

STA 506 2.0 Linear Regression Analysis

Lecture 13: Variable Selection and Model Building

Dr Thiyanga S. Talagala

Contents

Recap	2
Qualitative variable with more than 2 levels	3
Your turn	3
Different in both intercept and slope	5
Regression line differ in intercept only	5
Regression line both differ in slope and intercept	5
Variable Selection: Introduction	6
Correlation matrix	8
All possible regression	9
Intercept-only regression model	10
Full model	10
Computational Techniques for Variable Selection	12
Criteria for evaluating subset regression models	12
Forward selection	13
Backward elimination	16
Stepwise regression	18
Model adequacy checking	20
Note	24
Acknowledgement	24

Recap

1. Simple linear regression
2. Multiple linear regression
3. Variable transformations
4. Detection and treatment of outliers: leverage and influence
5. Indicator variables

Qualitative variable with more than 2 levels

In general, a qualitative variable with k levels is represented by $k - 1$ indicator variables, each taking the values 0 and 1.

	IQ	BMI	headcir	D1	D2
1	10	Normal	50.2	0	1
2	20	Normal	50.5	0	1
3	100	Obese	58.5	0	0
4	98	Obese	55.0	0	0
5	100	Underweight	54.9	1	0
6	11	Underweight	40.0	1	0
7	50	Underweight	48.5	1	0
8	70	Underweight	50.0	1	0

D_1	D_2	Description
1	0	observation is from underweight
0	1	observation is from normal
0	0	observation is from Obese

Your turn

Write the regression equations for the three levels.

	IQ	headcir	D1	D2
1	10	50.2	0	1
2	20	50.5	0	1
3	100	58.5	0	0
4	98	55.0	0	0
5	100	54.9	1	0
6	11	40.0	1	0
7	50	48.5	1	0
8	70	50.0	1	0

$$D_{1i} = \begin{cases} 1 & \text{if underweight} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$D_{2i} = \begin{cases} 1 & \text{if normal} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Let x_i be the head circumference

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_{1i} + \beta_3 D_{2i} + \epsilon_i$$

For underweight

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 + \epsilon_i$$

For normal

$$y_i = \beta_0 + \beta_1 x_i + \beta_3 + \epsilon_i$$

For overweight

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Different in both intercept and slope

	IQ	Gender	BMI
1	10	Male	20.2
2	20	Male	20.5
3	100	Male	18.5
4	98	Male	25.0
5	100	Female	24.9
6	11	Female	31.0
7	50	Female	18.5
8	70	Female	20.0

Indicator variable for Gender

$$D_i = \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases} \quad (3)$$

The choice of 0 and 1 to identify the levels of a qualitative variable is arbitrary.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + \epsilon_i,$$

Regression line differ in intercept only

Regression equation for **males**, $D_i = 1$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 + \epsilon_i,$$

Regression equation for **females**, $D_i = 0$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

Regression line both differ in slope and intercept

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + \beta_3 x_i D_i + \epsilon_i,$$

Regression equation for **males**, $D_i = 1$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 + \beta_3 x_i + \epsilon_i,$$

$$y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_i + \epsilon_i,$$

Regression equation for **females**, $D_i = 0$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

Variable Selection: Introduction

Variable selection: Finding an appropriate subset of regression for the model.

```
library(tidyverse)
realestate <- read.csv("real-estate.csv")
head(realestate)
```

```
   ID Price Sqft Bedroom Bathroom Airconditioning Garage Pool YearBuild Quality
1  1 360000 3032      4         4              1      2    0      1972      2
2  2 340000 2058      4         2              1      2    0      1976      2
3  3 250000 1780      4         3              1      2    0      1980      2
4  4 205500 1638      4         2              1      2    0      1963      2
5  5 275500 2196      4         3              1      2    0      1968      2
6  6 248000 1966      4         3              1      5    1      1972      2
  Lot AdjHighway
1 22221         0
2 22912         0
3 21345         0
4 17342         0
5 21786         0
6 18902         0
```

```
glimpse(realestate)
```

```
Rows: 522
Columns: 12
$ ID          <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, ~
$ Price       <int> 360000, 340000, 250000, 205500, 275500, 248000, 229900~
$ Sqft        <int> 3032, 2058, 1780, 1638, 2196, 1966, 2216, 1597, 1622, ~
$ Bedroom     <int> 4, 4, 4, 4, 4, 4, 3, 2, 3, 3, 7, 3, 5, 5, 3, 5, 2, 3, ~
$ Bathroom    <int> 4, 2, 3, 2, 3, 3, 2, 1, 2, 3, 5, 4, 4, 4, 3, 5, 2, 4, ~
$ Airconditioning <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, ~
$ Garage      <int> 2, 2, 2, 2, 2, 5, 2, 1, 2, 1, 2, 3, 3, 2, 2, 2, 2, 2, ~
$ Pool        <int> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, ~
$ YearBuild   <int> 1972, 1976, 1980, 1963, 1968, 1972, 1972, 1955, 1975, ~
$ Quality     <int> 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 1, 1, 1, 2, 2, 2, 2, ~
$ Lot         <int> 22221, 22912, 21345, 17342, 21786, 18902, 18639, 22112~
$ AdjHighway  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
```

```
summary(realestate)
```

ID	Price	Sqft	Bedroom
Min. : 1.0	Min. : 84000	Min. : 980	Min. : 0.000
1st Qu.: 131.2	1st Qu.: 180000	1st Qu.: 1701	1st Qu.: 3.000
Median : 261.5	Median : 229900	Median : 2061	Median : 3.000
Mean : 261.5	Mean : 277894	Mean : 2261	Mean : 3.471
3rd Qu.: 391.8	3rd Qu.: 335000	3rd Qu.: 2636	3rd Qu.: 4.000
Max. : 522.0	Max. : 920000	Max. : 5032	Max. : 7.000
Bathroom	Airconditioning	Garage	Pool
Min. : 0.000	Min. : 0.0000	Min. : 0.0	Min. : 0.00000
1st Qu.: 2.000	1st Qu.: 1.0000	1st Qu.: 2.0	1st Qu.: 0.00000

Median :3.000	Median :1.0000	Median :2.0	Median :0.00000
Mean :2.642	Mean :0.8314	Mean :2.1	Mean :0.06897
3rd Qu.:3.000	3rd Qu.:1.0000	3rd Qu.:2.0	3rd Qu.:0.00000
Max. :7.000	Max. :1.0000	Max. :7.0	Max. :1.00000
YearBuild	Quality	Lot	AdjHighway
Min. :1885	Min. :1.000	Min. : 4560	Min. :0.00000
1st Qu.:1956	1st Qu.:2.000	1st Qu.:17205	1st Qu.:0.00000
Median :1966	Median :2.000	Median :22200	Median :0.00000
Mean :1967	Mean :2.184	Mean :24370	Mean :0.02107
3rd Qu.:1981	3rd Qu.:3.000	3rd Qu.:26787	3rd Qu.:0.00000
Max. :1998	Max. :3.000	Max. :86830	Max. :1.00000

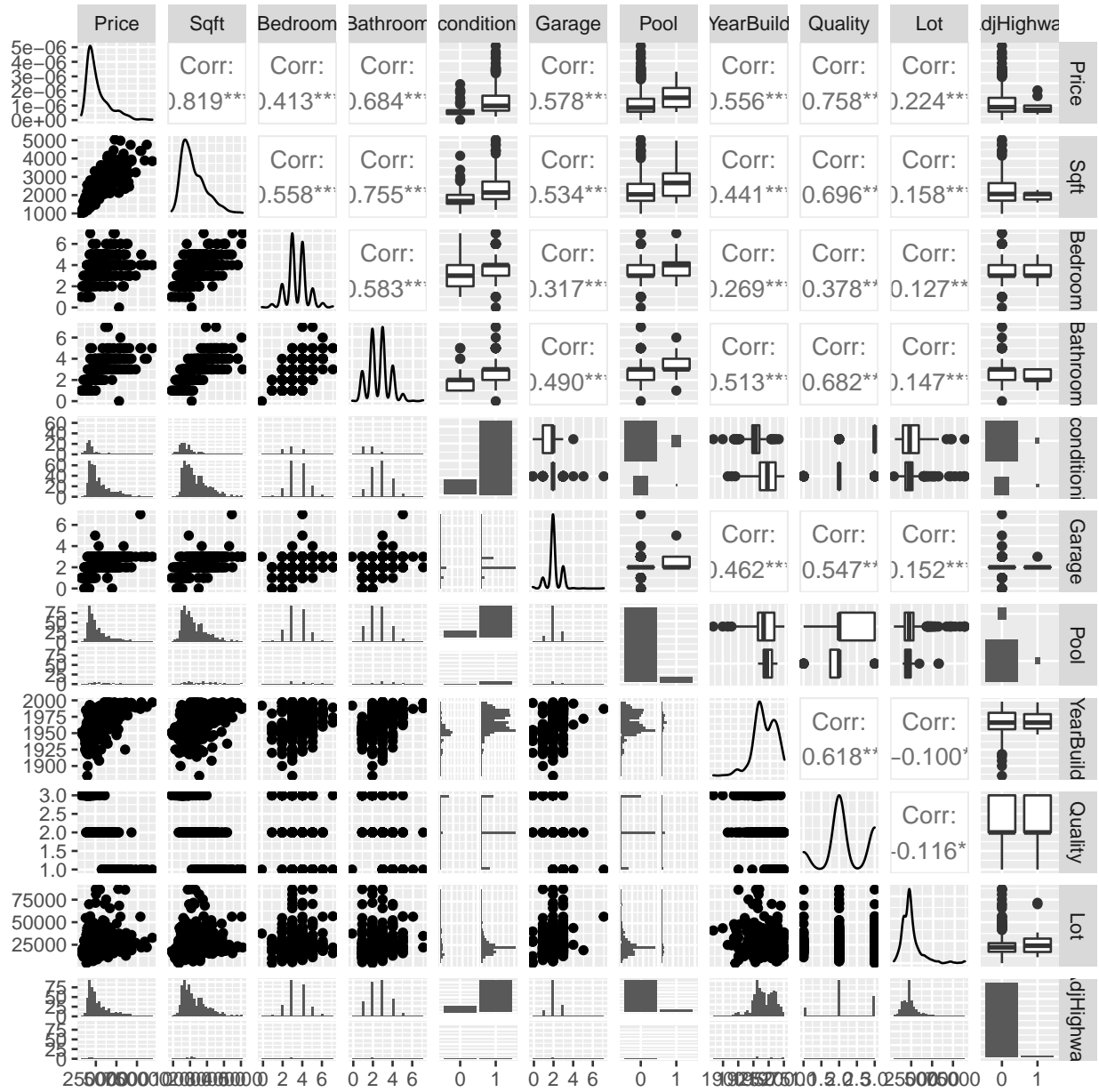
```

realestate$Airconditioning <- factor(realestate$Airconditioning)
realestate$Pool <- factor(realestate$Pool)
realestate$AdjHighway <- factor(realestate$AdjHighway)
summary(realestate)

```

ID	Price	Sqft	Bedroom
Min. : 1.0	Min. : 84000	Min. : 980	Min. :0.000
1st Qu.:131.2	1st Qu.:180000	1st Qu.:1701	1st Qu.:3.000
Median :261.5	Median :229900	Median :2061	Median :3.000
Mean :261.5	Mean :277894	Mean :2261	Mean :3.471
3rd Qu.:391.8	3rd Qu.:335000	3rd Qu.:2636	3rd Qu.:4.000
Max. :522.0	Max. :920000	Max. :5032	Max. :7.000
Bathroom	Airconditioning	Garage	Pool
Min. :0.000	0: 88	Min. :0.0	0:486
1st Qu.:2.000	1:434	1st Qu.:2.0	1: 36
Median :3.000		Median :2.0	
Mean :2.642		Mean :2.1	
3rd Qu.:3.000		3rd Qu.:2.0	
Max. :7.000		Max. :7.0	
Quality	Lot	AdjHighway	YearBuild
Min. :1.000	Min. : 4560	0:511	Min. :1885
1st Qu.:2.000	1st Qu.:17205	1: 11	1st Qu.:1956
Median :2.000	Median :22200		Median :1966
Mean :2.184	Mean :24370		Mean :1967
3rd Qu.:3.000	3rd Qu.:26787		3rd Qu.:1981
Max. :3.000	Max. :86830		Max. :1998

Correlation matrix



All possible regression

In-class

Intercept-only regression model

```
realty.lm.minimal <- lm(Price ~ 1, data=realestate)
realty.lm.minimal
```

Call:

```
lm(formula = Price ~ 1, data = realestate)
```

Coefficients:

```
(Intercept)
  277894
```

Full model

```
realty.lm.all <- lm(Price ~ . , data=realestate.var)
realty.lm.all
```

Call:

```
lm(formula = Price ~ . , data = realestate.var)
```

Coefficients:

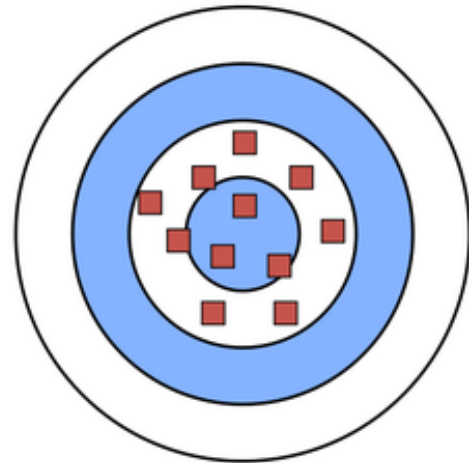
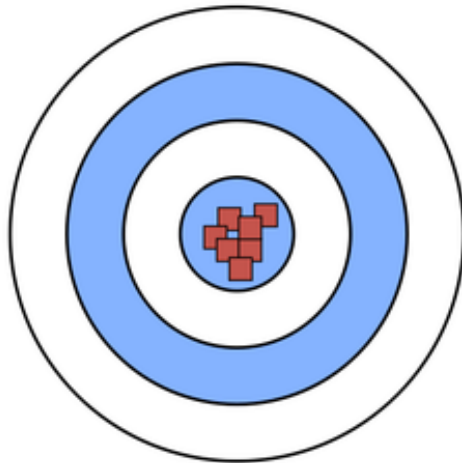
(Intercept)	Sqft	Bedroom	Bathroom
-2.390e+06	1.075e+02	-9.712e+03	-1.067e+02
Airconditioning1	Garage	Pool1	YearBuild
-1.222e+04	1.732e+04	1.249e+04	1.279e+03
Quality	Lot	AdjHighway1	
-5.390e+04	1.422e+00	-2.717e+04	

We need model to include as few independent variables as possible because the variance of the the predictions increases as the number of independent variables increases.

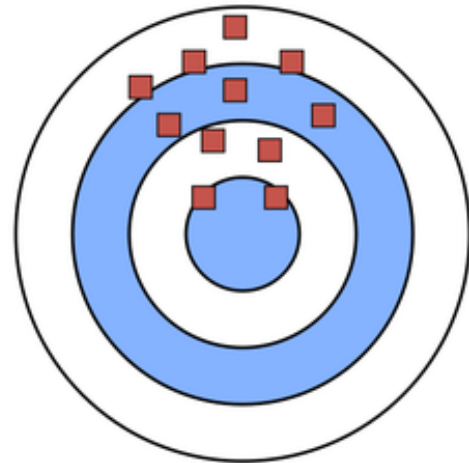
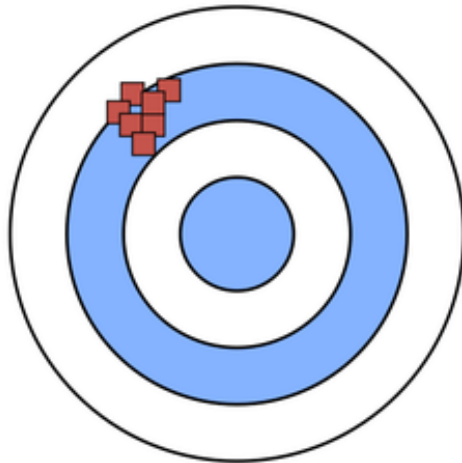
Low Variance
(Precise)

High Variance
(Not Precise)

Low Bias
(Accurate)



High Bias
(Not Accurate)



This work by Sebastian Raschka is licensed under a Creative Commons Attribution 4.0 International License.

Computational Techniques for Variable Selection

1. All possible regression
2. Stepwise regression methods
 - 2.1 Forward selection
 - 2.2 Backward elimination
 - 2.3 Stepwise regression

Criteria for evaluating subset regression models

- lowest p-value
- highest adjusted R_{adj}^2
- lowest Mallows's C_p
- lowest AIC
- lowest residual mean square
- lowest score under cross-validation, etc.

Forward selection

- Starts with intercept-only regression model.
- Then we gradually add one more variable at a time (or add main effects first, then interactions).

Intercept-only regression model

```
realty.lm.minimal <- lm(Price ~ 1, data=realestate)
realty.lm.minimal
```

Call:

```
lm(formula = Price ~ 1, data = realestate)
```

Coefficients:

```
(Intercept)
  277894
```

Forward selection

```
step(realty.lm.minimal, scope=list(upper = realty.lm.all,
lower= realty.lm.minimal), direction="forward",
trace=0)
```

Call:

```
lm(formula = Price ~ Sqft + Quality + YearBuild + Lot + Garage +
  Bedroom, data = realestate)
```

Coefficients:

```
(Intercept)      Sqft      Quality      YearBuild      Lot      Garage
-2.233e+06  1.093e+02 -5.223e+04  1.191e+03  1.415e+00  1.665e+04
  Bedroom
-1.007e+04
```

```
step(realty.lm.minimal, scope=list(upper = realty.lm.all,
lower= realty.lm.minimal), direction="forward",
trace=1)
```

Start: AIC=12356.17

Price ~ 1

	Df	Sum of Sq	RSS	AIC
+ Sqft	1	6.6555e+12	3.2554e+12	11777
+ Quality	1	5.6956e+12	4.2153e+12	11912
+ Bathroom	1	4.6326e+12	5.2783e+12	12029
+ Garage	1	3.3086e+12	6.6023e+12	12146
+ YearBuild	1	3.0585e+12	6.8524e+12	12166
+ Bedroom	1	1.6931e+12	8.2178e+12	12260
+ Airconditioning	1	8.2546e+11	9.0855e+12	12313
+ Lot	1	4.9804e+11	9.4129e+12	12331
+ Pool	1	2.1303e+11	9.6979e+12	12347

<none> 9.9109e+12 12356
 + AdjHighway 1 2.5746e+10 9.8852e+12 12357

Step: AIC=11777.02
 Price ~ Sqft

	Df	Sum of Sq	RSS	AIC
+ Quality	1	6.7926e+11	2.5762e+12	11657
+ YearBuild	1	4.6302e+11	2.7924e+12	11699
+ Garage	1	2.7313e+11	2.9823e+12	11733
+ Bathroom	1	9.6767e+10	3.1587e+12	11763
+ Lot	1	9.1880e+10	3.1635e+12	11764
+ Airconditioning	1	5.0865e+10	3.2046e+12	11771
+ Bedroom	1	2.7613e+10	3.2278e+12	11775
<none>			3.2554e+12	11777
+ Pool	1	1.8642e+09	3.2536e+12	11779
+ AdjHighway	1	1.6494e+07	3.2554e+12	11779

Step: AIC=11656.86
 Price ~ Sqft + Quality

	Df	Sum of Sq	RSS	AIC
+ YearBuild	1	1.0457e+11	2.4716e+12	11637
+ Garage	1	8.8087e+10	2.4881e+12	11641
+ Lot	1	8.7374e+10	2.4888e+12	11641
+ Bedroom	1	2.3350e+10	2.5528e+12	11654
<none>			2.5762e+12	11657
+ Airconditioning	1	2.2920e+09	2.5739e+12	11658
+ Bathroom	1	1.4980e+09	2.5747e+12	11659
+ AdjHighway	1	8.5057e+08	2.5753e+12	11659
+ Pool	1	8.3092e+08	2.5753e+12	11659

Step: AIC=11637.23
 Price ~ Sqft + Quality + YearBuild

	Df	Sum of Sq	RSS	AIC
+ Lot	1	1.4256e+11	2.3290e+12	11608
+ Garage	1	5.7571e+10	2.4140e+12	11627
+ Bedroom	1	2.7902e+10	2.4437e+12	11633
+ Airconditioning	1	1.6548e+10	2.4550e+12	11636
<none>			2.4716e+12	11637
+ AdjHighway	1	2.0662e+09	2.4695e+12	11639
+ Pool	1	1.3587e+09	2.4702e+12	11639
+ Bathroom	1	3.3406e+08	2.4713e+12	11639

Step: AIC=11608.22
 Price ~ Sqft + Quality + YearBuild + Lot

	Df	Sum of Sq	RSS	AIC
+ Garage	1	3.6990e+10	2.2920e+12	11602
+ Bedroom	1	3.5910e+10	2.2931e+12	11602
<none>			2.3290e+12	11608
+ AdjHighway	1	7.1918e+09	2.3218e+12	11609
+ Airconditioning	1	7.0188e+09	2.3220e+12	11609

```

+ Pool          1 4.1055e+09 2.3249e+12 11609
+ Bathroom     1 2.5720e+09 2.3265e+12 11610

```

Step: AIC=11601.86

Price ~ Sqft + Quality + YearBuild + Lot + Garage

```

          Df Sum of Sq      RSS   AIC
+ Bedroom      1 3.7251e+10 2.2548e+12 11595
+ Airconditioning 1 1.1182e+10 2.2809e+12 11601
<none>                2.2920e+12 11602
+ AdjHighway    1 7.3877e+09 2.2847e+12 11602
+ Pool          1 3.2365e+09 2.2888e+12 11603
+ Bathroom     1 3.0575e+09 2.2890e+12 11603

```

Step: AIC=11595.31

Price ~ Sqft + Quality + YearBuild + Lot + Garage + Bedroom

```

          Df Sum of Sq      RSS   AIC
<none>                2.2548e+12 11595
+ AdjHighway    1 7444567926 2.2473e+12 11596
+ Airconditioning 1 6973094059 2.2478e+12 11596
+ Pool          1 4676597321 2.2501e+12 11596
+ Bathroom     1  35810304 2.2548e+12 11597

```

Call:

```

lm(formula = Price ~ Sqft + Quality + YearBuild + Lot + Garage +
    Bedroom, data = realestate)

```

Coefficients:

```

(Intercept)      Sqft      Quality  YearBuild      Lot      Garage
-2.233e+06  1.093e+02 -5.223e+04  1.191e+03  1.415e+00  1.665e+04
Bedroom
-1.007e+04

```

Backward elimination

- we start with the full model and gradually delete variables one at a time.

```
step(reealty.lm.all, direction="backward", trace=0)
```

Call:

```
lm(formula = Price ~ Sqft + Bedroom + Garage + YearBuild + Quality +  
    Lot, data = realestate.var)
```

Coefficients:

```
(Intercept)      Sqft      Bedroom      Garage      YearBuild      Quality  
-2.233e+06  1.093e+02 -1.007e+04  1.665e+04  1.191e+03 -5.223e+04  
    Lot  
1.415e+00
```

```
step(reealty.lm.all, direction="backward", trace=1)
```

Start: AIC=11598.65

```
Price ~ Sqft + Bedroom + Bathroom + Airconditioning + Garage +  
    Pool + YearBuild + Quality + Lot + AdjHighway
```

	Df	Sum of Sq	RSS	AIC
- Bathroom	1	2.1979e+06	2.2348e+12	11597
- Pool	1	4.9762e+09	2.2397e+12	11598
- AdjHighway	1	7.7808e+09	2.2425e+12	11598
- Airconditioning	1	8.1526e+09	2.2429e+12	11599
<none>			2.2348e+12	11599
- Bedroom	1	3.0812e+10	2.2656e+12	11604
- Garage	1	4.0949e+10	2.2757e+12	11606
- Lot	1	1.2533e+11	2.3601e+12	11625
- YearBuild	1	1.3792e+11	2.3727e+12	11628
- Quality	1	2.1663e+11	2.4514e+12	11645
- Sqft	1	9.9451e+11	3.2293e+12	11789

Step: AIC=11596.65

```
Price ~ Sqft + Bedroom + Airconditioning + Garage + Pool + YearBuild +  
    Quality + Lot + AdjHighway
```

	Df	Sum of Sq	RSS	AIC
- Pool	1	4.9985e+09	2.2398e+12	11596
- AdjHighway	1	7.7816e+09	2.2426e+12	11596
- Airconditioning	1	8.1628e+09	2.2429e+12	11597
<none>			2.2348e+12	11597
- Bedroom	1	3.4220e+10	2.2690e+12	11603
- Garage	1	4.0949e+10	2.2757e+12	11604
- Lot	1	1.2607e+11	2.3608e+12	11623
- YearBuild	1	1.4205e+11	2.3768e+12	11627
- Quality	1	2.2740e+11	2.4622e+12	11645
- Sqft	1	1.1742e+12	3.4089e+12	11815

Step: AIC=11595.82

```
Price ~ Sqft + Bedroom + Airconditioning + Garage + YearBuild +
```


Quality + Lot + AdjHighway

	Df	Sum of Sq	RSS	AIC
- Airconditioning	1	7.5771e+09	2.2473e+12	11596
- AdjHighway	1	8.0486e+09	2.2478e+12	11596
<none>			2.2398e+12	11596
- Bedroom	1	3.2942e+10	2.2727e+12	11601
- Garage	1	4.1888e+10	2.2817e+12	11604
- Lot	1	1.2320e+11	2.3630e+12	11622
- YearBuild	1	1.3949e+11	2.3793e+12	11625
- Quality	1	2.2897e+11	2.4687e+12	11645
- Sqft	1	1.1908e+12	3.4306e+12	11816

Step: AIC=11595.58

Price ~ Sqft + Bedroom + Garage + YearBuild + Quality + Lot +
AdjHighway

	Df	Sum of Sq	RSS	AIC
- AdjHighway	1	7.4446e+09	2.2548e+12	11595
<none>			2.2473e+12	11596
- Bedroom	1	3.7308e+10	2.2847e+12	11602
- Garage	1	3.8532e+10	2.2859e+12	11602
- YearBuild	1	1.3231e+11	2.3797e+12	11623
- Lot	1	1.3423e+11	2.3816e+12	11624
- Quality	1	2.2142e+11	2.4688e+12	11643
- Sqft	1	1.2209e+12	3.4682e+12	11820

Step: AIC=11595.31

Price ~ Sqft + Bedroom + Garage + YearBuild + Quality + Lot

	Df	Sum of Sq	RSS	AIC
<none>			2.2548e+12	11595
- Bedroom	1	3.7251e+10	2.2920e+12	11602
- Garage	1	3.8331e+10	2.2931e+12	11602
- YearBuild	1	1.2861e+11	2.3834e+12	11622
- Lot	1	1.2923e+11	2.3840e+12	11622
- Quality	1	2.2231e+11	2.4771e+12	11642
- Sqft	1	1.2399e+12	3.4947e+12	11822

Call:

lm(formula = Price ~ Sqft + Bedroom + Garage + YearBuild + Quality +
Lot, data = realestate.var)

Coefficients:

(Intercept)	Sqft	Bedroom	Garage	YearBuild	Quality
-2.233e+06	1.093e+02	-1.007e+04	1.665e+04	1.191e+03	-5.223e+04
Lot					
1.415e+00					

Stepwise regression

```
step(realty.lm.minimal, scope=list(upper = realty.lm.all,  
lower= realty.lm.minimal), direction="both", trace=0)
```

Call:

```
lm(formula = Price ~ Sqft + Quality + YearBuild + Lot + Garage +  
Bedroom, data = realestate)
```

Coefficients:

```
(Intercept)      Sqft      Quality      YearBuild      Lot      Garage  
-2.233e+06  1.093e+02 -5.223e+04  1.191e+03  1.415e+00  1.665e+04  
Bedroom  
-1.007e+04
```

```
step(realty.lm.minimal, scope=list(upper = realty.lm.all,  
lower= realty.lm.minimal), direction="both", trace=1)
```

Start: AIC=12356.17

Price ~ 1

	Df	Sum of Sq	RSS	AIC
+ Sqft	1	6.6555e+12	3.2554e+12	11777
+ Quality	1	5.6956e+12	4.2153e+12	11912
+ Bathroom	1	4.6326e+12	5.2783e+12	12029
+ Garage	1	3.3086e+12	6.6023e+12	12146
+ YearBuild	1	3.0585e+12	6.8524e+12	12166
+ Bedroom	1	1.6931e+12	8.2178e+12	12260
+ Airconditioning	1	8.2546e+11	9.0855e+12	12313
+ Lot	1	4.9804e+11	9.4129e+12	12331
+ Pool	1	2.1303e+11	9.6979e+12	12347
<none>			9.9109e+12	12356
+ AdjHighway	1	2.5746e+10	9.8852e+12	12357

Step: AIC=11777.02

Price ~ Sqft

	Df	Sum of Sq	RSS	AIC
+ Quality	1	6.7926e+11	2.5762e+12	11657
+ YearBuild	1	4.6302e+11	2.7924e+12	11699
+ Garage	1	2.7313e+11	2.9823e+12	11733
+ Bathroom	1	9.6767e+10	3.1587e+12	11763
+ Lot	1	9.1880e+10	3.1635e+12	11764
+ Airconditioning	1	5.0865e+10	3.2046e+12	11771
+ Bedroom	1	2.7613e+10	3.2278e+12	11775
<none>			3.2554e+12	11777
+ Pool	1	1.8642e+09	3.2536e+12	11779
+ AdjHighway	1	1.6494e+07	3.2554e+12	11779
- Sqft	1	6.6555e+12	9.9109e+12	12356

Step: AIC=11656.86

Price ~ Sqft + Quality

	Df	Sum of Sq	RSS	AIC
+ YearBuild	1	1.0457e+11	2.4716e+12	11637
+ Garage	1	8.8087e+10	2.4881e+12	11641
+ Lot	1	8.7374e+10	2.4888e+12	11641
+ Bedroom	1	2.3350e+10	2.5528e+12	11654
<none>			2.5762e+12	11657
+ Airconditioning	1	2.2920e+09	2.5739e+12	11658
+ Bathroom	1	1.4980e+09	2.5747e+12	11659
+ AdjHighway	1	8.5057e+08	2.5753e+12	11659
+ Pool	1	8.3092e+08	2.5753e+12	11659
- Quality	1	6.7926e+11	3.2554e+12	11777
- Sqft	1	1.6391e+12	4.2153e+12	11912

Step: AIC=11637.23

Price ~ Sqft + Quality + YearBuild

	Df	Sum of Sq	RSS	AIC
+ Lot	1	1.4256e+11	2.3290e+12	11608
+ Garage	1	5.7571e+10	2.4140e+12	11627
+ Bedroom	1	2.7902e+10	2.4437e+12	11633
+ Airconditioning	1	1.6548e+10	2.4550e+12	11636
<none>			2.4716e+12	11637
+ AdjHighway	1	2.0662e+09	2.4695e+12	11639
+ Pool	1	1.3587e+09	2.4702e+12	11639
+ Bathroom	1	3.3406e+08	2.4713e+12	11639
- YearBuild	1	1.0457e+11	2.5762e+12	11657
- Quality	1	3.2082e+11	2.7924e+12	11699
- Sqft	1	1.6214e+12	4.0930e+12	11898

Step: AIC=11608.22

Price ~ Sqft + Quality + YearBuild + Lot

	Df	Sum of Sq	RSS	AIC
+ Garage	1	3.6990e+10	2.2920e+12	11602
+ Bedroom	1	3.5910e+10	2.2931e+12	11602
<none>			2.3290e+12	11608
+ AdjHighway	1	7.1918e+09	2.3218e+12	11609
+ Airconditioning	1	7.0188e+09	2.3220e+12	11609
+ Pool	1	4.1055e+09	2.3249e+12	11609
+ Bathroom	1	2.5720e+09	2.3265e+12	11610
- Lot	1	1.4256e+11	2.4716e+12	11637
- YearBuild	1	1.5976e+11	2.4888e+12	11641
- Quality	1	2.6818e+11	2.5972e+12	11663
- Sqft	1	1.4919e+12	3.8210e+12	11865

Step: AIC=11601.86

Price ~ Sqft + Quality + YearBuild + Lot + Garage

	Df	Sum of Sq	RSS	AIC
+ Bedroom	1	3.7251e+10	2.2548e+12	11595
+ Airconditioning	1	1.1182e+10	2.2809e+12	11601
<none>			2.2920e+12	11602
+ AdjHighway	1	7.3877e+09	2.2847e+12	11602

```

+ Pool          1 3.2365e+09 2.2888e+12 11603
+ Bathroom     1 3.0575e+09 2.2890e+12 11603
- Garage       1 3.6990e+10 2.3290e+12 11608
- Lot          1 1.2198e+11 2.4140e+12 11627
- YearBuild    1 1.2203e+11 2.4141e+12 11627
- Quality      1 2.3090e+11 2.5229e+12 11650
- Sqft        1 1.2961e+12 3.5882e+12 11834

```

Step: AIC=11595.31

Price ~ Sqft + Quality + YearBuild + Lot + Garage + Bedroom

	Df	Sum of Sq	RSS	AIC
<none>			2.2548e+12	11595
+ AdjHighway	1	7.4446e+09	2.2473e+12	11596
+ Airconditioning	1	6.9731e+09	2.2478e+12	11596
+ Pool	1	4.6766e+09	2.2501e+12	11596
+ Bathroom	1	3.5810e+07	2.2548e+12	11597
- Bedroom	1	3.7251e+10	2.2920e+12	11602
- Garage	1	3.8331e+10	2.2931e+12	11602
- YearBuild	1	1.2861e+11	2.3834e+12	11622
- Lot	1	1.2923e+11	2.3840e+12	11622
- Quality	1	2.2231e+11	2.4771e+12	11642
- Sqft	1	1.2399e+12	3.4947e+12	11822

Call:

```
lm(formula = Price ~ Sqft + Quality + YearBuild + Lot + Garage +
    Bedroom, data = realestate)
```

Coefficients:

(Intercept)	Sqft	Quality	YearBuild	Lot	Garage
-2.233e+06	1.093e+02	-5.223e+04	1.191e+03	1.415e+00	1.665e+04
Bedroom					
-1.007e+04					

In this example stepwise regression reaches the same answer as only doing forward selection.

Model adequacy checking

```

modell1 <- lm(Price ~ Sqft + Quality + YearBuild + Lot + Garage + Bedroom, data=realestate)
modell1

```

Call:

```
lm(formula = Price ~ Sqft + Quality + YearBuild + Lot + Garage +
    Bedroom, data = realestate)
```

Coefficients:

(Intercept)	Sqft	Quality	YearBuild	Lot	Garage
-2.233e+06	1.093e+02	-5.223e+04	1.191e+03	1.415e+00	1.665e+04
Bedroom					
-1.007e+04					

```
summary(model1)
```

Call:

```
lm(formula = Price ~ Sqft + Quality + YearBuild + Lot + Garage +  
    Bedroom, data = realestate)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-166119	-41432	-2654	32273	348313

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-2.233e+06	4.392e+05	-5.084	5.18e-07	***
Sqft	1.093e+02	6.496e+00	16.828	< 2e-16	***
Quality	-5.223e+04	7.330e+03	-7.126	3.51e-12	***
YearBuild	1.191e+03	2.198e+02	5.420	9.18e-08	***
Lot	1.415e+00	2.604e-01	5.433	8.57e-08	***
Garage	1.665e+04	5.626e+03	2.959	0.00323	**
Bedroom	-1.007e+04	3.454e+03	-2.917	0.00369	**

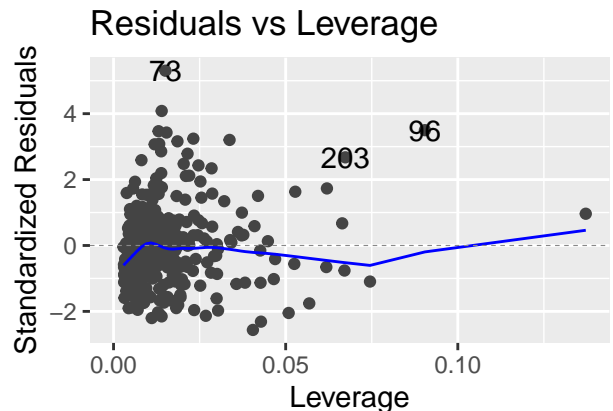
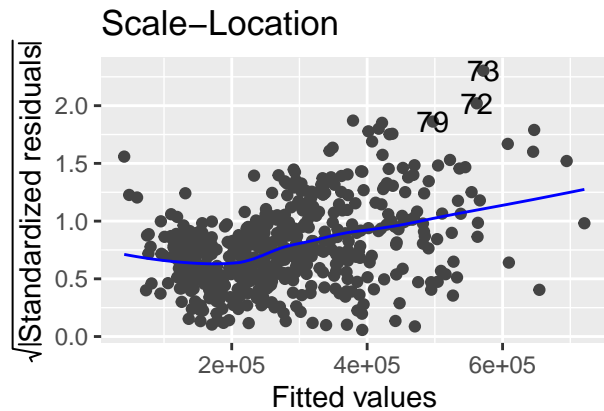
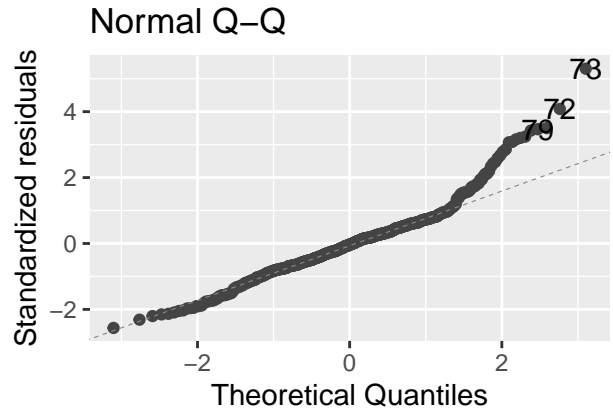
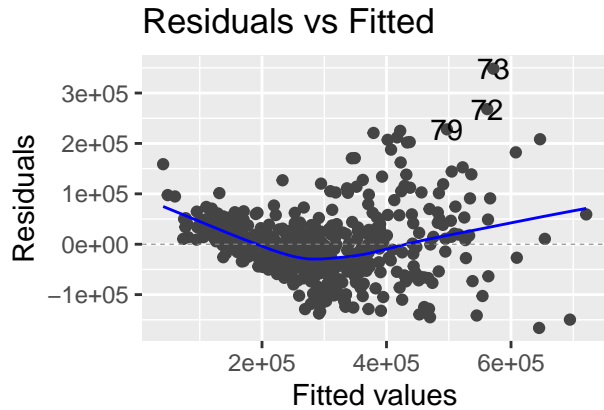
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 66170 on 515 degrees of freedom

Multiple R-squared: 0.7725, Adjusted R-squared: 0.7698

F-statistic: 291.4 on 6 and 515 DF, p-value: < 2.2e-16

```
library(ggfortify)  
autoplot(model1)
```



```
realestate$log.price <- log(realestate$Price)
model2 <- lm(log.price ~ Sqft + Quality + YearBuild + Lot + Garage + Bedroom, data=realestate)
model2
```

Call:

```
lm(formula = log.price ~ Sqft + Quality + YearBuild + Lot + Garage +
    Bedroom, data = realestate)
```

Coefficients:

```
(Intercept)      3.737e+00
Sqft            3.112e-04
Quality        -1.782e-01
YearBuild       4.138e-03
Lot             4.978e-06
Garage          5.000e-02
Bedroom         5.045e-03
```

```
summary(model2)
```

Call:

```
lm(formula = log.price ~ Sqft + Quality + YearBuild + Lot + Garage +
    Bedroom, data = realestate)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-0.58012 -0.11594 -0.00685  0.10988  0.50210
```

Coefficients:

```

      Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.737e+00  1.219e+00   3.065  0.00229 **
Sqft         3.112e-04  1.804e-05  17.253 < 2e-16 ***
Quality     -1.782e-01  2.035e-02  -8.755 < 2e-16 ***
YearBuild   4.138e-03  6.102e-04   6.781 3.27e-11 ***
Lot         4.978e-06  7.230e-07   6.885 1.69e-11 ***
Garage      5.000e-02  1.562e-02   3.201  0.00145 **
Bedroom     5.045e-03  9.588e-03   0.526  0.59902
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

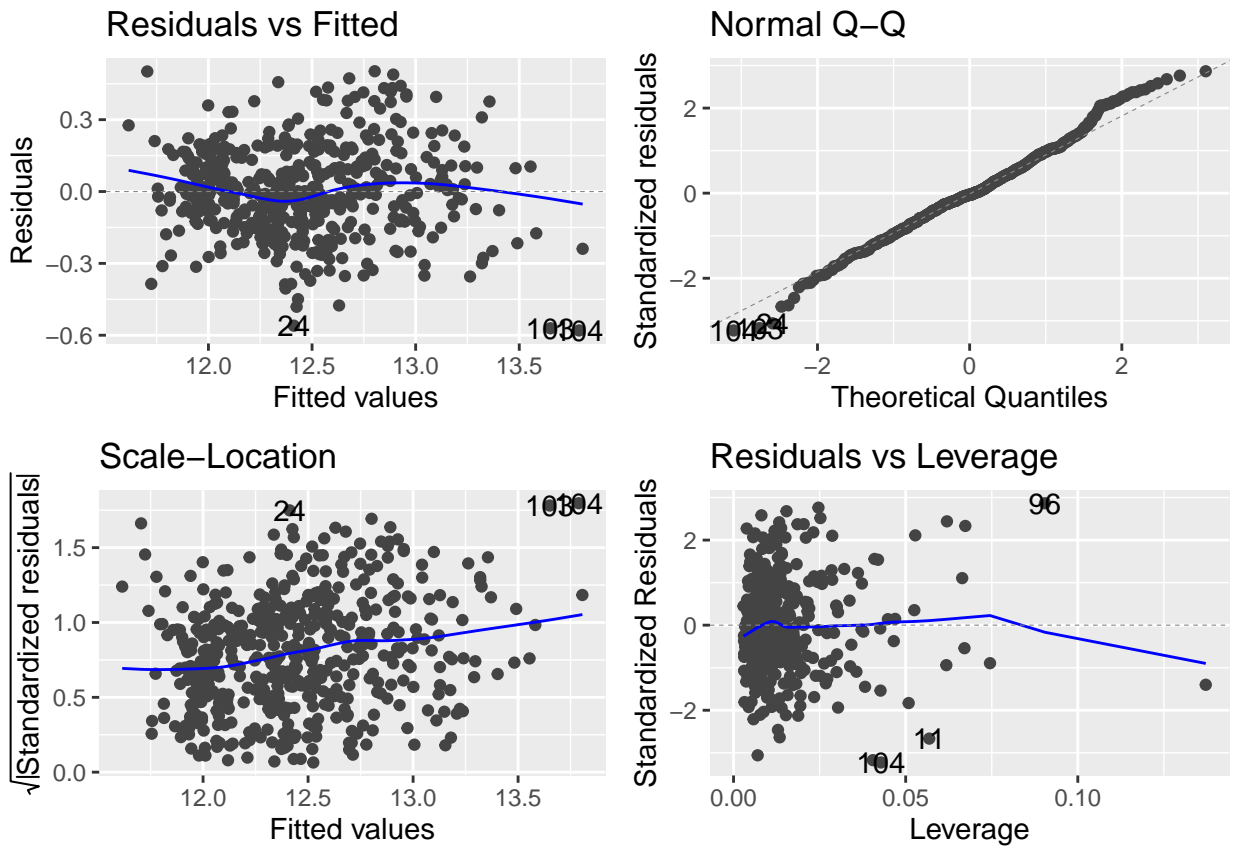
Residual standard error: 0.1837 on 515 degrees of freedom
Multiple R-squared:  0.821, Adjusted R-squared:  0.8189
F-statistic: 393.6 on 6 and 515 DF,  p-value: < 2.2e-16

```

```

library(ggfortify)
autoplot(model2)

```



Normality test

```

library(broom)
residout <- augment(model2)
residout

```

```
# A tibble: 522 x 13
```

```

log.price  Sqft Quality YearBuild  Lot Garage Bedroom .fitted  .resid
<dbl> <int> <int> <int> <int> <int> <int> <dbl> <dbl>
1      12.8  3032      2      1972 22221      2      4      12.7  0.0784
2      12.7  2058      2      1976 22912      2      4      12.4  0.304
3      12.4  1780      2      1980 21345      2      4      12.4  0.0746
4      12.2  1638      2      1963 17342      2      4      12.2  0.0130
5      12.5  2196      2      1968 21786      2      4      12.4  0.0897
6      12.4  1966      2      1972 18902      5      4      12.5 -0.0961
7      12.3  2216      2      1972 18639      2      3      12.4 -0.0933
8      11.9  1597      2      1955 22112      1      2      12.1 -0.220
9      12.2  1622      3      1975 14321      2      3      12.1  0.114
10     12.0  1976      3      1918 32358      1      3      12.0  0.00225
# ... with 512 more rows, and 4 more variables: .hat <dbl>, .sigma <dbl>,
#   .cooksd <dbl>, .std.resid <dbl>

```

```
shapiro.test(residout$.resid)
```

Shapiro-Wilk normality test

```
data: residout$.resid
W = 0.99333, p-value = 0.02065
```

Use level of significance: 0.01

Note

In variable selection it is usually assumed that the correct functional specification of the regressors is known ($1/x$, $\ln(Y)$), and that no outliers or influential observations are present. However, in practice these assumptions are rarely met. Hence, in practice we often use i) a particular variable selection strategy is employed, and then ii) the resulting model is checked for model adequacy, outliers, and influential cases and update the model accordingly.

Acknowledgement

Introduction to Linear Regression Analysis, Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining

Data: <http://www.stat.cmu.edu/~cshalizi/mreg/15/hw/08/real-estate.csv>