PHP 2511
Spring Midterm Examination Solutions
Wednesday, 16 March 2016

Name: _____

The Data

These data are from a randomized study to evaluate an antiepileptic drug in reducing the frequency of simple or partial seizures. The treatment variable is coded 0 for placebo and 1 for the drug under study, seizures is the outcome variable and indicates the number of seizures experienced over a two-week period during treatment. The baseline variable is the pre-treatment seizure rate over 8 weeks prior to randomization and is a potential adjustment variable in addition to the age of the patient (variable: age).

Variable	Description
Seizures	Number of seizures over 2 week period.
Treatment	Type of Treatment
0	Placebo
1	Drug
Baseline	Seizure Rate pre-treatment.
age	age of patient.

1. (5 points) We are interested in predicting the number of seizures. We run simple linear regressions on the variables in the data set and the results are as reported below:

	Estimate	95% CI of Estimate	p-Value	R^2
treatment	-0.776	(0 , 960.901)	0.84289899	0.00069
baseline	0.439	(1.422 , 1.692)	0	0.63218
age	0.018	(0.552 , 1.876)	0.95465464	6e-05

Which variable explains number of seizures best on its own? How do you know this?

From this output we can see that the `baseline` rate of seizures actually explains the number of seizures over the 2 week period of the trial best. We can see this from the R^2 value which is 0.63. This means that `baseline` explains 63% of the variation in `seizures`.

Points	Reason for Points
1	Stating that baseline explains best.
4	Explanation

2. (11 points) We decide to build a multiple regression model. *(See the R output for the summary of this)*

   a. (6 points) Are there any changes in the estimates from the simple linear regressions? Explain.

   We can see that the estimate of `baseline` has a slight increase. We can also see that estimate of `treatment` also has a slight increase however it remains insignificant. The largest change comes with the estimate of `age` which is now significant and had a very large increase in the effect size.

Points	Reason for Points
2	Stating `baseline` change or lack of change
2	Stating `treatment` change or lack of change
2	Stating `age` change or lack of change

b. (5 points) Interpret the effect of `baseline` in the context of this problem.

If we were to compare 2 subjects in this study one who has a `baseline` rate 1 unit higher than the other, the subject who has higher `baseline` would expect on average to have 0.46 more seizures during the 2 week period compared to the subject with the lower `baseline`.

Points	Reason for Points
5	Correct Interpretation

3. (5 points) One of your colleagues suggest that the reason `treatment` is proving to have been insignificant is that there may be an interaction between the `baseline` and `treatment`. What would interaction mean in this case?

   Interaction in this case would mean that the effect of `baseline` on `seizures` would be different for those who receive the placebo compared to those who receive the drug. In other words we would end up with 2 different regression lines one for those who have `treatment`=1 and one for those who have `treatment`=0.

Points	Reason for Points
5	Correct Explanation

4. (13 points) You decide to run the model with the interaction term added in. *(See the R output for the summary of this)*

   a. (5 points) Does adding the interaction term in improve the model? Explain.

   Interaction does improve the model fit. We see that in question 2 we have an adjusted R^2 of 0.6399. However with the addition of the interaction term we now have an adjusted R^2 value of 0.7161. This is almost a 0.08 increase. On top of this increase in adjusted R^2 we can see that all of the covariates in the model now have a significant estimated effect.

Points	Reason for Points
3	Comparing Adjusted R^2
2	Comparing P-values

   b. (8 points) Write out the 2 models that we now have to explain these variables effect on the number of seizures. *(No need to actually add any coefficients just write out the equations and leave the additions).*

   Model for Placebo:

   `seizures` = -14.38252 + 0.28504`baseline` + 0.51602`age`

   Model for Drug:

   `seizures` = (-14.38252 - 10.33219) + (0.28504 + 0.31535)`baseline` + 0.51602`age`

Points	Reason for Points
3	Correct model for placebo
5	Correct model for Drug

5. (9 points) You decide to keep moving with the model you have above since it uses all of the variables in the model and has a decent R^2. You move onto testing the model with residual plots. What do these residual plots tell you about the fit? *(See the R output for the plots.)*

   We can see that the `age` residual plot that our residuals are spread out around zero and appear to be randomly spread. There may be a slight pattern with `age` suggesting that it may not be a completely linear relationship with `seizure`.

   We can see that the `baseline` residual plot that our residuals are spread out around zero and appear to be random. There may be an issue with an outlier here given the last value is much further away than the rest.

   We can see that the `treatment` residual plot that our residuals are spread out around zero. We may find there there is larger variation in those that are on the drug than others. This could signify an issue with our regression assumption of constant variance. Once again there seems to be a value that may be an outlier.

Points	Reason for Points
3	Description of age residual plot
3	Description of baseline residual plot
3	Description of treatment residual plot

6. (8 points) You then move onto marginal model plots. What does this marginal model plot tell you about how this model fits the data? *(See the R output for the plots.)*

   We can see that the `baseline` variable fits the data well but there is an issue with one point pulling the data away from the model at the last point.

   We can see that the `treatment` variable fits the data really well and there is no distinguishment between the data and the model.

   We can see that the `age` variable fits the data well. There is a gap between the model and the data after about age 32. It does not appear to be a large gap though.

   We can see that overall the fit is not terrible but the model does overfit some of the data early on and then the point that could be an outlier pulls the data above the model at the end.

Points	Reason for Points
2	Description of age plot
2	Description of baseline plot
2	Description of treatment plot
2	Description of fitted values plot

7. (8 points) You decide that before moving on you wish to test for outliers. You use a Cook's Distance plot. Do you find any outliers? If so how many are clearly outliers? *(See the R output for the plots.)*

   From Cook's Distance we can see that most of the data fall below the cutoff line. Observations 29 and 38 may be outliers but they seem to close the cutoff to call from this graph alone. There is a clear outlier with observation 49. It has a cooks distance well above all of the others.

Points	Reason for Points
4	Mention of 29, 38
2	Clear outlier of 49
2	Overall explanation of plot

8. (8 points) You remove the largest outlier and proceed from there. Does the summary change from before the outlier was removed? If so, in what way? *(See the R output for the summary.)*

   We first notice that only the effects of `age` and `baseline` are significant now. The magnitude of `baseline` has not changed, however `age` is about half of what it was before. Finally we notice that the adjusted R^2 has actually gone down to 0.5542. We now see that a large amount of variation before was due to the outlier that was present.

Points	Reason for Points
2	Noting change of significance
2	Noting lack of change in baseline
2	Noting change in age
2	Noting change in R^2 and/or adjusted R^2

9. (10 points) Do the plots displayed show that this model excluding the outlier has a better fit? Explain. *(See the R output for the plots.)*

   With the outlier removed we see that all of our residual plots have improved. There is more random spread in `age` and `baseline`. There still may be a slight patter in `age` but this is hard to tell from the plot. `treatment` seems to have improved in constant variance. Overall they are a better fit.

   With the marginal model plots we see that `baseline` fits the data better and has hardly an deviation between the model and the data. We can also see that `treatment` also fits the data similar to before, only a slight gap between the model and the data. We see that `age` still has some issuesit would appear that it fits the data worse than it did before the outlier was removed however this is due to a change in the scale of the. Finally the fitted values fit better than they did before the outlier was removed. It appears it was a good idea to remove this outlier.

Points	Reason for Points
5	Discussion of Residual plots.
5	Discussion of Marginal Model plots.

10. (8 points) The PI on the study is concerned that there was a problem with the randomization. They feel the treatment does do something but someone made an error in the design. If there is no relationship between treatment and baseline or treatment and age, then we can rule out randomization as being a flaw. Luckily you know logistic regression and you run 2 simple logistic regressions. *(See the R output for the summaries.)*

a. (4 points) Is the PI right? Do you find associations between treatment and age/baseline?

Given the outputs from R we can see that neither model produces a significant effect of either `age` or `baseline` on `treatment`. This means that in our data there is no association between these variables. The issue with treatment not being effective is not due to the randomization it appears.

Points	Reason for Points
2	Mention of age significance
2	Mention of baseline significance

b. (4 points) Interpret the effect of baseline on treatment. (*Do this regardless of significance.*)

If we compare 2 patients who differ in baseline seizure rate by only 1 unit, the patient with a higher rate of baseline seizures would have an odds of being treat 0.993 times that of the odds for the patient with the lower baseline rate.

Points	Reason for Points
4	Correct Interpretation