictor*. Let $\boldsymbol{predictor}$ be a column vector of the non-reference indicator variables associated with predictor *predictor*. The MLR equation selected is

$$\beta_0 = 2837.264$$

$$\beta_{lwt} = 4.242$$

$$\boldsymbol{\beta}_{race} = \begin{bmatrix} -475.058 \\ -348.150 \end{bmatrix}$$

$$\boldsymbol{\beta}_{smoke} = \begin{bmatrix} -356.321 \end{bmatrix}$$

$$\boldsymbol{\beta}_{ht} = \begin{bmatrix} -585.193 \end{bmatrix}$$

$$\boldsymbol{\beta}_{ui} = \begin{bmatrix} -525.524 \end{bmatrix}$$

$$bwt = \beta_0 + \beta_{lwt}\ lwt + \boldsymbol{\beta}_{race} \cdot \boldsymbol{race} + \boldsymbol{\beta}_{smoke} \cdot \boldsymbol{smoke} + \boldsymbol{\beta}_{ht} \cdot \boldsymbol{ht} + \boldsymbol{\beta}_{ui} \cdot \boldsymbol{ui}$$

2. The data for this question are 36 monthly observations on variables affecting sales of a product. The objective is to determine an efficient model for predicting and explaining market share sales, *Share*, which is the average monthly market share for a product, in percent. The predictors are average monthly price in dollars, *price*, amount of advertising exposure based on gross Nielsen rating, *nielsen*, whether a discount price was in effect, *discount* (1 if a discount was in effect, 0 otherwise), whether a package promotion was in effect, *promo* (1 if a promotion was in effect, 0 otherwise), and time in months, *time*.

   (a) Output in the prompt for this homework was obtained after using the `step` function using forward selection, starting with a model with just the intercept term. What is the model selected based on forward selection?

   The model selected based on forward selection is

   $$Share = \beta_0 + \beta_{price}\ price + \boldsymbol{\beta}_{discount} \cdot \boldsymbol{discount} + \boldsymbol{\beta}_{promo} \cdot \boldsymbol{promo}$$

   (b) Your client asks you to explain what each step in the output shown above means. Explain the forward selection procedure to your client, for this output.

   The forward selection procedure determines an efficient multiple linear regression model for predicting and explaining market share sales, *Share*, as a function of various predictors, by considering a base / initial model (e.g., the intercept-only model) and models created by adding predictors, up to a model with a subset of predictors (e.g., the full model). The selected model is described in part 2a. Forward selection is an iterative process. The suboutput for each iteration

consists of a model m to which to consider adding predictors, the Akaike Information Criterion (AIC) for m, and a table with a column of predictors p and a column of the AIC of the model resulting from adding p to m. Considering one suboutput, the predictor p whose addition to m results in the lowest AIC is listed first in the suboutput's table and is added to m. Adding $< none >$, or no predictors to m results in no change of AIC. Adding a predictor higher in the table than $< none >$, which would result in a decrease in AIC for m, is preferred to adding no predictors. Adding no predictors is preferred to adding a predictor lower in the table than $< none >$, which would result in an increase in AIC for m. If $< none >$ is listed first in the table, forward selection ends, and the model associated with the suboutput is selected.

For example, the first suboutput contains the intercept-only model $m : Share = \beta_0$. The AIC for m is -94.8. A table begins with predictor *discount* and AIC -128.137. Predictor *discount* is added to m. The second suboutput contains model \$m: $Share = \beta_0 + \boldsymbol{\beta}_{discount} \cdot \boldsymbol{discount}$. The AIC for m is -128.14. A table begins with predictor *promo* and AIC -129.69. Predictor *promo* is added to m. The last suboutput contains the model in part 2a. The AIC for m is -132.94. Because a table begins with $< none >$, no predictor is added to the model and forward selection ends, with the model in part 2a selected.

(c) Your client asks if he should go ahead and use the model selected in part 2a. What advice to you have for your client?

The model selected in part 2a by forward selection is a first-order model that doesn't consider interactions or higher order terms. You may we to consider interactions or higher-order terms. Regression assumptions for the model selected may not be met. You may wish to check if the following regression assumptions are met. The model selected may not be the best model of the relationship between predictors and/or response, or for prediction. You may wish to consider all MLR models. Since forward selection was used, predictors are added only to a multiple linear regression model; predictors may become insignificant in the context of the multiple linear regression model / all predictors. Performing forward selection using an F statistic instead of an Akaike Information Criterion (AIC), backward selection, bidirection selection, or all possible regressions analysis may yield models different from the model selected.

3. Your client asks you to compare and contrast between R^2 and adjusted R^2, specifically: name one advantage of R^2 over adjusted R^2, and name one advantage of adjusted R^2 over R^2.

   The coefficient of determination R^2 is the proportion of variation in a response that is explained by a multiple linear relationship / predictors. The adjusted R^2 is a corrected goodness-of-fit (model accuracy) measure for multiple linear models that penalizes us for adding terms to the model that are not helpful. R^2 tends to optimistically estimate the fit of the linear regression. It always increases as the number of predictors in the model increases. Adjusted R^2 attempts to correct for this overestimation. Adjusted R^2 might decrease if a specific predictor does not improve the model. Adjusted R^2 is calculated by dividing the residual mean square error (which is the sample variance of the target field) by the total mean square error; The result is then subtracted from 1. Adjusted R^2 is always less than or equal to R^2. A value of 1 indicates a model that perfectly predicts values in the target field. A value that is less than or equal to 0.09 indicates a model that has no predictive value. In the real world, adjusted R^2 lies between 0 and 1.

4. Include the function you wrote to compute the PRESS statistic (Question 2 in Guided Question Set).

   https://github.com/tslever/Tom_Levers_Git_Repository/blob/main/R/TomLeversRPackage/R/calculate_PRESS.R

```
calculate_PRESS
```

```
## function (linear_model)
## {
##     vector_of_residuals <- linear_model$residuals
##     list_of_quantities_for_diagnostics_for_checking_quality_of_regression_fits <- lm.influence(l
```

```
##     diagonal_of_hat_matrix <- list_of_quantities_for_diagnostics_for_checking_quality_of_regress
##     PRESS <- sum((vector_of_residuals/(1 - diagonal_of_hat_matrix))^2)
##     return(PRESS)
## }
## <bytecode: 0x127552190>
## <environment: namespace:TomLeversRPackage>
```