OPIM 5603
Fall 2019

Simple Linear Regression

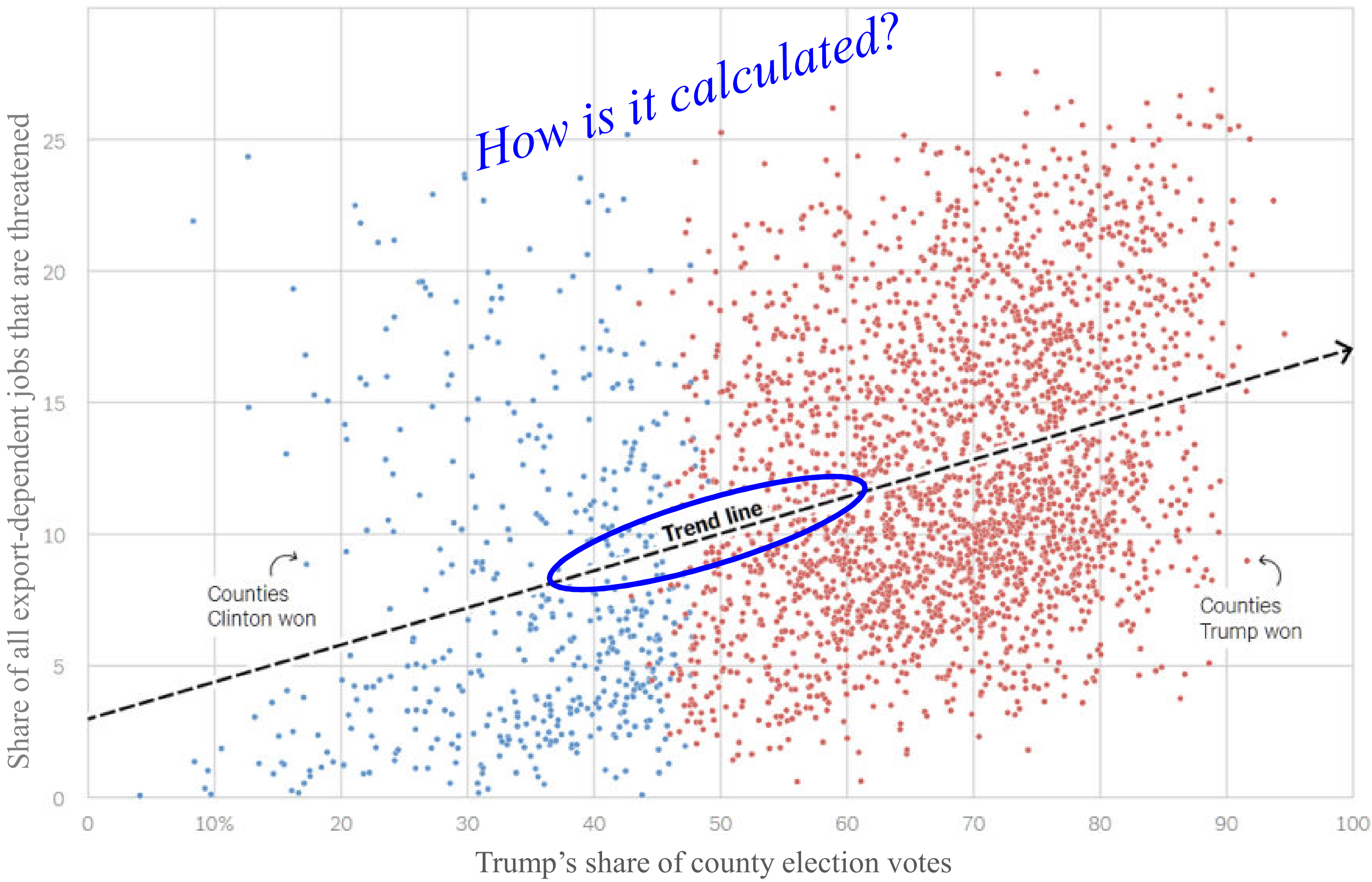
Selected Topics

1. The Least Squares Method
2. Simple Linear Regression
3. Measures of Variation
4. Residuals
5. Regression Variance
6. Regression Coefficients
7. Briefly on Multiple Linear Regression

Trend Line?

OPIM 5603 Fall 2019

Slide 3.3

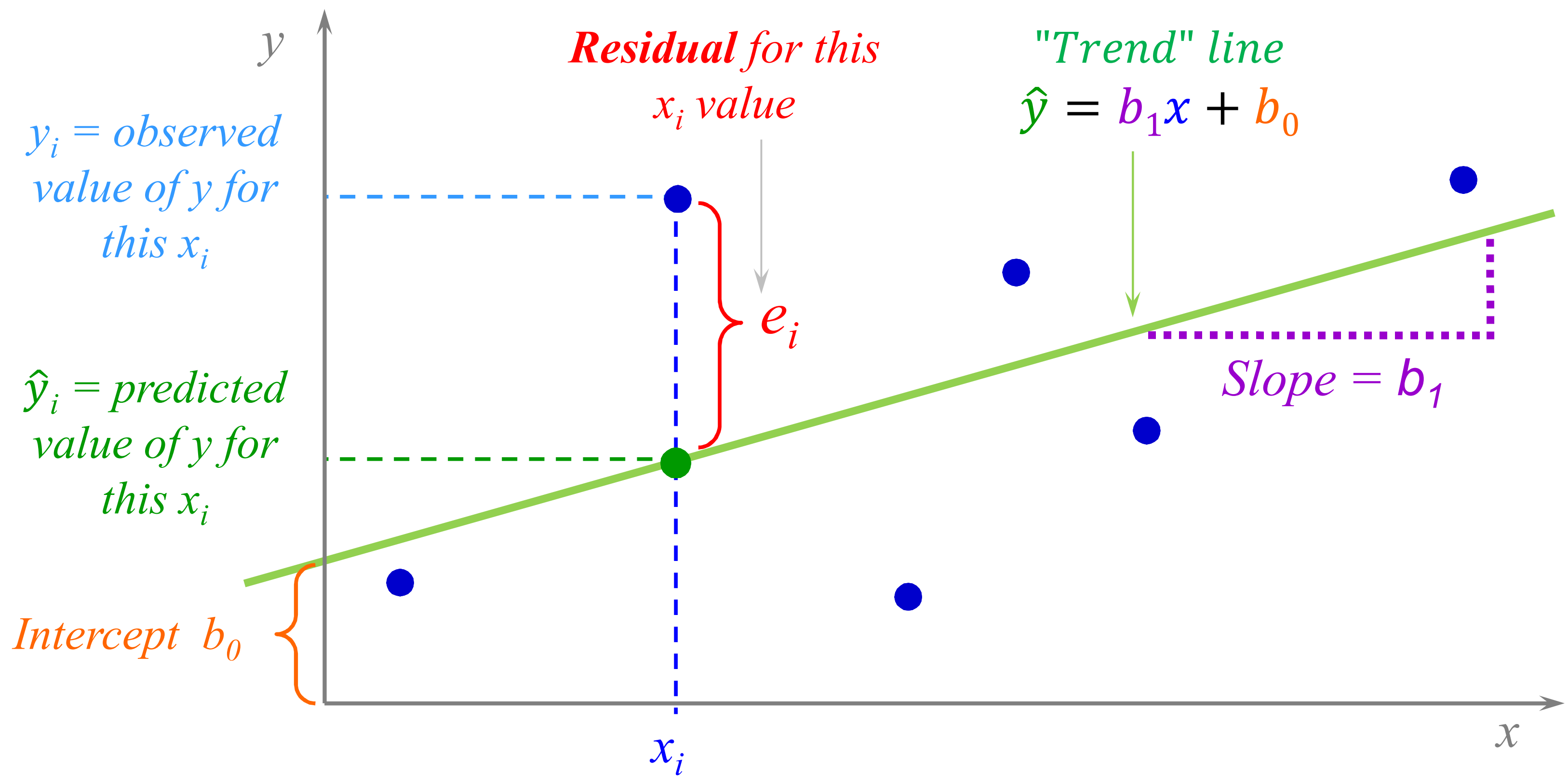


Percentage of export-dependent jobs affected by retaliatory tariffs, by U.S. counties

Tariffs That Send a Political Message, The New York Times, October 3rd 2018

The Least Squares Method

Given pairs of points (x_i, y_i), $i = 1,\ldots n$, calculate *intercept* b_0 and *slope* b_1 so that the line $\hat{y} = b_1 x + b_0$ is "the best" linear representative for points (x_i, y_i).



The coefficients b_0 and b_1 are computed from points (x_i, y_i) in such a way that *they minimize the sum of the residuals squared!*

OPIM 5603
Fall 2019

Slide 3.5

The Least Squares Method

"...they minimize the sum of the residuals squared":

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_1 x_i + b_0))^2$$

function $Q(a,m)$

$$= \underbrace{\min}_{\text{over all } a,m} \boxed{\sum_{i=1}^{n} (y_i - (m x_i + a))^2}$$

Rephrased: Among all lines $mx + a$ that can be used to predict y as a linear function of x the prediction line

$$\hat{y} = b_1 x + b_0$$

has the smallest sum of the residuals squared.

Question: How can we compute b_0 and b_1 ?

Basic Calculus: To find the minimum of the function $Q(a,m)$ take derivatives with respect to a and m and set them equal to zero.

The Least Squares Method

The solutions are $b_1 = m_{min} = \dfrac{\boxed{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Note: *cor(x,y)* and the slope b_1 have the same sign.

$$b_0 = a_{min} = \bar{y} - b_1\bar{x}$$

Note 1: Point $(\bar{x}, \bar{y})$ lies on the prediction line $\hat{y} = b_1 x + b_0$:

$$b_0 + b_1\bar{x} = \bar{y} - b_1\bar{x} + b_1\bar{x} = \bar{y}$$

Note 2: Average of the predictions $\hat{y}_1, \ldots, \hat{y}_n$ is $\bar{y}$:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(b_1 x_i + b_0) = b_0 + b_1\frac{1}{n}\sum_{i=1}^{n}x_i = b_0 + b\bar{x},$$

which equals $\bar{y}$ by Note 1.

Note 3: Sum of the residuals $e_1, e_2, \ldots, e_n$ is zero:

$$\frac{1}{n}\sum_{i=1}^{n}e_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \underbrace{\boxed{\frac{1}{n}\sum_{i=1}^{n}y_i}}_{\bar{y}} - \underbrace{\boxed{\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i}}_{=\bar{y}\text{, by Note 2}} = \bar{y} - \bar{y} = 0,$$

Measures of Variation

Another corollary of the *Least Square Method* is that

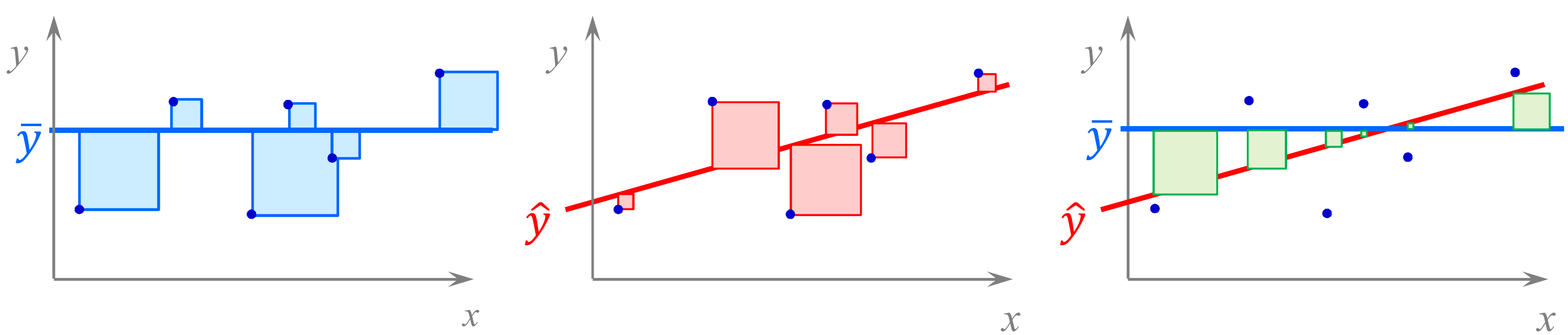
$$\underset{\sim \text{ Sample Variance of } y\text{'s}}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + \underset{\sim \text{ Sample Variance of } \hat{y}\text{'s}}{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}$$

SST
Total sum of squares
(Total Variation)

SSE
Error sum of squares
(Unexplained Variation)

SSR
Regression sum of squares
(Explained Variation)

(proportional to variation of x_i's)



Measures of Variation

All three quantities, SST, SSE, and SSR, are non-negative and

$$0 \le SSR \le SST \quad \text{i.e.,} \quad 0 \le \frac{SSR}{SST} \le 1$$

Larger the ratio the prediction is better.

$$0 \le SSE \le SST \quad \text{i.e.,} \quad 0 \le \frac{SSE}{SST} \le 1$$

Smaller the ratio the prediction is better.

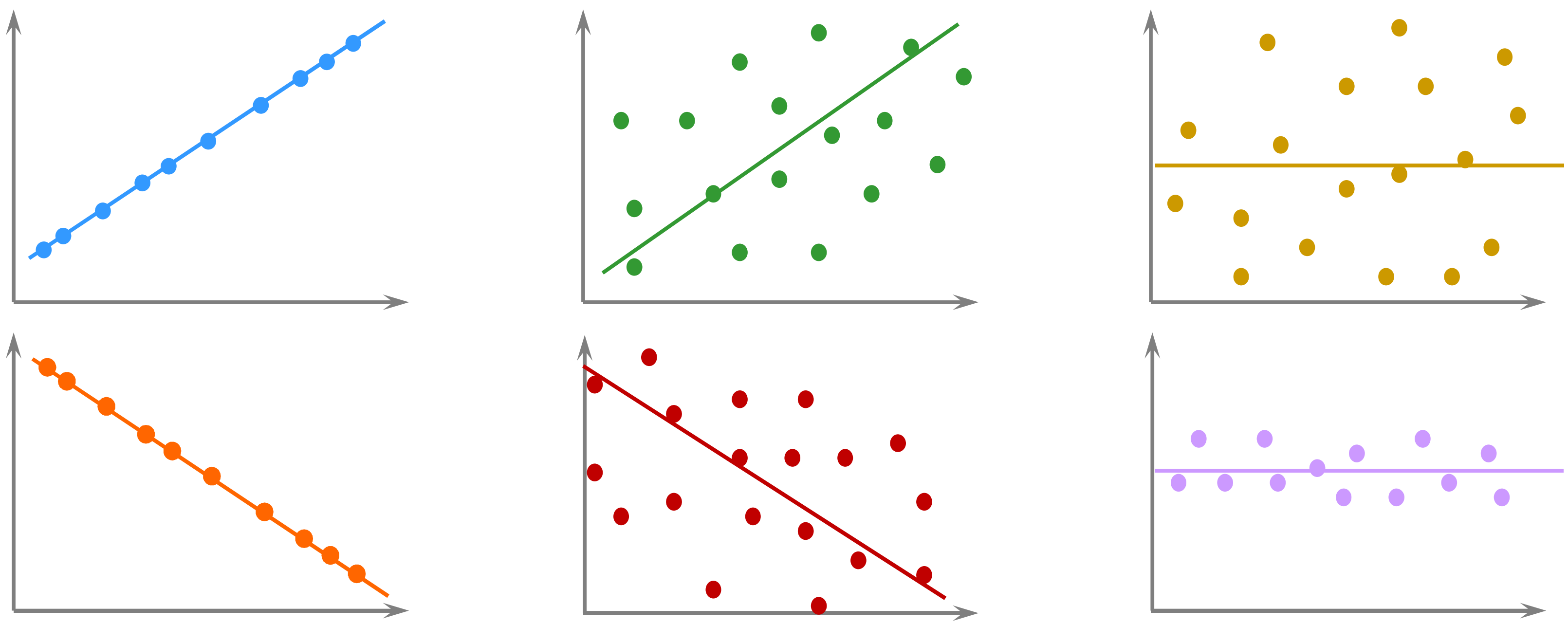
We define *Coefficient of Determination* $r^2 = \frac{SSR}{SST}$.

Clearly $0 \le r^2 \le 1$, and the prediction is better for r^2's closer to 1.

Notice that $r^2 = 1 - \frac{SSE}{SST}$ (since $SST = SSE + SSR$).

Thus r^2 is the portion of total variation in the dependent variable that is explained by variation in the independent variable.

Examples of r^2



$r^2 = 1$

Perfect linear relationship between x and y:
100% of the variation in y is explained by variation in x.

($SSE = 0$ since $\hat{y}_i = y_i$)

$0 < r^2 < 1$

Weaker linear relationships between x and y:
Some but not all of the variation in y is explained by variation in x.

$r^2 = 0$

NO linear relationship between x and y:
The value of y does not depend on x. (None of the variation in y is explained by variation in x).

($SSR = 0$ since $\hat{y}$ = const. $\bar{y}$)

Correlation, r^2, Adjusted r^2

Another corollary of the *Least Square Method* is:

$$r^2 = cor(x,y)^2 = cor(\hat{y},y)^2$$

Note: identity makes sense when we have only one predictor x.

Adjusted r^2 is primarily designed for multiple predictors:

$$r^2_{\text{adj}} = 1 - \frac{n-1}{n-p-1}(1 - r^2),$$

where p is the number of predictors excluding the constant term.

It takes into account the fact that r^2 automatically increases when additional predictors are added to the model.

On Least Squares Method Optional

$$\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = \sum_{i=1}^{n}(b_0 + b_1 x_i - y_i)^2 = \underbrace{\min}_{\text{over all } a,m} \underbrace{\boxed{\sum_{i=1}^{n}(a + m x_i - y_i)^2}}_{\text{function } Q(a,m)}$$

Partial derivatives:

$$\frac{\partial Q}{\partial a}(a,m) = 2\sum_{i=1}^{n}(a + m x_i - y_i)$$

$$\frac{\partial Q}{\partial m}(a,m) = 2\sum_{i=1}^{n}x_i(a + m x_i - y_i)$$

b_0 and b_1 are the values of a and m when the two equations above are set to 0.

$$\sum_{i=1}^{n}(b_0 + b_1 x_i - y_i) = 0 \quad \Rightarrow \sum_{i=1}^{n}(\hat{y}_i - y_i) = 0 \quad (1)$$

Note: The sum of residuals is 0.

$$\sum_{i=1}^{n}x_i(b_0 + b_1 x_i - y_i) = 0 \quad \Rightarrow \sum_{i=1}^{n}x_i(\hat{y}_i - y_i) = 0 \quad (2)$$

Corollary: SST = SSE + SSR Optional

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} \left((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \right)^2$$

$$= \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{SSE} + \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{2 \sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{\text{“Cross term” equals } 0}$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n} (y_i - y_i)(b_1 x_i + b_0 - \bar{y})$$

$$= b_1 \underbrace{\sum_{i=1}^{n} x_i (y_i - \hat{y}_i)}_{= 0, \text{ by } (2)} + (b_0 - \bar{y}) \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)}_{= 0, \text{ by } (1)} = 0$$

When should Least Square Method be used?

Recap: Given pairs of points (x_i, y_i), $i = 1, \ldots n$, the *Least Squares Method* calculates *intercept* b_0 and *slope* b_1 so that the line $\hat{y} = b_1 x + b_0$ has the *smallest sum of the residuals squared* among all lines that can be used to predict y as a *linear function* of x.

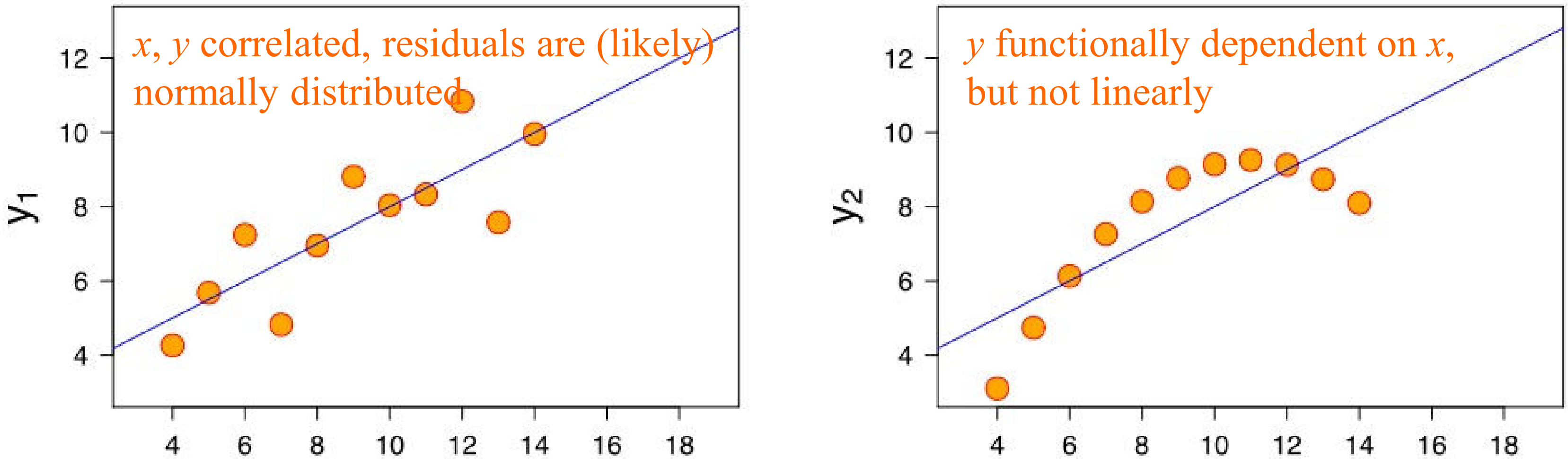
Hence the intent is to use it for cases when there is a linear relationship between x and y.

Think: why would you use a line to predict y as a function of x (or vice versa) if their relationship was not linear?

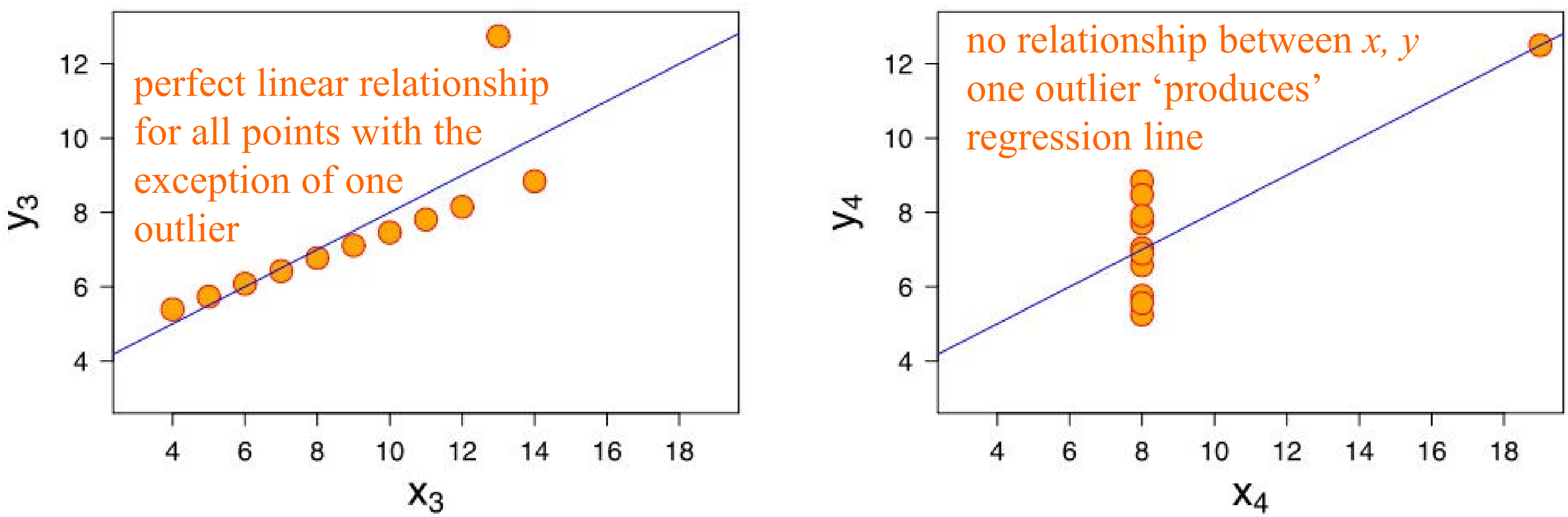
The method, however, is "blind" to this requirement: it will calculate the intercept and the slope no matter what the relationship between x and y is.

OPIM 5603
Fall 2019

Caution: Anscombe's Quartet

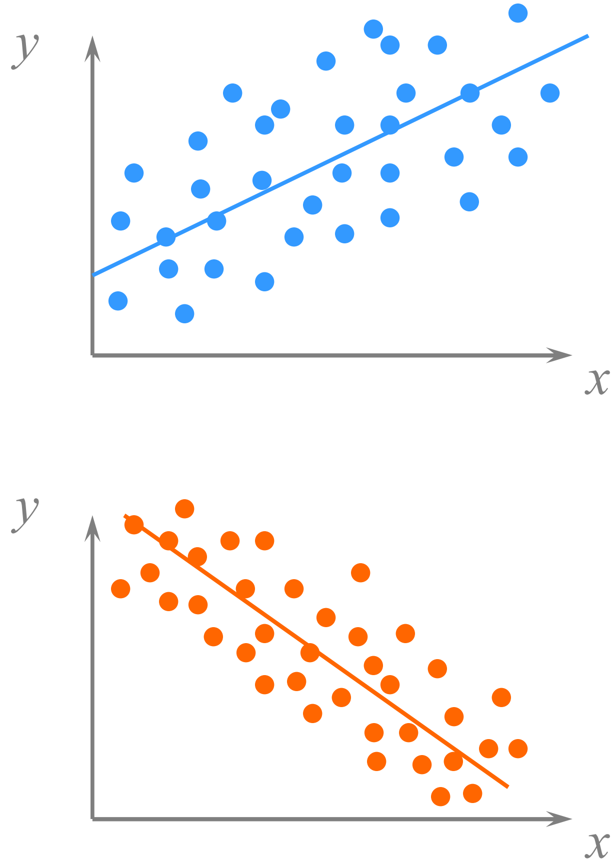


The four *y samples* have the same mean of 7.5, variance 4.12, correlation (with *x*) of 0.816 and regression line $y = 3 + 0.5x$ (example by Francis Anscombe).

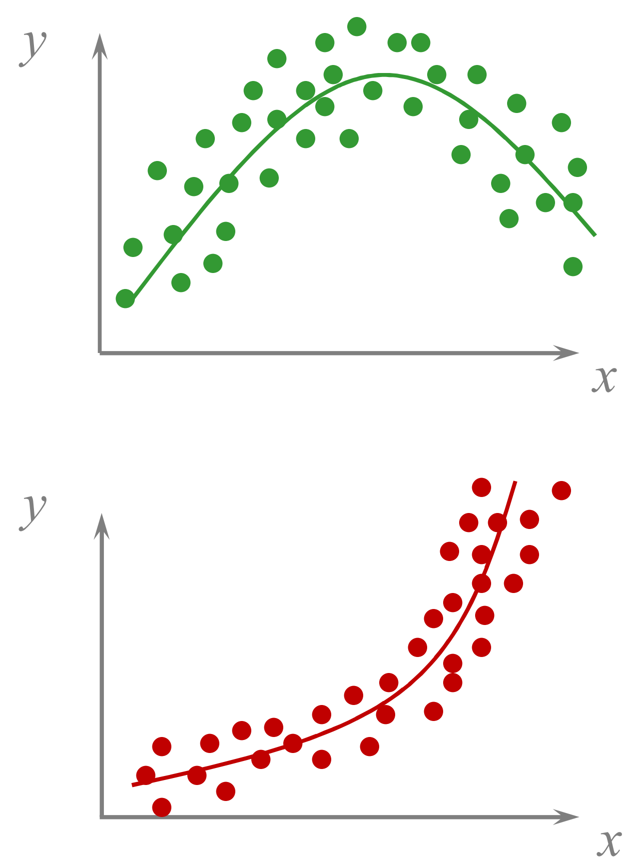


Types of Relationships

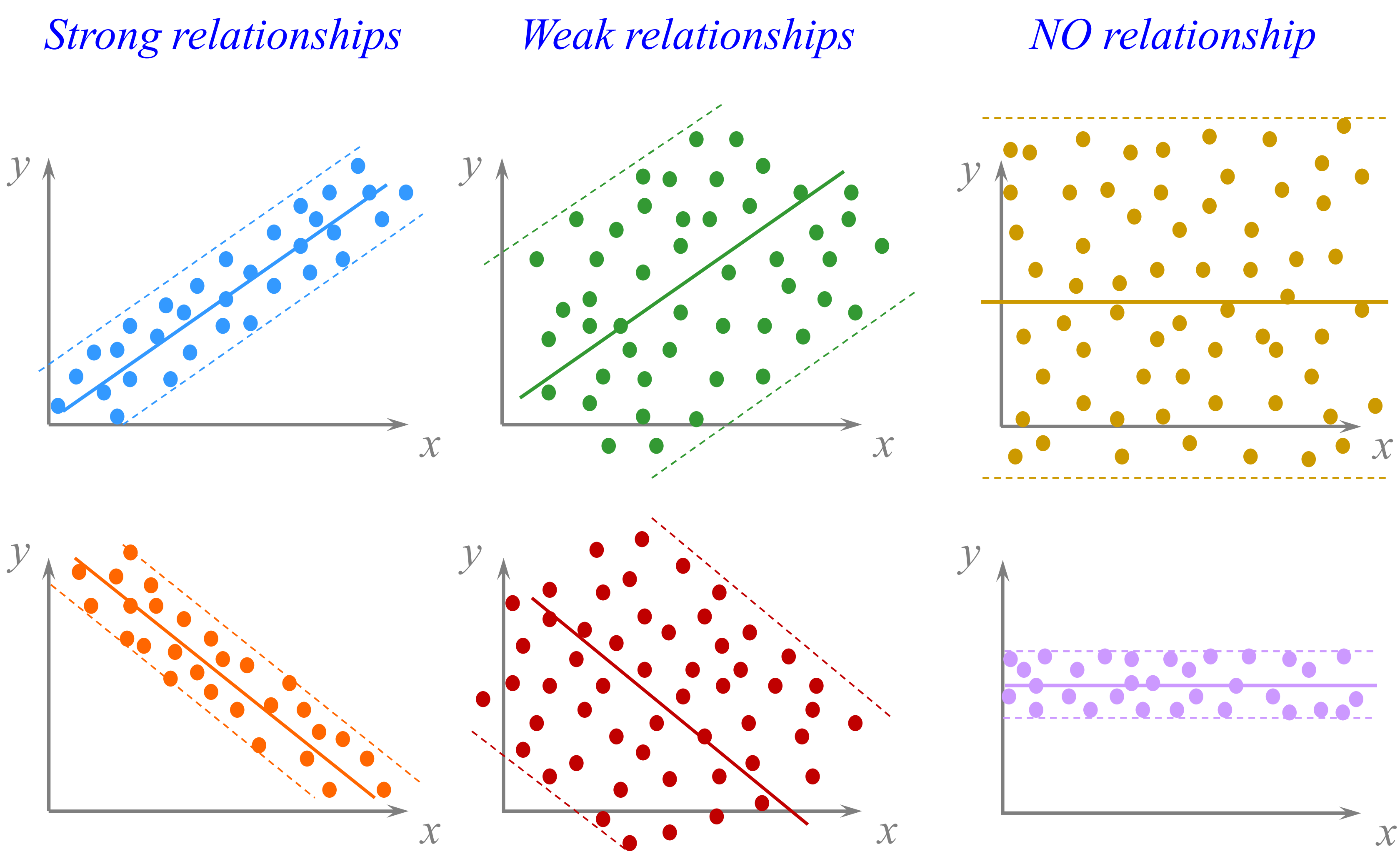
Linear relationships



Non-linear relationships



Linear Relationships



r^2 'closer' to 1

r^2 'closer' to 0

$r^2 = 0$

Simple Linear Regression

Linear relationship between ***two*** random variables X and Y:

Dependent Variable · *y-intercept* · *Slope* · *Predictor (Independent Variable)* · *Noise*

$$Y = \underbrace{\beta_0 + \beta_1 X}_{\textit{Linear component } \hat{Y}} + \underbrace{\varepsilon}_{\textit{Random Error component}}$$

The main assumption is that population Y linearly depends on population X. Further assumptions about the random error component will be elaborated later.

Note: There are three *random variables* above: X and Y represent their respective populations and the random error made by linear approximation is captured by ε. The coefficients *Slope* and *Intercept* are numbers.

Simple Linear Regression and the LSM

The *LSM* deals with pairs of numbers; *SLR* involves two random variables.

Suppose $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ is a sample from the (*joint*) distribution of (X, Y). Recall the 'representatives' from chapter 2:

$$
\begin{array}{ccccccccc}
(X_1^{(1)}, Y_1^{(1)}) & (X_2^{(1)}, Y_2^{(1)}) & \ldots & (X_n^{(1)}, Y_n^{(1)}) & \xrightarrow{LSM} & b_0^{(1)} & \text{and} & b_1^{(1)} \\
(X_1^{(2)}, Y_1^{(2)}) & (X_2^{(2)}, Y_2^{(2)}) & \ldots & (X_n^{(2)}, Y_n^{(2)}) & \longrightarrow & b_0^{(2)} & \text{and} & b_1^{(2)} \\
 & & \vdots & & & & \vdots & \\
(X_1^{(N)}, Y_1^{(N)}) & (X_2^{(N)}, Y_2^{(N)}) & \ldots & (X_n^{(N)}, Y_n^{(N)}) & \longrightarrow & b_0^{(N)} & \text{and} & b_1^{(N)}
\end{array}
$$

It would be reasonable to expect that the average of all b_1's is 'close' to the true regression slope β_1 (and the same for the intercepts).

To meet these 'reasonable expectations' we need to impose additional requests on the simple regression model.

Simple Linear Regression Assumptions

$$\underbrace{Y}_{\text{Dependent Variable}} = \underbrace{\underbrace{\beta_0}_{\text{y-intercept}} + \underbrace{\beta_1}_{\text{Slope}} \underbrace{X}_{\text{Predictor (Independent Variable)}}}_{\text{Linear component}} + \underbrace{\underbrace{\varepsilon}_{\text{Noise}}}_{\text{Random Error component}}$$

The relationship between Y and X is assumed to follow *LINE:*

Linearity: Relationship between X and Y is linear.

Independence of Errors: Given sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ the errors

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

are a sample from ε, hence independent.

Normality of Error: ε has a normal distribution with population mean zero.

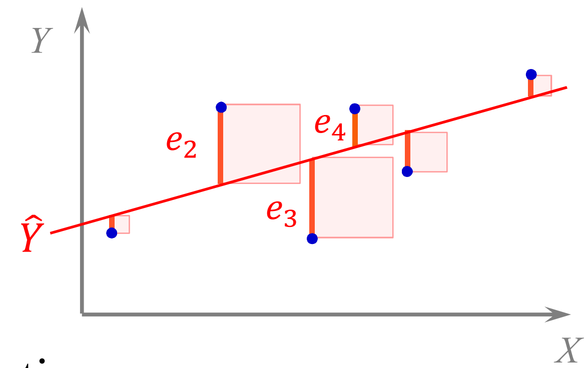
Equal Variance: (homoscedasticity) The variance of ε is constant with respect to X.

$\varepsilon \sim N(0, \sigma^2)$ with unknown variance σ^2 called *regression variance*.

Residuals

Recall: residuals $e_i = Y_i - \hat{Y}_i$.

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



For sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ notice the distinction:

Errors: $\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$

Sample from ε
($\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent r.v.'s)

Residuals: $e_i = Y_i - (b_0 + b_1 X_i)$

Dependent r.v.'s with sum 0 (from LSM), since b_0 and b_1 are r.v.'s completely dependent on (X_i, Y_i) (i.e., computed from them)

Problem: We do not know the values of the error terms ε_i and we only know the residuals e_i which approximate the error terms.

Given X_i's and Y_i's, we check the regression assumptions by examining various plots of residuals for *linearity*, *independence*, *homoscedasticity,* and finally for *normality* assumption.

The most practical way to conduct this is the *Graphical Analysis of Residuals*: scatter-plot of the residuals vs. values of X_i's or the fitted values $\hat{Y}_i$'s.

Graphical Analysis of Residuals

OPIM 5603 Fall 2019

These are mainly scatter plots, typically of the residuals vs. values of X_i's or the fitted values $\hat{Y}_i$'s.

Unfortunately these visual inspections are typically better at telling when the model assumption is not valid than when it is.

Typically the order of checking is *linearity*, *independence*, *homoscedasticity,* and lastly *normality*.

Linearity check: plot the residuals vs. values of X_i's (simple regression).

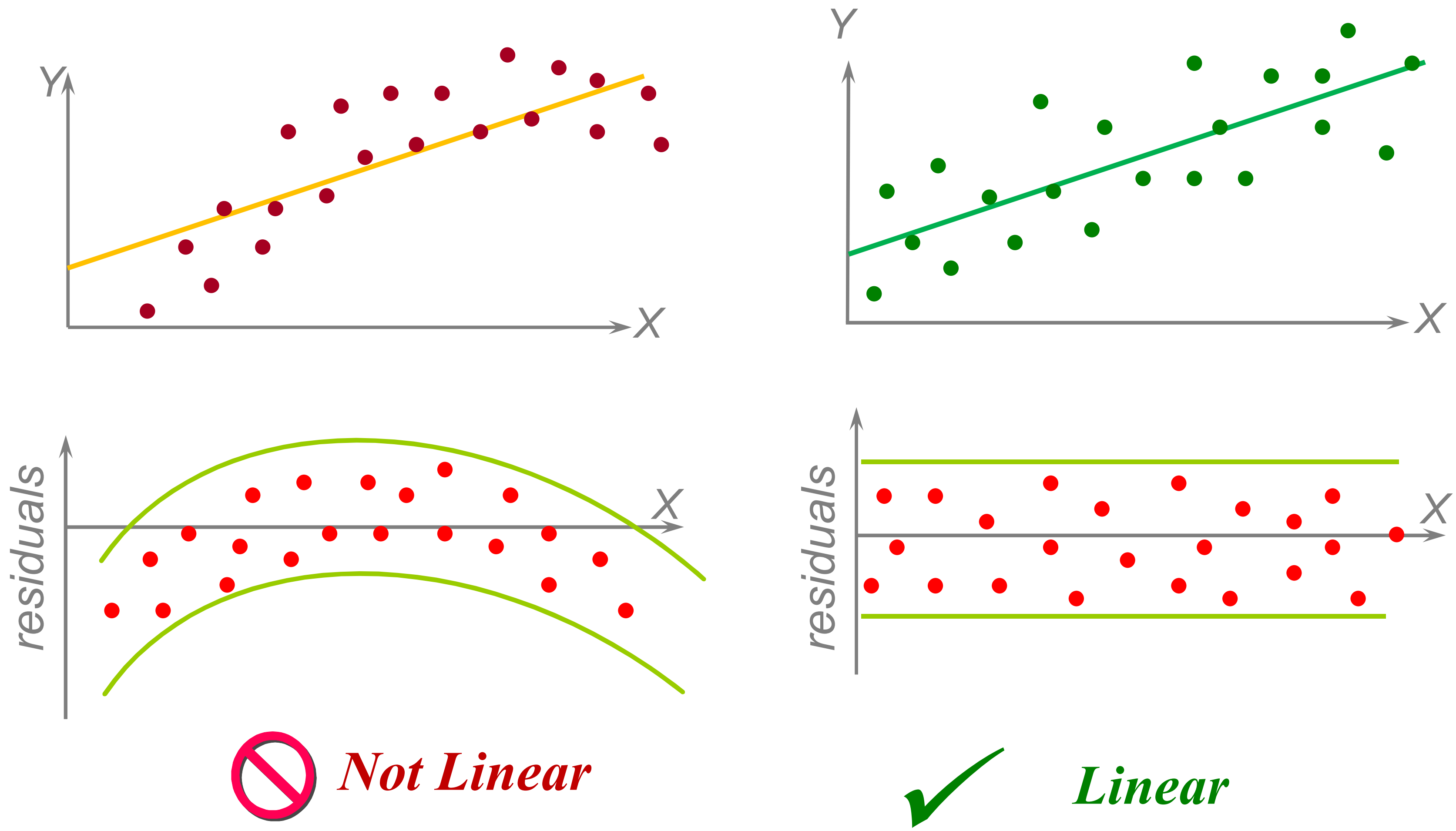
Independence check: plot residuals against any variables used in the technique: X_i's, $\hat{Y}_i$'s or Y_i's. A pattern that is not random suggests lack of independence.

Keep in mind that the residuals sum to zero and they are not independent so the plot is really a very rough approximation!

Homoscedasticity check: plot residuals against fitted values $\hat{Y}_i$'s and observe if the spread is changing in different ranges of $\hat{Y}_i$'s.

Normality check: Create a *quantile* (*q*-*q*) *plot*.

Residual Analysis for Linearity

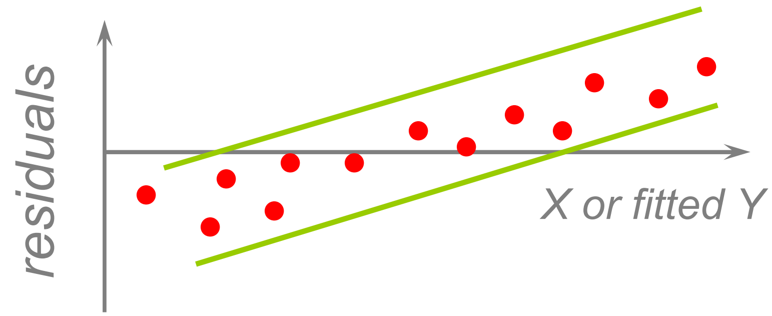


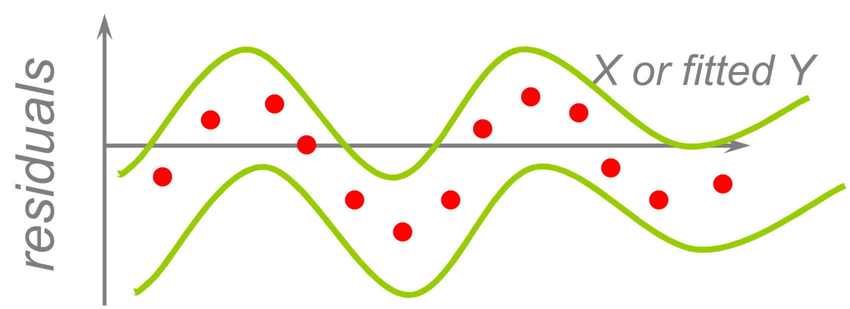
Residual Analysis for Independence

OPIM 5603 Fall 2019

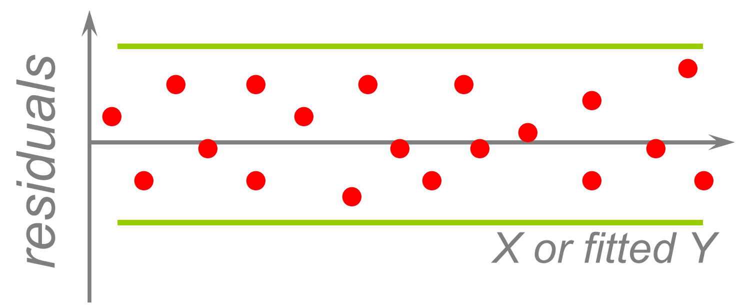
Slide 3.23

Not Independent

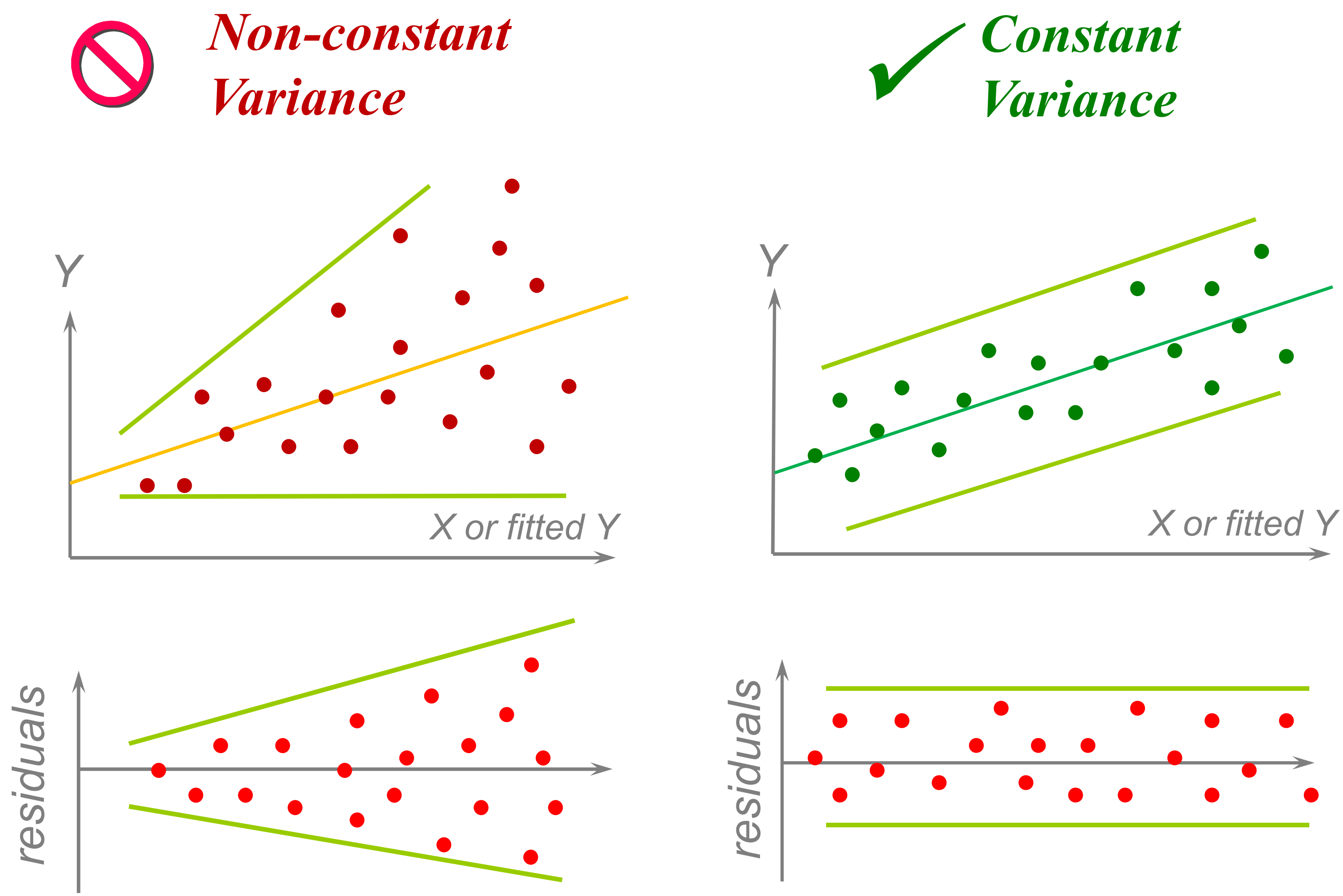




(might be) ***Independent***

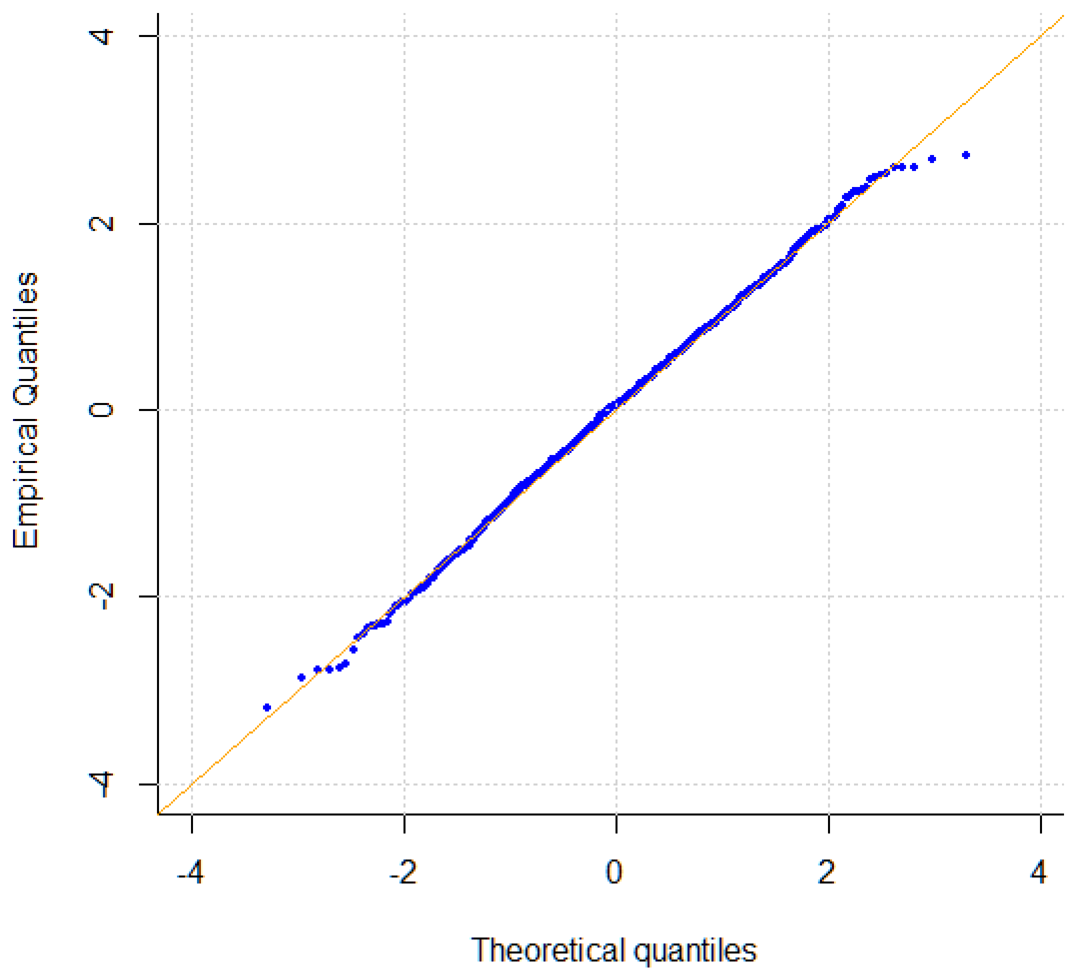


Residual Analysis for Homoscedasticity



Residual Analysis for Normality

Commonly accepted way to check whether the sample is taken from the (normal) distribution is the *(Normal) quantile plot*: normal errors will approximately display in a straight line:



R function *lm*

```
Call: lm(formula = weight ~ height, data = women)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7333 -1.1333 -0.3833  0.7417  3.1167

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -87.51667    5.93694  -14.74  1.71e-09 ***
height        3.45000    0.09114   37.85  1.09e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.525 on 13 degrees of freedom
Multiple R-squared:  0.991,     Adjusted R-squared:  0.9903
F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

Regression Variance & Standard Error

Simple regression: $Y = \beta_0 + \beta_1 X + \varepsilon$, with $\varepsilon \sim N(0,\sigma^2)$.

σ^2 is called the regression variance and (usually) is unknown.

Given a sample (X_i, Y_i) the errors ε_i are a sample from ε.
They are unknown: the residuals e_i are our best estimates.

Thus the residual *sample variance* can be used as an estimator for σ^2:

$$\frac{1}{n-1}\sum_{i=1}^{n}(e_i - \cancel{\bar{e}}^{\,0})^2 = \frac{1}{n-1}\sum_{i=1}^{n} e_i^{\,2} = \frac{SSE}{n-1}.$$

The estimator used in practice is *Sample Regression Variance* defined as

$$S^2 = \frac{SSE}{n-2}.$$

Division by $n-2$ instead of $n-1$ is justified by the fact that two parameters (*slope* and *intercept*) are involved in obtaining this estimator, hence two *degrees of freedom* are 'lost'.

*Regression (*or *Residual) Standard Error* (S) is the square root of this quantity.

Probability Space for Linear Regression

$$\underbrace{Y(\omega_1,\omega_2)}_{\text{Dependent Variable}} = \underbrace{\overbrace{\beta_0}^{\text{y-intercept}} + \overbrace{\beta_1}^{\text{Slope}}\overbrace{X(\omega_1)}^{\text{Predictor (Independent Variable)}}}_{\text{Linear component } \hat{Y}} + \underbrace{\overbrace{\varepsilon(\omega_2)}^{\text{Noise}}}_{\text{Random Error component}}$$

The coefficients *Slope* and *Intercept* are numbers.

Note: *X* lives in an underlying *population space*. Since errors can have different values for identical *X* values it is reasonable to postulate that they *live* in another *space*.

Think of it this way: There are two separate sources of underlying randomness: one for *X* and another for the error *ε*.

Regression Coefficients

Simple regression: $Y = \beta_0 + \beta_1 X + \varepsilon$ with $\varepsilon \sim N(0,\sigma^2)$ and β_0 and β_1 numbers

Given a sample (X_1,ε_1), (X_2,ε_2), …, (X_n,ε_n) from (X,ε), suppose we "freeze" the randomness of X_i's while keeping the ε_i's random.

Then X_i's become the numbers x_i while Yi's are still the random variables with randomness inherited from ε_i's: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Within this construct the *Least Square Method* produces

$$b_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})}{\sum_{k=1}^{n}(x_k-\bar{x})^2} = \frac{SxY}{SSx} \quad \text{and} \quad b_0 = \bar{Y} - b_1\bar{x}.$$

Note: b_0 and b_1 are random variables dependent on errors ε_i (a sample from ε).

Then: (1) b_0 and b_1 are normal random variables,

(2) $E(b_1) = \beta_1$ and $Var(b_1) = \dfrac{\sigma^2}{SSx}$

(3) $E(b_0) = \beta_0$ and $Var(b_0) = \dfrac{\bar{x}^2\sigma^2}{SSx}$

Note: b_0 depends on b_1 so it is enough to prove (1) for b_1.

Proof

Optional

$$SSx = \sum_{k=1}^{n}(x_k - \bar{x})^2 \quad = \sum_{k=1}^{n}(x_k{}^2 - 2x_k\bar{x} + \bar{x}^2)$$

$$= \sum_{k=1}^{n}(x_k{}^2 - x_k\bar{x}) + \sum_{k=1}^{n}(\bar{x}^2 - x_k\bar{x})$$

$$= \sum_{k=1}^{n} x_k(x_k - \bar{x}) + \bar{x}\cancel{\boxed{\sum_{k=1}^{n}(\bar{x} - x_k)}} \quad = \sum_{k=1}^{n} x_k(x_k - \bar{x})$$

$$= n\bar{x} - \sum_{k=1}^{n} x_k \quad = 0$$

$$SxY = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) \quad = \sum_{i=1}^{n}(x_i - \bar{x})Y_i - \cancel{\boxed{\left[\sum_{i=1}^{n}(x_i - \bar{x})\right]\bar{Y}}}$$

$$= \sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i)$$

$$= \beta_0 \cancel{\boxed{\sum_{i=1}^{n}(x_i - \bar{x})}} + \beta_1 \overset{= SSx}{\boxed{\sum_{i=1}^{n} x_i(x_i - \bar{x})}} + \sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i$$

$$= \beta_1 SSx + \sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i$$

$$b_1 = \frac{SxY}{SSx} = \frac{\beta_1 SSx + \sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{SSx} \quad = \beta_1 + \sum_{i=1}^{n} \overset{c_i}{\boxed{\frac{x_i - \bar{x}}{SSx}}}\varepsilon_i$$

Proof cont.

Optional

$$b_1 = \beta_1 + \sum_{i=1}^{n} \frac{x_i - \bar{x}}{SSx} \varepsilon_i = \beta_1 + \sum_{i=1}^{n} c_i \varepsilon_i$$

Notice that β_1 is a number and so are the x_i's (randomness of Xi's is "frozen"). Consequently so are $\bar{x}$ (average of numbers) and SSx. Thus c_i's defined above are numbers.

Note: this shows that the randomness of b_1 is instigated by the errors ε_i's.

Since ε_i's are independent $N(0,\sigma^2)$ (they are a sample from ε)

$$\Rightarrow \sum_{i=1}^{n} c_i \varepsilon_i \sim N\left(0, \sigma^2 \sum_{i=1}^{n} c_i^{\,2}\right)$$

$$\Rightarrow b_1 = \beta_1 + \sum_{i=1}^{n} c_i \varepsilon_i \sim N\left(\beta_1, \sigma^2 \sum_{i=1}^{n} c_i^{\,2}\right)$$

This proves (1). Clearly $E(b_1) = \beta_1$ and

$$Var(b_1) = \sigma^2 \sum_{i=1}^{n} c_i^{\,2} = \sigma^2 \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{SSx}\right)^2 = \frac{\sigma^2}{SSx^2} \overbrace{\boxed{\sum_{i=1}^{n} (x_i - \bar{x})^2}}^{= SSx} = \frac{\sigma^2}{SSx}$$

This proves (2).

Proof cont. Optional

$$E(b_0) = E(\bar{Y} - b_1\bar{x}) = E(\bar{Y}) - E(b_1\bar{x}) = E(\bar{Y}) - E(b_1)\bar{x} = E(\bar{Y}) - \beta_1\bar{x}$$

Notice that $\bar{Y} = \beta_0 + \beta_1\bar{x} + \bar{\varepsilon}$, where $\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$.

$$\Rightarrow E(\bar{Y}) = \beta_0 + \beta_1\bar{x} + \underbrace{\boxed{E(\bar{\varepsilon})}}_{=0} = \beta_0 + \beta_1\bar{x}$$

$$\Rightarrow E(b_0) = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$$

Finally,

$$Var(b_0) = Var(\bar{Y} - b_1\bar{x}) = Var(-\bar{x}b_1) = (-\bar{x})^2 Var(b_1) = \bar{x}^2\frac{\sigma^2}{SSx}$$

This proves (3).

More on Regression Slope

Hence $b_1 \sim N\left(\beta_1, \boxed{\frac{\sigma^2}{SSx}}\right)$. *Slope Variance*

The *regression variance* σ^2 is unknown, but its unbiased estimator is the *Sample Regression Variance* $S^2 = \frac{SSE}{n-2}$.

Since SSx is a number in this context (does not depend on error terms), the unbiased estimator for the slope variance is

$$\frac{S^2}{SSx} = \frac{SSE}{(n-2)SSx} \qquad \textit{Sample Slope Variance}$$

The square root of this quantity is *Slope Standard Error* (S_{b1}).

The following results should not be surprising:

$$\frac{b_1 - \beta_1}{\frac{\sigma}{\sqrt{SSx}}} \sim N(0,1) \qquad \frac{b_1 - \beta_1}{S_{b1}} \sim t(n\text{–}2)$$

Two-tailed t-test for Regression Slope

$$t_{stat} = \frac{b_1 - \beta_1}{S_{b_1}} = (b_1 - \beta_1)\sqrt{\frac{(n-2)\,SSx}{SSE}} \sim t(n\text{–}2)$$

Given (x_1,Y_1), (x_2,Y_2), …, (x_n,Y_n), formulate the hypotheses

$$H_0\text{: } \boxed{\beta_1 = 0} \qquad H_a\text{: } \beta_1 \neq 0$$

and compute:

the *t-statistic* $\qquad t_{stat} = \dfrac{b_1 - \beta_1}{S_{b_1}} = (b_1 - 0)\sqrt{\dfrac{(n-2)\,SSx}{SSE}}$

the *p-value* $\qquad P\big(t(n-2) \le -\,|\,t_{stat}\,|\big) + P\big(t(n-2) \ge |\,t_{stat}\,|\big)$

Given the agreed significance level α,

if p value > α, accept the null hypothesis!

if p value ≤ α, reject the null hypothesis!

Conclusion: *R* function *lm*

OPIM 5603 Fall 2019

Slide 3.35

```
Call: lm(formula = weight ~ height, data = women)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7333 -1.13
```

Slope Standard Error (see 3.33)

t_{stat} (see 3.34)

p-value (see 3.34)

```
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) -87.51667     5.93694   -14.74  1.71e-09 ***
height        3.45000     0.09114    37.85  1.09e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.525 on 13 degrees of freedom
Multiple R-squared:  0.991,    Adjusted R-squared:  0.9903
F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

Regression Standard Error (see 3.28)

Multiple Linear Regression

Dependent Variable (*Outcome* or *Response Variable*): Y

Predictors (*Independent*, *Input Variables*, *Repressors*): $X_1, \ldots, X_p$

Assumes the following relationship:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon,$$

where β's are the *coefficients* and ε is the *noise*.

As in simple regression, for the given sample from the random vector

$$(X_1, X_2, \ldots, X_p, Y)$$

the coefficients are obtained via the *Least Squares Method*.

(involves solving p+1 equations with partial derivatives set to zero)

There are also regression models with multiple dependent variables; these are usually called *Multivariate Linear Regression* models.

Regression via *lm()*

OPIM 5603 Fall 2019

Slide 3.37

Symbol	Usage
$\sim$	Separates response variables on the left from the explanatory variables on the right. For example, a prediction of y from x, z, and w would be coded $y \sim x + z + w$.
$+$	Separates predictor variables.
:	Denotes an interaction between predictor variables. A prediction of y from x, z, and *the interaction between* x *and* z would be coded $y \sim x + z + x{:}z$.
*	A shortcut for denoting all possible interactions. The code $y \sim x * z * w$ expands to $y \sim x + z + w + x{:}z + x{:}w + z{:}w + x{:}z{:}w$.
^	Denotes interactions up to a specified degree. The code $y \sim (x + z + w)\hat{}2$ expands to $y \sim x + z + w + x{:}z + x{:}w + z{:}w$.
.	A place holder for all other variables in the data frame except the dependent variable. E.g., if a data frame contained the variables x, y, z, and w, then the code $y \sim .$ would expand to $y \sim x + z + w$.
$-$	A minus sign removes a variable from the equation. For example, $y \sim (x + z + w)\hat{}2 - x{:}w$ expands to $y \sim x + z + w + x{:}z + z{:}w$.
-1	Suppresses the intercept. For example, the formula $y \sim x - 1$ fits a regression of y on x, and forces the line through the origin at $x = 0$.
I( )	Elements within the parentheses are interpreted arithmetically. For example, $y \sim x + (z + w)\hat{}2$ would expand to $y \sim x + z + w + z{:}w$. In contrast, the code $y \sim x + \mathrm{I}((z + w)\hat{}2)$ would expand to $y \sim x + h$, where h is a new variable created by squaring the sum of z and w.
function	Mathematical functions can be used in formulas. For example, $\log(y) \sim x + z + w$ would predict $\log(y)$ from x, z, and w.

Example: Multiple Regression 'Boston' data

OPIM 5603 Fall 2019

Slide 3.38

```
Call: lm(formula = medv ~ ., data = Boston)
   ...
Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144   3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287   0.001087 **
zn           4.642e-02  1.373e-02   3.382   0.000778 ***
indus        2.056e-02  6.150e-02   0.334   0.738288
chas         2.687e+00  8.616e-01   3.118   0.001925 **
nox         -1.777e+01  3.820e+00  -4.651   4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116    < 2e-16 ***
age          6.922e-04  1.321e-02   0.052   0.958229
dis         -1.476e+00  1.995e-01  -7.398   6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613   5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280   0.001112 **
ptratio     -9.527e-01  1.308e-01  -7.283   1.31e-12 ***
black        9.312e-03  2.686e-03   3.467   0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347    < 2e-16 ***
  ...
Residual standard error: 1.525 on 13 degrees of freedom
Multiple R-squared:  0.991, Adjusted R-squared:  0.9903
```

at any reasonable significance we can conclude that *medv* does not depend linearly on *indus*

t test H_0: *Slope*(*age*) = 0

H_0 is accepted at 5%, 10%, 20% (i.e., any reasonable) significance.

$\Rightarrow$ We can conclude (at any reasonable significance) that *medv* does not depend linearly on *age*

Regression Models

Optional

Linear Regression models are used to fit a linear relationship between a *continuous* dependent variable *Y* and a set of one or more *continuous* predictors;

Logistic Regression models measure the relationship between a *categorical* dependent variable *Y* and a set of one or more *continuous* predictors;

Simple Regression: only one predictor

Multiple Regression: multiple predictors