Linear Regression

In 1991, Orley Ashenfelter, an economics professor at Princeton University, stunned the wine world with a bold prediction. He predicted that the 1990 vintage of Bordeaux wines would be the "wine of the century," even better than the prized 1961 vintage. Furthermore, he made this prediction without tasting even a drop of the wine, which had been placed in oak barrels just months earlier.

How did Ashenfelter predict the quality of the wine without tasting it? He used data on past vintages to come up with the following formula for predicting wine quality:

$$\begin{aligned} \widehat{\text{wine quality}} = -7.8 &+ 0.62 \cdot (\text{average summer temperature}) \\ &+ 0.0012 \cdot (\text{winter rainfall}) \\ &- 0.0037 \cdot (\text{harvest rainfall}) \\ &+ 0.024 \cdot (\text{age of the wine}) \end{aligned} \tag{1}$$

The variable on the left-hand side of this expression, wine quality, is what we are trying to predict and is called the *target* (or *label*). (The hat symbol over "wine quality" indicates that the values are predicted instead of observed.) The variables on the right-hand side, such as "average summer temperature" and "harvest rainfall," are called *features* and are the inputs used to predict the target. Although Ashenfelter had no way of knowing the quality of the 1990 wines, he did have the values of the features in 1990, so to make a prediction, all he had to do was plug those values into the equation above. In this way, he arrived at the following prediction for the quality of the 1990 Bordeaux, after they had been aged for 31 years (like the 1961 Bordeaux had been at the time):

$$\begin{aligned} -7.8 &+ 0.62 \cdot (18.7) \\ &+ 0.0012 \cdot (468) \\ &- 0.0037 \cdot (80) \\ &+ 0.024 \cdot (31) = 4.8. \end{aligned} \tag{2}$$

For comparison, the quality of the prized 1961 vintage was 4.6.

You can imagine the uproar from wine experts, who had spent years refining their palates to distinguish good wines from bad. Robert Parker, the most influential wine critic in America, called Ashenfelter's predictions "ludicrous and absurd", comparing him to a "movie critic who never goes to see the movie but tells you how good it is based on the actors and the director." It did not help that Ashenfelter had also openly challenged Parker's rating of the 1986 Bordeaux. Parker thought they would be "very good and sometimes exceptional." But according to Ashenfelter's formula, the low summer temperatures and high harvest rainfalls in 1986 doomed the vintage.

Who was right? Thirty years later, Robert Parker ranks the 1986 Bordeaux well, but the 1990 Bordeaux wines are exceptional, with three of the six wines scoring a 98 on a 100-point scale.

We will reproduce Ashenfelter's analysis, which is an example of *machine learning*. Machine learning is concerned with the general problem of how to use data to make predictions. The process of producing a model like Ashenfelter's from data is called *fitting* a model (although the terms *training* or *learning* are also used), and the data that is used to fit the model is the *training data*.

Getting Familiar with the Data

First, we read in the historical data that Ashenfelter used. The observational unit in this data set is the vintage, so we index this `DataFrame` by the year.

In [1]:
```
import pandas as pd
data_dir = ""
bordeaux_df = pd.read_csv("bordeaux.csv",index_col="year")
bordeaux_df.head()
```

Out[1]:

year	price	summer	har	sep	win	age
1952	37.0	17.1	160	14.3	600	40
1953	63.0	16.7	80	17.3	690	39
1955	45.0	17.1	130	16.8	502	37
1957	22.0	16.1	110	16.2	420	35
1958	18.0	16.4	187	19.1	582	34

The **price** column is in 1981 dollars, normalized so that the 1961 Bordeaux has a price of 100. Price is a reasonable proxy for the quality of the wine. The **summer** column contains the average summer temperature (in degrees Celsius), while the **har** and **win** columns contain the harvest and winter rainfalls (in millimeters). The **sep** column stores the average temperature in September, which Ashenfelter did not include in his model.

Let us also take a peek at the end of this `DataFrame`.

In [2]:
```
bordeaux_df.tail()
```

Out[2]:

year	price	summer	har	sep	win	age
1987	NaN	17.0	115	18.9	452	5
1988	NaN	17.1	59	16.8	808	4
1989	NaN	18.6	82	18.4	443	3
1990	NaN	18.7	80	19.3	468	2
1991	NaN	17.7	183	20.4	570	1

We see that the `DataFrame` also contains data for vintages where the price is missing (including 1990, the vintage for which Ashenfelter made his prediction). In fact, prices are only available up to 1980, as it takes several years before wine quality can be estimated with much reliability), so only part of the `DataFrame` can be used for training. The rest of the data, where the features are known but the target is not, is called the *test data*. Machine learning fits a model to the training data, which is then used to predict the targets in the test data. The following code splits the `DataFrame` into the training and test sets.

In [3]:
```
bordeaux_train = bordeaux_df.loc[:1980].copy()
bordeaux_test = bordeaux_df.loc[1981:].copy()
```

Warm-Up: A Model with One Feature

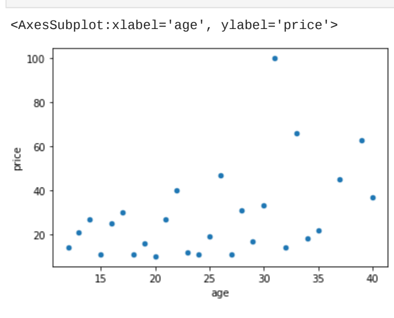
Before fitting a model that uses all of the features, we first consider a model that uses only the age of the wine to predict the price. That is, we fit a model of the form

$$\widehat{\text{price}} = b + c \cdot \text{age}, \tag{3}$$

where b and c are numbers that we will learn from the training data. Models of the form above are called *linear regression* models. (The way in which this model is "linear" will become apparent in a moment.) This model only involves two variables, **age** and **price**, so we can visualize the data easily using a scatterplot (see Chapter 3).

In [4]:
```
bordeaux_train.plot.scatter(x="age", y="price")
```

Out[4]: <AxesSubplot:xlabel='age', ylabel='price'>



Now, to fit models like the above to the training data, we use the scikit-learn package, which was used in Chapter 3 for transforming variables and calculating distances. However, its main purpose is to fit machine learning models, including linear regression. All models in scikit-learn are used in essentially the same way, following the three-step pattern:

1. Declare the model.
2. Fit the model to training data.
3. Use the model to predict on test data.

In the case of the linear regression model above, the code is as follows.

In [5]:
```
from sklearn.linear_model import LinearRegression

X_train = bordeaux_train[["age"]]
X_test = bordeaux_test[["age"]]
y_train = bordeaux_train["price"]

model = LinearRegression()
model.fit(X=X_train, y=y_train)
model.predict(X=X_test)
```

Out[5]:
```
array([12.41648163, 11.26046336, 10.1044451 ,  8.94842683,  7.79240856,
        6.6363903 ,  5.48037203,  4.32435376,  3.1683355 ,  2.01231723,
        0.85629897])
```

The parameters of `.fit()` are `X` for the features and `y` for the targets, which are assumed to be 2-D and 1-D arrays of numbers, respectively. So even when there is only one feature, as in this case, we still need to supply a 2-D array with one column---hence, the double brackets around `"age"` when defining `X_train` and `X_test`.

By contrast, `.predict()` only has one parameter, `X` for the features. That is because its job is to predict the targets `y` for the given features. Note that the predictions will always be returned in the form of `numpy` arrays, no matter the type of the input data---so although we supplied `pandas` objects, `sklearn` still returned the predicted values as `numpy` arrays. The predictions are in the same order as the rows of `X`.

Because there are only two variables involved, the model above is a rare example of a machine learning model we can visualize. A general way to do this is to generate a fine grid of `X` values using `np.linspace()` and call `model.predict()` to get the predicted target at each of these values. We can then use these predictions to draw a curve which depicts the predicted value of `y` at each value of `X`. In the code below, we put the predictions in a `pandas` `Series`, indexed by the `X` values, and then call `.plot.line()`.

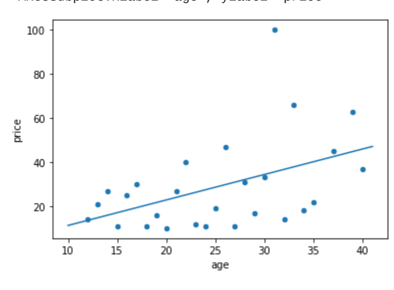
In [6]:
```
import numpy as np

X_new = pd.DataFrame()
# create a sequence of 200 evenly spaced numbers from 10 to 41
X_new["age"] = np.linspace(10, 41, num=200)

# create a Series out of the predicted values
# (trailing underscore indicates fitted values)
y_new_ = pd.Series(
    model.predict(X_new), # y values in Series.plot.line()
    index=X_new["age"]    # x values in Series.plot.line()
)

# plot the data, then the model
bordeaux_train.plot.scatter(x="age", y="price")
y_new_.plot.line()
```

Out[6]: <AxesSubplot:xlabel='age', ylabel='price'>



The resulting plot is shown above. Notice that the curve is a straight line, which is why this model is called *linear* regression. In hindsight, this is obvious from the model equation: b is simply the intercept and c the slope of this line. All linear regression does is choose the intercept and slope to minimize the total squared distance between the points and the line---that is, between the observed and predicted prices. In mathematical terms, b and c are chosen to minimize

$$\underset{\text{over training data}}{\text{sum of }(\text{price} - \widehat{\text{price}})^2} \tag{4}$$
$$= \underset{\text{over training data.}}{\text{sum of }(\text{price} - (b + c \cdot \text{age}))^2} \tag{5}$$

Since `sklearn` does this optimization for us, it is not necessary to understand the details of this process to extract useful insights out of linear regression. However, the math is explained in the appendix of this lesson for those who are curious.

What to Do about Nonlinearity

One question is whether the relationship between age and price is truly linear. In the graph above, it seems that the points deviate more from the line when prices are high than when they are low. To correct this, we need to spread out low prices and rein in high prices. Previously, we learned that this can be achieved by applying a log transformation to the prices. Let's add a column to the training data for the log-price.

In [7]:
```
bordeaux_train["log(price)"] = np.log(bordeaux_train["price"])
```

Now, we will fit a linear regression model to predict this new target. That is, in contrast to the previous model, we now fit the model

$$\widehat{\log(\text{price})} = b + c \cdot \text{age}, \tag{6}$$

where b and c are chosen to minimize

$$\text{sum of } (\log(\text{price}) - \widehat{\log(\text{price})})^2 \text{ over training data} \tag{7}$$

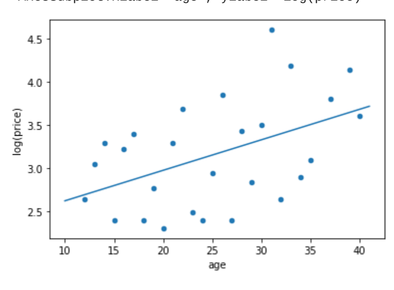
over the training data. The code below fits this model.

In [8]:
```
log_price_model = LinearRegression()
log_price_model.fit(X=bordeaux_train[["age"]],
                    y=bordeaux_train["log(price)"])

X_new = pd.DataFrame()
X_new["age"] = np.linspace(10, 41, num=200)
y_new_ = pd.Series(
    log_price_model.predict(X_new),
    index=X_new["age"]
)

bordeaux_train.plot.scatter(x="age", y="log(price)")
y_new_.plot.line()
```

Out[8]: <AxesSubplot:xlabel='age', ylabel='log(price)'>



The points are more evenly spread out when the target is log-price instead of price. For this reason, Ashenfelter chose log-price to be the measure of "wine quality" in his linear regression model.

Fitting Ashenfelter's Model

We are now ready to reproduce Ashenfelter's analysis. To do so, we will need to fit a linear regression model that predicts the log-price from the average summer temperature, winter rainfall, harvest rainfall, and the age of the wine. In other words, the model is of the form

$$\begin{aligned} \widehat{\log(\text{price})} = b &+ c_1 \cdot (\text{average summer temperature}) \\ &+ c_2 \cdot (\text{winter rainfall}) \\ &+ c_3 \cdot (\text{harvest rainfall}) \\ &+ c_4 \cdot (\text{age of the wine}), \end{aligned} \tag{8}$$

where b, c_1, c_2, c_3, c_4 are chosen to minimize

$$\text{sum of } (\log(\text{price}) - \widehat{\log(\text{price})})^2 \text{ over training data.} \tag{9}$$

This is still a *linear regression* model, albeit a more complicated one.

The code to fit this model is the natural extension of the code we wrote to fit the earlier models in this lesson. Instead of passing `bordeaux_train[["age"]]` for `X`, we now supply a `DataFrame` containing all of the features we want to be in the model.

In [9]:
```
ashen_model = LinearRegression()
ashen_model.fit(
    X=bordeaux_train[["summer", "win", "har", "age"]],
    y=bordeaux_train["log(price)"]
)
```

Out[9]: LinearRegression()

This model is much harder to visualize, since it involves five variables: four features, plus the target. Nevertheless, we can obtain predictions from it just as we did with the simpler models above. We just need to supply the values of all of the features in the model, in the same order as in the training data.

In [10]:
```
ashen_model.predict(
    X=bordeaux_test[["summer", "win", "har", "age"]]
)
```

Out[10]:
```
array([3.17926885, 3.4231464 , 3.71919787, 2.83391541, 3.48195778,
       2.4330387 , 2.91879638, 3.5924235 , 3.97294747, 4.04789338,
       3.14087609])
```

Communication Corner: Interpreting the Model

Even though we cannot visualize Ashenfelter's model, we can still interpret the model by examining the values of the *intercept* b and the *coefficients* c_1, c_2, c_3, c_4.

The coefficients are saved in the `.coef_` attribute, after the model has been fitted. (As above, the trailing underscore in `.coef_` reminds us that these are fitted values.)

In [11]:
```
ashen_model.coef_
```

Out[11]:
```
array([ 0.61871092,  0.00119721, -0.00374825,  0.02435187])
```

These coefficients are in the same order as the columns of `X`. So 0.61871092 is the coefficient for **summer**, 0.00119721 the coefficient for **win**, and so on. If you compare these values with the model at the beginning of this lesson, you will see that they are exactly the coefficients that Ashenfelter obtained.

A positive coefficient means that the predicted target *increases* as that feature increases, while a negative coefficient means that it *decreases* as that feature increases. Since **win** has a positive coefficient (0.0012) and **har** has a negative coefficient (-0.0037), we conclude from the model that Bordeaux wines tend to be best when winter rainfall is high and harvest rainfall is low.

Another essential component of a linear regression model is the *intercept*, which is stored in the `.intercept_` attribute, separately from the coefficients.

In [12]:
```
ashen_model.intercept_
```

Out[12]: -7.831137841446707

In principle, the intercept is the predicted value when all of the features are equal to 0. However, this interpretation is often purely hypothetical, since it may be impossible for some features to be 0. For example, to interpret the intercept of -7.8 in the model above, we would have to set **summer** equal to 0. That is, we would have to imagine a summer in Bordeaux, France where the average temperature was 0°C (i.e., freezing), which would be so catastrophic that the quality of red wine would be the least of our worries!

Exercises

Exercises 1-3 ask you to fit linear regression models to the Ames housing data set (AmesHousing.txt), which contains information about homes in Ames, Iowa.

1. Fit a linear regression model that predicts the price of a home (**SalePrice**) using square footage (**Gr Liv Area**) as the only feature. Then, make a graph of the fitted model (this is possible because there is only one feature in this model). Do this the way we did it in the lesson, by creating a grid of `X` values and calling `model.predict()` on those `X` values.

In [13]:
```
house_df = pd.read_csv("AmesHousing.txt", index_col="Yr Sold", sep="\t")
display(house_df)

house_model = LinearRegression()
house_model.fit(X=house_df[["Gr Liv Area"]], y=house_df["SalePrice"])
house_model.predict(X=house_df[["Gr Liv Area"]])
```

Yr Sold	Order	PID	MS SubClass	MS Zoning	Lot Frontage	Lot Area	Street	Alley	Lot Shape	Land Contour	...	Screen Porch	Pool Area	Pool QC	Fence	Misc Feature	Misc Val	Mo Sold	Sale Type	Sale Condition	SalePrice
2010	1	526301100	20	RL	141.0	31770	Pave	NaN	IR1	Lvl	...	0	0	NaN	NaN	NaN	0	5	WD	Normal	215000
2010	2	526350040	20	RH	80.0	11622	Pave	NaN	Reg	Lvl	...	120	0	NaN	MnPrv	NaN	0	6	WD	Normal	105000
2010	3	526351010	20	RL	81.0	14267	Pave	NaN	IR1	Lvl	...	0	0	NaN	NaN	Gar2	12500	6	WD	Normal	172000
2010	4	526353030	20	RL	93.0	11160	Pave	NaN	Reg	Lvl	...	0	0	NaN	NaN	NaN	0	4	WD	Normal	244000
2010	5	527105010	60	RL	74.0	13830	Pave	NaN	IR1	Lvl	...	0	0	NaN	MnPrv	NaN	0	3	WD	Normal	189900
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2006	2926	923275080	80	RL	37.0	7937	Pave	NaN	IR1	Lvl	...	0	0	NaN	GdPrv	NaN	0	3	WD	Normal	142500
2006	2927	923276100	20	RL	NaN	8885	Pave	NaN	IR1	Low	...	0	0	NaN	MnPrv	NaN	0	6	WD	Normal	131000
2006	2928	923400125	85	RL	62.0	10441	Pave	NaN	Reg	Lvl	...	0	0	NaN	MnPrv	Shed	700	7	WD	Normal	132000
2006	2929	924100070	20	RL	77.0	10010	Pave	NaN	Reg	Lvl	...	0	0	NaN	NaN	NaN	0	4	WD	Normal	170000
2006	2930	924151050	60	RL	74.0	9627	Pave	NaN	Reg	Lvl	...	0	0	NaN	NaN	NaN	0	11	WD	Normal	188000

2930 rows × 81 columns

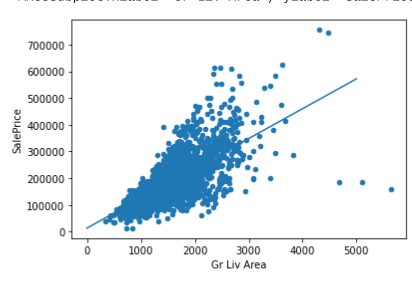
Out[13]:
```
array([198254.89978528, 113367.45913335, 161730.96150478, ...,
       121632.81519683, 168432.60155624, 236677.63608036])
```

This array represents the predicted Sale price of a home in dollars for each home, given the square footage.

In [14]:
```
X_new = pd.DataFrame()
X_new["Gr Liv Area"] = np.linspace(0, 5000)
y_new_ = pd.Series(
    house_model.predict(X_new), # y values in Series.plot.line()
    index=X_new["Gr Liv Area"]    # x values in Series.plot.line()
)

# plot the data, then the model
house_df.plot.scatter(x="Gr Liv Area", y="SalePrice")
y_new_.plot.line()
```

Out[14]: <AxesSubplot:xlabel='Gr Liv Area', ylabel='SalePrice'>

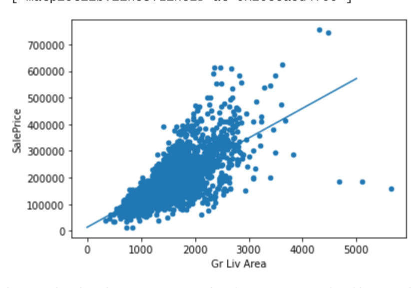


2. There is another way to graph a fitted linear regression model: extract the intercept and coefficient and draw a line with that intercept and slope. Verify that this gives the same graph as Exercise 2.

In [15]:
```
import matplotlib.pyplot as plt
house_df.plot.scatter(x="Gr Liv Area", y="SalePrice")
print("Slope is ",house_model.coef_, "Intercept is ",house_model.intercept_)
y_pred_ = house_model.coef_ * X_new["Gr Liv Area"] + house_model.intercept_
plt.plot(X_new["Gr Liv Area"], y_pred_)
```

```
Slope is  [111.69400086] Intercept is  13289.63436475952
```

Out[15]: [<matplotlib.lines.Line2D at 0x20eeacd4700>]



Yes, the graphs match. The slope represents that for every 1 Sq. ft. of house, there is an associated $111.69 increase in the sale price.

3. Fit a linear regression model that predicts the price of a home using square footage, number of bedrooms (**Bedroom AbvGr**), number of full bathrooms (**Full Bath**), and number of half bathrooms (**Half Bath**). Interpret the coefficients. Then, use your fitted model to predict the price of a home that is 1500 square feet, with 3 bedrooms, 2 full baths, and 1 half bath.

In [16]:
```
house_model = LinearRegression()
house_model.fit(
    X = house_df[["Gr Liv Area", "Bedroom AbvGr", "Full Bath", "Half Bath"]],
    y = house_df["SalePrice"]
)
house_model.predict(
    X = house_df[["Gr Liv Area", "Bedroom AbvGr", "Full Bath", "Half Bath"]]
)
```

Out[16]:
```
array([179259.36600924, 119498.42279868, 141911.84752569, ...,
        98242.85629925, 177721.65791096, 247885.39263788])
```

These coefficients represent the predicted prices of a home, with the given four variables.

In [17]:
```
coef = house_model.coef_
print(coef)
price = house_model.intercept_ + coef[0]*1500 + coef[1]*3 + coef[2]*2 + coef[3]*1
print(price)
```

```
[   118.09986838 -29994.95675968  26728.53342793   1271.13847733]
188835.45844692792
```

So the predicted price of the home is $188,835.46. I suspect that the -30K coefficient for num. bedrooms is caused by the 3 outliers (large sq. ft. low price) that is skewing this coefficient so much.

Exercises 4-5 ask you to fit linear regression models to the tips data (tips.csv), which contains information about tips collected by a waiter.

In [18]:
```
tips_df = pd.read_csv("tips.csv")
display(tips_df)
```

	obs	totbill	tip	sex	smoker	day	time	size
0	1	16.99	1.01	F	No	Sun	Night	2
1	2	10.34	1.66	M	No	Sun	Night	3
2	3	21.01	3.50	M	No	Sun	Night	3
3	4	23.68	3.31	M	No	Sun	Night	2
4	5	24.59	3.61	F	No	Sun	Night	4
...	...	...	...	...	...	...	...	...
239	240	29.03	5.92	M	No	Sat	Night	3
240	241	27.18	2.00	F	Yes	Sat	Night	2
241	242	22.67	2.00	M	Yes	Sat	Night	2
242	243	17.82	1.75	M	No	Sat	Night	2
243	244	18.78	3.00	F	No	Thu	Night	2

244 rows × 8 columns

```
4\. Suppose you want to predict how much a male diner will tip on a Sunday bill of \$40.00. Fit a linear regression model to the tips data to answer this question. (Hint: You will need to convert
categorical variables to quantitative variables. asZaqAZ)
```

In [19]:
```
#Set sex and day to dummy variables
#pd.set_option("display.max_rows",None, "display.max_columns", None)
newValues = {"sex": {'F': 0, 'M': 1}}
df_tips_new = tips_df
df_tips_new = tips_df.replace(newValues)
newValues = {'day': {'Thu': 0, 'Fri': 1, 'Sat':2, 'Sun':3}}
df_tips_new = df_tips_new.replace(newValues)

#tips_df["sex"] = pd.factorize(tips_df["sex"])[0] # F is 0, M is 1
#display(tips_df["sex"])
#tips_df["day"] = pd.factorize(tips_df["day"])[0] # Sun is 0, Sat 1, Thurs 2, Friday 3
#display(tips_df["day"])

tips_model = LinearRegression()
tips_model.fit(
    X = df_tips_new[["sex", "day", "totbill"]],
    y = df_tips_new["tip"]
)
coef = tips_model.coef_
print(coef)
tip = tips_model.intercept_ + coef[0]*1+coef[1]*3+coef[2]*40
print(tip)
```

```
[-0.03969456  0.02587309  0.10475113]
5.134548274120359
```

So for a male diner on a Sunday with a bill of $40.00, one can expect a tip of$ 5.13.

5. Fit a linear regression model, with no intercept, that predicts the tip from the total bill. That is, we want our predictions to be of the form

$$\widehat{\text{tip}} = c \cdot (\text{total bill}).$$

where c is some coefficient to be learned from the training data.

(*Hint:* `LinearRegression()` has a parameter, `fit_intercept=`, which is `True` by default.)

Plot the data and the fitted model. In practical terms, what assumption is being made when we fit a model with no intercept?

In [20]:
```
tips_model = LinearRegression(fit_intercept=False)
tips_model.fit(X=tips_df[["totbill"]], y=tips_df["tip"])
X_new = pd.DataFrame()
X_new["totbill"] = np.linspace(0, 70)
# create a Series out of the predicted values
# (trailing underscore indicates fitted values)
y_new_ = pd.Series(
    tips_model.predict(X_new), # y values in Series.plot.line()
    index=X_new["totbill"]    # x values in Series.plot.line()
)

# plot the data, then the model
tips_df.plot.scatter(x="totbill", y="tip")
y_new_.plot.line()
```

Out[20]: <AxesSubplot:xlabel='totbill', ylabel='tip'>

With no fit intercept, I assume that with a total bill of $0, there will be a tip of$ 0. Which of course makes locical sense.