piecewise reg

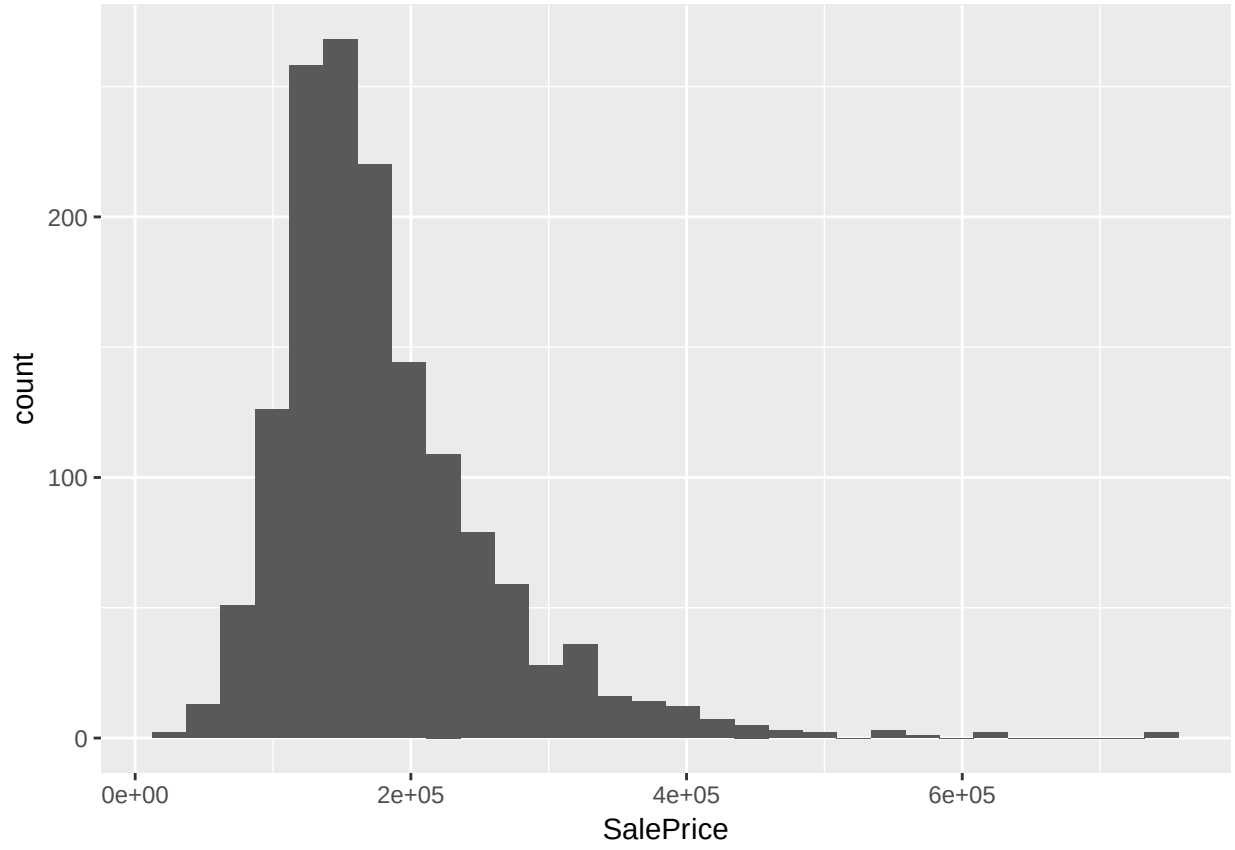
wonilChoi

2018년 1월 30일

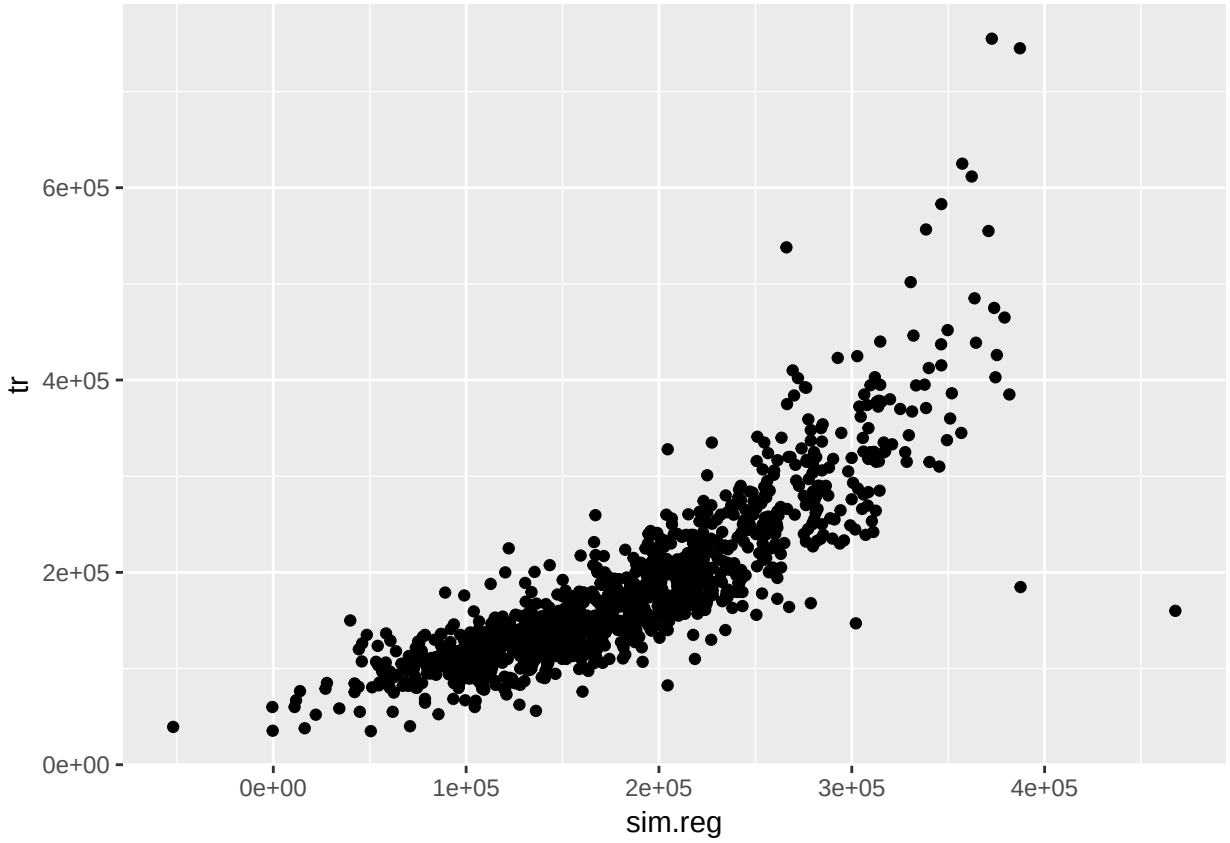
Season 0 - Preseason - Regression

```
library(data.table)

tr <- fread("train.csv", select=c("SalePrice","SaleCondition","MiscVal","GarageCars","Fireplaces","Overa
ts <- fread("test.csv")

library(ggplot2)
base <- ggplot(data=tr)
base+geom_histogram(aes(x=SalePrice))
```

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




```
tr <- na.omit(tr)

sim.reg <- lm(SalePrice~.,data=tr)
tr.err <- data.table(tr=tr$SalePrice)
tr.err$sim.reg <- sim.reg$fitted.values
qplot(data=tr.err,x=sim.reg,y=tr)
```

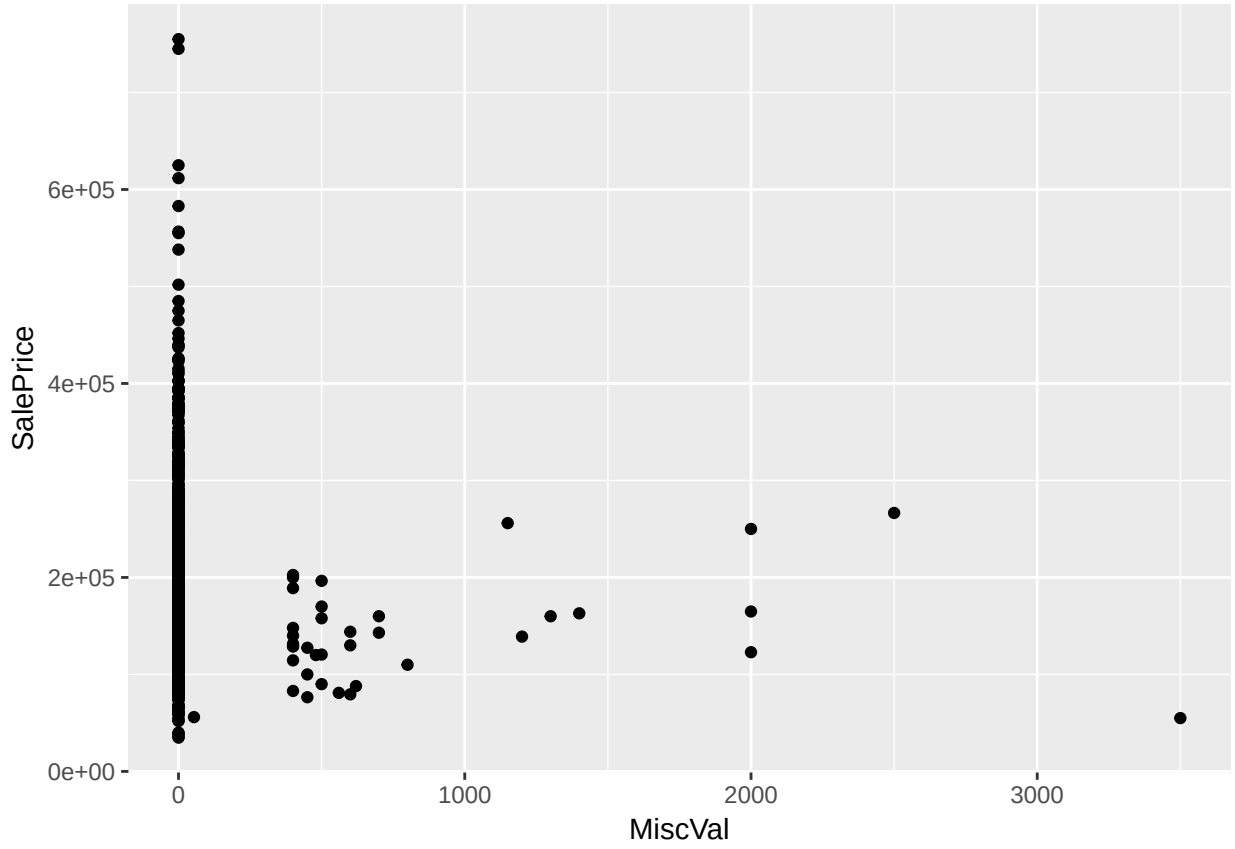



```
summary(sim.reg)
```

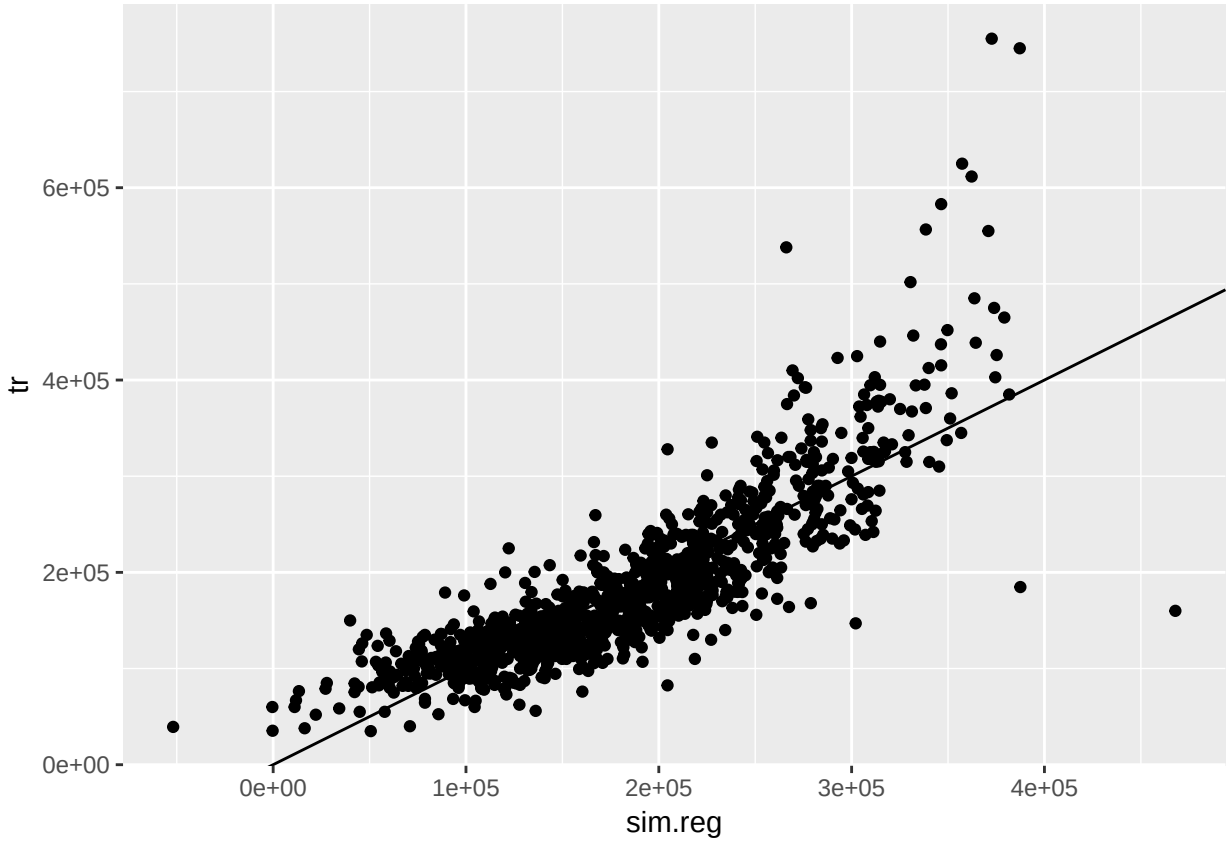
```
##
## Call:
## lm(formula = SalePrice ~ ., data = tr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -307810  -25076   -3227   19524  382357
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.126e+05  1.058e+04 -10.637  < 2e-16 ***
## SaleConditionAdjLand  1.767e+04  2.281e+04   0.775   0.4388
## SaleConditionAlloca   5.774e+03  1.489e+04   0.388   0.6983
## SaleConditionFamily  -1.656e+04  1.153e+04  -1.436   0.1512
## SaleConditionNormal   3.576e+03  5.102e+03   0.701   0.4835
## SaleConditionPartial  2.946e+04  6.626e+03   4.446 9.55e-06 ***
## MiscVal               1.188e+00  6.797e+00   0.175   0.8613
## GarageCars            2.242e+04  2.184e+03  10.269  < 2e-16 ***
## Fireplaces            1.830e+04  2.311e+03   7.918 5.49e-15 ***
## OverallCond           2.085e+03  1.214e+03   1.718   0.0861 .
## OverallQual           3.277e+04  1.242e+03  26.384  < 2e-16 ***
## LotFrontage           3.612e+02  5.617e+01   6.431 1.84e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 44300 on 1189 degrees of freedom
## Multiple R-squared:  0.7204, Adjusted R-squared:  0.7178
## F-statistic: 278.5 on 11 and 1189 DF,  p-value: < 2.2e-16
```

```
qplot(data=tr,x=MiscVal,y=SalePrice)
```




```
tr <- tr[,!"MiscVal"]
sim.reg <- lm(SalePrice~.,data=tr)
tr.err$sim.reg <- sim.reg$fitted.values
qplot(data=tr.err,x=sim.reg,y=tr)+geom_abline(slope=1,intercept=0)
```




```
summary(sim.reg)
```

```
##
## Call:
## lm(formula = SalePrice ~ ., data = tr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -307889  -25078   -3155   19474  382336
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -112577.66   10580.49 -10.640  < 2e-16 ***
## SaleConditionAdjLand   17648.85   22797.48   0.774   0.4390
## SaleConditionAlloca     5772.71   14885.92   0.388   0.6982
## SaleConditionFamily   -16573.52   11524.61  -1.438   0.1507
## SaleConditionNormal     3597.91    5098.21   0.706   0.4805
## SaleConditionPartial   29476.45    6622.61   4.451 9.35e-06 ***
## GarageCars             22411.54    2181.64  10.273  < 2e-16 ***
## Fireplaces             18327.14    2305.35   7.950 4.31e-15 ***
## OverallCond             2097.25    1211.21   1.732   0.0836 .
## OverallQual            32758.78    1239.73  26.424  < 2e-16 ***
## LotFrontage              361.38      56.14   6.438 1.76e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 44280 on 1190 degrees of freedom
## Multiple R-squared:  0.7204, Adjusted R-squared:  0.7181
## F-statistic: 306.6 on 10 and 1190 DF,  p-value: < 2.2e-16
```

```
library(splines)
sp.reg <- lm(data=tr.err, formula=tr~ns(tr.err$sim.reg, df=2))
bs.reg <- lm(data=tr.err, formula=tr~bs(tr.err$sim.reg, df=2))
```

```
## Warning in bs(tr.err$sim.reg, df = 2): 'df' was too small; have used 3
```

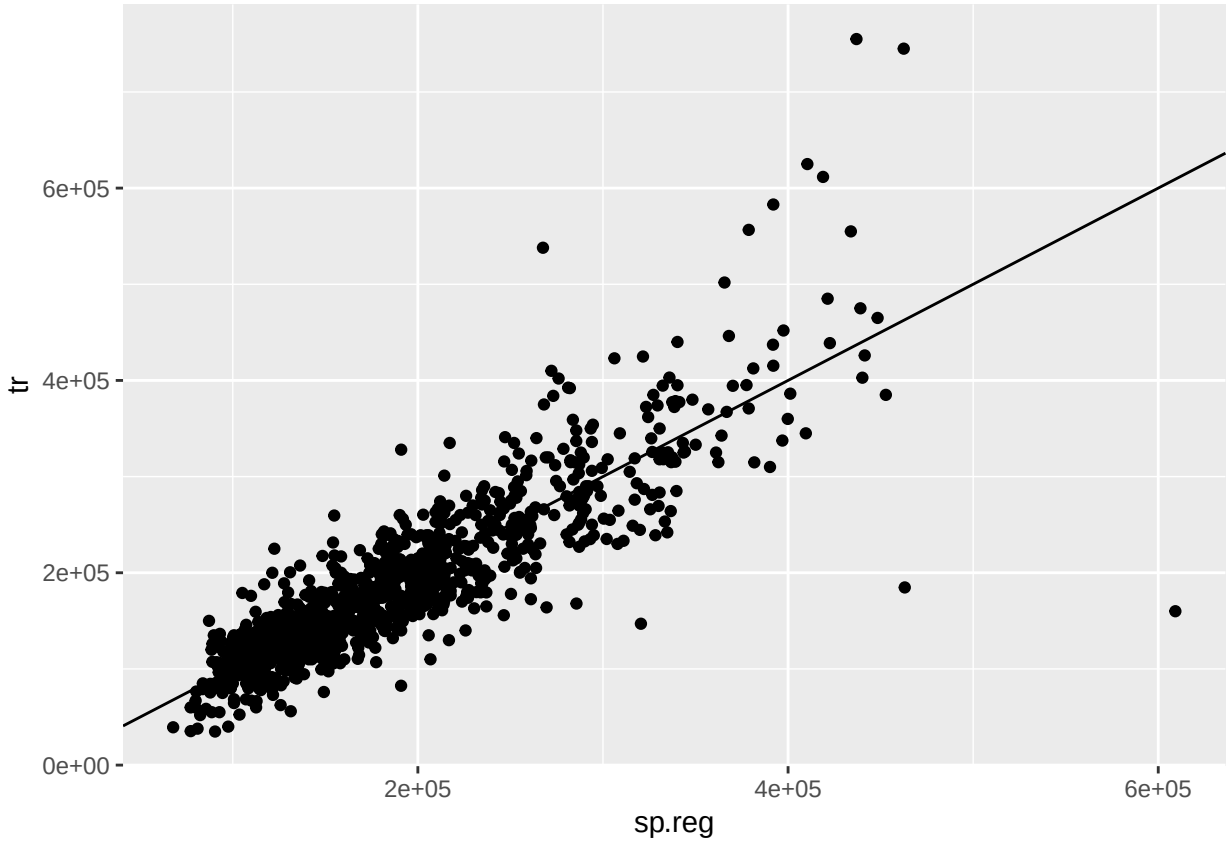
```
summary(sp.reg)
```

```
##
## Call:
## lm(formula = tr ~ ns(tr.err$sim.reg, df = 2), data = tr.err)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -449231  -20131    -688   17235  318056
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     67797       9008   7.526 1.02e-13 ***
## ns(tr.err$sim.reg, df = 2)1   390100      16438  23.731  < 2e-16 ***
## ns(tr.err$sim.reg, df = 2)2   523375      10046  52.097  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40560 on 1198 degrees of freedom
## Multiple R-squared:  0.7638, Adjusted R-squared:  0.7635
## F-statistic:  1937 on 2 and 1198 DF,  p-value: < 2.2e-16
```

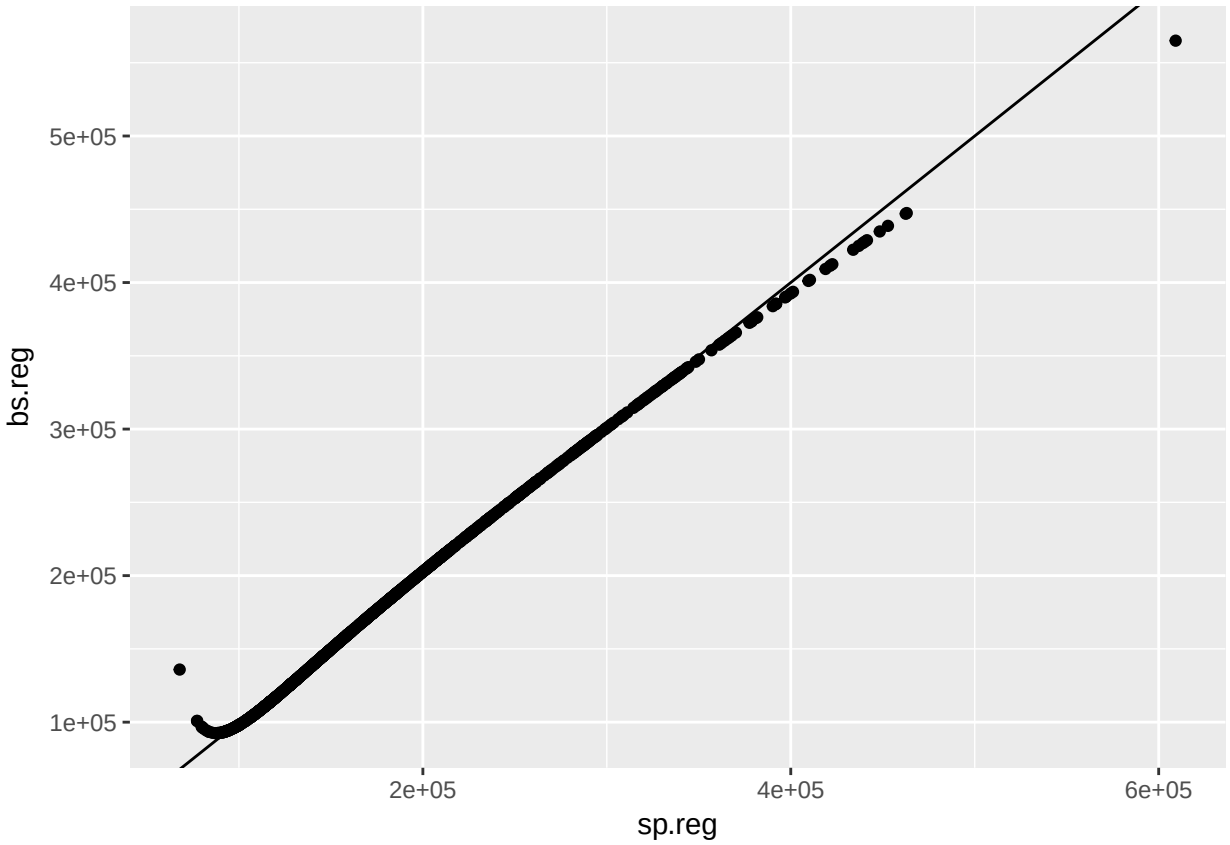
```
summary(bs.reg)
```

```
##
## Call:
## lm(formula = tr ~ bs(tr.err$sim.reg, df = 2), data = tr.err)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -405063  -20631    -539   18128  330001
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    135892      19312   7.037 3.31e-12 ***
## bs(tr.err$sim.reg, df = 2)1  -167725      43097  -3.892 0.000105 ***
## bs(tr.err$sim.reg, df = 2)2   188285      17933  10.500  < 2e-16 ***
## bs(tr.err$sim.reg, df = 2)3   429172      38092  11.267  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40730 on 1197 degrees of freedom
## Multiple R-squared:  0.762,  Adjusted R-squared:  0.7614
## F-statistic:  1278 on 3 and 1197 DF,  p-value: < 2.2e-16
```

```
tr.err$sp.reg <- sp.reg$fitted.values
tr.err$bs.reg <- bs.reg$fitted.values
base+geom_point(data=tr.err,aes(x=sp.reg,y=tr))+geom_abline(slope=1,intercept=0)
```




```
base+geom_point(data=tr.err,aes(x=sp.reg,y=bs.reg))+geom_abline(slope=1,intercept=0)
```




```
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
plot_ly(data=tr.err,y=~tr,x=~sp.reg)%>% add_markers()%>%add_lines(x = ~tr, y = ~tr)
```

```
breaks <- with(tr.err, sim.reg[which(sim.reg >= 310*10^3 & sim.reg <= 370*10^3)])
mse <- numeric(length(breaks))
for(i in 1:length(breaks)){
    piecewise1 <- lm(tr ~ sim.reg*(sim.reg < breaks[i]) + sim.reg*(sim.reg>=breaks[i]),data=tr.err)
    mse[i] <- summary(piecewise1)[6]
}
mse <- as.numeric(mse)
breaks[which(mse==min(mse))]
```

```
## [1] 357298.6
```

```r
# knot.reg <- lm(tr~sim.reg:(sim.reg>357298.6)+sim.reg:(sim.reg<=357298.6),data=tr.err)
library(SiZer)
```

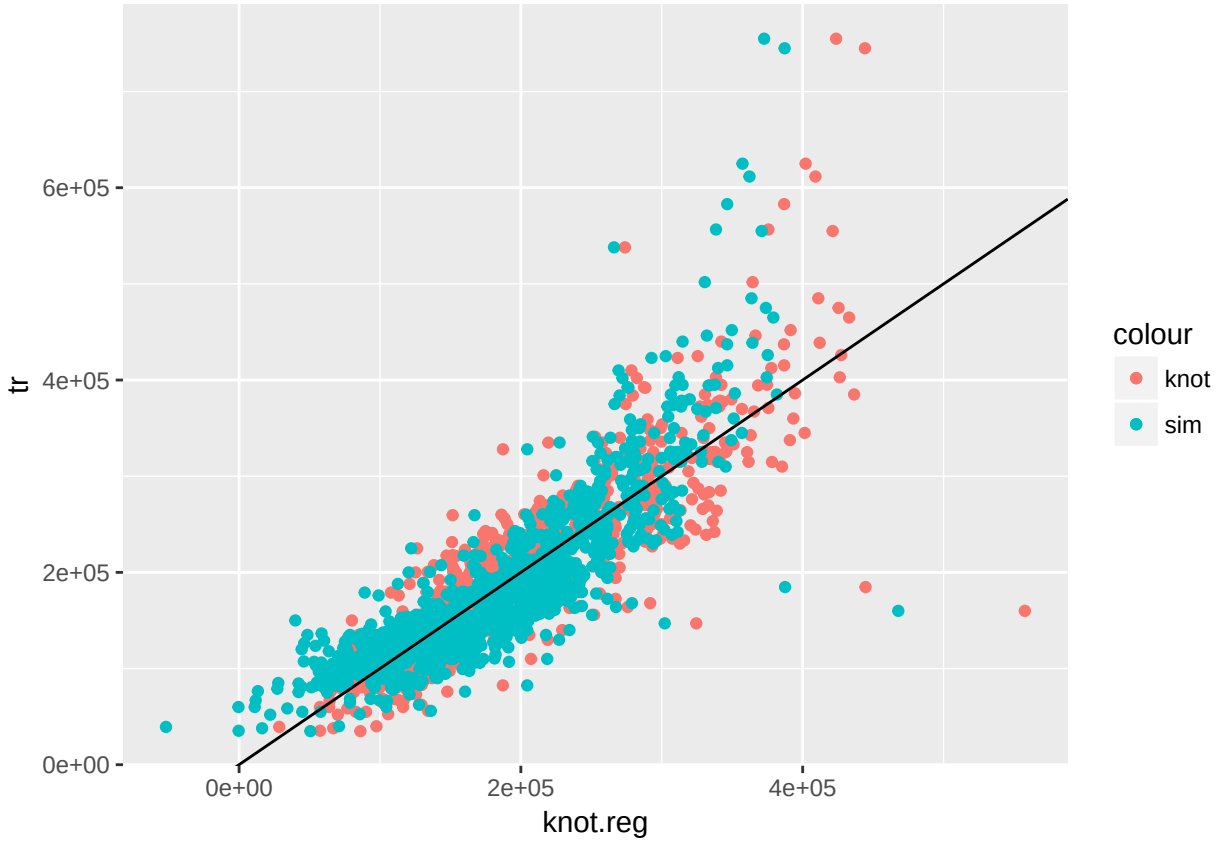
```
## Warning: package 'SiZer' was built under R version 3.4.3
```

```
## Loading required package: boot
```

```r
knot.reg <- with(tr.err,piecewise.linear(sim.reg, tr, middle = 1, CI = FALSE,
                 bootstrap.samples = 1000, sig.level = 0.05))
summary(knot.reg$model)
```

```
##
## Call:
## lm(formula = y ~ x + w)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -397683  -19911    -987   16759  331259
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.767e+04  4.977e+03   11.59   <2e-16 ***
## x           5.620e-01  3.331e-02   16.87   <2e-16 ***
## w           8.446e-01  5.581e-02   15.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40430 on 1198 degrees of freedom
## Multiple R-squared:  0.7653, Adjusted R-squared:  0.7649
## F-statistic:  1953 on 2 and 1198 DF,  p-value: < 2.2e-16
```

```r
tr.err <- cbind(tr.err, knot.reg=knot.reg$model$fitted.values)
base+geom_point(data=tr.err,aes(x=knot.reg,y=tr, col='knot'))+
    geom_point(data=tr.err,aes(x=sim.reg,y=tr, col='sim'))+
    geom_abline(slope=1,intercept=0)
```




```
tr.err[,sim.reg.err2:=(tr-sim.reg)^2]
tr.err[,sp.reg.err2:=(tr-sp.reg)^2]
tr.err[,knot.reg.err2:=(tr-knot.reg)^2]
sort(apply(tr.err,2,mean))
```

```
##            tr       sim.reg         sp.reg         bs.reg       knot.reg
##      180770.5      180770.5      180770.5      180770.5      180770.5
## knot.reg.err2   sp.reg.err2  sim.reg.err2
##  1630862017.5  1640811447.1  1942642595.3
```

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:boot':
##
##     logit
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
describe(tr.err)
```

```
##               vars    n         mean           sd      median     trimmed
## tr               1 1201     180770.5 8.338952e+04    159500.0    169540.0
## sim.reg          2 1201     180770.5 7.077818e+04    175539.8    177777.6
## sp.reg           3 1201     180770.5 7.288095e+04    162179.4    170842.5
```

```
## bs.reg           4 1201     180770.5 7.279355e+04    163003.0    171131.2
## knot.reg         5 1201     180770.5 7.294924e+04    156320.5    170941.8
## sim.reg.err2     6 1201 1942642595.3 7.864007e+09 500363323.4 818547676.0
## sp.reg.err2      7 1201 1640811447.1 8.087383e+09 349339962.1 620590980.5
## knot.reg.err2    8 1201 1630862017.5 7.392256e+09 339341870.2 623873313.4
##                       mad       min          max        range  skew
## tr                 58414.44  34900.00 7.550000e+05 7.201000e+05  1.89
## sim.reg            69558.14 -51860.33 4.678885e+05 5.197488e+05  0.37
## sp.reg             61852.80  67796.80 6.092315e+05 5.414347e+05  1.35
## bs.reg             65956.35  92503.29 5.650635e+05 4.725602e+05  1.21
## knot.reg           51920.45  28519.90 5.576828e+05 5.291629e+05  1.27
## sim.reg.err2  686715371.89    608.17 1.461810e+11 1.461810e+11 12.24
## sp.reg.err2   493052496.09    137.03 2.018089e+11 2.018089e+11 16.60
## knot.reg.err2 476085125.91     82.02 1.581516e+11 1.581516e+11 13.85
##                kurtosis           se
## tr                 6.22      2406.25
## sim.reg            0.04      2042.34
## sp.reg             2.14      2103.02
## bs.reg             1.40      2100.49
## knot.reg           1.46      2104.99
## sim.reg.err2     180.10 226919784.95
## sp.reg.err2      349.07 233365418.54
## knot.reg.err2    234.38 213307194.71
```