PHP 2511 Midterm Exam 2018

March 14, 2018

Name: _____

Instructions

- Write your solutions under each of the questions.
- You have 80 minutes for the examination, from 10:30 to 11:50. Papers will be collected at 11:50, no exceptions.
- This exam is closed book. You may not use any resources.
- This exam has ***16*** Problems and ***10*** pages. Some problems span several pages, so read the exam carefully. The last page has been left blank for scrap work.
- Show your work and **explain your reasoning**. The final answer is not as important as the process.
- All interpretations must be in context to the original problem including units.
- All R output is in this exam.
- **All answers must be in complete sentences**

Scoring

Problem	Point Value	Problem Grade
1	8 ‘	_____ ‘ ‘
2	3 ‘	_____ ‘ ‘
3	6 ‘	_____ ‘ ‘
4	4 ‘	_____ ‘ ‘
5	3 ‘	_____ ‘ ‘
6	5 ‘	_____ ‘ ‘
7	6 ‘	_____ ‘ ‘
8	4 ‘	_____ ‘ ‘
9	7 ‘	_____ ‘ ‘
10	3 ‘	_____ ‘ ‘
11	3 ‘	_____ ‘ ‘
12	8 ‘	_____ ‘ ‘
13	5 ‘	_____ ‘ ‘
14	3 ‘	_____ ‘ ‘
15	6 ‘	_____ ‘ ‘
16	5 ‘	_____ ‘ ‘
	Total	79

The Data

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.

The data presented here are taken from Mendenhall and Sincich (1992) and are a subset of the data produced by the Federal Trade Commission.

For more information, see the article Using Cigarette Data for an Introduction to Multiple Regression by Lauren McIntyre in Volume 2, Number 1, of the **Journal of Statistics Education**.

These data contain the following variables:

Variable	Description
brand	Brand name
tar	Tar content (mg)
nicotine	Nicotine content (mg)
weight	Weight (g)
CO	Carbon monoxide content (mg)

Conceptual Questions: PHP 2511 Only

1. **(8 points)** List the Assumptions of a linear Regression. Briefly explain what they are.

2. **(3 points)** Why do we need to use the `logit()` function for logistic regression instead of just regressing this like we do with linear regression?

3. **(6 points)** Explain what a confounder is and why need to control for it.

4. **(4 points)** List the Assumptions of a Generalized Linear Model.

5. **(3 points)** How can we test for homogeneity of the variances?

6. **(5 points)** Why is it considered a problem if our residuals are not normally distributed? What kind of incorrect inferences can you make?

Data Analysis Question

7. **(6 points)** Consider the Simple Linear Regressions below:

term	estimate	p.value	conf.low	conf.high
tar	0.801	0.000	0.697	0.905
nicotine	12.395	0.000	10.215	14.576
weight	25.068	0.019	4.422	45.714

Which variables are significant? Comment on whether or not the significant variables lead to an increase or decrease in Carbon Monoxide Output.

8. **(4 points)** The table below represents the fit statistics of the above simple linear regressions. Which variable alone explains the most variation? How much Variation does it explain?

	r.squared	adj.r.squared	sigma	statistic	p.value
tar	0.917	0.913	1.40	253.37	0.000
nicotine	0.857	0.851	1.83	138.27	0.000
weight	0.215	0.181	4.29	6.31	0.019

9. **(7 points)** Consider the Multiple Linear Regression below:

	term	estimate	p.value	conf.low	conf.high
2	tar	0.963	0.001	0.459	1.47
3	nicotine	-2.632	0.507	-10.743	5.48
4	weight	-0.130	0.974	-8.210	7.95

How did the coefficients change from the simple linear regressions? How did the significance change from the simple linear regressions? Why do you think there was so much change?

10. **(3 points)** Consider the fit statistics of the above regression. What can you conclude compared to the simple linear regressions?

r.squared	adj.r.squared	sigma	statistic	p.value
0.919	0.907	1.45	79	0

11. **(3 points)** Consider the VIF outputs below:

```
##      tar nicotine   weight
##    21.63    21.90     1.33
```

What does this tell you about the variables?

12. **(8 points)** We decide not to drop Nicotine from the model based on what we know about cigarettes and nicotine. Instead we create a binary Nicotine with levels "high" and "low". Consider the model below which is:

$$E[\hat{CO}] = \hat{\beta}_0 + \hat{\beta}_1 Tar + \hat{\beta}_2 Nicotine + \hat{\beta}_3 Tar * Nicotine$$

term	estimate	p.value	conf.low	conf.high
(Intercept)	1.518	0.038	0.093	2.942
tar	0.915	0.000	0.794	1.036
nicotine_binhigh	7.924	0.004	2.792	13.055
tar:nicotine_binhigh	-0.443	0.002	-0.697	-0.188

What does this regression tell you? Comment on significance of terms and give a notion for what the interaction term means.

13. **(5 points)** Consider if we ran just the main effects model for problem 10. We then compute an F-test comparing the models. What are the hypothesis for this test and what can you conclude?

```
## Analysis of Variance Table
##
## Model 1: CO ~ tar * nicotine_bin
## Model 2: CO ~ tar + nicotine_bin
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1     21 27.5
## 2     22 44.6 -1     -17.1 13.1 0.0016 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

14. **(3 points)** Consider the fit statistics for this model:

r.squared	adj.r.squared	sigma	statistic	p.value
0.949	0.942	1.14	130	0

What does the adjusted R^2 tell you about this compared to all the previous models?

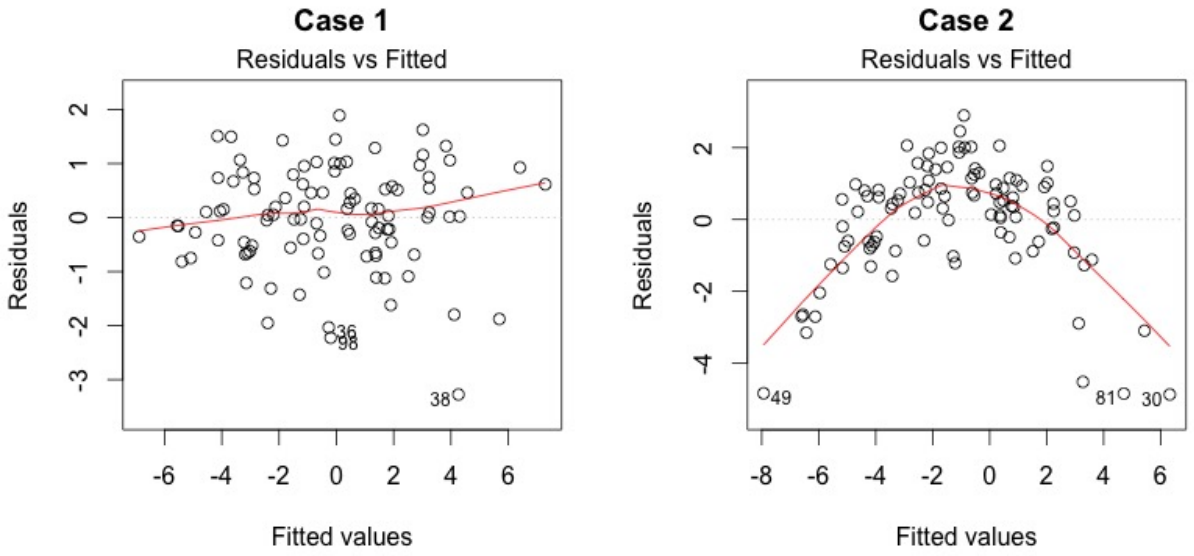

Figure 1:

15. **(6 points)** Given the residual plots above. What do they suggest about the assumptions for case 1 and case 2?

16. **(5 points)** From the model in question 12, Interpret the effect of `tar`, make sure it is in context to the problem.