CE 472 – Computer Applications in Highway and Traffic Engineering
Homework 3 – Estimating Models in R

1)

Figure. 1

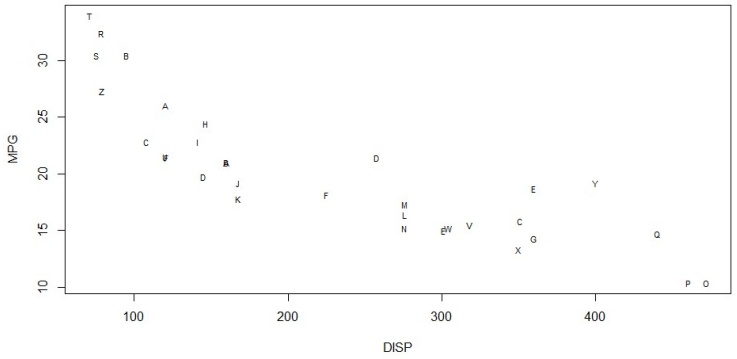


Figure. 2

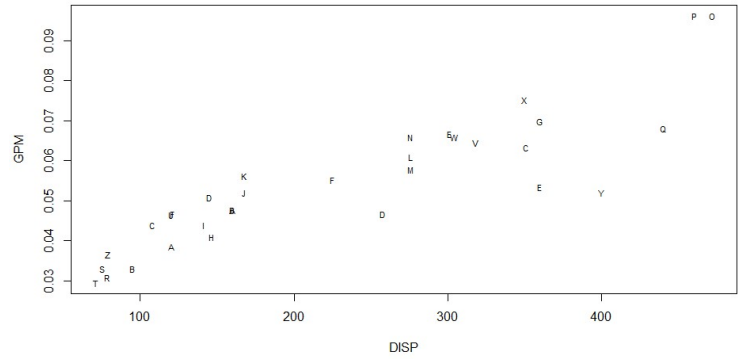


Figure. 3

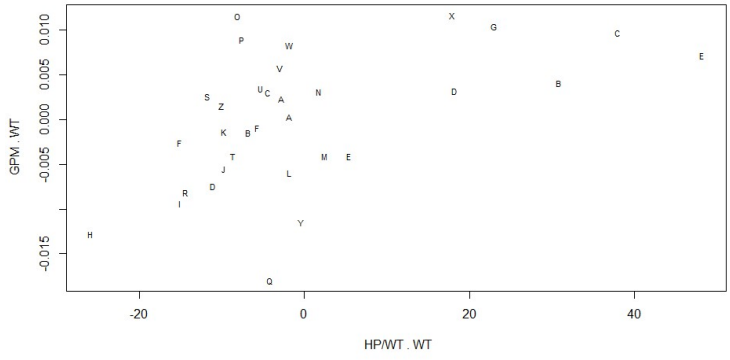
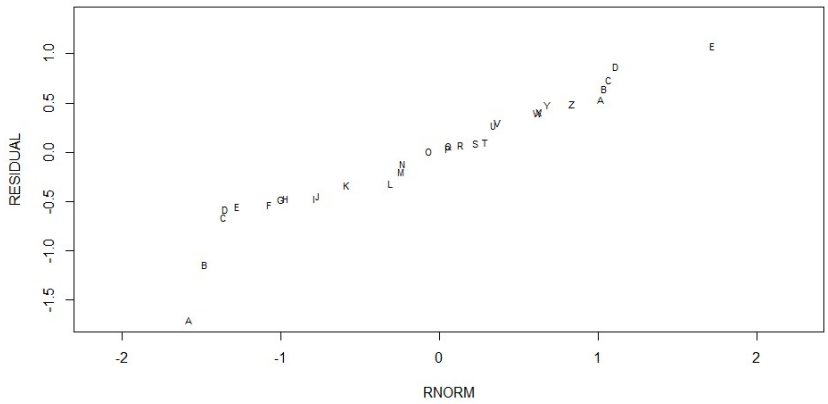


Figure. 4



Model with GPM as dependent variable (Table 2)

```
Call:
lm(formula = gpm ~ wt + hp_per_wt)

Residuals:
     Min       1Q   Median       3Q      Max
-1.69714 -0.46822  0.05312  0.42744  1.35097

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.401534   0.512044  -0.784  0.43929
wt           1.472176   0.121554  12.111 7.24e-13 ***
hp_per_wt    0.023997   0.007302   3.286  0.00266 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6612 on 29 degrees of freedom
Multiple R-squared:  0.8484,    Adjusted R-squared:  0.8379
F-statistic: 81.13 on 2 and 29 DF,  p-value: 1.322e-12
```

Results for Shapiro-Wilk normality test on the residuals of the above model:

- W = 0.98071
- p-value = 0.82

Model with MPG as dependent variable

```
Call:
lm(formula = mpg ~ wt_inverse + wt_per_hp)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5407 -1.4045 -0.3689  1.1423  5.2369

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.923      1.589  -0.581    0.566
wt_inverse    44.435      3.215  13.819 2.73e-14 ***
wt_per_hp    236.657     48.502   4.879 3.55e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.091 on 29 degrees of freedom
Multiple R-squared:  0.8874,    Adjusted R-squared:  0.8796
F-statistic: 114.3 on 2 and 29 DF,  p-value: 1.765e-14
```

Results for Shapiro-Wilk normality test on the residuals of the above model:

- W = 0 .97233
- p-value = 0.5659

Observations:

- For models with MPG and GPM, the residuals are normally distributed. So, we can safely assume the assumptions of linear regression has been satisfied.
- R-squared value for both the models are close to a certain extent but the Residual standard error of GPM model is just 0.6612 while MPG model has a Residual standard error of 2.091. So, we can conclude that GPM is easier to predict than MPG.

2)

Data Collection:

Each individual was given 9 hypothetical situations. Each individual was given 9 different combination for travel time, cost, etc. For these different set options, their choice of transport mode car/train/swiss metro (given that both train and swiss metro are available all the time) along with some personal info such as income, age, gender etc where noted. The choice of car was only offered to car owners.

i. For original train users: 470 people were interviewed through household or telephone surveys out of which only 441 usable surveys were compiled to form of a database.

ii. For original car users: Household or telephone survey was deemed impractical for car trips. By means of video recorders a total of 10529 relevant license plates on motorways were recorded during September 1997 out of which 9658 surveys were mailed to owners of these cars with the help of central Swiss car license agency and got 1758 replies out of which only 750 usable surveys were compiled to form of a database

Independent variables used and Coefficient interpretation (setting utility of Car as zero):

1) **SURVEY (0/1):** Current choice for transport mode (0-train & 1-car) – before introduction of swiss metro – This is variable has high contribution because people may tend to stick to their original transport choice even after metro innovation.
    i. SM: coeff is -ve, if a person is a car user, utility of SM decreases. Therefore, we can say that original train users prefer SM over car.
    ii. Train: coeff is -ve, if a person is a car user, utility of Train decreases. Therefore, we can say that original train users prefer train over car.

2) **LUGGAGE_none (0/1) & LUGGAGE_one_piece (0/1):** If both of these variables are zero it means luggage has more than one piece. People with heavy luggage might prefer cars over train or metro.
    i. SM: coeff of LUGGAGE_none is greater than LUGGAGE_one_piece (both coeff are +ve). Therefore, the heavier the luggage lesser the magnitude of SM's utility. As a result, we can conclude that people with more luggage don't prefer SM and rather prefer car.
    ii. Train: coeff of LUGGAGE_none is greater than LUGGAGE_one_piece (both coeff are +ve). Therefore, the heavier the luggage lesser the magnitude of train's utility. As a result, we can conclude that people with more luggage don't prefer train and rather prefer car.

3) **is_Old:** 0 – if age is lesser than 65 and 1 – otherwise. Older people might prefer car over train/metro.
    i. SM: coeff is -ve, so old people seem to prefer Car and Train over metro
    ii. Train: coeff is +ve, so old people seem to prefer Train over metro and car

4) **is_Annual_Ticket & GA:** 1 – if they have an annual ticket & 0 – otherwise.
    i. SM: coeff is +ve for both "is_Annual_Ticket" & "GA", so probability of choosing SM increases if people have annual ticket.
    ii. Train: coeff is +ve for both "is_Annual_Ticket" & "GA", so probability of choosing train increases if people have annual ticket.

5) **HEADWAY_diff:** Headway of SM – Headway of Train. If the difference is +ve, people will prefer SM over train and vice versa.
    i. SM: coeff is -ve, if headway of SM greater than train "HEADWAY_diff" is more which means waiting time for SM is greater. Therefore, if waiting time of SM increases, people will prefer train over SM.
    ii. Train: coeff is +ve, if headway of SM greater than train "HEADWAY_diff" is more which means waiting time for train is lesser. Therefore, if waiting time of train increases, people will prefer SM over train.

6) **Travel time (TT):** Travel time from origin to destination – Alternate variable – coefficient is -ve. Therefore, people will choose the transport mode with the least travel time.

7) **Cost (CO):** Cost of travel – Alternate variable – coefficient is -ve, so as cost increases utility of the transport mode decreases. Therefore, people will choose the cheapest transport mode.

Independent variables which didn't have much effect on CHOICE as expected:

The following variables were chosen based on intuition and dropped later.

1) **AV.CAR:** 1 – if a person owns a car and 0 – otherwise. Persons with car will be biased and choose car as their transportation choice more often (might be correlated with 'SURVEY')
2) **INCOME:** People with lesser income might choose train/metro. After doing some exploratory data analysis, I found that around 10% of the people didn't provide details about their income. As a result, INCOME can't be used for prediction.
3) **WHO:** If the person's job employer apart from salary offers additional beneficiary such as a free metro pass etc., then the person is more likely to choose metro more often. But there is a high chance that 'WHO' will explain the same variation explained by

'is_Annual_Ticket & GA' (i.e., they might be correlated). So, it makes sense of using either one in our model. Also 0.3% of the data is unknown.

VALUE OF TIME = Coeff. Of TT / Coeff. Of CO = -9.4241e-03 / -8.7919e-04
= 10.71907 CHF/minute

Conclusion:

From the survey we can see that around 58% of the people chose metro over train and car. So we can conclude that metro is a attractive and popular choice.

i. Coefficient of LUGGAGE_none (+ve) > LUGGAGE_one_piece (+ve) – that means people with heavier luggage hesitate to use metro. So, no. of available porters must be increased or an automated luggage carrying system like the ones available in airports may be implemented (if viable).
ii. Coefficient of is_Old is negative for SM and positive for train – so measures to attract older people must be undertaken.
iii. Coefficient of "GA" is positive. So, providing more offers and discounted prices on annual season tickets will help
iv. Coefficient of HEADWAY_diff for SM is -ve, which implies people choose the mode with less travel time. Therefore, increasing the frequency of metro arrival will make the metro more attractive.