Quantile Regression

Gabriel Vasquez

Case Scenario

As part of a data analyst team at a school district, my job is to collect, manage, analyze, and interpret educational data to inform administrative and instructional decisions. My responsibilities include maintaining student information systems and identifying trends in student performance. There have been recent concerns from parents and faculty about their children's high school academic performance. The district collected information on all of their students from the past academic year from two different schools, Gabriel Pereira and Mousinho da Silveira high schools. The data spans over only one academic school year, which is utilizing the quarter system (3 quarters = 1 academic school year). Each student record from this dataset comprises 33 features which includes information about the student's demographics, grade per quarter, and parental information. The district has given my team and I this data to draw up any useful information about these student's academic performance.

Question: We're curious to discover which student features (sex, age, study habits, attendance, family background, support programs, prior failures, etc.) are most significant to a student's GPA performance, and how do these associations differ across the lower (10%), middle (50%), and upper (90%) parts of the GPA distribution?

The Dataset

The dataset is comprised of 649 students from a school district, with 33 different features for each student. There are 16 numerical variables and 17 categorical variables.

Table 1: Variable definitions (subset of dataset)

Variable Name	**Type**	**Description**
school	Categorical	Student's school (binary: 'GP' – Gabriel Pereira or 'MS' – Mousinho da Silveira)
sex	Binary	Student's sex (binary: 'F' – female, 'M' – male)
age	Integer	Student's age (numeric: 15–22)
address	Categorical	Home address type (binary: 'U' – urban, 'R' – rural)
famsize	Categorical	Family size (binary: 'LE3' – "$\leq$ 3", 'GT3' – "$>$ 3")
Pstatus	Categorical	Parent cohabitation status (binary: 'T' – living together, 'A' – apart)
Medu	Integer	Mother's education (0 = none, 1 = primary, 2 = 5th–9th grade, 3 = secondary, 4 = higher)
Fedu	Integer	Father's education (0 = none, 1 = primary, 2 = 5th–9th grade, 3 = secondary, 4 = higher)
Mjob	Categorical	Mother's job (nominal: teacher, health care, civil services, at_home, other)
Fjob	Categorical	Father's job (nominal: teacher, health care, civil services, at_home, other)

Variable Name	Type	Description
reason	Categorical	Reason to choose school (home proximity, reputation, course preference, other)
guardian	Categorical	Student's guardian (mother, father, other)
traveltime	Integer	Home–school travel time (1 = <15 min, 2 = 15–30 min, 3 = 30–60 min, 4 = >1 h)
studytime	Integer	Weekly study time (1 = <2 h, 2 = 2–5 h, 3 = 5–10 h, 4 = >10 h)
failures	Integer	Number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	Binary	Extra educational support (yes/no)
famsup	Binary	Family educational support (yes/no)
paid	Binary	Extra paid classes within course subject (yes/no)
activities	Binary	Extra-curricular activities (yes/no)
walc	Integer	Weekend alcohol consumption (from 1=very low to 5=very high)
nursery	Binary	Attended nursery school (yes/no)

Mathematical Logic

What are quantiles? Think of quantiles as slices of a distribution. The q^{th} quantile (or slice) is the value below which q% of observations fall(ex: median = 50th percentile). Quantiles describe the shape of the distribution, not just its center.

Why quantile regression over OLS? As we'll see, OLS estimates the expected outcome (Y) given a set of parameters (X), but classical OLS inference assumes approximately normal, homoscedastic residuals. In skewed or heteroskedastic data, a single mean line hides how effects differ at the low vs high ends. Quantile regression allows us to see beyond the mean line at various quantile lines (like, the bottom 10% or top 10%).

For a chosen quantile $\tau \in (0, 1)$, the goal is to model the conditional τ^{th} quantile as:

$$Q_y(\tau|X) = X^T\beta_\tau$$

where X^T is the transpose of the predictor variables and β_τ is the coefficient vector at each τ. We interpret this as "holding all other features fixed, a one-unit change in a predictor shifts the τ^{th} conditional quantile of Y by the corresponding coefficient."

The estimated $\hat{\beta}_\tau$ values are calculated by using the quantile loss method, instead the sum of squared errors method. The objective function is as follows:

$$\min_{\beta_\tau} \sum_{i=1}^{n} \rho_\tau(Y_i - X_i^\top \beta_\tau)$$

- n = sample size.
- i = 1,2,3, ..., n = index for observation i.
- Y_i = response (outcome) for observation i.
- X_i = the (1 x p) vector of predictors for observation i. An (n x p) design matrix represents the stack of all observations.
- β_τ = the (p x 1) coefficient vector (a different β_τ is estimated differently at each τ).
- τ = the target quantile level (between 0 and 1).

$$\rho_\tau(\mu) = (\tau - \mathbb{1}_{\mu<0})\mu = \begin{cases} \tau\mu, & \mu \geq 0 \\ (\tau-1)|\mu|, & \mu < 0 \end{cases}$$

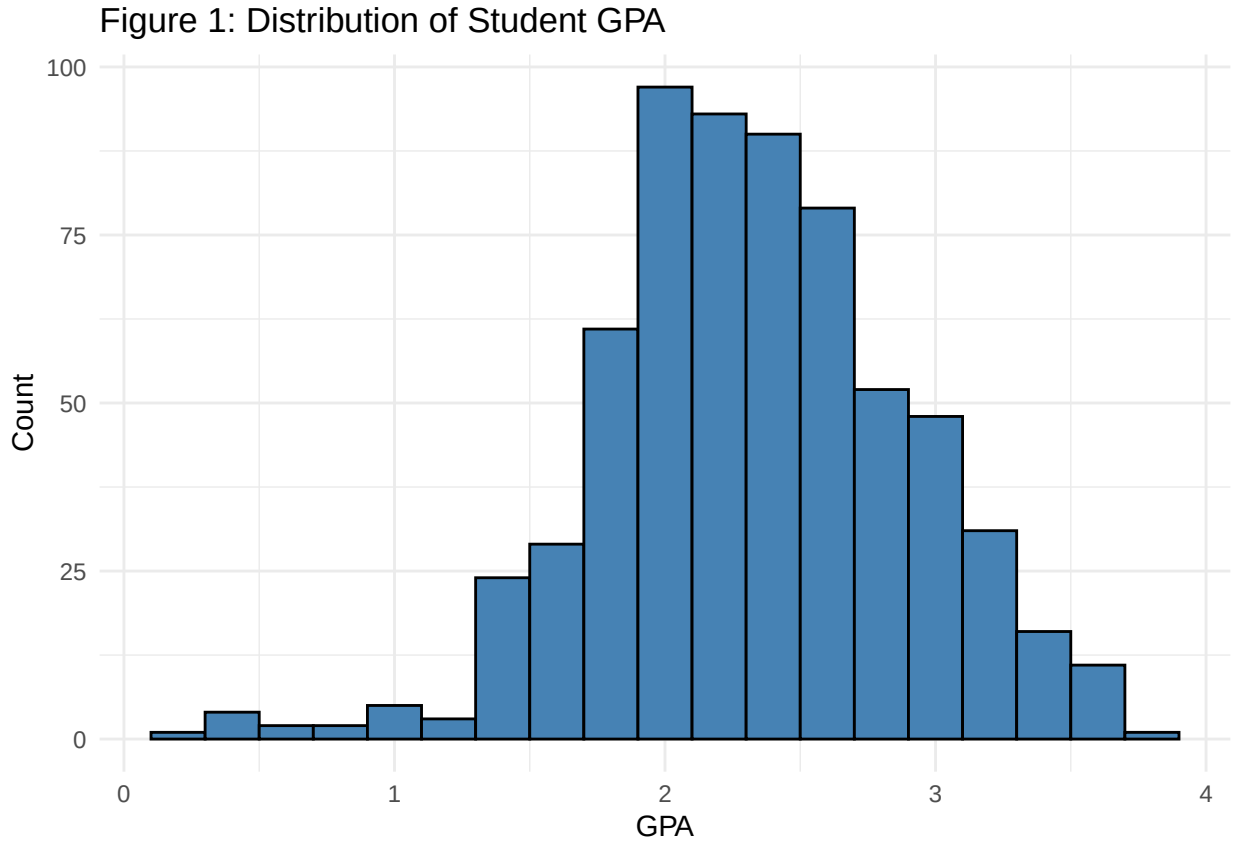
- $\rho_\tau(\mu)$ = the quantile loss (or pinball loss) which is an asymmetrically weighted absolute loss.
- μ = the residual from $\mu_i = Y_i - X_i^T\beta_\tau$.

Nonnegative residuals are weighted by τ; negative residuals by (τ-1). Minimizing the sum of these losses yields $\widehat{\beta}_\tau$.The asymmetric weights adjust the model fit so that about a fraction τ of residuals are ≤ 0 at the solution, targeting the conditional τ^{th} quantile rather than the mean.

No normality required. Heteroskedasticity is not a concern. More robust than OLS.

OLS Model

Before fitting an OLS model to the dataset, we observed from exploratory data analysis that the response variable (GPA) appears skewed. We also observed that all dependent variables from the dataset have moderate to less-than-moderate correlation with GPA. We annotated variables we believed to be most important based from our analysis.



Our exploratory data analysis and careful study behind each feature suggested to go with the following predictors for the OLS model:

$$GPA_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11}$$

where i = student i (1,2,3,...,649) and

- x_1 = walc, weekend alcohol consumption quantity (1=Very Low to 5=Very High).
- x_2 = studytime, hours per week of study (1,2,3,4).
- x_3 = sex (male, female).
- x_4 = schoolsup (extra educational support, yes, no).
- x_5 = school, ('GP'=Gabriel Pereira or 'MS'=Mousinho da Silveira).
- x_6 = higher, student wants to pursue a higher level of education (yes, no).
- x_7 = fedu, father's education (0=None, 1=Elementary, 2=Middle Education, 3=Secondary Education, 4=Higher Education).

- x_8, x_9, x_{10} = Failures, number of past class failures (0,1,2,3). Baseline is failures='0'.
- $x_1 1$ = absences, number of school absences (0 to 93).

We ran the OLS model to see if it could provide valuable information to answer our question.

```
ols_model <- lm(GPA ~ Walc + studytime + sex + schoolsup + school
                + higher + Fedu + failures + absences,
                data = df)
summary(ols_model)
```

```
##
## Call:
## lm(formula = GPA ~ Walc + studytime + sex + schoolsup + school +
##     higher + Fedu + failures + absences, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.07294 -0.29032 -0.02411  0.27814  1.47930
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.111341   0.095033  22.217  < 2e-16 ***
## Walc         -0.039916   0.015125  -2.639 0.008516 **
## studytime     0.077486   0.023138   3.349 0.000859 ***
## sexM         -0.097551   0.039593  -2.464 0.014008 *
## schoolsupyes -0.272407   0.059853  -4.551 6.39e-06 ***
## schoolMS     -0.261297   0.040770  -6.409 2.85e-10 ***
## higheryes     0.322992   0.063305   5.102 4.44e-07 ***
## Fedu          0.057423   0.017086   3.361 0.000823 ***
## failures1    -0.467440   0.061229  -7.634 8.34e-14 ***
## failures2    -0.502074   0.118379  -4.241 2.55e-05 ***
## failures3    -0.483442   0.126844  -3.811 0.000152 ***
## absences     -0.009105   0.004100  -2.221 0.026714 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4562 on 637 degrees of freedom
## Multiple R-squared:  0.3631, Adjusted R-squared:  0.3521
## F-statistic: 33.01 on 11 and 637 DF,  p-value: < 2.2e-16
```

$$GPA_i = 2.11 - 0.04x_1 + 0.078x_2 - 0.098x_3 - 0.272x_4 - 0.261x_5 + 0.323x_6 + 0.057x_7 - 0.467x_8 - 0.502x_9 - 0.483x_{10} - 0.009x_{11}$$

Figure 2: Density of OLS Model Residuals

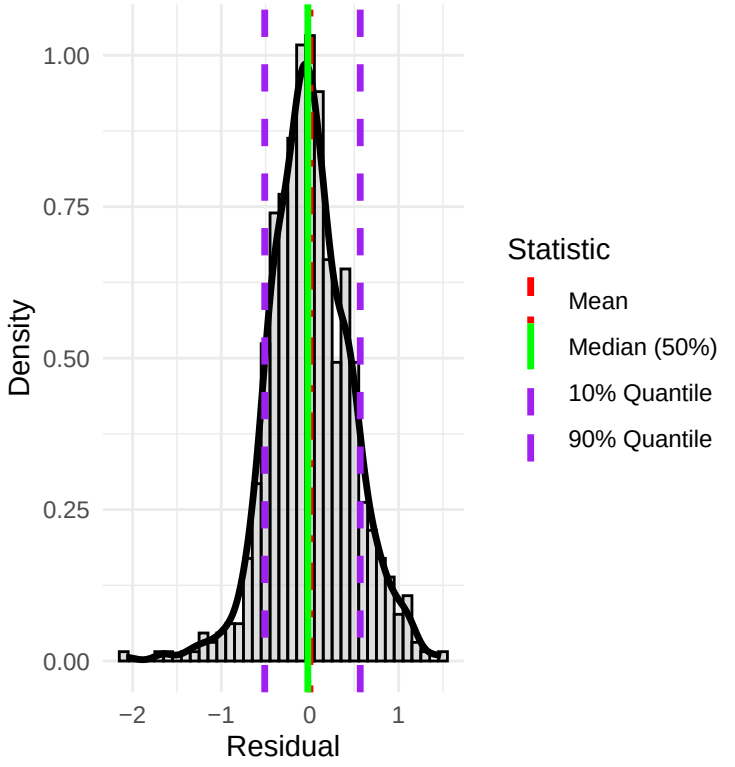
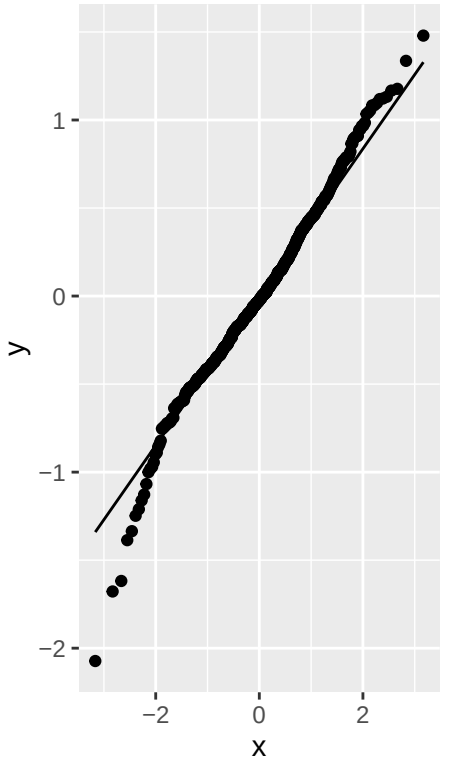


Figure 3: Normal Q−Q Plo



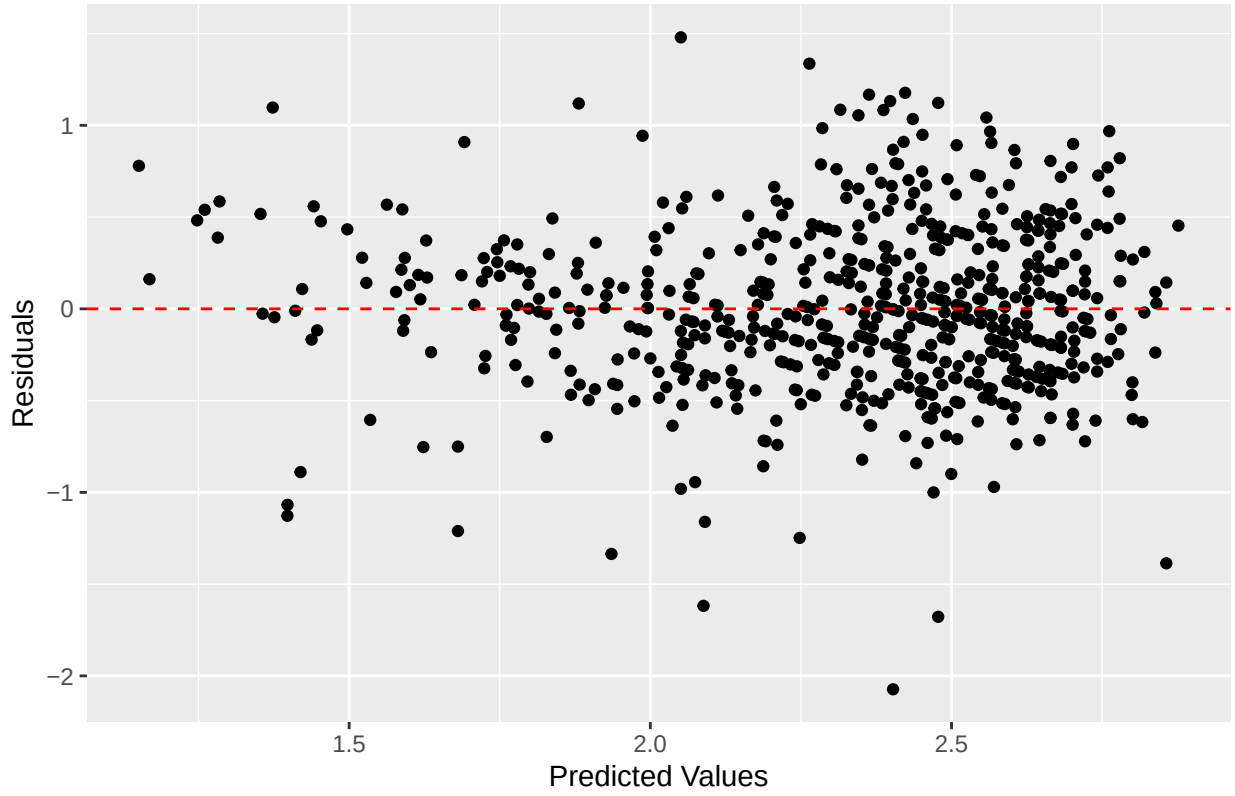

```
if (mean(res) > median(res)) {
  print(paste0("The mean (", round(mean(res), 21), ") > the median (", round(median(res), 4), ")."))
} else {
  print(paste0("The mean (", round(mean(res), 21), ") < the median (", round(median(res), 4), ")."))
}
```

```
## [1] "The mean (-1.2378e-17) > the median (-0.0241)."
```

The distribution almost appears to be normal. However, we note that the mean is greater than the median; therefore, the distribution is slightly more right skewed. We could remove the small set of outlier students (with GPA = 0) from the dataset to help normalize the distribution. But this is not ethically ideal because removing students would be doing a disservice to the district, the parents, and the students. So, we leave all students. The Shapiro-Wilk test testifies that the residuals from the OLS model do not follow normality because its (p-value=2.067e-05) $<$ ($\alpha = 0.05$).

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(ols_model)
## W = 0.98732, p-value = 2.067e-05
```

Figure 4: Residuals vs Predicted Values



The constant variance assumption does not appear satisfied because the predicted values between 0 and 1.75 have residuals plot closer around zero, while predicted values after 1.75 have their residuals fanning out more. These do not appear randomly scattered, therefore we observe a pattern.

Mathematical Logic: Failed Constant Variance

If we let Y represent the $(n \times 1)$ vector of response variable GPA, the $(p \times 1)$ vector $\hat{\beta}$ represent all estimated beta coefficients from our model, the $(n \times p)$ design matrix X represent the each student's predictor values (rows = students, columns = predictors), and the vector $\hat{r} = Y - X^T\hat{\beta}$ represent the model's residuals, then the true OLS model can be written as $Y = X\beta + \epsilon$ where ϵ represents the theoretical, unobservant deviations of the true linear model.

Since coefficients in an OLS model are solved with $\hat{\beta} = (X^TX)^{-1}X^TY$, then:

$$\hat{\beta} = (X^TX)^{-1}X^T \times (X\beta + \epsilon) = \beta + (X^TX)^{-1}X^T\epsilon$$

Classical OLS inference assumes its expected errors to be zero $(\epsilon|X)$=0 and constant variance $Var(\epsilon|X) = \sigma^2 I$ where I is the identity matrix $(X^TX)^{-1}$. Under the constant variance assumption, we have:

$$Var(\hat{\beta}|X) = \sigma^2 I = \sigma^2(X^TX)^{-1}$$

Since in our case the true student population variance (σ^2) is unknown, we have to estimate it from our the sample dataset provided by the district. The formula for estimating the sample variance is $(\hat{\sigma^2})$:

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-p} = \frac{\sum_{i=1}^{n} \hat{r}_i^2}{n-p}$$

where n=sample size (649 students) and p=size of vector $\hat{\beta}$ = 1 + number of predictors in the OLS model.

So, if the variance formula is dependent on the estimated $\widehat{\sigma^2}$, then it is also dependent on $\hat{r}$. If the residuals are not randomly spread, then the constant variance assumption is not satisfied. Our OLS model's standard errors, t-tests, and confidence intervals are only reliable if the assumptions are satisfied because of how dependent they are on the residuals. In our case, constant variance was violated because the $\hat{r}$ used to calculate the $\widehat{\sigma^2}$ is computing a $Var(\hat{\beta}|X) \neq \sigma^2(X^TX)^{-1}$.

OLS inference relies on the default covariance formula $Var(\hat{\beta}|X) \neq \sigma^2(X^TX)^{-1}$ with $\hat{\sigma}^2 = \frac{RSS}{(n-p)}$; that result is only valid when errors are uncorrelated and have constant variance. So if those assumptions fail the usual standard errors, t-tests, and confidence intervals are mis-sized. In our dataset, diagnostics show skewed GPA and residuals, Shapiro–Wilk rejects normality, and the residuals-vs-fitted plot fans out with groupwise tests indicating heteroskedasticity—conditions under which a single conditional-mean line can mask different relationships for low- vs high-performing students. Quantile regression addresses both issues by estimating conditional quantiles τ={.10,.50,.90} with weaker distributional assumptions and robustness to heteroskedasticity and skewness, letting us see which student features matter most across the lower, middle, and upper GPA ranges.

Quantile Regression Model

Warning: Solution may be nonunique

When attempting to fit a quatile regression model with the command ***rq(GPA ~ Walc + ... + absences, data = df, tau = 0.1, method = "fn")***, we received the following error message:

"Warning in rq.fit.br(x, y, tau = tau, ci = TRUE, ...) : Solution may be nonunique"

We received this same warning message for all three quantiles. This warning stems from many equal GPAs and a high-dimensional dummy-rich X^T, which makes the β solutions nonunique. Recall that the quantile regression model is optimizing the $\beta_{\tau,i,p}$ value, which is the coefficient estimate for student i's p^{th} predictor at τ. This is a linear programming problem. The objective is to find $\beta_{\tau ip}$ by minimizing the residuals. The linear programming solver searches for the convex optimum which makes roughly a τ fraction of residuals below the fitted quantile and a $(1-\tau)$ fraction above the fitted quantile so that $X^T\hat{\beta}_\tau$ estimates the τ^{th} conditional quantile of the student GPA given predictors X. When many GPAs are identical, the objective can develop nonconvex like regions which makes it difficult finding the minimum. This means multiple β's achieve the same minimum. We observe the data to determine the quantity of distinct GPA values, the average shared GPAs (ties), and

```
# See the quantity of distinct GPA values
length(unique(df$GPA))
```

```
## [1] 48
```

```
# Percent of students with same GPA (ties)
sum(duplicated(df$GPA))/length(df$GPA)
```

```
## [1] 0.9260401
```

```
# See frequency of repeated rounded values
table(df$GPA)
```

```
##
## 0.27 0.33 0.47 0.53  0.6  0.8 0.87 0.93    1 1.07 1.13 1.27 1.33  1.4 1.47 1.53
##    1    2    2    1    1    1    1    3    1    1    2    1    5    8   11    8
##  1.6 1.67 1.73  1.8 1.87 1.93    2 2.07 2.13  2.2 2.27 2.33  2.4 2.47 2.53  2.6
##    9   12   20   20   21   26   42   29   37   31   25   23   29   38   25   33
## 2.67 2.73  2.8 2.87 2.93    3 3.07 3.13  3.2 3.27 3.33  3.4 3.47 3.53  3.6 3.73
##   21   22   15   15   21   14   13   12   12    7    2    7    7    5    6    1
```

GPA takes only 48 distinct values in our sample and 92.6% of students share a GPA with at least one other student, creating many ties. We are at risk of the check loss objective becoming more flat than convex, leading to multiple equally optimal solutions. Therefore, we use the Frisch–Newton solver in our model fitting (to assist the dummy variables) and bootstrap standard errors (to assist with the flattening optima), which are robust to this degeneracy.

```
# Quantile Regession Model Fitting
set.seed(123)

QR_Model_10 <- rq(GPA ~ Walc + studytime + sex + schoolsup + school
                  + higher + Fedu + failures + absences,
                  data = df, tau = 0.1, method = "fn")

QR_Model_50 <- rq(GPA ~ Walc + studytime + sex + schoolsup + school
                  + higher + Fedu + failures + absences,
                  data = df, tau = 0.5, method = "fn")

QR_Model_90 <- rq(GPA ~ Walc + studytime + sex + schoolsup + school
                  + higher + Fedu + failures + absences,
                  data = df, tau = 0.9, method = "fn")
```

Bottom 10%

```
summary(QR_Model_10, se = "boot", R = 1000)
```

```
##
## Call: rq(formula = GPA ~ Walc + studytime + sex + schoolsup + school +
##     higher + Fedu + failures + absences, tau = 0.1, data = df,
##     method = "fn")
##
## tau: [1] 0.1
##
## Coefficients:
##              Value    Std. Error t value  Pr(>|t|)
## (Intercept)   1.84455  0.13490   13.67310  0.00000
## Walc         -0.04091  0.02356   -1.73670  0.08292
## studytime     0.01636  0.03566    0.45885  0.64650
## sexM         -0.16636  0.06713   -2.47825  0.01346
## schoolsupyes -0.22273  0.09199   -2.42133  0.01574
## schoolMS     -0.46000  0.08985   -5.11977  0.00000
## higheryes     0.30818  0.08764    3.51648  0.00047
## Fedu          0.02636  0.02525    1.04407  0.29685
## failures1    -0.49771  0.15303   -3.25231  0.00121
## failures2    -0.49727  0.13998   -3.55253  0.00041
## failures3    -0.35818  0.26746   -1.33919  0.18099
## absences     -0.00182  0.00684   -0.26590  0.79040
```

In the bottom decile, GPA is most strongly explained by school location (Mousinho da Silveira = −0.46), prior failures (−0.497), male (−0.166),higher education intention (0.308), and school support (-0.223); weekend alcohol use, absences, and studytime are not robust at $\tau = .10$ under the parsimonious specification.

Median 50%

```
summary(QR_Model_50, se = "boot", R = 1000)
```

```
##
## Call: rq(formula = GPA ~ Walc + studytime + sex + schoolsup + school +
##     higher + Fedu + failures + absences, tau = 0.5, data = df,
##     method = "fn")
##
## tau: [1] 0.5
##
## Coefficients:
##              Value    Std. Error t value  Pr(>|t|)
## (Intercept)   2.17867  0.09399   23.18006  0.00000
## Walc         -0.02800  0.01611   -1.73803  0.08269
## studytime     0.05800  0.02919    1.98722  0.04733
## sexM         -0.11033  0.04377   -2.52087  0.01195
## schoolsupyes -0.21700  0.05945   -3.64988  0.00028
## schoolMS     -0.26433  0.04785   -5.52477  0.00000
## higheryes     0.21700  0.06273    3.45943  0.00058
## Fedu          0.06667  0.02093    3.18518  0.00152
## failures1    -0.38400  0.06657   -5.76825  0.00000
## failures2    -0.59924  0.12543   -4.77741  0.00000
## failures3    -0.46148  0.08902   -5.18422  0.00000
## absences     -0.01217  0.00500   -2.43481  0.01517
```

At the median, absences and study time become more strongly influence to a student's GPA (they both become statistically significant). School location (Mousinho da Silveira) and failures still bring greatest risk to a student's GPA. Also, GPA decreases more when a student seeks higher education at the middle compared to the bottom 10%.

If we observe male students, compared to the bottom 10%, we see they are slightly more associated with lower GPAs than the middle. Holding all other variables fixed, the 10^{th}-quantile GPA for male students is about 0.166 units lower than for comparable females. However, at the 50^{th}-quantile, the median GPA for males is about 0.110 units lower than for comparable females.

If we observe the father's education, compared to the bottom 10%, we see they are slightly more associated with middle GPAs than with lower GPAs. We can say that a student's father's education is more meaningful to their academic performance in the middle of the distribution because the p-value is less than the significance level 0.05 at the 50% quantile, compared to the 10% quantile which is greater than 0.05.

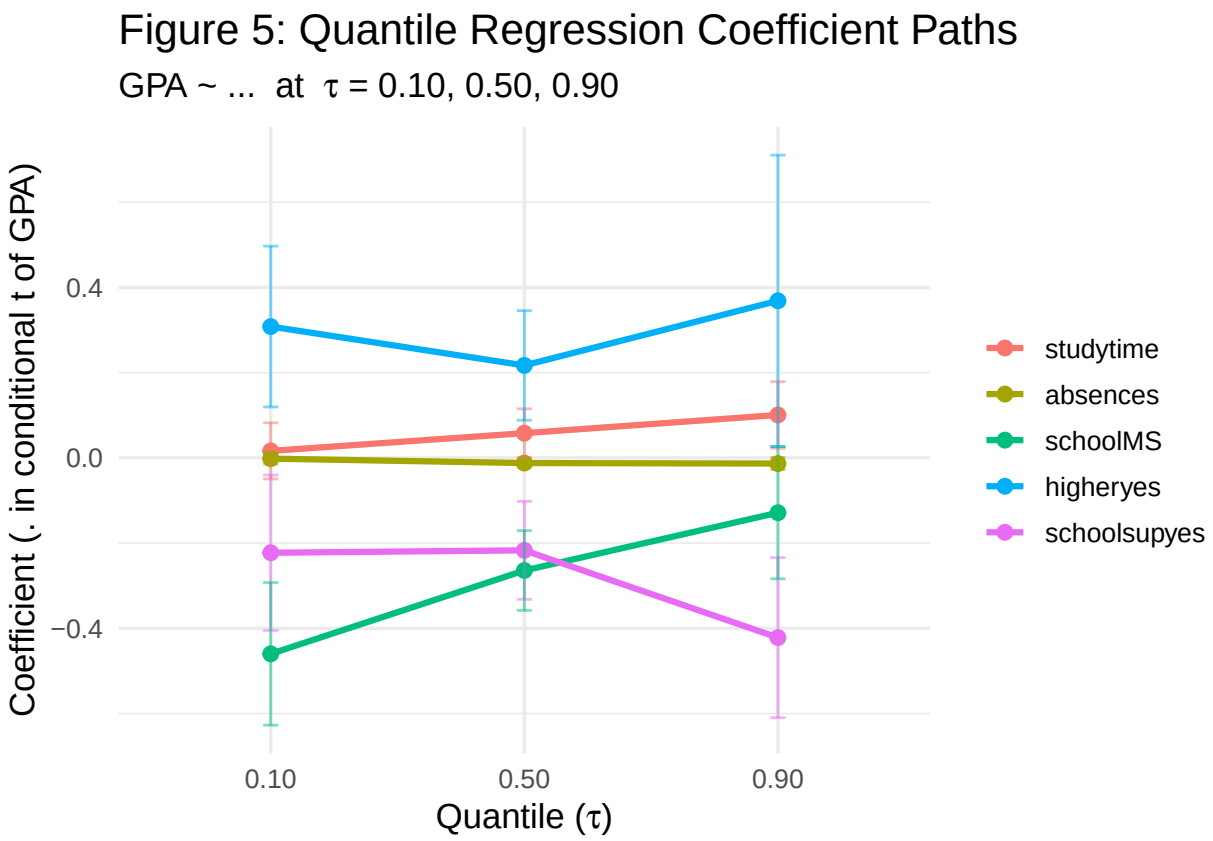
Top 90%

```r
summary(QR_Model_90, se = "boot", R = 1000)
```

```
##
## Call: rq(formula = GPA ~ Walc + studytime + sex + schoolsup + school +
##     higher + Fedu + failures + absences, tau = 0.9, data = df,
##     method = "fn")
##
## tau: [1] 0.9
##
## Coefficients:
##              Value    Std. Error t value  Pr(>|t|)
## (Intercept)   2.57529  0.21205   12.14477  0.00000
## Walc         -0.05529  0.02384   -2.31929  0.02069
## studytime     0.10064  0.03762    2.67492  0.00767
## sexM         -0.04257  0.07642   -0.55708  0.57767
## schoolsupyes -0.42186  0.10109   -4.17289  0.00003
```

```
## schoolMS     -0.12871  0.08083   -1.59238  0.11180
## higheryes     0.36864  0.17065    2.16024  0.03113
## Fedu          0.05936  0.03626    1.63692  0.10214
## failures1    -0.46735  0.16745   -2.79103  0.00541
## failures2    -0.47229  0.30753   -1.53573  0.12510
## failures3    -0.61471  0.24493   -2.50974  0.01233
## absences     -0.01336  0.00684   -1.95244  0.05132
```

School support, school location (although now insignificant), higher education intentions, and study time have a higher impact to the 90% GPAs. Absence becomes insignificant, but its change in effect to the GPA didn't change much. Father education and a student's sex have remain insignificant.



Quantile regression using the OLS-aligned specification shows clear heterogeneity across the GPA distribution. The school location (Mousinho da Silveira)is pronounced in the bottom decile (−0.46) but increase at the median (−0.26) and is only marginal at the top (−0.13), indicating school location matters most for at-risk students. In contrast, the benefit of study time grows with performance (0.016 to 0.058 to 0.10 per level). Absences is insignificant at the bottom, clearly matters at the median, and is marginal at the top. Prior failures remain strongly negative at all quantiles, and higher-education intention is positive throughout, with the largest association in the top decile (0.37). School support is negatively associated at all quantiles (consistent with selection into support), and male is lower at the bottom and median but not at the top. Overall, school context and failure history dominate risk in the lower tail, study time and aspiration differentiate higher performers, and absences are a mid-to-upper-tail lever.

Results and Conclusion

We built three quantile regression models

$$Q_y(\tau|X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11}$$

with the OLS-aligned specification at $\tau = \{.10, .50, .90\}$. We were set out to answer the question of which student features are most significant to GPA performance, and how those associations differ across the lower (10%), middle (50%), and upper (90%) parts of the GPA distribution. The quantile regression analysis shows that some factors matter consistently—most notably, prior failures exert large negative effects across all quantiles—while others vary by performance level. School location (Mousinho da Silveira) is strongly negative in the bottom decile but diminishes by the top, indicating school location effects are most consequential for at-risk students. Study time provides little lift at the bottom but grows steadily more beneficial toward the 90th percentile, while absences mainly depress the median and upper tail. Higher-education aspirations are positive at all levels, with the strongest impact among top performers. School support remains negatively associated across quantiles, consistent with selection into support among struggling students, and male students worse at the bottom and median but not at the top. Collectively, these results confirm that risk and success factors differ across the distribution: failures and school location dominate at the bottom, attendance and steady study time matter in the middle, and aspiration and study intensity distinguish high achiever