

Multiple Regression: A Model for the Mean

REVIEW THE NOTATION AND CONCEPTS OF
REGRESSION

Terminology & Goals

- There is a lot of notation and vocabulary involved in linear regression
- The core goal behind simple linear regression is estimate a relationship between
 - an input known as the EXPLANATORY VARIABLE and..
 - another measurement known as the RESPONSE VARIABLE
- **Etymology:**
 - **Linear:** We model this relationship as linear for simplicity and interpretability. We must check this modeling assumption.
 - **Regression:** Charles Darwin's cousin, Francis Galton, studied heritability of traits. He found that extra tall people tend to have less tall offspring and extra short people tend to have less short offspring
→ Regression

Differing Terminology & Causation

You will hear or read about alternative terms for “explanatory” and “response” variables:

Explanatory:

Independent variable
Exogenous variable
Predictor variable
Covariate
Feature
Input

Response:

Dependent variable
Endogenous variable
Supervisor
Output

It is dangerously tempting to interpret regression as X **CAUSING** Y even with an observational study → use “association”

Notation for the Mean

- Y is the response variable
- x_1, \dots, x_p are the explanatory variables
- $\mu\{Y|X\}$ is the “mean of Y as a function of $X = (x_1, \dots, x_p)$ ”

This gets estimated from data:

$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$

via solving for the “least squares” solution:

$$\hat{\mu}\{Y|X\} = \text{minimizer}(\sum_{i=1}^n (Y_i - \mu\{Y_i|X_i\})^2)$$

over functions $\mu\{Y|X\} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p = \beta_0 + \sum_{j=1}^p \beta_j x_j$

The quality of this solution depends on whether the assumptions are reasonably met

Checking Assumptions

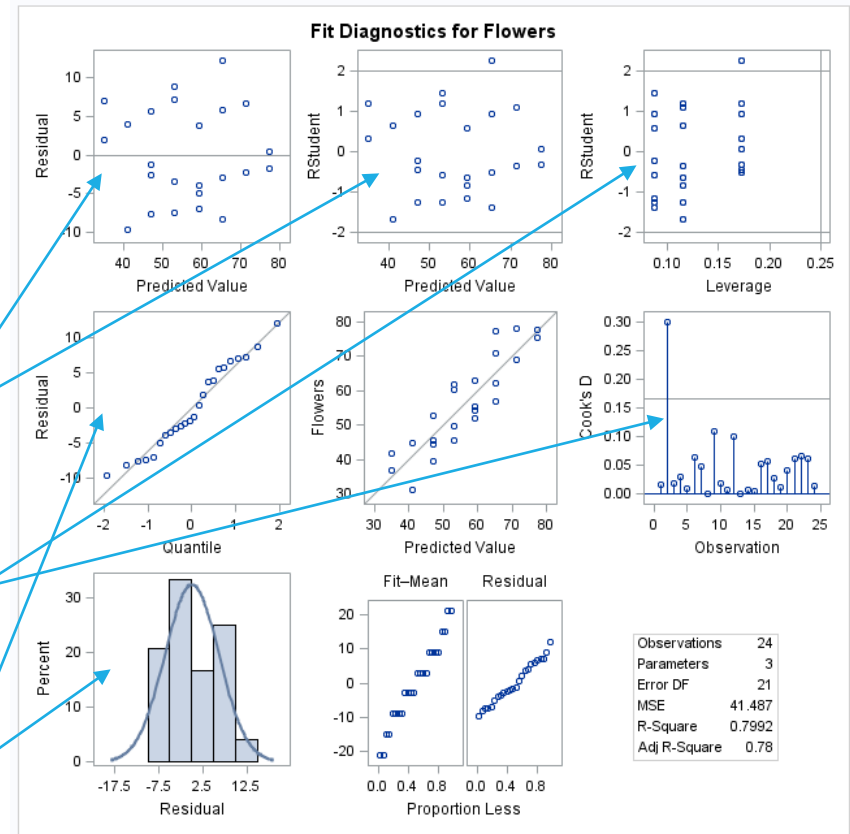
Residuals: No evidence of changing variance nor model misfit

(what does “Predicted Value” mean?)

Influence: No evidence of extreme or influential points

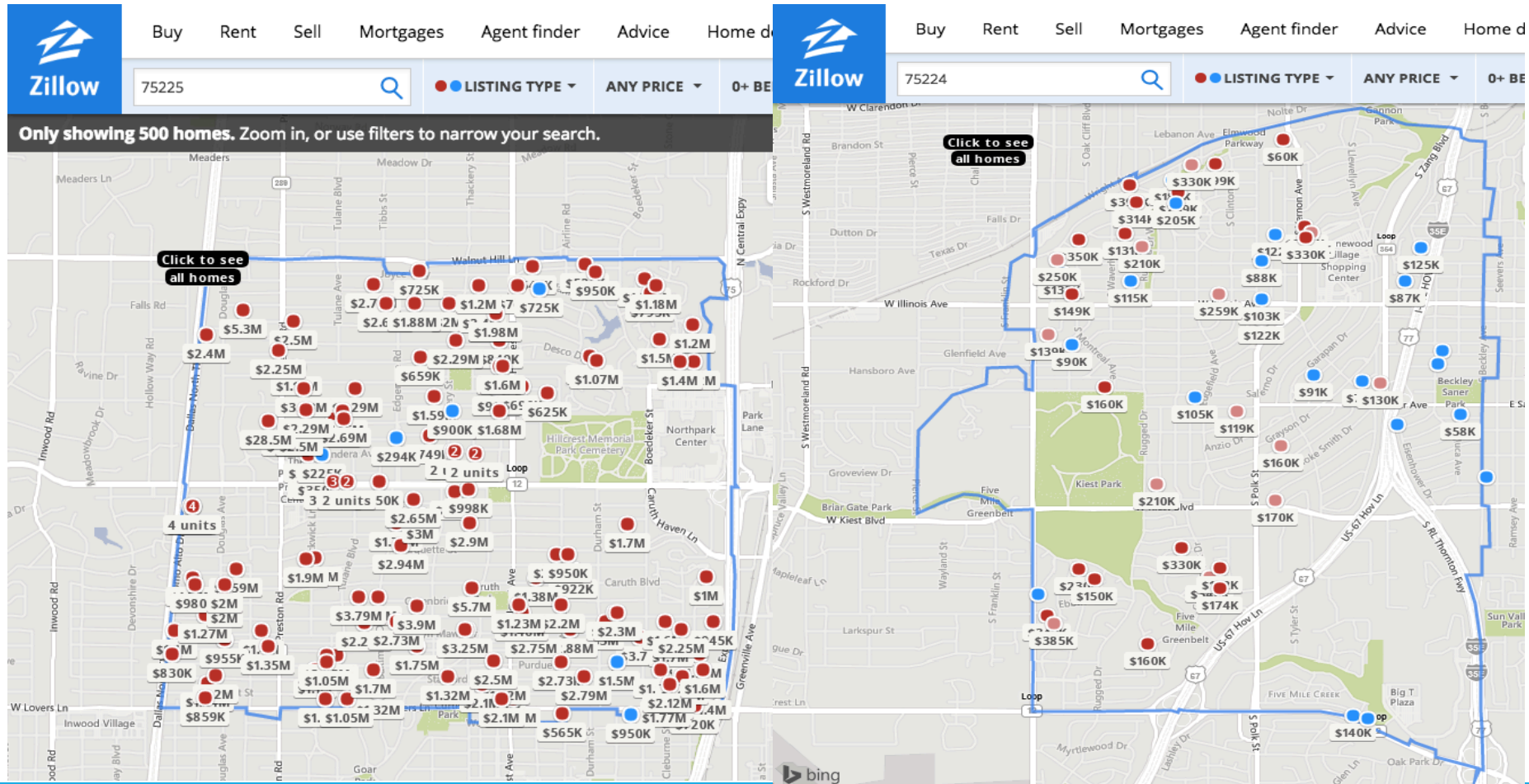
(High leverage + large residual)

Normality: No evidence of any issue



Review: Housing Data

Toward Building a Model for Housing Price



A Scientific Question

- It is very important to do statistics from the perspective of a scientist
- This means always keeping in mind the scientific question(s) of interest
- What are some scientific questions of interest for the housing data?
- Perhaps I'm a real estate company that has a few competing house designs:
 - How many bathrooms should I include?
 - How many bedrooms?
 - A few larger houses vs. many smaller houses?

A survey of housing data + multiple regression could be used

Converting a Scientific Question to Statistics

- For our survey, key questions:
 - What is a/the relevant population of interest?
 - Do we need/want to make causal conclusions?
- What is a/the relevant population of interest?

Perhaps the real estate company wants to draw conclusions about houses for sale in a particular area of Dallas during a particular period of time

Then a sample can be drawn with this population in mind

- Do we need/want to make causal conclusions?

In this case, an experiment establishing a causal relationship would be very hard to do (why?)

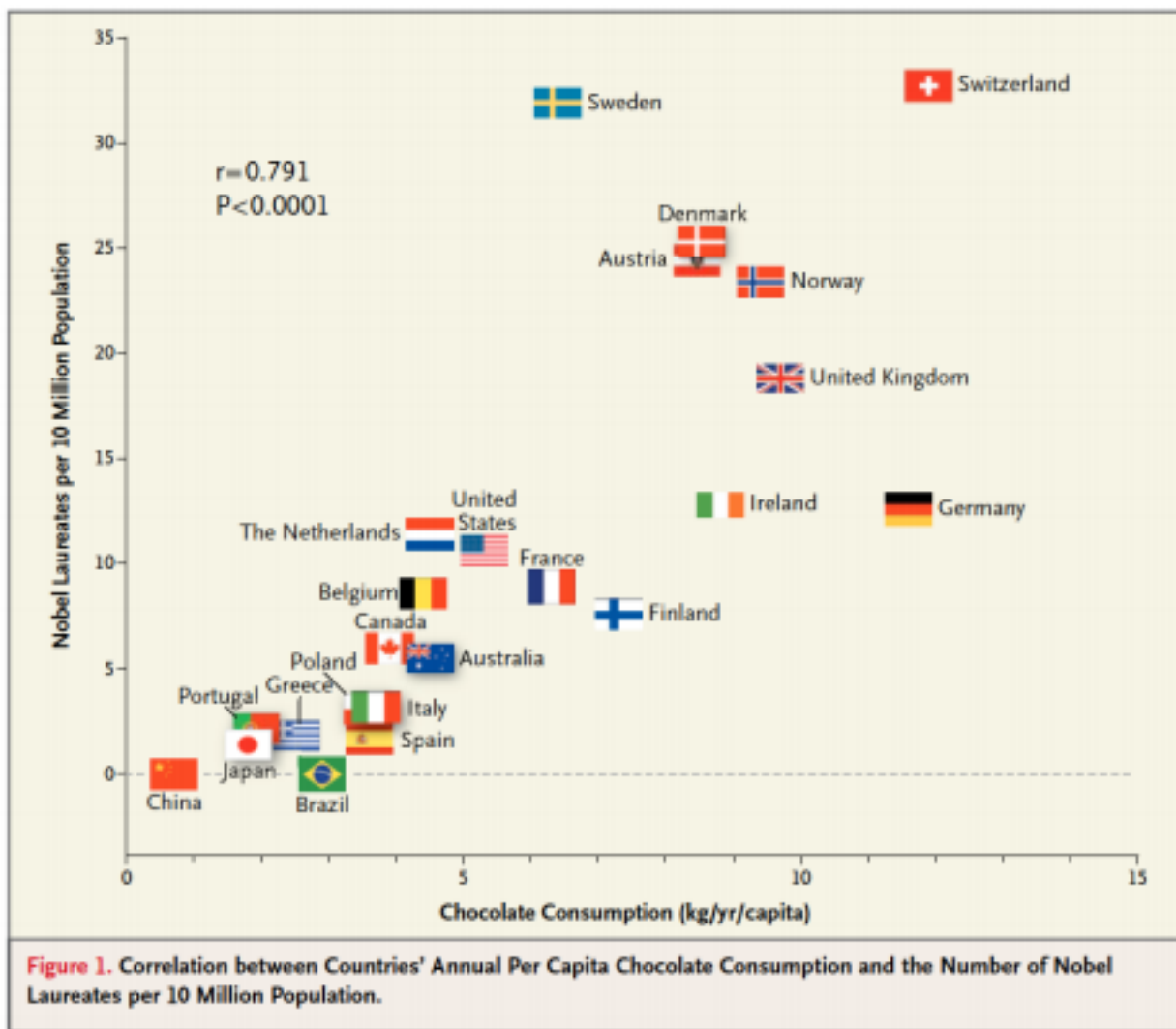
Hence, we will be forced to argue informally about possible confounders

Chocolate Study

Important:

- We can say that chocolate consumption is **predictive** of Nobel prizes
- We cannot say that changing chocolate consumption will **change** the number of Nobel prizes

A major reason for this distinction is the possibility of **CONFOUNDING VARIABLES**



Messerli (2012), New England Journal of Medicine

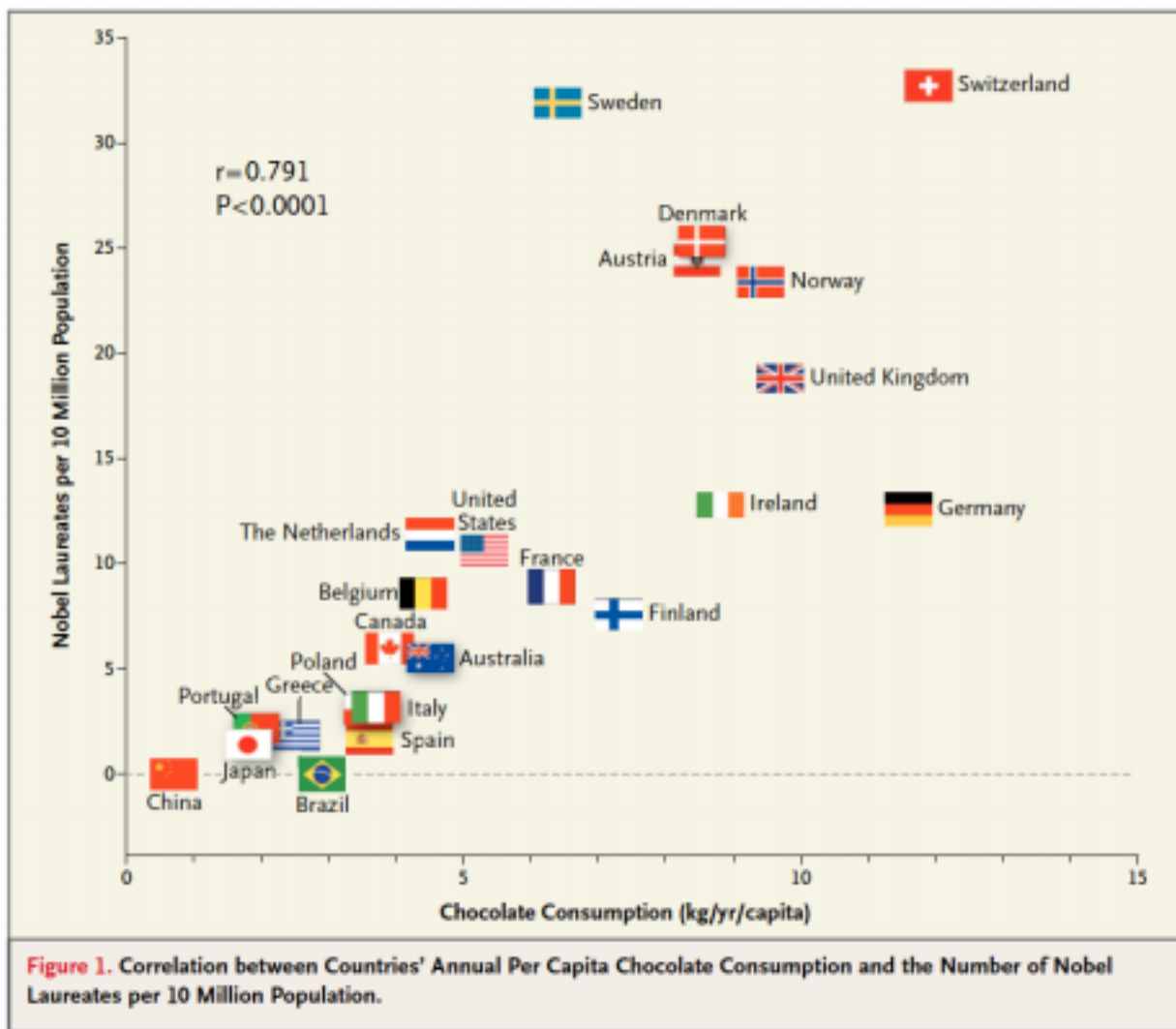
Chocolate Study

Important:

Prediction: Predict Y after **observing** X

Causation: Predict Y after **setting** X

(note that randomization breaks any confounding relationships and hence allows for causal inference)



Messerli (2012), New England Journal of Medicine

Chocolate makes you smarter, study suggests

People who eat chocolate at least once a week see their memory and abstract thinking improve, researchers say

Chocolate: 10 health reasons you should eat more of it

Why Chocolate Makes You Smarter: It's Proven!

10/29/2012 03:21 pm ET | Updated Dec 29, 2012


Wednesday 6 April 2016

 Life Newsletter

 Life Health Fo

Chocolate make you smarter, new research suggests



Patricia Murphy 
[EMAIL](#)

PUBLISHED
08/03/2016 | 14:07



Boosting Brain Power -- With Chocolate

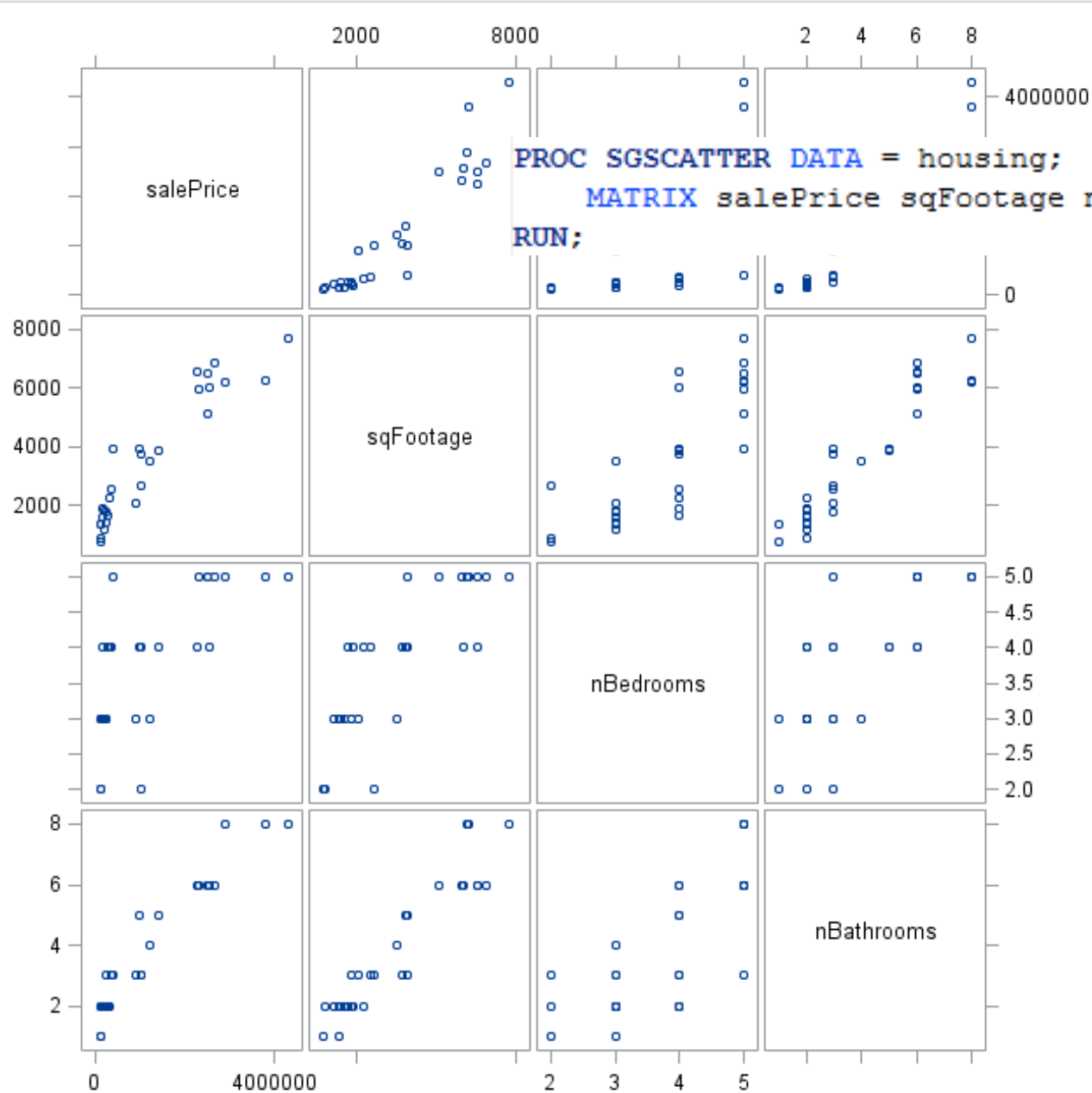
Date: February 22, 2007

Source: University of Nottingham

Summary: Eating chocolate could help to sharpen up the mind and give a short-term boost to cognitive skills, a University of Nottingham expert has found.

Converting a Scientific Question to Statistics

- Let's consider: How many bathrooms should I include?
- We want to estimate a relationship between bathrooms and sales price
- We know that we cannot formally control for confounders via sampling
- Even if sales price is larger if a house has an extra bathroom, is it due to:
 - the house being larger,
 - the house having more bedrooms (since baths and bedrooms are correlated)
 - the house being in a better neighborhood
 -
- We can informally control for confounders by estimating a relationship that “fixes” the levels of some (but not all) of these confounders



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

“Scatterplot matrix”
Or
“Draftsmen plot”
Or
“Pairs plot”

Converting a Scientific Question to Statistics

- Perhaps we conduct a survey and we gather:
 - Sale price (\$)
 - Square Footage (ft^2)
 - Zip code (we randomly select two zipcodes)
 - Number of bedrooms
 - Number of bathrooms
- A multiple regression gets at informally addressing some of the concerns about confounding:

$$\begin{aligned} &\mu\{salePrice|ft^2, zipcode, nBedrooms, nBaths\} \\ &= \beta_0 + \beta_1 ft^2 + \beta_2 zipcode + \beta_3 nBedrooms + \beta_4 nBaths \end{aligned}$$

- The house having more bathrooms for fixed sq. ft. and nBaths $\rightarrow \beta_4$

Specifying a Multiple Regression in SAS

- We can do this in at least two ways:

```
DATA housing;  
  SET housing;  
  IF zipCode = 75224 THEN zipCode_ind = 0;  
  IF zipCode = 75225 THEN zipCode_ind = 1;  
RUN;
```

(We will use this output unless otherwise indicated)



```
PROC GLM DATA = housing PLOTS=all;  
  CLASS zipCode (ref = '75224');  
  MODEL salePrice = sqFootage zipCode nBedrooms nBathrooms / SOLUTION CLPARM;  
RUN;
```

```
PROC REG DATA = housing PLOTS=all;  
  MODEL salePrice = sqFootage zipCode_ind nBedrooms nBathrooms;  
RUN;
```

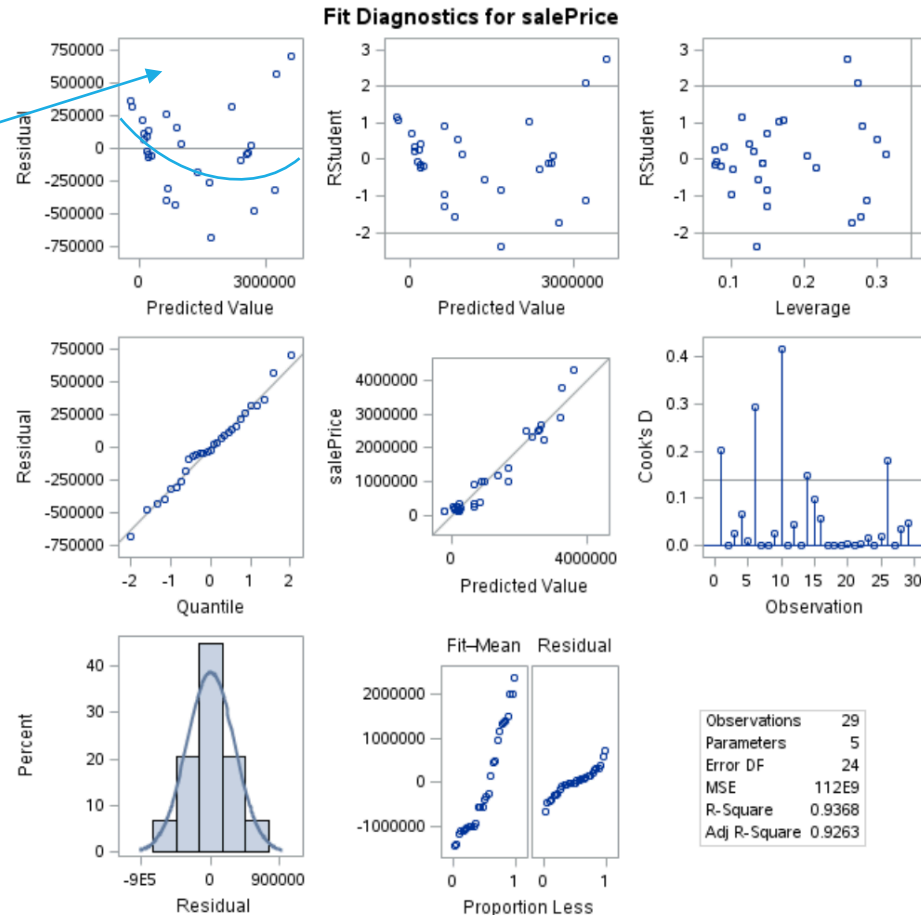
$$\begin{aligned} &\mu\{\text{salePrice} | \text{ft}^2, \text{zipcode}, \text{nBedrooms}, \text{nBaths}\} \\ &= \beta_0 + \beta_1 \text{ft}^2 + \beta_2 \text{zipcode} + \beta_3 \text{nBedrooms} + \beta_4 \text{nBaths} \end{aligned}$$

Checking Assumptions

Only concern: moderate indication of a "smiley face" in the residuals

(we will return to this in a bit)

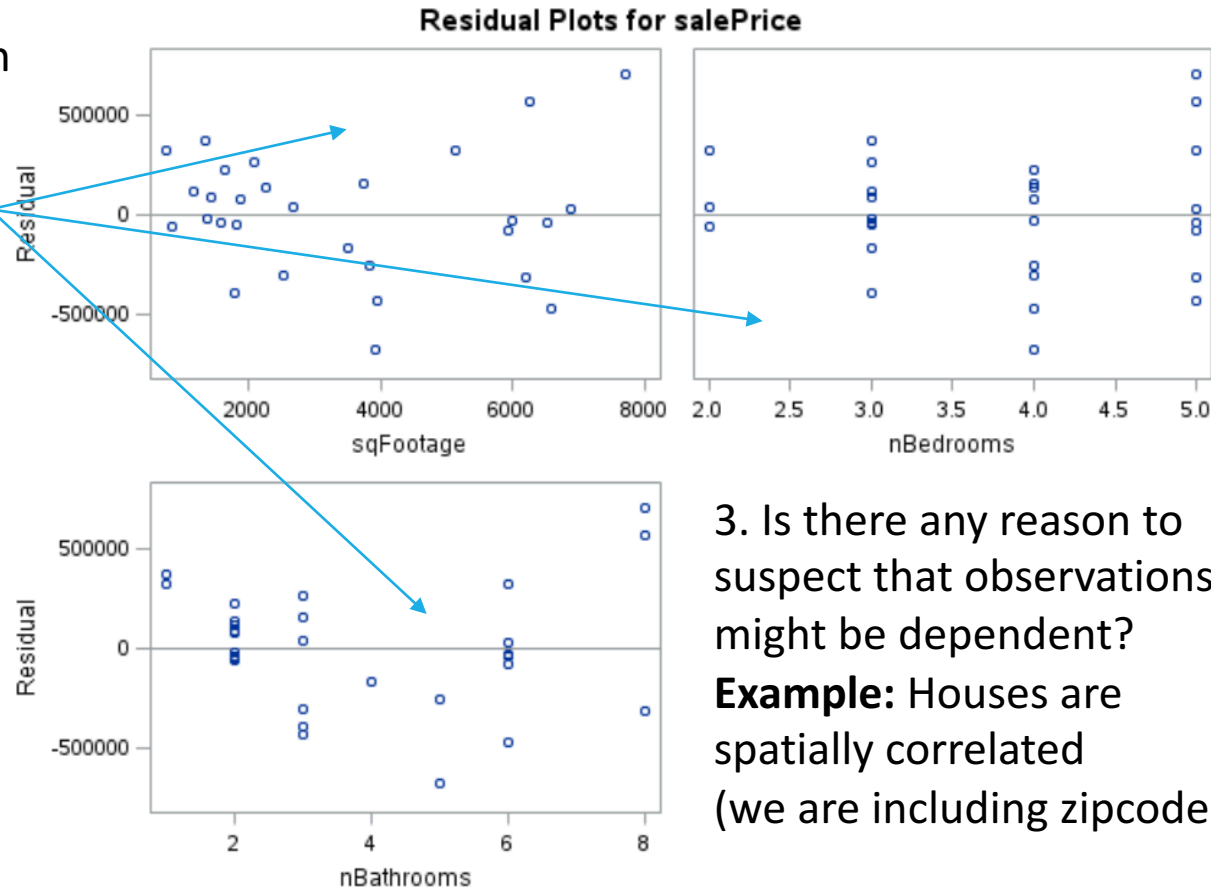
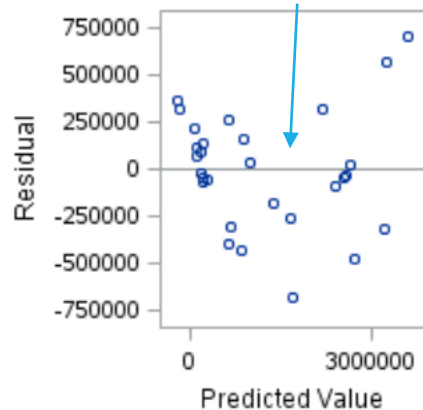
There is one assumption not included on this plot..



Assumption: Independence

Independence gets checked in three ways:

1. Do residuals appear to be related to an explanatory variable?
2. Do residuals appear to be related to each other?



3. Is there any reason to suspect that observations might be dependent?
Example: Houses are spatially correlated (we are including zipcode)

Implication of independence violation: Underestimate variability (e.g. p-values too small)

Converting a Scientific Question to Statistics

- Suppose we are satisfied that the assumptions are satisfactorily met
- We have an estimate of the model:

$$\begin{aligned} &\mu\{salePrice|ft^2, zipcode, nBedrooms, nBaths\} \\ &= \beta_0 + \beta_1 ft^2 + \beta_2 zipcode + \beta_3 nBedrooms + \beta_4 nBaths \end{aligned}$$

(Once the model is estimated, the parameters are written with a “hat” to indicate functions of the data)

$$\begin{aligned} &\hat{\mu}\{salePrice|ft^2, zipcode, nBedrooms, nBaths\} \\ &= \hat{\beta}_0 + \hat{\beta}_1 ft^2 + \hat{\beta}_2 zipcode + \hat{\beta}_3 nBedrooms + \hat{\beta}_4 nBaths \end{aligned}$$

Parameter	Estimate		Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	-413793.0938	B	316100.7504	-1.31	0.2029	-1066192.978	238606.7902
sqFootage	252.8059		108.2575	2.34	0.0282	29.3733	476.2384
zipcode 75225	-77819.1511	B	241444.0896	-0.32	0.7500	-576135.2603	420496.9581
zipcode 75224	0.0000	B
nBedrooms	-181430.9105		122758.9588	-1.48	0.1524	-434792.9490	71931.1280
nBathrooms	380639.9690		92117.7525	4.13	0.0004	190518.2722	570761.6658

(See pages 276-278 for a discussion on this output)

- What we do from here depends a lot on the scientific question at hand

Converting a Scientific Question to Statistics

Parameter	Estimate		Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	-413793.0938	B	316100.7504	-1.31	0.2029	-1066192.978	238606.7902
sqFootage	252.8059		108.2575	2.34	0.0282	29.3733	476.2384
zipcode 75225	-77819.1511	B	241444.0896	-0.32	0.7500	-576135.2603	420496.9581
zipcode 75224	0.0000	B
nBedrooms	-181430.9105		122758.9588	-1.48	0.1524	-434792.9490	71931.1280
nBathrooms	380639.9690		92117.7525	4.13	0.0004	190518.2722	570761.6658

(A one-sided test is perhaps more appropriate for this question)

- Our question: How many bathrooms should I include?
- Preliminary question: Does number of bathrooms seem to be related to sales price **given the other terms in the model?**

$$H_0: \beta_4 = 0$$

$$H_A: \beta_4 \neq 0$$

“There is evidence that the number of bathrooms is associated with mean sales price given the other terms in the model. A range of plausible values are a difference in mean sales price between \$190518 to \$570561 for houses with the same sq. ft, # of bedrooms, and zipcode but 1 bathroom difference.”

(See pages 244-246 & 278-289 in book for an discussion on interpreting regression output)

Converting a Scientific Question to Statistics

Parameter	Estimate		Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	-413793.0938	B	316100.7504	-1.31	0.2029	-1066192.978	238606.7902
sqFootage	252.8059		108.2575	2.34	0.0282	29.3733	476.2384
zipcode 75225	-77819.1511	B	241444.0896	-0.32	0.7500	-576135.2603	420496.9581
zipcode 75224	0.0000	B
nBedrooms	-181430.9105		122758.9588	-1.48	0.1524	-434792.9490	71931.1280
nBathrooms	380639.9690		92117.7525	4.13	0.0004	190518.2722	570761.6658

- Now for a more nuanced question: Should we keep all the terms in the model?
- The answer to this question depends on our goal and the types of errors we are willing to make

Multiple Regression and “Wrong” Explanatory Variables

There are roughly four possible (unknown) outcomes:

1. Model is correctly specified. All relevant explanatory variables are included
 2. Model is under specified. Some important explanatory variables are omitted
 3. Model includes unimportant explanatory variables
 4. Model is over specified. Some redundant (or nearly redundant) explanatory variables are included
1. Ideal case. Estimates are unbiased
 2. Estimates are biased and we are overestimating σ . We can get incorrect inferences due to confounding
 3. Estimates are unbiased. However, the inclusion of irrelevant parameters decreases power
 4. We'll return to this during multicollinearity

Converting a Scientific Question to Statistics

Parameter	Estimate		Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	-413793.0938	B	316100.7504	-1.31	0.2029	-1066192.978	238606.7902
sqFootage	252.8059		108.2575	2.34	0.0282	29.3733	476.2384
zipcode 75225	-77819.1511	B	241444.0896	-0.32	0.7500	-576135.2603	420496.9581
zipcode 75224	0.0000	B
nBedrooms	-181430.9105		122758.9588	-1.48	0.1524	-434792.9490	71931.1280
nBathrooms	380639.9690		92117.7525	4.13	0.0004	190518.2722	570761.6658

- So, should be keep all the terms in the model?
- Going by the four considerations from the previous slides, I would leave all the explanatory variables in the model as:
 - I am concerned about confounding as my scientific question of interest is directly about inference (as opposed to prediction)
 - It is more conservative to leave them in there (the degrees of freedom for the t-test is smaller)

Housing Data: Revisiting the Regression Fit

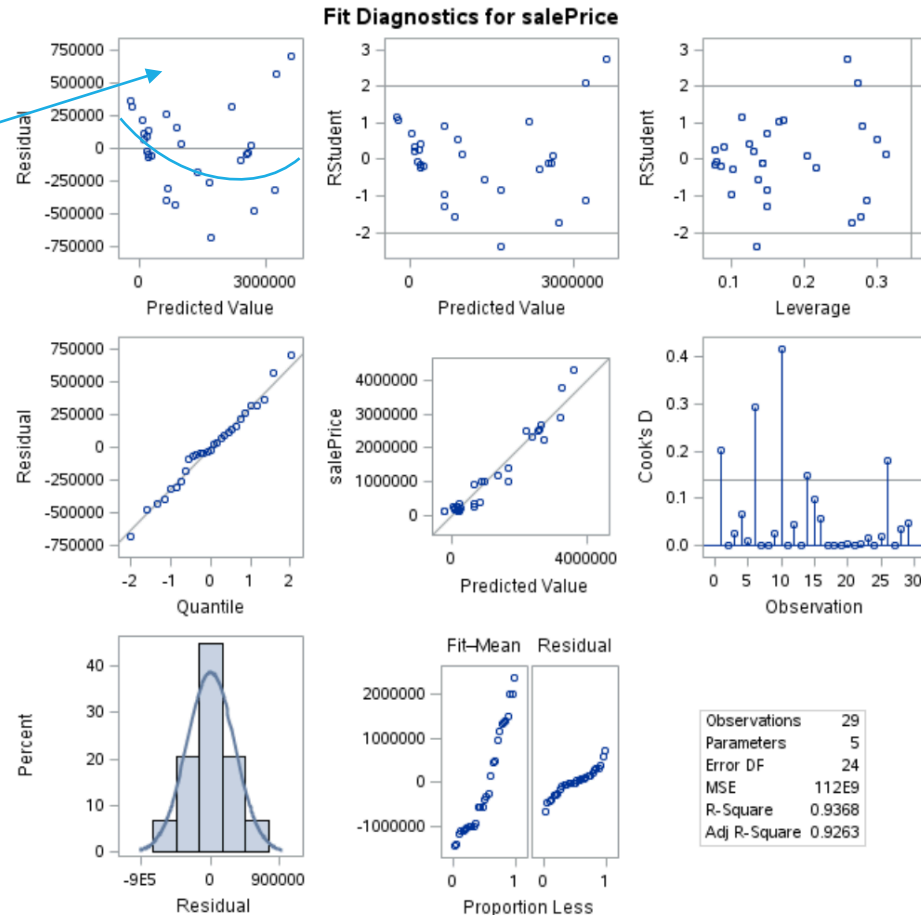
Re-Checking Assumptions

Only concern: moderate indication of a "smiley face" in the residuals

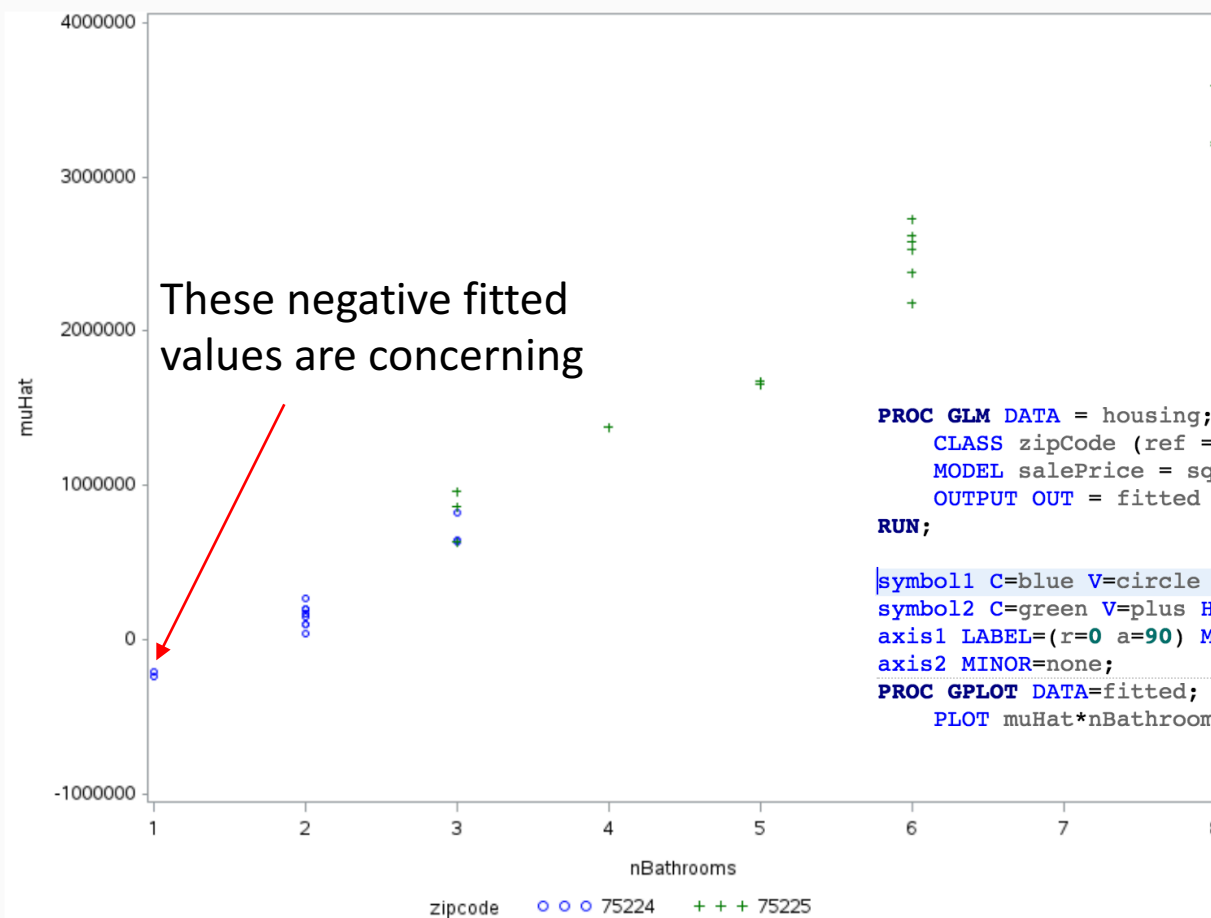
(we will return to this in a bit)

Let's see if we can address this pattern in the residuals

First, let's look at a plot of the fitted values vs. nBathrooms



Fitted Values vs. nBathrooms

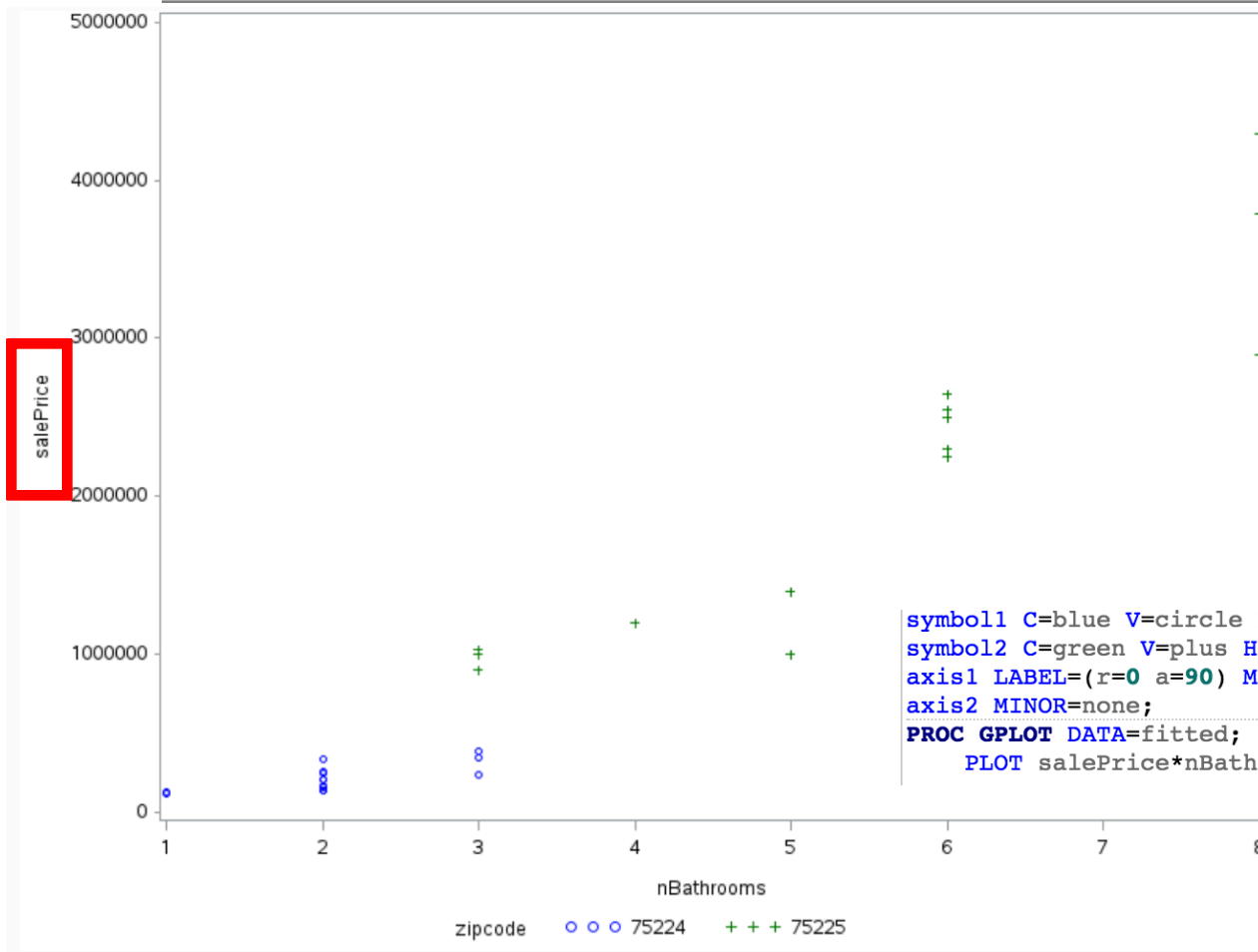


Definition: The **FITTED VALUES** are the estimated mean at the data: $\hat{\mu}\{Y_i|X_i\}$

```
PROC GLM DATA = housing;
  CLASS zipCode (ref = '75224');
  MODEL salePrice = sqFootage zipCode nBedrooms nBathrooms
  OUTPUT OUT = fitted PREDICTED = muHat;
RUN;

symbol1 C=blue V=circle H=0.8;
symbol2 C=green V=plus H=0.8;
axis1 LABEL=(r=0 a=90) MINOR=none;
axis2 MINOR=none;
PROC GPLOT DATA=fitted;
  PLOT muHat*nBathrooms=zipcode / VAXIS=axis1 HAXIS=axis2;
```

Sales Price vs. nBathrooms



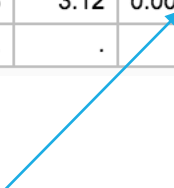
There is some indication here that a different slope is required for the different zipcodes

```
symbol1 C=blue V=circle H=0.8;  
symbol2 C=green V=plus H=0.8;  
axis1 LABEL=(r=0 a=90) MINOR=none;  
axis2 MINOR=none;  
PROC GPLOT DATA=fitted;  
    PLOT salePrice*nBathrooms=zipcode / VAXIS=axis1 HAXIS=axis2;
```

Adding an interaction term

```
PROC GLM DATA = housing PLOTS=all;  
  CLASS zipCode (ref = '75224');  
  MODEL salePrice = sqFootage zipCode nBedrooms nBathrooms nBathrooms*zipCode/ SOLUTION CLPARM;  
RUN;
```

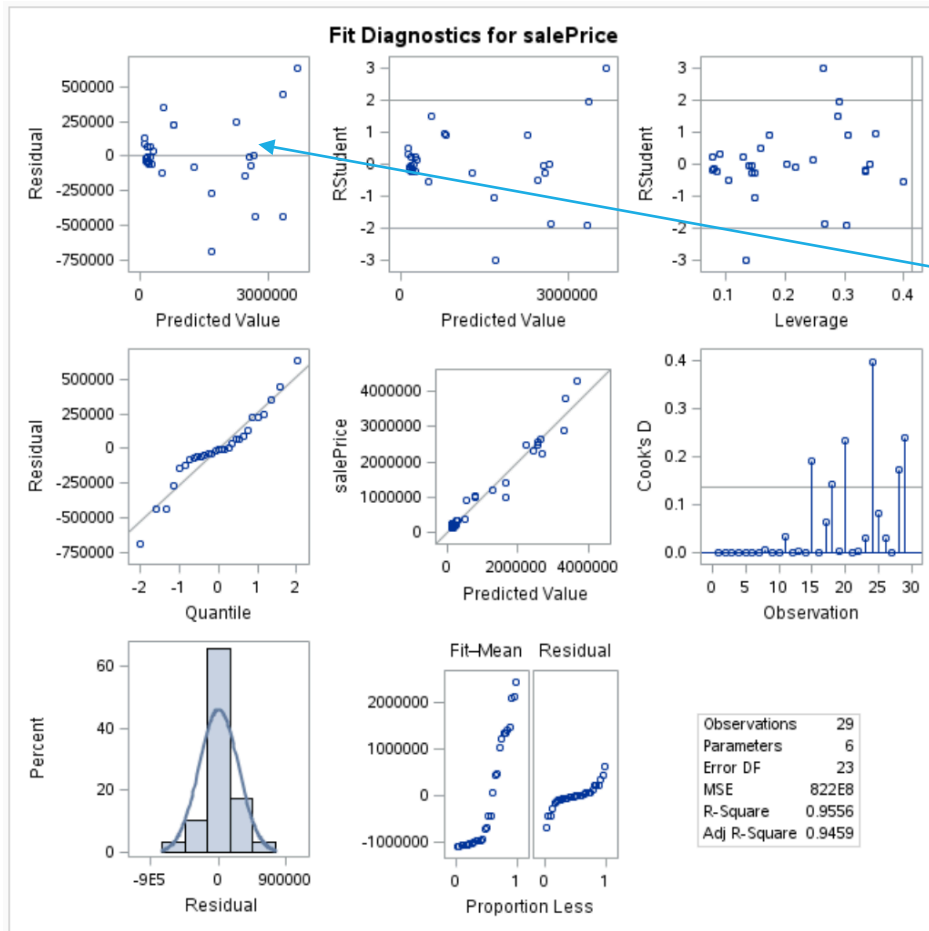
Parameter	Estimate		Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	219885.216	B	338511.8432	0.65	0.5224	-480379.884	920150.317
sqFootage	221.119		93.2692	2.37	0.0265	28.177	414.061
zipcode 75225	-1075799.128	B	381054.9625	-2.82	0.0096	-1864071.376	-287526.880
zipcode 75224	0.000	B
nBedrooms	-105504.679		107916.4652	-0.98	0.3384	-328746.896	117737.538
nBathrooms	-19026.733	B	150513.2001	-0.13	0.9005	-330387.010	292333.543
nBathrooms*zipcode 75225	437258.280	B	140237.3223	3.12	0.0048	147155.276	727361.284
nBathrooms*zipcode 75224	0.000	B



Evidence for a nonzero interaction term.

Reminder: Only look at p-value for interaction, not for main effects.

Adding an interaction term



There is moderate evidence of a lack of model fit

There is both a pattern in the residuals and evidence of increasing variance

In this case, a log transform can sometimes help

Analysis Detour: Inference with Interactions

Inference for Interactions

Our investigation of the residuals indicates poor model fit for the interaction model:

$$\begin{aligned} &\mu\{salePrice|ft^2, zipcode, nBedrooms, nBaths\} \\ &= \beta_0 + \beta_1 ft^2 + \beta_2 zipcode + \beta_3 nBedrooms + \beta_4 nBaths + \beta_5 nBaths * zipcode \end{aligned}$$

However, it is a good opportunity to discuss inference for interactions

As always, only attempt inference with a model that satisfies the assumptions

In this interaction model, there are two parameters for nBaths:

β_4 for zipcode = 75224

$\beta_4 + \beta_5$ for zipcode = 75225

Inference for these parameters involves testing for linear combinations

Inference for Interactions

We've already discussed a **test** for the interaction of nBathrooms and zipcode

Parameter	Estimate		Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	219885.216	B	338511.8432	0.65	0.5224	-480379.884	920150.317
sqFootage	221.119		93.2692	2.37	0.0265	28.177	414.061
zipcode 75225	-1075799.128	B	381054.9625	-2.82	0.0096	-1864071.376	-287526.880
zipcode 75224	0.000	B
nBedrooms	-105504.679		107916.4652	-0.98	0.3384	-328746.896	117737.538
nBathrooms	-19026.733	B	150513.2001	-0.13	0.9005	-330387.010	292333.543
nBathrooms*zipcode 75225	437258.280	B	140237.3223	3.12	0.0048	147155.276	727361.284
nBathrooms*zipcode 75224	0.000	B

When the categorical variable has more than two levels, we need to do an “extra sums of squares F test” (Chapter 10.3.2 in book and the next set of lecture notes)

We can find a confidence interval for nBathrooms when zipcode = 75224 on the default output (the parameter is β_4)

However, a confidence interval for nBathrooms when zipcode = 75225 requires more work (the parameter is $\beta_4 + \beta_5$)

Inference for Interactions: Linear Combinations

Looking at the model: $\mu\{salePrice|ft^2, zipcode, nBedrooms, nBaths\} = \beta_0 + \beta_1 ft^2 + \beta_2 zipcode + \beta_3 nBedrooms + \beta_4 nBaths + \beta_5 nBaths * zipcode$

We see that:

$$\beta_4 + \beta_5 = 0 * \beta_0 + 0 * \beta_1 + 0 * \beta_2 + 0 * \beta_3 + 1 * \beta_4 + 1 * \beta_5$$

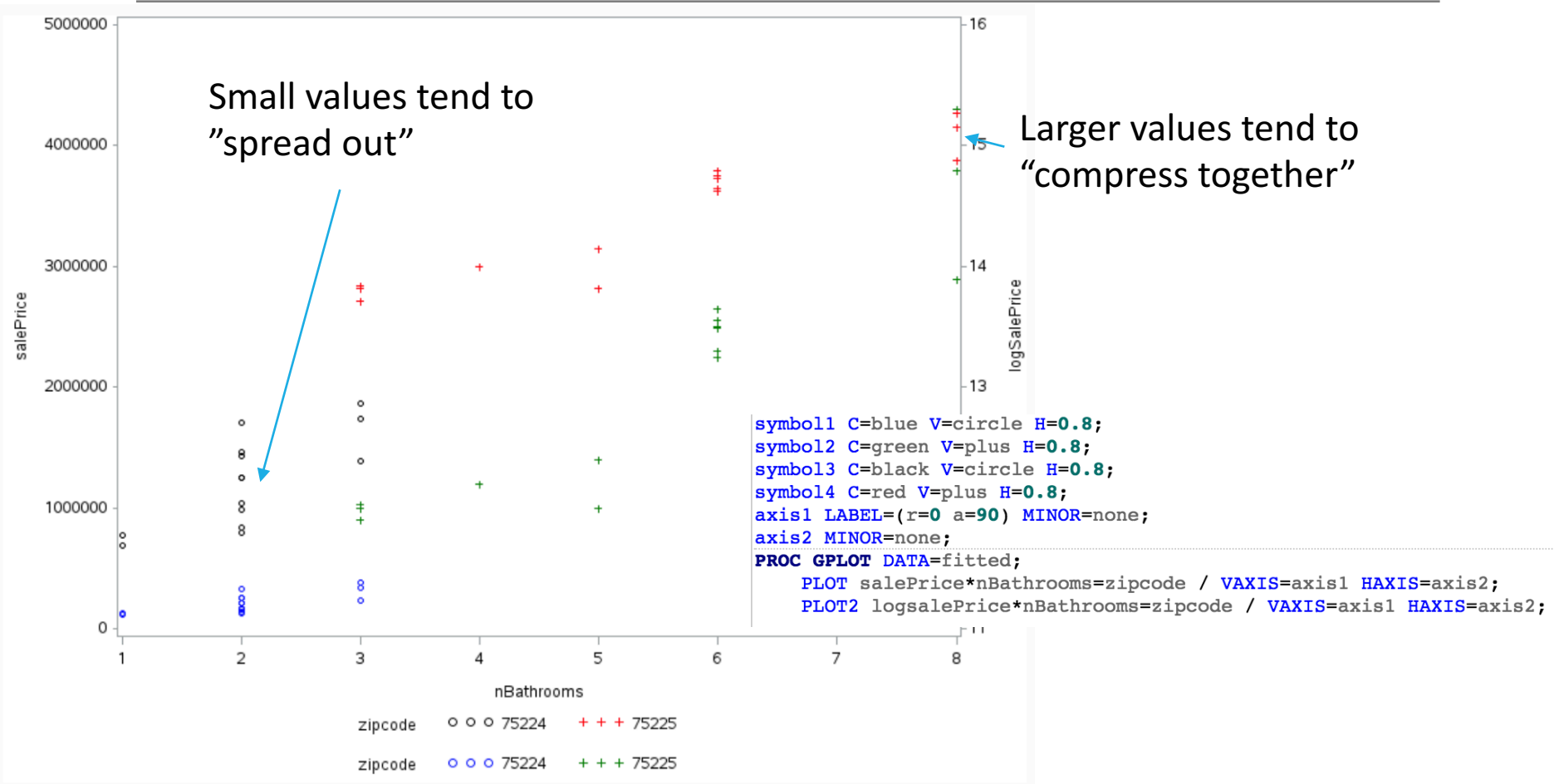
```
PROC GLM DATA = housing;
CLASS zipCode (ref = '75224');
MODEL salePrice = sqFootage zipCode nBedrooms nBathrooms nBathrooms*zipCode/ SOLUTION CLPARM;
ESTIMATE 'nBathrooms effect for zip = 75225' sqFootage 0 zipCode 0 nBedrooms 0 nBathrooms 1 nBathrooms*zipCode 1;
RUN;
```

Parameter	Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits
nBathrooms effect for zip = 75225	418231.547	79807.3899	5.24	<.0001	253137.383 583325.711

Parameter	Estimate		Standard Error	t Value	Pr > t	95% Confidence Limits
Intercept	219885.216	B	338511.8432	0.65	0.5224	-480379.884 920150.317
sqFootage	221.119		93.2692	2.37	0.0265	28.177 414.061
zipcode 75225	-1075799.128	B	381054.9625	-2.82	0.0096	-1864071.376 -287526.880
zipcode 75224	0.000	B
nBedrooms	-105504.679		107916.4652	-0.98	0.3384	-328746.896 117737.538
nBathrooms	-19026.733	B	150513.2001	-0.13	0.9005	-330387.010 292333.543
nBathrooms*zipcode 75225	437258.280	B	140237.3223	3.12	0.0048	147155.276 727361.284
nBathrooms*zipcode 75224	0.000	B

Return to Regular Analysis: Transformed Response

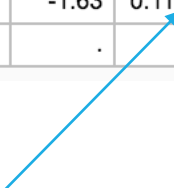
Considering a Log Transformation



Using a Transformed Response

```
PROC GLM DATA = housing PLOTS=all;  
  CLASS zipCode (ref = '75224');  
  MODEL logSalePrice = sqFootage zipCode nBedrooms nBathrooms nBathrooms*zipCode/ SOLUTION CLPARM;  
  OUTPUT OUT = fitted PREDICTED = muHat;  
RUN;
```

Parameter	Estimate		Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	11.15318264	B	0.22650050	49.24	<.0001	10.68463065	11.62173462
sqFootage	0.00014851		0.00006241	2.38	0.0260	0.00001941	0.00027761
zipcode 75225	1.55709971	B	0.25496638	6.11	<.0001	1.02966157	2.08453786
zipcode 75224	0.00000000	B
nBedrooms	0.07642440		0.07220762	1.06	0.3009	-0.07294843	0.22579724
nBathrooms	0.27333994	B	0.10070937	2.71	0.0124	0.06500673	0.48167315
nBathrooms*zipcode 75225	-0.15280988	B	0.09383371	-1.63	0.1170	-0.34691970	0.04129995
nBathrooms*zipcode 75224	0.00000000	B



No evidence for a nonzero interaction term. We should consider a simpler model without an interaction

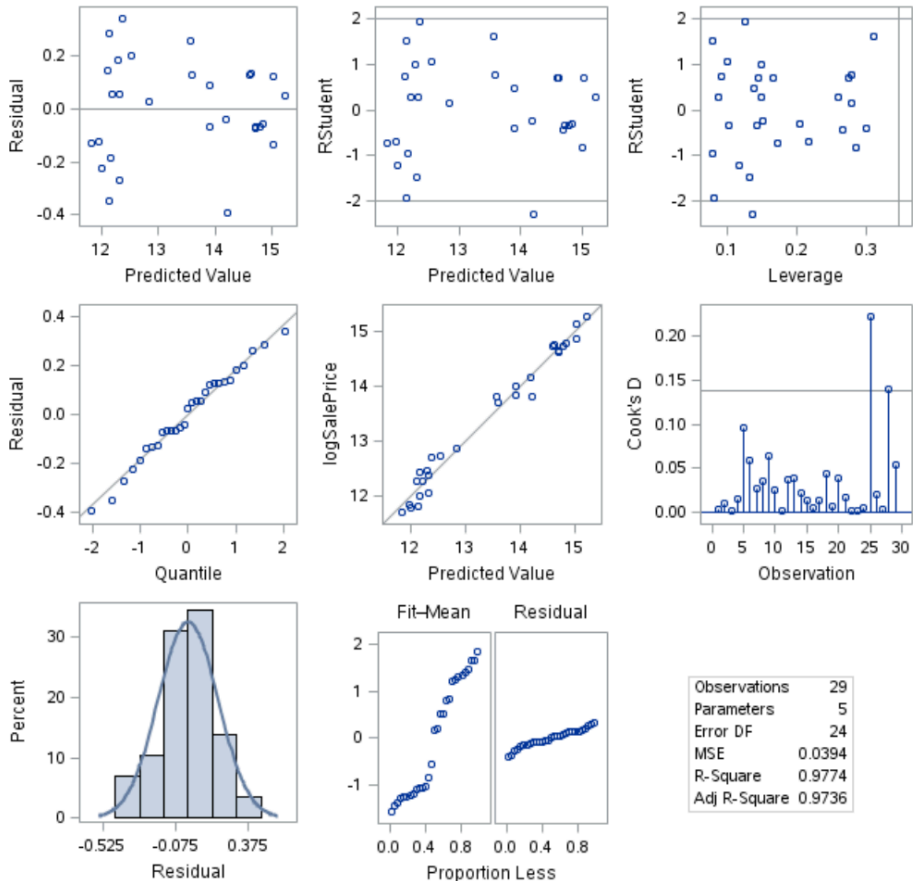
Using a Transformed Response: Remove Interaction

```
PROC GLM DATA = housing PLOTS=all;  
  CLASS zipCode (ref = '75224');  
  MODEL logSalePrice = sqFootage zipCode nBedrooms nBathrooms / SOLUTION CLPARM;  
  OUTPUT OUT = fitted PREDICTED = muHat;  
RUN;
```

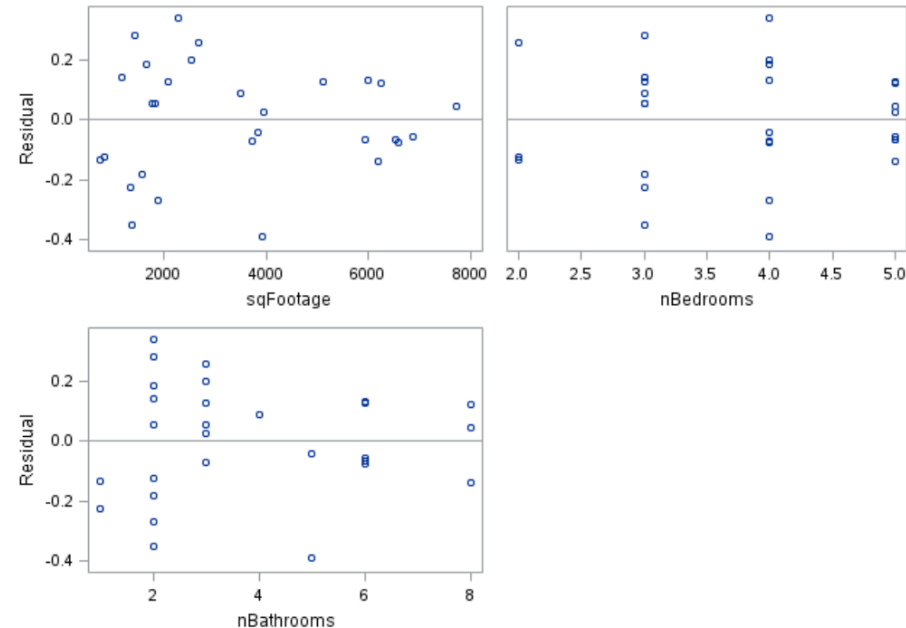
Parameter	Estimate		Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	11.37463597	B	0.18726780	60.74	<.0001	10.98813422	11.76113772
sqFootage	0.00013744		0.00006414	2.14	0.0425	0.00000507	0.00026981
zipcode 75225	1.20833285	B	0.14303890	8.45	<.0001	0.91311507	1.50355064
zipcode 75224	0.00000000	B
nBedrooms	0.10295855		0.07272618	1.42	0.1697	-0.04714091	0.25305801
nBathrooms	0.13366730		0.05457339	2.45	0.0220	0.02103336	0.24630124

Re-Checking the Assumptions

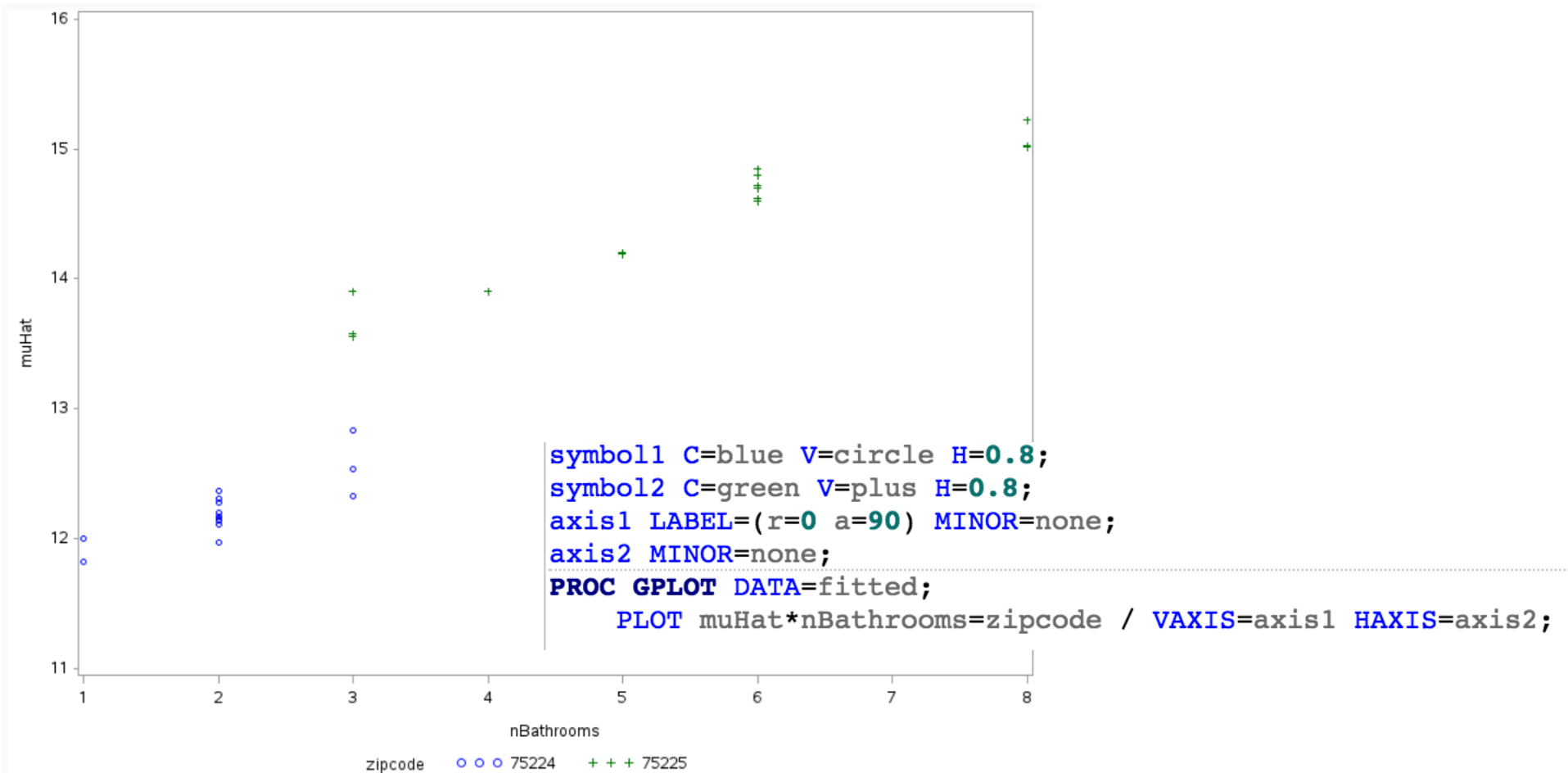
Fit Diagnostics for logSalePrice



Residual Plots for logSalePrice



One Last Look at Fitted Values



Interpreting the Model

```
PROC GLM DATA = housing PLOTS=all;  
  CLASS zipCode (ref = '75224');  
  MODEL logSalePrice = sqFootage zipCode nBedrooms nBathrooms / SOLUTION CLPARM;  
  OUTPUT OUT = fitted PREDICTED = muHat;  
RUN;
```

Parameter	Estimate		Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	11.37463597	B	0.18726780	60.74	<.0001	10.98813422	11.76113772
sqFootage	0.00013744		0.00006414	2.14	0.0425	0.00000507	0.00026981
zipcode 75225	1.20833285	B	0.14303890	8.45	<.0001	0.91311507	1.50355064
zipcode 75224	0.00000000	B
nBedrooms	0.10295855		0.07272618	1.42	0.1697	-0.04714091	0.25305801
nBathrooms	0.13366730		0.05457339	2.45	0.0220	0.02103336	0.24630124

“There is evidence that the number of bathrooms is associated with median sales price given the other terms in the model. A range of plausible values is a multiplicative change in median sales price between 1.02 and 1.28 for houses with the same sq. ft, # of bedrooms, and zipcode but 1 bathroom difference.”

(See pages 250-251 in book for an interpretation discussion)

Correlation and R-Square

Correlation and R-Square

The sample correlation coefficient describes the “degree of linear association between X and Y”

It is commonly denoted “r” and must be between -1 and 1

It is symmetric with respect to X and Y (unlike regression)

Often, we write $R^2 = r^2$ instead which is between 0 and 1

(Interpretation: R^2 is the proportion of the total variation in Y explained by it's least squares fit on X)

$$R^2 = (\text{Explained Sums of Squares})/(\text{Total Sums of Squares})$$

Correlation and R-Square

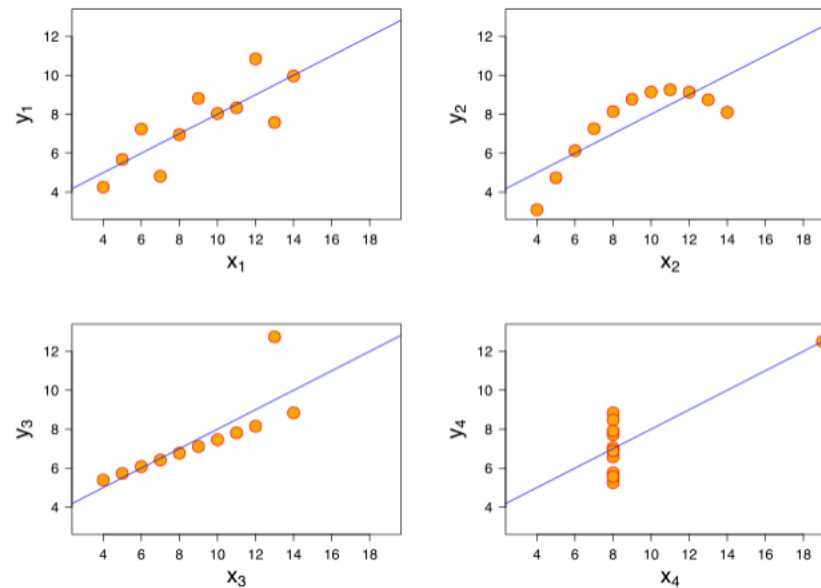
```
PROC GLM DATA = housing PLOTS=all;  
  CLASS zipCode (ref = '75224');  
  MODEL logSalePrice = sqFootage zipCode nBedrooms nBathrooms / SOLUTION CLPARM;  
  OUTPUT OUT = fitted PREDICTED = muHat;  
RUN;
```

R-Square	Coeff Var	Root MSE	logSalePrice Mean
0.977396	1.482217	0.198370	13.38336

- The discussed model has an R-Square of approx. 0.977
- The practical interpretation of R-Square is complicated:
 - $R\text{-Square} \approx 0$ could be poor model fit or a difficult problem
 - $R\text{-Square} \approx 1$ could be good model fit or too many parameters

Correlation and R-Square

It is important to not read too much into it as a single indicator of model fit
(see Chapter 10.4.1)



All these have same R^2 (from Wikipedia entry)