Regression Models Course Project - Motor Trend Data = ‘mtcars’ Miles Per Gallon Analysis

james c walmsley

12/1/2016

Executive Summary:

Using linear regression model variations, including the step function, to gauge model fit, we identified model fstp <- lm(mpg ~ (wt + qsec + am), data= mtcars) that provides an 84.966 Rˆ2 value indicating a reasonalby good fit and another model fit fnm6 <- lm(mpg ~ I(cyl + disp + hp + drat + wt + factor(am), data = mtcars) devloped using the multivariate nested approach followed by and anova table test to check for multicollinearity which also produced a reasonably good fit with an Rˆ2 value of 85.13%. It must also be noted that vehicle weight is highly correlated (-86.77%) with mpg ratings and transmission type is relatively highly correlated with vehicle wieght at (69%).

Problem Statement:
Backround information, problem statement & questions of interest: Background situation: As a member of a team of data analysts for the Motor Trend Magazine we have been given a data set called “mtcars” and asked to answer some questions of interest concerning differences between automatic and manual transmissoin types in regards to associated mpg or miles per gallon ratings in this data set. Assumptions: The given data set (a sample of a unknown larger population) for this analysis consists of (iid) independent and identically distrubuted random varialbles for 32 subjects (vehicles) of 11 observations or variables.

```
Questions of interest for Motor Trend Magazine:

Q1 "Is an automatic or manual transmission better for 'mpg'"
        or which type of transmission is associated with better mpg or gas mileage ratings?

        A1. The mean "mpg" rating of all vehicle models including both transmission types
        is 20.09 mpg with a 95% confidence interval of 17.92 mpg to 22.26 mpg.

Q2 "Quantify the mpg difference between automatic and manual transmissions"
        What is the expected difference in mpg rating and how accurate
        is this estimate based on the given data?

        A2. The mean "mpg" of models with automatic transmisions is 17.15 mpg, with a 95%
        confidence interval of between 14.85 mpg to 19.44 mpg and vehicles with manual
        transmisions have a mean of 24.39 mpg for a difference of 7.24 mpg with a 95 %
        confidence interval of between 18.49 mpg and 30.29 mpg
```

Analysis Considerations:

```
    Descriptive - any(is.na()), str() & summary(),
Exploratory - pairsPlots(), histograms(), boxPlots(), barPlots() QQ_Plots & multiple plots
Regression Models Analysis -
        SLR, BiVariate Regression, Multivariate Linear Regression, model selection,
        adjustments, residuals (predict fit, residual fit (-1)), coeficients, correlation,
        confidence intervals, influence & leverage,
Diagnostics
        See section on Diagnostics
Final Model Selection strategy
        Beginning with the simmple linear regression using just one predictor (am),
        Use a bivariate model, Use a multivariate model, Use an intercept adjusted multivariate mode
        Use a multivariate model removing a suspected key regressor, Use the nested multivariate pro
        Use the step(function) both directions process, Usee different combinations and leave out fi
        the results of the step(function), Choose the best fit which is understandable and easy to e
```

Technical Environment:

```
System - session Info; Set the Working Directory; Record the System & Session Info; Check for requi
```

Raw Data:

```
Clean up work space, import the data & check for missing values
Overview: Motor Trend 'mtcars' data set:
```

A data frame with 32 observations on 11 variables.

[, 1] mpg Miles/(US) gallon [, 2] cyl Number of cylinders (4,6,8) [, 3] disp Displacement (cu.in.) [, 4] hp Gross horsepower [, 5] drat Rear axle ratio [, 6] wt Weight (1000 lbs) [, 7] qsec 1/4 mile time [, 8] vs V/S (0 = vee-block, 1 = straight-block) [, 9] am Transmission (0 = automatic, 1 = manual) [,10] gear Number of forward gears (3:5) [,11] carb Number of carburetors (1:4,6,8)

Processed Data:

```
Factor columns 2 & 8:11 (cyl,vs,am,gear,carb) into levels
```

Descriptive Statistics:

Exploratory Analysis:

```
See Appendix A. Figures (pairs-plot, histogram, box-plot)
```

Statistical Modeling:

```
Multivarite Linear Model Finding Best Fit with Step function:
```

Preliminary findings:

```
# Quesions of interest: & interpretation of results:
A Revisit the Question - Considering all regressors:
        A. Is an automatic or manual transmission better for mpg
                The results of using multiple linear regression techniques
                sugggest that manual transmissions are associated with better
                mpg ratings than automatic transmissions
        B. Quantify the MPG difference between automatic and manual transmissions
                On average manual transmissions provides 24.39 mpg which is 7.24 mpg
                more than the 17.15 mpg average of the automatic transmission models
B Primary result
        A. Are any other regressors significantly correlated with mpg rating?
                a. model fnm6 = factor(am) + cyl + disp + hp + drat + wt
                this model has an R^2 value of 85.13%
        B. Further testing
                a. using the step function in both directions selects wt, qsec and am
                as good predictors with an 84.96% R^2 value indicating very good
                predictability using this set of regresssors

C Direction, Magnitude, Uncertainty
        A.
```

Multivarite Linear Model Finding Best Fit with Step function:

```r
library(stats);library(MASS);library(ggplot2)
fstp <- lm(mpg ~ ., data = mtcars)
stp <- step(fstp, trace = FALSE)
coef(summary(stp))
```

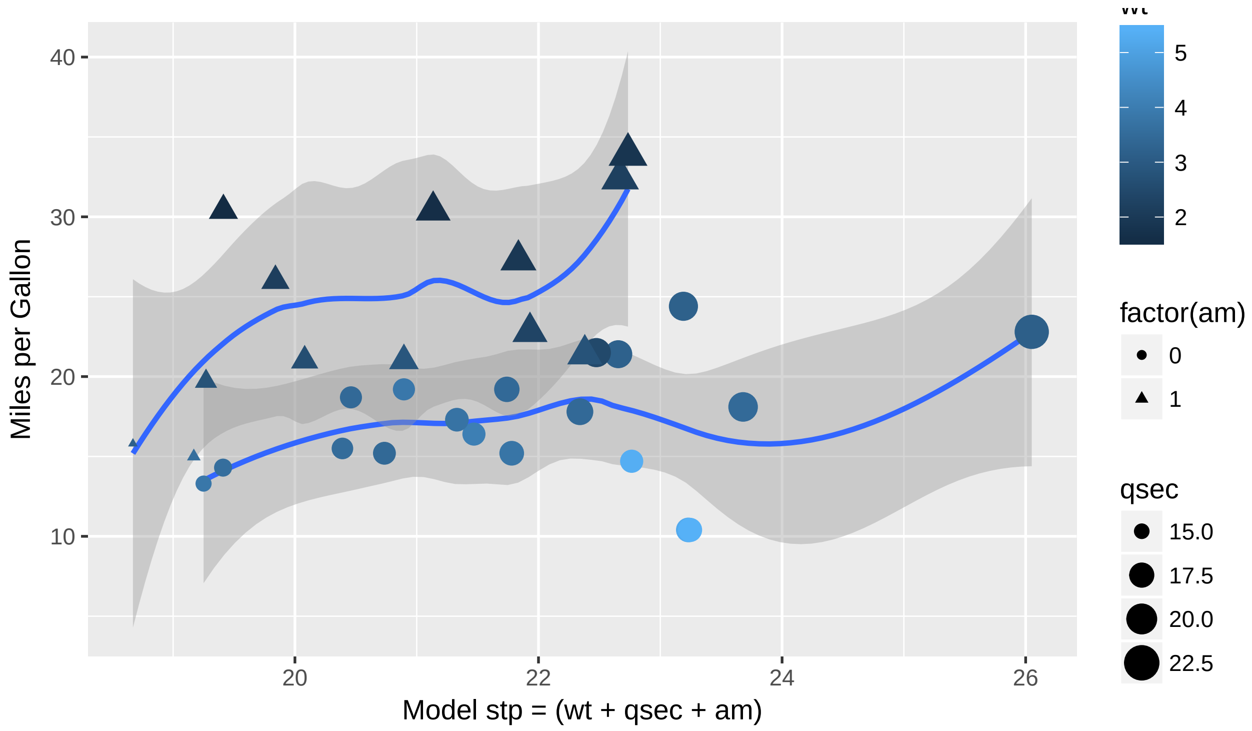
```
         Estimate Std. Error   t value     Pr(>|t|)
```

(Intercept) 9.617781 6.9595930 1.381946 1.779152e-01 wt -3.916504 0.7112016 -5.506882 6.952711e-06 qsec 1.225886 0.2886696 4.246676 2.161737e-04 am 2.935837 1.4109045 2.080819 4.671551e-02

```r
summary(stp)$r.squared
```

[1] 0.8496636

```r
x <- (mtcars$wt + mtcars$qsec + mtcars$am)
par(mfrow = c(1, 1), mar = c(4,4,4,2))
g <- ggplot(mtcars, aes(x = x, y = mpg),)
g <- g + xlab("Model stp = (wt + qsec + am)")
g <- g + ylab("Miles per Gallon")
g <- g + geom_smooth(aes(method = "lm", shape = factor(am)))
g <- g + geom_point(aes(shape = factor(am), size=qsec, colour=wt))
g
```



D Context A. It should be noted that vehicle weight has a strong negative correlation to mpg ratings (-86.76%) and the weight of vehicle models with manual transmissions range from 1.513tons to 3.570 tons and the weight of vehicles with automatic transmissions range from 2.465 tons to 5.424 tons

```
E Implications - Congruence with existing knowledge?
        A. Sedan, Sports, Luxury
        Generally accepted expectations of mpg ratings are that sports and luxury models
        typically will have lower mpg ratings than sedans
```

Diagnostics:
Diagnostic tests were conducted on model results in accordance with the plan for analysis considerations. Besides several vehicle models exhibiting influence as denoted in the Residuals vs Fitted plot such as the Chrysler Imperial, Fiat 128 & the Toyota Corrolla.

```
In the Normal QQ plot of Theoretical Quantiles vs Standardized residuals the Ford Pantera L
fell outside minus two standard deviations and the Chrysler Imperial and Fiat 128 fell
outside the range of positive two standard deviations of the Theoretical Quantiles.

Looking at the results of the Scale-LOcation plot of the sqrt of standardized residuals vs
the Fitted Values three models exhibited notable results beyond 1.2 sd from center of fit.

On the plot for cook's distance; Ford Panter L exhibits a significant cook's distance value
```

Hypothesis Test:

```
        H0 = mean(mpg[am==automatic])  =   mean(mpg[am==manual]) (REJECT)
        Ha = mean(automatic transmission)mpg  !=  mean(manual transmission)mpg (ACCEPT)
```

```r
t.test(mpg ~ factor(am), paired = FALSE, var.equal=FALSE, data = mtcars)
```

```
Welch Two Sample t-test
```

data: mpg by factor(am) t = -3.7671, df = 18.332, p-value = 0.001374 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -11.280194 -3.209684 sample estimates: mean in group 0 mean in group 1 17.14737 24.39231 Welch Two Sample t-test

data: mpg by factor(am) t = -3.7671, df = 18.332, p-value = 0.001374 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -11.280194 -3.209684 sample estimates: mean in group 0 mean in group 1 17.14737 24.39231

Inference & Prediction:

```r
data("mtcars")
levels(mtcars$wt) <- 2
fw <- lm(mpg ~ wt, data = mtcars)
new.weights <- c(1.750, 2.125, 2.750, 3.475, 4.125, 4.700)
predict(fw, newdata = data.frame(wt = new.weights))
```

```
##        1        2        3        4        5        6
## 27.93230 25.92812 22.58783 18.71309 15.23918 12.16611
```

Interpretation of Results:

Appendix A. Figures:

need to review again possibly condense to one plot