Homework 4

Ethan Ashby

3/19/2020

Question 1: We know that $SSTO = SSE + SSR$ (regardless of dimension). What we know we have is an F test for the null of H_0: $\beta_i = 0$ for all $i \geq 1$ with $SSR/p-1$ in the numerator. Similar to the question on homework 3 about SSLF, argue $SSR/p-1$ is estimating σ^2 if the null is true. Show that $SSR = \sum(\hat{y}_i - \bar{y})^2$. Do a similar composition by adding 0 creatively to SSTO (see HW3 solutions), and use HW1PR4 to show the necessary sum is 0.

(i) Show that $\frac{SSR}{p-1} = \sigma^2$ under the null hypothesis of the F test (H_0: $\beta_i = 0$ for all $i \geq 1$). Under the null, we know $\frac{\sum(y_i-\bar{y})^2}{n-1} = \sigma^2$ and $\frac{\sum(y_i-\hat{y})^2}{n-p} = \sigma^2$. We also know that $SSTO = \sum(y_i - \bar{y})^2$ and $SSE = \sum(y_i - \hat{y})^2$. Thus, $SSTO = \sigma^2 * (n-1)$ and $SSE = \sigma^2 * (n-p)$.

Thus, since $SSR = SSTO - SSE$, $SSR = \sigma^2 * (n-1) - \sigma^2 * (n-p) = (p-1)\sigma^2$. Q.E.D.

(ii) Show that $SSR = \sum(\hat{y}_i - \bar{y})^2$.

We know $SSR = SSTO - SSE$. Thus,

$$\begin{aligned}SSR = &\sum(y_i - \bar{y})^2 - (y_i - \hat{y}_i)^2\\ =&\sum(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 - (y_i - \hat{y})^2\\ =&\sum(y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 - (y_i - \hat{y})^2 \ *\\ =&\sum(\hat{y}_i - \bar{y})^2\end{aligned}$$

We've shown what we've intended to show, although our claim in (*) may appear specious at first. Let's double click on this step and see that what we've done is actually permissable.

In order for our claim in (*) to hold, $\sum 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$. This expression is equivalent to $2 * \sum e(\hat{y}_i - \bar{y}) = 0$ where e represents the residuals. From HW1PR4, if the sum is 0, then the correlation between our $\hat{y}_i$ and our residuals is 0. If we are doing our regression correctly, this should hold, because our fitted values should not correlate with our residuals (this is actually a diagnostic plot that we use to assess the quality of our model). Thus, assuming our linear model works, $\sum 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$ and our formula for SSR works out.

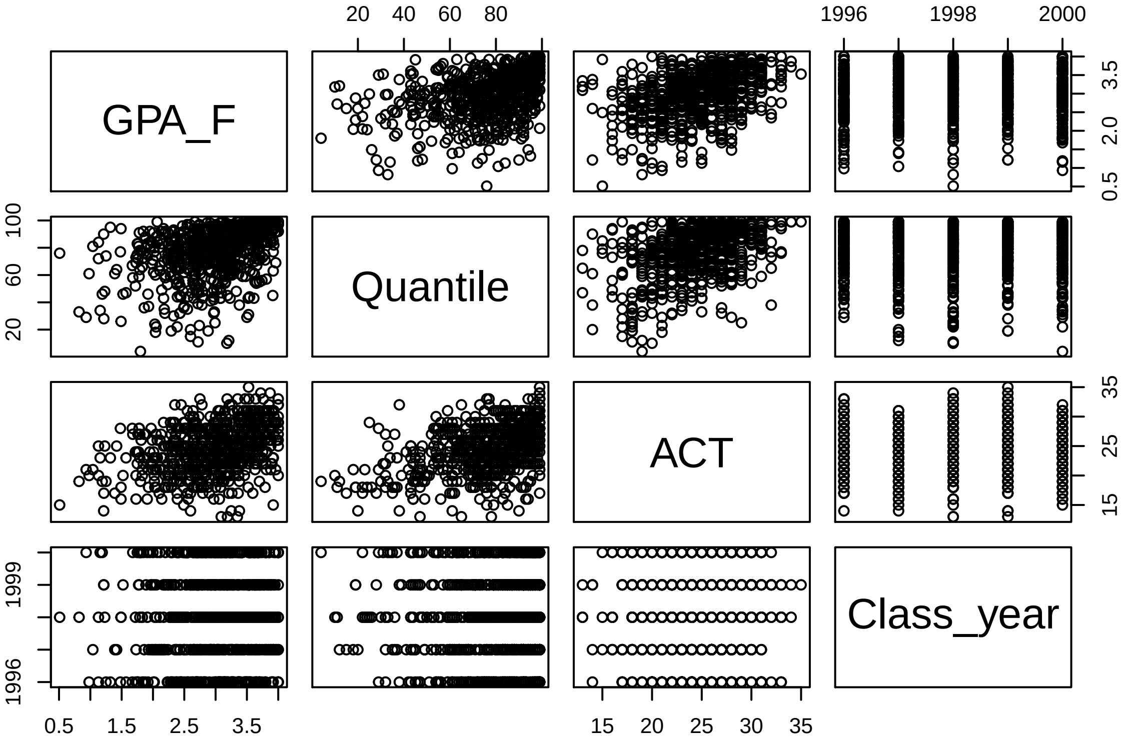
Question 2:

```
#load packages and data
library(tidyverse)
library(MASS)
library(lattice)
gpadata<-read.csv("HW4_Data", header=FALSE)
colnames(gpadata)<-c("ID", "GPA_F", "Quantile", "ACT", "Class_year")
gpadata<-gpadata %>% dplyr::select(-ID)

#create scatterplot matrix and fit full model
fit1<-lm(GPA_F~., data=gpadata)
summary(fit1)
```

```
##
## Call:
## lm(formula = GPA_F ~ ., data = gpadata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.10367 -0.31236  0.07643  0.40387  1.30266
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.248338  30.438148  -0.698    0.485
## Quantile      0.010046   0.001280   7.848 1.58e-14 ***
## ACT           0.037101   0.005943   6.243 7.42e-10 ***
## Class_year    0.011282   0.015235   0.741    0.459
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5673 on 701 degrees of freedom
## Multiple R-squared:  0.204,  Adjusted R-squared:  0.2006
## F-statistic: 59.87 on 3 and 701 DF,  p-value: < 2.2e-16
```

```r
plot(gpadata)
```



When we fit the full model, we see that ACT and quantile carry signal (have low p-values) but class year carries no appreciable signal. So we can exclude class year from all future models. This is supported by the scatterplot matrix, which seems to show that Quantile and ACT carry signal (albeit a lot of noise as well). There is no obvious transformation apparent as well, so we proceed with the most interpretable form of the explanatory variables (untransformed).

We fit our reduced model (quantile and ACT as explanatory variables) and then see if we can refine our model to improve the fit.

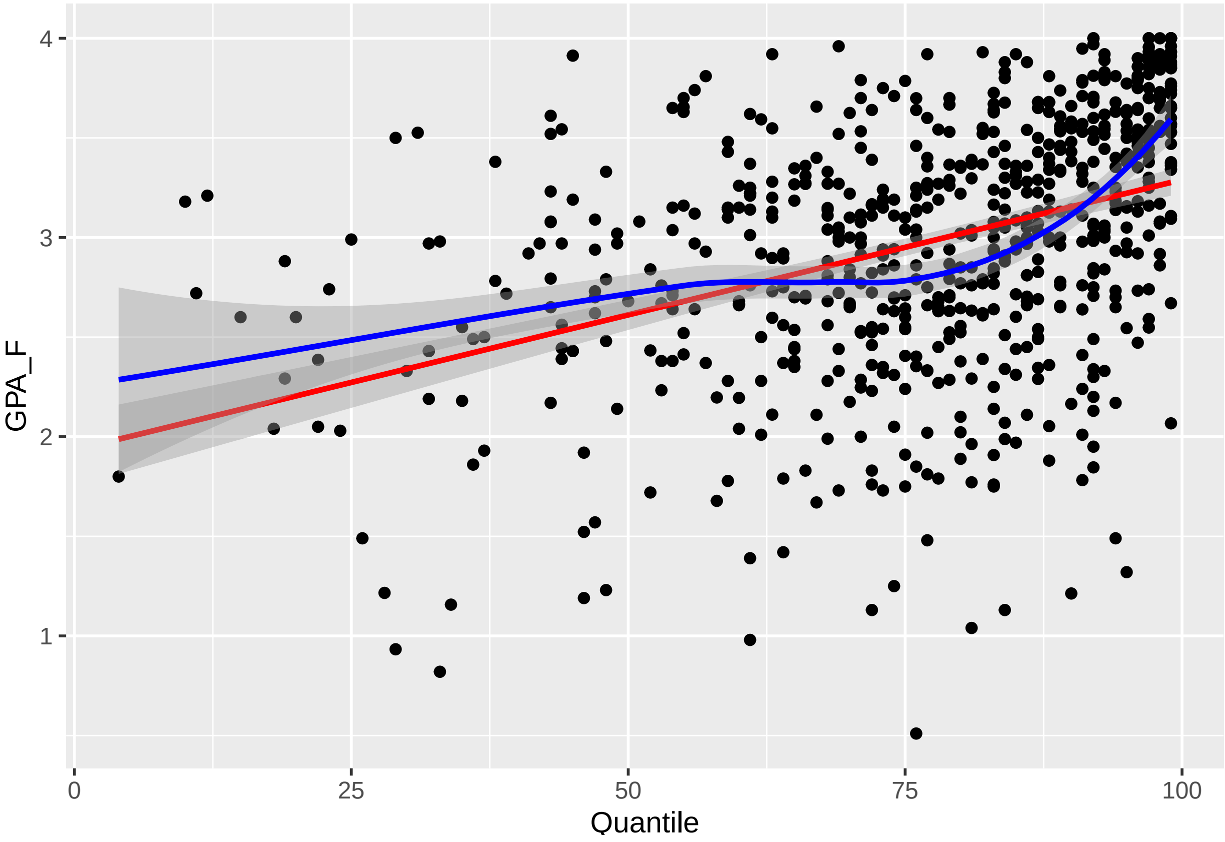
```r
#data looks pretty nosy but there's some signal in quantile and ACT, year carries no info. Fit this red
fit.r<-lm(GPA_F~ACT+Quantile, data=gpadata)
#the F-test confirms that year's coefficient was not significantly different from 0
anova(fit1, fit.r)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: GPA_F ~ Quantile + ACT + Class_year
## Model 2: GPA_F ~ ACT + Quantile
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    701 225.64
## 2    702 225.81 -1  -0.17653 0.5484 0.4592
```

```
#can we do better?

#When we plot quantile vs GPA, we see that the slope changes if you're in the >80 quantile
ggplot(gpadata)+geom_point(aes(x=Quantile, y=GPA_F), color="black")+geom_smooth(method="lm", aes(x=Quan
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

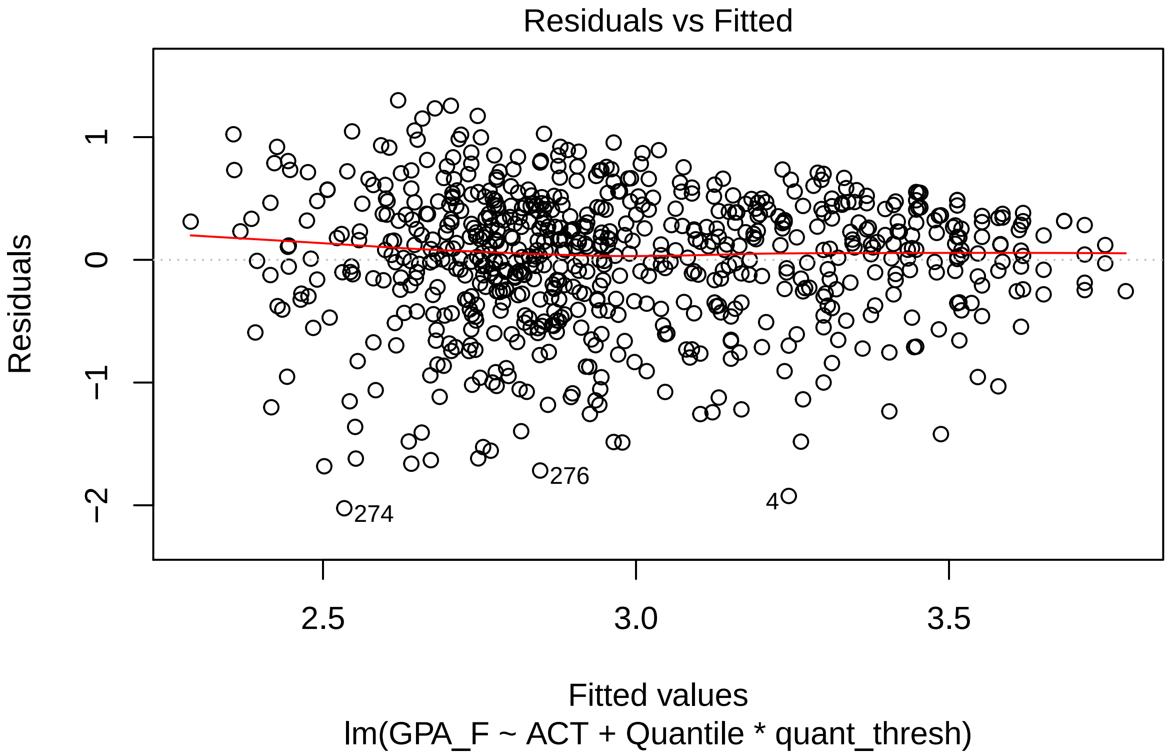


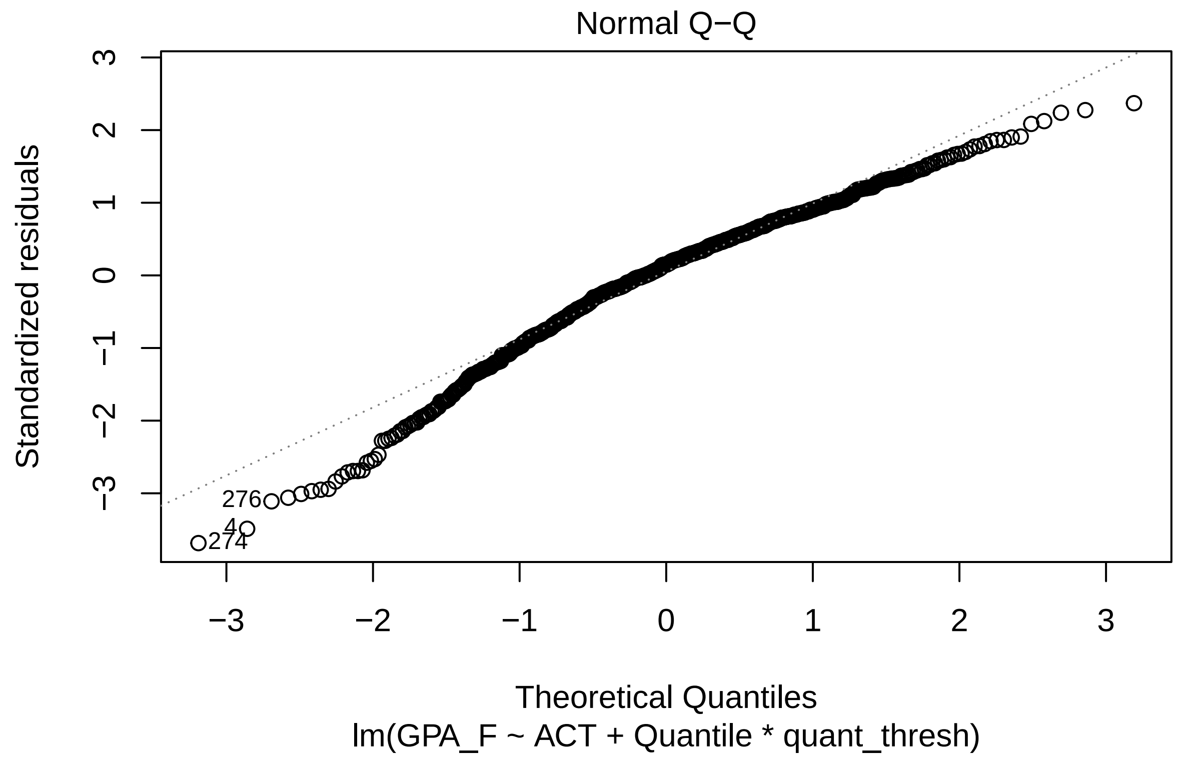
```
#thus, let's create two models depending on if we attain our quantile threshold
gpadata1<-gpadata %>% mutate(quant_thresh=as.numeric(Quantile>80))

fit.r<-lm(GPA_F~ACT+Quantile*quant_thresh, data=gpadata1)

summary(fit.r)
```

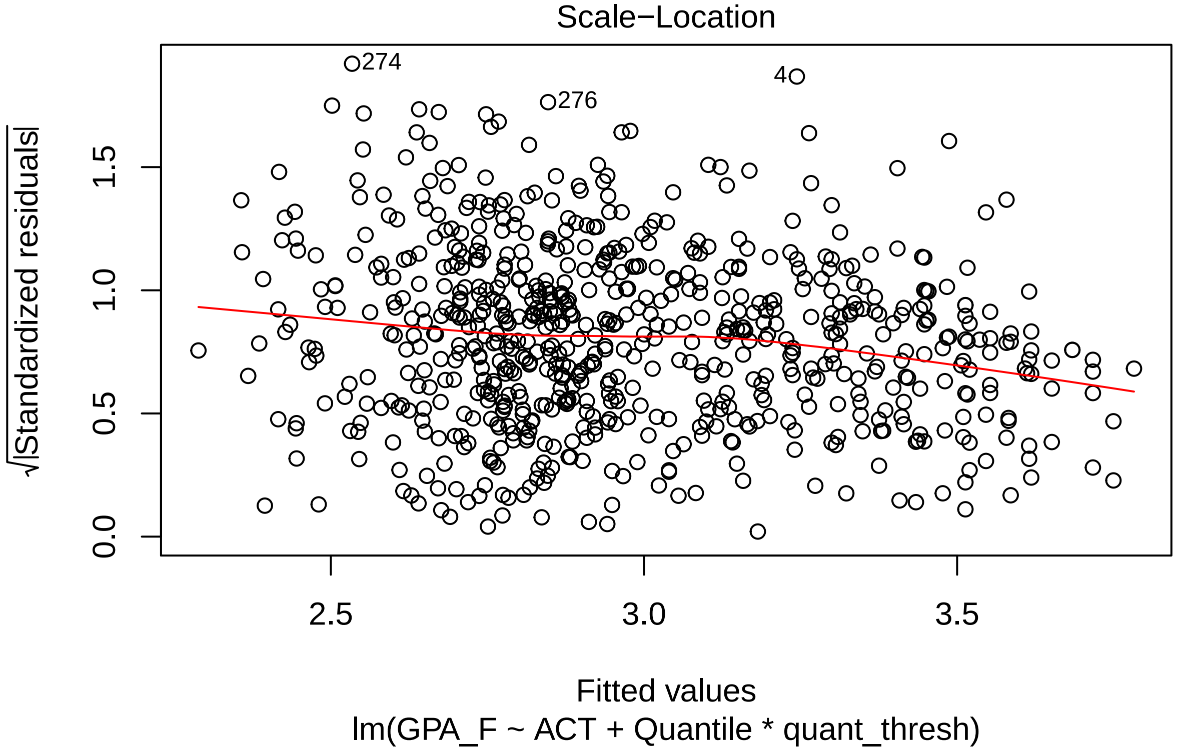
```
##
## Call:
## lm(formula = GPA_F ~ ACT + Quantile * quant_thresh, data = gpadata1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.02399 -0.31810  0.08215  0.37837  1.29997
##
## Coefficients:
```

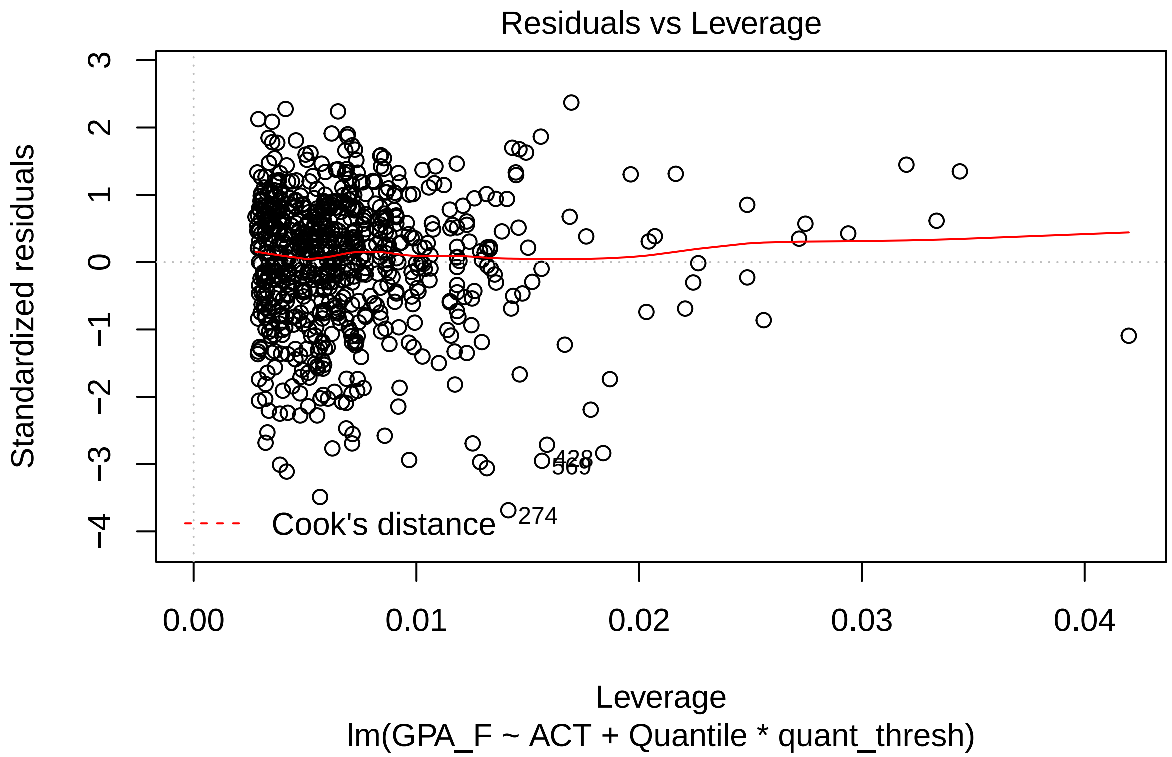
```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.753365   0.162113  10.816  < 2e-16 ***
## ACT                      0.032810   0.005835   5.623 2.71e-08 ***
## Quantile                 0.003796   0.001923   1.973   0.0488 *
## quant_thresh            -2.698518   0.492392  -5.480 5.93e-08 ***
## Quantile:quant_thresh    0.032358   0.005566   5.814 9.29e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5531 on 700 degrees of freedom
## Multiple R-squared:  0.2446, Adjusted R-squared:  0.2403
## F-statistic: 56.66 on 4 and 700 DF,  p-value: < 2.2e-16
```

```
plot(fit.r)
```




Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
276
4
274
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(GPA_F ~ ACT + Quantile * quant_thresh)


Scale−Location
$\sqrt{|\text{Standardized residuals}|}$
0.0
0.5
1.0
1.5
274
276
4
2.5
3.0
3.5
Fitted values
lm(GPA_F ~ ACT + Quantile * quant_thresh)



```
#the anova confirms that our reduced model's added coefficients were significantly different from 0
anova(fit1, fit.r)
```

```
## Analysis of Variance Table
##
## Model 1: GPA_F ~ Quantile + ACT + Class_year
## Model 2: GPA_F ~ ACT + Quantile * quant_thresh
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    701 225.64
## 2    700 214.12  1    11.512 37.635 1.427e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When we plot quantile vs gpa and fit a smooth function to the data, we see that when we reach the top 20% of the class (quantile greater than or equal to 80), the slope of function steepens. Thus, we should build a more flexible model that allows for different slopes based on which quantile group the students are in. This boosts our R^2 (0.2446) and we get a number of significant coefficients (seen in summary and the ANOVA comparing the full model and this reduced model).

When we consult our diagnostic plots, our residuals vs fitted values plot (plot 1) shows relative homoskedasticity and strong linearity. The Normal Q-Q plot (plot 2) shows that the residuals are approximately normally distributed, so another condition for the linear model checks out. The Scale-Location plot (plot 3) shows a slight downward trend in the standardized residuals, however the residuals are evenly spaced and the trend isn't appreciably large. The residuals vs leverage plot shows that there are no points of high leverage that are skewing our model. Thus our model linear model looks like a good fit and captures signal!

```
#interesting confidence and prediction intervals
newACT <- c(34)
newRank <- c(90)
newquantthresh<-c(1)
new.data <- data.frame(ACT=newACT, Quantile=newRank, quant_thresh=newquantthresh)
#CI
```

```r
predict(fit.r, new.data, interval="confidence")
```

```
##        fit      lwr     upr
## 1 3.424187 3.313703 3.53467
```

```r
#PI
predict(fit.r, new.data, interval="predict")
```

```
##        fit      lwr      upr
## 1 3.424187 2.332696 4.515677
```

Pomona College's Class of 2023 has a median ACT score of 34 and 90% of admitted students were in the top decile of their high school class. I use my model to construct a confidence interval for the true freshman year GPA is of a typical PO admittee (ACT 34 and 90th percentile): the 95% CI predicts the true mean freshman year GPA will lie between [3.46, 3.59]. Then I created a prediction interval, which generated a range of plausible GPA values for another student. We are 95% certain that the another student with these academic stats will have a freshman year GPA within these bounds: [2.34, 4.59].