simple case 3.31

```
setwd('/Users/quebec/Playground/ALSM/case')
# 需要的包
pacman::p_boot()
pacman::p_load(MASS,car,lmtest,alr3,data.table,ramify)
```

```
# 载入数据
.data<-fread('./data/APPENC07.txt')
names(.data)<-c('id','price','feet','bedrooms','bathrooms','air','garage','pool','year','quality',
set.seed(43)
rownames(.data)<-.data$id
.data$id<-NULL
.d<-.data[sample(1:nrow(.data),200)] # 取 price 和 feet
names(.d)[c(1,2)]<-c('y','x')
```

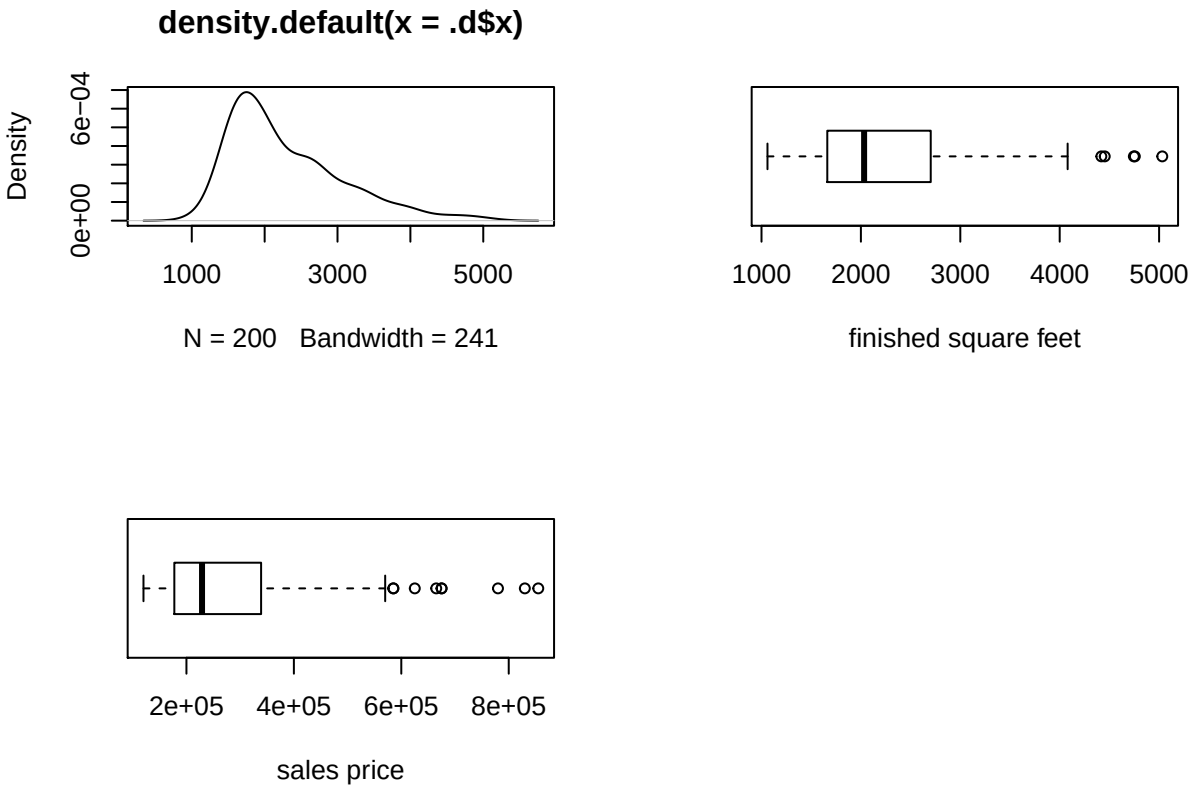
EDA

```
summary(.d[,1:2])
```

```
##        y                x
##  Min.   :120000   Min.   :1060
##  1st Qu.:177675   1st Qu.:1667
##  Median :229050   Median :2032
##  Mean   :280054   Mean   :2289
##  3rd Qu.:338500   3rd Qu.:2702
##  Max.   :855000   Max.   :5032
```

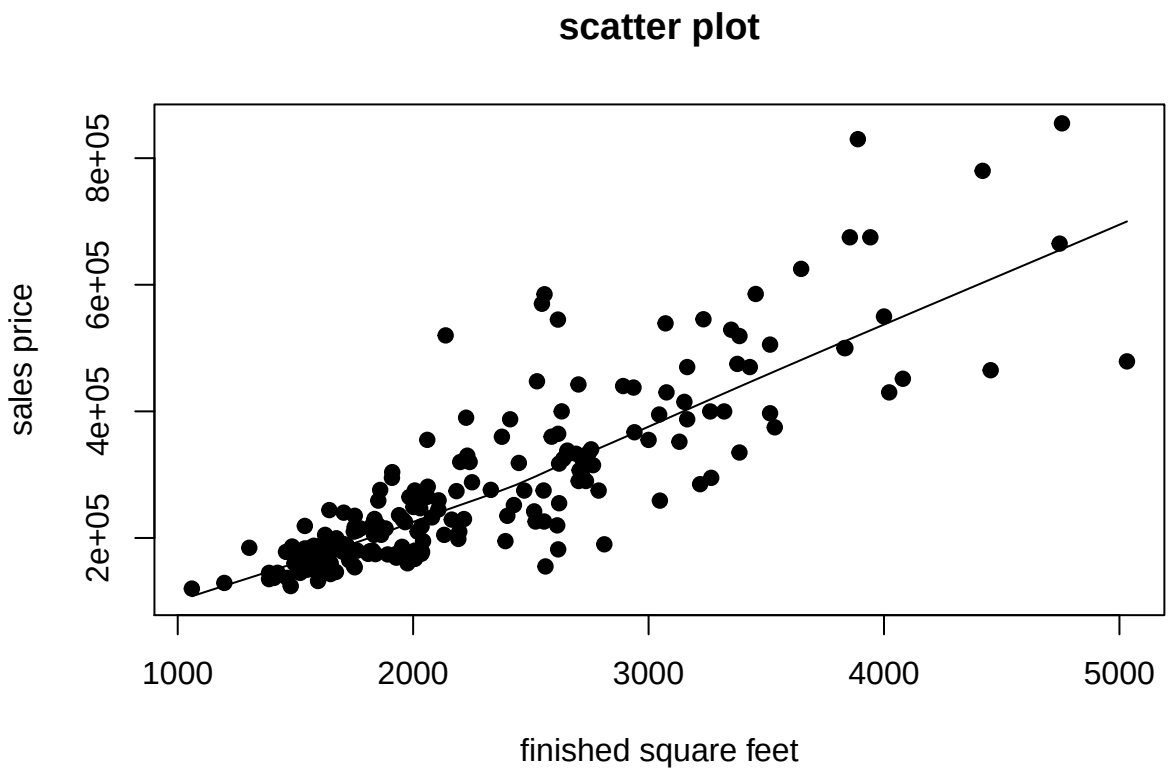
观察一下 x 的情况

```
par(mfrow=c(2,2))
plot(density(.d$x))
boxplot(.d$x, horizontal = TRUE, xlab='finished square feet')
boxplot(.d$y, horizontal = TRUE, xlab='sales price')
```



x 明显右偏，此外 x 与 y 都有离群值，但是 y 的离群值比 x 要多.

```
with(.d,scatter.smooth(x,y,pch=19,ann=F))
title(main='scatter plot',xlab='finished square feet',ylab='sales price')
```



右下角的 loess 图，发现有线性关系，但略微有点曲线. 此外看到 megaphone shape，提示我们之后可能需要对 Y 做幂变换.

线性模型

变换前

```
fit0<-lm(y~x,.d)
summary(fit0)
```

```
##
## Call:
## lm(formula = y ~ x, data = .d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -232554  -37465   -3607   22329  298227
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -79921.599  16413.096  -4.869 2.28e-06 ***
```

```
## x              157.288      6.794  23.152  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74340 on 198 degrees of freedom
## Multiple R-squared:  0.7302, Adjusted R-squared:  0.7289
## F-statistic:   536 on 1 and 198 DF,  p-value: < 2.2e-16
```

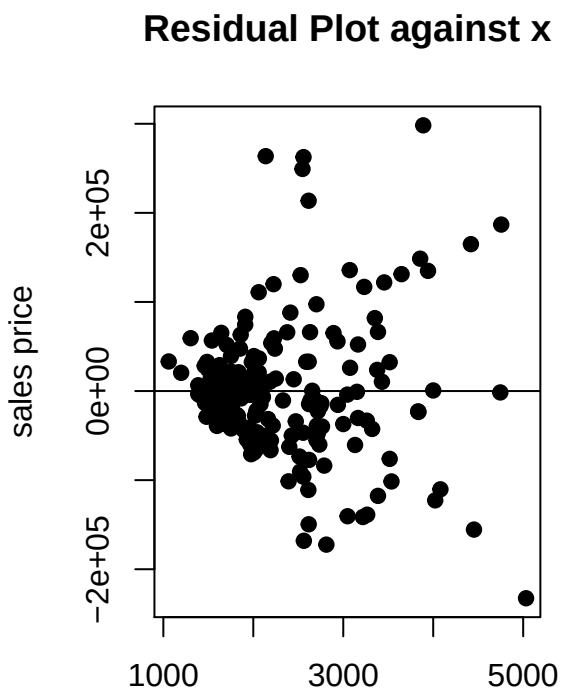
从 p 值来看，有明显的线性关系，但是 R^2 不算很大.

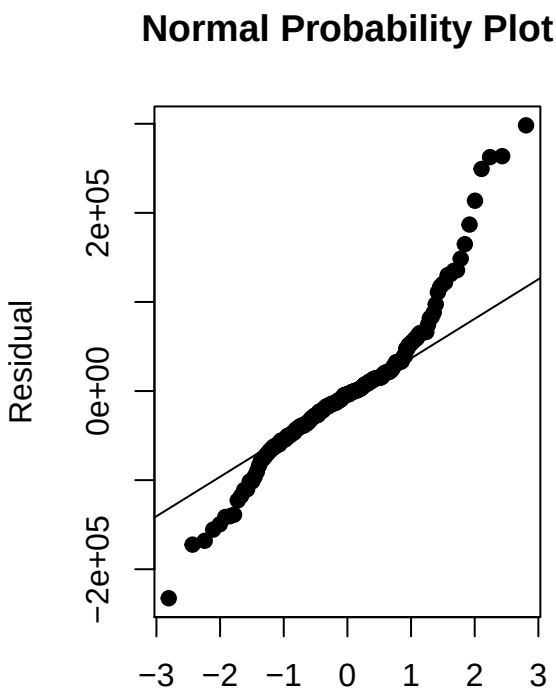
残差图:

```
plot_resid<-function(fit) {
  par(mfrow = c(1, 2), pch = 19)

  plot(.d$x, resid(fit), xlab='finished square feet',ylab='sales price')
  title("Residual Plot against x")
  abline(0,0)

  #boxplot(resid(fit), horizontal = TRUE, xlab = "Residual")
  #title("(c) Box Plot")
  qqnorm(resid(fit), xlab = "Expected", ylab = "Residual", main = "")
  title("Normal Probability Plot")
  qqline(resid(fit))
}
plot_resid(fit0)
```





如我们在 loess 图中发现的，等方差假设并不成立，误差项正态性假设也不成立.

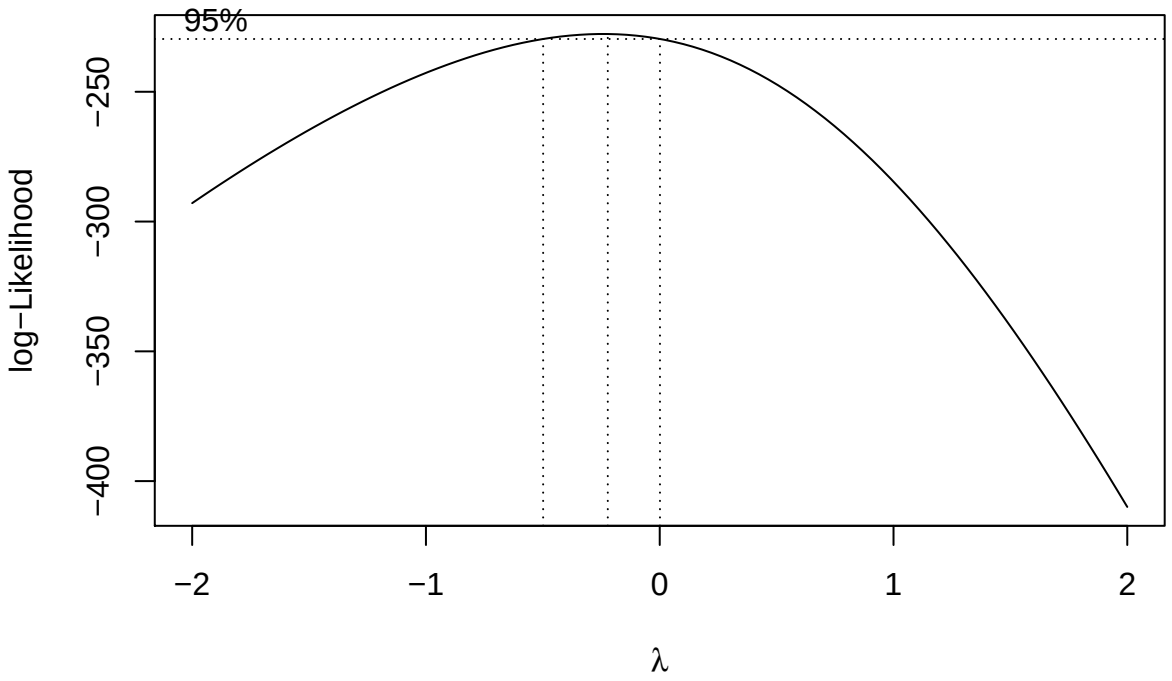
但是这些可能是受了模型不正确的影响，所以做一下差拟检验.

```
ano0<-pureErrorAnova(fit0)
1-pf(ano0[3,3]/ano0[4,3],20,178)
```

```
## [1] 0.01757458
```

模型表现不佳，考虑对 y 做幂变换.

```
boxcox(fit0)
```



出于可解释性的考虑，选择对数变换.

对数变换

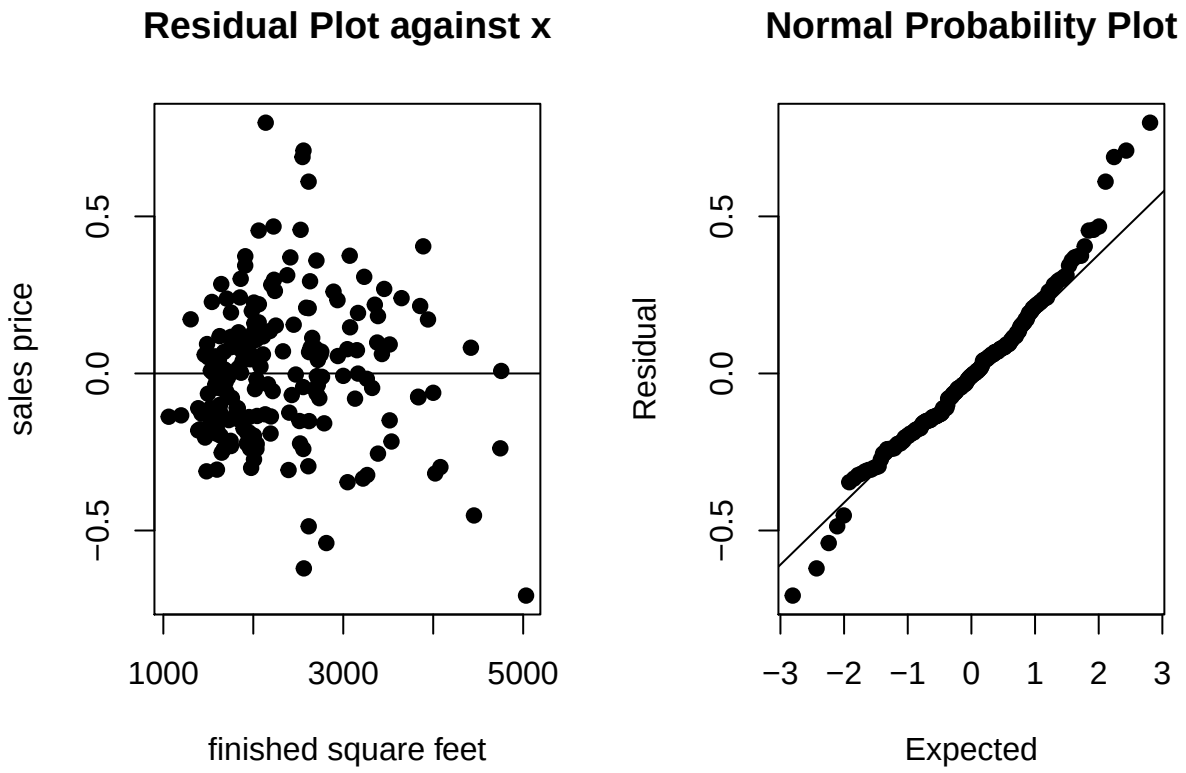
```
# tmp <- boxcox(fit0,plotit = FALSE)
# lambda<-tmp$x[tmp$y==max(tmp$y)]
# .d$y.tran<-ifelse(lambda==0,ln(.d$y),.d$y^lambda)
.d$y.tran<-log(.d$y)
fit1<-lm(y.tran~x,.d)
summary(fit1)
```

```
##
## Call:
## lm(formula = y.tran ~ x, data = .d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70716 -0.14949 -0.00807  0.11691  0.79832
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.131e+01  4.947e-02  228.68   <2e-16 ***
## x           4.918e-04  2.048e-05   24.02   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.224 on 198 degrees of freedom
## Multiple R-squared:  0.7445, Adjusted R-squared:  0.7432
## F-statistic:   577 on 1 and 198 DF,  p-value: < 2.2e-16
```

R^2 变化不大，略有提高.

```
plot_resid(fit1)
```



可以看到不等方差的情况和非正态性的情况有明显改善，做一下检验.

```
bptest(fit1,studentize = FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  fit1
## BP = 17.54, df = 1, p-value = 2.814e-05
```

```
shapiro.test(resid(fit1))
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  resid(fit1)
## W = 0.97804, p-value = 0.003166
```

```
durbinWatsonTest(fit1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.01621993      1.957986   0.726
##  Alternative hypothesis: rho != 0
```

正态性和等方差一如既往地不成立. 差拟检验:

```
ano1<-pureErrorAnova(fit1)
ano1[3,3]/ano1[4,3]
```

```
## [1] 1.232726
```

```
1-pf(ano1[3,3]/ano1[4,3],20,178)
```

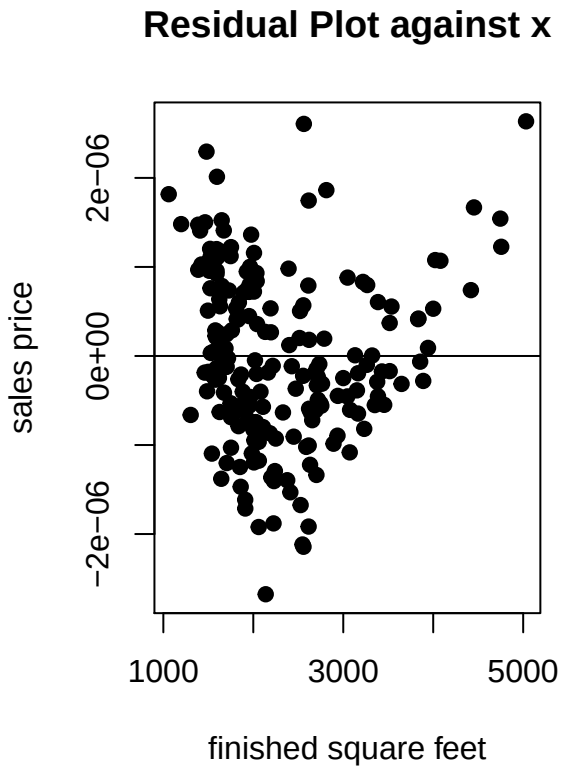
```
## [1] 0.2323454
```

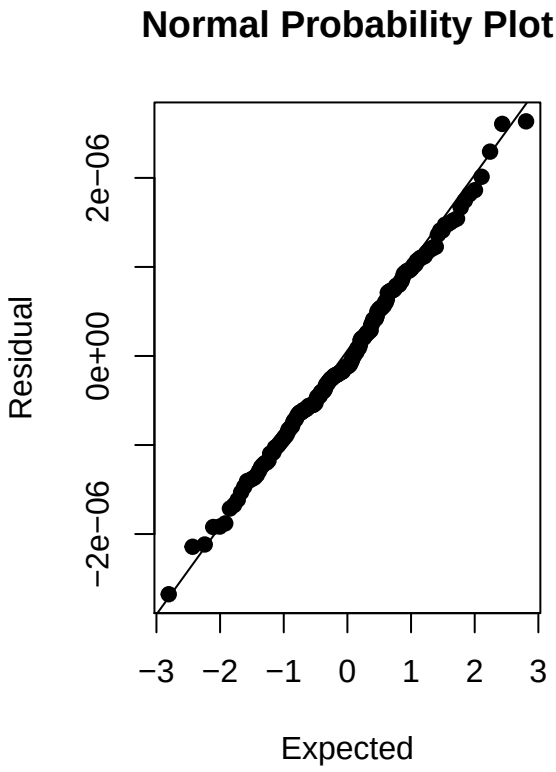
p 值显著提高，说明拟合得更好了. 这符合预期，因为 Box-Cox 变换的让 SSE 最小

倒数变换

再换一种变换，对于 megaphone pattern，一个常用的变换是 1/Y

```
.d$y.tran.1<-1/.d$y
fit2<-lm(y.tran.1~x,.d)
plot_resid(fit2)
```





看起来也不错.

```
shapiro.test(resid(fit2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit2)
## W = 0.99582, p-value = 0.861
```

```
bptest(fit2,studentize = FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  fit2
## BP = 0.29408, df = 1, p-value = 0.5876
```

正态性和等方差都满足，而且 p 值很大.

```
ano2<-pureErrorAnova(fit2)
1-pf(ano2[3,3]/ano2[4,3],20,178)
```

```
## [1] 0.3052491
```

差拟检验也没有问题.

对 y 做变换真的好？

但真正的残差结果却令人吃惊：

```
sum( (1/predict(fit2)-.d$y)^2)
```

```
## [1] 1.044073e+15
```

残差是非常大的，倒数变换之所以效果显得不错 (差拟检验)，是因为 y 太大，1/y 太小而造成的假象，这其中机器精度有很大影响，提示我们如果 y 很大，那么不要用倒数变换. 对数变换也有类似地问题对数变换：

```
sum((exp(predict(fit1))-.d$y)^2)-sum((predict(fit0)-.d$y)^2)
```

```
## [1] 265812157656
```

```
sum((predict(fit0)-.d$y)^2)# 与 fit1 作对比
```

```
## [1] 1.094149e+12
```

原因就是 ln 和 exp 的放缩效果太明显了，模型掩盖了这些，但是真实的数据表现就很差了. 因此还是用不变形的模型. 尽管幂变换可能使模型更接近假设，但是效果却不一定好. 当 y 小的时候可能还不错，但 y 很大的时候，预测效果就很不理想了.

模型评价

预测能力

由于数据中没有对应的，所以找出最接近的作为参考

```
query<-c(1100,4900)
predict(fit0,data.frame(x=query))
```

```
##         1         2
##  93095.62 690791.49
```

```
nearest<-function(tar) argmin(matrix(abs(.data$feet-tar)),F)
.data[sapply(query,nearest),]
```

```
##     price feet bedrooms bathrooms air garage pool year quality style   lot
## 1: 120000 1060        2         1   0      2    0 1947       3     1 15001
## 2: 545000 4973        6         6   1      3    1 1987       1     7 56139
##    hignway
## 1:       0
## 2:       0
```

我们发现，误差还是不小的. 相对误差在 x=1100 的时候越为 0.0225，在 4900 时为 0.2214286

优缺点

优点：SSE 是比较小的，与对 Y 做变换比起来，OLS 给出了不错的 SSE. 缺点：不满足正态假设，所以做区间估计效力不大，特别对于预测更是如此. 预测准确度也并不很高，受离群值影响明显.