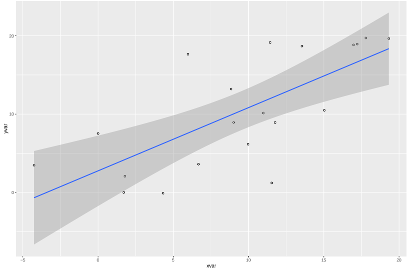
Simple Linear Regression

Zack W Almquist

April 17, 2018



Road Map to Today's Lecture

- Overview
  - Motivation (Why models?, why SLR?)
  - Example: Housing prices regressed on house size
  - Estimator (Least Squares)
  - Properties of Least Squares
  - ANOVA for Regression (R^2)
- Appendix: R Code and other information

Introduction: Simple Linear Regression

Motivation

- How do we theorize about the world?

Introduction: Simple Linear Regression

Motivation

- How do we theorize about the world?
- One way is to posit a relationship between two measurements

Introduction: Simple Linear Regression
Motivation

- How do we theorize about the world?
- One way is to posit a relationship between two measurements
  - E.g., Parent and Child Height

Introduction: Simple Linear Regression

Motivation

- How do we theorize about the world?
- One way is to posit a relationship between two measurements
  - E.g., Parent and Child Height
  - E.g., Beer Cost and Beer Sales

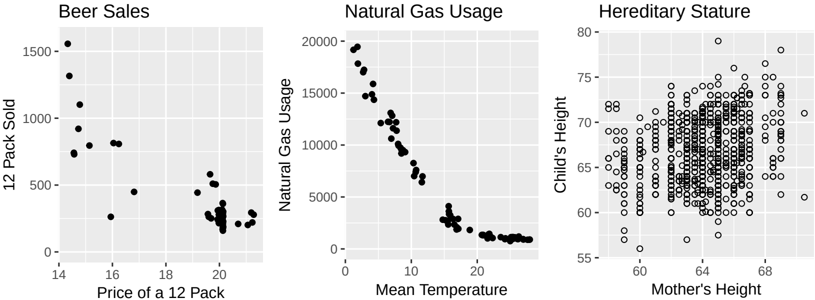
Introduction: Simple Linear Regression

Motivation

- How do we theorize about the world?
- One way is to posit a relationship between two measurements
  - E.g., Parent and Child Height
  - E.g., Beer Cost and Beer Sales
  - E.g., Natural Gas Usage and Temperature

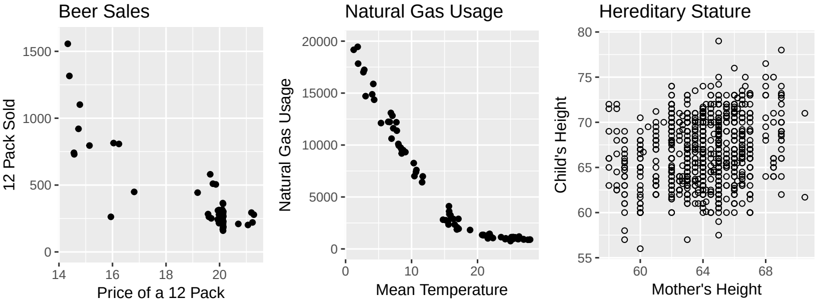
Introduction: Simple Linear Regression
Motivation

- How do we theorize about the world?



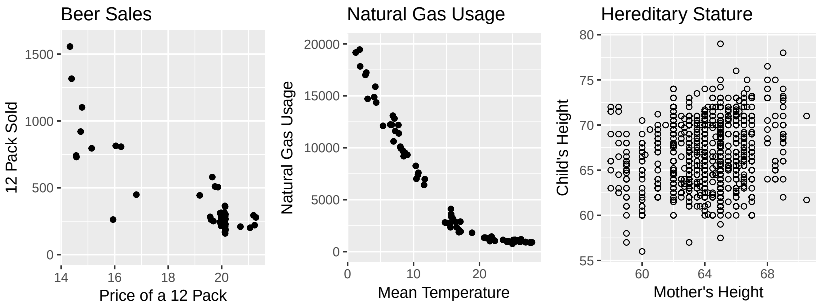
Introduction: Simple Linear Regression
Motivation

- How do we theorize about the world?
  - How does Beer Sales (Y) relate to Beer Cost (X)?



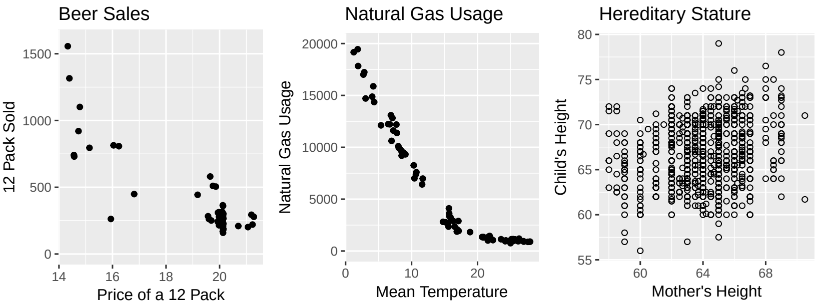
Introduction: Simple Linear Regression
Motivation

- How do we theorize about the world?
  - How does Beer Sales (Y) relate to Beer Cost (X)?
    - This can be written as $Y = f(X)$



Introduction: Simple Linear Regression
Motivation

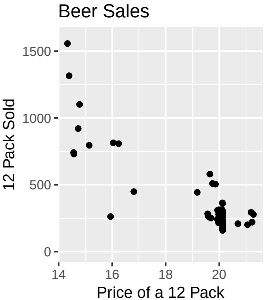
- How do we theorize about the world?
  - How does Beer Sales (Y) relate to Beer Cost (X)?
    - This can be written as $Y = f(X)$
  - Start by looking at the linear relationship between two objects

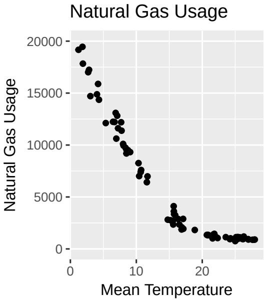


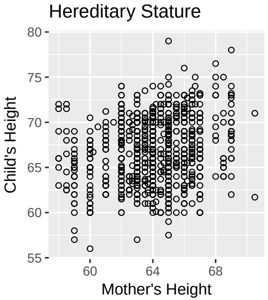
Introduction: Simple Linear Regression

Motivation

- How do we theorize about the world?
  - How does Beer Sales (Y) relate to Beer Cost (X)?
    - This can be written as $Y = f(X)$
  - Start by looking at the linear relationship between two objects
    - $Y = b_0 + b_1 X$

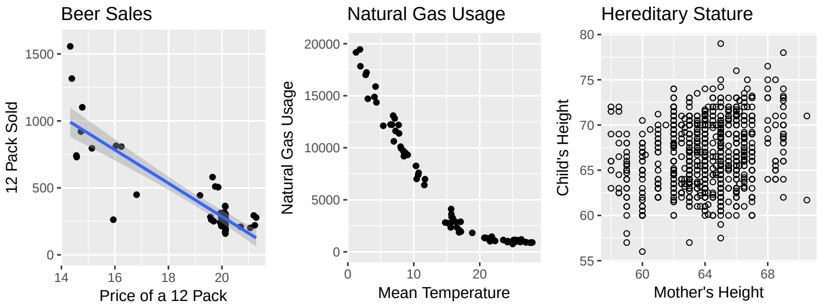






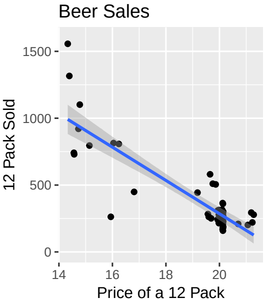
Introduction: Simple Linear Regression
Motivation

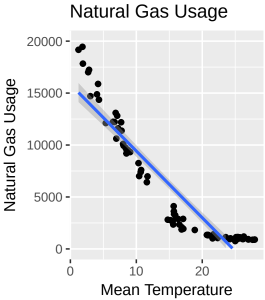
- How do we theorize about the world?
  - How does Beer Sales (Y) relate to Beer Cost (X)?
    - This can be written as $Y = f(X)$
  - Start by looking at the linear relationship between two objects
    - $Y = b_0 + b_1 X$

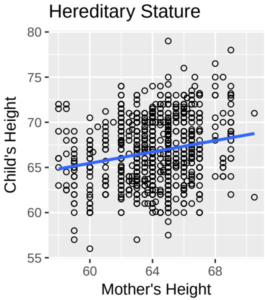


Introduction: Simple Linear Regression
Motivation

- How do we theorize about the world?
  - How does Beer Sales (Y) relate to Beer Cost (X)?
    - This can be written as $Y = f(X)$
  - Start by looking at the linear relationship between two objects
    - $Y = b_0 + b_1 X$

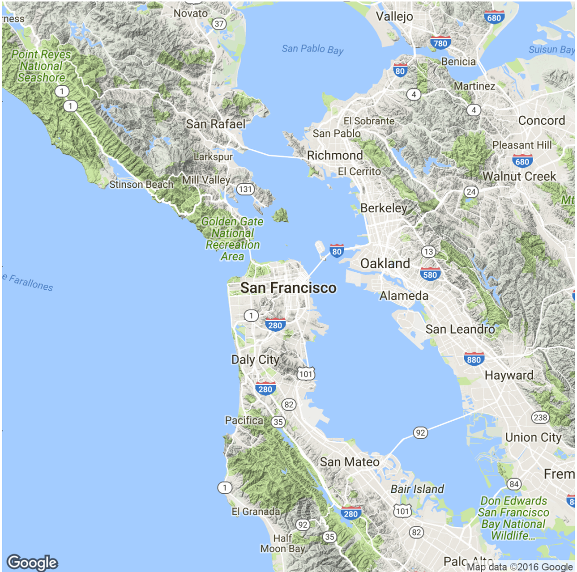






Let's Build Some Intuition

Housing Cost: A Case Study in San Francisco



Let's Build Some Intuition

- Let's start with a practical problem:
  - Housing prices (Somthing we can appreciate in SF!)
- We will use this to develop our Regression model
  - Focus today is on the "least squares" estimator
    - Note: Population parameters (Greek, e.g., β)
    - Note: Estimators of the "true" parameters (Roman, e.g., b)

Predicting Housing Prices

Problem:

- Predict market price based on observed characteristics
  - E.g., How much doe I expect to pay for a house?

Solution:

- Look at property sales data where we know the price and some observed characteristics
- Build a decision rule that predicts price as a function of the observed characteristics

What characteristics do we use?

We have to define the variables of interest and develop a specific quantitative measure of these variables

- Many factors or variables affect the price of a house
  - Size of house
  - Number of baths
  - Garage, air conditioning, etc.
  - Size of land (lot size)
  - Neighborhood location (e.g., the Mission)
- Easy to quantify price and size but what about other variables such as aesthetics, workmanship, etc?

Predicting Housing Prices

The value that we seek to predict is called the

- Dependent (or output) variable
  - Y = price of house (e.g. thousands of dollars)

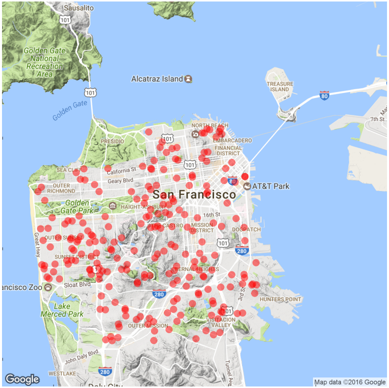
The variable that we use to guide prediction is the

- Explanatory (or input) variable
  - X = size of the house (in thousands of square feet)

To keep things simple we will focus only on the size of the house

Zillow.com Data

- Let's consider data from Zillow.com for San Francisco
- Focus on a random sample (county tax lot database)
  - Here we will consider a sample of 204 houses
- Focus on X (house size in SqFt) and Y (Zillow est. of price)



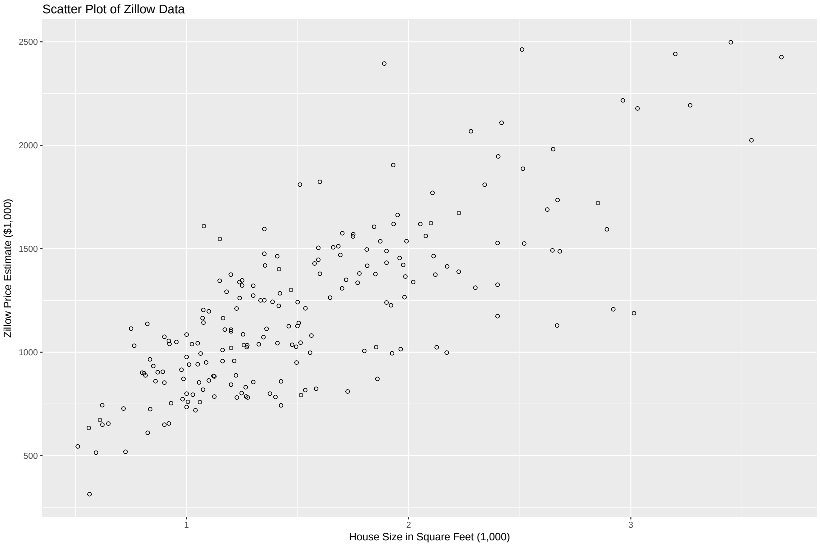
Zillow.com Data

Sample of data from Zillow.com

	House Size (SqFt)	Zillow Estimate
255 DOWNEY ST	1.683	1511.798
4845 17TH ST	2.650	1981.286
1158 CAPITOL AVE	1.125	882.611
184 BURLWOOD DR	1.254	1086.365
79 GAVIOTA WAY	1.593	1446.701
2342 41ST AVE	2.680	1487.369

- House Size (in 1,000 SqFt) and Zillow Estimate (in $1,000)

Zillow.com Data



Simple Linear Regression Model

The general relationship approximated by:

- $Y = f(X) + e$
  - Y is the response outcome variable
  - X is the explanitory or input variables
  - e represents anything left over, not described by f

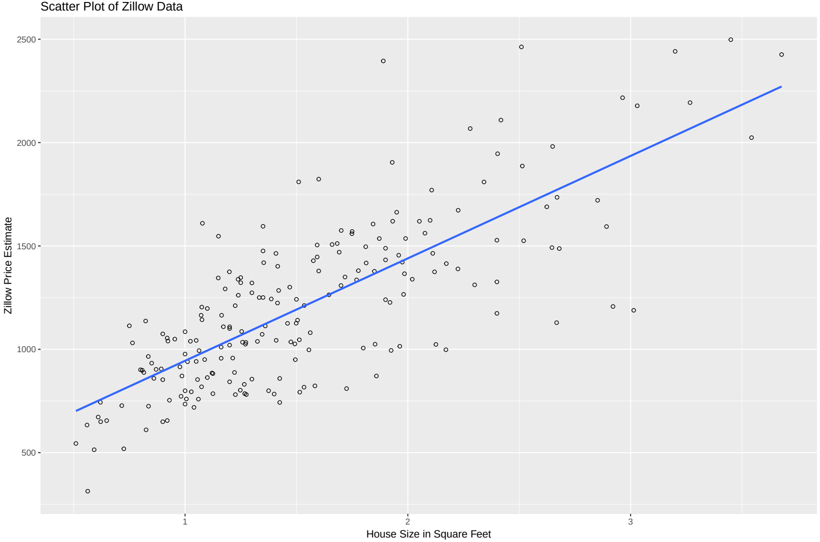
And, in particular, a linear relationship is written as:

- $Y = b_0 + b_1 X + e$

Zillow Data

Appears to be a linear relationship between price and size:

- As size goes up, price goes up



Geometry of a line

Recall how the slope (b_1) and intercept (b_0) work together

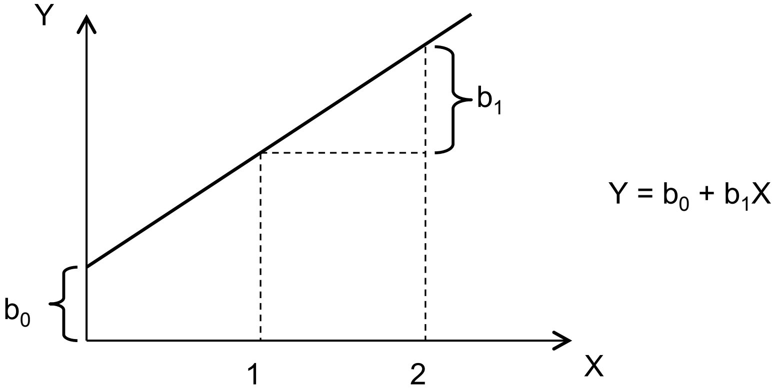


Figure 1:

Linear Model

Recall that the equation of a line is

$$Y = b_0 + b_1 X$$

where b_0 is the intercept and b_1 is the slope

- The intercept value is in units of Y (\$1000)
- The slope is in units of Y per units of X (\$1000/(1000 SqFt))

In the house price example

- The line is $b_0 = 449.001$, $b_1 = 495.482$

Linear Prediction

We can now predict the price of a house when we know only the size

- just read the value off the line that we've drawn

For example, given a house with size $X = 1$ (i.e., a 1,000 SqFt house), the predicted price would be

$$\hat{Y} = 449.001 + 495.482 \cdot 1 = 944.483 \ (K\$)$$

- Note: Conversion from 1,000 SqFt to \$1,000 is done for us by the slope coefficient (b_1)

What is a good line?

We desire a strategy for estimating the slope and intercept parameters in the model $\hat{Y} = b_0 + b_1 X$

That involves:

- Choosing a criterion, i.e., quantifying how well a line fits the observed data
- And matching that with a solution, i.e., finding the best line subject to that criteria

Criteria

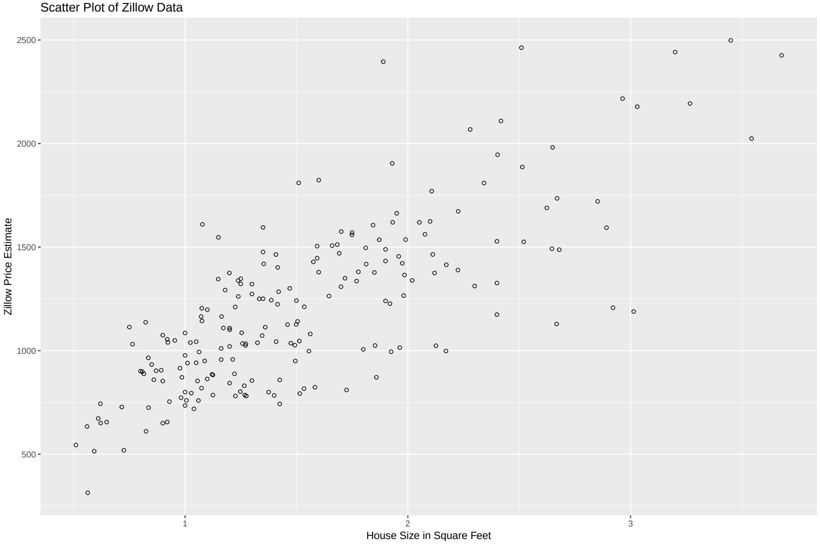
Although there are lots of ways to choose a criterion

- Only a small handful lead to solutions that are "easy" to compute (our focus today)
- And which have nice statistical properties (more later)

Most reasonable criteria involve measuring the amount by which the *fitted value* obtained from the line *differs* from the *observed value* of the response value(s) in the data

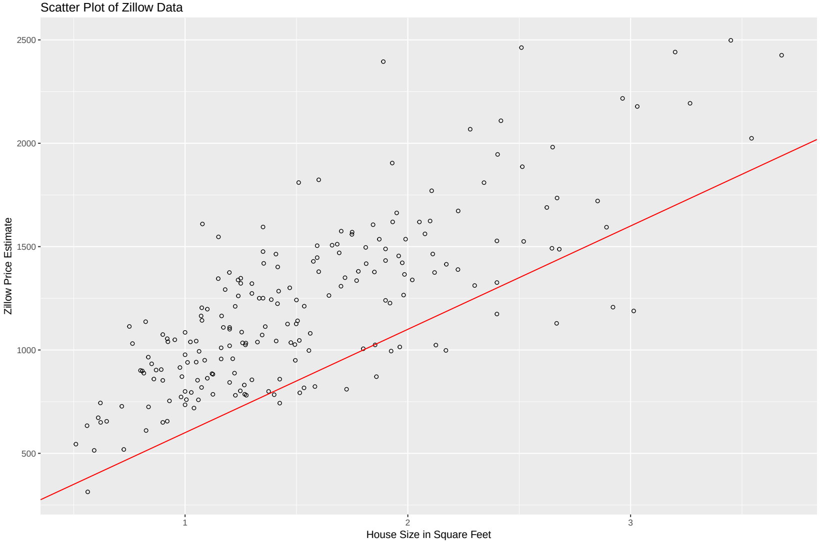
- This amount is called the residual
- Good lines produce small residuals

Criteria



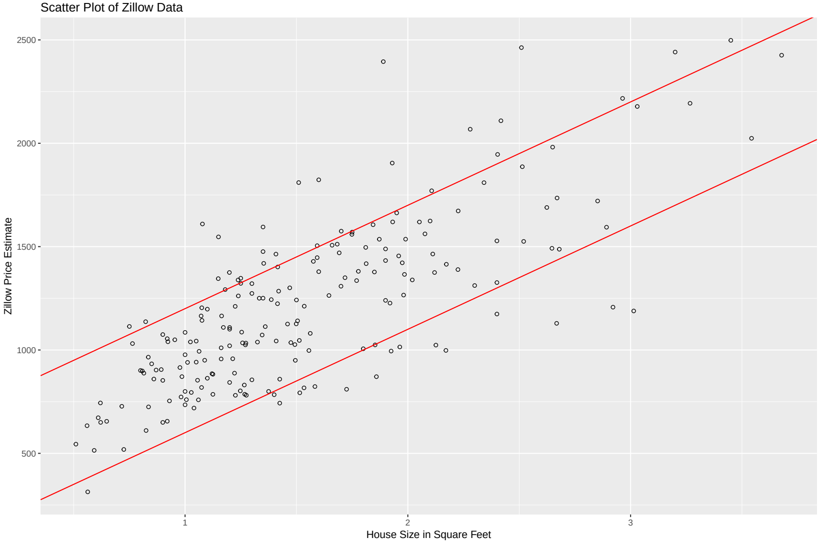
Scatter Plot of Zillow Data
Zillow Price Estimate
2500
2000
1500
1000
500
1
2
3
House Size in Square Feet

Criteria

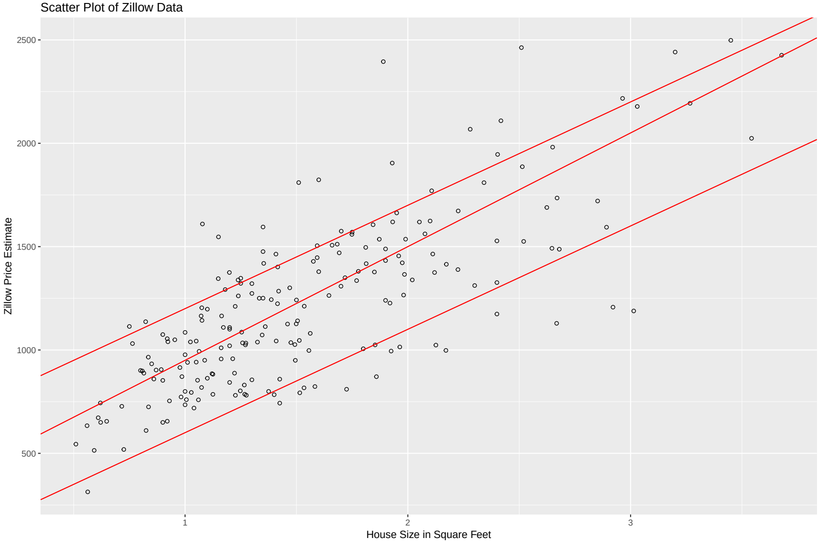


Scatter Plot of Zillow Data
Zillow Price Estimate
House Size in Square Feet
2500
2000
1500
1000
500
1
2
3

Criteria

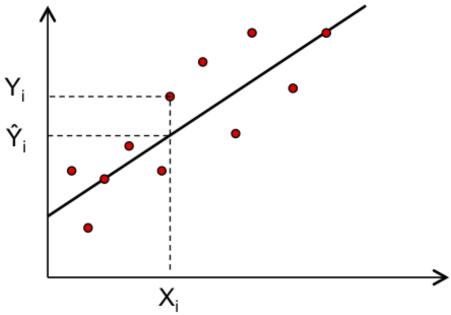

Scatter Plot of Zillow Data
Zillow Price Estimate
2500
2000
1500
1000
500
1
2
3
House Size in Square Feet

Criteria



Scatter Plot of Zillow Data
Zillow Price Estimate
2500
2000
1500
1000
500
1
2
3
House Size in Square Feet

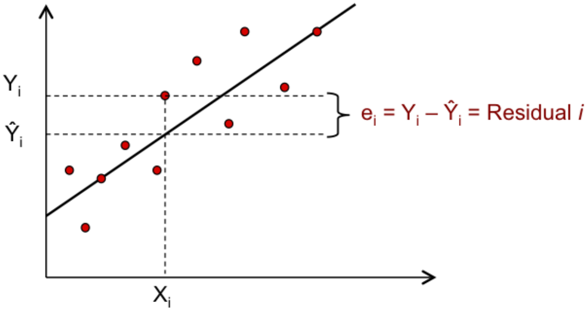
Fitted Values



The dots are the observed values and the line represents our fitted values given by

$$\hat{Y}_i = b_0 + b_1 X_i$$

Fitted Values



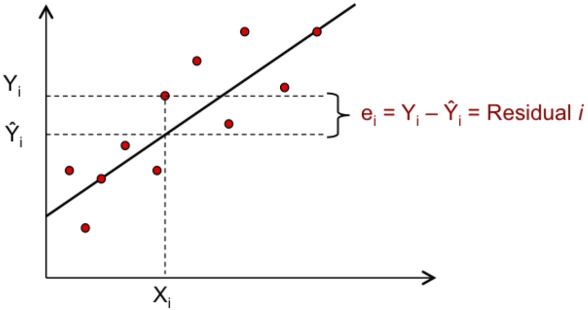
- The residual e_i is the discrepency between the fitted $\hat{Y}_i$ and observed Y_i values
  - Note that we can write

$$\begin{aligned} Y_i &= \hat{Y}_i + (Y_i - \hat{Y}_i) \\ &= \hat{Y}_i + e_i \end{aligned}$$

Least Squares

A reasonable goal is to minimize the size of all residuals:

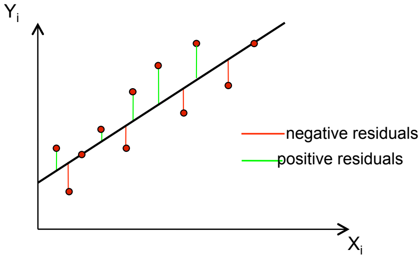
- This should guarantee that if Y and X are perfectedly related our residuals will be zero
  - And we would have perfect line!
- For imperfect data we end up with a trade-off between *moving closer* to some points and at the same time *moving away* from other points



Least Squares

Since some residuals are *positive* and some are *negative*, we need a measure to minimize the errors:

- Absolute Error: $|e_i|$ - treats positives and negatives equally
- Squared Error: So does e_i^2 - has nicer mathematical properties



Least squares chooses b_0 *and* b_1 *to minimize* $\sum_{i=1}^{n} e_i^2$

Least Squares

Choose the line to minimize the sum of the squares of the residuals,

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - [b_0 + b_1 X_i])^2$$

How do we minimize $\sum_{i=1}^{n} (Y_i - [b_0 + b_1 X_i])^2$?

- We need to find the *optimal* b_0 and b_1

Least Squares

Choose the line to minimize the sum of the squares of the residuals,

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - [b_0 + b_1 X_i])^2$$

How do we minimize $\sum_{i=1}^{n}(Y_i - [b_0 + b_1 X_i])^2$?

- We need to find the *optimal* b_0 and b_1
- This is a classic calculus optimization problem

Least Squares

Choose the line to minimize the sum of the squares of the residuals,

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - [b_0 + b_1 X_i])^2$$

How do we minimize $\sum_{i=1}^{n} (Y_i - [b_0 + b_1 X_i])^2$?

- We need to find the *optimal* b_0 and b_1
- This is a classic calculus optimization problem
- $$b_1 = r_{xy} \frac{S_y}{S_x}$$

Least Squares

How do we minimize $\sum_{i=1}^{n}(Y_i - [b_0 + b_1 X_i])^2$?

- We need to find the *optimal* b_0 and b_1
- This is a classic calculus optimization problem
- $$b_1 = r_{xy}\frac{S_y}{S_x}$$

$$r_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$$

$$S_X = \sqrt{\frac{1}{n}\sum(X_i - \bar{X})^2}$$

$$S_Y = \sqrt{\frac{1}{n}\sum(Y_i - \bar{Y})^2}$$

Least Squares

How do we minimize $\sum_{i=1}^{n}(Y_i - [b_0 + b_1 X_i])^2$?

$$b_1 = r_{xy}\frac{S_y}{S_x}$$
$$b_0 = \bar{Y} - b_1\bar{X}$$

Exercise

- Use Household Size to predict Last Sold Price for Zillow data
- Google Sheets Example
- Google Sheets In Class Assignment
  - Compute $\bar{X}$, $\bar{Y}$, $S(X)$, $S(Y)$, $corr(X, Y)$, b_0, b_1
  - And $\hat{Y}$, e, e^2
- Extra (You can do SLR in Google Sheets automatically):
  - Try highlighting columns: "House Size [X] (in 1000 SqFt)" and "Last Sold Price [Y] (in $1000)"
  - Insert → Chart → Scatter → Customization → Trendline → linear → Label → Use equation → Check box R^2

$$\hat{Y}_i = b_0 + b_1 X_i \qquad\qquad b_1 = r_{xy}\frac{S_y}{S_x}$$

$$e_i = Y_i - \hat{Y}_i \qquad\qquad b_0 = \bar{Y} - b_1\bar{X}$$

Steps in a regression analysis

Basic Recipe for SLR problems

We have covered the first 5!

1. State the problem
2. Select potentially relevant variables
3. Data collection
4. Model specification (simple linear)
5. Model fitting (least squares)
6. Model validation and criticism
7. Answering the posed questions

This is a simplification ...

`- it is more iterative, and it can be an art ...`

Model Assessment and Properties of Least Squares Fit

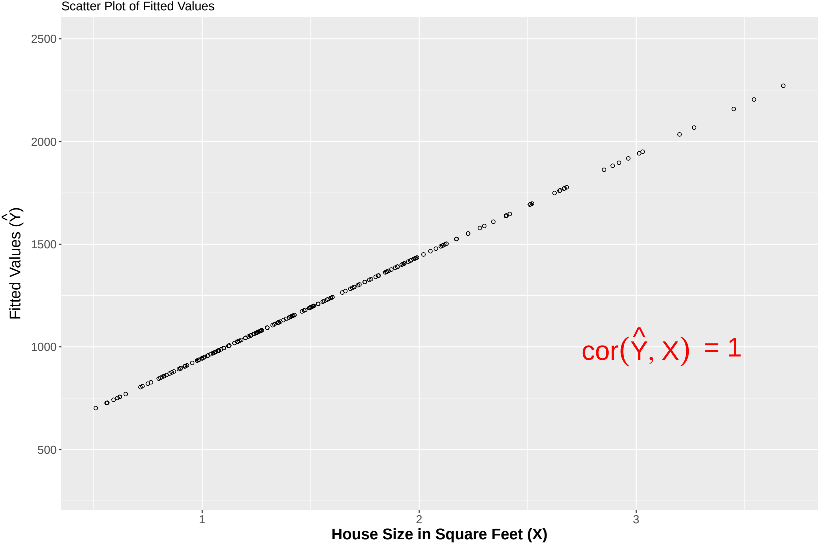
Properties of the least squares fit

Developing techniques for model validation and criticism requires a *deeper understanding* of the least squares line

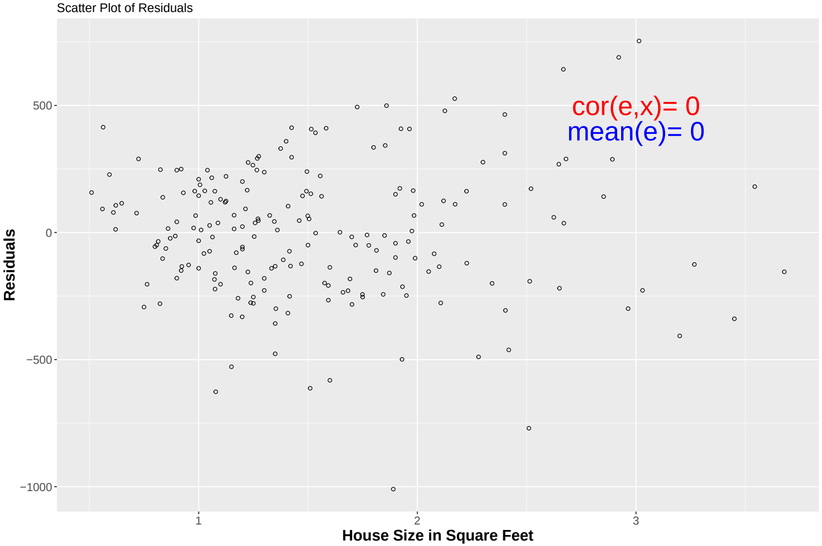
The fitted values ($\hat{Y}_i$) and "residuals" (e_i) obtained from the least squares line have some special properties

Lets look at the housing data analysis to figure out what some of these properties are...

The fitted values are perfectly correlated with the inputs



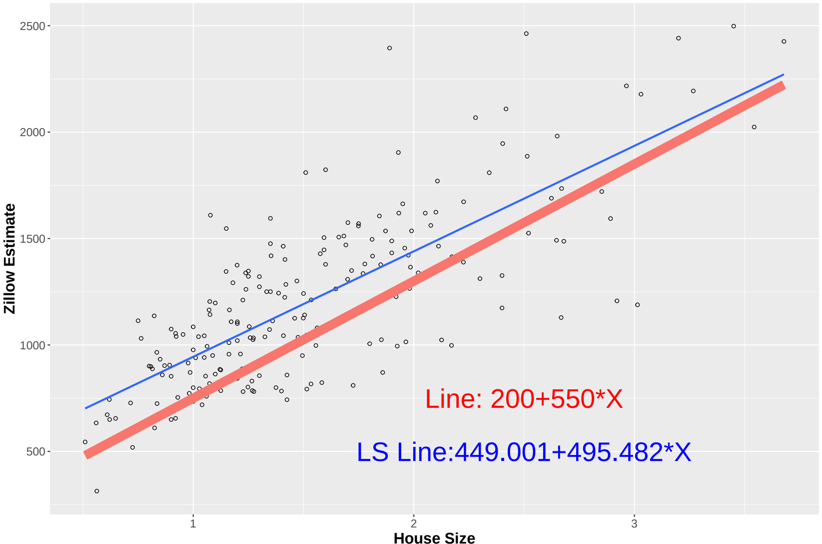
The residuals are “stripped of all linearity”



Relationship between $\hat{Y}$, e and X

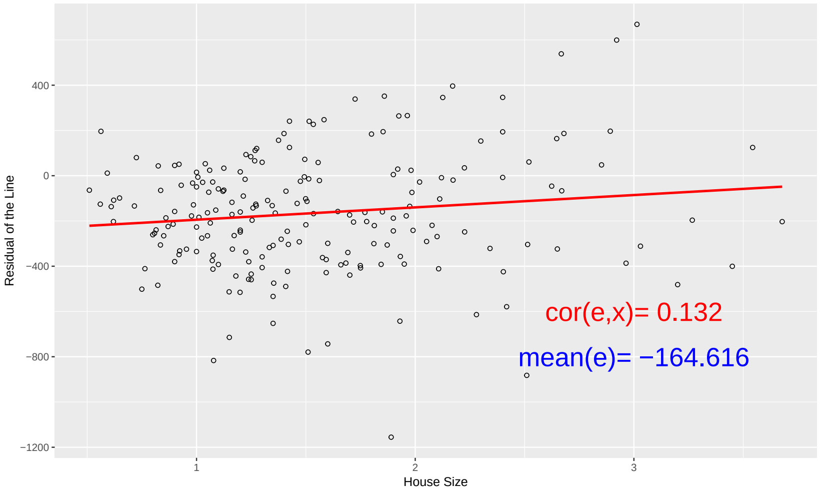
What is the intuition for the relationship between $\hat{Y}$, e and X?

- Lets consider an alternative line:



Relationship between $\hat{Y}$, e and X

This is a bad fit! We are underestimating the value of small houses and overestimating the value of big houses



Clearly, we have left some predictive ability on the table!

Relationship between $\hat{Y}$, e and X

As long as the correlation between e and X is non-zero, we could always adjust our prediction rule to do better

We need to exploit all of the predictive prower in the X values and put this into $\hat{Y}$,

- Leaving no "Xness" in the residuals

In summary: $Y = \hat{Y} + e$ where:

- $\hat{Y}$ is "made from X"; $corr(X, \hat{Y}) = 1$
- e is unrelated to X; $corr(X, e) = 0$
- $\bar{e} = 0$

An alternate derivation

Suppose we turned things around and

- Insisted on these qualities
- Rather than serendipitiously obtaining them as a consequence

$$\frac{1}{n}\sum_{i=1}^{n} e_i = 0$$

An alternate derivation

Suppose we turned things around and

- Insisted on these qualities
- Rather than serendipitously obtaining them as a consequence

$$\frac{1}{n}\sum_{i=1}^{n} e_i = 0 \Rightarrow \frac{1}{n}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)$$

An alternate derivation

Suppose we turned things around and

- Insisted on these qualities
- Rather than serendipitiously obtaining them as a consequence

$$\frac{1}{n}\sum_{i=1}^{n} e_i = 0 \Rightarrow \frac{1}{n}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)$$
$$\Rightarrow \bar{Y} - b_0 - b_1 \bar{X} = 0$$

An alternate derivation

Suppose we turned things around and

- Insisted on these qualities
- Rather than serendipitiously obtaining them as a consequence

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} e_i = 0 \Rightarrow \frac{1}{n}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) \\ \Rightarrow \bar{Y} - b_0 - b_1 \bar{X} = 0 \\ \Rightarrow b_0 = \bar{Y} - b_1 \bar{X}\end{aligned}$$

An alternate derivation

Suppose we turned things around and

- Insisted on these qualities
- Rather than serendipitiously obtaining them as a consequence

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} e_i = 0 &\Rightarrow \frac{1}{n}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) \\
&\Rightarrow \bar{Y} - b_0 - b_1\bar{X} = 0 \\
&\Rightarrow b_0 = \bar{Y} - b_1\bar{X}
\end{aligned}
$$

- Gives us our intercept!

An alternate derivation

Suppose $0 = corr(e, X)$, then

$$\Rightarrow 0 = corr(e, X)$$

An alternate derivation

$$0 = corr(e, X) = \frac{1}{SD(X)SD(e)} \cdot \sum_{i=1}^{n} (e_i - \bar{e})(X_i - \bar{X})$$

An alternate derivation

$$0 = corr(e, X) = \frac{1}{SD(X)SD(e)} \cdot \sum_{i=1}^{n} (e_i - \bar{e})(X_i - \bar{X}) = \sum_{i=1}^{n} e_i(X_i - \bar{X})$$

An alternate derivation

$$0 = corr(e, X) = \frac{1}{SD(X)SD(e)} \cdot \sum_{i=1}^{n} (e_i - \bar{e})(X_i - \bar{X}) = \sum_{i=1}^{n} e_i (X_i - \bar{X})$$

$$\Rightarrow 0 = \sum_{i=1}^{n} e_i (X_i - \bar{X})$$

An alternate derivation

$$0 = corr(e, X) = \frac{1}{SD(X)SD(e)} \cdot \sum_{i=1}^{n} (e_i - \bar{e})(X_i - \bar{X}) = \sum_{i=1}^{n} e_i (X_i - \bar{X})$$

$$\Rightarrow 0 = \sum_{i=1}^{n} e_i (X_i - \bar{X})$$

$$= \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)(X_i - \bar{X})$$

An alternate derivation

$$0 = corr(e, X) = \frac{1}{SD(X)SD(e)} \cdot \sum_{i=1}^{n}(e_i - \bar{e})(X_i - \bar{X}) = \sum_{i=1}^{n} e_i(X_i - \bar{X})$$

$$\Rightarrow 0 = \sum_{i=1}^{n} e_i(X_i - \bar{X})$$

$$= \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)(X_i - \bar{X})$$

$$= \sum_{i=1}^{n}(Y_i - (\bar{Y} - b_1\bar{X}) - b_1 X_i)(X_i - \bar{X})$$

[note: replace b_0 with b_1 solved for earlier]

An alternate derivation

$$0 = corr(e, X) = \frac{1}{SD(X)SD(e)} \cdot \sum_{i=1}^{n}(e_i - \bar{e})(X_i - \bar{X}) = \sum_{i=1}^{n} e_i(X_i - \bar{X})$$

$$\begin{aligned}
\Rightarrow 0 &= \sum_{i=1}^{n} e_i(X_i - \bar{X}) \\
&= \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)(X_i - \bar{X}) \\
&= \sum_{i=1}^{n}(Y_i - (\bar{Y} - b_1\bar{X}) - b_1 X_i)(X_i - \bar{X}) \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y} - b_1(X_i - \bar{X})(X_i - \bar{X}))
\end{aligned}$$

An alternate derivation

$$0 = corr(e, X) = \frac{1}{SD(X)SD(e)} \cdot \sum_{i=1}^{n}(e_i - \bar{e})(X_i - \bar{X}) = \sum_{i=1}^{n} e_i(X_i - \bar{X})$$

$$\Rightarrow 0 = \sum_{i=1}^{n} e_i(X_i - \bar{X})$$

$$= \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)(X_i - \bar{X})$$

$$= \sum_{i=1}^{n}(Y_i - (\bar{Y} - b_1\bar{X}) - b_1 X_i)(X_i - \bar{X})$$

$$= \sum_{i=1}^{n}(Y_i - \bar{Y} - b_1(X_i - \bar{X})(X_i - \bar{X}))$$

Solving $\Rightarrow b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = r_{XY}\frac{S_Y}{S_X}$

An alternate derivation

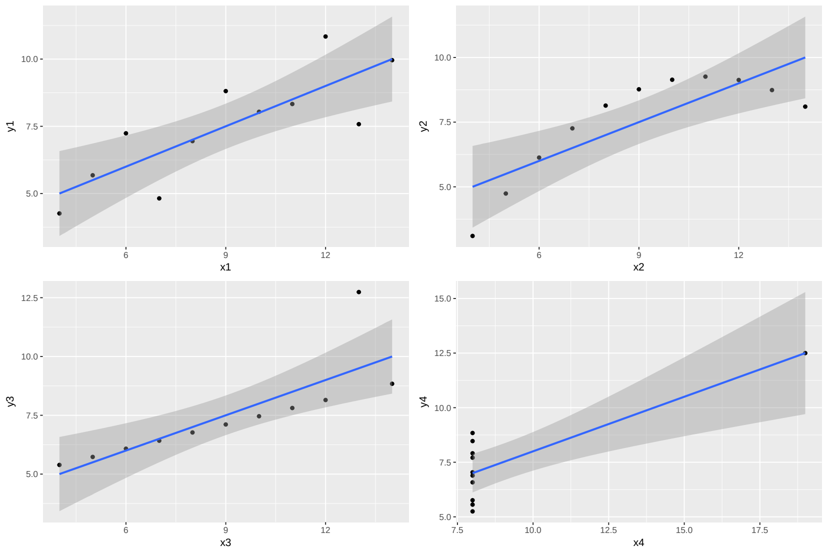
$$0 = corr(e, X) = \frac{1}{SD(X)SD(e)} \cdot \sum_{i=1}^{n}(e_i - \bar{e})(X_i - \bar{X}) = \sum_{i=1}^{n} e_i(X_i - \bar{X})$$

$$\begin{aligned}
\Rightarrow 0 &= \sum_{i=1}^{n} e_i(X_i - \bar{X}) \\
&= \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)(X_i - \bar{X}) \\
&= \sum_{i=1}^{n}(Y_i - (\bar{Y} - b_1\bar{X}) - b_1 X_i)(X_i - \bar{X}) \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y} - b_1(X_i - \bar{X})(X_i - \bar{X}))
\end{aligned}$$

$$\text{Solving } \Rightarrow b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = r_{XY}\frac{s_Y}{s_X}$$

- Gives us our slope!

Model Fit: When is your line good enough?



Model Fit: When is your line good enough?

- What is simplest baseline ($\bar{Y}$)
  - (Remember) Null hypothesis testing?
    - $\bar{Y} = 1215.925$ (in \$1000s)
  - We can compare our model (SLR) to a simple mean model ($\bar{Y}$)
- We can also think of this as variance decomposition of Y
  - $\Rightarrow$ ANOVA

Decomposing the Variance

How well does the least squares line explain variation in Y?

Since $\hat{Y}$ and e are independent (i.e. $cov(\hat{Y}, e) = 0$),

$$var(Y) = var(\hat{Y} + e) = var(\hat{Y}) + var(e)$$

This leads to ANOVA for regression:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2$$

Decomposing the Variance

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\substack{\text{Total}\\ \text{Sum of Squares}\\ \text{(SST)}}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\substack{\text{Regression}\\ \text{Sum of Squares}\\ \text{(SSR)}}} + \underbrace{\sum_{i=1}^{n} e_i^2}_{\substack{\text{Error}\\ \text{Sum of Squares}\\ \text{(SSE)}}}$$

SSR: Variation in Y explained by the regression line

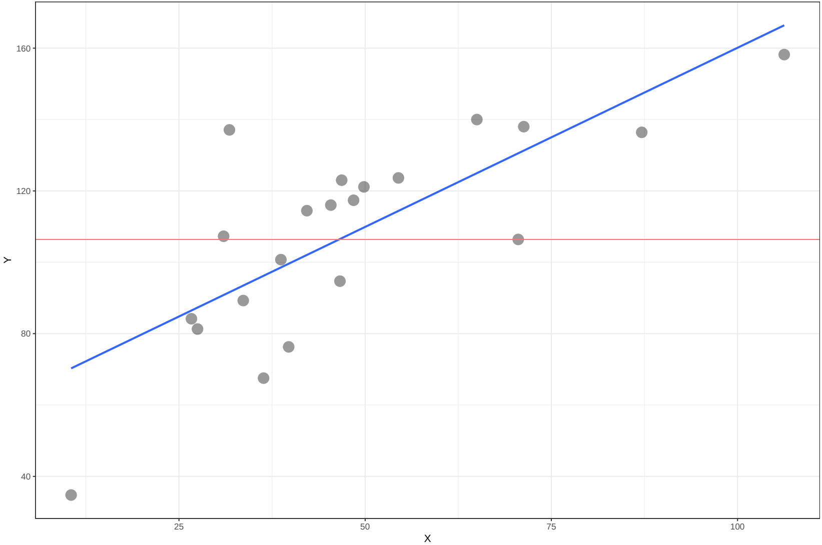
SSE: Variation in Y that is left unexplained

$$SSR = SST \Rightarrow \text{ perfect fit}$$

Be careful of similar acronyms, e.g. SSR for “residual” SS

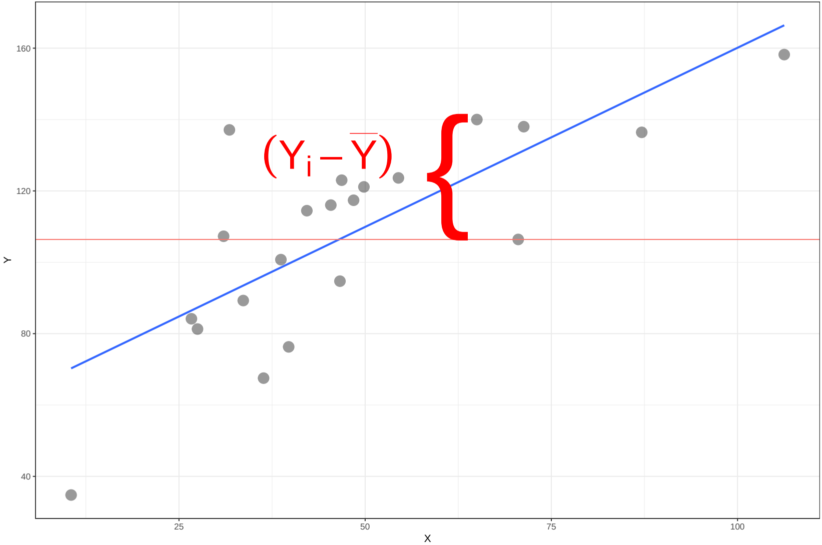
Decomposing the Variance

How does that breakdown look on a scatterplot?



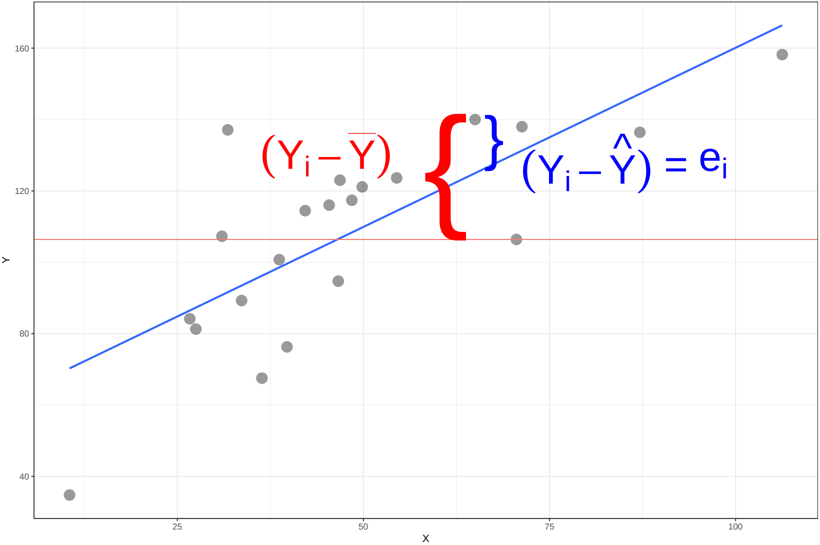
Decomposing the Variance

How does that breakdown look on a scatterplot?



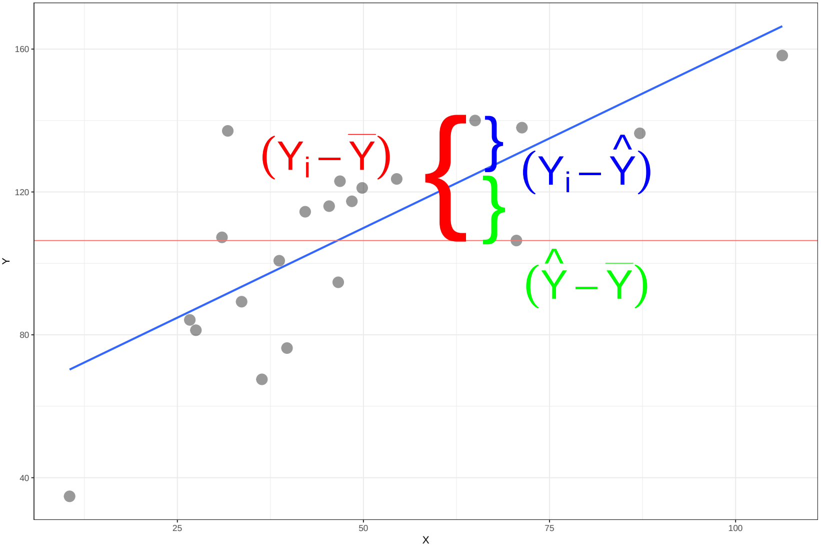
Decomposing the Variance

How does that breakdown look on a scatterplot?



Decomposing the Variance

How does that breakdown look on a scatterplot?



A goodness of fit measure: R^2

The coefficient of determination, denoted by R^2, measures goodness-of-fit:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- $0 < R^2 < 1$
- The closer R^2 is to 1, the better the fit

A goodness of fit measure: R^2

An interesting fact: $R^2 = r^2_{XY}$

- i.e., R^2 is squared correlation

$$
\begin{aligned}
R^2 &= \frac{\sum_{i=1}^n (\hat{Y} - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \frac{\sum_{i=1}^n (b_0 + b_1 X_i - b_0 - b_1 \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \frac{b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{b_1^2 S_X^2}{S_Y^2} = r^2_{xy}
\end{aligned}
$$

- No surprise: the higher the sample correlation between X and Y, the better you are doing in your regression

Summarizing/Back to the Housing Data

Source	df	Sum Sq	Mean Sq	F-value	p-va
Regress	p-1	SSR	SSR/(p-1)	(MSR/MSE)	p*
Error	n-p	SSE	SSE/(n-p)		
Total	n-1	SST	SST/(n-1)		

Summarizing/Back to the house data

```r
lm <- lm(Zillow_Estimate ~ House_Size_SqFt, data = sf_house)
anova(lm)
```

```
Analysis of Variance Table

Response: Zillow_Estimate
                 Df   Sum Sq  Mean Sq F value    Pr(>F)
House_Size_SqFt   1 20158987 20158987   294.9 < 2.2e-16 ***
Residuals       202 13808267    68358
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Careful:
  - Residuals SS is our Error SS (SSE)
  - Size SS is our Regression SS (SSR)

Summarizing/Back to the house data

```
Call:
lm(formula = Zillow_Estimate ~ House_Size_SqFt, data = sf_house)

Residuals:
    Min      1Q  Median      3Q     Max 
-753.61 -162.70   -7.83  165.67 1009.50 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       449.00      48.27   9.303   <2e-16 ***
House_Size_SqFt   495.48      28.85  17.173   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 261.5 on 202 degrees of freedom
Multiple R-squared:  0.5935,    Adjusted R-squared:  0.5915 
F-statistic: 294.9 on 1 and 202 DF,  p-value: < 2.2e-16
```

Exercise

- Use Household Size to predict Last Sold Price for the Zillow House data (ANOVA Test)
- Google Sheets Example
- Google Sheets in Class Assignment
  - Compute $(Y_i - \bar{Y})^2$ and $(\hat{Y}_i - \bar{Y})^2$
  - Fill in the ANOVA table and R^2 from our previous work

Source	df	Sum Sq	Mean Sq	F-value	p-va
Regress	p-1	SSR	SSR/(p-1)	(MSR/MSE)	p*
Error	n-p	SSE	SSE/(n-p)		
Total	n-1	SST	SST/(n-1)		

$$R^2 = SSR/SST$$

Prediction and the modeling goal

Slides for SLR as a Probability Model

Prediction and the modeling goal

A prediction rule is any function where you input X and it outputs $\hat{Y}$ as a predicted response at X

The least squares line is a prediction rule:

$$\hat{Y} = f(x) = b_0 + b_1 X$$

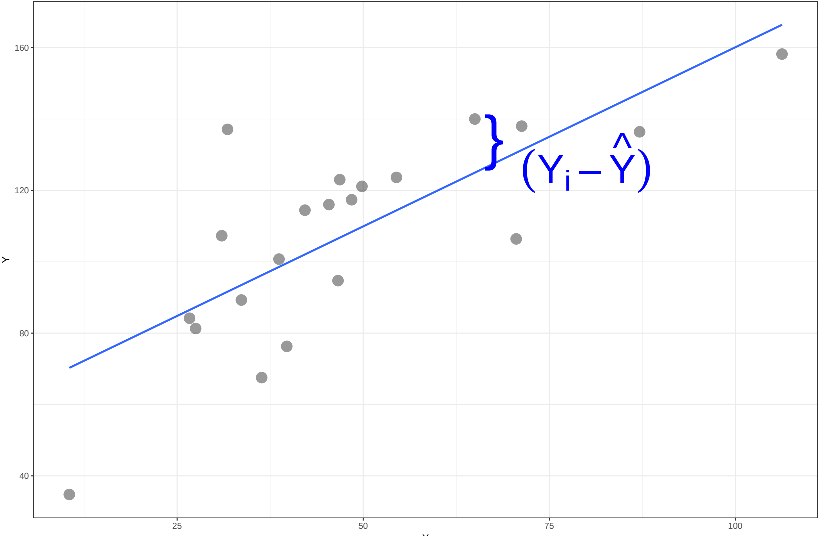
This rule tells us what to do when a new X comes along

- Run it through the formula above and obtain a guess $\hat{Y}$

Prediction and the modeling goal

$\hat{Y}$ is just what we expect for a given X

- It is not going to be a perfect prediction



Prediction and the modeling goal

We need to devise a notion of forecast accuracy

- How sure are we about our forecast?
- Or how different could Y be from what we expect?

Forecasts are useless without some kind of uncertainty qualification/quantification

One method is to specify a range of Y values that are likely, given an X value

- A prediction interval: a probable range for Ys given X

Prediction and the modeling goal

Key insight: to construct a prediction interval, we will have to assess the likely range of residual values corresponding to a Y value that has not yet been observed!

We must "invest" in a **probability model** (e.g., a normal distribution)

- Only then we can say something like "with 95% probability the residuals (the error term) will be no less than -$50,000 or larger than $50,000"

We must also acknowledge that the "fitted" line may be fooled by particular realizations of the residuals

- I.e., that our estimated coefficients b_0 & b_1 are random

SLR model

Here it is:

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Similar but with important differences

- It is a model, so we are assuming this relationship holds for some **true but unknown** values of β_0, β_1
- Greek letters remind us they are *not* the same as the LS estimates of b_0 and b_1

The error ϵ is independent, additive, "idosyncratic noise"

- Its distribution is *known* up to its spread σ^2
- Greek letters remind us that ϵ is not the same as e

SLR model

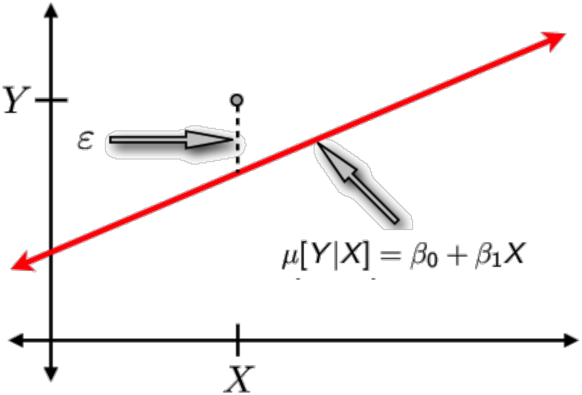
Why do we have $\epsilon \sim N(0, \sigma^2)$?

- $\mu[\epsilon] = 0 \iff \mu[Y|X] = \beta_0 + \beta_1 X$
  ($\mu[Y|X]$ is "conditional expectation of Y given X")
- Many distributions are close to normal (central limit theorem)
- MLE estimates for β's are the same as the LS b's, giving us a handle on computation
- We can estimate the spread of σ^2
- It works! *This is a very robust model!*

SLR model

Before looking at any data, the model specifies

- How Y varies with X on average: $\mu[Y|X] = \beta_0 + \beta_1 X$
- And the influence of factors other than X, $\epsilon \sim N(0, \sigma^2)$ independently of X



SLR model

Context from the house data example

Think of $\mu[Y|X]$ as the average price of houses with X, and σ^2 is the spread around that average

When we specify the SLR model we say that

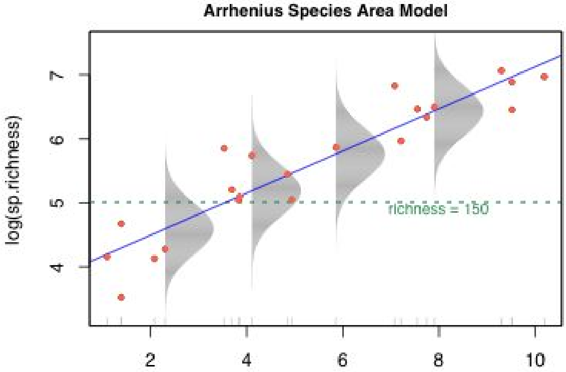
- The average house price is linear in its size, but we don't know the coefficients
- Some houses could have a higher than expected value, some lower, but the amount by which they differ from average is unknown but
  - Is independent of the size
  - And is normal

SLR model

We think about the data as being one possible realization of data that *could* have been **generated from the model**

$$Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$$

- σ^2 controls the dispersion of Y around $\beta_0 + \beta_1 X$



Prediction intervals in the true model

You are told (without looking at the data) that

$$\beta_0 = 40;\ \beta_1 = 45;\ \sigma = 10$$

and you are asked to predict the price of a 1500 SqFt house

What do you know about Y from the model?

$$\begin{aligned} Y &= 40 + 45(1.5) + \epsilon \\ &= 107.5 + \epsilon \end{aligned}$$

Thus our prediction for price is

$$Y \sim N(107.5, 10^2)$$

Prediction intervals in the true model

The model $(449 + 495.48X)$ says that the mean value of a 1000 SqFt house is \$495931.06 and the deviation from mean is within $\approx$ \$28808.114

We are 95% sure that

- $\$ - 56463.904 < \epsilon < \$56463.904$
- $\$439127.88 < Y < \$552734.25$

In general, the 95% Prediction Interval is $PI = \beta_0 + \beta_1 X \pm 1.96\sigma_{\hat{Y}}$

- Note that here I am using a t-distribution with $df = 202$, such that $495931.064 \pm 1.9717774 \cdot 28808.114$

SE of $\hat{Y}$

- SE of $\hat{y}$ at x is:

$$\hat{\sigma}_{\hat{Y}} = SE(\hat{Y}) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where $\hat{\sigma} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y})^2}$ (Also known as $rMSE$)

- Example $x = 1000$
- $\hat{Y} = \beta_0 + \beta_1 x = 449 + 495.48 \times 1000 = 495929$
- $\hat{\sigma} = 261.45$
- $\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{996906.73}{82.11}$
- $\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = 110.19$
- $\hat{\sigma}_{\hat{Y}} = 28809.3$
- 95% CI: $495929 \pm qt(1 - .05/2, n-2) \times 28809.3$
- 95% CI: $495929 \pm 1.97 \times 28809.3$

Summary of Simple Linear Regression

Assume that all observations are drawn from our regression model and that errors on those observations are independent

The model is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

wher ϵ_i is independent and identically distributed $N(0, \sigma^2)$

The SLR has 3 basic parameters

- β_0, β_1 (linear pattern)
- σ (variation around the line)

Summary of Simple Linear Regression

What are the key characterstics of SLR?

- I.e., how do we describe the model in words?

We assume that

- The mean of Y is linear in X
- The error terms (deviations from the line) are *normally distributed*
  - Very few deviations are more than 2 sd away from the regression mean
- And the error terms have constant variance

Extra Notes

Conditional vs Marginal Distributions

More on the conditional distribution:

$Y|X \sim N(\mu(Y|X], Var(Y|X))$

- Mean is $\mu[Y|X] = \mu[\beta_0 + \beta_1 X + \epsilon|X] = \beta_0 + \beta_1 X$
- Variance is

$$Var(Y|X) = Var(\beta_0 + \beta_1 X + \epsilon|X) = Var(\epsilon) = \sigma^2$$

- $\sigma^2 < Var(Y)$ if X and Y are related

And the ANOVA for regression:

- The bigger $[1 - \sigma^2/Var(Y)]$, the more X matters!

R Addendum

- Key R Code:

```r
help(lm)
help(predict.lm)
help(anova)
help(plot.lm)
help(package = "ggplot2")
```

Example of R SLR: In Class Exercise

- R Studio

Example of R SLR: In Class Exercise

```
## Read Data and load ggplot2
library(ggplot2)
hdata <- read.csv("https://goo.gl/F2q8Vy")
```

Example of R SLR: In Class Exercise

```
ggplot(hdata, aes(y = ZillowEstimate, x = houseSize..sq.ft.)) + geom_point(shape = 1) +
    geom_smooth(method = lm)
```

Example of R SLR: In Class Exercise

```r
## Redo Minus top 50
hdata2 <- hdata[-order(hdata$ZillowEstimate, decreasing = TRUE)[1:50], ]

ggplot(hdata2, aes(y = ZillowEstimate, x = houseSize..sq.ft.)) + geom_point(shape = 1) +
    geom_smooth(method = lm)
```

Example of R SLR: In Class Exercise

```
lm1 <- lm(ZillowEstimate ~ houseSize..sq.ft., data = hdata)
summary(lm1)
```

Example of R SLR: In Class Exercise

```
anova(lm1)
```

Example of R SLR: In Class Exercise

```
## Predict house price if 500 sq ft, 1000 sq ft, 1500 sq ft
predict(lm1, data.frame(houseSize..sq.ft. = c(500, 1000, 1500)), interval = "confidence",
    level = 0.05)
```

Example of R SLR: In Class Exercise

```r
## Predict house price if 500 sq ft, 1000 sq ft, 1500 sq ft
predict(lm1, data.frame(houseSize..sq.ft. = c(500, 1000, 1500)))
```

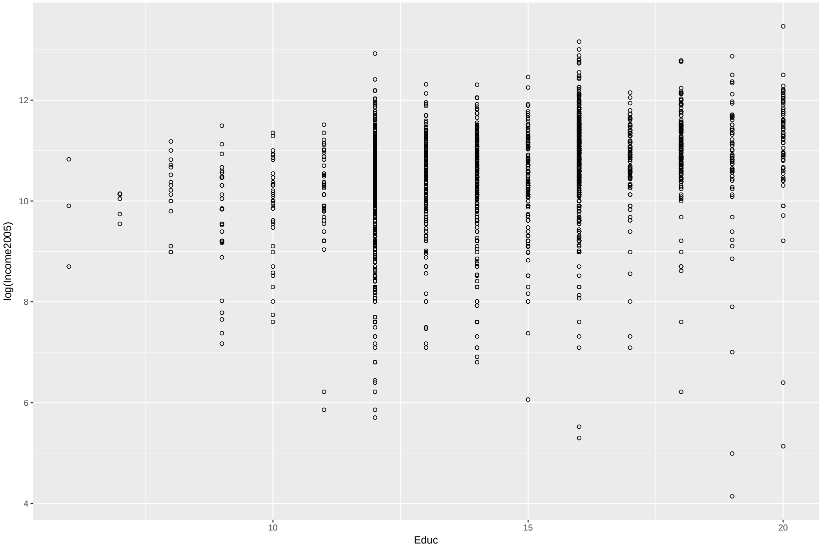
Extra Example of R SLR: Data

```r
library(Sleuth3)
data(ex0828)
head(ex0828)
```

```
  Subject   AFQT Educ Income2005
1       2  6.841   12       5500
2       6 99.393   16      65000
3       7 47.412   12      19000
4       8 44.022   14      36000
5       9 59.683   14      65000
6      13 72.313   16       8000
```

Example of R SLR: Validity of the regression model

```
ggplot(ex0828, aes(y = log(Income2005), x = Educ)) + geom_point(shape = 1) +
    geom_smooth(method = lm)
```



Example of R SLR: Data

```r
lm_1 <- lm(log(Income2005) ~ Educ, data = ex0828)
summary(lm_1)
```

```
Call:
lm(formula = log(Income2005) ~ Educ, data = ex0828)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.8671 -0.3487  0.1441  0.5772  2.6981 

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)    
(Intercept) 8.880995   0.103473   85.83 <0.0000000000000002 ***
Educ        0.112067   0.007331   15.29 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9328 on 2582 degrees of freedom
Multiple R-squared:  0.08299,	Adjusted R-squared:  0.08264 
F-statistic: 233.7 on 1 and 2582 DF,  p-value: < 0.00000000000000022
```