Week 5, February 10th 2017

New variables?

Assignment #2

- Introduction = why it matters
  - Many did not comment on the public health relevance of radon
  - This makes Sarah sad ☹
- Methods = what you did
  - Many reported some results in the methods section (i.e. unequal variance between groups)
  - Very few supplied regression equation
- Results = what you found
  - Many neglected to give the simplest possible summary of the data - how many observations in each group?
  - What effect are we measuring when we take the exponentiated values for the regression coefficients on log-transformed data?
- Discussion = what it MEANS
  - Many did not not give any possible reasons for why hypothesized relationships were not found
  - Strengths and limitations
  - Discussion section not of much value without critical thinking about the results section
- Be not dismayed!

Midterm

DETAILS:

You should be prepared to cover the following material:

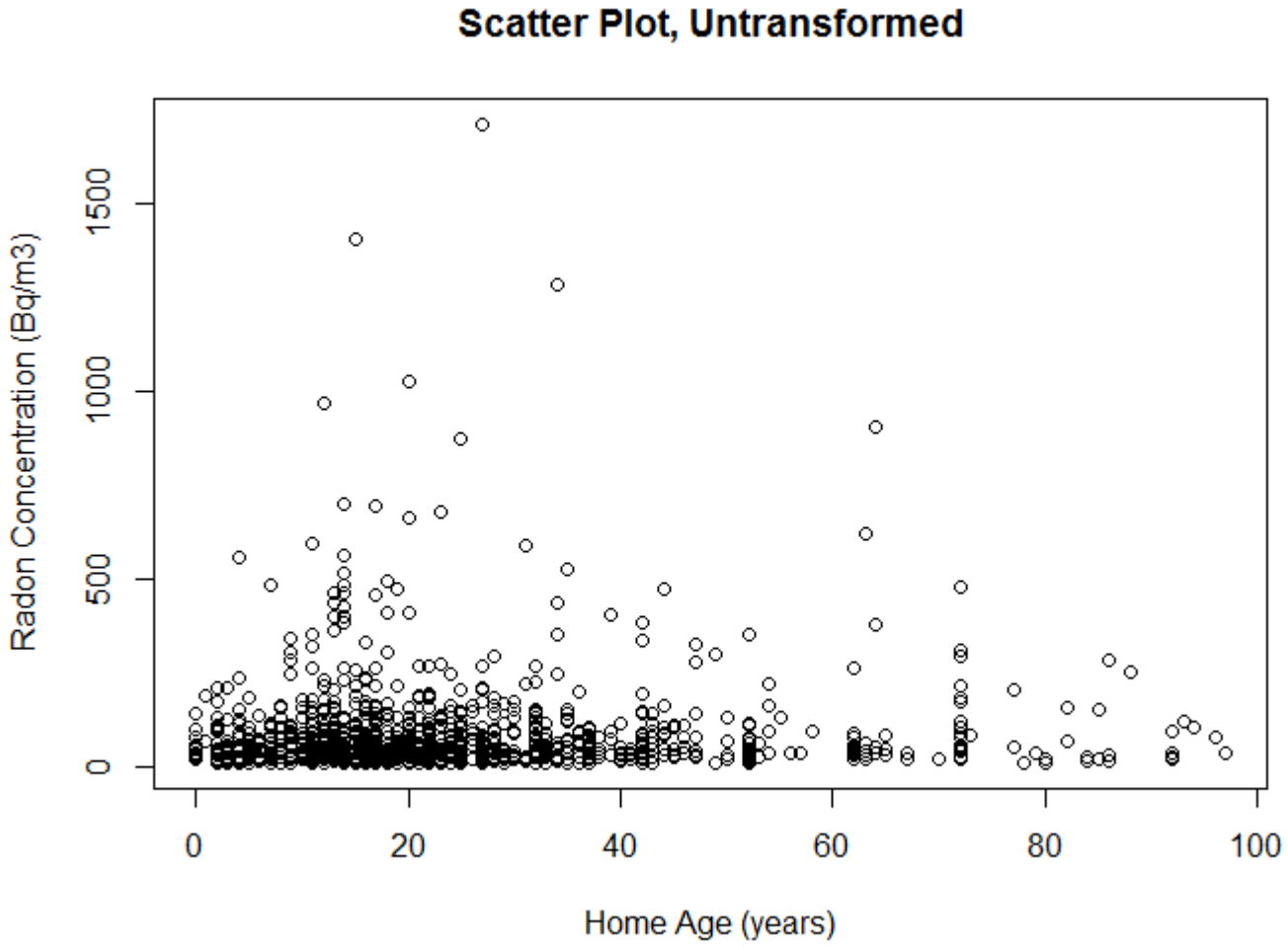
- Calculating the mean, standard deviation, geometric mean, and geometric standard deviation of a sample.
- The difference between the sample mean and the population mean.
- What to do with samples under the Limit of Detection.
- Properties of the perfect normal and log-normal distributions.
- Methods for visualizing, assessing, and quantifying the association between a continuous dependent variable and dichotomous, categorical, and continuous independent variables.
- Interpreting the results of linear regression and calculating confidence intervals.
- The meaning of statistical significance.
- Hypothesizing about the relationship between indoor radon concentrations and a variable that is not currently part of the dataset.

Continuous Variables

- What are they?
- Which variables in the radon dataset (as provided) are continuous?
- What hypotheses do we have about the association between these variables and radon concentrations?
- What other continuous variables would be nice to have in the dataset?
- How much data are we omitting due to missing information?

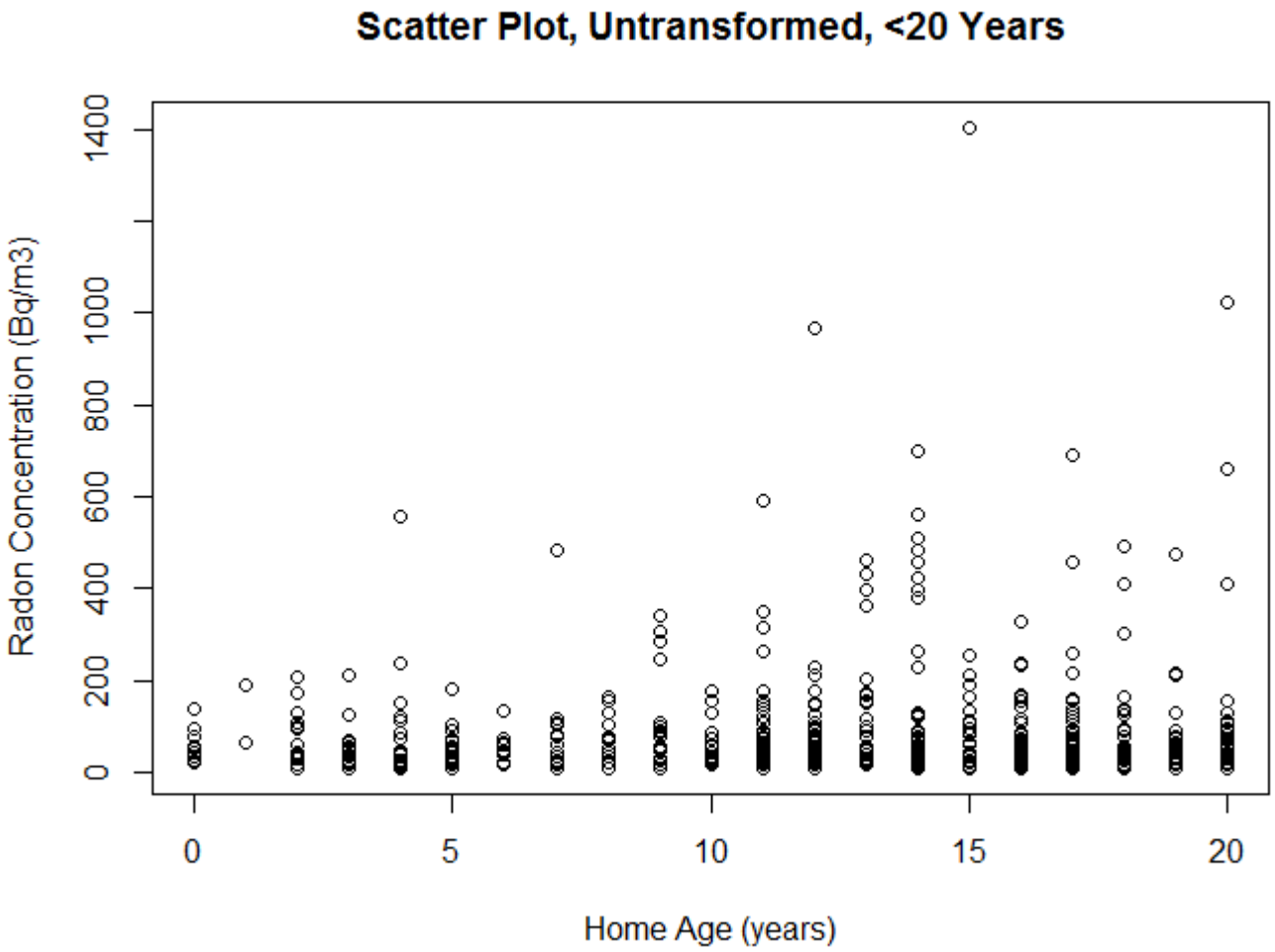
Scatter Plots

- Does it look like there is a relationship here?
- What if we limited the data we were working with to homes <20 years?



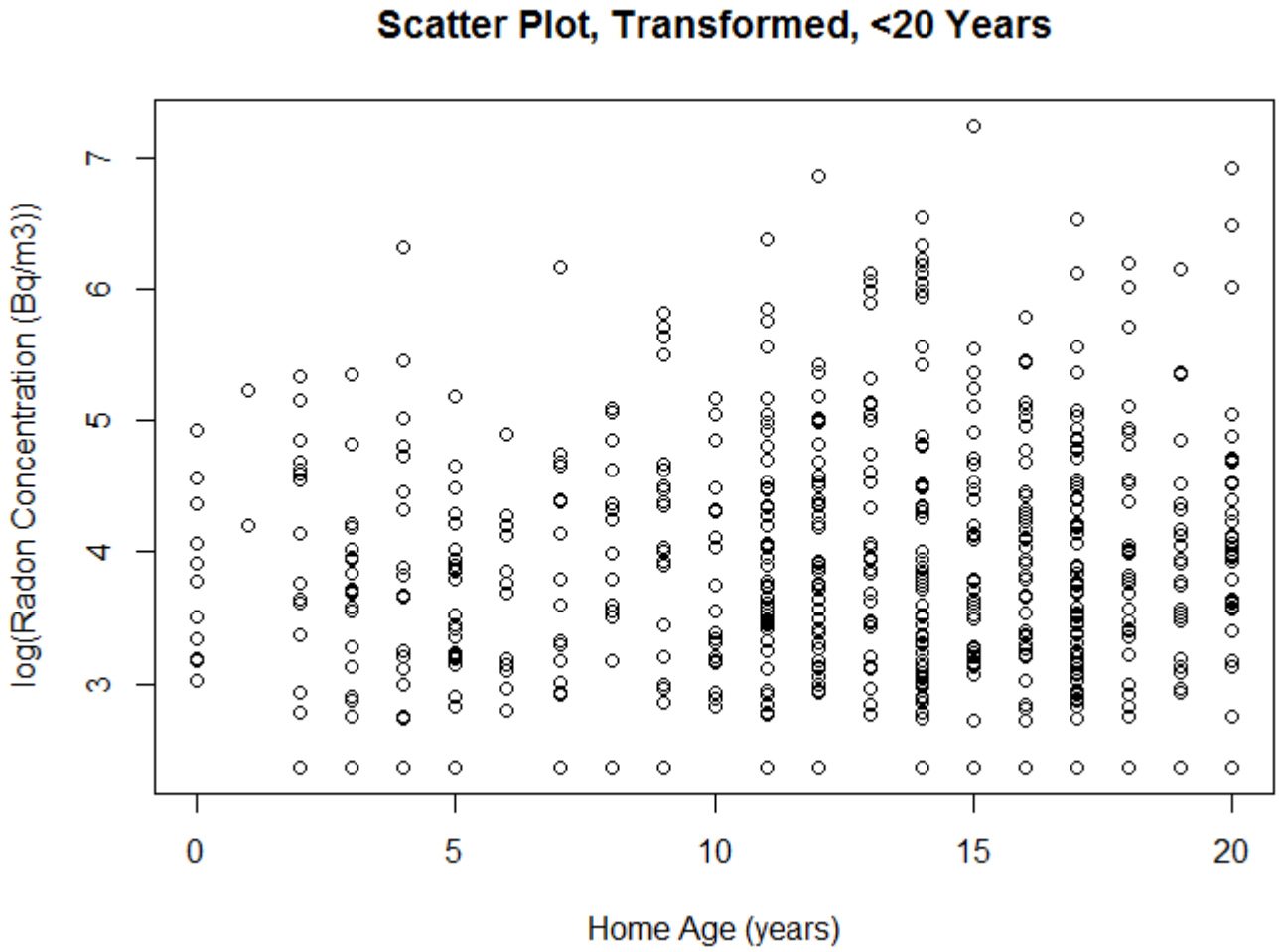
Scatter Plots

- What does it look like now?
- What would it mean for our results if we restricted our data this way?



Scatter Plots

- How does it look now?
- Do you think there is a significant relationship?



Pearson's Correlation (*r*)

- Used to test the association between two continuous variables
- The value of *r* can range from -1 to 1
- H_0 = the true value of *r* is not different from 0
- We test this hypothesis with another statistic that follows our old friend, the t distribution

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Correlation between untransformed radon concentrations and home age. Is it significant?

```
        Pearson's product-moment correlation

data:  radon$MainRadon and radon$HomeAge1990
t = 1.5482, df = 1132, p-value = 0.1219
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.01228041  0.10390189
sample estimates:
       cor
0.04596618
```

Pearson's Correlation

```
        Pearson's product-moment correlation

data:  radon$MainRadon[which(radon$HomeAge1990 <= 20)] and
radon$HomeAge1990[which(radon$HomeAge1990 <= 20)]
t = 1.9348, df = 629, p-value = 0.05346
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.001140402  0.154045601
sample estimates:
       cor
0.07691844
```

Correlation between untransformed radon concentrations and home age for homes <= 20 years.

```
        Pearson's product-moment correlation

data:  radon$LogRadon and radon$HomeAge1990
t = 2.3424, df = 1132, p-value = 0.01933
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.01128409 0.12715186
sample estimates:
      cor
0.0694522
```

Correlation between log-transformed radon concentrations and home age.

```
        Pearson's product-moment correlation

data:  radon$LogRadon[which(radon$HomeAge1990 <= 20)] and
radon$HomeAge1990[which(radon$HomeAge1990 <= 20)]
t = 1.4961, df = 629, p-value = 0.1351
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.0185920  0.1369618
sample estimates:
       cor
0.05954641
```

Correlation between log-transformed radon concentrations and home age for homes <= 20 years.

Simple Linear Regression

- What does the intercept mean now?
- What is the interpretation for the Home Age coefficient?
- How is *r* related to R^2 when we have two continuous variables in a simple regression model?

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.821018   0.046809  81.630   <2e-16 ***
radon$HomeAge1990 0.003684   0.001573   2.342   0.0193 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9417 on 1132 degrees of freedom
Multiple R-squared:  0.004824,  Adjusted R-squared:  0.003944
F-statistic: 5.487 on 1 and 1132 DF,  p-value: 0.01933
```

- GM of radon for a home of 0 years is exp(3.821018+0.003684***0**) = 45.65 Bq/m^3
- GM of radon for a home of 1 years is exp(3.821018+0.003684***1**) = 45.82 Bq/m^3
- GM of radon for a home of 2 years is exp(3.821018+0.003684***2**) = 45.99 Bq/m^3
- 45.82/45.65 = 1.0037
- 46.99/45.82 = 1.0037
- exp(0.003684***1**) = 1.0037
- exp(0.003684***2**) = 1.0074

Each **1-unit** increase in home age is associated with a 1.0037-fold increase in geometric mean radon concentrations.

Report What Makes Sense

- What does the intercept mean now?
- What is the interpretation for the Home Age coefficient?
- How is *r* related to R^2 when we have two continuous variables in a simple regression model?

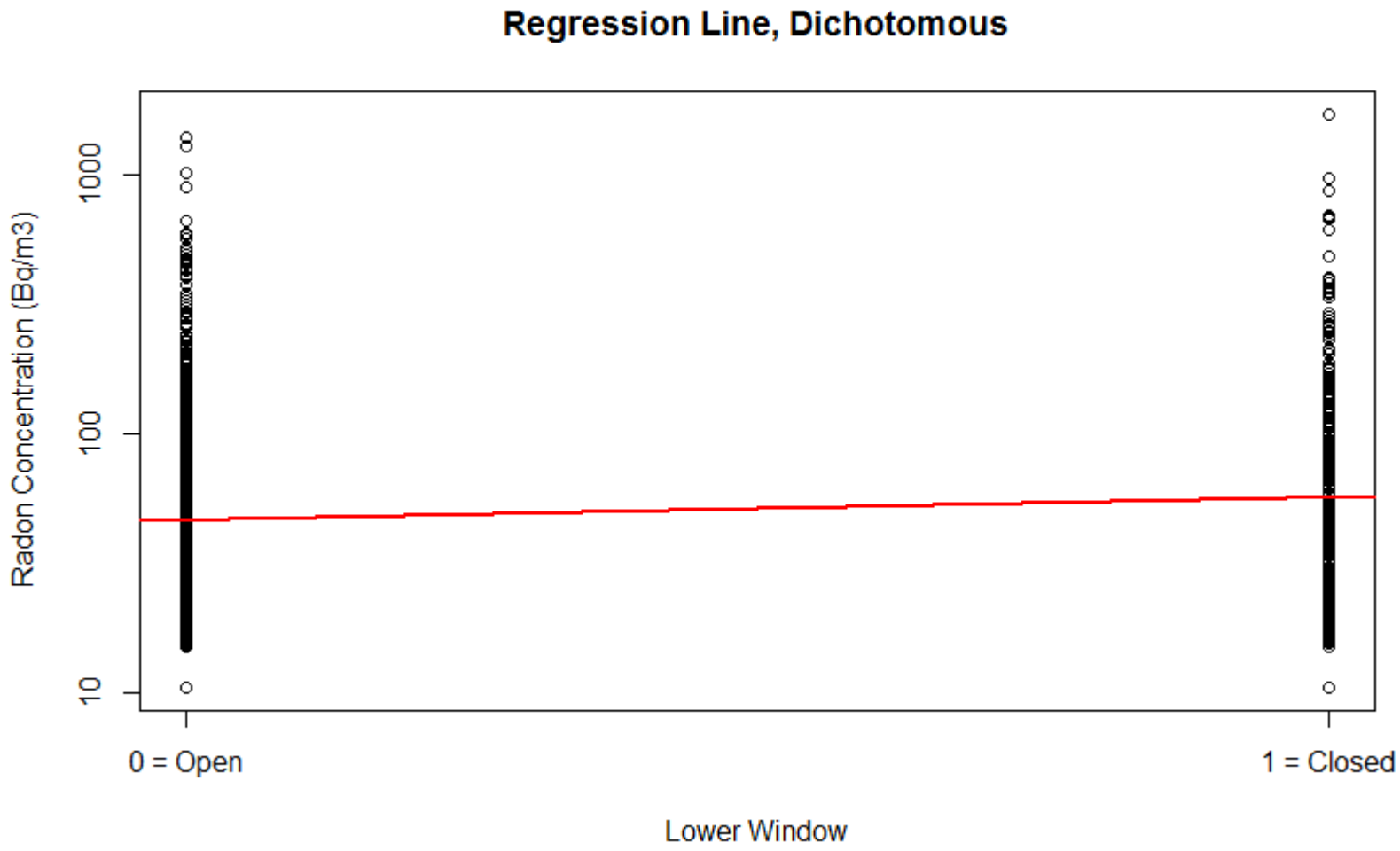
```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.821018   0.046809  81.630   <2e-16 ***
radon$HomeAge1990 0.003684   0.001573   2.342   0.0193 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9417 on 1132 degrees of freedom
Multiple R-squared:  0.004824,  Adjusted R-squared:  0.003944
F-statistic: 5.487 on 1 and 1132 DF,  p-value: 0.01933
```

- GM of radon for a home of 0 years is exp(3.821018+0.003684***0**) = 45.65 Bq/m^3
- GM of radon for a home of 1 years is exp(3.821018+0.003684***10**) = 47.36 Bq/m^3
- GM of radon for a home of 2 years is exp(3.821018+0.003684***20**) = 49.14 Bq/m^3
- 47.36/45.65 = 1.037
- 49.14/47.36 = 1.037
- exp(0.003684***10**) = 1.037
- exp(0.003684***20**) = 1.076

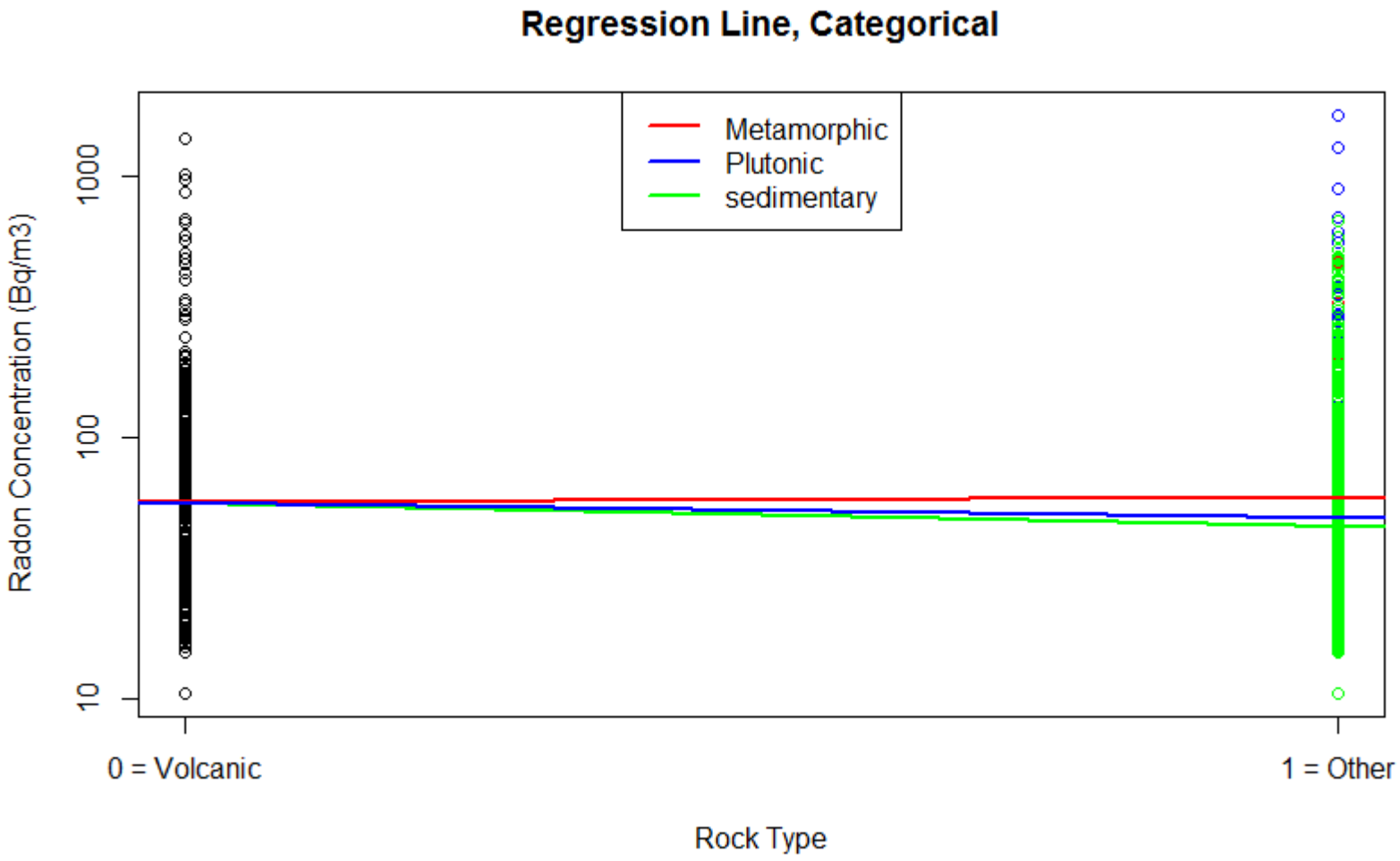
Each **10-unit** increase in home age is associated with a 1.037-fold increase in geometric mean radon concentrations.

Regression Line, Continuous

Regression Line, Dichotomous

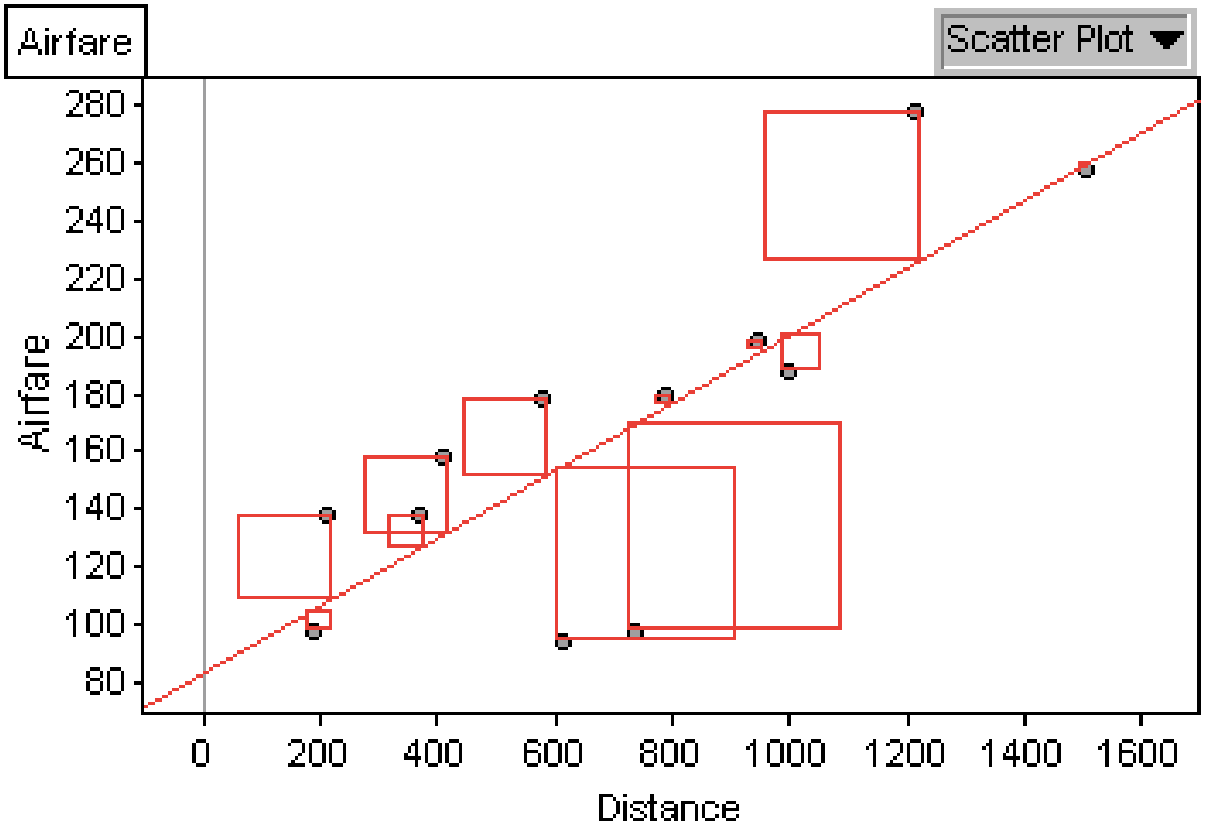


Regression Line, Categorical



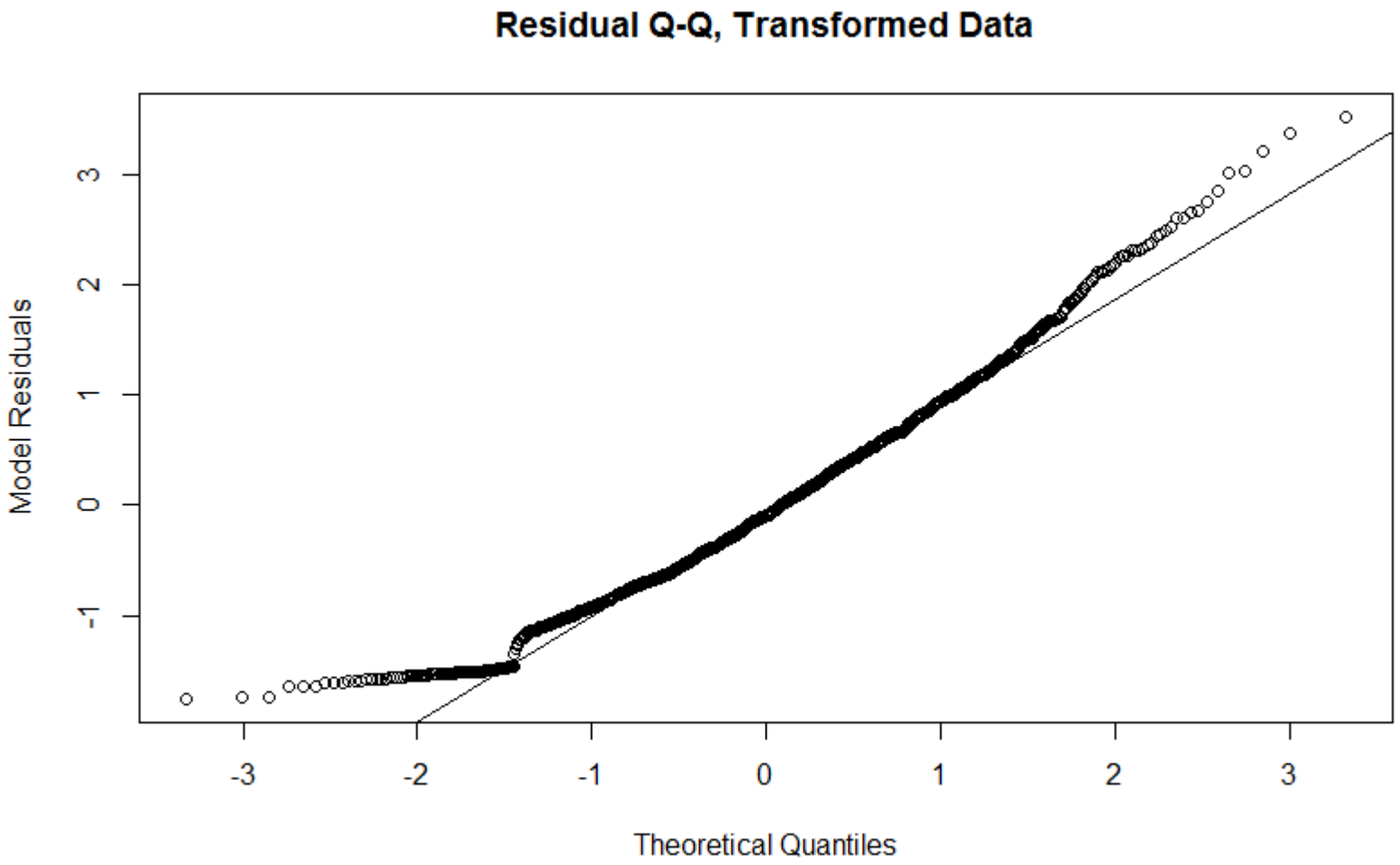
Least Squares

- All of our regression lines to date have been fitted this way?
- Why do we take squares instead of just lines?
- This is the most simple way to fit models…can you think of another way, based on what we've done so far in class and assignments?



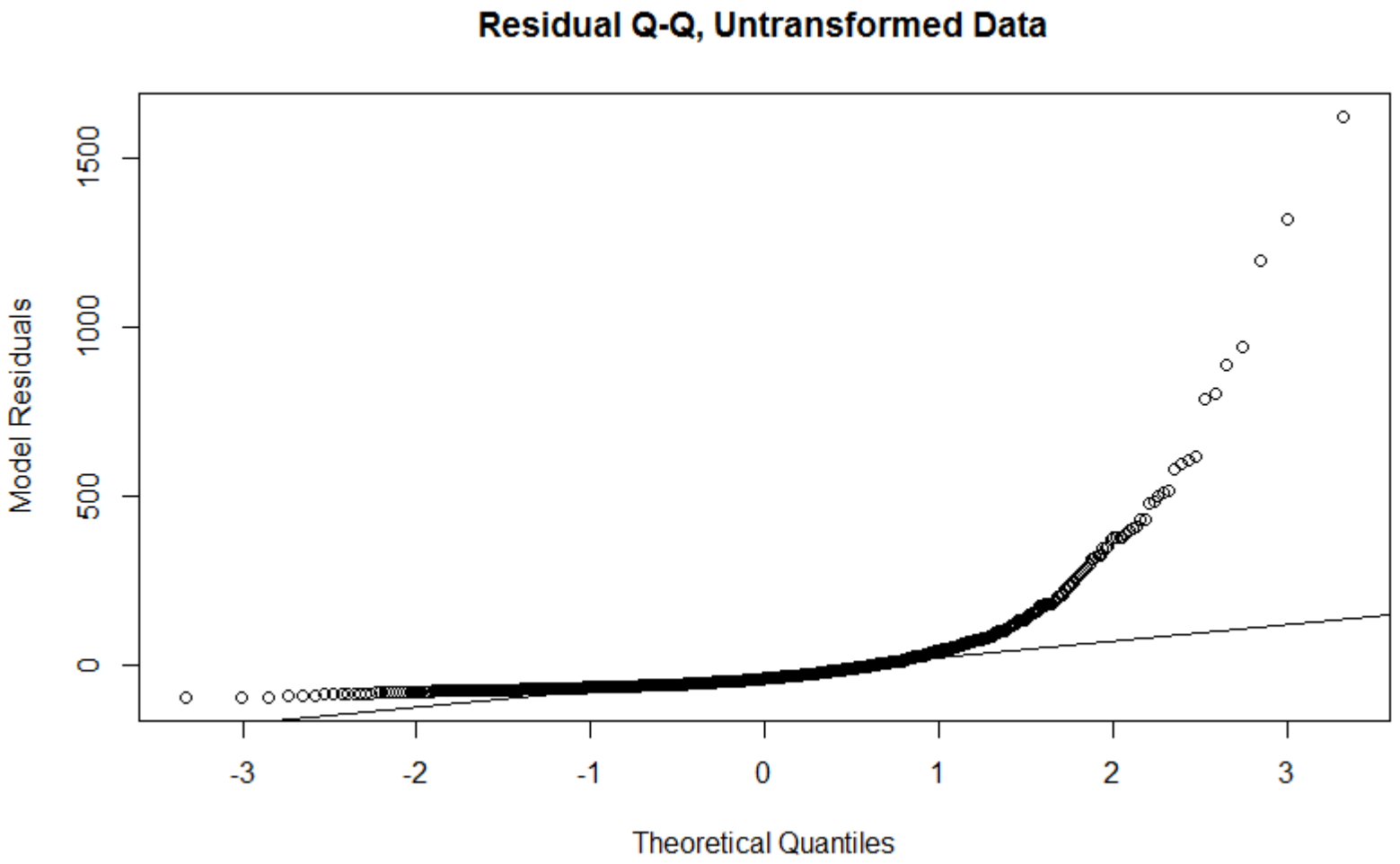
Diagnostic Plots: Q-Q

- What is the fundamental assumption of linear regression?
- How can we test if we have violated that assumption?



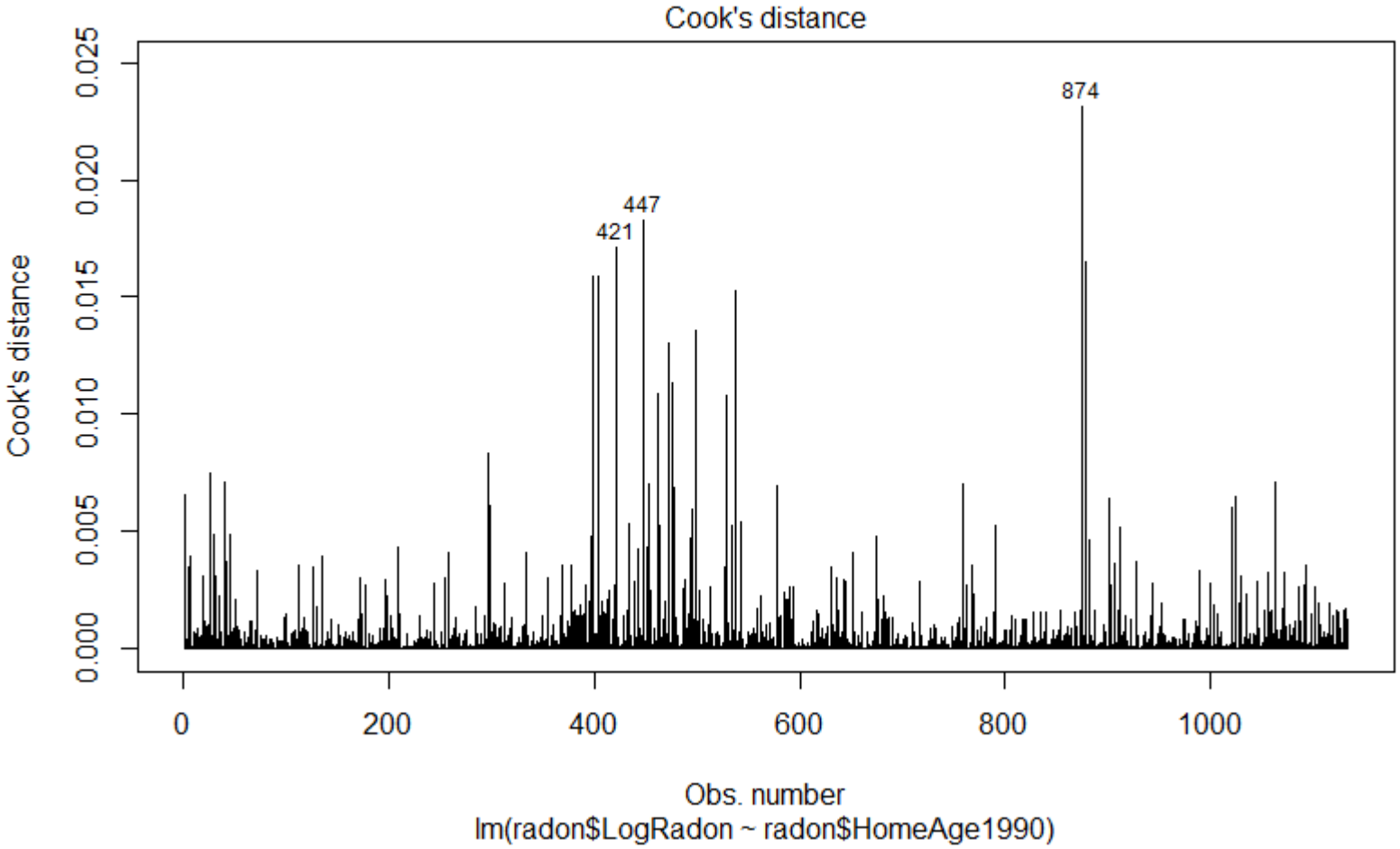
Diagnostic Plots: Q-Q

- From herein you must demonstrate to me that you are not violating the assumptions of linear regression, either in figures or in words.



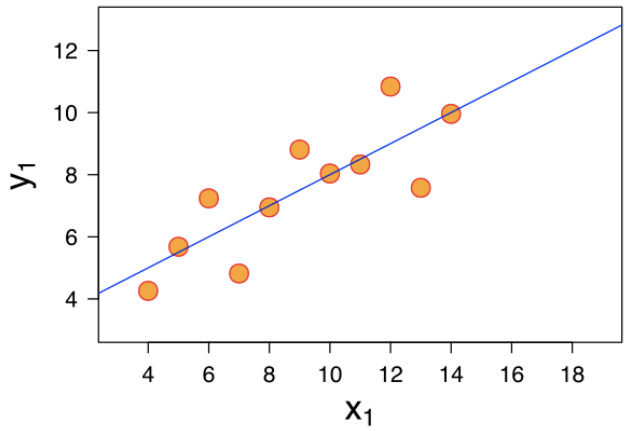
Diagnostic Plots: Cook's D

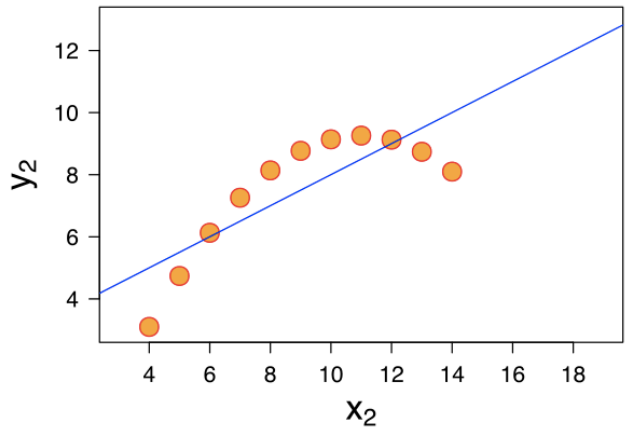
- Used to detect influential observations
- Removing an influential observation may drastically change your results
- Why are influential points clustered together here?
- Would it be appropriate to remove?

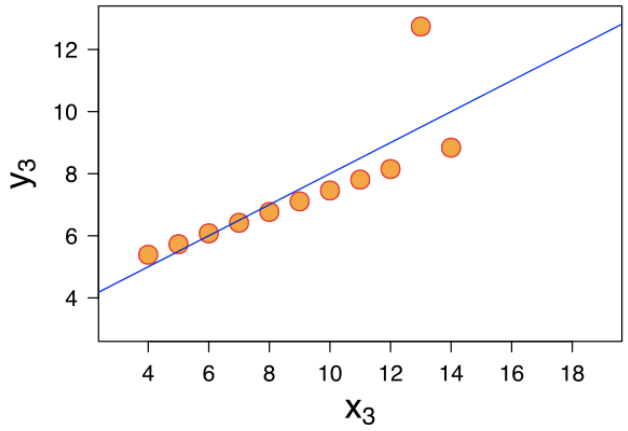


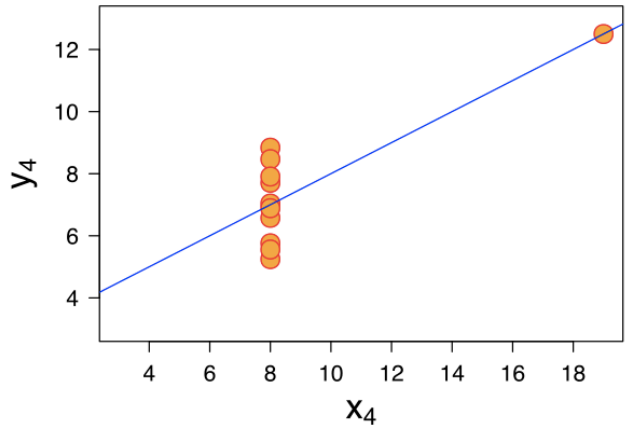
Get Intimate with Your Data

- Before you begin any regression analysis you must conduct exploratory analyses
- Understand the univariate relationships between in your dependent and independent variables
- Think of Abscombe's Quartet!









Univariate Exploration

How do we?	Dichotomous vs. Continuous	Categorical vs. Continuous	Continuous vs. Continuous
Visualize?			
Test for an association?			
Quantify the effect?			

Multiple Regression

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots$ how far could we go?
- We are now exploring the COMBINED effects of a SET OF INDEPENDENT VARIABLES on the dependent variable
- The effect of each variable in the model is ADJUSTED for the effect of all of the other variables in the model

```
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             3.83792    0.03400  112.87  < 2e-16 ***
radon$LowerWindowClosed 0.20993    0.06121    3.43 0.000627 ***
```

Model 1

```
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              4.03952    0.05753  70.213  < 2e-16 ***
radon$RockTypeMetamorphic  0.05216    0.13341   0.391  0.69591
radon$RockTypePlutonic    -0.13794    0.08399  -1.642  0.10080
radon$RockTypeSedimentary -0.20825    0.06972  -2.987  0.00288 **
```

Model 2

```
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              3.96980    0.06139  64.669  < 2e-16 ***
radon$LowerWindowClosed    0.22259    0.06148   3.621 0.000307 ***
radon$RockTypeMetamorphic  0.07080    0.13319   0.532 0.595157
radon$RockTypePlutonic    -0.12367    0.08463  -1.461 0.144230
radon$RockTypeSedimentary -0.22803    0.07041  -3.239 0.001237 **
```

Model 3

Fill this in for Effect on **GM** with p-values

Coefficient	Unadjusted Effect	Adjusted for Lower Window	Adjusted for Rock Type	Adjusted for Home Age	Adjusted for both other variables
Window Closed		----------			
Metamorphic Rock			----------		
Plutonic Rock			----------		
Sedimentary Rock			----------		
Home Age				----------	

How to Interpret

- Write the equation on the board
- What does the Intercept mean?
- What is the effect of closed windows on geometric mean radon?

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 3.906806   0.068705  56.864  < 2e-16 ***
radon$LowerWindowClosed     0.207216   0.061858   3.350 0.000836 ***
radon$RockTypeMetamorphic   0.042381   0.133737   0.317 0.751381
radon$RockTypePlutonic     -0.168095   0.087299  -1.926 0.054423 .
radon$RockTypeSedimentary  -0.231490   0.070331  -3.291 0.001029 **
radon$HomeAge1990           0.003351   0.001651   2.030 0.042581 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9339 on 1099 degrees of freedom
  (29 observations deleted due to missingness)
Multiple R-squared:  0.02646,   Adjusted R-squared:  0.02204
F-statistic: 5.975 on 5 and 1099 DF,  p-value: 1.836e-05
```

Closed basement windows are associated with a 1.23-fold [1.09, 1.39] increase in GEOMETRIC MEAN radon compared with open basement windows, WHEN ALL OTHER VARIABLES ARE HELD CONSTANT.

Prediction

What is geometric mean radon when: (write your equations)

1. Windows closed, Metamorphic Rock, 100 years old
2. Windows open, Volcanic Rock, 20 years old
3. Windows closed, Sedimentary Rock, 12 years old
4. Windows open, Plutonic Rock, 76 years old
5. Windows closed, Metamorphic Rock, 2 years old

Variable Selection

- Try every possible model and pick the best one based on variable p-values and model PARSIMONY
- Select best univariate predictors
- Backward elimination
  - Start will all possible predictors
  - Remove the one one with the highest p-value
  - Continue until all have p-values under some threshold
- Forward selection
  - Start with strongest predictor
  - Add other predictors one by one and retain only those with small p-values
- Stepwise selection
  - A computer-automated hybrid of backward and forward selection
- Theory and common sense
  - Avoid collinearity
  - Include factors that should be adjusted for, regardless of significance
  - Parsimony

Next Week

- Logistic regression on a binary dependent variables
- Model comparisons