Michael Purtle Model Stuff

#Exploring Key Perdictors in Mobility

In our analysis of the factors associated with mobility, we employed both exploratory data analysis (EDA) and a multiple linear regression model, and a beta regression model to determine which variables exhibit a strong relationship with our outcome measure. Based on the regression output and correlation values from our EDA, we identified a subset of variables that appear particularly influential in explaining mobility.

mobility <- read.csv("mobility-all.csv", header = TRUE, stringsAsFactors = TRUE)  
  
quals <- c("ID","Name", "State", "Latitude", "Longitude")  
  
mobility <- mobility[,!(names(mobility) %in% quals)]  
  
bad\_cols <- c("Colleges","Tuition", "Graduation", "HS\_dropout") # +100 NULL  
mobility <- mobility[,!(names(mobility) %in% bad\_cols)]  
  
before <- nrow(mobility)  
mobility <- drop\_na(mobility)  
dropped <- before - nrow(mobility)  
  
print("Data reduced by: ")

## [1] "Data reduced by: "

print((dropped/before))

## [1] 0.145749

lm.1 <- lm(formula = Mobility ~ .,   
 data = mobility)  
  
summary(lm.1)

##   
## Call:  
## lm(formula = Mobility ~ ., data = mobility)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.072790 -0.014110 -0.001634 0.011117 0.155600   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.408e-01 5.642e-02 2.496 0.01284 \*   
## Population 1.265e-09 1.485e-09 0.852 0.39463   
## Urban 1.523e-03 3.434e-03 0.444 0.65753   
## Black 6.422e-02 2.302e-02 2.790 0.00544 \*\*   
## Seg\_racial -4.778e-02 1.475e-02 -3.238 0.00127 \*\*   
## Seg\_income -2.523e-01 7.164e-01 -0.352 0.72481   
## Seg\_poverty -6.192e-02 3.826e-01 -0.162 0.87148   
## Seg\_affluence 2.447e-01 3.649e-01 0.671 0.50269   
## Commute 5.881e-02 2.005e-02 2.933 0.00348 \*\*   
## Income 4.698e-07 4.603e-07 1.021 0.30787   
## Gini 2.227e+00 2.734e+00 0.814 0.41574   
## Share01 -2.247e-02 2.735e-02 -0.822 0.41148   
## Gini\_99 -2.341e+00 2.731e+00 -0.857 0.39158   
## Middle\_class 1.117e-01 3.488e-02 3.204 0.00143 \*\*   
## Local\_tax\_rate 2.019e-01 1.670e-01 1.209 0.22714   
## Local\_gov\_spending 8.749e-07 1.665e-06 0.525 0.59945   
## Progressivity 5.321e-03 8.745e-04 6.085 2.08e-09 \*\*\*  
## EITC -3.454e-04 3.526e-04 -0.980 0.32773   
## School\_spending 1.299e-03 1.509e-03 0.861 0.38941   
## Student\_teacher\_ratio 4.133e-04 7.256e-04 0.570 0.56918   
## Test\_scores -8.006e-05 2.294e-04 -0.349 0.72726   
## Labor\_force\_participation -3.130e-02 3.705e-02 -0.845 0.39854   
## Manufacturing -1.725e-01 2.094e-02 -8.237 1.11e-15 \*\*\*  
## Chinese\_imports -8.960e-04 7.309e-04 -1.226 0.22072   
## Teenage\_labor -1.819e+00 1.747e+00 -1.041 0.29819   
## Migration\_in -5.001e-01 2.464e-01 -2.030 0.04281 \*   
## Migration\_out 4.755e-02 2.990e-01 0.159 0.87369   
## Foreign\_born 1.930e-02 4.011e-02 0.481 0.63061   
## Social\_capital -3.523e-03 1.985e-03 -1.775 0.07645 .   
## Religious 4.655e-02 9.886e-03 4.709 3.10e-06 \*\*\*  
## Violent\_crime -4.199e+00 1.321e+00 -3.179 0.00156 \*\*   
## Single\_mothers -3.575e-01 6.966e-02 -5.132 3.89e-07 \*\*\*  
## Divorced -1.735e-01 1.183e-01 -1.467 0.14282   
## Married 5.582e-04 5.074e-02 0.011 0.99123   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.02658 on 599 degrees of freedom  
## Multiple R-squared: 0.7201, Adjusted R-squared: 0.7047   
## F-statistic: 46.7 on 33 and 599 DF, p-value: < 2.2e-16

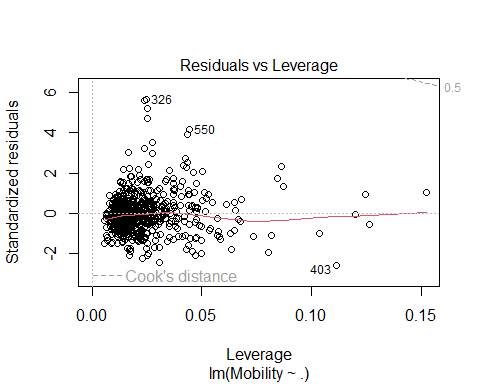
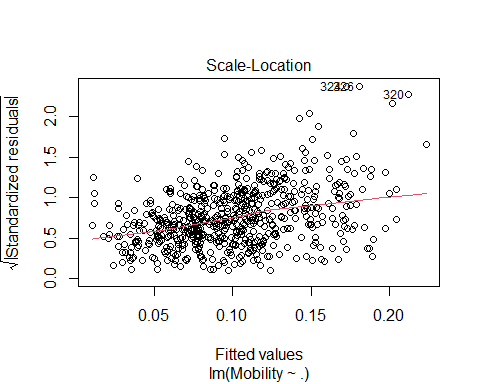
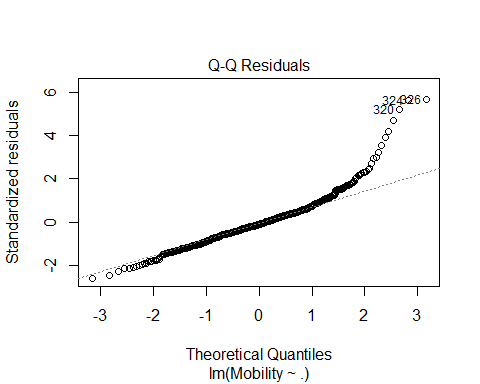
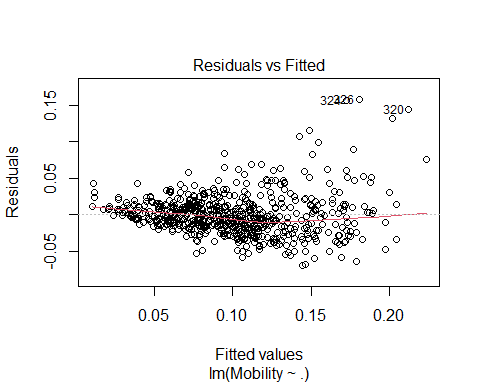
Taking into account the exploratory data analysis and the linear model from above, we now have a list of possible predictors that we can use. 1. Educational Spending + Scool Spending: Higher per-pupil expenditures can signal greater investment in educational resources, facilities, and student support programs, thereby fostering improved long-term mobility outcomes. + Test Scores: These serve as a proxy for overall educational quality and student achievement, and they often correlate positively with economic and social mobility. + Colleges: The presence and density of colleges in an area can enhance the availability of higher education and skill development opportunities, contributing to upward mobility. 2. Social and Demographic Factors + Black & Seg\_racial: The proportion of Black residents and the degree of racial segregation in a community are crucial indicators, reflecting underlying social structures and potential barriers or pathways to mobility. Areas with higher racial segregation often experience limited socioeconomic opportunities. + Commute: Longer or more prevalent commutes could signal broader labor markets or suburban sprawl. Where commutes are feasible, individuals may have access to a wider range of job opportunities. + Gini: Income inequality is often inversely related to mobility; higher inequality can concentrate disadvantage and limit pathways for advancement. + Middle\_class & Progressivity: Communities with a robust middle class or more progressive tax structures may provide support systems (e.g., social services, quality infrastructure) that encourage upward mobility. + Manufacturing: A heavy reliance on manufacturing might limit economic diversification; in many cases, deindustrialization or technological shifts in manufacturing can hinder long-term mobility prospects. + Migration\_in: Areas experiencing inbound migration may be more economically dynamic, suggesting better job prospects and growth, which in turn can improve mobility opportunities. + Religious: Higher religiosity can be associated with stronger community networks or social support, potentially facilitating resource sharing and stability that foster mobility. + Violent\_crime: High crime rates often correspond to reduced social cohesion and economic investment, negatively affecting people’s prospects for advancement. + Single\_mothers & Divorced: Family structure measures can indicate economic vulnerability or instability, impacting children’s outcomes and the intergenerational transmission of opportunity. 3. Government Spending + Local\_tax\_rate: The percentage of local taxes levied on residents and businesses. While higher rates can reduce disposable income, they can also fund public services that may boost economic mobility. + Local\_goc\_spending: Reflects how much local authorities invest in public services, infrastructure, and community programs. Effective spending can expand opportunities and resources, potentially promoting upward mobility.

edu\_corval <- c("School\_spending", "Test\_scores", "Colleges") #for edu stuff  
social\_cor <- c("Black", "Seg\_racial", "Commute", "Gini", "Middle\_class",   
 "Progressivity", "Manufacturing", "Migration\_in", "Religious",   
 "Violent\_crime", "Single mothers", "Divorced") #social stuff  
gov\_spending\_stuff <- c("Local\_tax\_rate", "Local\_goc\_spending")  
  
high\_cor <- mobility[, names(mobility) %in% c("Mobility",edu\_corval, social\_cor, gov\_spending\_stuff)]

From a methodological standpoint, these variables as signigicant predictors in our regression and/or showed string correlation in EDA phase. To refine this model, we plan to:

* Check for Multicollinearity using Variance Inflation Factors (VIFs) to ensure no subset of variables is overly redundant.
* Check for Heteroskedasticity see if any of the variable have varying variance
* How influential are outliners in the model and see if we need to remove them (More here no?)

yikes <- lm(Mobility ~ ., data = high\_cor)  
  
plot(yikes)

 We fitted the model, now let look for co-linearity between the variables then we can decide on how to deal with them.

vif(yikes)

## Black Seg\_racial Commute Gini Middle\_class   
## 2.394938 1.320021 2.932599 2.684005 4.097972   
## Local\_tax\_rate Progressivity School\_spending Test\_scores Manufacturing   
## 1.712259 1.229092 1.655802 2.264826 1.723279   
## Migration\_in Religious Violent\_crime Divorced   
## 1.389331 1.787569 1.613089 1.765527

Since the above facotrs are around 1-3, which suggest that col-linearity isn’t a huge concern. The highest VIF score is the “Middle\_class” (about 4.097), which is borderline but still not extremely high. It’s something to keep an eye on, but not a immediate concern.

coeftest(yikes, vcov = vcovHC(yikes, type = "HC3"))

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.05698998 0.03222521 1.7685 0.0774722 .   
## Black -0.04578197 0.01188349 -3.8526 0.0001291 \*\*\*  
## Seg\_racial -0.08977309 0.01695497 -5.2948 1.658e-07 \*\*\*  
## Commute 0.01104064 0.01722973 0.6408 0.5218967   
## Gini -0.01518672 0.02609392 -0.5820 0.5607773   
## Middle\_class 0.18999556 0.02912158 6.5242 1.424e-10 \*\*\*  
## Local\_tax\_rate 0.56297971 0.24900925 2.2609 0.0241137 \*   
## Progressivity 0.00463411 0.00109617 4.2276 2.720e-05 \*\*\*  
## School\_spending -0.00063444 0.00158336 -0.4007 0.6887862   
## Test\_scores -0.00019584 0.00022102 -0.8861 0.3759231   
## Manufacturing -0.16876108 0.02165467 -7.7933 2.777e-14 \*\*\*  
## Migration\_in -0.34146670 0.12459017 -2.7407 0.0063077 \*\*   
## Religious 0.04806408 0.00958220 5.0160 6.905e-07 \*\*\*  
## Violent\_crime -2.50481983 1.39186687 -1.7996 0.0724096 .   
## Divorced -0.52885532 0.08737476 -6.0527 2.471e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The above result is a summary of each predictors estimated coefficient, its standard error, the t-statistic (coefficient / standard error), and the associated p-value. In short, there are variables like Black, Seg\_Racial, Middle\_class, Progressivity, Manufacturing, Religious, Divorced, and Local\_tax\_rate show significant relationships while some of the others do not reach conventional significance thresholds. And we in fact use the more significant variables when choosing our next model below.

yikes\_refit <- lm(Mobility ~ Black + Seg\_racial + Middle\_class + Progressivity + Manufacturing + Migration\_in + Religious + Divorced + Local\_tax\_rate, data = high\_cor)  
coeftest(yikes\_refit, vcov = vcovHC(yikes\_refit, type = "HC3"))

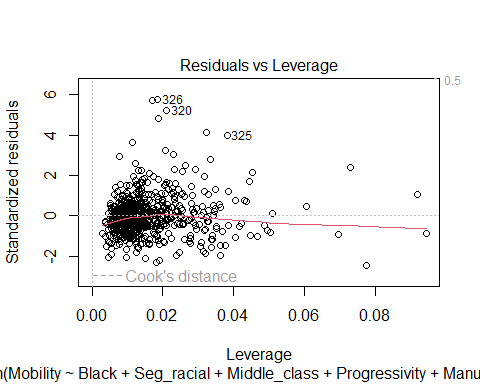
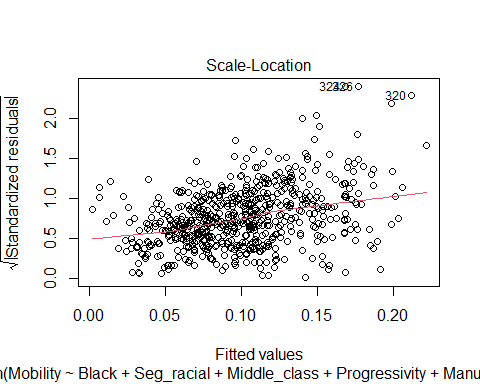
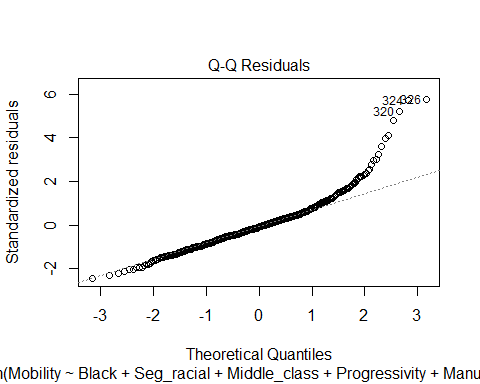
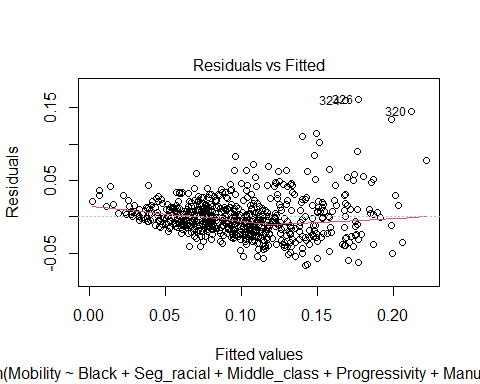
##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.04179063 0.01956206 2.1363 0.0330432 \*   
## Black -0.04962636 0.01073611 -4.6224 4.614e-06 \*\*\*  
## Seg\_racial -0.09329019 0.01407567 -6.6278 7.386e-11 \*\*\*  
## Middle\_class 0.20544945 0.02397602 8.5690 < 2.2e-16 \*\*\*  
## Progressivity 0.00425447 0.00088835 4.7892 2.095e-06 \*\*\*  
## Manufacturing -0.17453221 0.01996929 -8.7400 < 2.2e-16 \*\*\*  
## Migration\_in -0.39467805 0.10803502 -3.6532 0.0002807 \*\*\*  
## Religious 0.05151437 0.00954470 5.3972 9.634e-08 \*\*\*  
## Divorced -0.53260576 0.08454566 -6.2996 5.641e-10 \*\*\*  
## Local\_tax\_rate 0.51959734 0.21268622 2.4430 0.0148414 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(yikes\_refit)

##   
## Call:  
## lm(formula = Mobility ~ Black + Seg\_racial + Middle\_class + Progressivity +   
## Manufacturing + Migration\_in + Religious + Divorced + Local\_tax\_rate,   
## data = high\_cor)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.066467 -0.016442 -0.002381 0.012039 0.161265   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.0417906 0.0183809 2.274 0.023330 \*   
## Black -0.0496264 0.0128633 -3.858 0.000126 \*\*\*  
## Seg\_racial -0.0932902 0.0118153 -7.896 1.31e-14 \*\*\*  
## Middle\_class 0.2054495 0.0203783 10.082 < 2e-16 \*\*\*  
## Progressivity 0.0042545 0.0007718 5.512 5.19e-08 \*\*\*  
## Manufacturing -0.1745322 0.0160459 -10.877 < 2e-16 \*\*\*  
## Migration\_in -0.3946780 0.1261807 -3.128 0.001843 \*\*   
## Religious 0.0515144 0.0088718 5.807 1.02e-08 \*\*\*  
## Divorced -0.5326058 0.0834021 -6.386 3.33e-10 \*\*\*  
## Local\_tax\_rate 0.5195973 0.1362840 3.813 0.000151 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.02829 on 623 degrees of freedom  
## Multiple R-squared: 0.6702, Adjusted R-squared: 0.6654   
## F-statistic: 140.6 on 9 and 623 DF, p-value: < 2.2e-16

Having refined our model to include only statistically significant variables, we can now delve deeper into the data by identifying influential observations. The Q-Q plot below helps us pinpoint these observations, highlighting points that deviate from the expected distribution. These deviations may indicate outliers or data points that have a disproportionate impact on the model’s performance, warranting further investigation.

plot(yikes\_refit )



After examining the diagnostic plots (Residuals vs. Fitted, Q-Q Plot, Scale-Location, and Residuals vs. Leverage), we identified observations 320, 325, 326, and others as potential outliers or influential points. These points appeared to deviate substantially from the overall trend, suggesting they might exert an undue influence on the regression results.

Why Remove Outliers?

1. Influence on Parameter Estimates:

* Outliers can disproportionately affect the estimated coefficients, leading to skewed interpretations.

1. Violation of Model Assumptions:

* If extreme points violate assumptions of normality or homoscedasticity, they can compromise the validity of the model’s inferences.

1. Model Fit:

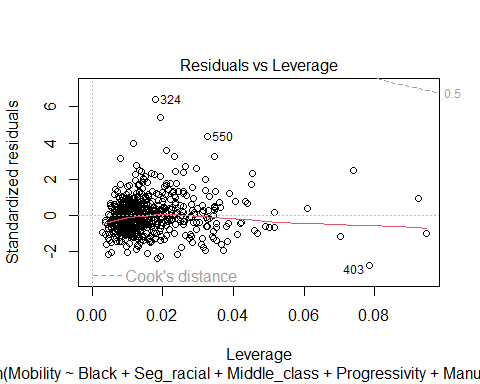
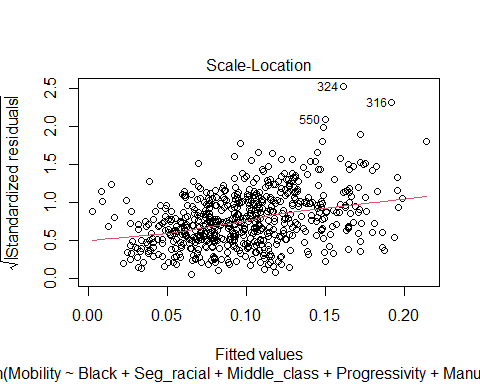
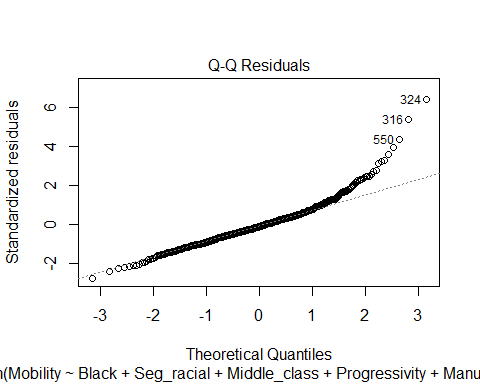
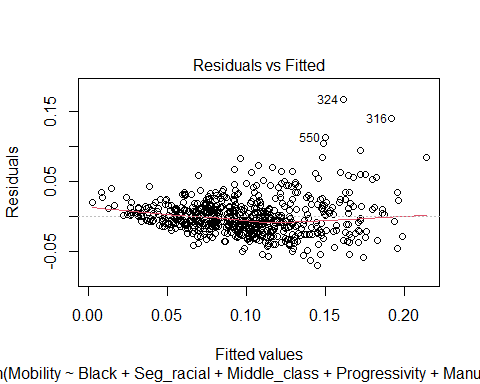
* Removing influential points may improve the model fit and reveal a clearer relationship among the remaining data.

Now er just have to remove the outliers and refit the model.

high\_cor\_clean <- high\_cor[-c(320, 325, 326, 385, 391, 393), ]  
yikes\_clean <- lm(Mobility ~ Black + Seg\_racial + Middle\_class + Progressivity + Manufacturing + Migration\_in + Religious + Divorced + Local\_tax\_rate, data = high\_cor\_clean)  
summary(yikes\_clean)

##   
## Call:  
## lm(formula = Mobility ~ Black + Seg\_racial + Middle\_class + Progressivity +   
## Manufacturing + Migration\_in + Religious + Divorced + Local\_tax\_rate,   
## data = high\_cor\_clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.069629 -0.015856 -0.002656 0.012126 0.167709   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.0480667 0.0172375 2.789 0.00546 \*\*   
## Black -0.0549853 0.0120468 -4.564 6.05e-06 \*\*\*  
## Seg\_racial -0.0985409 0.0111992 -8.799 < 2e-16 \*\*\*  
## Middle\_class 0.1890656 0.0191673 9.864 < 2e-16 \*\*\*  
## Progressivity 0.0036112 0.0007258 4.975 8.45e-07 \*\*\*  
## Manufacturing -0.1581657 0.0151367 -10.449 < 2e-16 \*\*\*  
## Migration\_in -0.3455706 0.1183115 -2.921 0.00362 \*\*   
## Religious 0.0479549 0.0083284 5.758 1.34e-08 \*\*\*  
## Divorced -0.5178954 0.0782790 -6.616 8.01e-11 \*\*\*  
## Local\_tax\_rate 0.5675728 0.1278860 4.438 1.07e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.02646 on 617 degrees of freedom  
## Multiple R-squared: 0.68, Adjusted R-squared: 0.6753   
## F-statistic: 145.7 on 9 and 617 DF, p-value: < 2.2e-16

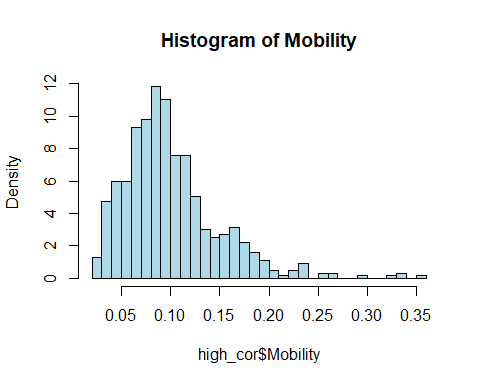
plot(yikes\_clean)



This is likely the best we can achieve using a simple linear regression model, given that we’re attempting to predict a probability (mobility). Because probabilities are bounded between 0 and 1, linear regression can struggle to provide accurate predictions or valid inferences in this context. As a result, we may need a more advanced modeling technique—such as logistic regression or another specialized method—to capture the probabilistic nature of mobility. This modeling mismatch is probably why we’ve encountered multiple issues when trying to predict mobility with a simple linear approach.

For the next model, I plan to use beta regression because it is specifically designed for response variables that lie between 0 and 1. This approach assumes the response follows a beta distribution, making it well-suited for modeling probabilities or other bounded outcomes. But first we need to see of mobility displays a beta distribution.

hist(high\_cor$Mobility, breaks = 30, probability = TRUE, main = "Histogram of Mobility", col = "lightblue")

 The histogram above does not provide sufficient evidence that mobility follows a beta distribution. Consequently, we will use the Kolmogorov-Smirnov (K-S) test next. Essentially, the K-S test asks: ‘How likely is it that we would observe these two samples if they were drawn from the same probability distribution?’

library(MASS)

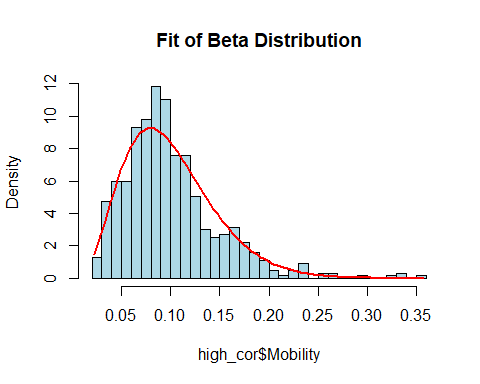
##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

# Fit beta distribution parameters  
mobility\_fit <- fitdistr(high\_cor$Mobility, "beta", start = list(shape1 = 2, shape2 = 5))

## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced  
## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced  
## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced  
## Warning in densfun(x, parm[1], parm[2], ...): NaNs produced

# Generate theoretical beta values  
x\_vals <- seq(min(high\_cor$Mobility), max(high\_cor$Mobility), length.out = 100)  
beta\_vals <- dbeta(x\_vals, mobility\_fit$estimate[1], mobility\_fit$estimate[2])  
  
# Plot the histogram and beta density curve  
hist(high\_cor$Mobility, breaks = 30, probability = TRUE, col = "lightblue", main = "Fit of Beta Distribution")  
lines(x\_vals, beta\_vals, col = "red", lwd = 2)



ks.test(high\_cor$Mobility, "pbeta", mobility\_fit$estimate[1], mobility\_fit$estimate[2])

## Warning in ks.test.default(high\_cor$Mobility, "pbeta",  
## mobility\_fit$estimate[1], : ties should not be present for the one-sample  
## Kolmogorov-Smirnov test

##   
## Asymptotic one-sample Kolmogorov-Smirnov test  
##   
## data: high\_cor$Mobility  
## D = 0.052811, p-value = 0.05856  
## alternative hypothesis: two-sided

The p-value from the one-sample Kolmogorov-Smirnov test is 0.05856, which is slightly above the conventional 0.05 threshold. This suggests that we do not have strong evidence to reject the null hypothesis that our data follow a beta distribution. However, before proceeding with beta regression, we need to verify that all values of our response variable lie strictly within the (0, 1) interval—a key assumption for beta regression.

We can confirm this in R with the following code:

all(high\_cor$Mobility > 0 & high\_cor$Mobility < 1)

## [1] TRUE

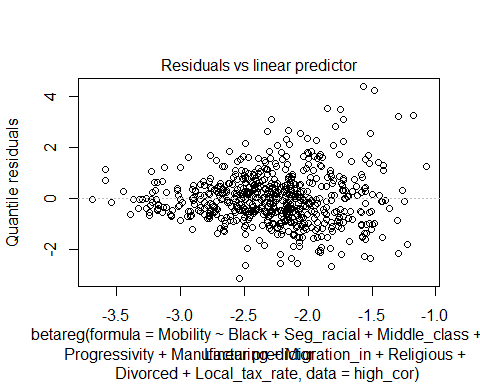
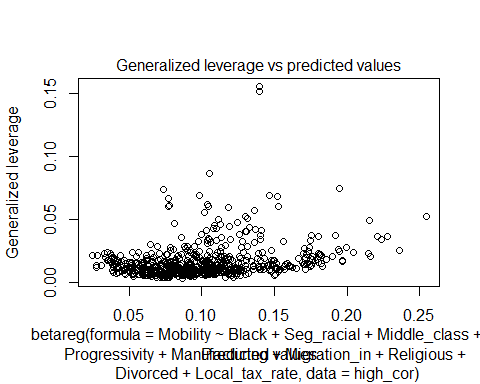
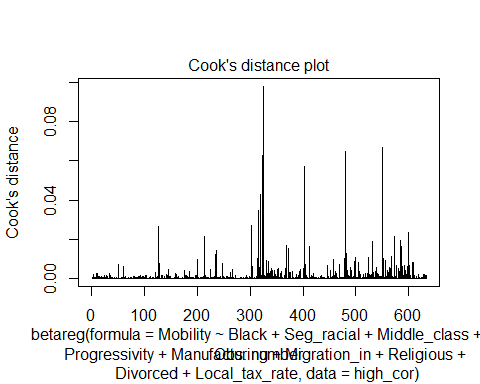
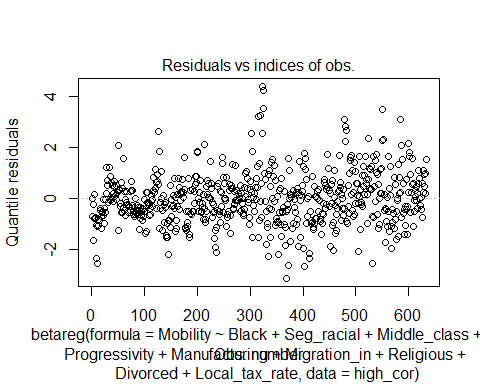
If this returns TRUE, it confirms that all values meet the (0, 1) requirement, and we can confidently move forward with the beta regression analysis.

Beta regression:

yikes\_beta <- betareg(Mobility ~ Black + Seg\_racial + Middle\_class + Progressivity + Manufacturing + Migration\_in + Religious + Divorced + Local\_tax\_rate,   
 data = high\_cor)  
summary(yikes\_beta)

##   
## Call:  
## betareg(formula = Mobility ~ Black + Seg\_racial + Middle\_class + Progressivity +   
## Manufacturing + Migration\_in + Religious + Divorced + Local\_tax\_rate,   
## data = high\_cor)  
##   
## Quantile residuals:  
## Min 1Q Median 3Q Max   
## -3.1124 -0.6053 -0.0501 0.5301 4.4049   
##   
## Coefficients (mean model with logit link):  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.80529 0.17635 -15.908 < 2e-16 \*\*\*  
## Black -1.26750 0.14661 -8.645 < 2e-16 \*\*\*  
## Seg\_racial -0.90970 0.11543 -7.881 3.25e-15 \*\*\*  
## Middle\_class 1.93784 0.19057 10.169 < 2e-16 \*\*\*  
## Progressivity 0.04289 0.00703 6.101 1.05e-09 \*\*\*  
## Manufacturing -1.66994 0.15308 -10.909 < 2e-16 \*\*\*  
## Migration\_in -2.21954 1.20252 -1.846 0.0649 .   
## Religious 0.50536 0.08251 6.125 9.09e-10 \*\*\*  
## Divorced -4.80191 0.78567 -6.112 9.85e-10 \*\*\*  
## Local\_tax\_rate 4.68461 1.19788 3.911 9.20e-05 \*\*\*  
##   
## Phi coefficients (precision model with identity link):  
## Estimate Std. Error z value Pr(>|z|)   
## (phi) 153.809 8.661 17.76 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1   
##   
## Type of estimator: ML (maximum likelihood)  
## Log-likelihood: 1497 on 11 Df  
## Pseudo R-squared: 0.7531  
## Number of iterations: 21 (BFGS) + 2 (Fisher scoring)

plot(yikes\_beta)



Overall, the beta regression model appears to be performing well based on several diagnostic indicators. First, the significance of the predictors in the mean model is high, with all coefficients (except Migration\_in, which is still below the 0.05 threshold) showing very low p-values. This suggests that each variable contributes meaningfully to explaining the variability in Mobility. Furthermore, the pseudo R-squared value of approximately 0.75 indicates that the model accounts for a substantial portion of the variation in the data—quite high for a beta regression.

Looking at the diagnostic plots, the residuals versus observation indices and the residuals versus the linear predictor do not exhibit strong patterns, suggesting no obvious violations of model assumptions. The Cook’s distance plot reveals that no single observation exerts undue influence on the model, and generalized leverage values are mostly low, indicating limited leverage points. Lastly, the estimated precision (phi ≈ 150) is relatively large, pointing to a tightly clustered distribution around the fitted values. Taken together, these diagnostics imply that the beta regression model is a good fit, although you may want to confirm its predictive performance through additional fit metrics or cross-validation.

model\_aic <- AIC(yikes\_beta)  
model\_bic <- BIC(yikes\_beta)  
  
cat("AIC:", model\_aic, "\n")

## AIC: -2972.293

cat("BIC:", model\_bic, "\n")

## BIC: -2923.338

set.seed(123)  
  
n\_folds <- 5  
  
  
folds <- sample(1:n\_folds, size = nrow(high\_cor), replace = TRUE)  
  
  
rmse\_vals <- numeric(n\_folds)  
  
for (k in 1:n\_folds) {  
   
  
 train\_idx <- which(folds != k)  
 test\_idx <- which(folds == k)  
   
  
 fit\_k <- betareg(  
 Mobility ~ Black + Seg\_racial + Middle\_class + Progressivity +   
 Manufacturing + Migration\_in + Religious + Divorced,  
 data = high\_cor[train\_idx, ]  
 )  
   
  
 preds <- predict(fit\_k, newdata = high\_cor[test\_idx, ], type = "response")  
 actuals <- high\_cor$Mobility[test\_idx]  
   
  
 rmse\_vals[k] <- sqrt(mean((preds - actuals)^2))  
}  
  
  
mean\_rmse <- mean(rmse\_vals)  
  
cat("Mean RMSE (5-fold CV):", mean\_rmse, "\n")

## Mean RMSE (5-fold CV): 0.02762679

The model’s fit is strong, as indicated by the AIC (-2972.293) and BIC (-2923.338), which suggest excellent in-sample fit, with lower values signifying a better model. The mean RMSE from the 5-fold cross-validation is 0.0276, meaning that, on average, the model’s predictions are off by just 2.8 percentage points—indicating good predictive accuracy. Overall, these results suggest that your beta regression model is well-calibrated, performing well both in terms of fit and prediction. However, comparing these metrics across different models can provide additional insights into the relative performance.

The negative AIC and BIC values indicate that the model has a high likelihood relative to its complexity, though these measures are most informative when comparing across multiple models. Meanwhile, the mean RMSE of approximately 0.028 suggests that, on average, the model’s predictions deviate from the actual values by about 2.8 percentage points on the 0–1 scale, which is fairly small. Taken together, these results suggest that the beta regression model fits the data well, although further comparisons with alternative models or additional diagnostics would provide a more comprehensive assessment.