Fan Bu
Duke University - Department of Statistical Science

Application Exercise 9: Multiple linear regression - KEY

Write your responses in the spaces provided below. WRITE LEGIBLY and SHOW ALL WORK! Concise and coherent are best!

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)

1. Suppose the full model uses the following explanatory variables: nicotine, tar, length, filter, pack, strength, and type. Describe, briefly, in your own words, how you would carry out model selection using the backwards elimination method based on adjusted R^2.

   Fit full model, record its adjusted R^2. Then, drop one variable at a time, record each model's adjusted R^2. Then, move to the model with the highest adjusted R^2. Stop when adjusted R^2 stops increasing.

2. The output of the model resulting from backwards elimination with adjusted R^2 is shown below. Evaluate the slopes of NIC and TAR variables. Are these results surprising? Why, or why not? Make sure to use appropriate terminology in your answer. *Hint:* The pairs plot will at the end of this document can be helpful for determining whether the results are surprising or not.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.5489	0.5395	-1.02	0.3092
NIC	-4.0406	0.4342	-9.31	0.0000
TAR	1.0485	0.0441	23.80	0.0000
LEN	0.0350	0.0055	6.38	0.0000
FLTRNF	-6.4925	0.3577	-18.15	0.0000
PACKSOFT	0.5128	0.1046	4.90	0.0000
STRENGTHLIGHT	1.6804	0.2110	7.96	0.0000
STRENGTHMEDIUM	0.7339	0.4607	1.59	0.1114
STRENGTHREGULAR	0.2801	0.3059	0.92	0.3600
STRENGTHFULL FLAVOR	2.2447	0.3287	6.83	0.0000

Yes, according to this model output, as nicotine increases CO emission decreases. But we know from the pairs plot that nicotine and CO are positively associated. This is likely due to multi-collinearity in the model – nicotine and tar are strongly correlated, resulting in unreliable slope estimates.

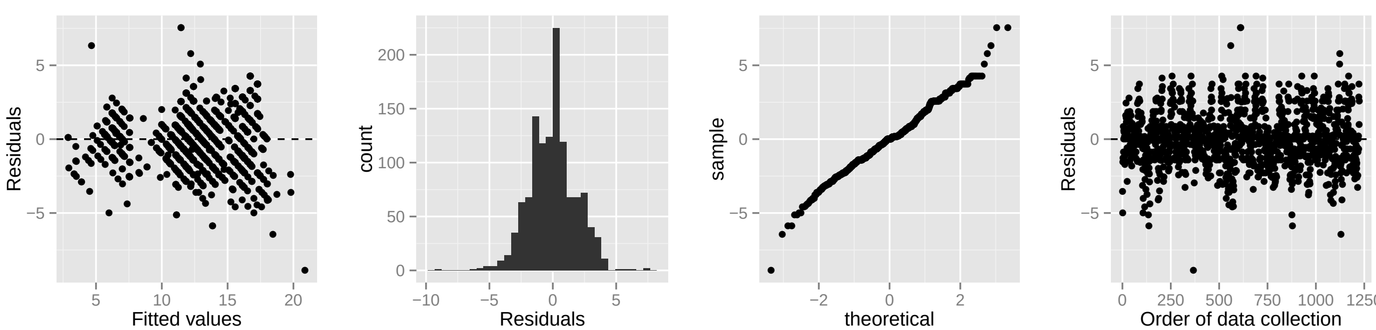
3. Next, we try the following two models, and obtain the following adjusted R^2 values:

   - Option 1, remove TAR: `lm(CO ~ NIC + LEN + FLTR + PACK + STRENGTH, data = cig07)`, adjusted $R^2 = 0.7066$
   - Option 2, remove NIC: `lm(CO ~ TAR + LEN + FLTR + PACK + STRENGTH, data = cig07)`, adjusted $R^2 = 0.7857$

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

   Remove nicotine, higher adjusted R^2.

4. In the remainder of the application exercise we will complete some inferential tasks based on the final model. Use the following plots to check conditions before to determine whether we can proceed with these tasks.



1. Linearity – residuals scattered around 0, but there is some structure, may not be met.

2. Nearly normal residuals – mostly met.

3. Constant variance – residuals plot does not have an apparent fan shape.

4. *Independence – no apparent structure due to order of data collection, but we would like to know about data collection methods (i.e. how were these cigarettes selected)*

5. Provided below is the final model output. Construct a 95% confidence interval for the slope of the pack variable (PACKSOFT), and interpret it in context.

	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

Residual standard error: 1.836 on 1216 degrees of freedom

Multiple R-squared: 0.7871, Adjusted R-squared: 0.7857

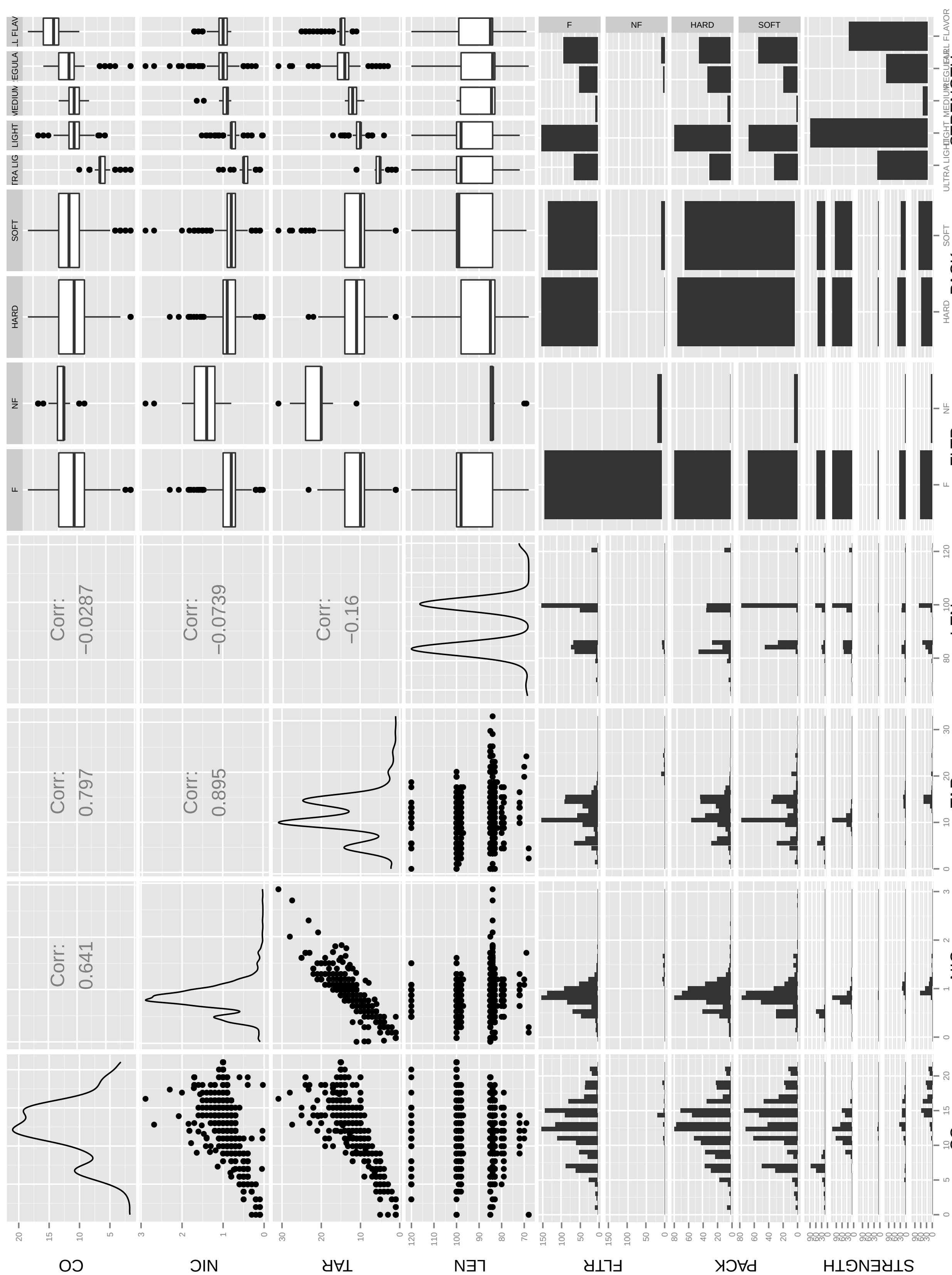
F-statistic: 561.8 on 8 and 1216 DF, p-value: ¡ 2.2e-16

$df = 1216,\ t^\star_{1216} \approx 1.96$
$0.5597 \pm 1.96 \times 0.1081 = (0.35, 0.77)$
We are 95% confident that, on average, soft pack cigarettes emit 0.35 mg to 0.77 mg more CO than hard pack cigarettes.

6. The ANOVA output below shows the sum of squares attributed to each variable separately. Based on this output which predictor is able to explain the highest portion of the variability in CO emission of cigarettes? What percent of the variability in CO emission does this variable explain?

	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		

$R^2_{TAR} = \frac{12216.31}{12216.31+194.02+1675.48+169.17+900.44+4100.50} \approx 0.63$


CO
NIC
TAR
LEN
FLTR
PACK
STRENGTH
Corr: 0.641
Corr: 0.797
Corr: -0.0287
Corr: 0.895
Corr: -0.0739
Corr: -0.16
F
NF
HARD
SOFT