Multiple Regression

ANALYZING THE HOUSING DATA

Zillow Data

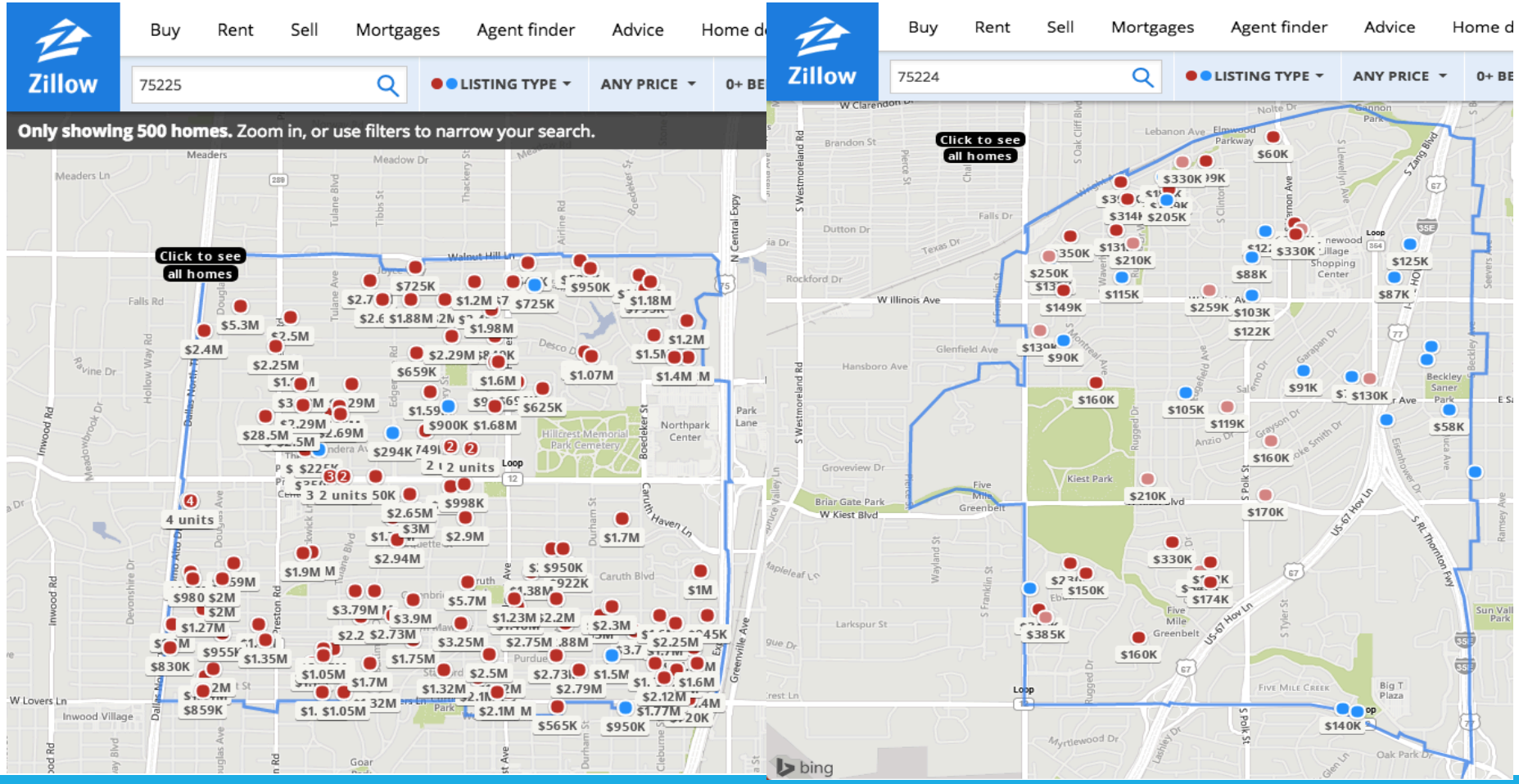
- Many companies are currently trying to make online businesses for providing information to prospective buyers/sellers of houses

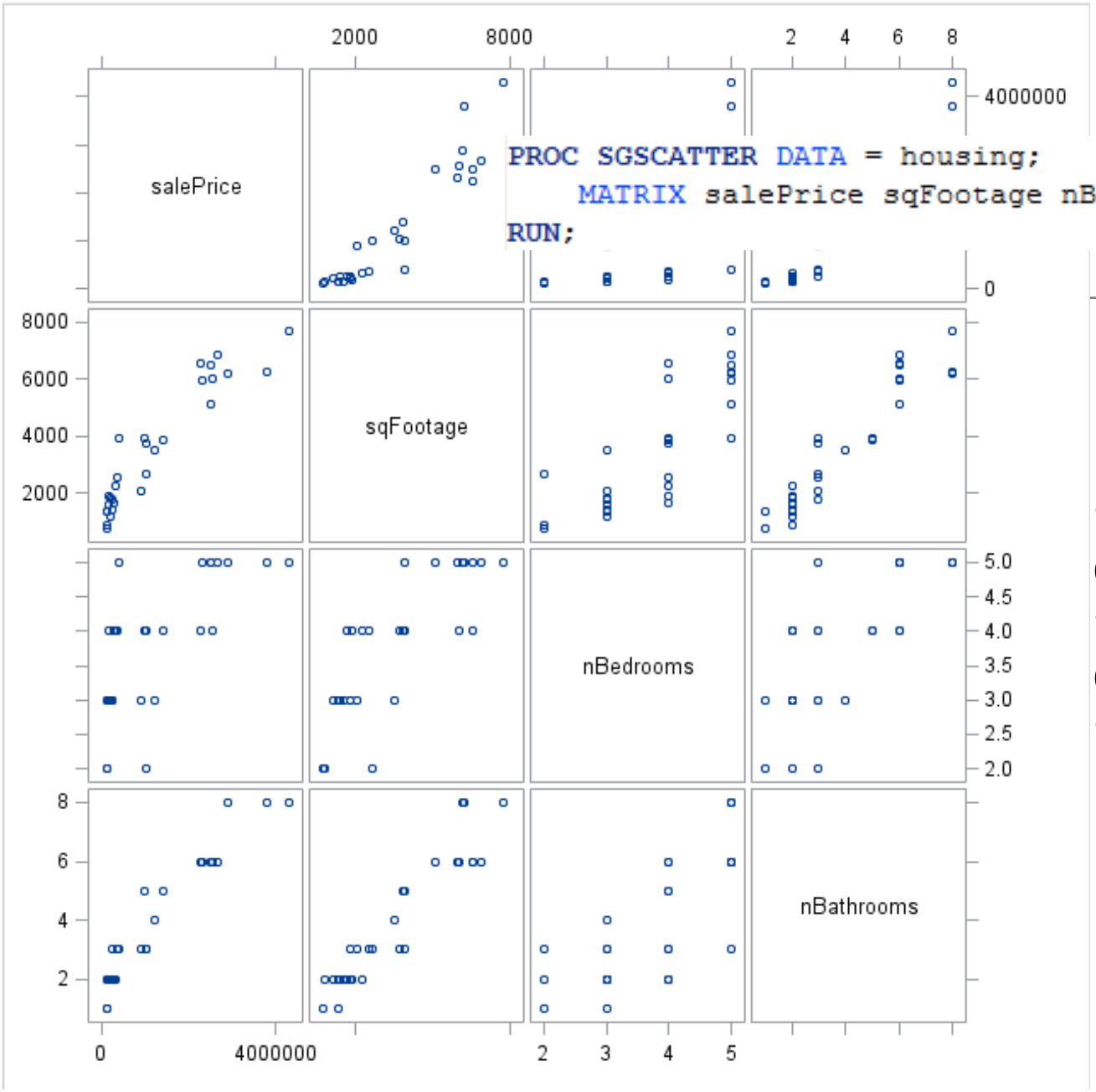
  (zillow.com, trulia.com, redfin.com, realtor.com,…)

- One of the core parts of the business is to analyze/predict the sale price of houses based on factors such as square footage, number of bedrooms/bathrooms, location, school district, ...

- We will look at data from zillow to build a sale price model using multiple regression

Involving Zipcode






```
PROC SGSCATTER DATA = housing;
    MATRIX salePrice sqFootage nBedrooms nBathrooms;
RUN;
```

"Scatterplot matrix"
Or
"Draftsmen plot"
Or
"Pairs plot"

Involving Zipcode and Using "PROC GLM"

This class statement is crucial (why?)

```
ODS GRAPHICS ON;
PROC GLM DATA = housing PLOTS=(ALL);
    CLASS zipCode (ref = '75224');
    MODEL salePrice = sqFootage zipCode / SOLUTION CLPARM;
RUN;
ODS GRAPHICS OFF;
```

$$\mu\{salePrice|ft^2, zipcode\} = \beta_0 + \beta_1 ft^2 + \beta_2 zipcode$$

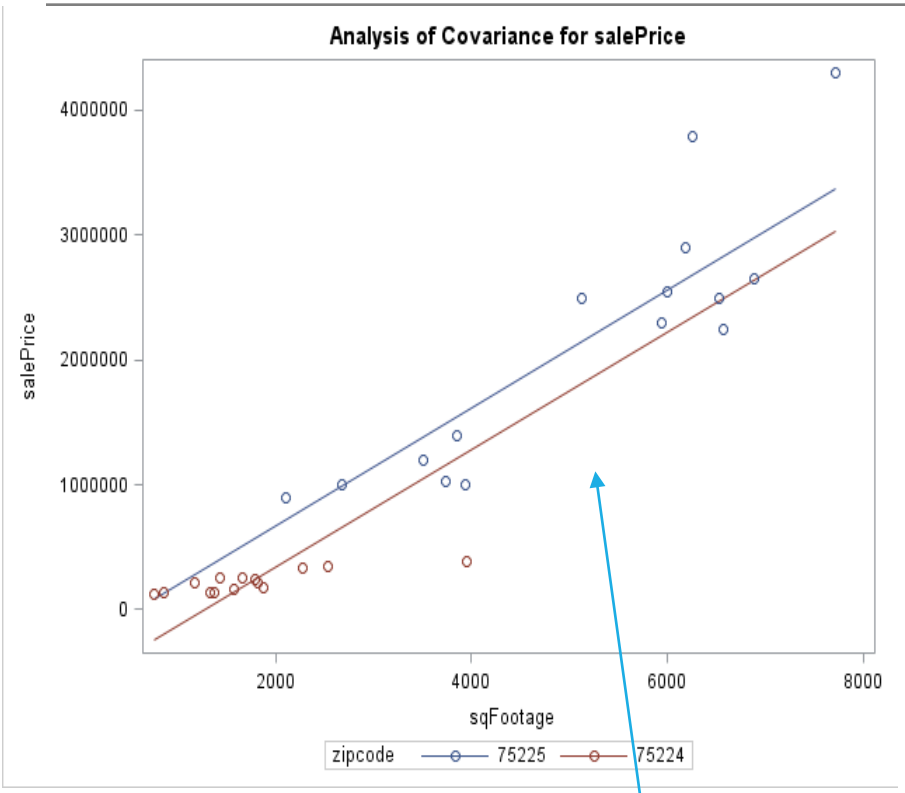
```
/* sq ft with UNequal slopes*/
ODS GRAPHICS ON;
PROC GLM DATA = housing PLOTS=(ALL);
    CLASS zipCode (ref = '75224');
    MODEL salePrice = sqFootage zipCode zipCode*sqFootage / SOLUTION CLPARM;
RUN;
ODS GRAPHICS OFF;
```

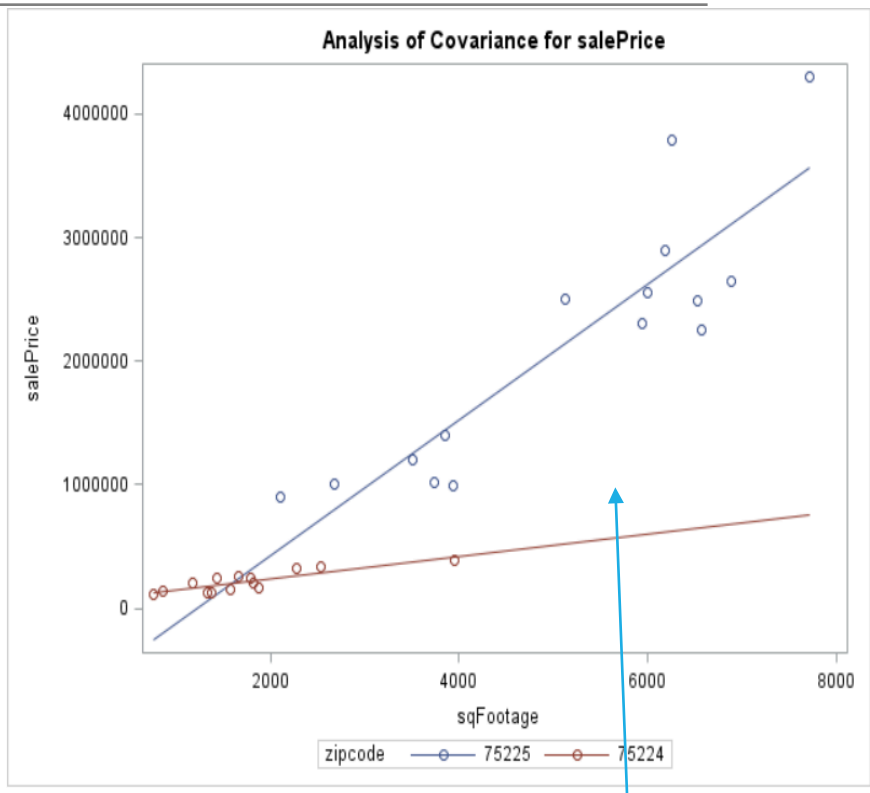
$$\mu\{salePrice|ft^2, zipcode\} = \beta_0 + \beta_1 ft^2 + \beta_2 zipcode + \beta_3 ft^2 * zipcode$$

These are known as **MAIN EFFECTS**

This is known as an **INTERACTION TERM**

Involving Zip Code



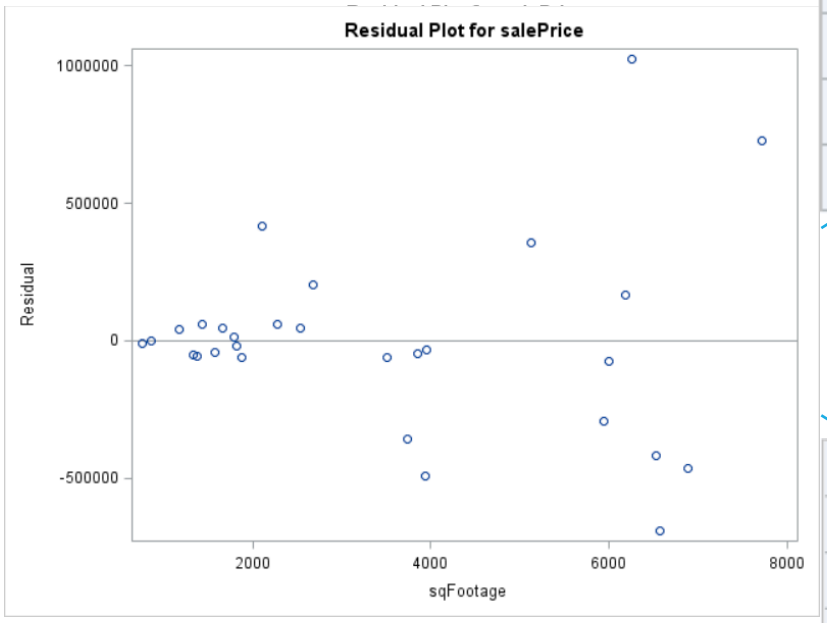


$$\hat{\mu}\{salePrice|ft^2, zipcode\} = \widehat{\beta_0} + \widehat{\beta_1} ft^2 + \widehat{\beta_2} zipcode$$

$$\hat{\mu}\{salePrice|ft^2, zipcode\} = \widehat{\beta_0} + \widehat{\beta_1} ft^2 + \widehat{\beta_2} zipcode + \widehat{\beta_3} ft^2 * zipcode$$

Involving Zip Code: Model Fit and Residuals

(Ignore these)



Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-602754.4865	B	156690.1763	-3.85	0.0007
sqFootage	471.4142		61.3530	7.68	<.0001
zipcode 75225	333602.6291	B	261744.3432	1.27	0.2137
zipcode 75224	0.0000	B	.	.	.

$$\mu\{salePrice|ft^2, zipcode\} = \beta_0 + \beta_1 ft^2 + \beta_2 zipcode$$

$$\mu\{salePrice|ft^2, zipcode\} = \beta_0 + \beta_1 ft^2 + \beta_2 zipcode + \beta_3 ft^2 * zipcode$$

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	61408.9606	B	240645.5881	0.26	0.8007
sqFootage	90.6333	B	126.1889	0.72	0.4793
zipcode 75225	-733067.9288	B	391300.1388	-1.87	0.0727
zipcode 75224	0.0000	B	.	.	.
sqFootage*zipcode 75225	459.2229	B	138.5784	3.31	0.0028
sqFootage*zipcode 75224	0.0000	B	.	.	.

Involving Zip Code: Interpretation

If the p-value for the interaction term is significant, interpret it even if the p-values for the main effects are not significant

There is evidence that sqFootage depends on zipcode (two-sided p-value 0.0028 for the interaction). We estimate every 1 sq. ft. increase is associated with an $90.6 increase for zip=75224 or 90.6+459.2 = $549.8 increase for zip=75225 in mean sale price.

(To get confidence intervals for this effect we can use a contrasts)

$$\mu\{salePrice|ft^2, zipcode\} = \beta_0 + \beta_1 ft^2 + \beta_2 zipcode + \beta_3 ft^2 * zipcode$$

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	61408.9606	B	240645.5881	0.26	0.8007
sqFootage	90.6333	B	126.1889	0.72	0.4793
zipcode 75225	-733067.9288	B	391300.1388	-1.87	0.0727
zipcode 75224	0.0000	B	.	.	.
sqFootage*zipcode 75225	459.2229	B	138.5784	3.31	0.0028
sqFootage*zipcode 75224	0.0000	B	.	.	.

Multiple Regression: Additional variables

We saw that the residuals show evidence of an assumption violation

Hence, our results/interpretation are highly suspect

We can try and add additional explanatory variables to improve the linearity assumption

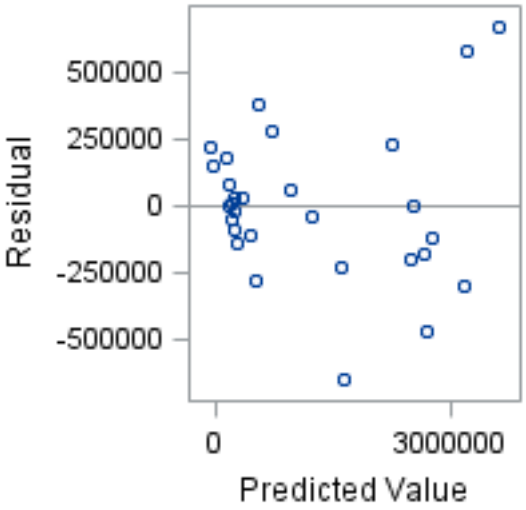
$$\mu\{salePrice|ft^2, zipcode, nBaths, nBeds\}$$
$$= \beta_0 + \beta_1 ft^2 + \beta_2 zipcode + \beta_3 ft^2 * zipcode + \beta_4 nBaths + \beta_5 nBeds$$

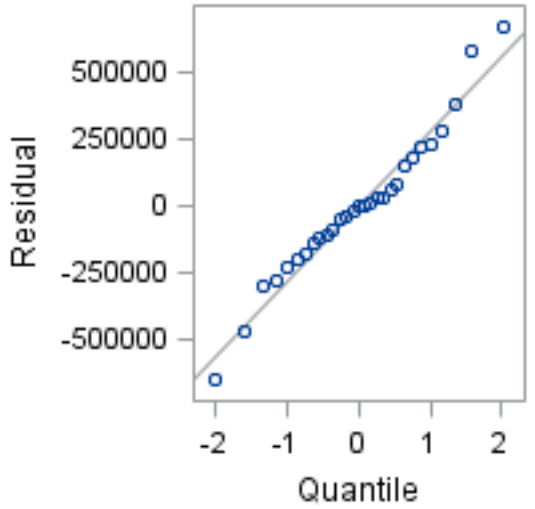
Multiple Regression: Model Statement

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	-242417.9238	B	296441.9769	-0.82	0.4219	-855654.8754	370819.0278
sqFootage	-47.0805	B	157.8412	-0.30	0.7682	-373.6000	279.4389
zipcode 75225	-674673.2583	B	329453.2933	-2.05	0.0522	-1356199.321	6852.8043
zipcode 75224	0.0000	B	.	.	.	.	.
sqFootage*zipcode 75225	334.1428	B	137.3190	2.43	0.0231	50.0768	618.2087
sqFootage*zipcode 75224	0.0000	B	.	.	.	.	.
nBathrooms	311379.2010		88611.2275	3.51	0.0019	128072.9110	494685.4911
nBedrooms	-30730.1490		127832.3057	-0.24	0.8122	-295171.4210	233711.1230

$$\mu\{salePrice|ft^2, zipcode, nBaths, nBeds\}$$
$$= \beta_0 + \beta_1 ft^2 + \beta_2 zipcode + \beta_3 ft^2 * zipcode + \beta_4 nBaths + \beta_5 nBeds$$

Multiple Regression: Residuals





No strong evidence of assumption violations

Multiple Regression: Interpretation

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	-242417.9238	B	296441.9769	-0.82	0.4219	-855654.8754	370819.0278
sqFootage	-47.0805	B	157.8412	-0.30	0.7682	-373.6000	279.4389
zipcode 75225	-674673.2583	B	329453.2933	-2.05	0.0522	-1356199.321	6852.8043
zipcode 75224	0.0000	B	.	.	.	.	.
sqFootage*zipcode 75225	334.1428	B	137.3190	2.43	0.0231	50.0768	618.2087
sqFootage*zipcode 75224	0.0000	B	.	.	.	.	.
nBathrooms	311379.2010		88611.2275	3.51	0.0019	128072.9110	494685.4911
nBedrooms	-30730.1490		127832.3057	-0.24	0.8122	-295171.4210	233711.1230

Evidence that nBedrooms isn't an important explanatory variable **given** the other explanatory variables are in the model

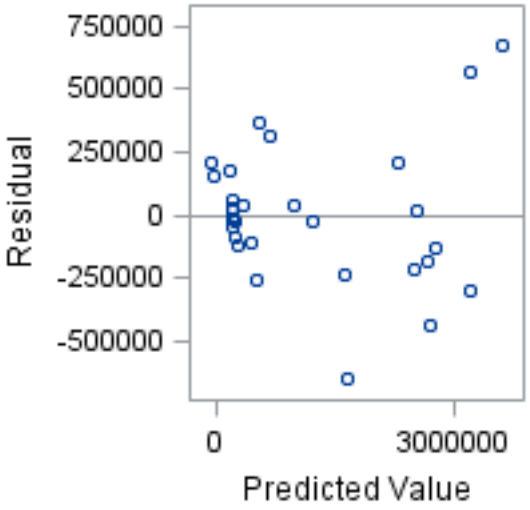
Let's refit/recheck the model without nBedrooms

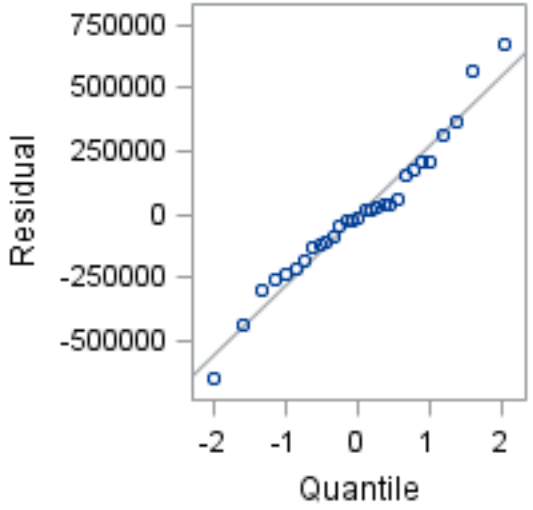
Multiple Regression: Without nBedrooms

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	-288934.0937	B	220126.6280	-1.31	0.2017	-743253.1246	165384.9372
sqFootage	-72.9738	B	113.0904	-0.65	0.5249	-306.3808	160.4332
zipcode 75225	-681603.9196	B	321682.7404	-2.12	0.0446	-1345524.465	-17683.3745
zipcode 75224	0.0000	B	.	.	.	.	.
sqFootage*zipcode 75225	350.1357	B	117.7455	2.97	0.0066	107.1208	593.1505
sqFootage*zipcode 75224	0.0000	B	.	.	.	.	.
nBathrooms	306893.9357		84907.2381	3.61	0.0014	131654.0092	482133.8622

$$\mu\{salePrice|ft^2, zipcode, nBaths\}$$
$$= \beta_0 + \beta_1 ft^2 + \beta_2 zipcode + \beta_3 ft^2 * zipcode + \beta_4 nBaths$$

Multiple Regression: Residuals





These look pretty good, no strong evidence of assumption violations

Multiple Regression: Interpretation

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	-288934.0937	B	220126.6280	-1.31	0.2017	-743253.1246	165384.9372
sqFootage	-72.9738	B	113.0904	-0.65	0.5249	-306.3808	160.4332
zipcode 75225	-681603.9196	B	321682.7404	-2.12	0.0446	-1345524.465	-17683.3745
zipcode 75224	0.0000	B	.	.	.	.	.
sqFootage*zipcode 75225	350.1357	B	117.7455	2.97	0.0066	107.1208	593.1505
sqFootage*zipcode 75224	0.0000	B	.	.	.	.	.
nBathrooms	306893.9357		84907.2381	3.61	0.0014	131654.0092	482133.8622

$$\mu\{salePrice|ft^2, zipcode, nBaths\}$$
$$= \beta_0 + \beta_1 ft^2 + \beta_2 zipcode + \beta_3 ft^2 * zipcode + \beta_4 nBaths$$

sqFootage: There is evidence to suggest that sqFootage is associated with mean sale price and that this association depends on zip code (p-value 0.0066). We estimate every 1 sq. ft. increase is associated with a -$73 increase for zip=75224 or a -73+350 = $277 increase for zip=75225 in mean sale price, given the other explanatory variables in the model

Multiple Regression: Interpretation

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	-288934.0937	B	220126.6280	-1.31	0.2017	-743253.1246	165384.9372
sqFootage	-72.9738	B	113.0904	-0.65	0.5249	-306.3808	160.4332
zipcode 75225	-681603.9196	B	321682.7404	-2.12	0.0446	-1345524.465	-17683.3745
zipcode 75224	0.0000	B	.	.	.	.	.
sqFootage*zipcode 75225	350.1357	B	117.7455	2.97	0.0066	107.1208	593.1505
sqFootage*zipcode 75224	0.0000	B	.	.	.	.	.
nBathrooms	306893.9357		84907.2381	3.61	0.0014	131654.0092	482133.8622

$$\mu\{salePrice|ft^2, zipcode, nBaths\}$$
$$= \beta_0 + \beta_1 ft^2 + \beta_2 zipcode + \beta_3 ft^2 * zipcode + \beta_4 nBaths$$

nBathrooms: There is evidence to suggest that nBathrooms is associated with mean sale price (p-value 0.0014). We estimate every additional bathroom is associated with a $306,893 increase in mean sale price, given the other explanatory variables in the model (95% confidence interval of [$131654,$482133]).

Multiple Regression: Interpretation

Suppose I'm a homeowner in one of these two zip codes

1. I read this report and decide that by converting a portion of my existing house to add an additional bathroom I will increase my house's sale price.

2. I read this report and decide if I add on to my house to add an additional bathroom, I will cause my house's sale price to increase between [$131654,$482133]

What interpretation errors am I making in each case?

$$\mu\{salePrice|ft^2, zipcode, nBaths\}$$
$$= \beta_0 + \beta_1 ft^2 + \beta_2 zipcode + \beta_3 ft^2 * zipcode + \beta_4 nBaths$$

nBathrooms: There is evidence to suggest that nBathrooms is associated with mean sale price (p-value 0.0014). We estimate every additional bathroom is associated with a $306,893 increase in mean sale price, given the other explanatory variables in the model (95% confidence interval of [$131654,$482133]).