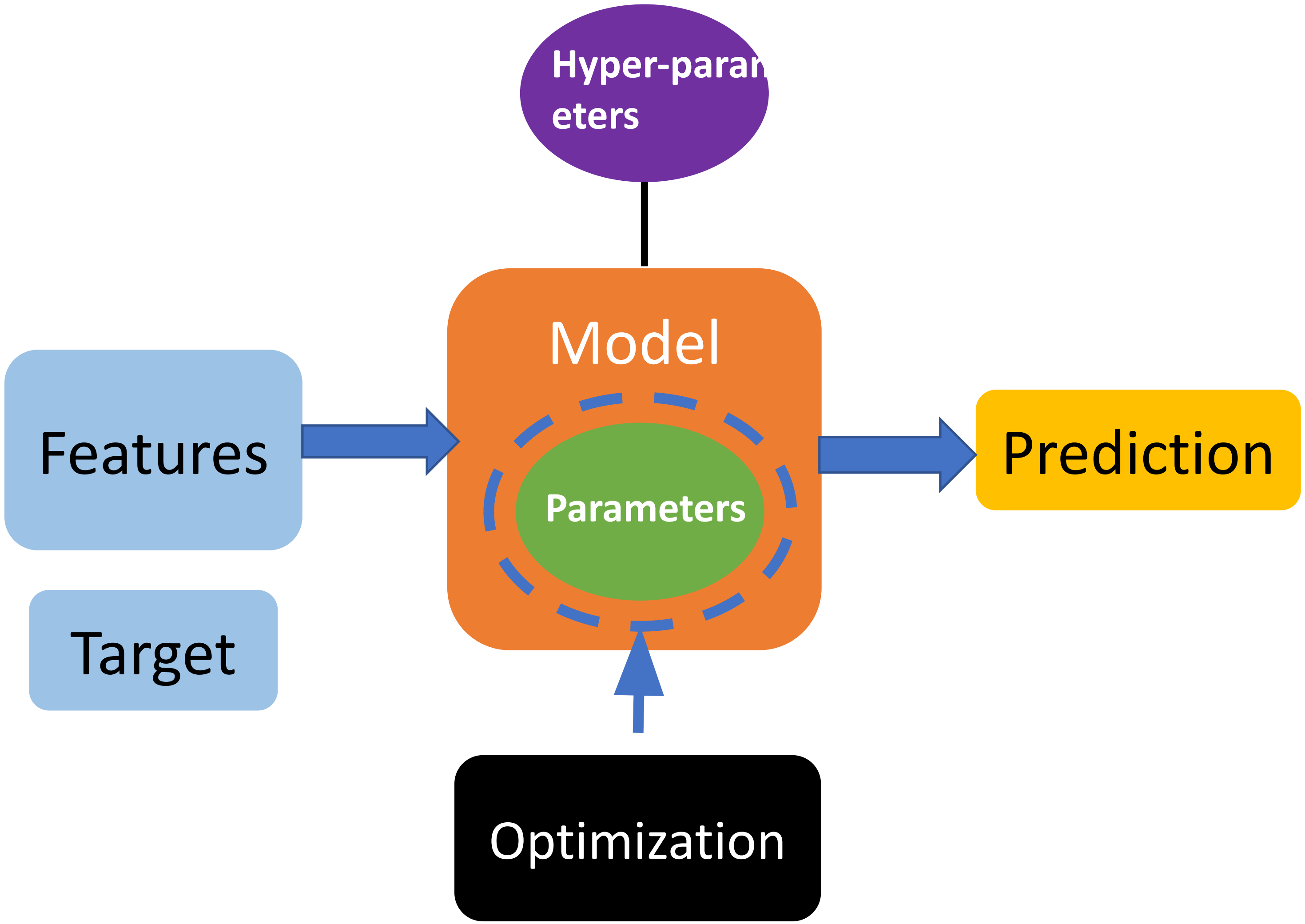
Linear Regression

How Supervised Learning Works



What is Linear Regression

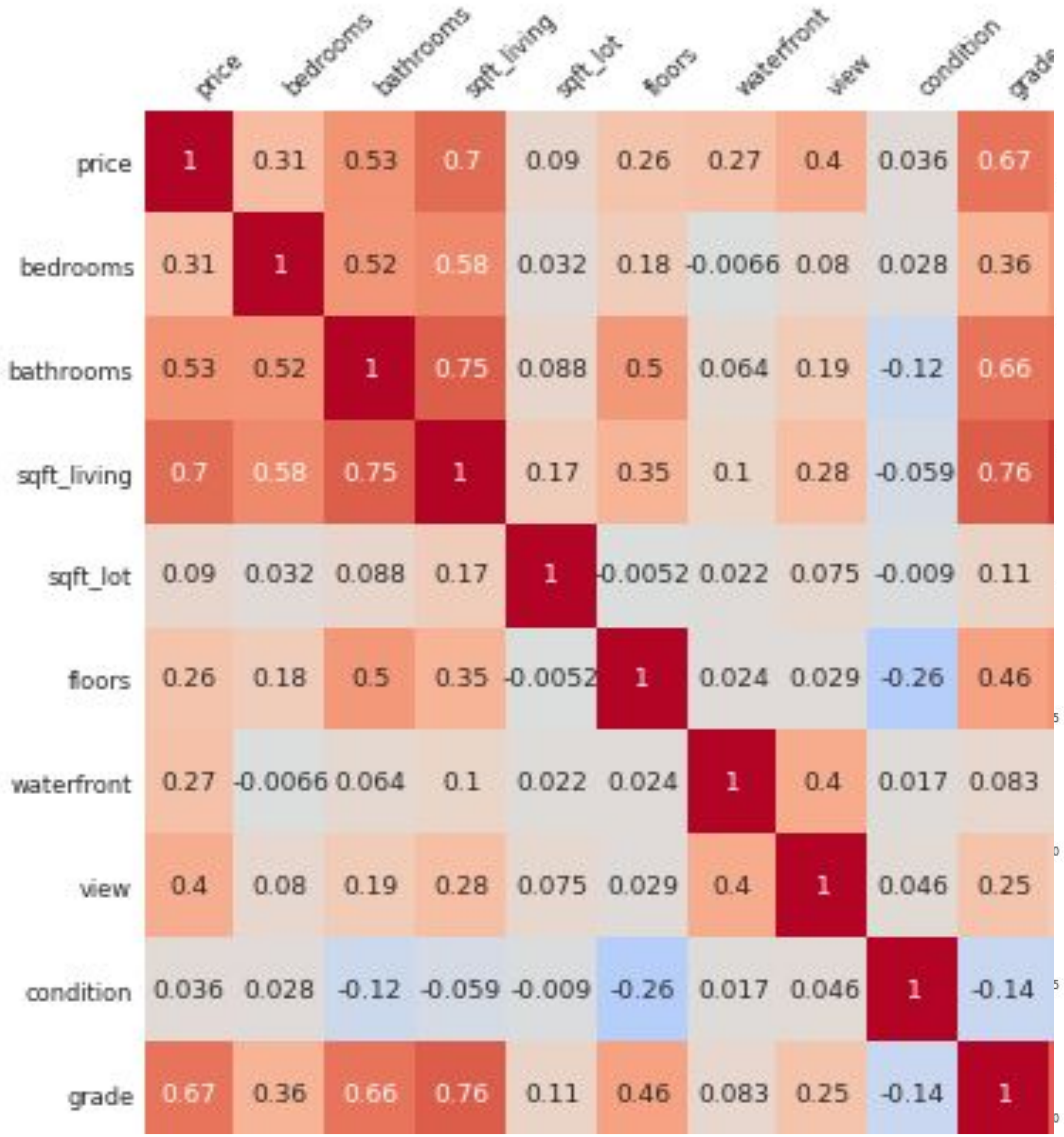
- Supervised learning model
- Predictive task- real valued numbers
- Parametric model
- No hyperparameters
- Features have linear relationship to the target variable

Example

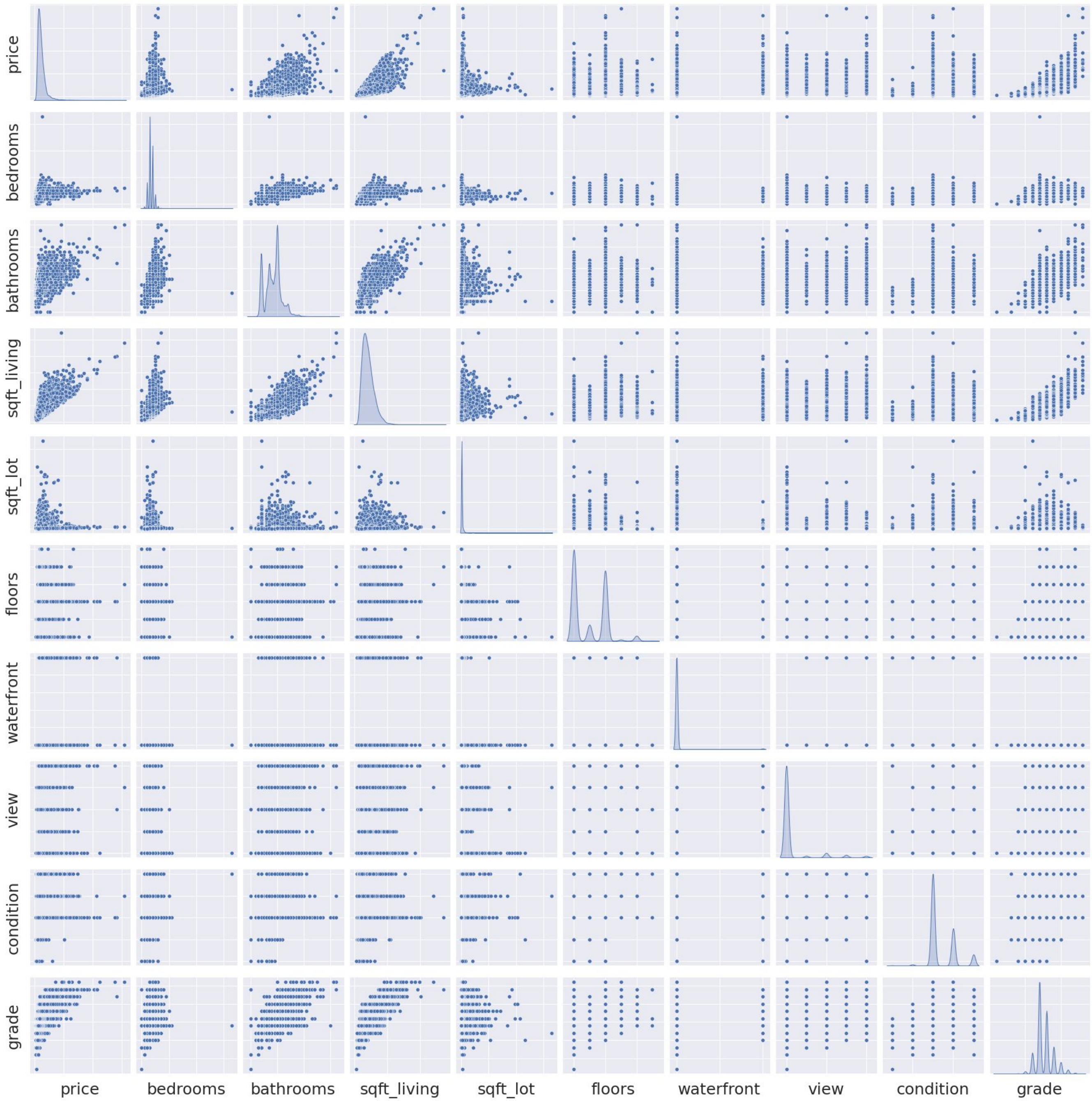
An example using House sales data from Kaggle
https://www.kaggle.com/harlfoxem/housesalesprediction/download

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	...	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



price
bedrooms
bathrooms
sqft_living
sqft_lot
floors
waterfront
view
condition
grade

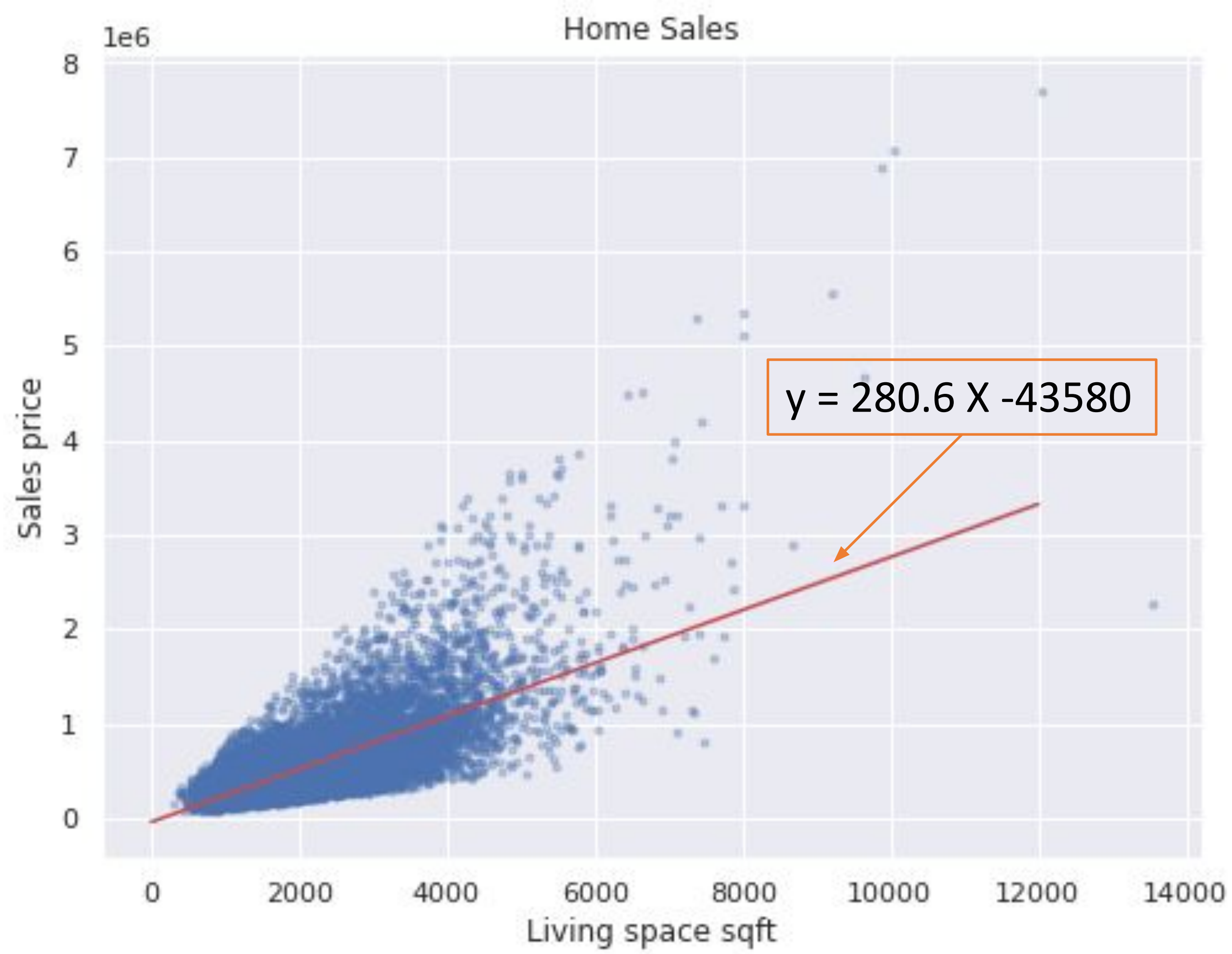
Univariate Linear Regression

Intercept Slope Residual

$$Y = \underbrace{\beta_0}_{\text{Intercept}} + \underbrace{\beta_1}_{\text{Slope}} X + \underbrace{\epsilon}_{\text{Residual}}$$

Coefficients, or Parameters

Using `statsmodel`’s `OLS` (ordinary least squares) package



OLS Regression Results

Dep. Variable:	price	R-squared:	0.493
Model:	OLS	Adj. R-squared:	0.493
Method:	Least Squares	F-statistic:	2.100e+04
Date:	Thu, 25 Feb 2021	Prob (F-statistic):	0.00
Time:	23:11:09	Log-Likelihood:	-3.0027e+05
No. Observations:	21613	AIC:	6.005e+05
Df Residuals:	21611	BIC:	6.006e+05
Df Model:	1		
Covariance Type:	nonrobust		

	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

Omnibus:	14832.4	Durbin-Watson:	1.983
Prob(Omnibus):	0.0	Jarque-Bera (JB):	546444.709
Skew:	2.8	Prob(JB):	0.00
Kurtosis:	26.9	Cond. No.	5.63e+03

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?

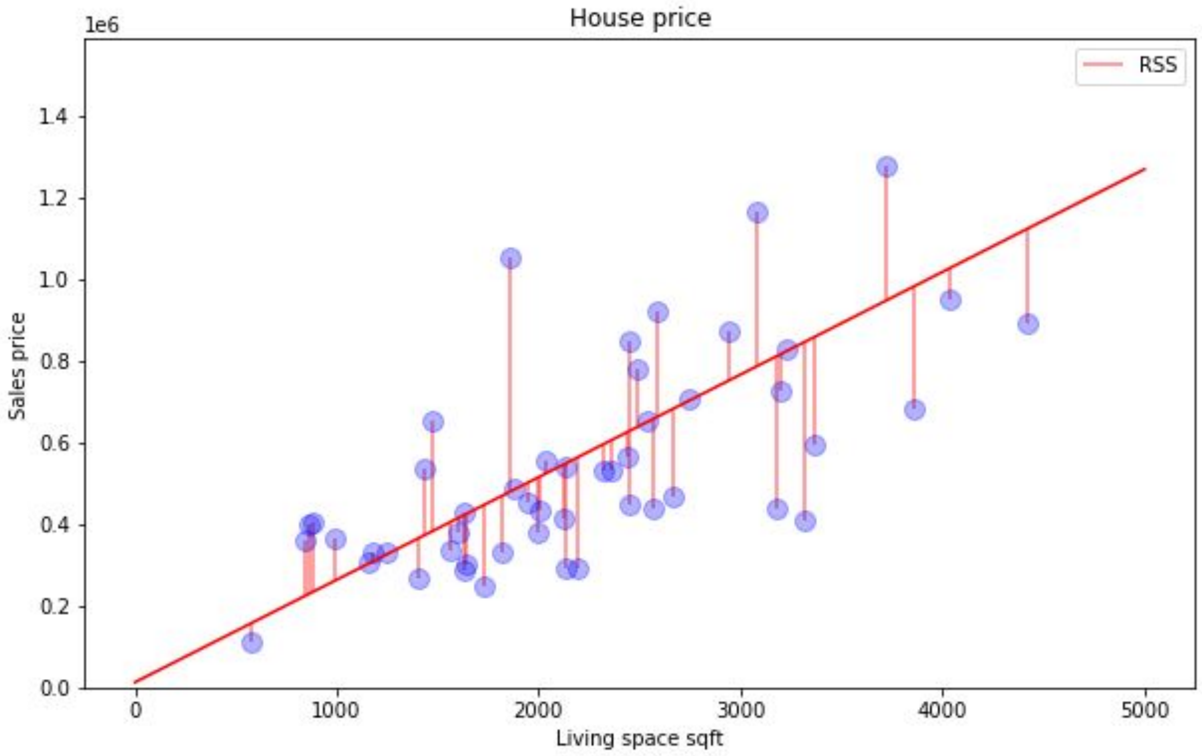
Q1. How do we find the coefficients?

$$Y_i = \overbrace{\boxed{\beta_0 + \beta_1 X_i}}^{\hat{y}_i} + \epsilon_i$$

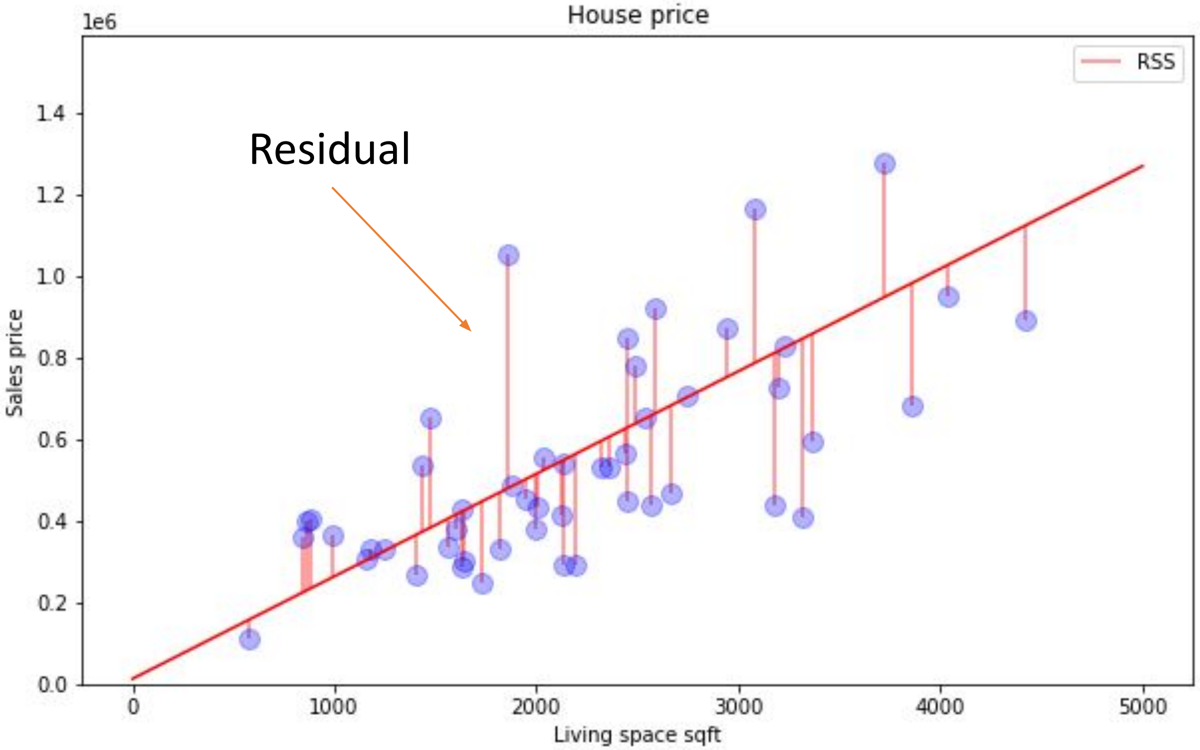
Coefficients

Residual

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



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



Mean Absolute Error (MAE)

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

Mean Percent Absolute Error (MAPE)

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

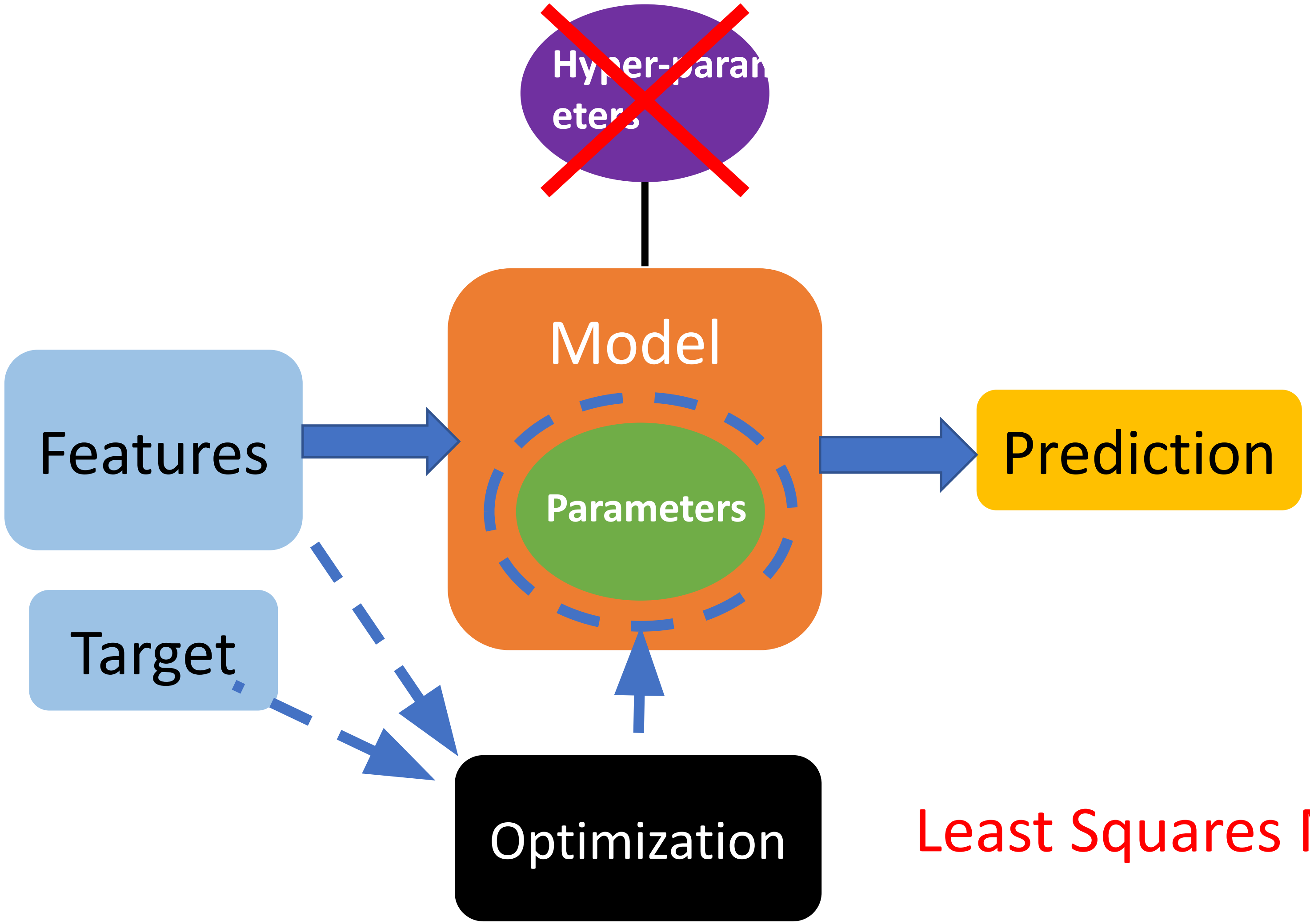
Mean Squared Error (MSE)

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

Root Mean Squared Error (RMSE)

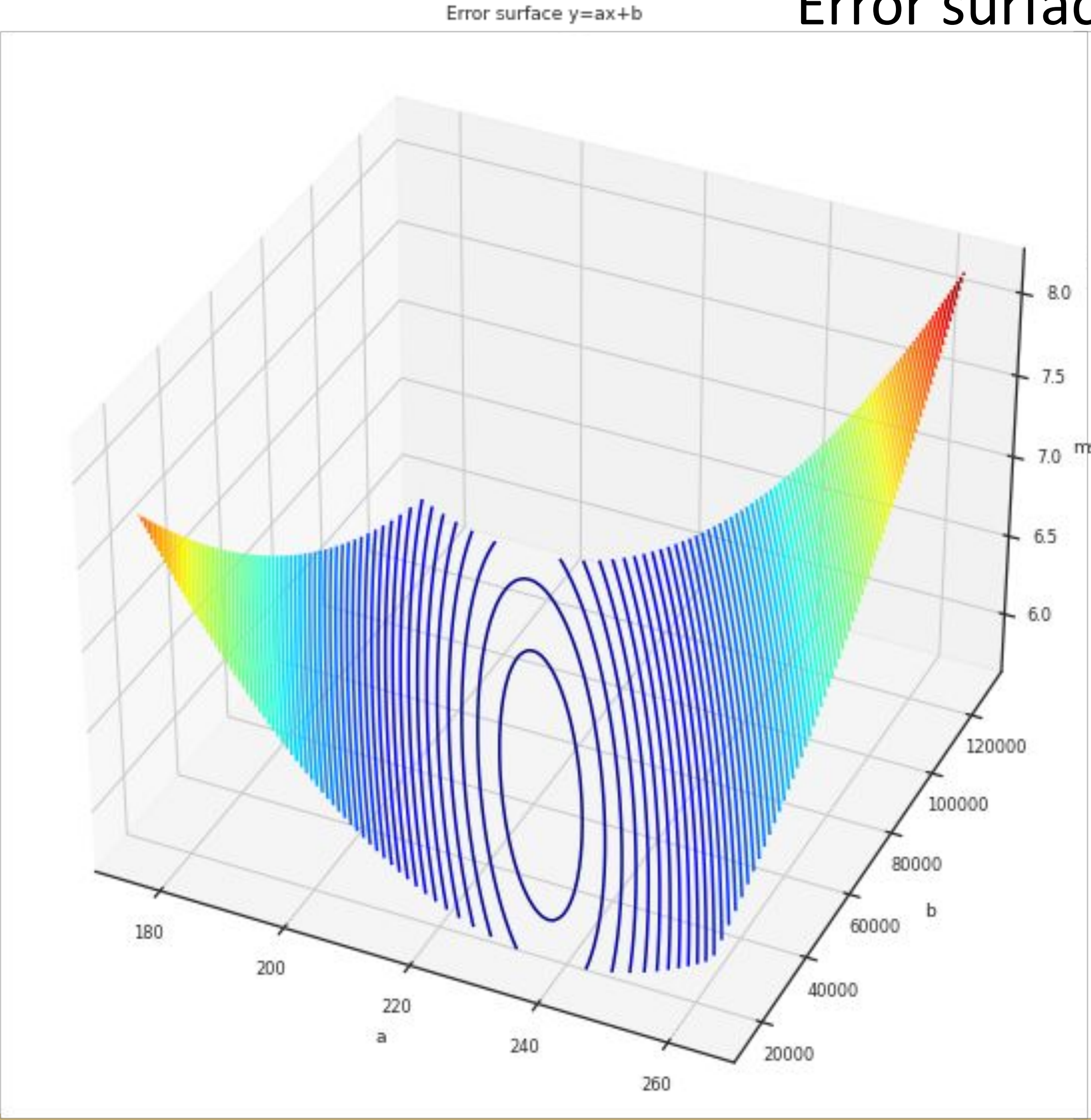
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

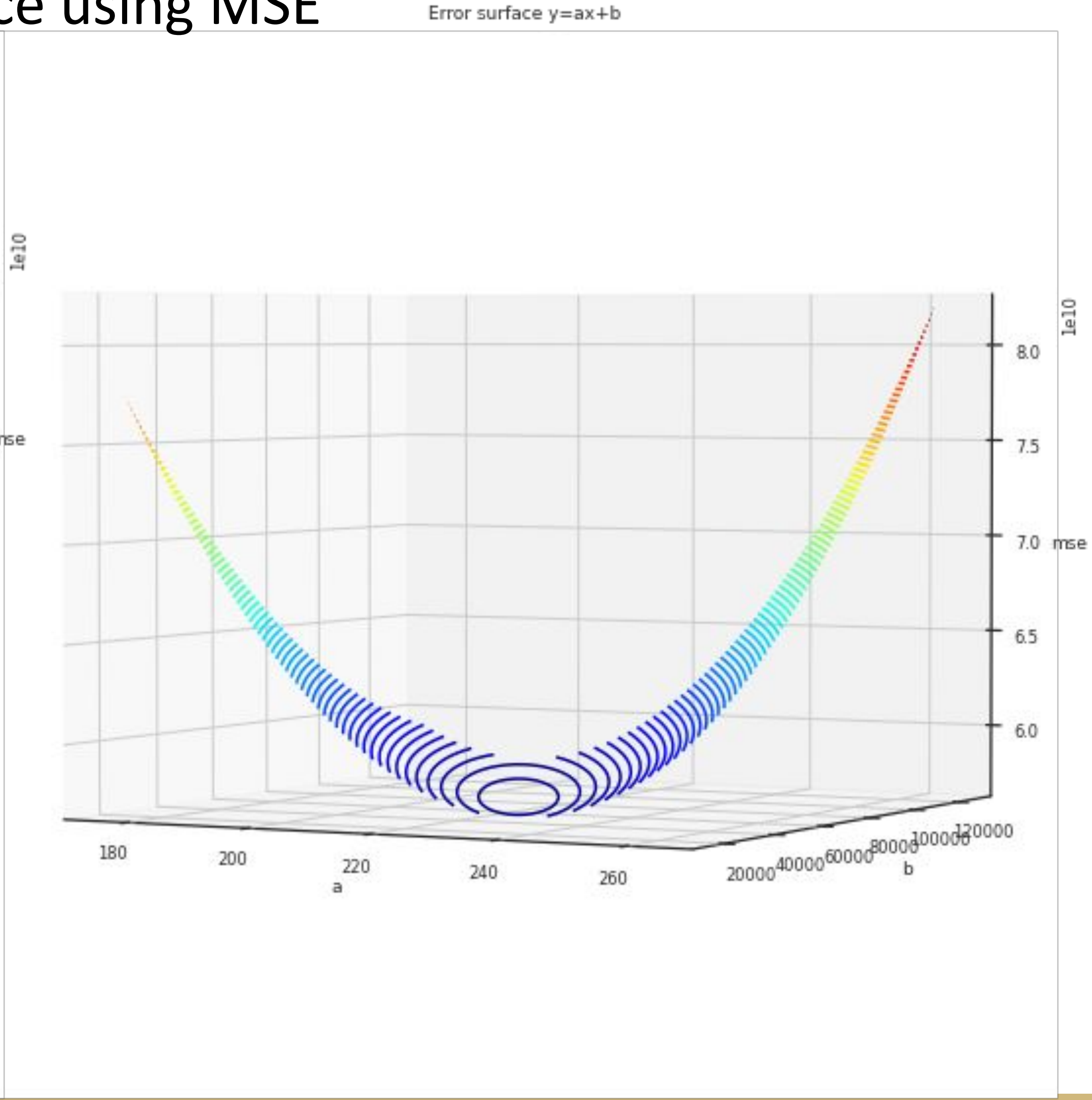
Optimization in Linear Regression



Least Squares Method

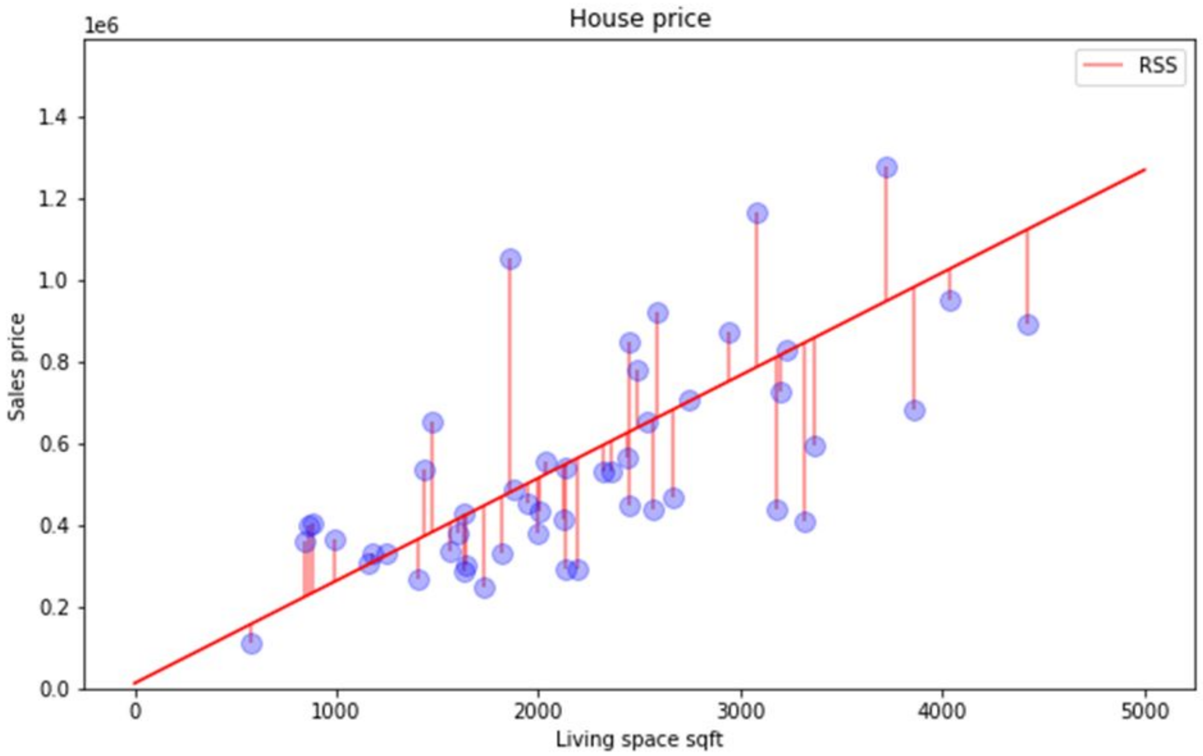
Error surface using MSE





Least Squares Method

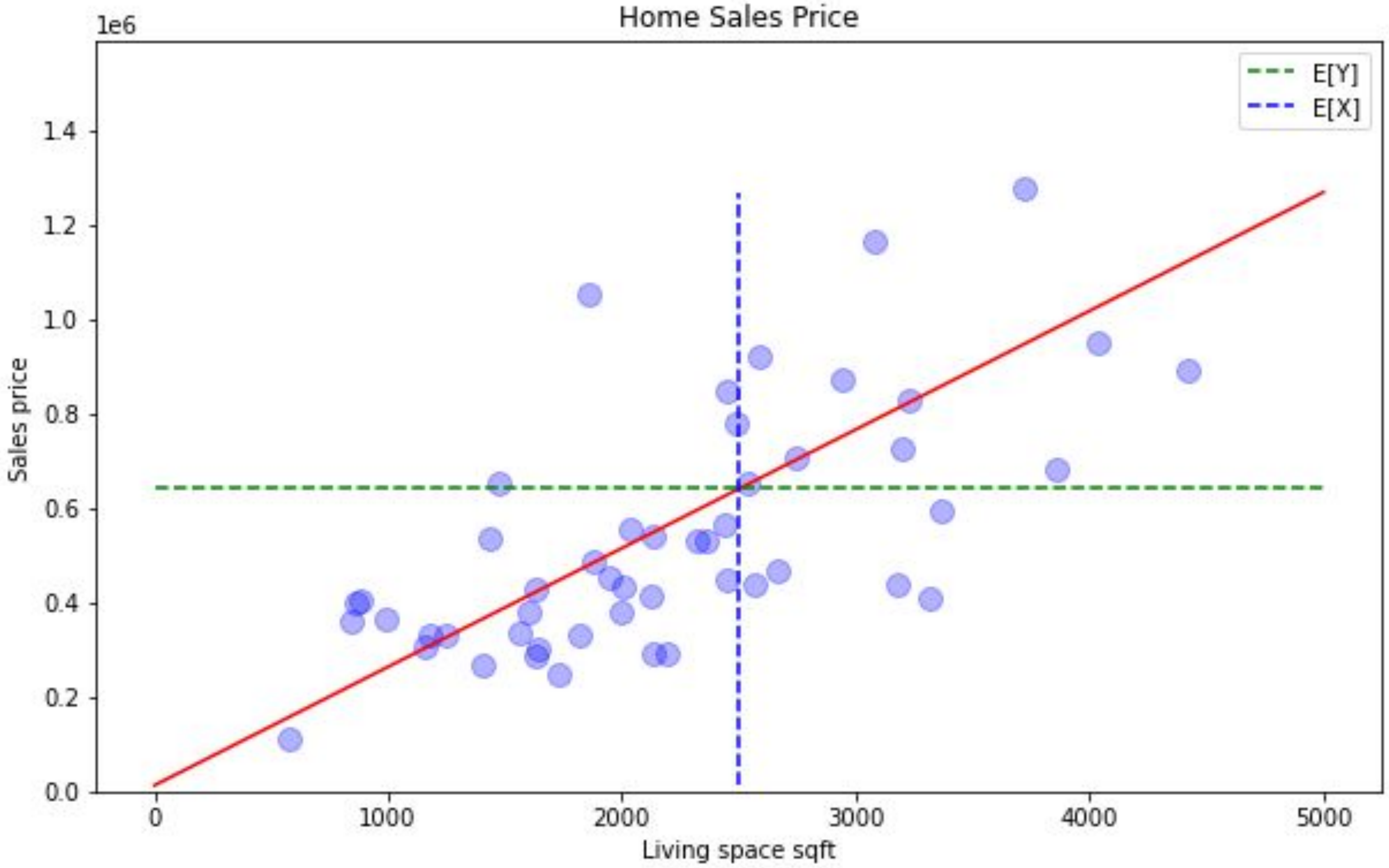
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



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

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

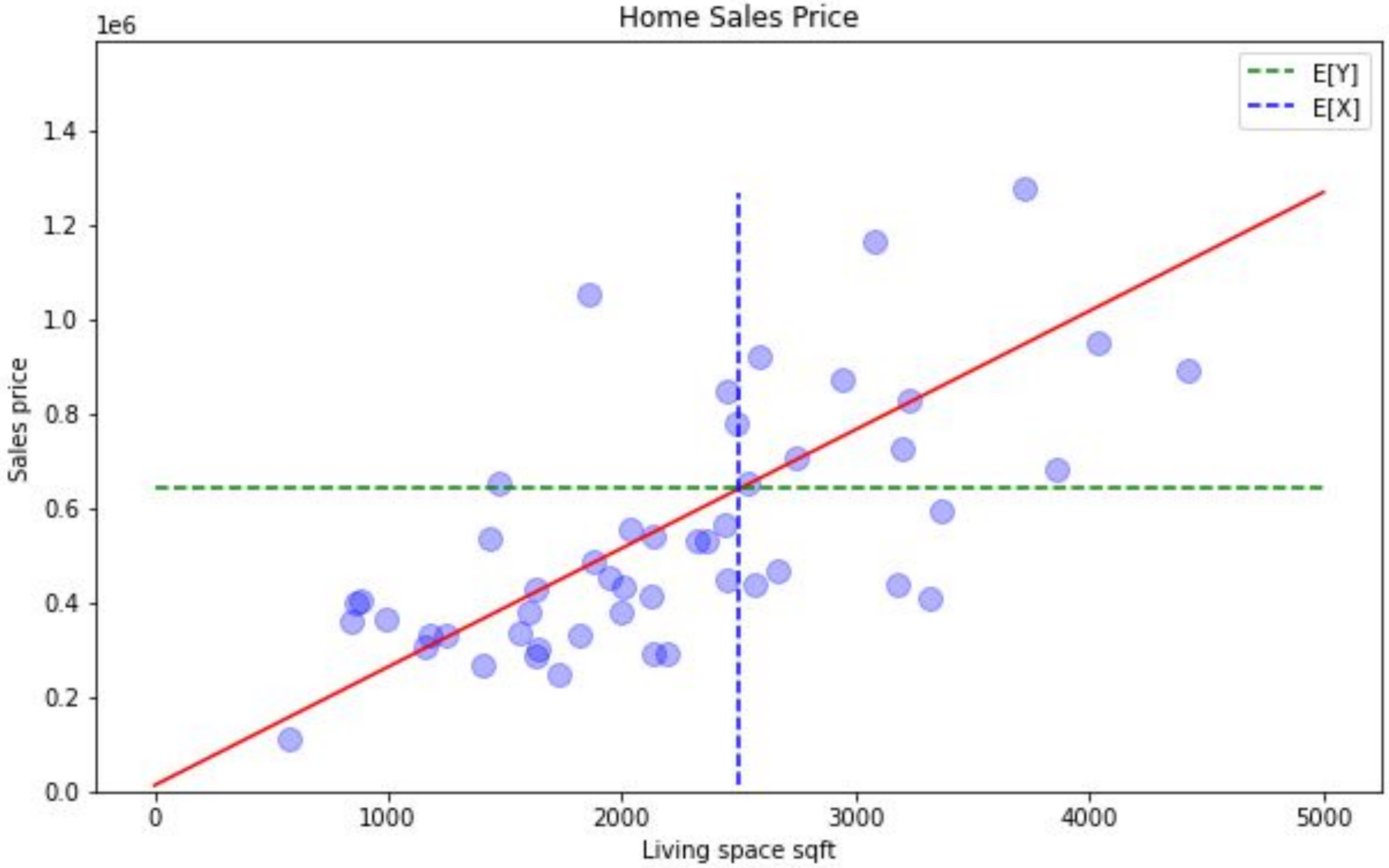
Least Squares Method



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

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

What happens when scaling variables?



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

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

Least Squares Method in Multivariate case

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

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

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

Q2. How well does the model fit?

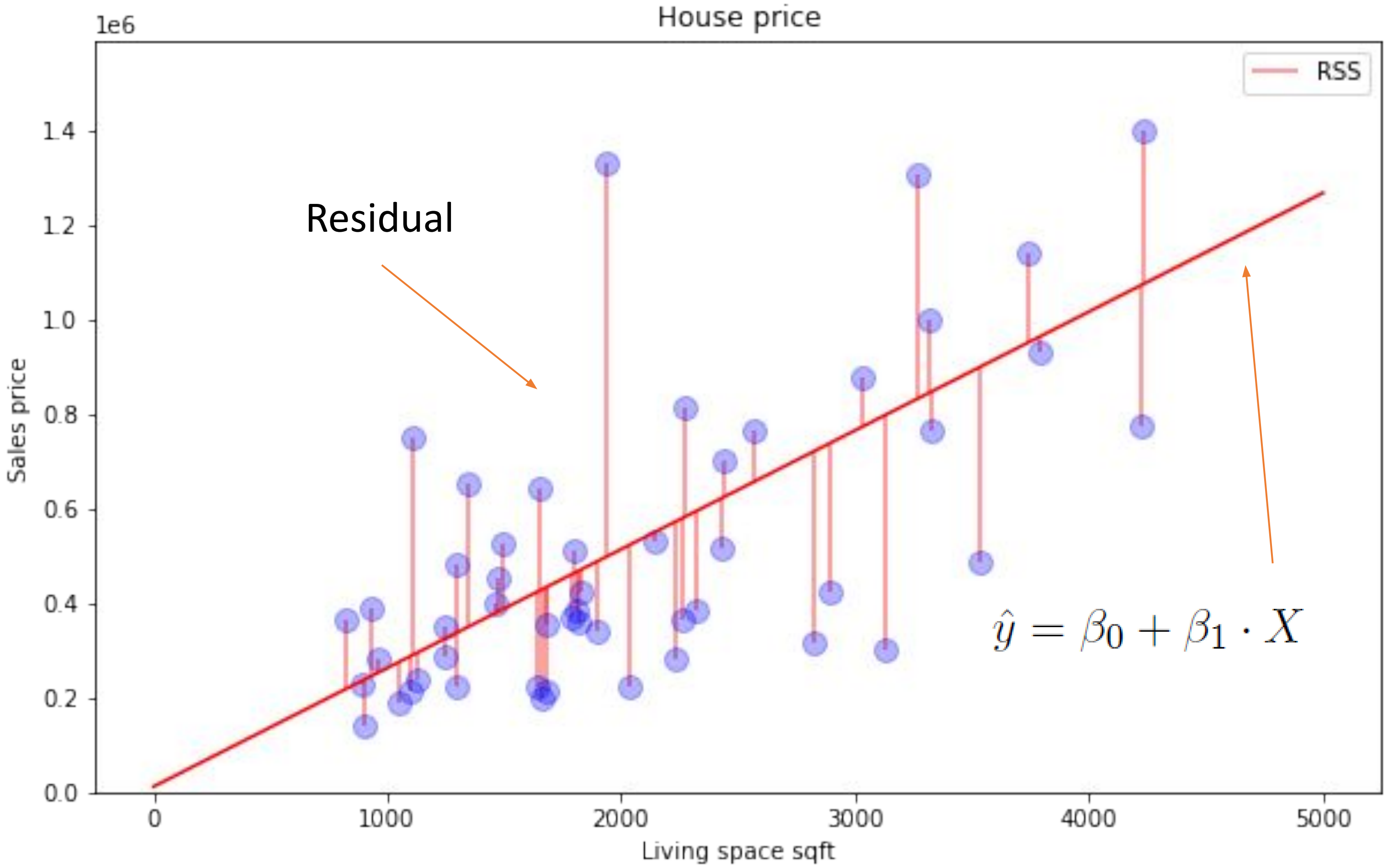
OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.493
Model:	OLS	**Adj. R-squared:**	0.493
Method:	Least Squares	**F-statistic:**	2.100e+04
Date:	Thu, 25 Feb 2021	**Prob (F-statistic):**	0.00
Time:	23:11:09	**Log-Likelihood:**	-3.0027e+05
No. Observations:	21613	**AIC:**	6.005e+05
Df Residuals:	21611	**BIC:**	6.006e+05
Df Model:	1		
Covariance Type:	nonrobust		

	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

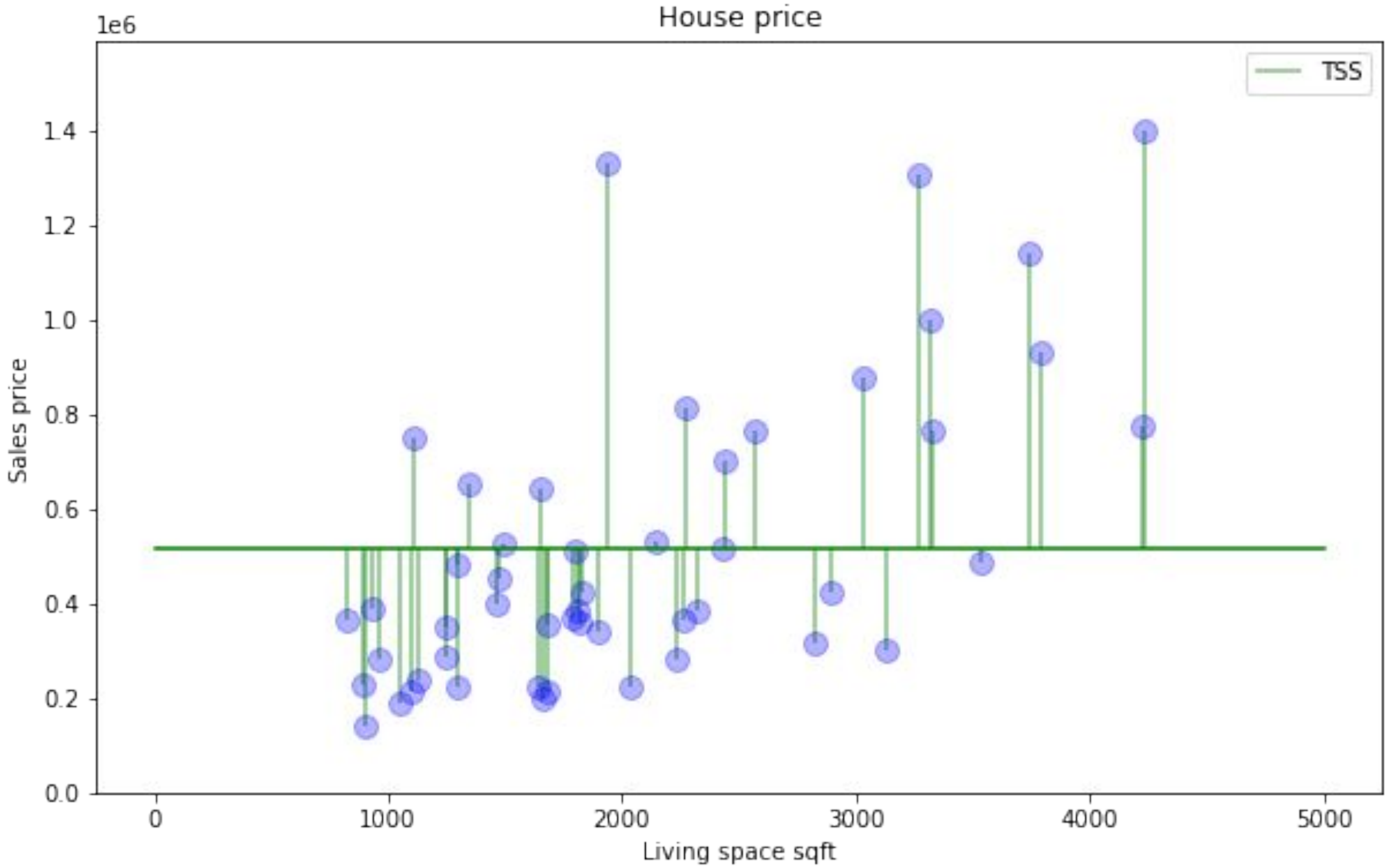
Omnibus:	14832.490	**Durbin-Watson:**	1.983
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	546444.709
Skew:	2.824	**Prob(JB):**	0.00
Kurtosis:	26.977	**Cond. No.**	5.63e+03

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



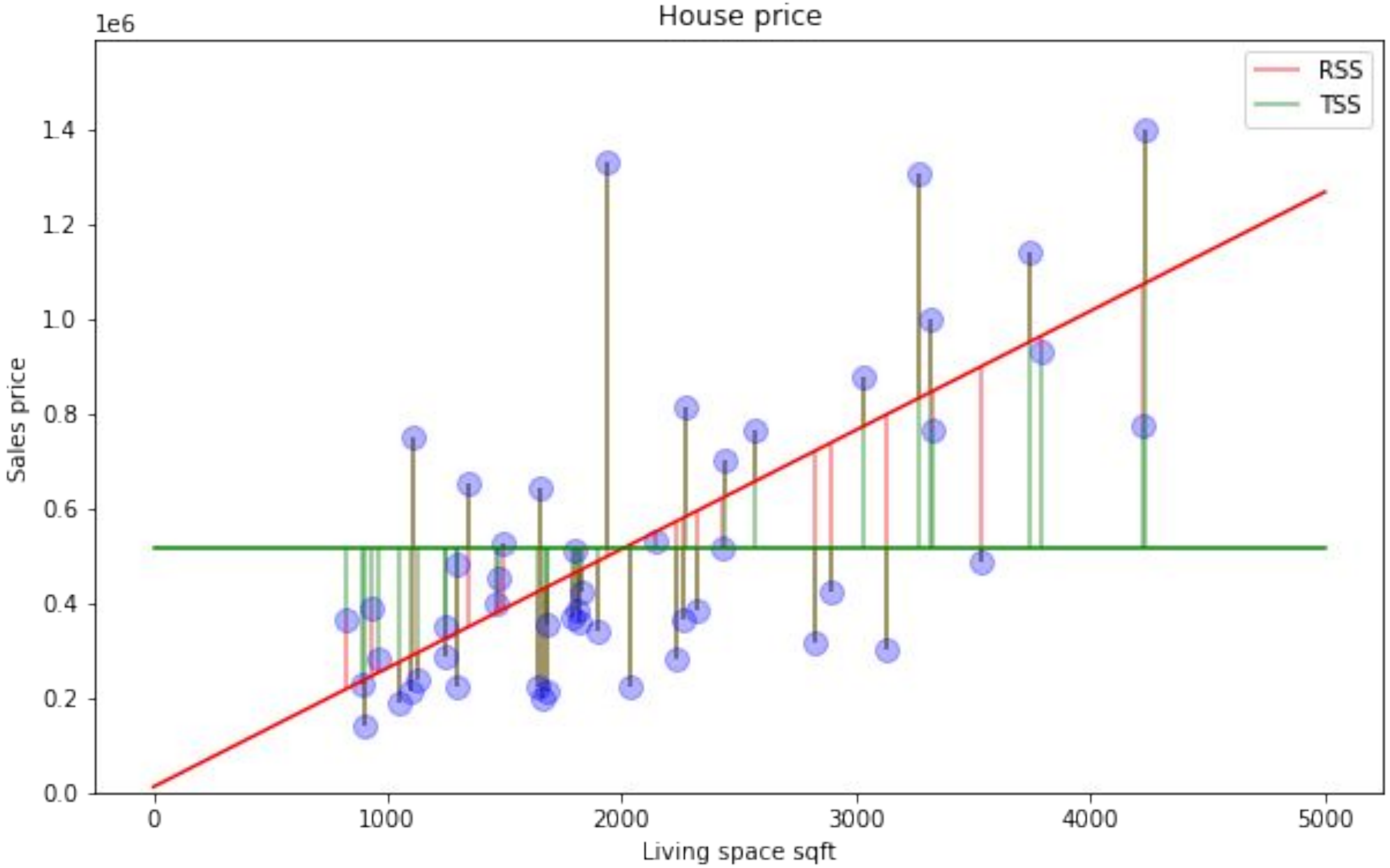
$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

"Residual Sum of Squares"



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

“Total Sum of Squares”



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

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

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

When there is no intercept

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.493
Model:	OLS	**Adj. R-squared:**	0.493
Method:	Least Squares	**F-statistic:**	2.100e+04
Date:	Thu, 25 Feb 2021	**Prob (F-statistic):**	0.00
Time:	23:11:09	**Log-Likelihood:**	-3.0027e+05
No. Observations:	21613	**AIC:**	6.005e+05
Df Residuals:	21611	**BIC:**	6.006e+05
Df Model:	1		
Covariance Type:	nonrobust		

	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

Omnibus:	14832.490	**Durbin-Watson:**	1.983
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	546444.709
Skew:	2.824	**Prob(JB):**	0.00
Kurtosis:	26.977	**Cond. No.**	5.63e+03

Kurtosis:	30.013	**Cond. No.**	1.00

When there is no intercept

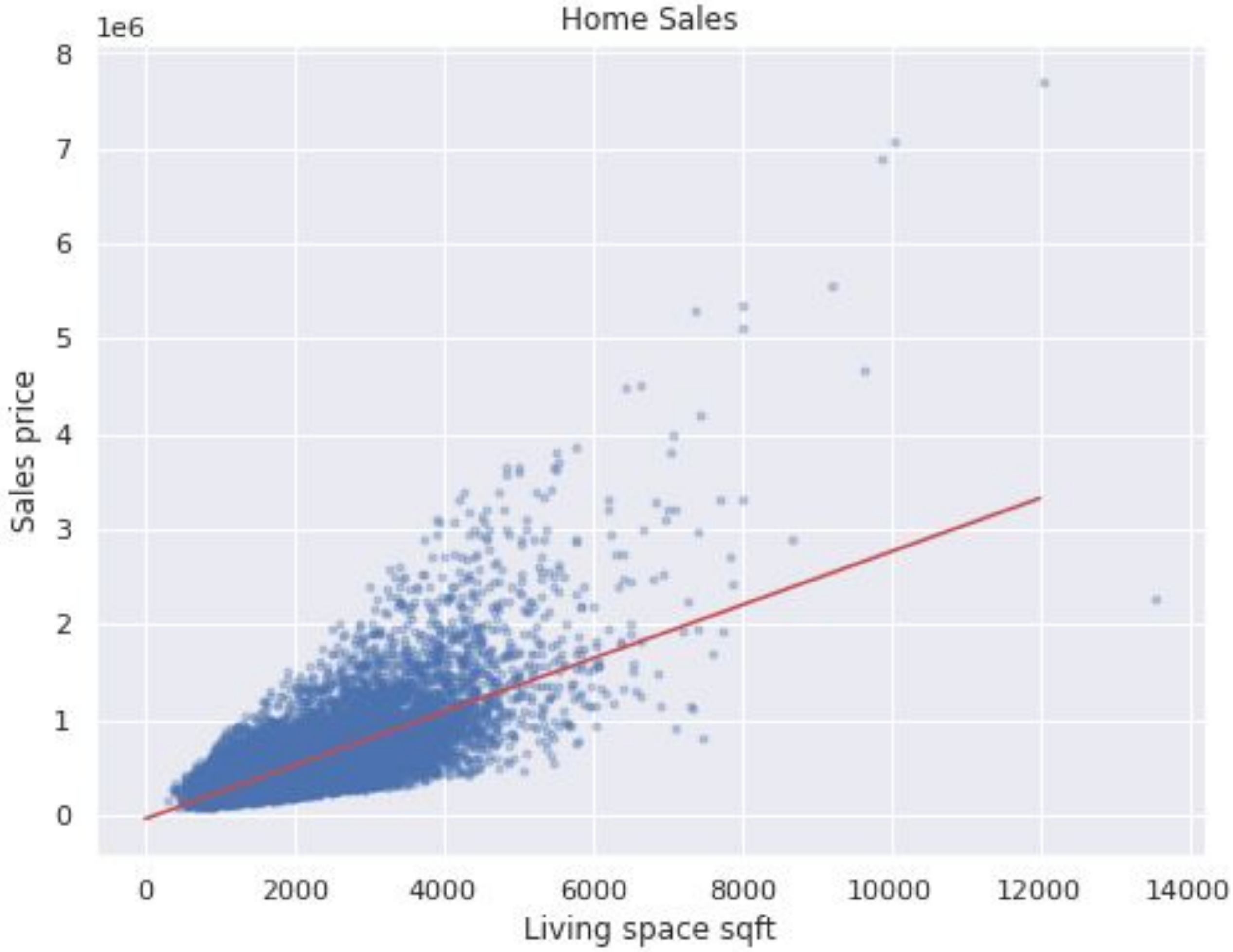
OLS Regression Results

Dep. Variable:	price	**R-squared (uncentered):**	0.839
Model:	OLS	**Adj. R-squared (uncentered):**	0.839
Method:	Least Squares	**F-statistic:**	1.126e+05
Date:	Thu, 25 Feb 2021	**Prob (F-statistic):**	0.00
Time:	23:09:13	**Log-Likelihood:**	-3.0032e+05
No. Observations:	21613	**AIC:**	6.006e+05
Df Residuals:	21612	**BIC:**	6.006e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
sqft_living	263.0892	0.784	335.597	0.000	261.553	264.626

Omnibus:	16043.334	**Durbin-Watson:**	1.980
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	692411.844
Skew:	3.130	**Prob(JB):**	0.00
Kurtosis:	30.013	**Cond. No.**	1.00

Q3. How significant are the estimated coefficients?



OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.493
Model:	OLS	**Adj. R-squared:**	0.493
Method:	Least Squares	**F-statistic:**	2.100e+04
Date:	Thu, 25 Feb 2021	**Prob (F-statistic):**	0.00
Time:	23:11:09	**Log-Likelihood:**	-3.0027e+05
No. Observations:	21613	**AIC:**	6.005e+05
Df Residuals:	21611	**BIC:**	6.006e+05
Df Model:	1		
Covariance Type:	nonrobust		

	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

Omnibus:	14832.490	**Durbin-Watson:**	1.983
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	546444.709
Skew:	2.824	**Prob(JB):**	0.00
Kurtosis:	26.977	**Cond. No.**	5.63e+03

From homoscedasticity assumption

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

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

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

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

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

Bootstrap

p-value

Null hypothesis $H_0 : \beta_1 = 0$

Alternative hypothesis $H_A : \beta_1 \neq 0$

t-statistics

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

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

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.493
Model:	OLS	**Adj. R-squared:**	0.493
Method:	Least Squares	**F-statistic:**	2.100e+04
Date:	Thu, 25 Feb 2021	**Prob (F-statistic):**	0.00
Time:	23:11:09	**Log-Likelihood:**	-3.0027e+05
No. Observations:	21613	**AIC:**	6.005e+05
Df Residuals:	21611	**BIC:**	6.006e+05
Df Model:	1		
Covariance Type:	nonrobust		

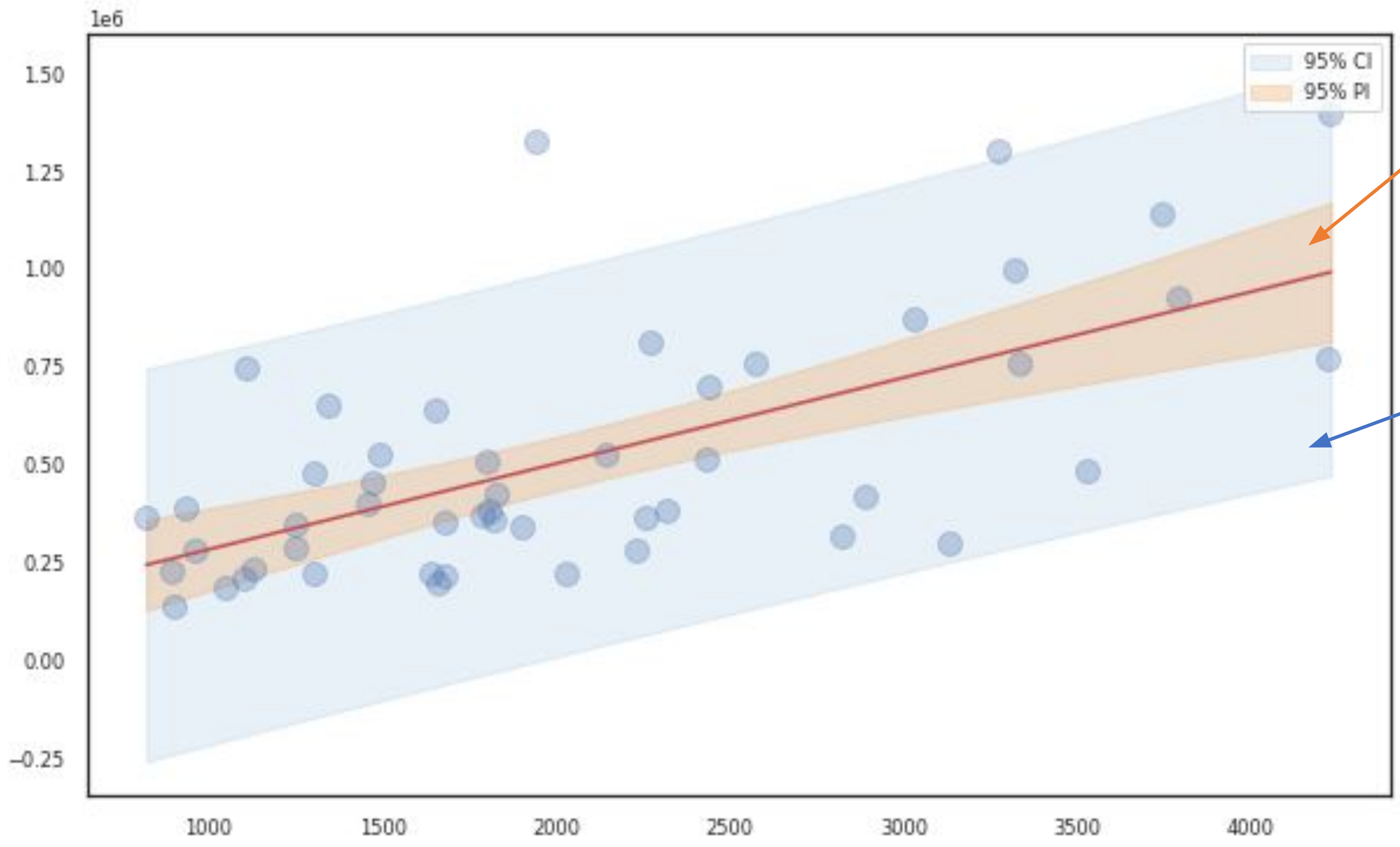
95% CI for the coefficients

	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

$$\left[\hat{\beta}_1 - 2 \cdot \boxed{\mathrm{SE}(\hat{\beta}_1)},\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

Omnibus:	1		
Prob(Omnibus):			
Skew:	2.824	**Prob(JB):**	0.00
Kurtosis:	26.977	**Cond. No.**	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)

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

Mean Percent Absolute Error (MAPE)

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

Mean Squared Error (MSE)

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

Root Mean Squared Error (RMSE)

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

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?