DATA 303/473 Assignment 1 Solutions

Due 1159pm Friday 17 March 2023

Assignment 1 Solutions

Q1. (33 marks) Car sales in India

a. **(3 marks)** Data summary

```
## 'data.frame':    8028 obs. of  10 variables:
##  $ price  : num  75 99 122 200 300 ...
##  $ make   : Factor w/ 8 levels "Ford","Honda",..: 6 6 6 6 6 6 6 6 6 6 ...
##  $ kms    : num  90 100 60 80 200 50 31.8 53 120 31.8 ...
##  $ fuel   : Factor w/ 2 levels "Diesel","Petrol": 1 1 1 1 1 1 1 1 1 1 ...
##  $ seller : Factor w/ 3 levels "Dealer","Individual",..: 2 2 2 2 2 1 1 1 2 1 ...
##  $ tx     : Factor w/ 2 levels "Automatic","Manual": 2 2 2 2 2 1 1 1 1 1 ...
##  $ owner  : Factor w/ 3 levels "First ","Second ",..: 2 2 2 3 2 1 1 1 1 1 ...
##  $ mileage: num  12.8 12.8 12.8 13.5 20.1 ...
##  $ esize  : int  1489 1995 1995 1995 1461 1968 1968 2967 1968 1968 ...
##  $ power  : num  35.5 52 52 52 75 ...
```

Table 1: Summary of variables in car sales dataset (continued below)

price	make	kms	fuel
Min. : 30.0	Maruti :2378	Min. : 1.00	Diesel:4401
1st Qu.: 260.0	Other :1401	1st Qu.: 35.00	Petrol:3627
Median : 450.0	Hyundai :1393	Median : 60.00	NA
Mean : 640.4	Mahindra: 772	Mean : 69.77	NA
3rd Qu.: 680.0	Tata : 733	3rd Qu.: 98.00	NA
Max. :10000.0	Toyota : 488	Max. :2360.46	NA
NA	(Other) : 863	NA	NA

Table 2: Table continues below

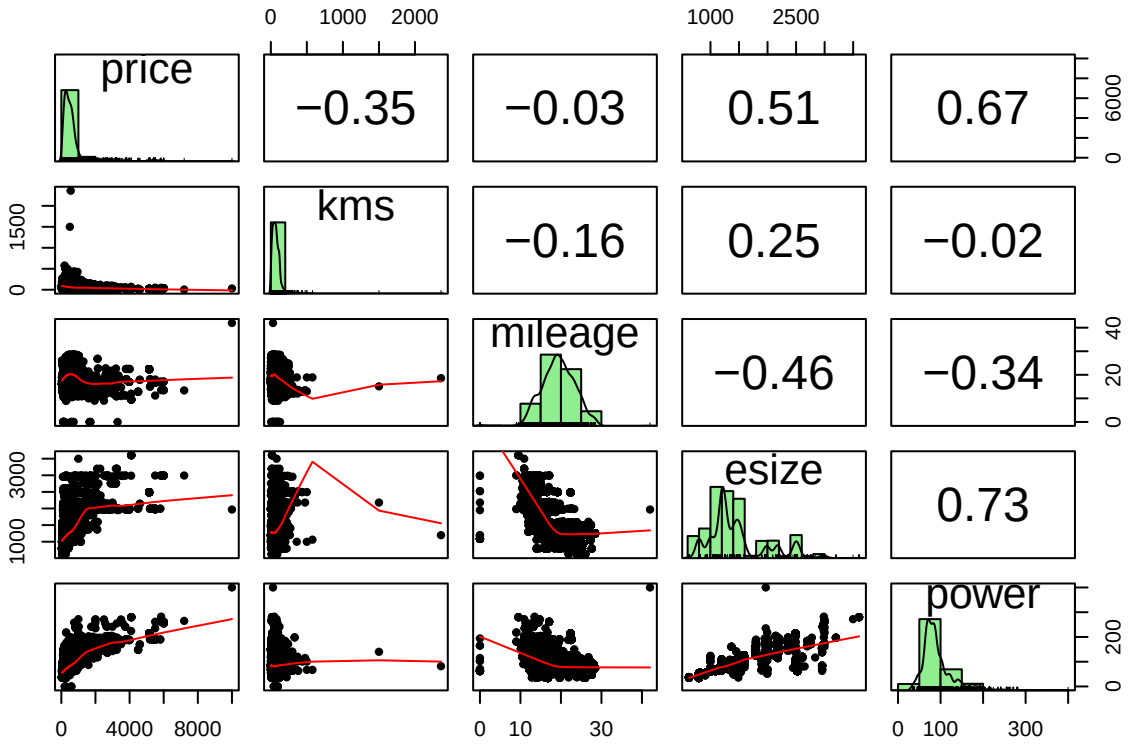
seller	tx	owner	mileage
Dealer :1119	Automatic:1046	First :5238	Min. : 0.00
Individual :6673	Manual :6982	Second :2073	1st Qu.:16.78
Trustmark Dealer: 236	NA	Third or above: 717	Median :19.30
NA	NA	NA	Mean :19.39
NA	NA	NA	3rd Qu.:22.32
NA	NA	NA	Max. :42.00
NA	NA	NA	NA's :214

esize	power
Min. : 624	Min. : 0.00

esize	power
1st Qu.:1197	1st Qu.: 68.85
Median :1248	Median : 82.40
Mean :1463	Mean : 91.82
3rd Qu.:1582	3rd Qu.:102.00
Max. :3604	Max. :400.00
NA's :214	NA's :208

i) `mileage` and `power` have values of 0. These are obviously incorrect and should be excluded
ii) `mileage`, `esize` and `power` have missing values.

b. **(4 marks)** Scatterplot matrix



i) All numerical variables `kms`, `mileage`, `esize` and `power` have a non-linear relationship with `price`.

ii) Yes. `price` is highly skewed. We may need to consider a transformation of price if the model indicates non-normal residuals.

c. **(3 marks)** Linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-745.8	77.74	-9.594	1.114e-21
makeHonda	-151	31.91	-4.73	2.281e-06
makeHyundai	18.5	26.51	0.698	0.4852
makeMahindra	108.3	30.47	3.554	0.0003814
makeMaruti	149.1	25.6	5.822	6.032e-09
makeOther	262.9	26.62	9.876	7.217e-23
makeTata	-50.09	28.61	-1.751	0.08006
makeToyota	235.5	34.48	6.829	9.207e-12
kms	-1.586	0.1033	-15.35	2.193e-52
fuelPetrol	-5.7	15.65	-0.3642	0.7157
sellerIndividual	-235.5	16.39	-14.37	2.994e-46
sellerTrustmark Dealer	-284.3	34.71	-8.191	3.003e-16
txManual	-414.8	19.71	-21.04	1.174e-95
ownerSecond	-112.9	12.72	-8.874	8.609e-19
ownerThird or above	-129.5	19.79	-6.547	6.254e-11
mileage	30.9	2.08	14.85	3.248e-49
esize	0.09409	0.02463	3.82	0.0001344
power	13.83	0.2601	53.16	0

Table 5: Fitting linear model: price ~ make + kms + fuel + seller + tx + owner + mileage + esize + power

Observations	Residual Std. Error	R^2	Adjusted R^2
7797	451.2	0.6901	0.6894

$\hat{\sigma}^2 = 451.2^2 = 203581.4$

d. **(2 marks)** `make`=Ford, `kms`=0, `fuel`=Diesel, `seller`=Dealer, `tx`=Automatic, `owner`=First, `mileage`=0, `esize`=0, `power`=0.

e. **(4 marks)** Interpreting coefficients:

i) `txManual`: The expected price of a car with a manual transmission is 414 800 INR lower than that of a car with an automatic transmission when all other predictors are held constant.

ii) `mileage`: An increase in `mileage` by 1 kmpl is associated with an increase in expected price of 30 900 INR, when all other predictors are held constant.

f. **(3 marks)** 95% confidence and prediction intervals for the last three observations in the dataset.

```
##Get predictor values to predict for as a dataframe.
xdata<-subset(tail(cars2,3), select = -c(price))
## Select last 3 rows (1501:1503) and exclude response variable

library(pander)
pander(predict(fit1, newdata=xdata, interval="confidence"),
       caption="Confidence intervals")
```

Table 6: Confidence intervals

	fit	lwr	upr
7795	2294	2256	2332

	fit	lwr	upr
7796	2623	2581	2665
7797	2034	1987	2081

```
pander(predict(fit1, newdata=xdata, interval="prediction"),
       caption="Prediction intervals")
```

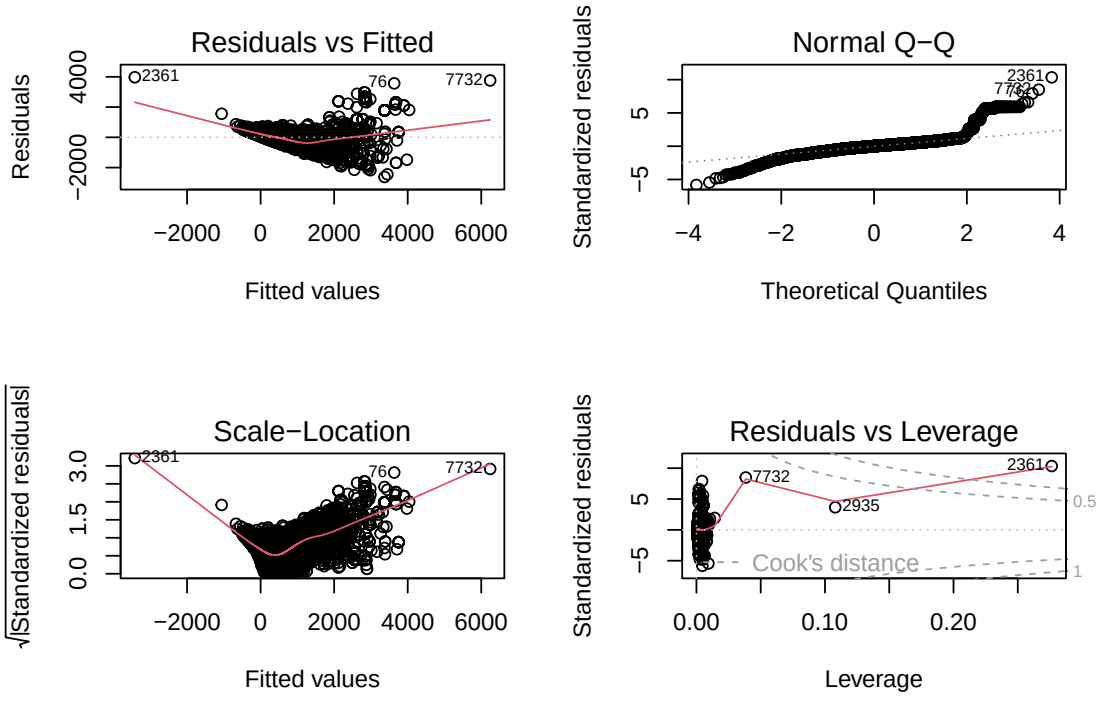
Table 7: Prediction intervals

	fit	lwr	upr
7795	2294	1408	3179
7796	2623	1737	3508
7797	2034	1148	2920

The prediction intervals are substantially wider than the confidence intervals, reflecting greater uncertainty about the predicted price of an individual car compared to uncertainty about the predicted average price of several cars.

g. **(4 marks)** Residual diagnostics

```
par(mfrow=c(2,2))
plot(fit1)
```



- Residuals vs Fitted plot shows curvate indicating non-linearity
- Q-Q plot shows points deviate from a straight line indicating non-normality
- Scale-location plot shows non-random scatter of points indicating non-constant variance
- Residuals vs Leverage plot shows one point outside Cook's Distance threshold, indicating there is one highly influential observation (2361).

h. **(4 marks)** Hypothesis tests for regression assumptions.

```
ks.test(fit1$res, 'pnorm')
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  fit1$res
## D = 0.49948, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
library(lmtest)
bptest(fit1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit1
## BP = 2600.1, df = 17, p-value < 2.2e-16
```

Kolmogorov-Smirnov test

- *Null Hypothesis (H_0)*: The sample comes from a normal distribution.
- *Alternative Hypothesis (H_1)*: The sample does not come from a normal distribution.

Breush-Pagan test

- *Null Hypothesis (H_0)*: Homoscedasticity is present (the residuals are distributed with equal variance)
- *Alternative Hypothesis (H_1)*: Heteroscedasticity is present (the residuals are not distributed with equal variance)

Both tests have p-value$< 2.2 \times 10^{-16}$. We therefore reject the null hypothesis in both cases, confirming departure from the assumptions of normality and constant variance.

i. **(2 marks)** Multicollinearity check

```
library(car)
car::vif(fit1)
```

```
##             GVIF Df GVIF^(1/(2*Df))
## make    3.502941  7        1.093674
## kms     1.324188  1        1.150734
## fuel    2.321487  1        1.523643
## seller  1.470636  2        1.101226
## tx      1.706962  1        1.306508
## owner   1.224059  2        1.051842
## mileage 2.523910  1        1.588682
## esize   5.904424  1        2.429902
## power   3.303755  1        1.817623
```

There is no evidence of severe multicollinearity among predictors since all values of `GVIF^(1/(2*Df))` are less than 10.

j. **(4 marks)** Global usefulness test

```
summary(fit1)
```

```
##
## Call:
## lm(formula = price ~ make + kms + fuel + seller + tx + owner +
##     mileage + esize + power, data = cars2)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2620.0  -188.9     3.0   171.2  3970.6
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -745.84341   77.73880  -9.594  < 2e-16 ***
## makeHonda              -150.95095   31.91104  -4.730 2.28e-06 ***
## makeHyundai              18.50214   26.50858   0.698 0.485218
## makeMahindra            108.30980   30.47386   3.554 0.000381 ***
## makeMaruti              149.07261   25.60333   5.822 6.03e-09 ***
## makeOther               262.86123   26.61647   9.876  < 2e-16 ***
## makeTata                -50.08944   28.61375  -1.751 0.080065 .
## makeToyota              235.48398   34.48390   6.829 9.21e-12 ***
## kms                      -1.58575    0.10333 -15.346  < 2e-16 ***
## fuelPetrol               -5.70032   15.65198  -0.364 0.715725
## sellerIndividual       -235.50855   16.38584 -14.373  < 2e-16 ***
## sellerTrustmark Dealer -284.31597   34.71053  -8.191 3.00e-16 ***
## txManual               -414.75405   19.70954 -21.043  < 2e-16 ***
## ownerSecond            -112.88321   12.72001  -8.874  < 2e-16 ***
## ownerThird or above    -129.54557   19.78783  -6.547 6.25e-11 ***
## mileage                  30.89608    2.08048  14.850  < 2e-16 ***
## esize                     0.09409    0.02463   3.820 0.000134 ***
## power                    13.82643    0.26011  53.156  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 451.2 on 7779 degrees of freedom
## Multiple R-squared:  0.6901, Adjusted R-squared:  0.6894
## F-statistic:  1019 on 17 and 7779 DF,  p-value: < 2.2e-16
```

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{17} = 0$

against

H_1 : At least one $\beta_j \neq 0, j = 1, \ldots, 17$

We find $F = 1019$ with 17 and 7779 d.o.f and $p-$value$< 2.2 \times 10^{-16}$. There is very strong evidence to reject H_0 and insufficient evidence that all regression coefficients are zero in the population. Therefore it is worth going on to further analyse and interpret a model of `price` against the predictors.

Q2. (7 marks) Olive oil

a. **(2 marks)** Oleic acid is strongly correlated (r=-0.84) with linoleic acid. Adding oleic acid as a predictor to `fit2` results in multicollinearity, which can cause signs of regression coefficients to change.

b. **(2 marks)** Predictions

```
library(pander)
pander(predict(fit3,
            newdata=data.frame(linoleic=0.3, stearic=2.2, oleic=73.0),
            interval="confidence"),digits=3,caption="Confidence intervals")
```

Table 8: Confidence intervals

fit	lwr	upr
18.9	18.7	19.1

```
pander(predict(fit3,
               newdata=data.frame(linoleic=0.3, stearic=2.2, oleic=73.0),
               interval="prediction"), digits=3,caption="Prediction intervals")
```

Table 9: Prediction intervals

fit	lwr	upr
18.9	18.2	19.6

c. **(1 mark)** Values for the each of the three predictors should be within the range of values in the original sample.
d. **(2 marks)** Independence. Oils from the same region are likely to have the same growing conditions, meaning their constituents are more similar compared to samples from a different region. Therefore spatial autocorrelation would be present.

Assignment total: 40 marks