A Three Day Course in Applied Regression Analysis

[1] Understanding data, conditional distributions, estimation and prediction in linear regression

Matt Taddy, University of Chicago Booth School of Business

faculty.chicagobooth.edu/matt.taddy/teaching

Goals

We're here to make sure you have the necessary tools for real-world business intelligence and data analysis.

- Solid understanding of the essential statistics principles.
- Concrete analysis ability and best-practice guidelines.

Set-Up

Our three days will be a constant mix of review and lecture.

- I'll be covering the essential core of regression analysis.
- Optional ungraded homework problems are available online and we'll answer them together in class.
- We'll work together to get the most out of these sessions.

Ask questions!

Computing

Your regression class used Minitab, future classes/work will use a variety (Stata, Eviews, Matlab), and I use R.

It shouldn't matter.

All of the different software ouputs are similarly easy to understand, you can use whatever you want at home, and we'll be going over examples together on the whiteboard.

The key to these three sessions will be practical understanding.

But...

Computing with R

If you want, take advantage of this opportunity to learn R.

This is the real deal: industrial strength software for data analysis. It's free, cross platform, and widely used.

You can download R from www.r-project.org. Links, video tutorials, and NY times article are up on my web-page.

Computing with R

The barrier for R is its command line interface:

you type commands to get what you want.

All of the code for lectures and homework will be available online. The best way to learn software is through imitation.

A good book is *A Modern Approach to Regression with R*, by Sheather. For other options, just search 'R' at `Amazon.com`.

The Super Quick Guide to R

To load data, type `mydata = read.csv('mydata.csv')`.
Any file in excel can be saved as '.csv', and vice versa.

`mydata` is then a matrix with names. You can access variables with `$` or by matrix indices: e.g., `mydata$Y` or `mydata[,1]`.

To see the 8^{th} observation, type `mydata[8,]`. To see the first variable for this observation, do `mydata[8,1]`.

The Super Quick Guide to R

All calculator functions are as you'd expect (e.g., `*, /, +,-`).

Use functions to create new variables: `lY = log(mydata$Y)`.

And add them to your data: `mydata$lY = lY`.

R is all about assigning names, and `<-` works the same as `=`.

Plotting is super intuitive:

Use `plot(mydata$X, lY)` or `plot(lY ~ mydata$X)`

The Super Quick Guide to R

To do regression in R, you only need one command:

```
reg = lm( Y ~ X, data=mydata)
```

The object `reg` is a list of useful things (type `names(reg)`).

For example, we'll often want to see

```
plot(reg$residuals ~ reg$fitted).
```

Finally, `summary(reg)` prints almost everything you need.

So, now that you know all about R...

Regression in a nutshell

Regression is always and only the construction of a model for how some response Y changes as a function of covariates $\mathbf{X}$.

- Choosing $\mathbf{X}$: you want all important variables, but unnecessary X's lead to bad predictions.
- Estimation of model parameters: Least Squares.
- Quantifying uncertainty: you need error bounds around your predictions to understand what is at risk. This step relies upon your making model assumptions.

Regression Model

$Y =$ response or outcome variable

$X1, X2, X3, \ldots, Xp =$ covariates or input variables

The general relationship approximated by:

$$Y = f(X_1, X_2, \ldots, X_p) + e$$

And a linear relationship is written

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p + e$$

Regression: Prediction and Estimation

Competing aims abound:

- Hypothesis testing and response prediction
- Chasing R^2 but avoiding overcomplicated models
- Selecting variables and satisfying model assumptions
- Colinearity, ommited variables, and the economist

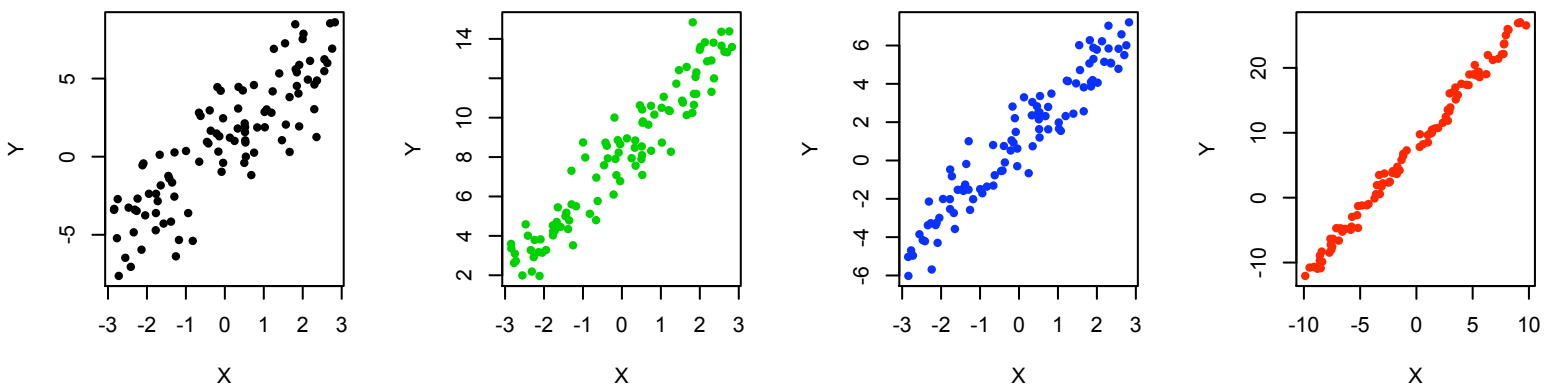
A good practical strategy: find and fit the model that is best for prediction, and make inferences by interpreting this model.

A preference for simplicity

Where to Start: Understand your Data

It is easy to get lost in the statistics details and forget what you already know about your data and problem.

This is why it is so important to plot and explore your data.



The most basic unit in regression is the scatterplot.

Data Visualization

Consider some data about pickups for sale on Craigslist:

```
> data <- read.csv("pickup.csv")
> attach(data)
> names(data)
[1] "year"  "miles" "price" "make"
```

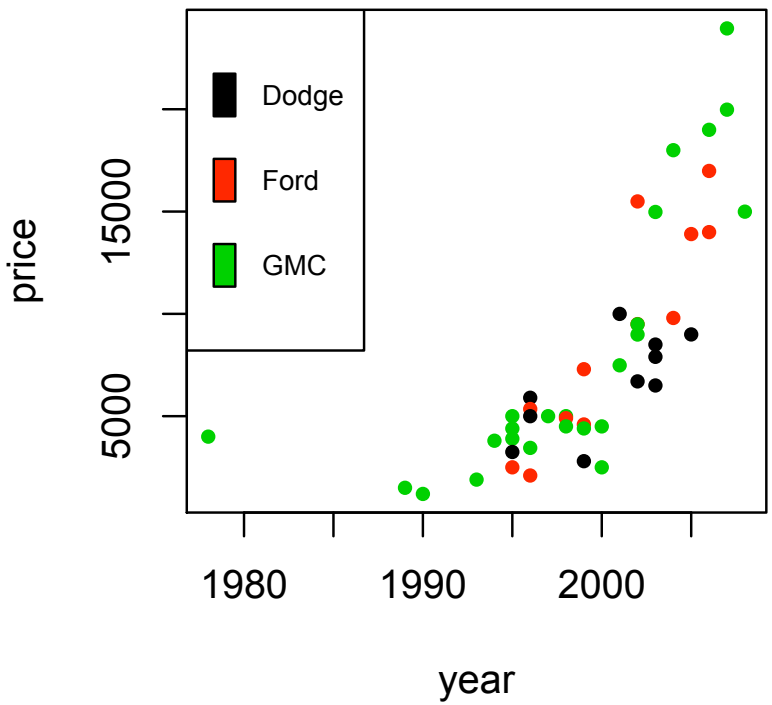
We have 4 dimensions to consider. A simple summary is

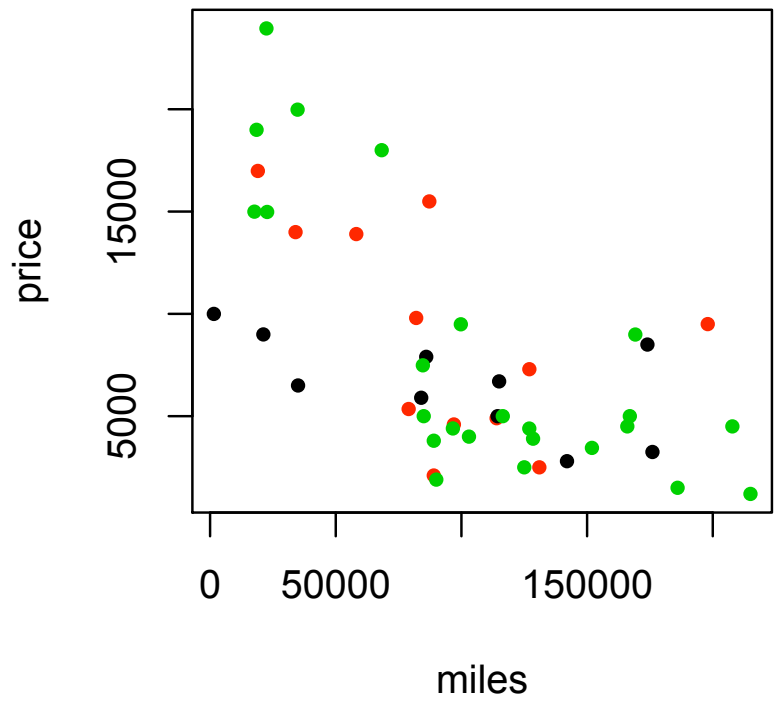
```
> summary(data)
      year          miles            price          make
 Min.   :1978   Min.   :  1500   Min.   : 1200   Dodge:10
 1st Qu.:1996   1st Qu.: 70958   1st Qu.: 4099   Ford :12
 Median :2000   Median : 96800   Median : 5625   GMC  :24
 Mean   :1999   Mean   :101233   Mean   : 7910
 3rd Qu.:2003   3rd Qu.:130375   3rd Qu.: 9725
 Max.   :2008   Max.   :215000   Max.   :23950
```

Data Visualization

We can plot and compare trucks over four dimensions.

```
> par(mfrow=c(1,2))
> plot(price ~ year, col=make, data=pickup, pch=20)
> plot(price ~ miles, col=make, data=pickup, pch=20)
> legend("topright", fill=c(1:3), legend=levels(pickup$make))
```

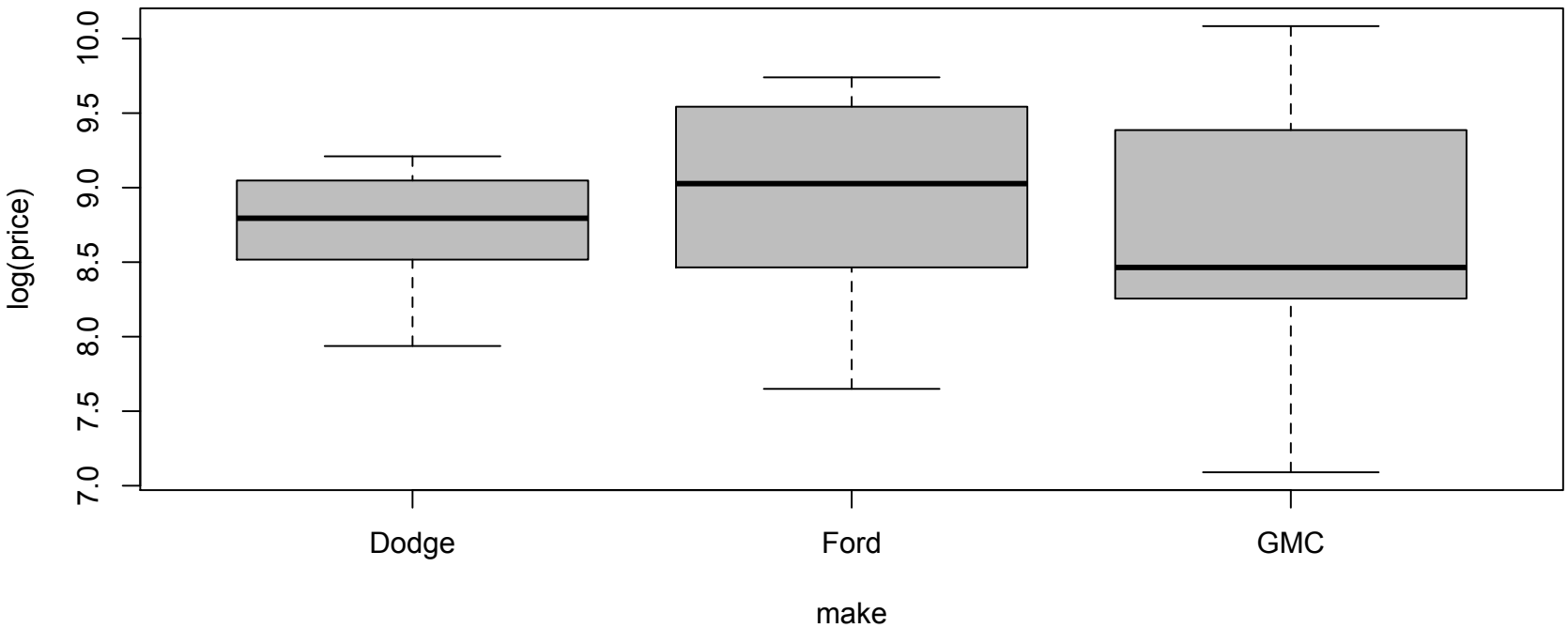




Data Visualization

For discrete factors, the scatterplot becomes a boxplot.

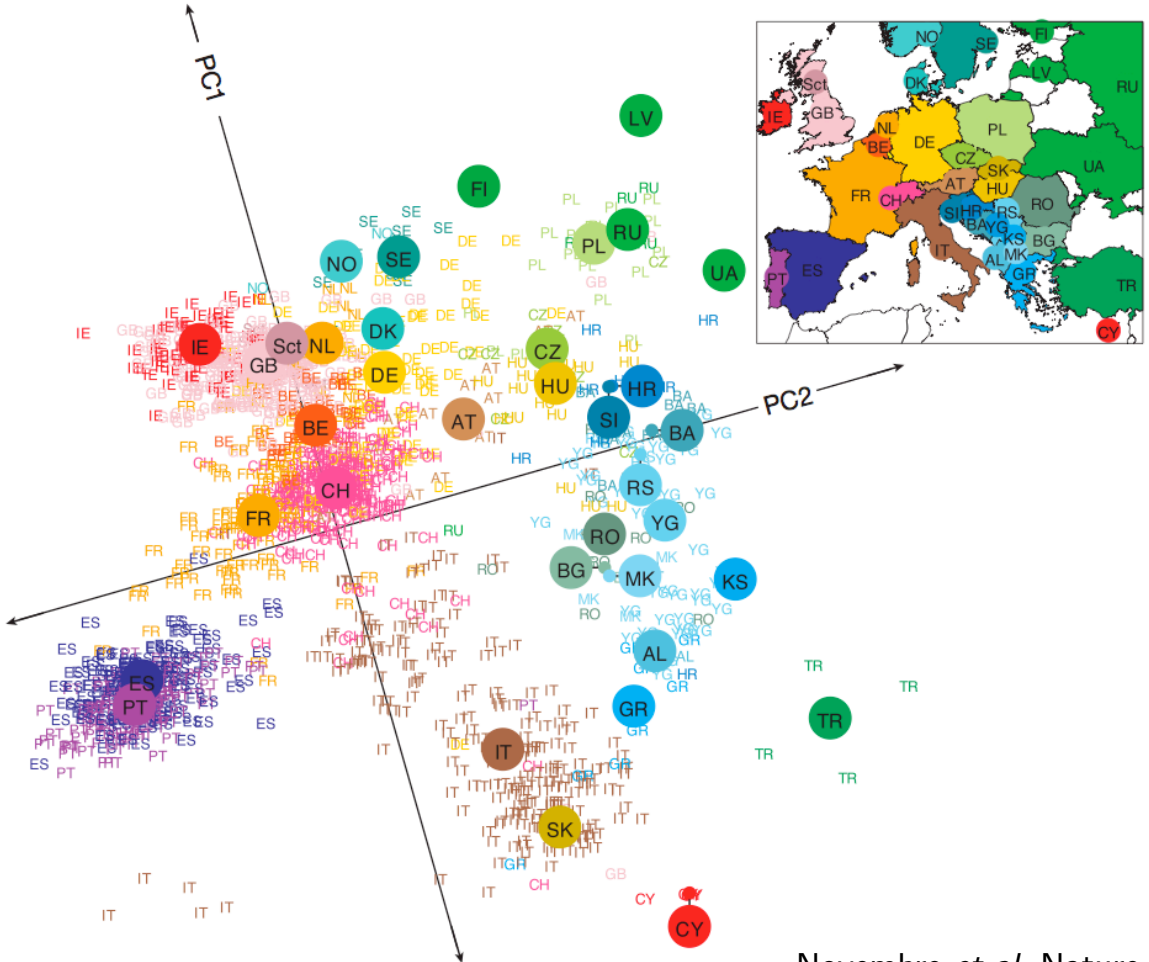
```
> plot(log(price) ~ make, data=pickup, col=8)
```



The box is the Interquartile Range (IQR; i.e., 25^{th} to 75^{th} %), with the median in bold. The whiskers extend to the most extreme point which is no more than 1.5 times the IQR width from the box.

Data Visualization

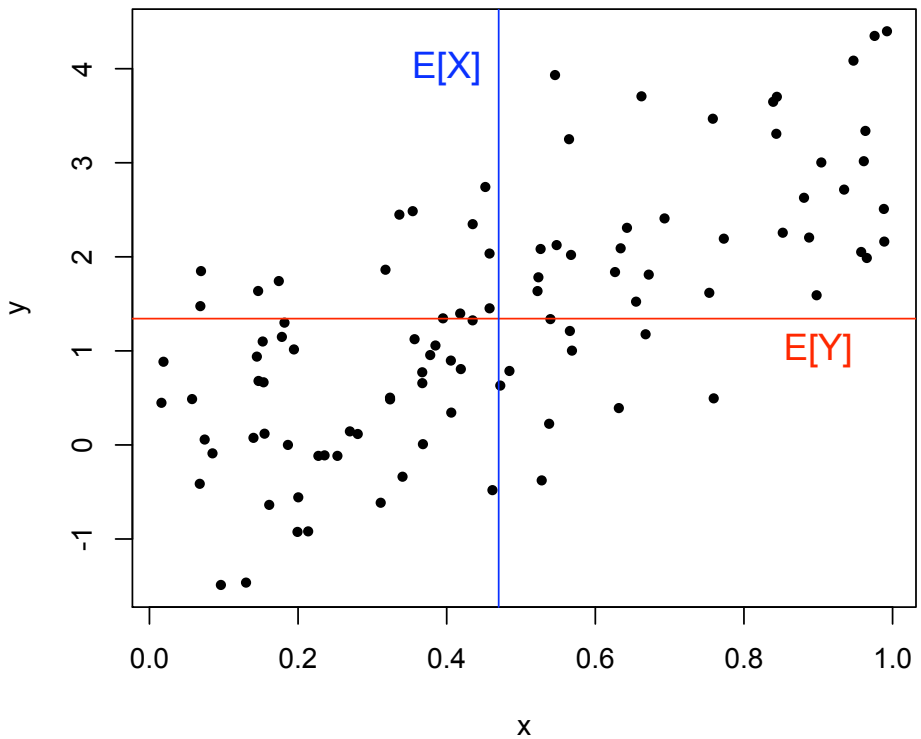
Data mining and the search for a scatterplot: finding reduced dimension representations for very high dimensional information.



Novembre *et al*, Nature 456 (2008)

Correlation and Covariance

$$\text{Covariance}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$



X and Y vary with each other around their means.

Correlation and Covariance

Correlation is the standardized covariance:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{\text{cov}(X, Y)}{\text{sd}(X)\text{sd}(Y)}$$

The correlation is scale invariant and the units of measurement don't matter: It is always true that $-1 \leq \text{corr}(X, Y) \leq 1$.

This gives the direction (- or +) and strength ($0 \to 1$)
of the linear relationship between X and Y.

Sample Correlation and Standard Deviation

Recall:

- Sample Covariance is $s_{xy} = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$.

  (in units X times units Y)

- Sample Standard Deviation is $s_x = \sqrt{\dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$.

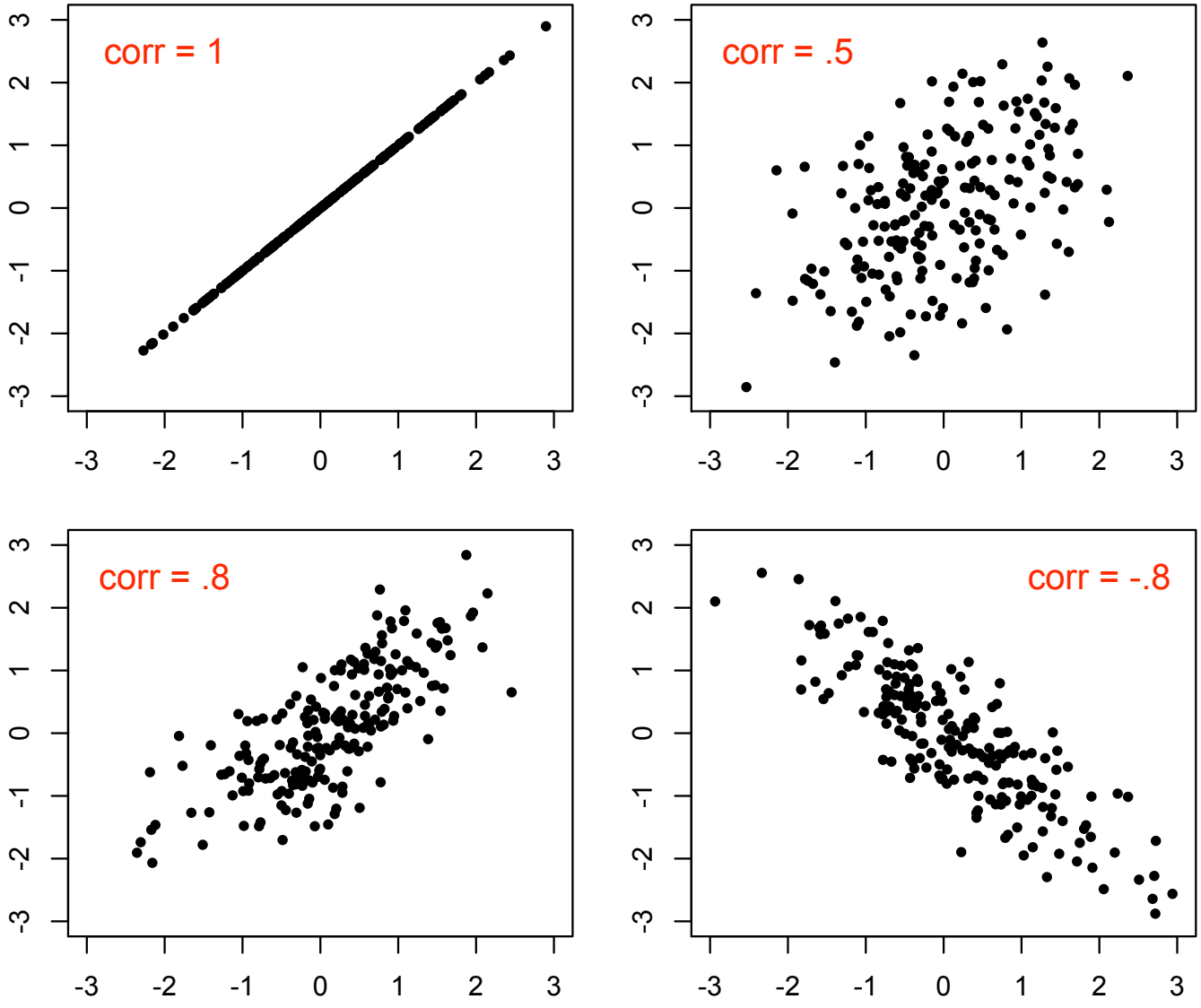
  (in units X)

- Sample Correlation is

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})}{s_x} \frac{(Y_i - \bar{Y})}{s_y}$$

(correlation is scale free!)

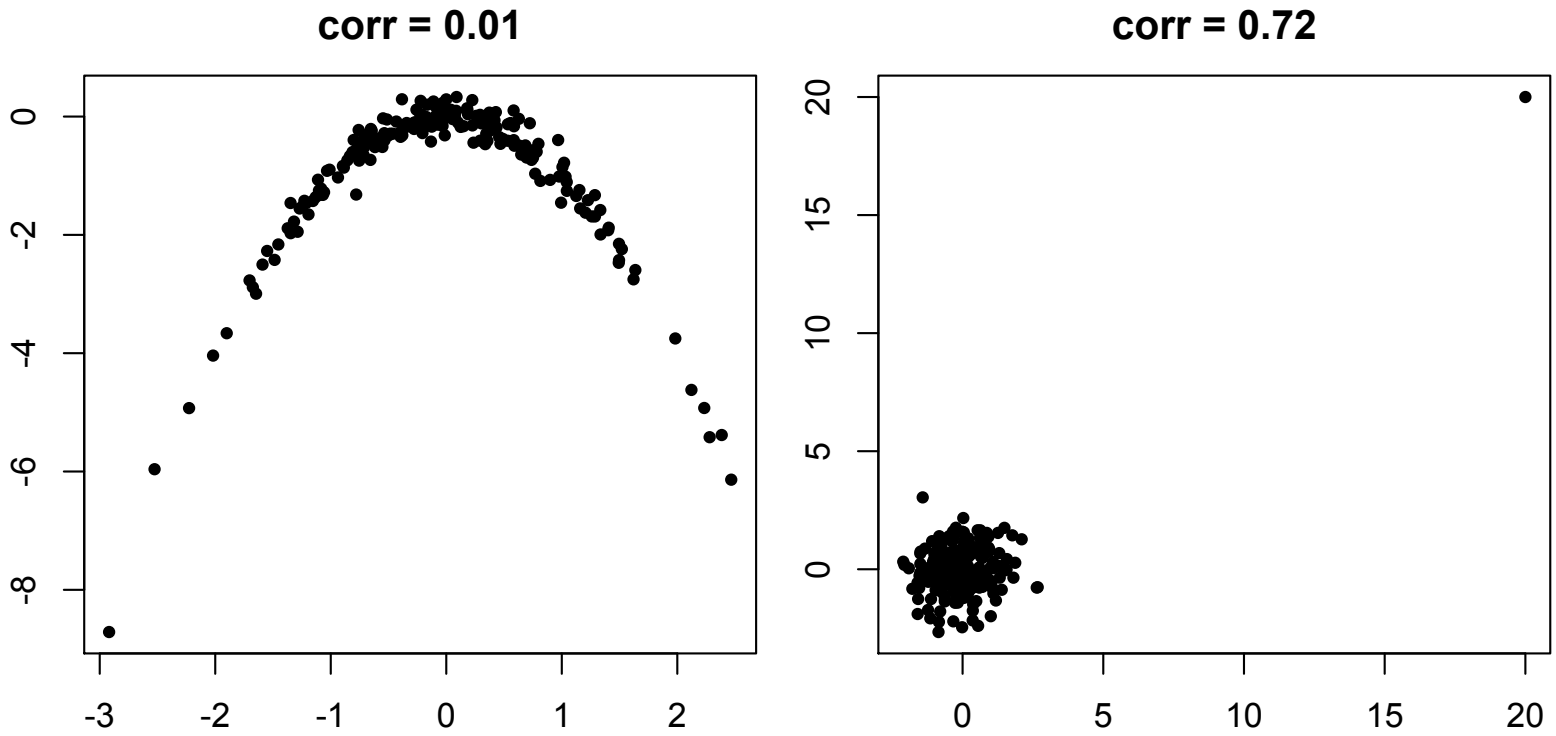
Correlation



Correlation

Only measures linear relationships:

$\text{corr}(X, Y) = 0$ does not mean the variables are not related!



Also be careful with influential observations.

Correlation and Regression

"Imagine" that $Y = b_0 + b_1 X + e$:

$$\begin{aligned} \operatorname{cov}(X, Y) &= \operatorname{cov}(X, b_0 + b_1 X + e) \\ &= \operatorname{cov}(X, b_1 X) \\ &= b_1 \operatorname{var}(X) \end{aligned}$$

$$\text{Thus } \operatorname{corr}(X, Y) = b_1 \frac{s_x}{s_y} \Leftrightarrow b_1 = r_{xy} \frac{s_y}{s_x}.$$

That is, b_1 is correlation times units Y per units X.

We can just choose b_0 to center the line: $b_0 = \bar{Y} - b_1 \bar{X}$.

1st Example: Predicting House Prices

Problem:

- Predict market price based on observed characteristics

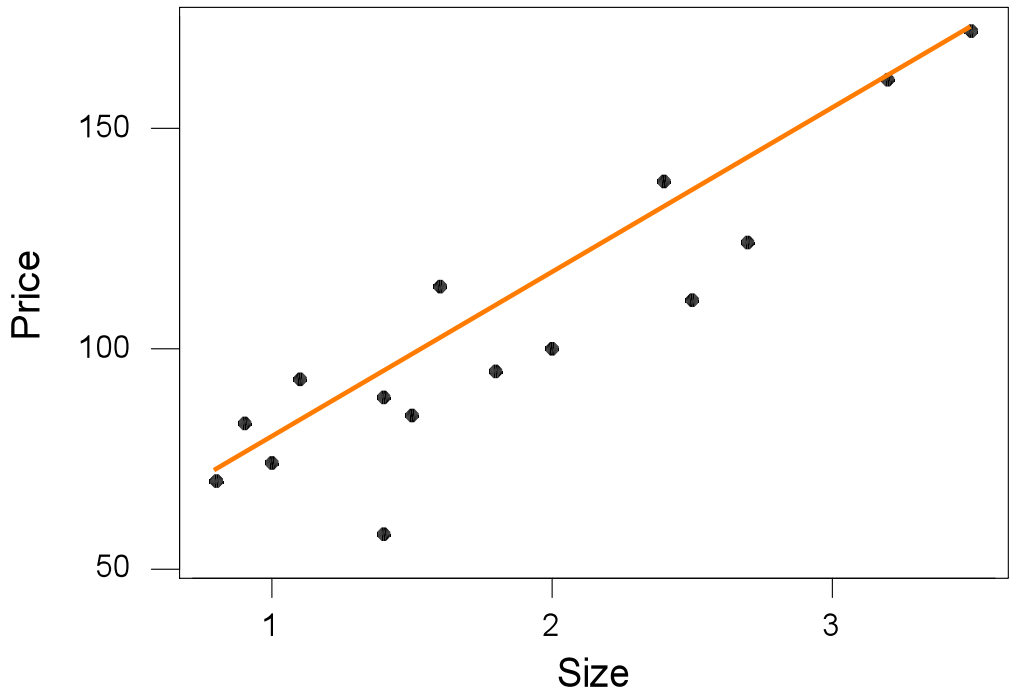
Solution:

- 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.

Linear Prediction

Appears to be a linear relationship between price and size:

As size goes up, price goes up.



The line shown was fit by the "eyeball" method.

Linear Prediction

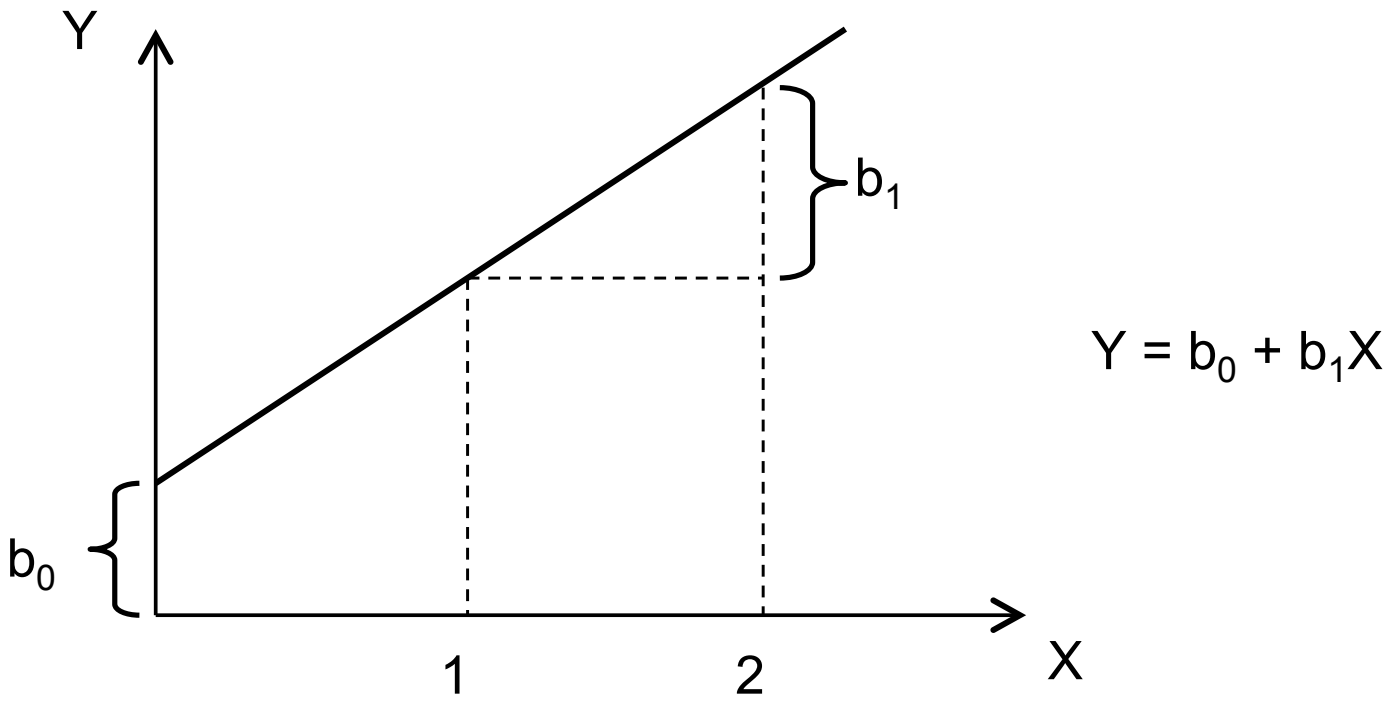
Recall that the equation of a line is:

$$Y = b_0 + b_1 X$$

Where b_0 is the intercept and b_1 is the slope.

The intercept value is in units of Y (\$1,000).
The slope is in units of Y *per* units of X (\$1,000/1,000 sq ft).

Linear Prediction



$Y = b_0 + b_1X$

Our "eyeball" line has $b_0 = 35$, $b_1 = 40$.

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 of size $X = 2.2$.

Predicted price $\hat{Y} = 35 + 40(2.2 * 1,000 \text{ sq ft}) = \$123,000$.

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

Linear Prediction

Can we do better than the eyeball method?

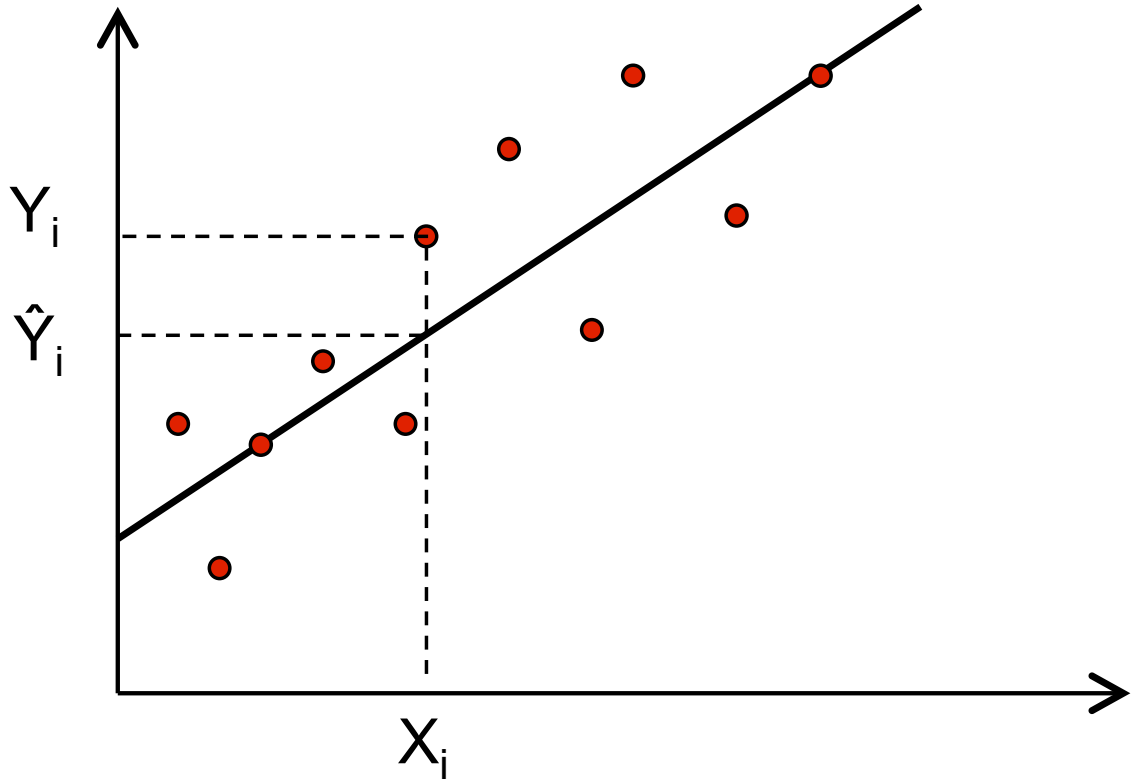
We desire a strategy for estimating the slope and intercept parameters in the model $\hat{Y} = b_0 + b_1 X$

A reasonable way to fit a line is to minimize the amount by which the fitted value differs from the actual value.

This amount is called the residual.

Linear Prediction

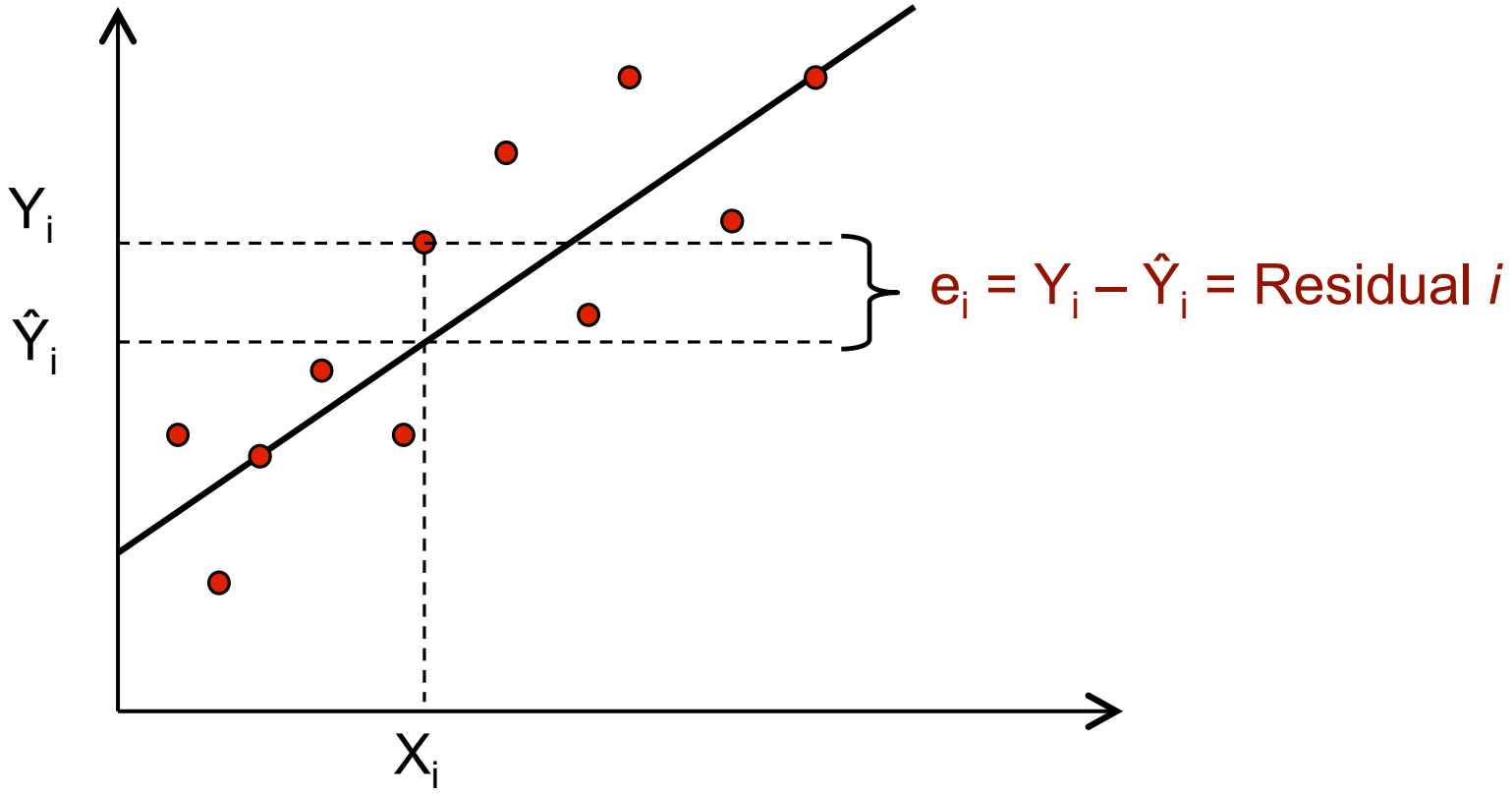
What is the "fitted value"?



The dots are the observed values and the line represents our fitted values given by $\hat{Y}_i = b_0 + b_1 X_1$.

Linear Prediction

What is the "residual"' for the ith observation'?



We can write $Y_i = \hat{Y}_i + (Y_i - \hat{Y}_i) = \hat{Y}_i + e_i$.

Least Squares

Ideally we want to minimize the size of all residuals:

- If they were all zero we would have a perfect line.
- Trade-off between moving closer to some points and at the same time moving away from other points.

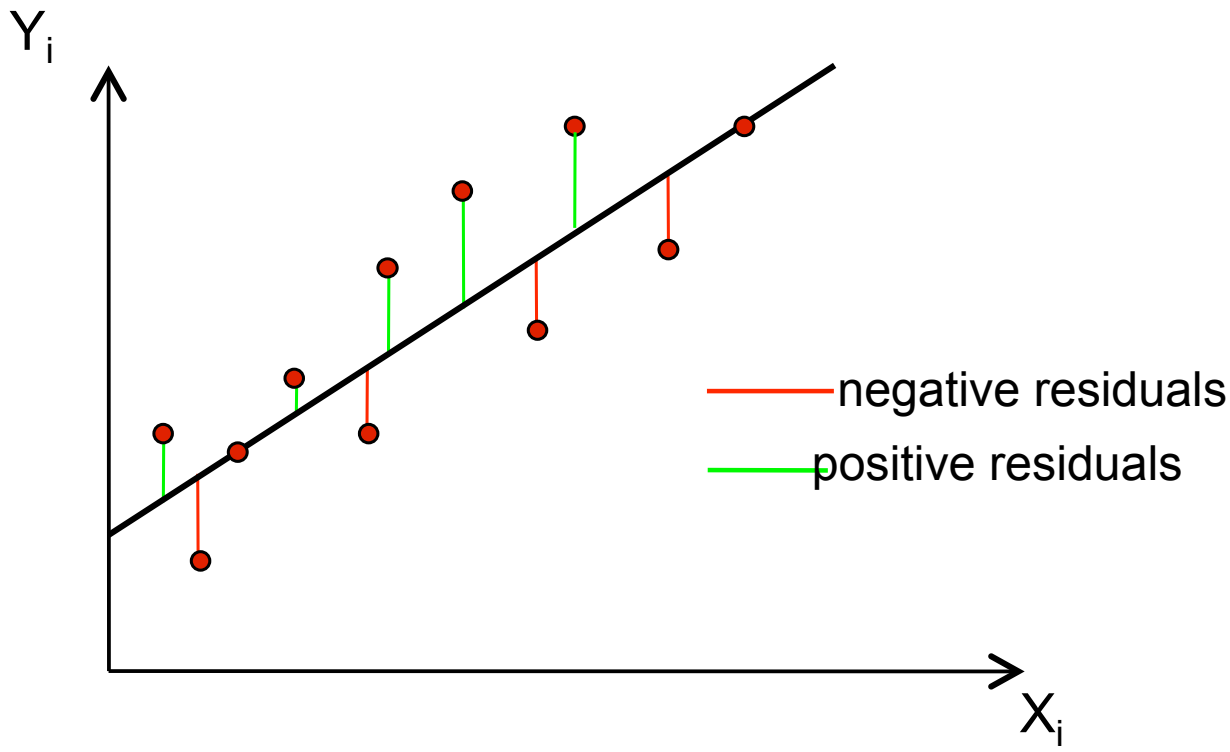
The line fitting process:

- Give weights to all of the residuals.
- Minimize the "total" of residuals to get best fit.

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,

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

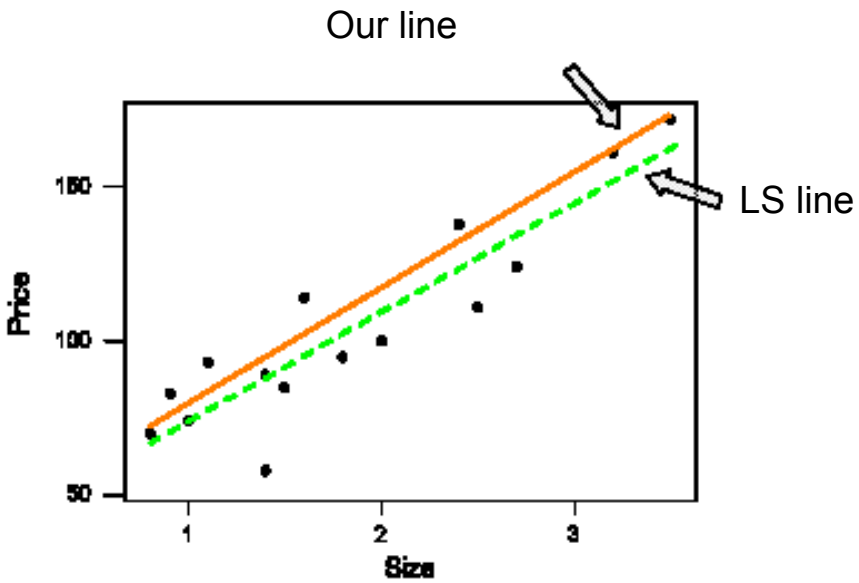
Least Squares

LS chooses a different line from ours:

```
> lm(price ~ size)

Call:
lm(formula = price ~ size)

Coefficients:
(Intercept)         size
      38.88        35.39
```



Least Squares

It turns out that we already know how to do least squares:

```
> b1 <- cor(price,size)*sd(price)/sd(size)
> b0 <- mean(price) - mean(size)*b1
> cbind(b0,b1)
           b0       b1
[1,] 38.88468 35.38596
```

Thus our scaled correlation and "means on the line" approach is the same as the least squares estimates.

Least Squares

To summarize:

R's `lm(Y ~ X)` function fits the "least squares" line $\hat{Y} = b_0 + b_1 X$ to minimize $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$.

The least squares formulas are

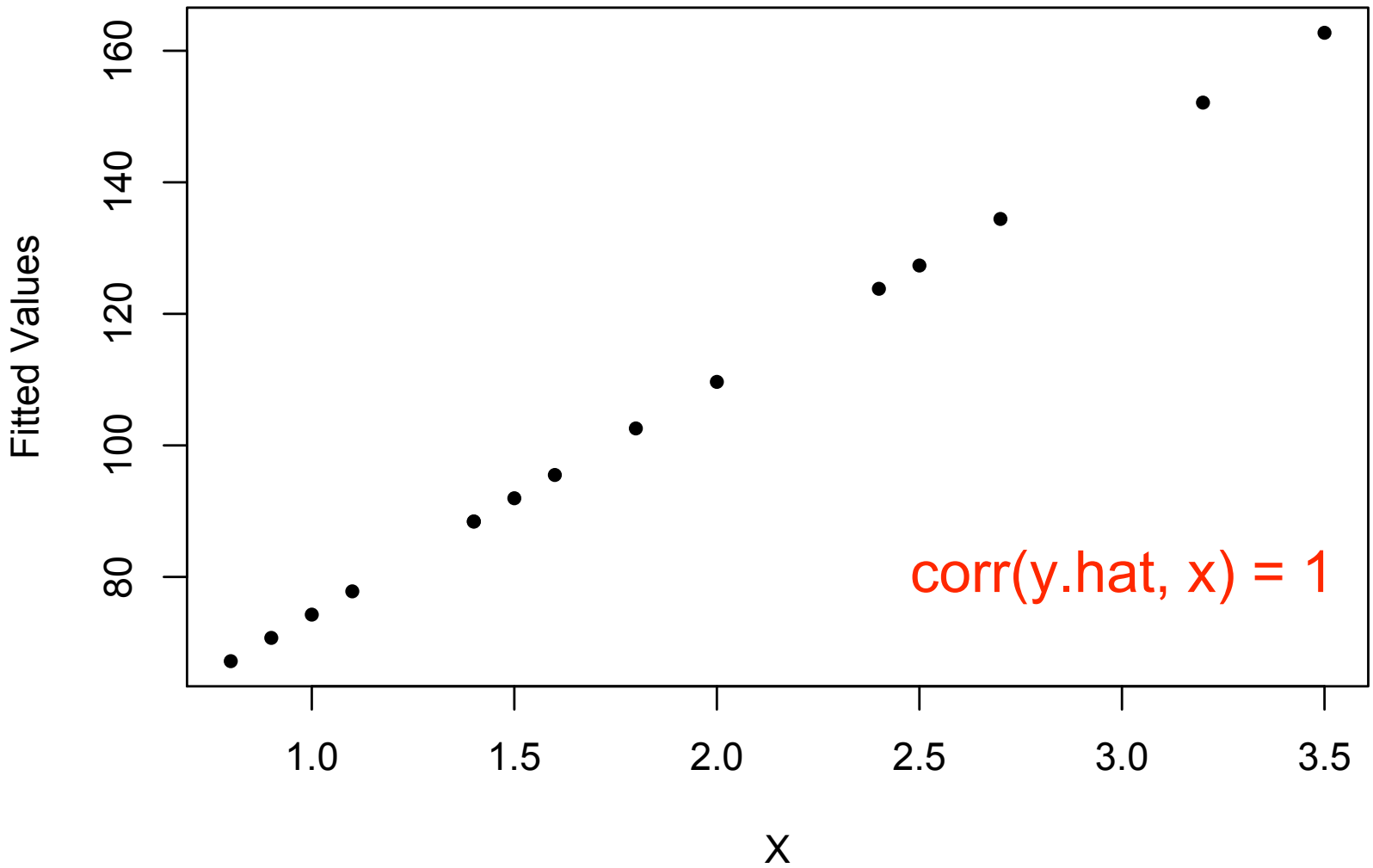
$$b_1 = r_{xy}\frac{s_y}{s_x} \quad \text{and} \quad b_0 = \bar{Y} - b_1\bar{X}$$

More on Least Squares

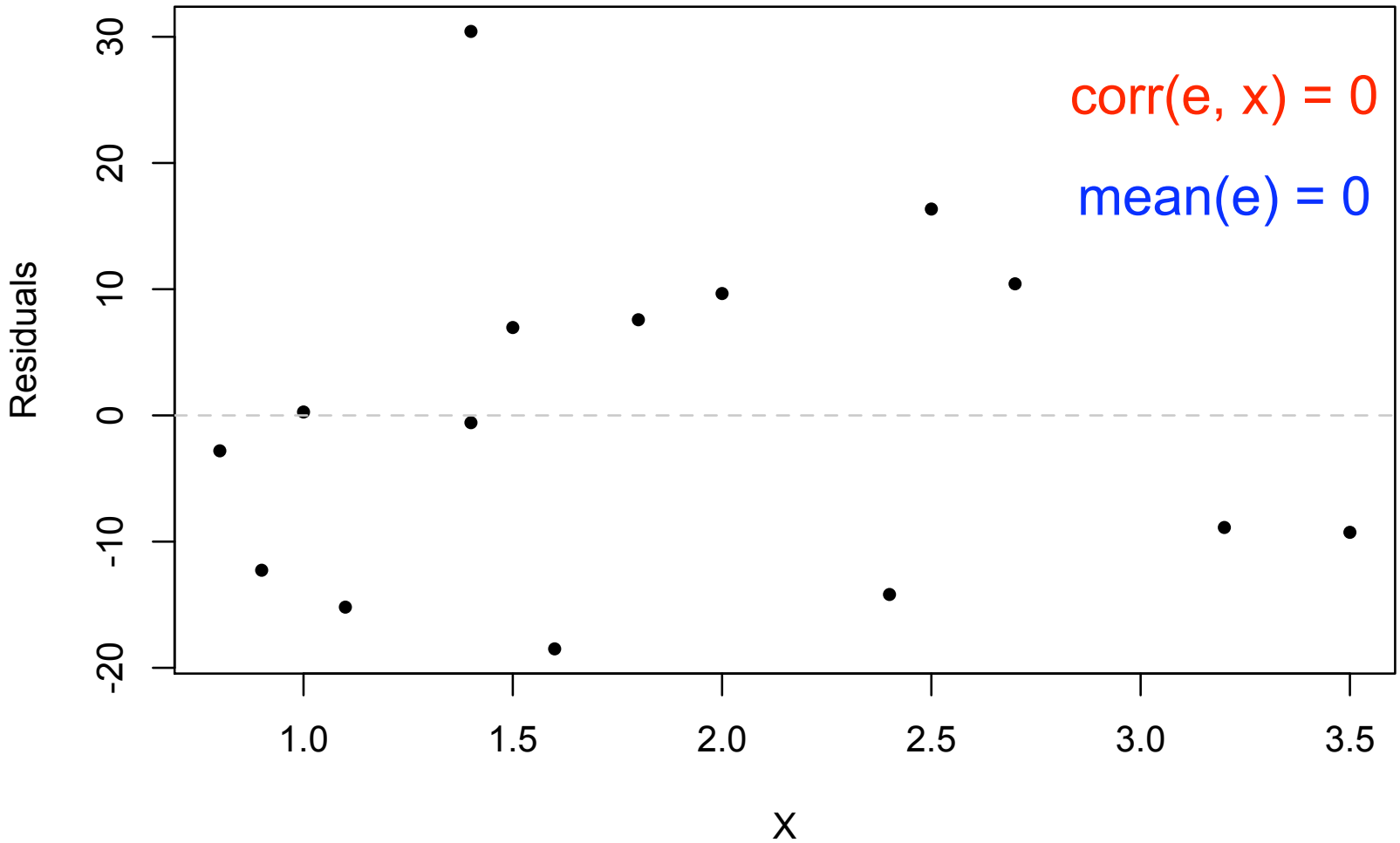
From now on, terms "fitted values" ($\hat{Y}_i$) and "residuals" (e_i) refer to those obtained from the least squares line.

The fitted values and residuals have some special properties. Lets look at the housing data analysis to figure out what these properties are...

The Fitted Values and X

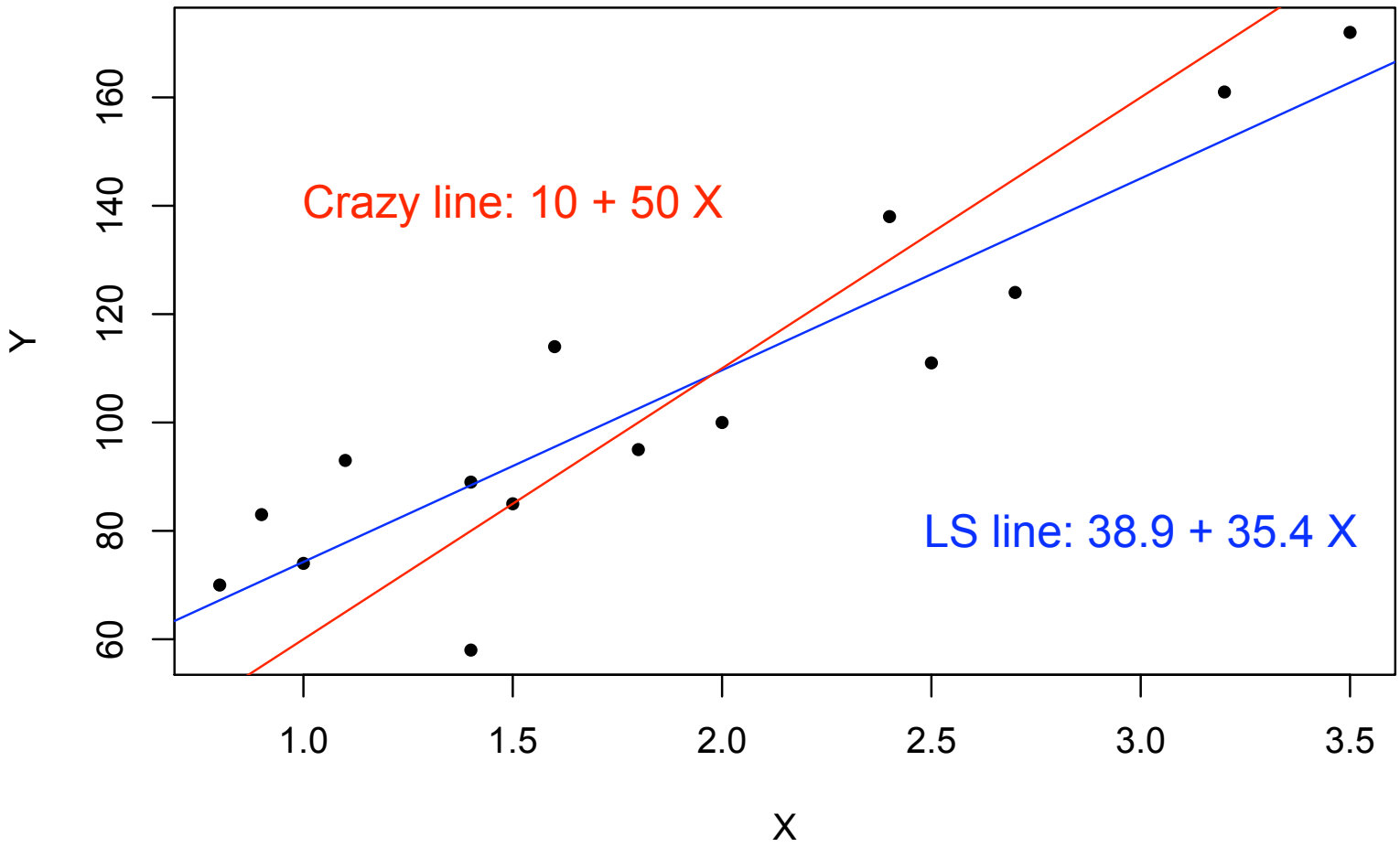


The Residuals and X



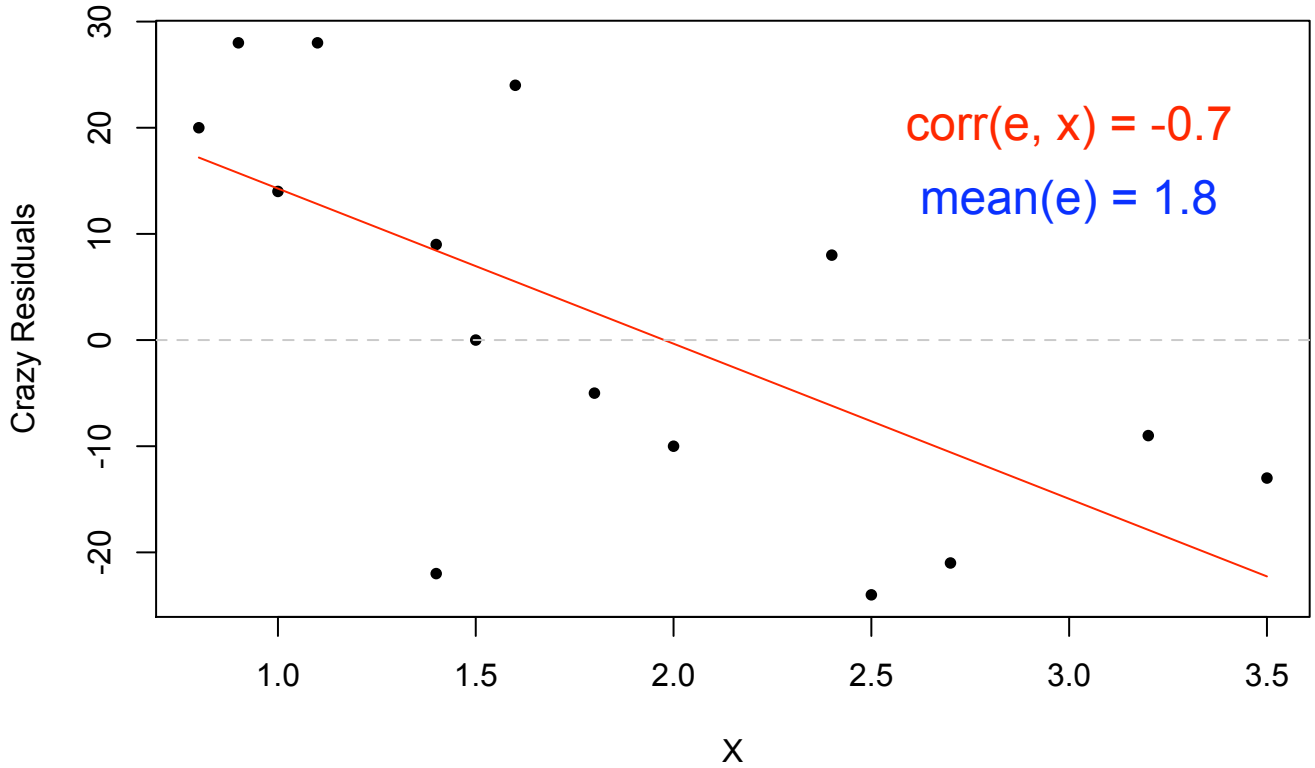
Why?

What is the intuition for the relationship between $\hat{Y}$ and e and X? Lets consider some "crazy" alternative line:



Fitted Values and Residuals

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!

Fitted Values and Residuals

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 power 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"; $\text{corr}(X, \hat{Y}) = 1$.
- e is unrelated to X; $\text{corr}(X, e) = 0$.

Decomposing the Variance

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

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

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

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 – ANOVA Tables

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

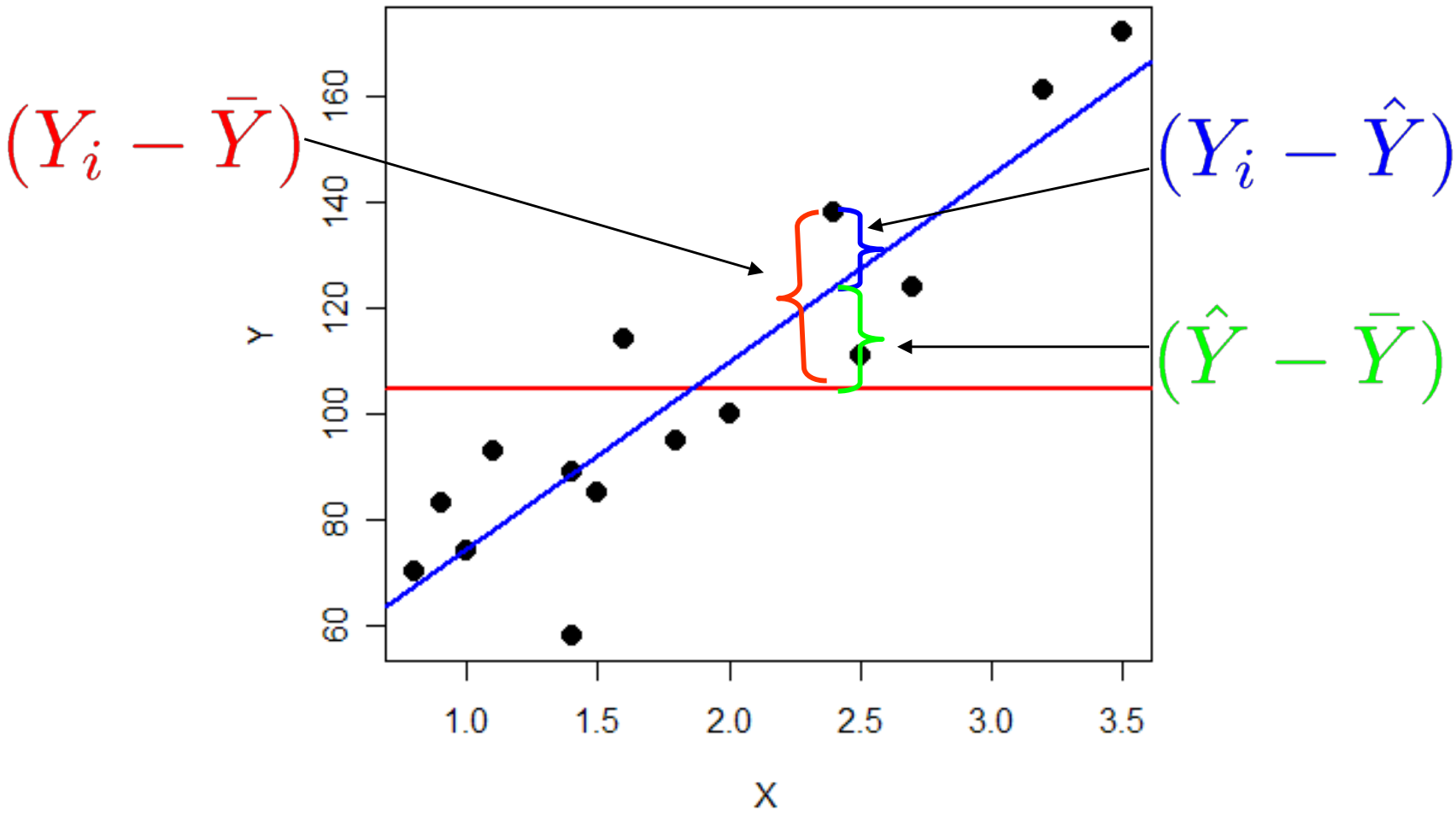
SSR: Variation in Y explained by the regression line.

SSE: Variation in Y that is left unexplained.

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

Be careful of similar acronyms; e.g. SSR for "residual" SS.

Decomposing the Variance – ANOVA Tables



A Goodness of Fit Measure: R^2

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

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

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

A Goodness of Fit Measure: R^2

An interesting fact: $R^2 = r_{xy}^2$ (i.e., R^2 is squared correlation).

$$
\begin{aligned}
R^2 &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \frac{\sum_{i=1}^n (b_0 + b_1 X_i - b_0 - b_1 \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \frac{b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{b_1^2 s_x^2}{s_y^2} = r_{xy}^2
\end{aligned}
$$

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

Summarizing Regression: Back to the House Data

```
> summary(reg)

Call:
lm(formula = price ~ size)

Residuals:
    Min      1Q  Median      3Q     Max
-30.425  -8.618   0.575  10.766  18.498

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   38.885      9.094   4.276 0.000903 ***
size          35.386      4.494   7.874 2.66e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.14 on 13 degrees of freedom
Multiple R-squared: 0.8267, Adjusted R-squared: 0.8133
F-statistic:    62 on 1 and 13 DF,  p-value: 2.66e-06

> var(reg$fitted)/var(price)
[1] 0.8266628
```

Prediction and the Modelling Goal

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

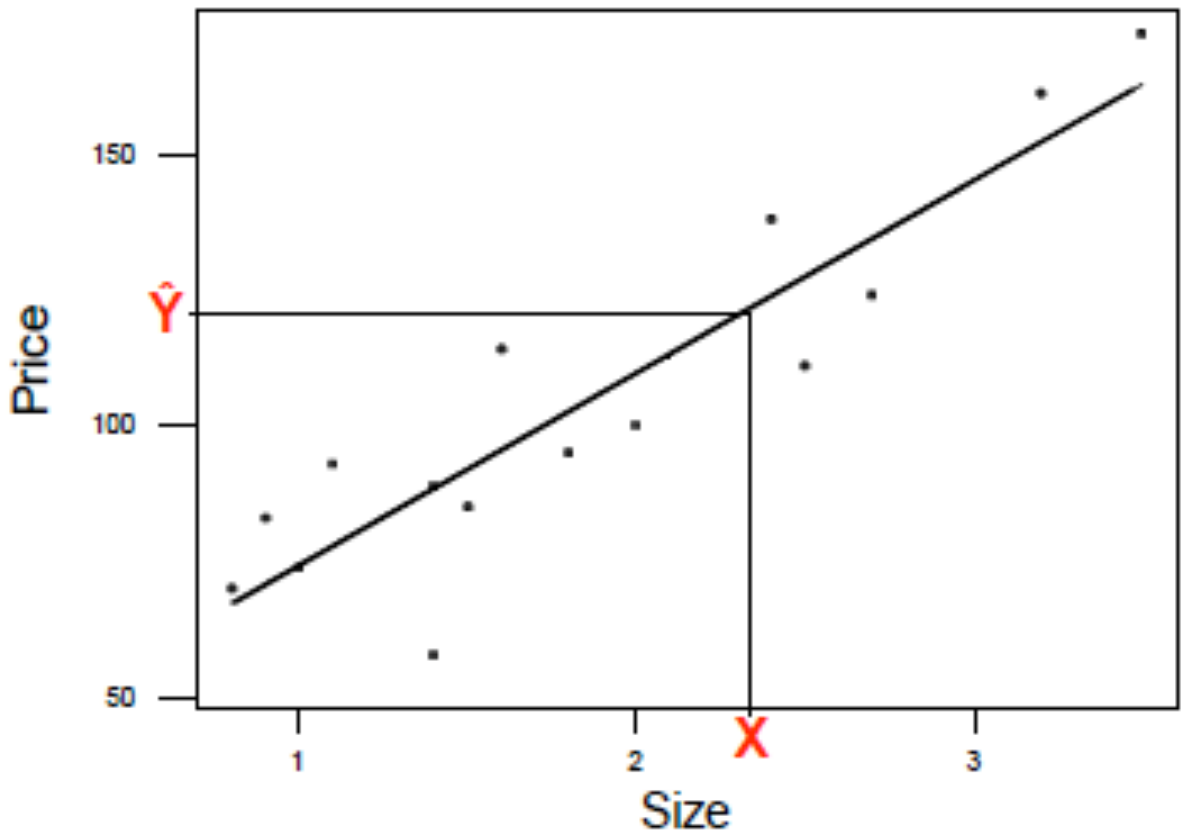
The least squares line is a prediction rule:

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

Prediction and the Modelling Goal

$\hat{Y}$ is not going to be a perfect prediction.

We need to devise a notion of forecast accuracy.



The Simple Linear Regression Model

The power of statistical inference comes from the ability to make precise statements about the accuracy of the forecasts.

In order to do this we must invest in a probability model.

Simple Linear Regression Model: $Y = \beta_0 + \beta_1 X + \varepsilon$

$$\varepsilon \sim \mathrm{N}(0, \sigma^2)$$

The error term ε is independent "idosyncratic noise".

Independent Normal Additive Error

Why do we have $\varepsilon \sim \mathrm{N}(0, \sigma^2)$?

- $\mathbb{E}[\varepsilon] = 0 \Leftrightarrow \mathbb{E}[Y \mid X] = \beta_0 + \beta_1 X$
  ($\mathbb{E}[Y \mid X]$ is "conditional expectation of Y given X").
- Many things are close to Normal (central limit theorem).
- MLE estimates for β's are the same as the LS b's.
- It works! This is a very robust model for the world.

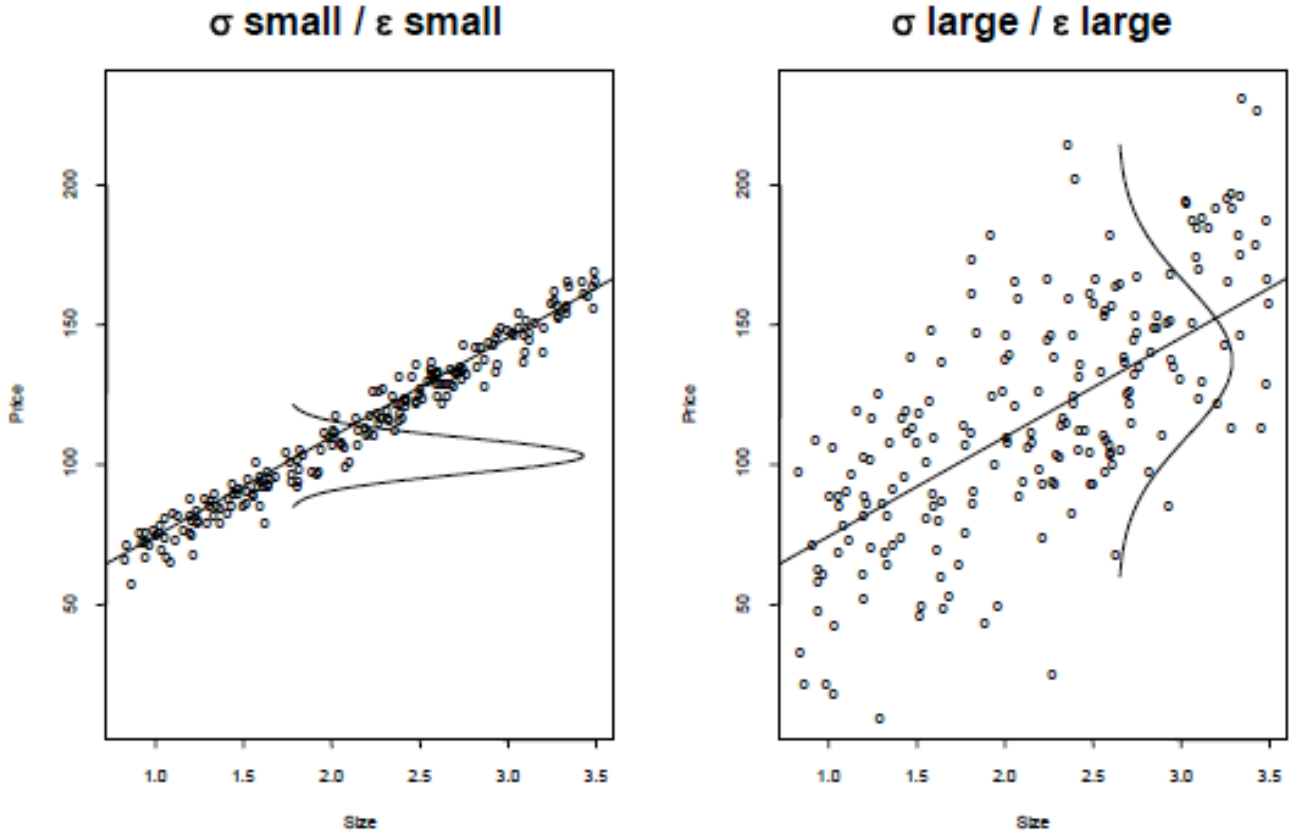
We can think of $\beta_0 + \beta_1 X$ as the "true" regression line.

Conditional Distributions

The conditional distribution for Y given X is Normal:

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

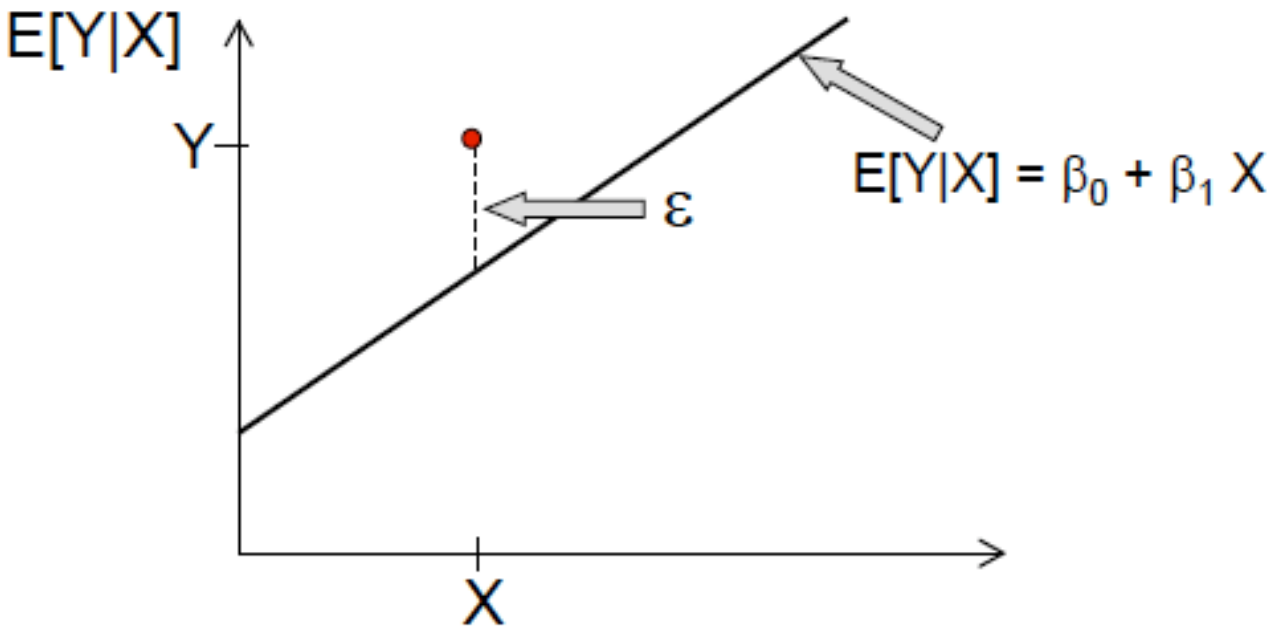
σ controls dispersion:



The Regression Model and our House Data

Think of $E[Y|X]$ as the average price of houses with size X:
Some houses could have a higher than expected value, some lower, and the true line tells us what to expect on average.

The error term represents influence of factors other X.



Prediction Intervals with 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 price of a 1500 square foot house.

What do you know about Y from the model?

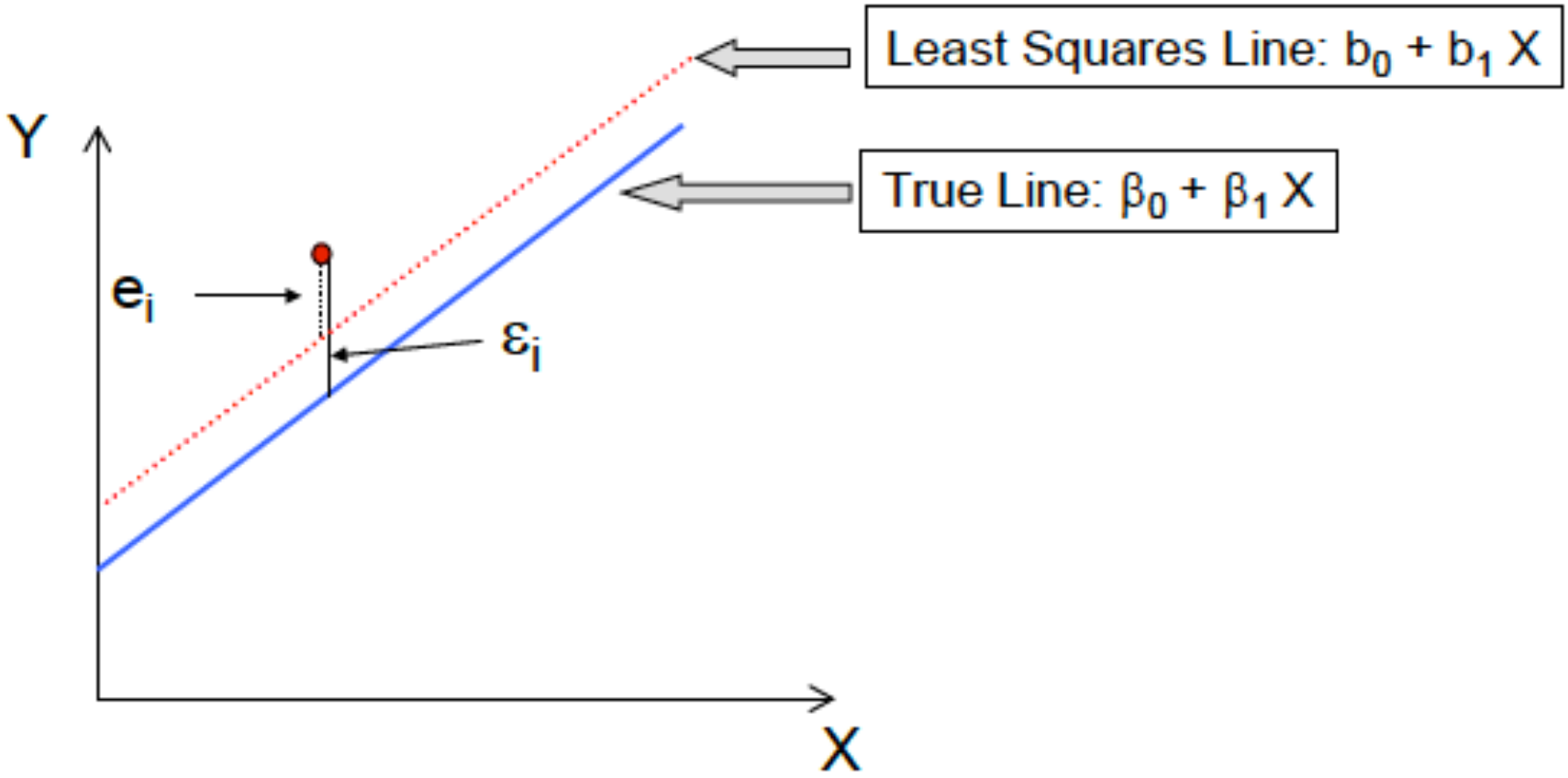
$$\begin{aligned} Y &= 40 + 45(1.5) + \varepsilon \\ &= 107.5 + \varepsilon \end{aligned}$$

Thus our prediction for price is

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

Estimation for the SLR Model

NOTE!!: β_0 **is not** b_0, β_1 **is not** b_1 **and** ε_i **is not e**



Estimation of Error Variance

Recall that $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, and that σ drives the width of the prediction intervals:

$$\sigma^2 = \text{var}(\varepsilon_i) = \mathbb{E}[(\varepsilon_i - \mathbb{E}[\varepsilon_i])^2] = \mathbb{E}[\varepsilon_i^2]$$

A sensible strategy would be to estimate the average for squared errors with the sample average squared residuals:

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{n} e_i^2$$

Estimation of Error Variance

However, this is not an unbiased estimator of σ^2. We have to alter the denominator slightly:

$$s^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2 = \frac{SSE}{n-2}$$

(p is the number of regression coefficients; i.e. 2 for $\beta_0 + \beta_1$).

We have $n - p$ degrees of freedom because 2 have been "used up" in the estimatiation of b_0 and b_1.

We usually use $s = \sqrt{SSE/(n-p)}$, in the same units as Y.

Degrees of Freedom

Degrees of Freedom is the number of times you get to observe useful information about the variance you're trying to estimate.

For example, consider $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$:

- If $n = 1$, $\bar{Y} = Y_1$ and $SST = 0$: since Y_1 is "used up" estimating the mean, we haven't observed any variability!
- For $n > 1$, we've only had $n - 1$ chances for deviation from the mean, and we estimate $s_y^2 = SST/(n-1)$.

In regression with p coefficients (e.g., $p = 2$ in SLR), you only get $n - p$ real observations of variability $\Rightarrow DoF = n - p$.

Estimation of Error Variance

Where is *s* in the output?

```
summary(reg)
```

```
Residual standard error: 14.14 on 13 degrees of freedom
Multiple R-squared: 0.8267, Adjusted R-squared: 0.8133
F-statistic:    62 on 1 and 13 DF,  p-value: 2.66e-06


>
> summary(reg)$sigma
[1] 14.13840
> sqrt(sum(reg$resid^2)/(n-2))
[1] 14.13840
```

Remember that whenever you see "standard error" read it as estimated standard deviation: σ is the standard deviation.

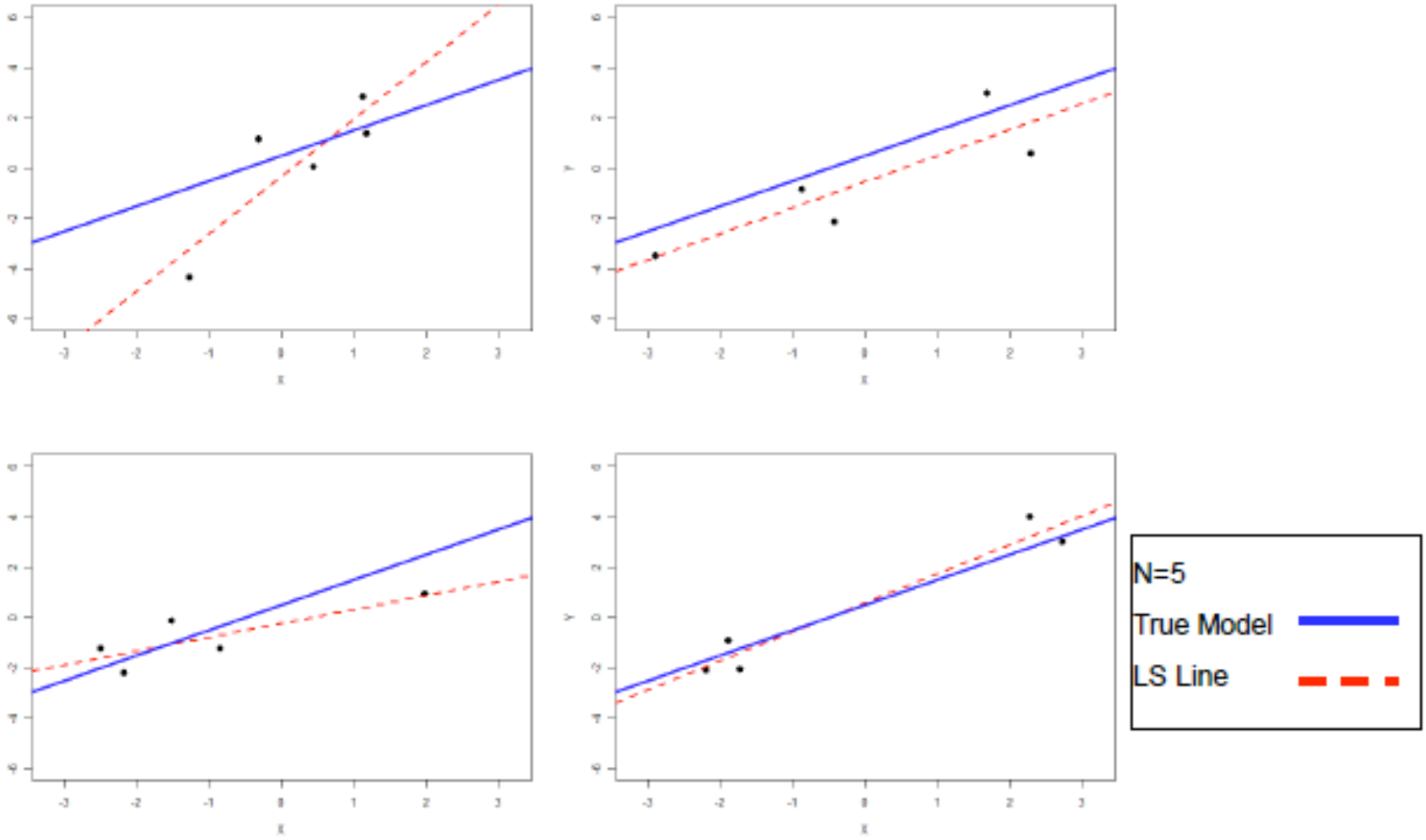
Sampling Distribution of Least Squares Estimates

How much do our estimates depend on the particular random sample that we happen to observe? Imagine:

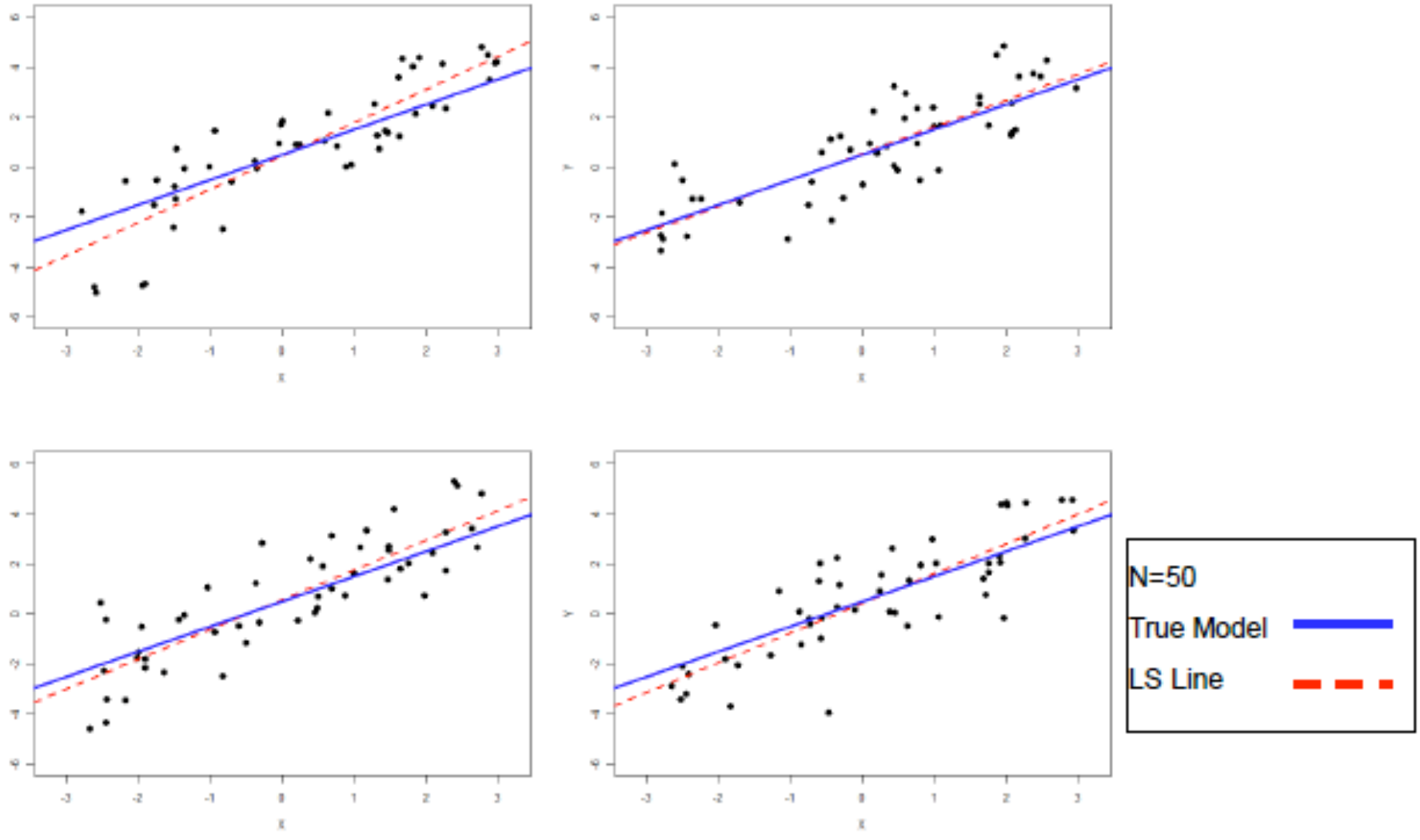
- Randomly draw different samples of the same size.
- For each sample, compute the estimates b_0, b_1, and s.

If the estimates don't vary much from sample to sample, then it doesn't matter which sample you happen to observe.
If the estimates do vary a lot, then it matters which sample you happen to observe.

Sampling Distribution of Least Squares Estimates



Sampling Distribution of Least Squares Estimates



Sampling Distribution of Least Squares Estimates

LS lines are much closer to the true line when $n = 50$.

For $n = 5$, some lines are close, others aren't:

we need to get "lucky"

Review: Sampling Distribution of Sample Mean

Step back for a moment and consider the mean for an *iid* sample of n observations of a random variable $\{X_1, \ldots, X_n\}$

Suppose that $\mathbb{E}(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$

- $\mathbb{E}(\bar{X}) = \frac{1}{n}\sum \mathbb{E}(X_i) = \mu$
- $\text{var}(\bar{X}) = \text{var}\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n^2}\sum \text{var}\left(X_i\right) = \dfrac{\sigma^2}{n}$

If X is normal, then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

If X is not normal, we have the central limit theorem!

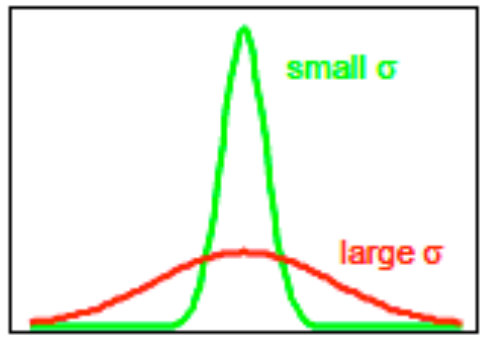
Sampling Distribution of b_1

The sampling distribution of b_1 describes how estimator $b_1 = \hat{\beta}_1$ varies over different samples with the X values fixed.

It turns out that b_1 is normally distributed: $b_1 \sim N(\beta_1, \sigma^2_{b_1})$.

- b_1 is unbiased: $\mathbb{E}[b_1] = \beta_1$.
- Sampling sd σ_{b_1} determines precision of b_1.

The variance term determines how close the estimate will be to the true value. Remember: large σ is bad!



Sampling Distribution of b_1

Can we intuit what should be in the formula for σ_{b_1}?

- How should σ figure in the formula?
- What about n?
- Anything else?

$$\text{var}(b_1) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2} = \frac{\sigma^2}{(n-1)s_x^2}$$

Three Factors:

sample size (n), error variance ($\sigma^2 = \sigma_\varepsilon^2$), and X-spread (s_x).

Optional: Derivation of Var(b_1)

First off, $b_1 = \dfrac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \dfrac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2}$

since $\sum(X_i - \bar{X}) = \sum X_i - n\bar{X} = 0$. Thus

$$b_1 = \sum \frac{X_i - \bar{X}}{D} Y_i = \sum w_i Y_i$$

where $D = \sum(X_i - \bar{X})^2$ is the sum of squares for X. Finally,

$$var(b_1) = \sum w_i^2 var(Y_i \mid X_i) = \sigma^2 \sum w_i^2 = \sigma^2 \frac{D}{D^2} = \frac{\sigma^2}{D}$$

(Note: b_1 is more heavily weighted by X_i that are far from $\bar{X}$.)

Sampling Distribution of b_0

The intercept is also normal and unbiased: $b_0 \sim N(\beta_0, \sigma^2_{b_0})$.

$$\sigma^2_{b_0} = \mathrm{var}(b_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s^2_x} \right)$$

What is the intuition here?

$$\mathrm{var}(\bar{Y} - \bar{X} b_1) = \mathrm{var}(\bar{Y}) - \bar{X}^2 \mathrm{var}(b_1) \; + \; 2\mathrm{cov}(\bar{Y}, b_1)$$

$\bar{Y}$ and b_1 are uncorrelated because the slope (b_1) is invariant if you shift the data up or down ($\bar{Y}$).

Joint Distribution of b_0 and b_1

We know that b_0 and b_1 are dependent: $\mathbb{E}[(b_0 - \beta_0)(b_1 - \beta_1)]$.

This means that an estimation error in the slope is correlated with the estimation error in the intercept.

$$\operatorname{cov}(b_0, b_1) = -\sigma^2 \left(\frac{\bar{X}}{(N-1)s_x^2} \right)$$

- Usually, if the slope estimate is too high, the intercept estimate is too low (negative correlation).
- The correlation decreases with more X spread (s_x^2).

Estimated Variance

We estimate variation with "sample standard deviations":

$$s_{b_1} = \sqrt{\frac{s^2}{(n-1)s_x^2}} \quad s_{b_0} = \sqrt{s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_x^2}\right)}$$

Recall that $s = \sqrt{\sum e_i^2/(n-p)}$ is the estimator for $\sigma = \sigma_\varepsilon$.
Hence, $s_{b_1} = \hat{\sigma}_{b_1}$ and $s_{b_0} = \hat{\sigma}_{b_0}$ are estimated coefficient sd's.

A high level of info/precision/accuracy means small s_b values.

Normal and Student's t

Recall what *Student* discovered:

If $\theta \sim N(\mu, \sigma^2)$, but you estimate $\sigma^2 \approx s^2$ based on $n - p$ degrees of freedom, then $\theta \sim t_{n-p}(\mu, s^2)$.

For example:

- $\bar{Y} \sim t_{n-1}(\mu, s_y^2/n)$.
- $b_0 \sim t_{n-2}\left(\beta_0, s_{b_0}^2\right)$ and $b_1 \sim t_{n-2}\left(\beta_1, s_{b_1}^2\right)$

The t distribution is just a fat-tailed version of the normal. As $n - p \longrightarrow \infty$, our tails get skinny and the t becomes normal.

Forecasting

The conditional forecasting problem: Given covariate X_f and sample data $\{X_i, Y_i\}_{i=1}^n$, predict the "future" observation y_f.

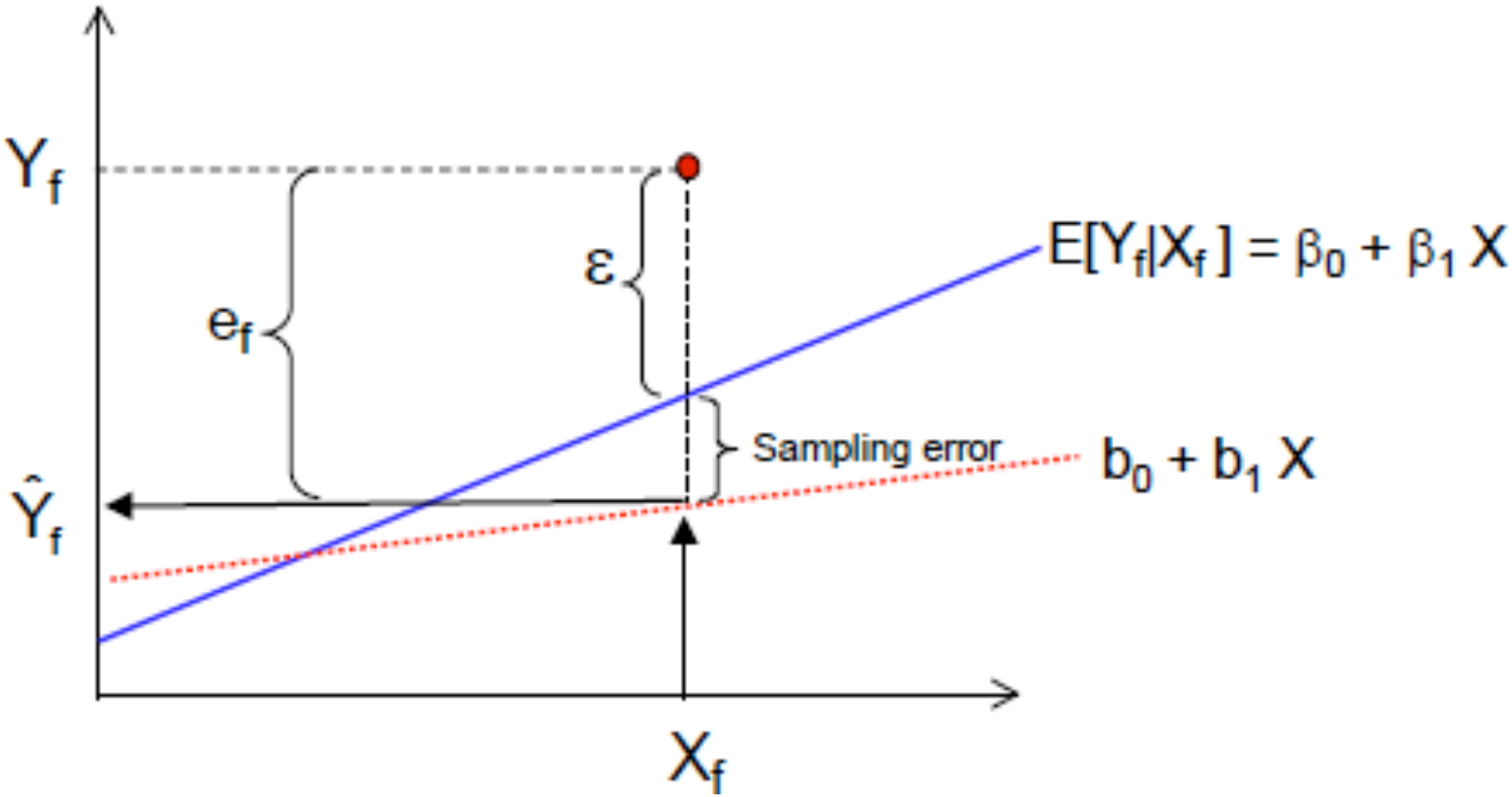
The solution is to use our LS fitted value: $\hat{Y}_f = b_0 + b_1 X_f$.

This is the easy bit. The hard (and very important!) part of forecasting is assessing uncertainty about our predictions.

Forecasting

If we use $\hat{Y}_f$, our prediction error is

$$e_f = Y_f - \hat{Y}_f = Y_f - b_0 - b_1 X_f$$



Forecasting

We can decompose e_f into two sources of error

- Inherrent idiosyncratic randomness (due to ε).
- Estimation error in the intercept and slope (i.e., discrepency between our line and "the truth").

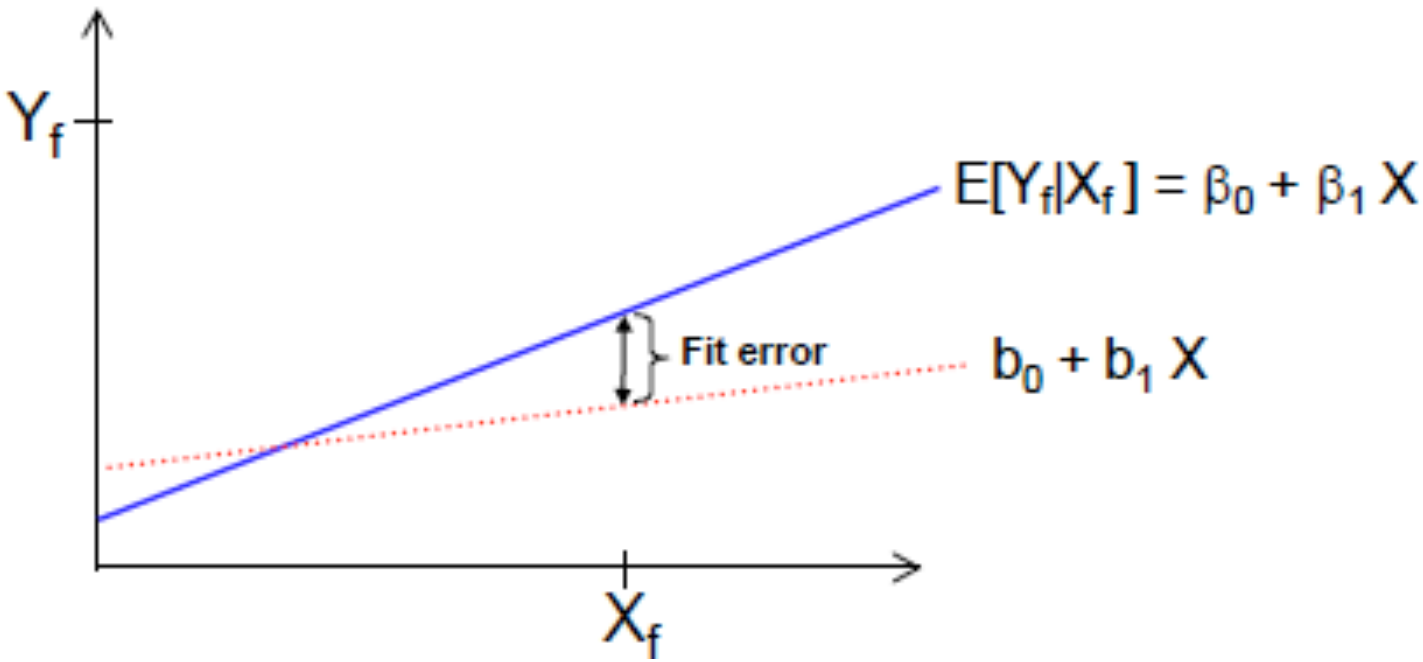
$$\begin{aligned} e_f &= Y_f - \hat{Y}_f = (Y_f - \mathbb{E}[Y_f|X_f]) + \mathbb{E}[Y_f|X_f] - \hat{Y}_f \\ &= \varepsilon_f + (\mathbb{E}[Y_f|X_f] - \hat{Y}_f) \\ &= \varepsilon_f + (\beta_0 - b_0) + (\beta_1 - b_1)X_f. \end{aligned}$$

Forecasting

The variance of our prediction error is thus

$$\mathrm{var}(e_f) = \mathrm{var}(\varepsilon_f) + \mathrm{var}(\mathbb{E}[Y_f|X_f] - \hat{Y}_f) = \sigma^2 + \mathrm{var}(\hat{Y}_f)$$

We know $\mathrm{var}(\varepsilon_f) = \sigma^2 \approx s^2$, but what about the fit error?



Forecasting

From the sampling distributions derived earlier, $\mathrm{var}(\hat{Y}_f)$ is

$$\begin{aligned}\mathrm{var}(b_0 + b_1 X_f) &= \mathrm{var}(b_0) + X_f^2 \mathrm{var}(b_1) + 2X_f \mathrm{cov}(b_0, b_1) \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)s_x^2} \right]\end{aligned}$$

And hence the variance of our predictive error is

$$\mathrm{var}(e_f) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)s_x^2} \right]$$

Forecasting

Putting it all together, we have that

$$Y_f \sim \mathrm{N}\left(\hat{Y}_f, \sigma^2\left[1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)s_x^2}\right]\right)$$

(sums of normals are normal) or, with estimated variance,

$$Y_f \sim t_{n-p}\left(\hat{Y}_f, s^2\left[1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)s_x^2}\right]\right).$$

Finally, a $(1-\alpha)100\%$ confidence interval for Y_f is thus

$$b_0 + b_1 X_f \pm t_{n-2,\alpha/2}\left(s\sqrt{1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)s_x^2}}\right)$$

Forecasting

Looking closer at what we'll call

$$s_{pred} = s\sqrt{1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)s_x^2}}.$$

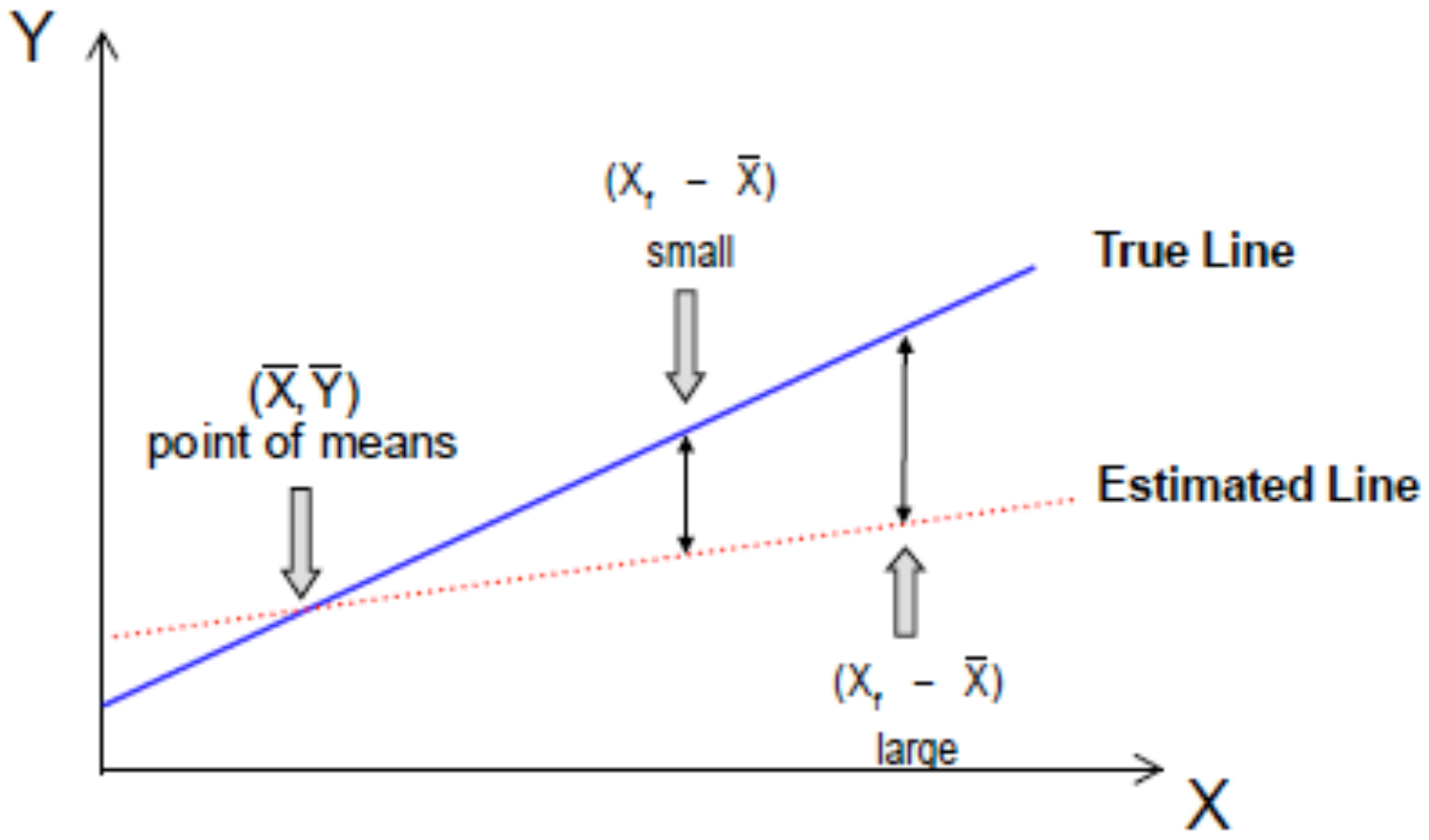
A large predictive error variance (high uncertainty) comes from

- Large s (i.e., large ε's).
- Small n (not enough data).
- Small s_x (not enough observed spread in covariates).
- Large $(X_f - \bar{X})$.

The first three are familiar... what about the last one?

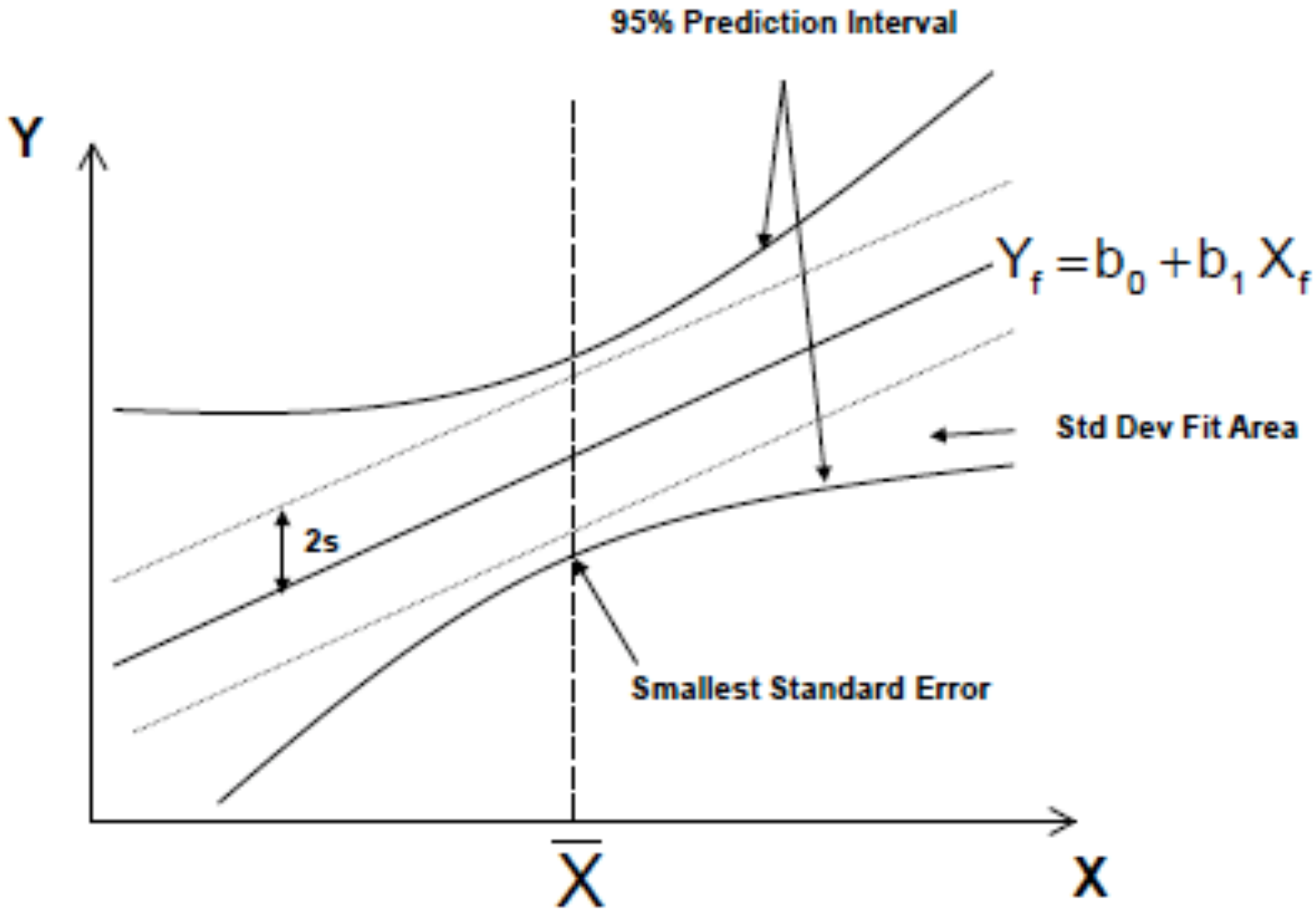
Forecasting

For X_f far from our $\bar{X}$, the space between lines is magnified...



Forecasting

$\Rightarrow$ The prediction (conf.) interval needs to widen away from $\bar{X}$



Glossary of Symbols

b_0 - least squares estimate of the intercept

b_1 - least squares estimate of the slope

e_i - least squares residual for observation i

$\hat{Y}$ - fitted value

$r_{xy} = r =$ sample correlation coefficient

SST - total sum of squares

SSR - regression sum of squares

SSE - error sum of squares

R^2 - coefficient of determination, goodness of fit, SSR/SST.

Glossary of Symbols

Regression model parameters:

- β_0 - true line intercept
- β_1 - true slope
- σ - error standard deviations
- ε - additive error term

Fitted value: $\hat{Y}_i = b_0 + b_1 X_i$
Residual: $e_i = Y_i - \hat{Y}_i$
LS Estimators: $b_1 = r_{xy}\dfrac{s_y}{s_x} = \dfrac{s_{xy}}{s_x^2}$ and $b_0 = \bar{Y} - b_1\bar{X}$.

Glossary and Equations

- $\hat{Y}_i = b_0 + b_1 X_i$ is the ith fitted value.
- $e_i = Y_i - \hat{Y}_i$ is the ith residual.
- s: standard error of regression residuals ($\approx \sigma = \sigma_\varepsilon$).

$$s^2 = \frac{1}{n-2} \sum e_i^2$$

- s_{b_j}: standard error of regression coefficients.

$$s_{b_1} = \sqrt{\frac{s^2}{(n-1)s_x^2}} \qquad s_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_x^2}}$$

Glossary and Equations

- α is the significance level (prob of type 1 error).
- $t_{n-p,\alpha/2}$ is the value such that for $Z_{n-p} \sim t_{n-p}(0,1)$,

$$\mathrm{P}(Z_{n-p} > t_{n-p,\alpha/2}) = \mathrm{P}(Z_{n-p} < -t_{n-p,\alpha/2}) = \alpha/2.$$

- $z_{b_j} \sim t_{n-p}(0,1)$ is the standardized coefficient t-value:

$$z_{b_j} = \frac{b_j - \beta_j^0}{s_{b_j}} \quad (= b_j/s_{b_j} \textit{ most often})$$

- The $(1-\alpha) * 100\%$ for β_j is $b_j \pm t_{n-p,\alpha/2} s_{b_j}$.
- $\varphi = 2\mathrm{P}(Z_{n-p} > z_{b_j})$ is the coefficient p-value.

Glossary and Equations

- $\hat{Y}_f = b_0 + X_f b_1$ is a forecast prediction.

$$\mathrm{sd}(\hat{Y}_f) = s_{fit} = s\sqrt{\frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)s_x^2}}$$

- Forecast residual is $e_f = Y_f - \hat{Y}_f$ and $\mathrm{var}(e_f) = s^2 + s_{fit}^2$. That is, the predictive standard error is

$$s_{pred} = s\sqrt{1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)s_x^2}}.$$

and $\hat{Y}_f \pm t_{n-p,\alpha/2} s_{pred}$ is the $(1-\alpha)100\%$ PI at X_f.