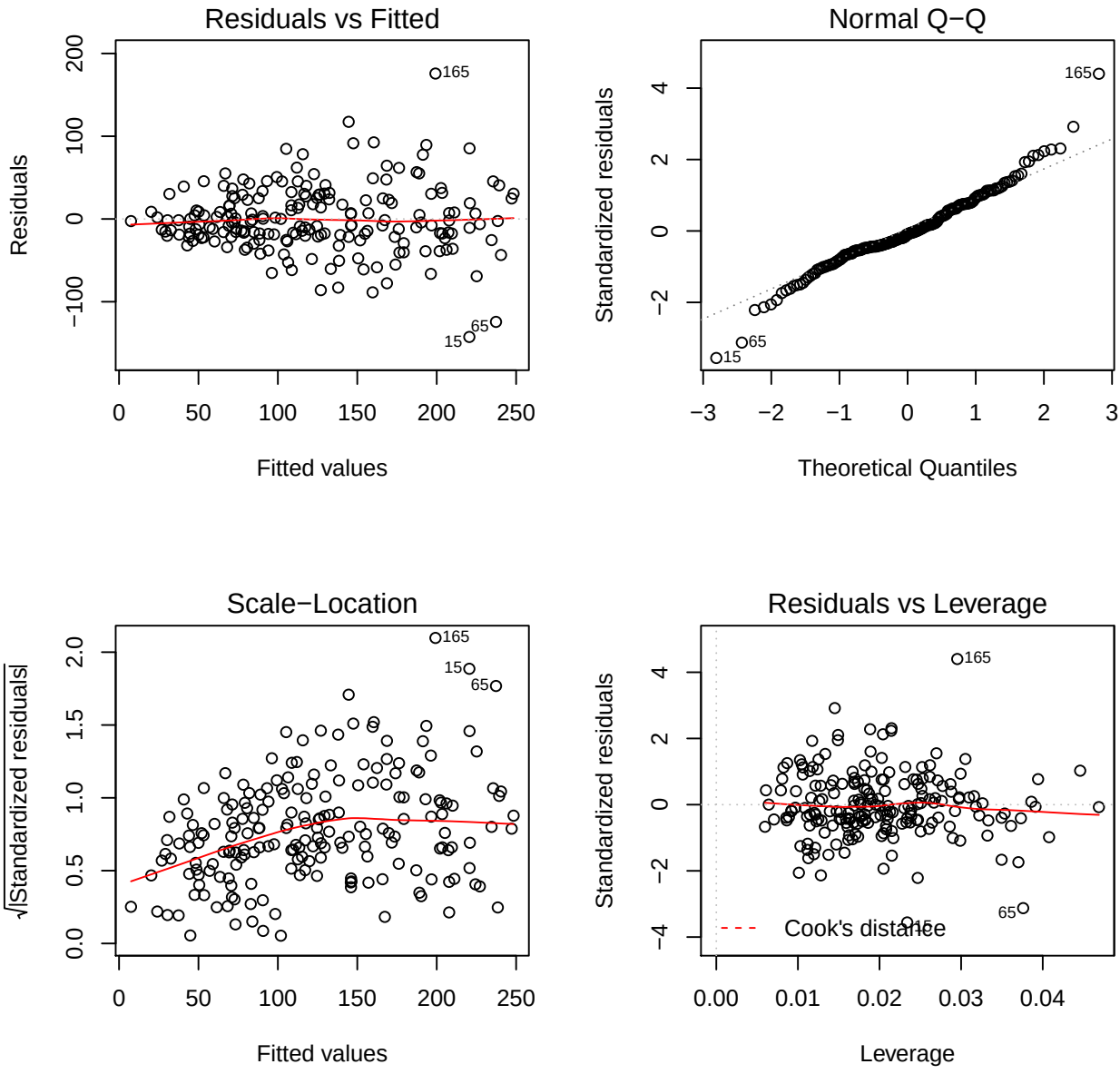
Exam 2 - STAT 757 **Name:** _____

Instructions: You may use any resources to take this exam (books, web resources, etc.), except for help from people other than the instructor. Each problem is worth 20 points. The exam will be graded out of 100 (5 extra credit points are possible). You are encouraged to ask the instructor questions, and to consult any "Hints" posted on the course website. Partial credit WILL be awarded where sufficient details have been provided. **Due before class on Tuesday, 3 May.**

1. Which of these plot should be your primary graphical tool to assess the MLR assumption of constant variance? For this particular example, explain whether or not this assumption holds. If it does not, briefly describe the next step you would take to try and correct the problem.



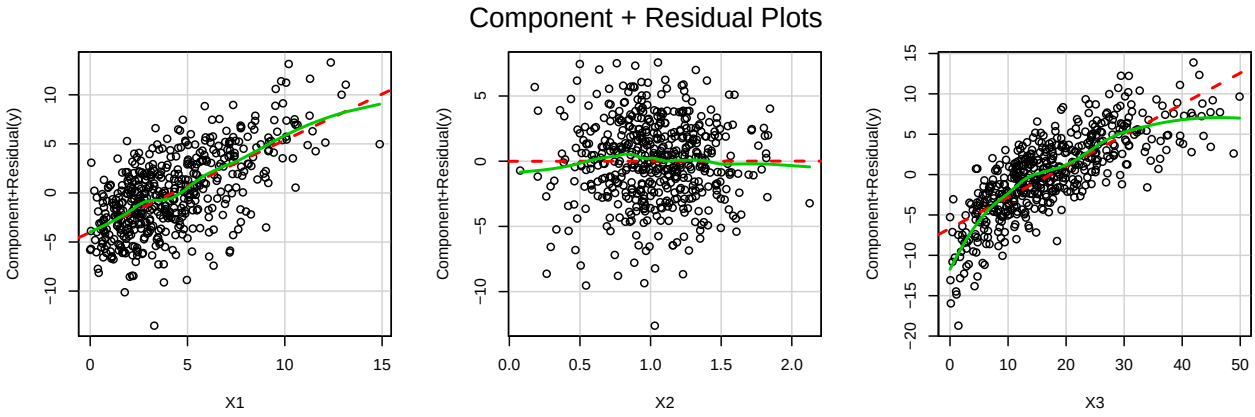
2. Describe (briefly) the difference between how you use **Added Variable Plots** and **Marginal Model Plots** as diagnostic tools for Multiple Linear Regression. What sorts of problems (i.e., assumption violations) does each reveal? For each of those problems, give at least 1 action someone might take to correct them?

3. Another diagnostic plot is the **Component + Residual Plot**, also called **Partial Residual Plots**.

   a. Look up the definition and use of Partial Residual Plots, and describe which assumption violations they can help detect.

   b. What changes would you make to the regression model $y_i = \beta_0 + \beta_1\, x_{1,i} + \beta_2\, x_{2,i} + \beta_3\, x_{3,i} + e_i$ (as implemented in `ex2-prob3.R`) based on the following **Partial Residual** plots?

   Bonus (5 pts): Improve the model above by implementing those changes, then compare 95% prediction intervals to the observations in `ex2-prob3-ec.csv` (see problem 4).



4. Suppose you are trying to model the number of social contacts made by a population (N=950) of teenagers (ages 14-18) using as predictors their gender, age, and number of individuals living in their household. You randomly divide these data into a group of 900 (your *training data*) and a group of 50 (*test data*), and plan to look at how well a regression model (based on the training data) can predict the observed number of contacts in the test data. Work through the R script `ex2-prob4.R` and explain how you use the given diagnostic plots to determine which model fits the data best. (If you think it necessary, you may do additional analyses and include those results in your discussion).

5. Checking the *variance inflation factor* for each covariate in the regression conducted in ~~`ex2-prob4.R`~~ `ex2-prob5.R` reveals the following *colinearity* problem. Describe steps you would take to correct the problem, and modify the analysis accordingly (i.e., modify the R code and reanalyze the data). Describe how your modification alters the outcome in terms of how well the resulting model fits the data, and in terms of the regression coefficients.

```
fit=lm(y~.,mydat)
summary(fit)

##
## Call:
## lm(formula = y ~ ., data = mydat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.911 -13.759  -2.354  15.021  51.190
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  32.5524     6.6496   4.895 4.01e-06 ***
## X1           -0.7406     0.1222  -6.059 2.74e-08 ***
## X2            0.6276     0.3408   1.841  0.06868 .
## X3            0.9616     0.3252   2.957  0.00392 **
## X4           -0.3850     0.3175  -1.212  0.22836
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.98 on 95 degrees of freedom
## Multiple R-squared:  0.427,Adjusted R-squared:  0.4029
## F-statistic:  17.7 on 4 and 95 DF,  p-value: 6.924e-11

vif(fit)

##        X1        X2        X3        X4
##  1.041841  8.852165  8.721592 17.214683
```