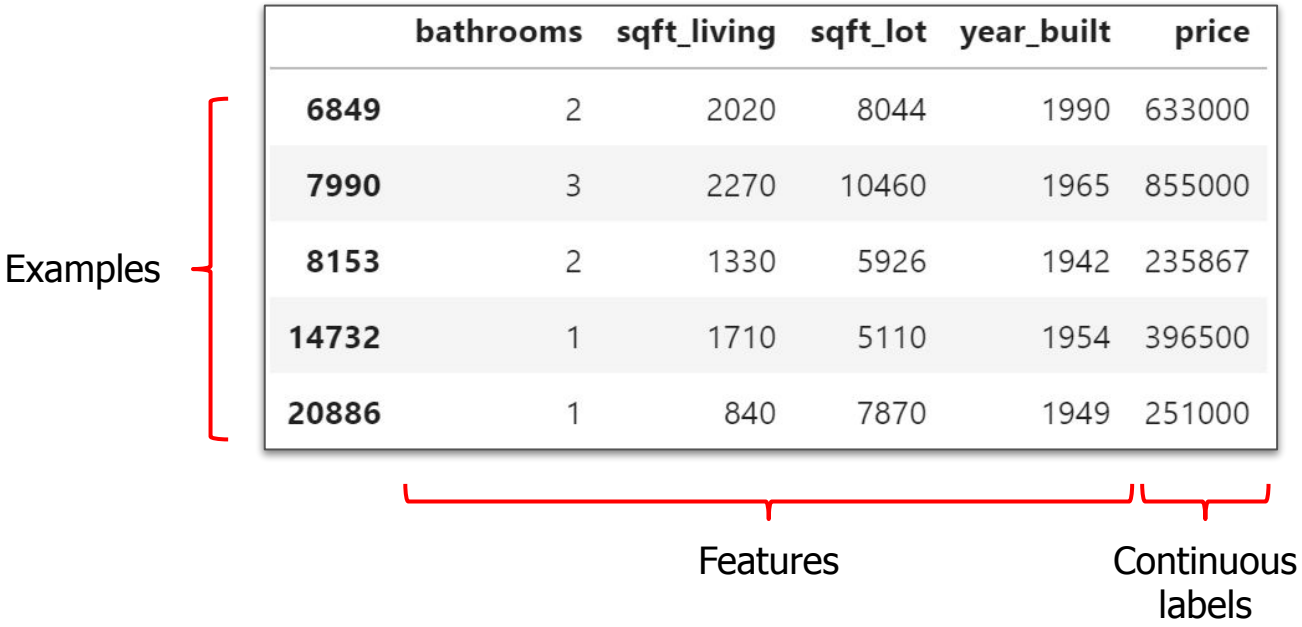
Linear Regression

Concept Module 11

Template by: Laurent Lessard

What is regression?

- Regression is simply classification with continuous labels



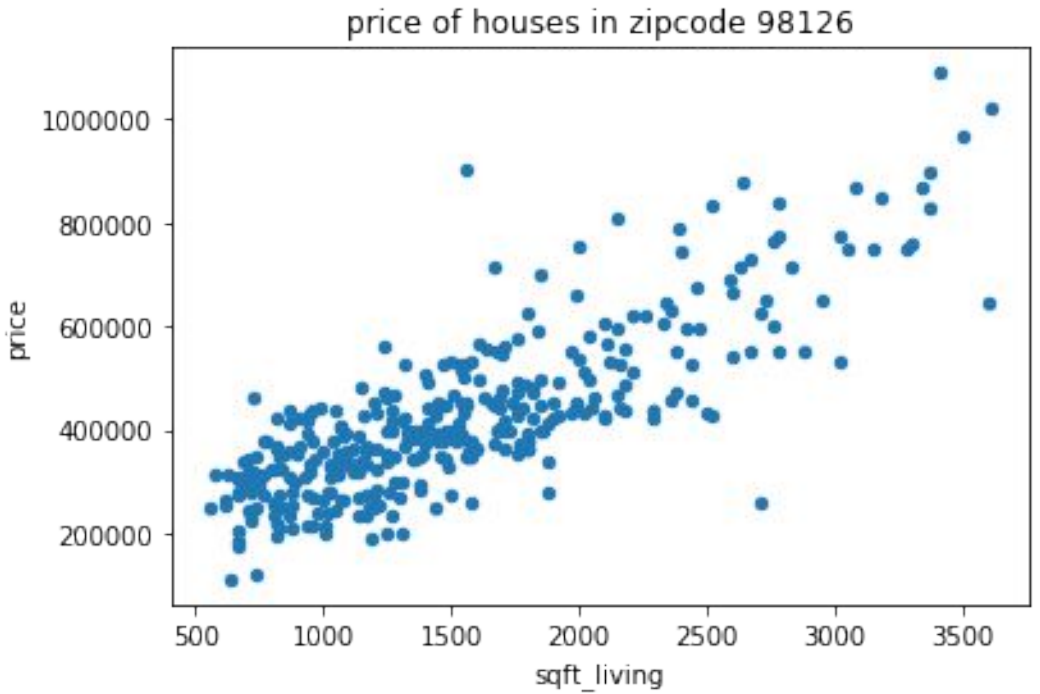
	bathrooms	sqft_living	sqft_lot	year_built	price
6849	2	2020	8044	1990	633000
7990	3	2270	10460	1965	855000
8153	2	1330	5926	1942	235867
14732	1	1710	5110	1954	396500
20886	1	840	7870	1949	251000

Regression with one feature

Feature: `sqft_living`, Label: `price`

Make a scatter plot!

Our labels are continuous.
Our predictions should be continuous as well!



Regression with one feature

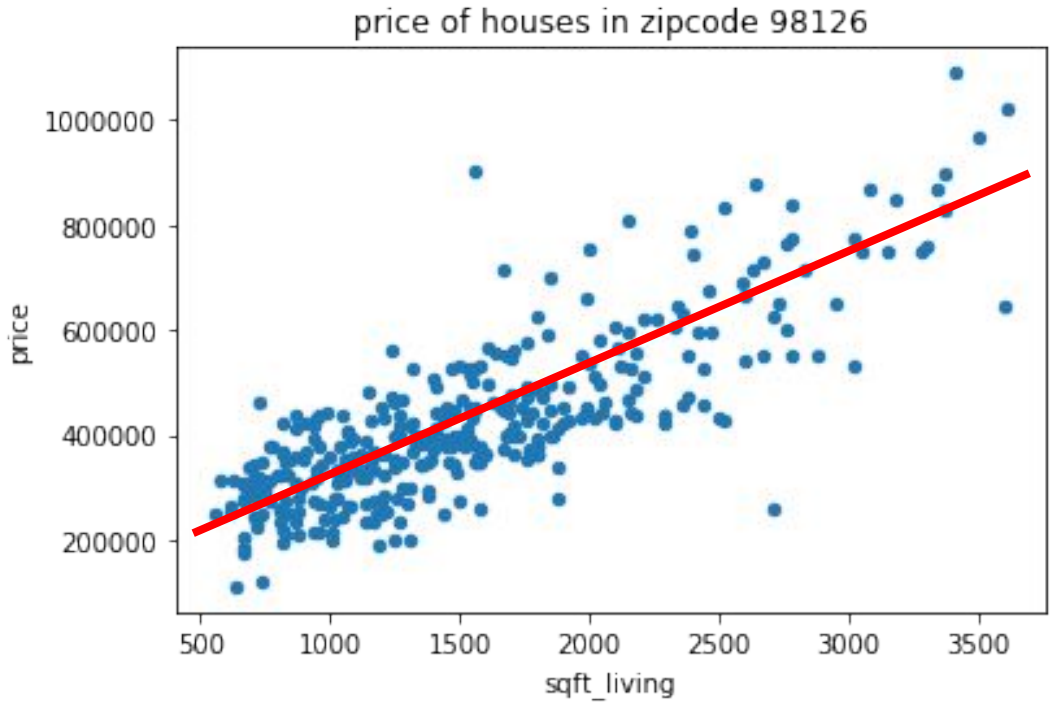
Main idea: find a line that best fits the data

Equation of line:

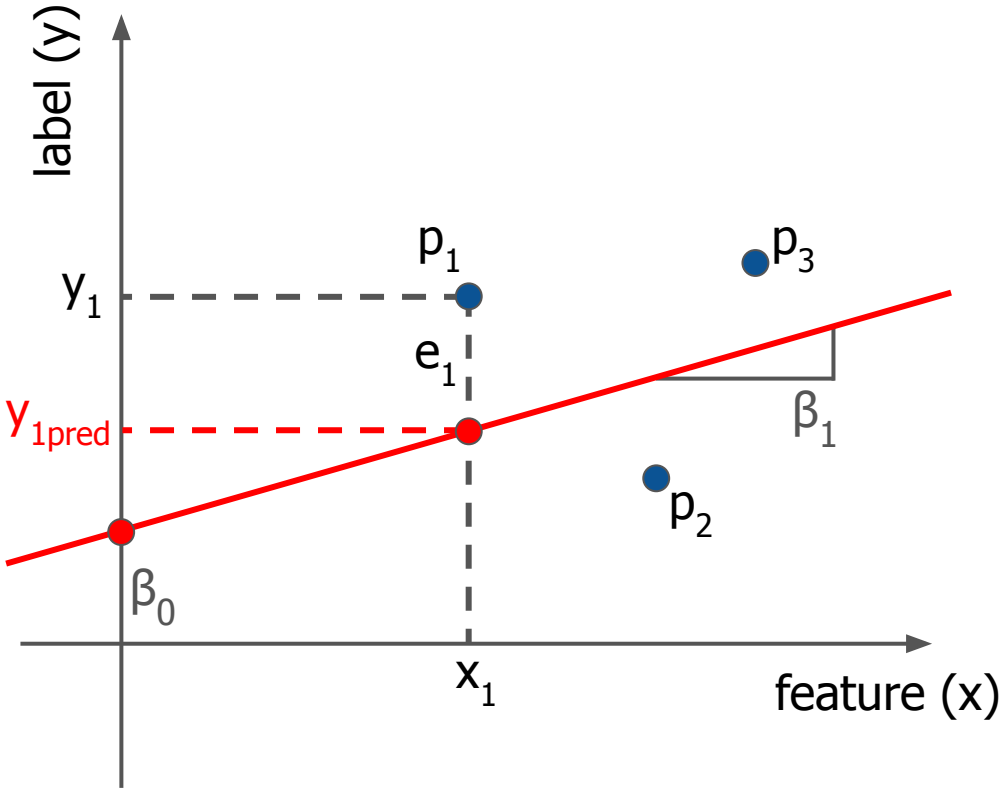
$$(\text{price}) = \beta_0 + \beta_1(\text{sqft\_living})$$

intercept

slope



Geometry of linear regression

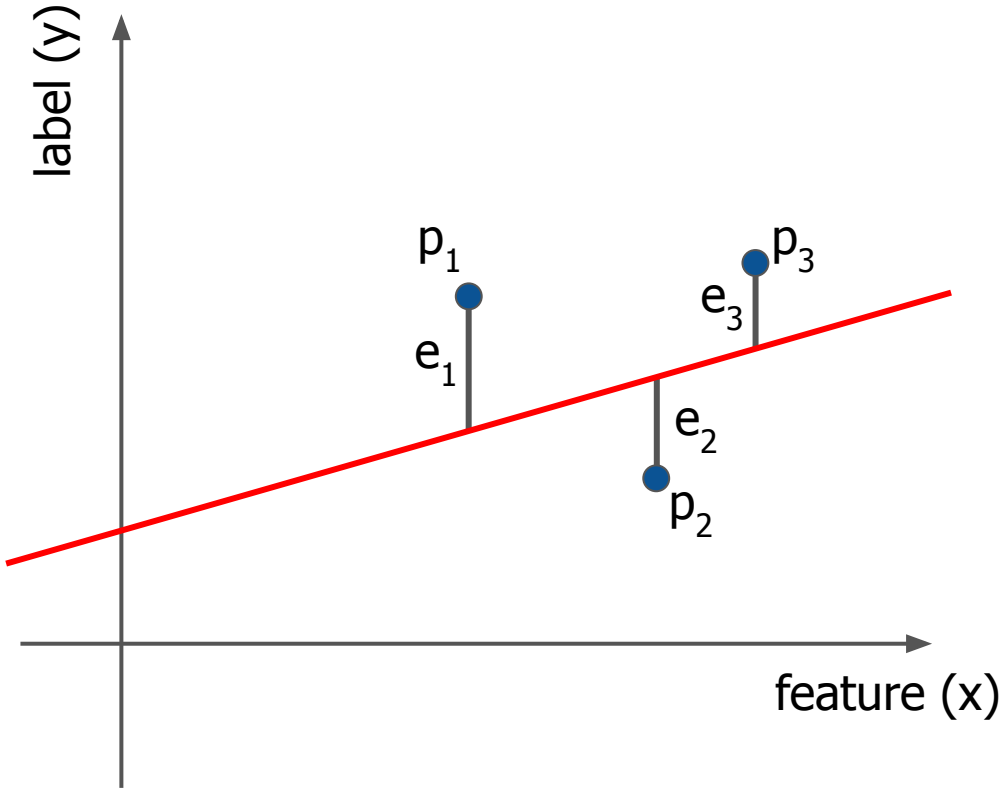


Equation of line:

$y = \beta_0 + \beta_1 x$

The residual for point k is $|e_k|$

Geometry of linear regression

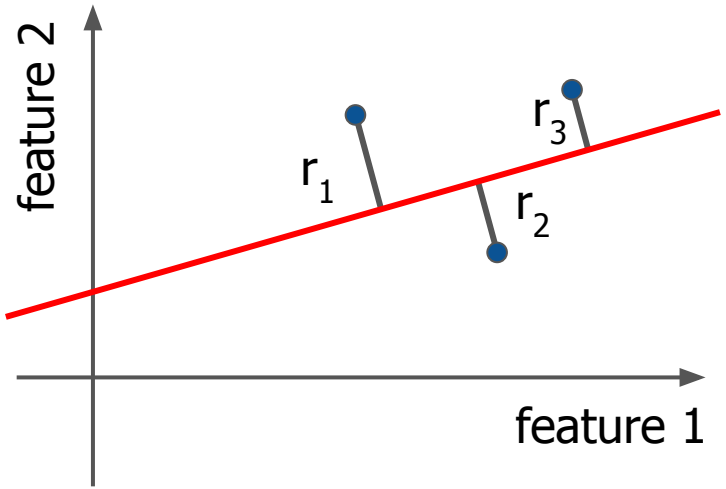


Goal: choose β_0, β_1 to minimize the residual sum of squares:

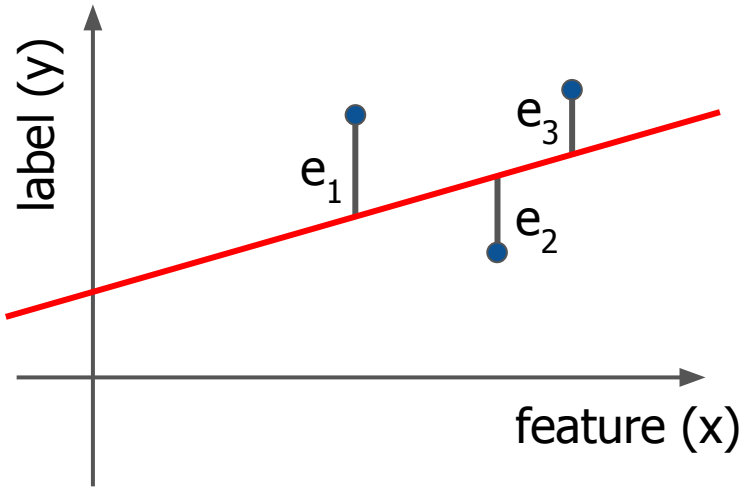
$$RSS = (e_1)^2 + \ldots + (e_n)^2$$

PCA vs linear regression

PCA: 2-D features, no labels.
minimize (perpendicular distance)^2



Regression: 1-D features,
continuous labels, minimize RSS

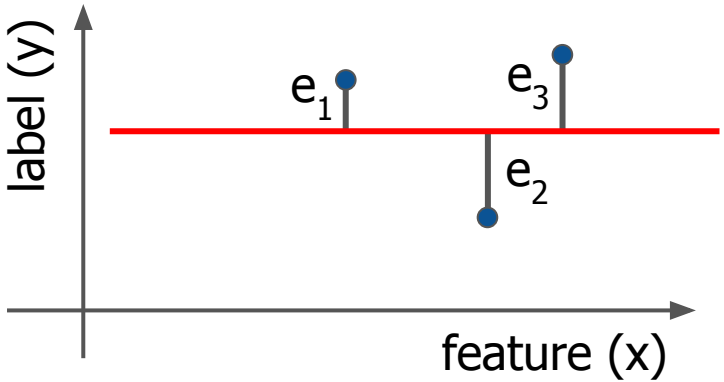


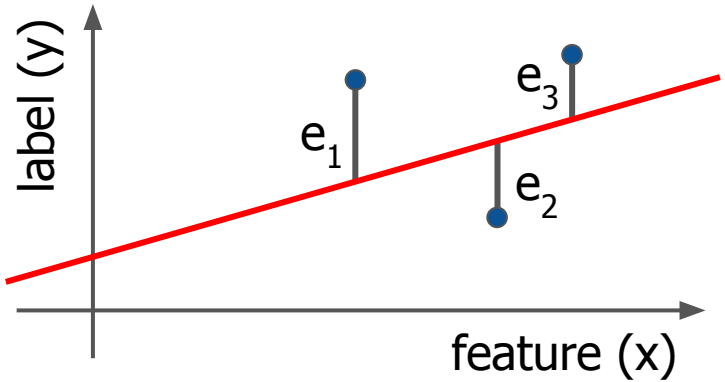
Total vs Residual sum of squares

If we force $\beta_1=0$ (no slope), then the best we can do is to set β_0 = mean(y). Then RSS = variance of y. We call this the "total sum of squares" (TSS).

TSS = total variance of the labels

If we choose β_0 and β_1 optimally, we will further improve RSS. So we will have 0 < RSS < TSS.





Coefficient of determination

R^2 or r^2 a.k.a. "R-squared" is:

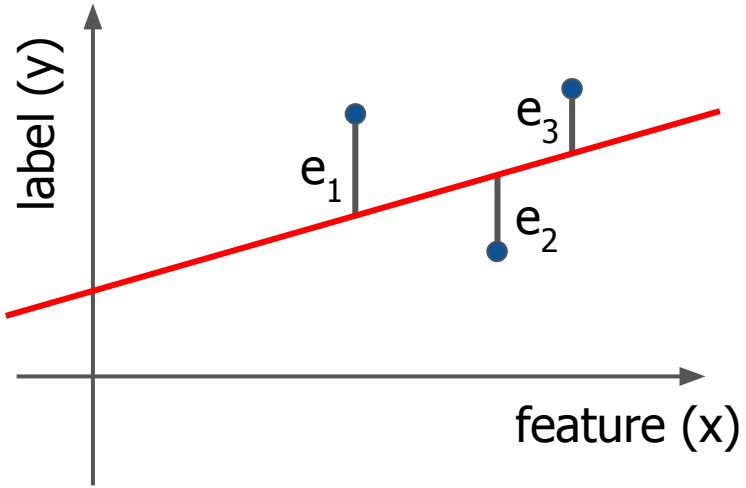
the proportion of the variance predictable from the feature x.

$$R^2 = 1 - \frac{RSS}{TSS}$$

If RSS is small: the residuals are small, and R^2 is close to 1.

If RSS large (close to TSS), then R^2 is close to 0.

$$0 < RSS < TSS$$



Linear regression in Python

```
from sklearn.linear_model import LinearRegression

X = df[['sqft_living']]  # feature (must be a column)
y = df['price']          # labels

regr = LinearRegression()
regr.fit(X,y)
```

Slope and intercept

```
# Obtain intercept
regr.intercept_

# Obtain slope
regr.coef_
```

Predict new data

```
# Predict labels for
# new unlabeled data
ytest = regr.predict(Xtest)
```

Get R^2 score

```
# Obtain R-squared
regr.score(X,y)
```

Housing data result

```
# Obtain intercept
regr.intercept_
```

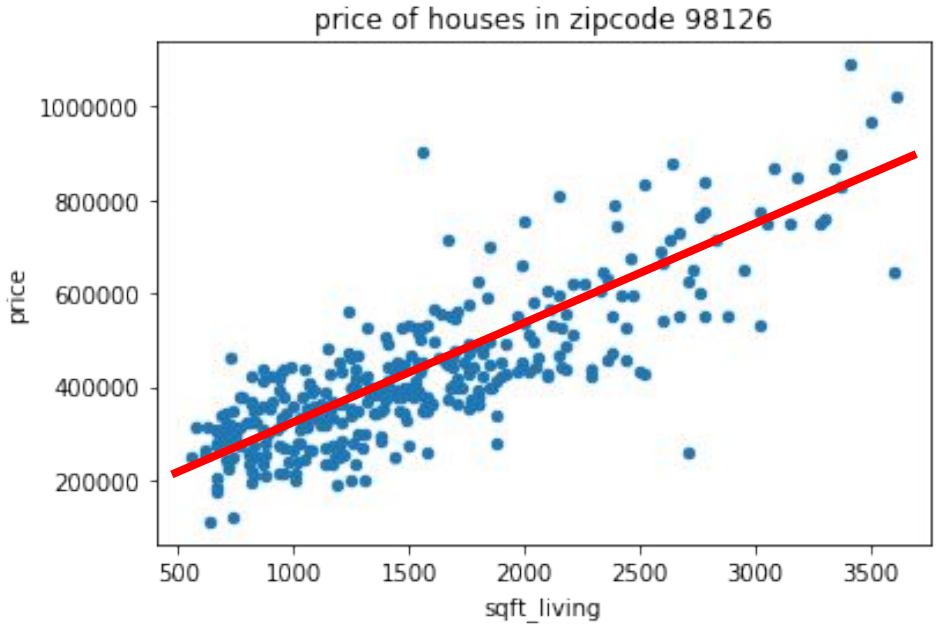
```
123936.34087573813
```

```
# Obtain slope
regr.coef_
```

```
array([ 194.47792472])
```

```
# Obtain R-squared
regr.score(X,y)
```

```
0.65741621635804526
```



Formula for line of best fit:

price = $123,936 + ($194 / sq.ft.) × (sq.ft. living)

Summary

- Regression is classification, but for continuous labels
- When there is one feature, we can plot label vs feature and visualize the regression as a line on this plot
- Line is characterized by slope and intercept.
- Linear regression minimizes sum of squared residuals (RSS)
- Quality of fit given by R^2 value (0=bad, 1=good).