

Stat 6021: Homework Set 9

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, Mass during 1986. The data contain information regarding weights of newborn babies as well as a number of potential predictors. Before proceeding, be sure to read the documentation about the data set by typing `?birthwt`. The goal of the data set is to relate the birthweight of newborns with the characteristics of their mothers during pregnancy.
 - (a) Which of these variables are 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.
 - (b) A classmate of yours 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.
 - (c) Based on your answer to part 1b, perform all possible regressions using the `regsubsets()` function from the `leaps` package (use `nbest=1`). Write down the predictors that lead to a first-order model having the best
 - i. adjusted R^2 ,
 - ii. Mallow’s C_p ,
 - iii. BIC.
 - (d) Based on your answer to part 1b, use backward selection to find the best model according to AIC. Start with the first-order model with all the predictors. What is the regression equation selected?
2. (No R required) 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 the product, in percent. The predictors are average monthly price in dollars, *price*, amount of advertising exposure based on gross Nielson rating, *nielsen*, whether a discount price was in effect, *discount* (1 if discount, 0 otherwise), whether a package promotion was in effect, *promo* (1 if promotion, 0 otherwise), and time in months, *time*.

- (a) The output below is 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?

```
> start<-lm(Share~1, data=data)
> end<-lm(Share~.,data=data)
> result.f<-step(start, scope=list(lower=start,
+ upper=end), direction="forward")
Start:  AIC=-94.8
Share ~ 1
```

	Df	Sum of Sq	RSS	AIC
+ discount	1	1.52953	0.91672	-128.137
+ promo	1	0.22756	2.21870	-96.318
<none>			2.44626	-94.803
+ price	1	0.08693	2.35933	-94.105
+ nielsen	1	0.01288	2.43337	-92.993
+ time	1	0.00469	2.44156	-92.872

Step: AIC=-128.14

Share ~ discount

	Df	Sum of Sq	RSS	AIC
+ promo	1	0.086097	0.83063	-129.69
+ price	1	0.080864	0.83586	-129.46
+ time	1	0.058506	0.85822	-128.51
<none>			0.91672	-128.14
+ nielsen	1	0.041559	0.87516	-127.81

Step: AIC=-129.69

Share ~ discount + promo

	Df	Sum of Sq	RSS	AIC
+ price	1	0.112673	0.71795	-132.94
+ time	1	0.075200	0.75543	-131.10
<none>			0.83063	-129.69
+ nielsen	1	0.025277	0.80535	-128.80

Step: AIC=-132.94

Share ~ discount + promo + price

	Df	Sum of Sq	RSS	AIC
<none>			0.71795	-132.94
+ time	1	0.0110210	0.70693	-131.49
+ nielsen	1	0.0003132	0.71764	-130.95

- (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.
 - (c) Your client asks if he should go ahead and use the models selected in part 2a. What advice do you have for your client?
3. (No R required) Your client asks you to compare and contrast between R^2 and the adjusted R^2 , specifically: name one advantage of R^2 over the adjusted R^2 , and name one advantage of the adjusted R^2 over R^2 .
 4. Include the function your group wrote to compute the PRESS statistic (Question 5 in Guided Question Set).