Housing price in king county, Washington in 2014,2015

Carl Zou, Chris Liu, Jun Zhang

DATASET

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	condition	lat	long	age
1	221900	3	3.00	1180	5650	1	0	3	47.5112	-122.257	63
2	538000	3	6.75	2570	7242	2	0	3	47.7210	-122.319	27
3	180000	2	2.00	770	10000	1	0	3	47.7379	-122.233	85
4	604000	4	12.00	1960	5000	1	0	5	47.5208	-122.393	53
5	510000	3	6.00	1680	8080	1	0	3	47.6168	-122.045	31
6	1225000	4	18.00	5420	101930	1	0	3	47.6561	-122.005	17
7	257500	3	6.75	1715	6819	2	0	3	47.3097	-122.327	23

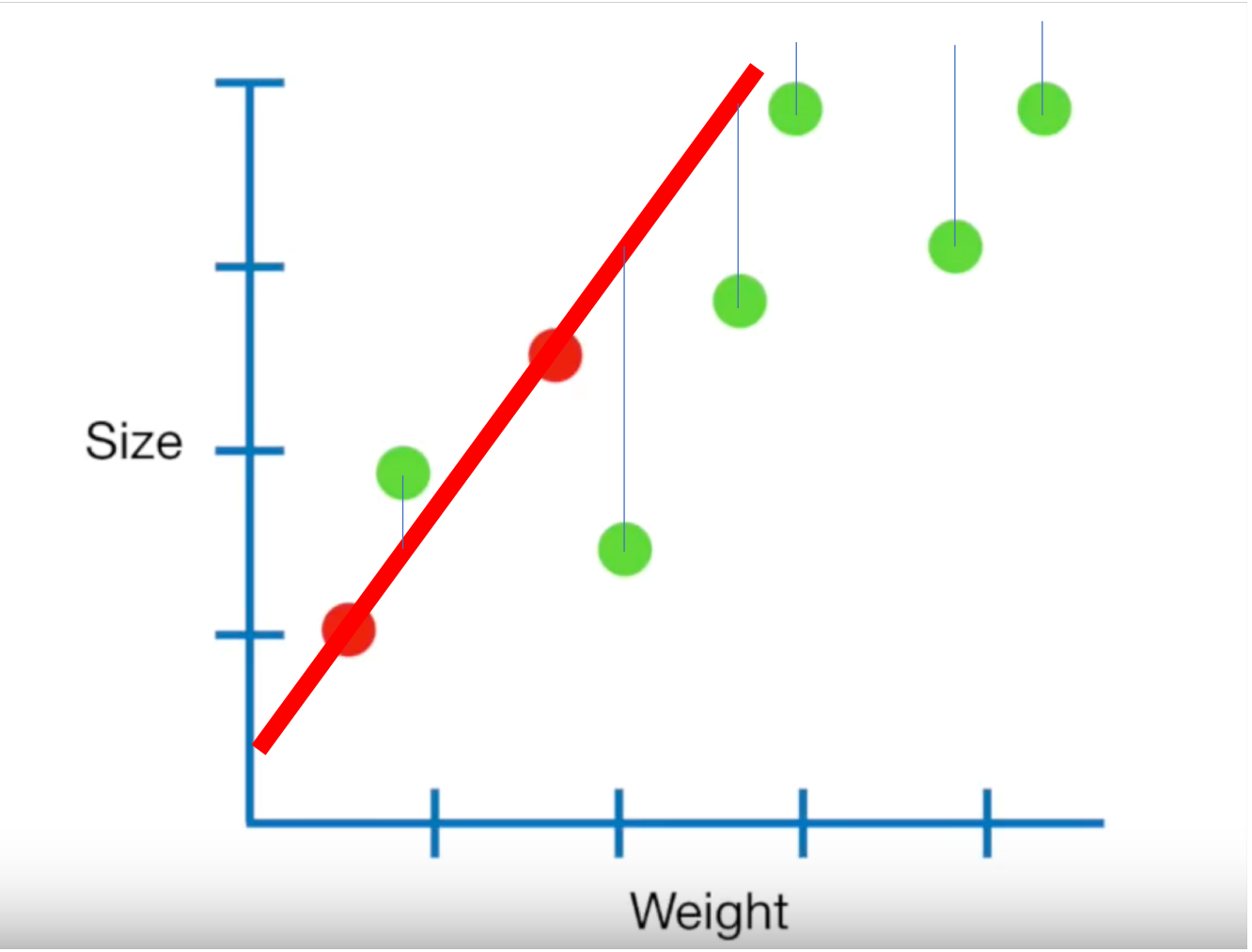
y=price x1=bedrooms x2=bathrooms x3=sqft_living x4=sqft_lot x5=floors
x6=waterfront x7=condition x8=age x9=lat x10=long

Bathroom
Full = Tub, sink, toilet
3/4 = Shower without tub, sink, toilet
1/2 = Sink and toilet

Floors
1.5 floors= lower floor is twice the area of upper floor

SFU

TRAINING AND TESTING



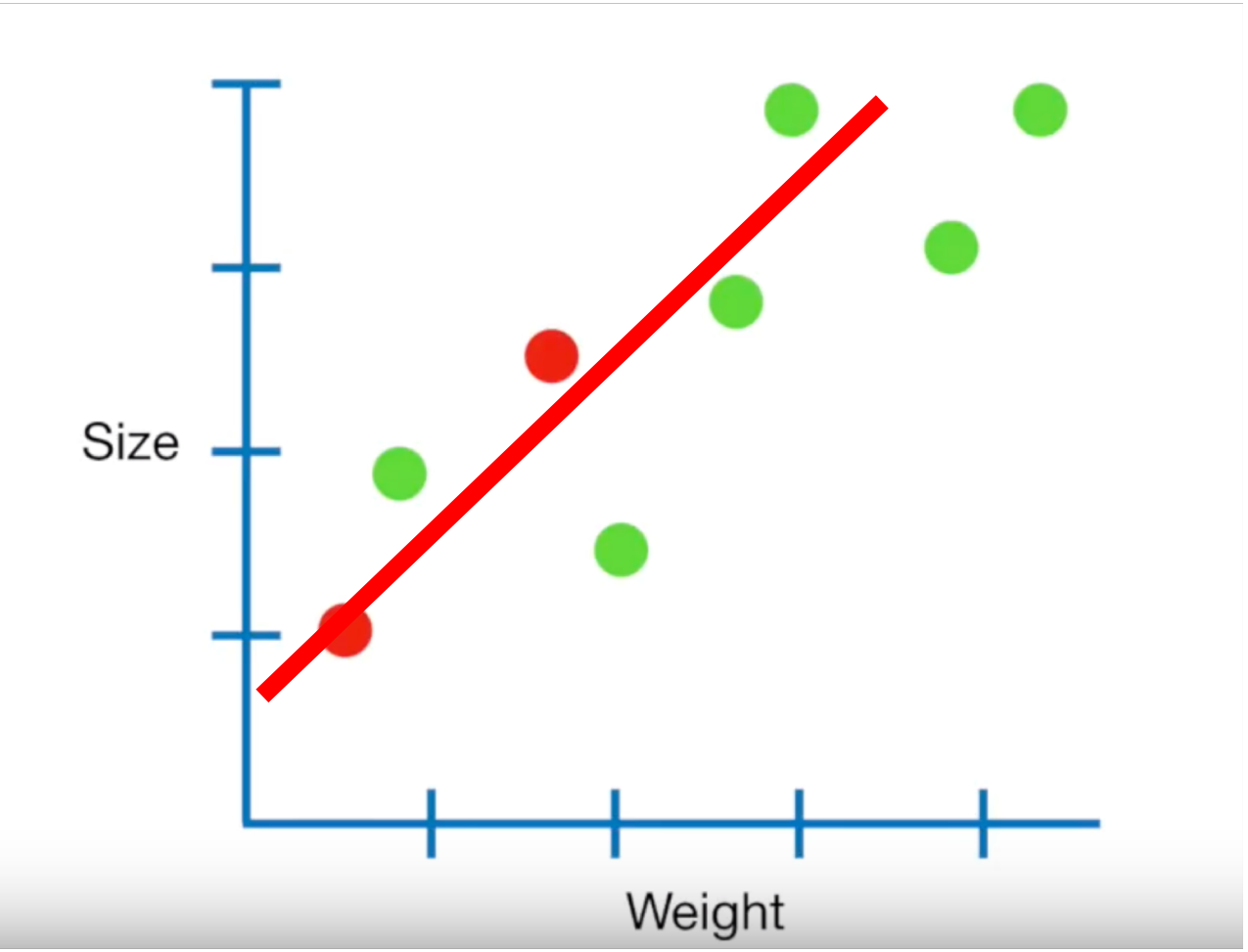
Training data

Testing data

- Fit a line to training dataset using least square regression line.
- The least square regression line has low bias, but high variance

SFU

TRAINING AND TESTING



- Fit a line using Lasso and Ridge regression may look something like this!
- In our example, 1/3 into testing set, 2/3 into training set

- Increase small amount of bias, but decrease in variance

SFU

FORMULA FOR LASSO AND RIDGE

- Lasso:

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

- Ridge regression:

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Lambda can be from 0 to positive infinity
- Lambda can be found by cross validation

SFU

DIFFERENCES BETWEEN LASSO AND RIDGE

- If the Lambda=0

➢ Lasso and Ridge regression line will be the same as least squares line

- If the Lambda=∞

➢Ridge regression can only shrink the slope close to zero, but never equal to 0.

➢Lasso can shrink the slope all the way to 0.

- Lasso Regression can select useful variable from the model.
- Ridge Regression tends to do better job when most of the variables are significant

SFU

Full model

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10)

Residuals:
     Min       1Q   Median       3Q      Max
-1954697  -110138   -12297    81639  3758105

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.132e+07  1.570e+06 -32.689  < 2e-16 ***
x1          -9.276e+04  3.007e+03 -30.845  < 2e-16 ***
x2           1.603e+04  8.499e+02  18.867  < 2e-16 ***
x3           2.826e+02  2.760e+00 102.399  < 2e-16 ***
x4          -4.588e-02  3.959e-02  -1.159  0.24646
x51.5        2.593e+04  5.740e+03   4.517 6.30e-06 ***
x52          1.347e+04  4.120e+03   3.269  0.00108 **
x52.5        1.862e+05  1.768e+04  10.530  < 2e-16 ***
x53          6.047e+04  1.003e+04   6.032 1.65e-09 ***
x53.5        1.913e+05  8.319e+04   2.300  0.02146 *
x61          7.740e+05  1.751e+04  44.207  < 2e-16 ***
x72          3.316e+04  4.418e+04   0.751  0.45293
x73          4.591e+04  4.098e+04   1.120  0.26257
x74          8.051e+04  4.098e+04   1.965  0.04946 *
x75          1.089e+05  4.121e+04   2.643  0.00821 **
x8           1.218e+03  7.928e+01  15.369  < 2e-16 ***
x9           6.377e+05  1.129e+04  56.493  < 2e-16 ***
x10         -1.718e+05  1.228e+04 -13.996  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 219800 on 21577 degrees of freedom
Multiple R-squared:  0.6418,    Adjusted R-squared:  0.6416
F-statistic:  2275 on 17 and 21577 DF,  p-value: < 2.2e-16
```

- Most regressors are significant and reasonable

SFU

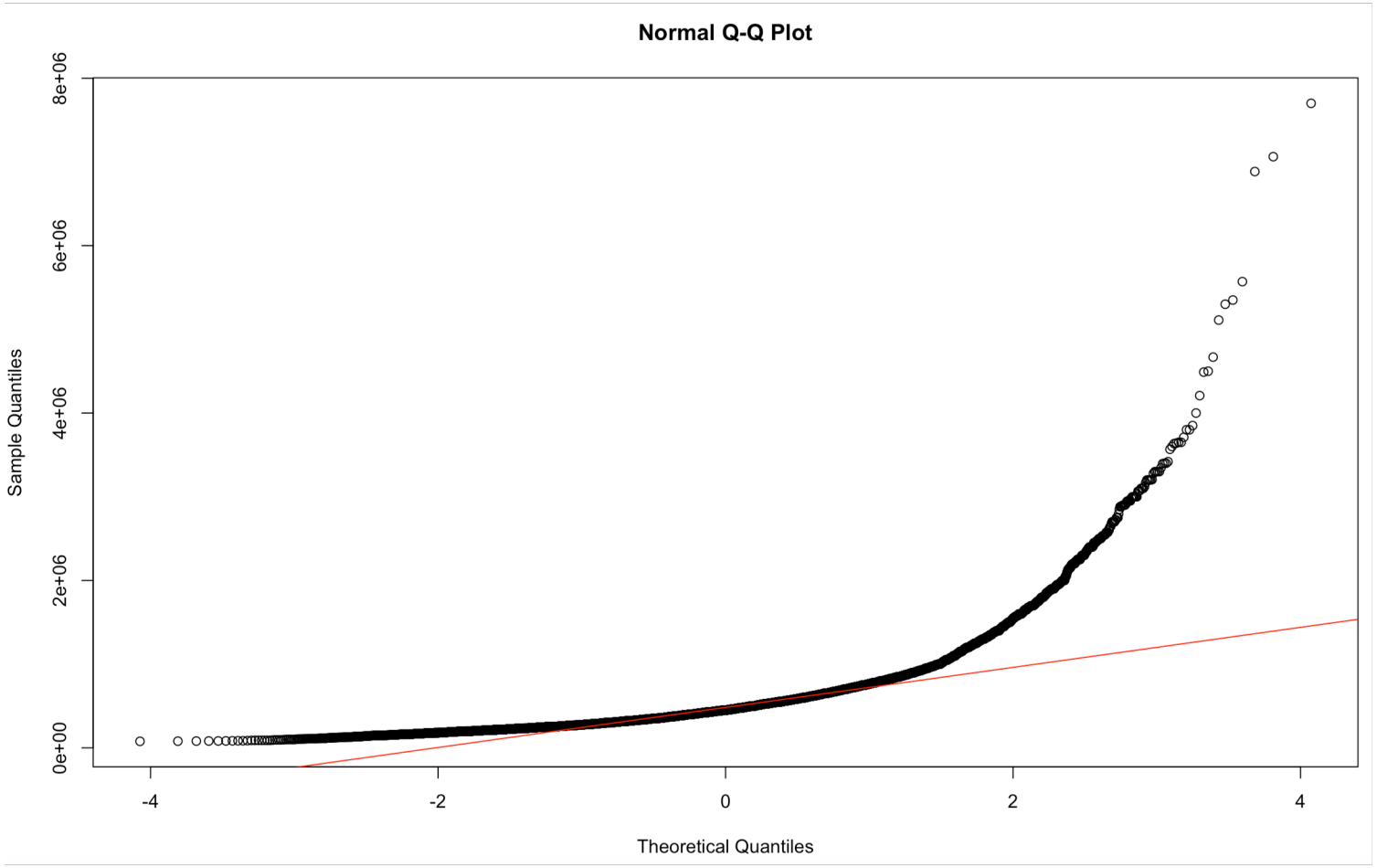
Model Selection

```
Step:  AIC=531276.7
y ~ x1 + x2 + x3 + x5 + x6 + x7 + x8 + x9 + x10

       Df  Sum of Sq        RSS    AIC
<none>                 1.0426e+15 531277
+ x4    1 6.4906e+10 1.0425e+15 531277
- x7    4 7.5819e+12 1.0502e+15 531425
- x5    5 7.8727e+12 1.0505e+15 531429
- x10   1 1.0224e+13 1.0528e+15 531485
- x8    1 1.1372e+13 1.0540e+15 531509
- x2    1 1.7218e+13 1.0598e+15 531628
- x1    1 4.5942e+13 1.0885e+15 532206
- x6    1 9.4405e+13 1.1370e+15 533147
- x9    1 1.5573e+14 1.1983e+15 534281
- x3    1 5.1999e+14 1.5626e+15 540013
```

- $y \sim x1 + x2 + x3 + x5 + x6 + x7 + x8 + x9 + x10$

SFU

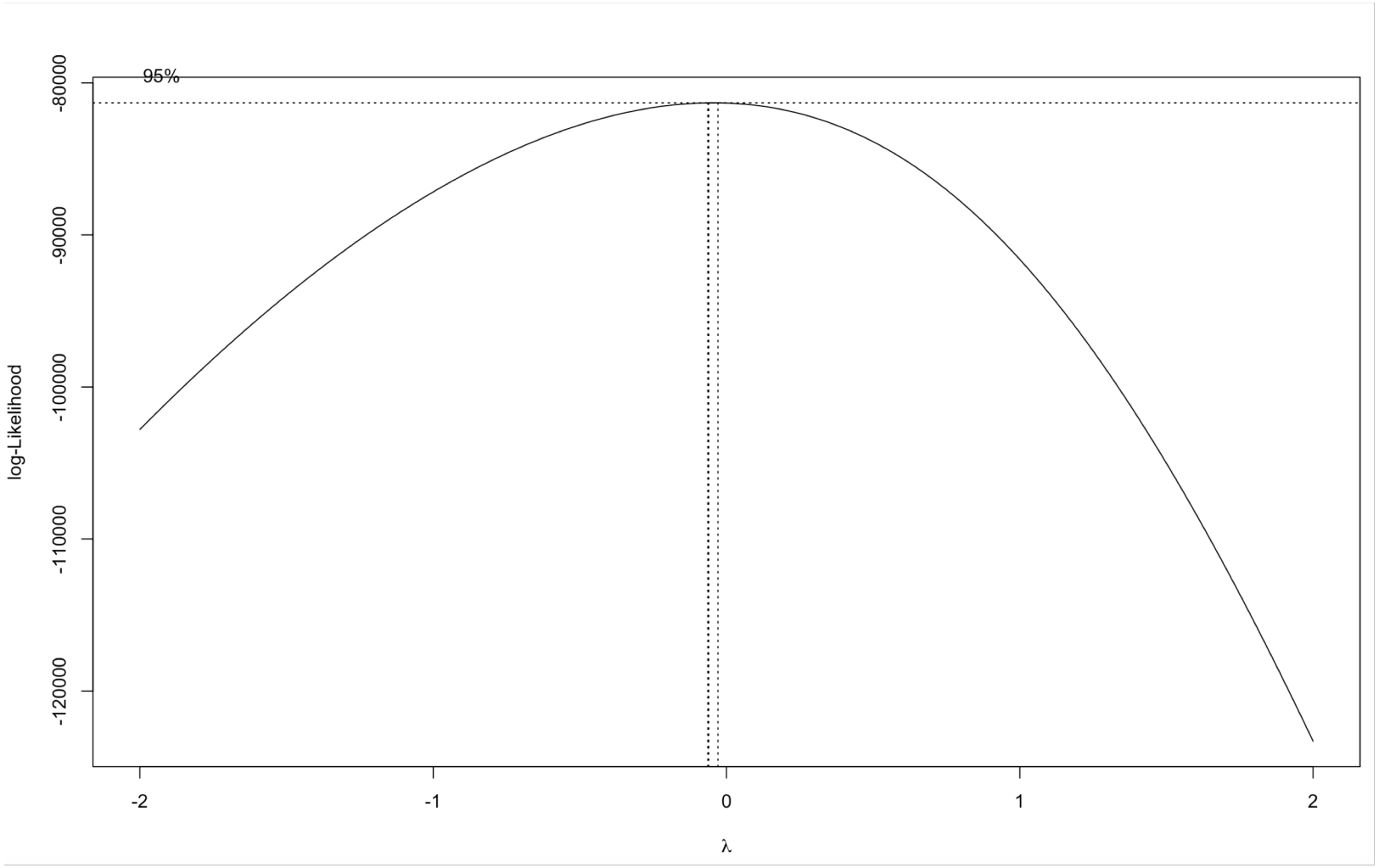
Normal Q-Q plot of respond variable



- Y is not normal
- Needs transformation

SFU

Box-Cox Transformation



- Best lambda= -0.0606
- y -> y^ (-0.0606)

Final model

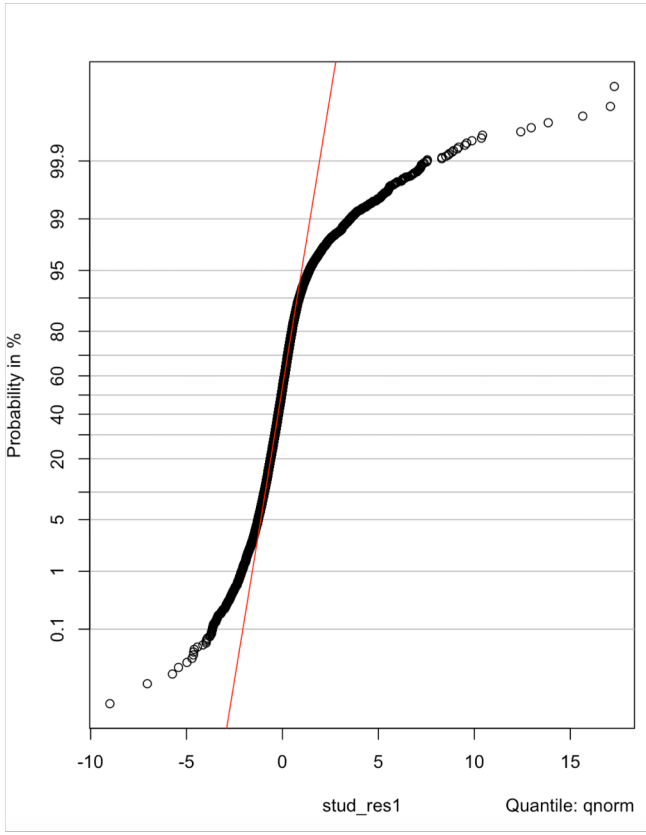
```
Call:
lm(formula = yp ~ x1 + x2 + x3 + x5 + x6 + x7 + x8 + x9 + x10)

Residuals:
      Min        1Q    Median        3Q       Max
-0.039264 -0.004941  0.000008  0.004987  0.086385

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  2.898e+00  5.621e-02   51.551  < 2e-16 ***
x1           1.696e-03  1.103e-04   15.383  < 2e-16 ***
x2          -2.652e-04  3.123e-05   -8.493  < 2e-16 ***
x3          -1.020e-05  9.991e-08 -102.085  < 2e-16 ***
x51.5       -2.714e-03  2.108e-04  -12.872  < 2e-16 ***
x52         -2.815e-03  1.510e-04  -18.645  < 2e-16 ***
x52.5       -5.591e-03  6.498e-04   -8.604  < 2e-16 ***
x53         -3.910e-03  3.684e-04  -10.613  < 2e-16 ***
x53.5       -3.699e-03  3.057e-03   -1.210    0.226
x61         -1.631e-02  6.434e-04  -25.348  < 2e-16 ***
x72         -2.647e-03  1.624e-03   -1.630    0.103
x73         -7.641e-03  1.506e-03   -5.075 3.92e-07 ***
x74         -9.679e-03  1.506e-03   -6.428 1.32e-10 ***
x75         -1.129e-02  1.514e-03   -7.460 8.99e-14 ***
x8          -2.212e-05  2.912e-06   -7.594 3.22e-14 ***
x9          -4.206e-02  4.135e-04 -101.728  < 2e-16 ***
x10          3.396e-03  4.415e-04    7.692 1.51e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.008077 on 21578 degrees of freedom
Multiple R-squared:  0.6848,    Adjusted R-squared:  0.6846
F-statistic:  2930 on 16 and 21578 DF,  p-value: < 2.2e-16
```

- Adjusted R-square rises from 0.6416 to 0.6846

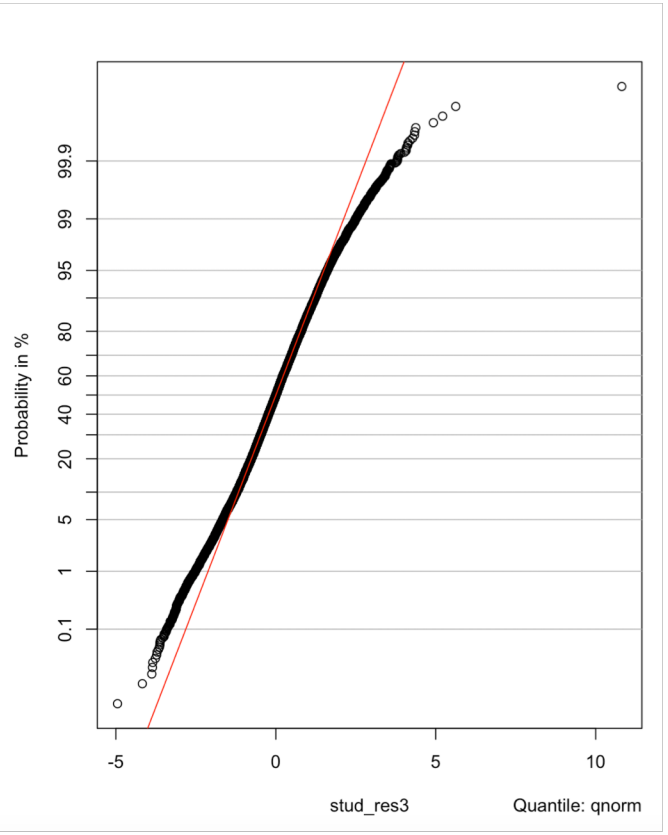
SFU

Comparisons

Full model



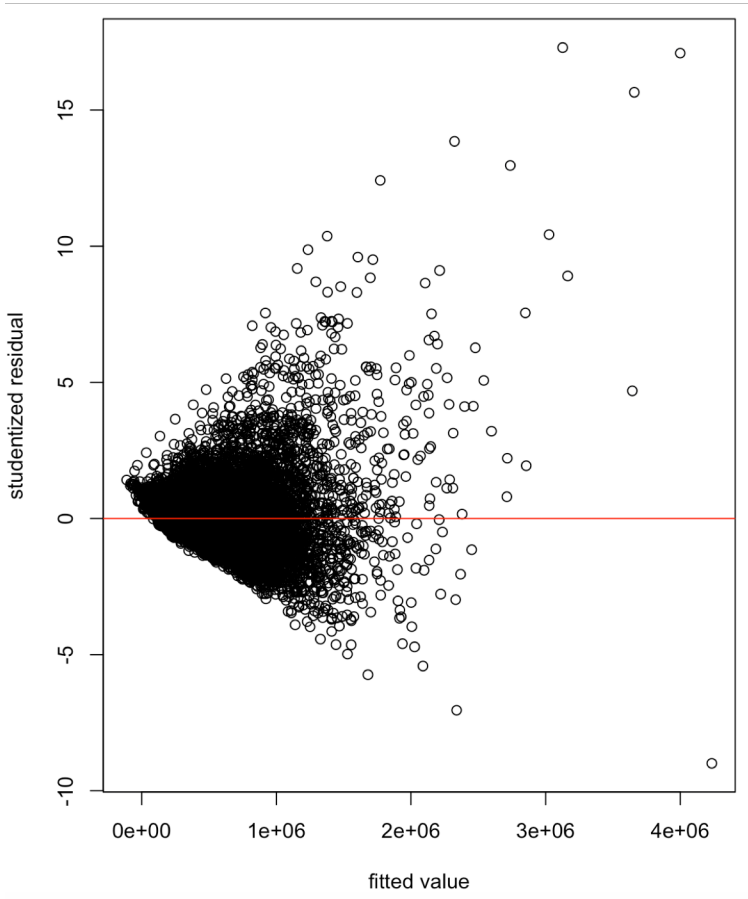
Final model



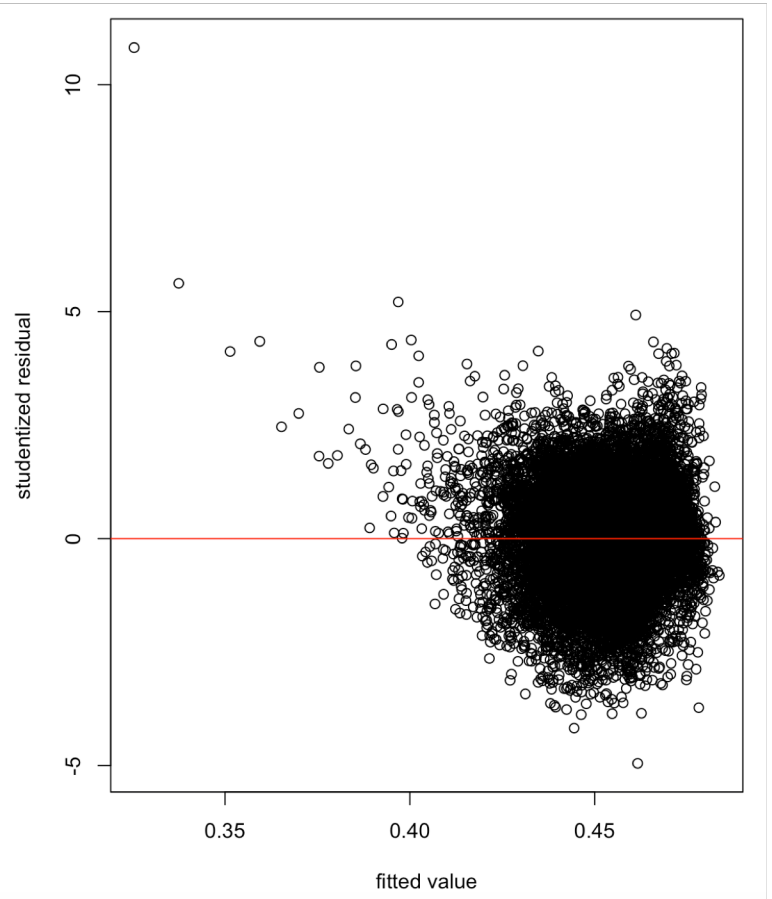
- Normality has improved

Comparisons

Full Model



Final Model



- Constant variance improved a lot

Final model

Final Model: y^(-0.0606) = 2.898 + 0.001696 * x1 - 0.0002652 * x2
- 0.00001020 * x3 - 0.002714 * (x5=1.5) - 0.002815 * (x5=2)
- 0.005591 * (x5=2.5) - 0.003910 * (x5=3) - 0.003699 * (x5=3.5) - 0.01631 * x6
- 0.002647 * (x7=2) - 0.007641 * (x7=3) - 0.009679 * (x7=4) - 0.01129 * (x7=5)
- 0.00002212 * x8 - 0.04206 * x9 + 0.003396 * x10

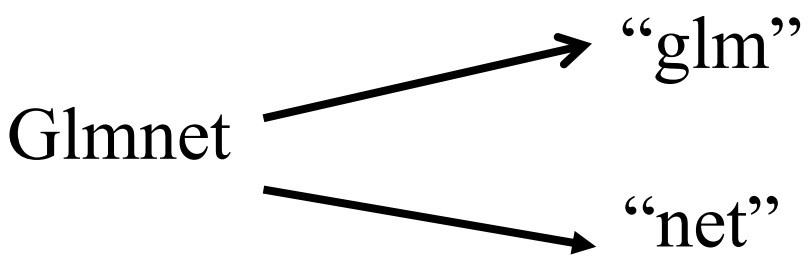
Example: a house with 3 bedrooms, 4 bathrooms, 2000 sqft of living, 2 floors, not waterfront, condition 3, 50 years, 47.50000 latitudes, -122.300 longitudes

The predicted price will be $410919

SFU

Lasso and Ridge Regression

- **library**(glmnet)



$$\text{the sum of the squared residuals} + \lambda \times \left[\alpha \times \left(|\text{variable}_1| + \ldots + |\text{variable}_X|\right) + (1-\alpha) \times \left(\text{variable}_1^2 + \ldots + \text{variable}_X^2\right)\right]$$

#Lasso

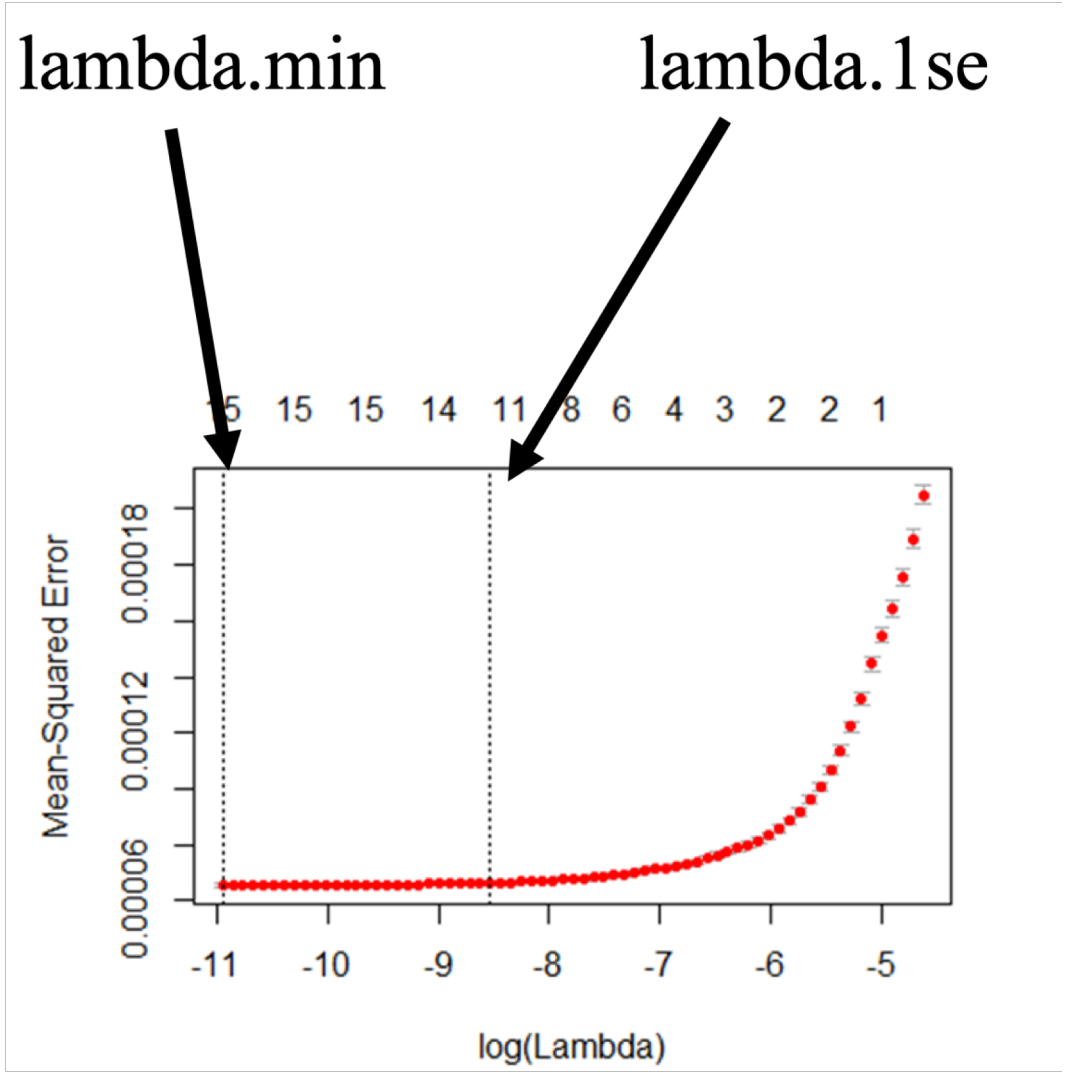
fit1=cv.glmnet(x_training, y_training, type.measure="mse", **alpha=1**)

#Ridge

fit2=cv.glmnet(x_training, y_training, type.measure="mse", **alpha=0**)

SFU

Output



- Lambda.1se will result in the simplest model

SFU

Model comparison

LASSO

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                              1
## (Intercept)  2.847394e+00
## x1           3.500088e-04
## x2           .
## x3          -9.927207e-06
## x4          -3.594295e-09
## x51          2.461796e-03
## x51.5        .
## x52          .
## x52.5       -3.687071e-04
## x53          .
## x53.5        .
## x61         -1.574125e-02
## x72          4.928422e-03
## x73          1.650219e-03
## x74          .
## x75         -1.466760e-03
## x8          -5.976679e-06
## x9          -4.193058e-02
## x10          3.123485e-03
```

RIDGE

```
19 x 1 sparse Matrix of class "dgCMatrix"
                              1
(Intercept)  2.694659e+00
x1           7.526647e-04
x2          -4.006358e-04
x3          -7.799480e-06
x4          -1.217446e-08
x51          1.904197e-03
x51.5       -7.774922e-04
x52         -1.424133e-03
x52.5       -3.962884e-03
x53         -1.759043e-03
x53.5       -1.498042e-03
x61         -1.737917e-02
x72          6.243013e-03
x73          9.805143e-04
x74         -9.466162e-04
x75         -2.470070e-03
x8          -1.756865e-05
x9          -3.855130e-02
x10          3.192170e-03
```

- $y = X\beta$
- Lasso can do model selection
- But Ridge cannot

SFU

Final conclusion

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

- the smaller the better

```
mse_AICs

## [1] 6.642516e-05

mse_lasso

## [1] 6.572709e-05

mse_ridge

## [1] 6.740143e-05
```

SFU

Thank you for listening

King County

Any Questions?

SFU