Exam 2

Kevin Linares

2024-11-08

Notes

- Label each answer with the appropriate question number in your R Markdown document (e.g., 1.1, 1.2, etc.).
- Clearly demonstrate your work. Where applicable, include any R code pertinent to your answer.
- Submit a single pdf file via Canvas by the deadline (930 am October 8).
- You may consult any reference materials except tools that utilize AI (e.g., ChatGPT, Github Copilot, etc.)
- This is not a group assignment so do not consult with your classmates. Your answers should be based on your own, individual work.
- The point value for each question is given in square brackets.

1. Why should we be cautious in using tests like the Shapiro-Wilk test of normality in assessing whether residuals are normally distributed? Be sure to describe how this is related to sample size [15pt].

- The Shapiro-Wilk test is a formal test for normality as the Null hypothesis being tested is that residuals come from a normal distribution. However, as the sample size increases mild deviations from non-normality may appear, suggesting that the likelihood of rejecting the Null is more likely as the sample size increases. On the other hand, with small samples there may not be enough power to reject the Null hypothesis, meaning that the test could fail to detect non-normal distributions. Finally, we have to take caution using these tests of normality since the p-value is not very helpful as an indicator of how to approach the violation of non-normal residuals. Therefore, also examining QQ-plots is recommended for a visual inspection of normality of the residuals. Of course, this violation, according to Gelman 2021, really only matters if using the model coefficients to make predictions.

2.1 Using the Hitters dataset from the ISLR2 package, fit a linear regression model with Salary as a function of Division and CHits. Save it as mod1. Refit the same model, but this time add an interaction term between Division and CHits. Create 2 scatter plots (one for each model) showing the observations and include line plots showing the regression lines of the respective model. Your outcome should be on the y axis.

- We fit two linear regressions with annual Salary (in thousands) as a function of the number of player's division, and number of hits during his career. In the second model we include an interaction term between division and number of hits during their career. The interaction coefficient in model 2 is significant and we reject the null hypothesis, suggesting that the relationship between the two predictors is not additive but rather the impact of one predictor varies depending on the other predictor's levels.

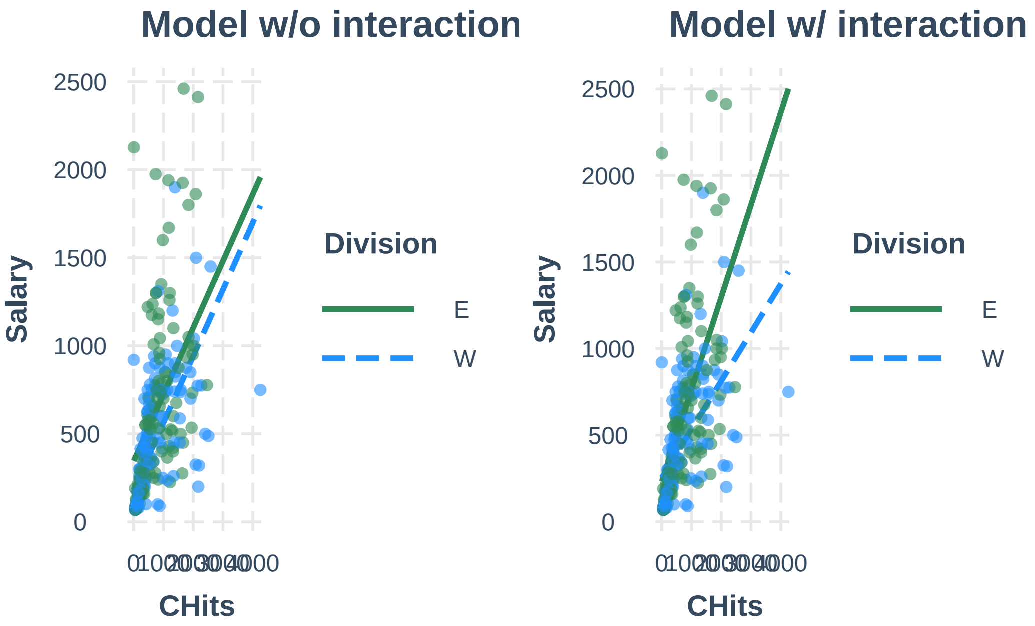
```r
data("Hitters")
mod1 <- lm(Salary ~ Division + CHits, Hitters)
mod2 <- lm(Salary ~ Division + CHits + Division*CHits, Hitters)

export_summs(mod1, mod2)
```

	Model 1	Model 2
(Intercept)	344.61 ***	230.64 ***
	(41.66)	(51.65)
DivisionW	-161.80 ***	22.24
	(45.61)	(67.96)
CHits	0.38 ***	0.53 ***
	(0.04)	(0.06)
DivisionW:CHits		-0.25 ***
		(0.07)
N	263	263
R2	0.33	0.37

*** p < 0.001; ** p < 0.01; * p < 0.05.

- We can see when we plot the observations against the prediction lines that Salary begins to widen between division at higher CHits, meaning that when we include the interaction term the slopes for division are different.

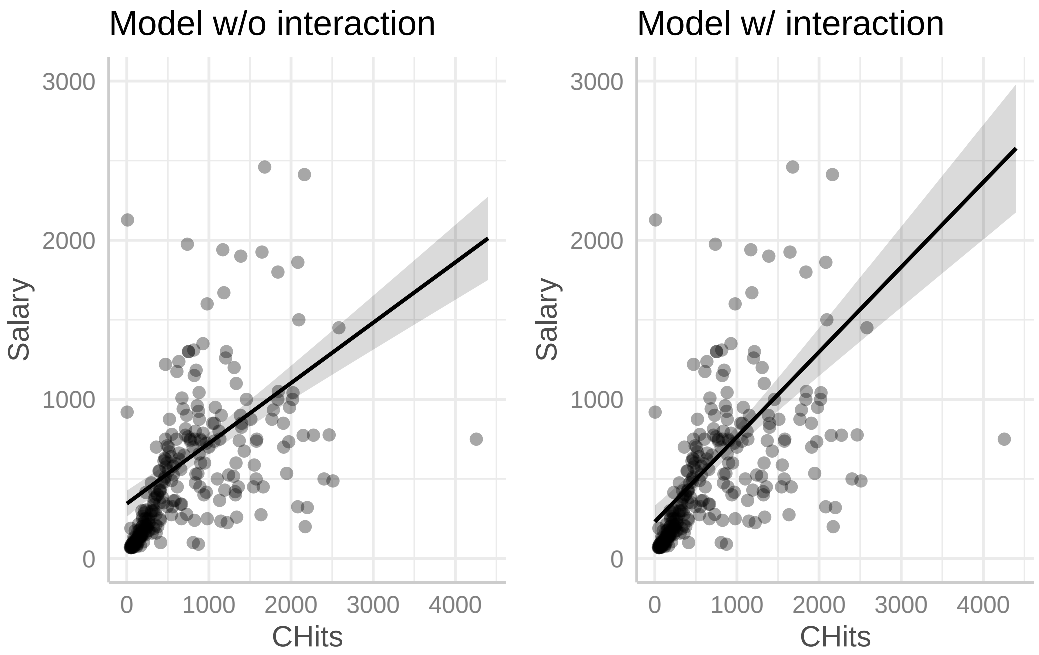
```
grid.arrange(
interact_plot(mod1, pred = CHits, modx = Division,
              plot.points = TRUE, centered = "none", jitter=.2,
              main.title = "Model w/o interaction",
              colors = c("seagreen", "dodgerblue")),

interact_plot(mod2, pred = CHits, modx = Division,
              plot.points = TRUE, centered = "none", jitter=.2,
              main.title= "Model w/ interaction",
              colors = c("seagreen", "dodgerblue")),
ncol=2)
```



2.2 Interpret the relationship between CHits and Salary in each model. Be sure to give your interpretation in the proper units for CHits, Salary, and Division (if applicable). [10pt]

- Model with no interaction: For each additional baseball hit in a player's career, we expect salary to increase by .38 on average, or in 1000s in translates to $380 USD annually.
- Model with interaction: For each additional baseball hit in a player's career, we expect salary to increase by .53 on average, or in 1000s in translates to $530 USD annually.
- For both models in the plot below (observed salary and CHits with model line of best fit) we see a positive linear relationship between these two variables, suggesting that they are associated, as one variable increases so does the other one. We also observe a stepper slope for the model with the interaction term.

```
grid.arrange(
  ggpredict(mod1, terms = c("CHits"), jitter=.2) |>
    plot(show_data=TRUE, show_ci=TRUE) +
    ggtitle("Model w/o interaction") +
    ylim(0,3000),
  ggpredict(mod2, terms = c("CHits"), jitter=.2) |>
    plot(show_data=TRUE, show_ci=TRUE) +
    ggtitle("Model w/ interaction") +
    ylim(0,3000),
  ncol=2)
```



2.3 Use the global F test to compare the models and state which one is preferable and why (based only on the result of the test).[5pt]

- We use a General F test to examine which model to keep by testing the Null hypothesis that $H_0 : RSS_{reduced} = RSS_{full}$ and we reject the Null hypothesis suggesting that the Residual Sum of Squares for the model with no interaction is statistically higher than the model with the interaction. Therefore, the model with the interaction minimizes the residuals a bit more and we prefer this model.

```
anova(mod1, mod2)
```

```
Analysis of Variance Table

Model 1: Salary ~ Division + CHits
Model 2: Salary ~ Division + CHits + Division * CHits
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1    260 35534342
2    259 33850527  1   1683815 12.883 0.0003964 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. Use PATH_MIDTERM.csv.

- Read in the data.

```
path <- read_csv("~/UMD/classes/stat_mod_ML_1_SURV615/exams/PATH_MIDTERM.csv")
```

- The data comes from a project called "Positive Attitudes Towards Health" in 2017 that targeted persons who inject drugs in Southeast Michigan.
    - The dataset includes 408 cases and the following variables:
    - SAMPLEID: Id of respondents
    - AGE: Age in years
    - MALE:
        - 1. Male;
        - 0. Female
    - BLACK:
        - 1. Race black;
        - 0. Other than black
    - EDUC:
        - 1. <=High school (HS) Education,
        - 2. HS Education;
        - 3. >HS Education
    - LIFESAT: Life satisfaction score summarized from 5 questions; Higher scores mean higher satisfaction
    - AGE_DIFF: Age one feels (from a question,"How old do you feel?") minus Actual age; Positive values mean feeling older than actual age; Negative values mean feeling younger than actual age.

3.1 Fit a linear model that regresses AGE_DIFF on AGE, MALE, BLACK, EDUC and LIFESAT. Interpret the results from the estimated model coefficients (including the intercept) in layman's terms. Treat each as if it were the exposure of interest. [10pt]

- β_0: When AGE and life satisfaction equals 0 among non-Black females with less than a high school education, the expected average perceived age one feels is 17.6 years older than their actual age.
  - Without centering age, the intercept age=0 is realistic as infants are 0 years old; however, the construct "age_diff" does not make sense for this sub population of infants.
- β_1: For every year increase in age, the expected average perceived age decreases by .3 years from one's actual age, while holding other predictors constant.
- β_2: .9 years is the expected mean difference in perceived age difference from actual age among males and females, while holding other predictors constant, suggesting that males on average perceive their age almost one year older than they actually are compared to females.
- β_3: 6.3 years is the expected mean difference in perceived age difference from actual age among black and other than black individuals, while holding other predictors constant, suggesting that on average blacks perceive their age to be 6.3 years below their actual age compared to non-blacks.
- β_4: 1.4 years is the expected mean difference in perceived age difference from actual age among high school educated and less than high school educated individuals, while holding other predictors constant, suggesting that those with a high school education perceive themselves to be almost one and half years older than their actual age, compared to those with less than a high school education.
- β_5: 1.2 years is the expected mean difference in perceived age difference from actual age among individuals with beyond a high school education and less than high school, while holding other predictors constant, suggesting that those with beyond a high school education perceive themselves to be almost one and a quarter years younger than their actual age, compared to those with less than a high school education.
- β_6: For every year increase in life satisfaction scores, the expected average perceived age decreases by 1 year from one's actual age, while holding other predictors constant.

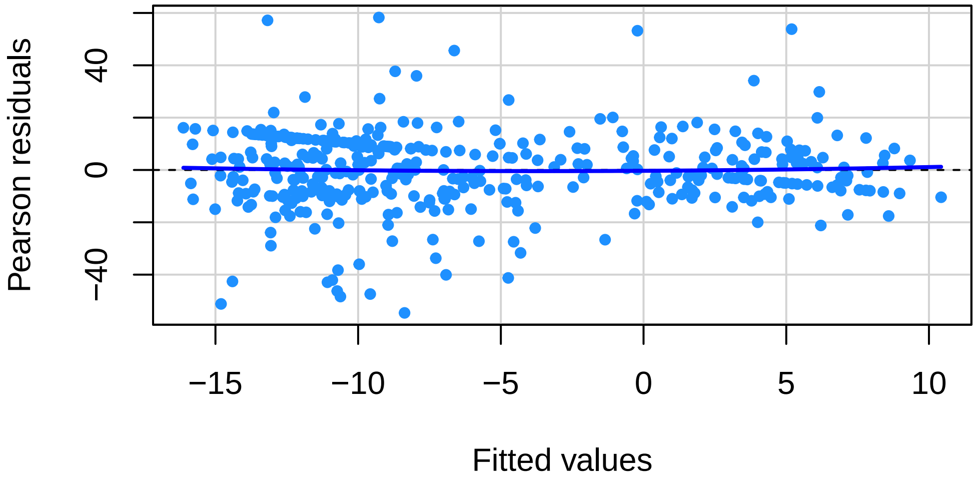
```
summm(mod_age_dif <- lm(AGE_DIFF ~ AGE + MALE + BLACK + EDUC + LIFESAT, path))
```

```
MODEL INFO:
Observations: 408
Dependent Variable: AGE_DIFF
Type: OLS linear regression

MODEL FIT:
F(6,401) = 15.09, p = 0.00
R² = 0.18
Adj. R² = 0.17

Standard errors:OLS
------------------------------------------------
                     Est.   S.E.   t val.      p
----------------- ------- ------ -------- ------
(Intercept)         17.55   3.75     4.68   0.00
AGE                 -0.31   0.08    -3.92   0.00
MALE                 0.89   1.61     0.55   0.58
BLACK               -6.29   2.14    -2.94   0.00
EDUC2.HS             1.39   1.77     0.78   0.43
EDUC3.>HS           -1.16   1.92    -0.61   0.55
LIFESAT             -1.01   0.48    -2.11   0.04
------------------------------------------------
```

3.2 Plot residuals against fitted values. Comment on the zero mean error assumption and the constant error variance assumption. [5pt]

- We plot $\hat{\epsilon}$ against $\hat{y}$ and observe that in general they are distributed around 0. This plot overlays a purple line that we expect to be flat and horizontal, and while it is we also observe a cone shape pattern suggesting non-constant error variance. Lower fitted values under -5 appear to have lots of variation but center around 0, while fitted values over 0 have a few values above 0. We can speculate that some values do not always center around 0, as well as the variation in values varies depending on the fitted values.
  - Zero mean error $(E(\epsilon_i) = 0 for i = 1, ..., n)$ assumption means the average error value should be 0 and determined my random chance. Non-zero mean error can be a result of error dependence, non-normality, or non-constant error variance.
  - Constant error variance ($V(\epsilon_i) = \sigma^2$, aka homoscedasticity) assumption refers to the residual variance being constant meaning that the error term does not vary much as the predictor variance changes. Generally, we expect the variance of the data points to be about the same for all data points. If we do not have constant variance, meaning the residuals are heteroscedastic, the error terms may increase with the value of the response. Graphically this shows up as a funnel shape or some pattern in the residual plot. Ignoring this assumption leads to biased standard errors, confidence intervals, predictions, and hypothesis testing.

```
residualPlots(mod_age_dif, pch=20, lwd=2, col="dodgerblue", terms = ~1, tests=F)
```



3.3. Report the mean of residuals for the entire data and for the fitted values >-5 vs. <=-5. What do you conclude about the zero mean error assumption? Make sure to incorporate your observations from #3.2 in your answer. [5pt]

- As in 3.2 we observe and mention that in general the values appear to be centered at zero, and using summary statistics we confirm this with a residual mean of 0. We also confirm that for different sets of values across the fitted value they are not always centered at zero, fitted values greater than -5 have a mean of -.02 while values less than or equal to -5 have a mean of .01.

```r
path_augmented <- augment(mod_age_dif) |>
  mutate(subset_fitted = ifelse(.fitted > -5, "greater -5",
                                ifelse(.fitted <= -5,
                                       "less or equal -5", NA)))

path_augmented |>
  summarise(resid_mean = round(mean(.resid), 3)) |> kable()
```

resid_mean
0

```r
path_augmented |>
  summarise(resid_mean =
              round(mean(.resid), 3), .by=subset_fitted) |> kable()
```

subset_fitted	resid_mean
less or equal -5	0.010
greater -5	-0.016

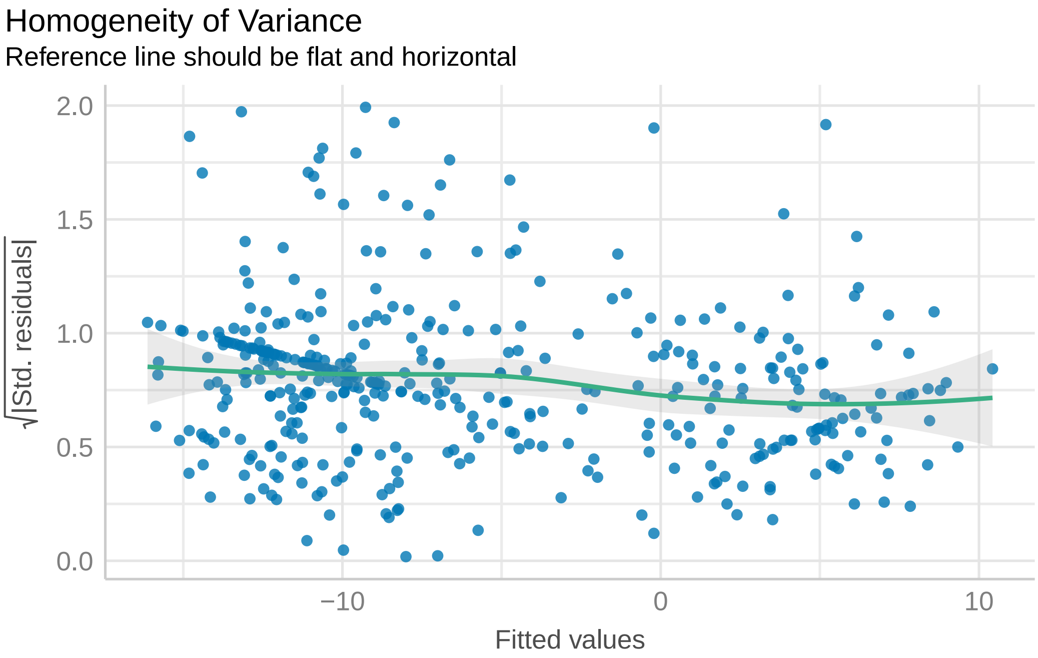
3.4. Report the variance of residuals for the entire data and for the fitted values >-5 vs. <=-5. Include formal testing. What do you conclude about the constant error variance assumption? Make sure to incorporate your observations from #3.2 in your answer. [10pt]

- We examine the constant variance assumption by plotting $\sqrt{\hat{\epsilon}}$ against $\hat{y}$ (Faraway, 2014, p. 75) and it seems that we have non-constant variance based on the plot below. Our reference line in green should be flat and horizontal, which has subtle movements across the fitted values range.

```
path_augmented |> summarise(resid_var = round(var(.resid), 3)) |>
  mutate(subset_fitted="ALL") |>
  relocate(subset_fitted, .before = "resid_var") |>
  add_row(path_augmented |> summarise(
  resid_var = round(var(.resid), 3), .by=subset_fitted)) |> kable()
```

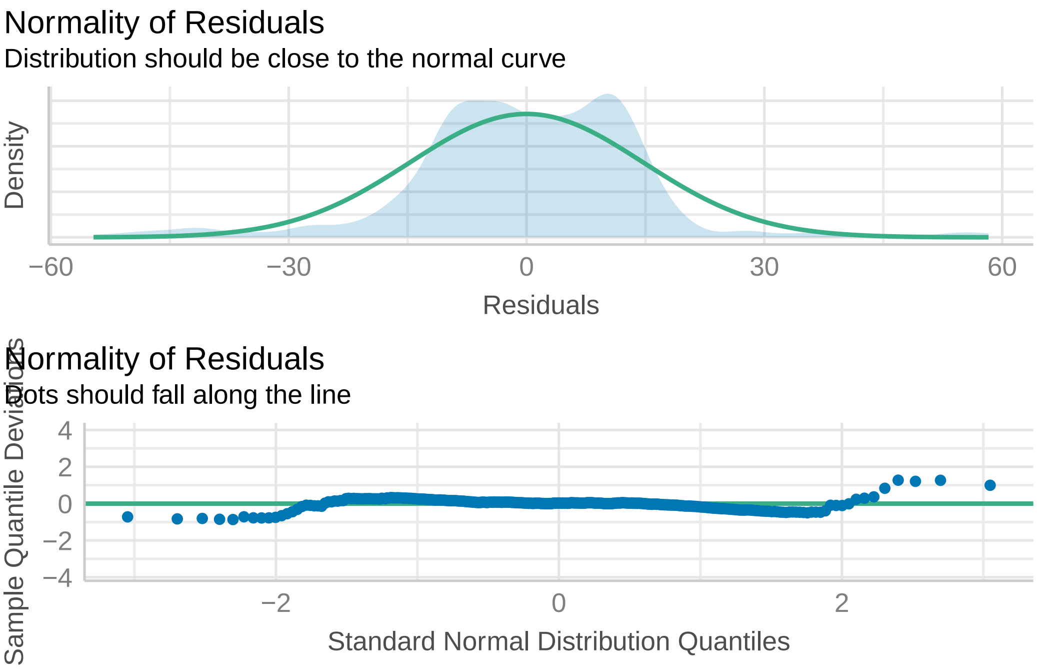
subset_fitted	resid_var
ALL	216.951
less or equal -5	251.361
greater -5	160.876

```
plot(check_heteroskedasticity(mod_age_dif))
```



- Graphically we would like to get a better idea of the shape of the distribution of the residuals through a density and QQplot, which we can see below a bi modal distribution and non-normality in the residual variances. These graphics also helps us examine the error normality assumptions, and provides us with evidence that we should continue examining the constant error assumption statistically.

```r
grid.arrange(
plot(check_normality(mod_age_dif), type="density"),
plot(check_normality(mod_age_dif), type="qq")
)
```



- We perform a t test and fail to reject the null hypothesis that the mean of our $\hat{\epsilon}$ is equal to 0, our confidence interval contains 0, suggesting that our residual mean is statistically close to 0. We conclude that our residuals have a 0 mean; however, we have to consider that in 3.2 we observed a funnel shape in the residuals and fitted values plot which can be a result of constant error variance being violated, or heteroscedasticity as described in 3.2.

```r
t.test(resid(mod_age_dif))
```

```

    One Sample t-test

data:  resid(mod_age_dif)
t = 0.00000000000000036731, df = 407, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1.433481  1.433481
sample estimates:
               mean of x
0.000000000000000002678447
```

- Numerically we take note of a weak correlation between the residuals and predictors, but when we model these values we see that for every one unit increase in fitted values, residuals decrease by about .03, suggesting that there is statistical evidence for non-constant variance, or heteroscedasticity.

```r
cor(fitted(mod_age_dif), resid(mod_age_dif))
```

```
[1] -0.0000000000000000002462352
```

```r
summ(lm(sqrt(abs(residuals(mod_age_dif))) ~ fitted(mod_age_dif)))
```

```
MODEL INFO:
Observations: 408
Dependent Variable: sqrt(abs(residuals(mod_age_dif)))
Type: OLS linear regression

MODEL FIT:
F(1,406) = 10.71, p = 0.00
R² = 0.03
Adj. R² = 0.02

Standard errors:OLS
--------------------------------------------------------
                             Est.   S.E.   t val.      p
------------------------- -------- ------ --------- ------
(Intercept)                  2.83   0.08    33.31   0.00
fitted(mod_age_dif)         -0.03   0.01    -3.27   0.00
--------------------------------------------------------
```

- We finally examine statistically equal variances for fitted values greater than -5 and values less than or equal to -5. The F-test for equal variances we see that the variances in the residuals are not equal based on that the confidence intervals (95% [.55, .95]) does not contain 0, meaning that we reject the null that true ratio of variances = 1 in favor of the alternative hypothesis = > unequal variances, according to our F-test. As we suggested in 3.2, this is evidence for non-constant variance, or heteroscedasticity, and we find that the error term varies as the predictor variance changes, in other words we do not observe variance of the data points to be about the same across all data points.

```r
var.test(resid(mod_age_dif)[fitted(mod_age_dif)>-5],
+ resid(mod_age_dif)[fitted(mod_age_dif)<=5])
```

```
    F test to compare two variances

data:  resid(mod_age_dif)[fitted(mod_age_dif) > -5] and +resid(mod_age_dif)[fitted(mod_age_di
F = 0.71969, num df = 152, denom df = 364, p-value = 0.0198
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5543752 0.9484241
sample estimates:
ratio of variances
         0.7196946
```

3.5. Identify top 3 influential observations through Cook's distance. What do you conclude about these observations? What makes them stand out? [5pt]

- Using cooks distance we observe that the top three observations that stand out as influential are 25, 332, 31, which we also see these points in the halfnorm plot.

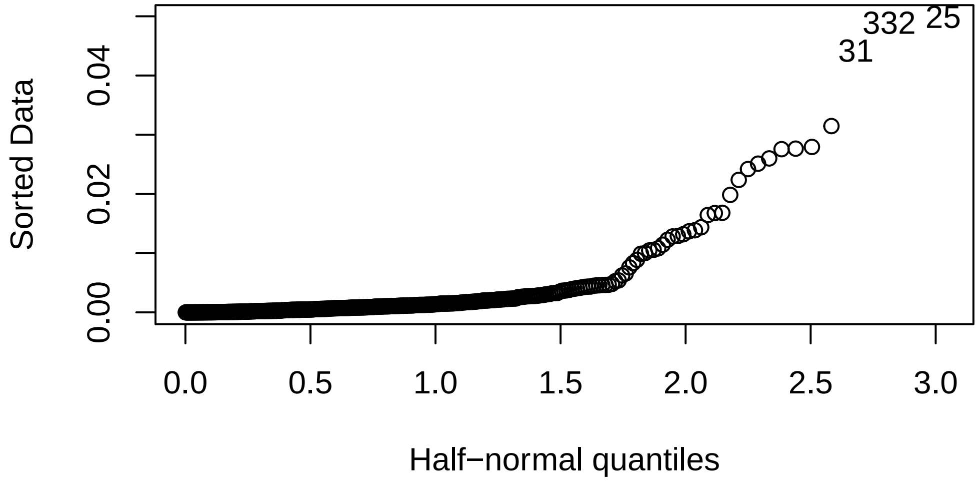
```
cookd <- cooks.distance(mod_age_dif)
summary(cookd)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
0.0000000 0.0002418 0.0008135 0.0024763 0.0017801 0.0498857 
```

```
sort(cookd, decreasing=TRUE) |> head()
```

```
        25        332         31         59        299        360 
0.04988574 0.04889600 0.04418868 0.03146099 0.02793497 0.02764950 
```

```
halfnorm(cookd, nlab = 3)
```



3.6. Fit the model in #3.1 without the top 3 influential observations identified in #3.5. Based on one of the goodness of fit metrics we've discussed, state whether you would remove these observations. [10pt]

- We fit a linear model (same as 3.1) to the data without the 3 influential points called out in 3.5. In the full model, the adjusted r-square was .17 and after dropping the influential points we observe the adjusted r-squared to increase to .19, or 19% of the variation in age_diff can be explained by the predictors in the model. Given the increase in adjusted r-square simply by removing three influential points out of a sample of 405, we would go ahead and keep these points out of the final model.

```r
path_no_influential <- path |>
  rename(id=1) |>
  filter(!id %in% c(25, 332, 31))

summ(mod_age_dif_no_influential <-
       lm(AGE_DIFF ~ AGE + MALE + BLACK + EDUC + LIFESAT, path_no_influential))
```

```
MODEL INFO:
Observations: 405
Dependent Variable: AGE_DIFF
Type: OLS linear regression

MODEL FIT:
F(6,398) = 16.96, p = 0.00
R² = 0.20
Adj. R² = 0.19

Standard errors:OLS
------------------------------------------------
                     Est.   S.E.   t val.      p
----------------- -------- ------ --------- ------
(Intercept)         17.18   3.56     4.83   0.00
AGE                 -0.31   0.08    -4.08   0.00
MALE                 2.27   1.54     1.48   0.14
BLACK               -6.76   2.03    -3.32   0.00
EDUC2.HS             0.23   1.68     0.14   0.89
EDUC3.>HS           -1.79   1.83    -0.98   0.33
LIFESAT             -1.09   0.46    -2.38   0.02
------------------------------------------------
```

3.7. Based on the model in #3.1, predict AGE_DIFF for the mean 45 year old black female with high school education and mean life satisfaction and construct its appropriate 95% interval. What does this interval mean in layman's terms? [15pt]

- For an average 45 year old Black women with a high school education and average life satisfaction scores we provide a 95% confidence interval and predict their average perceived age difference to be almost 5.5 (between 1.4 and 9.6) years lower than their actual age. If we took 100 repeated samples from the same population of the same size, we would expect the 95% CIs of the samples to contain the true mean 95% of the time.

```
age_diff_pred <- tibble(
  AGE=45,
  BLACK=1,
  MALE = 0,
  EDUC="2.HS",
  LIFESAT = mean(path$LIFESAT)
        )

age_diff_pred |> kable()
```

AGE	BLACK	MALE	EDUC	LIFESAT
45	1	0	2.HS	4.088725

```
predict(mod_age_dif,
newdata = age_diff_pred, interval = "confidence") |>
kable()
```

fit	lwr	upr
-5.474057	-9.549875	-1.39824

3.8. Now predict AGE_DIFF for a 45 year old black female with high school education and mean life satisfaction and construct its appropriate 95% interval. How does this interval compare to the one from 3.7? [5pt]

- Our prediction interval is different from the confidence interval reported in 3.7., given that a prediction interval is a range of values that is likely to contain a single new observation based on a specified subset of predictor values. We observe the same prediction value of 5.5 but take note that the our confidence interval is much more narrower than the prediction interval of -34.9 to 24.0 which also happens to contain zero. The difference is that with confidence intervals we assess variability around the mean in an estimated quantity on the predictor, while the prediction interval looks at a single new observation, not the average, so we have to account for the variability of the predictor in addition to the variability of our estimate of the mean. In this case we have a lot of uncertainty on our prediction score for a 45 year old Black women with a high school education and average life satisfaction scores.

```r
predict(mod_age_dif,
newdata = age_diff_pred, interval = "prediction") |>
kable()
```

fit	lwr	upr
-5.474057	-34.92943	23.98131