08 | Linear Regression

Ivan Corneillet

Data Scientist

Learning Objectives

After this lesson, you should be able to:

- Define simple linear regression
- Build a linear regression model using *statsmodels*
- Evaluate model fit using statistical analysis (t-tests, p-values, t-values, confidence intervals)

Simple Linear Regression

Simple Linear Regression

- The simple linear regression model captures a linear relationship between a single feature variable x and a response variable y

$$y = \beta_0 + \beta_1 \cdot x + \varepsilon$$

- y is the **response** variable (what we want to predict); also called *dependent* variable, *endogenous* variable, or *regressand*
- x is the **feature** variable (what we use to train the model); also called *explanatory* variable, *independent* variable, *exogenous* variable, or *regressor*
- β_0 and β_1 are the **regression's coefficients**; also called the *model's parameters*
  - β_0 is the line's intercept; β_1 is the line's slope
- ε is the **error** term; also called the residual

Simple Linear Regression (cont.)

- Given $x = \left(x^{(1)}, x^{(2)}, \ldots, x^{(n)}\right)$, $y = \left(y^{(1)}, y^{(2)}, \ldots, y^{(n)}\right)$, and $\varepsilon = \left(\varepsilon^{(1)}, \varepsilon^{(2)}, \ldots, \varepsilon^{(n)}\right)$, we can formulate the linear model as

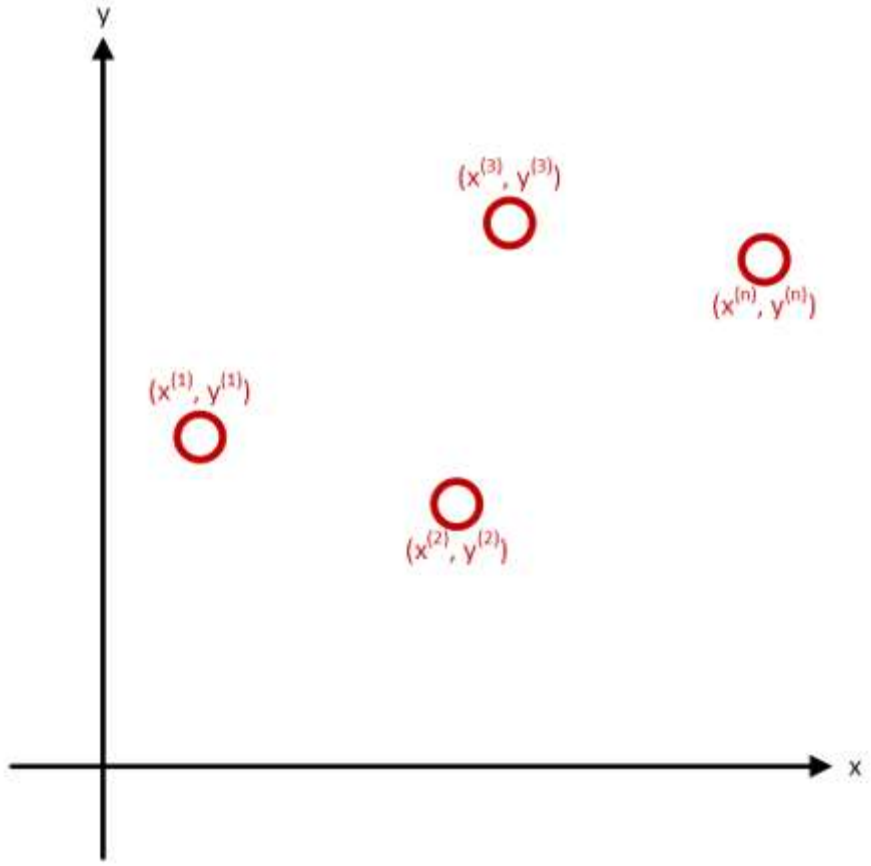
$$y^{(i)} = \beta_0 + \beta_1 \cdot x^{(i)} + \varepsilon^{(i)}$$

- In words, this equation says that for each sample i, $y^{(i)}$ can be explained by $\beta_0 + \beta_1 \cdot x^{(i)}$

- In our Python environment, x and y represent *pandas Series* and $x^{(i)}$ and $y^{(i)}$ their values at index i
- E.g. (SF housing dataset),
  - x is the property's size (`df.Size`)
  - y is the property's sale price (`df.SalePrice`)

Simple Linear Regression (cont.)

- ε is a "white noise" disturbance which we **do not observe**
  - ε models how the observations deviate from the exact slope-intercept relation
- We **do not observe** the constants β_0 or β_1 either, so we have to estimate them



Simple Linear Regression (cont.)

- E.g. (SF housing dataset),

$$\widehat{SalePrice} = \hat{\beta}_0 + \hat{\beta}_1 \cdot Size$$

How to interpret *statsmodels* report?

Dep. Variable:	SalePrice	**R-squared:**	0.236
Model:	OLS	**Adj. R-squared:**	0.235
Method:	Least Squares	**F-statistic:**	297.4
Date:		**Prob (F-statistic):**	2.67e-58
Time:		**Log-Likelihood:**	-1687.9
No. Observations:	967	**AIC:**	3380.
Df Residuals:	965	**BIC:**	3390.
Df Model:	1		
Covariance Type:	nonrobust		

The model's fit

Is the model's fit significant?

The estimated coefficients $\widehat{\beta}_0$ (the intercept) and $\widehat{\beta}_1$ (the slope; "size")

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	0.1551	0.084	1.842	0.066	-0.010 0.320
Size	0.7498	0.043	17.246	0.000	0.664 0.835

Are these estimated significant? (i.e., are they meaningful?; do they make sense?)

Omnibus:	1842.865	**Durbin-Watson:**	1.704
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3398350.943
Skew:	13.502	**Prob(JB):**	0.00
Kurtosis:	292.162	**Cond. No.**	4.40

Simple Linear Regression

Interpreting the regression's coefficients $\hat{\beta}$

Interpreting the regression's coefficients

$\hat{\beta}_0 = .155$

- What's the unit of $\hat{\beta}_0$?
  - $[\hat{\beta}_0] = [\widehat{SalePrice}] = \$M$
- How to interpret $\hat{\beta}_0$?
  - $\hat{\beta}_0 = .155\ (\$M) = \$155k$
  - $\widehat{SalePrice}(Size = 0) =$ $\hat{\beta}_0 + \hat{\beta}_1 0 = \hat{\beta}_0$
  - The model predicts that a property of 0 sqft would cost $155k

$\hat{\beta}_1 = .750$

- What's the unit of $\hat{\beta}_1$?
  - $[\hat{\beta}_1] = \frac{[SalePrice]}{[Size]} = \frac{\$M}{1{,}000\ sqft}$
- How to interpret $\hat{\beta}_1$?
  - $\hat{\beta}_1 = .750\ (\$M/1{,}000\ sqft)$
  - $= \$750k/1{,}000\ sqft$
- The model predicting that each additional 1,000 sqft costs $750k

Simple Linear Regression

Are the regression's coefficients $\hat{\beta}$ significant?

Are the regression's coefficients $\hat{\beta}$ significant?

The β coefficients follow a normal distribution:

$$\boldsymbol{\mu_\beta \sim N\left(\beta, \left(X^T X\right)^{-1}\sigma^2\right)}$$

(or)

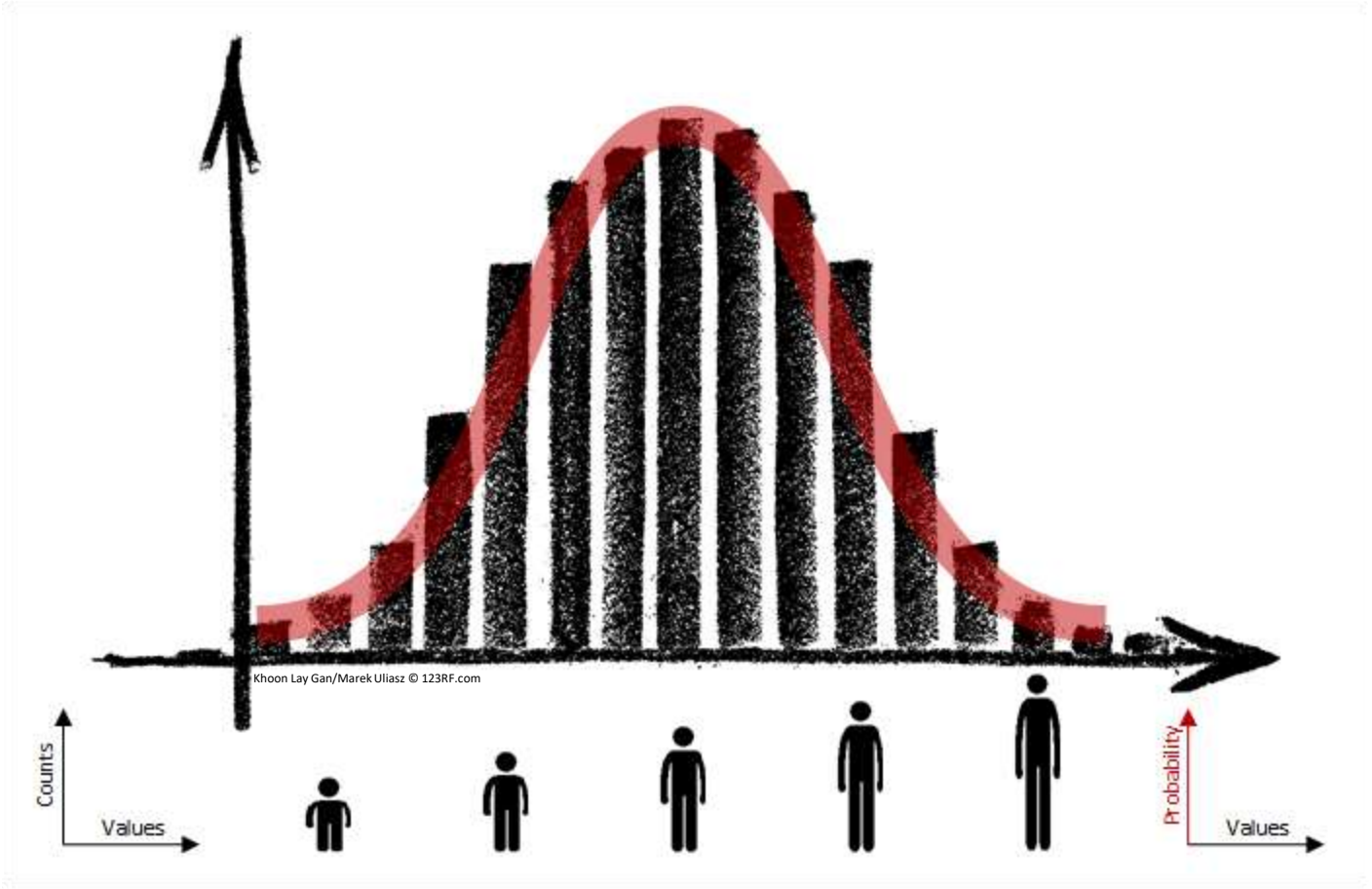
$$\boldsymbol{\mu_{\beta_j} \sim N\left(\beta_j, v_j\sigma^2\right)}$$

$$\left(X^T X\right)^{-1} = \begin{pmatrix} v_0 = v_{0,0} & \cdots & v_{0,j} & \cdots \\ \vdots & \ddots & & \\ v_{j,0} & & v_j = v_{j,j} & \\ \vdots & & & \ddots \end{pmatrix}$$

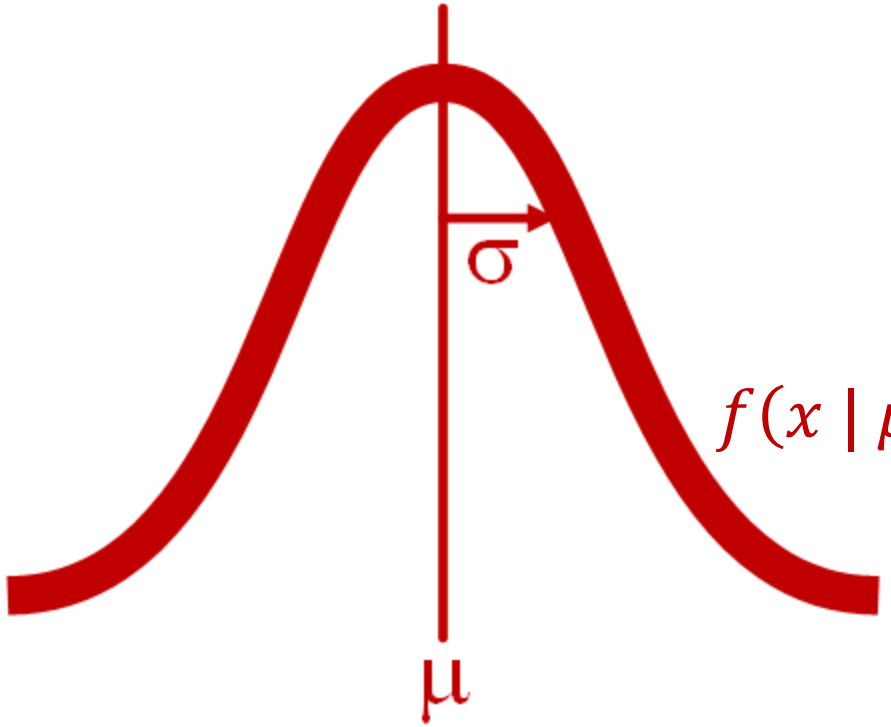
(v_j is the j^{th} diagonal element of $\left(X^T X\right)^{-1}$)

The Normal Distribution

People’s height follows a bell shape distribution. (For men in the US, the average height is around 70 inches (5’10) with a standard deviation of 4 inches; few people are shorter than 67 inches; few are as tall as 73 inches)

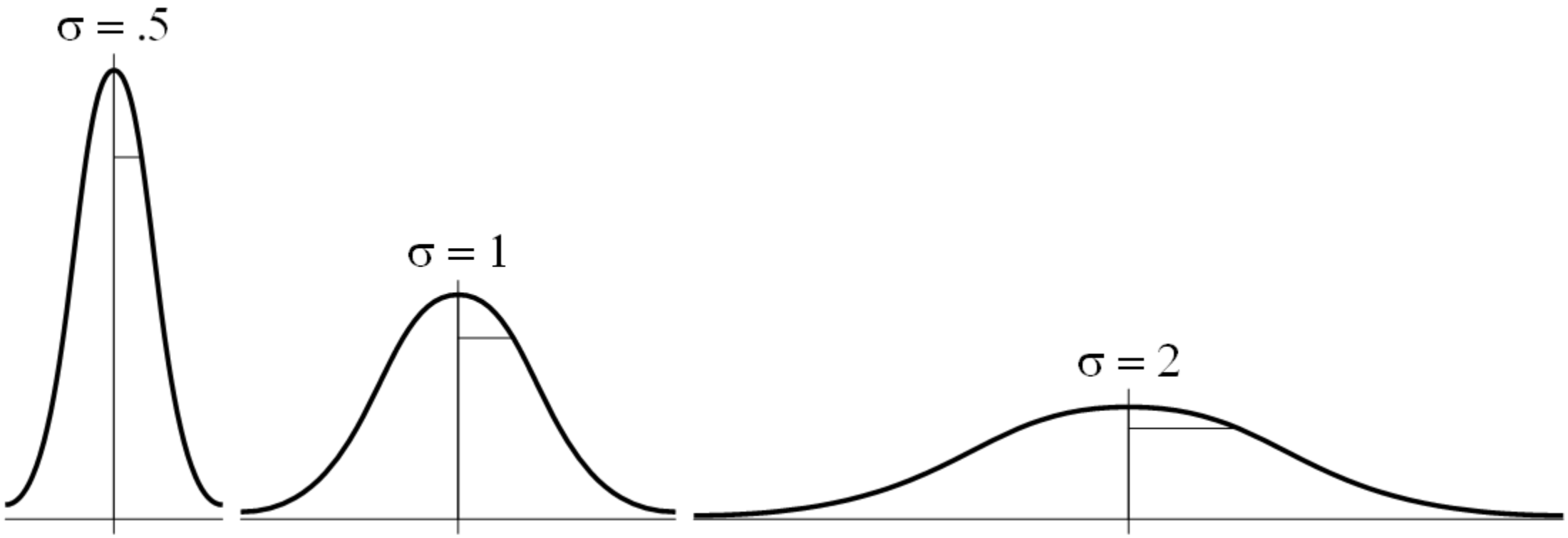


The Normal Distribution

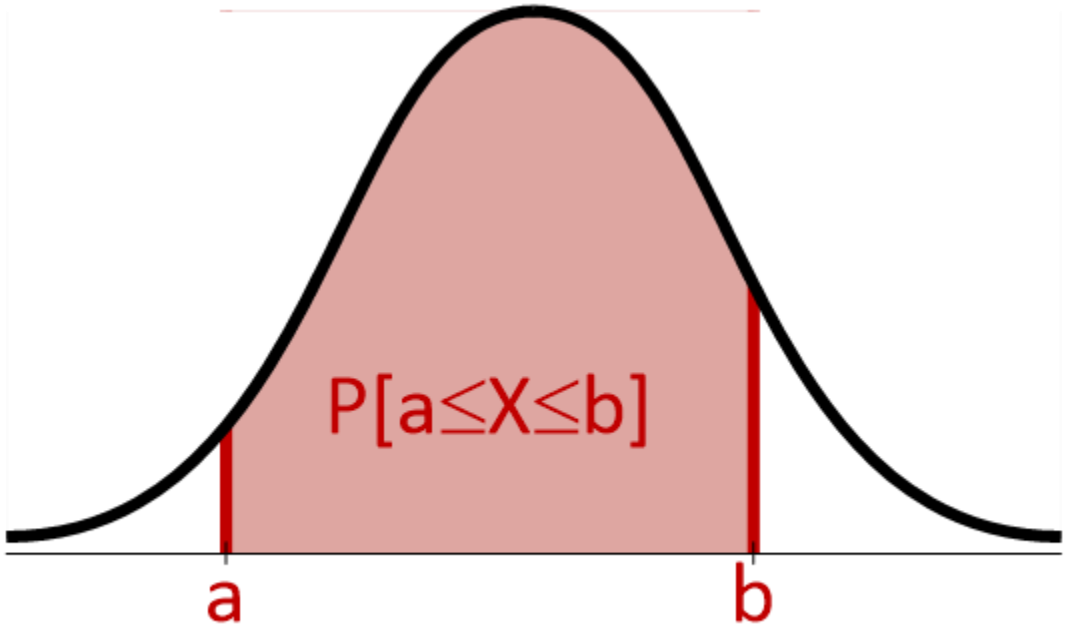


$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

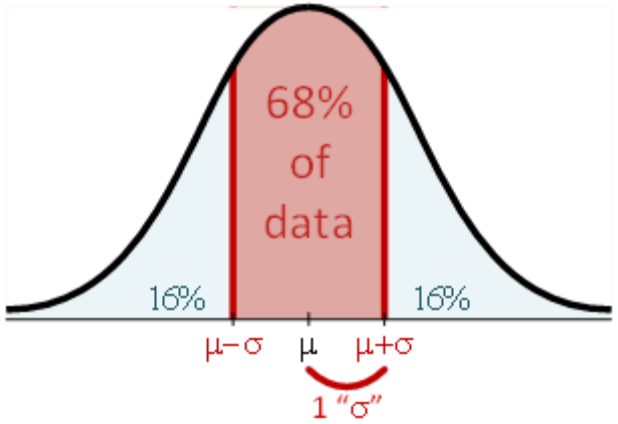
This bell-shaped curve is a probability density function (PDF): The area under the curve is always 1 (for any σ) (cont.)

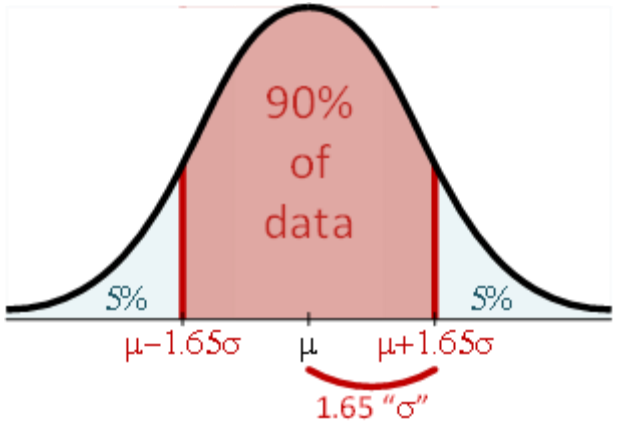


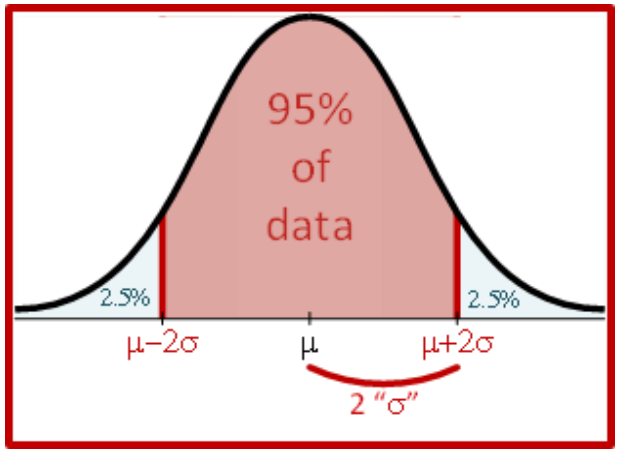
The area under the curve is called a Cumulative Distribution Function (CDF)

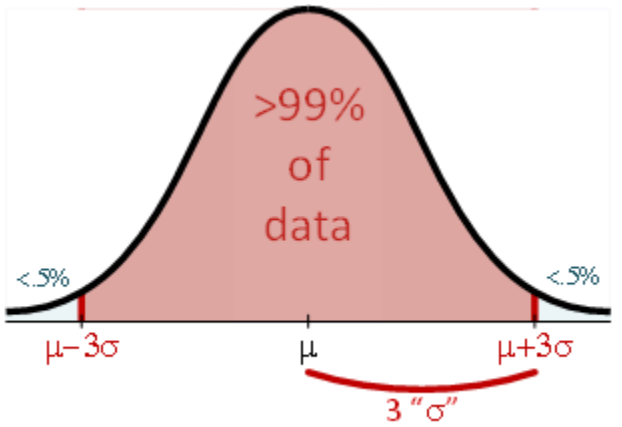


The 68 – 90 – 95 – 99.7 Rule (cont.)







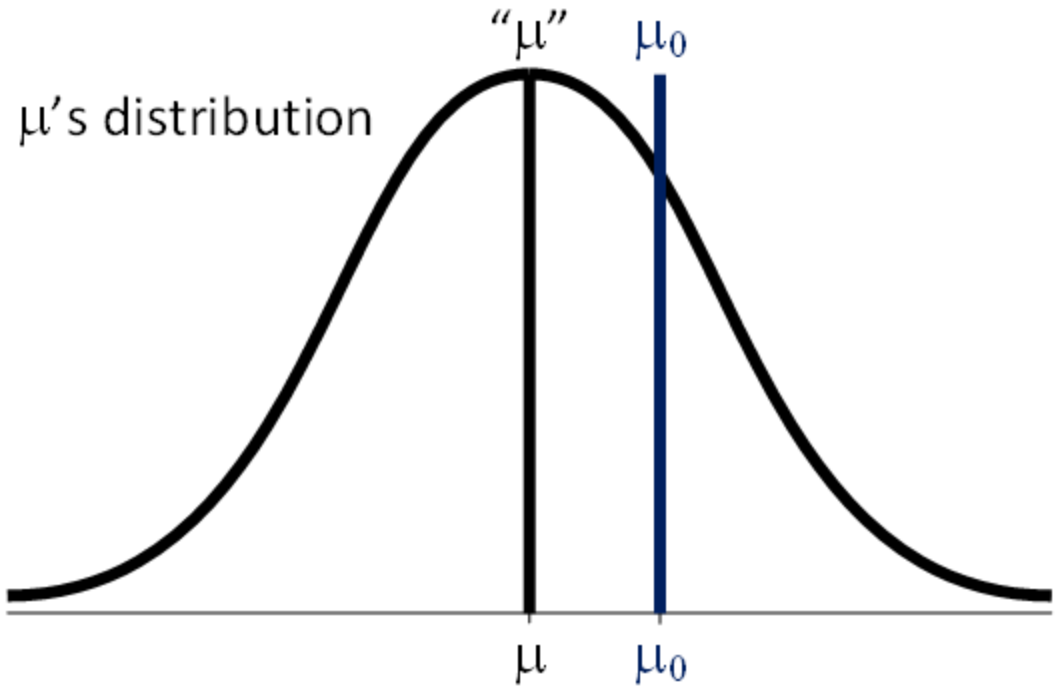


Hypothesis Testing

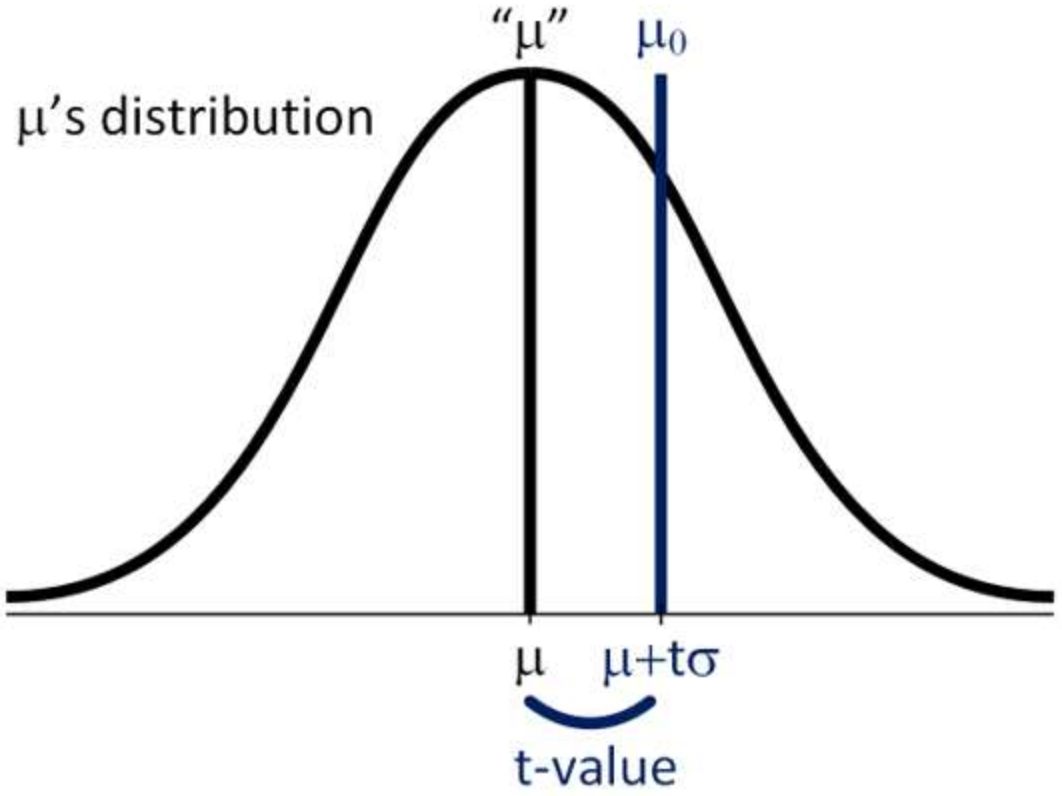
Hypothesis Testing

- A hypothesis is an assumption about a population parameter. E.g.,
  - μ_{β_0} = <a specific value, e.g. .155>
  - μ_{β_1} = <a specific value, e.g. .750>
- In both cases, we made a statement about a population parameter that may or may not be true
- The purpose of hypothesis testing is to make a statistical conclusion about **rejecting** or **failing to reject** such statement

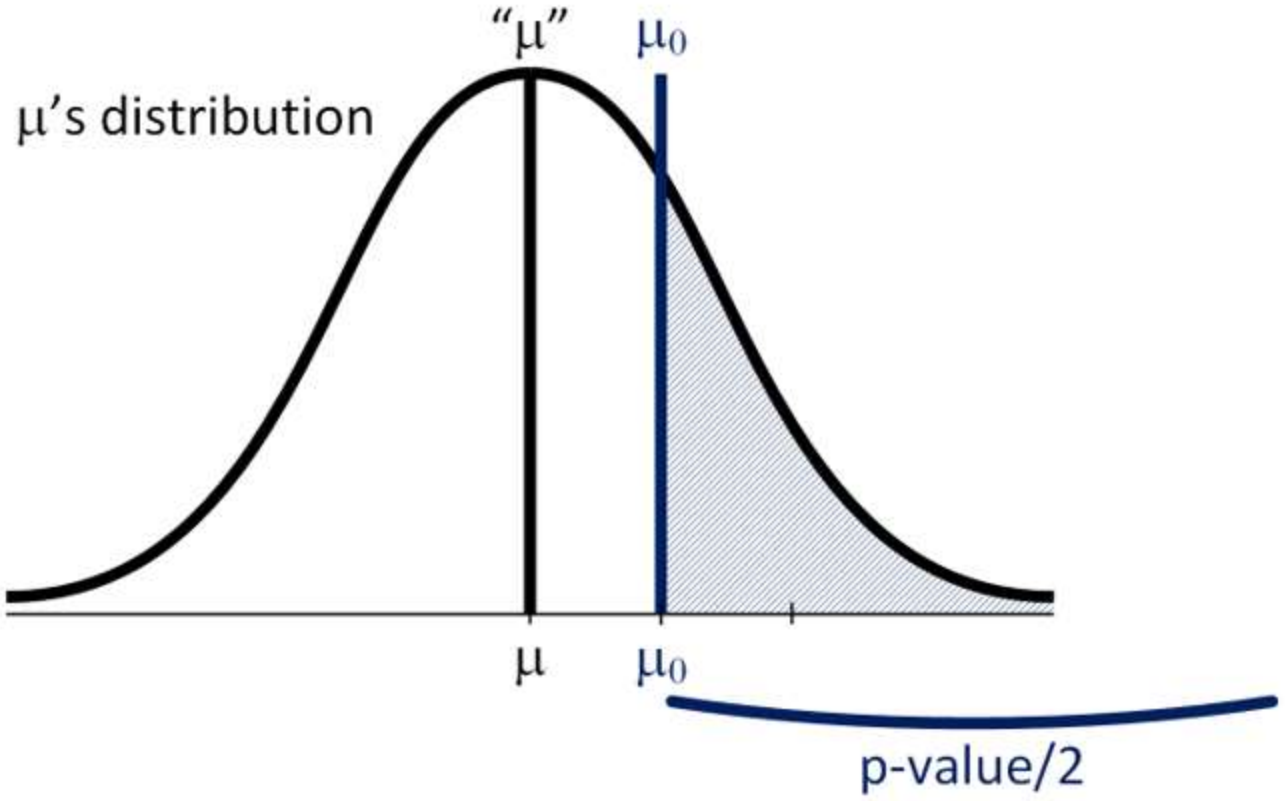
Two-Tail Hypothesis Testing (cont.)



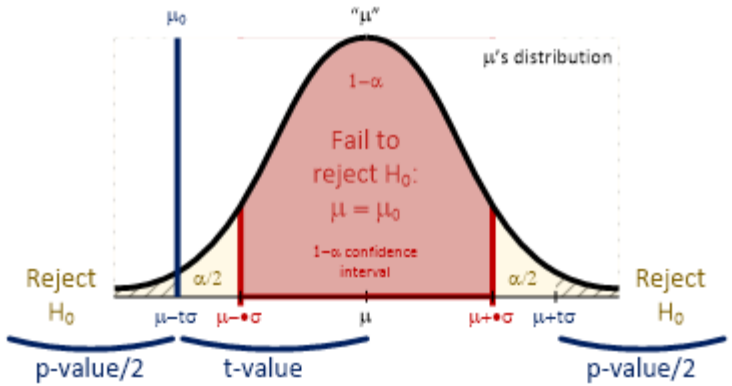
t-value measures the difference to μ_0 in σ. t-values of large magnitudes (either negative or positive) are less likely. The far left and right "tails" of the distribution curve represent instances of obtaining extreme values of t, far from μ

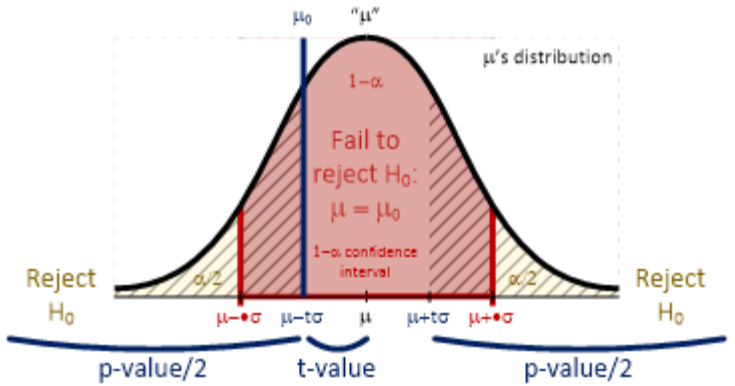


p-value determines the probability (assuming H_0 is true) of observing a more extreme test statistic in the direction of H_a than the one observed



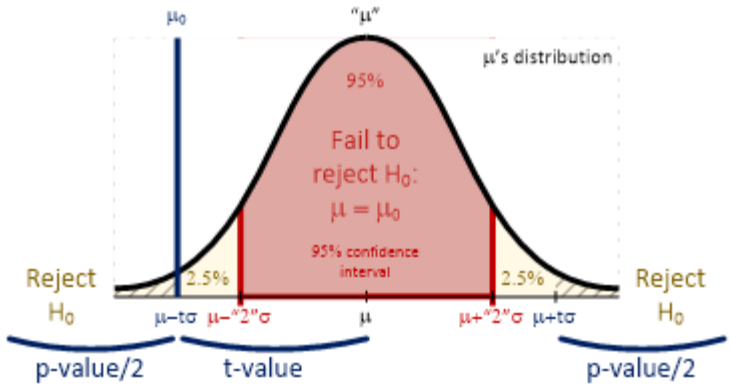
Two-Tail Hypothesis Testing *(simplified)* (cont.)

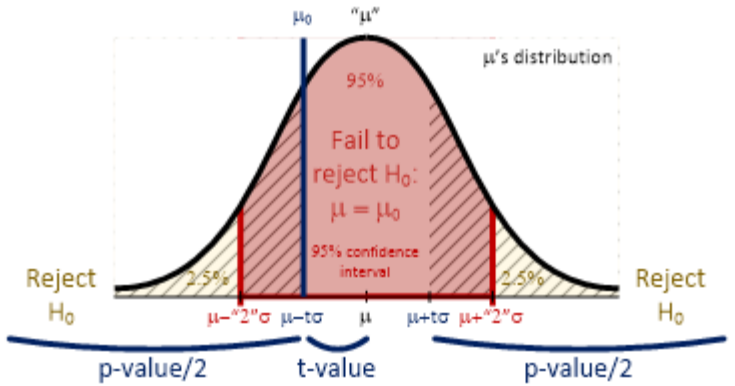




|t-value|	p-value	$1-\alpha$ Confidence Interval ($[\mu -\cdot \sigma, \mu +\cdot \sigma]$)	H_0 / H_a	Outcome
$< \cdot$	$> \alpha$	μ_0 is inside	Did not find evidence that $\mu \neq \mu_0$: Fail to reject H_0	$\mu = \mu_0$
$\geq \cdot$	$\leq \alpha$	μ_0 is outside	Found evidence that $\mu \neq \mu_0$: Reject H_0	$\mu \neq \mu_0$

Two-Tail Hypothesis Testing *(simplified)* ($\alpha = .05$) (cont.)





|t-value|	p-value	95% Confidence Interval ($[\mu - 2\sigma, \mu + 2\sigma]$)	H_0 / H_a	Outcome
$<$ "~ 2"(*) (*) (check t-table)	$> .05$	μ_0 is inside	Did not find evidence that $\mu \neq \mu_0$: Fail to reject H_0	$\mu = \mu_0$
$\geq$ "~ 2"	$\leq .05$	μ_0 is outside	Found evidence that $\mu \neq \mu_0$: Reject H_0	$\mu \neq \mu_0$

Simple Linear Regression

Are the regression's coefficients $\hat{\beta}$ significant? (cont.)

What β_1 would make our multiple linear regression model useless?

- (the simple linear regression model again, without intercept to keep things simple)

$$y = \beta_1 \cdot x + \varepsilon$$

- Answer: If $\beta_1 = 0$, we don't have a linear model
  - ($y = 0$ isn't very exciting, is it?)

Is the regression's coefficient $\hat{\beta}$ significant?

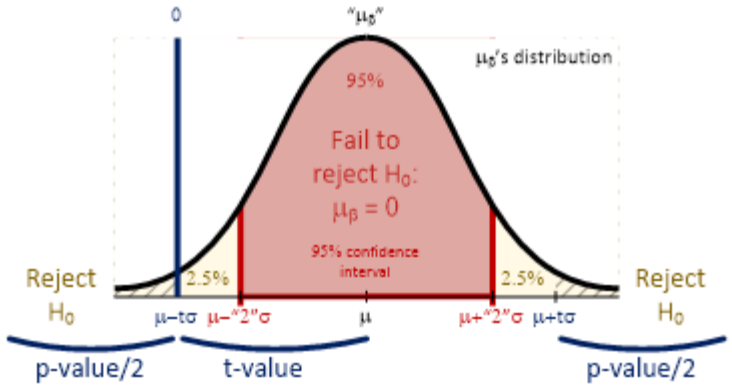
- The *null hypothesis* (H_0) represents the status quo; that the mean of the regression's coefficient β is equal to 0, i.e. that β is not significant:

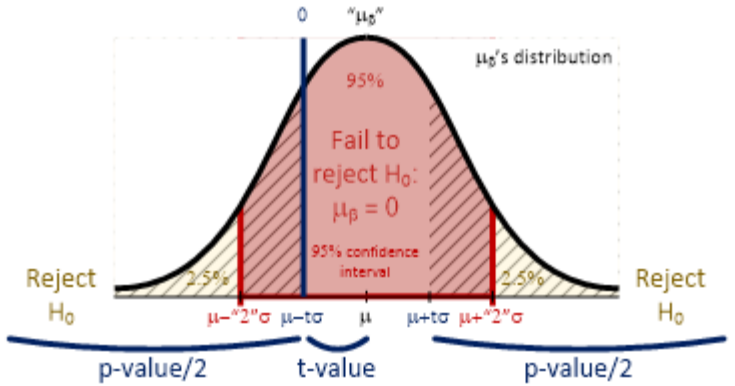
$$H_0\colon\ \mu_\beta = 0$$

- The *alternate hypothesis* (H_a) represents the opposite of the null hypothesis and holds true if the *null hypothesis* is found to be false; that the mean of the regression's coefficient β is not equal to 0, i.e. that β is significant:

$$H_a\colon \mu_\beta \neq 0$$

Is the regression's coefficient $\hat{\beta}$ significant? (at the 5% significance level)





|t-value|	p-value	95% Confidence Interval ($[\mu_\beta - 2\sigma, \mu_\beta + 2\sigma]$)	H_0 / H_a	Outcome
$<$ "~ 2"(*) (*) (check t-table)	$> .05$	0 is inside	Did not find evidence that $\mu_\beta \neq 0$: Fail to reject H_0	$\mu_\beta = 0$; the coefficient β is not significant
$\geq$ "~ 2"	$\leq .05$	0 is outside	Found evidence that $\mu_\beta \neq 0$: Reject H_0	$\mu_\beta \neq 0$; the coefficient β is significant

Simple Linear Regression

Are the regression's coefficients $\hat{\beta}$ significant? (cont.)

`SalePrice` as a function of `Size` (cont.)

Dep. Variable:	SalePrice	**R-squared:**	0.236
Model:	OLS	**Adj. R-squared:**	0.235
Method:	Least Squares	**F-statistic:**	297.4
Date:		**Prob (F-statistic):**	2.67e-58
Time:		**Log-Likelihood:**	-1687.9
No. Observations:	967	**AIC:**	3380.
Df Residuals:	965	**BIC:**	3390.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	0.1551	0.084	1.842	0.066	-0.010 0.320
Size	0.7497	0.043	17.246	0.000	0.664 0.835

Omnibus:	1842.865	**Durbin-Watson:**	1.704
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3398350.943
Skew:	13.502	**Prob(JB):**	0.00
Kurtosis:	292.162	**Cond. No.**	4.40

$$SalePrice\ [\$M] = \underbrace{.155}_{\widehat{\beta}_0} + \underbrace{.750}_{\widehat{\beta}_1} \times Size\ [1{,}000\ sqft]$$

(the slope is significant but not the intercept)

`SalePrice ~ 0 + Size` (‘0’ meaning the intercept is forced to 0) (cont.)

Dep. Variable:	SalePrice	**R-squared:**	0.565
Model:	OLS	**Adj. R-squared:**	0.565
Method:	Least Squares	**F-statistic:**	1255.
Date:		**Prob (F-statistic):**	7.83e-177
Time:		**Log-Likelihood:**	-1689.6
No. Observations:	967	**AIC:**	3381.
Df Residuals:	966	**BIC:**	3386.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Size	0.8176	0.023	35.426	0.000	0.772 0.863

Omnibus:	1830.896	**Durbin-Watson:**	1.722
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3370566.094
Skew:	13.300	**Prob(JB):**	0.00
Kurtosis:	291.005	**Cond. No.**	1.00

$$SalePrice\ [\$M] = \underbrace{0.}_{\widehat{\beta}_0} + \underbrace{.810}_{\widehat{\beta}_1} \times Size\ [1{,}000\ sqft]$$

(the slope is significant)

`SalePrice ~ Size` (with outliers removed) (cont.)

Dep. Variable:	SalePrice	**R-squared:**	0.200
Model:	OLS	**Adj. R-squared:**	0.199
Method:	Least Squares	**F-statistic:**	225.0
Date:		**Prob (F-statistic):**	1.41e-45
Time:		**Log-Likelihood:**	-560.34
No. Observations:	903	**AIC:**	1125.
Df Residuals:	901	**BIC:**	1134.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
Intercept	0.7082	0.032	22.152	0.000	0.645 0.771
Size	0.2784	0.019	15.002	0.000	0.242 0.315

Omnibus:	24.647	**Durbin-Watson:**	1.625
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	53.865
Skew:	0.054	**Prob(JB):**	2.01e-12
Kurtosis:	4.192	**Cond. No.**	4.70

$$SalePrice\ [\$M] =$$

$$\underbrace{.708}_{(was\ .155)} + \underbrace{.278}_{(was\ .750)} \times Size\ [1{,}000\ sqft]$$

(both intercept and slope are now significant)

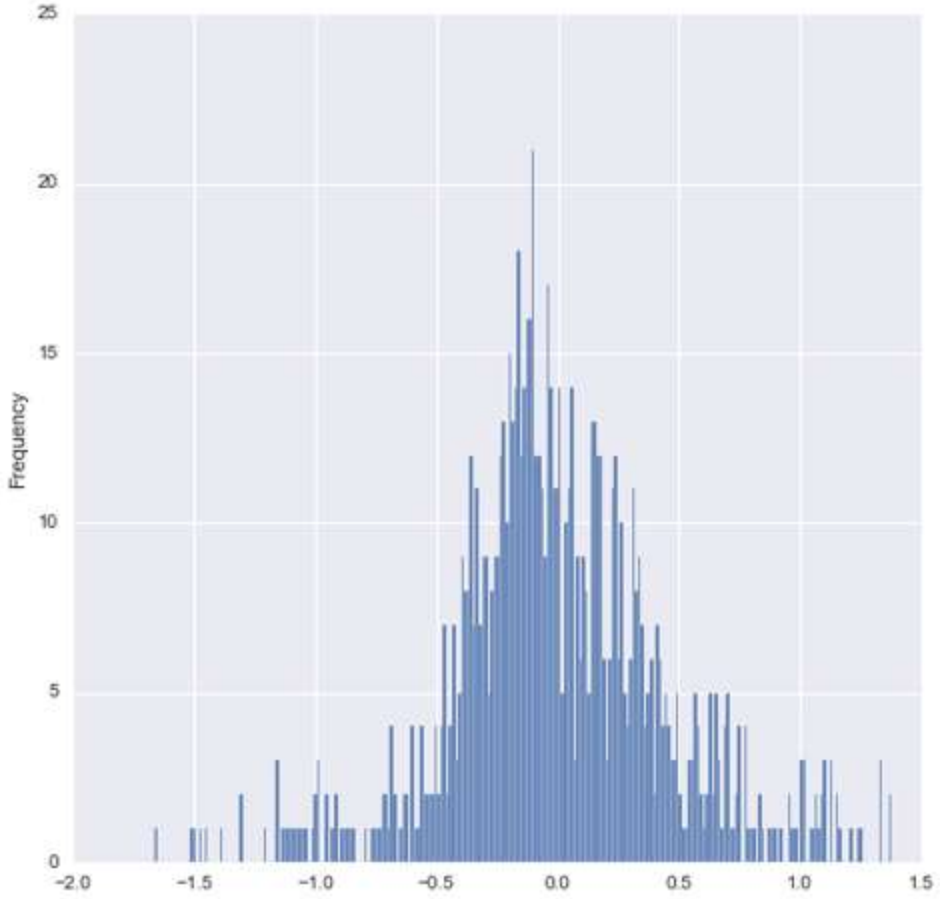
Simple Linear Regression

Common Regression Assumptions

Common Regression Assumptions (Part 1)

- The model is linear
  - x significantly explains y
- $\varepsilon \sim N(0, \ \cdot)$
  - Specifically, we expect ε to be 0 on average, i.e., $\mu_\varepsilon = 0$
- x and ε are independent
  - $\rho(x, \varepsilon) = 0$

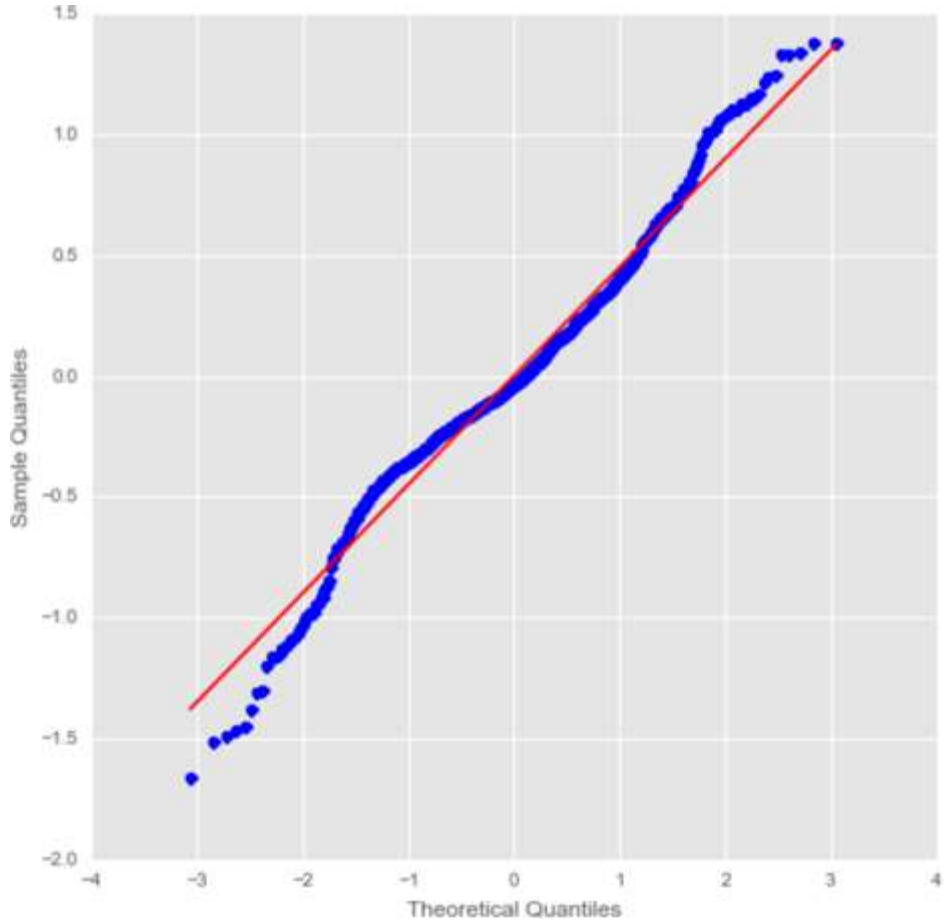
Is $\varepsilon \sim N(0, \cdot)$?



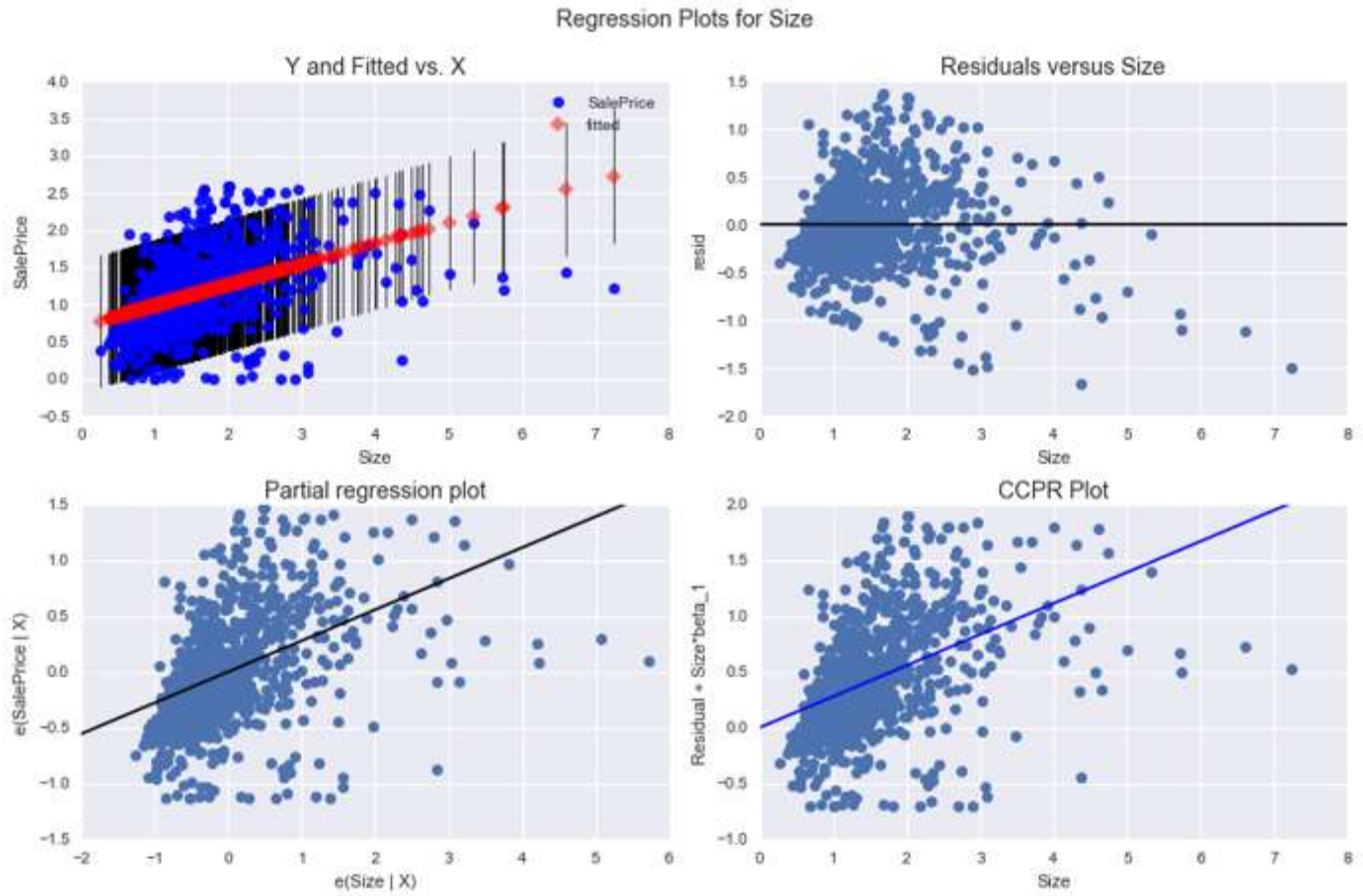
$\varepsilon \sim N(0, \cdot)$: `.qqplot()`

- "Quantile-Quantile (q-q) Plot"
- Graphical technique for determining if two datasets come from populations with a common distribution
- Plot of the quantiles of the first dataset (vertically) against the quantiles of the second's (horizontally)
- If unspecified, the second dataset will default to $N(0, 1)$
- If the two datasets come from a population with the same distribution, the points should fall approximately along a 45-degree reference line
- The greater the departure from this reference line, the greater the evidence for the conclusion that the datasets have come from populations with different distributions

$\varepsilon \sim N(0, \cdot)$: `.qqplot()` (with `line = 's'`) (cont.)



x and ε are independent: `.plot_regress_exog()`



x and ε are independent: `.plot_regress_exog()` (cont.)

- Scatterplot of observed values (y) compared to fitted values ($\hat{y}$) with confidence intervals against the regressor (x)
  - `.plot_fit()`

- "Residual Plot"
  - Scatterplot of the model's residuals ($\hat{\varepsilon}$) against the regressor (x)

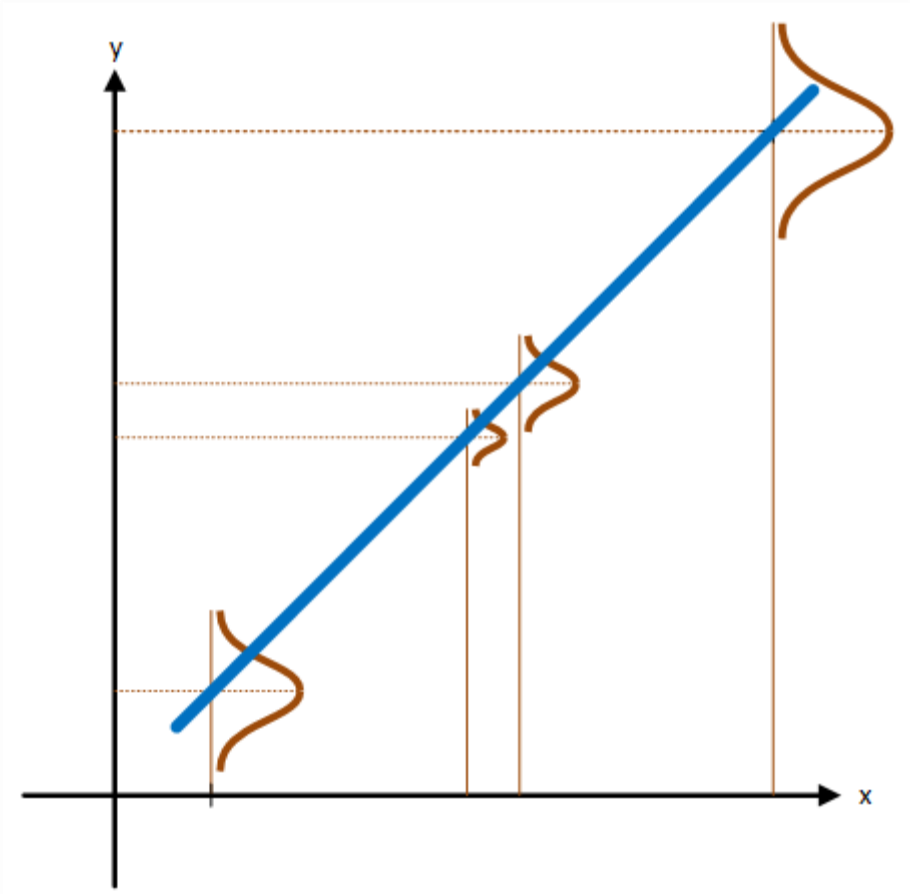
- "Partial Regression Plot" and "CCPR Plot (Component and Component-Plus-Residual)"
  - (useful for multiple regression)

Simple Linear Regression

Assessing the model's fit with R^2

Fit and Inference

- The deviations of the data from the best fitting line are normally distributed about the line. Since $\mu_\varepsilon = 0$, we "expect" that on average, the line will be correct
- How confident we are about how well the relationship holds depends on σ_ε^2



Assessing the model's fit with R^2

- When a measure of how much of the total variation in y, $\sigma_y^2 = \beta^2\sigma_x^2 + \sigma_\varepsilon^2$, is explained by the portion associated with the explanatory variable x, $\sigma_{\hat{y}}^2 = \beta^2\sigma_x^2$; also called systematic variation (the variation explained by your model)

$$R^2 = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2} = \frac{\beta^2\sigma_x^2}{\beta^2\sigma_x^2 + \sigma_\varepsilon^2}$$

- $0 \leq R^2 \leq 1$ (since $-1 \leq \rho_{xy} \leq 1$)

- $1 - R^2 = \frac{\sigma_\varepsilon^2}{\beta^2\sigma_x^2+\sigma_\varepsilon^2}$ is the idiosyncratic variation

  (the variation left unexplained by your model)

Assessing the model's fit with R^2 (cont.)

When x significantly explains y	**When x does not significantly explains y**
❑ The fit is **better**	❑ The fit is **worse**
❑ The **explained** systematic variation dominates	❑ The **unexplained** idiosyncratic variation dominates
❑ σ_ε^2 is low (and/or $\beta^2\sigma_x^2$ is high)	❑ σ_ε^2 is high (and/or $\beta^2\sigma_x^2$ is low)
❑ $\boldsymbol{R^2 = \frac{1}{1+\underbrace{\frac{\sigma_\varepsilon^2}{\beta^2\sigma_x^2}}_{\cong 0}}}$ **is closer to 1**	❑ $\boldsymbol{R^2 = \frac{1}{1+\underbrace{\frac{\sigma_\varepsilon^2}{\beta^2\sigma_x^2}}_{\gg 1}}}$ **is closer to 0**