1. Data analysis - nonconstant variance I. Sample R code for this problem and the next problem is provided below.

(a) As hinted in part (b), the residual plot shows a pattern in that the variance of the residuals appears to increase as the fitted value increases.

(b) Handling non-constant variance

i. Transformations - The square root transformation yields a better residual plot and higher $R^2(=0.9333)$ than the log transformation ($R^2 = 0.8407$). For the square root transformation most of the same variables are significant though common area size is now significant and the number of desk hours just misses. Note that $R^2$ in the original untransformed scale is higher (0.9613) but the residual plot indicates the assumptions are not satisfied on the original scale.

ii. Weighted least squares - The residuals actually show increasing variance against almost all predictors so we could have used any here. The WLS residual plot does not show increasing variance; it actually seems to show decreasing variance. Perhaps the weighting is too extreme. If also changes which variables are significant, only number of rooms and common area are significant at the .05 level (capacity is close to significance).

iii. Sandwich estimator - The vcovHC function provides a number of options. The HC2 option uses residuals divided by $(1 - h_{ii})$, which creates more nearly constant variance, in the middle of the sandwich. (The HC option uses the observed residuals.) Examining the new standard errors show that the they have increased. It does not change which variables are significant but should provide more reliable confidence intervals.

(c) The advantage of the transformation approach is that it (arguably) creates a data set that satisfies our linear regression assumptions. The main disadvantage is that it is harder to interpret. The weighted least squares approach and sandwich estimator approach allow us to maintain the nice linear model interpretation. WLS works sometimes but doesn’t seem very compelling here as its hard to know why only checkins should impact the variance. The robust standard error approach doesn’t really have a disadvantage in this case so might be preferred here.

2. Data analysis - nonconstant variance II.

(a) Here we go back to the original, untransformed regression. For those data, observations 22, 23, 24 are highly influential (Cook’s D > 1) and have high leverages. Observation 23 has an especially high Cook’s D. The studentized residuals for observations 23, 24 are also high. It is a good idea to examine the data. In doing so, we find that 22, 23, 24 are all large buildings and that for building 23 the avg occupancy is greater than its capacity (this seems odd!).

(b) Without case 23 the residual plot is improved. There is still some slight evidence of increasing variance but it is minor. Notice that observations 22, 23 (which used to be 22, 24) are still high leverage but not nearly so influential. Also note that avg occupancy and checkin are significant which makes good sense.

(c) The writeups should focus on the question that motivated the study. Short write-ups should describe the problem under study, the key results (including sign and magnitude of important relationships), and any limitations or future directions. Here’s a possible paragraph for these data – This study attempts to develop a formula or approach for predicting US Navy dormitory manpower requirements. The analysis suggests that it is critical for the Navy to determine the nature of the buildings under consideration. In particular one observation in the current data set does not appear to be like the others and this observation plays a key role in the analysis. If observation 23 has been misclassified or the data recorded in error, then it should be removed from the data set. In that case, there is a fairly natural model forecasting expected manpower (in man-hours required per month) as $21\times$average daily occupancy plus 1.4*average monthly checkin. If on the other hand we must find a model that appears to work across this full set of dormitories, then it might be best to rely on a model which explains square root of manpower in terms of capacity, number of rooms, average monthly checkins, and weekly hours of desk operation.

R code

```r
library(tidyverse)
datainput <- read_csv("H://HAL/Courses/Stat210//navydorm.csv")
navydorm <- data.frame(datainput)
#
# fit regression
#
manpower_reg <- lm(manpower ~ avgoccup + checkin + desk + area + wings + capacity + rooms, data=navydorm)
summary(manpower_reg)
manpower_reg$stdres <- rstandard(manpower_reg)
ggplot() +
  geom_point(data=navydorm, mapping=aes(x=manpower_reg$fitted.values, y=manpower_reg$stdres))
qqnorm(manpower_reg$stdres)
```
# try transformations
#
navydorm$logmanp <- log(navydorm$manpower)
navydorm$sqrtmanp <- sqrt(navydorm$manpower)

logmanp_reg <- lm(logmanp ~ avgoccup + checkin + desk + area + wings + capacity + rooms, data=navydorm)
summary(logmanp_reg)
logmanp_reg$stdres <- rstandard(logmanp_reg)

ggplot() +
  geom_point(data=navydorm, mapping=aes(x=logmanp_reg$fitted.values, y=logmanp_reg$stdres))
qqnorm(logmanp_reg$stdres)

sqrtmanp_reg <- lm(sqrtmanp ~ avgoccup + checkin + desk + area + wings + capacity + rooms, data=navydorm)
summary(sqrtmanp_reg)
sqrtmanp_reg$stdres <- rstandard(sqrtmanp_reg)

ggplot() +
  geom_point(data=navydorm, mapping=aes(x=sqrtmanp_reg$fitted.values, y=sqrtmanp_reg$stdres))
qqnorm(sqrtmanp_reg$stdres)

# try WLS
#
manpower_wlsreg <- lm(manpower ~ avgoccup + checkin + desk + area + wings + capacity + rooms, weights = 1/checkin, data=navydorm)
summary(manpower_wlsreg)
manpower_wlsreg$stdres <- rstandard(manpower_wlsreg)
plot(manpower_wlsreg$fitted.values, manpower_wlsreg$stdres)

# try sandwich estimator (type=const gives ordinary least-squares variance estimates; type=HC2 is a sandwich estimator)
#
install.packages("sandwich")
library(sandwich)
sand_0 <- vcovHC(manpower_reg, type = "const")
sand_v2 <- vcovHC(manpower_reg, type = "HC2")
summary(manpower_reg)
chbind(sqrt(diag(sand_0)), sqrt(diag(sand_v2)))

# case diagnostics (for untransformed regression)
#
manpower_reg$dffits <- dffits(manpower_reg)
manpower_reg$cooksd <- cooks.distance(manpower_reg)
manpower_reg$hat <- hatvalues(manpower_reg)
manpower_reg$studentr <- rstudent(manpower_reg)
diagnostics <- cbind(navydorm$manpower, manpower_reg$fitted.values, manpower_reg$residuals, manpower_reg$stdres,
  manpower_reg$studentr, manpower_reg$hat, manpower_reg$dffits, manpower_reg$cooksd)
dimnames(diagnostics)[[2]] <- c("manpower", "fitted", "resid", "std.res", "stud.res", "leverage", "dffits", "cooksd")
round(diagnostics, 4)

# rerun regression without case 23
#
navydorm2 <- navydorm[-23,]
manpower2_reg <- lm(manpower ~ avgoccup + checkin + desk + area + wings + capacity + rooms, data=navydorm2)
summary(manpower2_reg)
manpower2_reg$stdres <- rstandard(manpower2_reg)

3. Model assumptions (theory): incorrect model specification

(a) Let’s start by calculating the SS(Residuals). It turns out that

\[ SS(Residuals) = Y^T(I - H)Y = (X\beta + Z\alpha + \epsilon)^T(I - H)(X\beta + Z\alpha + \epsilon) \]

\[ = (X\beta + Z\alpha + \epsilon)^T(I - H)X\beta + \beta^T X^T(I - H)Z\alpha + \beta^T X^T(I - H)\epsilon + \alpha^T Z^T(I - H)Z\alpha + \alpha^T Z^T(I - H)\epsilon + \epsilon^T(I - H)^TZ\alpha + \epsilon^T(I - H)^T\epsilon. \]

The last equality is obtained by noticing that \((I - H)X = X^T(I - H) = 0\) (see HW 5) which makes the first three terms zero. The first term in the last expression is a constant and the second term has mean zero since \(\epsilon\) has mean
zero. Thus \( E(\text{SS(Residuals)}) = \alpha^T Z^T (I - H) Z \alpha + E(\epsilon^T (I - H) \epsilon) \). The final expected value can be calculated by multiplying this quadratic form out and taking term by term expectations or by noting that \( E(\epsilon^T (I - H) \epsilon) = E(\text{tr}(\epsilon^T (I - H) \epsilon)) = E(\text{tr}((I - H) E(\epsilon \epsilon^T))) = \text{tr}(\sigma^2 (I - H) I) = \sigma^2 \text{tr}(I - H) = \sigma^2 (n - (k + 1)) \).

The latter argument uses properties of the “trace” which is the sum of the diagonal elements of a matrix and recalls that \( \sum h_{ii} = (k + 1) \). Putting everything together yields \( E(MS(\text{Residuals})) = E(\text{SS(Residuals)}) / (n - (k + 1)) = \sigma^2 + \frac{1}{n-(k+1)} \alpha^T Z^T (I - H) Z \alpha \) as advertised.

(b) If you have some knowledge about \( \sigma^2 \) then you can compare your regression \( MS(\text{Residuals}) \) to the known value of \( \sigma^2 \). Many people wondered about how you could carry out a test (e.g., by dividing the MS(Residuals by \( \sigma^2 \)) and whether it would be an F-statistic. It is not possible to answer this without knowing where the information about \( \sigma^2 \) came from. I intended a less formal answer. If the \( MS(\text{Residuals}) \) is greater than the value of \( \sigma^2 \) that you were expecting, then this is evidence that there are some variables effecting \( \eta \) that are not included in your model (these are the \( Z \)'s). Note that this argument is similar to the one used for the lack of fit test from question 5 on HW 4. It is not very often that we would have such information. The point of this question is to conceptually understand what \( \sigma^2 \) measures and how our estimate of it is impacted by unmeasured predictors.

4. Model assumptions (theory): non-constant variance

(a) The problem specifies that we believe a linear model is appropriate for individuals. Let’s use the subscript \( ij \) to denote the \( j \)th individual in family \( i \). Then we believe \( Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_i \) with \( \epsilon_i \sim N(0, \sigma^2) \). We only have the family average values \( Z_i = \frac{1}{n_i} \sum_j Y_{ij} \) and \( W_i = \frac{1}{n_i} \sum_j X_{ij} \) where \( n_i \) is the number of members in the family. By averaging our model for \( Y_{ij} \) over the members of the family, we find that \( Z_i = \frac{1}{n_i} \sum_j Y_{ij} = \frac{1}{n_i} \sum_j (\beta_0 + \beta_1 X_{ij} + \epsilon_{ij}) = \beta_0 + \beta_1 W_i + \eta_i \) where \( \eta_i = \frac{1}{n_i} \sum_j \epsilon_{ij} \) is \( N(0, \sigma^2/n_i) \). It follows from this last piece that the model for household averages will have the same intercept and slope as in the individual model but will have non-constant variance with bigger households providing more precise information. It is important to note that you will not get the same estimates for \( \beta_0 \) and \( \beta_1 \) from the family average data and from the individual level data. Thus the “estimands” are the same but the “estimates” are not.

(b) As we discussed in class the least squares regression estimate of the slope is still unbiased when there is non-constant variance. See the proof provided in class or in the text.

(c) In this case we actually have a good idea about the cause of the non-constant variance and therefore how to adjust for it. We should do weighted least squares with the weight for family \( i \) equal to the number of individuals in the family \( (n_i) \).

(d) This last part of the problem was just provided as a caveat about being sure that your model accurately reflects what is happening in the data.

i. Within each family \( Y \) increases linearly as \( X \) increases.
ii. If you look at the family-level data (there are only two families), then you find that families with higher values of \( X \) (family 2) have lower values of \( Y \).
iii. The argument in parts (a)-(c) do not apply here because there is not a single common regression model that is valid for the individuals in both family 1 and family 2. Thus in this case regression on the family data is not estimating the same quantity as a regression on the individual data. The conclusions of the family model would be relevant for families but not for individuals.