Multiple-Linear-regression-with-dummy-variables.R

Finley

2025-04-08

library(ggplot2)  
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

library(dplyr)

##   
## Attaching package: 'dplyr'

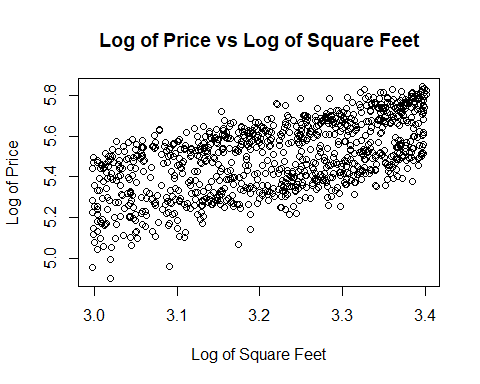
## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

house\_dta <- read.csv("C:\\Users\\finle\\OneDrive\\University\\Econometrics and Statistics\\Computer stuff\\house\_dta.csv", header= TRUE)  
  
summary(house\_dta)

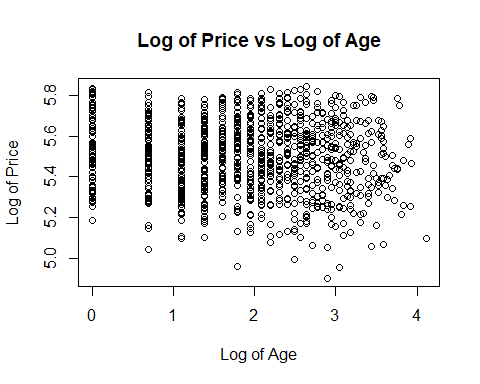
## id price sqft age   
## Min. : 1.0 Min. :134.3 Min. :20.03 Min. : 0.000   
## 1st Qu.: 250.8 1st Qu.:215.6 1st Qu.:22.83 1st Qu.: 3.000   
## Median : 500.5 Median :245.8 Median :25.36 Median : 6.000   
## Mean : 500.5 Mean :247.7 Mean :25.21 Mean : 9.392   
## 3rd Qu.: 750.2 3rd Qu.:278.3 3rd Qu.:27.75 3rd Qu.:13.000   
## Max. :1000.0 Max. :345.2 Max. :30.00 Max. :60.000   
## school fplace   
## Min. :0.000 Min. :0.000   
## 1st Qu.:0.000 1st Qu.:0.000   
## Median :1.000 Median :1.000   
## Mean :0.519 Mean :0.518   
## 3rd Qu.:1.000 3rd Qu.:1.000   
## Max. :1.000 Max. :1.000

house\_dta <- house\_dta %>%  
 mutate(  
 lnprice = log(price),  
 lnsqft = log(sqft),  
 lnage = log(age + 1)  
 )  
  
  
  
plot(lnprice ~ lnsqft, data = house\_dta,   
 main = "Log of Price vs Log of Square Feet",  
 xlab = "Log of Square Feet",  
 ylab = "Log of Price")



Check for Linear relationship

plot(lnprice ~ lnage, data = house\_dta,   
 main = "Log of Price vs Log of Age",  
 xlab = "Log of Age",  
 ylab = "Log of Price")



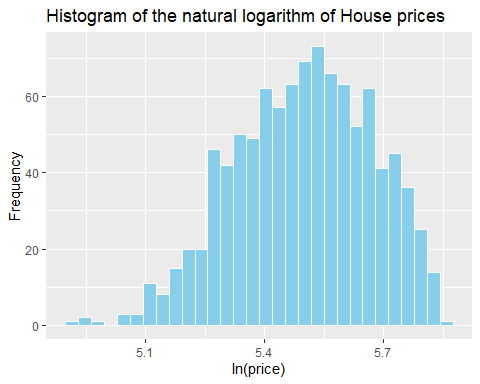
house\_dta %>%  
 select(lnsqft, lnage, school, fplace) %>%  
 cor()

## lnsqft lnage school fplace  
## lnsqft 1.00000000 -0.02025511 0.020936605 0.099248978  
## lnage -0.02025511 1.00000000 -0.033040947 0.021269163  
## school 0.02093661 -0.03304095 1.000000000 -0.007378108  
## fplace 0.09924898 0.02126916 -0.007378108 1.000000000

Check correlation of independent variables

ggplot(house\_dta, aes(x = lnprice)) +  
 geom\_histogram(fill = "skyblue", color = "white") +   
 labs(title = "Histogram of the natural logarithm of House prices",  
 x = "ln(price)",   
 y = "Frequency"  
 )

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



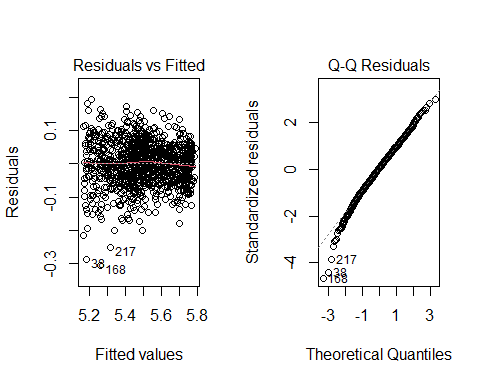
Check normality of dependent variable

house\_dta.lm <- lm(formula= lnprice ~ lnsqft + lnage + school + fplace, data = house\_dta)  
summary(house\_dta.lm)

##   
## Call:  
## lm(formula = lnprice ~ lnsqft + lnage + school + fplace, data = house\_dta)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.304377 -0.040248 0.002187 0.042951 0.193432   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.649309 0.057078 46.416 < 2e-16 \*\*\*  
## lnsqft 0.848146 0.017693 47.937 < 2e-16 \*\*\*  
## lnage -0.007949 0.002137 -3.720 0.00021 \*\*\*  
## school 0.247842 0.004141 59.857 < 2e-16 \*\*\*  
## fplace 0.006140 0.004159 1.476 0.14014   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.06537 on 995 degrees of freedom  
## Multiple R-squared: 0.8594, Adjusted R-squared: 0.8589   
## F-statistic: 1521 on 4 and 995 DF, p-value: < 2.2e-16

High coefficient of determination

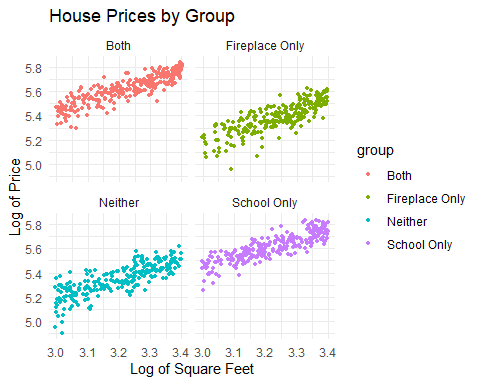
par(mfrow = c(1,2))  
plot(house\_dta.lm,1:2)



Check for homoskedasticity of residuals and normality of residuals.

house\_dta$group <- with(house\_dta,   
 ifelse(school == 1 & fplace == 1, "Both",  
 ifelse(school == 1, "School Only",  
 ifelse(fplace == 1, "Fireplace Only", "Neither"))))  
  
  
  
  
ggplot(house\_dta, aes(x = lnsqft, y = lnprice, color = group)) +  
 geom\_point(size = 1) +  
 labs(title = "House Prices by Group",  
 x = "Log of Square Feet",  
 y = "Log of Price") +  
 facet\_wrap(~ group) + # This will create separate plots for each group  
 theme\_minimal()

Create groups for indicator variables



Plot for different indicator variables.