



# Model Selection

Class of 2024

# Review

- Simple Linear Regression
- Multiple Linear Regression
- With many explanatory variables, how do we know which ones are most informative?

Ames Housing Data

Sale Price =

Second Floor SF

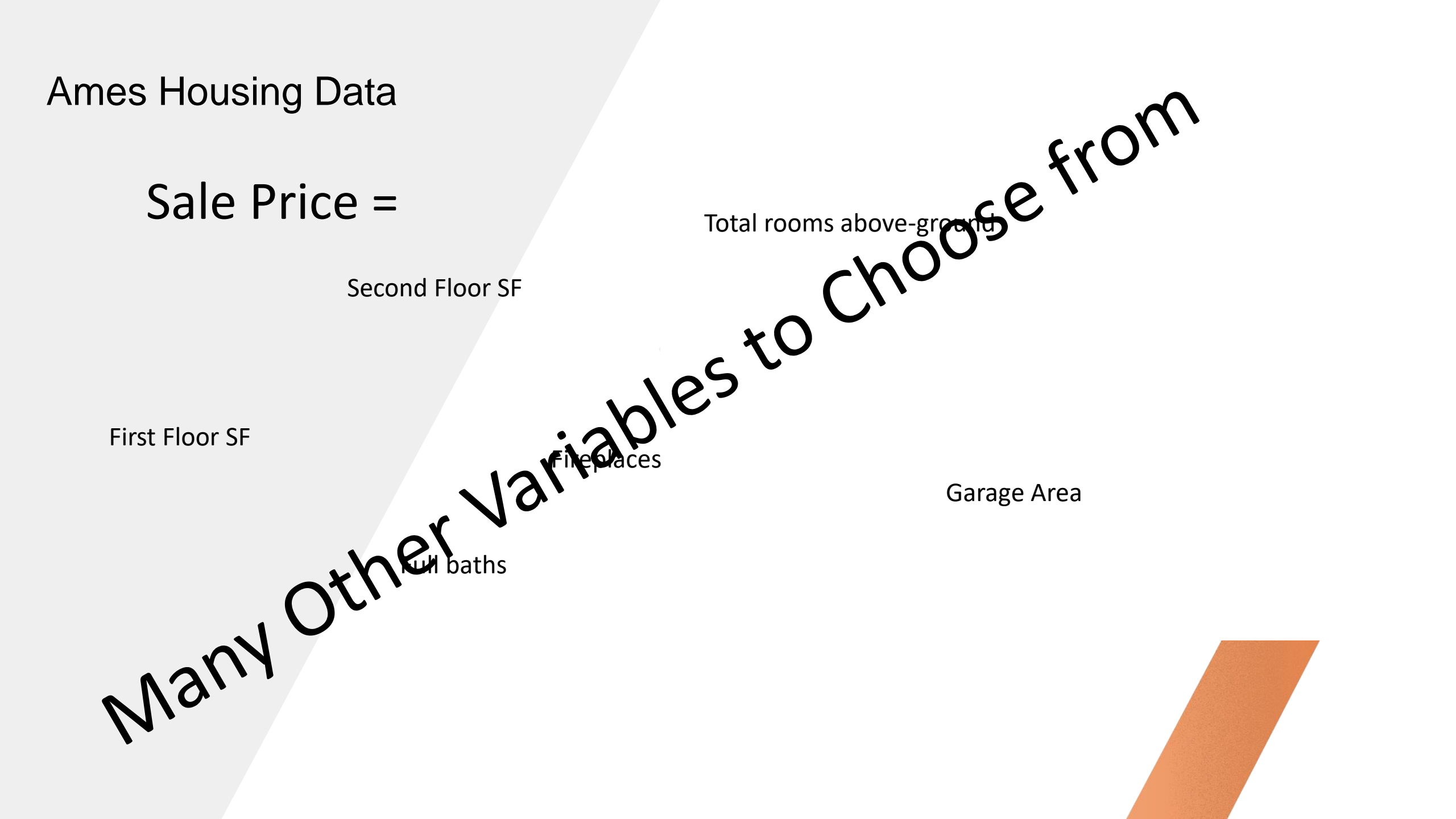
Total rooms above-ground

First Floor SF

Fireplaces

Garage Area

Full baths



Many Other Variables to Choose from

# Model Building

- Information Criteria
- Selection Algorithms
  - Forward Selection
  - Backward Elimination
  - Stepwise
- Discussion of p-values

# Model Building

- Should ALWAYS be done with training data!

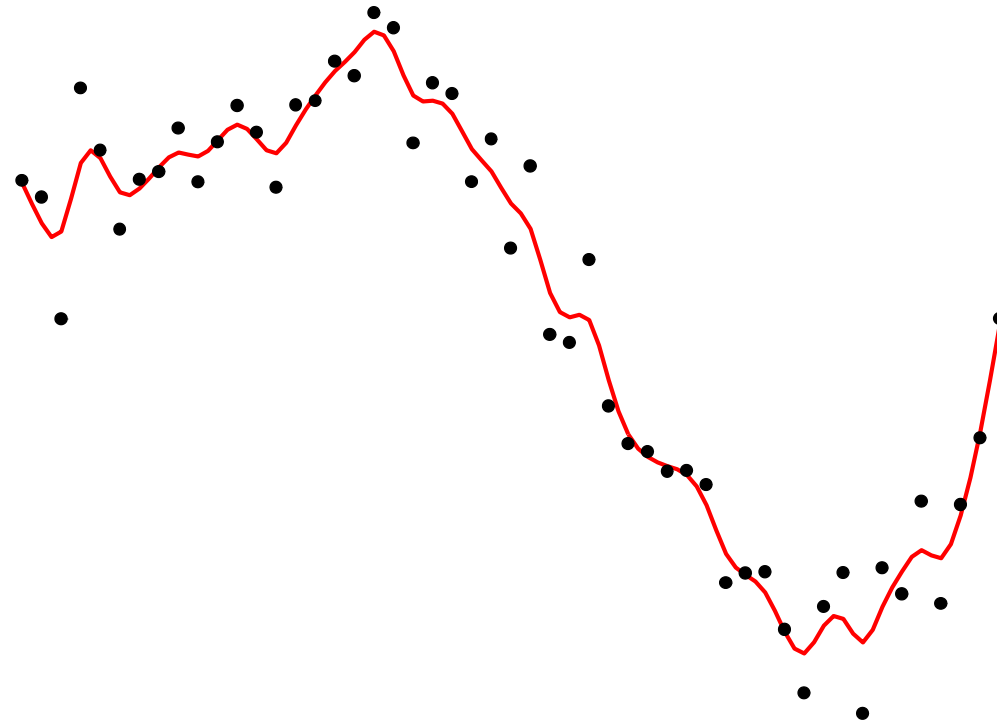
# Model Building

- Should ALWAYS be done with training data!



# Model Building

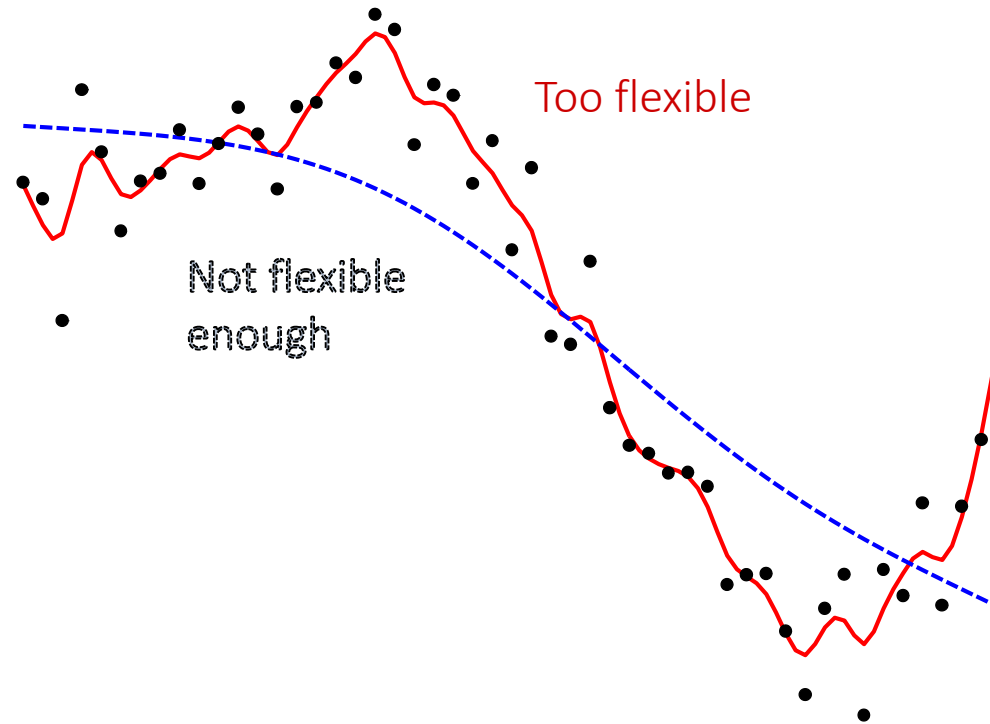
- Should ALWAYS be done with training data!



...

# Model Building

- Should ALWAYS be done with training data!

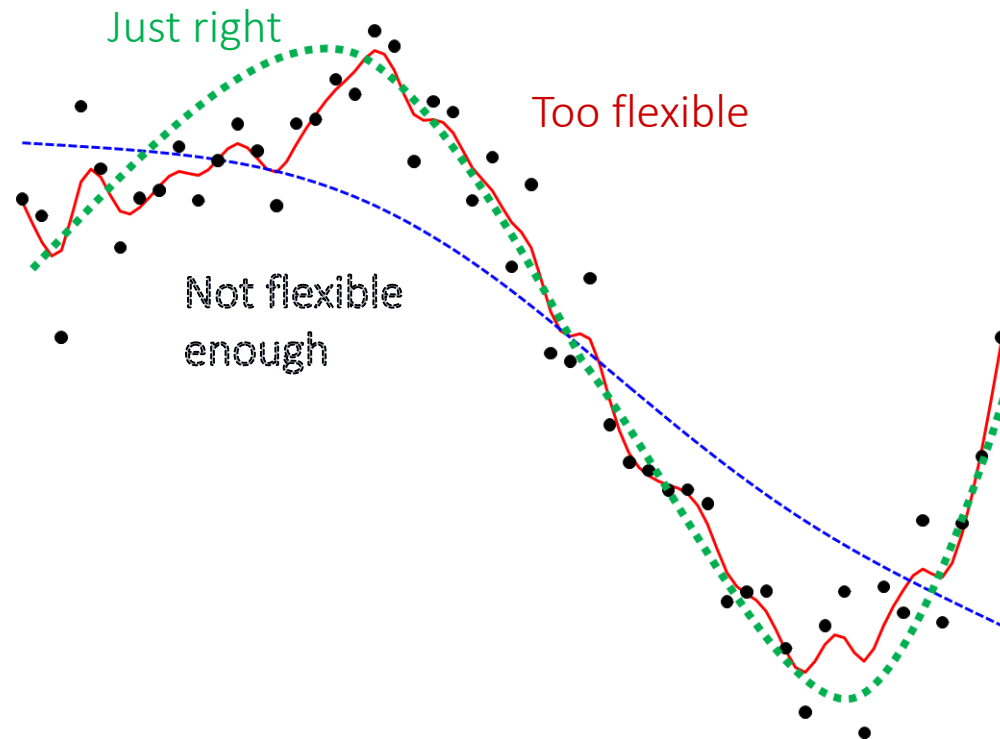


...



# Model Building

- Should ALWAYS be done with training data!





# Information Criteria

## Information Criteria

- Good to compare models (no one value “cutoff”)
- Uses the likelihood of the data with some penalty
- Two of the most common Information criteria are:
  - AIC: Akaike Information Criteria
  - BIC: Bayesian Information Criteria
- Smaller is better!!

## More on Information Criteria

- Different criteria have different penalties
  - AIC penalty:  $2p$
  - BIC penalty:  $p \log(n)$

Where  $p$  = # estimated parameters and  $n$  = sample size



# Forward Selection

# Forward Selection

In forward selection, we start with a “null” model (just the intercept) and systematically build the model (one variable at a time)

0. Start with a null model, this is the base model
1. For each variable not in model, create a linear regression model with the base model plus this variable
2. See which linear regression is best (based on criterion)
3. Is this regression better than the base model?
  - a. Yes, then continue on to step 4
  - b. No, exit the algorithm with the base model as the chosen model
4. Update the base model to be the one selected in Step 3. Using this as your new base model, go back to step 1 and continue.

# Forward Selection

0



There are 12 potential variables

# Forward Selection

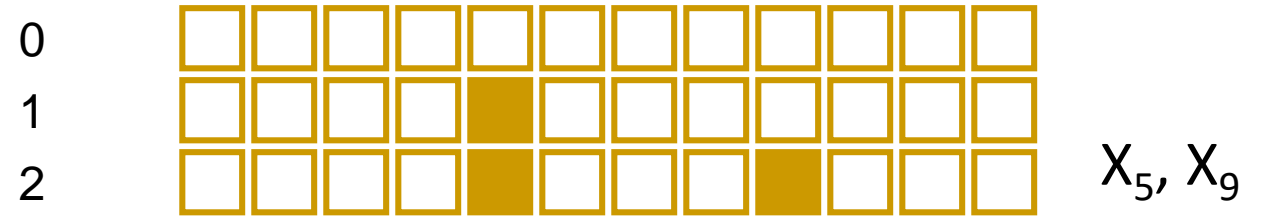
0

1

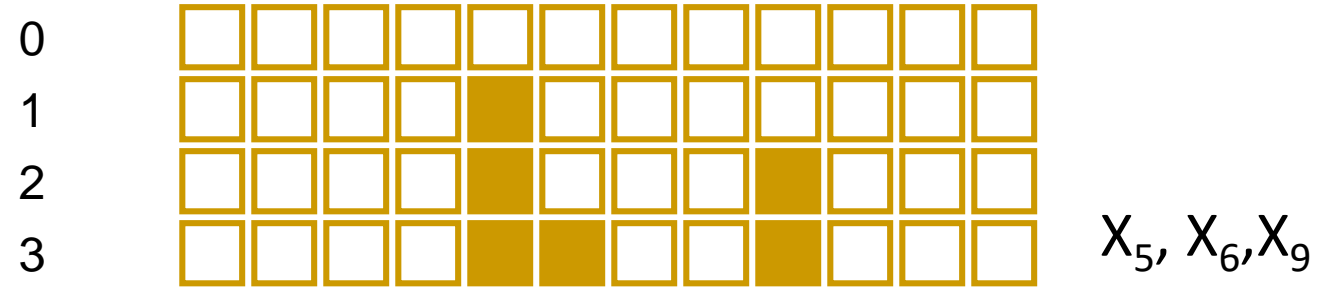

$$X_5$$



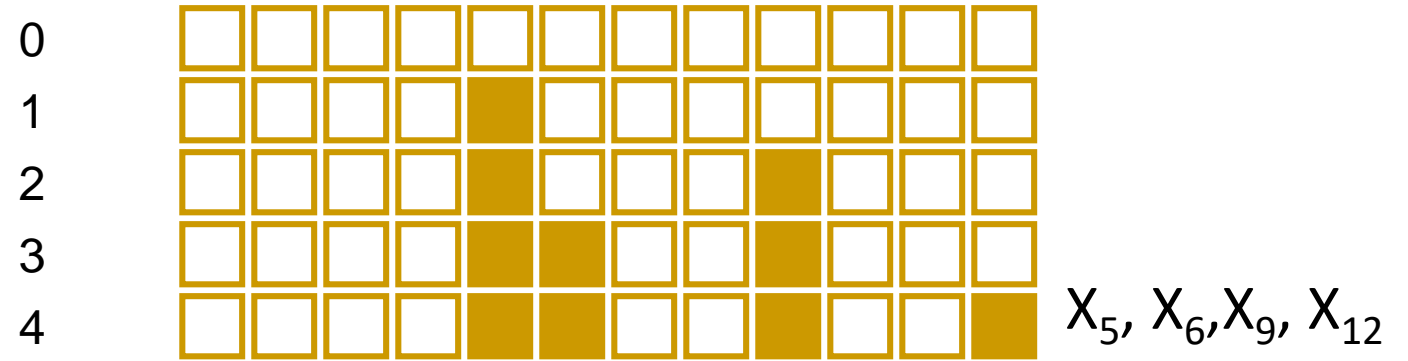
# Forward Selection



# Forward Selection



# Forward Selection



# Forward Selection

0 1 2 3 4 5

$X_5, X_6, X_8, X_9, X_{12}$

## Forward Selection

0											
1											
2											
3											
4											
5											
Stop											

Final Model includes:  $X_1, X_5, X_6, X_8, X_9, X_{12}$

Ames Housing data



# Data

```
train_sel = train %>%  
  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`) %>%  
  replace(is.na(.), 0)
```

# Forward Selection Code

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

# k = 2 for AIC selection
for.model <- step(empty.model,
                  scope = list(lower = empty.model,
                              upper = full.model),
                  direction = "forward", k = 2)
```



## Step 1

Start: AIC=46323.64

Sale\_Price ~ 1

	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

## Step 2

Step: AIC=43816.66  
Sale\_Price ~ Overall\_Qual

	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

Step: AIC=43210.24

Sale\_Price ~ Overall\_Qual + Gr\_Liv\_Area

## Step 3

	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

# Exit algorithm

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

- These algorithms are referred to as stepwise because at each step, ONLY one variable can be added or taken away
- You can also use p-values as your selection criteria (penalty is a  $\chi^2$  quantile)
- Adjusted  $R^2$  (unfortunately, R does not give you this option)

## Other Criteria

```
# k = log(n) for BIC selection
for.model2 <- step(empty.model,
  scope = list(lower = empty.model,
    upper = full.model),
  direction = "forward", k = log(nrow(train_sel)))
```

```
# k = qchisq(alpha, 1, lower.tail = FALSE) for p-value with alpha selection
alpha.f=0.05
for.model3 <- step(empty.model,
  scope = list(lower = empty.model,
    upper = full.model),
  direction = "forward", k = qchisq(alpha.f, 1, lower.tail = FALSE))
```



# Backward Elimination

# Backward Elimination

Backward elimination systematically removes variables “not informative” in the model

0. Start with full model with all predictor variables in it, this is the base model and calculate the criterion on this model
1. Create models such that each model has exactly one predictor variable removed from it and calculate the criterion for each model
2. From models in step 1, find the best model based on the criterion
3. Is this regression model better than the base model?
  - a. Yes, then continue on to step 4
  - b. No, exit the algorithm with the base model as the chosen model
4. Update the base model to be the model selected in Step 2. Using this as your new base model, go back to step 1 and continue.



0



Start with all 12 variables in model

# Backward Elimination

0

1



Remove  $x_2$

Backward Elimination

0

1

2


Remove  $x_{11}$

Backward Elimination

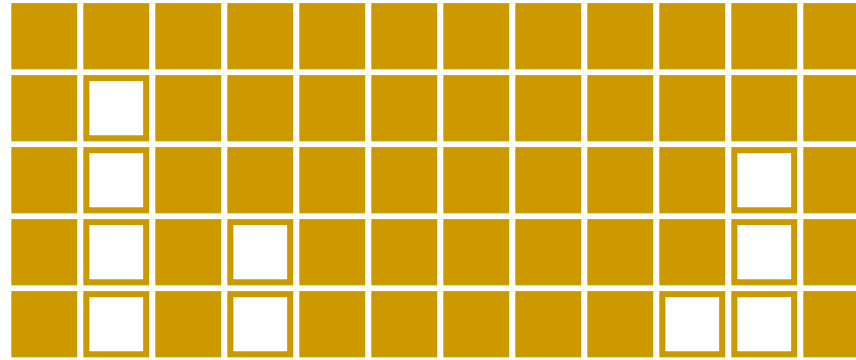
# Backward Elimination

0												
1												
2												
3												

Remove  $x_4$

# Backward Elimination

0  
1  
2  
3  
4



Remove  $x_{10}$

# Backward Elimination

0											
1											
2											
3											
4											
5											

Remove  $x_7$

# Backward Elimination

0  
1  
2  
3  
4  
5  
6


Remove  $x_3$

# Backward Elimination

0												
1												
2												
3												
4												
5												
6												
Stop												

Remove  $x_6$

Final Model includes:  $x_1, x_5, x_8, x_9, x_{12}$



Ames Housing data



# Code

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

# k = 2 for AIC selection
back.model <- step(full.model,
  scope = list(lower = empty.model,
    upper = full.model),
  direction = "backward", k = 2)
```

# Step 1

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

## Step 2

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

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

## Step 3 (last step)

	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

```
# k = log(n) for BIC selection
back.model2 <- step(full.model,
  scope = list(lower = empty.model,
    upper = full.model),
  direction = "backward", k = log(nrow(train_sel)))
```

```
# k = qchisq(alpha, 1, lower.tail = FALSE) for p-value with alpha selection
alpha.f=0.05
back.model3 <- step(full.model,
  scope = list(lower = empty.model,
    upper = full.model),
  direction = "backward", k = qchisq(alpha.f, 1, lower.tail = FALSE))
```



# Stepwise selection

# Stepwise Selection

Stepwise selection can systematically add or delete one variable from the model

0. Start with empty model with only the intercept in it, this is the base model and calculate the criterion on this model
1. For each variable not in model, create a linear regression model with the base model plus this one variable
2. For each variable in the model, create models with the base model taking away one variable at a time
3. See which linear regression is best from Step 1 and Step 2(based on criterion)
4. Is this regression better than the base model?
  - a. Yes, then continue on to step 5
  - b. No, exit the algorithm with the base model as the chosen model
5. Update the base model to the one selected in Step 3 and go back to step 1 and continue.



# Stepwise selection

0




































Start with null model

# Stepwise selection

0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

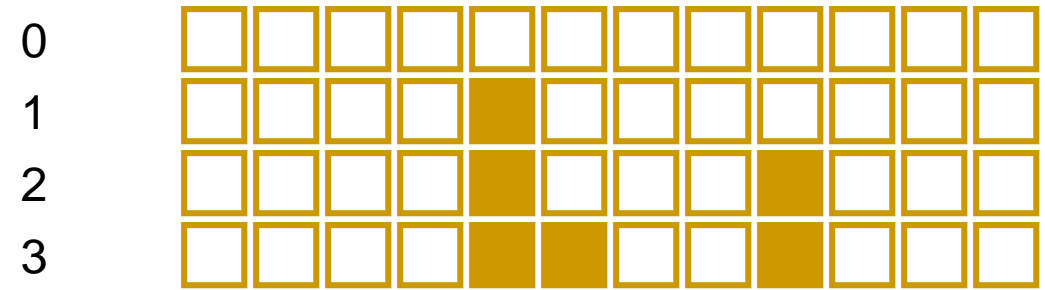
Add  $X_5$

# Stepwise selection

0											
1											
2											

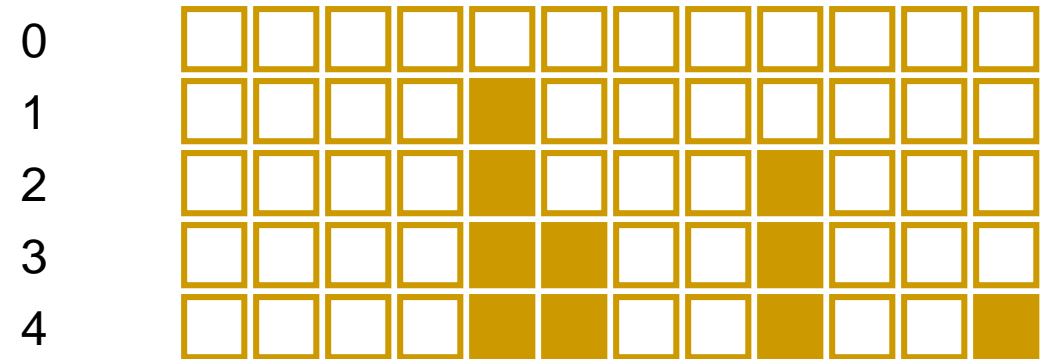
Add  $X_9$

# Stepwise selection



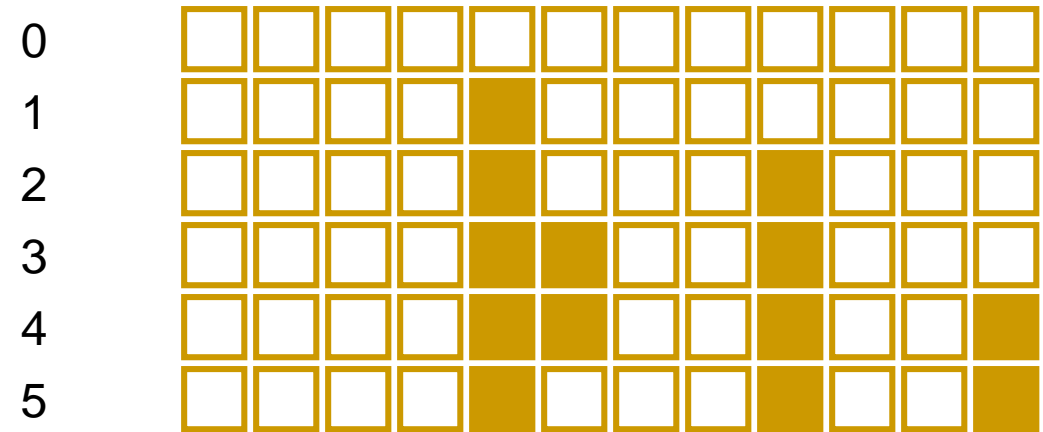
Add  $X_6$

# Stepwise selection



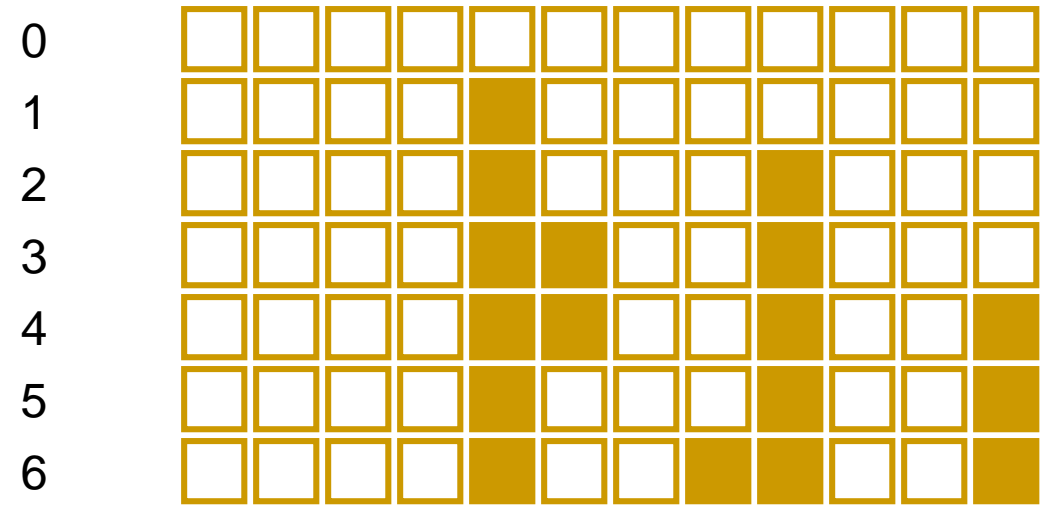
Add  $X_{12}$

## Stepwise selection



Remove  $X_6$

# Stepwise selection



Add  $X_8$

## Stepwise selection

0											
1											
2											
3											
4											
5											
6											
Stop											

Add  $X_1$

Final Model includes:  $X_1, X_5, X_8, X_9, X_{12}$



Ames Housing data



# Code

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

# k = 2 for AIC selection
step.model <- step(empty.model,
                    scope = list(lower = empty.model,
                                  upper = full.model),
                    direction = "both", k = 2)
```

Start: AIC=46323.64

Sale\_Price ~ 1

## Step 1

	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

## Step 2

Step: AIC=43816.66  
Sale\_Price ~ Overall\_Qual

	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

## Step 3

Step: AIC=43210.24

Sale\_Price ~ Overall\_Qual + Gr\_Liv\_Area

	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

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

## Exit algorithm

	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

```
# k = log(n) for BIC selection
for.model2 <- step(empty.model,
  scope = list(lower = empty.model,
    upper = full.model),
  direction = "both", k = log(nrow(train_sel)))
```

```
# k = qchisq(alpha, 1, lower.tail = FALSE) for p-value with alpha selection
alpha.f=0.05
for.model3 <- step(empty.model,
  scope = list(lower = empty.model,
    upper = full.model),
  direction = "both", k = qchisq(alpha.f, 1, lower.tail = FALSE))
```

# 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

Source: Adrian Raftery, 1994

Sample Size				
Evidence	30	50	100	1000
Weak	.076	.053	.032	.009
Fair	.028	.019	.010	.003
Strong	.005	.003	.001	.0003
Very Strong	.001	.0005	.0001	.00004

# Wrap-up

- 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



# Break out and Lab 7

Don't forget to take the lab check on Moodle!