Linear Regression

We have thus far concentrated on the "surface level" aspects of Linear Regression.

That is, we focused on the equation

$$\hat{\mathbf{y}} = \Theta \cdot \mathbf{x}$$

But we have not considered how to interpret, or analyze the utility, of individual features $\mathbf{x}_j$.

Magnitude of $\mathbf{\Theta}_j$

Does a large value of a coefficient Θ_j mean the associated feature $\mathbf{x}_j$ is "important" ?

No !

Consider the true model

$$p_{\text{data}}(\mathbf{y} \mid \mathbf{x})$$

is

$$\mathbf{y} = \Theta_1 * \mathbf{x}_1$$

What happens if I change the magnitude of $\mathbf{x}$

- e.g., from units of "dollars" to units of "thousands of dollars"

$$\mathbf{x}'_1 = \frac{\mathbf{x}_1}{1000}$$

The relationship becomes

$$
\begin{aligned}
\mathbf{y} &= (\Theta_1 * 1000) * \frac{\mathbf{x}_1}{1000} && \text{coefficient increases to offset decrease in feature} \\
&= \Theta'_1 * \mathbf{x}'_1 && \text{where } \Theta'_1 = \Theta_1 * 1000
\end{aligned}
$$

Re-denominating the feature $\mathbf{x}_1$

- causes the coefficient to increase by a factor of 1000
- but does not change the fundamental underlying relationship
    - a unit change in $\mathbf{x}_1$ (equivalently: a change in $\mathbf{x}'_1$ of $.001$)
    - changes prediction $\mathbf{y}$ by Θ_1

The larger Θ'_1 is no more important than Θ_1.

That is

- the magnitude of a coefficient
- depends on the magnitude of the feature

Don't conflate magnitude with importance.

Mathematically

$$
\begin{aligned}
\Theta_j &= \frac{\partial \mathbf{y}}{\partial \mathbf{x}_j} \\
\Theta_{j'} &= \frac{\partial \mathbf{y}}{\partial \mathbf{x}_{j'}}
\end{aligned}
$$

So a unit change in $\mathbf{x}_j$

- causes a larger change in $\mathbf{y}$
- than a unit change in $\mathbf{x}_{j'}$
- when $\Theta_j > \Theta_{j'}$

But the larger "impact" on $\mathbf{y}$ does not make feature $\mathbf{x}_j$ more "important".

This is critical when we have more than one feature

- $\Theta_j > \Theta_{j'}$ may be an *artifact* of $\mathbf{x}_j$ and $\mathbf{x}_{j'}$ being on different scales
- **not** an indication of greater importance of feature $\mathbf{x}_j$ versus feature $\mathbf{x}_{j'}$

Regularization

In fitting a Linear Regression model

- adding more features won't adversely affect Loss
- but might adversely affect out-of-sample generalization

An irrelevant feature won't increase in-sample Loss.

But it might capture meaningless "noise" in the training data

- that causes out-of-sample prediction to become worse

So we need to trade-off

- the desire to include potentially relevant features
- with the potential adverse impact on generalization

One attempt at managing this trade-off is via *Regularization*.

Recall that we can add a *regularization* term to the Loss Function $\mathcal{L}$ for Linear Regression

- a *penalty* term
- that forces parameters coefficients toward 0

For example uses the penalty, the standard MSE loss

$$\mathcal{L}_{\text{MSE}} = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{y^{(i)}} - \Theta \cdot \mathbf{x})^2$$

can be augmented with a penalty

$$Q = \sum_{n=1}^{N} \Theta_n^2$$

to give the loss

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda Q$$

Regularization is an attempt

- to not prematurely exclude a potentially important feature
- but to omit it "after the fact" by forcing its coefficient to zero

Hyper-parameter λ expresses a trade-off between

- reducing the magnitude of Θ
- and the resulting increase in $\mathcal{L}_{\text{MSE}}$

Beware !

The coefficients that are made smaller by regularization

- do not necessarily correspond to "unimportant" features

As we have observed above, special attention should be paid

- to the *scale* of each feature
- when Regularization will be applied
- as the scale of the corresponding parameter moves inversely to the scale of the feature

Statistical significance of Θ_j

Consider the true model

$$p_{\text{data}}(\mathbf{y} \mid \mathbf{x})$$

is

$$\mathbf{y} = \Theta * \mathbf{x}_1$$

In general

- we **don't know** the true model
- we only have access to the training dataset $\langle \mathbf{X}, \mathbf{y} \rangle$
- which is a *sample* from the true model joint distribution of $\mathbf{y}$ and $\mathbf{x}$
- and we hypothesize (and test) theories for what the true model is

The Θ^* obtained from fitting a model to the training dataset

- depends on the particular sample $\langle \mathbf{X}, \mathbf{y} \rangle$ we observe in the training dataset
- a different sample would lead to a different Θ^*

By drawing many possible samples of $\langle \mathbf{X}, \mathbf{y} \rangle$ from the true $p_{\text{data}}(\mathbf{y} \mid \mathbf{x})$

- we can estimate a *distribution* of the values for Θ^* we obtain by fitting

That is

- our fitting is an *estimate* of the true Θ
- that depends on the distribution of Θ^*

Let σ_j denote the first moment of the distribution of Θ^*_j

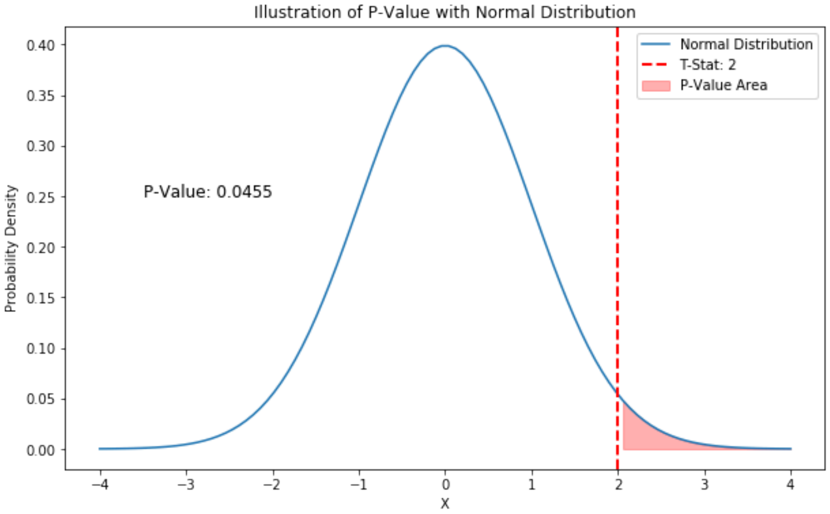
If we know

- the distributional form of Θ_j^*
  - typically: Student-t
- and the distribution's mean
- and the first moment σ_j
- we can calculate how likely it is
  - to draw the measured $\Theta_j^* \neq 0$ from the distribution

Here is a picture

In [4]:
```
fig_pval
```

Out[4]:



The process is as follows.

We start off with the hypothesis (that we hope to contradict)

- that the true mean of $\Theta_j^* = 0$.

We then calculate how far our measured $\Theta_j^* \neq 0$ is from 0.

- the farther away it is, the less likely it is that we will draw $\Theta_j^* \neq 0$ from the zero-mean distribution

If it is very unlikely (e.g., with probability p a small number)

- then we **reject** our hypothesis that the true mean of $\Theta_j^* = 0$
- **accept** our measured $\Theta_j^* \neq 0$ as being *significantly different than 0*
  - hence: there **is** a true relationship between $\mathbf{y}$ and $\mathbf{x}_j$
- and **we will be wrong in doing so** with probability no greater than p

This gives us a basis for deciding

- whether to accept
- that there is a true non-zero relationship between target $\mathbf{y}$ and feature $\mathbf{x}_j$

We should include features $\mathbf{x}_j$

- when the probability of being wrong in accepting the relationship to $\mathbf{y}$
- is small

The *t-stat* and *p-value* are computed values that express different ways

- of allowing us to accept that our measured $\Theta^*_j \neq 0$
- is significantly different than 0
- and hence, we should accept that $\mathbf{y}$ is truly related to $\mathbf{x}_j$

Given a particular estimate Θ^*_j from our fitting

- we measure its distance from 0, called the *t-stat*
- expressed in units of "number of first moments"

$$t_j = \frac{\Theta^*_j}{\sigma_j}$$

The *p-value* is the probability of drawing the measured $\Theta^*_j \neq 0$ from a zero-mean distribution.

So, it is *possible* to draw $\Theta^*_j \neq 0$ from a zero-mean distribution

- but only with probability p

By rejecting the hypothesis that $\Theta^*_j = 0$

- and accepting a relationship between $\mathbf{y}$ and $\mathbf{x}_j$
- we will be *wrong* with probability p

Thus

- the *t-stat*
- and *p-value*

are complementary ways of allowing us to accept that $\mathbf{y}$ and $\mathbf{x}_j$ are truly related.

In the diagram above, you can observe (for a Normal distribution) the

- t-stat
- p-value

In general, it might be best

- to exclude $\mathbf{x}_j$ from the model
- if the probability of it **not** being significantly different than zero is too large.
    - low t-stat
    - large p-value

In [5]:

```
print("Done")
```

```
Done
```