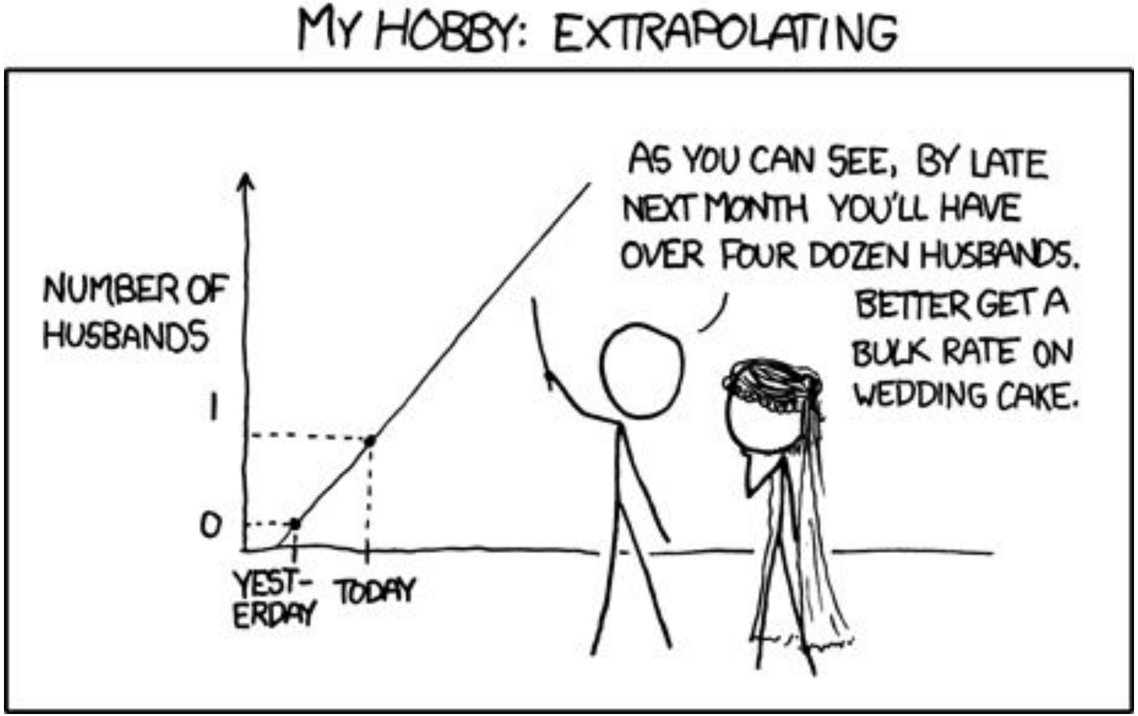
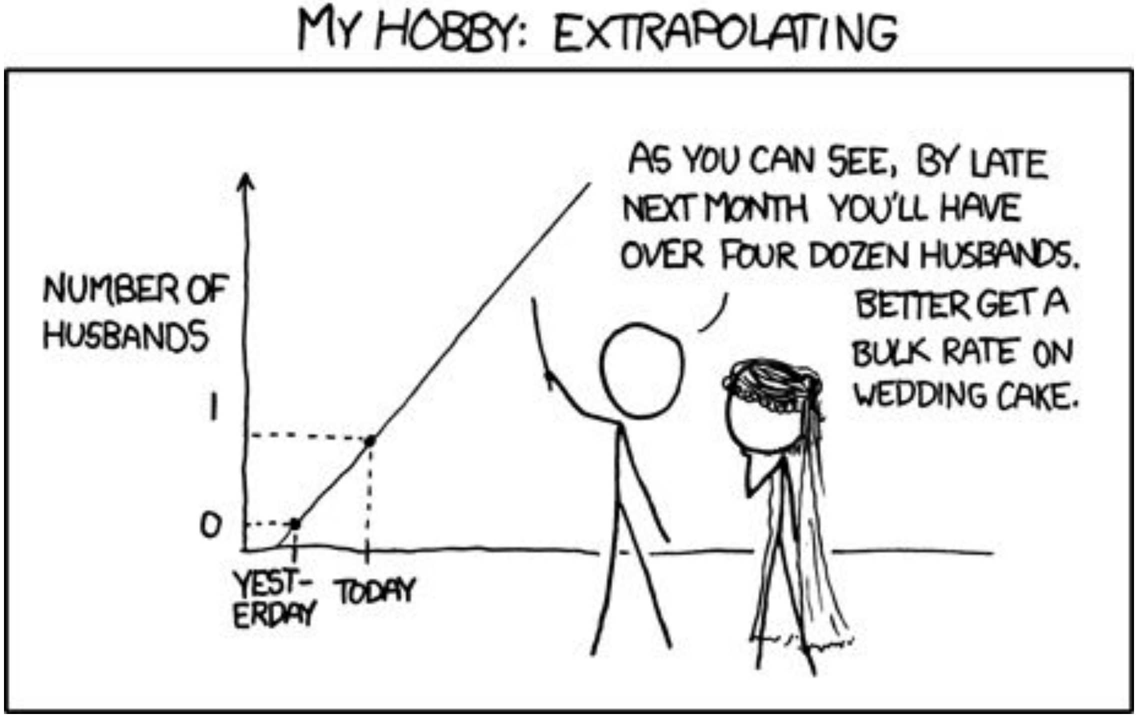
Lecture 20: Introduction to Regression



Announcements and reminders

- HW 4 due **Today! at 5 PM**
- HW 5 will be posted soon



Statistical modeling

So far, we've talked about…

- Descriptive statistics: "this is the way my sample ***is***"
- Inferential statistics: "This is what I can **conclude** from my sample [with 100(1-α)% **confidence**]"

Today: **predictive statistics**

Linear regression for prediction

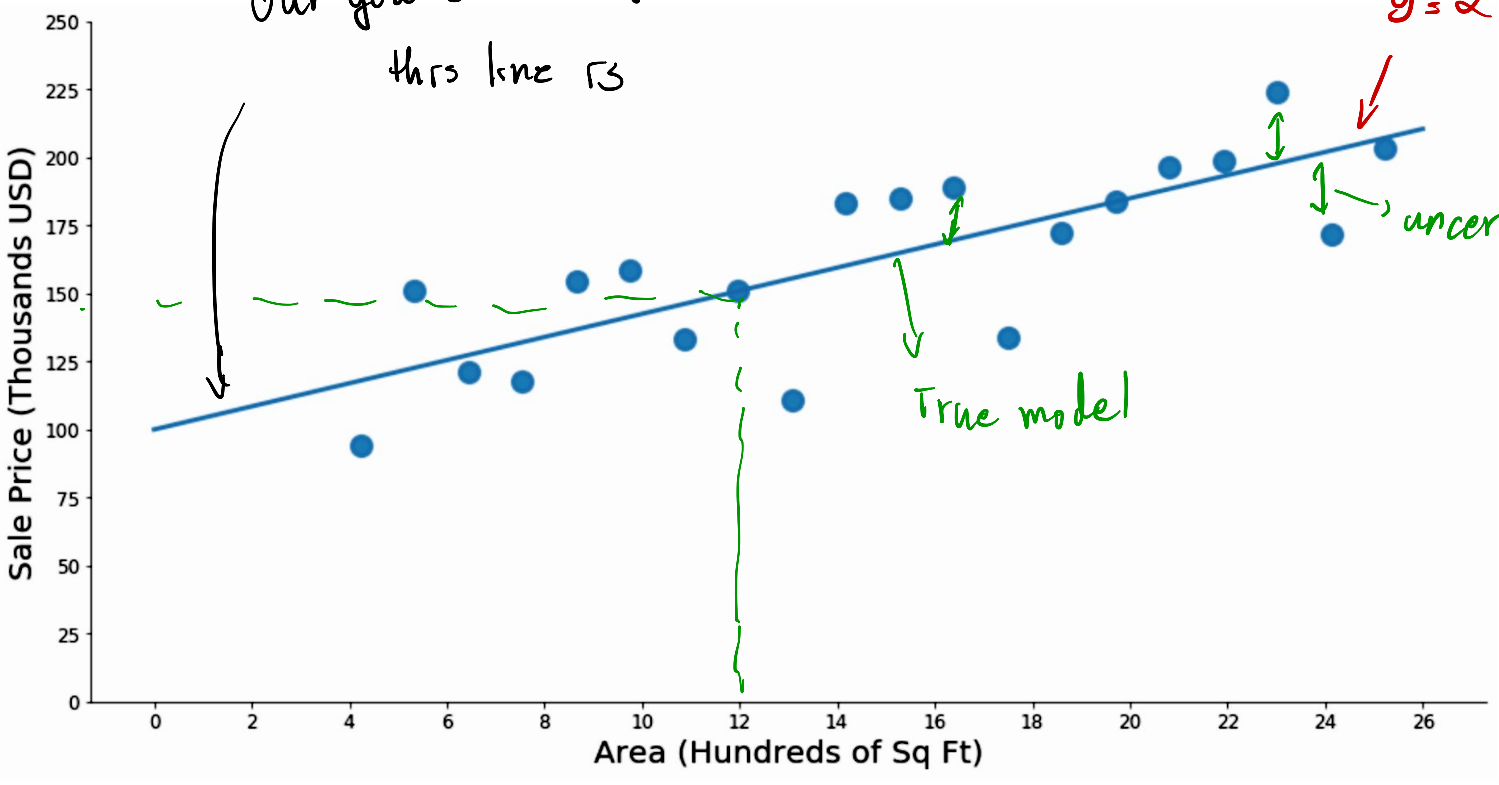
Examples:

- Given a person's age and gender, predict their height
- Given the area of a house, predict its sale price
- Given unemployment, inflation, number of wars and economic growth, predict the president's approval rating
- Given a person's browser history, predict how long they'll stay on a product page
- Given the advertising budget expenditures in various media markets, predict the number of products they'll sell

Area as predictor for house price

our goal is to figure out what the equation of this line is

$y = \alpha + \beta x$



uncertainty

True model

Simple linear regression (SLR) model

Definitions and **Assumptions** of SLR model:

1.

2.

3.

$y = \alpha + \beta x + \varepsilon$ characterize uncertainty

If she loves you more each and every day,
by linear regression she hated you before you met.

Simple linear regression (SLR) model

Definitions and **Assumptions** of SLR model:

1. $y_i = \alpha + \beta x_i + \epsilon_i$ ← i is index of a data point
2. Each of the ϵ_i are independent

   uncertainties residuals are independent of another
3. $\epsilon_i \sim N(0, \sigma^2)$

If she loves you more each and every day,
by linear regression she hated you before you met.

Simple linear regression (SLR) model $Y = \alpha + \beta X + \epsilon$

SLR model **vocabulary**:

- X: the independent variable, the predictor, the explanatory variable, the **feature**
- Y: the dependent variable, the **response** variable thing we want to predict
- ϵ: the random deviation or **random error**

latex: \epsilon

Question: What is ϵ doing?

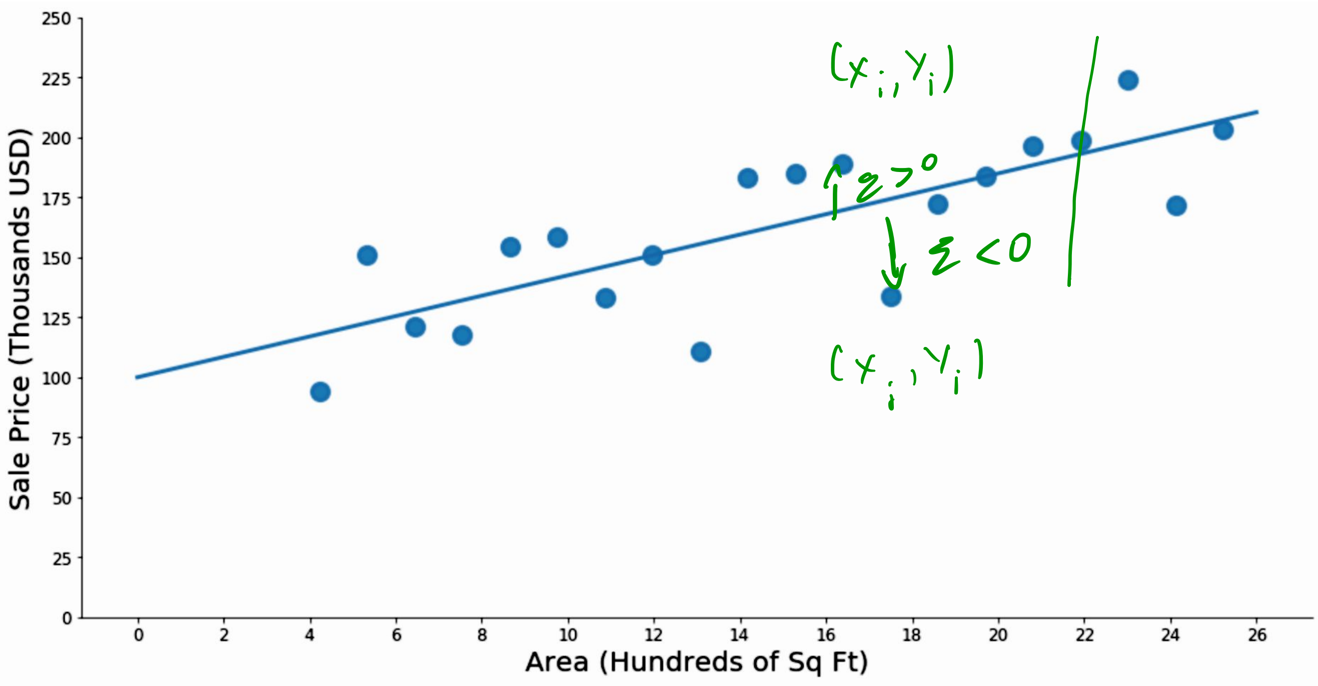
accounting for the fact that the world isn't certain; that there are random deviations around the true process

Simple linear regression (SLR) model

$Y = \alpha + \beta X + \epsilon$

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

The points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ resulting from n independent observations will be scattered about the true regression line

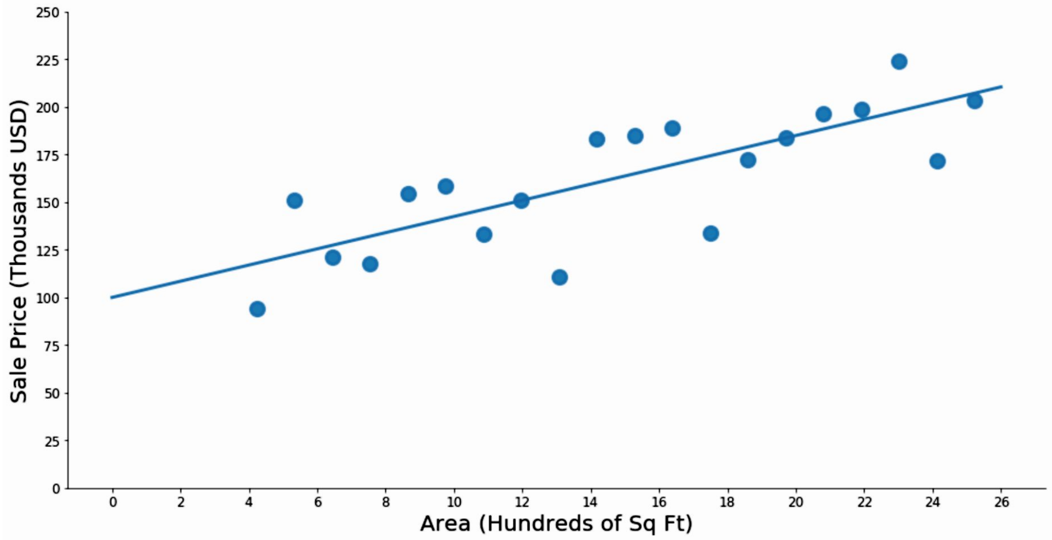


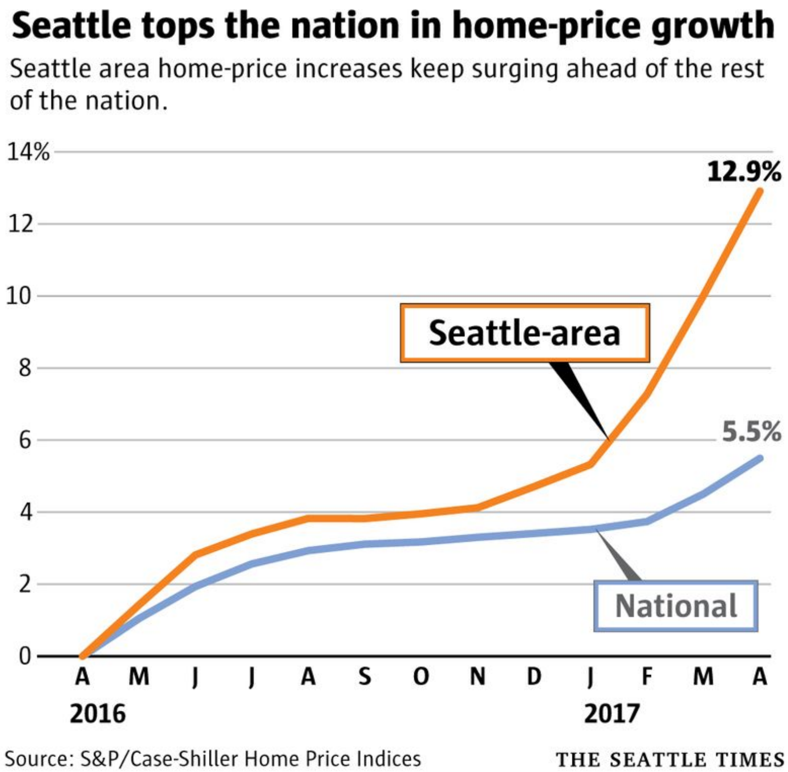
Simple linear regression (SLR) theory

$Y = \alpha + \beta X + \epsilon$

Question: how do we know that the SLR model is appropriate?

* experience
* eyeball metric
* later: R^2 & R^2 adj





Interpreting SLR parameters

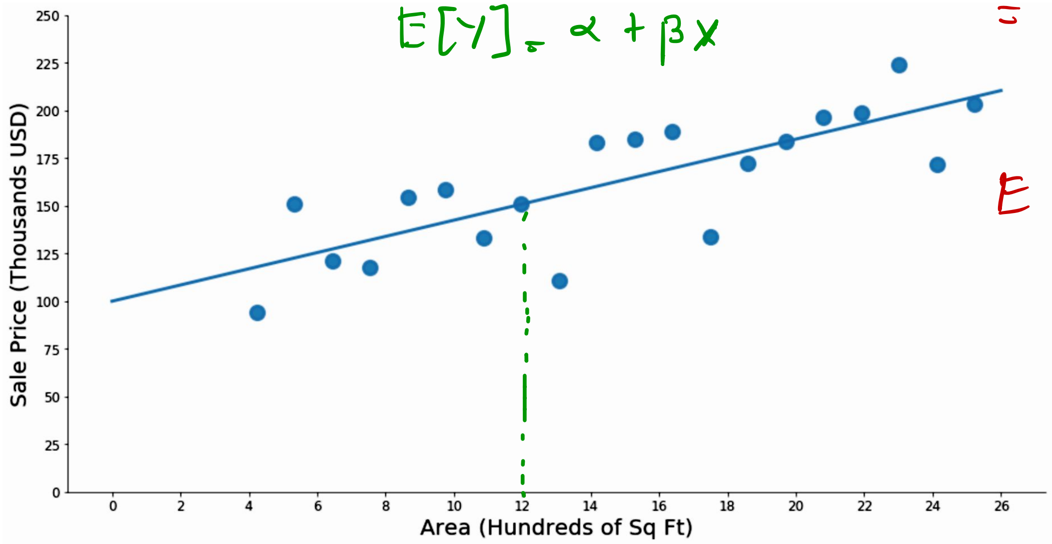
$Y = \alpha + \beta X + \epsilon$

Y is a **random variable**. → What is its **expectation**?

$$E[Y] = E[\alpha + \beta X + \varepsilon]$$

$$= \underbrace{E[\alpha]}_{\alpha} + \underbrace{E[\beta X]}_{\beta X} + \underbrace{E[\varepsilon]}_{0}$$

$$E[Y] = \alpha + \beta X$$



Interpreting SLR parameters

$Y = \alpha + \beta X + \epsilon$

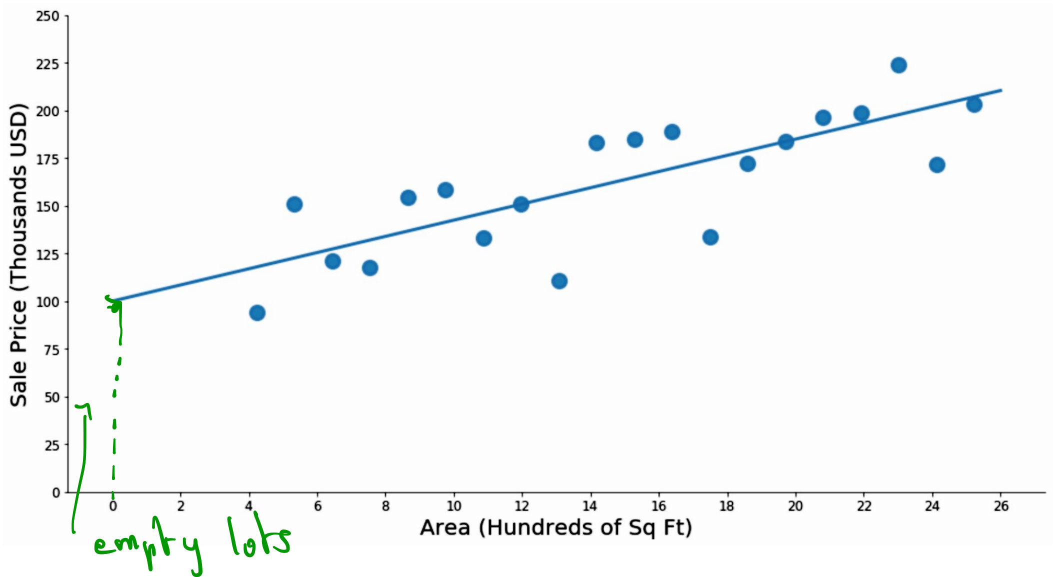
α is the intercept of the true regression line (i.e., the baseline average)

Why "baseline"?

response when feature = 0

$Y = \alpha + \beta \cdot 0$ feature = 0

$\alpha = 100$



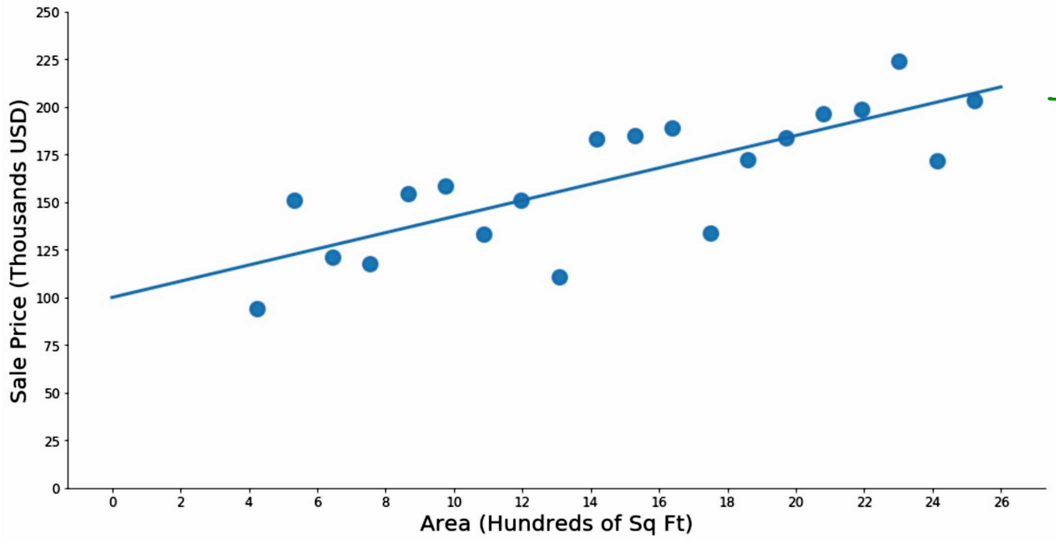
Interpreting SLR parameters

$Y = \alpha + \beta X + \epsilon$

β is the slope of the true regression line

β = increase in our response from a unit increase in our feature

$\beta > 0 \rightarrow$ feature $\rightarrow$ positive about response

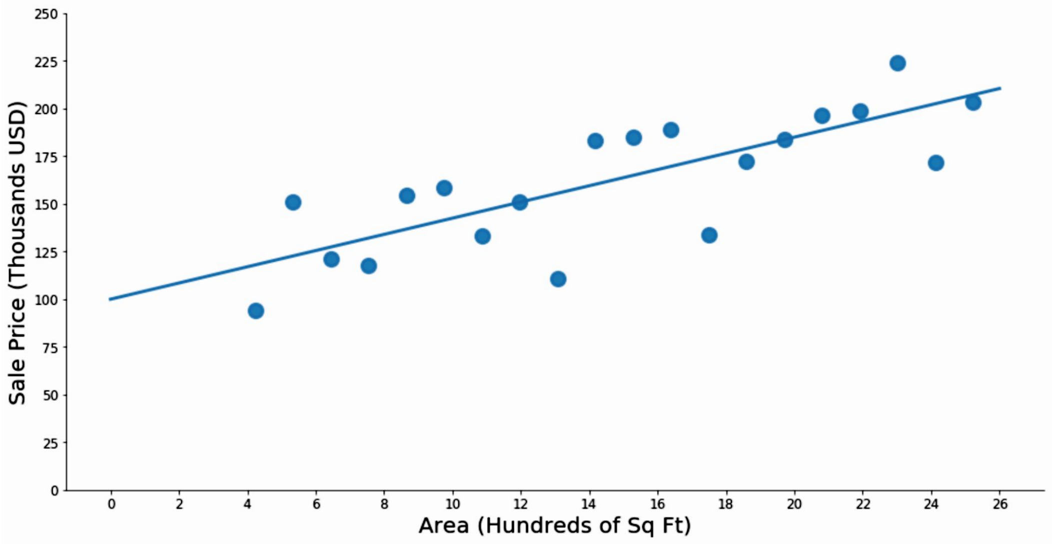


$$
\begin{aligned}
& Y(x+1) = \alpha + \beta(x+1) + \varepsilon \\
- \; & Y(x) = \alpha + \beta x + \varepsilon \\
\hline
& = \beta(x+1) - \beta x \\
& = \beta x + \beta - \beta x = \beta
\end{aligned}
$$

Interpreting the error term

$$Y = \alpha + \beta X + \epsilon$$

The variance parameter σ^2 determines the extent to which each normal curve spreads about the true regression line



Exploration

Get in groups, open up your computer and load the in-class notebook **nb 20**

Let's have a look at how these uncertainties affect our linear regression.

- Work through Exercise 1 -- **fitting noisy lines**

Directional considerations

So far, we've come up with a framework where we can choose the model parameters and then generate random data.

→ Called a **generative model**

But we *really* want to run this process in reverse. We have data, and we want to **find/learn/estimate** the parameters that explain the data.

→ **Inference!**

$\text{Pois}(\hat{\lambda})$ → resamples

General Model + Parameters → Data

General Model + Parameters ← $\boxed{\text{Data}}$ $(x_1, y_1)\ (x_2, y_2) \ldots$

α? β? σ^2?

How can we estimate parameters from some data?

Game plan: The variance of our model σ^2 will be smallest if the differences between the estimate of the true regression line and each point is the smallest.

→ This is the goal: **minimize σ^2**

We will use our sample data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ to estimate the parameters of the regression line.

Super duper important question: What are we assuming about each of the observations?

(x_i, y_i) data points $\square$ collected indep. of the other

Estimating model parameters

So we've got data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

And we want to figure out our regression line: $y_i = \alpha + \beta x_i$

… how could we minimize **residuals**, $\epsilon_i = y_i - (\alpha + \beta x_i)$?

pick α & β so this $(\alpha + \beta x_i)$ is close to Y_i

sum of squared errors:

$\hat{Y}_i$ = model's prediction is using

$$SSE = \sum_{i=1}^{n} \left(Y_i - \underbrace{(\hat{\alpha} + \hat{\beta} x_i)}_{\hat{Y}_i}\right)^2$$

Estimating model parameters

$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$

Definition: The **sum of squared-errors** for the points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ to the regression line is given by

$$\boxed{SSE = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2}$$

Definition: The point-estimates (single value estimates from the data) of the slope and intercept parameters are called the **least-squares estimates**, and are defined to be the values that minimize the SSE:

↳ take deriv.

↳ set it = 0

↳ solve for $\hat{\alpha}$ & $\hat{\beta}$

True variable : α & β

$f(x, y) = 2x^2 + xy^2$

$\frac{\partial f}{\partial y} = 0 + 2xy$ // $\frac{df}{dx} = 4x^2 + y^2$

Estimating model parameters

Definition: The **fitted regression line** or the **least-squares line** is then the line given by:

$$\hat{y} = \hat{\alpha} + \hat{\beta} x$$

Question: Great! So… how do we actually find the parameter estimates?

→ minimize SSE w.r.t α & β

$$SSE = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

→ $\dfrac{\partial SSE}{\partial \alpha} = 0$ & $\dfrac{\partial SSE}{\partial \beta} = 0$

1 2

Estimating model parameters

$$SSE = \sum_{i=1}^{n} (Y_i - (\alpha + \beta x_i))^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

$$\frac{\partial SSE}{\partial \alpha} = \sum_{i=1}^{n} \frac{\partial}{\partial \alpha}\left[(\underbrace{Y_i}_{0} - \alpha - \underbrace{\beta x_i}_{0})^2\right]$$

$$= \sum_{i=1}^{n} 2(Y_i - \alpha - \beta x_i)(-1) = \frac{-2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)}{-2} \overset{(!)}{=} \frac{0}{-2}$$

$$\rightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \alpha - \sum_{i=0}^{n} \beta x_i \overset{(!)}{=} 0$$

$$\rightarrow \sum_{i=1}^{n} Y_i - \alpha \sum_{i=1}^{n} 1 - \beta \sum_{i=1}^{n} x_i \overset{(!)}{=} 0$$

$$\rightarrow \sum Y_i - \alpha n - \beta \sum x_i = 0$$

→ super close to $\bar{y}$ & $\bar{x}$

→ just divide by n

$$\frac{1}{n}\sum y_i - \alpha - \beta \frac{1}{n} \sum x_i = 0$$

$$\bar{Y} - \alpha - \beta \bar{x} = 0$$

$$\Rightarrow \boxed{\alpha = \bar{Y} - \beta \bar{x}}$$

Estimating model parameters

So. How does this work in practice? What can we do?!

… let’s crack open **nb 20** and find out!

- Work through Exercise 2 -- **fitting least squares regression lines**

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

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

Residuals

- The **fitted** or **predicted values** $\hat{Y} = \hat{\alpha} + \hat{\beta} x$ are obtained by plugging in the independent data variables into the fitted model.

- The **residuals** are the difference between the observed and the predicted responses:

  $r_i = Y_i - \hat{Y}_i$

- **Claim:** The residuals r_i are estimates of the (unknown) **true error** ϵ_i

  $Y_i = \alpha + \beta x_i + \underline{\underline{\varepsilon_i}}$

  $r \approx \varepsilon \sim N(0, \alpha^2)$

Maximum likelihood estimation

- An alternative method for estimating model parameters is to create a **likelihood function** that quantifies the **goodness-of-fit** between the model and the data, and choose the values of the parameters that maximizes it

- Turns out, we've done this before! But we didn't call it Maximum Likelihood Estimation at the time.

what is $\hat{p}$? → $\hat{p} = \frac{5}{6}$?

Example: S'pose you have a biased coin. You flip it 6 times, and get 5 Heads and 1 Tails. Estimate the parameter p for the coin.

likelihood fcn = prob. of observing our data, assuming our s'posed parameter(s) are true

$$L(Y|P) = P(5H, 1T | P) = \binom{6}{5} p^5 (1-p)^1 = 6p^5(1-p) = 6p^5 - 6p^6$$

indep. variable

$$\frac{dL}{dp} = \frac{30p^4}{36p^5} - \frac{36p^5}{36p^5} \stackrel{set}{=} 0 \longrightarrow p = \frac{30}{36} = \frac{5}{6}$$