Model Selection

Institute for Advanced Analytics
MSA Class of 2025

Review of Modeling Up Until Now…

- **Simple Linear Regression** – one predictor variable for continuous target.
- **Multiple Linear Regression** – many predictor variables (continuous or categorical) for continuous target.
- With many explanatory variables, how do we know which ones are most informative?

Ames Housing Data

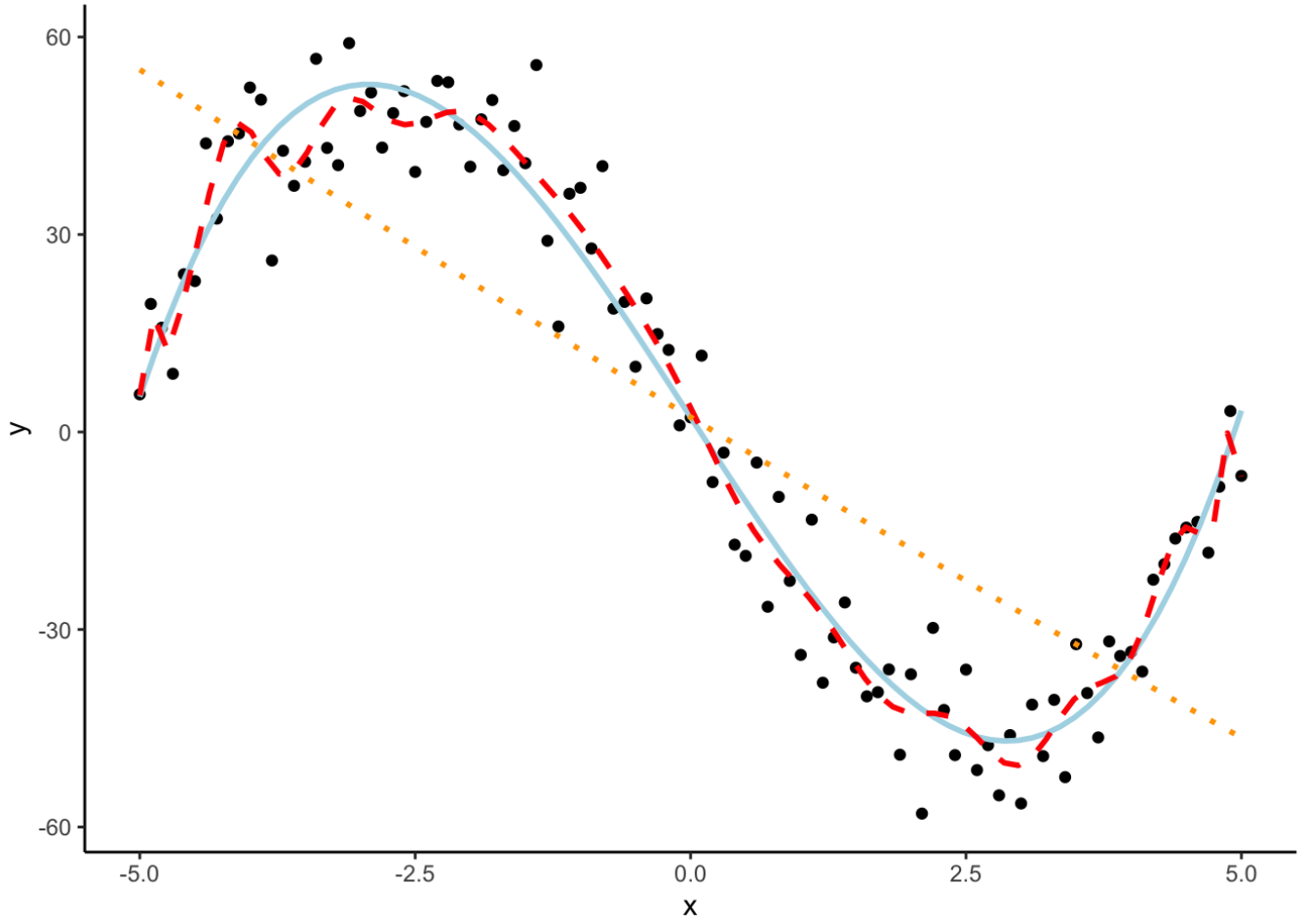
- Sale Price predicted with…
  - Greater Living Area
  - Lot Size
  - Central Air
  - Heating Quality
  - Number of Rooms
  - Number of Bathrooms
  - And many, many, more…

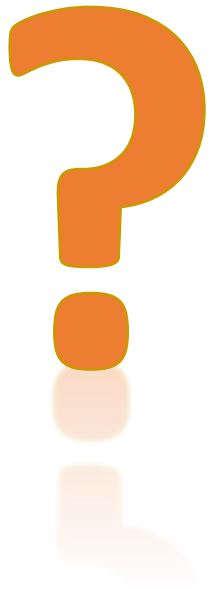
Model Building Concepts

- **Information Criteria** – commonly used to ”select” variables for the model.
- **Selection Algorithm** – automated technique to quickly evaluate variables based on some selection criteria.
  - Stepwise Selection (forward, backward, stepwise)
  - All-regression Selection (R^2, R_a^2, Mallow's C_p)

Fear of Overfitting

- Model selection should always be done with training data.
- Will hold out validation / testing data to help evaluate (honest assessment) if we have overfit our data.
- In machine learning, we will use cross-validation.





Information Criteria

AIC and BIC

- AIC and BIC approximate out-of-sample prediction error by applying a penalty for model complexity:
  - **AIC (Akaike Information Criterion)** – crude, large-sample approximation of leave-one-out cross-validation.
  - **BIC (Bayesian Information Criterion)** – favors smaller models/penalizes model complexity more.
- Lower values “better” than higher.
- No amount of lower is “better” enough.
- May not always agree, but neither is necessarily better.

AIC and BIC

- AIC (Akaike Information Criterion)

$$\mathrm{AIC} = -2\log(L) + 2k$$

$$\mathrm{AIC} = n\log\left(\frac{SSE}{n}\right) + 2k$$

- BIC (Bayesian Information Criterion)

$$\mathrm{BIC} = -2\log(L) + k\log(n)$$

$$\mathrm{BIC} = n\log\left(\frac{SSE}{n}\right) + k\log(n)$$

AIC and BIC

- AIC (Akaike Information Criterion)

$$\mathrm{AIC} = -2\log(L) + 2k$$

$$\mathrm{AIC} = n\log\left(\frac{SSE}{n}\right) + 2k$$

- BIC (Bayesian Information Criterion)

$$\mathrm{BIC} = -2\log(L) + k\log(n)$$

$$\mathrm{BIC} = n\log\left(\frac{SSE}{n}\right) + k\log(n)$$

Forward Selection

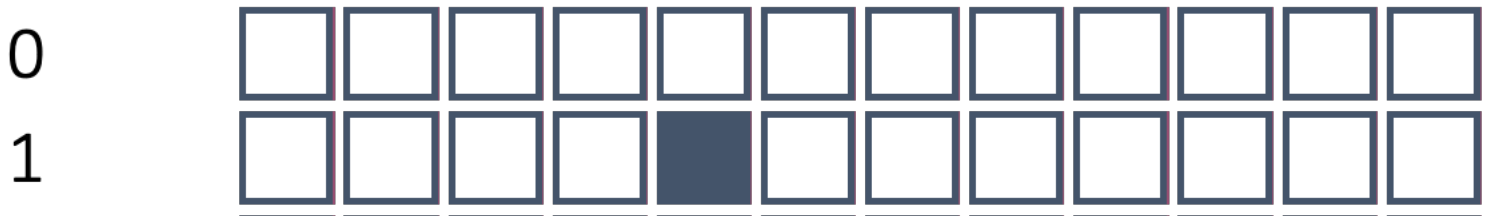
Process

- Start with a “null” model (just the intercept) and systematically build the model (one variable at a time).

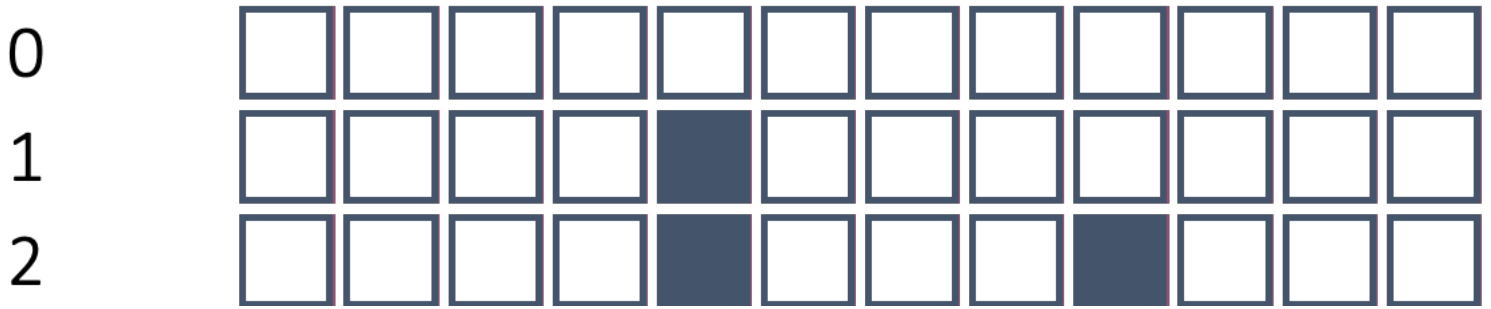
1. Start with intercept only model (this is the base model)
2. For each variable not in model, create a linear regression model with the base model plus this variable
3. See which linear regression is best (based on criterion)
4. If better than base model, continue to next step…
   otherwise STOP (this is your model)
5. Update base model to this new model, repeat whole process…

Process

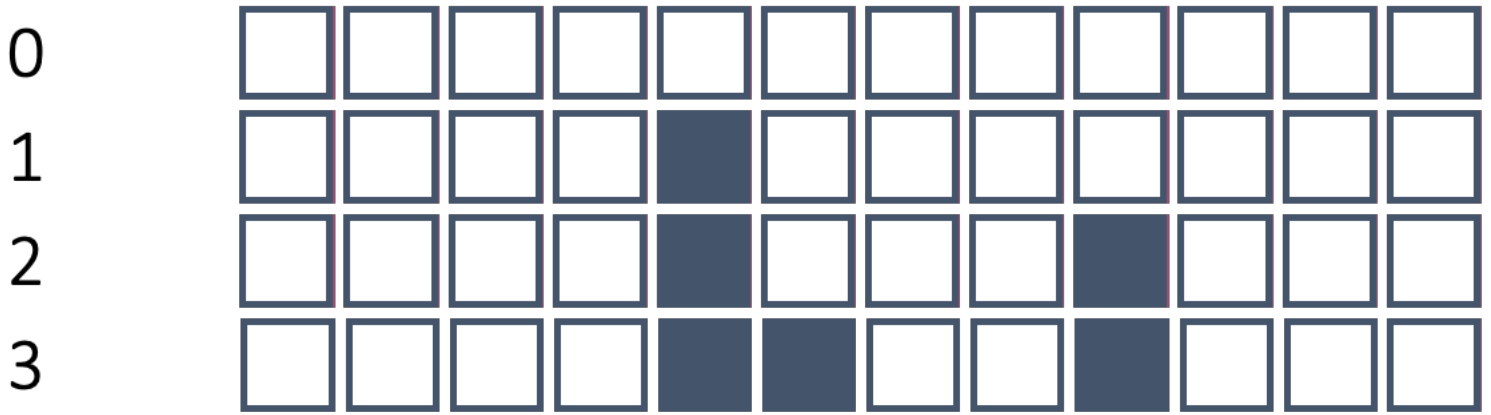
Process



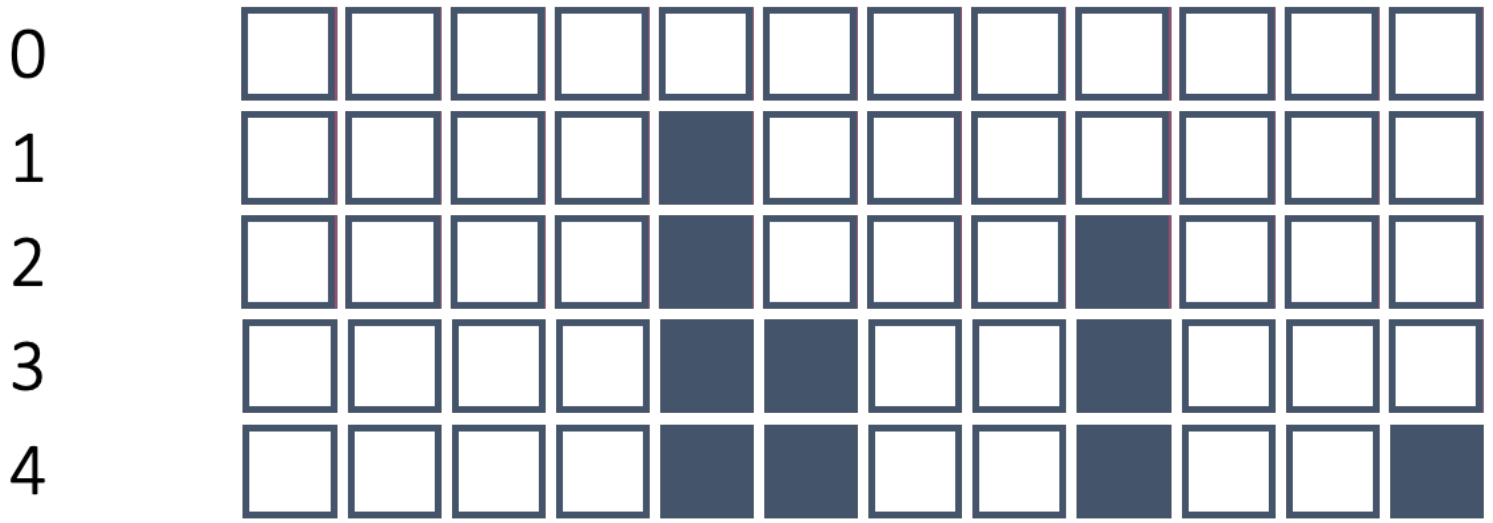
Process



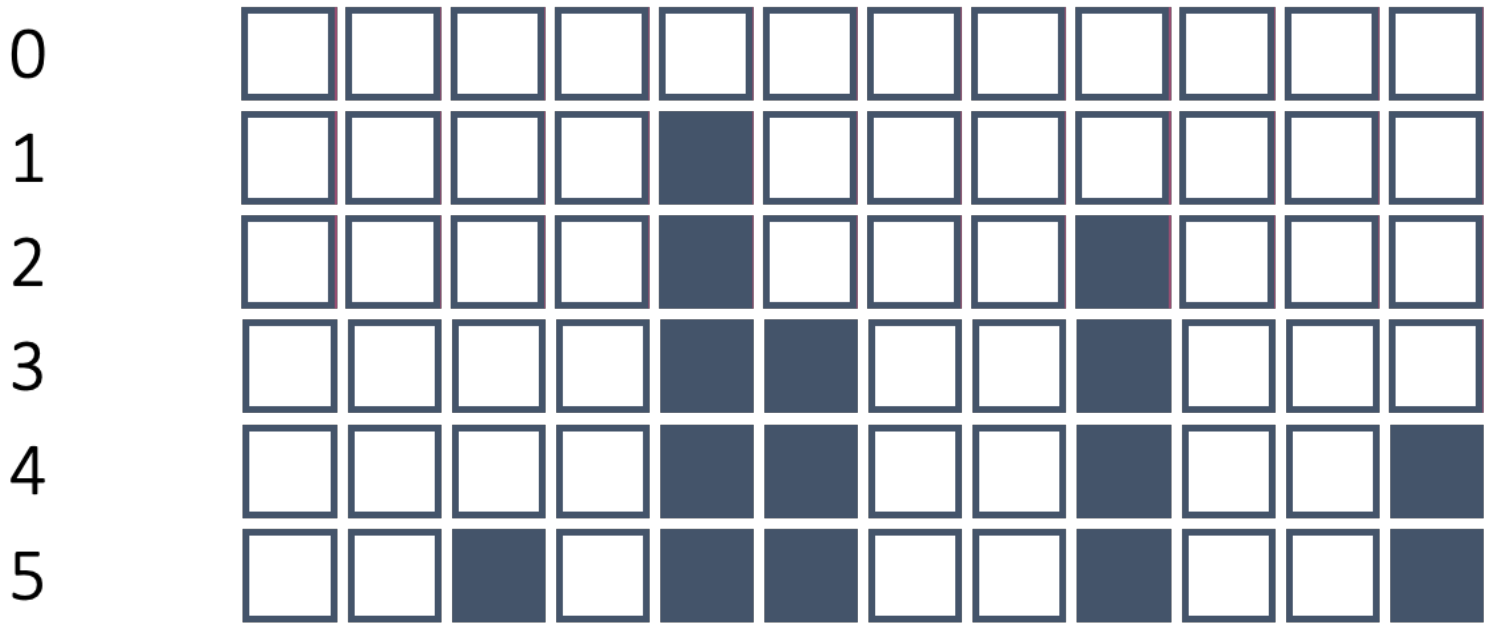
Process



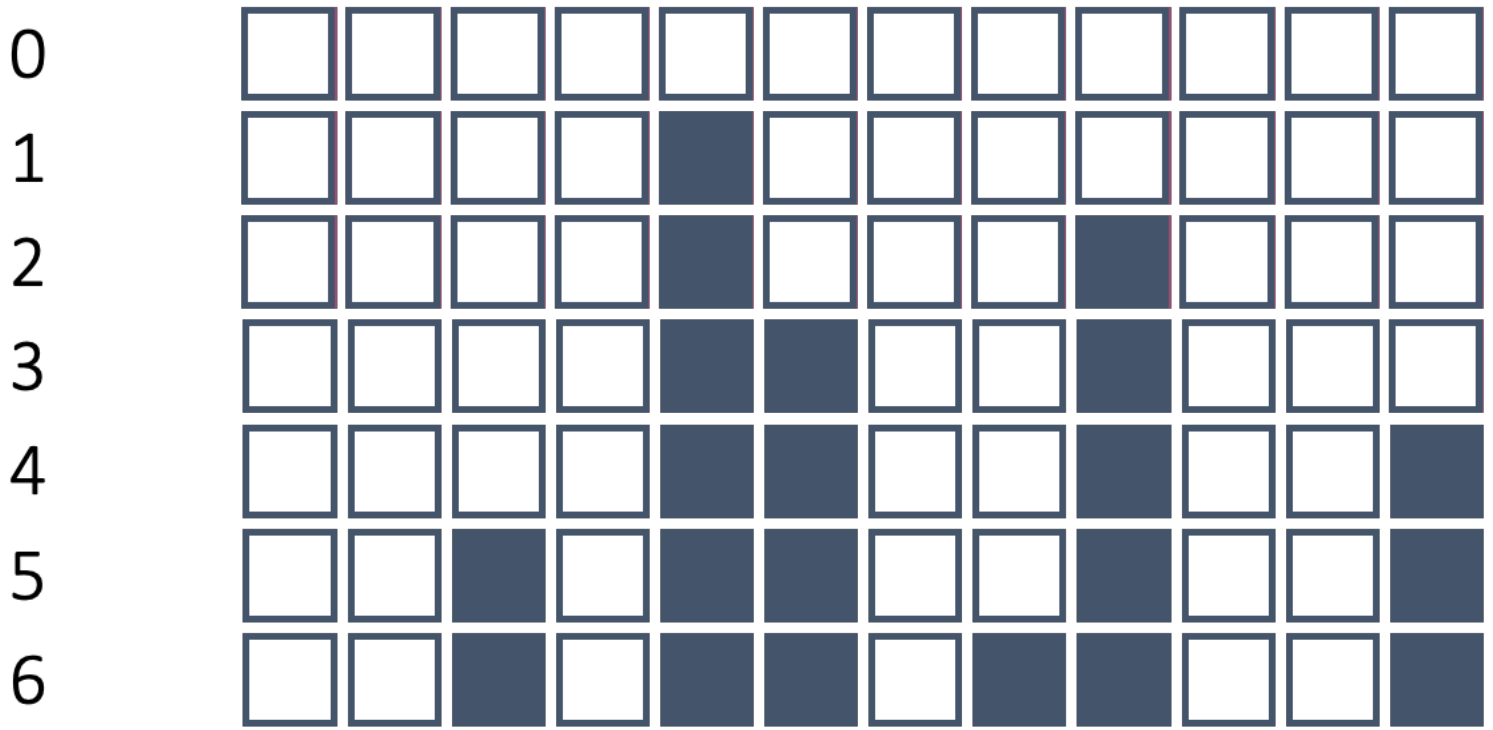
Process



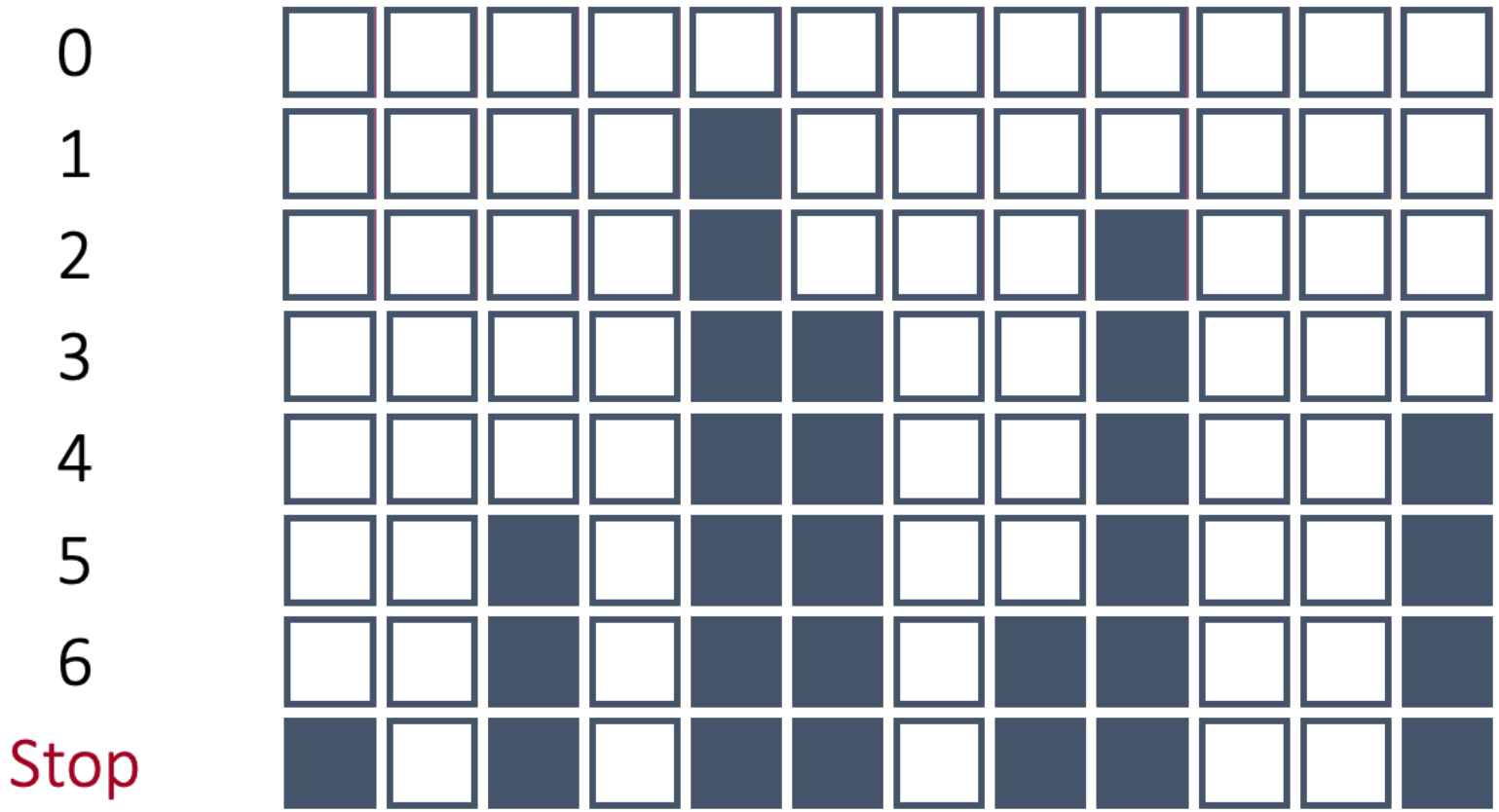
Process



Process



Process



Selecting Some Variables

```
train_sel <- train %>%
              dplyr::select(Sale_Price, Lot_Area, Street,
                            Bldg_Type, House_Style, Overall_Qual,
                            Roof_Style, Central_Air, First_Flr_SF,
                            Second_Flr_SF, Full_Bath, Half_Bath,
                            Fireplaces, Garage_Area, Gr_Liv_Area,
                            TotRms_AbvGrd) %>%
             mutate_if(is.numeric, ~replace_na(., mean(., na.rm = TRUE)))
```

Selecting Some Variables

```
train_sel <- train %>%
                dplyr::select(Sale_Price, Lot_Area, Street,
                              Bldg_Type, House_Style, Overall_Qual,
                              Roof_Style, Central_Air, First_Flr_SF,
                              Second_Flr_SF, Full_Bath, Half_Bath,
                              Fireplaces, Garage_Area, Gr_Liv_Area,
                              TotRms_AbvGrd) %>%
                mutate_if(is.numeric, ~replace_na(., mean(., na.rm = TRUE)))
```

Replacing all missing with mean of the column for now.
Will discuss different imputations at later date!

Forward Selection

```
full.model <- lm(Sale_Price ~ ., data = train_sel)
empty.model <- lm(Sale_Price ~ 1, data = train_sel)

for.model <- step(empty.model,
                  scope = list(lower = empty.model,
                               upper = full.model),
                  direction = “forward”, k = 2)
```

Forward Selection

```
full.model <- lm(Sale_Price ~ ., data = train_sel)
empty.model <- lm(Sale_Price ~ 1, data = train_sel)

for.model <- step(empty.model,
                  scope = list(lower = empty.model,
                               upper = full.model),
                  direction = “forward”, k = 2)
```

AIC Selection

Forward Selection

```
Start:  AIC=46323.64
Sale_Price ~ 1
```

- Step 1
- AIC from base model at top of output.
- Shows the addition (“+”) of each variable and the “new” AIC from that model.
- Best variable at top.

```
                Df  Sum of Sq        RSS   AIC
+ Overall_Qual   9 9.3437e+12 3.8531e+12 43817
+ Gr_Liv_Area    1 6.4389e+12 6.7578e+12 44953
+ Garage_Area    1 5.3561e+12 7.8407e+12 45258
+ First_Flr_SF   1 4.8867e+12 8.3100e+12 45377
+ Full_Bath      1 3.7827e+12 9.4141e+12 45633
+ TotRms_AbvGrd  1 3.2304e+12 9.9663e+12 45750
+ Fireplaces     1 2.9715e+12 1.0225e+13 45802
+ Half_Bath      1 1.1209e+12 1.2076e+13 46144
+ Roof_Style     5 1.0724e+12 1.2124e+13 46160
+ Central_Air    1 9.6147e+11 1.2235e+13 46170
+ House_Style    7 1.0245e+12 1.2172e+13 46172
+ Second_Flr_SF  1 9.4611e+11 1.2251e+13 46173
+ Lot_Area       1 9.0332e+11 1.2293e+13 46180
+ Bldg_Type      4 4.6434e+11 1.2732e+13 46258
+ Street         1 3.1752e+10 1.3165e+13 46321
<none>                        1.3197e+13 46324
```

Forward Selection

```
Step:  AIC=43816.66
Sale_Price ~ Overall_Qual
```

- Step 2
- AIC from new base model at top of output.
- Shows the addition (“+”) of each variable and the “new” AIC from that model.
- Best variable at top.
- Notice the <none>…

```
                  Df  Sum of Sq        RSS   AIC
+ Gr_Liv_Area      1 9.8905e+11 2.8640e+12 43210
+ First_Flr_SF     1 5.2665e+11 3.3264e+12 43517
+ Garage_Area      1 4.6644e+11 3.3866e+12 43554
+ TotRms_AbvGrd    1 4.6123e+11 3.3918e+12 43557
+ Full_Bath        1 4.1206e+11 3.4410e+12 43587
+ Fireplaces       1 4.0551e+11 3.4476e+12 43591
+ Lot_Area         1 3.8148e+11 3.4716e+12 43605
+ Bldg_Type        4 2.3715e+11 3.6159e+12 43694
+ Second_Flr_SF    1 1.7555e+11 3.6775e+12 43723
+ Half_Bath        1 1.3948e+11 3.7136e+12 43743
+ Central_Air      1 9.1322e+10 3.7617e+12 43769
+ House_Style      7 6.1815e+10 3.7912e+12 43797
+ Roof_Style       5 5.1448e+10 3.8016e+12 43799
<none>                         3.8531e+12 43817
+ Street          1 1.9573e+06 3.8531e+12 43819
```

Forward Selection

```
Step:  AIC=43210.24
Sale_Price ~ Overall_Qual + Gr_Liv_Area
```

- Step 3
- AIC from new base model at top of output.
- Shows the addition (“+”) of each variable and the “new” AIC from that model.
- Best variable at top.
- Watch the <none>…

```
                Df  Sum of Sq        RSS   AIC
+ House_Style    7 2.5351e+11 2.6105e+12 43034
+ Garage_Area    1 2.1638e+11 2.6476e+12 43051
+ Lot_Area       1 1.3097e+11 2.7330e+12 43116
+ First_Flr_SF   1 1.2210e+11 2.7419e+12 43123
+ Fireplaces     1 1.1069e+11 2.7533e+12 43131
+ Central_Air    1 1.1050e+11 2.7535e+12 43132
+ Second_Flr_SF  1 1.0207e+11 2.7619e+12 43138
+ Bldg_Type      4 1.0299e+11 2.7610e+12 43143
+ Roof_Style     5 6.0726e+10 2.8033e+12 43176
+ Full_Bath      1 3.2970e+10 2.8310e+12 43188
+ TotRms_AbvGrd  1 2.4688e+10 2.8393e+12 43194
<none>                       2.8640e+12 43210
+ Half_Bath      1 4.0261e+07 2.8640e+12 43212
+ Street         1 2.2632e+07 2.8640e+12 43212
```

Forward Selection

- Step 15
- Exit the algorithm since <none> is the highest step.
- Poor Street variable…

```
Step:  AIC=42676.1
Sale_Price ~ Overall_Qual + Gr_Liv_Area +
House_Style + Garage_Area + Bldg_Type +
Fireplaces + Full_Bath + Half_Bath + Lot_Area +
Roof_Style + Central_Air + Second_Flr_SF +
TotRms_AbvGrd +
First_Flr_SF

           Df Sum of Sq        RSS   AIC
<none>                  2.1542e+12 42676
+ Street  1 1.028e+09 2.1532e+12 42677
```

Other Criteria – BIC

```
full.model <- lm(Sale_Price ~ ., data = train_sel)
empty.model <- lm(Sale_Price ~ 1, data = train_sel)

for.model <- step(empty.model,
                  scope = list(lower = empty.model,
                               upper = full.model),
                  direction = “forward”, k = log(nrow(train_sel))
```

BIC Selection

Other Criteria – P-value Selection (α = 0.05)

```
full.model <- lm(Sale_Price ~ ., data = train_sel)
empty.model <- lm(Sale_Price ~ 1, data = train_sel)

for.model <- step(empty.model,
                  scope = list(lower = empty.model,
                               upper = full.model),
                  direction = "forward", k = qchisq(0.05, 1, lower.tail = FALSE))
```

P-value selection with alpha = 0.05

Can easily change alpha to any number…

Backward Selection

Process

- Systematically removes variables “not informative” in the model (one variable at a time).

1. Start with full model with all variables (this is the base model)
2. Create models such that each model has exactly one predictor variable removed from it and calculate the criterion for each model
3. See which linear regression is best (based on criterion)
4. If better than base model, continue to next step…
   otherwise STOP (this is your model)
5. Update base model to this new model, repeat whole process…

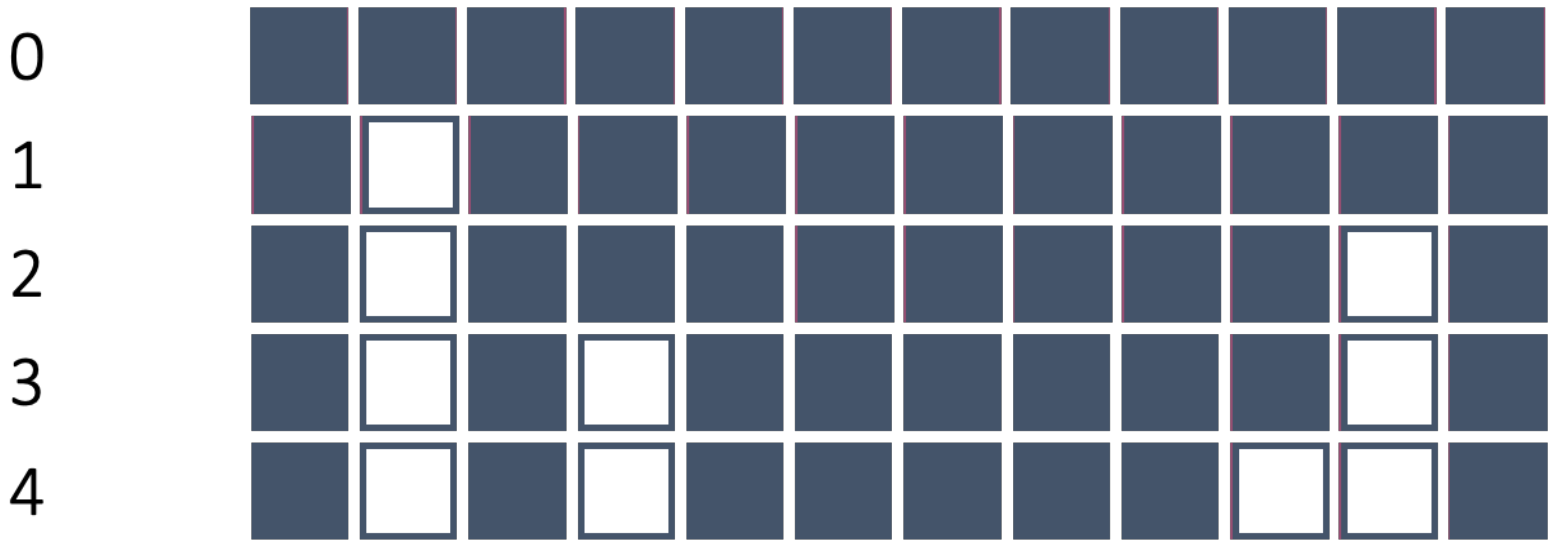
Process

Process

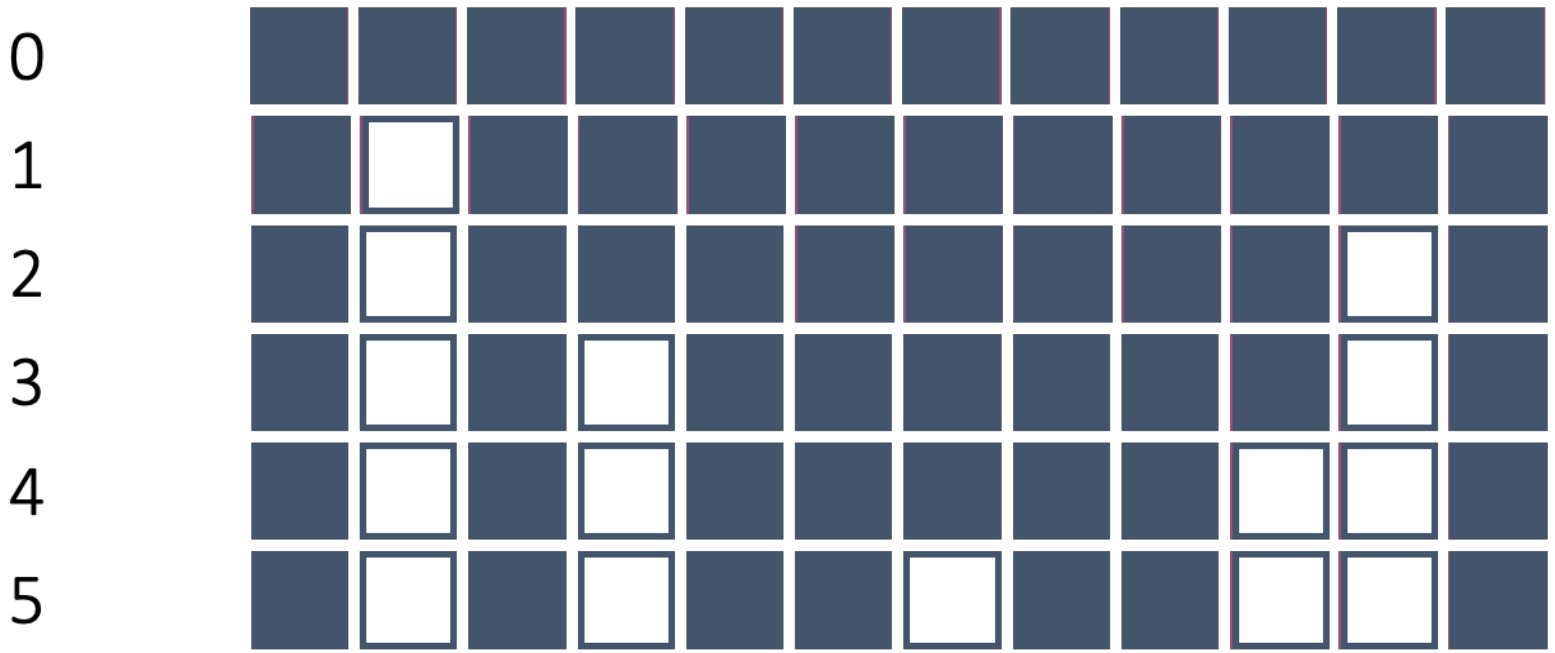
Process

Process

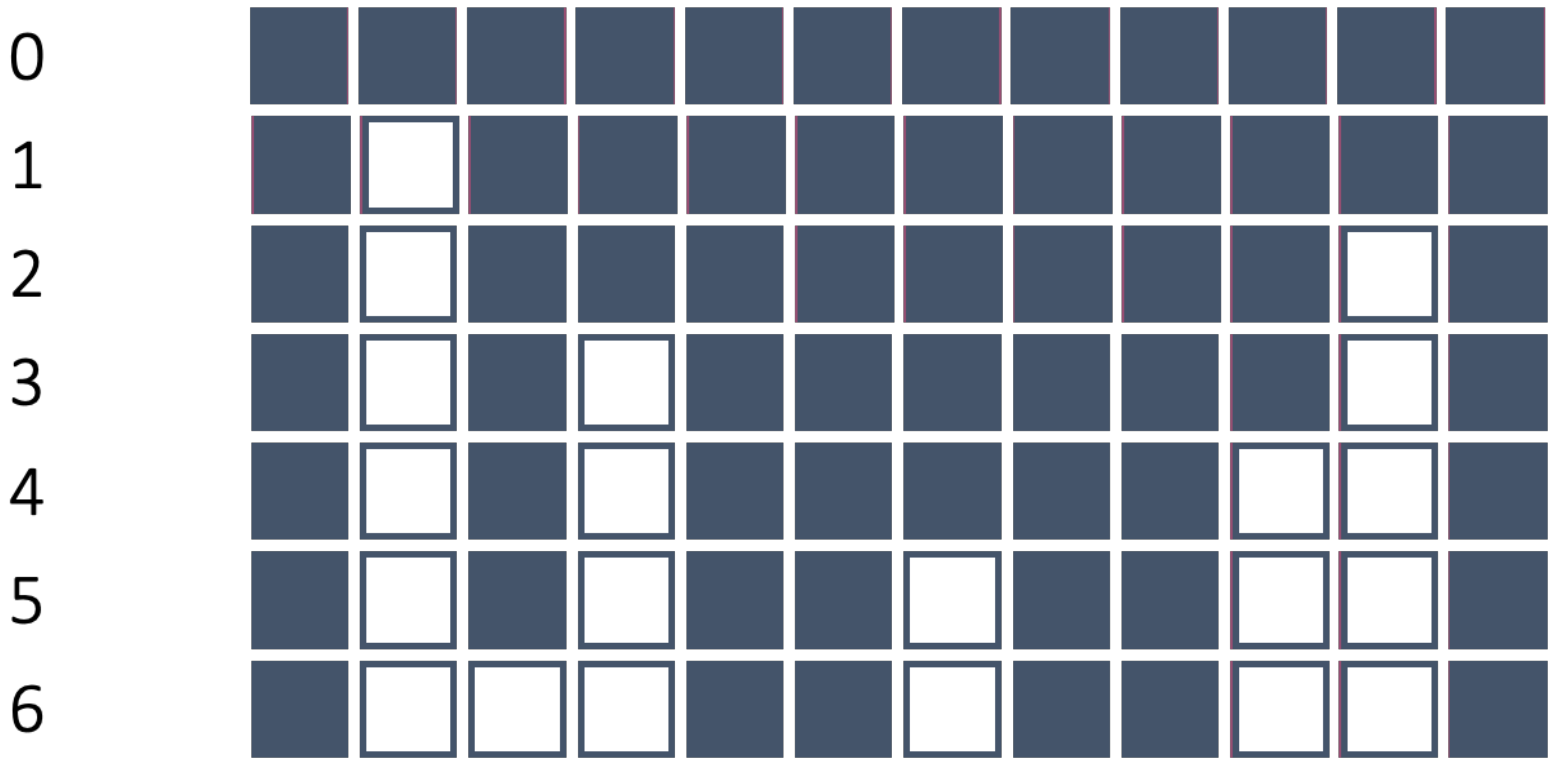
Process



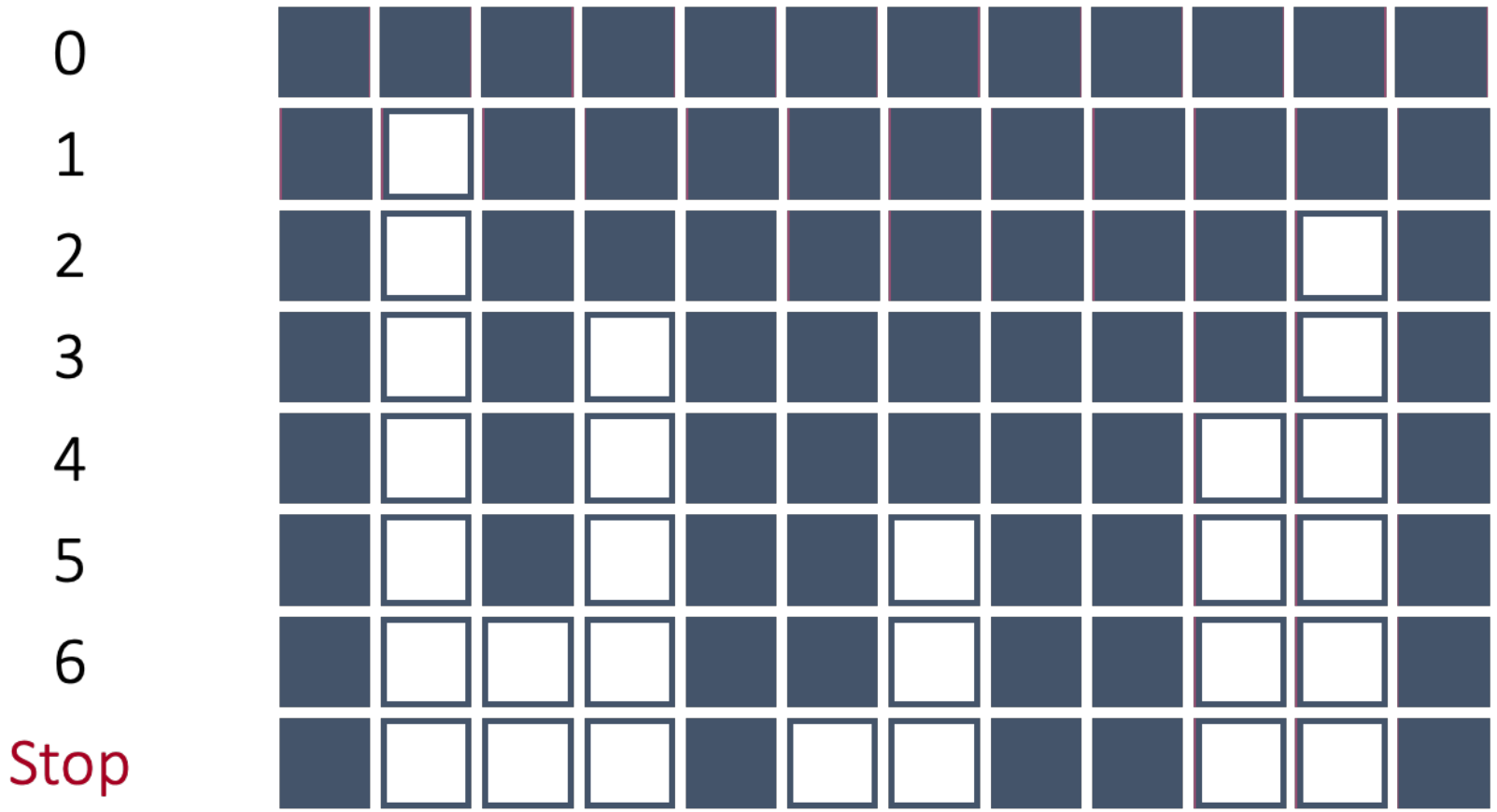
Process



Process



Process



Backward Selection

```
full.model <- lm(Sale_Price ~ ., data = train_sel)
empty.model <- lm(Sale_Price ~ 1, data = train_sel)

back.model <- step(full.model,
                   scope = list(lower = empty.model,
                                upper = full.model),
                   direction = “backward”, k = 2)
```

Backward Selection

```
full.model <- lm(Sale_Price ~ ., data = train_sel)
empty.model <- lm(Sale_Price ~ 1, data = train_sel)

back.model <- step(full.model,
                   scope = list(lower = empty.model,
                                upper = full.model),
                   direction = “backward”, k = 2)
```

AIC Selection

Backward Selection

- Step 1
- AIC from base model at top of output.
- Shows the removal (“-”) of each variable and the “new” AIC from that model.
- Worst variable (most likely to remove) at top.

```
Start:  AIC=42677.12
Sale_Price ~ Lot_Area + Street + Bldg_Type +
House_Style + Overall_Qual + Roof_Style + Central_Air
+ First_Flr_SF + Second_Flr_SF + Full_Bath +
Half_Bath + Fireplaces + Garage_Area + Gr_Liv_Area +
TotRms_AbvGrd

                Df  Sum of Sq        RSS   AIC
- Gr_Liv_Area    1 4.9138e+08 2.1537e+12 42676
- Street         1 1.0280e+09 2.1542e+12 42676
<none>                       2.1532e+12 42677
- First_Flr_SF   1 3.1548e+09 2.1563e+12 42678
- TotRms_AbvGrd  1 3.4112e+09 2.1566e+12 42678
- Second_Flr_SF  1 6.4939e+09 2.1597e+12 42681
- Central_Air    1 1.6533e+10 2.1697e+12 42691
- Roof_Style     5 2.8786e+10 2.1820e+12 42694
- Half_Bath      1 3.5009e+10 2.1882e+12 42708
- Lot_Area       1 3.5997e+10 2.1892e+12 42709
- Fireplaces     1 3.6853e+10 2.1900e+12 42710
- House_Style    7 7.0980e+10 2.2241e+12 42730
- Garage_Area    1 6.4143e+10 2.2173e+12 42735
- Bldg_Type      4 7.1274e+10 2.2244e+12 42736
- Full_Bath      1 6.8198e+10 2.2214e+12 42739
- Overall_Qual   9 1.7183e+12 3.8715e+12 43862
```

Backward Selection

- Step 2
- AIC from base model at top of output.
- Shows the removal (“-”) of each variable and the “new” AIC from that model.
- Worst variable (most likely to remove) at top.
- Watch the <none>…

```
Step:  AIC=42675.59
Sale_Price ~ Lot_Area + Street + Bldg_Type +
House_Style + Overall_Qual + Roof_Style + Central_Air
+ First_Flr_SF + Second_Flr_SF + Full_Bath +
Half_Bath + Fireplaces + Garage_Area + TotRms_AbvGrd

                Df  Sum of Sq        RSS   AIC
- Street         1 1.0581e+09 2.1547e+12 42675
<none>                        2.1537e+12 42676
- TotRms_AbvGrd  1 3.1247e+09 2.1568e+12 42677
- Central_Air    1 1.6456e+10 2.1701e+12 42689
- Roof_Style     5 2.8773e+10 2.1824e+12 42693
- Half_Bath      1 3.5031e+10 2.1887e+12 42707
- Lot_Area       1 3.6074e+10 2.1897e+12 42708
- Fireplaces     1 3.6944e+10 2.1906e+12 42708
- House_Style    7 7.2205e+10 2.2259e+12 42729
- Garage_Area    1 6.4018e+10 2.2177e+12 42734
- Bldg_Type      4 7.1756e+10 2.2254e+12 42735
- Full_Bath      1 6.9016e+10 2.2227e+12 42738
- Second_Flr_SF  1 1.2417e+11 2.2778e+12 42789
- First_Flr_SF   1 1.4119e+11 2.2949e+12 42804
- Overall_Qual   9 1.7192e+12 3.8728e+12 43861
```

Backward Selection

- Step 3
- Exit the algorithm since <none> is the highest step.

```
Step:  AIC=42674.6
Sale_Price ~ Lot_Area + Bldg_Type + House_Style +
Overall_Qual + Roof_Style + Central_Air +
First_Flr_SF + Second_Flr_SF + Full_Bath + Half_Bath
+ Fireplaces + Garage_Area + TotRms_AbvGrd

                 Df  Sum of Sq        RSS   AIC
<none>                          2.1547e+12 42675
- TotRms_AbvGrd   1 2.9784e+09 2.1577e+12 42675
- Central_Air     1 1.7247e+10 2.1720e+12 42689
- Roof_Style      5 2.8560e+10 2.1833e+12 42692
- Half_Bath       1 3.4751e+10 2.1895e+12 42705
- Lot_Area        1 3.5041e+10 2.1898e+12 42706
- Fireplaces      1 3.6680e+10 2.1914e+12 42707
- House_Style     7 7.3149e+10 2.2279e+12 42729
- Garage_Area     1 6.3520e+10 2.2182e+12 42732
- Bldg_Type       4 7.3044e+10 2.2278e+12 42735
- Full_Bath       1 6.8973e+10 2.2237e+12 42737
- Second_Flr_SF   1 1.2513e+11 2.2798e+12 42788
- First_Flr_SF    1 1.4221e+11 2.2969e+12 42804
- Overall_Qual    9 1.7202e+12 3.8749e+12 43860
```

Other Criteria – BIC

```
full.model <- lm(Sale_Price ~ ., data = train_sel)
empty.model <- lm(Sale_Price ~ 1, data = train_sel)

back.model <- step(full.model,
                   scope = list(lower = empty.model,
                                upper = full.model),
                   direction = “backward”, k = log(nrow(train_sel)))
```

BIC Selection

Other Criteria – P-value Selection (α = 0.05)

```
full.model <- lm(Sale_Price ~ ., data = train_sel)
empty.model <- lm(Sale_Price ~ 1, data = train_sel)

back.model <- step(full.model,
                   scope = list(lower = empty.model,
                                upper = full.model),
                   direction = "backward", k = qchisq(0.05, 1, lower.tail = FALSE))
```

P-value selection with alpha = 0.05

Can easily change alpha to any number…

Stepwise Selection

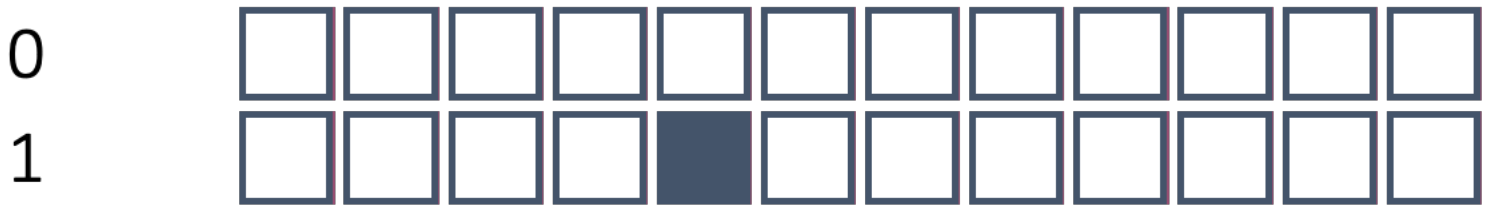
Process

- Start with a “null” model (just the intercept) and build the model (one variable at a time), but can also **delete** variables.

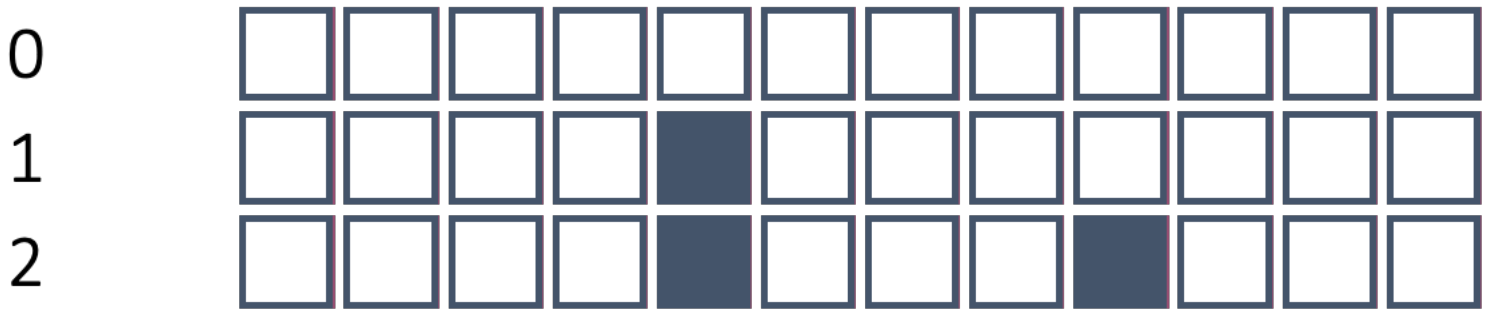
1. Start with intercept only model (this is the base model)
2. For each variable not in model, create a linear regression model with the base model plus this variable
3. For each variable in the model, create models with the base model taking away one variable at a time
4. See which linear regression is best (based on criterion)
5. If better than base model, continue to next step…
   otherwise STOP (this is your model)
6. Update base model to this new model, repeat whole process…

Process

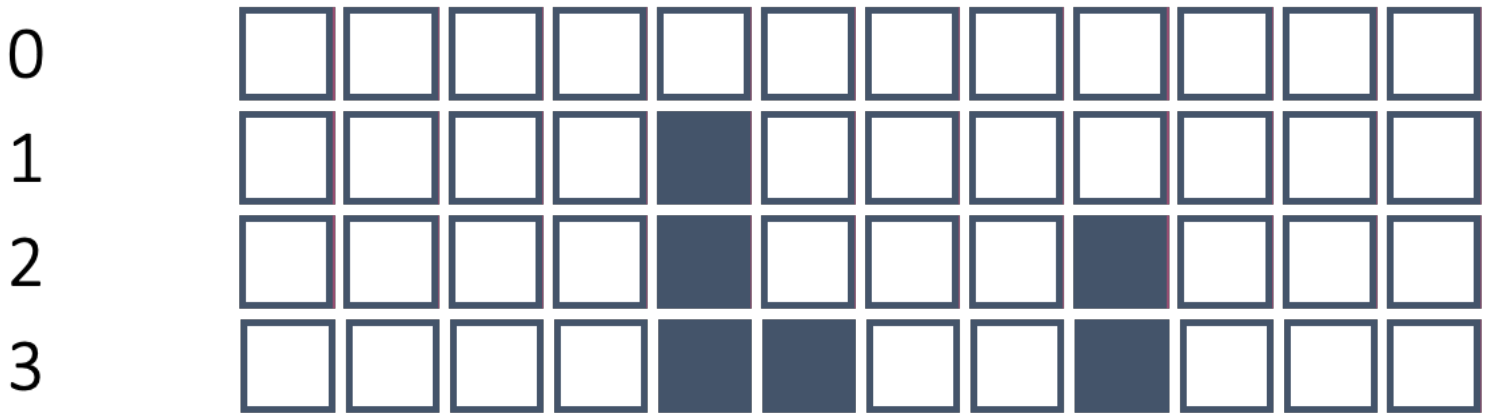
Process



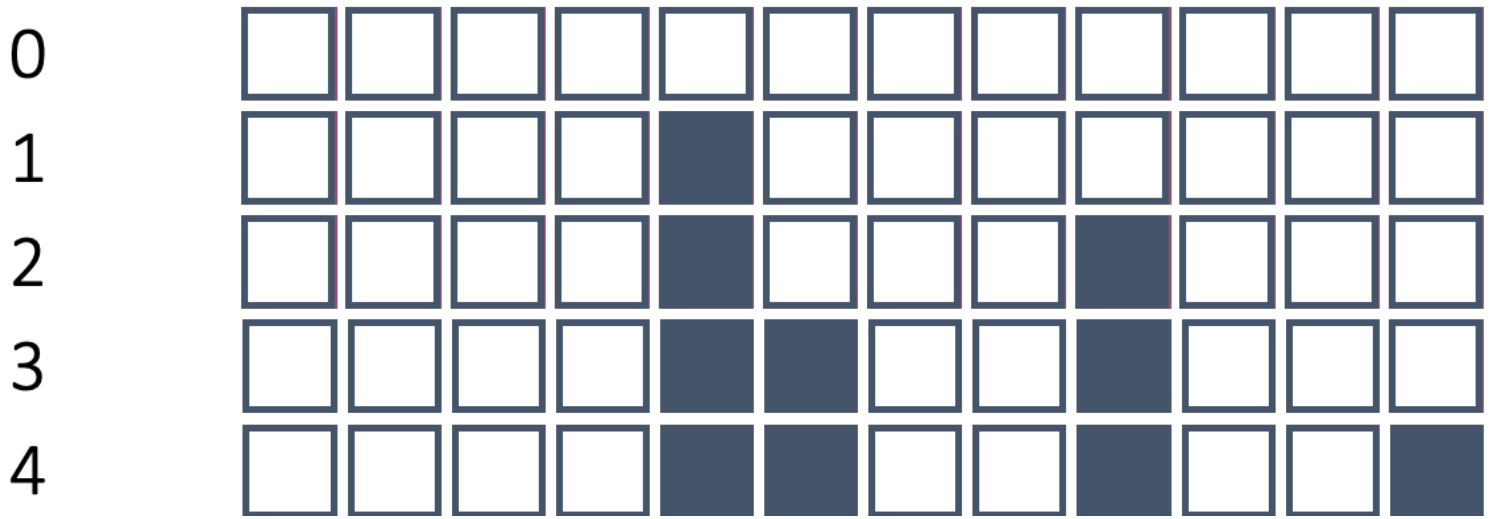
Process



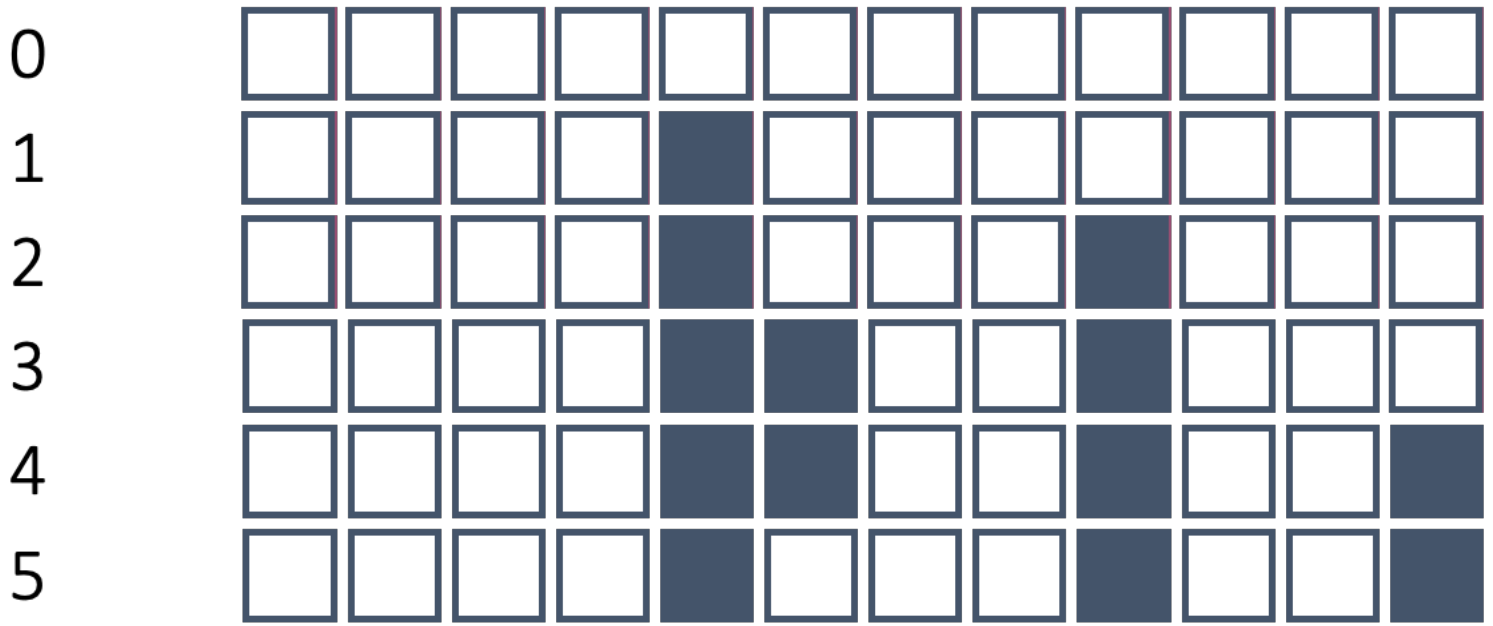
Process



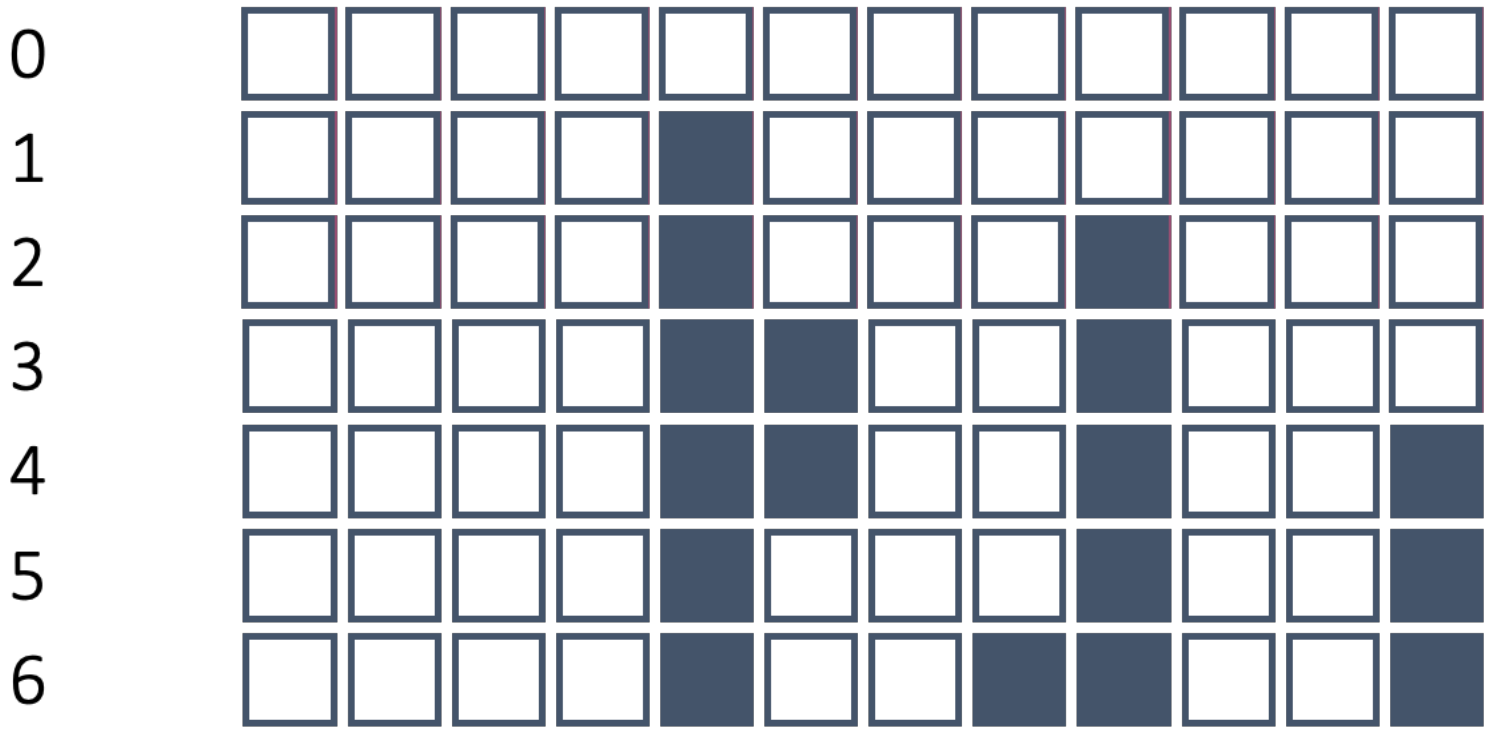
Process



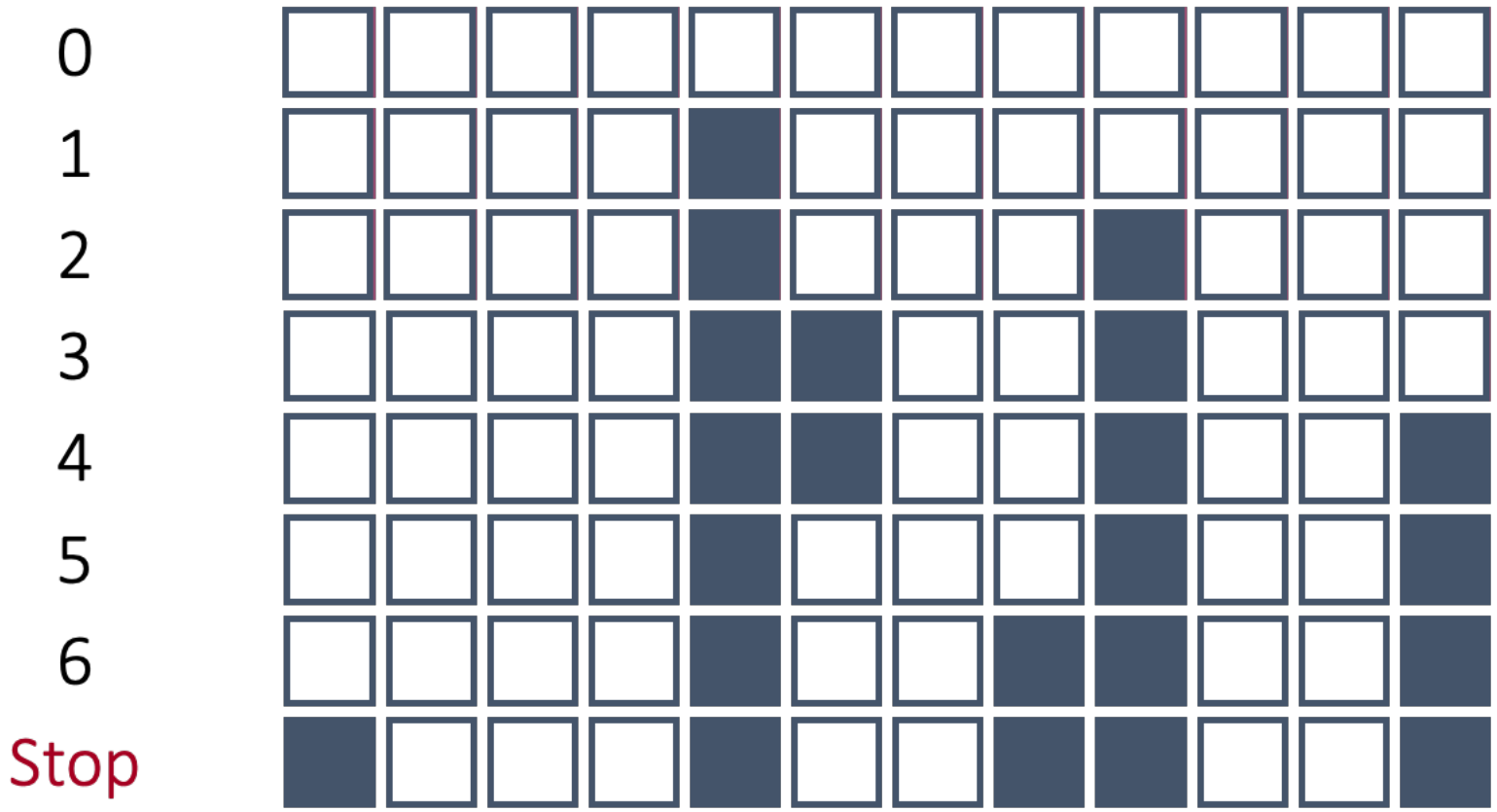
Process



Process



Process



Stepwise Selection

```
full.model <- lm(Sale_Price ~ ., data = train_sel)
empty.model <- lm(Sale_Price ~ 1, data = train_sel)

step.model <- step(empty.model,
                   scope = list(lower = empty.model,
                                upper = full.model),
                   direction = “both”, k = 2)
```

Stepwise Selection

```
full.model <- lm(Sale_Price ~ ., data = train_sel)
empty.model <- lm(Sale_Price ~ 1, data = train_sel)

step.model <- step(empty.model,
                   scope = list(lower = empty.model,
                                upper = full.model),
                   direction = “both”, k = 2)
```

AIC Selection

Stepwise Selection

```
Start:  AIC=46323.64
Sale_Price ~ 1
```

- Step 1
- AIC from base model at top of output.
- Shows the addition (“+”) of each variable and the “new” AIC from that model.
- Best variable at top.

```
                  Df  Sum of Sq        RSS   AIC
+ Overall_Qual     9 9.3437e+12 3.8531e+12 43817
+ Gr_Liv_Area      1 6.4389e+12 6.7578e+12 44953
+ Garage_Area      1 5.3561e+12 7.8407e+12 45258
+ First_Flr_SF     1 4.8867e+12 8.3100e+12 45377
+ Full_Bath        1 3.7827e+12 9.4141e+12 45633
+ TotRms_AbvGrd   1 3.2304e+12 9.9663e+12 45750
+ Fireplaces       1 2.9715e+12 1.0225e+13 45802
+ Half_Bath        1 1.1209e+12 1.2076e+13 46144
+ Roof_Style       5 1.0724e+12 1.2124e+13 46160
+ Central_Air      1 9.6147e+11 1.2235e+13 46170
+ House_Style      7 1.0245e+12 1.2172e+13 46172
+ Second_Flr_SF   1 9.4611e+11 1.2251e+13 46173
+ Lot_Area         1 9.0332e+11 1.2293e+13 46180
+ Bldg_Type        4 4.6434e+11 1.2732e+13 46258
+ Street           1 3.1752e+10 1.3165e+13 46321
<none>                          1.3197e+13 46324
```

Stepwise Selection

```
Step:  AIC=43816.66
Sale_Price ~ Overall_Qual
```

- Step 2
- AIC from base model at top of output.
- Shows the addition (“+”) or removal (“-”) of each variable and the “new” AIC from that model.
- Best choice at top.
- Watch the <none>…

```
                  Df  Sum of Sq        RSS   AIC
+ Gr_Liv_Area      1 9.8905e+11 2.8640e+12 43210
+ First_Flr_SF     1 5.2665e+11 3.3264e+12 43517
+ Garage_Area      1 4.6644e+11 3.3866e+12 43554
+ TotRms_AbvGrd   1 4.6123e+11 3.3918e+12 43557
+ Full_Bath       1 4.1206e+11 3.4410e+12 43587
+ Fireplaces      1 4.0551e+11 3.4476e+12 43591
+ Lot_Area        1 3.8148e+11 3.4716e+12 43605
+ Bldg_Type       4 2.3715e+11 3.6159e+12 43694
+ Second_Flr_SF   1 1.7555e+11 3.6775e+12 43723
+ Half_Bath       1 1.3948e+11 3.7136e+12 43743
+ Central_Air     1 9.1322e+10 3.7617e+12 43769
+ House_Style     7 6.1815e+10 3.7912e+12 43797
+ Roof_Style      5 5.1448e+10 3.8016e+12 43799
<none>                        3.8531e+12 43817
+ Street          1 1.9573e+06 3.8531e+12 43819
- Overall_Qual    9 9.3437e+12 1.3197e+13 46324
```

Stepwise Selection

```
Step:  AIC=43210.24
Sale_Price ~ Overall_Qual + Gr_Liv_Area
```

- Step 3
- AIC from base model at top of output.
- Shows the addition (“+”) or removal (“-”) of each variable and the “new” AIC from that model.
- Best choice at top.
- Watch the <none>…

	Df	Sum of Sq	RSS	AIC
+ House_Style	7	2.5351e+11	2.6105e+12	43034
+ Garage_Area	1	2.1638e+11	2.6476e+12	43051
+ Lot_Area	1	1.3097e+11	2.7330e+12	43116
+ First_Flr_SF	1	1.2210e+11	2.7419e+12	43123
+ Fireplaces	1	1.1069e+11	2.7533e+12	43131
+ Central_Air	1	1.1050e+11	2.7535e+12	43132
+ Second_Flr_SF	1	1.0207e+11	2.7619e+12	43138
+ Bldg_Type	4	1.0299e+11	2.7610e+12	43143
+ Roof_Style	5	6.0726e+10	2.8033e+12	43176
+ Full_Bath	1	3.2970e+10	2.8310e+12	43188
+ TotRms_AbvGrd	1	2.4688e+10	2.8393e+12	43194
<none>			2.8640e+12	43210
+ Half_Bath	1	4.0261e+07	2.8640e+12	43212
+ Street	1	2.2632e+07	2.8640e+12	43212
- Gr_Liv_Area	1	9.8905e+11	3.8531e+12	43817
- Overall_Qual	9	3.8938e+12	6.7578e+12	44953

Stepwise Selection

```
Step:  AIC=42674.6
Sale_Price ~ Overall_Qual + House_Style +
Garage_Area + Bldg_Type + Fireplaces + Full_Bath
+ Half_Bath + Lot_Area + Roof_Style +
Central_Air + Second_Flr_SF + TotRms_AbvGrd +
First_Flr_SF
```

- Step 14
- Exit the algorithm since <none> is the highest step.

```
                  Df  Sum of Sq        RSS   AIC
<none>                          2.1547e+12 42675
- TotRms_AbvGrd    1 2.9784e+09 2.1577e+12 42675
+ Street           1 1.0581e+09 2.1537e+12 42676
+ Gr_Liv_Area      1 5.2156e+08 2.1542e+12 42676
- Central_Air      1 1.7247e+10 2.1720e+12 42689
- Roof_Style       5 2.8560e+10 2.1833e+12 42692
- Half_Bath        1 3.4751e+10 2.1895e+12 42705
- Lot_Area         1 3.5041e+10 2.1898e+12 42706
- Fireplaces       1 3.6680e+10 2.1914e+12 42707
- House_Style      7 7.3149e+10 2.2279e+12 42729
- Garage_Area      1 6.3520e+10 2.2182e+12 42732
- Bldg_Type        4 7.3044e+10 2.2278e+12 42735
- Full_Bath        1 6.8973e+10 2.2237e+12 42737
- Second_Flr_SF    1 1.2513e+11 2.2798e+12 42788
- First_Flr_SF     1 1.4221e+11 2.2969e+12 42804
- Overall_Qual     9 1.7202e+12 3.8749e+12 43860
```

Other Criteria – BIC

```
full.model <- lm(Sale_Price ~ ., data = train_sel)
empty.model <- lm(Sale_Price ~ 1, data = train_sel)

step.model <- step(empty.model,
                   scope = list(lower = empty.model,
                                upper = full.model),
                   direction = “both”, k = log(nrow(train_sel)))
```

BIC Selection

Other Criteria – P-value Selection (α = 0.05)

```
full.model <- lm(Sale_Price ~ ., data = train_sel)
empty.model <- lm(Sale_Price ~ 1, data = train_sel)

step.model <- step(empty.model,
                   scope = list(lower = empty.model,
                                upper = full.model),
                   direction = “both”, k = qchisq(0.05, 1, lower.tail = FALSE))
```

P-value selection with alpha = 0.05

Can easily change alpha to any number…

Issues with Automatic Search Algorithms

- Automated model selection results in the following:
  - Biases in parameter estimates, predictions, and standard errors
  - Incorrect calculation of degrees of freedom (p-value method)
  - P-values that tend to err on the side of overestimating significance (increasing Type I Error probability)
- Can result in locally best model (not global)
- DO NOT blindly use result from automatic search algorithm as final model!!

Significance Levels

Conservative P-values (Adrian Raftery, 1994)

	Sample Size					
Evidence	**30**	**50**	**100**	**1,000**	**10,000**	**100,000**
Weak	.076	.053	.032	.009	.002	.0007
Fair	.028	.019	.010	.003	.0008	.0002
Strong	.005	.003	.001	.0003	.0001	.00003
Very Strong	.001	.0005	.0001	.00004	.00001	.000004

In Summary…

- Automatic stepwise search algorithms can help provide a subset of potential variables
- **NO** model chosen from one of these algorithms should be blindly selected as the final model (**always** explore other potential models and **investigate model assumptions**)
- If you use p-values for your selection, be sure to ***adjust your p-values*** if you have a large sample size