MAS_Thesis_RD_GP_RDD

Devin Reeh

2025-07-02

```
## Skipping install of 'gpss' from a github remote, the SHA1 (5d7c08ff) has not changed since last insta
##   Use `force = TRUE` to force installation

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
cupc_path <- "~/personal-projects/ucla-masds-thesis/data/cupc2122-k12.xlsx"
upc_data <- read_excel(cupc_path, sheet="School-Level CALPADS UPC Data", col_names = T, skip = 1)
colnames(upc_data) <- c(
  "academic_year",
  "county_code",
  "district_code",
  "school_code",
  "county_name",
  "district_name",
  "school_name",
  "district_type",
  "school_type",
  "educational_option_type",
  "nslp_provision_status",
  "charter_school_yn",
  "charter_number",
  "charter_funding_type",
  "irc",
  "low_grade",
  "high_grade",
  "total_enrollment",
  "frpm_program",
  "foster",
  "tribal_foster_youth",
  "homeless",
  "migrant_program",
  "direct_certification",
  "undup_frpm_eligible_count",
  "english_learner",
```

```
  "calpads_upc",
  "calpads_certified_yn"
)
glimpse(upc_data)
```

```
## Rows: 10,549
## Columns: 28
## $ academic_year             <chr> "2021-2022", "2021-2022", "2021-2022", "2021~
## $ county_code               <chr> "01", "01", "01", "01", "01", "01", "01", "0~
## $ district_code             <chr> "10017", "10017", "10017", "10017", "10017",~
## $ school_code               <chr> "0130419", "0130401", "0130625", "0137448", ~
## $ county_name               <chr> "Alameda", "Alameda", "Alameda", "Alameda", ~
## $ district_name             <chr> "Alameda County Office of Education", "Alame~
## $ school_name               <chr> "Alameda County Community", "Alameda County ~
## $ district_type             <chr> "County Office of Education (COE)", "County ~
## $ school_type               <chr> "County Community", "Juvenile Court Schools"~
## $ educational_option_type   <chr> "County Community School", "Juvenile Court S~
## $ nslp_provision_status     <chr> "N/A", "N/A", "N/A", "N/A", "N/A", "N/A", "N~
## $ charter_school_yn         <chr> "N", "N", "Y", "Y", "Y", "Y", "Y", "Y", "Y",~
## $ charter_number            <chr> "N/A", "N/A", "0398", "1908", "1284", "1881"~
## $ charter_funding_type      <chr> "N/A", "N/A", "Directly funded", "Directly f~
## $ irc                       <chr> "N", "N", "Y", "Y", "Y", "Y", "Y", "Y", "Y",~
## $ low_grade                 <chr> "K", "K", "9", "6", "K", "K", "P", "4", "K",~
## $ high_grade                <chr> "12", "12", "12", "8", "8", "12", "5", "12",~
## $ total_enrollment          <dbl> 57, 64, 150, 164, 202, 514, 516, 411, 141, 4~
## $ frpm_program              <dbl> 24, 0, 107, 147, 163, 101, 269, 263, 92, 234~
## $ foster                    <dbl> 2, 7, 0, 1, 0, 0, 3, 3, 0, 0, 0, 9, 0, 0, 0,~
## $ tribal_foster_youth       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ homeless                  <dbl> 0, 16, 8, 1, 1, 6, 2, 0, 0, 2, 0, 2, 2, 1, 3~
## $ migrant_program           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ direct_certification      <dbl> 34, 27, 63, 84, 106, 93, 255, 156, 35, 203, ~
## $ undup_frpm_eligible_count <dbl> 47, 41, 134, 148, 177, 124, 360, 288, 93, 30~
## $ english_learner           <dbl> 18, 11, 93, 51, 100, 11, 268, 77, 45, 247, 4~
## $ calpads_upc               <dbl> 49, 64, 142, 150, 185, 130, 421, 301, 102, 3~
## $ calpads_certified_yn      <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y",~
```

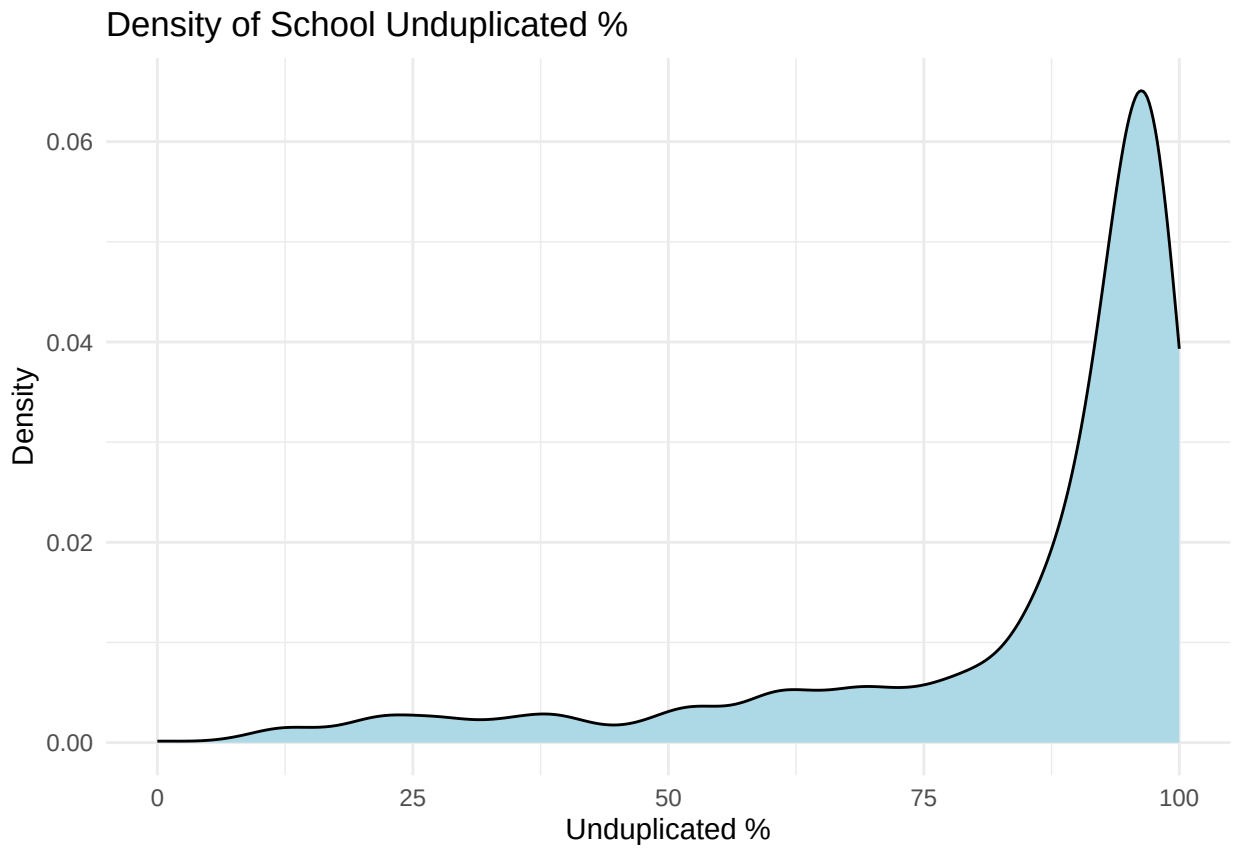
```
lausd_cupc <- upc_data %>%
  filter(
    academic_year == "2021-2022",
    district_code == "64733"     # LAUSD county-district prefix
  ) %>%
  mutate(
    undup_pct = 100 * (as.numeric(calpads_upc) / as.numeric(total_enrollment))
  ) %>%
    select(
      school_code,
      undup_pct
    )

summary(lausd_cupc$undup_pct)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   80.21   93.49   83.97   96.88  100.00       1
```

```r
ggplot(lausd_cupc, aes(x = undup_pct)) +
  geom_density(fill = "lightblue") +
  labs(title = "Density of School Unduplicated %", x = "Unduplicated %", y = "Density") +
  theme_minimal()
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_density()`).
```



Data Summmary

```r
glimpse(df_clean)
```

```
## Rows: 1,001
## Columns: 36
## $ school_name.x                          <chr> "ABRAHAM LINCOLN SENIOR~
## $ school_code                            <chr> "1935121", "0120097", "~
## $ frpm_rate                              <dbl> 0.9525166, 0.8165138, 0~
## $ schoolcode                             <chr> "1935121", "0120097", "~
## $ Academic.Year                          <chr> "2021-22", "2021-22", "~
## $ Aggregate.Level                        <chr> "S", "S", "S", "S", "S"~
## $ County.Code                            <int> 19, 19, 19, 19, 19, 19,~
```

```
## $ District.Code                                        <int> 64733, 64733, 64733, 64~
## $ County.Name                                          <chr> "Los Angeles", "Los Ang~
## $ district_name                                        <chr> "Los Angeles Unified", ~
## $ school_name.y                                        <chr> "Abraham Lincoln Senior~
## $ Charter.School                                       <chr> "No ", "Yes", "No ", "N~
## $ DASS                                                 <chr> "No ", "No ", "No ", "N~
## $ Reporting.Category                                   <chr> "TA", "TA", "TA", "TA",~
## $ ChronicAbsenteeismEligibleCumulativeEnrollment       <dbl> 1127, 471, 528, 300, 11~
## $ ChronicAbsenteeismCount                              <dbl> 369, 224, 187, 56, 21, ~
## $ chronic_absenteeism                                  <dbl> 32.7, 47.6, 35.4, 18.7,~
## $ treated                                              <dbl> 1, 1, 0, 0, 1, 1, 1, 1,~
## $ running                                              <dbl> 0.202516619, 0.06651376~
## $ total_enroll                                         <dbl> 1090, 439, 503, 290, 10~
## $ pct_hispanic                                         <dbl> 75.87156, 98.63326, 55.~
## $ pct_black                                            <dbl> 2.1100917, 0.9111617, 1~
## $ pct_white                                            <dbl> 1.0091743, 0.4555809, 2~
## $ pct_asian                                            <dbl> 20.27522936, 0.00000000~
## $ pct_two_or_more                                      <dbl> 0.2752294, 0.0000000, 1~
## $ pct_other                                            <dbl> 0.45871560, 0.00000000,~
## $ undup_pct                                            <dbl> 95.50459, 86.78815, 66.~
## $ SchoolYear                                           <chr> "2021-2022", NA, NA, "2~
## $ BeforeSchoolGrantProgram                             <int> 0, NA, NA, 0, NA, NA, N~
## $ AfterSchoolGeneralFundedPrograms                     <int> 0, NA, NA, 1, NA, NA, N~
## $ btb_participation                                    <dbl> 0, NA, NA, 1, NA, NA, N~
## $ frpm_percent                                         <dbl> 95.25166, 81.65138, 65.~
## $ food_eligible                                        <dbl> 1, 1, 1, 0, 1, 1, 1, 1,~
## $ btb                                                  <int> 0, 0, 0, 1, 0, 0, 0, 0,~
## $ county                                               <fct> 19, 19, 19, 19, 19, 19,~
## $ district                                             <fct> 64733, 64733, 64733, 64~
```

RDD chronic_abseentism ~ FRPM % - 40 cut off

```
#########################################################
# GP RDD
# chronic_abseentism ~ FRPM % - 40 cut off
#########################################################
rdd_res_absenteeism_frpm_40_cutoff <- gp_rdd(
  df_clean$frpm_percent,
  df_clean$chronic_absenteeism,
  40
)
rdd_res_absenteeism_frpm_40_cutoff$tau     # estimated effect
```

```
## [1] 3.475961
```

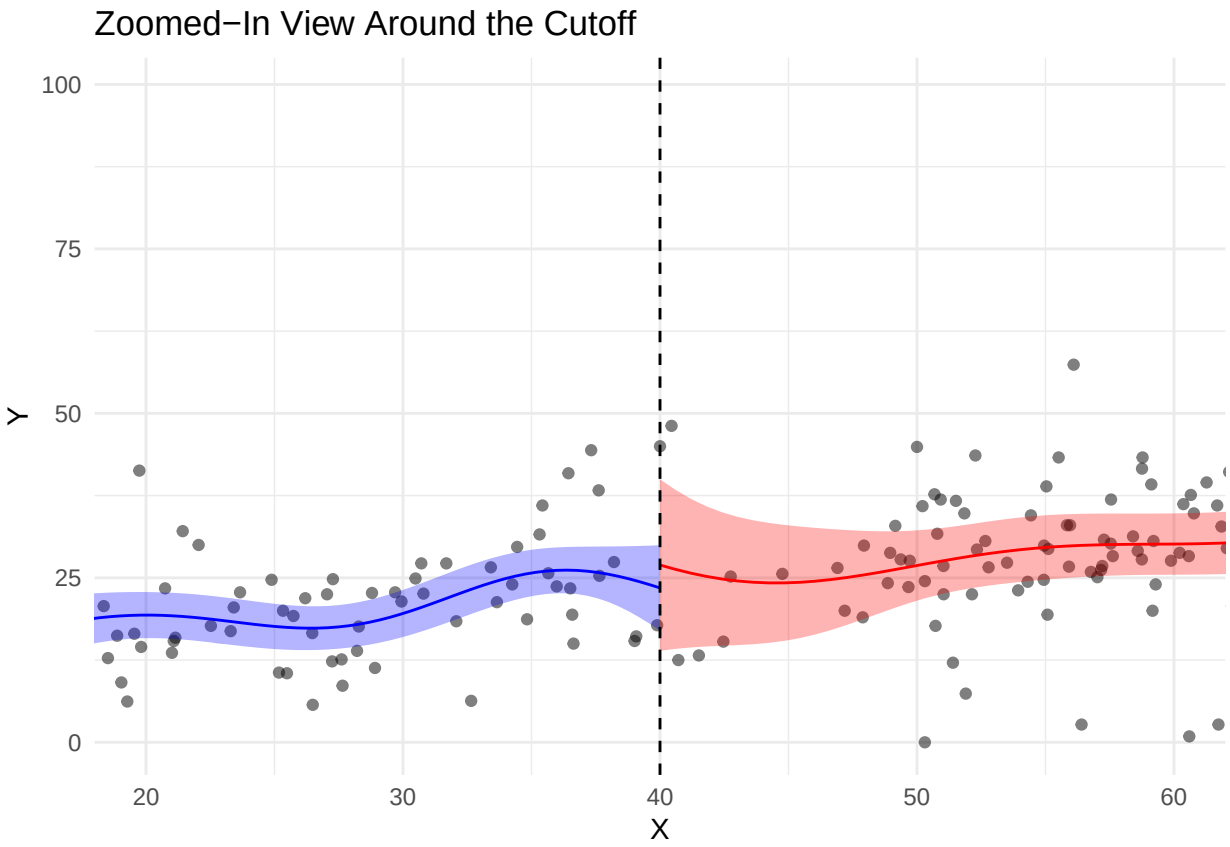
```
rdd_res_absenteeism_frpm_40_cutoff$se      # standard error
```

```
## [1] 7.428552
```

```
rdd_res_absenteeism_frpm_40_cutoff$ci      # confidence interval
```

```
##     lower     upper
## -11.08373  18.03565
```

```
rdd_result_plot_1 <- gp_rdd_plot(rdd_res_absenteeism_frpm_40_cutoff) +
  geom_vline(xintercept = 40, linetype = "dashed") +
  coord_cartesian(xlim = c(20, 60)) +
  labs(title = "Zoomed-In View Around the Cutoff")
print(rdd_result_plot_1)
```



GP RDD - chronic_abseentism ~ FRPM % - 75 cut off

```
##########################################################
# GP RDD
# chronic_abseentism ~ FRPM % - 75 cut off
##########################################################
# Example using formula interface:
rdd_res_absenteeism_frpm_75_cutoff <- gp_rdd(
  df_clean$frpm_percent,
  df_clean$chronic_absenteeism,
  75
```

```
)
rdd_res_absenteeism_frpm_75_cutoff$tau      # estimated effect
```

```
## [1] -0.07402481
```

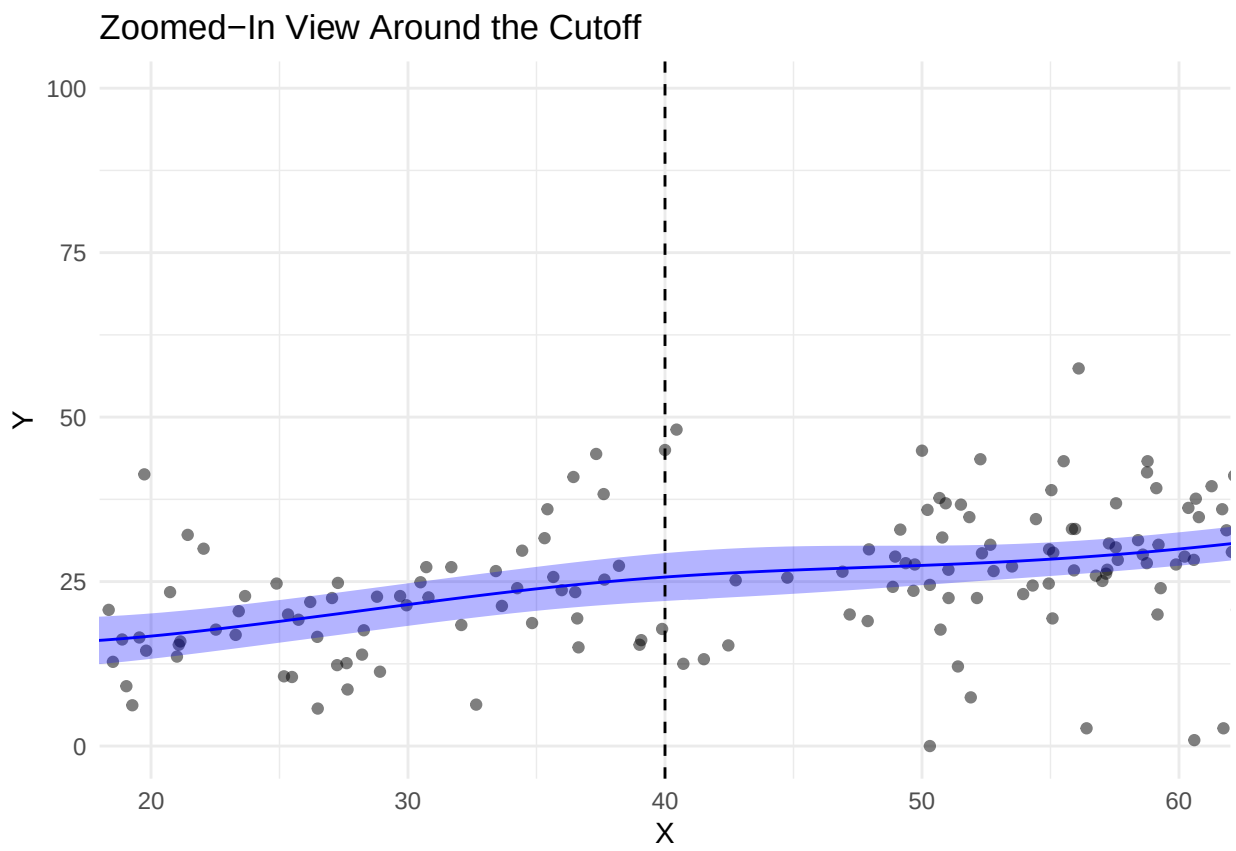
```
rdd_res_absenteeism_frpm_75_cutoff$se       # standard error
```

```
## [1] 6.059793
```

```
rdd_res_absenteeism_frpm_75_cutoff$ci       # confidence interval
```

```
##     lower     upper
## -11.95100  11.80295
```

```
rdd_result_plot_2 <- gp_rdd_plot(rdd_res_absenteeism_frpm_75_cutoff) +
  geom_vline(xintercept = 40, linetype = "dashed") +
  coord_cartesian(xlim = c(20, 60)) +
  labs(title = "Zoomed-In View Around the Cutoff")
print(rdd_result_plot_2)
```



GP RDD BTB ~ FRPM % - 40 cut off

```
#########################################################
# GP RDD
# BTB ~ FRPM % - 40 cut off
#########################################################
rdd_res_absenteeism_BTB_40_cutoff <- gp_rdd(
  df_clean$frpm_percent,
  df_clean$btb,
  40
)
rdd_res_absenteeism_BTB_40_cutoff$tau     # estimated effect
```

```
## [1] -0.3867966
```

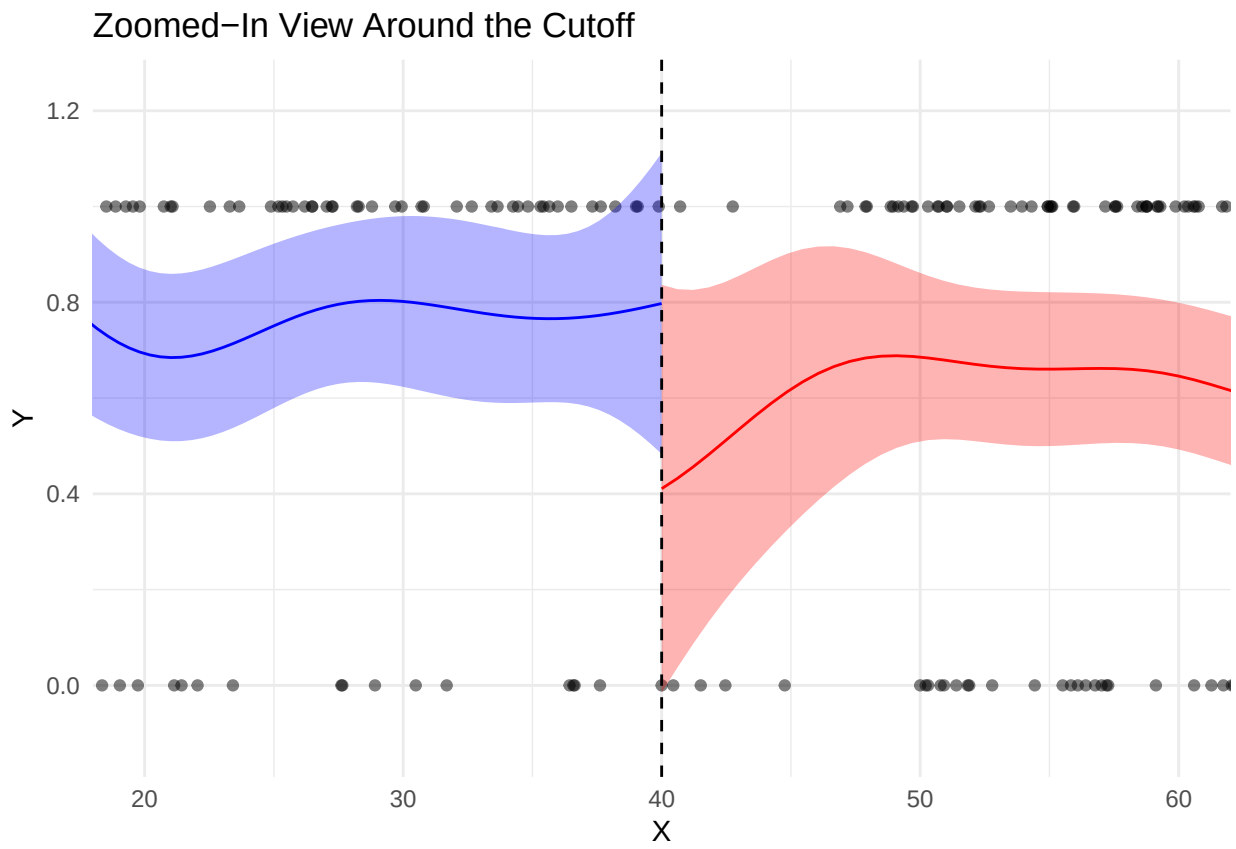
```
rdd_res_absenteeism_BTB_40_cutoff$se      # standard error
```

```
## [1] 0.2699848
```

```
rdd_res_absenteeism_BTB_40_cutoff$ci      # confidence interval
```

```
##      lower      upper
## -0.9159571  0.1423639
```

```
rdd_result_plot_3 <- gp_rdd_plot(rdd_res_absenteeism_BTB_40_cutoff) +
  geom_vline(xintercept = 40, linetype = "dashed") +
  coord_cartesian(xlim = c(20, 60)) +
  labs(title = "Zoomed-In View Around the Cutoff")
print(rdd_result_plot_3)
```



GP RDD BTB ~ FRPM % - 75 cut off

```
#########################################################
# GP RDD
# BTB ~ FRPM % - 75 cut off
#########################################################
rdd_res_absenteeism_BTB_75_cutoff <- gp_rdd(
  df_clean$frpm_percent,
  df_clean$btb,
  75
)
rdd_res_absenteeism_BTB_75_cutoff$tau     # estimated effect
```

```
## [1] -0.1589799
```

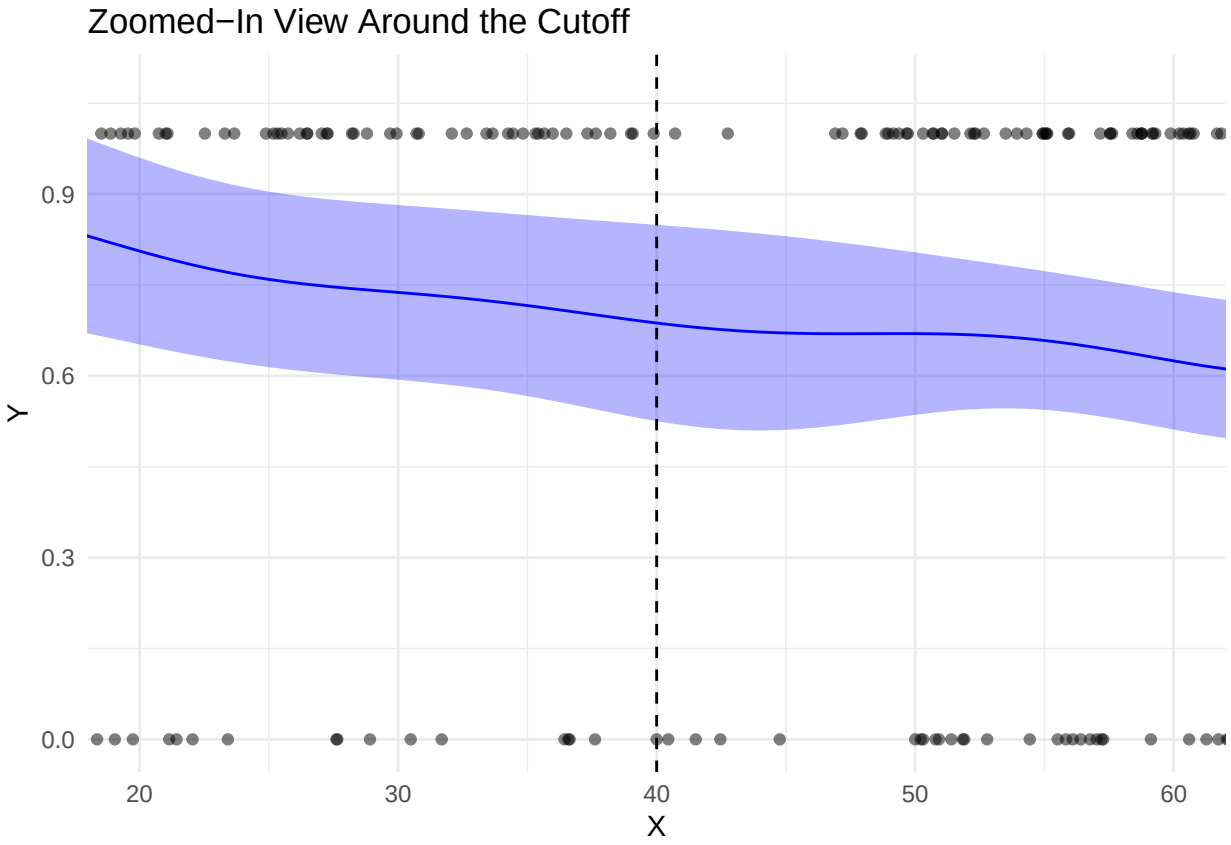
```
rdd_res_absenteeism_BTB_75_cutoff$se      # standard error
```

```
## [1] 0.2079251
```

```
rdd_res_absenteeism_BTB_75_cutoff$ci      # confidence interval
```

```
##      lower      upper
## -0.5665055  0.2485458
```

```
rdd_result_plot_4 <- gp_rdd_plot(rdd_res_absenteeism_BTB_75_cutoff) +
                     geom_vline(xintercept = 40, linetype = "dashed") +
                     coord_cartesian(xlim = c(20, 60)) +
                     labs(title = "Zoomed-In View Around the Cutoff")
print(rdd_result_plot_4)
```



Balance Tests with Xs

```
library(dplyr)
library(tidyr)

#
# Balance tests for continuous covariates
#
run_continuous_balance_tests <- function(df, cutoff = 0.75, bandwidth = 0.1) {
  df_band <- df %>%
    filter(frpm_rate >= (cutoff - bandwidth), frpm_rate <= (cutoff + bandwidth)) %>%
    mutate(
      treated = as.factor(if_else(frpm_rate >= cutoff, 1, 0))
    )
```

```r
  covariates <- c(
    "pct_hispanic", "pct_black", "pct_white", "pct_asian",
    "pct_two_or_more", "pct_other", "total_enroll"
  )

  results <- lapply(covariates, function(var) {
    if (nlevels(df_band$treated) == 2) {
      formula <- as.formula(paste(var, "~ treated"))
      test <- t.test(formula, data = df_band)
      tibble(
        variable = var,
        test_type = "t-test",
        p_value = test$p.value,
        mean_treated = mean(df_band[[var]][df_band$treated == "1"], na.rm = TRUE),
        mean_control = mean(df_band[[var]][df_band$treated == "0"], na.rm = TRUE)
      )
    } else {
      tibble(
        variable = var,
        test_type = "t-test",
        p_value = NA,
        mean_treated = NA,
        mean_control = NA
      )
    }
  })

  bind_rows(results)
}

#
#  Balance tests for binary categorical covariates
#
run_binary_balance_tests <- function(df, cutoff = 0.75, bandwidth = 0.1) {
  df_band <- df %>%
    filter(frpm_rate >= (cutoff - bandwidth), frpm_rate <= (cutoff + bandwidth)) %>%
    mutate(
      treated = as.factor(if_else(frpm_rate >= cutoff, 1, 0)),
      DASS = factor(trimws(DASS)),
      Charter = factor(trimws(Charter.School))
    )

  results <- list()

  for (var in c("DASS", "Charter")) {
    tab <- table(df_band[[var]], df_band$treated)

    if (nrow(tab) > 1 && ncol(tab) > 1) {
      test <- chisq.test(tab)
      result <- tibble(
        variable = var,
        test_type = "chi-squared",
        p_value = test$p.value,
```

```
        prop_treated = round(100 * prop.table(tab, 2)[, "1"], 1),
        prop_control = round(100 * prop.table(tab, 2)[, "0"], 1)
      )
    } else {
      result <- tibble(
        variable = var,
        test_type = "chi-squared",
        p_value = NA,
        prop_treated = NA,
        prop_control = NA
      )
    }

    results[[var]] <- result
  }

  bind_rows(results)
}

#
# Tests at both cutoffs
#

# For 75% cutoff
balance_75_cont <- run_continuous_balance_tests(df_clean, cutoff = 0.75)
balance_75_cat  <- run_binary_balance_tests(df_clean, cutoff = 0.75)
```

```
## Warning in chisq.test(tab): Chi-squared approximation may be incorrect
```

```
balance_75 <- bind_rows(balance_75_cont, balance_75_cat)

# For 40% cutoff
balance_40_cont <- run_continuous_balance_tests(df_clean, cutoff = 0.40)
balance_40_cat  <- run_binary_balance_tests(df_clean, cutoff = 0.40)
balance_40 <- bind_rows(balance_40_cont, balance_40_cat)

#
# Results
#
print("Balance tests around 75% cutoff:")
```

```
## [1] "Balance tests around 75% cutoff:"
```

```
print(balance_75)
```

```
## # A tibble: 11 x 7
##    variable        test_type   p_value mean_treated mean_control prop_treated
##    <chr>           <chr>         <dbl>        <dbl>        <dbl>        <dbl>
##  1 pct_hispanic    t-test       0.0377        73.0         65.8          NA
##  2 pct_black       t-test       0.884         10.4         10.0          NA
##  3 pct_white       t-test       0.0306         6.84        11.9          NA
##  4 pct_asian       t-test       0.602          3.07         3.51         NA
```

```
##  5 pct_two_or_more t-test       0.155          1.73         2.14         NA
##  6 pct_other       t-test       0.187          4.92         6.60         NA
##  7 total_enroll    t-test       0.998        644.         644.           NA
##  8 DASS            chi-squared  0.249         NA           NA            89.9
##  9 DASS            chi-squared  0.249         NA           NA            10.1
## 10 Charter         chi-squared  0.103         NA           NA            62.9
## 11 Charter         chi-squared  0.103         NA           NA            37.1
## # i 1 more variable: prop_control <dbl>
```

```
print("Balance tests around 40% cutoff:")
```

```
## [1] "Balance tests around 40% cutoff:"
```

```
print(balance_40)
```

```
## # A tibble: 10 x 7
##    variable        test_type     p_value mean_treated mean_control prop_treated
##    <chr>           <chr>           <dbl>        <dbl>        <dbl>        <dbl>
##  1 pct_hispanic    t-test       0.000371        44.9         31.4            NA
##  2 pct_black       t-test       0.946            7.47         7.30           NA
##  3 pct_white       t-test       0.0113          28.0         37.8            NA
##  4 pct_asian       t-test       0.529            7.86         9.55           NA
##  5 pct_two_or_more t-test       0.000137         4.76         8.12           NA
##  6 pct_other       t-test       0.409            7.00         5.80           NA
##  7 total_enroll    t-test       0.282         1007.         629.            NA
##  8 DASS            chi-squared NA               NA           NA              NA
##  9 Charter         chi-squared  1               NA           NA              50
## 10 Charter         chi-squared  1               NA           NA              50
## # i 1 more variable: prop_control <dbl>
```

```
# Optional: write to CSV
# write.csv(balance_75, "balance_test_75.csv", row.names = FALSE)
# write.csv(balance_40, "balance_test_40.csv", row.names = FALSE)
```

Covariate-Adjusted RD $Y \sim D \mid X$

```
df_clean %>%
  summarise(
    charter_n  = n_distinct(Charter.School),
    dass_n     = n_distinct(DASS),
    county_n   = n_distinct(County.Code),
    district_n = n_distinct(District.Code)
  )
```

```
## # A tibble: 1 x 4
##   charter_n dass_n county_n district_n
##       <int>  <int>    <int>      <int>
## 1         2      2        1          1
```

```r
# Convert categorical covariates to factors
df_clean <- df_clean %>%
  mutate(across(
    c(Charter.School, DASS),
    ~ as.factor(trimws(.x))
  ))

# GP-RDD with covariate adjustment (demographic variables that showed imbalance)
model_y_adj <- gpss(
  formula = chronic_absenteeism ~ treated + running +
    Charter.School + DASS +
    pct_hispanic + pct_white + pct_two_or_more,
  data = df_clean
)

# Summary output
summary(model_y_adj)
```

```
## Basic Model Information
## formula: chronic_absenteeism ~ treated + running + Charter.School + DASS +
##     pct_hispanic + pct_white + pct_two_or_more
## number of observations:  1001
## number of covariates:  7
## mixed data (containing a categorical variable?): FALSE
##
## Hyperparrameters
## b (bandwidth): 4.322543
## s2 (noise variance): 0.3
##
## Scaling information
## scaled: FALSE
##
## Usage Example
## e.g. fit <- gpss(Y~X)to extract SEs of fitted values: sqrt(diag(fit$post_cov_orig))
```