# The Effect of New Housing Projects on Expenditures in Two New York Municipalities

Andrew Bates

October 25, 2018

**Executive Summary**

In this paper we estimate future expenditures for two New York municipalities, Warwick and Monroe, based on various projected demographic and income-related factors. The desire for expenditure estimates was triggered by proposals for new housing construction. Warwick and Monroe wish to determine if they should increase funds to compensate for increased expenditures related to the housing projects. The estimates are obtained from a linear regression model chosen via a stepwise model selection procedure. The variables included in the model are wealth per person, population, percent intergovernmental funding, and growth rate. The estimated expenditures for Warwick are $261 per person for 2005 and $271 per person for 2025. The estimated expenditures for Monroe are $273 per person for both 2005 and 2025. Based on the model, we can infer that with increases in population and wealth per person, we can expect increases in expenditures. Conversely, decreases in expenditures are associated with increases in intergovernmental funding and growth rate

## 1 Introduction

Two New York towns, Warwick and Monroe, would like to estimate future expenditures triggered by new housing construction proposals. They are primarily interested in determining whether they should increase funds to compensate for increased expenditures related to the housing projects. To construct these estimates, Warwick and Monroe obtained data on expenditures along with various demographic and income-related variables from several New York municipalities.

The data used in this study consisted of 916 observations of seven variables. Each observation corresponds to a New York municipality for which each of the variables were collected. The response variable is expenditure per person. The covariates are as follows: wealth per person, population, revenue from state and federal grants, population density, mean income per person, and growth rate. Two observations were removed because they contained missing values.

For reasons of simplicity and interpretability, a linear regression model was chosen to estimate future expenditures. The correlation between population and population density was high (0.67) so to prevent this from leading to colinearity issues, only population was considered in the analysis. The variables in this data set (including the response) were heavily skewed, so log transformations were applied to each. To ensure a linear relationship between the predictors and the response, the data was subsetted to include only those observations for which the population was larger than $4,000$. The projected covariates for Warwick and Monroe fall within the range of this subsetted data so we felt this method was appropriate. In addition, this method was favored over a more complicated method like including a quadratic term on log population which would be more difficult to interpret. The regression model was constructed via stepwise regression using Akaike Information Criteria (AIC) as the selection criterion. This model was then used to estimate future expenditures.

# 2 Analysis

## 2.1 Exploratory Analysis

In this data set there were two measures of the size of a municipality: population and population density. Unsurprisingly, the correlation between these two measures was relatively high (0.67). In the interest of parsimony and to mitigate possible colinearity issues, we decided to consider only one of these variables to construct our model. Population density had a moderate correlation with mean income per person (0.49) whereas population had a comparably low correlation with income (0.29). To hedge against problems with colinearity, we decided to consider population for the model building process.

Upon an initial visual examination of the data, it was evident that transformations would be necessary. Each of the covariates, and the response, were skewed (see Appendix A.1) and some were heavily skewed. As this might pose problems with linearity, we performed transformations on the variables. To remain in line with our goal of having a comprehensible model, we favored log transformations over a more complex procedure. To that end, we performed log transformations on all variables in the data set. However, a straightforward application of the logarithm was not possible for one covariate. Growth rate contains some negative values as well as some zero values. To remedy this, we used the following pseudo-log transformation:

$$\text{p-}\log(\text{growth rate}) = \begin{cases} \log(\text{growth rate} + 0.15) & \text{if growth rate} > 0 \\ -\log(-\text{growth rate} + 0.15) & \text{if growth rate} \leq 0. \end{cases}$$

To ensure the linearity assumption of our model was satisfied, we examined the relationship between log expenditure and the log of each of the covariates. All covariates had an approximately linear relationship with log expenditure except for log population. Figure 1 is a plot of log expenditure vs. log population with the addition of a smoothing line. It shows what appears to be a quadratic relationship. However, notice that on either side of the vertical reference line (corresponding to a log population of 8.3), the relationship is approximately linear. So we were essentially faced with two choices. We could try to include the square of log population as an additional predictor or we could subset the data and have a roughly linear relationship between log expenditure and log population.

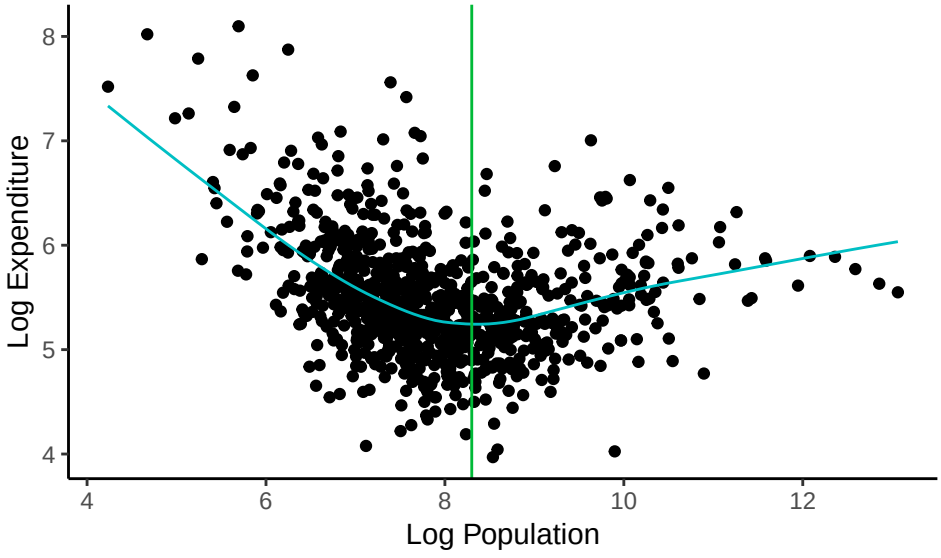


Figure 1: Plot of log expenditure vs. log population with LOWESS smooth line

We chose the latter method. As mentioned previously, we favored a simpler model for the interpretability that comes with it. Also, the projected covariates we would be predicting with fall within the range of the

covariates in the subsetted data. For these reasons, we subsetted the data to include only those observations that had a population greater than $4,000$ (log population above 8.3). This subsetted data set included 266 of the original observations. Figure 2 is a plot of log expenditure vs. log population for this subsetted data set along with a smoothing line. The relationship between the two variables is roughly linear. We do see minor deviation from linearity for high values of log population but overall the trend is approximately linear. Additionally, it is much more linear than the relationship we see in Figure 1.

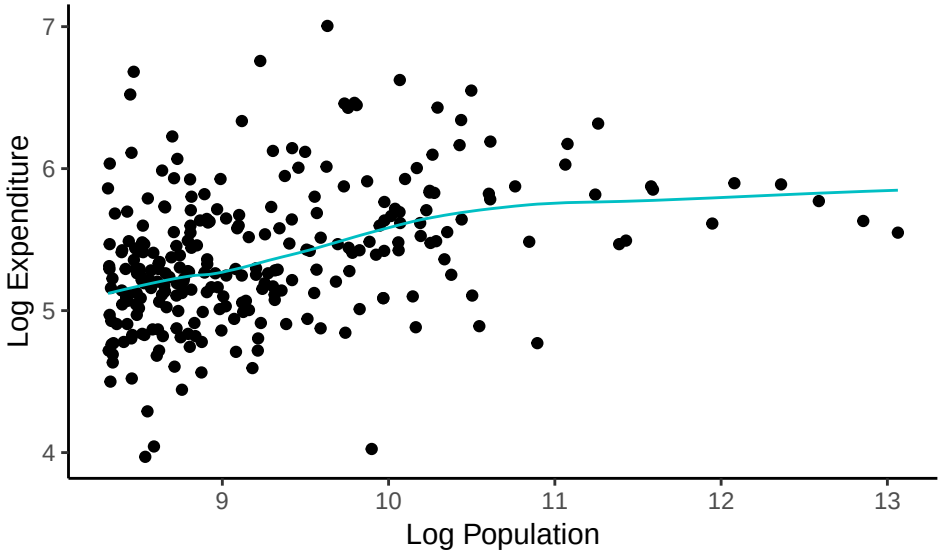


Figure 2: Plot of log expenditure vs. log population with LOWESS smooth line from subsetted data

After subsetting, we followed on with an additional exploratory analysis on the new data set. We plotted histograms of the untransformed and log transformed data as well as scatterplots of log expenditure vs. the log of each covariate. These plots are available in Appendix A.2. The results are similar to the unsubsetted data. All variables were skewed and thus suggested a log transformation. After log transformations, the response was reasonably close to normal and the covariates were symmetric. The scatterplots show a linear relationship between expenditures and each of the predictor variables.

## 2.2 Modeling and Diagnostics

After performing variable transformations and data subsetting, our regression model was ready to be constructed. We used backward stepwise regression to build the model. AIC was used as our model selection criterion. This was partly due to the ubiquity of AIC and partly because it penalizes larger models. This penalization aligns itself with our preference for simpler models. That being said, we did investigate the inclusion of squared log transformations for all the covariates ($\log^2$(income), etc.). However, there was not much of an improvement that could be gained by including these terms. The AIC for the model without the quadratic log terms was $-611$ while the AIC for the model with the extra terms was $-622$, only slightly better. Furthermore, the model with the quadratic variables included contained some quadratic terms without their corresponding linear components. We simply could not justify such a model. The final model used log expenditure as the response with the following covariates: log wealth, log population, log percent intergovernmental funding, and log growth rate.

Diagnostic plots from the fitted model are shown in Figure 3. The left figure is a plot of studentized residuals vs. fitted values. The horizontal line is a reference line at a residual of zero. There are no apparent patterns and the points are scattered roughly equally about the reference line. Nevertheless, there is a troublesome point with an unusually small value. This point corresponds to a municipality with a rather

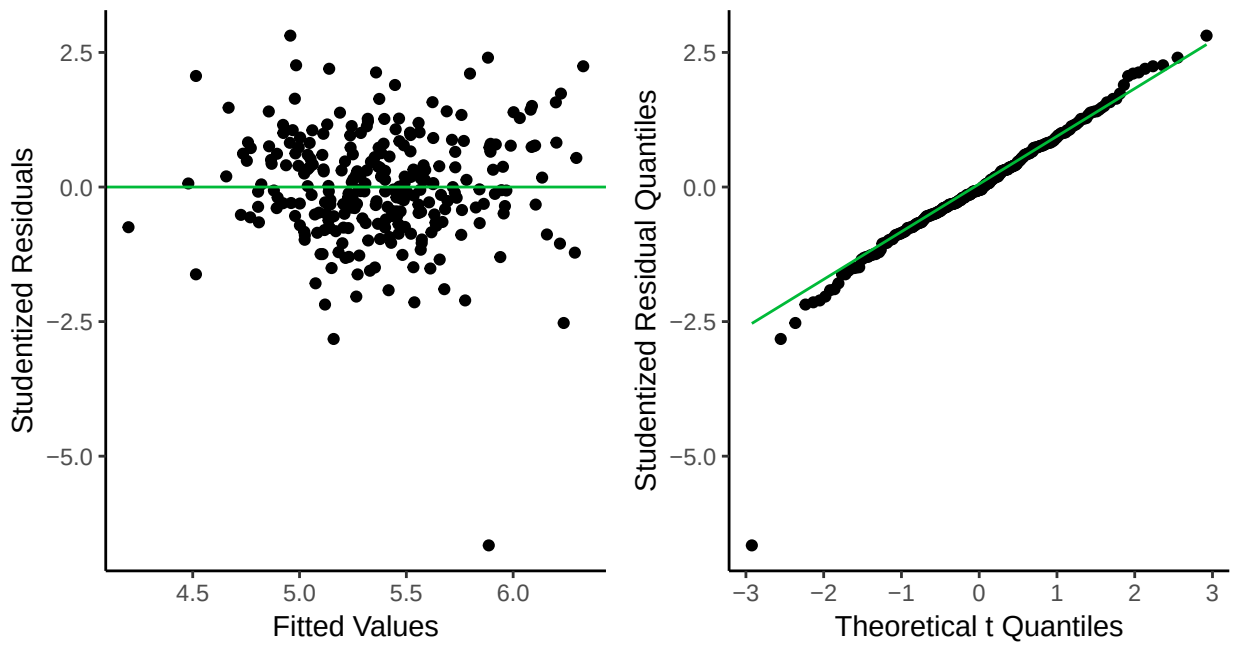


Figure 3: Diagnostic plots for regression model. The left plot shows studentized residuals vs. fitted values. The right plot shows a QQ plot of the studentized residuals vs. t-distribution quantiles.

large value of wealth per person. It's not entirely surprising that such a municipality exists as wealth related data often contains abnormally high observations. This data point also contains the smallest value of growth rate for all the municipalities so we can expect it to be pesky. Although this observation is a bit outside the range of most of the data, it is nonetheless a valid data point and this municipality may also wish to estimate their future expenditures at some point. For this reason, we decided to not remove this data point. The right-hand side of figure 3 provides an alternative view of the model residuals. Here we have plotted quantiles of the studentized residuals against quantiles of a $t$ distribution which is the distribution that they should theoretically follow. The green line is a reference line, indicating what the theoretical relationship should be. Aside from the point just discussed, the residuals follow their theoretical distribution quite well. As an additional diagnostic assessment, we computed variance inflation factors which were all less than 1.2. Considering our diagnostic plots and metrics as a whole, we were satisfied with this model.

A summary of the regression model can be found in Table 1. Since both the response and predictor variables were log transformed, the interpretation of the coefficients is different than the standard case when there are no transformations. Instead of interpreting a coefficient in terms of a unit increase in the respective dependent variable, we interpret it in terms of a percent increase. We illustrate this procedure for wealth here but the interpretation for the other variables is similar. Suppose wealth per person increases by 10%. Then, since the coefficient of wealth is 0.49, we take $1 + 10\%$ and raise it to the power of 0.49. This gives 1.047. Subtract one from this and multiply by 100 and we have 4.78. This tells us that if all other variables remain constant, a 10% increase in wealth per person corresponds to a roughly 5% increase in expenditure per person. This may seem odd in that one might expect wealthier people to require less money from their local government. But perhaps the increase in expenditures comes from providing additional community resources, infrastructure improvements, or increased funding for public schools. Increased expenditures on items such as these tend to be more common in wealthier areas compared to their less wealthy counterparts.

| | Estimate | SE | p-value | 95% CI |
|---|---|---|---|---|
| Intercept | 0.13 | 0.44 | 0.769 | ( -0.731 , 0.987 ) |
| Log Wealth | 0.49 | 0.04 | 0.000 | ( 0.423 , 0.563 ) |
| Log Population | 0.08 | 0.02 | 0.001 | ( 0.032 , 0.123 ) |
| Log % Intergov Funding | -0.28 | 0.04 | 0.000 | ( -0.358 , -0.198 ) |
| Log Growth Rate | -0.02 | 0.01 | 0.031 | ( -0.046 , -0.002 ) |

Table 1: Summary table for regressing log expenditure on log wealth, log population, log percent intergovernmental funding, and log growth rate.

### 2.3 Predictions

The projections for Warwick and Monroe in 2005 and 2025, along with expenditure predictions and prediction intervals, can be found in Table 2. The most obvious thing to note is the the estimated expenditure for Monroe is the same in 2005 and 2025. The increased expenditures from rising population and wealth from 2005 to 2025 is probably offset by the decrease in expenditure from rising intergovernmental funding rate. The net effect being that the predicted expenditures remain the same. Warwick on the other hand, is projected to see larger increases in population and wealth over the twenty year period than Monroe will. These large increases are not offset by the decrease in expenditure due to rising intergovernmental funding and growth rate which leads to an increase in expenditures over the same period of $10 per person. Although Warwick should expect to see their expenditures rise, by 2025 we expect to see roughly the same expenditure per person for both towns.

| | Population | Wealth | % Intergov Funding | Growth Rate | Estimate | 95 % PI |
|---|---|---|---|---|---|---|
| Warwick-2005 | 20442 | 85000 | 25.0 | 35.0 | 261 | ( 139 , 489 ) |
| Warwick-2025 | 31033 | 89000 | 26.0 | 40.0 | 271 | ( 144 , 509 ) |
| Monroe-2005 | 10496 | 58000 | 8.8 | 35.0 | 273 | ( 147 , 510 ) |
| Monroe-2025 | 13913 | 60000 | 10.0 | 35.0 | 273 | ( 147 , 510 ) |

Table 2: Projections for Warwick and Monroe along with predicted expenditures and prediction intervals.

## 3 Conclusion

In this paper we developed a linear regression model in order to estimate future expenditures for two New York municipalities, Warwick and Monroe. The model was constructed via a stepwise regression procedure with the final model containing the following covariates: population, wealth per person, percent intergovernmental funding, and growth rate. The model suggests that a rise in either population or growth rate is associated with a rise in expenditures. On the other hand, a rise in intergovernmental funding or growth rate is connected to a decrease in expenditures. The predicted expenditures for Warwick are $261 per person in 2005 and $271 per person for 2025. The estimated expenditures for Monroe are $273 per person for both 2005 and 2025. These predictions should be handled carefully. While they give an estimate of future expenditures, they should be seen as just that, estimates. They should be seen as more of a ballpark figure than a true representation of what future expenditures will be. The time between data collection and 2025 is over three decades and a lot can happen in that time. Moreover, the expenditure estimates obtained here are themselves based on estimates of the future demographic and income status of the two towns. Warwick and Monroe may want to consider updating their projections and model at a later date.

# Appendix A Supplementary Plots

## A.1 Histograms and Scatter Plots Before Subsetting

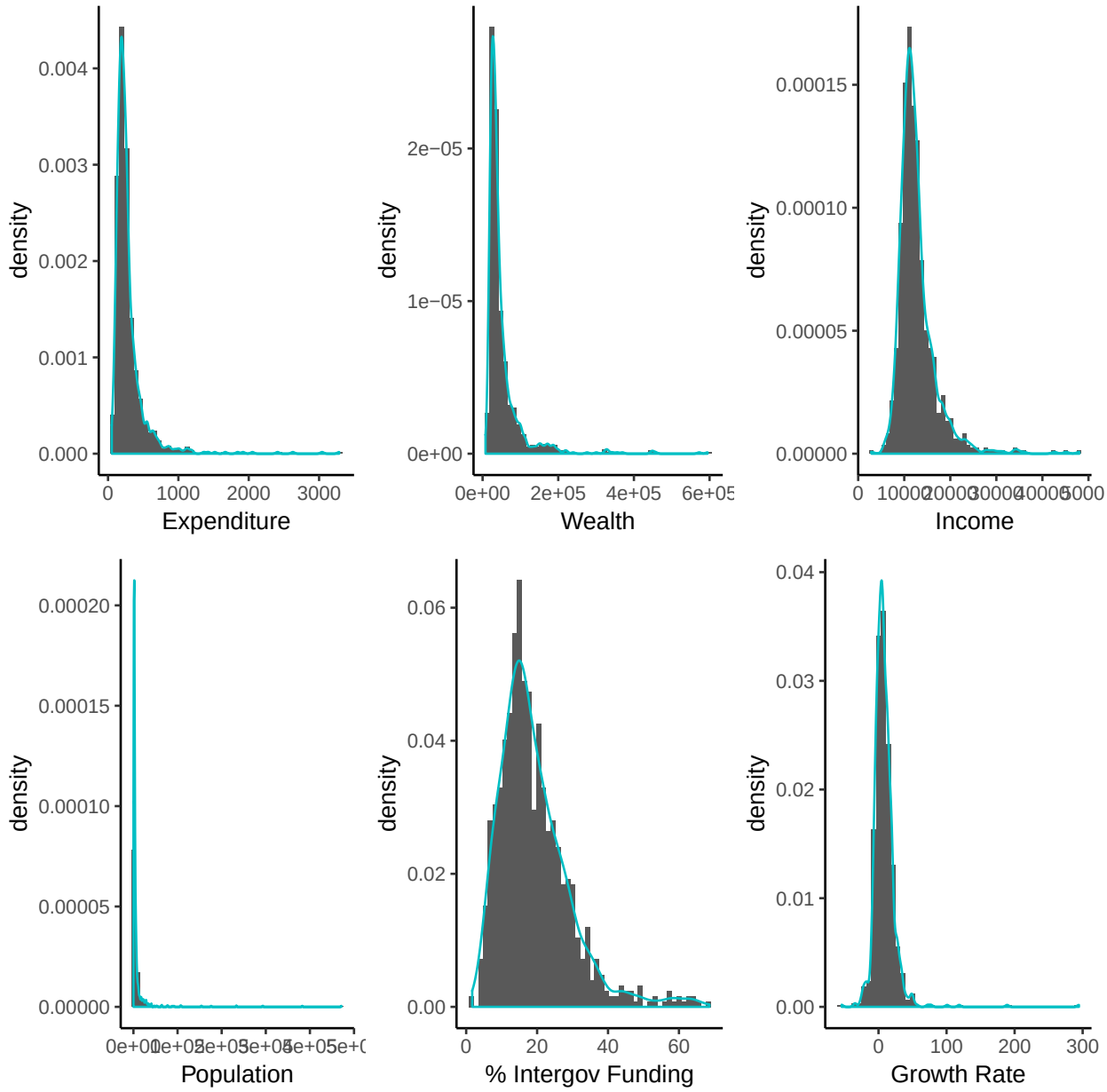


Figure 4: Histograms of all variables with densities overlayed.

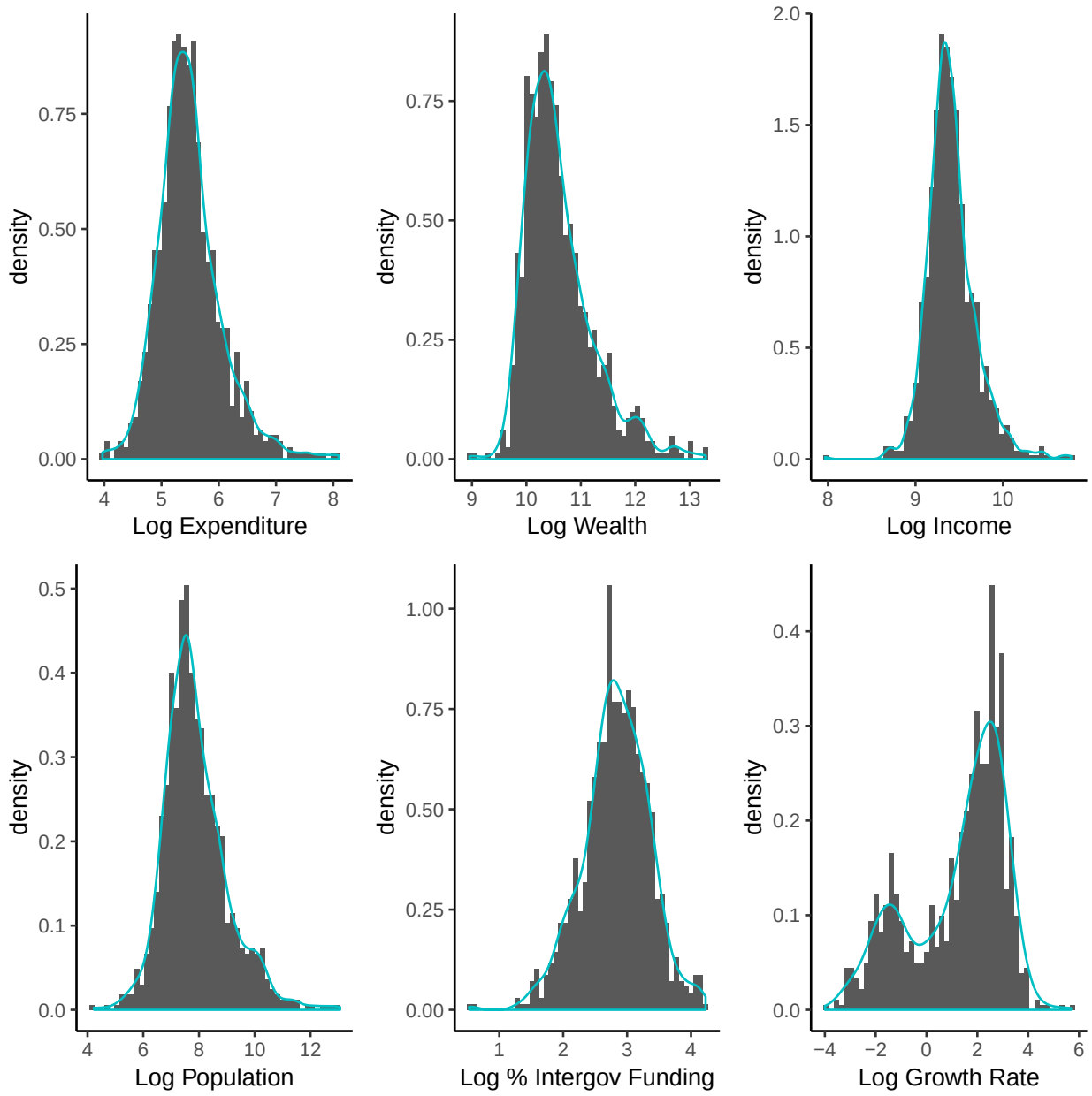


Figure 5: Histograms of log transformed variables with densities overlayed.

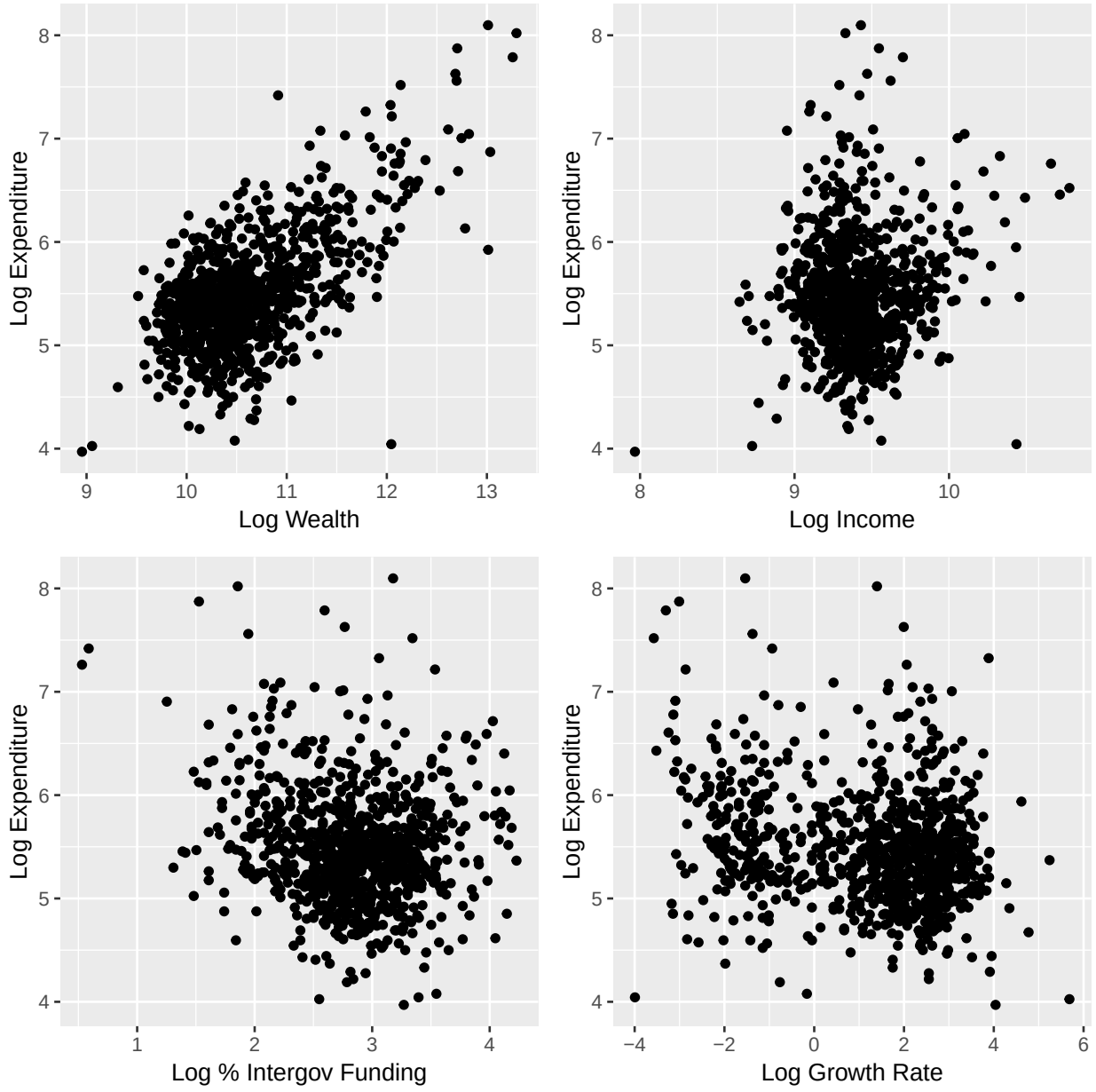


Figure 6: Scatter plots of log expenditure vs. log covariate for each covariate except population (see main text).

## A.2 Histograms and Scatter Plots After Subsetting

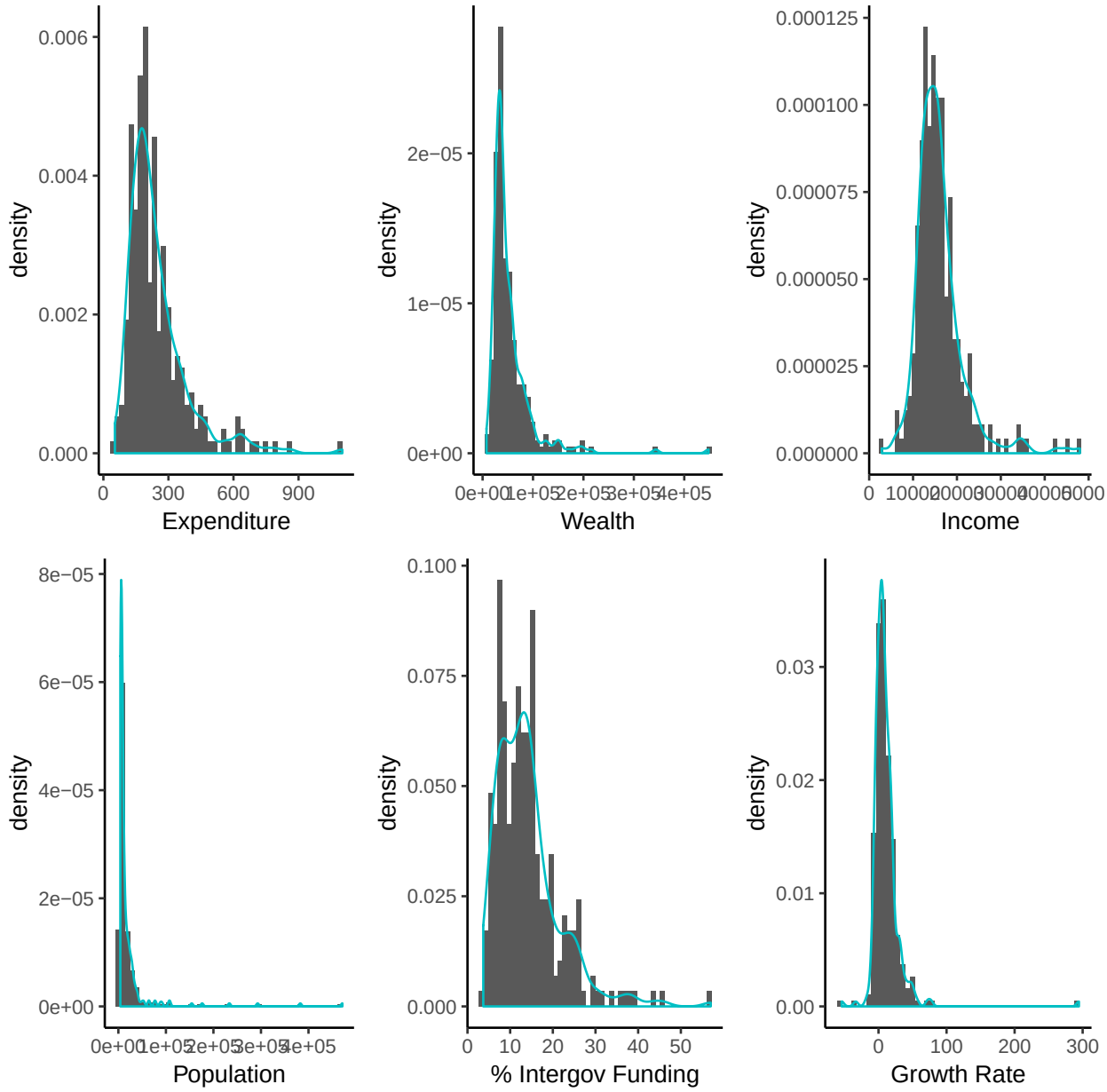


Figure 7: Histograms of all variables after subsetting with densities overlayed.

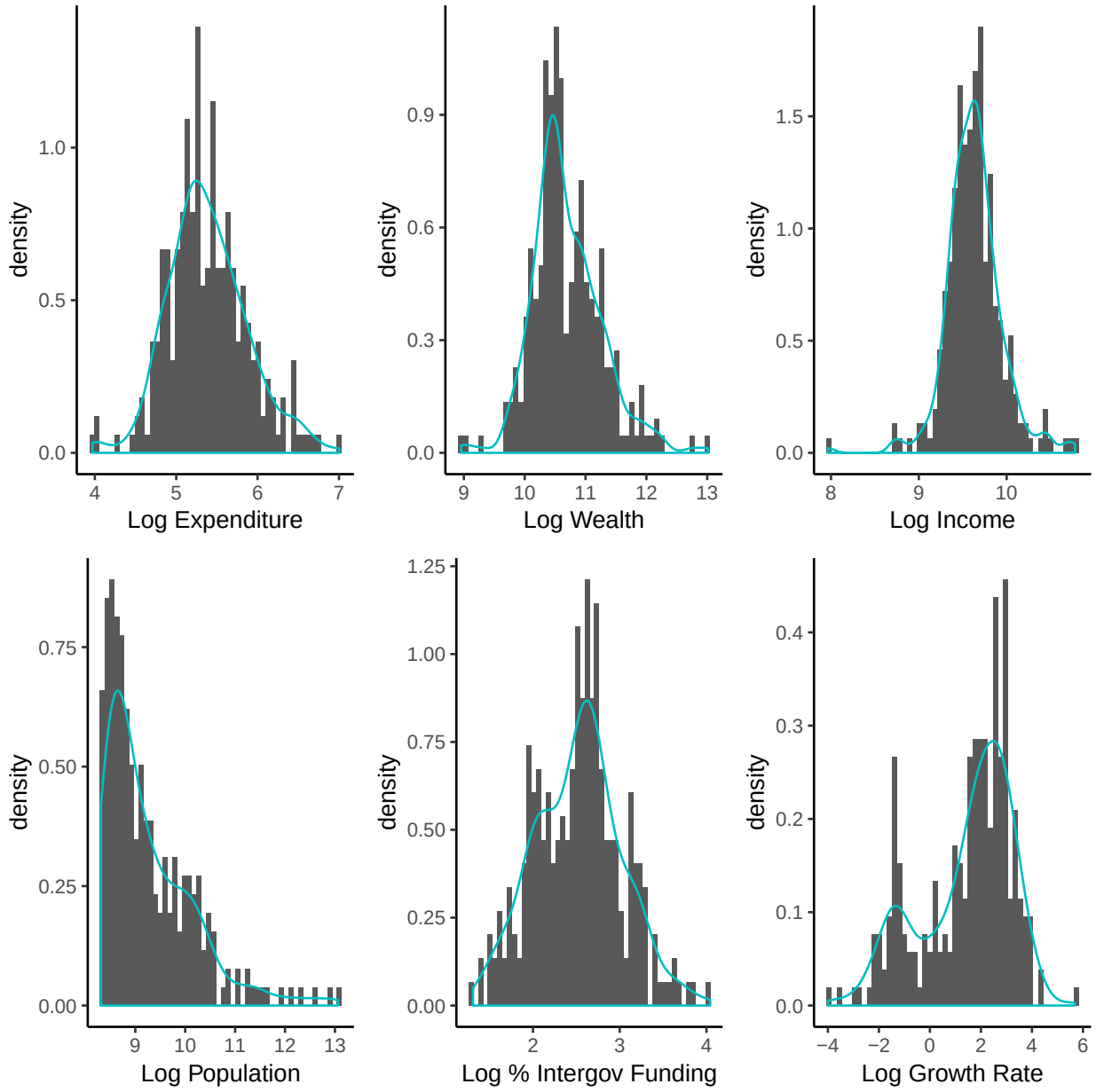


Figure 8: Histograms of log transformed variables after subsetting with densities overlayed.

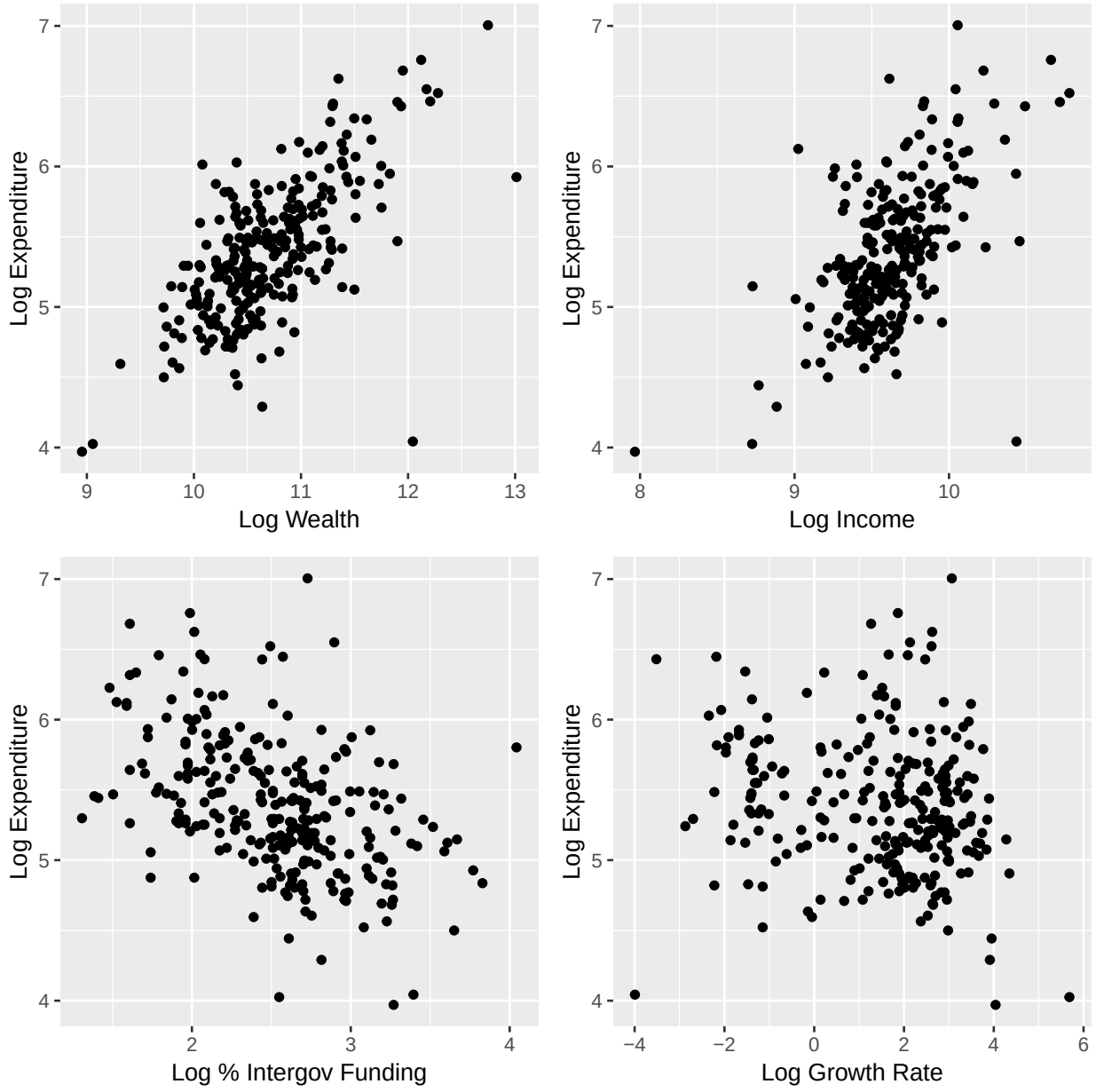


Figure 9: Scatter plots of log expenditure vs. log covariate for each covariate except population (see main text) after subsetting.

# Appendix B R Code

```
library(readr) # read in data
library(dplyr) # data manipulation
library(corrplot) # correlation plotting
library(ggplot2) # plotting
library(MASS) # stepwise regression
library(car)
library(broom) # tidy model output
library(here) # file path helper
library(xtable) # table formatting

ny <- read_csv(here("reports/ny_expenditure/ny_expend.csv"))

# ======================================================
# ---------------- EDA ----------------------
# ======================================================

# ---- prior to subsetting ----

# ------ numerical summaries --------
ny <- filter(ny, !is.na(expenditure)) # remove 2 rows with missing values
dim(ny) # 914 x 7
summary(ny)
corr_mat <- cor(ny)
corrplot(corr_mat, type = "upper", diag = FALSE) # correlation matrix plot

# highest correlations
cor(ny$pop, ny$pop_dens) # 0.67
cor(ny$income, ny$pop_dens) # 0.49
cor(ny$income, ny$perc_intergov) # -0.30
cor(ny$income, ny$pop) # 0.29

ny <- ny %>%
  dplyr::select(-pop_dens) # remove population density

# ---  histograms and density plots ----

# function to create histogram and density given variable name
# as is, it's really only useful in this script
hist_dens <- function(var){
  plot_var <- enquo(var)
  ggplot(ny, aes(x = !! plot_var, y = ..density..)) +
    geom_histogram(bins = 50) +
    geom_density()
}

hist_dens(expenditure)
hist_dens(wealth)
hist_dens(income)
hist_dens(pop)
```

```r
hist_dens(perc_intergov)
hist_dens(grow_rate)

# all variables are right-skewed so we log transform them all
# however, since growth rate contains negative values,
# we use a modified log transform
ny <- ny %>%
  mutate(
    log_expenditure = log(expenditure),
    log_wealth = log(wealth),
    log_income = log(income),
    log_pop = log(pop),
    log_perc_intergov = log(perc_intergov),
    log_grow_rate = ifelse(grow_rate > 0,
                           log(grow_rate + 0.15),
                           -log(-grow_rate + 0.15))
  )

hist_dens(log_expenditure)
hist_dens(log_wealth)
hist_dens(log_income)
hist_dens(log_pop)
hist_dens(log_perc_intergov)
hist_dens(log_grow_rate)


# ---  scatterplots: response vs. predictors ----

# function to create scatterplot of log expenditure vs. log covariate given
# covariate name. also adds a scatter plot smooth
scatter_smooth_log <- function(var){
  x_var <- enquo(var)
  ggplot(ny, aes(x = !! x_var, y = log_expenditure)) +
    geom_point() +
    geom_smooth(method = loess, formula = y ~ x) # see also method = lm
}

scatter_smooth_log(log_wealth)
scatter_smooth_log(log_income)
scatter_smooth_log(log_pop)
scatter_smooth_log(log_perc_intergov)
scatter_smooth_log(log_grow_rate)

# all the above plots show an approximatley linear relationship
# except log expenditure vs. log population
# in the following plot, it looks like if we might be able to
# split log population into 2 groups and get an approximately linear relation

lpop_smooth <- lowess(ny$log_pop, ny$log_expenditure)

ggplot(ny, aes(x = log_pop, y = log_expenditure)) +
```

```
  geom_point() +
  geom_line(aes(x = lpop_smooth$x, y = lpop_smooth$y), color = "#00BFC4") +
  geom_segment(aes(x = 8.3, xend = 8.3, y = 4, yend = 5.24), color = "#00BA38")


# ------ data subsetting ------
ny_low <- ny %>%
  filter(log_pop <= 8.3)

ny_high <- ny %>%
  filter(log_pop > 8.3)
dim(ny_high) # 266  12

ggplot(ny_low, aes(x = log_pop, y = log_expenditure)) +
  geom_point() +
  geom_smooth(method = loess, formula = y ~ x)

lpop_high_smooth <- lowess(ny_high$log_pop, ny_high$log_expenditure)

ggplot(ny_high, aes(x = log_pop, y = log_expenditure)) +
  geom_point() +
  geom_line(aes(x = lpop_high_smooth$x,
                y = lpop_high_smooth$y), color = "#00BFC4")


# ----- histograms and density plots -----------

hist_dens_high <- function(var){
  plot_var <- enquo(var)
  ggplot(ny_high, aes(x = !! plot_var, y = ..density..)) +
    geom_histogram(bins = 50) +
    geom_density(color = "#00BFC4") +
    theme_classic()
}

hist_dens_high(expenditure)
hist_dens_high(wealth)
hist_dens_high(income)
hist_dens_high(pop)
hist_dens_high(perc_intergov)
hist_dens_high(grow_rate)

# ----- log transformed ----
hist_dens_high(log_expenditure)
hist_dens_high(log_wealth)
hist_dens_high(log_income)
hist_dens_high(log_pop)
hist_dens_high(log_perc_intergov)
hist_dens_high(log_grow_rate)
```

```r
# ----- scatterplots: response vs. predictors -----

scatter_smooth_log_high <- function(var){
  x_var <- enquo(var)
  ggplot(ny_high, aes(x = !! x_var, y = log_expenditure)) +
    geom_point() +
    ylab("Log Expenditure")
}

scatter_smooth_log_high(log_wealth) + xlab("Log Wealth")
scatter_smooth_log_high(log_income) + xlab("Log Income")
scatter_smooth_log_high(log_perc_intergov) +
  xlab("Log % Intergov Funding")
scatter_smooth_log_high(log_grow_rate) + xlab("Log Growth Rate")




# based on the above plots, it appears that the relationship between
# log expenditure and log population is approximately piecewise linear

# ========================================================
# ---------------- model building ----------------------
# ========================================================

# we will use stepwise regression for model selection
# with AIC as our criteria
# we consider quadratic log terms in the model
# mostly just to see what the results are
for_mod_fit_high <- ny_high %>%
  dplyr::select(log_expenditure,
                log_wealth,
                log_income,
                log_pop,
                log_perc_intergov,
                log_grow_rate) %>%
  mutate(
    log_wealth_2 = log_wealth^2,
    log_income_2 = log_income^2,
    log_pop_2 = log_pop^2,
    log_perc_intergov_2 = log_perc_intergov^2,
    log_grow_rate_2 = log_grow_rate^2
  )

# fit model without quadratic terms
# best AIC is -611.3
fit_high <- lm(log_expenditure ~ log_wealth +
                 log_income +
                 log_pop +
                 log_perc_intergov +
                 log_grow_rate,
```

```
               data = for_mod_fit_high)
best_fit_high <- stepAIC(fit_high)
best_fit_high

# fit model with quadratic terms
# best AIC is -622.91
# this model is not used because:
# there isn't much of an improvement in AIC compared to the simpler model above
# this model includes quadratic terms without their respective linear terms
# we just can not justify such a model
full_fit <- lm(log_expenditure ~., data = for_mod_fit_high)
best_full_fit <- stepAIC(full_fit)
best_full_fit

final_fit <- lm(log_expenditure ~
                  log_wealth +
                  log_pop +
                  log_perc_intergov +
                  log_grow_rate,
                data = ny_high)

# ======================================================
# ---------------- model diagnostics -------------------
# ======================================================


# --------------- diagnostic plots ----------------

# add fitted values, residuals, etc. to orignal data frame used to fit the model
ny_high <- augment(final_fit, data = ny_high) # (broom package)

# add studentized residuals to data frame
ny_high <- ny_high %>%
  mutate(stud_res = rstudent(final_fit))

# studentized residuals vs. fitted
ggplot(ny_high, aes(x = .fitted, y = stud_res)) +
  geom_point() +
  geom_hline(yintercept = 0)

# qq plot studentized residuals vs. t distribution
dist_pars = list(df = final_fit$df.residual)
ny_high %>%
  mutate(stud = rstudent(final_fit)) %>%
  ggplot(aes(sample = stud)) +
  stat_qq(distribution = qt, dparams = dist_pars[["df"]]) +
  stat_qq_line(distribution = qt, dparams = dist_pars[["df"]])

# row with smallest studentized residual
ny_high %>%
  filter(stud_res == min(stud_res))
```

```r
# this row has a wealth per person as well as the smallest growth rate
summary(ny_high$wealth)
min(ny_high$grow_rate)

ny_high %>%
  filter(stud_res == min(stud_res)) %>%
  dplyr::select(stud_res)

# ----- regression summaries -------

# all p-values (except intercept) are significant at the 0.05 level
# F test significant at 0.05 level
summary(final_fit)

# all VIF's are less than 1.2
vif(final_fit)  # (car package)


# the following code is mostly just formatting
ci_bounds = formatC(signif(confint(final_fit), digits = 6),
                    digits = 2, format = "f", flag = "#")

regress_tbl <- tidy(final_fit) %>% # from broom
  dplyr::select(-statistic) %>%
  bind_cols(`95% CI` = paste("(", ci_bounds[,1], ",", ci_bounds[,2], ")")) %>%
  mutate(
    Estimate = formatC(signif(estimate, digits = 6),
                       digits = 2, format = "f", flag = "#"),
    SE = formatC(signif(std.error, digits = 6),
                 digits = 2, format = "f", flag = "#"),
    `p-value` = formatC(signif(p.value, digits = 6),
                        digits = 3, format = "f", flag = "#")

  ) %>%
  dplyr::select(Estimate, SE, `p-value`, `95% CI`) %>%
  as.data.frame()

rownames(regress_tbl) = c("Intercept", "Log Wealth", "Log Population",
                          "Log % Intergov Funding", "Log Growth Rate")

regress_tbl
# convert to latex table
xtable(regress_tbl, label = "tbl:regress", align = "|l|rrrr|")

# ========================================================
# ---------------- predictions -------------------------
# ========================================================

# projected values for Warwick and Monroe for 2005 & 2025
projected <- data.frame(
```

```
  pop = c(20442L, 31033L, 10496L, 13913L),
  wealth = c(85000L, 89000L, 58000L, 60000L),
  perc_intergov = c(24.7, 26.0, 8.8, 10.1),
  grow_rate = c(35.0, 40.0, 35.0, 35.0)
)

log_projected <- projected %>%
  mutate(
    log_pop = log(pop),
    log_wealth = log(wealth),
    log_perc_intergov = log(perc_intergov),
    log_grow_rate = ifelse(grow_rate > 0,
                           log(grow_rate + 0.15),
                           -log(-grow_rate + 0.15))
  ) %>%
  dplyr::select(-(pop:grow_rate))

sd_fit <- sd(final_fit$resid)

pred <- exp(predict(final_fit,
                    newdata = log_projected,
                    interval = "prediction") + sd_fit^2/2)
pred <- formatC(signif(pred, digits = 6), format = "d", flag = "#")

# format for latex output via xtable
pred_tbl <- projected %>%
  mutate(
    perc_intergov = formatC(signif(perc_intergov, digits = 2),
                            digits = 1, format = "f", flag = "#"),
    grow_rate = formatC(signif(grow_rate, digits = 2),
                        digits = 1, format = "f", flag = "#")
  ) %>%
  rename(
    Population = pop,
    Wealth = wealth,
    `% Intergov Funding` = perc_intergov,
    `Growth Rate` = grow_rate
  ) %>%
  bind_cols(Estimate = pred[,1]) %>%
  bind_cols(`95 % PI` = paste("(", pred[,2], ",", pred[,3], ")"))
rownames(pred_tbl) <- c("Warwick-2005", "Warwick-2025 ",
                        "Monroe-2005", "Monroe-2025")

xtable(pred_tbl, align = "|l|rrrrrr|", label = "tbl:pred",
       caption = "Projections for Warwick and Monroe along with predicted
       expenditures and prediction intervals.")
```