STAT 135
12. Linear predictions (Least Squares)

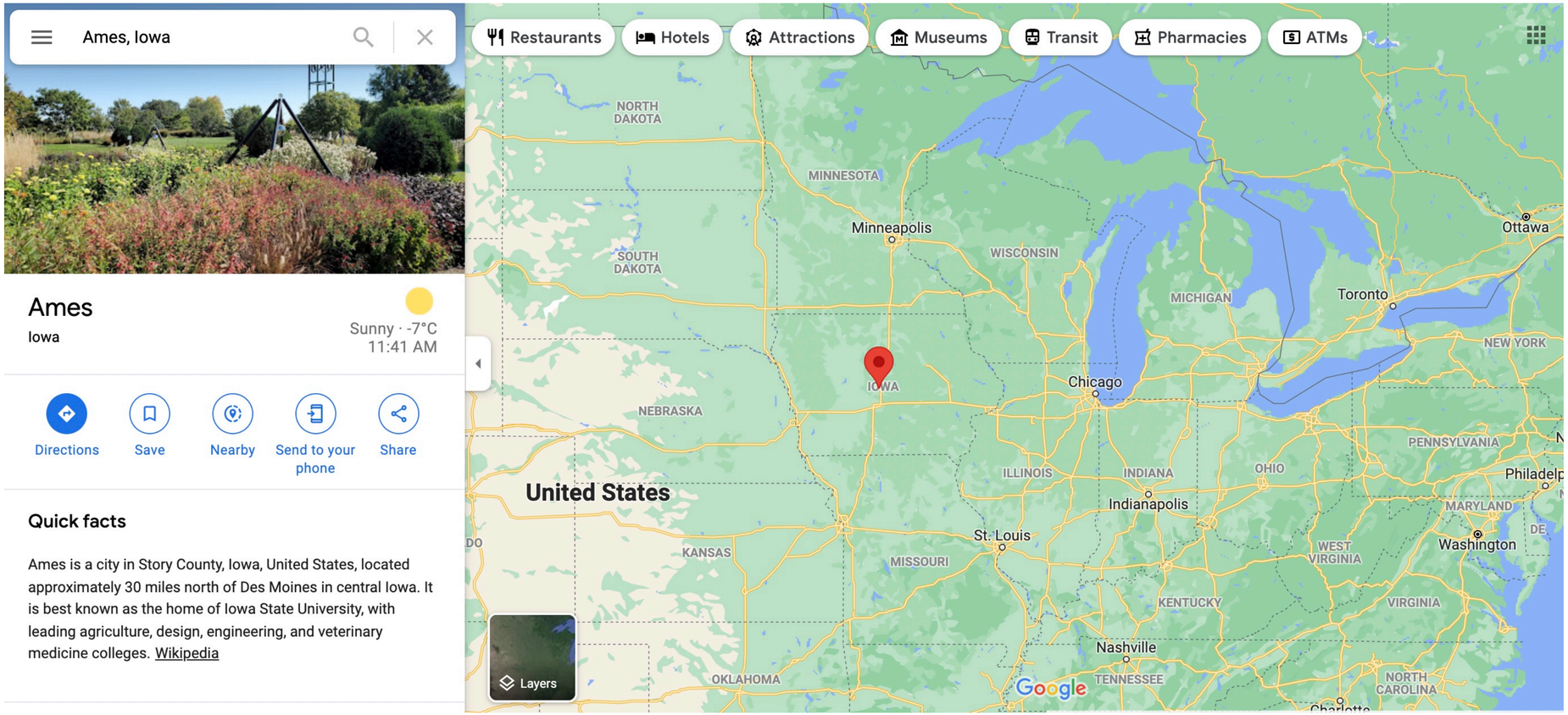
Spring 2022

Lecturer: Dr Rebecca Barter (*she/her*)
Office hours: Tu 9:30-10:30 (in person), Th 4-5pm (virtual)
Office: Evans 339

Email: rebeccabarter@berkeley.edu
Twitter: @rlbarter
GitHub: rlbarter

Prediction problems

Predicting house prices in Ames, Iowa



Predicting house prices in Ames, Iowa

Photo by Sieuwert Otterloo on Unsplash

Goal 1 (**prediction**):

We want to **predict the price of future houses** on the market in Ames using features such as lot size, year built, quality score

Goal 2 (**inference**):

We want to **understand the relationship between the sale price and features** (such as lot size, year built, quality score) of the houses in Ames

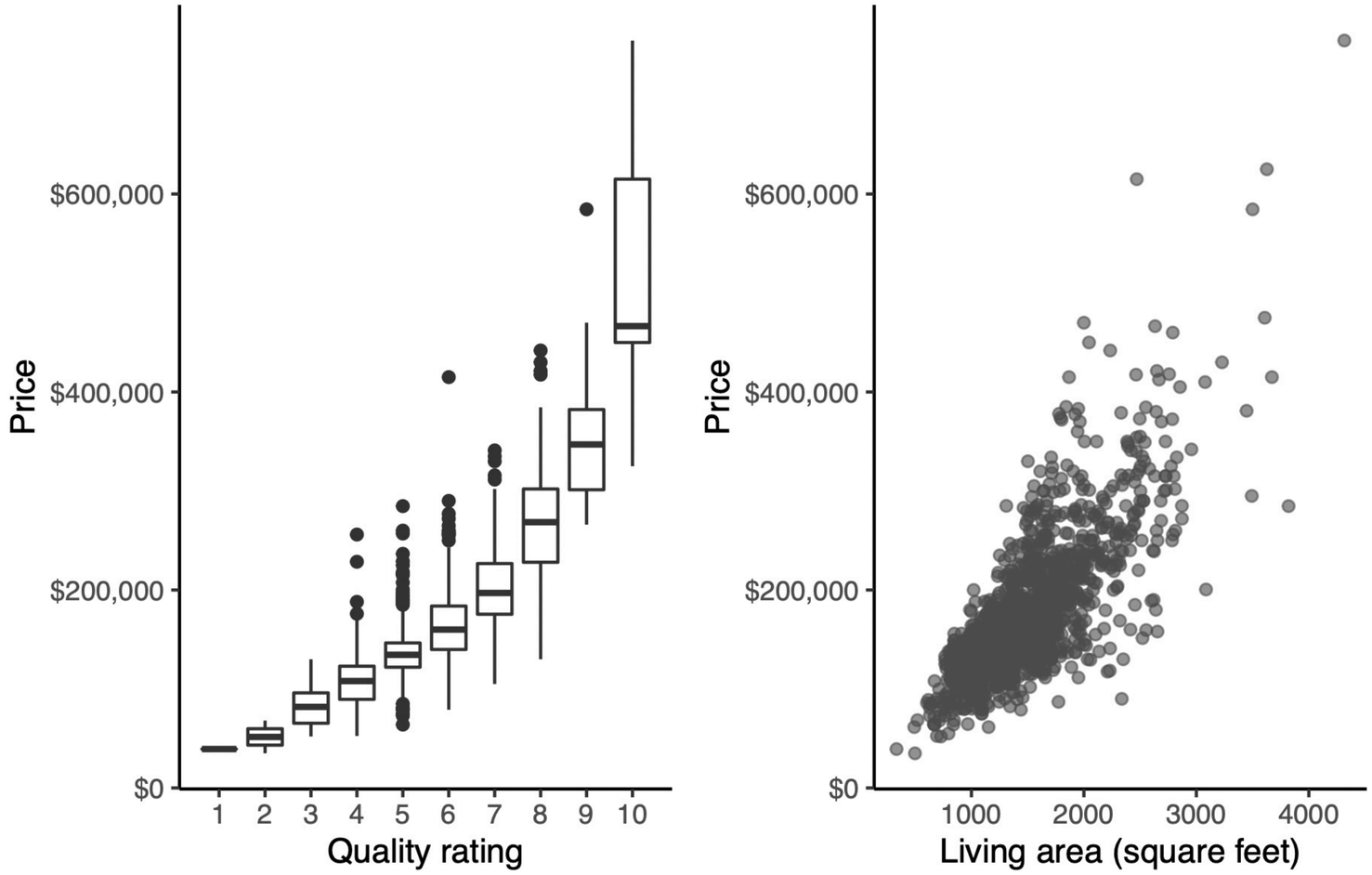
De Cock's Ames house price data

Dean De Cock obtained the data from the Ames City Assessor's Office (http://jse.amstat.org/v19n3/decock.pdf).

Data from:

- 2930 houses sold in Ames between 2006 and 2010
- With 82 variables (**sale price**, quality rating, living area, year built, etc)

EDA: how is house price related to the other features in the data



Predicting house prices in Ames, Iowa

The data we have:

Data from houses that have previously been sold in Ames

Price ($)	Living area	Quality score	Year built
298,500	1,550	8	1965
223,100	1,238	7	1956

Observed price: $298,500

Observed price: $223,100

Fit a model:

Quantify the relationship between sale price and features of the house:

Price $\approx f($**Living area**, **Quality**, **Year**, $\ldots)$

Price ($)		Living area	Quality score	Year built
298,500	f	1,550	8	1965
223,100		1,238	7	1956

"Response" (y)

"Features"/ "Covariates" (x)

Predict sale price of future houses:

Price ($)	Living area	Quality score	Year built
??	1,340	6	1980
??	1,112	7	1926

$f($

Living area	Quality score	Year built
1,340	6	1980
1,112	7	1926

$)$ $=$

Predicted price ($)
191,200
212,400

What kind of relationship, $y = f(x)$, should we seek?

Linear regression

Linear relationships

We assume that each observation can be represented as

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{A general linear relationship that describes the relationship between } y \text{ and } x \text{ across all observations}} + \underbrace{\epsilon_i}_{\text{An "error" term that allows for variation from the linear trend for individual houses}}$$

Where $E(\epsilon_i) = 0$, and $Var(\epsilon_i) = \sigma^2$ for all i

For example:

$$price_i = \beta_0 + \beta_1 area_i + \epsilon_i$$

Linear relationship assumptions

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

ϵ_i is **random** $\implies$ y_i is **random**

- ϵ_i are IID
- $E(\epsilon_i) = 0$, and $Var(\epsilon_i) = \sigma^2$ for all i

x_i is **fixed**

β_0, β_1 are **population parameters** that we need to estimate

Linear relationships assumptions

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

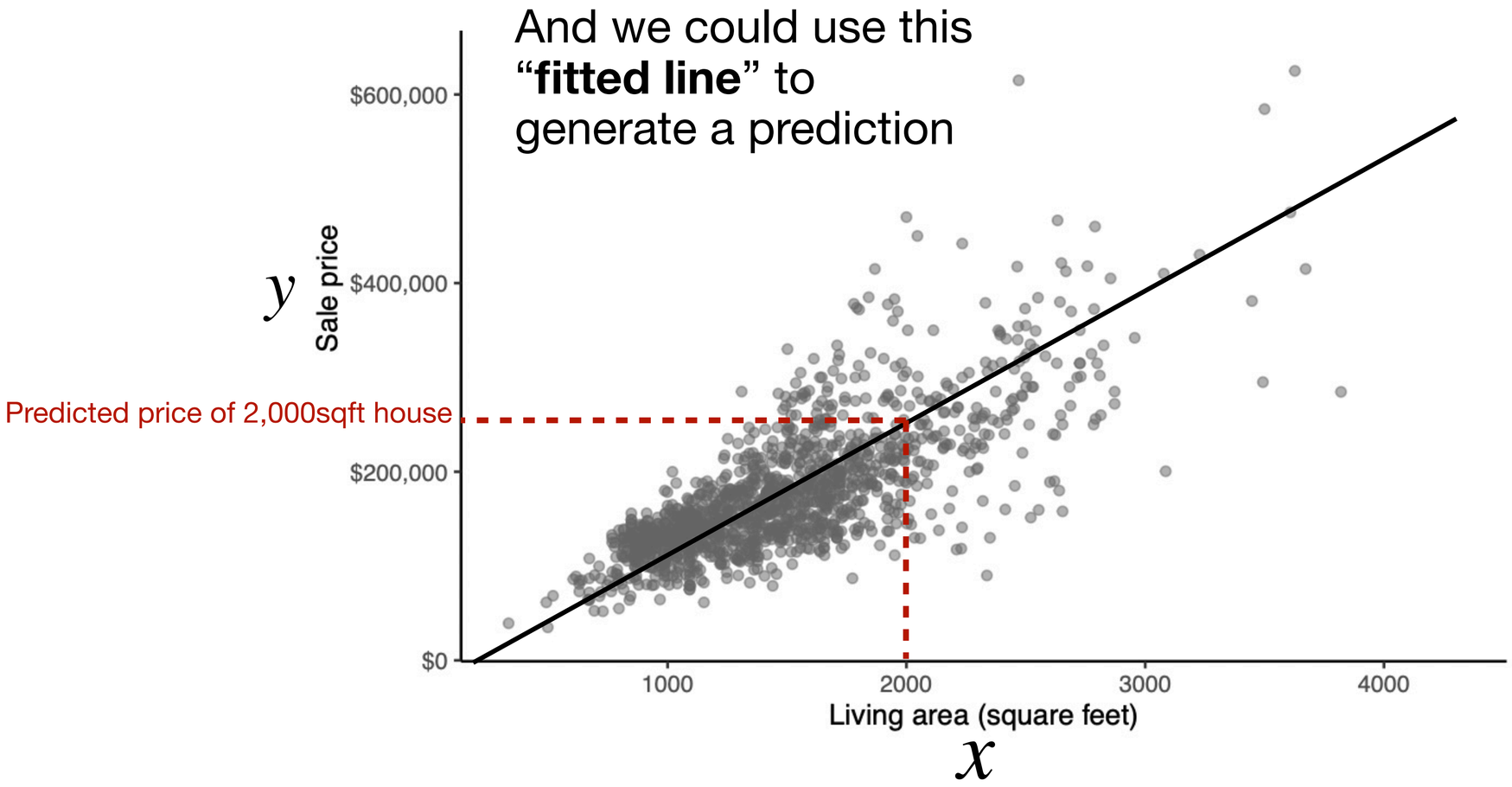
- ϵ_i are IID
- $E(\epsilon_i) = 0$, and $Var(\epsilon_i) = \sigma^2$ for all i

If this model is correct, then $$E[y_i] = \beta_0 + \beta_1 x_i$$

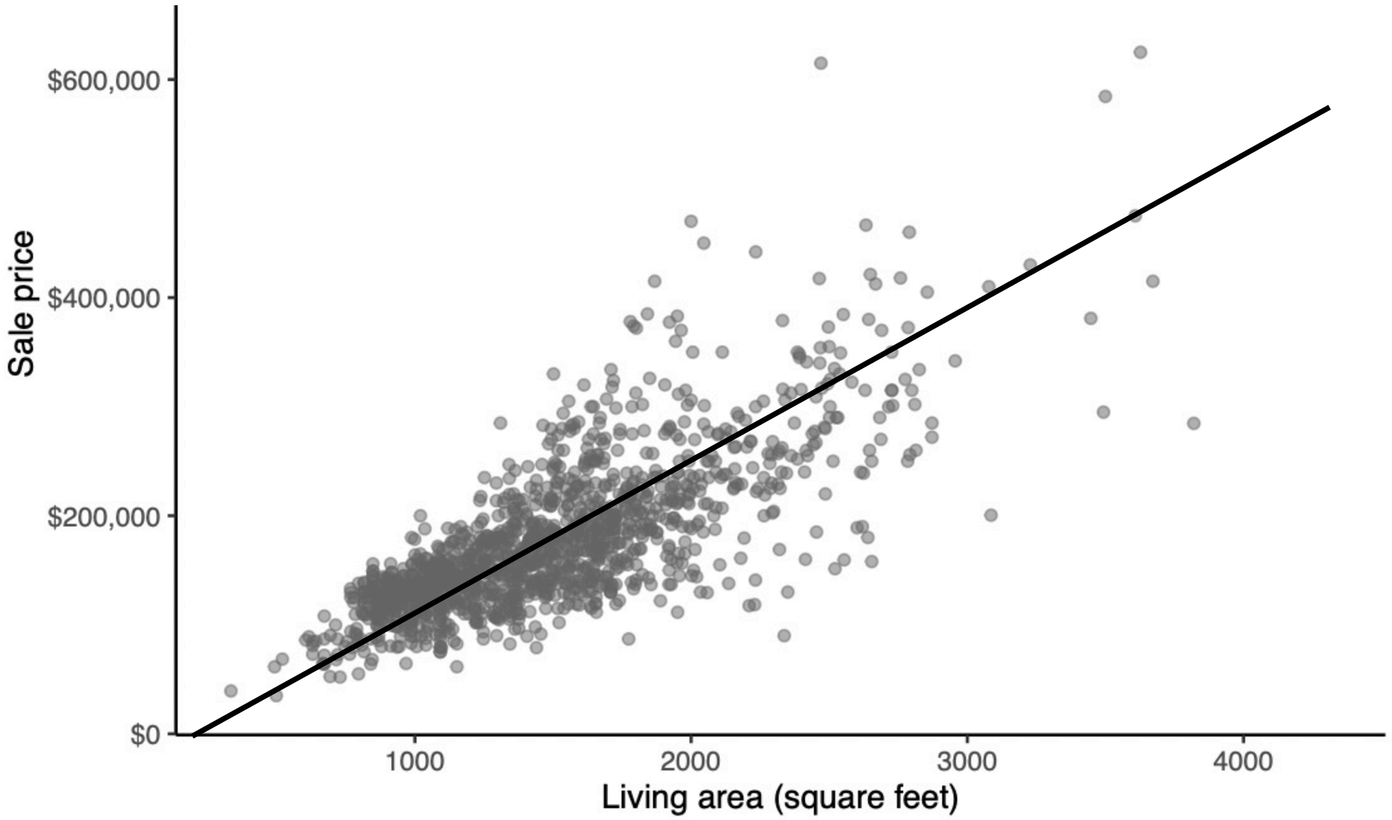
Which means that if we can estimate β_0 by $\hat{\beta}_0$ and β_1 by $\hat{\beta}_1$, we can use the "fitted line" $\hat{\beta}_0 + \hat{\beta}_1 x$ to predict the response for observations with covariate value x

Linear relationships

Let's predict **sale price** using **living area** only



Linear relationships



This fitted line can be written as:

$$\widehat{price} = 16{,}208 + 108 \times area$$

Predicted response

Covariate/feature/ predictor

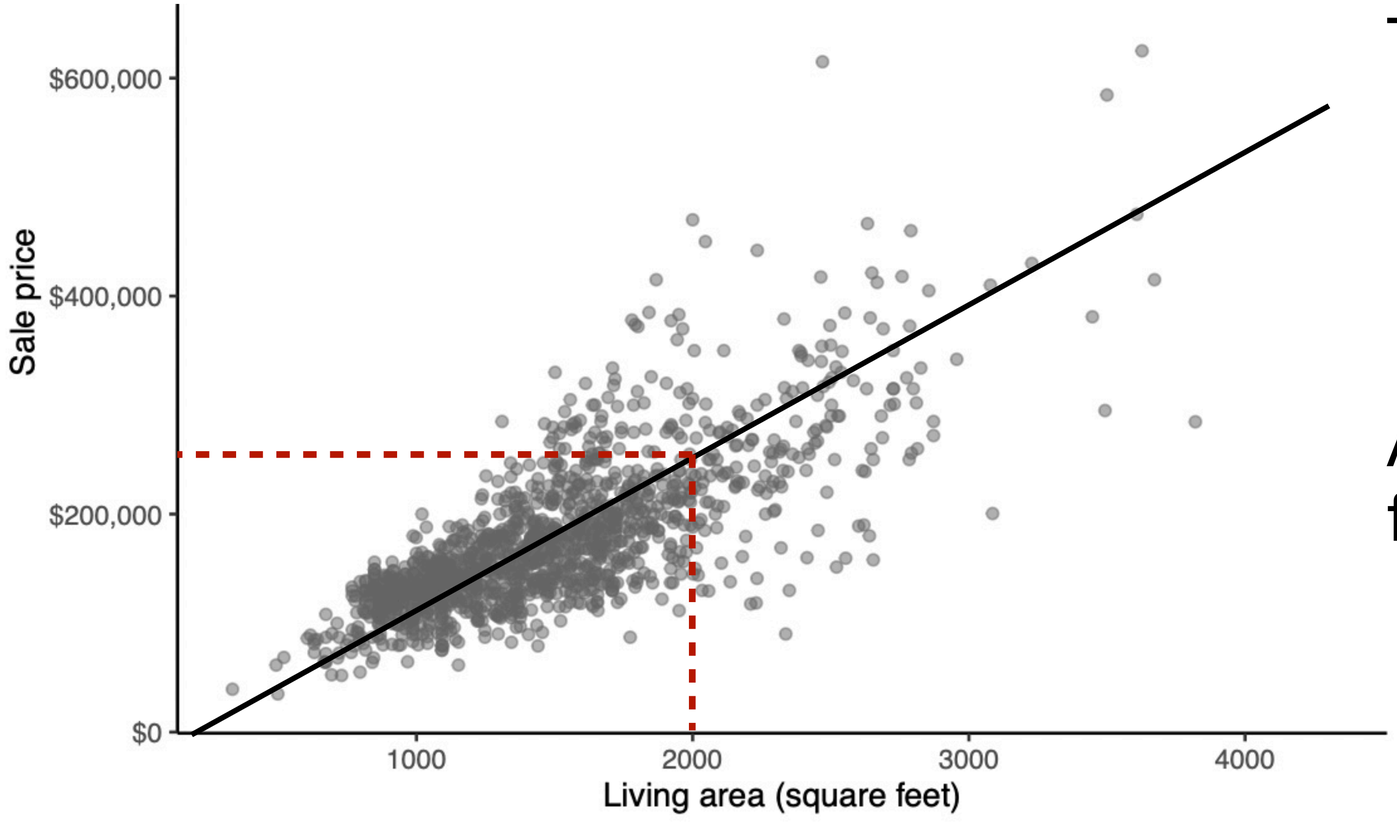
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Coefficients of the linear fit

The "**hat**", e.g. $\hat{y}, \hat{\beta}$, means that the quantity is a number that is **computed** ("estimated") from the data

$\hat{\beta}_0, \hat{\beta}_1$ are the **coefficients** of the the linear fit that we compute from the data

Generating a prediction from a linear fit



(The predictive feature/covariate)

This fitted line can be written as:

$$\widehat{price} = 16{,}208 + 108 \times area$$

A 2,000ft^2 house is predicted to sell for

$$\widehat{price} = 16{,}208 + 108 \times 2000$$
$$= 232{,}514$$

Interpreting the coefficients

If the assumed model is correct i.e. the data actually does follow:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

β_0 is the *base* response of observations with $x = 0$

(β_0 is the *base* house price of $0ft^2$ houses)

β_1 is the amount that the response increases when x increases by 1 unit

(β_1 is the amount that the price would increase if the living area of the house increased by $1ft^2$)

Interpreting the population coefficients

If the assumed model is indeed correct:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

β_1 is the amount that the response increases <u>for any individual data point</u> when x_i increases by 1 unit

$$\begin{aligned} y_i(x_i + 1) - y_i(x_i) &= \beta_0 + \beta_1(x_i + 1) + \epsilon_i) - (\beta_0 + \beta_1 x_i + \epsilon_i) \\ &= \beta_0 + \beta_1 x_i + \beta_1 + \epsilon_i - \beta_0 - \beta_1 x_i - \epsilon_i \\ &= \beta_1 \end{aligned}$$

Interpreting the population coefficients

If the assumed model is indeed correct:

$$y = \beta_0 + \beta_1 x + \epsilon$$

β_1 is the amount that the response increases <u>on average</u> when x increases by 1 unit

$$\begin{aligned} E[y(x+1)] - E[y(x)] &= \beta_0 + \beta_1(x+1) - (\beta_0 + \beta_1 x) \\ &= \beta_0 + \beta_1 x + \beta_1 - \beta_0 - \beta_1 x \\ &= \beta_1 \end{aligned}$$

Caution: causality

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

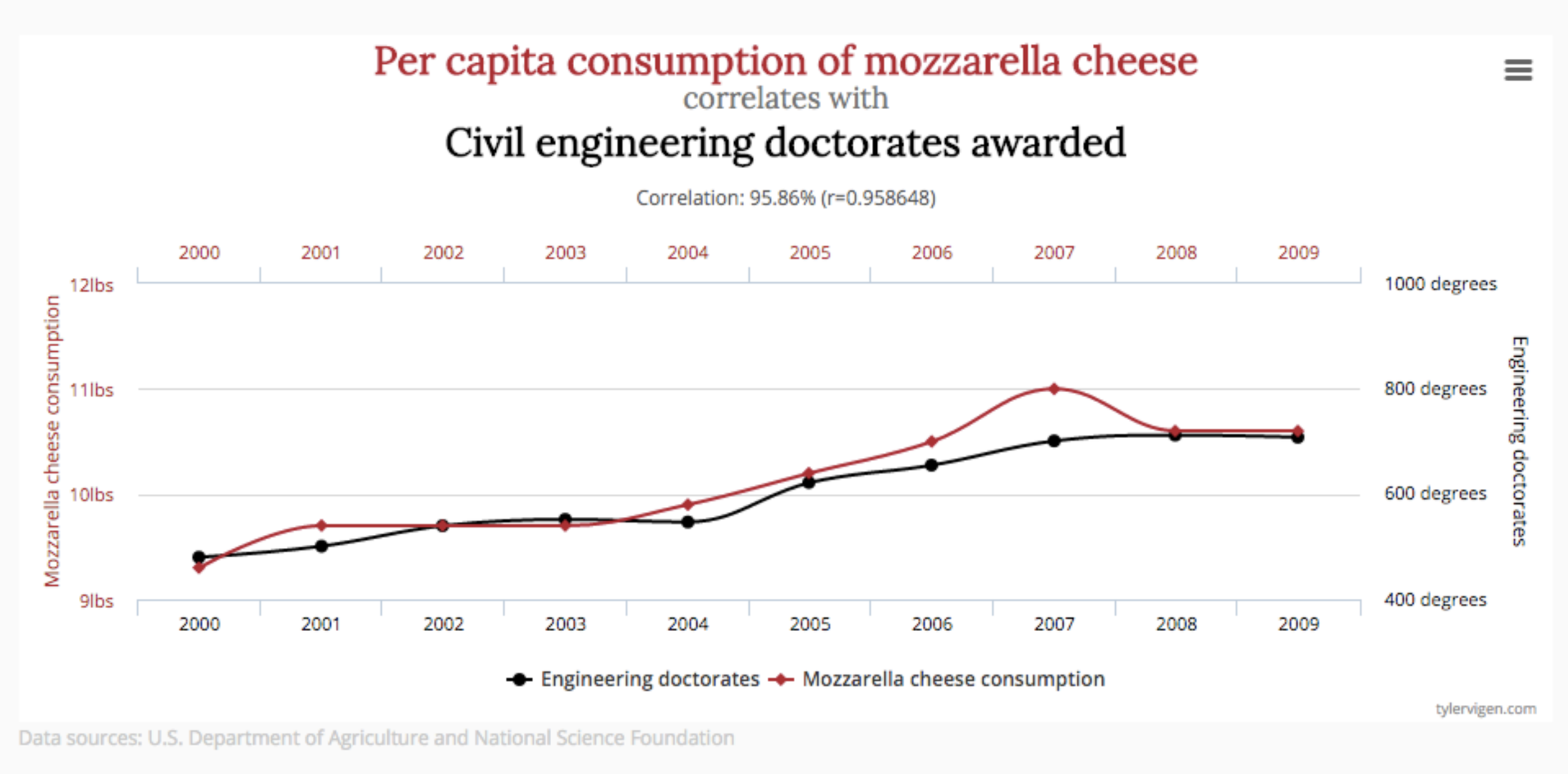
β_1 is the amount that the response increases when x_i increases by 1 unit

We aren't saying **anything** about causation.

We're not saying that increasing x by 1 unit **causes** y to increase by β_1

(There could be something else that causes both x and y to increase simultaneously)

Caution: causality



Conclusion: Eating a lot of mozzarella will get you a civil engineering PhD!

Interpreting the coefficients of the fitted line

For an estimate of the linear relationship from the data:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\widehat{price} = 16{,}208 + 108 \times area$$

$\hat{\beta}_0$ is the *base* predicted response of an observation with $x = 0$

($\hat{\beta}_0$ is the base predicted house price of an $0ft^2$ house)

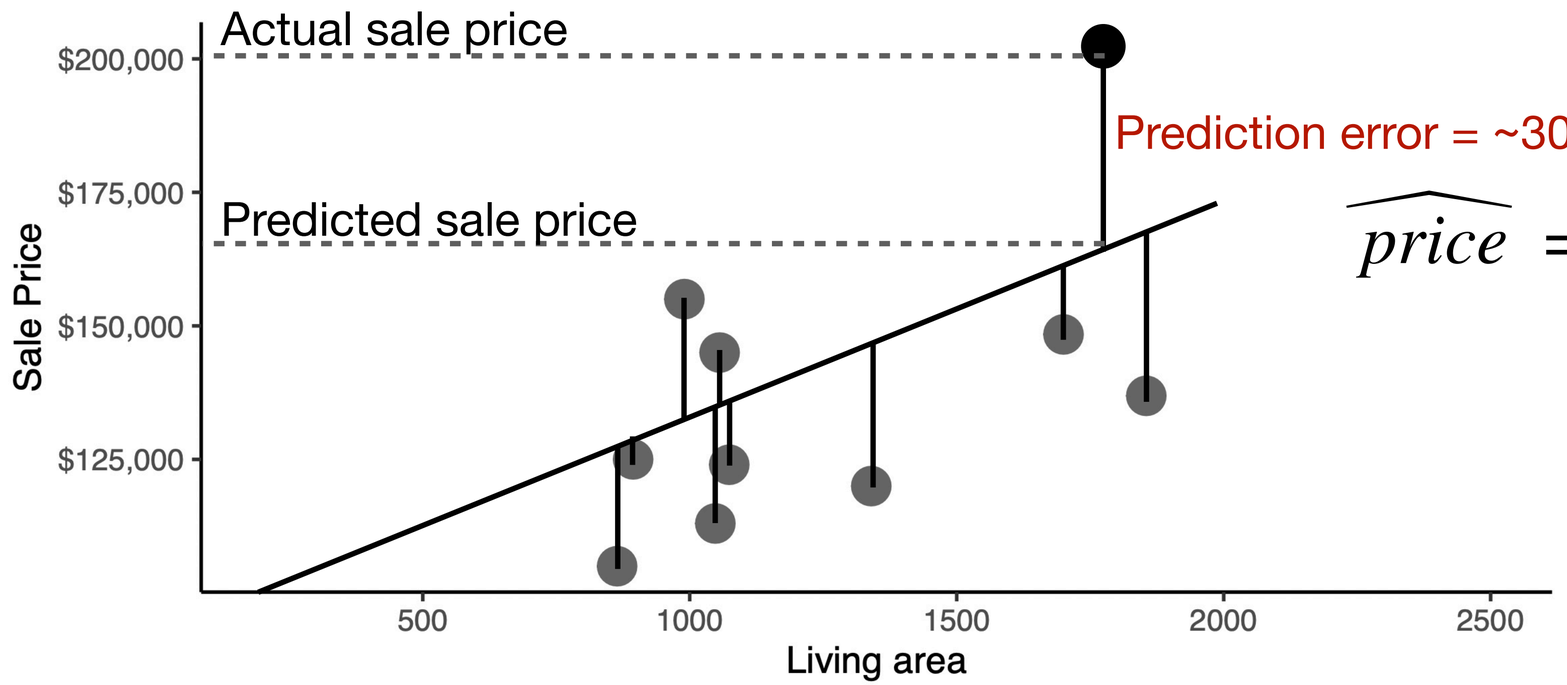
$\hat{\beta}_1$ is the amount that the predicted response increases when x increases by 1 unit

($\hat{\beta}_1$ is the amount that the predicted price would increase if the living area of the house increased by $1ft^2$)

The Least Absolute Deviation (LAD) algorithm

Estimating a linear relationship

How should we compute a fitted line?



$$\widehat{price} = 82{,}138 + 39 \times area$$

We could choose the line that has the <u>lowest *average/total*</u> prediction error

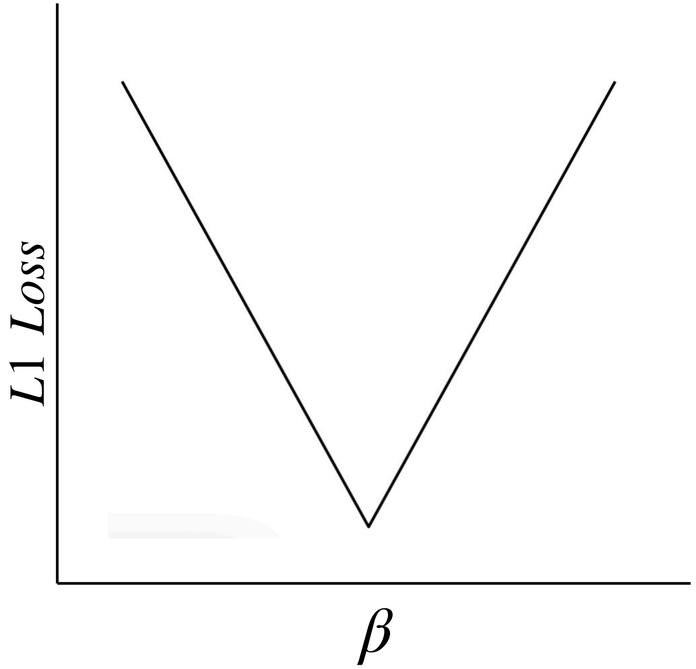
The Least Absolute Deviation (LAD) fitted line

Choosing the fitted line based on the values of absolute value error is called **Least Absolute Deviation (LAD):**

L1 loss function

$$\hat{\beta}_0 = \underset{\beta_0}{argmin} \boxed{\sum_{i=1}^{n} |\, y_i - (\boldsymbol{\beta_0} + \beta_1 x_i)\,|}$$

$$\hat{\beta}_1 = \underset{\beta_1}{argmin} \boxed{\sum_{i=1}^{n} |\, y_i - (\beta_0 + \boldsymbol{\beta_1} x_i)\,|}$$



Note that the L1 Loss is **not** differentiable at the minimum

I.e. the LAD is the line such that the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen so as to minimize the absolute difference between the observed and predicted response

How would you find the values of $\hat{\beta}_0, \hat{\beta}_1$ that minimize the L1 loss?

Since the L1 loss is not differentiable at the minimum, you could try plugging in various values of $\hat{\beta}_0, \hat{\beta}_1$ and see which values yield the lowest loss, e.g.:

$$\hat{\beta}_0 = 200, \quad \hat{\beta}_1 = 1, \qquad L1\ Loss = \sum_{i=1}^{10} |price_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)| = 3{,}656$$

$$\hat{\beta}_0 = 150, \quad \hat{\beta}_1 = 1.5, \qquad L1\ Loss = \sum_{i=1}^{10} |price_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)| = 8{,}097$$

$$\hat{\beta}_0 = 300, \quad \hat{\beta}_1 = 1, \qquad L1\ Loss = \sum_{i=1}^{10} |price_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)| = 3{,}856$$

$$\vdots \qquad\qquad\qquad \vdots$$

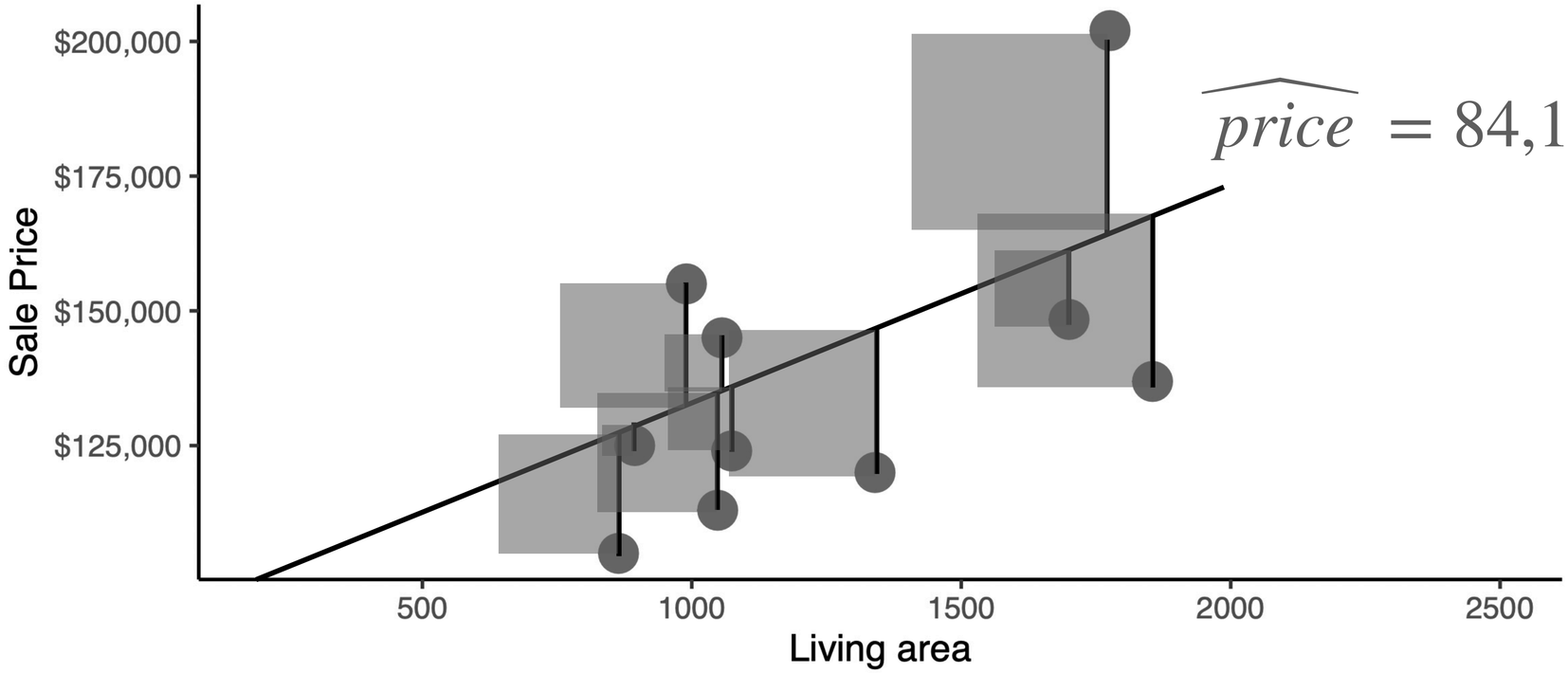
But this would be very inefficient and tedious…

The Least Squares (LS) algorithm

A different loss function: **Squared (L2) loss**

The L2 loss is the sum of the **areas of the squares** whose *edges* are the vertical distances from the point to the line

"Squared" **prediction error** (for a single house) $= (y - \hat{y})^2$



L2 loss function*

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

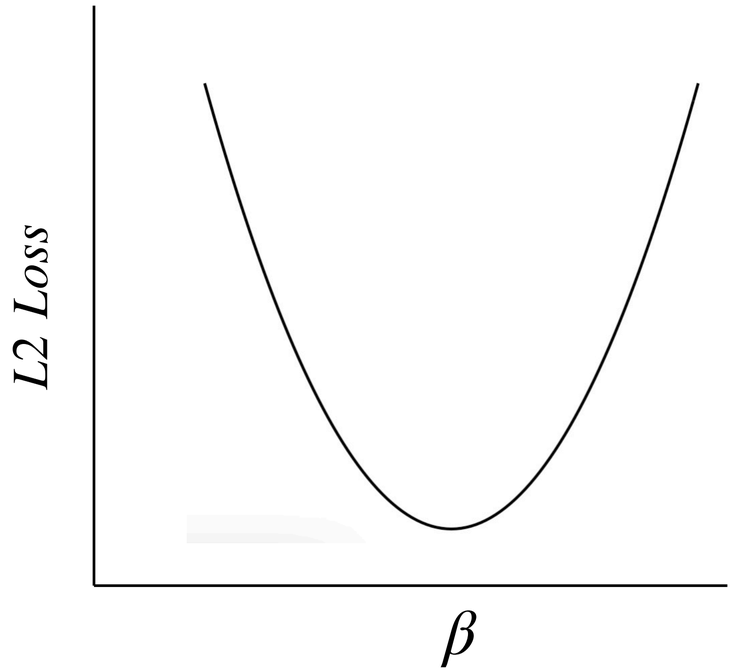
The Least Squares (LS) fitted line

Choosing the fitted line based on the values of **squared** error is called **Least Squares (LS):**

L2 loss function*

$$\hat{\beta}_0 = \underset{\beta_0}{argmin} \boxed{\sum_{i=1}^{n} (y_i - (\boldsymbol{\beta_0} + \beta_1 x_i))^2}$$

$$\hat{\beta}_1 = \underset{\beta_1}{argmin} \boxed{\sum_{i=1}^{n} (y_i - (\beta_0 + \boldsymbol{\beta_1} x_i))^2}$$



Note that the L2 Loss **is** differentiable at the minimum

I.e. the LS line is the line such that the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen so as to minimize the squared difference between the observed and predicted response

Computing the Least Squares fitted

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

The nice thing about **LS** is that it is possible to derive a **formula** for the values of β_0, β_1 that minimize the loss function:

$$\hat{\beta}_0 = \underset{\beta_0}{argmin} \sum_{i=1}^{n} (y_i - (\boldsymbol{\beta_0} + \beta_1 x_i))^2 \qquad \hat{\beta}_1 = \underset{\beta_1}{argmin} \sum_{i=1}^{n} (y_i - (\beta_0 + \boldsymbol{\beta_1} x_i))^2$$

Using calculus, we will show (in a few slides) that:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{Cov(x, y)}{Var(x)}$$

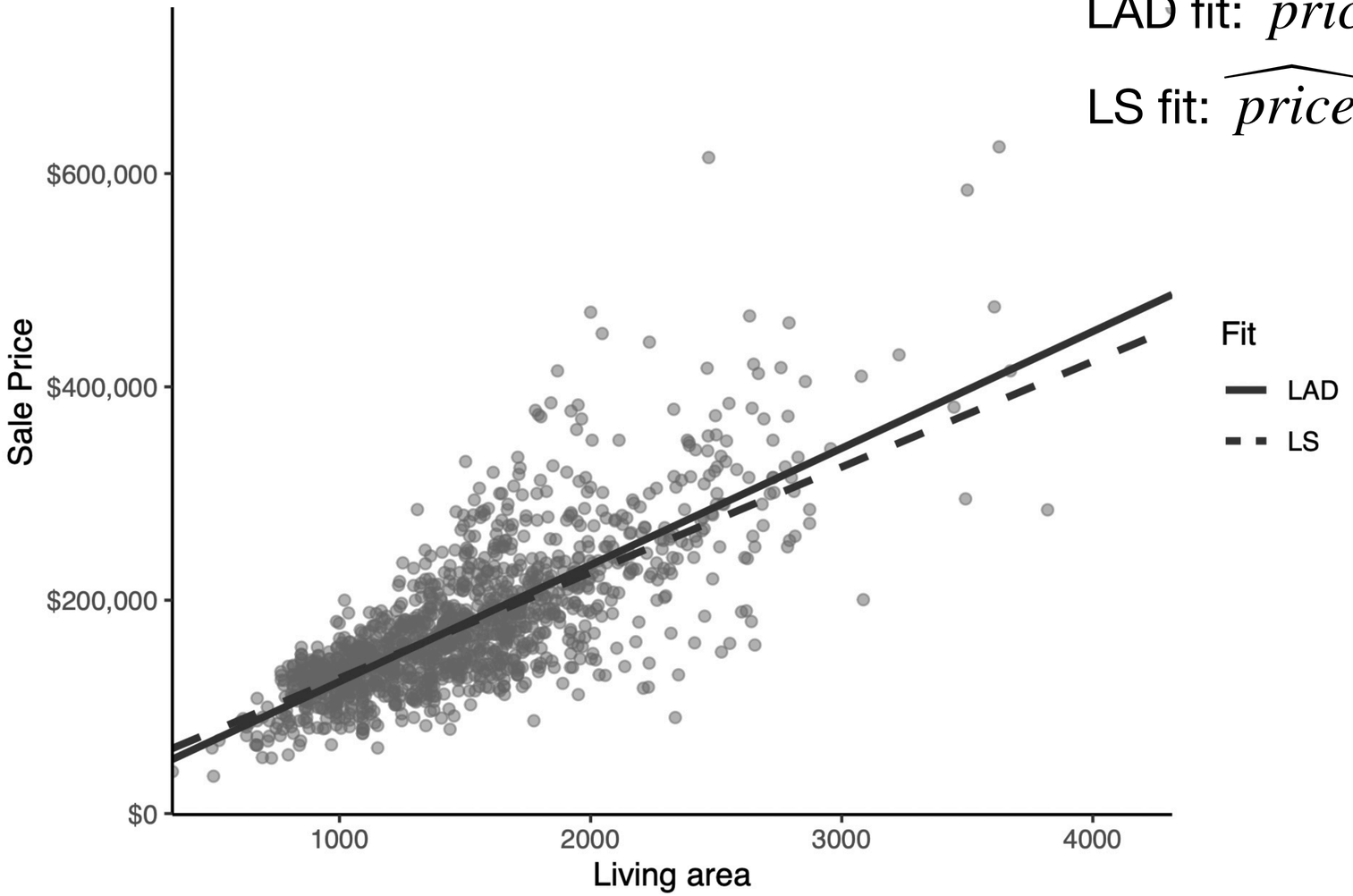
Plugging the data into these equations gives us the line, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, that minimizes the total area of the squares (squared error)

Empirically comparing the LS and LAD algorithm

Applying LS and LAD to the Ames house price prediction project

LAD fit: $\widehat{price} = 29{,}577 + 98 \times area$

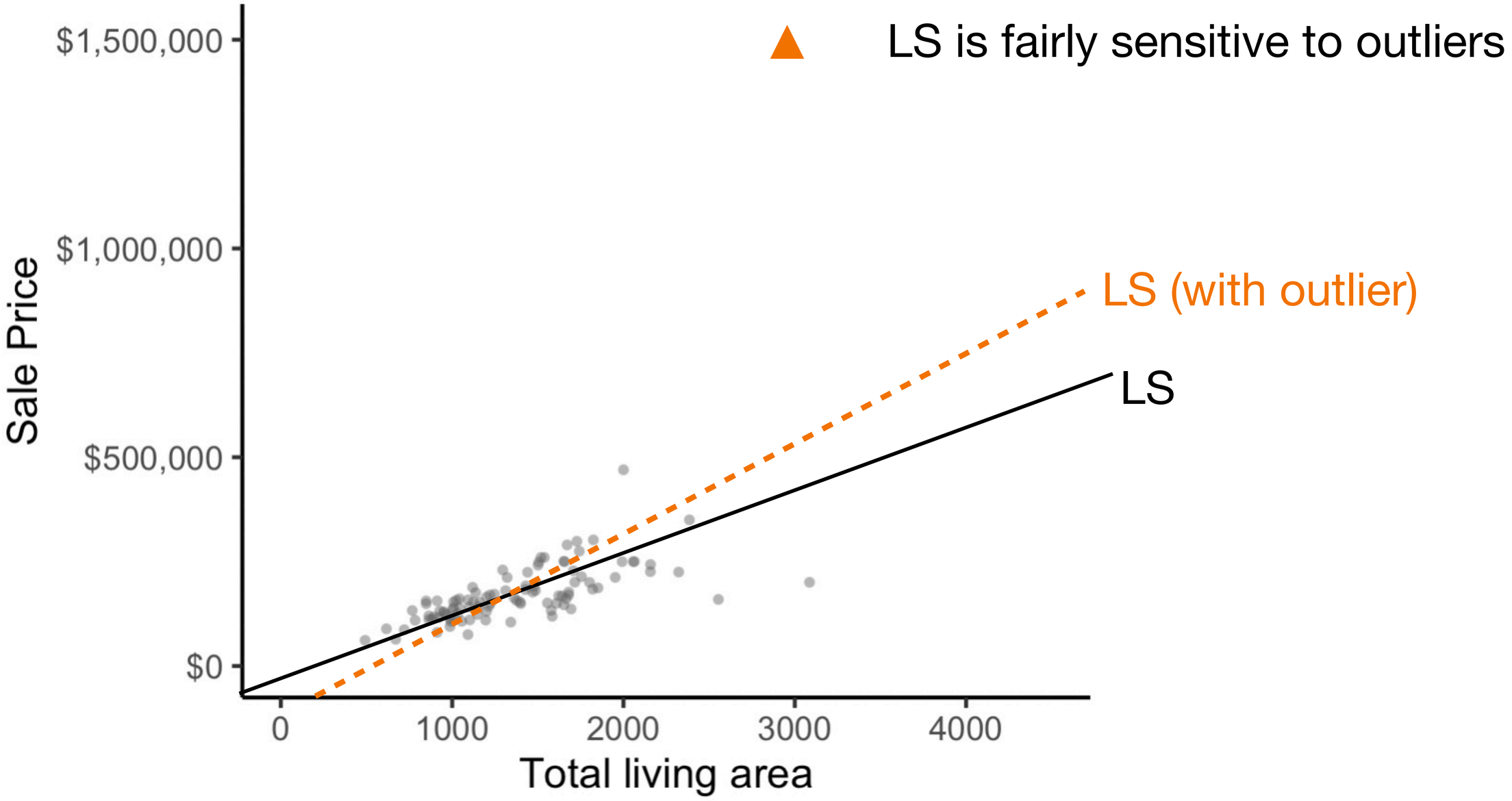
LS fit: $\widehat{price} = 16{,}208 + 108 \times area$



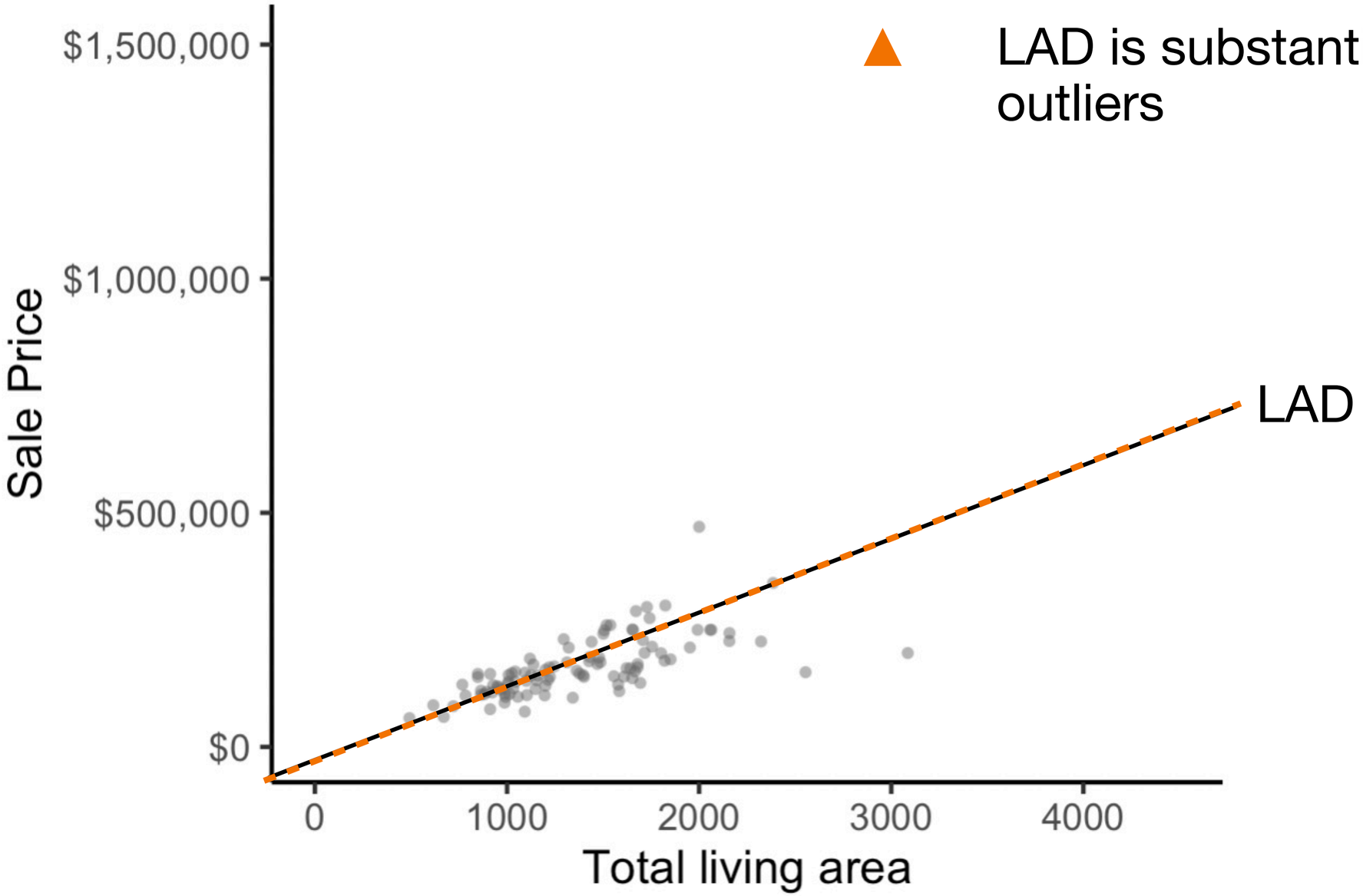
Living area	LAD pred	LS Pred
870	$114,500	$110,301
1200	$146,711	$145,991
1500	$175,995	$178,437

Which fit should we use?

Comparing LS and LAD: outliers

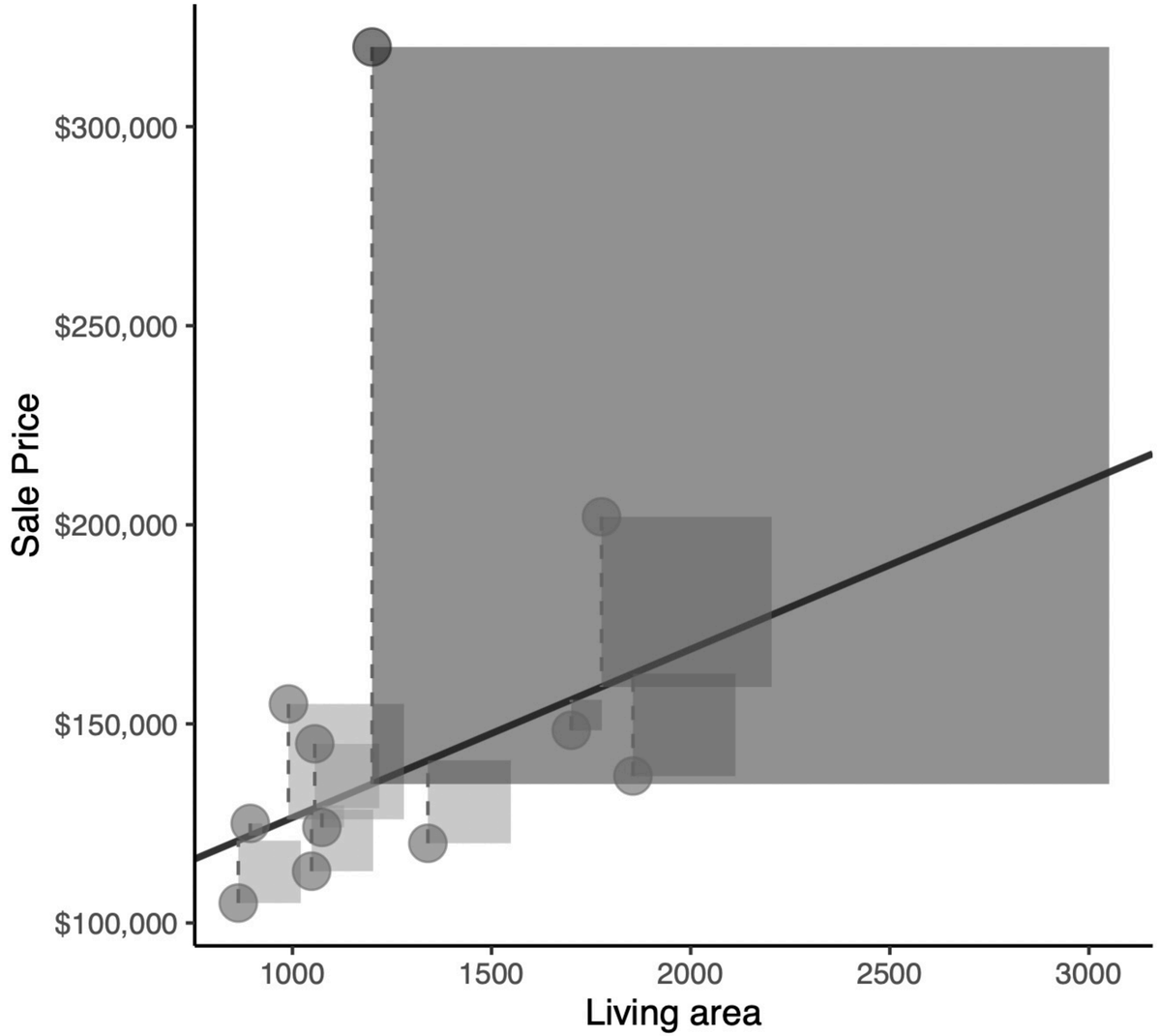


Comparing LS and LAD: outliers

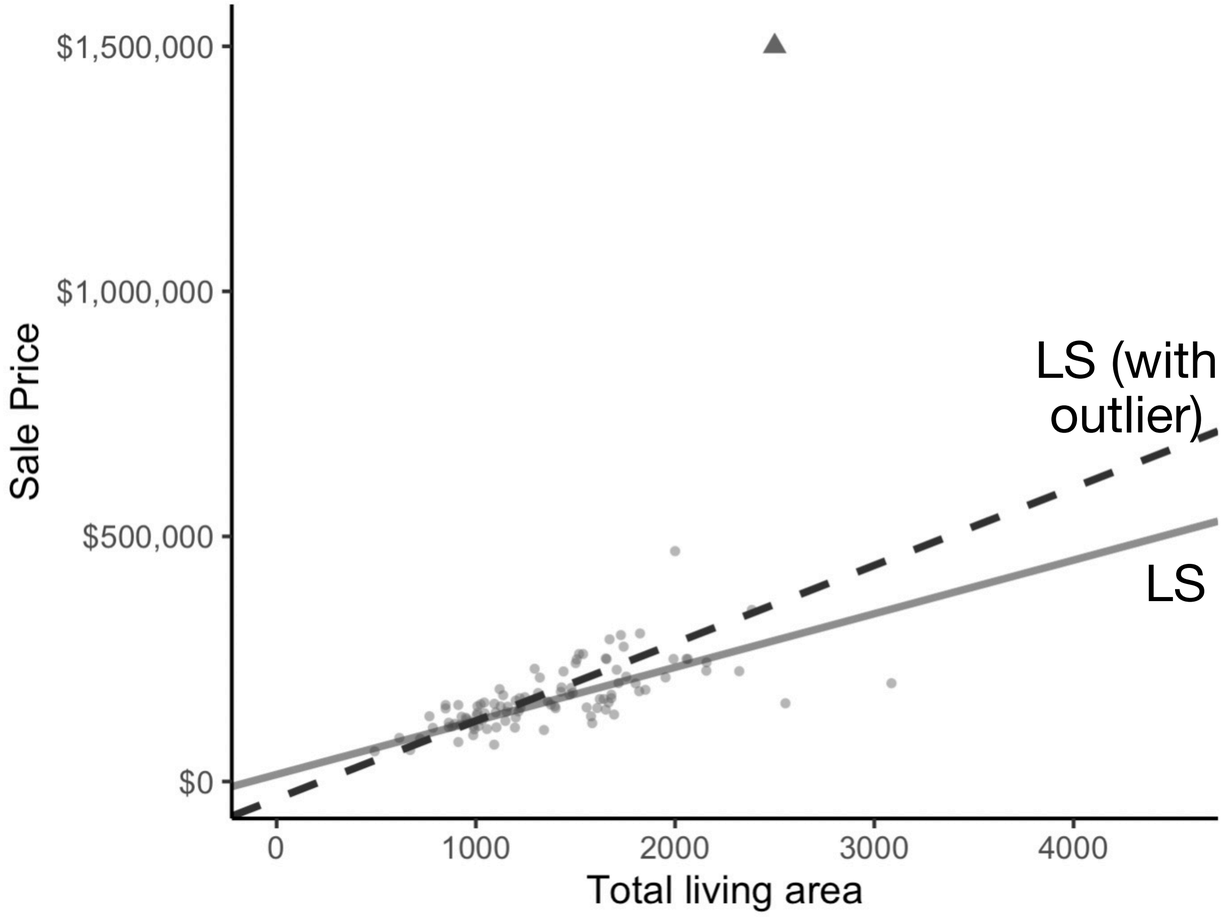


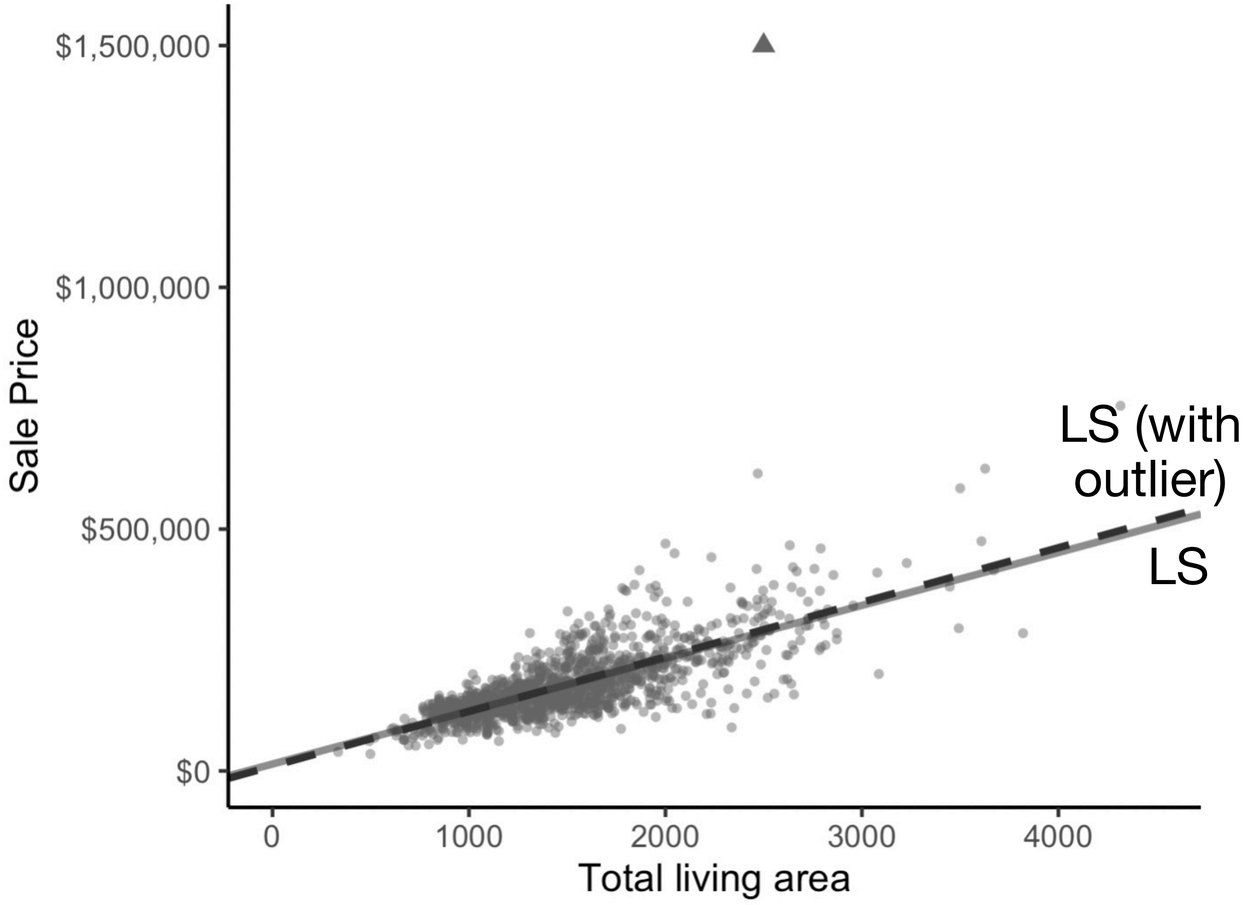
▲ LAD is substantially **less** sensitive to outliers

Understanding LS sensitivity to outliers



The effect of outliers on LS diminishes with the number of data points





Discussion:

What are some pros and cons of LS vs LAD?

Which would you use for the Ames house price prediction project?

(LAD +1) LAD is less sensitive to outliers than LS (like the median vs mean). Most people don't necessarily want the sale of rare mansions to influence the sale price predictions.

(LAD +1) The absolute error is more intuitive than the the squared error in the context of house prices.

(LS +1) LS is easier to compute and is more common in practice.

(LS +1) If our goal is inference, the LS fits have some useful properties that we will discuss later.

Try both! We will discuss later how to compare the performance of one algorithm to the other.

We will focus on the **LS** algorithm

Deriving the LS fitted line:

(Estimating the β_0 and β_1 parameters using LS)

The “Ordinary” Least Squares (OLS) coefficients

The line we are trying to fit is: $y = \beta_0 + \beta_1 x$

The values of β_0 and β_1 that minimize the L2 squared loss (LS algorithm):

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Are given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{Cov(x, y)}{Var(x)}$$

These are ***estimates*** of the population parameters β_0 and β_1

(There are no such *formulas* for the LAD algorithm)

LS coefficient proof ($\hat{\beta}_0$)

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

In order to find the values of $\beta_0$and β_1 that minimize $S(\beta_0, \beta_1)$, we need to set

$\frac{\partial S}{\partial \beta_0}$ and $\frac{\partial S}{\partial \beta_1}$ to zero, and solve for β_0 and β_1, respectively.

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

So $\frac{\partial S}{\partial \beta_0} = 0$ implies that $\quad 0 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \beta_1 x_i)$

$$\implies n\hat{\beta}_0 = \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i \qquad \implies \boxed{\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}}$$

LS coefficient proof ($\hat{\beta}_1$)

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

In order to find the values of $\beta_0$and β_1 that minimize $S(\beta_0, \beta_1)$, we need to set

$\frac{\partial S}{\partial \beta_0}$ and $\frac{\partial S}{\partial \beta_1}$ to zero, and solve for β_0 and β_1, respectively.

Next, $\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$

So $\frac{\partial S}{\partial \beta_1} = 0$ implies that $0 = \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$

Substituting β_0 with $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$: $0 = \sum_{i=1}^{n} x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$

Rearranging to make $\hat{\beta}_1$ the subject:

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

Formulations of the slope coefficient $(\hat{\beta}_1)$

The slope of the line, $\hat{\beta}_1$, is essentially the correlation between x and y:

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

$$= \frac{Cov(x,y)}{Var(x)} = \frac{\hat{\sigma}_{x,y}}{\hat{\sigma}_x^2}$$

Covariance

$$= Corr(x,y)\frac{\sqrt{Var(y)}}{\sqrt{Var(x)}} = \rho_{x,y}\frac{\hat{\sigma}_y}{\hat{\sigma}_x}$$

Correlation

Correlation and covariance are both measures of the strength of a linear relationship

Correlation and covariance

Covariance asks: when the value of one variable increases, to what extent does the other also tend to increase??

Population version: $$\sigma_{x,y} = Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

Data sample version: $$\hat{\sigma}_{x,y} = Cov(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

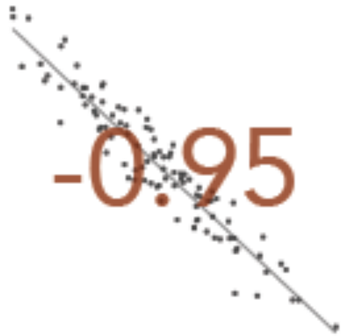
Correlation is like scaled covariance (it lies between -1 and 1)

Population version: $$\rho_{X,Y} = Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

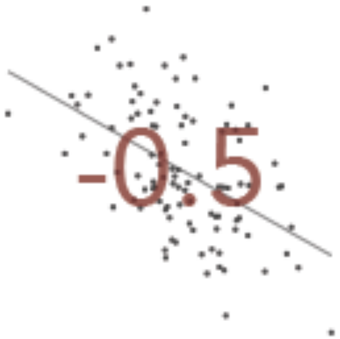
Data sample version:

$$\rho_{x,y} = Corr(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\hat{\sigma}_x \hat{\sigma}_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Correlation and covariance



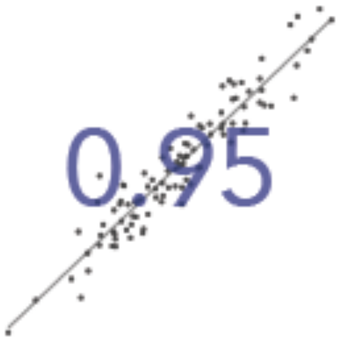
strong inverse correlation



inverse correlation

no correlation

correlation



strong correlation

LS coefficient example

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \rho_{x,y} \frac{\hat{\sigma}_y}{\hat{\sigma}_x}$$

For the Ames data

$$\bar{y} = \overline{price} = 176{,}172 \qquad \rho_{x,y} = \rho_{area,price} = 0.737$$

$$\bar{x} = \overline{area} = 1{,}479 \qquad \hat{\sigma}_x = \hat{\sigma}_{area} = 482.7$$

$$\hat{\sigma}_y = \hat{\sigma}_{price} = 70{,}797.2$$

$$\hat{\beta}_1 = \rho_{x,y} \frac{\hat{\sigma}_y}{\hat{\sigma}_x} = 0.737 \times \frac{70{,}797.2}{482.7} = 108.15$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 176{,}172 - 108.15 \times 1{,}479 = 16{,}207.9$$

So our fitted line is:

$$\widehat{price} = 16{,}207.9 + 108.15 \times area$$

LS and LAD code:

ls.R

Evaluating prediction performance

Evaluating predictions

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x$$

How can we provide a sense of how accurate our predictions are?

Compare the observed response to the predicted response

Predicted price ($)	Observed price ($)	
191,200	182,400	Absolute error: $8,800 Squared error: $77,440,000
212,400	213,700	Absolute error: $1,300 Squared error: $1,690,000

You *typically* want your definition of "error" to match the metric used to define your prediction

i.e., LS = squared error, LAD = absolute error

Mean Squared Error (MSE)

The LS fitted line is the line that minimizes the L2 loss

L2 loss function*

$$\hat{\beta}_0, \hat{\beta}_1 = \underset{\beta_0, \beta_1}{argmin} \boxed{\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2}$$

Mean Squared error: the value of the L2 loss when you plug in the estimated parameters

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2$$

Mean Squared Error (MSE)

Since MSE is on a *squared* scale, it is common to compute the ***root Mean Squared Error (rMSE)***

$$rMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2}$$

Which is on the same scale as the original data

Mean Absolute Deviation (MAD)

The LAD fitted line is the line that minimizes the L1 loss

L1 loss function*

$$\hat{\beta}_0, \hat{\beta}_1 = \underset{\beta_0, \beta_1}{argmin} \boxed{\sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)|}$$

Mean Absolute Deviation: the value of the L1 loss when you plug in the estimated parameters

$$MAD = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \hat{y}_i \right| = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right|$$

Comparing predictive algorithms with MSE/MAD

What metric(s) should you use to compare the LS fit to the LAD fit?

(a) Use the MSE to evaluate the LS and use the MAD to evaluate the LAD

(b) Use the MSE to evaluate both the LS and the LAD fits

(c) Use the MAD to evaluate both the LS and the LAD fits

You shouldn't really compare the MSE to the MAD…. (So (b) and (c) are reasonable, but (a) is not)

Comparing predictive algorithms with MSE/MAD

Our two algorithms evaluated using three metrics:

	LS	LAD
MSE	**2,341,650,645**	2,382,251,490
rMSE	**48,391**	48,808
MAD	34,047	**33,653**

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

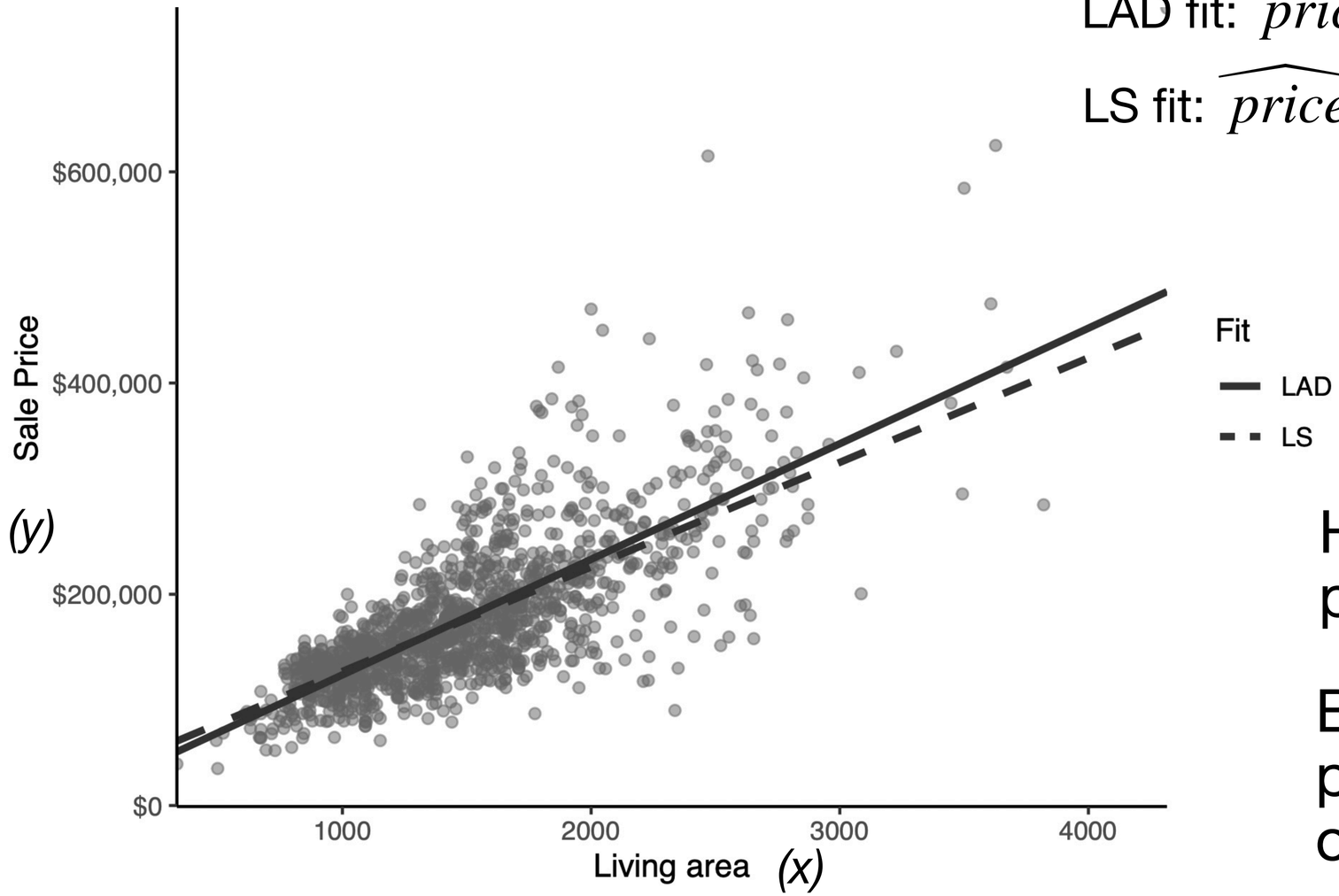
$$rMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$$

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

Visually comparing predictive algorithms

LAD fit: $\widehat{price} = 29{,}577 + 98 \times area$

LS fit: $\widehat{price} = 16{,}208 + 108 \times area$

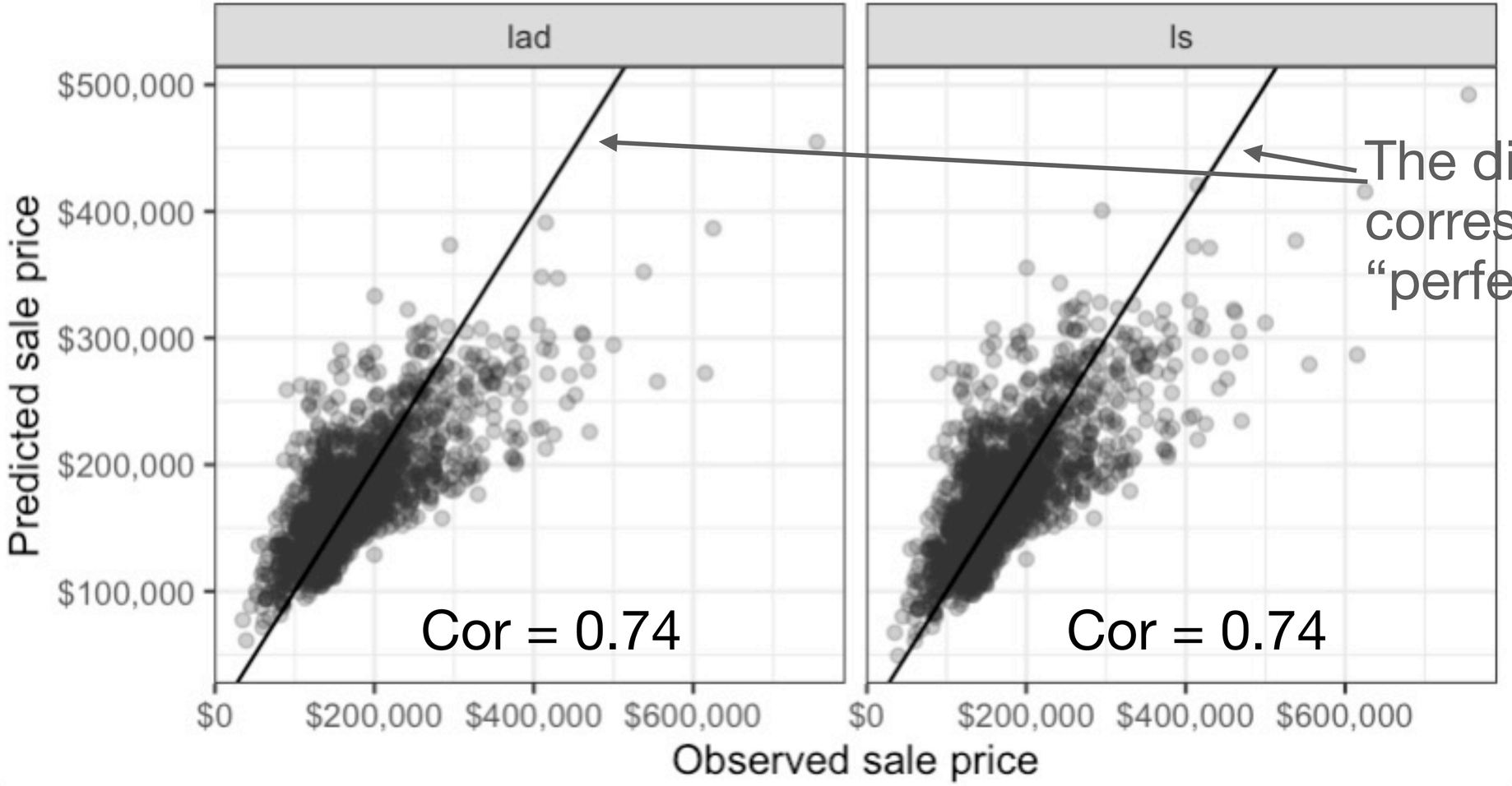


Here we are just viewing the predicted line in y-vs-x space

But we can also compare the predicted response versus the observed response

Visualizing predictive performance

Here are scatterplots of the predicted versus observed responses for the LAD and LS algorithms…. They look fairly similar



The diagonal line corresponds to a "perfect" prediction

R-squared value (coefficient of determination)

The R-squared value is the squared correlation between the observed and predicted response

$$R^2 = \rho^2_{y,\hat{y}}$$

R^2 lies between 0 and 1

R^2 corresponds to the **proportion of variability in the observed response that can be "explained" by the covariate**

$$R^2 = \frac{\hat{\sigma}^2_y - \hat{\sigma}^2_{\hat{e}}}{\hat{\sigma}^2_y}$$

(If $\hat{\sigma}^2_{\hat{e}}$ is very small, then $\hat{y}_i$ is very similar to y_i on average, which means x conveys a lot of information about y)

Where $\hat{e}_i = y_i - \hat{y}_i$

(The proof of this is just some algebra)

Comparing predictive algorithms with R squared

Our two algorithms evaluated using R squared:

	LS	LAD
R squared	**0.55**	**0.55**

So 55% of the variation in sale price can be explained by living area, using either algorithm

Training and test set

So far we have only evaluated our algorithms using the same data that we used to train our algorithms.

However, our algorithm will typically generate more accurate predictions on data it has seen than new data that it hasn’t seen.

If you plan to use the algorithm in practice, it is important to evaluate its performance on new data that was not used to train it.

Unfortunately, we usually don’t have access to new data!

Training and test set

Before you train a predictive algorithm, you need to split your data into:

Training data: the data you will use to train your algorithm

Testing data: the data you will use to evaluate your algorithm

Common proportions: 60% training/40% testing

70% training/30% testing

80% training/20% testing

Comparing predictive algorithms with MSE/MAD

Our two algorithms evaluated using three metrics:

		LS	LAD
MSE	Training	**2,341,650,645**	2,382,251,490
	Test	2,162,104,540	**2,138,177,484**
rMSE	Training	**48,391**	48,808
	Test	46,498	**46,240**
MAD	Training	34,047	**33,653**
	Test	32,934	**32,291**
Correlation	Training	0.743	0.743
	Test	0.726	0.726

Evaluating predictions:

ls.R

Using predictive algorithms in the real-world

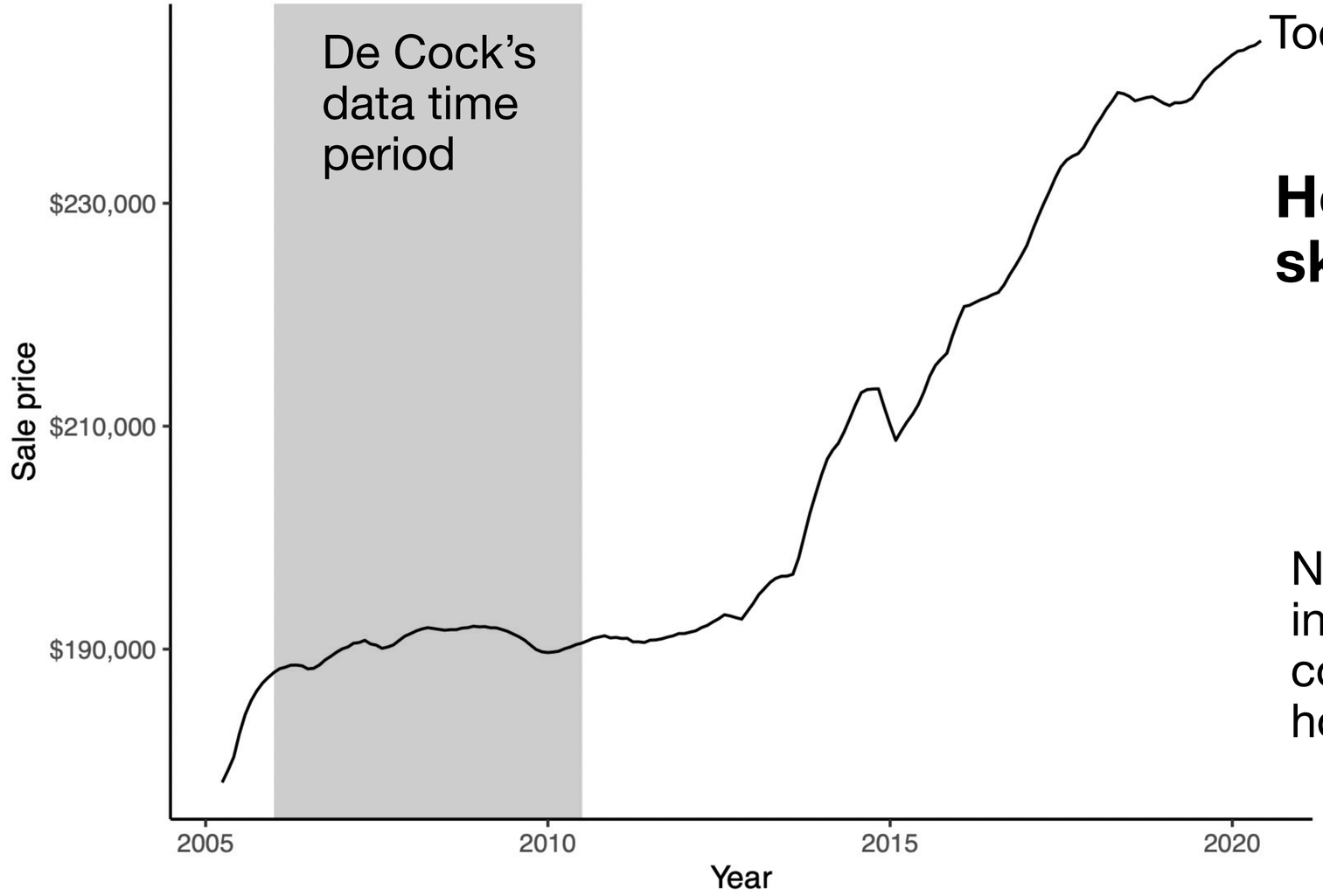
Discussion:

How could you check whether your house price predictions (based on De Cock's data) are **relevant** to and **accurate** in *today's* housing market in Ames?

1. Look to see whether house price trends in Ames have changed over time
2. Evaluate the predictions on a few houses that have recently been sold (requires access to new data)
3. …

An investigation of data relevance (Zillow)

Problem: De Cock's dataset ends in 2011, but we are living in 2022!



House prices in Ames have skyrocketed since 2013!

Note that this Zillow data only contains information on sale price, but it doesn't contain other information about the houses

Always view your data and models in the real-world context in which you intend to apply them!

Since we will just use this data for educational purposes, we will pretend that we are living in the blissful year of 2011.

But we would not recommend using our predictions in today’s market.