STAT 306 2024W2 - Project: Report

Group Project Report

Outline

1. fit linear model with all covariates
2. multi-collinearity check correlation among covariates (VIF)
3. assumptions
4. residual pattern -> transformation
5. model selection (forward R^2, backward max p-value, CP)
6. interaction term
7. (optional) outlier, leverage, influence
8. anova

Data Preparation

The response variable 'SalePrice' column is divided by 10,000, so the unit for response variable is 10 thousand dollar.

```
library(car)
```

```
Loading required package: carData
```

```
library(leaps)
library(knitr)
library(kableExtra)
library(olsrr)
```

```
Attaching package: 'olsrr'
```

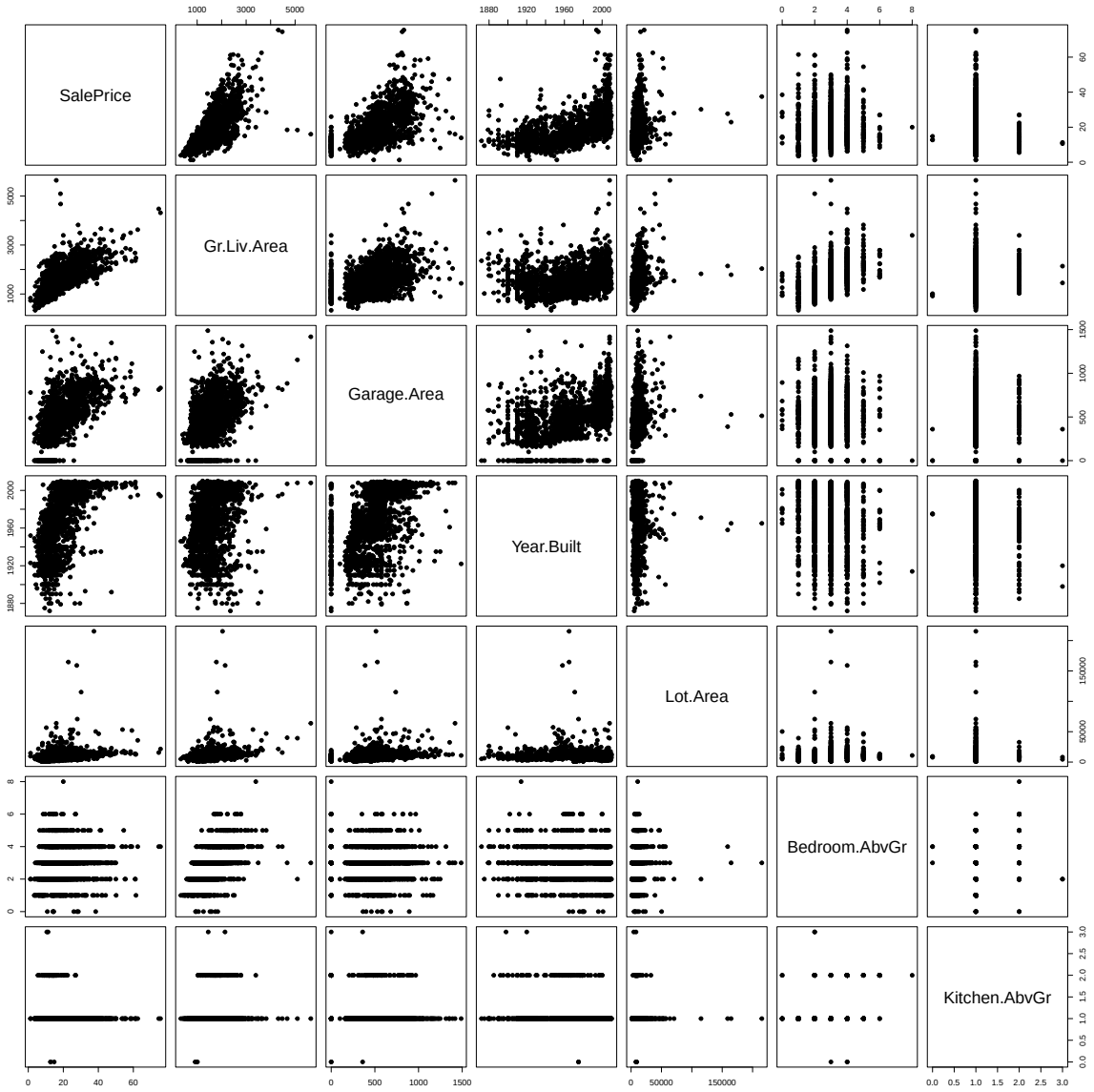
```
The following object is masked from 'package:datasets':

    rivers
```

```
# options(repr.matrix.max.rows = 20)

ames_housing <- read.table("https://raw.githubusercontent.com/AllenCheng5186/STAT306-G14-Grou

ames_housing$SalePrice <- ames_housing$SalePrice / 10000

house = ames_housing[, c("SalePrice","Overall.Qual", "Gr.Liv.Area", "Garage.Area",
                         "Year.Built", "Lot.Area", "Bedroom.AbvGr", "Kitchen.AbvGr", "Lot.Sha

house[sapply(house, is.character)] <- lapply(house[sapply(house, is.character)], as.factor)
house$Overall.Qual <- as.factor(house$Overall.Qual)
house[is.na(house)] <- 0

head(house)
```

```
  SalePrice Overall.Qual Gr.Liv.Area Garage.Area Year.Built Lot.Area
1     21.50            6        1656         528       1960    31770
2     10.50            5         896         730       1961    11622
3     17.20            6        1329         312       1958    14267
4     24.40            7        2110         522       1968    11160
5     18.99            5        1629         482       1997    13830
6     19.55            6        1604         470       1998     9978
  Bedroom.AbvGr Kitchen.AbvGr Lot.Shape
1             3             1       IR1
2             2             1       Reg
3             3             1       IR1
4             3             1       Reg
5             3             1       IR1
6             3             1       IR1
```

Based on the pairwise scatterplots for the continuous variables in the data (all variables except `Overall.Qual` and `Lot.Shape`), none of the plots show obvious linear patterns and so there do not appear to be strong linear associations.

```
options(repr.plot.width = 15, repr.plot.height = 15)

pairs(house[,-c(2, 9, 11)], pch=19)
```



There does not appear to be a strong correlation between any pair of continuous variables.

```
round(cor(house[,!(names(house) %in% c("Overall.Qual","Lot.Shape"))]), 3)
```

```
              SalePrice Gr.Liv.Area Garage.Area Year.Built Lot.Area
SalePrice         1.000       0.707       0.640      0.558    0.267
Gr.Liv.Area       0.707       1.000       0.484      0.242    0.286
Garage.Area       0.640       0.484       1.000      0.481    0.213
Year.Built        0.558       0.242       0.481      1.000    0.023
```

```
Lot.Area          0.267       0.286       0.213      0.023    1.000
Bedroom.AbvGr     0.144       0.517       0.073     -0.055    0.137
Kitchen.AbvGr    -0.120       0.118      -0.058     -0.138   -0.020
              Bedroom.AbvGr Kitchen.AbvGr
SalePrice             0.144        -0.120
Gr.Liv.Area           0.517         0.118
Garage.Area           0.073        -0.058
Year.Built           -0.055        -0.138
Lot.Area              0.137        -0.020
Bedroom.AbvGr         1.000         0.241
Kitchen.AbvGr         0.241         1.000
```

calculate VIF to verify that there is not strong correlation between any continuous variables.

why use VIF? What is the advantage of using VIF? Using thumb theory of 10 as cutoff, since every continuous variable does not over 10, so no multi-collinearity.

explain what is multi-collinearity -> any correlation among explanatory variables (X)

```r
house_numeric_only = house[, sapply(house, is.numeric)]
house_numeric_lreg = lm(SalePrice ~ ., data = house_numeric_only)

vif(house_numeric_lreg)
```

```
  Gr.Liv.Area   Garage.Area    Year.Built      Lot.Area Bedroom.AbvGr
     1.961712      1.667759      1.353542      1.111787      1.515821
Kitchen.AbvGr
     1.085901
```

Fit full model (using all covariates)

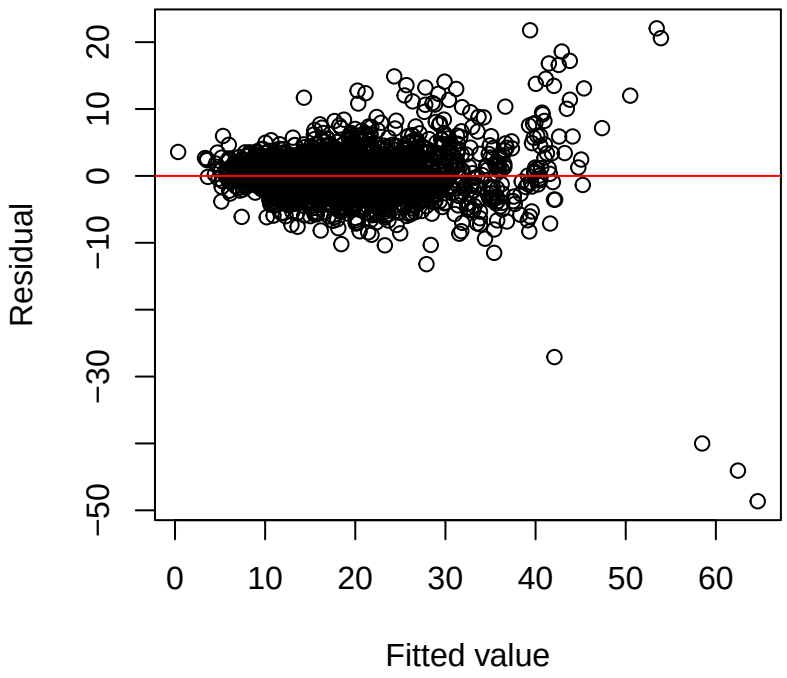
```r
full_linear_reg = lm(SalePrice~.,data = house)

summary(full_linear_reg)
```

```
Call:
lm(formula = SalePrice ~ ., data = house)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-48.642  -1.526  -0.078   1.327  22.062

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -8.731e+01  5.578e+00 -15.652  < 2e-16 ***
Overall.Qual2   2.341e+00  1.883e+00   1.243 0.213813
Overall.Qual3   3.916e+00  1.728e+00   2.266 0.023508 *
Overall.Qual4   4.889e+00  1.663e+00   2.940 0.003305 **
Overall.Qual5   6.367e+00  1.654e+00   3.850 0.000121 ***
Overall.Qual6   6.997e+00  1.657e+00   4.223 2.48e-05 ***
Overall.Qual7   8.628e+00  1.664e+00   5.186 2.29e-07 ***
Overall.Qual8   1.284e+01  1.675e+00   7.665 2.42e-14 ***
Overall.Qual9   2.036e+01  1.702e+00  11.962  < 2e-16 ***
Overall.Qual10  2.360e+01  1.793e+00  13.160  < 2e-16 ***
Gr.Liv.Area     5.988e-03  2.037e-04  29.403  < 2e-16 ***
Garage.Area     3.672e-03  3.758e-04   9.771  < 2e-16 ***
Year.Built      4.553e-02  2.720e-03  16.737  < 2e-16 ***
Lot.Area        8.676e-05  8.589e-06  10.102  < 2e-16 ***
Bedroom.AbvGr  -4.930e-01  9.624e-02  -5.122 3.22e-07 ***
Kitchen.AbvGr  -2.440e+00  3.030e-01  -8.055 1.15e-15 ***
Lot.ShapeIR2    3.486e-01  3.978e-01   0.876 0.380939
Lot.ShapeIR3   -3.693e+00  8.484e-01  -4.353 1.39e-05 ***
Lot.ShapeReg   -4.467e-01  1.386e-01  -3.224 0.001278 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.29 on 2911 degrees of freedom
Multiple R-squared:  0.8315,    Adjusted R-squared:  0.8304
F-statistic: 797.9 on 18 and 2911 DF,  p-value: < 2.2e-16
```

verify assumption: residual vs fitted value plot

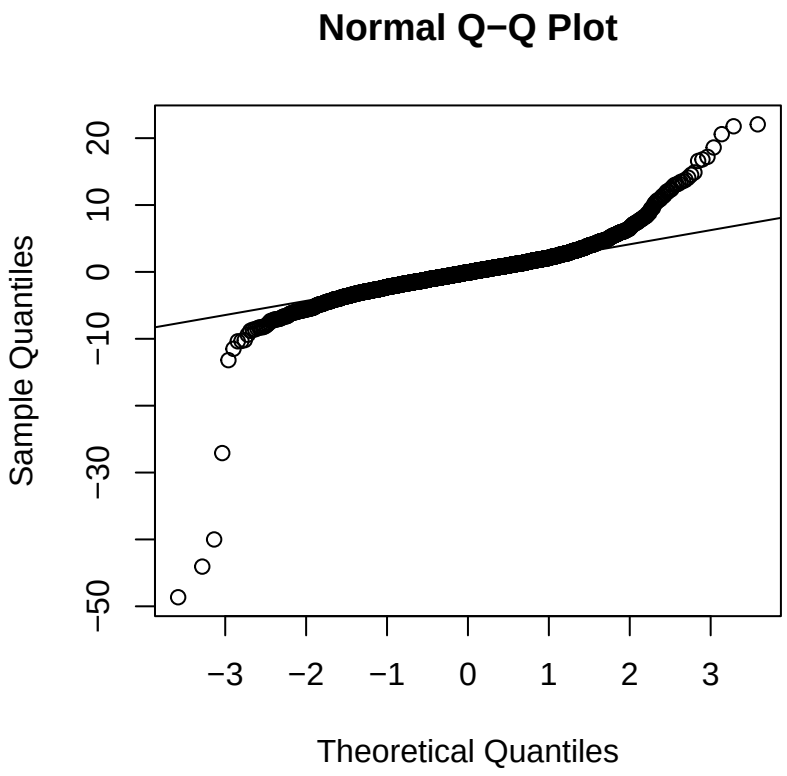
```r
# options(repr.plot.width = 7, repr.plot.height = 7)

plot(full_linear_reg$fitted.values, full_linear_reg$residuals,
     xlab="Fitted value", ylab="Residual")
abline(h = 0, col = "red")
```



funnel pattern appear

linear regression assumption: homoscedasticity $Var(\Sigma) = \sigma^2$ violated

```
# options(repr.plot.width = 7, repr.plot.height = 7)

qqnorm(full_linear_reg$residuals)
qqline(full_linear_reg$residuals)
```



heavy tail observed -> homoscedasticity $Var(\Sigma) = \sigma^2$ violated

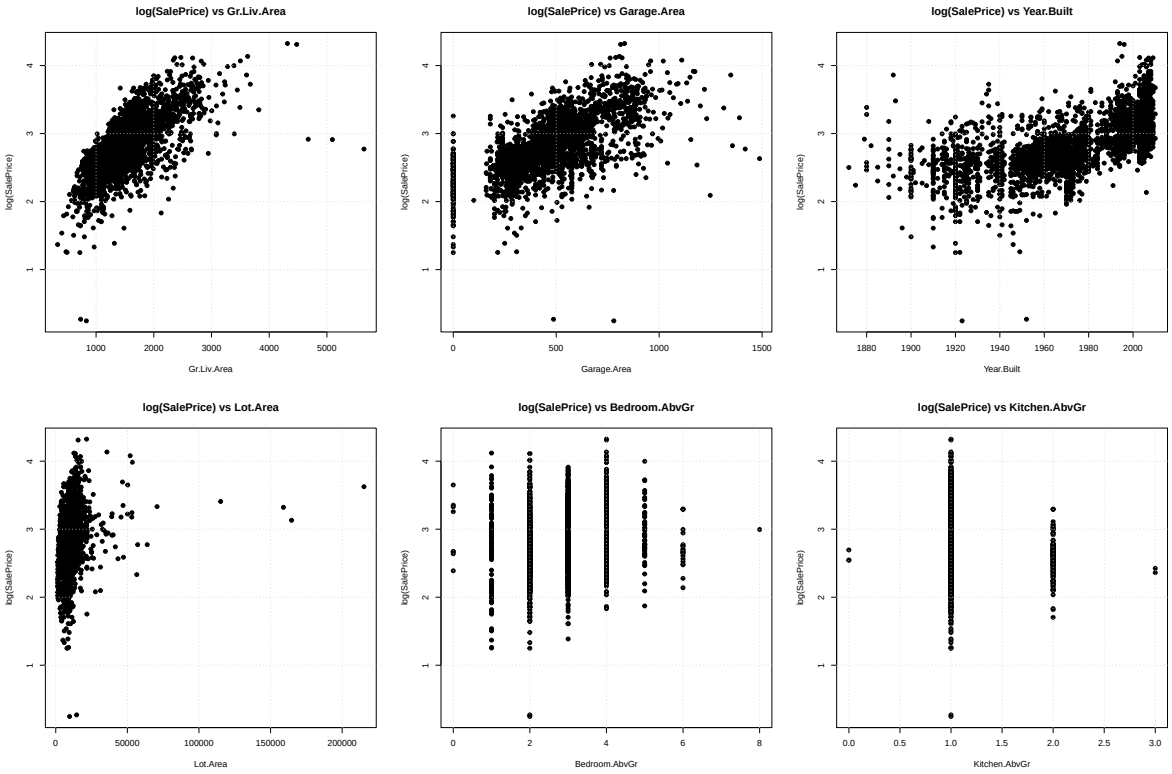
Log(Y) Transformation

take log on response variable Y, log(Y)

create scatterplot between log(Y) and each continuous variables of X

```
num_vars <- names(house)[sapply(house, is.numeric)]
predictors <- setdiff(num_vars, c("SalePrice", "log_SalePrice"))

# options(repr.plot.width = 15, repr.plot.height = 10)
par(mfrow = c(2, 3))

for (var in predictors) {
  plot(house[[var]], log(house$SalePrice), pch=19,
       xlab = var, ylab = "log(SalePrice)",
       main = paste("log(SalePrice) vs", var))
```

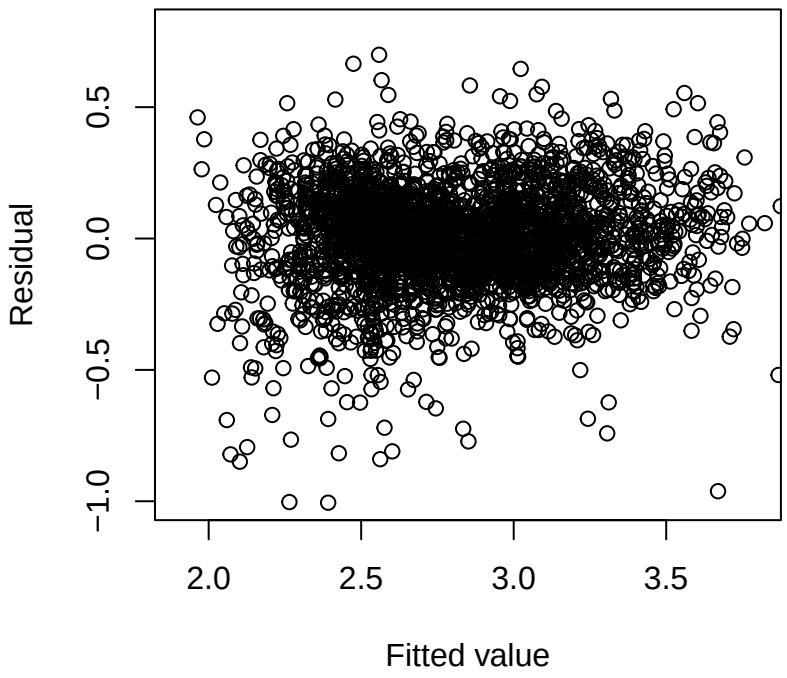
```
  grid()
}
```



every pair show linear relationship, except for `Year.Built` that still curved.

fit linear model again with log(Y) as response variable

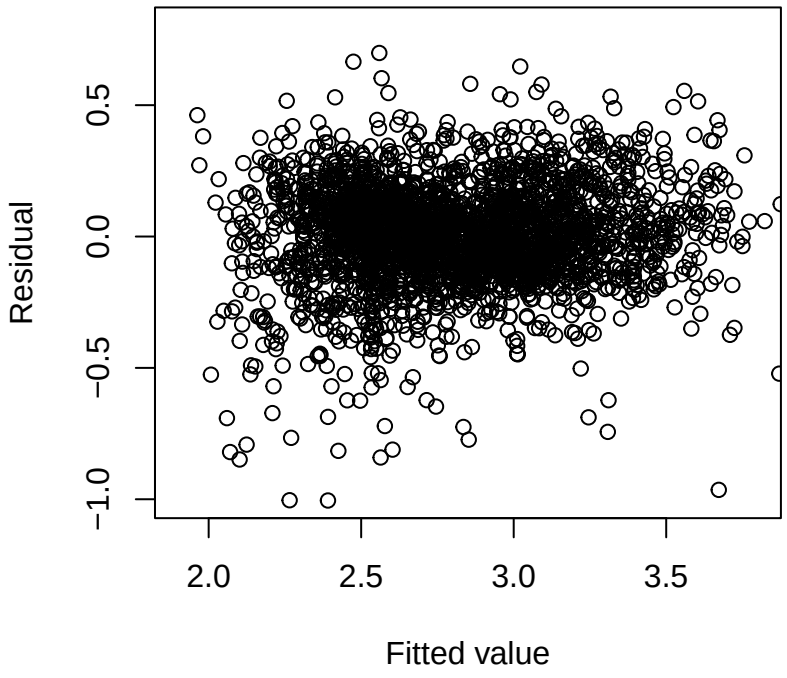
```
# options(repr.plot.width = 7, repr.plot.height = 7)

reg_logY = lm(log(SalePrice) ~ Gr.Liv.Area+ Garage.Area + Year.Built + Lot.Area +
              Bedroom.AbvGr + Kitchen.AbvGr, data = house)

plot(x = reg_logY$fitted.values, y = reg_logY$residuals,
     xlim = c(1.9, 3.8), ylim =c(-1.0, 0.8),
     xlab="Fitted value", ylab="Residual")
```



looks random -> assumptions aligned! -> conclusion will be valid

since `Year.Built` is curved which could have quadratic relationship with response variable, so fit log(Year.Built) into model again

```
reg_logY_logYear = lm(log(SalePrice) ~ Gr.Liv.Area+ Garage.Area + log(Year.Built) +
                        Lot.Area + Bedroom.AbvGr + Kitchen.AbvGr, data = house)

plot(x = reg_logY_logYear$fitted.values, y = reg_logY_logYear$residuals,
     xlim = c(1.9, 3.8), ylim =c(-1.0, 0.8),  #TODO outlier in residual plot
     xlab="Fitted value", ylab="Residual")
```




```
house$log_SalePrice = log(house$SalePrice)
house$log_Year.Built = log(house$Year.Built)

log_full_model = lm(log_SalePrice ~ Overall.Qual + Gr.Liv.Area + Garage.Area +
                    log_Year.Built + Lot.Area + Bedroom.AbvGr + Kitchen.AbvGr +
                    Lot.Shape, data = house)

summary(log_full_model)
```

```
Call:
lm(formula = log_SalePrice ~ Overall.Qual + Gr.Liv.Area + Garage.Area +
    log_Year.Built + Lot.Area + Bedroom.AbvGr + Kitchen.AbvGr +
    Lot.Shape, data = house)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-1.95901 -0.07910  0.00380  0.09036  0.67473

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4.239e+01  2.036e+00 -20.820  < 2e-16 ***
Overall.Qual2   2.970e-01  9.498e-02   3.127  0.00178 **
Overall.Qual3   7.180e-01  8.716e-02   8.237 2.64e-16 ***
Overall.Qual4   8.874e-01  8.388e-02  10.579  < 2e-16 ***
Overall.Qual5   1.060e+00  8.343e-02  12.700  < 2e-16 ***
Overall.Qual6   1.124e+00  8.358e-02  13.447  < 2e-16 ***
Overall.Qual7   1.213e+00  8.392e-02  14.450  < 2e-16 ***
Overall.Qual8   1.357e+00  8.449e-02  16.057  < 2e-16 ***
Overall.Qual9   1.547e+00  8.588e-02  18.008  < 2e-16 ***
Overall.Qual10  1.470e+00  9.046e-02  16.248  < 2e-16 ***
Gr.Liv.Area     2.929e-04  1.028e-05  28.504  < 2e-16 ***
Garage.Area     2.343e-04  1.896e-05  12.355  < 2e-16 ***
log_Year.Built  5.756e+00  2.684e-01  21.441  < 2e-16 ***
Lot.Area        4.453e-06  4.333e-07  10.278  < 2e-16 ***
Bedroom.AbvGr  -1.815e-02  4.855e-03  -3.737  0.00019 ***
Kitchen.AbvGr  -1.216e-01  1.529e-02  -7.957 2.50e-15 ***
Lot.ShapeIR2    5.635e-03  2.007e-02   0.281  0.77892
Lot.ShapeIR3   -2.118e-01  4.280e-02  -4.949 7.90e-07 ***
Lot.ShapeReg   -3.215e-02  6.991e-03  -4.598 4.44e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.166 on 2911 degrees of freedom
Multiple R-squared:  0.8352,    Adjusted R-squared:  0.8342
F-statistic: 819.7 on 18 and 2911 DF,  p-value: < 2.2e-16
```

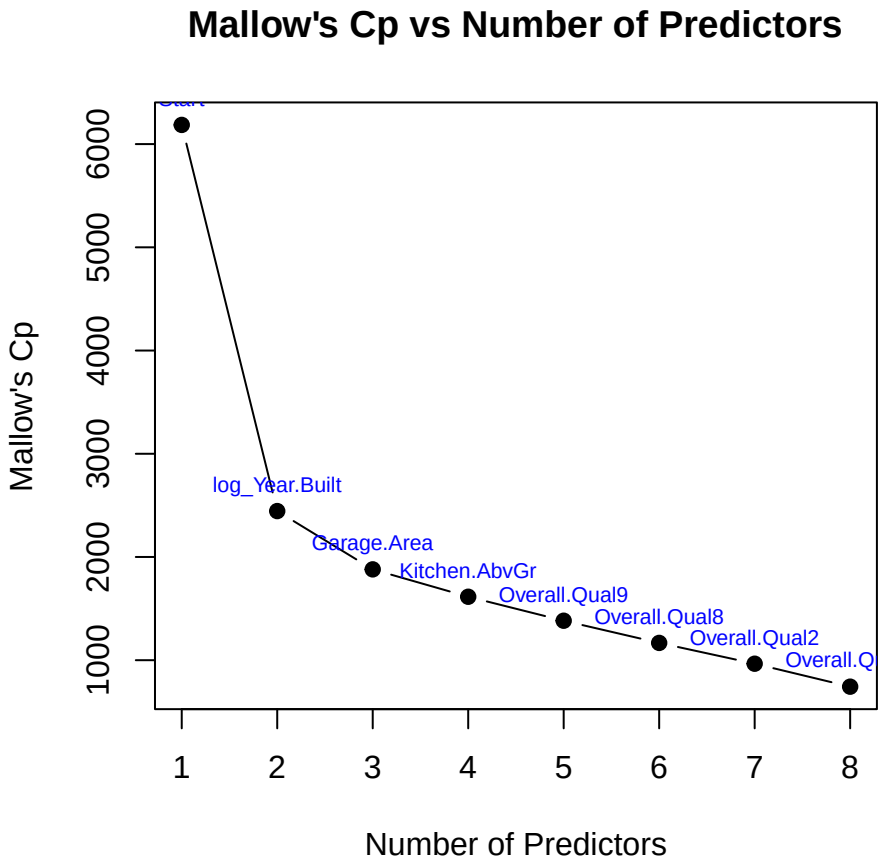
Model Selection

two forwarding selection functions from two different package have different result with categorical covariates.

I suggest we check with prof to decide which one use.

```
s <- regsubsets(log_SalePrice ~ Overall.Qual + Gr.Liv.Area + Garage.Area +
                    log_Year.Built + Lot.Area + Bedroom.AbvGr + Kitchen.AbvGr +
                    Lot.Shape, data = house, method = "forward")

ss <- summary(s)
```

```r
# s_summary <- as.data.frame(summary(s)$outmat)
# kable(s_summary, format = "latex", booktabs = TRUE) %>%
#   kable_styling(latex_options = c("scale_down", "hold_position"))

cps <- ss$cp
num_predictors <- apply(ss$which, 1, function(x) sum(x) - 1)

model_matrix <- ss$which[, -1, drop = FALSE]
model_names <- colnames(model_matrix)

added_vars <- character(nrow(model_matrix))
for (i in 2:nrow(model_matrix)) {
  prev <- model_matrix[i - 1, ]
  curr <- model_matrix[i, ]
  new_var <- setdiff(model_names[curr & !prev], model_names[prev])
  added_vars[i] <- ifelse(length(new_var) > 0, new_var, "")
}
added_vars[1] <- "Start"

plot(num_predictors, cps, type = "b", pch = 19,
     xlab = "Number of Predictors", ylab = "Mallow's Cp",
     main = "Mallow's Cp vs Number of Predictors")
text(num_predictors, cps, labels = added_vars, pos = 3, cex = 0.7, col = "blue")
abline(0, 1, col = "red", lty = 2)
```




```
forward_sel <- ols_step_forward_r2(log_full_model)

forward_sel$metrics
```

```
  step       variable        r2    adj_r2        aic        sbc       sbic
1    1   Overall.Qual 0.6891653 0.6882072  -347.0641  -281.2538  -8680.345
2    2    Gr.Liv.Area 0.7657285 0.7649260 -1173.6115 -1101.8185  -9506.514
3    3 log_Year.Built 0.8077547 0.8070300 -1750.8976 -1673.1217 -10083.013
4    4    Garage.Area 0.8199908 0.8192503 -1941.5867 -1857.8281 -10273.365
5    5       Lot.Area 0.8275419 0.8267731 -2065.1483 -1975.4070 -10396.591
6    6  Kitchen.AbvGr 0.8318250 0.8310173 -2136.8353 -2041.1112 -10468.009
7    7      Lot.Shape 0.8344302 0.8334636 -2176.5777 -2062.9053 -10511.530
8    8  Bedroom.AbvGr 0.8352208 0.8342019 -2188.6025 -2068.9473 -10523.465
```

But top four covariates output are same

1 Overall.Qual
2 Gr.Liv.Area
3 Year.Built
4 Lot.Area

Now, we fit the model with selected top 4 covariates (use log(Year.Built) for Year.Built)

```
best_model = lm(log_SalePrice ~ Overall.Qual + Gr.Liv.Area + log_Year.Built +
                Lot.Area, data = house)

summary(best_model)
```

```
Call:
lm(formula = log_SalePrice ~ Overall.Qual + Gr.Liv.Area + log_Year.Built +
    Lot.Area, data = house)

Residuals:
     Min       1Q   Median       3Q      Max
-2.03932 -0.08321  0.00916  0.09804  0.60691

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -5.116e+01  2.021e+00 -25.311  < 2e-16 ***
Overall.Qual2   3.199e-01  9.968e-02   3.209  0.00134 **
Overall.Qual3   7.192e-01  9.144e-02   7.865 5.15e-15 ***
Overall.Qual4   8.957e-01  8.799e-02  10.180  < 2e-16 ***
Overall.Qual5   1.078e+00  8.746e-02  12.323  < 2e-16 ***
Overall.Qual6   1.161e+00  8.764e-02  13.242  < 2e-16 ***
Overall.Qual7   1.264e+00  8.800e-02  14.362  < 2e-16 ***
Overall.Qual8   1.448e+00  8.848e-02  16.365  < 2e-16 ***
Overall.Qual9   1.664e+00  8.985e-02  18.517  < 2e-16 ***
Overall.Qual10  1.617e+00  9.439e-02  17.128  < 2e-16 ***
Gr.Liv.Area     2.766e-04  8.263e-06  33.469  < 2e-16 ***
log_Year.Built  6.897e+00  2.666e-01  25.870  < 2e-16 ***
Lot.Area        5.496e-06  4.293e-07  12.802  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1742 on 2917 degrees of freedom
Multiple R-squared:  0.818, Adjusted R-squared:  0.8172
F-statistic:  1092 on 12 and 2917 DF,  p-value: < 2.2e-16
```

very high adjust R square

```
ols_mallows_cp(best_model, log_full_model)
```

```
[1] 311.5513
```

Mallow's CP score also show it is the lowest one compared to the above plot.