STAT 425 — Spring 2016 — Sections 1UG, 1GR

Homework 3

Please submit your assignment *on paper*, following the Formatting Guidelines for Homework Submission. (Even if correct, answers might not receive credit if they are too difficult to read.) Remember to include relevant computer output.

1. Consider the `MinnWater` data set from R package `alr4`.

   (a) [2 pts] Produce a *pairs* plot of all variables in the data set.

   (b) [2 pts] Three of the variables appear to have especially high sample correlations with each other. Which are they?

   (c) [2 pts] Fit the linear regression model with `muniUse` as the response, and *all* other variables as the regressors. Produce a summary of the model fit.

   (d) [2 pts] Compute the variance inflation factors (VIFs) for the variables. Using the threshold of 10 to determine if a VIF indicates a problem of (approximate) collinearity, which variables have a VIF indicating a possible problem?

   (e) [2 pts] Fit the linear regression model with `muniUse` as the response, but only `allUse`, `irrUse`, `muniPrecip`, and `statePop` as the regressors (the variables that were individually significant in the original analysis). Produce a summary of the model fit. Are all of those variables still significant?

   (f) [2 pts] Compute the VIFs for the reduced set of variables in the previous part. Relative to the original fit, have the VIFs for those variables increased? Decreased? Stayed the same? Using the threshold of 10, which have a VIF indicating a possible problem?

2. Nineteenth century economist W. Stanley Jevons was concerned about the loss of value in coins due to their loss in weight while in circulation. He collected, cleaned, and weighed 274 gold sovereigns, then grouped them roughly according to age (in decades):

Age	Number	Average Weight (g)	Standard Dev.
1	123	7.9725	0.01409
2	78	7.9503	0.02272
3	32	7.9276	0.03426
4	17	7.8962	0.04057
5	24	7.8730	0.05353

   (a) [2 pts] Perform the usual (unweighted) simple linear regression of Average Weight on Age. Produce a summary of the results.

   (b) [2 pts] Form a 95% confidence interval for the regression slope, that is, the average weight change per decade. (You may use the function `confint`.)

   (c) [2 pts] Perform a *weighted* regression of Average Weight on Age, using the group sizes as weights (for weighted least squares). Produce a summary of the results.

   (d) [2 pts] What are the weight matrix $\boldsymbol{W}$ and the matrix $\boldsymbol{\Sigma}$ for the analysis of the previous part?

(e) [2 pts] Under this weighted model, form a 95% confidence interval for the slope. (Again, you may use the function `confint`, which also works for weighted models.)

(f) [2 pts] Notice that the standard deviations seem to be different for different groups. Perform another *weighted* regression of Average Weight on Age, this time using weights that also account for the different standard deviations of the different groups. (Refer to the notes on *Weighted Least Squares* for a formula that incorporates both group sizes and standard deviations.) Produce a summary of the results.

(g) [2 pts] What are the weight matrix $\boldsymbol{W}$ and the matrix $\boldsymbol{\Sigma}$ for the analysis of the previous part?

(h) [2 pts] Under this new weighted model, form a 95% confidence interval for the slope.

3. The `lakemary` data from R package `alr4` contains ages and lengths of 78 bluegill fish captured from Lake Mary, Minnesota.

(a) [2 pts] Fit a simple linear regression with `Length` as the response and `Age` as the predictor. Produce a summary of the regression results.

(b) [2 pts] Produce a plot of `Length` versus `Age`, and include a fitted regression line (from your least-squares fit in the previous part) on the same plot.

(c) [2 pts] Briefly explain why it *is* possible to perform a lack-of-fit test for this data, even though the error variance is unknown.

(d) [2 pts] Perform a lack-of-fit test for the simple linear regression. (Show both your R code and the results.) Interpret your results. (What can you say about lack of fit?)

(e) [2 pts] Compute an estimate $\tilde{\sigma}^2$ of the *pure error variance*. (This is the sum of squares for pure error divided by its degrees of freedom.)

(f) [2 pts] Now fit a *quadratic* regression: Use the formula `Length ~ Age + I(Age^2)`. Produce a summary of the regression results.

(g) [2 pts] You can perform a lack-of-fit test for this quadratic regression model in the same way as for the simple linear regression. Perform that lack-of-fit test and interpret the results. (Show your code and output.)

4. [GRADUATE SECTION ONLY] Consider the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\mathrm{E}(\boldsymbol{e}) = \boldsymbol{0}$ and $\mathrm{Var}(\boldsymbol{e}) = \sigma^2\boldsymbol{\Sigma}$, for known, invertible $\boldsymbol{\Sigma}$.

(a) [2 pts] Find $\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right)$ for the ordinary least squares estimator $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.

(b) [2 pts] Find $\mathrm{Var}\left(\hat{\boldsymbol{\beta}}_G\right)$ for the generalized least squares estimator $\hat{\boldsymbol{\beta}}_G = \left(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$. (Simplify your answer.)

5. [GRADUATE SECTION ONLY] Revisit the weighted model you fit in 2(c).

(a) [2 pts] Compute the unweighted residuals as $\hat{\boldsymbol{e}}_W = \boldsymbol{y} - \hat{\boldsymbol{y}}_W$, where $\hat{\boldsymbol{y}}_W = \boldsymbol{X}\hat{\boldsymbol{\beta}}_W$ and $\hat{\boldsymbol{\beta}}_W$ is the weighted least squares estimate.

(b) [2 pts] Compute the weighted residuals as $\widehat{\boldsymbol{\delta}} = \boldsymbol{W}^{1/2}\hat{\boldsymbol{e}}_W$, where $\boldsymbol{W}^{1/2}$ is the diagonal matrix with non-negative elements such that $\boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2} = \boldsymbol{W}$.

(c) [2 pts] What values does the R `residuals` function compute for your model from 2(c)? Are they unweighted? Weighted? Something else?

Some reminders:

- Unless otherwise stated, all data sets are either automatically available or can be found in either the `alr4` package or the `faraway` package in R.
- Unless otherwise stated, use a 5% level ($\alpha = 0.05$) in all tests.