```
################################################################################################
#############################

###              HYDRO QUEBEC TRANSMISSION DATA PROCESING FOR ECONOMETRIC BENCHMARKING
###

################################################################################################
#############################

# Date: February 2021

# Author: Rebecca Kavan

################################################################################################
#############################


# Initialize environment

rm(list = ls())


# set width for output to prevent line wrapping

options(width = 120)


################################################################################################
#############################

### Define Paths

################################################################################################
#############################


dataPath =  "[dir]/HQTDB (Confidential).xlsx"

pegDataOutPath = "[dir]/EconometricData_Transformed.xlsx"

funPath = "[dir]/SingleEqFUNs.R"


################################################################################################
#############################

### Load Packages and Functions

################################################################################################
#############################
```

```
require(foreign)

require(gtools)

require(dplyr)

require(openxlsx)

source(funPath)


################################################################################################################################

### Import Data and Prepare Dataset

################################################################################################################################

startyr = 2004


# Import the dataset

txcostdataraw = read.xlsx(dataPath)

names(txcostdataraw)[1] <- "snlid"


# Add trend variable

txcostdataraw$trend <- txcostdataraw$year - 1995


# Levelize costs

txcostdataraw$rtc <- txcostdataraw$c/txcostdataraw$wndxl


txcostdataraw$rck <- txcostdataraw$ck/txcostdataraw$wklvl


txcostdataraw$rcom <- txcostdataraw$com/txcostdataraw$womlvl


# Pull back values of construction standards index to earlier years
```

```
txcostdataraw <- txcostdataraw %>%
  group_by(snlid) %>%
  arrange(year) %>%
  mutate(load_tx = if_else(year < 2004, load_tx[year == 2004], load_tx))

txcostdataraw <- as.data.frame(txcostdataraw)

# Pull forward values of construction standards index to end of sample
txcostdataraw$load_tx[txcostdataraw$year > 2016 & txcostdataraw$snlid != 1] <-
txcostdataraw$load_tx[txcostdataraw$year == 2016 & txcostdataraw$snlid != 1]

# Extend forestation variable
txcostdataraw$pforgis1[txcostdataraw$year != 2007] <- txcostdataraw$pforgis1[txcostdataraw$year ==
2007]

# Pull back values of substations and mva to earlier years
txcostdataraw$nsub0919[txcostdataraw$year < 2009] <- txcostdataraw$nsub0919[txcostdataraw$year
== 2009]
txcostdataraw$mva0919[txcostdataraw$year < 2009] <- txcostdataraw$mva0919[txcostdataraw$year
== 2009]

# Add mva/nsub variable
txcostdataraw <- txcostdataraw %>%
  mutate(mva0919pernsub0919 = mva0919/nsub0919)

# Create moving substations per line mile variable
txcostdataraw <- txcostdataraw %>%
  group_by(snlid) %>%
  arrange(year) %>%
  mutate(nsub0919perym = nsub0919/ym)
```

```
txcostdataraw <- as.data.frame(txcostdataraw)


# define dataset
txcostdata = txcostdataraw[which(txcostdataraw$exclude == 0 & txcostdataraw$year >= startyr),]


txcostdata = as.data.frame(txcostdata)
################################################################################################################################
### Demean and Log Appropriate Variables
################################################################################################################################


# Specify the variables to log only, and the variables to log and demean
varlist1 <- c("rtc",
              "rcom",
              "rck")
varlist2 <-
  c(
    "ym",
    "yptx",
    "load_tx",
    "pforgis1",
    "pctptx_peg",
    "pctpoh",
    "nsub0919",
    "mva0919",
    "mva0919pernsub0919",
    "nsub0919perym",
```

```
    "rto"
  )
ltxcostvarlist <- c(varlist1, varlist2)
dtxcostvarlist <- c(varlist2)

# Mean scale the appropriate variables in the data
dtxcostout <-
  meanscale(
    fulldata = txcostdata,
    varlist = dtxcostvarlist,
    verbose = F,
    incfc = F
  )

# Get the mean-scaled data from the output of the meanscale function
dtxcostdata <- dtxcostout[[1]]

# Get the variable means from the output of the meanscale function
txcostmeans <- dtxcostout[[2]]

# Log the appropriate variables in the data
ldtxcostout <-
  logvars(fulldata = dtxcostdata,
          varlist = ltxcostvarlist,
          verbose = F)

# Get the logged and mean-scaled data from the output of the logvars function
ldtxcostdata <- ldtxcostout[[1]]
```

```
# Get the information on when ln(1+x) was used instead of ln(x) due to zeroes
txcostinfo <- ldtxcostout[[2]]

# Define row names
row.names(ldtxcostdata) <- as.character(c(1:dim(ldtxcostdata)[1]))



# Create quadratic and interaction terms post-log treatment
ldtxcostdata$ym2 = ldtxcostdata$ym*ldtxcostdata$ym/2
ldtxcostdata$yptx2 = ldtxcostdata$yptx*ldtxcostdata$yptx/2
ldtxcostdata$ymyptx = ldtxcostdata$ym*ldtxcostdata$yptx

# Save transformed dataset
write.xlsx(ldtxcostdata, pegDataOutPath)
```