# BIOSTAT 650 SEC 002 HW7

Jaehoon Kim

2024-11-04

## 1.

(a)

$$T = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

The contrast matrix is rank 3 thus the df associated with the test statistic is 3 and n-5.

(b)

$$T = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

The contrast matrix is rank 2, thus the df associated with the test statistic is 2 and n-5.

(c) (a) Tries to test if the mean effect (slope) of all covariates are the same or meaningfully different whilst (b) tries to test whether the mean effects of $X_1$ and $X_2$, $X_3$ and $X_4$ are the same or meaningfully different respectively.

## 2.

$$Depression_i = \beta_0 + \beta_1 Age1_i + \beta_2 Age2_i + \beta_3 Age3_i + \beta_4 Fatalism + \beta_5 Sex + \beta_6 RE + \epsilon_i$$

```r
mySASData = read.sas7bdat("completedata.sas7bdat")

var = c(
"Depression",
"Age",
"Sex",
"R_E",
"Fatalism"
)

df <- mySASData[ ,var]
typeof(df$Depression)
```

```
## [1] "double"
```

```r
#(a)
df <- df |>
  mutate(Age_Cat = cut(Age,breaks = c(-Inf, quantile(Age,c(0.25, 0.5, 0.75)), Inf), labels = c(0, 1, 2,
df<- df |>
  mutate(
    Fatalism_c = (df$Fatalism-mean(df$Fatalism))/IQR(df$Fatalism)
  )
```

```r
#A0 = as.numeric(df$Age_Cat==0)
A1 = as.numeric(df$Age_Cat==1)
A2 = as.numeric(df$Age_Cat==2)
A3 = as.numeric(df$Age_Cat==3)
model = lm(Depression~Age_Cat+Fatalism_c+Sex+R_E, df)
summary(model)
```

```
##
## Call:
## lm(formula = Depression ~ Age_Cat + Fatalism_c + Sex + R_E, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.905 -3.368 -1.278  2.076 20.837
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.6002     0.5341  12.358  < 2e-16 ***
## Age_Cat1     -1.6394     0.5650  -2.901  0.00385 **
## Age_Cat2     -0.7899     0.5709  -1.384  0.16698
## Age_Cat3     -3.5238     0.6051  -5.824 9.35e-09 ***
## Fatalism_c    2.2117     0.3193   6.927 1.10e-11 ***
## Sex           0.5206     0.4075   1.278  0.20189
## R_E           0.2789     0.4250   0.656  0.51193
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.987 on 605 degrees of freedom
## Multiple R-squared:  0.1236, Adjusted R-squared:  0.1149
## F-statistic: 14.22 on 6 and 605 DF,  p-value: 3.433e-15
```

```r
#No intercept?
```

(a) $H_0 : \beta_1 = 0, \beta_2 = \beta 3$ $H_1 : \beta_1 \neq 0, \beta_2 \neq \beta_3$

$$T = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 \end{bmatrix}$$

$F_2, 605 = 14.893$ above critical value at 3.07 at significance level $\alpha = 0.05$ Reject the null hypothesis that mean depression score among all below the median age is the same, and the mean depression score among all those above the median age is the same.

```r
Contrast.T = matrix(c(0,1,0,0,0,0,0,0,0,1,-1,0,0,0), byrow=T, nrow=2)
car::linearHypothesis(model=model,hypothesis.matrix=Contrast.T)
```

```
##
## Linear hypothesis test:
## Age_Cat1 = 0
## Age_Cat2 - Age_Cat3 = 0
##
## Model 1: restricted model
## Model 2: Depression ~ Age_Cat + Fatalism_c + Sex + R_E
##
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    607 15786
## 2    605 15046  2    740.74 14.893 4.856e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b) $H_0 : \beta_1 = \beta_2 = \beta 3$ $H_1 : \beta_1 \neq \beta_2$ or $\beta_2 \neq \beta_3$

$$T = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 \end{bmatrix}$$

$F_2, 605 = 11.131$ above critical value at 3.07 at significance level $\alpha = 0.05$ Reject the null hypothesis that mean depression scores among those older than the first quartile are the same.

```
Contrast.T = matrix(c(0,1,-1,0,0,0,0,0,0,1,-1,0,0,0), byrow=T, nrow=2)
car::linearHypothesis(model=model,hypothesis.matrix=Contrast.T, ths=c(0,0))
```

```
##
## Linear hypothesis test:
## Age_Cat1 - Age_Cat2 = 0
## Age_Cat2 - Age_Cat3 = 0
##
## Model 1: restricted model
## Model 2: Depression ~ Age_Cat + Fatalism_c + Sex + R_E
##
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    607 15600
## 2    605 15046  2    553.66 11.131 1.789e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## 3.

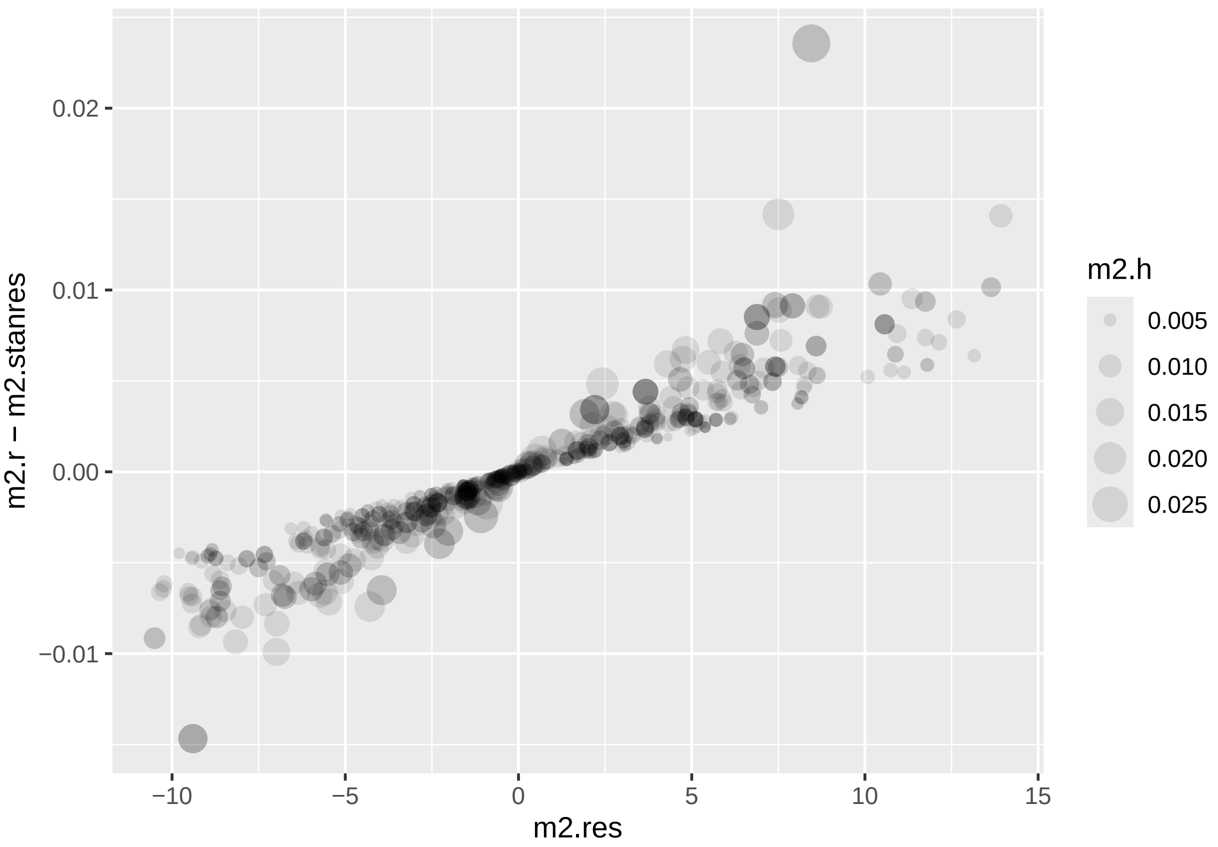
(a)

```
var = c(
"Fatalism",
"Age",
"Sex",
"R_E",
"Comorbidity1"
)

df2 <- mySASData[ ,var]
m2 = lm(Fatalism~Age+Sex+R_E+Comorbidity1, df2)
m2.yhat = m2$fitted.values
m2.res = m2$residuals #regular residuals
m2.stanres = m2$residuals/summary(m2)$sigma #standardized residuals
m2.h = hatvalues(m2) #leverage
m2.r = rstandard(m2) #internally studentized residuals
m2.rr = rstudent(m2) #externally studentized residuals
```

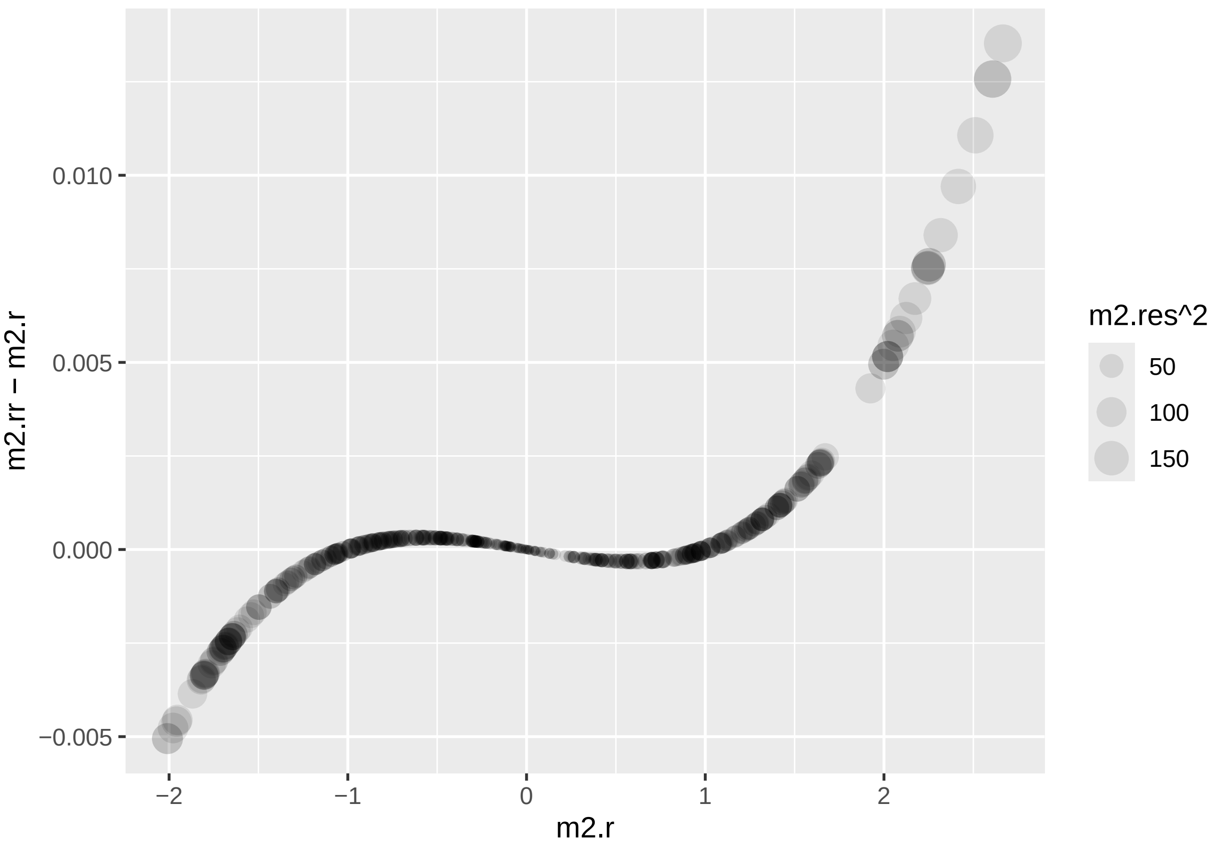
**(b)** Residuals for which the leverage values are high are farthest from 0. This is due to the fact $h_i i$ directly affects the value of studentized residuals.

```
ggplot(df2, aes(x=m2.res , y=m2.r-m2.stanres, size=m2.h))+
  geom_point(alpha=0.1)
```



Residuals for which the $\hat{\epsilon}_i^2$ are high are farthest from 0. This is likely due to the fact that externally studentized residual inflates $\hat{\sigma}^2$ for large $\hat{\epsilon}_i^2$.

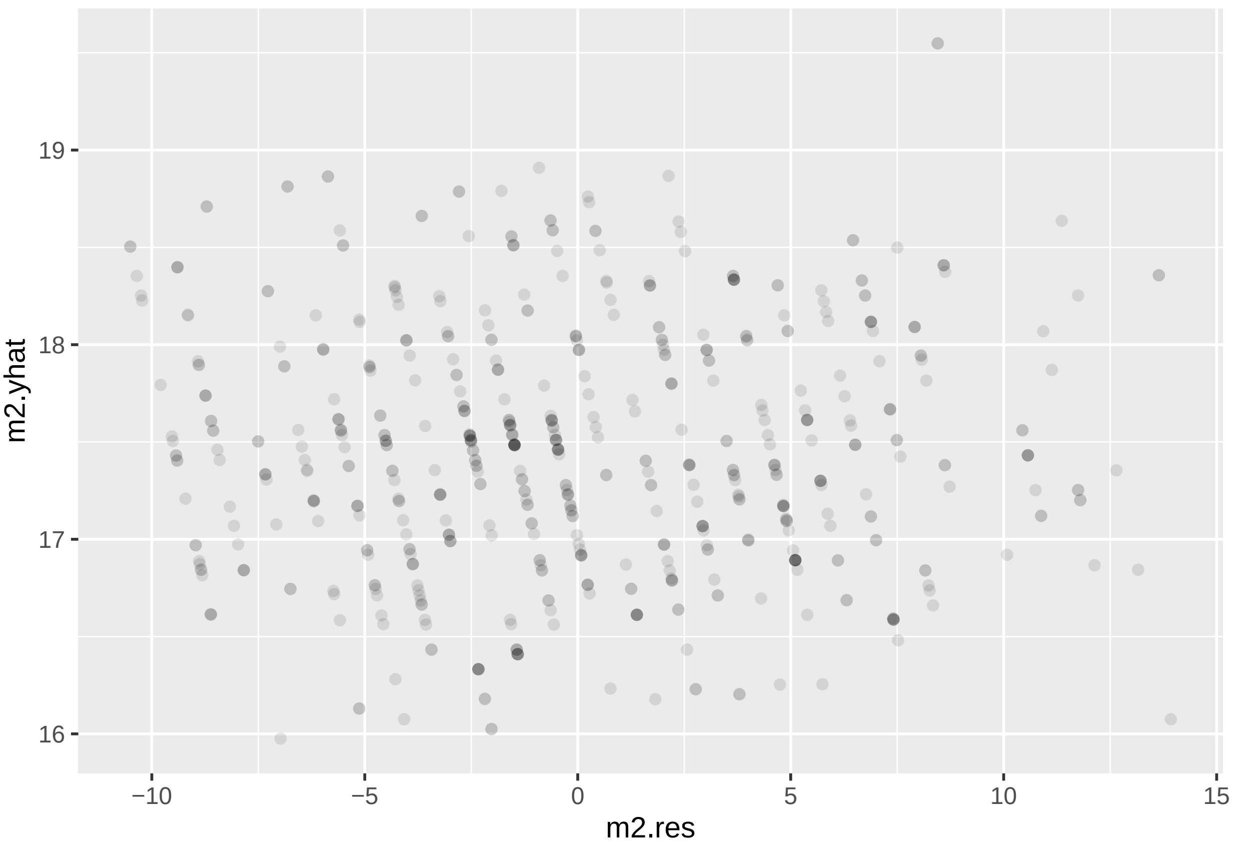
```
ggplot(df2, aes(x=m2.r , y=m2.rr-m2.r, size=m2.res^2))+
  geom_point(alpha=0.1)
```



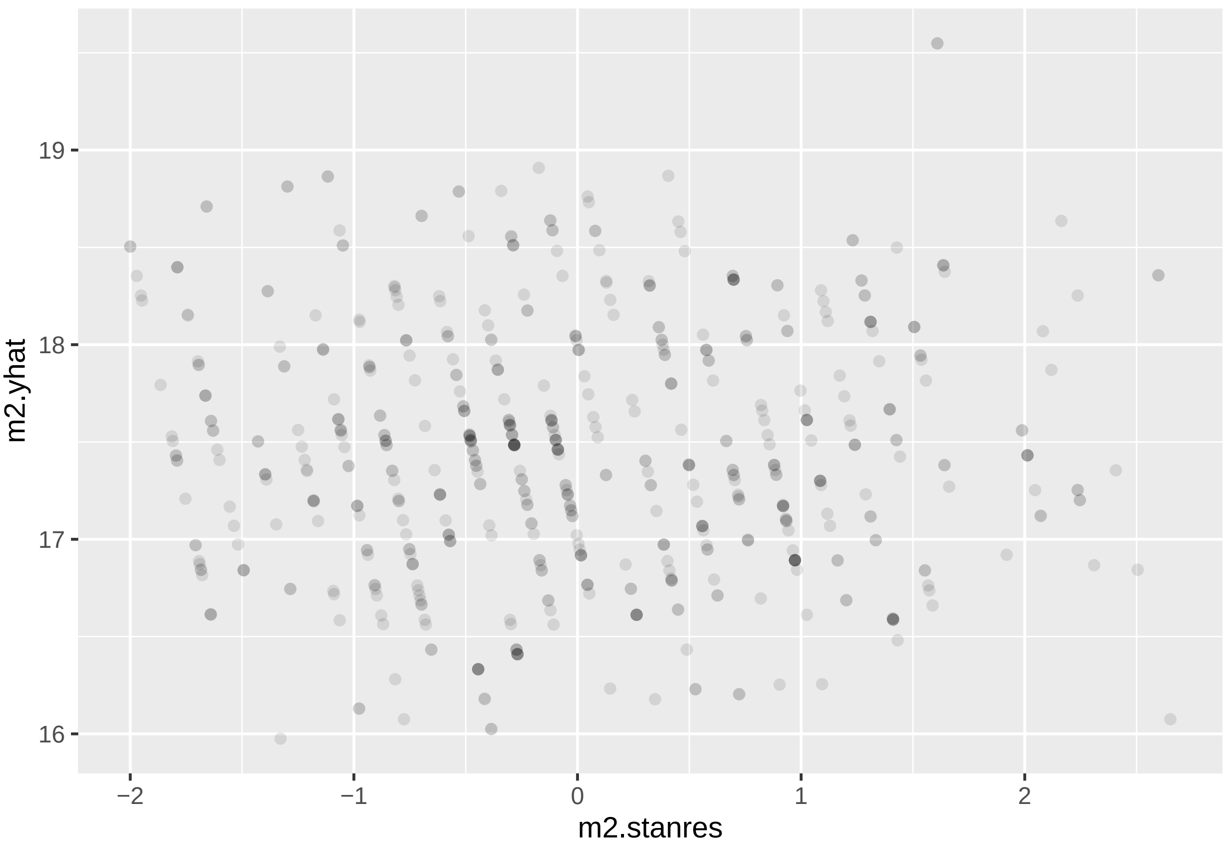
The difference between internally and externally studentized residuals are more smaller as they are by definition quite similar, but across the all residuals, the magnitude of differences are pretty much negligible in absolute terms.

(c) The constant variance assumption holds across all types of residuals as there are no positive relationships observed.

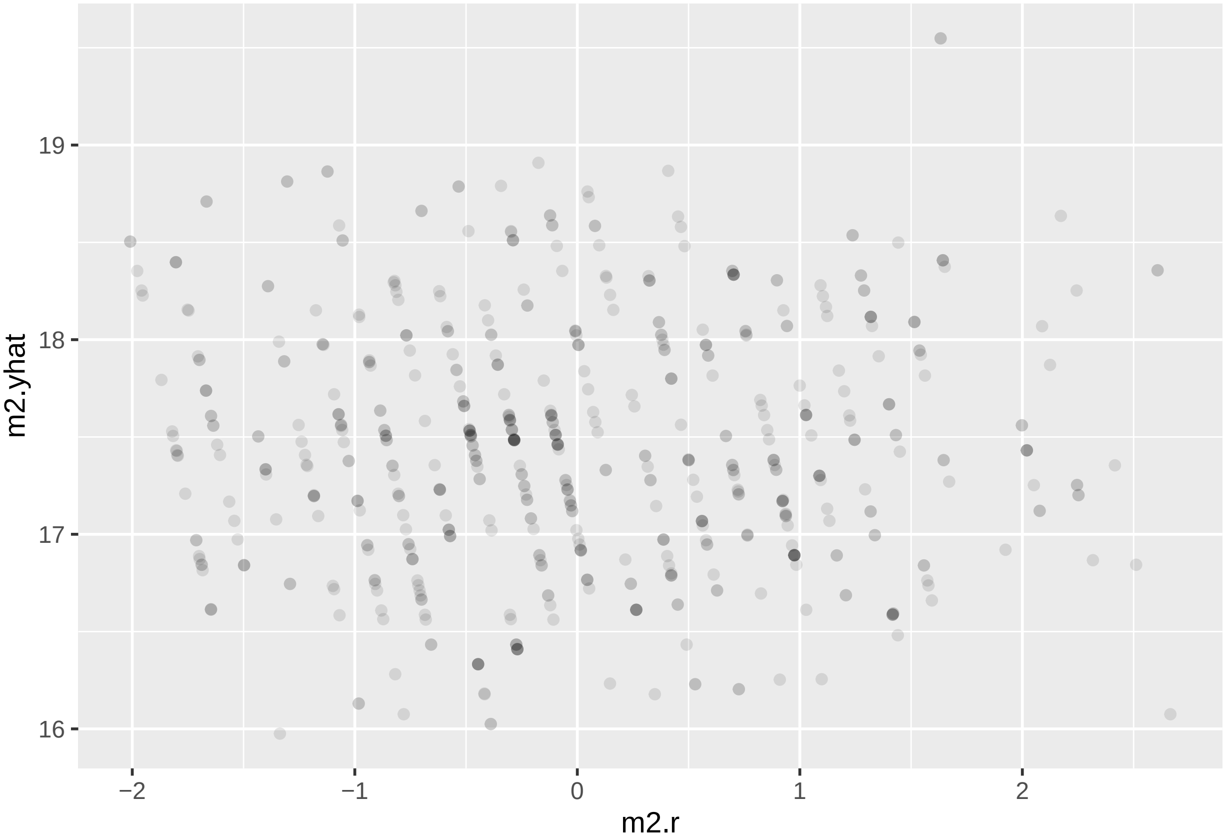
```
ggplot(df2, aes(x=m2.res , y=m2.yhat))+
  geom_point(alpha=0.1)
```



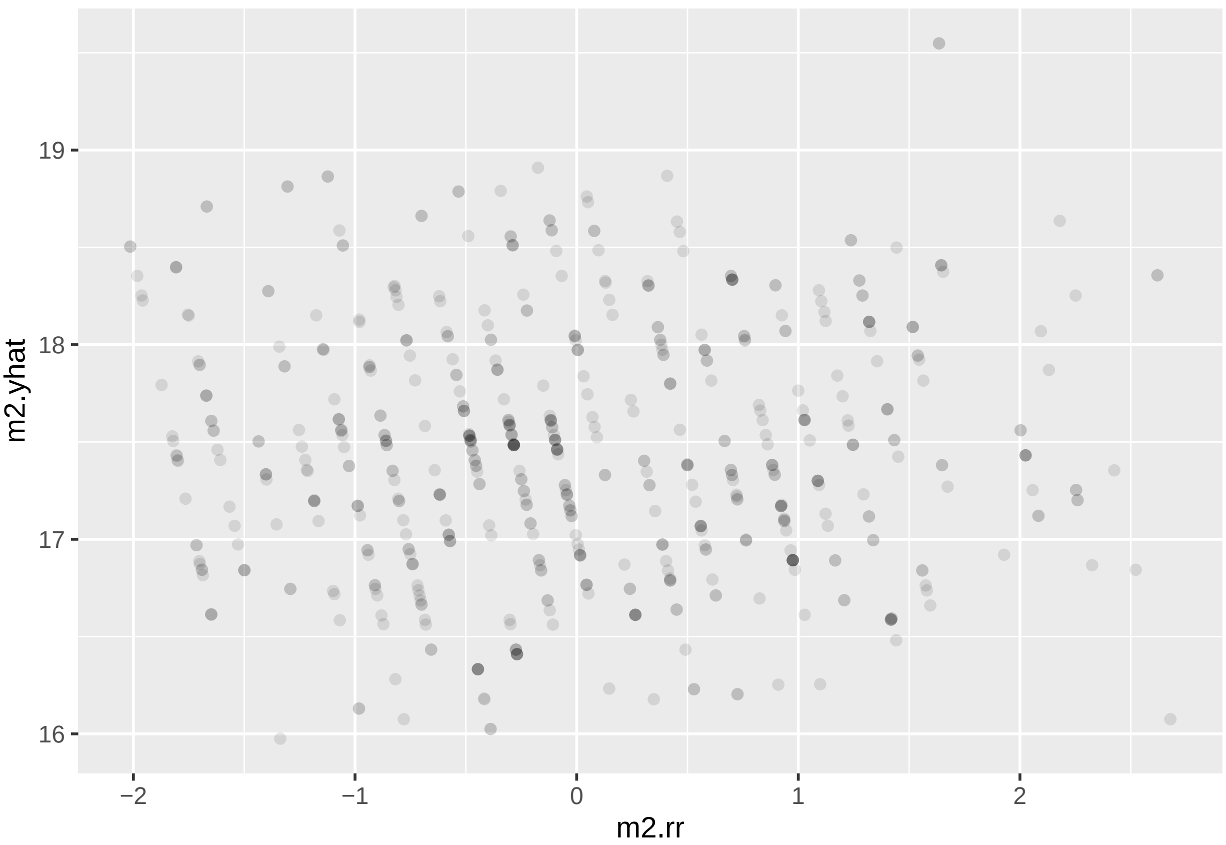
```r
ggplot(df2, aes(x=m2.stanres , y=m2.yhat))+
  geom_point(alpha=0.1)
```



```
ggplot(df2, aes(x=m2.r , y=m2.yhat))+
  geom_point(alpha=0.1)
```



```
ggplot(df2, aes(x=m2.rr , y=m2.yhat))+
  geom_point(alpha=0.1)
```



(d) There isn't a really noticeable difference which residual you use. The differences will start to occur with small to moderately sized samples as the sensitivy to outliers start to get pronounced for the studentized residuals which might more readily detect by displaying points that deviate from the horizontal line.

## 4.

(a)

$$\mathbb{E}[Y_i|X_i = a] = \mathbb{E}[\beta_0 + \beta_1 \log_{10} a + \epsilon_i] = \beta_0 + \beta_1 \log_{10} a$$

(b)

$$\mathbb{E}[Y_i|X_i = 10a] = \mathbb{E}[\beta_0 + \beta_1(\log_{10} 10 + \log_{10} a) + \epsilon_i] = \beta_0 + \beta_1 + \beta_1 \log_{10} a$$

(c)

$$\mathbb{E}[Y_i|X_i = 10a] - \mathbb{E}[Y_i|X_i = a] = \beta_1$$

(d)

Agree,as demonstrated in (c)

(e) Agree

(f)

$$\mathbb{E}(Y|X_j = 1.05a) - \mathbb{E}(Y|X_j = a) = \beta_0 + \beta_1 \ln 1.05a - \beta_0 - \beta_1 \ln a = \beta_1 \ln 1.05$$

(g) A 5% higher $X$ is associated with a difference of $\beta_1 \ln 1.05$ units in $Y$, 95%CI. Formula for lower bound: $\beta_1 \ln 1.05 - 1.96 * \ln 1.05 SE(\beta_1)$. Formula for upper bound: $\beta_1 \ln 1.05 + 1.96 * \ln 1.05 SE(\beta_1)$