Regression Analysis

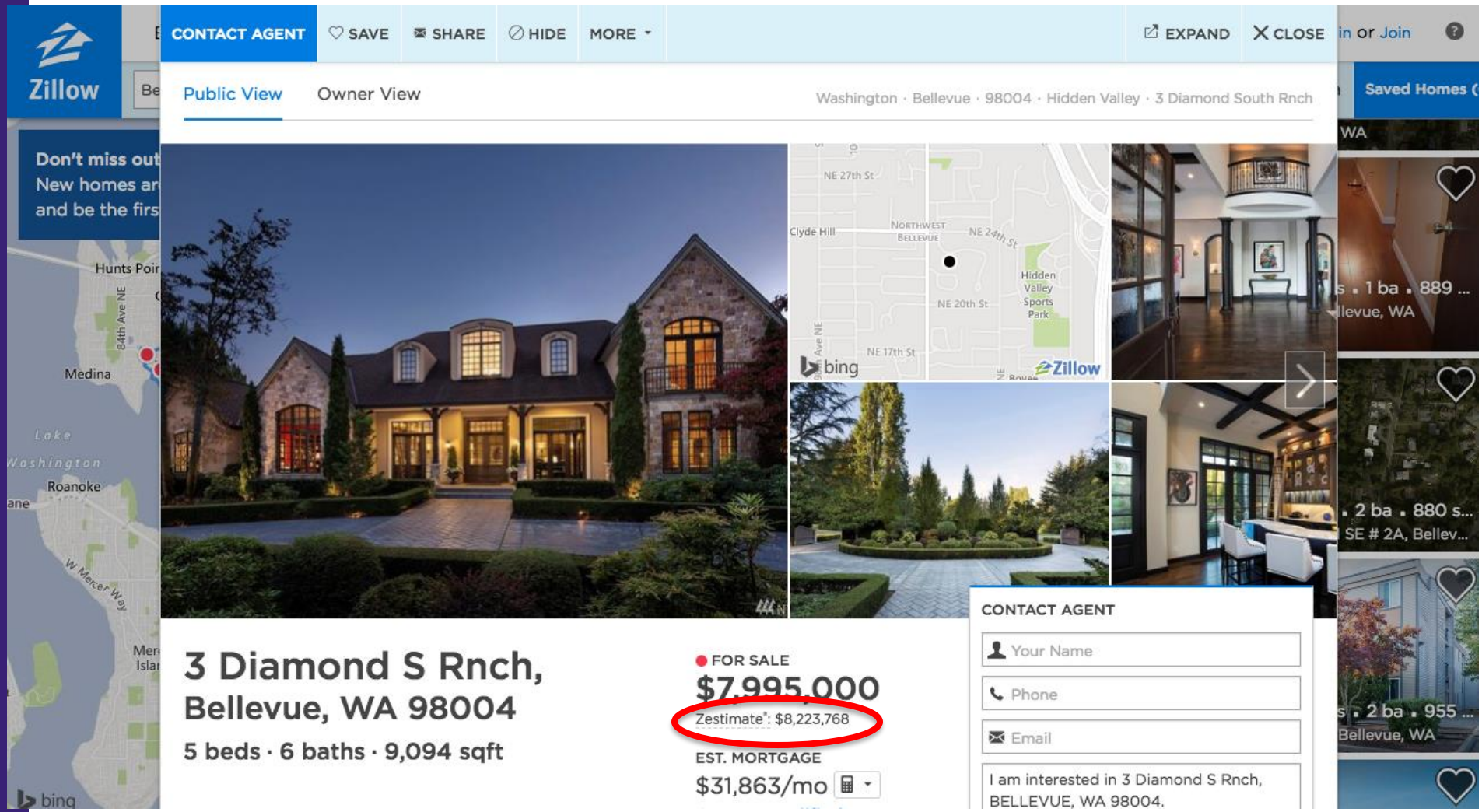
Foster
School of Business
UNIVERSITY of WASHINGTON

Outline

- Linear Regression

- Performance Evaluation

Business Understanding



Pre-collected Data

TABLE 2.1 DESCRIPTION OF VARIABLES IN WEST ROXBURY (BOSTON) HOME VALUE D

TOTAL VALUE	Total assessed value for property, in thousands of USD
TAX	Tax bill amount based on total assessed value multiplied by the tax rate, in USD
LOT SQ FT	Total lot size of parcel in square feet
YR BUILT	Year the property was built
GROSS AREA	Gross floor area
LIVING AREA	Total living area for residential properties (ft^2)
FLOORS	Number of floors
ROOMS	Total number of rooms
BEDROOMS	Total number of bedrooms
FULL BATH	Total number of full baths
HALF BATH	Total number of half baths
KITCHEN	Total number of kitchens
FIREPLACE	Total number of fireplaces
REMODEL	When the house was remodeled (Recent/Old/None)

Data: Flat Files

	A	B	C	D	E	F	G	H	I
1	TOTAL VALU	TAX	LOT SQFT	YR BUILT	GROSS AREA	LIVING AREA	FLOORS	ROOMS	BEDROOMS
2	344.2	4330	9965	1880	2436	1352	2	6	3
3	412.6	5190	6590	1945	3108	1976	2	10	4
5	498.6	6272	13773	1957	5032	2608	1	9	5
6	331.5	4170	5000	1910	2370	1438	2	7	3
7	337.4	4244	5142	1950	2124	1060	1	6	3
8	359.4	4521	5000	1954	3220	1916	2	7	3
9	320.4	4030	10000	1950	2208	1200	1	6	3
10	333.5	4195	6835	1958	2582	1092	1	5	3
11	409.4	5150	5093	1900	4818	2992	2	8	4
12	313	3937	5000	1960	2624	1485	1.5	6	3
13	344.5	4333	6768	1958	2844	1460	1.5	6	3
14	315.5	3968	5000	1889	2196	1290	2	6	3
15	575	7233	12288	2004	4616	2378	2	9	4
16	326.2	4103	5000	1954	2536	1272	1.5	6	3
17	298.2	3751	5000	1940	2129	864	1	7	3
18	313.1	3938	6949	1880	2612	1438	1.5	7	3
19	344.9	4338	10000	1950	2099	1445	1	7	3
20	330.7	4160	5000	1910	2408	1470	2	7	3
21	348	4377	9001	1875	2840	1632	2	7	3
22	317.5	3994	4450	1920	1400	1232	2	7	3
23	330.8	4161	5000	1889	2560	1302	1.5	6	2
24	357.8	4501	12255	1944	2631	1275	1.5	6	3
25	414.7	5216	12972	1892	3796	2054	1.5	6	3

instance, sample, example, record, observation (e.g., customer, house, applicant)

attribute, feature, predictor

Regression Analysis

- Goal: Predict a single numeric "target" or "outcome" variable
- Training data, where **target value is known**

	A	B	C	D	E	F	G	H	I
1	TOTAL VALU	TAX	LOT SQFT	YR BUILT	GROSS AREA	LIVING AREA	FLOORS	ROOMS	BEDROOMS
2	344.2	4330	9965	1880	2436	1352	2	6	3
3	412.6	5190	6590	1945	3108	1976	2	10	4
4	330.1	4152	7500	1890	2294	1371	2	8	4
5	498.6	6272	13773	1957	5032	2608	1	9	5
6	331.5	4170	5000	1910	2370	1438	2	7	3
7	337.4	4244	5142	1950	2124	1060	1	6	3
8	359.4	4521	5000	1954	3220	1916	2	7	3
9	320.4	4030	10000	1950	2208	1200	1	6	3
10	333.5	4195	6835	1958	2582	1092	1	5	3
11	409.4	5150	5093	1900	4818	2992	2	8	4
12	313	3937	5000	1960	2624	1485	1.5	6	3
13	344.5	4333	6768	1958	2844	1460	1.5	6	3
14	315.5	3968	5000	1889	2196	1290	2	6	3
15	575	7233	12288	2004	4616	2378	2	9	4
16	326.2	4103	5000	1954	2536	1272	1.5	6	3
17	298.2	3751	5000	1940	2129	864	1	7	3
18	313.1	3938	6949	1880	2612	1438	1.5	7	3
19	344.9	4338	10000	1950	2099	1445	1	7	3
20	330.7	4160	5000	1910	2408	1470	2	7	3
21	348	4377	9001	1875	2840	1632	2	7	3
22	317.5	3994	4450	1920	1400	1232	2	7	3
23	330.8	4161	5000	1889	2560	1302	1.5	6	2
24	357.8	4501	12255	1944	2631	1275	1.5	6	3
25	414.7	5216	12972	1892	3796	2054	1.5	6	3

Target, outcome

(e.g. sales, revenue, Performance)

Regression Analysis

- Goal: Predict a single numeric “target” or “outcome” variable
- Training data, where **target value is known**
- Scoring new observations, where value is not known

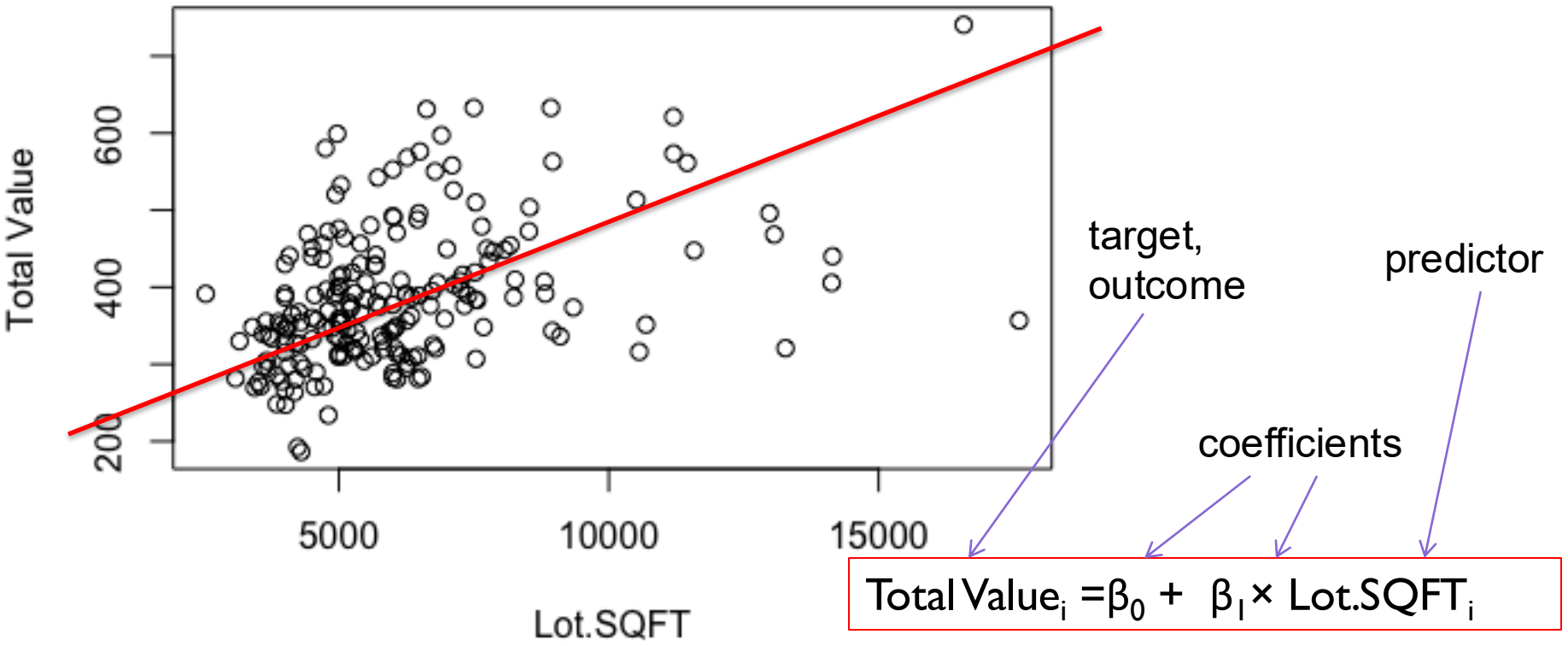
House on sale?

A	C	D	E	F	G	H	I
TOTAL VALU	LOT SQFT	YR BUILT	GROSS AREA	LIVING AREA	FLOORS	ROOMS	BEDROOMS
?	6733	1990	2880	1792	2	7	3

Foster School of Business

Simple Linear Regression

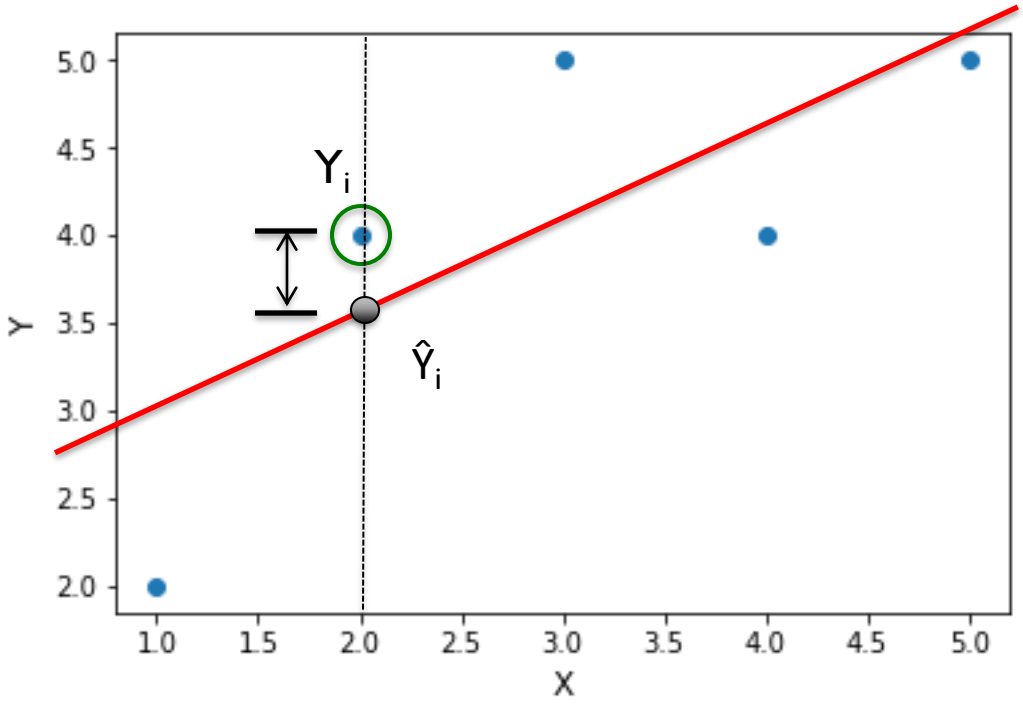
- Simple linear regression is a linear approach to model the relationship between a numeric target variable and one explanatory variables (predictors).



Intuition behind the Regression Line

> Sample dataset

$e_i = Y_i - \hat{Y}_i$

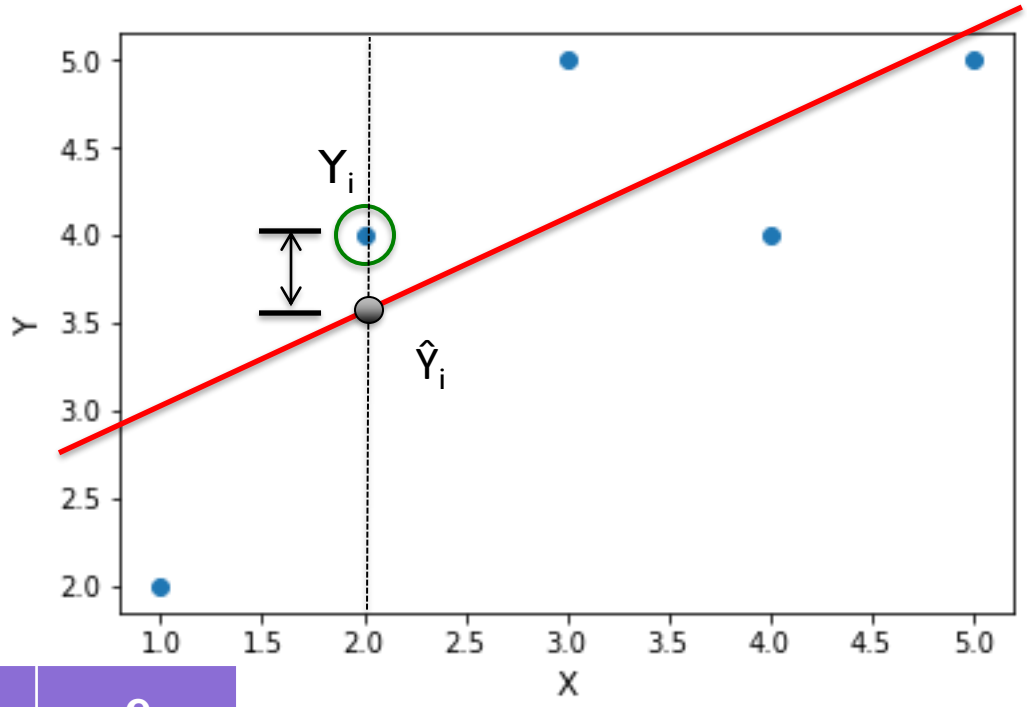


Predictor	Target
X	Y
1	2
2	4
3	5
4	4
5	5

Intuition behind the Regression Line

> Sample dataset

$e_i = Y_i - \hat{Y}_i$

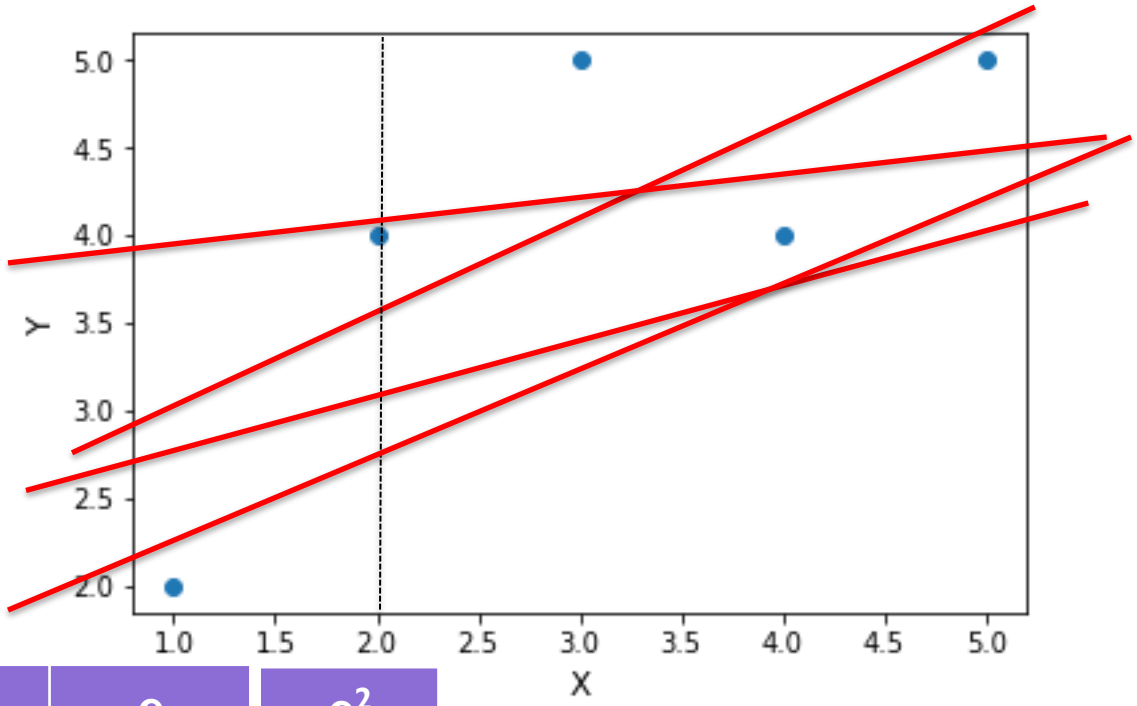


Predictor	Target		
X	Y	$\hat{Y}$	e
1	2	2.8	-0.8
2	4	3.4	0.6
3	5	4	1
4	4	4.6	-0.6
5	5	5.2	-0.2

Intuition behind the Regression Line

- Sample dataset
- Goodness of fit:

  $\Sigma (Y_i - \hat{Y}_i)^2$



Predictor	Target			
X	Y	$\hat{Y}$	e	e^2
1	2	2.8	-0.8	0.64
2	4	3.4	0.6	0.36
3	5	4	1	1
4	4	4.6	-0.6	0.36
5	5	5.2	-0.2	0.04

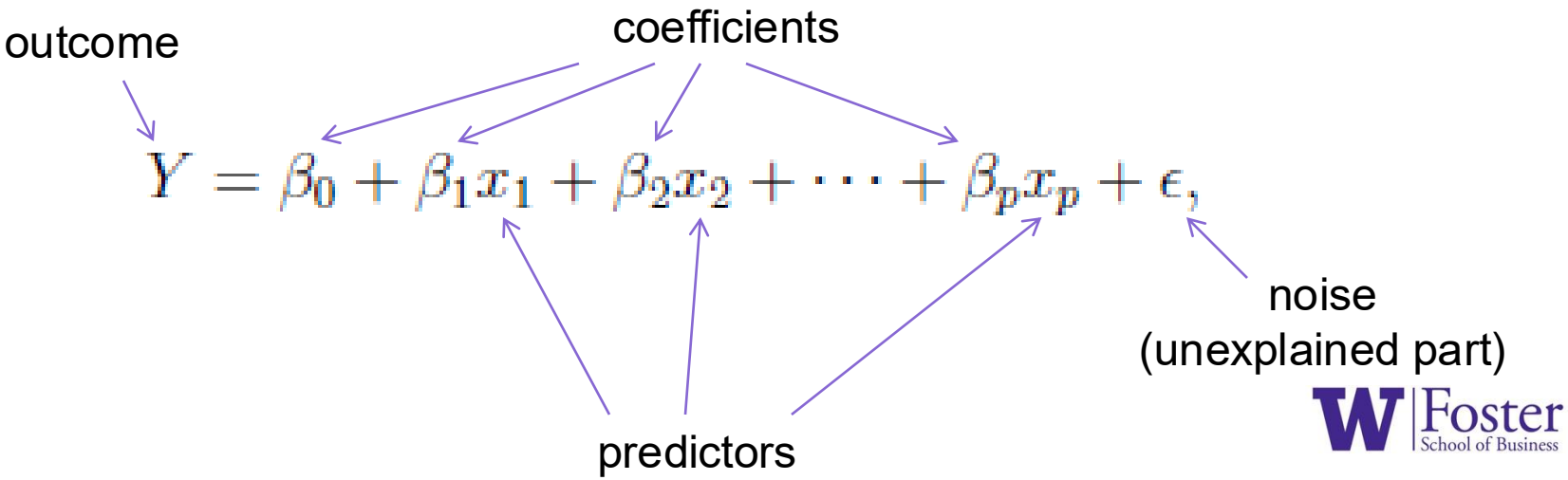
Estimate coefficients from data (fit model)

For any β values, we can compute the predicted value $\hat{Y}_i$ for each data point i

Find the best values that minimizes the sum of squared errors, $\Sigma(Y_i - \hat{Y}_i)^2$

Multiple Linear Regression

- The most popular model for making predictions.
- This model is used to fit a linear relationship between a numerical outcome variable Y (*target, dependent variable*), and a set of predictors x_1, x_2,…,x_p (also referred to as *independent variables, explanatory variables, input variables, predictors*).



Multiple Linear Regression

- Choose predictors based on business understanding

$$\text{TOTAL VALUE} = \beta_0 + \beta_1 \times \text{LOT SQFT} + \beta_2 \times \text{LIVING AREA} + \beta_3 \times \text{ROOMS}$$

- Variables in the linear equation must be numeric

- Estimate coefficients from data (fit model)
  - For any β values, we can compute the predicted value $\hat{Y}_i$ for each data point i
  - Find the best $\hat{\boldsymbol{\beta}}$ values that minimizes the sum of squared errors, $\Sigma(Y_i - \hat{Y}_i)^2$

Outline

- Linear Regression

- Performance Evaluation

Foster
School of Business
UNIVERSITY of WASHINGTON

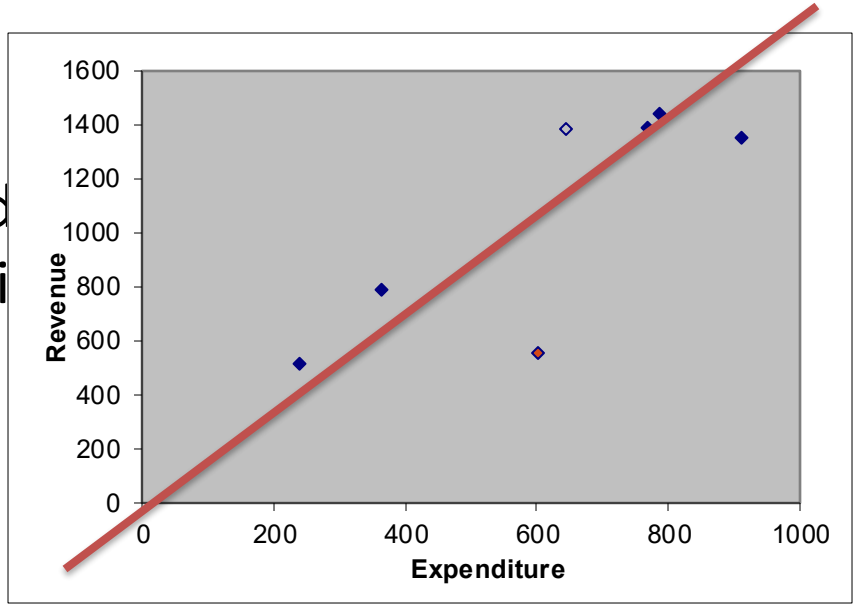
Prediction Accuracy Measures

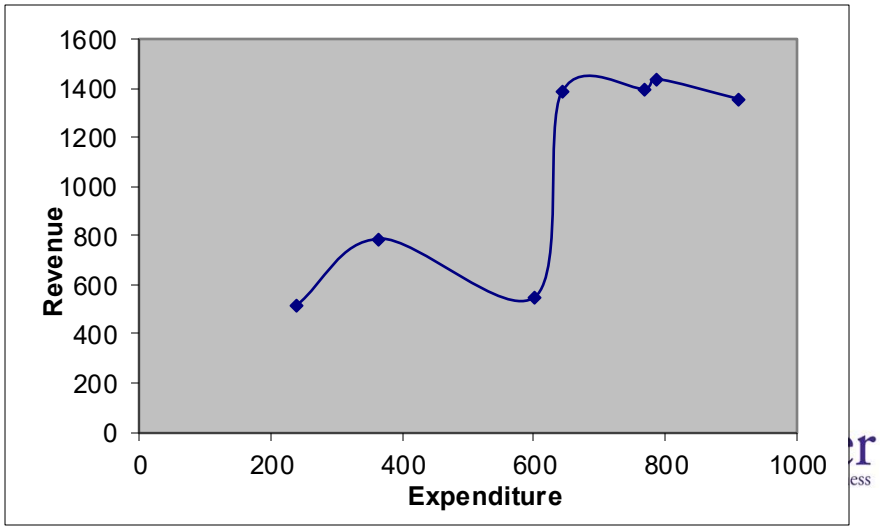
Why Evaluate?

- Multiple methods are available
- For each method, multiple choices of variables to use
- To choose the best model, need to assess each model's performance

The Problem of **Overfitting**

- Data mining models can prod
  of relationships between vari
- The "fit" may be excellent





The Problem of **Overfitting**

- Data mining models can produce highly complex explanations of relationships between variables
- The "fit" may be excellent
- **When used with new data, models of great complexity usually do not perform so well; that is, Models of great complexity over-fits the noises/randomness in data**
  - e.g. 100% fit – not useful for new data

- Consequence: model is too specialized and adapted to the training data that it is unable to generalize and make correct predictions on new data.
- Deployed model will not work as well as expected with completely new data.

Partition the Data for Performance Evaluation

	A	B	C	D	E	F	G	H	I
1	TOTAL VALU	TAX	LOT SQFT	YR BUILT	GROSS AREA	LIVING AREA	FLOORS	ROOMS	BEDROOMS
2	344.2	4330	9965	1880	2436	1352	2	6	3
3	412.6	5190	6590	1945	3108	1976	2	10	4
4	330.1	4152	7500	1890	2294	1371	2	8	4
5	498.6	6272	13773	1957	5032	2608	1	9	5
6	331.5	4170	5000	1910	2370	1438	2	7	3
7	337.4	4244	5142	1950	2124	1060	1	6	3
8	359.4	4521	5000	1954	3220	1916	2	7	3
9	320.4	4030	10000	1950	2208	1200	1	6	3
10	333.5	4195	6835	1958	2582	1092	1	5	3
11	409.4	5150	5093	1900	4818	2992	2	8	4
12	313	3937	5000	1960	2624	1485	1.5	6	3
13	344.5	4333	6768	1958	2844	1460	1.5	6	3
14	315.5	3968	5000	1889	2196	1290	2	6	3
15	575	7233	12288	2004	4616	2378	2	9	4
16	326.2	4103	5000	1954	2536	1272	1.5	6	3
17	298.2	3751	5000	1940	2129	864	1	7	3
18	313.1	3938	6949	1880	2612	1438	1.5	7	3
19	344.9	4338	10000	1950	2099	1445	1	7	3
20	330.7	4160	5000	1910	2408	1470	2	7	3
21	348	4377	9001	1875	2840	1632	2	7	3
22	317.5	3994	4450	1920	1400	1232	2	7	3
23	330.8	4161	5000	1889	2560	1302	1.5	6	2
24	357.8	4501	12255	1944	2631	1275	1.5	6	3
25	414.7	5216	12972	1892	3796	2054	1.5	6	3

New Unseen Data

A	C	D	E	F	G	H	I
TOTAL VALU	LOT SQFT	YR BUILT	GROSS AREA	LIVING AREA	FLOORS	ROOMS	BEDROOMS
?	6733	1990	2880	1792	2	7	3

Partition the Data for Performance Evaluation

	A	B	C	D	E	F	G	H	I
1	TOTAL VALU	TAX	LOT SQFT	YR BUILT	GROSS AREA	LIVING AREA	FLOORS	ROOMS	BEDROOMS
2	344.2	4330	9965	1880	2436	1352	2	6	3
3	412.6	5190	6590	1945	3108	1976	2	10	4
4	330.1	4152	7500	1890	2294	1371	2	8	4
5	498.6	6272	13773	1957	5032	2608	1	9	5
6	331.5	4170	5000	1910	2370	1438	2	7	3
7	337.4	4244	5142	1950	2124	1060	1	6	3
8	359.4	4521	5000	1954	3220	1916	2	7	3
9	320.4	4030	10000	1950	2208	1200	1	6	3
10	333.5	4195	6835	1958	2582	1092	1	5	3
11	409.4	5150	5093	1900	4818	2992	2	8	4
12	313	3937	5000	1960	2624	1485	1.5	6	3
13	344.5	4333	6768	1958	2844	1460	1.5	6	3
14	315.5	3968	5000	1889	2196	1290	2	6	3
15	575	7233	12288	2004	4616	2378	2	9	4
16	326.2	4103	5000	1954	2536	1272	1.5	6	3
17	298.2	3751	5000	1940	2129	864	1	7	3
18	313.1	3938	6949	1880	2612	1438	1.5	7	3
19	344.9	4338	10000	1950	2099	1445	1	7	3
20	330.7	4160	5000	1910	2408	1470	2	7	3
21	348	4377	9001	1875	2840	1632	2	7	3
22		3994	4450	1920	1400	1232	2	7	3
23		4161	5000	1889	2560	1302	1.5	6	2
24		4501	12255	1944	2631	1275	1.5	6	3
25		5216	12972	1892	3796	2054	1.5	6	3

Training Data

Predicted Value
300
320
380
420

Test Data

Partition the Data for Performance Evaluation

	A	B	C	D	E	F	G	H	I
1	TOTAL VALU	TAX	LOT SQFT	YR BUILT	GROSS AREA	LIVING AREA	FLOORS	ROOMS	BEDROOMS
2	344.2	4330	9965	1880	2436	1352	2	6	3
3	412.6	5190	6590	1945	3108	1976	2	10	4
4	330.1	4152	7500	1890	2294	1371	2	8	4
5	498.6	6272	13773	1957	5032	2608	1	9	5
6	331.5	4170	5000	1910	2370	1438	2	7	3
7	337.4	4244	5142	1950	2124	1060	1	6	3
8	359.4	4521	5000	1954	3220	1916	2	7	3
9	320.4	4030	10000	1950	2208	1200	1	6	3
10	333.5	4195	6835	1958	2582	1092	1	5	3
11	409.4	5150	5093	1900	4818	2992	2	8	4
12	313	3937	5000	1960	2624	1485	1.5	6	3
13	344.5	4333	6768	1958	2844	1460	1.5	6	3
14	315.5	3968	5000	1889	2196	1290	2	6	3
15	575	7233	12288	2004	4616	2378	2	9	4
16	326.2	4103	5000	1954	2536	1272	1.5	6	3
17	298.2	3751	5000	1940	2129	864	1	7	3
18	313.1	3938	6949	1880	2612	1438	1.5	7	3
19	344.9	4338	10000	1950	2099	1445	1	7	3
20	330.7	4160	5000	1910	2408	1470	2	7	3
21	348	4377	9001	1875	2840	1632	2	7	3
22	317.5	3994	4450	1920	1400	1232	2	7	3
23	330.8	4161	5000	1889	2560	1302	1.5	6	2
24	357.8	4501	12255	1944	2631	1275	1.5	6	3
25	414.7	5216	12972	1892	3796	2054	1.5	6	3

Training Data

Training error: e_i

Goodness-of-fit

Predicted Value
300
320
380
420

Test Data

Test error: e_i

Predictive Power

Random Partition

	A	B	C	D	E	F	G	H	I
1	TOTAL VALU	TAX	LOT SQFT	YR BUILT	GROSS AREA	LIVING AREA	FLOORS	ROOMS	BEDROOMS
2	344.2	4330	9965	1880	2436	1352	2	6	3
3	412.6	5190	6590	1945	3108	1976	2	10	4
4	330.1	4152	7500	1890	2294	1371	2	8	4
5	498.6	6272	13773	1957	5032	2608	1	9	5
6	331.5	4170	5886	1918	2370	1438	2	7	3
7	337.4	4244	5142	1950	2124	1060	1	6	3
8	359.4	4521	5000	1954	3220	1916	2	7	3
9	320.4	4030	10000	1950	2208	1200	1	6	3
10	333.5	4195	6835	1958	2582	1092	1	5	3
11	409.4	5150	5093	1900	4818	2992	2	8	4
12	313	3937	5000	1960	2624	1485	1.5	6	3
13	344.5	4333	6768	1958	2844	1460	1.5	6	3
14	315.5	3968	5000	1889	2196	1290	2	6	3
15	575	7233	12288	2004	4616	2378	2	9	4
16	326.2	4103	5000	1954	2536	1272	1.5	6	3
17	298.2	3751	5000	1940	2129	864	1	7	3
18	313.1	3938	6949	1880	2612	1438	1.5	7	3
19	344.9	4338	10000	1950	2099	1445	1	7	3
20	330.7	4160	5000	1910	2408	1470	2	7	3
21	348	4377	9001	1875	2840	1632	2	7	3
22	317.5	3994	4450	1920	1400	1232	2	7	3
23	330.8	4161	5000	1889	2560	1302	1.5	6	2
24	357.8	4501	12255	1944	2631	1275	1.5	6	3
25	414.7	5216	12972	1892	3796	2054	1.5	6	3

Training Data

Test Data

Requirement for Success in Learning

- The training set must be
  - (i) large enough to yield meaningful results
  - (ii) representative of the dataset as a whole
  - (iii) including accurate target values

	TOTAL VALU	TAX	LOT SQFT	YR BUILT	GROSS AREA	LIVING AREA	FLOORS	ROOMS	BEDROOMS
2	344.2	4330	9965	1880	2436	1352	2	6	3
3	412.6	5190	6590	1945	3108	1976	2	10	4
4	330.1	4152	7500	1890	2294	1371	2	8	4
5	498.6	6272	13773	1957	5032	2608	1	9	5
6	331.5	4170	5000	1910	2370	1438	2	7	3
7	337.4	4244	5142	1950	2124	1060	1	6	3
8	359.4	4521	5000	1954	3220	1916	2	7	3
9	320.4	4030	10000	1950	2208	1200	1	6	3
10	333.5	4195	6835	1958	2582	1092	1	5	3
11	409.4	5150	5093	1900	4818	2992	2	8	4
12	313	3937	5000	1960	2624	1485	1.5	6	3
13	344.5	4333	6768	1958	2844	1460	1.5	6	3
14	315.5	3968	5000	1889	2196	1290	2	6	3
15	575	7233	12288	2004	4616	2378	2	9	4
16	326.2	4103	5000	1954	2536	1272	1.5	6	3

Foster School of Business

Measures of Error

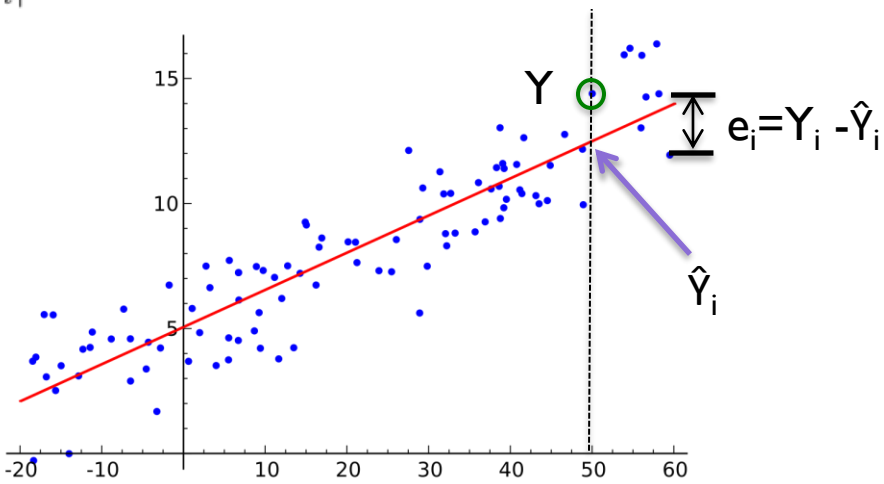
- Key component of most measures is the difference between actual Y and the predicted $\hat{Y}$, which also termed "error" or "residual": $e_i = Y_i - \hat{Y}_i$

- Mean Error: $\frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)$

- MAE (mean absolute error): $\frac{1}{n}\sum_{i=1}^{n} |e_i|$

- RMSE (root mean squared error):

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$$



Application of Linear Regression

Product price

Predict what would be the price of a used car

Application of Linear Regression

Score prediction

Predict the number of runs a player would score in the coming matches based on previous performance

https://www.degruyter.com/document/doi/10.1515/jqas-2015-0027/html