
Linear Model Evaluation

— Boston University CS 506 - Lance Galletti —

Evaluating Our Regression Model

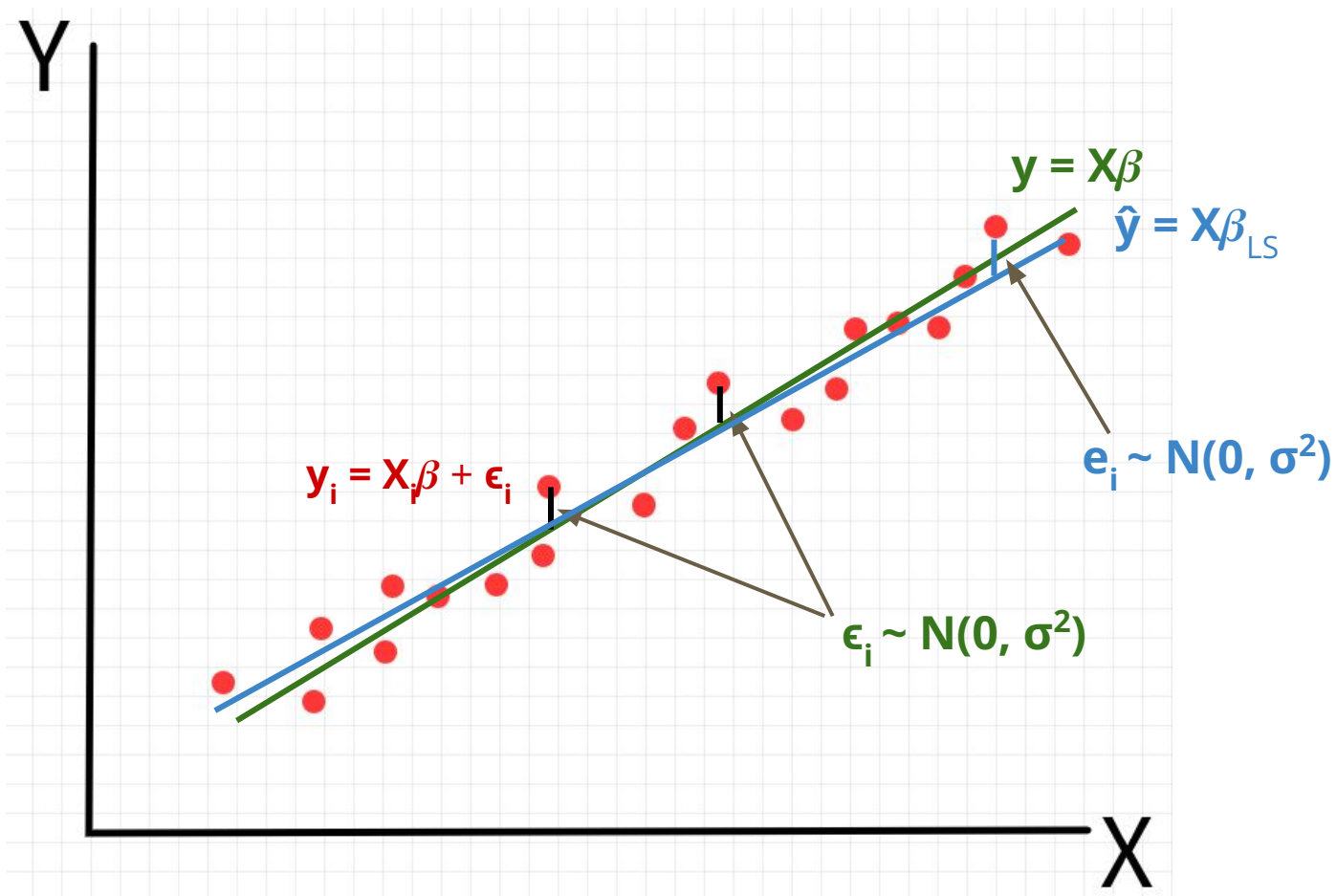
Some Notation:

\mathbf{y}_i is the “true” value from our data set (i.e. $\mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$)

$\hat{\mathbf{y}}_i$ is the estimate of y_i from our model (i.e. $\mathbf{x}_i\boldsymbol{\beta}_{LS}$)

$\bar{\mathbf{y}}$ is the sample mean all \mathbf{y}_i

$\mathbf{y}_i - \hat{\mathbf{y}}_i$ are the estimates of $\boldsymbol{\epsilon}_i$ and are referred to as residuals



Metric for evaluation the fit of our model?

Is the value of the loss function sufficient? i.e. no
=> different model, different number
of data point, different scale of data

$$\|y - X\beta\|_2^2 = \sum_i^n (y_i - \hat{y}_i)^2$$

Evaluating Our Regression Model

$$TSS = \sum_i^n (y_i - \bar{y})^2$$

← This is a measure of the spread of y_i around the mean of y

Evaluating Our Regression Model

$$TSS = \sum_i^n (y_i - \bar{y})^2$$

← This is a measure of the spread of y_i around the mean of y

$$ESS = \sum_i^n (\hat{y}_i - \bar{y})^2$$

← This is a measure of the spread of our model's estimates of y_i around the mean of y

Evaluating Our Regression Model

$$TSS = \sum_i^n (y_i - \bar{y})^2$$

$$ESS = \sum_i^n (\hat{y}_i - \bar{y})^2$$

$$R^2 = \frac{ESS}{TSS}$$

$0 < R^2 < 1$, $R^2 == 1$ is good
when $y_i == \hat{y}_i$, $R^2 == 1$
 R^2 measures the fraction of variance that is
explained by \hat{y} (our model)

<https://medium.com/@saishruthi.tn/is-r-sqaure-value-always-between-0-to-1-36a8d17807d1>,
see why R^2 is between 0 and 1

Evaluating Our Regression Model

$$TSS = \sum_i^n (y_i - \bar{y})^2 \qquad R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_i^n (y_i - \hat{y}_i)^2$$

← This is what our linear model is minimizing

$$ESS = \sum_i^n (\hat{y}_i - \bar{y})^2$$

Exercise

Show that $TSS = ESS + RSS$

$$\begin{aligned} TSS &= \sum_i (y_i - \bar{y})^2 \\ &= \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2 \sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\ &= ESS + RSS + 2 \sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}). \end{aligned}$$

$$\begin{aligned} \sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_i (y_i - \hat{y}_i)\hat{y}_i - \bar{y} \sum_i (y_i - \hat{y}_i) \\ &= \hat{\beta}_0 \sum_i (y_i - \hat{y}_i) + \hat{\beta}_1 \sum_i (y_i - \hat{y}_i)x_i - \bar{y} \sum_i (y_i - \hat{y}_i) \end{aligned}$$

Assume for simplicity that $\hat{y}_i = \beta_0 + \beta_1 x_i$. Since β_0 and β_1 are least squares estimates, we know they minimize

$$\sum_i (y_i - \hat{y}_i)^2$$

By taking derivatives of the above with respect to β_0 and β_1 we discover that

$$\sum_i (y_i - \hat{y}_i) = 0 \text{ and } \sum_i (y_i - \hat{y}_i)x_i = 0$$

Evaluating our Regression Model

OLS Regression Results						
=====						
Dep. Variable:	y	R-squared:	0.840			
Model:	OLS	Adj. R-squared:	0.836			
Method:	Least Squares	F-statistic:	254.1			
Date:	Sun, 20 Mar 2022	Prob (F-statistic):	2.72e-39			
Time:	11:36:16	Log-Likelihood:	-482.37			
No. Observations:	100	AIC:	970.7			
Df Residuals:	97	BIC:	978.5			
Df Model:	2					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	2.1912	3.162	0.693	0.490	-4.085	8.467
x1	29.3912	3.274	8.977	0.000	22.893	35.889
x2	78.1391	3.594	21.741	0.000	71.006	85.272
=====						
Omnibus:	1.279	Durbin-Watson:	1.824			
Prob(Omnibus):	0.527	Jarque-Bera (JB):	1.065			
Skew:	0.253	Prob(JB):	0.587			
Kurtosis:	2.999	Cond. No.	1.38			
=====						

Evaluating our Regression Model

OLS Regression Results						
Dep. Variable:	y	R-squared:	0.840			
Model:	OLS	Adj. R-squared:	0.836			
Method:	Least Squares	F-statistic:	254.1			
Date:	Sun, 20 Mar 2022	Prob (F-statistic):	2.72e-39			
Time:	11:36:16	Log-Likelihood:	-482.37			
No. Observations:	100	AIC:	970.7			
Df Residuals:	97	BIC:	978.5			
Df Model:	2					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	2.1912	3.162	0.693	0.490	-4.085	8.467
x1	29.3912	3.274	8.977	0.000	22.893	35.889
x2	78.1391	3.594	21.741	0.000	71.006	85.272
Omnibus:	1.279		Durbin-Watson:		1.824	
Prob(Omnibus):	0.527		Jarque-Bera (JB):		1.065	
Skew:	0.253		Prob(JB):		0.587	
Kurtosis:	2.999		Cond. No.		1.38	

Evaluating our Regression Model

```

=====
OLS Regression Results
=====
Dep. Variable:          y      R-squared:                0.840
Model:                  OLS    Adj. R-squared:           0.836
Method:                 Least Squares    F-statistic:             254.1
Date:                  Sun, 20 Mar 2022    Prob (F-statistic):      2.72e-39
Time:                  11:36:16    Log-Likelihood:          -482.37
No. Observations:      100    AIC:                     970.7
Df Residuals:          97    BIC:                     978.5
Df Model:              2
Covariance Type:       nonrobust
=====

```

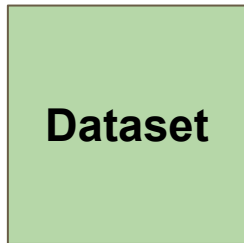
	coef	std err	t	P> t	[0.025	0.975]
const	2.1912	3.162	0.693	0.490	-4.085	8.467
x1	29.3912	3.274	8.977	0.000	22.893	35.889
x2	78.1391	3.594	21.741	0.000	71.006	85.272

```

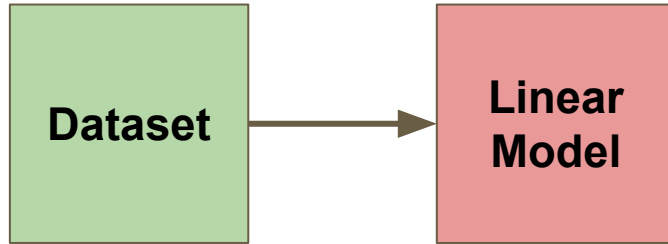
=====
Omnibus:              1.279    Durbin-Watson:           1.824
Prob(Omnibus):        0.527    Jarque-Bera (JB):        1.065
Skew:                 0.253    Prob(JB):                0.587
Kurtosis:             2.999    Cond. No.:               1.38
=====

```

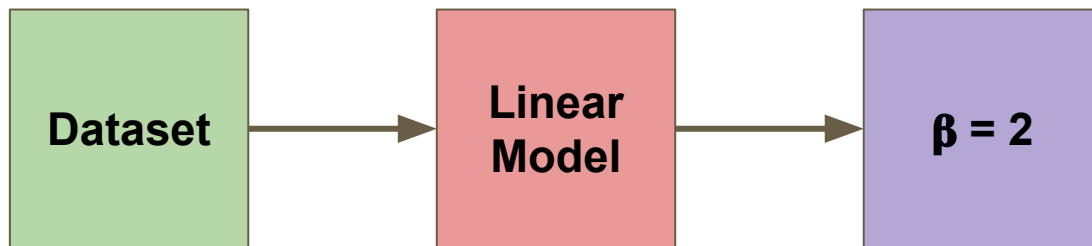
Hypothesis Testing



Hypothesis Testing

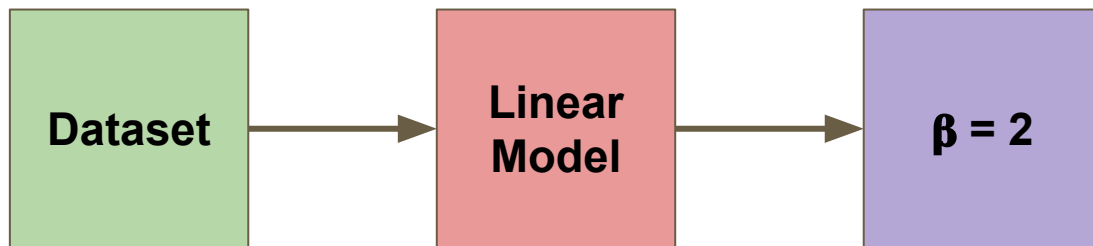


Hypothesis Testing



under the hypothesis, how is the probability to see this outcome

Hypothesis Testing



Could the
real beta
be 5?

HHHHHHHH



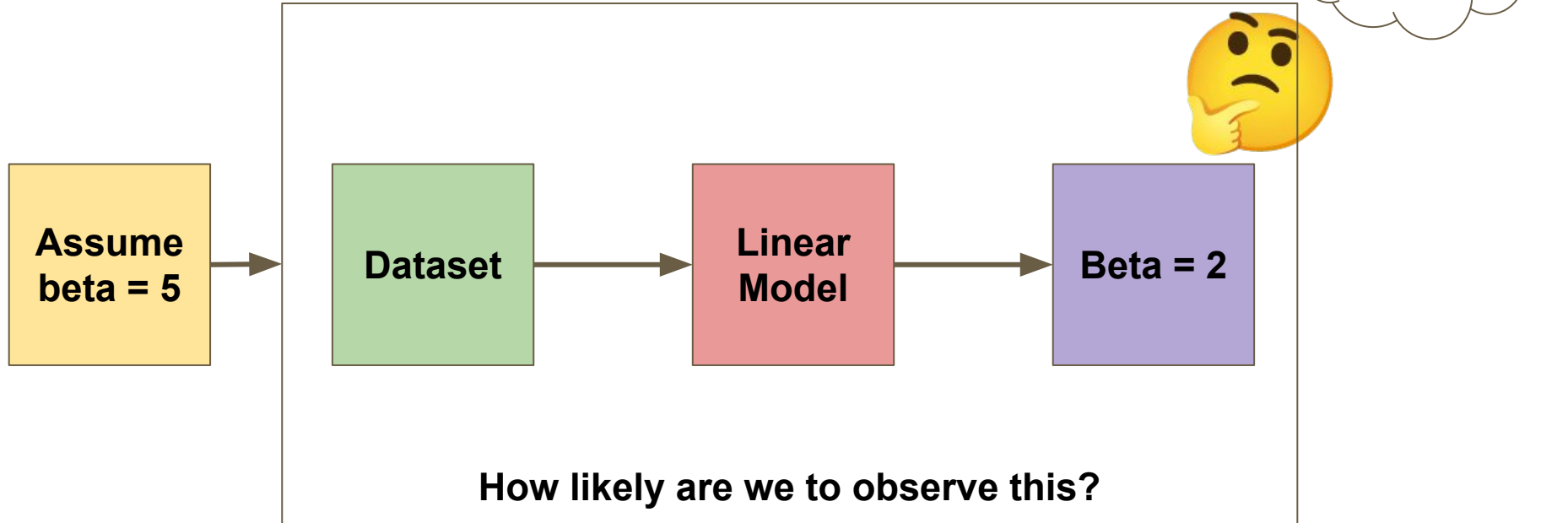
The coin is
probably
not fair

HTHTHTTT

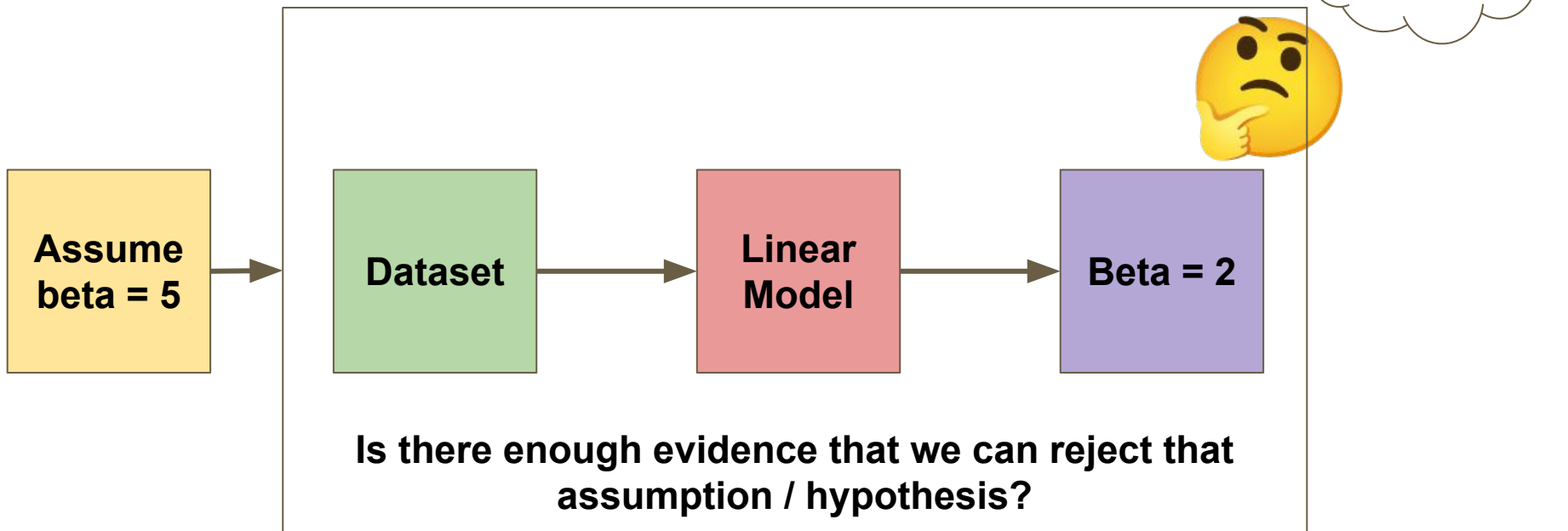


The coin
could be
fair

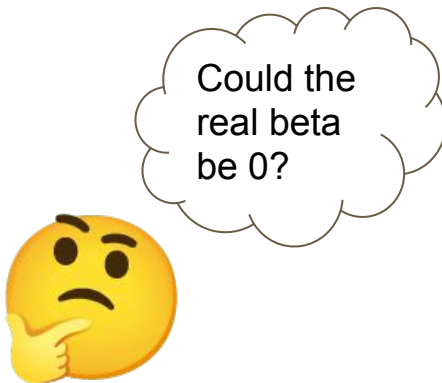
Hypothesis Testing



Hypothesis Testing



worksheet



Hypothesis Testing

Each parameter of an independent variable \mathbf{x} has an associated confidence interval and t-value + p-value.

If the parameter / coefficient is not significantly distinguishable from 0 then we cannot assume that there is a significant linear relationship between that independent variable and the observations \mathbf{y} (i.e. if the interval includes 0 or if the p-value is too large)

Hypothesis Test

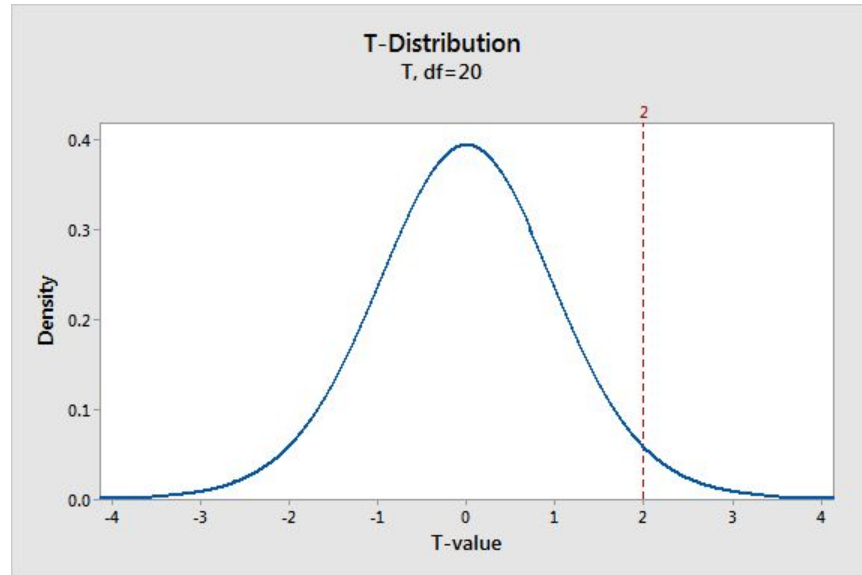
We want to know if there is evidence to reject the hypothesis $H_0 : \beta = 0$ (i.e. that there is no linear relation between X and Y) using the information from $\hat{\beta}$.

We want to know the largest probability of obtaining the data observed, under the assumption that the null hypothesis is correct.

How do we obtain that probability?

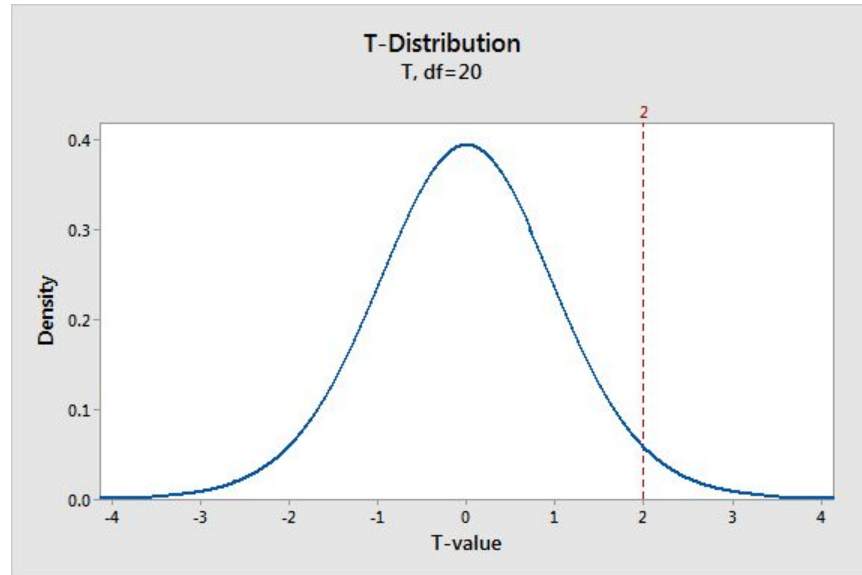
Hypothesis Test

Under the null hypothesis what should be the distribution of the normalized estimates? T-distribution (parametrized by the sample size)



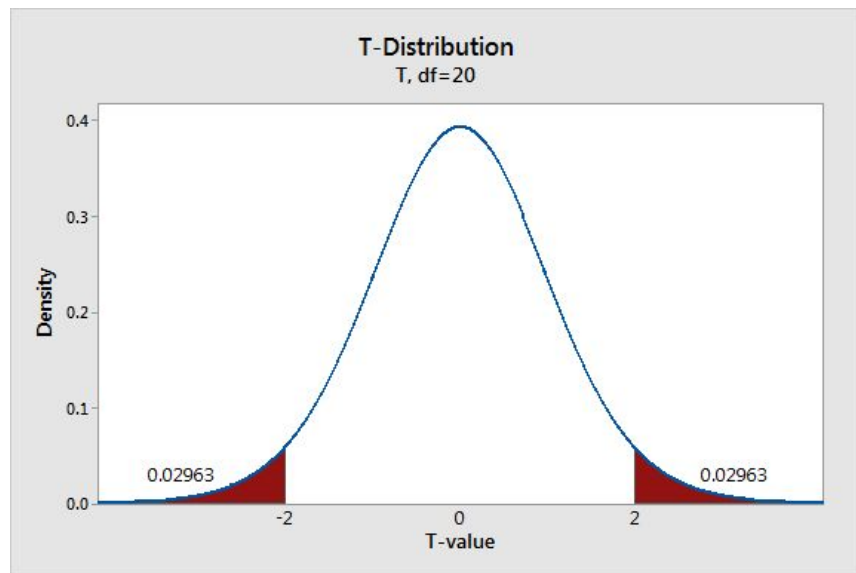
Hypothesis Test

We can then compute the t-value that corresponds to the sample we observed.



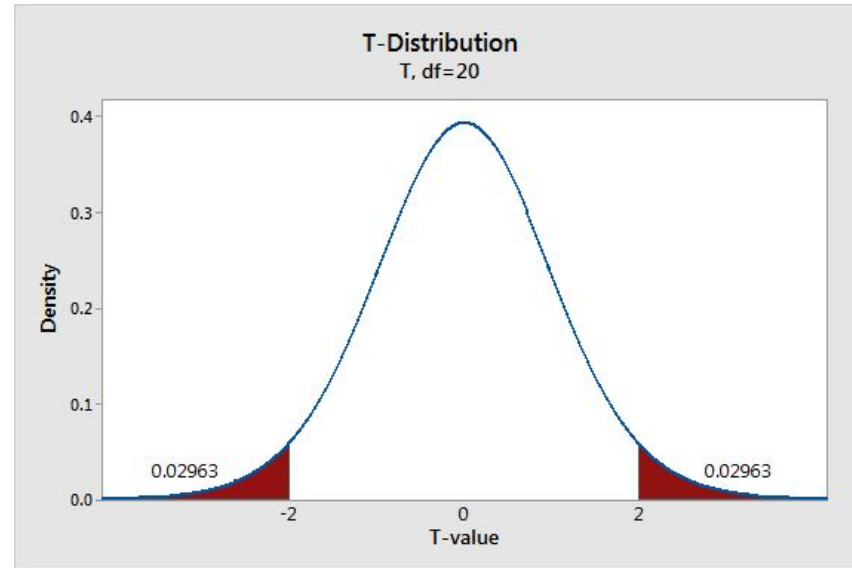
Hypothesis Test

And then compute the probability of observing estimates of β at least as extreme as the one observed. (i.e. trying to find evidence against H_0)



Hypothesis Test

This probability is called a p-value.



Hypothesis Test

A p-value **smaller than a given threshold** would mean the data was unlikely to be observed under H_0 so we can reject the hypothesis H_0 . If not, then we lack the evidence to reject H_0 .

	coef	std err	t	P> t	[0.025	0.975]
const	2.1912	3.162	0.693	0.490	-4.085	8.467
x1	29.3912	3.274	8.977	0.000	22.893	35.889
x2	78.1391	3.594	21.741	0.000	71.006	85.272

Hypothesis Test

Which parameters should we not include in our linear model?

	coef	std err	t	P> t	[0.025	0.975]
const	2.1912	3.162	0.693	0.490	-4.085	8.467
x1	29.3912	3.274	8.977	0.000	22.893	35.889
x2	78.1391	3.594	21.741	0.000	71.006	85.272

Evaluating our Regression Model

OLS Regression Results						
Dep. Variable:	y	R-squared:	0.840			
Model:	OLS	Adj. R-squared:	0.836			
Method:	Least Squares	F-statistic:	254.1			
Date:	Sun, 20 Mar 2022	Prob (F-statistic):	2.72e-39			
Time:	11:36:16	Log-Likelihood:	-482.37			
No. Observations:	100	AIC:	970.7			
Df Residuals:	97	BIC:	978.5			
Df Model:	2					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	2.1912	3.162	0.693	0.490	-4.085	8.467
x1	29.3912	3.274	8.977	0.000	22.893	35.889
x2	78.1391	3.594	21.741	0.000	71.006	85.272
Omnibus:	1.279	Durbin-Watson:	1.824			
Prob(Omnibus):	0.527	Jarque-Bera (JB):	1.065			
Skew:	0.253	Prob(JB):	0.587			
Kurtosis:	2.999	Cond. No.	1.38			

Confidence Intervals

An interval that describes the uncertainty around an estimate (here this could be $\hat{\beta}$).

Goal: for a given confidence level (let's say 90%), construct an interval around an estimate such that, if the estimation process were repeated indefinitely, the interval would contain the true value (that the estimate is estimating) 90% of the time.

	coef	std err	t	P> t	[0.025	0.975]
const	2.1912	3.162	0.693	0.490	-4.085	8.467
x1	29.3912	3.274	8.977	0.000	22.893	35.889
x2	78.1391	3.594	21.741	0.000	71.006	85.272

Z-values

These are the number of standard deviations from the mean of a $N(0,1)$ distribution required in order to contain a specific % of values were you to sample a large number of times.

To find the .95 z-value (the value z such that 95% of the observations lie within z standard deviations of the mean ($\mu \pm z * \sigma$)) you need to solve:

$$\int_{-z}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = .95$$

Z-values

The .95 z-value is 1.96.

This means 95% of observations from a $N(\mu, \sigma)$ lie within 1.96 standard deviations of the mean ($\mu \pm 1.960 * \sigma$)

If we get a sample from a $N(\mu, \sigma)$ of size n , how would we create a confidence interval around the estimated mean?

Confidence Intervals

How do we build a confidence interval?

Assume $\mathbf{Y}_i \sim \mathbf{N}(5, 25)$, for $1 \leq i \leq 100$ and $\mathbf{y}_i = \mu + \epsilon$ where $\epsilon \sim \mathbf{N}(0, 25)$. Then the Least Squares estimator of μ (μ_{LS}) is

the sample mean \bar{y}

What is the 95% confidence interval for μ_{LS} ?

$$\begin{aligned} CI_{.95} &= [\bar{y} - 1.96 \times SE(\mu_{LS}), \bar{y} + 1.96 \times SE(\mu_{LS})] \\ &= [\bar{y} - 1.96 \times .5, \bar{y} + 1.96 \times .5] \end{aligned}$$

$$\begin{aligned} SE(\mu_{LS}) &= \sigma_{\epsilon} / \sqrt{n} \\ &= 5 / \sqrt{100} \\ &= .5 \end{aligned}$$

Z-value for 95% Confidence Interval

Checking our Assumptions

1. Normal Distribution?
2. Constant Variance?

QQ plot

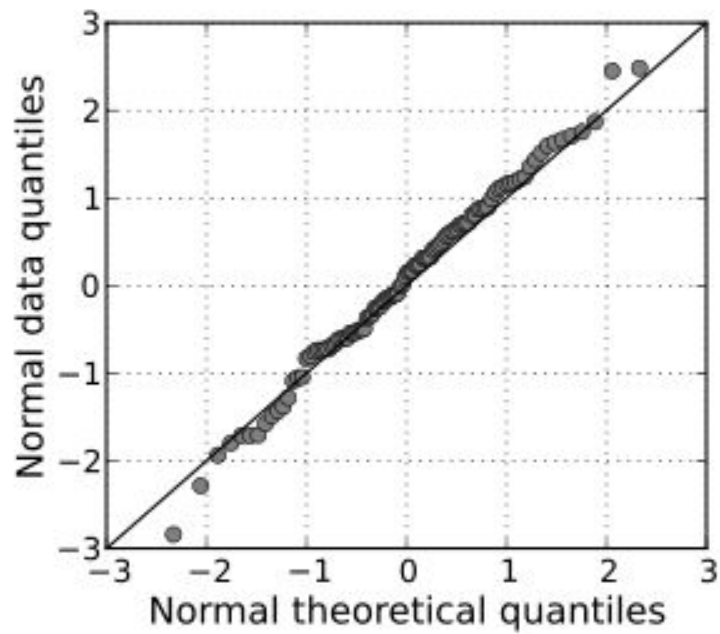
Quantiles are the values for which a particular % of values are contained below it.

For example the 50% quantile of a $N(0,1)$ distribution is 0 since 50% of samples would be contained below 0 were you to sample a large number of times.

QQ plot

For all quantiles q , if $\text{sample}.q == \text{known_distribution}.q$ then they have the same distribution.

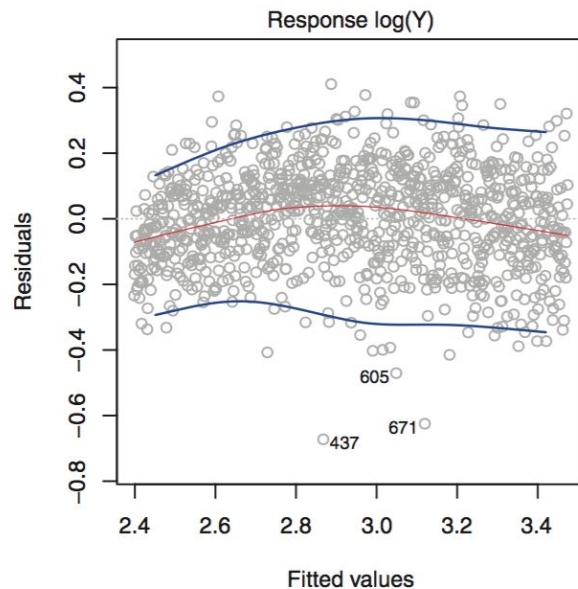
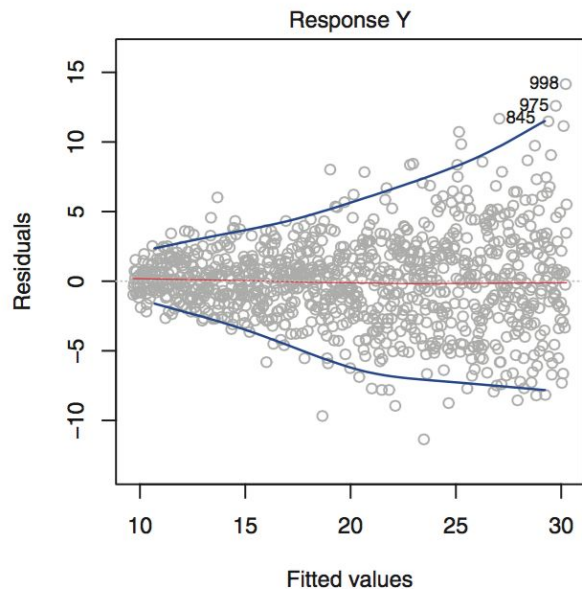
QQ plot



Constant Variance

One of our assumptions was that our noise had constant variance. How can we verify this?

We can plot residuals (noise estimates) for each fitted value \hat{y}_i



Extending our Linear Model

Changing the assumptions we made can drastically change the problem we are solving. A few ways to extend the linear model:

1. Non-constant variance - used in WLS (weighted least squares)
2. Distribution of error is not Normal - used in GLM (generalized linear models)