

# STAT 401A - Statistical Methods for Research Workers

## Model refinement analysis - Ames housing prices

Jarad Niemi (Dr. J)

Iowa State University

last updated: November 21, 2014

# Objective

Determine the fair market value of a house in Ames with asking price of \$240,000 with the following details:

- 2-story on 9201 ft<sup>2</sup> land
- 4 bedroom
- 2 bathroom
- 2 car attached garage
- built in 1975
- 2199 ft<sup>2</sup> above grade plus
- basement: 780 ft<sup>2</sup> (75% finished) with additional bath

based on houses sold in Ames from June 2010 to August 2011.

What does fair market value mean?

- What would people pay on average?
- What would we predict the next person would pay?

# Multiple regression

Decisions to make before performing analysis (you can change your mind)

- Data
  - June 2010 to August 2011
  - 2-story vs 1.5-story vs 1-story
- Response
  - Price (log?)
- Explanatory variables
  - continuous vs categorical, e.g. # of bathrooms
  - transformations (log)
  - higher order (squared terms)
  - interactions

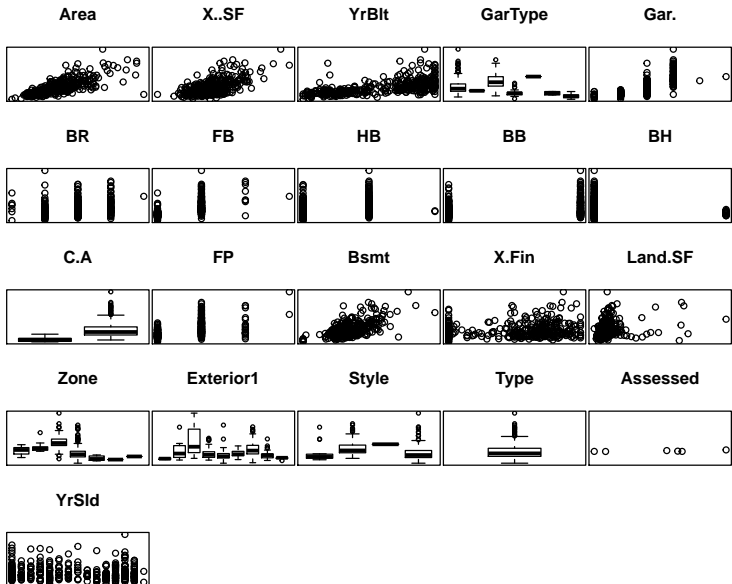
# Explanatory variables

- Year built
- Year/month sold
- Area of
  - Land
  - Above grade living space
  - Basement living space
- Number of
  - Bedrooms
  - Bathrooms above grade
  - Half bathrooms above grade
  - Bathrooms in basement
  - Half bathrooms in basement
  - Fireplaces
- Garage
  - size: number of cars
  - type: (attached, detached, etc)
- Exterior type: (VinylSd, HdBoard, etc)
- Number of levels
- Percent of basement that is finished

```

[1] 363 27
      Price      GarType      C.A      Style      Exterior1
Min.   : 62000   Attachd:271   No : 6   1.5 Fin: 16   VinylSd:163
1st Qu.:144250   Basment: 3   Yes:357  2-Story:123   HdBoard: 66
Median :180000   BuiltIn: 31   2.5 Unf: 1   Wd Sdng: 36
Mean   :204772   Detachd: 47   1-Story:223   MetalSd: 35
3rd Qu.:241750   2 Types: 1   Plywood: 28
Max.   :665000   No Data: 2   CemntBd: 18
      None : 8   (Other): 17
AsbShng BrkFace CemntBd HdBoard MetalSd Plywood VinylSd Wd Sdng WdShing
      2      8      18      66      35      28      163      36      7

```



# Exploratory analysis

Adjustments made:

- Slimmed data set to include only the following
  - Zoned residential low density
  - Sell price \$150-\$300k
  - Has central air
  - 1-story or 2-story (no 1.5 or 2.5 story)
  - Exterior not AsbShng, BrkFace, WdShing
- Created new variables:
  - Date sold (Year+Month/12)
  - Basement finished area
  - Basement unfinished area
- Variables eliminated
  - % of basement finished
  - Type
  - Assessed

Fit initial model with all explanatory variables, but no interaction

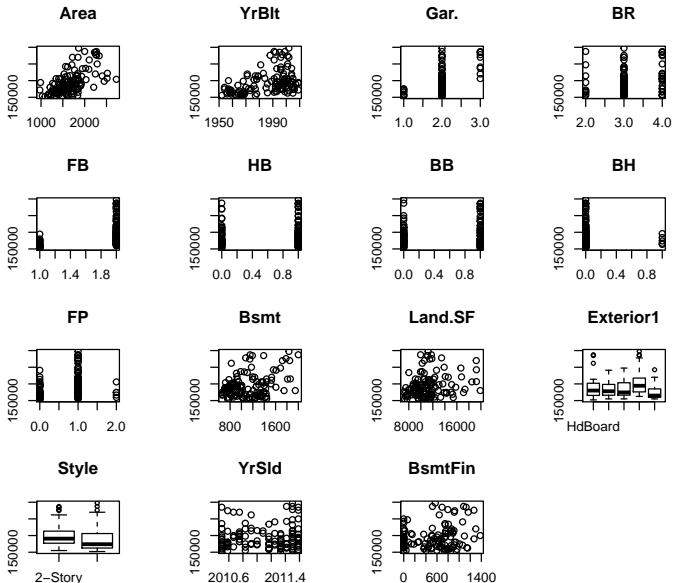
## Slimmed data set

```
[1] 126 16
      Price          Area          YrBlt          Gar.          BR          FB          HB
Min.   :152000    Min.   : 972    Min.   :1952    Min.   :1.000    Min.   :2.000    Min.   :1.000    Min.   :0.0000
1st Qu.:169000    1st Qu.:1410    1st Qu.:1969    1st Qu.:2.000    1st Qu.:3.000    1st Qu.:2.000    1st Qu.:0.0000
Median :185500    Median :1613    Median :1992    Median :2.000    Median :3.000    Median :2.000    Median :1.0000
Mean   :194279    Mean   :1659    Mean   :1986    Mean   :1.984    Mean   :3.111    Mean   :1.802    Mean   :0.6111
3rd Qu.:206800    3rd Qu.:1846    3rd Qu.:1999    3rd Qu.:2.000    3rd Qu.:3.000    3rd Qu.:2.000    3rd Qu.:1.0000
Max.   :297500    Max.   :2726    Max.   :2009    Max.   :3.000    Max.   :4.000    Max.   :2.000    Max.   :1.0000

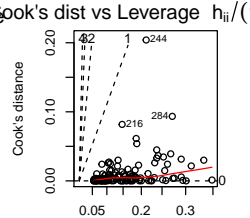
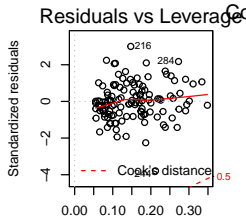
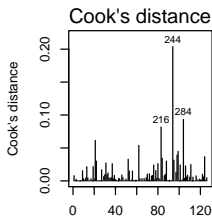
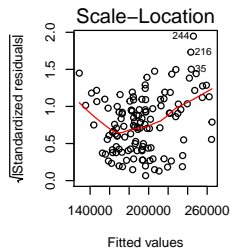
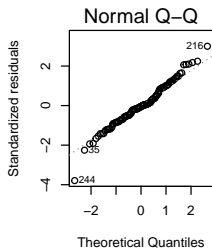
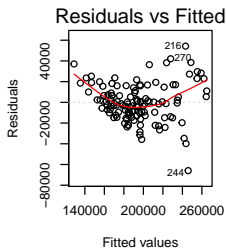
      BB          BH          FP          Bsmt          Land.SF          Exterior1          Style
Min.   :0.0000    Min.   :0.00000    Min.   :0.0000    Min.   : 654    Min.   : 7153    HdBoard:31    2-Story:64
1st Qu.:0.0000    1st Qu.:0.00000    1st Qu.:0.0000    1st Qu.: 832    1st Qu.: 9309    MetalSd:10    1-Story:62
Median :0.0000    Median :0.00000    Median :1.0000    Median :1045    Median :10785    Plywood:15
Mean   :0.4286    Mean   :0.03968    Mean   :0.7143    Mean   :1117    Mean   :11233    VinylSd:59
3rd Qu.:1.0000    3rd Qu.:0.00000    3rd Qu.:1.0000    3rd Qu.:1349    3rd Qu.:12114    Wd Sdng:11
Max.   :1.0000    Max.   :1.00000    Max.   :2.0000    Max.   :2000    Max.   :19900

      YrSld          BsmtFin
Min.   :2010    Min.   : 0.0
1st Qu.:2011    1st Qu.: 0.0
Median :2011    Median : 583.2
Mean   :2011    Mean   : 508.2
3rd Qu.:2012    3rd Qu.: 808.2
Max.   :2012    Max.   :1381.4
```





```
mod = lm(Price~., slim)
par(mfrow=c(2,3))
plot(mod,1:6, ask=F)
```



# Response and outliers

Comments from plots:

- There is some evidence that increased price leads to increased residuals, so consider logarithm of Price as the response.
- Some observations were flagged as important, but none are exerting an undue influence on the regression (Cook's distance less than 0.20).
- There is a concern that there are outlying observations and therefore heavier tails.

# Training-testing data sets

For example,

- 1 Randomly split your observations into two sets:
  - training
  - testing
- 2 Use the training data set to find model(s), e.g.
  - use a model selection procedure to find a model and
  - estimate the parameters in that model.
- 3 Use the testing data set to evaluate the model(s), e.g. calculate mean square prediction error in the testing data, i.e.

$$MSPE = \frac{1}{n'} \sum_{i=1}^{n'} (P_i - \hat{P}_i)^2$$

where

- $P_i$  is the actual sale price for house  $i$  in the testing data set and
- $\hat{P}_i$  is the predicted sale price from a particular model.

## Candidate models

Use all explanatory variables from earlier and allow the following models combinations:

- Response: Price and  $\log(\text{Price})$
- Interactions: Yes and No
- Selection criterion: AIC and BIC

For a particular combination,

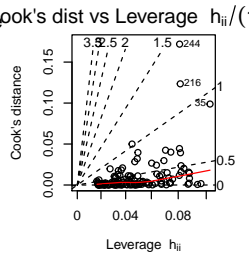
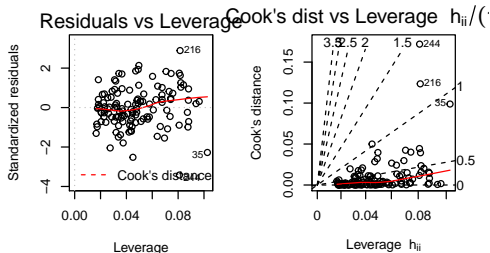
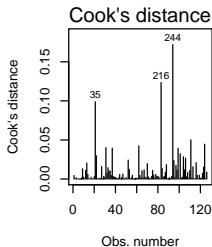
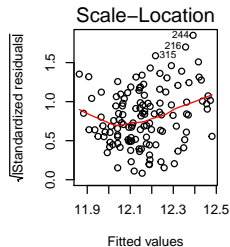
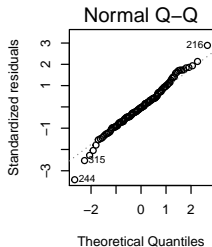
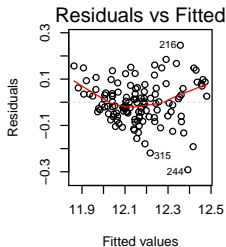
- 1 Initialize model to have all main effects.
- 2 Use stepwise selection to select a model.
- 3 Calculate the model's MSPE

Choose the model that has the lowest mean squared prediction error amongst all these models.

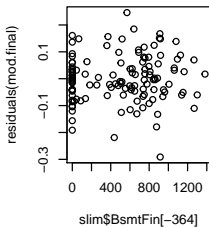
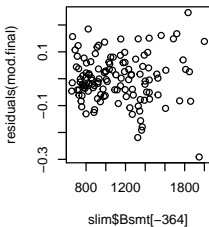
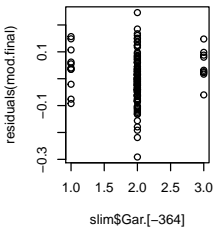
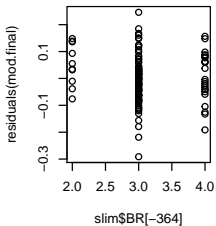
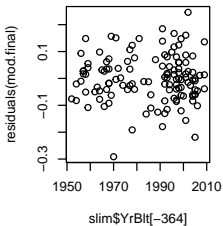
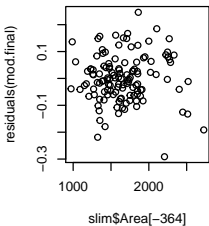
# Results

	Response	Interactions	Criterion	sqrt(MSPE)	Ratio
1	Price	No	AIC	16389	1.20
2	Price	Yes	AIC	519864	1202.62
3	log(Price)	No	AIC	15751	1.10
4	log(Price)	Yes	AIC	18866950	1583997.40
5	Price	No	BIC	17958	1.43
6	Price	Yes	BIC	16158	1.16
7	log(Price)	No	BIC	14991	1.00
8	log(Price)	Yes	BIC	15756	1.10

# Diagnostic plots for final model (7) using all data



# Quadratic terms?





# Final model when using all data

## Summary of the final model estimated using all observations

Call:

```
lm(formula = formula(mod[[id]]), data = slim)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.291314	-0.050565	-0.003477	0.052854	0.245974

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	3.484e+00	1.072e+00	3.251	0.001494	**
Area	2.569e-04	2.436e-05	10.549	< 2e-16	***
YrBlt	3.998e-03	5.456e-04	7.327	2.97e-11	***
Bsmt	1.409e-04	2.566e-05	5.492	2.26e-07	***
BB	5.808e-02	1.668e-02	3.482	0.000696	***
Gar.	6.636e-02	2.352e-02	2.822	0.005591	**

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

Residual standard error: 0.08883 on 120 degrees of freedom

Multiple R-squared: 0.7225, Adjusted R-squared: 0.711

F-statistic: 62.5 on 5 and 120 DF, p-value: < 2.2e-16

# Prediction

```
new <- read.csv("Ch12a-new.csv",header=T)
new$YrSld = 2012
exp(predict(mod.final, new, interval="confidence"))

      fit      lwr      upr
1 208047.7 199141.6 217352.1

exp(predict(mod.final, new, interval="prediction"))

      fit      lwr      upr
1 208047.7 173561.7 249386
```

One aspect that has been completely neglected is location of the properties which clearly has a large impact on the fair market value.

# Summary

Who would perform a regression like this?

- Buyer
- Seller
- Real estate agent
- Mortgage appraiser
- Tax assessor