Application exercise 7.2: Multiple linear regression

Submit your responses on Sakai, under the appropriate assignment. Only one submission per team is required. One team will be randomly selected and their responses will be discussed.

Cigarettes and CO

The Federal Trade Commission annually rates varieties of domestic cigarettes according to their tar, nicotine, and carbon monoxide content. The United States Surgeon General considers each of these substances hazardous to a smoker's health. Past studies have shown that increases in the tar and nicotine content of a cigarette are accompanied by an increase in the carbon monoxide emitted from the cigarette smoke.

In this exercise we will work with data from 2007 on cigarettes sold in the US. Each row in the dataset represents a cigarette. There are 11 variables in the dataset:

- BRAND_NAME
- TYPE: Type of cigarette, REGULAR or MENTHOL
- NIC: Nicotine content, in mg
- TAR: Tar content, in mg
- CO: Carbon monoxide, in mg
- LEN: Length of cigarette, in mm
- FLTR: Filter, F or NF
- PACK: Pack type, HARD or SOFT
- STRENGTH: Strength of cigarette, ULTRA LIGHT, LIGHT, MEDIUM, REGULAR FULL, or FLAVOR
- STYLE: Some information of style of cigarette (not available for all cigarettes, and not used in this analysis)
- OTHER: Other relevant information (not available for all cigarettes, and not used in this analysis)

To load the data use the following:

```
download("http://stat.duke.edu/~mc301/data/cig07.RData", destfile = "cig07.RData")
load("cig07.RData")
```

1. First we try a full model with all explanatory variables in the model. The output of this model is shown below. Interpret the slopes of nicotine and tar variables. Are these results surprising? Why, or why not? Make sure to use appropriate terminology in your answer, and use summary statistics and/or visualizations of the data to support your answer.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.5581	0.5406	-1.03	0.3021
NIC	-4.0506	0.4357	-9.30	0.0000
TAR	1.0492	0.0441	23.77	0.0000
LEN	0.0352	0.0055	6.37	0.0000
FLTRNF	-6.5050	0.3603	-18.06	0.0000
PACKSOFT	0.5130	0.1047	4.90	0.0000
STRENGTHLIGHT	1.6867	0.2121	7.95	0.0000
STRENGTHMEDIUM	0.7418	0.4616	1.61	0.1083
STRENGTHREGULAR	0.2880	0.3072	0.94	0.3487
STRENGTHFULL FLAVOR	2.2496	0.3292	6.83	0.0000
TYPEMENTHOL	-0.0332	0.1109	-0.30	0.7647

2. Next, we try the following two models, and obtain the following adjusted R^2 values:

   - Option 1, use TAR: `lm(CO ~ TAR + LEN + FLTR + PACK + STRENGTH + TYPE, data = cig07)`, adjusted $R^2 = 0.7855$
   - Option 2, use NIC: `lm(CO ~ NIC + LEN + FLTR + PACK + STRENGTH + TYPE, data = cig07)`, adjusted $R^2 = 0.7066$

   Based on these results which variable should we keep in our full model, NIC or TAR? Why?

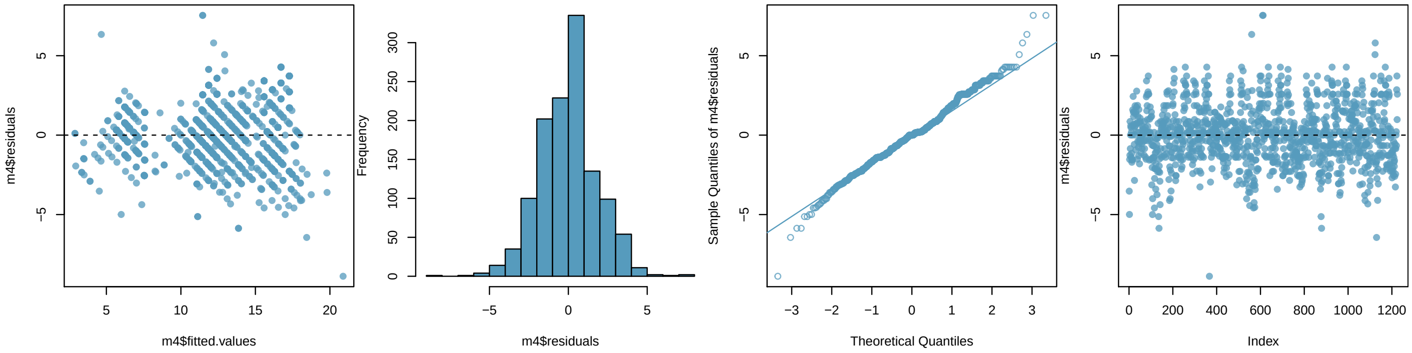
3. Suppose the final model (based on backwards p-value based elimination) as follows. Interpret the slope estimates associated with the strength variable. Based on this model, cigarettes with which strength level has the highest CO emissions?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0586	0.5555	-0.11	0.9160
TAR	0.7344	0.0293	25.07	0.0000
LEN	0.0267	0.0056	4.76	0.0000
FLTRNF	-6.1949	0.3686	-16.81	0.0000
PACKSOFT	0.5597	0.1081	5.18	0.0000
STRENGTHLIGHT	1.9077	0.2168	8.80	0.0000
STRENGTHMEDIUM	0.7900	0.4766	1.66	0.0976
STRENGTHREGULAR	0.5664	0.3149	1.80	0.0723
STRENGTHFULL FLAVOR	3.0920	0.3268	9.46	0.0000

:

Multiple R-squared: $[R^2]$, Adjusted R-squared: $[R^2_{adj}]$
F-statistic: $[F]$ on $[df_1]$ and $[df_2]$ DF, p-value: $<$ 2.2e-16

4. Use the following plots to check conditions for this model.



5. Construct a 95% confidence interval for the slope of the filter variable (FLTRNF), and interpret it in context.

6. Using the ANOVA output for the model provided below, calculate the [redacted] statistics on the regression output above. *Hint:* The ANOVA output below shows SS attributed to each variable separately, $SS_{Regression}$ is the sum of those.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TAR	1	12216.31	12216.31	3622.74	0.0000
LEN	1	194.02	194.02	57.54	0.0000
FLTR	1	1675.48	1675.48	496.86	0.0000
PACK	1	169.17	169.17	50.17	0.0000
STRENGTH	4	900.44	225.11	66.76	0.0000
Residuals	1216	4100.50	3.37		

7. In this model, which predictor is able to explain the highest portion of the variability in CO emission of cigarettes?

8. Using the regression model predict the CO emission for a cigarette with the following characteristics. Note that you may not need to use each attribute in your calculation.

- BRAND_NAME: Sir Smokes-a-Lot
- TYPE: MENTHOL
- NIC: 0.75 mg
- TAR: 12 mg
- LEN: 80 mm
- FLTR: F
- PACK: HARD
- STRENGTH: LIGHT

9. Now confirm your prediction from the previous question using the `predict` function in R. Note that your hand calculated prediction might be very slightly different from R's prediction, due to rounding of the coefficients on the regression output.

```
# fit the model
m = lm(CO ~ TAR + LEN + FLTR + PACK + STRENGTH, data = cig07)
# create the new data point
smokesalot = data.frame(TAR = 12, LEN = 80, FLTR = "F", PACK = "HARD", STRENGTH = "LIGHT")
# predict
predict(m, newdata = smokesalot)
```

Include your R code and your write up in your submission.

If you have questions about the R syntax, refer to https://stat.duke.edu/~mc301/R/Rcommands.html or ask.