LINEAR REGRESSION

Review

https://ai.lange-analytics.com/

WHAT WILL YOU LEARN/REVIEW

- Reviewing the basic idea behind linear regression
- Learning how to measure predictive quality with Mean Square Error (MSE).
- Calculating optimal OLS regression parameters using `tidymodels`
- Distinguish between unfitted and fitted models
- How to interpret the OLS regression parameters and their significance
- Using metrics to evaluate prediction quality on the testing

LOADING THE LIBRARIES AND THE DATA

▶ Code

```
     Price Sqft Bedrooms Condition
1 523633.4 2040        4         3
2 530960.7 2120        4         3
3 523466.8 2130        4         4
4 759747.7 3330        4         3
5 546377.8 2440        4         3
6 186536.6  900        3         4
```

SPLITTING IN TRAINING AND TESTING DATA:

```r
set.seed(Seed)
Split7030=initial_split(DataHouses,prop=0.7, strata = Price)
DataTrain=training(Split7030)
DataTest=testing(Split7030)
```

HOW MUCH IS A HOUSE WORTH IN KING COUNTY?

A house with average properties should be predicted with an average price!

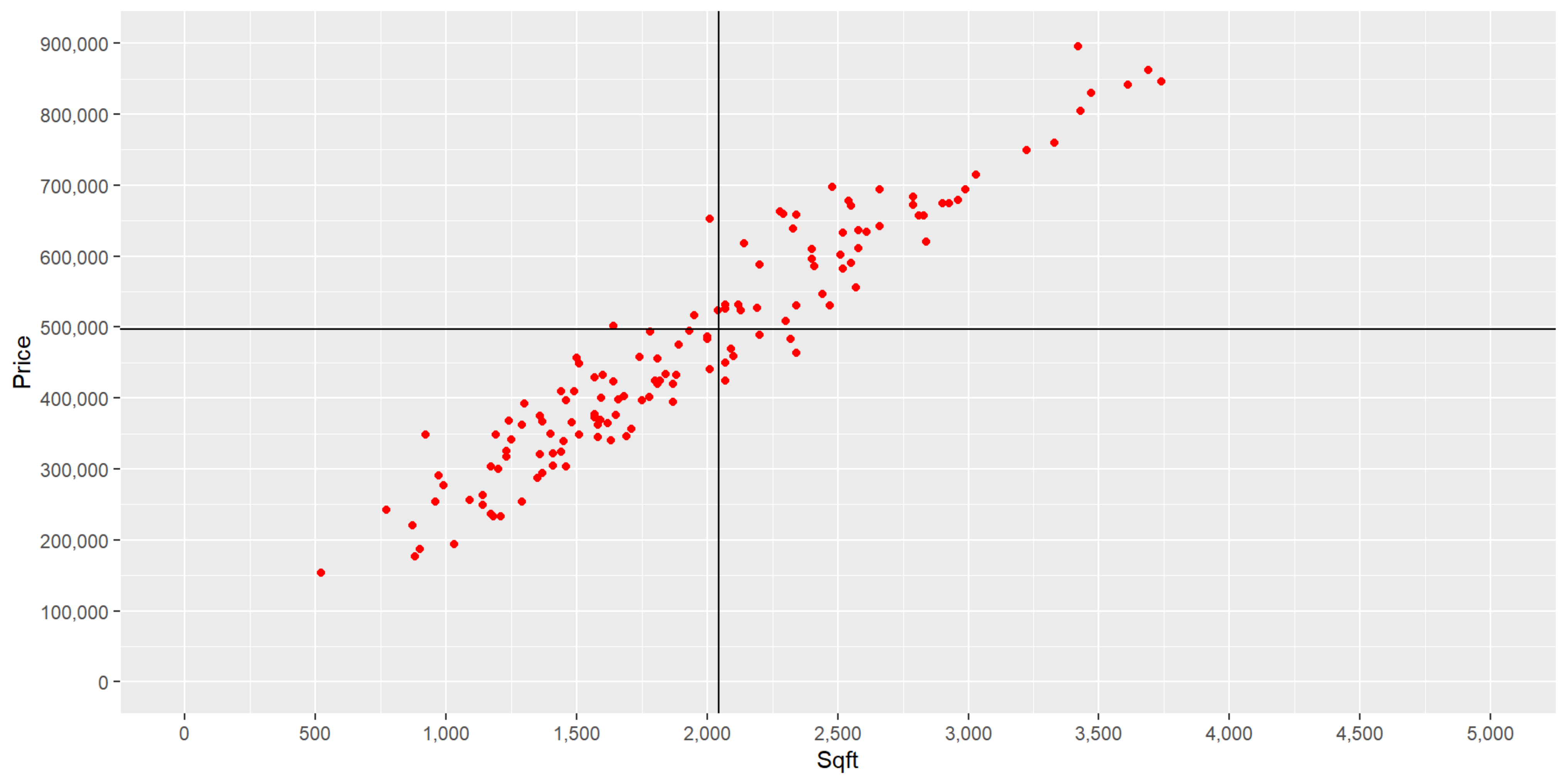
▶ Code

```
The mean square footage of a house in King county is: 2044.319
```

▶ Code

```
The mean price of a house in King county is: 497414.9
```

PREDICTING THE PRICE OF AN AVERAGE SIZED HOUSE AS THE AVERAGE OF ALL HOUSE PRICES



https://ai.lange-analytics.com/

HOW TO MEASURE PREDICTION QUALITY WITH THE MEAN SQUARED ERROR (MSE)

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

$$\iff$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\underbrace{\overbrace{\beta_1 x_i + \beta_0}^{\text{Prediction } i} - y_i}_{\text{Error } i})^2$$

Note, when the data are given (i.e., x_i and y_i are given), the MSE depends only on the choice of β_1 and β_0 »

INCLUDING SQFT AS DETERMINAT OF PRICE

PREPARING THE DATA

Blueprint for the data:

```r
RecipeHouses=recipe(Price~Sqft, data=DataTrain)
```

CHOOSING THE MODEL BLUEPRINT

Blueprint for the model:

```
ModelDesignOLS=linear_reg() |>
            set_engine("lm") |>
            set_mode("regression")
```

HOW DOES THE UNFITTED MODEL LOOKS LIKE?

$$\underbrace{Price}_{y} = \underbrace{\beta_1}_{m} \underbrace{Sqft}_{x} + \underbrace{\beta_0}_{b}$$

USING A WORKFLOW TO FIT THE MODEL TO THE DATA (FINDING THE OPTIMAL β_1 AND β_0 VALUES

$$\underbrace{Price}_{y} = \underbrace{\beta_1}_{m}\underbrace{Sqft}_{x} + \underbrace{\beta_0}_{b}$$

```r
WFModelHouses=workflow() |>
              add_recipe(RecipeHouses) |>
              add_model(ModelDesignOLS) |>
              fit(DataTrain)
```

UNFITTED MODEL VS FITTED WORKFLOW MODEL

Unfitted Model:

$$\underbrace{Price}_{y} = \underbrace{\beta_1}_{m} \underbrace{Sqft}_{x} + \underbrace{\beta_0}_{b}$$

▶ Code

```
# A tibble: 2 × 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    6584.  12191.       0.540 5.90e- 1
2 Sqft            238.      5.47    43.5   6.43e-66
```

Fitted Model:

$$\underbrace{Price}_{y} = \underbrace{238}_{m} \cdot \underbrace{Sqft}_{x} + \underbrace{6584}_{b}$$

Predict the price for a house with 1,000 sqft and send it to me in a private chat!

INTERPRETATION AND SIGNIFICANCE

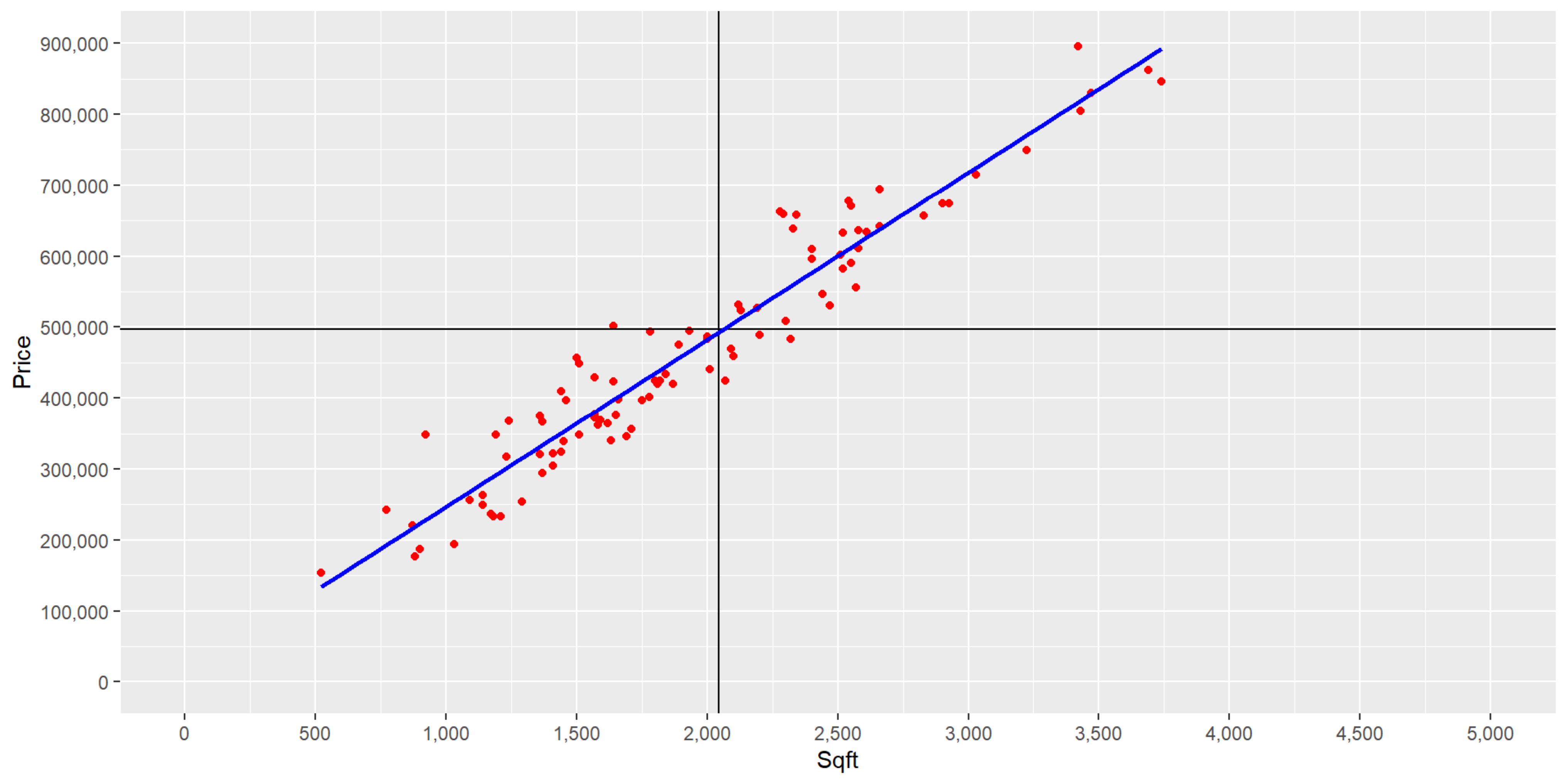
```
# A tibble: 2 × 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    6584.  12191.       0.540 5.90e- 1
2 Sqft            238.      5.47    43.5   6.43e-66
```

$$\begin{aligned}\widehat{Price} &= 238 \cdot Sqft + 6584 \\ (+238) &= 238 \cdot (+1) + (+0) \\ (+476) &= 238 \cdot (+2) + (+0) \\ (+714) &= 238 \cdot (+3) + (+0)\end{aligned}$$

For each extra $Sqft$ the predicted price increases by $238

The variable $Sqft$ is significant. I.e., the probability that the related coefficient β_1 equals zero is extremely small.

HOW DOES THE FITTED MODEL THAT CONSIDERS SQFT IMPROVES THE PREDICTION COMPARED TO A SIMPLE AVERAGE



EVALUATING PREDICTIVE QUALITY WITH THE TESTING DATASET

```r
DataTestWithPred=augment(WFModelHouses, new_data=DataTest)
metrics(DataTestWithPred, truth=Price, estimate=.pred)
```

```
# A tibble: 3 × 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 rmse    standard   50915.
2 rsq     standard       0.923
3 mae     standard   39748.
```

PROJECT: ANALYSIS WITH ALL VARIABLES

```r
library(rio)
library(janitor)
library(tidyverse)
DataHouses=import("https://ai.lange-analytics.com/data/HousingData.csv") |>
              clean_names("upper_camel") |>
              select(Price,Sqft=SqftLiving,Bedrooms,Condition)
```

https://ai.lange-analytics.com/