

# PSTAT126 Project Step-2

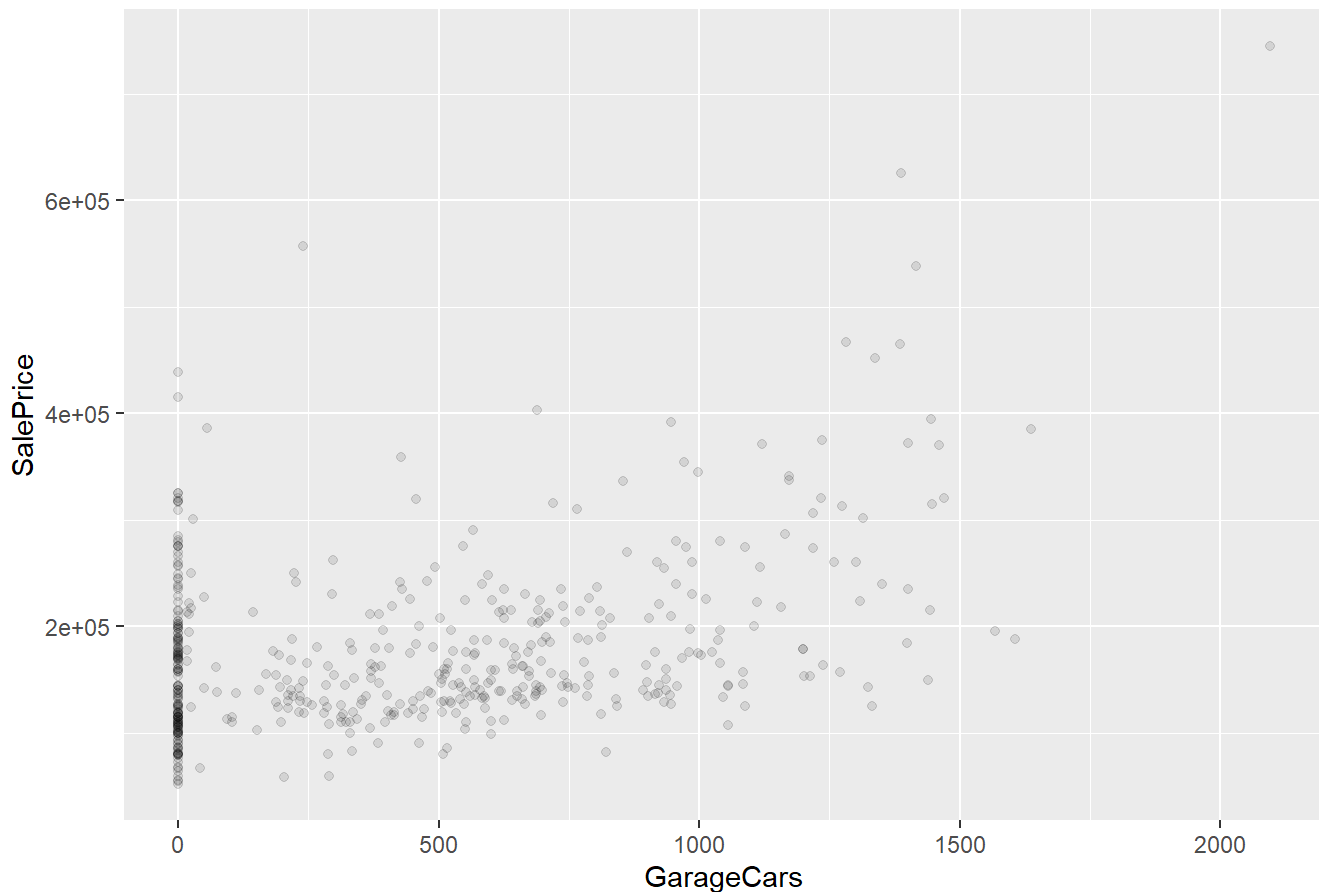
Liuqian Bao

2023-10-27

```
set.seed(12345)
train2_500 <- train2[sample(1:1460, 500, replace = FALSE),]
```

```
ggplot(data = train2_500, mapping = aes(x = BsmtFinSF1, y = SalePrice)) +
  geom_point(alpha = 0.1) +
  labs(title="SalePrice vs. GarageCars") +
  xlab("GarageCars") + ylab("SalePrice")
```

SalePrice vs. GarageCars



## Linear Model

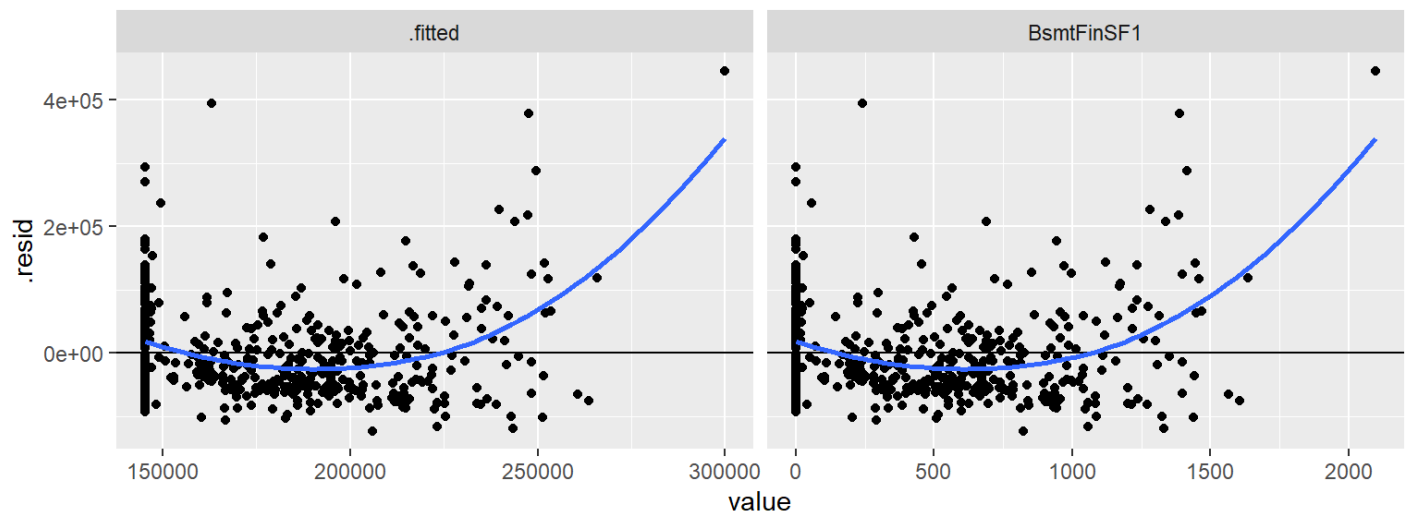
```
fit <- lm(SalePrice ~ ., train2_500)
fit_BsmtFinSF1 <- lm(SalePrice ~ BsmtFinSF1, train2_500)
summary(fit_BsmtFinSF1)
```

```
##
## Call:
## lm(formula = SalePrice ~ BsmtFinSF1, data = train2_500)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123399  -51622  -19782   32491  445210
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.453e+05  4.869e+03  29.846  <2e-16 ***
## BsmtFinSF1   7.370e+01  7.863e+00   9.372  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 75590 on 498 degrees of freedom
## Multiple R-squared:  0.1499, Adjusted R-squared:  0.1482
## F-statistic: 87.84 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
augment(fit_BsmtFinSF1, train2_500) %>% head(4)
```

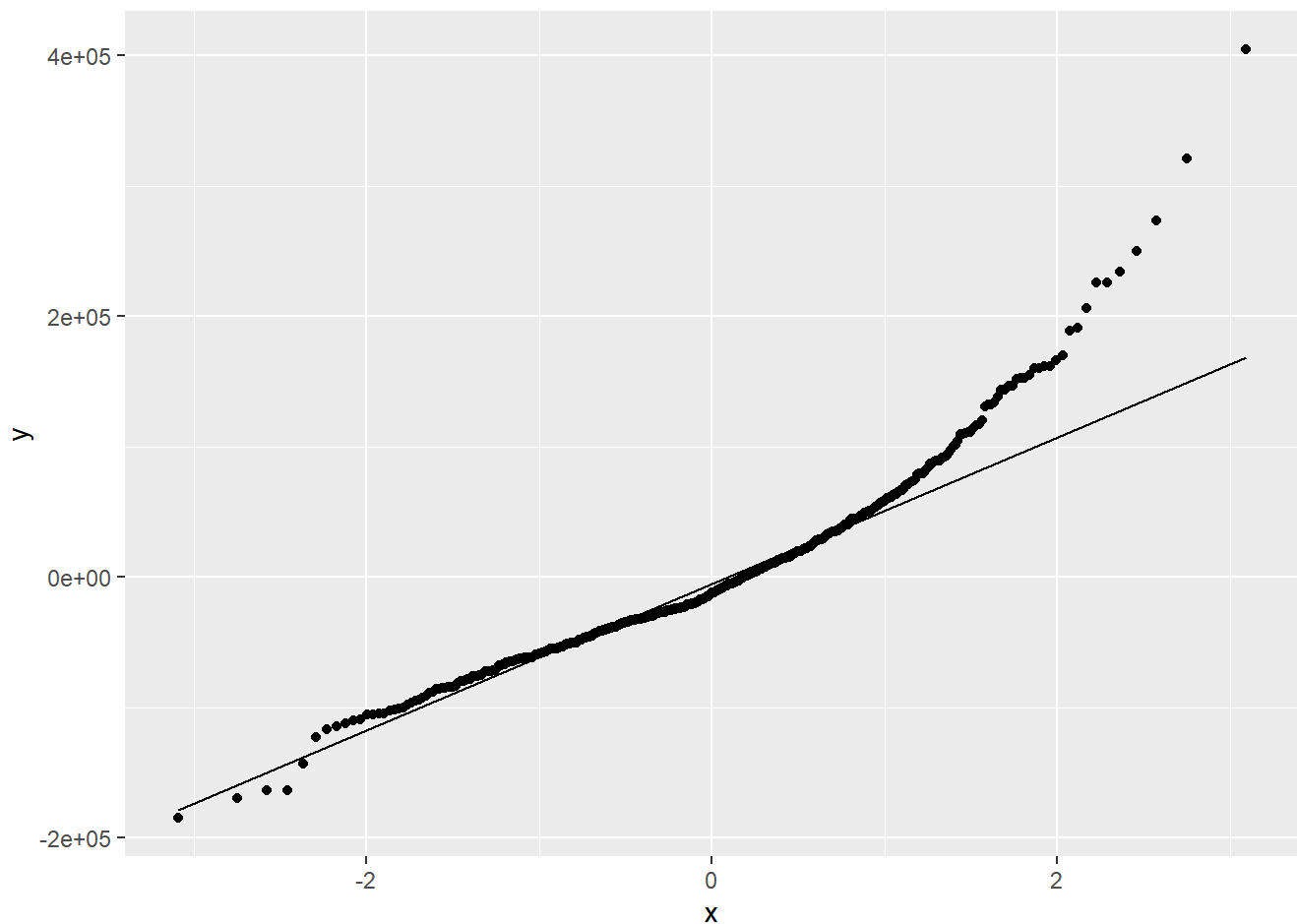
```
## # A tibble: 4 × 36
##       Id LotArea LotShape LotConfig Neighborhood HouseStyle RoofStyle Exterior1st
##   <dbl>   <dbl> <chr>      <chr>      <chr>          <chr>      <chr>      <chr>
## 1   142   11645 Reg       Inside    CollgCr       1Story     Gable     VinylSd
## 2    51   13869 IR2       Corner    Gilbert       2Story     Gable     VinylSd
## 3   720   9920  IR1       Inside    CollgCr       1Story     Gable     HdBoard
## 4   730   6240  Reg       Inside    IDOTRR       1.5Fin     Gable     MetalSd
## # i 28 more variables: MasVnrType <chr>, ExterQual <chr>, ExterCond <chr>,
## #   Foundation <chr>, BsmtQual <chr>, BsmtCond <chr>, BsmtFinType1 <chr>,
## #   BsmtFinSF1 <dbl>, BsmtFinType2 <chr>, BsmtUnfSF <dbl>, HeatingQC <chr>,
## #   Electrical <chr>, BsmtFullBath <dbl>, FullBath <dbl>, HalfBath <dbl>,
## #   KitchenQual <chr>, FireplaceQu <chr>, GarageType <chr>, GarageFinish <chr>,
## #   GarageCars <dbl>, OpenPorchSF <dbl>, SalePrice <dbl>, .fitted <dbl>,
## #   .resid <dbl>, .hat <dbl>, .sigma <dbl>, .cooks <dbl>, .std.resid <dbl>
```

```
# panel of residual plots
augment(fit_BsmtFinSF1, train2_500) %>%
  pivot_longer(cols = c(.fitted, BsmtFinSF1)) %>%
  ggplot(aes(y = .resid, x = value)) +
  facet_wrap(~ name, scales = 'free_x') +
  geom_point() +
  geom_hline(aes(yintercept = 0)) +
  geom_smooth(method = 'loess', formula = 'y ~ x', se = F, span = 1)
```



```
# add quadratic term in expenditure
fit_BsmtFinSF1_q <- lm(SalePrice ~ poly(BsmtFinSF1, 2, raw = T), data = train2_500)
```

```
# normality check
augment(fit_BsmtFinSF1_q, train2_500) %>%
  ggplot(aes(sample = .resid)) +
  geom_qq() +
  geom_qq_line()
```



t-test

```
n <- dim(train2_500)[1] # number of observations, or equivalently use nrow(statedata)
p <- 1 # number of predictors
round(coefficients(summary(fit_BsmtFinSF1_q)), 5)
```

```
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)    165198.49239  5071.61812  32.57313    0e+00
## poly(BsmtFinSF1, 2, raw = T)1    -86.72346    19.71717  -4.39837    1e-05
## poly(BsmtFinSF1, 2, raw = T)2     0.13461     0.01536   8.76337    0e+00
```

## Plot transformed

```
# ggplot(data = train2_500, mapping = aes(x = poly(BsmtFinSF1, 2, raw = T), y = SalePrice))+
#   geom_point(alpha = 0.1) +
#   labs(title="SalePrice vs. GarageCars") +
#   xlab("GarageCars") + ylab("SalePrice")
```

## TRansformed residual graph

```
# panel of residual plots
augment(fit_BsmtFinSF1_q, train2_500) %>%
  pivot_longer(cols = c(.fitted, BsmtFinSF1)) %>%
  ggplot(aes(y = .resid, x = value)) +
  facet_wrap(~ name, scales = 'free_x') +
  geom_point() +
  geom_hline(aes(yintercept = 0)) +
  geom_smooth(method = 'loess', formula = 'y ~ x', se = F, span = 1)
```

