Multilevel Regression Analysis of the Property Values in Santa Monica, Los Angeles County

Ta Yun Yang

06/17/2020

Global Environment

Notice that the follow research is using stan_glm. The value calculated in the result might be slightly different from the writeup because of the resampling. The values in the writeup should be considered as approximations.

Question 1. Introduction

Code 1: Load and proprocessing data

```
parcel <- read.csv("D:\\Project Data\\Data_Viz project data\\full_data30.csv")
selected_cols <- c("ZIPcode5", "AIN", "GeneralUseType", "TotalValue", "EffectiveYearBuilt",
                   "SQFTmain", "Bedrooms", "Bathrooms", "CT", "unemployment", "pop", "gini",
                   "median_income", "poverty_rate", "pop75", "Travel_Time701902",
                   "TotalValue2017")

# Limited to only commercial properties
parcel <- parcel[parcel$GeneralUseType == "Commercial", selected_cols]

for (col in colnames(parcel)){
  sub_parc <- subset(parcel, is.na(parcel[, col]))
  if (nrow(sub_parc) != 0){
    print(paste0("there is ", paste(nrow(sub_parc), " missing rows in parcels for col ",
                                    paste(col))))}
}
```

```
## [1] "there is 159  missing rows in parcels for col  ZIPcode5"
```

```
parcel$ZIPcode5 <- NULL
parcel <- parcel[(parcel$EffectiveYearBuilt != 0) & (parcel$TotalValue != 0) &
                   (parcel$TotalValue2017 != 0), ]
parcel$EffectiveYearBuilt <- 2018 - parcel$EffectiveYearBuilt
```

Response 1:

The inequality of property value between neighborhoods have been considered as one of the important questions in Economics, Sociology and industry. We are interested in the factors that drives the property value. In the following research, Santa Monica region in Los Angeles is selected to be researched since it possesses independent commercial district and the neighbored regions are limited. The factors are assumed to be divided to three groups: Housing characteristics, traveling distance to main commercial district and demographic factors. First of all, data for the housing characteristics and values are collected from parcel data of Assessor's Office of Los Angeles County. It includes property value of 2018 (denoted as TotalValue) which is used as dependent variable, and building square footage (denoted as SQFTmain), building age (denoted as EffectiveYearBuilt), number of Bedrooms and Bathrooms (denoted as Bedrooms and Bathrooms) that are used as individual level (per house parcel) predictors. Secondly, I collected regional level (Census Tract) traveling distance from main commercial district (denoted Travel_Time701902 where Census Tract No.701902 is the commercial district of Santa Monica) from Uber Movement data which is a appropriate approximation of traveling time in the region in LA since automobile are the main traveling tool. Finally, regional level (Census Tract) unemployment rate and median income data are collected from American Community Survey 2017, which could be used to represent the demographic patterns of neighborhoods.

In my data, there are 27 regional groups (Census Tract). In order to limit the number properties in the analysis under time constraint, I only include the commercial properties in the following analysis.

```
CTn <- parcel %>% group_by(CT) %>% summarize(n=n()) %>% arrange(-n)
print(paste0("there are ", paste(nrow(CTn)), " Census Tract in the data"))
```

```
## [1] "there are 27 Census Tract in the data"
```

The 5 groups with largest size and 5 groups (Census Tract) with smallest size are presented as the following

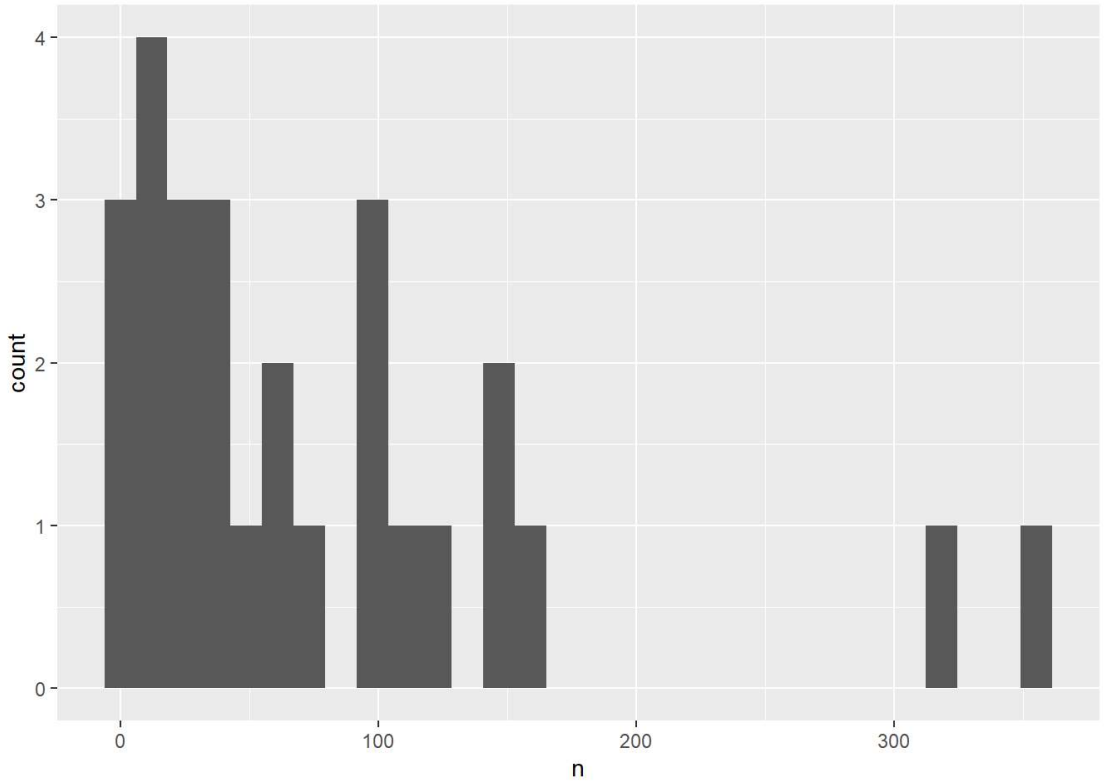
```
knitr::kable(list(head(CTn), tail(CTn)))
```

CT	n	CT	n
702300	360	271701	17
701902	315	269000	11
273100	160	271901	7
701702	148	271100	6
267800	146	269800	5
267200	122	271500	5

The distribution of the group size is presented as the following

```
ggplot(data=CTn, aes(n)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Multilevel model might be more effective than traditional pooled or unpooled model in this data structure since it could separate the variance of the regional level factors from the individual level factors more precisely. We could distinguish the effect within or between the communities when we are searching for the factors which influence the property value after controlling for the variables in individual level.

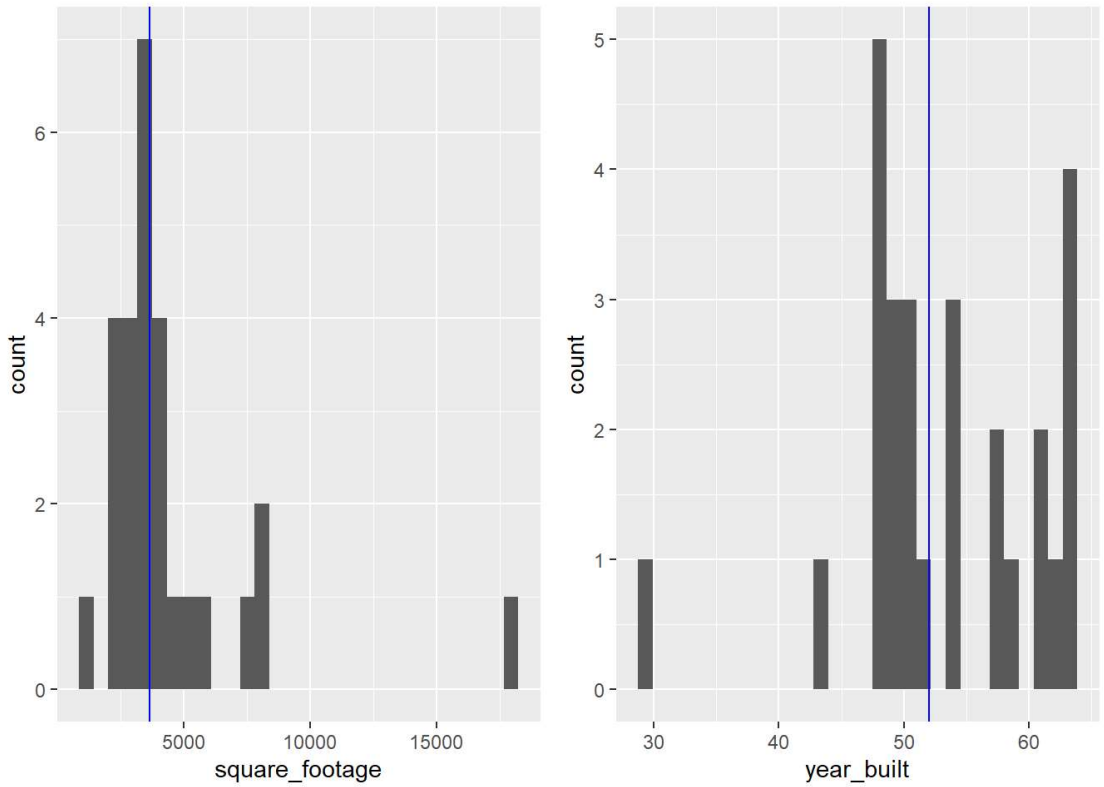
Question 2 Code 2:

```
CT_summary <- parcel %>% group_by(CT) %>%
              summarize(n=n(), square_footage=median(SQFTmain),
                        year_built = median(EffectiveYearBuilt),
                        median_income=median(median_income),
                        traveltime_to_com=median(Travel_Time701902),
                        median_property_value=median(TotalValue)) %>%
              arrange(-n)

plt1 <- ggplot(data=CT_summary, aes(square_footage)) + geom_histogram() +
    geom_vline(xintercept = median(CT_summary$square_footage), col="blue")

plt2 <- ggplot(data=CT_summary, aes(year_built)) + geom_histogram() +
    geom_vline(xintercept = median(CT_summary$year_built), col="blue")

plt3 <- ggplot(data=CT_summary, aes(median_income)) + geom_histogram() +
    geom_vline(xintercept = median(CT_summary$median_income), col="blue")

plt4 <- ggplot(data=CT_summary, aes(traveltime_to_com)) + geom_histogram() +
    geom_vline(xintercept = median(CT_summary$traveltime_to_com), col="blue")

plt5 <-ggplot(data=CT_summary, aes(median_property_value)) + geom_histogram() +
    geom_vline(xintercept = median(CT_summary$median_property_value), col="blue")

plt6 <- ggplot(CT_summary, aes(x=square_footage, y=median_property_value)) +
    geom_point(aes(size=median_income))
```

Response 2: Histograms of building square footage and building age (median of individual level in each group). The distribution of median house square footage is more diversified than the median building age.

```
grid.arrange(plt1, plt2, nrow = 1)
```

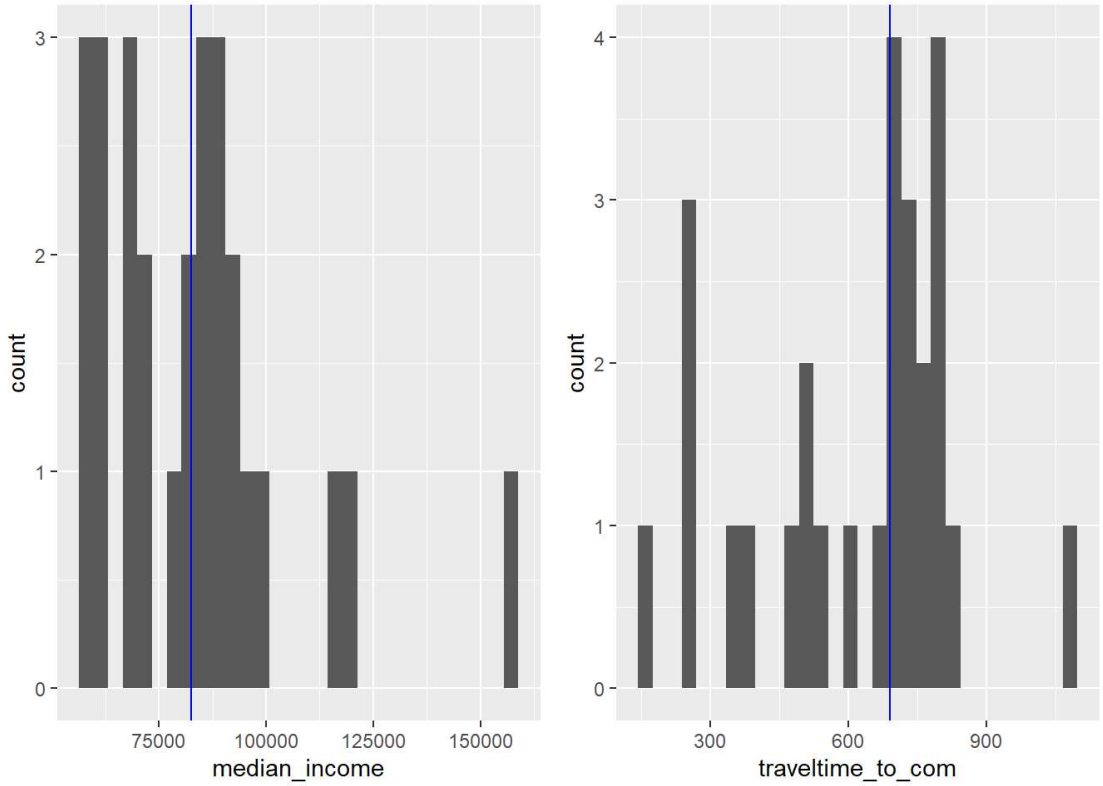
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Histogram of community median income and traveling time to commercial district (value in group level). Both group level traveling time and median income are distributed diversely.

```
grid.arrange(plt3, plt4, nrow = 1)
```

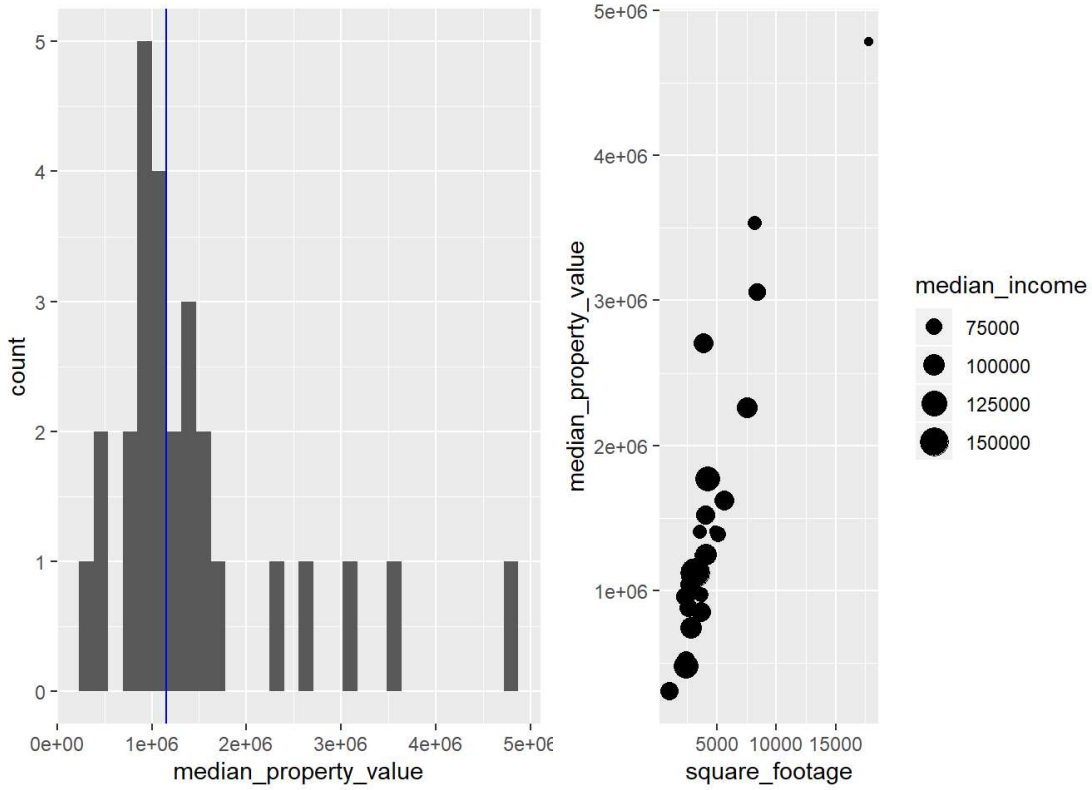
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Histogram of the property value (dependent variables, median of individual level in each group), and scatter plot (property value v.s. median income and median square footage of house). The distribution of median property value are close to normal distribution with some extremely high values. On the other hand, we could also observe a positive relation between building square footage and property value, although the influence from median income is more ambiguous.

```
grid.arrange(plt5, plt6, nrow = 1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Question 3

Code 3:

```
# Pooled Model
fit.pool <- stan_glm(TotalValue ~ EffectiveYearBuilt + SQFTmain + Bedrooms + Bathrooms
                     + unemployment + median_income + Travel_Time701902,
                     data=parcel)

# Unpooled Model
fit.unpool <- stan_glm(TotalValue ~ EffectiveYearBuilt + SQFTmain + Bedrooms + Bathrooms
                       + unemployment + median_income + Travel_Time701902 + as.factor(CT),
                       data=parcel)
```

Response 3.a Pooled Model

$$TotalValue_i = \alpha + \beta_1 SQFTmain_i + \beta_2 EffectiveYearBuilt_i + \beta_3 Bedrooms_i + \beta_4 Bathrooms$$

$$+\beta_5 MedianIncome_i + \beta_6 Unemployment_i + \beta_7 TravelTime701902_i + \epsilon_i$$

```
fit.pool
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      TotalValue ~ EffectiveYearBuilt + SQFTmain + Bedrooms + Bathrooms +
##     unemployment + median_income + Travel_Time701902
##  observations: 2156
##  predictors:   8
## ------
##                    Median    MAD_SD
## (Intercept)        -176590.6 1399657.7
## EffectiveYearBuilt  -20047.5    9677.2
## SQFTmain               424.4       4.8
## Bedrooms            -41145.2   37659.6
## Bathrooms            50267.3   40355.1
## unemployment        249240.6   88245.1
## median_income            0.9      11.4
## Travel_Time701902    -1594.0     907.6
##
## Auxiliary parameter(s):
##       Median    MAD_SD
## sigma 8442151.7  134121.8
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

According to the table of coefficients, we can discover that a year increase in building age is corresponding to 19600 dollars decrease in property value; a unit of square footage increase is corresponding to 424 dollars increase in property value; an unit increase of unemployment rate is corresponding to 250,000 dollars increase in property value. Other coefficients estimates have very high uncertainty especially for the intercept.

Response 3.b

$$TotalValue_i = \alpha + \beta_1 SQFTmain_i + \beta_2 EffectiveYearBuilt_i + \beta_3 Bedrooms_i + \beta_4 Bathrooms + \beta_5 MedianIncome_i$$

$$+\beta_6 Unemployment_i + \beta_7 TravelTime701902_i + CensusTract_i + \epsilon_i$$

where $CensusTract_i$ represents the (27 binary) indicators that mark which census tracts the house belongs to.

```
fit.unpool
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      TotalValue ~ EffectiveYearBuilt + SQFTmain + Bedrooms + Bathrooms +
##     unemployment + median_income + Travel_Time701902 + as.factor(CT)
##  observations: 2156
##  predictors:   34
## ------
##                     Median       MAD_SD
## (Intercept)         -118015516.5  249409584.5
## EffectiveYearBuilt      949226.9    4821401.7
## SQFTmain                  -490.7       1721.4
## Bedrooms              -2769708.7    5526260.8
## Bathrooms              -565580.9    8624150.5
## unemployment          -4634560.6   16272922.3
## median_income             1096.3       2586.9
## Travel_Time701902      -131371.5     138101.7
## as.factor(CT)267200  -22521491.1    7105133.3
## as.factor(CT)267600  -58154082.2   42660587.7
## as.factor(CT)267800   41322365.8   53797656.2
## as.factor(CT)269000   46169863.7   53605649.2
## as.factor(CT)269300   24498478.6   24852205.4
## as.factor(CT)269800   63362398.5   19200174.8
## as.factor(CT)269904   17153938.0   46689662.3
## as.factor(CT)271100   62776912.0   17215489.1
## as.factor(CT)271200  -17837223.4   25592236.0
## as.factor(CT)271300   -5570395.9    4176282.9
## as.factor(CT)271500   67748740.5   19424565.0
## as.factor(CT)271600     606438.0   67759206.8
## as.factor(CT)271701  -36100427.2   59766007.7
## as.factor(CT)271702   72958468.9   21111116.5
## as.factor(CT)271901    1491227.2   63207774.0
## as.factor(CT)273100   83743207.0    7969243.2
## as.factor(CT)701402   34149502.8   37382628.1
## as.factor(CT)701701  -22921307.0   58622792.2
## as.factor(CT)701702   27696970.2   68118374.9
## as.factor(CT)701801   -2310121.2   45086556.6
## as.factor(CT)701802   33755106.5   47224702.3
## as.factor(CT)701902   -9056052.1   37332843.9
## as.factor(CT)702201    4421200.4   45639191.2
## as.factor(CT)702202  -58852313.3   42146936.2
## as.factor(CT)702300   68595993.2   28746904.0
## as.factor(CT)702802   32274983.4   52383565.5
##
## Auxiliary parameter(s):
##       Median         MAD_SD
## sigma 202720507016.6  62739979994.1
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

According to the table of coefficients, we can discover that a year increase in building age is corresponding to 855,000 dollars increase in property value; an unit increase of median income is corresponding to 1,350 dollars decrease in property value. Other coefficients estimates have very high uncertainty. Most of the group indicators included are having high level of uncertainty, and also, in comparison to the pooled model, the sign of median income and building age have been changed and the effect of building square footage becomes ambiguous. This is likely to be caused by multi-collinearity of the variables that the estimates become fluctuating and uncertain.

Question 4

Code 4:

```
# Multilevel intercept
fit.model1 <- stan_glmer(TotalValue ~ EffectiveYearBuilt + SQFTmain + Bedrooms + Bathrooms
                         + (1 | CT) + unemployment + median_income + Travel_Time701902,
                         data=parcel)

# Multilevel intercept with previous (2017) assessed value (see Appendix.a for more information)
fit.model2 <- stan_glmer(TotalValue ~ EffectiveYearBuilt + SQFTmain + Bedrooms + Bathrooms
                         + (1 | CT) + unemployment + median_income + Travel_Time701902
                         + TotalValue2017,
                         data=parcel)

# Multilevel intercept and slope (see Appendix.b for more information)
# fit.model3 <- stan_glmer(TotalValue ~ EffectiveYearBuilt + SQFTmain + Bedrooms + Bathrooms
#                          + (1 + TotalValue2017| CT) + unemployment
#                          + median_income + Travel_Time701902 + TotalValue2017
#                          + TotalValue2017:unemployment + TotalValue2017:median_income
#                          + TotalValue2017:Travel_Time701902,
#                          data=parcel)

# Multilevel with two level of group
parcel$travel_dregion <- 1
index <- 2
q_list <- c(0.25, 0.5, 0.75)
for (q in q_list){
  parcel$travel_dregion <- ifelse(parcel$Travel_Time701902 > quantile(
                           parcel$Travel_Time701902, q)[[1]], index, parcel$travel_dregion)
  index <- index + 1
}

# Without previous (2017) assessed value
fit.model5 <- stan_glmer(TotalValue ~ EffectiveYearBuilt + SQFTmain + Bedrooms + Bathrooms
                         + unemployment + median_income + Travel_Time701902
                         + (1| CT) + (1|travel_dregion),
                         data=parcel)

# With previous (2017) assessed value ((see Appendix.a for more information))
fit.model6 <- stan_glmer(TotalValue ~ EffectiveYearBuilt + SQFTmain + Bedrooms + Bathrooms
                         + unemployment + median_income + Travel_Time701902 + TotalValue2017
                         + (1| CT) + (1|travel_dregion),
                         data=parcel)
```

Response 4.a

$$TotalValue_i = \alpha_{j[i]} + \beta_1 SQFTmain_i + \beta_2 EffectiveYearBuilt_i + \beta_3 Bedrooms_i + \beta_4 Bathrooms + \epsilon_i$$

$$\alpha_{j[i]} = \gamma_0 + \gamma_1 MedianIncome_j + \gamma_2 Unemployment_j + \gamma_3 TravelTime701902_j + \eta_j$$

where group level predictors (median income, unemployment rate and traveling time to commercial district) are used to estimate the group level varying intercept

```
fit.model1
```

```
## stan_glmer
##  family:       gaussian [identity]
##  formula:      TotalValue ~ EffectiveYearBuilt + SQFTmain + Bedrooms + Bathrooms +
##     (1 | CT) + unemployment + median_income + Travel_Time701902
##  observations: 2156
## ------
##                    Median    MAD_SD
## (Intercept)        -219736.6 1668311.8
## EffectiveYearBuilt  -21064.8    9828.1
## SQFTmain               423.7       4.7
## Bedrooms            -44238.8   37130.5
## Bathrooms            53356.6   40966.3
## unemployment        258501.2  111314.7
## median_income            0.4      13.5
## Travel_Time701902    -1439.9    1023.6
##
## Auxiliary parameter(s):
##       Median    MAD_SD
## sigma 8429346.2  132245.1
##
## Error terms:
##  Groups   Name        Std.Dev.
##  CT       (Intercept)  738298
##  Residual             8430964
## Num. levels: CT 27
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

According to the table of coefficient, we can discover that a year increase in building age is corresponding to 21,000 dollars decrease in property value; a unit of square footage increase is corresponding to 423 dollars increase in property value; an unit increase of unemployment rate is corresponding to 250,000 dollars increase in property value. Other coefficients estimates have very high uncertainty especially for the intercept. For the group level error term (standard deviation of group level σ_η), the census tract differs by $\$720,000$, and for the individual error term (standard deviation of individual level σ_y), the properties value differs by $\$8,430,000$.

In comparison to the models in Question 3, the uncertainty of σ_y lies between the pooled and unpooled model (130,000 is between 128,000 and 57,000,000) and the coefficient estimates are very similar between pooled model and model 4.a. The group level predictors are absorbing some of the unexplained group variance in the previous model. To conclude, the model includes variation of different group (census tract) and remain the precision of the estimation and prediction.

Response 4.b

$$TotalValue_i = \alpha^{CT}_{j[i]} + \beta_1 SQFTmain_i + \beta_2 EffectiveYearBuilt_i + \beta_3 Bedrooms_i + \beta_4 Bathrooms + \epsilon_i$$

$$\alpha^{CT}_{j[i]} \sim N(\gamma_{m[j]} + \gamma_1 MedianIncome_j + \gamma_2 Unemployment_j + \gamma_3 TravelTime701902_j \ , \ \sigma^2_{CT})$$

where group level predictors (median income, unemployment rate and traveling time to commercial district) are used to estimate the group level varying intercept

$$\alpha^{region}_m \sim N(0, \ \sigma^2_{region})$$

where second level region is defined as the distance level to commercial districts (categorized 1 - 4 from close to far), which is assumed to be varying for intercept in first-level group intercept estimates.

```
fit.model5
```

```
## stan_glmer
##  family:       gaussian [identity]
##  formula:      TotalValue ~ EffectiveYearBuilt + SQFTmain + Bedrooms + Bathrooms +
##     unemployment + median_income + Travel_Time701902 + (1 | CT) +
##     (1 | travel_dregion)
##  observations: 2156
## ------
##                     Median    MAD_SD
## (Intercept)        -430870.3 2015874.2
## EffectiveYearBuilt  -20871.6    9761.7
## SQFTmain               423.6       4.6
## Bedrooms            -44825.4   37297.8
## Bathrooms            55014.8   41201.9
## unemployment        260957.5  118275.1
## median_income            0.5      13.7
## Travel_Time701902    -1131.3    1652.6
##
## Auxiliary parameter(s):
##       Median    MAD_SD
## sigma 8425630.9  132789.3
##
## Error terms:
##  Groups         Name        Std.Dev.
##  CT             (Intercept)  649395
##  travel_dregion (Intercept) 1260514
##  Residual                   8426426
## Num. levels: CT 27, travel_dregion 4
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

According to the table of coefficient, we can discover that a year increase in building age is corresponding to 20,000 dollars decrease in property value; a unit of square footage increase is corresponding to 423 dollars increase in property value; an unit increase of unemployment rate is corresponding to 260,000 dollars increase in property value. Other coefficients estimates have very high uncertainty especially for the intercept. For the first group error term (standard deviation of group level σ_{CT}), the census tract differs by $\$650,000$, and for the second group error term (standard deviation of group level σ_{region}), the region differs by $\$1,400,000$. And for the individual error term (standard deviation of individual level σ_y), the properties value differs by $\$8,430,000$. In both cases in model 4.a and 4.b, between group variance is relatively low comparing to the within group variance.

In comparison to model 4.a, the estimates for the coefficients are very similar to model 4.a, and the estimates of the variance of σ_y (slightly decrease from 130,000). The the second level group level predictor is absorbing some of the unexplained variance in the first level group estimate and end up in reducing the uncertainty in the individual level (which causes σ_{CT} to decrease). To conclue, the model include variation of different group (census tract and regions) and remain the precision of the estimation and prediction.

Question 5

Code 5:

```
stan_summary <- function(fit){
  as.data.frame(fit) %>%
    gather("parameter") %>%
    group_by(parameter) %>%
    summarize(median=median(value), MAD_SD=mad(value))
}

plot_pred <- function(newdata, pred, xcolname){
  pred <- as.data.frame(pred)
  colnames(pred) <- newdata$CT

  plot_table <- stan_summary(pred)
  plot_table <- merge(newdata, plot_table, by.x="CT", by.y="parameter")
  plot_table$fit <- plot_table$median

  plt <- ggplot(data=plot_table,
                aes(x=plot_table[, xcolname],
                    y=fit,
                    ymin=fit-MAD_SD,
                    ymax=fit+MAD_SD)) +
                geom_point() +
                geom_errorbar() +
                geom_smooth(method="lm", formula=y~x) +
                labs(x = xcolname)
  return(plt)
}
```

For pooled model

```
pred0 <- predict(fit.pool, new_data0)
parcel$pred0 <- pred0

CT_summary0 <- parcel %>% group_by(CT) %>%
               summarize(EffectiveYearBuilt=median(EffectiveYearBuilt),
                         SQFTmain=median(SQFTmain),
                         pred0=median(pred0))

CT_summary0$pred_mad <- mad(pred0)


plt1 <- ggplot(data=CT_summary0,
       aes(x=EffectiveYearBuilt,
           y=pred0,
           ymin=pred0-pred_mad,
           ymax=pred0+pred_mad)) +
       geom_point() +
       geom_errorbar() +
       geom_smooth(method="lm", formula=y~x)

plt2 <- ggplot(data=CT_summary0,
       aes(x=SQFTmain,
           y=pred0,
           ymin=pred0-pred_mad,
           ymax=pred0+pred_mad)) +
       geom_point() +
       geom_errorbar() +
       geom_smooth(method="lm", formula=y~x)
```

For unpooled model

```
CT_summary4 <- parcel %>% group_by(CT) %>%
               summarize(EffectiveYearBuilt=median(EffectiveYearBuilt),
                         SQFTmain=median(SQFTmain),
                         Bedrooms=median(Bedrooms),
                         Bathrooms=median(Bathrooms),
                         unemployment=median(unemployment),
                         median_income=median(median_income),
                         Travel_Time701902=median(Travel_Time701902))

new_data4 <- data.frame(CT=CT_summary4$CT,
                        EffectiveYearBuilt=CT_summary4$EffectiveYearBuilt,
                        SQFTmain=CT_summary4$SQFTmain,
                        Bedrooms=CT_summary4$Bedrooms,
                        Bathrooms=CT_summary4$Bathrooms,
                        unemployment=CT_summary4$unemployment,
                        median_income=CT_summary4$median_income,
                        Travel_Time701902=CT_summary4$Travel_Time701902)

pred4 <- predict(fit.unpool, newdata=new_data4)
CT_summary4$pred4 <- pred4

plot1 <- ggplot(data=CT_summary4,
       aes(x=EffectiveYearBuilt,
           y=pred4)) +
       geom_point() +
       geom_smooth(method="lm", formula=y~x)

plot2 <- ggplot(data=CT_summary4,
       aes(x=SQFTmain,
           y=pred4)) +
       geom_point() +
       geom_smooth(method="lm", formula=y~x)
```

For model 4.a

```
CT_summary1 <- parcel %>% group_by(CT) %>%
               summarize(EffectiveYearBuilt=median(EffectiveYearBuilt),
                         SQFTmain=median(SQFTmain),
                         Bedrooms=median(Bedrooms),
                         Bathrooms=median(Bathrooms),
                         unemployment=median(unemployment),
                         median_income=median(median_income),
                         Travel_Time701902=median(Travel_Time701902))

new_data1 <- data.frame(CT=CT_summary1$CT,
                        EffectiveYearBuilt=CT_summary1$EffectiveYearBuilt,
                        SQFTmain=CT_summary1$SQFTmain,
                        Bedrooms=CT_summary1$Bedrooms,
                        Bathrooms=CT_summary1$Bathrooms,
                        unemployment=CT_summary1$unemployment,
                        median_income=CT_summary1$median_income,
                        Travel_Time701902=CT_summary1$Travel_Time701902)

pred1 <- posterior_linpred(fit.model1, newdata=new_data1)
plot1E <- plot_pred(new_data1, pred1, "EffectiveYearBuilt")
plot1S <- plot_pred(new_data1, pred1, "SQFTmain")
```

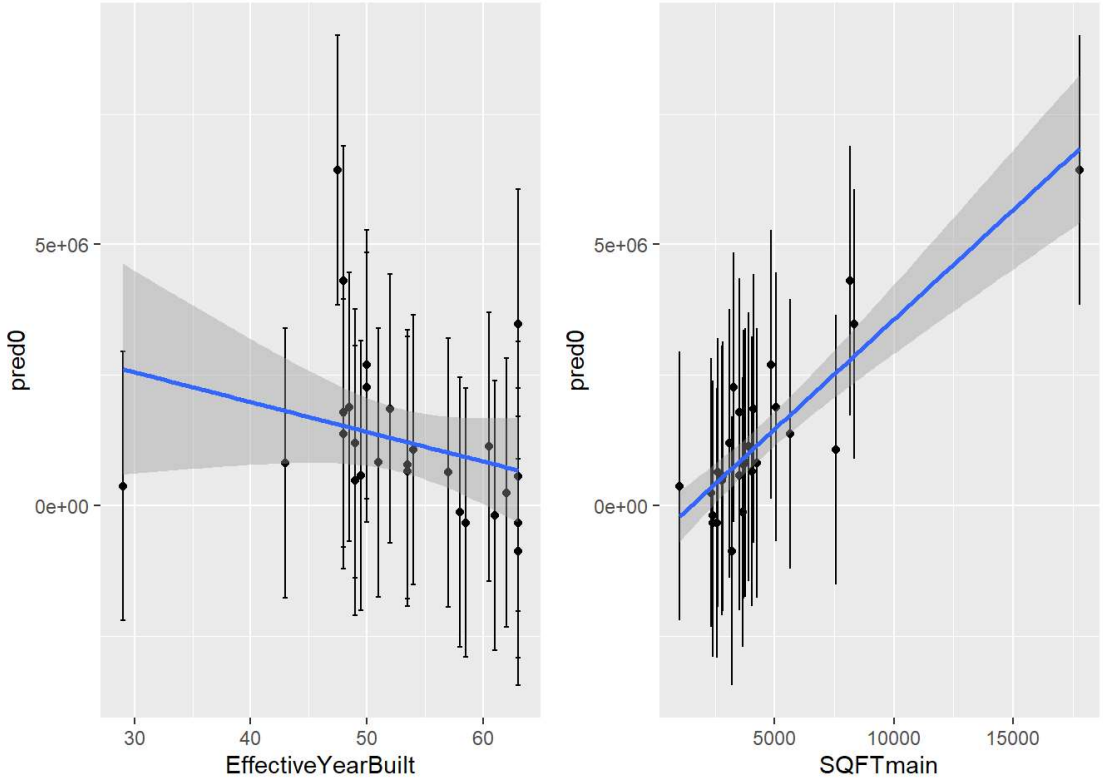
For model 4.b

```
CT_summary5 <- parcel %>% group_by(CT) %>%
  summarize(EffectiveYearBuilt=median(EffectiveYearBuilt),
            SQFTmain=median(SQFTmain),
            Bedrooms=median(Bedrooms),
            Bathrooms=median(Bathrooms),
            unemployment=median(unemployment),
            median_income=median(median_income),
            Travel_Time701902=median(Travel_Time701902))

new_data5 <- data.frame(CT=CT_summary5$CT,
                        EffectiveYearBuilt=CT_summary5$EffectiveYearBuilt,
                        SQFTmain=CT_summary5$SQFTmain,
                        Bedrooms=CT_summary5$Bedrooms,
                        Bathrooms=CT_summary5$Bathrooms,
                        unemployment=CT_summary5$unemployment,
                        median_income=CT_summary5$median_income,
                        Travel_Time701902=CT_summary5$Travel_Time701902,
                        travel_dregion=2)

pred5 <- posterior_linpred(fit.model5, newdata=new_data5)
plot5E <- plot_pred(new_data5, pred5, "EffectiveYearBuilt")
plot5S <- plot_pred(new_data5, pred5, "SQFTmain")
```

Response 5

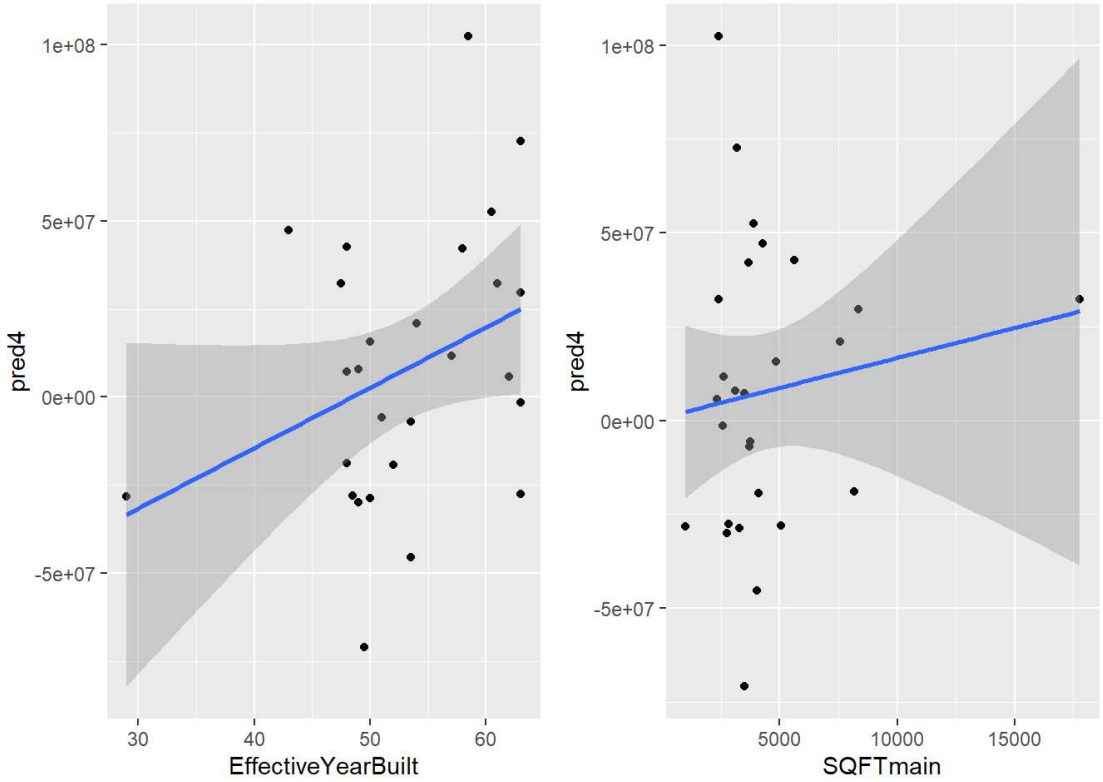
For pooled model

```
grid.arrange(plt1, plt2, nrow = 1)
```



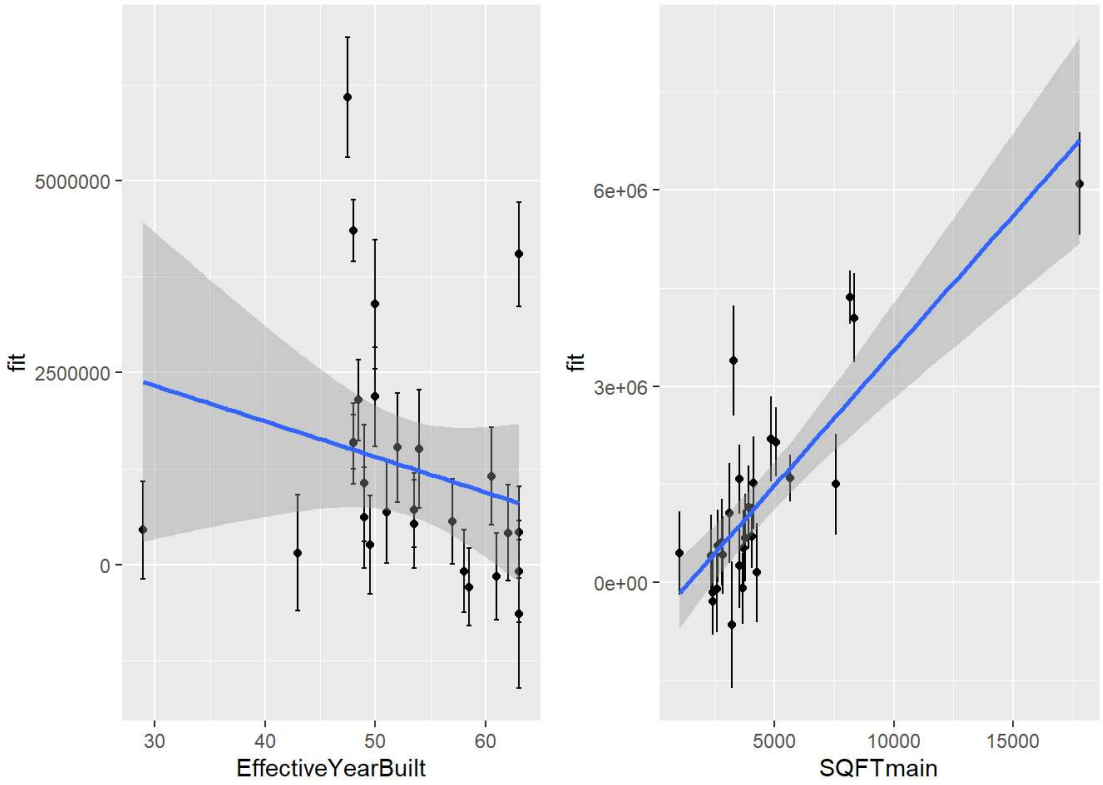
For unpooled model

```
grid.arrange(plot1, plot2, nrow = 1)
```



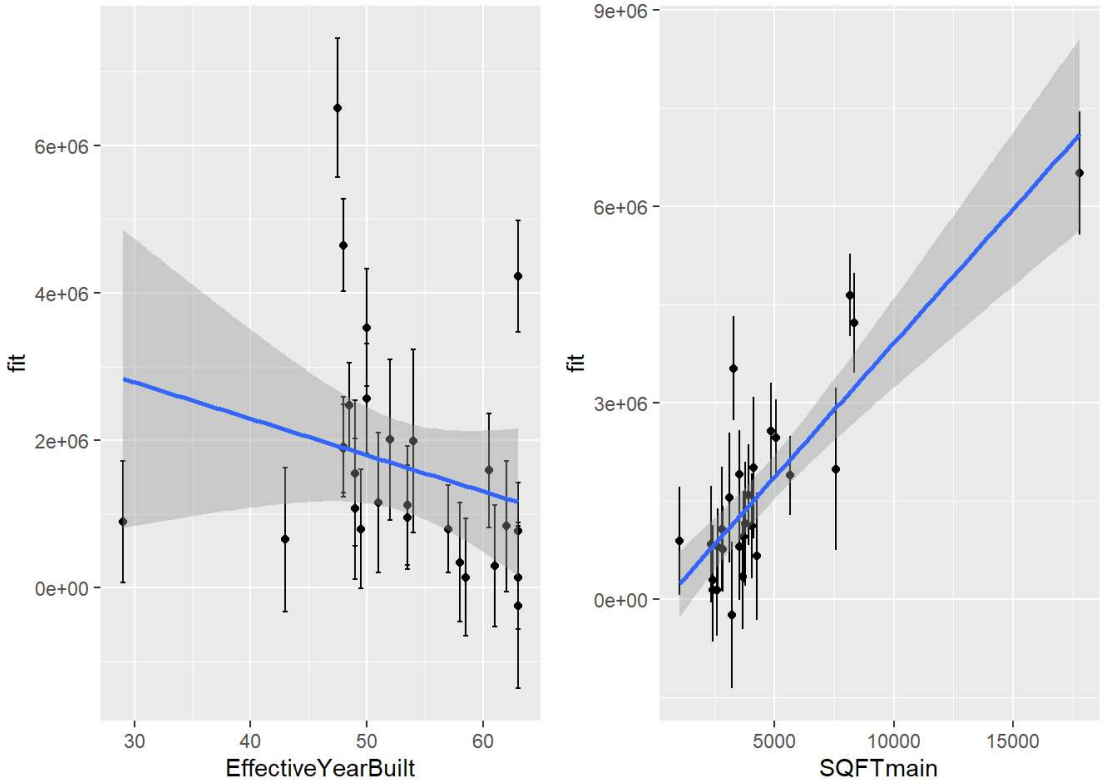
For model 4.a

```
grid.arrange(plot1E, plot1S, nrow = 1)
```



For model 4.b

```
grid.arrange(plot5E, plot5S, nrow = 1)
```



In the pooled model, model 4.a and model 4.b, the positive correlation between building square footage and property value, and the negative correlation between building age and property value are clear both between and within groups. On the contrary, due to the result of collinearity, the prediction and estimate of unpooled model are ambiguous that we cannot observe a consistent pattern of it.

Question 6

Code 6.a

```
n <- nrow(CT_summary0)
plot_alpha <- function(model, sum_data, xcolname) {

  alphas <- stan_summary(model)[(2:(n + 1)), ]
  sum_data$parameter <- paste0("b[(Intercept) CT:",
                               paste(as.character(sum_data$CT)), "]")
  full_sum <- merge(alphas, sum_data, by="parameter")
  full_sum$alphaj <- full_sum$median

  plt <- ggplot(data=full_sum,
                aes(x=full_sum[, xcolname],
                    y=alphaj,
                    ymin=alphaj-MAD_SD,
                    ymax=alphaj+MAD_SD)) +
         geom_point() +
         geom_errorbar() +
         geom_smooth(method="lm", formula=y~x) +
         labs(x = xcolname)

  # extract the percentage of property value
  CTv <- parcel %>% group_by(CT) %>% summarize(TotalValue=median(TotalValue))
  CTv_final <- merge(CTv, full_sum, by="CT")
  CTv_final$ratio <- CTv_final$MAD_SD / CTv_final$TotalValue
  print(paste0("the fluctuation of alpha j lies between ",
               paste(min(CTv_final$ratio) * 100), " and "))
  print(paste0(paste(max(CTv_final$ratio) * 100), " % of property value"))

  return(plt)
}
```

Response 6.a For model 4.a

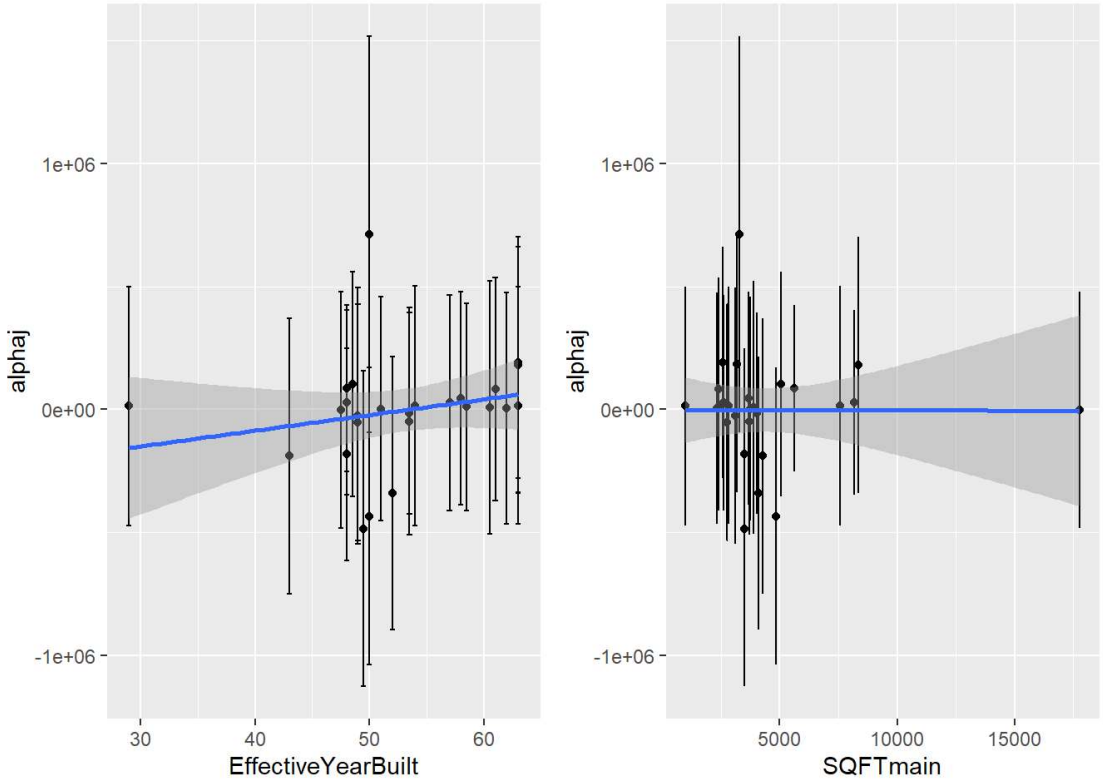
```
plotAE <- plot_alpha(fit.model1, CT_summary1, "EffectiveYearBuilt")
```

```
## [1] "the fluctuation of alpha j lies between 10.0570038584491 and "
## [1] "160.709610953923 % of property value"
```

```
plotAS <- plot_alpha(fit.model1, CT_summary1, "SQFTmain")
```

```
## [1] "the fluctuation of alpha j lies between 10.0570038584491 and "
## [1] "160.709610953923 % of property value"
```

```
grid.arrange(plotAE, plotAS, nrow = 1)
```



For model 4.b

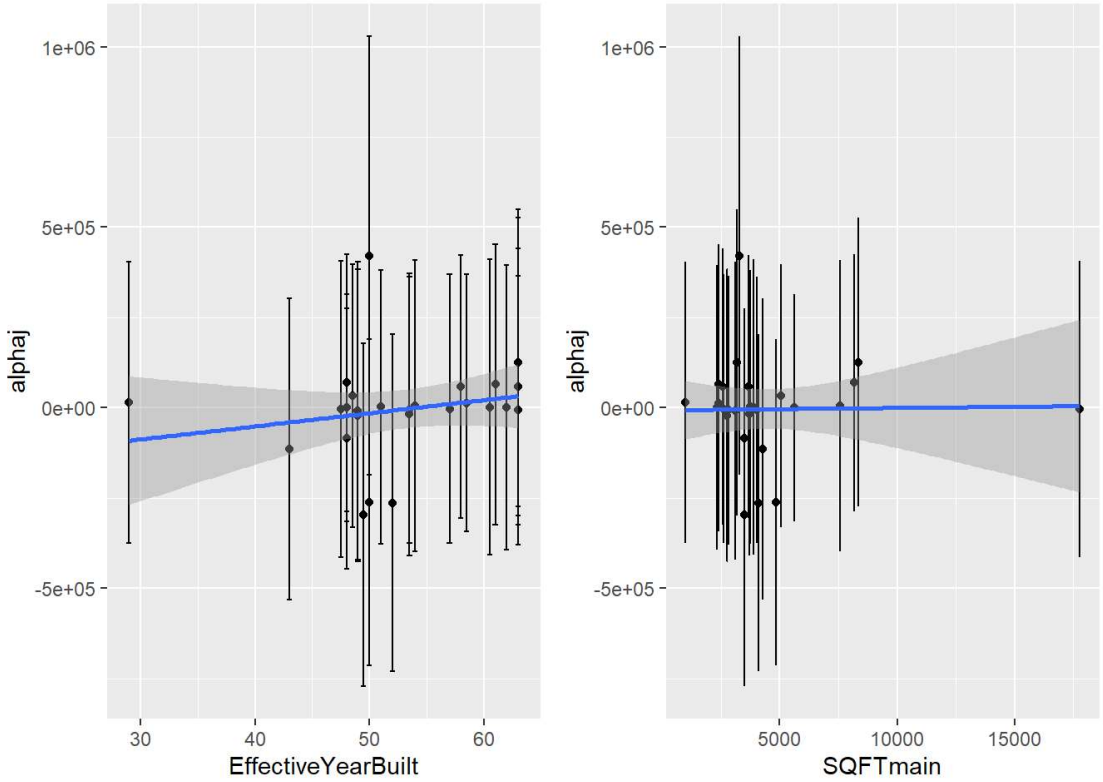
```
plotAE <- plot_alpha(fit.model5, CT_summary5, "EffectiveYearBuilt")
```

```
## [1] "the fluctuation of alpha j lies between 8.57749619241409 and "
## [1] "128.57460048928 % of property value"
```

```
plotAS <- plot_alpha(fit.model5, CT_summary5, "SQFTmain")
```

```
## [1] "the fluctuation of alpha j lies between 8.57749619241409 and "
## [1] "128.57460048928 % of property value"
```

```
grid.arrange(plotAE, plotAS, nrow = 1)
```



All the graphs above show the varying alpha for different groups and they are all demeaned. According to the information in the graph, the intercept estimates are fluctuating around the median with high level of uncertainty, and they are quite stable when individual level predictors increase. For model 4.a, the alpha j estimate fluctuates between 9 and 161 % of grouped median property value. In comparison, model 4.b has higher precision of alpha j estimate (with fluctuation between 8 and 131 % of grouped median property value)

Response 6.b

```
lambda <- function(fit) {
  eta <- as.data.frame(fit) %>%
    dplyr::select(starts_with('b[(Intercept)'))
  numerator <- var(colMeans(eta))
  denominator <- mean(apply(eta, 1, var))
  1 - numerator / denominator
}

lambda(fit.model1);lambda(fit.model5)
```

```
## [1] 0.8249176
```

```
## [1] 0.8781482
```

```
#lambda(fit.model2);lambda(fit.model6)
```

The pooling factor for model 4.b (0.88) is higher than 4.a (0.82) which leads to higher precision of the overall estimation and prediction.

Question 7

Code and Response 7:

```
options(mc.cores = 1)
loo_pool <- loo(fit.pool)
loo_model1 <- loo(fit.model1)
loo_model2 <- loo(fit.model2)
loo_model5 <- loo(fit.model5)
loo_model6 <- loo(fit.model6)
loo_unpool <- loo(fit.unpool)

loo_compare(loo_pool, loo_model1, loo_model5, loo_unpool)
```

```
##            elpd_diff se_diff
## fit.pool        0.0       0.0
## fit.model5     -2.2       3.6
## fit.model1     -6.5       7.7
## fit.unpool -20854.8     449.3
```

I apply the leave-one-out strategy with all of the models fitted for cross validation and find that model 4.b has the least deviation and highest precision of all of the models. Pooled model, model 4.a and model 4.b are relatively not very sensitive to the one left out sample in the model fitting.

```
R2_pool <- quantile(bayes_R2(fit.pool), c(.25, .5, .75))
R2_model1 <- quantile(bayes_R2(fit.model1), c(.25, .5, .75))
R2_model2 <- quantile(bayes_R2(fit.model2), c(.25, .5, .75))
R2_model5 <- quantile(bayes_R2(fit.model5), c(.25, .5, .75))
R2_model6 <- quantile(bayes_R2(fit.model6), c(.25, .5, .75))
R2_unpool <- quantile(bayes_R2(fit.unpool), c(.25, .5, .75))
```

```
R2_pool
```

```
##       25%       50%       75%
## 0.8037963 0.8078175 0.8117925
```

```
R2_model1
```

```
##       25%       50%       75%
## 0.8044380 0.8085325 0.8124670
```

```
R2_model5
```

```
##       25%       50%       75%
## 0.8045889 0.8087365 0.8125871
```

```
R2_unpool
```

```
##          25%          50%          75%
## 2.214224e-07 5.437015e-07 1.295812e-06
```

The R square value demonstrates the explained variances in all of the models. It is clear that the R square values are stable among sampling (between 25%, 50% and 75% percentile) in pooled model, model 4.a and 4.b. The R square value is low in unpooled model due to the fluctuation caused by collinearity. Overall, the R square of model 4.b is slightly higher than other models due to the fact that more variance in group level is explained by including more group level variables.

Question 8

Response 8:

There are several findings worth mentioning. First of all, according to the result in question 3 and 4, individual level predictor building square footage is positively correlated and building age is negatively correlated to the property value. These effects are reasonable since the housing quality and space are factors that drive the property value. However, there is positive effect of unemployment rate to property value which is against our intuition. This finding might be attributed to the fact that we only include commercial properties in this research. In Los Angeles, the function of a given block (or census tract) might be very specific, neighborhoods are formed with either high proportion of commercial properties or residential properties. The residential-commercial mixed neighborhoods are not common, especially in traditional cities in the United States. Thus, region with high-valued commercial properties tend to hire people who do not reside in the region and the unemployment rate is calculated among residents who are not able to afford sub-urban housing with better living qualities.

Secondly, the variance of within group variance dominates both model 4.a and 4.b that σ_y is much more higher than σ_{CT} and σ_{region}. Thus, adding group predictors might increase the precision of the model in estimation and prediction (according to the cross validation and graph from question 4, 6 and 7), but the varying intercept estimates are still very uncertain. Besides, according to the finding from appendix b, the coefficient estimate in group level for all varying slope (γ) are all 0. The influence of group factors in these models is relatively limited. To conclude, although we can observe divergence in property values and demographic patterns, housing

characteristics are the main factors which drive the property value. However, given the fact that only commercial properties are included in this research, the results of model fit might be very different for residential properties since the function of the property is more certain and the sample size is larger.

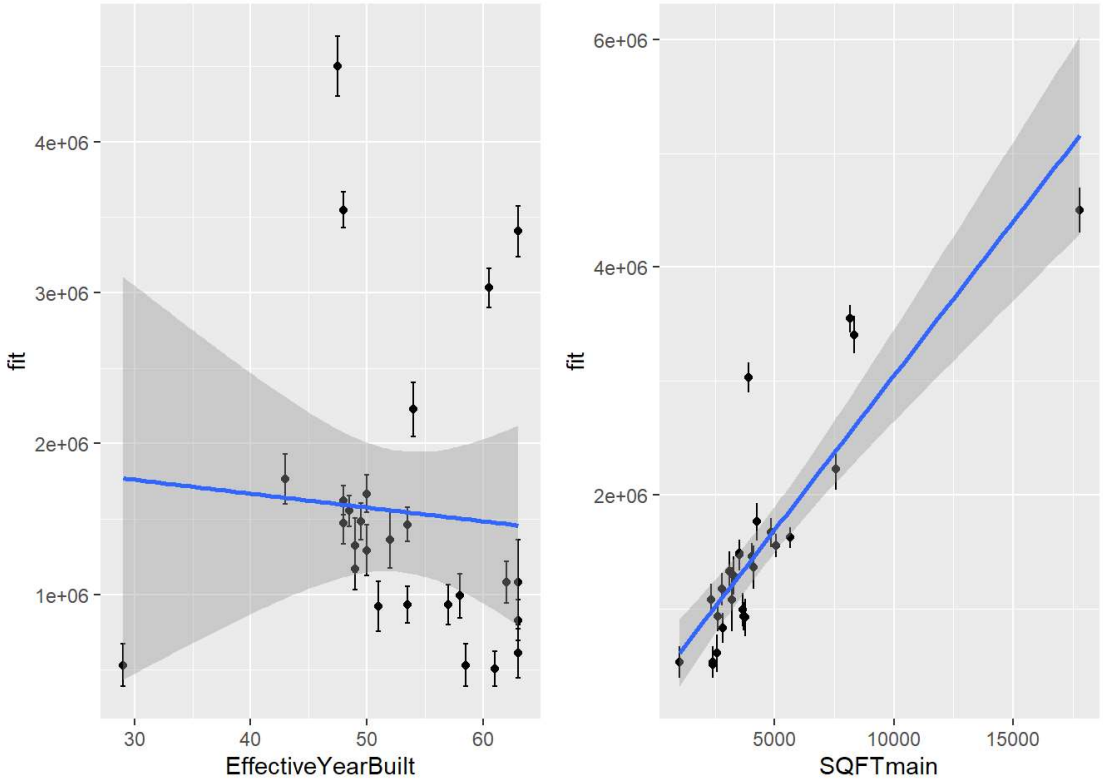
Appendix a: Model Overfitting

Alternative Model 4.a (with 2017 assessed value)

```
fit.model2
```

```
## stan_glmer
##  family:       gaussian [identity]
##  formula:      TotalValue ~ EffectiveYearBuilt + SQFTmain + Bedrooms + Bathrooms +
##     (1 | CT) + unemployment + median_income + Travel_Time701902 +
##     TotalValue2017
##  observations: 2156
## ------
##                     Median   MAD_SD
## (Intercept)        424639.6 473737.4
## EffectiveYearBuilt  -3939.7   3272.5
## SQFTmain              -30.7      3.6
## Bedrooms            -7860.9  12171.7
## Bathrooms           18411.7  13318.2
## unemployment        18218.4  31106.0
## median_income          -0.4      3.9
## Travel_Time701902    -260.6    310.7
## TotalValue2017          1.1      0.0
##
## Auxiliary parameter(s):
##       Median    MAD_SD
## sigma 2776785.8   41039.8
##
## Error terms:
##  Groups   Name        Std.Dev.
##  CT       (Intercept)  116789
##  Residual             2777816
## Num. levels: CT 27
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

```
CT_summary2 <- parcel %>% group_by(CT) %>%
               summarize(EffectiveYearBuilt=median(EffectiveYearBuilt),
                         SQFTmain=median(SQFTmain),
                         Bedrooms=median(Bedrooms),
                         Bathrooms=median(Bathrooms),
                         unemployment=median(unemployment),
                         median_income=median(median_income),
                         Travel_Time701902=median(Travel_Time701902),
                         TotalValue2017=median(TotalValue2017))

new_data2 <- data.frame(CT=CT_summary2$CT,
                        EffectiveYearBuilt=CT_summary2$EffectiveYearBuilt,
                        SQFTmain=CT_summary2$SQFTmain,
                        Bedrooms=CT_summary2$Bedrooms,
                        Bathrooms=CT_summary2$Bathrooms,
                        unemployment=CT_summary2$unemployment,
                        median_income=CT_summary2$median_income,
                        Travel_Time701902=CT_summary2$Travel_Time701902,
                        TotalValue2017=CT_summary2$TotalValue2017)

pred2 <- posterior_linpred(fit.model2, newdata=new_data2)
plotE <- plot_pred(new_data2, pred2, "EffectiveYearBuilt")
plotS <- plot_pred(new_data2, pred2, "SQFTmain")
grid.arrange(plotE, plotS, nrow = 1)
```



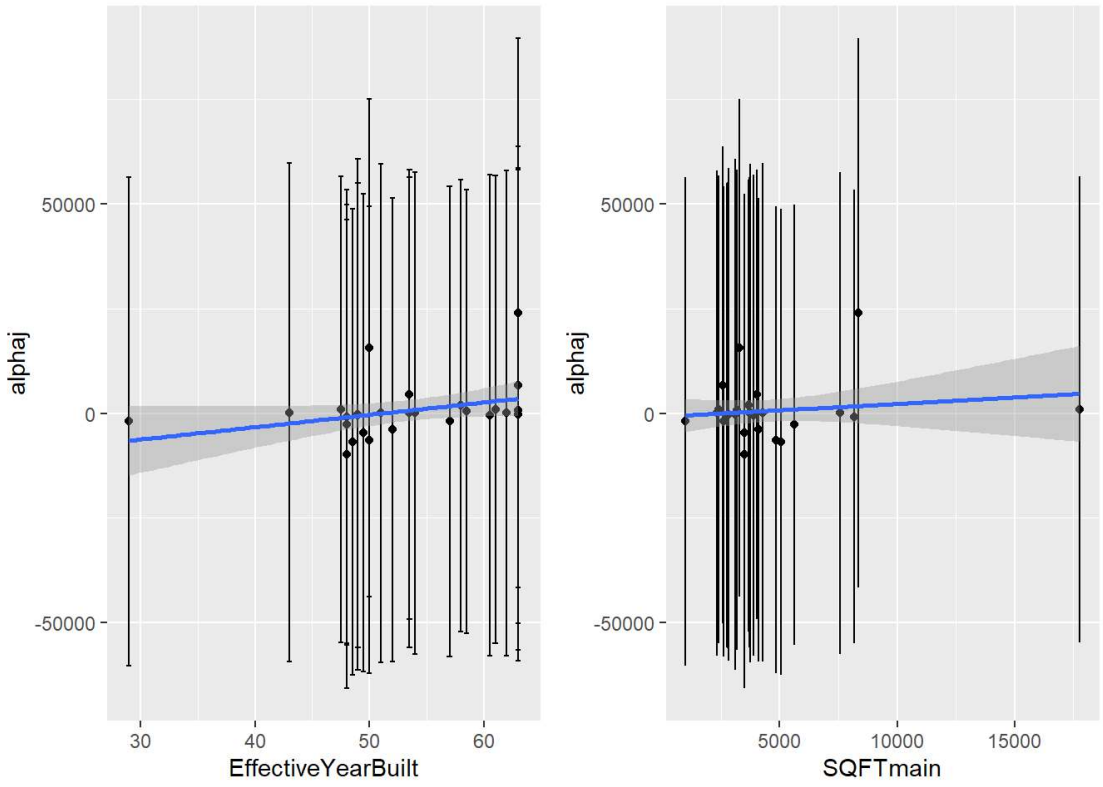

```
plotAE <- plot_alpha(fit.model2, CT_summary2, "EffectiveYearBuilt")
```

```
## [1] "the fluctuation of alpha j lies between 1.16288334142815 and "
## [1] "19.2460804260848 % of property value"
```

```
plotAS <- plot_alpha(fit.model2, CT_summary2, "SQFTmain")
```

```
## [1] "the fluctuation of alpha j lies between 1.16288334142815 and "
## [1] "19.2460804260848 % of property value"
```

```
grid.arrange(plotAE, plotAS, nrow = 1)
```

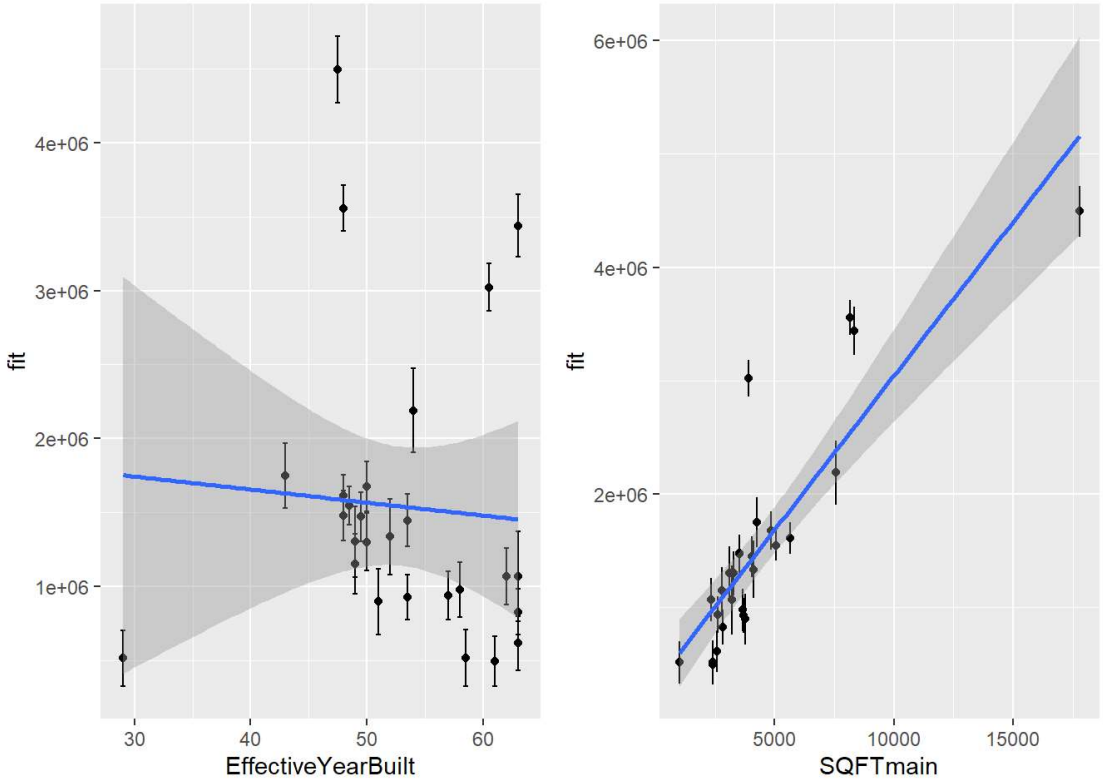


Alternative Model 4.b (with another level of group effect and with 2017 assessed value)

```
fit.model6
```

```
## stan_glmer
##  family:       gaussian [identity]
##  formula:      TotalValue ~ EffectiveYearBuilt + SQFTmain + Bedrooms + Bathrooms +
##     unemployment + median_income + Travel_Time701902 + TotalValue2017 +
##     (1 | CT) + (1 | travel_dregion)
##  observations: 2156
## ------
##                    Median   MAD_SD
## (Intercept)        463583.7 579444.8
## EffectiveYearBuilt  -3948.7   3326.9
## SQFTmain              -30.7      3.7
## Bedrooms            -7802.7  12755.9
## Bathrooms           18222.4  13455.0
## unemployment        16590.2  34089.1
## median_income          -0.4      4.0
## Travel_Time701902    -333.9    449.9
## TotalValue2017          1.1      0.0
##
## Auxiliary parameter(s):
##       Median    MAD_SD
## sigma 2777799.2   41156.4
##
## Error terms:
##  Groups         Name        Std.Dev.
##  CT             (Intercept)  123100
##  travel_dregion (Intercept)  303075
##  Residual                    2778555
## Num. levels: CT 27, travel_dregion 4
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

```
CT_summary6 <- parcel %>% group_by(CT) %>%
  summarize(EffectiveYearBuilt=median(EffectiveYearBuilt),
            SQFTmain=median(SQFTmain),
            Bedrooms=median(Bedrooms),
            Bathrooms=median(Bathrooms),
            unemployment=median(unemployment),
            median_income=median(median_income),
            Travel_Time701902=median(Travel_Time701902),
            TotalValue2017=median(TotalValue2017))

new_data6 <- data.frame(CT=CT_summary6$CT,
                        EffectiveYearBuilt=CT_summary6$EffectiveYearBuilt,
                        SQFTmain=CT_summary6$SQFTmain,
                        Bedrooms=CT_summary6$Bedrooms,
                        Bathrooms=CT_summary6$Bathrooms,
                        unemployment=CT_summary6$unemployment,
                        median_income=CT_summary6$median_income,
                        Travel_Time701902=CT_summary6$Travel_Time701902,
                        TotalValue2017=CT_summary6$TotalValue2017,
                        travel_dregion=2)

pred6 <- posterior_linpred(fit.model6, newdata=new_data6)
plotE <- plot_pred(new_data6, pred6, "EffectiveYearBuilt")
plotS <- plot_pred(new_data6, pred6, "SQFTmain")
grid.arrange(plotE, plotS, nrow = 1)
```



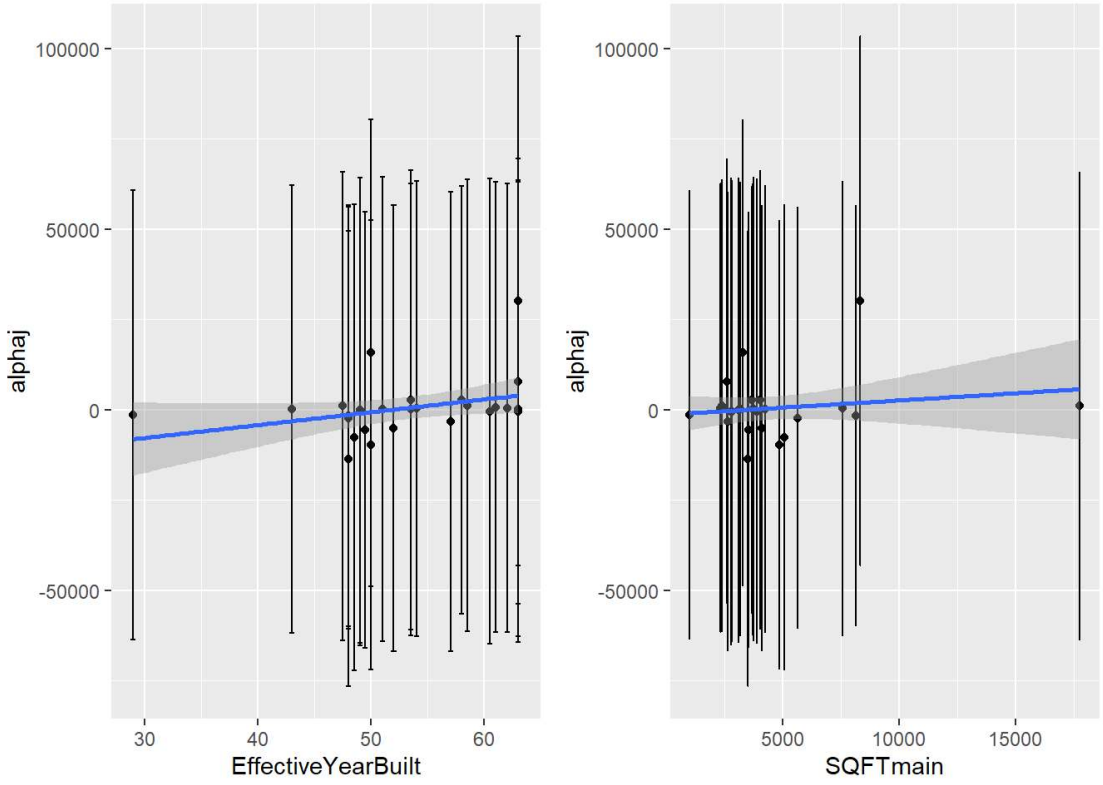

```
plotAE <- plot_alpha(fit.model6, CT_summary6, "EffectiveYearBuilt")
```

```
## [1] "the fluctuation of alpha j lies between 1.35449642701582 and "
## [1] "20.5377982972899 % of property value"
```

```
plotAS <- plot_alpha(fit.model6, CT_summary6, "SQFTmain")
```

```
## [1] "the fluctuation of alpha j lies between 1.35449642701582 and "
## [1] "20.5377982972899 % of property value"
```

```
grid.arrange(plotAE, plotAS, nrow = 1)
```



```
loo_compare(loo_model1, loo_model2, loo_model5, loo_model6)
```

```
##            elpd_diff se_diff
## fit.model2     0.0       0.0
## fit.model6    -5.3       2.5
## fit.model5 -2376.7     318.8
## fit.model1 -2381.0     322.1
```

```
lambda(fit.model1);lambda(fit.model2);lambda(fit.model5);lambda(fit.model6)
```

```
## [1] 0.8249176
```

```
## [1] 0.9638304
```

```
## [1] 0.8781482
```

```
## [1] 0.9666887
```

The above model (model 2 and model 6) include the previous property value (2017) as one of the predictor, we can observe that in both model, the precision of the prediction increase significantly. However, we might also face the problem of overfitting since property value of previous year are correlated to almost all of other predictors and do not provide information we need for coefficients interpretation.

Appendix b: Varying Slope Model Fitting

```
#Multilevel intercept and slope (see Appendix.b for more information)
#fit.model3 <- stan_glmer(TotalValue ~ EffectiveYearBuilt + SQFTmain + Bedrooms + Bathrooms
#                         + (1 + TotalValue2017| CT) + unemployment
#                         + median_income + Travel_Time701902 + TotalValue2017
#                         + TotalValue2017:unemployment + TotalValue2017:median_income
#                         + TotalValue2017:Travel_Time701902,
#                         data=parcel)

#all zero effect is problematic
#CT_summary3 <- parcel %>% group_by(CT) %>%
#               summarize(EffectiveYearBuilt=median(EffectiveYearBuilt),
#                         SQFTmain=median(SQFTmain),
#                         Bedrooms=median(Bedrooms),
#                         Bathrooms=median(Bathrooms),
#                         unemployment=median(unemployment),
#                         median_income=median(median_income),
#                         Travel_Time701902=median(Travel_Time701902),
#                         TotalValue2017=median(TotalValue2017))

#new_data3 <- data.frame(CT=CT_summary3$CT,
#                        EffectiveYearBuilt=CT_summary3$EffectiveYearBuilt,
#                        SQFTmain=CT_summary3$SQFTmain,
#                        Bedrooms=CT_summary3$Bedrooms,
#                        Bathrooms=CT_summary3$Bathrooms,
#                        unemployment=CT_summary3$unemployment,
#                        median_income=CT_summary3$median_income,
#                        Travel_Time701902=CT_summary3$Travel_Time701902,
#                        TotalValue2017=CT_summary3$TotalValue2017)

#plot_pred(fit.model3, new_data3, "median_income")
#plot_pred(fit.model3, new_data3, "Travel_Time701902")
```

Due to the time constraint, the analysis of the varying intercept model is not included in the main report. However, according to the result, the coefficient estimates in group level for varying slope estimate are all very close to 0.

Reference

PPHA41420 Multilevel Regression Modeling for Public Policy Course Materials from Eric Potash, The University of Chicago

American Community Survey (https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml (https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml))

Uber Movement Data (https://movement.uber.com/cities?lang=en-US (https://movement.uber.com/cities?lang=en-US))

Assessor Parcels Data from Los Angeles County Open Data Portal (https://data.lacounty.gov/Parcel-/Assessor-Parcels-Data-2006-thru-2019/9trm-uz8i (https://data.lacounty.gov/Parcel-/Assessor-Parcels-Data-2006-thru-2019/9trm-uz8i))