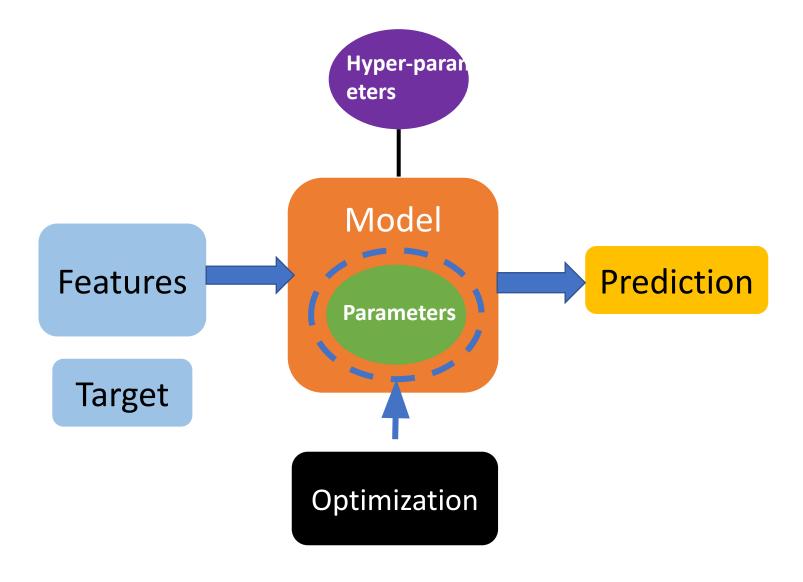


## **How Supervised Learning Works**



### What is Linear Regression

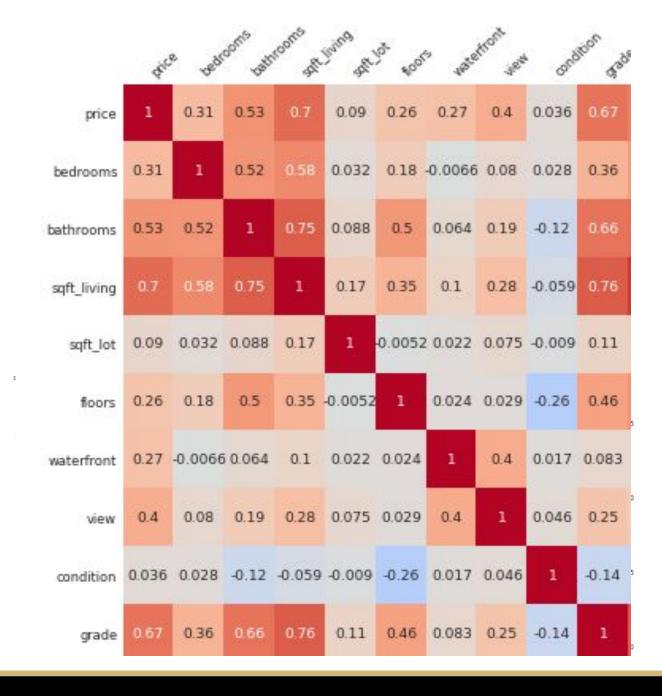
- Supervised learning model
- Predictive task- real valued numbers
- Parametric model
- No hyperparameters
- Features have linear relationship to the target variable

## Example

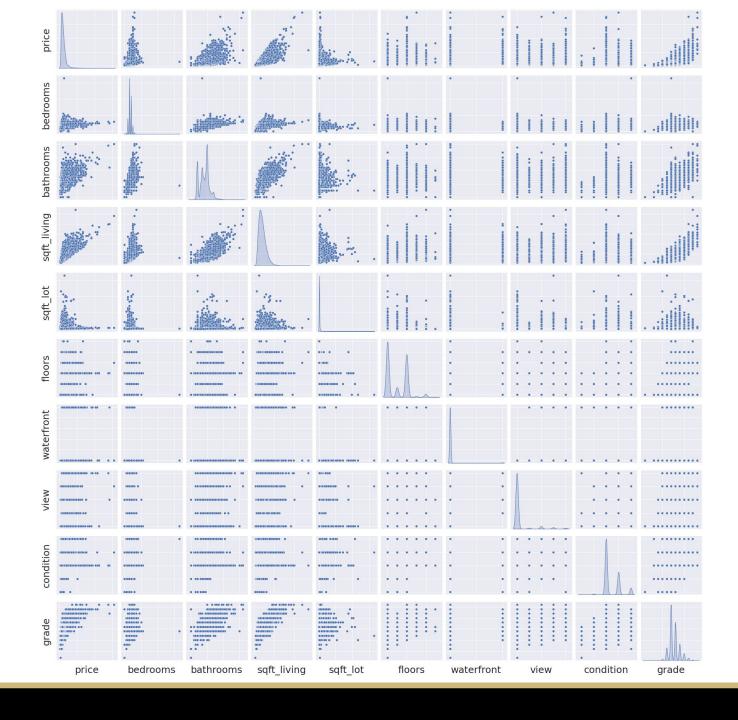
An example using House sales data from Kaggle https://www.kaggle.com/harlfoxem/housesalesprediction/download

price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view		grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode
221900	3	1.00	1180	5650	1.0	0	0		7	1180	0	1955	0	98178
538000	3	2.25	2570	7242	2.0	0	0	Kang	7	2170	400	1951	1991	98125
180000	2	1.00	770	10000	1.0	0	0		6	770	0	1933	0	98028
604000	4	3.00	1960	5000	1.0	0	0		7	1050	910	1965	0	98136
510000	3	2.00	1680	8080	1.0	0	0		8	1680	0	1987	0	98074

### **Correlation Matrix**



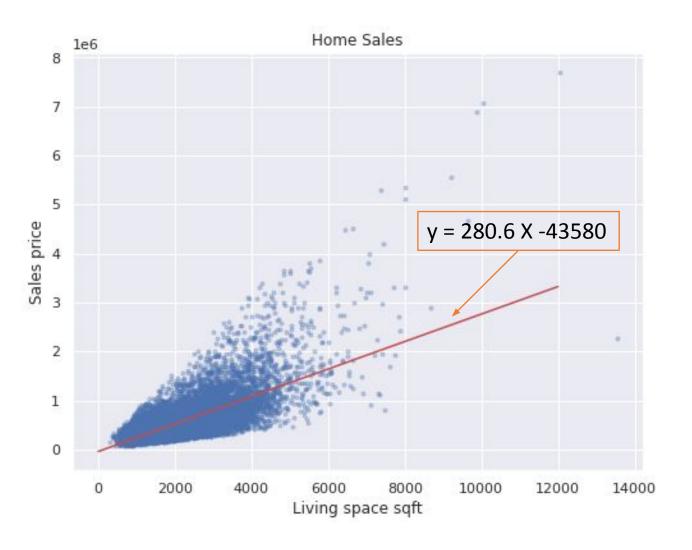
### Pair Plot



### **Univariate Linear Regression**

Intercept Slope Residual 
$$Y=eta_0+eta_1X+\epsilon$$
 Coefficients, or Parameters

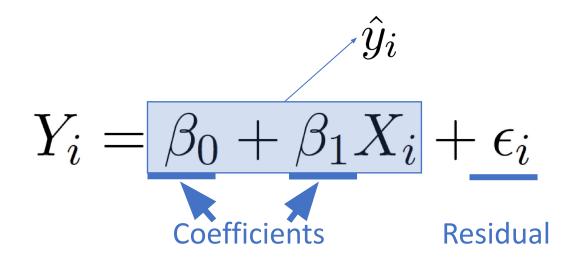
### Using statsmodel's OLS (ordinary least squares) package



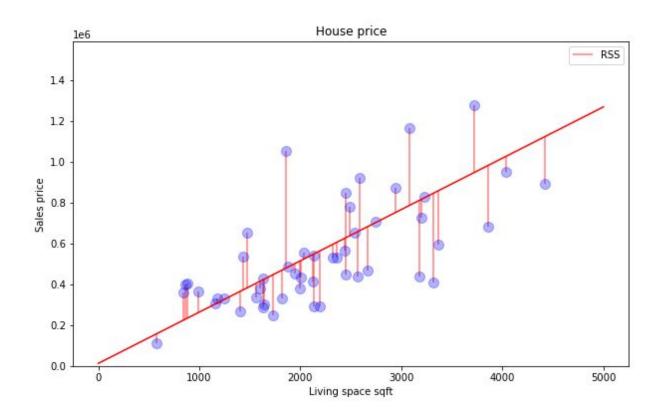
	coef std err		
Intercept -4.358	e <b>+04</b> 4402.690		
sqft_living 280.	<b>6236</b> 1.936		

- 1. How do we determine the coefficients?
- 2. How well does the model fit?
- 3. How significant are the coefficients?
- 4. How well does the model predict on unseen data?

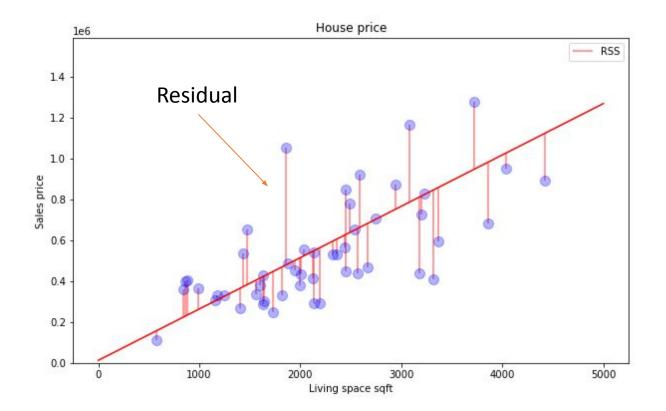
### Q1. How do we find the coefficients?



$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$



$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$



### Mean Absolute Error (MAE)

MAE = 
$$\frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

### Mean Percent Absolute Error (MAPE)

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

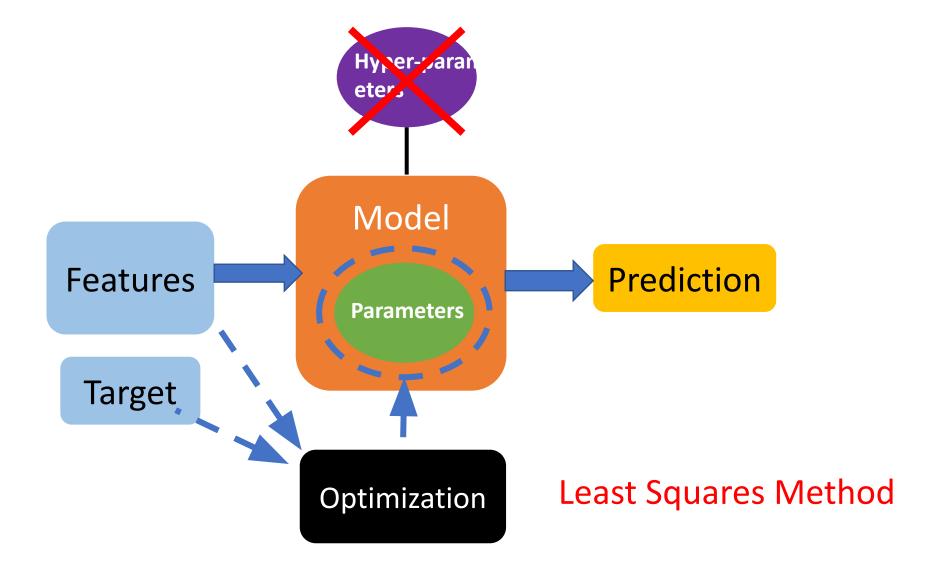
#### Mean Squared Error (MSE)

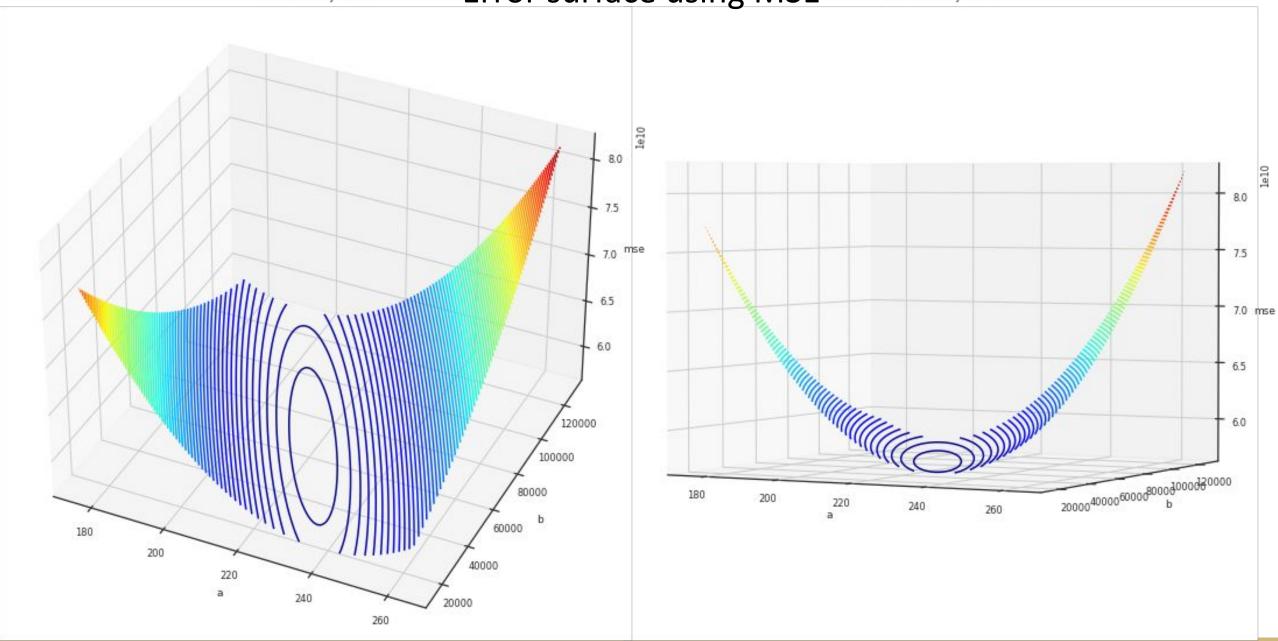
$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

### Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

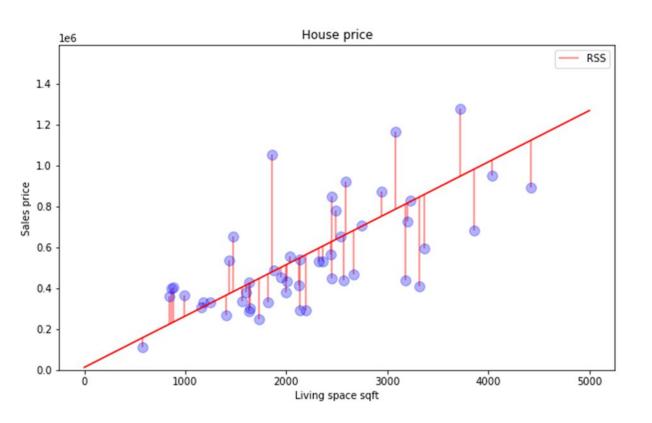
## **Optimization in Linear Regression**





## **Least Squares Method**

MSE = 
$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



$$\beta_1 = \frac{\operatorname{Cov}(X, Y)}{\operatorname{Var}(X)}$$

$$\beta_0 = \mathrm{E}[Y] - \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} \mathrm{E}[X]$$

## **Least Squares Method**



$$\beta_1 = \frac{\operatorname{Cov}(X, Y)}{\operatorname{Var}(X)}$$

$$\beta_0 = \mathrm{E}[Y] - \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} \mathrm{E}[X]$$

# What happens when scaling variables?



$$\beta_1 = \frac{\operatorname{Cov}(X, Y)}{\operatorname{Var}(X)}$$

$$\beta_0 = \mathrm{E}[Y] - \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} \mathrm{E}[X]$$

## Least Squares Method in Multivariate case

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$\frac{\partial \text{MSE}}{\partial \boldsymbol{\beta}} = 2\boldsymbol{X}^{\top} \boldsymbol{X} \boldsymbol{\beta} - 2\boldsymbol{X}^{\top} \boldsymbol{Y} + \boldsymbol{0} = \boldsymbol{0}$$

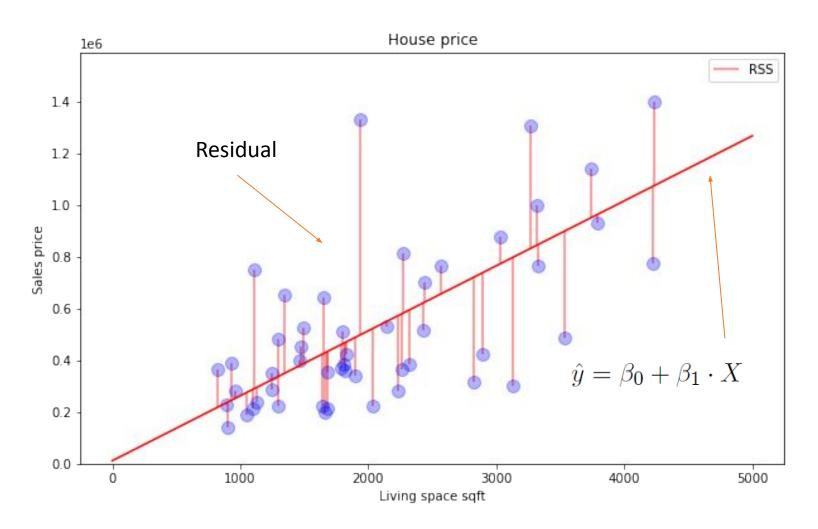
$$\boldsymbol{\beta} = (\boldsymbol{X}^{\top} \boldsymbol{X})^{-1} \boldsymbol{X}^{\top} \boldsymbol{Y}$$

### Q2. How well does the model fit?



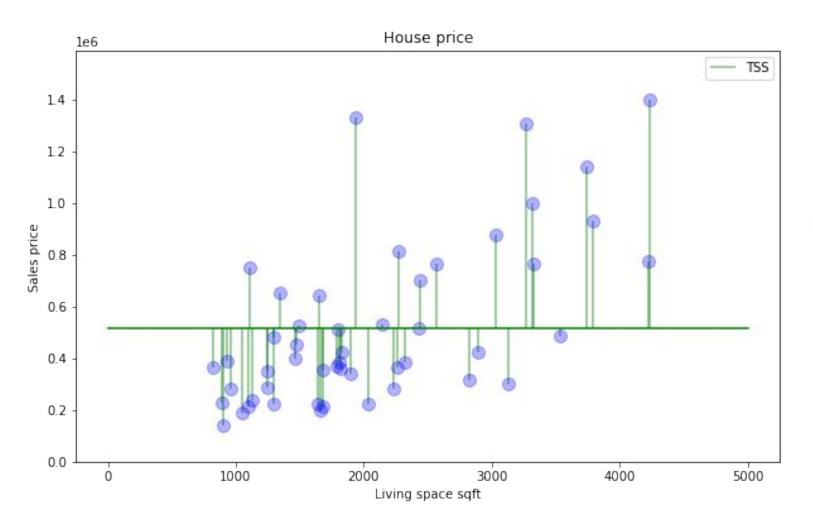
Dep. Variab	le:	price	e F	R-squar	ed:	0.493
Mod	el:	OLS	Adj. F	R-squar	ed:	0.493
Metho	od: Le	east Squares	3	F-statis	tic: 2.10	0e+04
Da	te: Thu,	25 Feb 2021	1 Prob (F	-statist	ic):	0.00
Tin	ne:	23:11:09	Log-l	ikeliho	od: -3.002	7e+05
No. Observation	ıs:	21613	3	A	MC: 6.00	5e+05
Df Residua	ls:	21611	1	Е	BIC: 6.00	6e+05
Df Mod	el:	•	1			
Covariance Typ	e:	nonrobus	t			
	coef	std err	t	P> t	[0.025	0.975
Intercept -4.3	358e+04	4402.690	-9.899	0.000	-5.22e+04	-3.5e+0
sqft_living 2	80.6236	1.936	144.920	0.000	276.828	284.41
Omnibus	14832	490 <b>Dur</b>	bin-Watso	on:	1.983	
Prob(Omnibus)	: 0.	000 Jarqu	e-Bera (J	B): 546	6444.709	
Skew	2.	824	Prob(J	B):	0.00	
Kurtosis	26.	977	Cond. N	lo.	5.63e+03	

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$



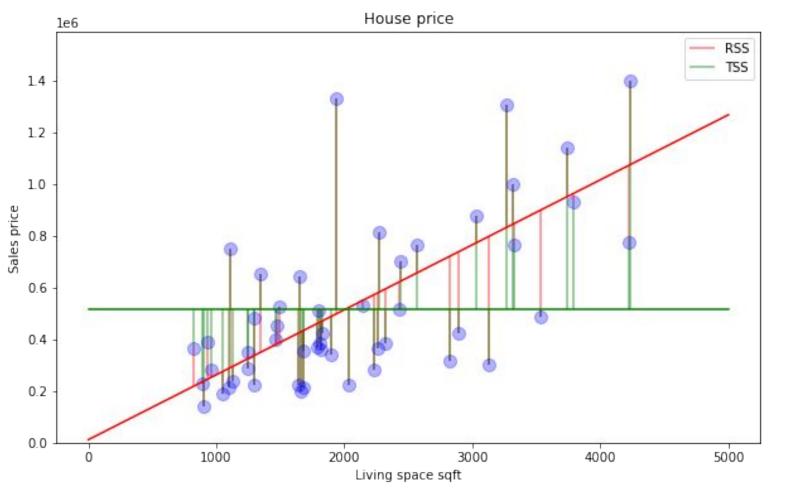
$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

"Residual Sum of Squares"



$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

"Total Sum of Squares"



$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

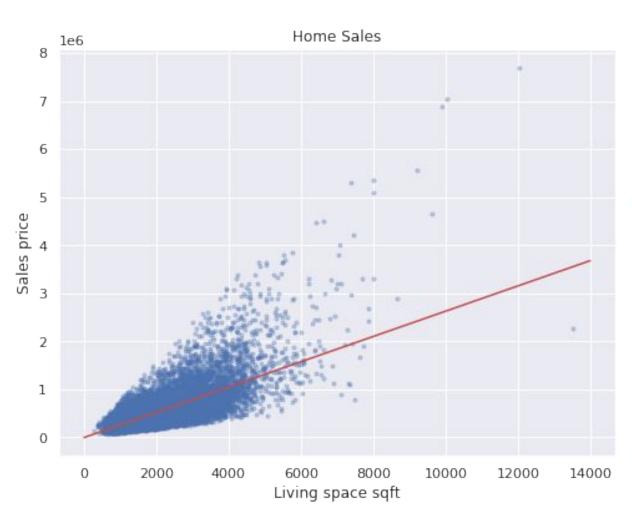
$$R^2 = 1 - \frac{RSS}{TSS}$$

### When there is no intercept

	0.493	(	ed:	quar	R-s		price		):	riable	Dep. Va
	0.493	(	ed:	quar	R-s	Adj.	OLS		l:	Mode	
	e+04	2.100	tic:	tatis	F-s		Squares	east	i: Le	etho	N
	0.00		ic):	atist	F-st	Prob (	eb <mark>202</mark> 1	25 F	: Thu,	Date	
	e+05	-3.0027	od:	eliho	Lik	Log	23:11:09		x:	Time	
	e+05	6.005	AIC:	A			21613		): 	ations	No. Observ
	e+05	6.006	BIC:	В			21611		<b>;</b> :	iduals	Df Res
							1		l:	Mode	Df
							nrobust	no	):	е Туре	Covarianc
75	0.9	[0.025		P> t	1	t	std err	5	coef		
+04	-3.5e	22e+04	-5.	.000	0	-9.899	2.690	440	58e+04	-4.3	Intercept
419	284.	276.828	1	.000	0	144.920	1.936		0.6236	28	sqft_living
		1.983	-		on:	in-Wats	Durl	490	14832	bus:	Omn
		4.709	6444	546	IB):	-Bera (	Jarque	000	0	ous):	Prob(Omni
		0.00			IB):	Prob(		824	2	kew:	S
		e+03	5.63		No.	Cond.		977	26	osis:	Kurt



### When there is no intercept



0.839	entered):	ared (unce	R-squ		price		e:	riable	Dep. Va
0.839	entered):	ared (unce	j. R-squ	Ad	OLS		l:	Mode	
1.126e+05	statistic:	F-s			quares	Least S	1:	lethod	N
0.00	tatistic):	Prob (F-s			b 2021	u, 25 Fe	: Thi	Date	
-3.0032e+05	celihood:	Log-Lik			3:09:13	23	e:	Time	
6.006e+05	AIC:				21613		S:	ations	No. Observ
6.006e+0	BIC:				21612		s:	iduals	Df Res
					1		l:	Mode	Df
					nrobust	non	e:	е Туре	Covarianc
	0.975]	[0.025	P> t	t	г	std err	coef		
	264.626	261.553	0.000	597	335.5	0.784	0892	263.	sqft_living
	80	1.9	latson:	in-W	Durb	3.334	1604	ibus:	Omn
	44	692411.8	a (JB):	-Ber	Jarque	0.000		bus):	Prob(Omni
	00	0.	ob(JB):	Pro		3.130		kew:	S
	00	1.	nd. No.	Co		0.013	3	osis:	Kurt

### Q3. How significant are the estimated coefficients?



Dep. Variable:	р	rice	R-square	ed:	0.493
Model:	(	OLS Adj.	R-square	e <mark>d:</mark>	0.493
Method:	Least Squa	ares	F-statist	tic: 2.100	e+04
Date:	Thu, 25 Feb 2	021 Prob	(F-statisti	ic):	0.00
Time:	23:1	1:09 <b>Log</b>	-Likeliho	od: -3.0027	e+05
No. Observations:	21	613	A	IC: 6.005	e+05
Df Residuals:	21	611	В	IC: 6.006	e+05
Df Model:		1			
Covariance Type:	nonrol	bust			
	coef std e	err	t P> t	[0.025	0.975
Intercept -4.358	8e+04 4402.69	90 -9.899	0.000	-5.22e+04	-3.5e+04
sqft_living 280	.6236 <b>1</b> .93	36 144.920	0.000	276.828	284.419
Omnibus:	14832.490 I	Durbin-Wats	son:	1.983	
Prob(Omnibus):	0.000 Ja	rque-Bera (	JB): 546	444.709	
Skew:	2.824	Prob(	JB):	0.00	
Kurtosis:	26.977	Cond.	No. 5	5.63e+03	

## From homoscedasticity assumption

$$\varepsilon \sim N(0, \sigma^2)$$

$$\operatorname{Var}(\hat{\beta}) = (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2$$

$$\hat{\sigma}^2 = \frac{1}{N - p - 1} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

# **Bootstrap**



### p-value

$$H_0: \beta_1 = 0$$

Alternative hypothesis

Null hypothesis

$$H_A:\beta_1\neq 0$$

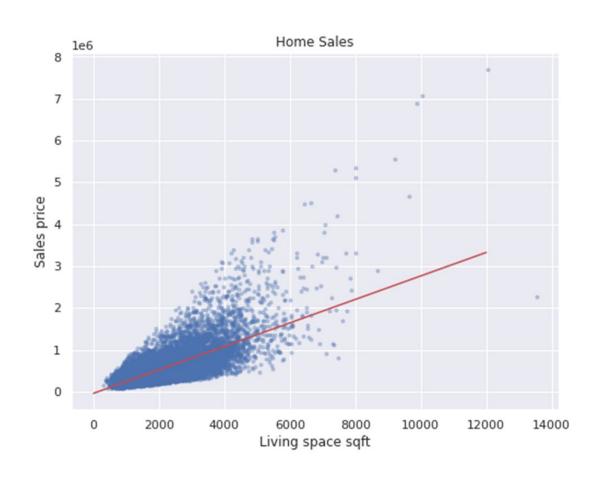
### *t*-statistics

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)}$$

*p*-value

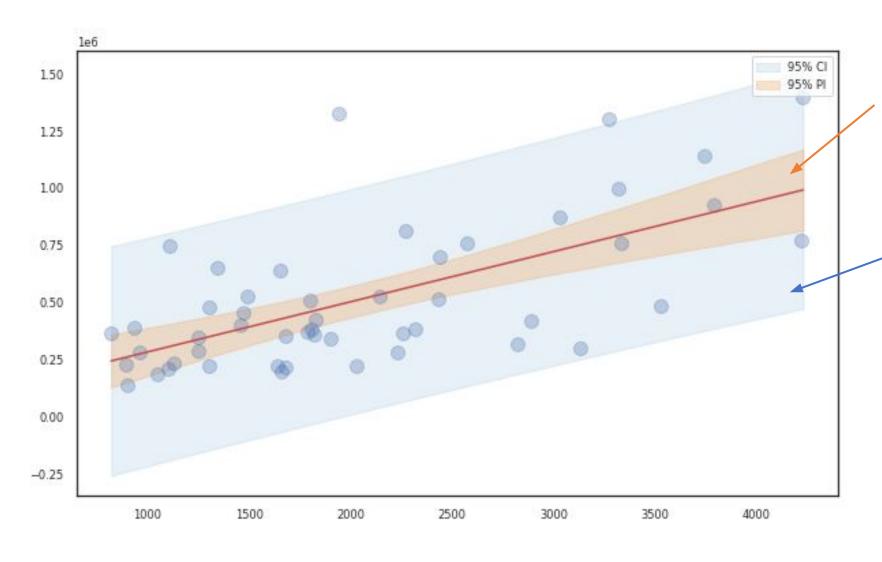
	coef	std err	t	P> t	[0.025	0.975]
Intercept	-4.358e+04	4402.690	-9.899	0.000	-5.22e+04	-3.5e+04
sqft_living	280.6236	1.936	144.920	0.000	276.828	284.419

## **Confidence Interval for Coefficients**



Dep. Variable:		price	F	R-square	ed:	0.493
Model:		OLS	Adj. F	R-square	ed:	0.493
Method:	Le	ast Squares	1	F-statist	tic: 2.10	00e+04
Date:	Thu, 2	25 Feb 2021	Prob (F	-statisti	ic):	0.00
Time:		23:11:09	Log-l	Likeliho	od: -3.002	27e+05
No. Observations:		21613		A	IC: 6.00	5e+05
Df Residuals:		21611		В	IC: 6.00	6e+05
Df Model:		1				
Covariance Type:		nonrobus	t	95%	CI for t	he coe
	coef	std err	t	P> t	[0.02	5 0.97
Intercept -4.358	e+04	4402.690	-9.899	0.000	-5.22e+04	4 -3.5e+0
sqft_living 280.	6236	1.936	144.920	0.000	276.828	3 284.41
Omnibus: 1	$\int \hat{\beta}_{1}$	-2	· SE(	$(\hat{eta}_1)$	$\hat{\beta}_1$	+2
Skew:	2.8	324	Prob(J	B):	0.00	
Kurtosis:	26.9	977	Cond. N	lo. 5	5.63e+03	

## **Confidence Intervals for Regression**



95% Confidence Interval (for the regression line)

95% Prediction Interval (for sample points)

### Q4. How well does the model predict on unseen data?

### Popular Error metrics

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Mean Percent Absolute Error (MAPE)

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Mean Squared Error (MSE)

MSE = 
$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Dataset	R2	MSE	MAPE
Train (80%)	0.492	6.632 E10	0.3598
Test (20%)	0.494	7.648 E10	0.3570

### Summary

- 1. How do we determine the coefficients?
- 2. How well does the model fit?

- 3. How significant are the coefficients?
- 4. How well does the model predict on unseen data?