Stat-302 Project (GitHub)

Miles Matthews

2024-01-02

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

### Importing data

CollegeDistance <- read.csv("CollegeDistance.csv")

## Some two-way relationships between variables

### Gender and distance:

#### Checking variables:

Checking what values are in the gender variable:

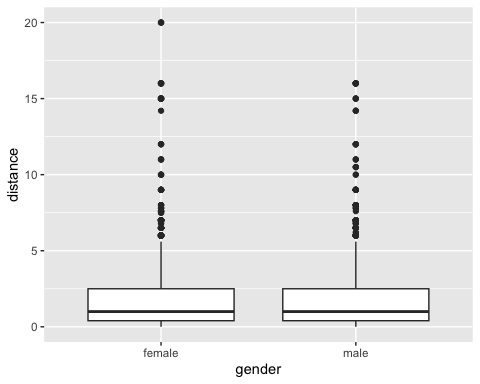
unique(CollegeDistance$gender)

## [1] "male" "female"

Since the relationship between a categorical variable and a quantitative variable is being examined, boxplots and histograms can be good ways to investigate this data.

#### Boxplot:

ggplot(data = CollegeDistance, mapping = aes(gender, distance))+  
 geom\_boxplot()



Given that the data for distance appears skewed, I will try transforming it from now on.

#### Transformed boxplot:

ggplot(data = CollegeDistance, mapping = aes(gender, log(distance)))+  
 geom\_boxplot()

## Warning: Removed 94 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).



It appears that the relationship between distance and gender is roughly the same for both men and women. The tail for women seems to contain larger values for distance than the tail for men, potentially indicating a gender disparity among the students who have the least access to a nearby college.

### Income and distance:

Checking what values are in the income variable:

unique(CollegeDistance$income)

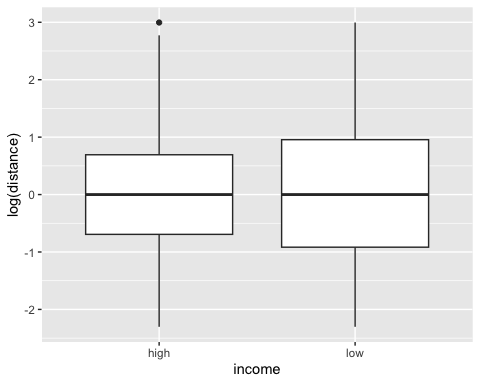
## [1] "high" "low"

Since a relationship between a categorical variable and quantitative variable is being examined, boxplots can be a good way to investigate this data.

#### Transformed boxplot:

ggplot(data = CollegeDistance, mapping = aes(income, log(distance)))+  
 geom\_boxplot()

## Warning: Removed 94 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).



Once again, the relationships between income and distance look roughly the same between different income levels, however the distance for low-income students appears to have more variability. There might be one outlier in the log of distance for high-income students.

## Multiple linear regression analysis:

Here I am preparing distance data by adding a new variable to college distance with the transformed distance data.

CollegeDistance2 <- CollegeDistance |>  
 mutate(log\_distance = log(distance+0.1))

### Full versus reduced model

Examining full model to choose variables to include in reduced model. In the reduced model, I will include the variables that R marks with one or more stars to indicate significant p-values.

mlr1 = lm(score ~ ., CollegeDistance2)  
summary(mlr1)

##   
## Call:  
## lm(formula = score ~ ., data = CollegeDistance2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.7629 -5.0831 0.0114 5.0365 24.3583   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.491e+01 1.188e+00 12.551 < 2e-16 \*\*\*  
## rownames 3.835e-05 2.448e-05 1.567 0.117285   
## gendermale 1.103e+00 2.078e-01 5.308 1.16e-07 \*\*\*  
## ethnicityhispanic 2.686e+00 3.639e-01 7.383 1.82e-13 \*\*\*  
## ethnicityother 6.380e+00 3.023e-01 21.104 < 2e-16 \*\*\*  
## fcollegeyes 1.225e+00 3.008e-01 4.071 4.75e-05 \*\*\*  
## mcollegeyes 1.050e+00 3.362e-01 3.123 0.001798 \*\*   
## homeyes 7.434e-01 2.750e-01 2.703 0.006895 \*\*   
## urbanyes -5.589e-01 2.676e-01 -2.089 0.036776 \*   
## unemp -8.917e-02 4.205e-02 -2.120 0.034024 \*   
## wage 1.812e-01 8.561e-02 2.116 0.034386 \*   
## distance 1.226e-01 8.200e-02 1.495 0.134903   
## tuition 2.245e+00 4.177e-01 5.374 8.06e-08 \*\*\*  
## education 1.927e+00 6.139e-02 31.386 < 2e-16 \*\*\*  
## incomelow 1.742e-01 2.506e-01 0.695 0.486864   
## regionwest 2.793e-01 3.998e-01 0.699 0.484843   
## log\_distance -6.243e-01 1.732e-01 -3.604 0.000316 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.086 on 4722 degrees of freedom  
## Multiple R-squared: 0.3392, Adjusted R-squared: 0.337   
## F-statistic: 151.5 on 16 and 4722 DF, p-value: < 2.2e-16

mlr2 = lm(score ~ gender + ethnicity + fcollege + mcollege + home + urban + unemp + wage + tuition + education + log\_distance, CollegeDistance2)  
summary(mlr2)

##   
## Call:  
## lm(formula = score ~ gender + ethnicity + fcollege + mcollege +   
## home + urban + unemp + wage + tuition + education + log\_distance,   
## data = CollegeDistance2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.6062 -5.0688 0.0456 5.0061 24.4732   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15.61445 1.11819 13.964 < 2e-16 \*\*\*  
## gendermale 1.08995 0.20766 5.249 1.60e-07 \*\*\*  
## ethnicityhispanic 2.76666 0.36237 7.635 2.72e-14 \*\*\*  
## ethnicityother 6.42706 0.29995 21.427 < 2e-16 \*\*\*  
## fcollegeyes 1.22494 0.29236 4.190 2.84e-05 \*\*\*  
## mcollegeyes 1.01673 0.33504 3.035 0.00242 \*\*   
## homeyes 0.70387 0.27388 2.570 0.01020 \*   
## urbanyes -0.57220 0.26660 -2.146 0.03190 \*   
## unemp -0.06936 0.04056 -1.710 0.08734 .   
## wage 0.20834 0.08436 2.470 0.01356 \*   
## tuition 1.66504 0.34179 4.872 1.14e-06 \*\*\*  
## education 1.91963 0.06102 31.457 < 2e-16 \*\*\*  
## log\_distance -0.43906 0.10584 -4.148 3.41e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.09 on 4726 degrees of freedom  
## Multiple R-squared: 0.3379, Adjusted R-squared: 0.3362   
## F-statistic: 201 on 12 and 4726 DF, p-value: < 2.2e-16

#### Testing to make sure that the reduced model explains score roughly as well as the full model:

Gathering ANOVA data:

anova(mlr1)

## Analysis of Variance Table  
##   
## Response: score  
## Df Sum Sq Mean Sq F value Pr(>F)   
## rownames 1 122 122 2.4354 0.1186884   
## gender 1 2315 2315 46.1135 1.254e-11 \*\*\*  
## ethnicity 2 47066 23533 468.7126 < 2.2e-16 \*\*\*  
## fcollege 1 13069 13069 260.2948 < 2.2e-16 \*\*\*  
## mcollege 1 2646 2646 52.6943 4.534e-13 \*\*\*  
## home 1 1061 1061 21.1337 4.395e-06 \*\*\*  
## urban 1 0 0 0.0100 0.9205227   
## unemp 1 0 0 0.0082 0.9278362   
## wage 1 837 837 16.6768 4.505e-05 \*\*\*  
## distance 1 1537 1537 30.6161 3.315e-08 \*\*\*  
## tuition 1 1982 1982 39.4734 3.625e-10 \*\*\*  
## education 1 50362 50362 1003.0790 < 2.2e-16 \*\*\*  
## income 1 24 24 0.4798 0.4885544   
## region 1 23 23 0.4569 0.4990922   
## log\_distance 1 652 652 12.9918 0.0003161 \*\*\*  
## Residuals 4722 237079 50   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(mlr2)

## Analysis of Variance Table  
##   
## Response: score  
## Df Sum Sq Mean Sq F value Pr(>F)   
## gender 1 2306 2306 45.8769 1.413e-11 \*\*\*  
## ethnicity 2 47174 23587 469.2800 < 2.2e-16 \*\*\*  
## fcollege 1 13090 13090 260.4316 < 2.2e-16 \*\*\*  
## mcollege 1 2647 2647 52.6654 4.600e-13 \*\*\*  
## home 1 1061 1061 21.1088 4.452e-06 \*\*\*  
## urban 1 0 0 0.0094 0.9226   
## unemp 1 0 0 0.0069 0.9337   
## wage 1 835 835 16.6105 4.665e-05 \*\*\*  
## tuition 1 2063 2063 41.0369 1.639e-10 \*\*\*  
## education 1 51195 51195 1018.5499 < 2.2e-16 \*\*\*  
## log\_distance 1 865 865 17.2087 3.408e-05 \*\*\*  
## Residuals 4726 237540 50   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Assigning values to relevant parts of ANOVA data:

full\_SSE <- 237079  
reduced\_SSE <- 237540   
ExtraSS <- reduced\_SSE - full\_SSE  
n <- 4738  
p <- 16  
q <- 12

Calculating F statistic:

F\_stat <- (ExtraSS / (p-q)) / (full\_SSE / (n-p))  
F\_stat

## [1] 2.295482

Calculating p-value:

anova(mlr1,mlr2)

## Analysis of Variance Table  
##   
## Model 1: score ~ rownames + gender + ethnicity + fcollege + mcollege +   
## home + urban + unemp + wage + distance + tuition + education +   
## income + region + log\_distance  
## Model 2: score ~ gender + ethnicity + fcollege + mcollege + home + urban +   
## unemp + wage + tuition + education + log\_distance  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 4722 237079   
## 2 4726 237540 -4 -460.9 2.295 0.05692 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The p-value (0.05692) is greater than 0.05, so we fail to reject that the reduced model is roughly as accurate as the full one for predicting score. However, it is barely over 0.05, we can proceed with the reduced model but we should proceed with caution.

Credit to <https://www.statology.org/how-to-calculate-the-p-value-of-an-f-statistic-in-r/> for reference.

### Stating assumptions:

1. Linearity
2. Independence of errors
3. Independence of independent variables
4. Homoskedasticity: similar variances for each group
5. Normality

Credit to <https://www.cfainstitute.org/en/membership/professional-development/refresher-readings/multiple-regression#:~:text=Five%20main%20assumptions%20underlying%20multiple,whether%20these%20assumptions%20are%20satisfied.> for reference.

### Does the data meet the assumptions?

#### Residual plots 1: scatterplot of the residuals versus explanatory variable

Note: I will only be checking the numerical explanatory variables (wage, tuition, education, and log of distance), as recommended by Dr. Kacar.

Plot for wage:

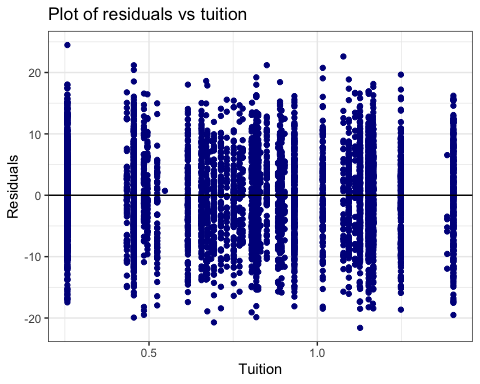
ggplot(mlr2, aes(x = CollegeDistance2$wage, y = .resid)) +   
 geom\_point(color = "dark blue") +  
 geom\_hline(yintercept = 0) +  
 xlab("Wage") +  
 ylab("Residuals") +  
 ggtitle("Residuals vs wage")+  
 theme\_bw()



The residuals seem evenly centered around zero with no concerning patterns, so we can assume that the data for wages meets the linearity condition.

Plot for tuition:

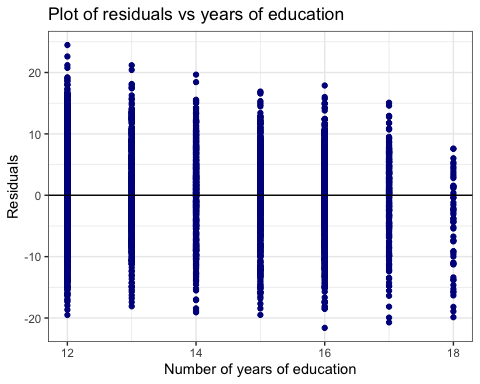
ggplot(mlr2, aes(x = CollegeDistance2$tuition, y = .resid)) +   
 geom\_point(color = "dark blue") +  
 geom\_hline(yintercept = 0) +  
 xlab("Tuition") +  
 ylab("Residuals") +  
 ggtitle("Plot of residuals vs tuition")+  
 theme\_bw()



The residuals seem evenly centered around zero with no concerning patterns, so we can assume that the data for tuition meets the linearity condition.

Plot for education:

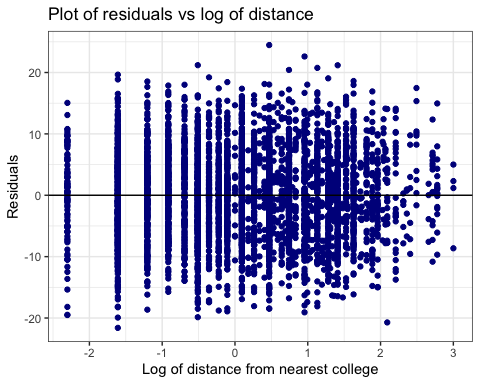
ggplot(mlr2, aes(x = CollegeDistance2$education, y = .resid)) +   
 geom\_point(color = "dark blue") +  
 geom\_hline(yintercept = 0) +  
 xlab("Number of years of education") +  
 ylab("Residuals") +  
 ggtitle("Plot of residuals vs years of education")+  
 theme\_bw()



The residuals seem roughly evenly centered around zero, however there does seem to be a downward trend in residuals as number of years of education increases. So, we should be cautious when assuming that the data in the model is linear for number of years of education. However, I don’t think that it is so concerning that we can’t continue with the test.

Plot for log of distance:

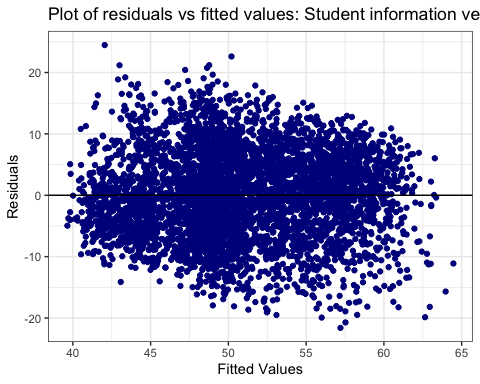
ggplot(mlr2, aes(x = CollegeDistance2$log\_distance, y = .resid)) +   
 geom\_point(color = "dark blue") +  
 geom\_hline(yintercept = 0) +  
 xlab("Log of distance from nearest college") +  
 ylab("Residuals") +  
 ggtitle("Plot of residuals vs log of distance")+  
 theme\_bw()



The residuals seem roughly evenly centered around zero. However they do seem condensed into the right side of the x axis. So, we should be cautious when assuming that the data in the model is linear for log of distance. However, I don’t think that it is so concerning that we can’t continue with the test.

#### Residual plot 2: scatterplot of the residuals versus fitted values

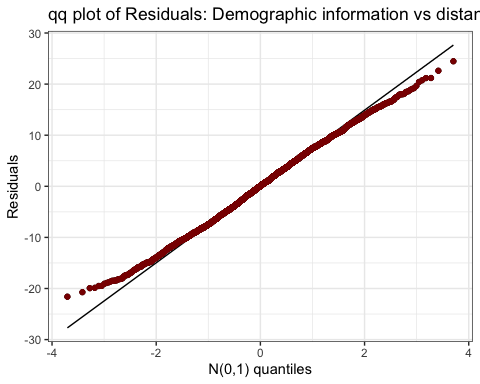
ggplot(mlr2, aes(x = .fitted, y = .resid)) +   
 geom\_point(color = "dark blue") +  
 geom\_hline(yintercept = 0) +  
 xlab("Fitted Values") +  
 ylab("Residuals") +  
 ggtitle("Plot of residuals vs fitted values: Student information versus student success")+  
 theme\_bw()



The residuals seem roughly evenly centered around zero with no concerning patterns, so we can assume that the data for the model meets the homoscedasticity condition.

#### Residual plot 3: Q-Q plot of the residuals

ggplot(mlr2, mapping = aes(sample = mlr2$residuals)) +   
 stat\_qq() +  
 stat\_qq\_line() +  
 geom\_qq( color = "dark red") +  
 ggtitle("qq plot of Residuals: Demographic information vs distance") +  
 ylab("Residuals") +  
 xlab("N(0,1) quantiles") +  
 theme\_bw()



The tails of the qq plot diverge somewhat from the line, showing that there may be some non-normality in the data. However, since there is a very large number of students, we can apply CLT. So, we can assume that the data for the model meets the normality condition.

Credit to <https://thestatsgeek.com/2013/08/07/assumptions-for-linear-regression/> for reference.

Credit to <https://www.statology.org/multiple-linear-regression-assumptions/> and to <https://www.r-bloggers.com/2020/10/residual-plots-and-assumption-checking/#> for reference.

## Running the test:

Here is the original reduced model:

mlr2 = lm(score ~ gender + ethnicity + fcollege + mcollege + home + urban + unemp + wage + tuition + education + log\_distance, CollegeDistance2)  
summary(mlr2)

##   
## Call:  
## lm(formula = score ~ gender + ethnicity + fcollege + mcollege +   
## home + urban + unemp + wage + tuition + education + log\_distance,   
## data = CollegeDistance2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.6062 -5.0688 0.0456 5.0061 24.4732   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15.61445 1.11819 13.964 < 2e-16 \*\*\*  
## gendermale 1.08995 0.20766 5.249 1.60e-07 \*\*\*  
## ethnicityhispanic 2.76666 0.36237 7.635 2.72e-14 \*\*\*  
## ethnicityother 6.42706 0.29995 21.427 < 2e-16 \*\*\*  
## fcollegeyes 1.22494 0.29236 4.190 2.84e-05 \*\*\*  
## mcollegeyes 1.01673 0.33504 3.035 0.00242 \*\*   
## homeyes 0.70387 0.27388 2.570 0.01020 \*   
## urbanyes -0.57220 0.26660 -2.146 0.03190 \*   
## unemp -0.06936 0.04056 -1.710 0.08734 .   
## wage 0.20834 0.08436 2.470 0.01356 \*   
## tuition 1.66504 0.34179 4.872 1.14e-06 \*\*\*  
## education 1.91963 0.06102 31.457 < 2e-16 \*\*\*  
## log\_distance -0.43906 0.10584 -4.148 3.41e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.09 on 4726 degrees of freedom  
## Multiple R-squared: 0.3379, Adjusted R-squared: 0.3362   
## F-statistic: 201 on 12 and 4726 DF, p-value: < 2.2e-16

All of the variables are significant except for unemployment. I will make a new reduced model that does not contain unemployment and test it against the original reduced model.

mlr3 = lm(score ~ gender + ethnicity + fcollege + mcollege + home + urban + wage + tuition + education + log\_distance, CollegeDistance2)  
summary(mlr3)

##   
## Call:  
## lm(formula = score ~ gender + ethnicity + fcollege + mcollege +   
## home + urban + wage + tuition + education + log\_distance,   
## data = CollegeDistance2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.836 -5.065 0.049 5.005 24.452   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15.52174 1.11710 13.895 < 2e-16 \*\*\*  
## gendermale 1.10367 0.20754 5.318 1.10e-07 \*\*\*  
## ethnicityhispanic 2.69637 0.36011 7.488 8.32e-14 \*\*\*  
## ethnicityother 6.43518 0.29998 21.452 < 2e-16 \*\*\*  
## fcollegeyes 1.25135 0.29202 4.285 1.86e-05 \*\*\*  
## mcollegeyes 1.04109 0.33480 3.110 0.00188 \*\*   
## homeyes 0.70460 0.27394 2.572 0.01014 \*   
## urbanyes -0.57752 0.26664 -2.166 0.03036 \*   
## wage 0.17539 0.08215 2.135 0.03281 \*   
## tuition 1.56364 0.33668 4.644 3.50e-06 \*\*\*  
## education 1.91660 0.06101 31.414 < 2e-16 \*\*\*  
## log\_distance -0.47500 0.10375 -4.578 4.81e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.091 on 4727 degrees of freedom  
## Multiple R-squared: 0.3375, Adjusted R-squared: 0.336   
## F-statistic: 218.9 on 11 and 4727 DF, p-value: < 2.2e-16

Now I will run the ANOVA test to make sure that the model without employment has not lost much accuracy.

anova(mlr2)

## Analysis of Variance Table  
##   
## Response: score  
## Df Sum Sq Mean Sq F value Pr(>F)   
## gender 1 2306 2306 45.8769 1.413e-11 \*\*\*  
## ethnicity 2 47174 23587 469.2800 < 2.2e-16 \*\*\*  
## fcollege 1 13090 13090 260.4316 < 2.2e-16 \*\*\*  
## mcollege 1 2647 2647 52.6654 4.600e-13 \*\*\*  
## home 1 1061 1061 21.1088 4.452e-06 \*\*\*  
## urban 1 0 0 0.0094 0.9226   
## unemp 1 0 0 0.0069 0.9337   
## wage 1 835 835 16.6105 4.665e-05 \*\*\*  
## tuition 1 2063 2063 41.0369 1.639e-10 \*\*\*  
## education 1 51195 51195 1018.5499 < 2.2e-16 \*\*\*  
## log\_distance 1 865 865 17.2087 3.408e-05 \*\*\*  
## Residuals 4726 237540 50   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(mlr3)

## Analysis of Variance Table  
##   
## Response: score  
## Df Sum Sq Mean Sq F value Pr(>F)   
## gender 1 2306 2306 45.8582 1.427e-11 \*\*\*  
## ethnicity 2 47174 23587 469.0891 < 2.2e-16 \*\*\*  
## fcollege 1 13090 13090 260.3257 < 2.2e-16 \*\*\*  
## mcollege 1 2647 2647 52.6440 4.650e-13 \*\*\*  
## home 1 1061 1061 21.1002 4.472e-06 \*\*\*  
## urban 1 0 0 0.0094 0.9226103   
## wage 1 761 761 15.1375 0.0001013 \*\*\*  
## tuition 1 1899 1899 37.7567 8.672e-10 \*\*\*  
## education 1 51097 51097 1016.1920 < 2.2e-16 \*\*\*  
## log\_distance 1 1054 1054 20.9591 4.812e-06 \*\*\*  
## Residuals 4727 237687 50   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Assigning the ANOVA output to the different components of the F test:

SSE1 <- 237540  
SSE2 <- 237687   
ExtraSS2 <- SSE2 - SSE1  
n <- 4738  
p2 <- 12  
q2 <- 11

Calculating the F statistic:

F\_stat2 <- (ExtraSS2 / (p2-q2)) / (SSE1 / (n-p2))  
F\_stat2

## [1] 2.924653

Calculating the p-value:

anova(mlr2, mlr3)

## Analysis of Variance Table  
##   
## Model 1: score ~ gender + ethnicity + fcollege + mcollege + home + urban +   
## unemp + wage + tuition + education + log\_distance  
## Model 2: score ~ gender + ethnicity + fcollege + mcollege + home + urban +   
## wage + tuition + education + log\_distance  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 4726 237540   
## 2 4727 237687 -1 -146.97 2.924 0.08734 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Since the p-value (0.08734) is greater than 0.05, we fail to reject the claim that the original reduced model is much more accurate than the model without income. So, we can go ahead and drop income from our model.

Here is our final model and summary statistics:

summary(mlr3)

##   
## Call:  
## lm(formula = score ~ gender + ethnicity + fcollege + mcollege +   
## home + urban + wage + tuition + education + log\_distance,   
## data = CollegeDistance2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.836 -5.065 0.049 5.005 24.452   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15.52174 1.11710 13.895 < 2e-16 \*\*\*  
## gendermale 1.10367 0.20754 5.318 1.10e-07 \*\*\*  
## ethnicityhispanic 2.69637 0.36011 7.488 8.32e-14 \*\*\*  
## ethnicityother 6.43518 0.29998 21.452 < 2e-16 \*\*\*  
## fcollegeyes 1.25135 0.29202 4.285 1.86e-05 \*\*\*  
## mcollegeyes 1.04109 0.33480 3.110 0.00188 \*\*   
## homeyes 0.70460 0.27394 2.572 0.01014 \*   
## urbanyes -0.57752 0.26664 -2.166 0.03036 \*   
## wage 0.17539 0.08215 2.135 0.03281 \*   
## tuition 1.56364 0.33668 4.644 3.50e-06 \*\*\*  
## education 1.91660 0.06101 31.414 < 2e-16 \*\*\*  
## log\_distance -0.47500 0.10375 -4.578 4.81e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.091 on 4727 degrees of freedom  
## Multiple R-squared: 0.3375, Adjusted R-squared: 0.336   
## F-statistic: 218.9 on 11 and 4727 DF, p-value: < 2.2e-16

As you can see every variable is now statistically significant!

## Conclusion and Interpretation:

The MLR test gave important information about how various factors correlate with student success.

Concerningly, the variable with the largest coefficient was ethnicity: other, with a coefficient of 6.43518, and the variable with the second largest coefficient was ethnicity: hispanic, with a coefficient of 2.69637. Strangely, the dataset only included three categories for ethnicity: African American, Hispanic, and Other.

unique(CollegeDistance2$ethnicity)

## [1] "other" "afam" "hispanic"

Presumably many people in the “Other” category would have been classified as White if the researchers had collected that data ([https://www.childstats.gov/americaschildren/tables/pop3.asp).](https://www.childstats.gov/americaschildren/tables/pop3.asp).) The coefficient of 6.43518 indicates Other (likely mostly White) students tend to do far better on the tests than African American students when all other variables are constant. The 2.69637 coefficient indicates that Hispanic students also tend to do much better on the tests than African American students when all other variables are held constant, though the difference is less than for Other/White students.

Unsurprisingly, the variable with the next largest coefficient was education, with a coefficient of 1.91660. This indicates when adjusting for all other variables, students with more years of education have higher test scores on average. Tuition had a coefficient of 1.56364, indicating that students who attended universities with higher tuition tended to have higher test scores, on average and when adjusting for other variables.

A student’s mother having attended college contributed a coefficient of 1.25135 to the model, while a student’s father having attended college contributed a slightly lower coefficient of 1.04109. Another concerning finding of our model was that male students, with all other variables held constant, had higher test scores on average, with a coefficient of 1.10367.

Other variables with significant p values included whether the student’s family owned their home (positive correlation, coefficient of 0.70460), whether the student lived in an urban area (negative correlation, coefficient of -0.57752), the log of their distance from the nearest college (negative correlation, coefficient of -0.47500) and the minimum wage in the students home state (positive correlation, coefficient of 0.17539).

The model also has an extremely high F statistic of 201, indicating that it is extremely likely that there is a relationship between the combination of variables and student test scores.

However, the model doesn’t explain very much of the variability in the data. It has a low Multiple R-squared value of 0.338. In other words, it doesn’t capture many of the variables that are related to student success. However, it can still be useful in examining the relationships that it does cover.

Our data was collected in 1980, and so is of limited use in observing current education trends; however it can provide a snapshot into some 1980s education dynamics that may continue to affect aspects of our society. While we can’t definitively say without using causal inference, our model’s very large coefficients for race may indicate deep racial inequities in 1980s education system and society; there are still striking racial disparities in education access to this day (<https://home.treasury.gov/news/featured-stories/post-5-racial-differences-in-educational-experiences-and-attainment>). This might indicate that the biggest thing that people can do to boost student achievement is to tackle racial disparities in education and in general. Years of education, and tuition rates at student’s universities, also appear to be strongly linked to student success. Additionally, while we would need to use causal inference to make statistical claims, our models coefficient for gender might indicate gender inequities in the 1980’s education system and society. These dynamics, and possible interventions, should be further explored through causal inference.