PHP 2511 Midterm Exam 2018 Solutions

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.

- ***Linearity:*** Function f is linear in the coefficients.
- Mean of error term is 0.
$$E(\varepsilon) = 0$$
- ***Independence:*** Error term is independent of covariate.
$$Corr(X, \varepsilon) = 0$$
- ***Homoscedacity:*** Variance of error term is same regardless of value of X.
$$Var(\varepsilon) = \sigma^2$$
- ***Normality:*** Errors are normally Distributed

Points were taken off for missing either the assumptions or missing the brief explanation of 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?

When we consider logistic regression we have a binary outcome which comes from a binomial distribution. With this distribution we are actually modeling the probability of a success. This poses an issue as a probability ranges from 0 to 1. However when we place this into a linear regression, the values range from $-\infty$ to ∞. So our model predictions will be outside of the interval in which we need. This is there the logit function comes in. When we take the log of the odds we expand our 0 to 1 range into $-\infty$ to ∞.

Points were taken off for not stating that the logit function transforms the probability from 0 to 1 range into $-\infty$ to ∞. Also points were taken off for stating that logit bound the systematic part of the regression to 0 to 1 values, as this is incorrect, the inverse of the logit function bounds the regression.

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

A confounder is a covariate which is associated to the outcome and a a covariate of interest in our model. It also must not be within the causal pathway between the covariate of interest and the outcome. By failing to control for confounding a regression may actually find an effect that is not truly present or give an incorrect estimate of an effect. Essentially failing to adjust for confounding leads to biased results.

Points were taken off for not explaning why we need to control for it.

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

- The data $Y_1, Y_2, \ldots, Y_2$ are independently distributed.
- The dependent variable Y_i is from an exponential family.
  - Normal (Gaussian)
  - Bernoulli
  - Binomial
  - Multinomial
  - Exponential
  - Poisson
- Linear Relationship between link function and systematic component.
- Errors are independent.

Points were taken off for missing one or more of these.

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

One simple method is to consider homogeneity is to plot a scatter plot of the residuals and the fitted values. Then we can look to see if there appears to be an increase or decrease in the vertical spread of the residuals. Basically, we can look for a cone shape happening. We can also use the `olrss` package in R and the `ols_score_test()` or `ols_f_test()` function.

Points were taken off by suggesting to look for any type of pattern in the data or incorrect tests. We could still have heteroskedasticity with a non-linear function.

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

When we fail to have normal residuals that we make incorrect inferences about the significance of the coefficients. We still can achieve good estimates themselves but the ability to test if they are significant is based on a t-test which requires normally distributed data. So if we do not have normality of the residuals than we cannot trust our variances or our p-values on the model. We will still have good estimates though.

Points were taken off for discussing how this means the data is not linear, this is not the same assumption and some other errors.

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.

Each of the variables are significant at the $\alpha = 0.05$ level, since the p-values are all less than 0.05. In addition to this all of the coefficients of the model are positive which suggests that an increase in any of these covariates, than this leads to an increase in carbon monoxide.

Points were taken off for suggesting these were odds ratios or incorrectly stating that some led to a decrease in average carbon monoxide levels. Some points were taken off for making a statement of which has the greatest impact on the data due to the coefficient size. For example, inches to feet leads to inches having an efffect that is 12 times larger than feet but this does not mean it is more important.

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

When we consider variation we are considering R^2 values. This would mean that tar explains the most variation as its $R^2 = 0.917$ is larger than the others. In addition, this would mean that tar explains 91.7% of the variation in carbon monoxide.

Points were taken off for incorrectly using the Adjusted R^2 and interpreting this as explained variance. Adjusted R^2 is only used for comparing models and does not allow for interpretation of the amount of explained variance.

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?

When we compare the multiple linear regression to the univariate regressions we can see that tar has increased slighlty and is still significant. In addition we can notice that the confidence interval of tar has increased in width as well. We can then notice that nicotine and weight have changed both in magnitude and direction.

These extreme changes will be found to be due to Multicollinearity. Yes confounding also may play a role but to see a complete change in direction and magnitude is most likely due to multicollinearity.

Points were taken off for forgetting to mention multicollinearity, failing to mention about the changes of magnitude and direction. Again confounding can lead to biased results but the main cause of drastic changes in magnitude and direction is multicollinearity as mentioned in class.

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

When we consider the fit and compare it to other models we need to compare the adjusted R^2 values. When we do this, we see that this model is better than the model with just nicotine or weight, however the adjusted R^2 of the multivariable model is 0.907 and tar alone has an adjusted R^2 of 0.913. This means that the model with tar alone is actually a better fit for the data.

Points were taken off for incorrectly interpreting the adjusted R^2. You must remember not to provide extra intrepretations that were not asked for as you can make big mistakes by doing this.

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?

With this model we can see that we have 2 VIF values over 21. This would mean that both tar and nicotine are highly correlated with each other. However, weight is not highly correlated with the rest of the variables.

Points were taken off for incorrectly interpretting VIF as being about explained variance or by suggesting that all values were highly correlated as weight is an acceptable level or correlation with other covariates.

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.

When we run this regression model we can see that tar, binary nicotine and the interaction between tar and binary nicotine are all significant. We also note that tar has a similar effect to the other previous regression models which we have seen. The interaction tells us that there are 2 differenct regression lines, one for low nicotine and the other for high nicotine. In addition, the interaction term suggests that the effect of tar is decreased in the model with high nicotine levels, however this can be difficult to interpret still because the intercept with high levels of nicotine is on average 7.9 units higher.

Points were taken off for suggesting that tar had a negative effect as the interaction is negative but not large enough to negate the effect of tar. Points were also taken off for general incorrect interpretations or not stating anything about the interaction.

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
```

When we consider this test we are testing the hypothesis:

$$H_0: \text{ The smaller model is just as good as the larger model}$$

vs

$$H_1: \text{ The smaller model is not as good as the larger model}$$

Then we find that the p-value is less than 0.05 which means that we reject the null in favor of the alternative. We need to go with the interaction model which is Model 1.

Points were taken off for incorrect hypothesis as well as choosing the wrong model. The R output shows which is the main effects and which is the interaction.

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?

The adjusted R^2 is higher than all of the previous models. This means that this is better than all models we have seen. Remember adjusted R^2 is not the amount of variation that is explained.

Points were taken off for incorrectly interpreting the results. Again do not provide extra details that were not asked for. It can lead to incorrect results on your work.

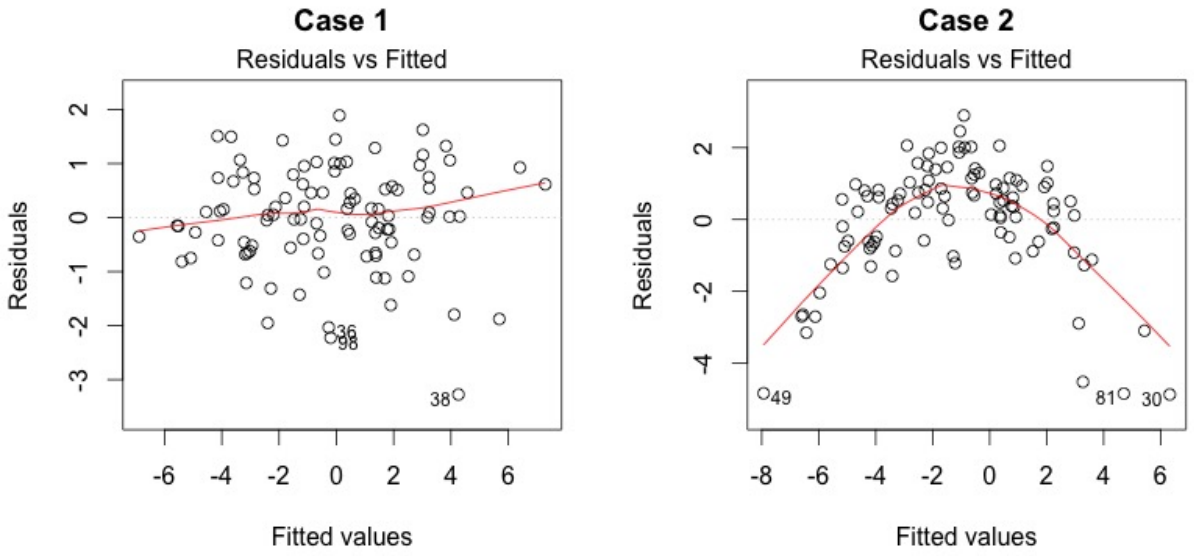

Figure 1:

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

For case 1, we can see that the residuals appear to be centered around 0. This would suggest that they have a mean of zero. In addition we can see that there are no major patterns which would suggest that a model with linear fit is reasonable. We can finally see that the range of values of residuals for any fitted value is similar suggesting that the variances are equal.

For case 2, we can see that there is a pattern in the residuals which is in the shape of a parabola. This suggests that the model is not linear. This does not tell us about normality, nor is it just due to outliers. We can see that the variances do appear to be similar and that we do not appear to have errors which are centered around 0.

Points were taken off for incorrect statements. For suggesting the pattern informed about normallity and other incorrect statements.

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

When we consider tar, we can only intrepret this single coefficient in the model where nicotine has low levels. Then if we consider 2 cigarettes, a cigarette with a 1 mg increased in tar will have on average 0.915 mg increase in Carbon monoxide compared to the cigarette with lower tar levels.

Points were taken off for incorrectly interpretting the results.