AMS 380.01: Quiz 3

Due on 03/23

Wei Zhu

Harris Temuri

Problem 1

The data set: For this quiz, we shall use the mtcars dataset, which comes with R. This dataset is on 32 models of car, taken from the 1974 Motor Trend magazine (US). For each car, we have 11 features, expressed in varying units (US units).

Please note that when we mention the mtcars data have 11 variables, we count each categorical variable as a single variable – even though some may be represented by multiple dummy variables. Please also note if at least one of its dummy variables is found significant, the corresponding original categorical variable should be included in the model, even though as a whole it is not significant.

Solution

(a) Please fit the general linear model with the response variable being ‘mpg’ and a single predictor being ‘am’. Please write down your regression model and point out which ‘am’ value group is the baseline group. Is ‘am’ a significant predictor of ‘mpg’ based on your analysis? Please report the p-value.

```r
# 1.1 Finding Linear fit for mpg ~ am
fit1 <- lm(mpg ~ am, data=df)
summary(fit1)
```

$$mpg = 7.245 * I(am1) + 17.147$$

The baseline group is 0. Yes ’am’ is a significant predictor of ’mpg’ as the p-value is 0.000285.

(b) Now you will note that the default baseline group used in part (a) for the categorical variable ‘am’ is the “0” group. Can you rerun part (a) with the “1” group as the baseline?

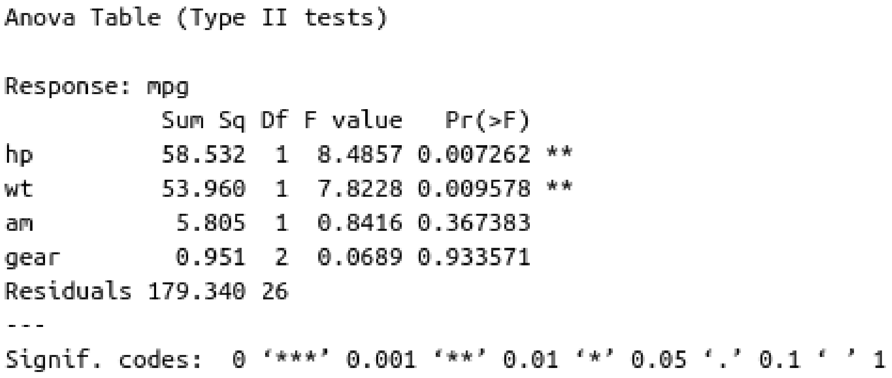
```r
# 1.2 Making 1 as am baseline
df$am <- relevel(df$am, ref = "1")
fit1 <- lm(mpg ~ am, data=df)
summary(fit1)
```

$$mpg = -7.245 * I(am0) + 24.392$$

The baseline group is 1. Yes ’am’ is a significant predictor of ’mpg’ as the p-value is 0.000285.

(c) The categorical variable ‘gear’ in the mtcars data has three levels: “3”, “4” and “5”. Now, please fit the general linear model with the response variable being ‘mpg’ and the predictors being ‘hp’, ‘wt’, ‘am’ and ‘gear’. Please use the Anova() function [in car package] to show the p-values of each variable. Which variables are significant at the significance level of $\alpha = 0.05$? Please use the summary() function to write down the entire regression equation.

```
# 1.3 Finding Linear fit for mpg ~ hp,wt,am,gear
fit2 <- lm(mpg ~ hp + wt + am + gear, data=df)
Anova(fit2)
summary(fit2)
```



```
Anova Table (Type II tests)

Response: mpg
          Sum Sq Df F value   Pr(>F)
hp        58.532  1  8.4857 0.007262 **
wt        53.960  1  7.8228 0.009578 **
am         5.805  1  0.8416 0.367383
gear       0.951  2  0.0689 0.933571
Residuals 179.340 26
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 1: Anova of fit2

The variables that are significant are hp and wt and the rest are not.

$$mpg = -0.03835 * I(hp) - 2.8 * I(wt) - 1.68602 * I(am0) + 0.40668 * I(gear4) + 0.80482 * I(gear5) + 32.55626$$

(d) Report the coefficient of determination for the model in part (c) above – does this statistic indicate a good linear model fit?

The value of the coefficient of determination is 0.8407. I would say that this indicates a pretty good linear model fit.

(e) What assumptions are necessary for the regression in part (c) above? Please test these assumptions.

We need to test for homoscedasticity and normality.

```
# 1.5 Testing assumptions for homoscedasticity and normality
shapiro.test(residuals(fit2))
lmtest::bptest(fit2)
```

The p-value for for the homoscedasticity test was above 0.05 so the variance is not constant and heteroscedasticity is present. The p-value for the normality test is 0.055 which is around 0.05 so the residuals are not normal.

Problem 2

Now we learn how to do variable selection using the best subset method, and the stepwise variable selection method. We shall use the same built-in R dataset 'mtcars' – and please be sure to tell R those categorical variables using the factor function

Solution

(a) We shall use 'mpg' as the response variables, and there are a total of 10 regressors to choose from. First, please use the R function regsubsets() [leaps package] for best-subset variable selection to identify different best models of different sizes ranging from 1 to 5.

```
# subset variable selection
 models <- regsubsets(mpg~., data=df, nvmax=5)
summary(models)
```

The best model with one variable is mpg wt

The best model with two variables is mpg hp + wt

The best model with three variables is mpg wt + qsec + am0

The best model with four variables is mpg cyl6 + hp + wt + am0

The best model with five variables is mpg cyl6 + hp + wt + vs1 + am0

(b) Please use the 5-fold cross-validation to select the best overall model from all 5 best subset models identified in part (a) above. Please write down the equation of this best overall model.

(c) Please use the R function stepAIC() [MASS package] to identify the best model using the stepwise variable selection method. Please write down the equation of this best overall model.

```
res.lm <- lm(mpg ~., data= df)
step <- stepAIC(res.lm, direction = "both", trace = FALSE)
step
```

The best function is:

$$mpg = 35.51754 - 3.03134 * \mathit{cyl6} - 2.16368 * \mathit{cyl8} - 0.03211 * hp - 2.49683 * wt - 1.80921 * \mathit{am0}$$