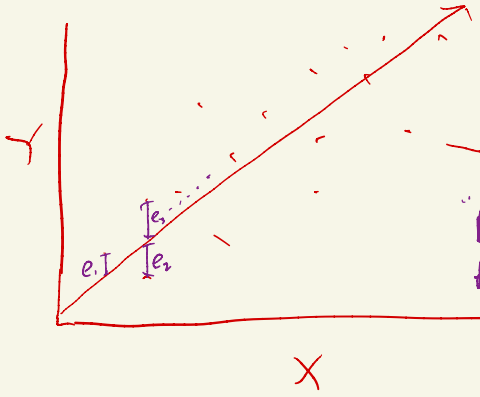
Prediction

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

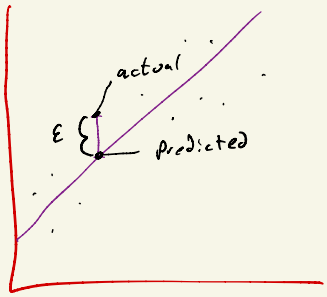
- Y: response
- β_0: intercept coefficient
- β_1: slope coefficient
- X: predictor variable
- ε: error

Intuitively



$Y = \beta_0 + \beta_1 X + \varepsilon$

"Draw a line that fits the points"



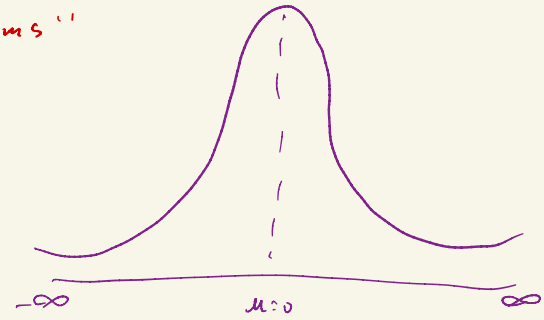
Mathematically

"minimize the squared error terms"

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ — Prediction

$Y = \beta_0 + \beta_1 X + \varepsilon$ — actual

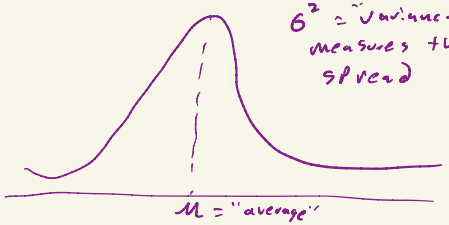
Find x such that $(Y - \hat{Y})^2$ is as small as possible.



Assumptions needed for linear regression

1. Linear relationship
2. (Normally) distributed errors (with mean zero)
3. Constant errors (Heteroscedasticity)
4. No outliers
5. No Multicolinearity (for multiple regression next week)

Normal Distribution



"Great"

"Bad" :(

Summary statistics for residuals

function

price = $\beta_0 + \beta_1$ Bed + β_2 bath + ...

```
Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + sqft_lot,
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1560041  -142669   -21995   102250  4165866

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.420e+04  7.510e+03   9.880  < 2e-16 ***
bedrooms    -5.728e+04  2.528e+03 -22.660  < 2e-16 ***
bathrooms    7.477e+03  3.827e+03   1.954   0.0507
sqft_living  3.115e+02  3.425e+00  90.966  < 2e-16 ***
sqft_lot    -3.604e-01  4.605e-02  -7.826 5.31e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 256400 on 18006 degrees of freedom
Multiple R-squared:  0.507,	Adjusted R-squared:  0.5069
F-statistic:  4629 on 4 and 18006 DF,  p-value: < 2.2e-16
```

Significance level

Statistic that determines the significance of each coefficient for the test H_0: $B_r = 0$, H_A: $B_r \neq 0$

Corresponding p-values

R^2 "Percent of variation in Y explained by X. (one number summary of goodness of fit.)

Don't worry about this for now.

Coefficient estimates

S.D. for those estimates (lower the better)

Statistic and p-val for H_0: Null model, H_A: This model