11/28/2023
Andrew Alford
IST 686 Quantitative Reasoning for Data Science

Homework 8

1. **The data sets package in R contains a small data set called mtcars that contains n = 32 observations of the characteristics of different automobiles. Create a new data frame from part of this data set using this command: myCars <- data.frame(mtcars[,1:6]).**

This question is quite simple, to create a data frame from the subset of the mtcars data set, I used the code outlined in the question:

```
> # Question 1
> data("mtcars")
> myCars <- data.frame(mtcars[,1:6])
> View(myCars)
>
```

2. **Create and interpret a bivariate correlation matrix using cor(myCars) keeping in mind the idea that you will be trying to predict the mpg variable. Which other variable might be the single best predictor of mpg?**

First things first, I created the bivariate correlation matrix using the function outlined in the question and got this result:

```
> cor(myCars)
            mpg        cyl       disp         hp       drat         wt
mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.6999381  0.7824958
disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.7102139  0.8879799
hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.4487591  0.6587479
drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.0000000 -0.7124406
wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.7124406  1.0000000
>
```

Regarding the interpretation of the matrix, it revealed several notable relationships with the 'mpg' variable. The variable 'wt' shows the strongest correlation with 'mpg', exhibiting a high negative correlation coefficient of -0.87. This suggests that as the weight of the car increases, its fuel efficiency tends to decrease. Other variables also demonstrate significant correlations with 'mpg'. 'Cyl' and 'disp' have correlation coefficients of -0.852 and -0.848 respectively,

indicating strong negative correlations. Meanwhile, 'hp' has a correlation of -.078, and 'drat' has a positive correlation of 0.68.
Considering an alternative single predictor for 'mpg', the variable 'cyl', representing the number of cylinders, emerges as a potential candidate. With a correlation coefficient of -0.85, 'cyl' shows a robust negative relationship with 'mpg', indicating that cars with more cylinders tend to have lower fuel efficiency. This strong correlation suggests that 'cyl' could be a viable predictor for 'mpg', almost as significant as 'wt', but slightly less impactful.

3. **Run a multiple regression analysis on the myCars data with lm(), using mpg as the dependent variable and wt (weight) and hp (horsepower) as the predictors. Make sure to say whether or not the overall R-squared was significant. If it was significant, report the value and say in your own words whether it seems like a strong result or not. Review the significance tests on the coefficients (B-weights). For each one that was significant, report its value and say in your own words whether it seems like a strong result or not.**

The code that I used to make the multiple regression as well as its statistical results is as follows:

```
> # Question 3
> lm(mpg ~ wt + hp, data = myCars)

Call:
lm(formula = mpg ~ wt + hp, data = myCars)

Coefficients:
(Intercept)           wt           hp
   37.22727     -3.87783     -0.03177

> summary(lm(mpg ~ wt + hp, data = myCars))

Call:
lm(formula = mpg ~ wt + hp, data = myCars)

Residuals:
   Min     1Q Median     3Q    Max
-3.941 -1.600 -0.182  1.050  5.854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
hp          -0.03177    0.00903  -3.519  0.00145 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.593 on 29 degrees of freedom
Multiple R-squared:  0.8268,    Adjusted R-squared:  0.8148
F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

In the multiple regression above, the results demonstrate a strong model fit. The multiple R-squared value of 0.826 signifies that approximately 82.68% of the variability in the mpg is explained by the combined effects of the car's weight and horsepower. This high R-squared value indicates a substantial correlation between these predictors and the fuel efficiency, suggesting that the model captures a significant portion of the factors affecting mpg. Furthermore, the F-statistic of 69.21 with a highly significant p-value of 9.109e-12 corroborates the overall statistical significance of the regression model, reinforcing the robustness of the predictors in explaining the variance in mpg.

Delving into the coefficients of the model, the wt variable emerges as the particularly influential predictor. The coefficient of -3.88 is statistically significant, with a t-value of -6.129 and a p-value of 1.12e-06, indicating a strong negative impact on mpg. This suggests that as the weight of the car increases, its fuel efficiency significantly decreases. 'Hp' also presents a notable effect, albeit smaller in magnitude, with a coefficient of -0.032. Its t-value of -3.519 and p-value of 0.00145 signify a statistically significant, albeit less pronounced, negative relationship with mpg.

4. **Using the results of the analysis from Exercise 2, construct a prediction equation for mpg using all three of the coefficients from the analysis (the intercept along with the two B-weights). Pretend that an automobile designer has asked you to predict the mpg for a car with 110 horsepower and a weight of 3 tons. Show your calculation and the resulting value of mpg.**

5. **Run a multiple regression analysis on the myCars data with lmBF(), using mpg as the dependent variable and wt (weight) and hp (horsepower) as the predictors. Interpret the resulting Bayes factor in terms of the odds in favor of the alternative hypothesis. If you did Exercise 2, do these results strengthen or weaken your conclusions?**

The code that I used to run this analysis are as follows:

```
> # Question 5
> lmBF(mpg ~ wt + hp, data = myCars)
Bayes factor analysis
--------------
[1] wt + hp : 788547604 ±0%

Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS

> summary(lmBF(mpg ~ wt + hp, data = myCars))
Bayes factor analysis
--------------
[1] wt + hp : 788547604 ±0%

Against denominator:
  Intercept only
---
Bayes factor type: BFlinearModel, JZS
```

In the regression analysis above, the resulting Bayes factor of approximately 788,547,604 provides exceptionally strong evidence in favor of the alternative hypothesis. This Bayes factor, which quantifies the evidence, suggests that the data are overwhelmingly more likely under a model that includes wt and hp as significant predictors compared to a model with only an intercept. Such a high Bayes factor indicates that the odds are substantially in favor of the relevance of wt and hp in predicting mpg. This finding is consistent with and significantly strengthens the conclusions drawn from the frequentist multiple regression analysis conducted earlier. Both the Bayesian and frequentist approaches robustly conclude that the weight and horsepower of a car are crucial factors in determining its fuel efficiency, as evidenced in the dataset.

6. **Run lmBF() with the same model as for Exercise 4, but with the options posterior=TRUE and iterations=10000. Interpret the resulting information about the coefficients.**

In order to see the results, I ran this code:

```
> summary(lmBF(mpg ~ wt + hp, data = myCars, posterior = TRUE, iterations = 10000))
|----|----|----|----|----|----|----|----|----|----|
**************************************************|

Iterations = 1:10000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

         Mean      SD  Naive SE Time-series SE
mu   20.09661 0.48246 0.0048246      0.0048246
wt   -3.78149 0.66326 0.0066326      0.0067988
hp   -0.03108 0.00945 0.0000945      0.0000945
sig2  7.49590 2.30661 0.0230661      0.0275433
g     3.66830 9.31911 0.0931911      0.0931911

2. Quantiles for each variable:

         2.5%      25%      50%      75%    97.5%
mu   19.14197 19.77831 20.09574 20.41866 21.03581
wt   -5.06673 -4.21270 -3.78380 -3.34540 -2.47351
hp   -0.04953 -0.03735 -0.03107 -0.02499 -0.01226
sig2  4.40567  5.99677  7.14028  8.56314 12.56035
g     0.35928  0.94938  1.67013  3.31311 18.23213
```

In the analysis above, the posterior distributions offer a comprehensive insight into the relationship between these variables. The empirical mean of the intercept (mu) at approximately 20.10, with a standard deviation of 0.48, suggests an average predicted mpg around 20.10 when wt and hp are at their mean values. The mean coefficient for wt if -3.78, accompanied by a standard deviation of 0.66, indicating a robust negative relationship between the car's weight and its fuel efficiency; as the weight increases, the mpg notably decreases. The coefficient for hp has a mean of around -0.031 with a standard deviation of 0.009, reflecting a negative, albeit less pronounced, impact of horsepower on mpg compared to weight.

The quantile information from the analysis further underscores these findings. The 95% credible intervals for wt and hp, ranging from approximately -5.07 to -2.47 and -0.0495 to -0.0123 respectively, reinforce the conclusion that both factors negatively affect mpg, with weight having a more significant influence. Additionally, the error variance (sig2), with a mean of around 7.50, reflects the variability of mpg values around the predicted values, indicating the extent of deviation from the regression line. The shrinkage parameter (g) and its distribution suggest the degree to which the model accounts for overfitting by shrinking coefficients towards zero.

Collectively, these results from the Bayesian analysis not only align with the conclusions drawn from the frequentist approach but also enrich them by providing a detailed probabilistic interpretation of the coefficients, capturing the uncertainties inherent in the model estimates.

7. **Run install.packages() and library() for the "car" package. The car package is "companion to applied regression" rather than more data about automobiles. Read the help file for the vif() procedure and then look up more information online about how to interpret the results. Then write down in your own words a "rule of thumb" for interpreting vif.**

After installing and engaging the car package, I was able to navigate to the vif part of the documentation:

vif {car} R Documentation

Variance Inflation Factors

Description

Calculates variance-inflation and generalized variance-inflation factors (VIFs and GVIFs) for linear, generalized linear, and other regression models.

Usage

```
vif(mod, ...)

## Default S3 method:
vif(mod, ...)

## S3 method for class 'lm'
vif(mod, type=c("terms", "predictor"), ...)

## S3 method for class 'merMod'
vif(mod, ...)

## S3 method for class 'polr'
vif(mod, ...)

## S3 method for class 'svyolr'
vif(mod, ...)
```

Arguments

`mod`	for the default method, an object that responds to `coef`, `vcov`, and `model.matrix`, such as a `glm` object.
`type`	for unweighted `lm` objects only, how to handle models that contain interactions: see Details below.
`...`	not used.

Details

If all terms in an unweighted linear model have 1 df, then the usual variance-inflation factors are calculated.

If any terms in an unweighted linear model have more than 1 df, then generalized variance-inflation factors (Fox and Monette, 1992) are calculated. These are interpretable as the inflation in size of the confidence ellipse or ellipsoid for the coefficients of the term in comparison with what would be obtained for

Vifs are used in regression analysis to diagnose to diagnose multicollinearity among predictor variables. In essence, a vif value quantifies how much the variance of a regression coefficient is inflated due to correlations among the predictors. A vif of 1 indicates no correlation and is ideal,

suggesting that the predictor is not linearly related to other predictors. Vif values between 1 and 5 generally indicate moderate correlation, typically not severe enough to require action, although the threshold can vary depending on the specific context and field of study. However, a vif of 5 or 10 and above is often considered a signal of significant multicollinearity, suggesting that the predictor in question is highly correlated with other variables in the model. Such high vif values warrant a closer examination of the model, potentially leading to the reconsideration of the inclusion of certain predictors, or to transformations of the data to reduce these correlations. It's important to interpret vif values not just as standalone numbers but in the context of the overall model, the data, and the specific objectives of the analysis, complemented by domain-specific knowledge.

8. **Run vif() on the results of the model from Exercise 2. Interpret the results. Then run a model that predicts mpg from all five of the predictors in myCars. Run vif() on those results and interpret what you find.**

The code that I used to get vif scores for each variable is as follows:

```
> #Question 8
> model <- lm(mpg ~ wt + hp, data = myCars)
> print(vif(model))
      wt       hp
1.766625 1.766625
> full_model <- lm(mpg ~ cyl + disp + hp + drat + wt, data = myCars)
> print(vif(full_model))
      cyl      disp        hp      drat        wt
 7.869010 10.463957  3.990380  2.662298  5.168795
>
```

In the first model, where mpg is predicted using wt and hp, the vif results for both predictors are 1.766625. These values fall well below the commonly used threshold of 5, indicating minimal multicollinearity between wt and hp. This suggests that each of these predictors contributes unique information to the model, without much overlap or linear dependency on each other. The relatively low vif values imply that the coefficients of wt and hp in predicting mpg are not significantly inflated due to multicollearity. Therefore, this model appears reliable in terms of the independence of its predictors, providing a clearer and more interpretable understanding of how wt and hp individually relate to mpg.

However, the scenario changes when considering the full model that includes all five predictors: cyl, disp, hp, drat, and wt. The vif results for cyl and disp are particularly high at 7.87 and 10.46 respectively, indicating significant multicollinearity. These high values suggest that cyl and disp are strongly linearly related to other predictors in the model, which could lead to unreliable and difficult-to-interpret coefficient estimates for these variables. Wt also shows a notable level of multicollinearity with a vif above 5. Although hp and drat exhibit lower vif values, suggesting moderate to low multicollinearity, the overall picture from the full model indicates that multicollinearity might be affecting the model's effectiveness. This necessitates a reassessment

of the predictor variables, potentially considering the removal, combination, or transformation of some variables, or applying dimensionality reduction techniques to mitigate the effects of multicollinearity on the model.