Prepare data: transformations

Transforming data (Recipe C.3) may be **the most important** step of the multi-step Recipe !

Recipe for Machine Learning

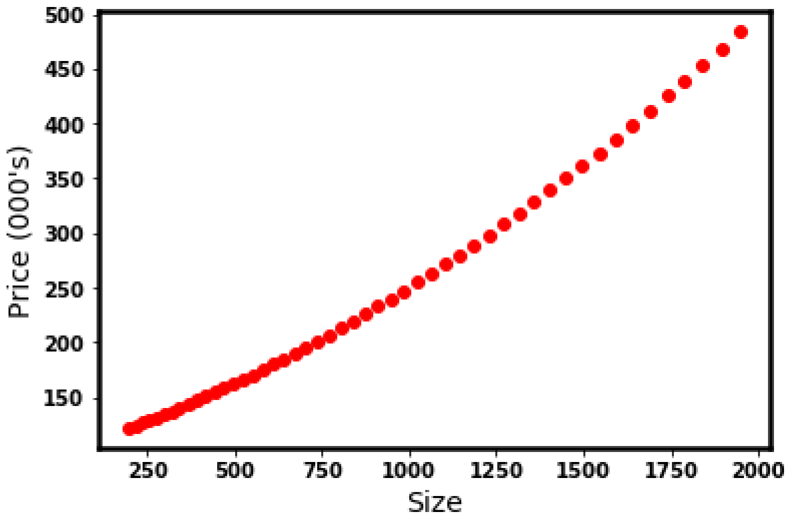
It is often the case that the "raw" features given to us don't suffice

- we may need to create "synthetic" features.
- This is called **feature engineering**.

Recall: our "curvy" data set from the previous lecture:

In [5]:

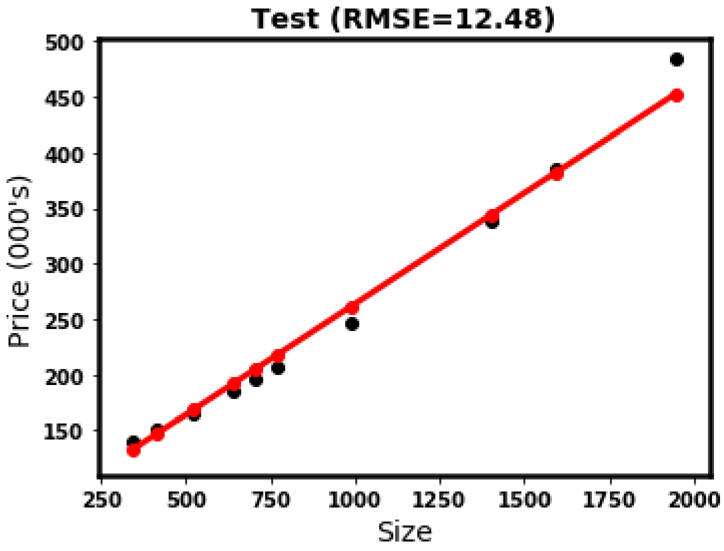
```
(xlabel, ylabel) = ("Size", "Price (000's)")
v1, a1 = 1, .005
v2, a2 = v1, a1*2
curv = recipe_helper.Recipe_Helper(v = v2, a = a2)
X_curve, y_curve = curv.gen_data(num=50)
_= curv.gen_plot(X_curve,y_curve, xlabel, ylabel)
```

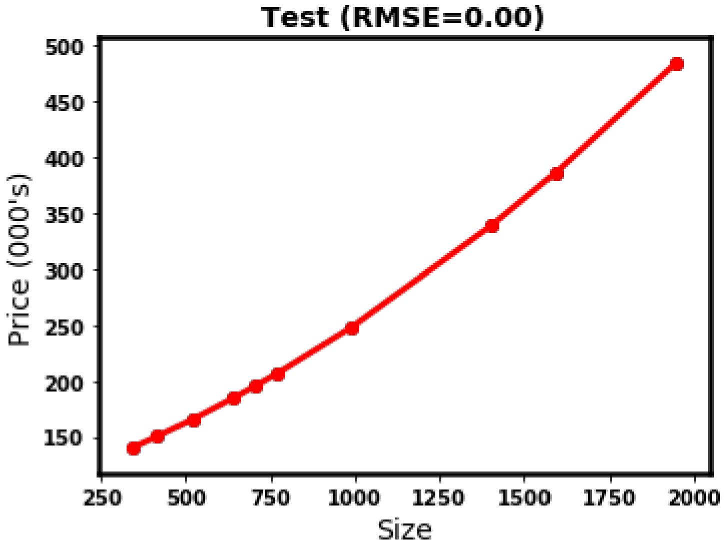


And compare the out of sample performance on this data set

- On a linear model (single, raw feature)
- On a model with a second feature (squared version of raw feature)

In [6]:
```
model_results = curv.compare_regress(X_curve, y_curve, xlabel=xlabel, ylabel=yla
bel, visible=True, plot_train=False)
```



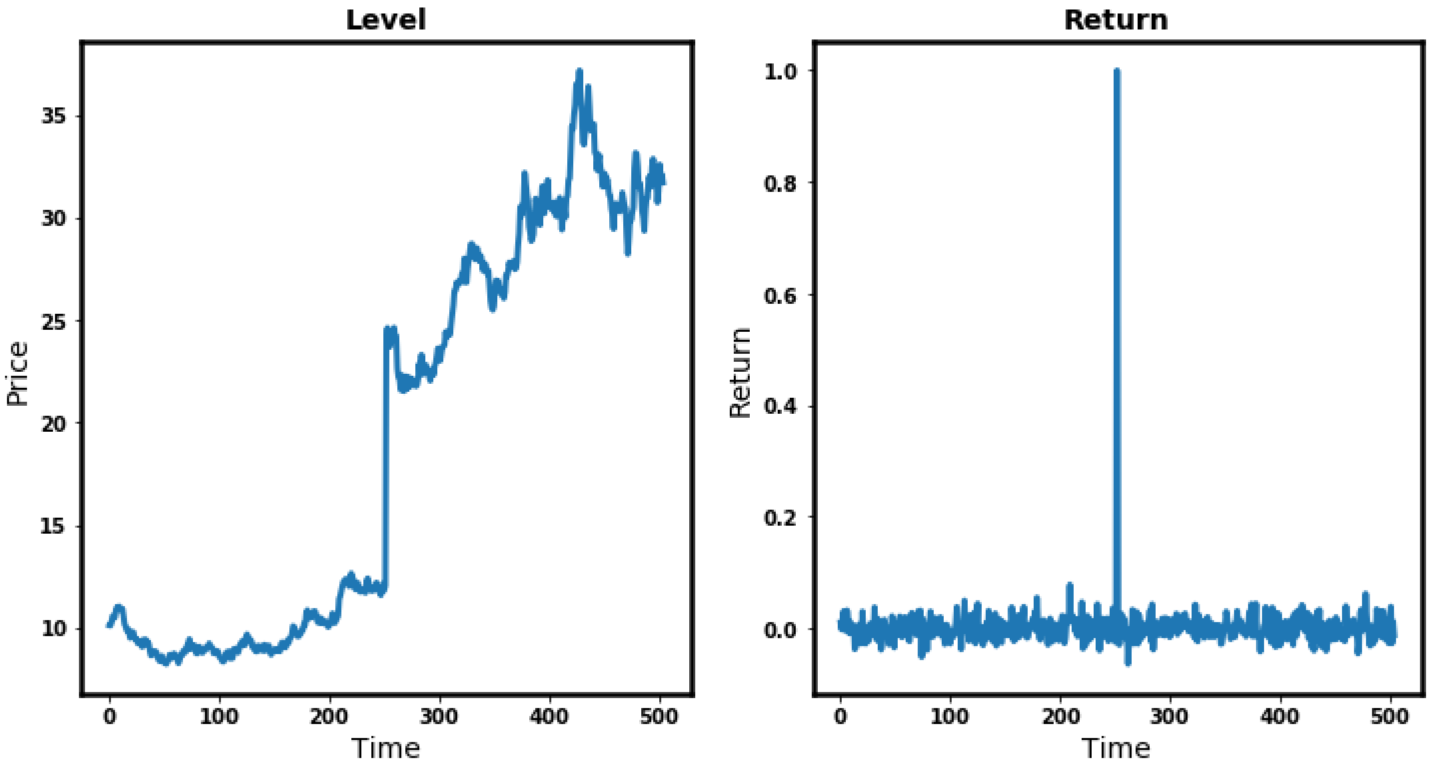


Adding the synthetic feature was key to better performance (lower RMSE).

Or recall the stock whose price jumped at one instant

In [7]:
```
fig_data
```

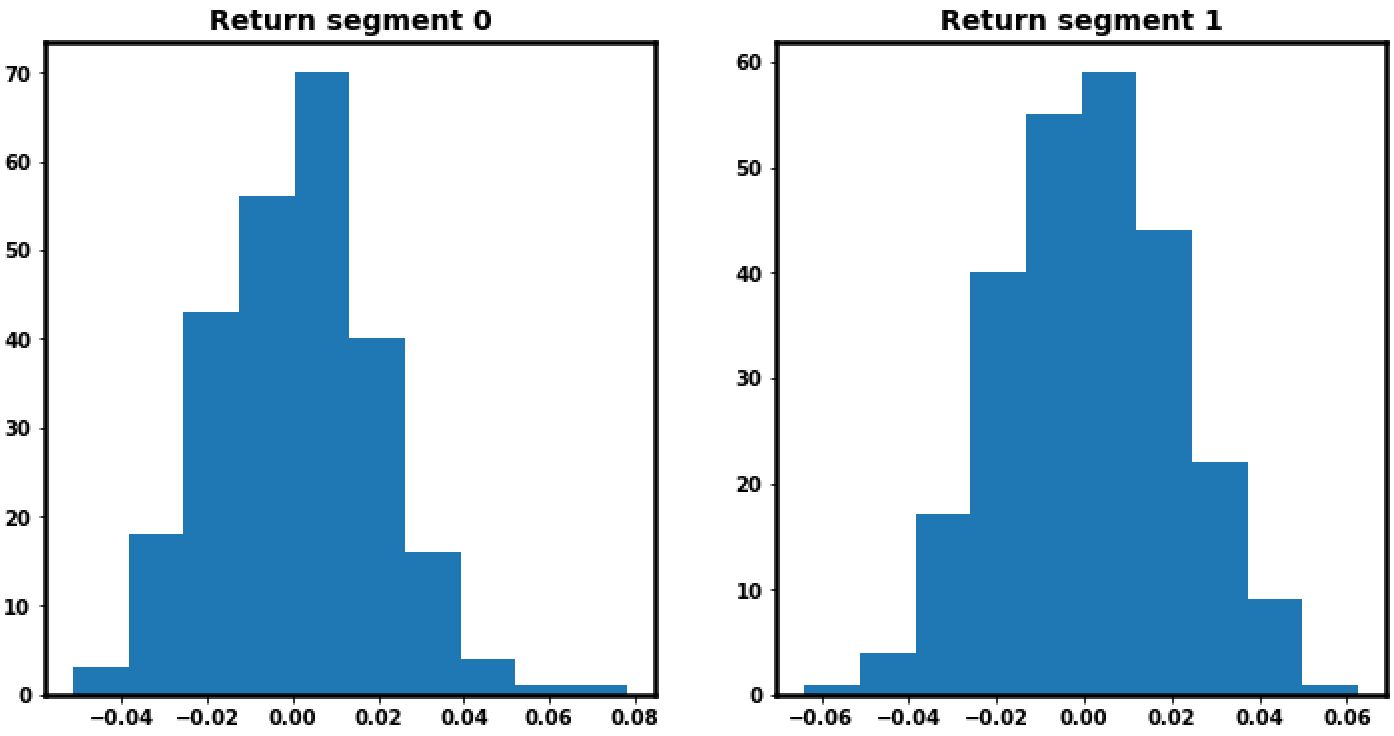
Out[7]:



In [8]:

```
fig_segs
```

Out[8]:



We would probably have better luck

- predicting future *returns* from past returns
- compared to predicting future *prices* from past prices

That is: the synthetic feature ("Return") replaces the raw feature ("Price").

We could also argue that adding an additional synthetic feature might facilitate using Price as a feature:

- a time index
- or indicator (true/false) that identifies examples as being either pre or post jump

It will probably be better using Return rather than Level and a segment indicator

- A jump can occur **within** the training data
    - or each example could drift weakly over time

In order to learn, it helps to have *training* data be more homogeneous

- Can more easily learn a pattern from many examples rather than a handful

Either way: transforming the raw features is key to successful modeling and prediction.

Feature engineering (transformations)

- takes an example: vector $\mathbf{x}^{(\mathbf{i})}$ with n features
- produces a new vector $\tilde{\mathbf{x}}^{(\mathbf{i})}$, with n' features

We ultimately fit the model with the transformed *training* examples.

We can apply multiple transformations, each

- Adding new synthetic features
- Further transforming synthetic features

Feature Engineering

The above diagram shows multiple transformations

- organized as a sequence (sometimes called a *pipeline*) of independent transformations $T_1, T_2, \ldots, T_t$

$$
\begin{aligned}
\tilde{\mathbf{x}}_{(1)} &= T_1(\mathbf{x}) \\
\tilde{\mathbf{x}}_{(2)} &= T_2(\tilde{\mathbf{x}}_{(1)}) \\
&\vdots \\
\tilde{\mathbf{x}}_{(l+1)} &= T_{(l+1)}(\tilde{\mathbf{x}}_{(l)})
\end{aligned}
$$

We write the final transformed $\tilde{\mathbf{x}}$ as a function T that is the composition of each transformation function

$$\tilde{\mathbf{x}} = T(\mathbf{x}) = T_t(\ T_{t-1}(\ldots T_1(\mathbf{x}) \ldots)\ )$$

The length of the final transformed vector $\tilde{\mathbf{x}}$ may differ from the n, the length of the input $\mathbf{x}$

- may add features
- may drop features

The predictions are now a function of $\tilde{\mathbf{x}}$ rather than $\mathbf{x}$

$$\hat{\mathbf{y}} = h_{\Theta}(\tilde{\mathbf{x}})$$

Example transformation: Missing data imputation

The first transformation we encountered added a feature ($\mathbf{x}^2$ term) that improved prediction.

Some transformations alter existing features rather than adding new ones.

Transformations in detail will be the subject of a separate lecture but let's cover the basics.

Let's consider a second reason for transformation: filling in (imputing) missing data for a feature.

$\boxed{\#}$	$\mathbf{x_1}$	$\mathbf{x_2}$
1	1.0	10
2	2.0	20
$\vdots$	$\vdots$	$\vdots$
i	2.0	NaN
$\vdots$	$\vdots$	$\vdots$
m	$\ldots$	

In the above: feature $\mathbf{x}_2$ is missing a value in example i: $\mathbf{x}_2^{(\mathbf{i})} = \mathrm{NaN}$

We will spend more time later discussing the various ways to deal with missing data imputation.

For now: let's adopt the common strategy of replacing it with the median of the defined values:

$$\mathrm{median}(\mathbf{x}_2) = \mathrm{median}(\{\mathbf{x}_2^{(\mathbf{i})} | 1 \le i \le m, \mathbf{x}_2^{(\mathbf{i})} \ne \mathrm{NaN}\})$$

This imputation is a kind of data transformation: replacing an undefined value.

Without this transformation: the algorithm that implements our model

- May fail
- May impute a less desirable value, since it lacks specific knowledge of our problem

"Fitting" transformations

The behavior of our models for prediction have parameters Θ.

It might not be obvious that transformations have parameters $\Theta_{\text{transform}}$ as well

$$\tilde{\mathbf{x}} = T_{\Theta_{\text{transform}}}(\mathbf{x})$$

For example: when missing data imputation for a feature substitutes the mean/median feature value

- $\Theta_{\text{transform}}$ stores this value

We use the term "fitting" to describe the process of solving for $\Theta_{\text{transform}}$

- Unlike Θ, one doesn't usually find a "optimal" value for $\Theta_{\text{transform}}$

Our prediction is thus

$$\begin{aligned}\hat{\mathbf{y}} &= h_{\Theta}(\tilde{\mathbf{x}}) \\ &= h_{\Theta}(\, T_{\Theta_{\text{transform}}}(\mathbf{x})\,)\end{aligned}$$

The process of Transformations is similar to fitting a model and predicting.

The parameters in $\Theta_\text{transform}$

- are "fit" by examining all training data $\mathbf{X}$
- once fit, we can transform ("predict") *any* example (whether it be training/validation or test)

Applying transformations consistently

Since the prediction is now

$$\hat{\mathbf{y}} \quad = \quad h_\Theta(\,\tilde{\mathbf{x}}\,) \quad \text{where } \tilde{\mathbf{x}} = T_{\Theta_\text{transform}}(\mathbf{x})$$

each and every input $\mathbf{x}$ must be transformed

- Training examples
- Test examples

That is: the transformation is applied consistently across all examples, regardless of their source

If we didn't apply the same transformation to both training and test examples

- We would violate the Fundamental Assumption of Machine Learning

However

- $\Theta_{\text{transform}}$ is fit **only** to training examples
- It is **not** recalculated on a set of test examples

Here's the picture

Feature engineering: fit, then transform

There are several reasons not to re-fit on test examples

- It would be a kind of "cheating" to see all test examples (required to fit)
- You should assume that you only encounter one test example at a time, not as a group

Pipelines in `sklearn`

We will see a real use case for Pipelines in a subsequent lecture.

For now, we only give a preview to illustrate the highlights.

Transformations in `sklearn` respond to the methods `fit` and `transform`

`sklearn` provides a `Pipeline` object

- a container for a list of objects that respond to `fit` and `transform` (e.g., Transformations)
- applying `fit` (resp., `transform`) to a `Pipeline` object will apply the method to each element of the list, in sequence

So the `Pipeline` object in `sklearn` is a convenient way of bundling multiple transformations.

This will make it easier to apply the entire set of transformations consistently (to in-sample and out of sample examples)

You may also recall that models in `sklearn` also respond to methods `fit` and `transform`.

We will see that you can also place a model object in a `Pipeline` (usually as the last element of the list).

One benefit of doing so is that the entire process of (transformations + modeling) is neatly wrapped into a single object (promoting consistency).

But we will also see that it facilitates the avoidance of the subtle problem of "cheating in cross validation".

Let's explore [Transformation pipelines in `sklearn` (Transformations_Pipelines.ipynb)](Transformations_Pipelines.ipynb)

We will see this in action within the notebook for Classification.

Using pipelines to avoid cheating in cross validation

Although we start off with the best intentions, it is easy to accidentally "cheat"

- When we combine transformations and cross-validation (to measure out of sample performance)
- Is surprisingly common !

k-fold cross-validation:

- Divides the training examples into k "folds"
- A model is fit k times

- Each fit

    - Uses $(k-1)$ folds for training
    - The remaining fold is considered "out of sample" for that fit

- This gives us k Performance Metrics: a distribution of out of sample performance

Cross Validation/Test split

Consider the difference between fitting $\Theta_{\text{transform}}$

- Once, on *all* the training examples, *before* applying cross-validation
- Separately for each of the k fits of Cross-Validation
    - Using the $(k-1)$ folds used for training in this fit

For example, when Fold_k is out of sample

$$\Theta_{\text{transform}} = f([\mathrm{Fold}_1, \mathrm{Fold}_2, \ldots \mathrm{Fold}_{k-1}, \mathrm{Fold}_k])$$

$$\text{versus}$$

$$\Theta_{\text{transform}} = f([\mathrm{Fold}_1, \mathrm{Fold}_2, \ldots \mathrm{Fold}_{k-1})$$

In the first case, we are cheating !

- Fold k is out of sample for this fit
- And should **not** influence $\Theta_{\text{transform}}$

The second case avoids this problem

- With seemingly a lot more work
- Fitting $\Theta_{\text{transform}}$ multiple times

Perhaps the cheating will become more apparent if we look at some code.

Here is code that "cheats" by fitting the transformation to **all** folds

Transform the data X_train = preprocess_pipeline.fit_transform(train_data) # Cross validation scores = cross_val_score(clf, X_train, y_train, cv=10) print("Model: {m:s} avg cross val score= {s:3.2f}\n".format(m=name, s=scores.mean())) # Fit the model using all training data. # cross_val_score does not run the fit on the complete data. # It runs a number of fits, each fit being on the data with one fold held out for validation. _= clf.fit(X_train, y_train)

And here is code that does not cheat: the transformation is fit **only** to the folds that are in-sample during cross validation

Combine the transformation pipeline with a final classification step model_pipeline = Pipeline(steps=[("transform", preprocess_pipeline), ("classify", clf)]) # Cross validation on the combined pipeline scores = cross_val_score(model_pipeline, train_data, y_train, cv=10) print("Model: {m:s} avg cross val score= {s:3.2f}\n".format(m=name, s=scores.mean())) # Fit the model using all training data. # cross_val_score does not run the fit on the complete data. # It runs a number of fits, each fit being on the data with one fold held out for validation. _= model_pipeline.fit(train_data, y_train)

`cross_val_score`

- Divides `train_data` into folds
- For each fold f
    - Splits `train_data` into
        - set of folds F **excluding** f
        - uses f as out of sample
    - Applies the first argument (e.g., `model_pipeline` rather than the model object `clf`) to F
    - Resulting in the `preprocess_pipeline` and model object being applied to all folds **except** f
- The result is that there is one score (Performance metric) computed for each fold (when that fold is out of sample)

In [9]:

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

```
Done
```