CSCI 3022-002 Intro to Data Science

Regression

$y = mx + b$ NO

Consider the graph below. Can you draw a straight line that passes through each and every point? What about a parabola?

$y = ax^2 + bx + c$

3

(a can be/not be 0)

pt 4?

(19, 12)

3

Mathematical Interpolation

These problems are known as *interpolation*. We actually don't care about them! As statisticians we ask a different problem: how can we *best* draw a line through data values if the process that's generated them is subject to **error**?

In that sense, what we do in data science is we take data that's highly variable - or noisy - and try to describe it with a simpler process. "Simpler" here means a couple of things: relying on less parameters/estimations, and nicer mathematically. The latter is often referred to as **smooth**, and it should come as no surprise that data scientists are **smooth** to the core.

Regression is the way that we smooth out data: take something that's noisy and has *lots* of observations and try to describe it with a single line!

Announcements and Reminders

- Short HW for next week. (posted w?)
- NB day Friday.

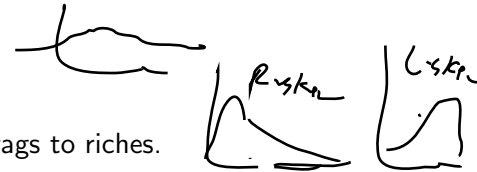
Where we at?

We're "done" with CI's and hypothesis testing (as standalones...)!
If:

1. We're estimating means:
   - 1.1 Underlying normal: use t
   - 1.2 Large sample: use z
   - 1.3 Small sample, no assumptions: bootstrap.

) if both: use t. ($\approx z$)
2. We're not estimating means:
   - 2.1 Underlying normal: we can use χ^2 to make CIs/tests on variances for one sample or F to make CIs/tests on comparisons of variances. (**Not done, we'll talk about these later!**)
   - 2.2 Otherwise, we bootstrap!

Bootstrapping Recap



Bootstrapping is an attempt to bring your data set from rags to riches.

When we aren't making assumptions like "underlying normal," we mean that **shape matters**. Bootstrapping is a way to understand how much the shape of a (maybe small) sample matters!

1. Take subsamples - with replacement! - of your original sample
2. Compute descriptive statistics from your subsample: anything from means, variances, maxima, to .72nd quantiles, and use those to get a feel for how those parameters behave from the underlying population!
3. **Example**: what's your best guess for the mean rainfall of Boulder in the last 10 years? Variance?

Bootstrapping

This process: simulating a data set, calculating a desired *sample statistic* from it, and then creating a *distribution* of that sample statistic is called a *non-parametric* bootstrap since it doesn't make distributional assumptions.

Definition: *parametric* statistics assume that sample data comes from a population that follows a probability distribution on a fixed set of parameters.

Examples:

1. μ and σ are the parameters of the Normal distribution.
2. λ is the parameters of the Poisson and Exponential distributions.
3. p is the parameter of the geometric and Bernoulli distributions.

Parametric Bootstrapping

Sometimes we really want to know about various statistics on e.g. the Poisson or Exponential *without* solving some challenging integral or sum or whatever else equations.

CI for λ of exponential.

Definition: *parametric* bootstraps estimate a CI for a desired property in two steps.

1. Estimate the parameters of the known distribution from your sample. estimate $\hat{\lambda}$
2. Draw bootstrap resamples from the distribution, *assuming* the estimated parameter
3. Compute a CI for the desired property from your resamples.

simulate from exp $(\hat{\lambda})$

note $E[X] = 1/\lambda$

$\hat{\lambda} = \frac{1}{\bar{x}}$?

Parametric Bootstrapping

Example: If we want to estimate the median of a sample that we *assume* is Poisson, we might:

1. Assume the data is Pois(λ). Estimate the parameter, e.g. $\lambda \approx \bar{X}$.
2. Simulate a bootstrapped sample from Pois($\bar{X}$).
3. Create a CI for the median from that pool of bootstrapped samples.

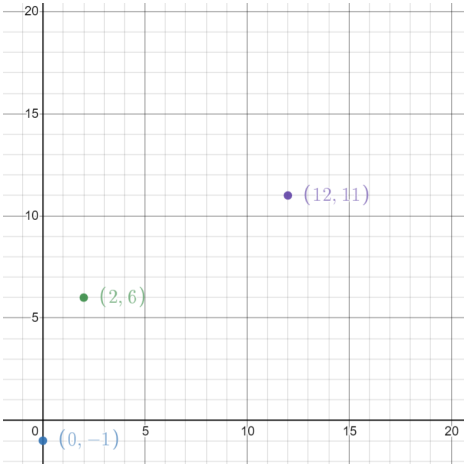
b/c $E[X] = \lambda$ for Poisson

Why make *more* assumptions, like assuming the distribution of the random variable at all? The advantage of the parametric bootstrap is that it can be shown to do a better job in particular scenarios. (if we're right)

The downside? The parametric bootstrap does a very poor job if the population does not have the same population as you assumed. This is called *model misspecification*, and is a risk **any** time we assume things have **any** underlying distribution, including in hypothesis testing!

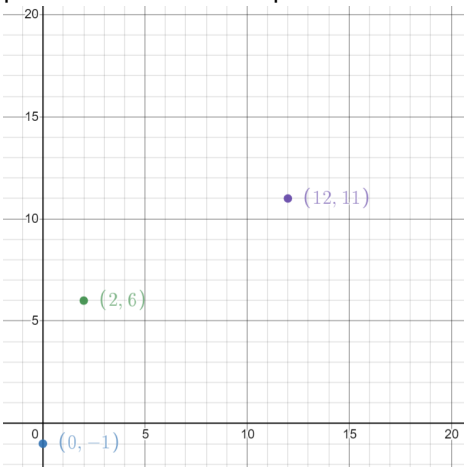
Opening

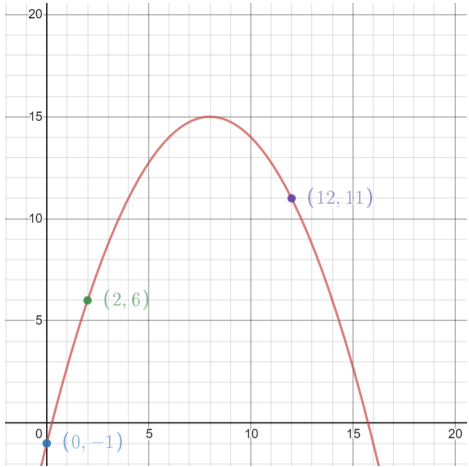
Consider the graph below. Can you draw a straight line that passes through each and every point? What about a parabola?



Opening

Consider the graph below. Can you draw a straight line that passes through each and every point? What about a parabola?





Mathematical Interpolation

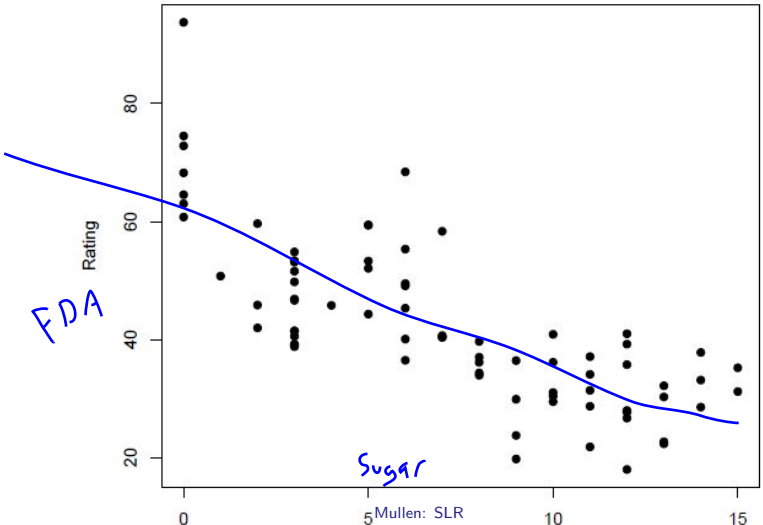
These problems are known as *interpolation*. We actually don't care about them! As statisticians we ask a different problem: how can we *best* draw a line through data values if the process that's generated them is subject to **error**?

In that sense, what we do in data science is we take data that's highly variable - or noisy - and try to describe it with a simpler process. "Simpler" here means a couple of things: relying on less parameters/estimations, and nicer mathematically. The latter is often referred to as **smooth**, and it should come as no surprise that data scientists are **smooth** to the core.

Regression is the way that we smooth out data: take something that's noisy and has *lots* of observations and try to describe it with a single line!

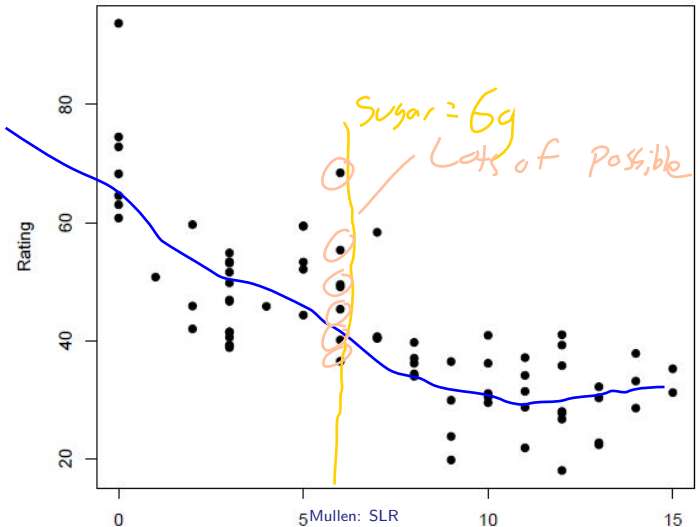
Simple Linear Regression

We see bivariate data: a set of points (X_i, Y_i).



Simple Linear Regression

...and we draw the "best" line through them.



Simple Linear Regression Model

$y = mx + b$

The simplest mathematical relationship between two variables x and y is a linear relationship:

Simple Linear Regression Model

The simplest mathematical relationship between two variables x and y is a linear relationship:

$$y = mx + b$$

The objective of this section is about equivalent linear probabilistic models.

Simple Linear Regression Model

The simplest mathematical relationship between two variables x and y is a linear relationship:

$$y = mx + b$$

The objective of this section is about equivalent linear probabilistic models.

If two random variables are probabilistically related, then for a fixed value of x, there is uncertainty in the value of the second variable.

Rating given Sugar is random

Simple Linear Regression Model

The simplest mathematical relationship between two variables x and y is a linear relationship:

$$y = mx + b$$

The objective of this section is about equivalent linear probabilistic models.

If two random variables are probabilistically related, then for a fixed value of x, there is uncertainty in the value of the second variable.

So we assume where ε is a random variable and: MODEL

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

$y = b + mx +$ error / probability

Simple Linear Regression Model

y depends on X

Definition: *Simple Linear Regression* (SLR)

The *Simple Linear Regression* model is a model of the form y random GIVEN

$$y = \underbrace{\text{intercept} + \text{slope} \cdot x}_{\text{non-random } x.} + \text{errors}$$

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

With 3 assumptions on ε:

iid 2 for 1!

Independence

Identical Dist

Normality

Y given X is random

Simple Linear Regression Model

Definition: *Simple Linear Regression* (SLR)

The *Simple Linear Regression* model is a model of the form

1.

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

With 3 assumptions on ε:

Simple Linear Regression Model

Definition: *Simple Linear Regression* (SLR)

The *Simple Linear Regression* model is a model of the form

1. $$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

   With 3 assumptions on ε:

2. $\text{Cov}[\varepsilon_i, \varepsilon_j] = 0 \qquad \forall i, j$

   Independence of errors

Covariance = 0 ⇔ independence

"Knowledge of error of observation i tells us nothing about the error of observation j"

Result: Data probably not X = time.

Simple Linear Regression Model

Definition: *Simple Linear Regression* (SLR)

The *Simple Linear Regression* model is a model of the form

1. $$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

   With 3 assumptions on ε:

2. $$\text{Cov}[\varepsilon_i, \varepsilon_j] = 0 \qquad \forall i, j$$

   Independence of errors

3. $$Var(\varepsilon_i) = \sigma^2 \qquad \forall i$$

   Homoskedasticity of errors

Identical errors

"size" of errors is their variance.

Simple Linear Regression Model

Definition: *Simple Linear Regression* (SLR)

The *Simple Linear Regression* model is a model of the form

1. LINE IS CORRECT

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

With 3 assumptions on ε:

2.

$$\text{Cov}[\varepsilon_i, \varepsilon_j] = 0 \qquad \forall i, j$$

Independence of errors

3.

$$Var(\varepsilon_i) = \sigma^2 \qquad \forall i$$

Homoskedasticity of errors

4. Normality for all i,

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

mean 0: line is correct model

Simple Linear Regression Model

Important Terminology:

- x: the independent variable, predictor, or explanatory variable (usually known). x is not random.
- Y: The dependent variable or response variable. For fixed x, Y is random.
- ε: The random deviation or random error term. For fixed x, ε is random.

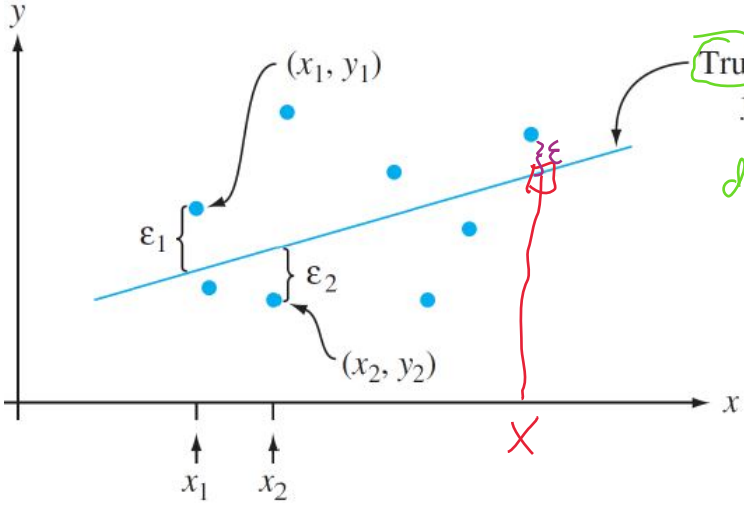
What exactly does ε do?

Simple Linear Regression: ε

Model: $y = \beta_0 + \beta_1 x + \varepsilon$

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

Reality:

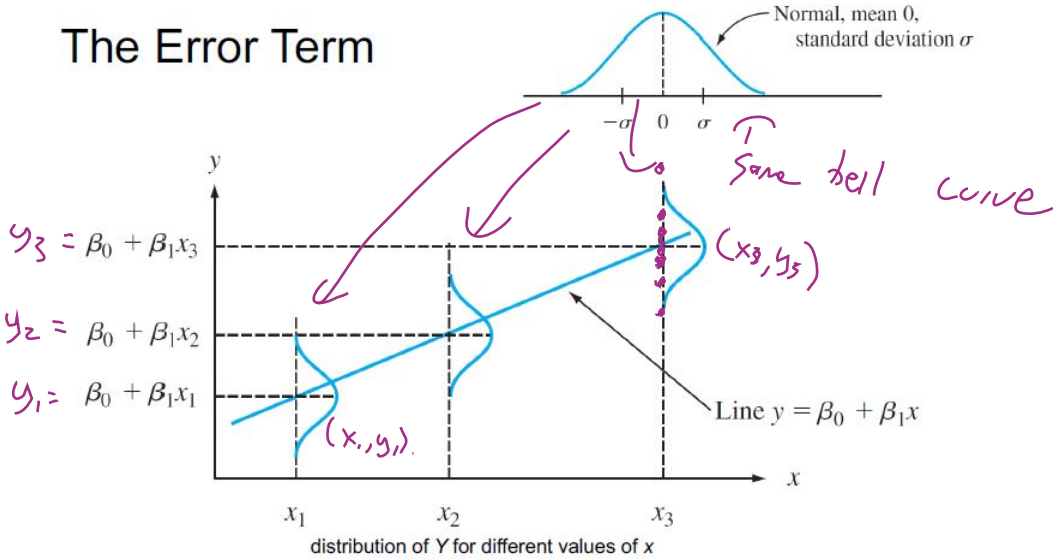


data comes from the line

ε_i: vertical distance between true line and the ith data value.

But we "assumed" a certain distribution here: $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

Simple Linear Regression: ε



Simple Linear Regression Model

In practice, using a linear model largely comes down to diagnosing and fixing deviations from these assumptions. What does that look like?

1. **Linearity:** Is a *straight* line actually appropriate? $y = \beta_0 + \beta_1 x$ or? close?
2. **Independence:** Are there patterns or structures in my data not covered by a straight line? clumps of over/under estimates.
3. **Homoskedasticity:** Do the sizes of my errors vary depending on where I am in the data set? Larger x, larger y?
4. **Normality:** Do the errors appear to be coming from a normal distribution?

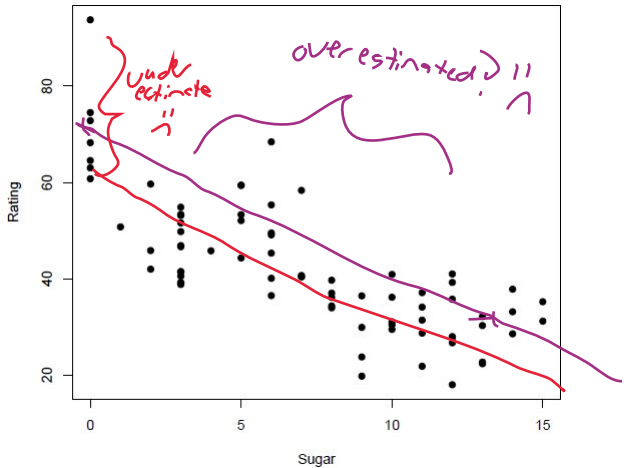
Simple Linear Regression: Assumptions

How do we know when the SLR model is appropriate?

Scatter plots

Is line ok?

Assumptions



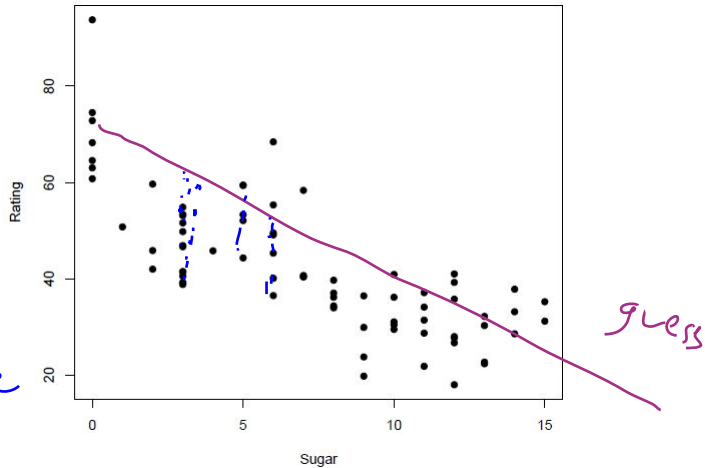
Simple Linear Regression: Assumptions

How do we know when the SLR model is appropriate?

Scatter plots
To check *Linearity*

Assumptions
To check assumptions on *errors*, we need to plot and test errors

depend on our line



guess

Simple Linear Regression: Interpretations

Interpreting parameters:

$y = \beta_0 + \beta_1 x + \text{error}$

$x = 0$

- β_0 : the intercept of the true regression line.
  The average value of Y when x is zero.
  Usually this is called the "baseline average".

- β_1 : the slope of the true regression line.
  The average change in Y associated with a 1-unit increase in the value of x.

Estimating SLR Parameters

Goal:

Given sample data, which consists of n observed pairs, $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$, construct an estimated "line of best fit":

This line can then be used to make predictions or provide explanations for unobserved phenomena.

How do we construct this line?

Estimating SLR Parameters

Goal:

Given sample data, which consists of n observed pairs, $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$, construct an estimated "line of best fit":

our intercept

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

our estimated slope

This line can then be used to make predictions or provide explanations for unobserved phenomena.

How do we construct this line?
Find estimates of the β values with nice statistical properties, and call them $\hat{\beta}$!

SLR Parameters

What are the properties of our ith data point, Y_i?

Consider $E[Y_i]$:

$$E[y_i] = E[\underbrace{\beta_0 + \beta_1 x_i}_{\text{not random}} + \underbrace{\varepsilon}_{\text{random}}] = \beta_0 + \beta_1 x_i + E[\varepsilon]$$

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

Consider $Var[Y_i]$:

What if we got a new data point, like $X_{new} = X_i + 1$?

SLR Parameters

What are the properties of our ith data point, Y_i?
Consider $E[Y_i]$:

$$E[Y_i] = E[\beta_0 + \beta_1 X_i + \varepsilon] = \beta_0 + \beta_1 X_i + E[\varepsilon] = \beta_0 + \beta_1 X_i$$

Consider $Var[Y_i]$:

$$Var[Y_i] = Var[\beta_0 + \beta_1 X_i + \varepsilon] = Var[\varepsilon] = \sigma^2$$

Only random bit!

What if we got a new data point, like $X_{new} = X_i + 1$?

$$E[Y_{new}] = E[\beta_0 + \beta_1 X_{new} + \varepsilon] = \beta_0 + \beta_1 X_{new} + E[\varepsilon] = \beta_0 + \beta_1 (X_i + 1) =$$

$$= \beta_0 + \beta_1 X_i + \beta_1$$

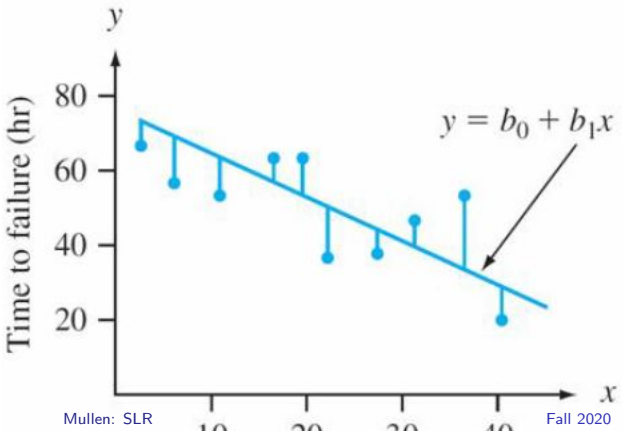
old Y value plus slope.

So we interpret β_1 as a regular slope: it's the *expected* change in Y per 1 unit change in X.

Estimating SLR Parameters

One way to define "best fit" line is motivated by the principle of least squares, which can be traced back to the German mathematician Gauss (1777–1855):

A line provides the **best fit** to the data if the sum of the squared vertical distances (deviations) from the observed points to that line is as small as it can be.



Estimating SLR Parameters

The sum of *squared vertical deviations* from the data points to the line $y = \beta_0 + \beta_1 x$ is then

GOAL:

Estimating SLR Parameters

"best" line: $\hat{\beta}_0 + \hat{\beta}_1 x$

The sum of *squared vertical deviations* from the data points to the line $y = \beta_0 + \beta_1 x$ is then

errors: $y - (\hat{\beta}_0 + \hat{\beta}_1 x)$

$$\sum_{i=1}^{n} \left(\overbrace{Y_i}^{\text{Data}} - \overbrace{\beta_0 - \beta_1 X_i}^{\text{Line}} \right)^2$$

Squared

sum

deviations

Estimating SLR Parameters

The sum of *squared vertical deviations* from the data points to the line $y = \beta_0 + \beta_1 x$ is then

$$\sum_{i=1}^{n} \left(\overbrace{Y_i}^{\text{Data}} - \overbrace{\beta_0 - \beta_1 X_i}^{\text{Line}} \right)^2$$

The point estimates of β_0 and β_1, denoted $\hat{\beta}_0; \hat{\beta}_1$ are called the *least squares estimates*. They are those values that minimize SSE or sum of squared errors.

Estimating SLR Parameters

The sum of *squared vertical deviations* from the data points to the line $y = \beta_0 + \beta_1 x$ is then

$$\sum_{i=1}^{n} \left(\overbrace{Y_i}^{\text{Data}} - \overbrace{\beta_0 - \beta_1 X_i}^{\text{Line}} \right)^2$$

The point estimates of β_0 and β_1, denoted $\hat{\beta}_0; \hat{\beta}_1$ are called the *least squares estimates*. They are those values that minimize SSE or sum of squared errors.

$$\hat{\beta}_0, \hat{\beta}_1 = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Estimating SLR Parameters: Pen and Paper

Goal: find the minimizers of the function $f(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$. Sounds like a Calculus problem!

Estimating SLR Parameters: Pen and Paper

Goal: find the minimizers of the function $f(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$. Sounds like a Calculus problem!

(two of 'em)

derivative

$$\frac{df}{d\beta_0} = \frac{d}{d\beta_0} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$ set $= 0$

$$\frac{df}{d\beta_1} = \frac{d}{d\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$ set $= 0$

2 eqns, 2 unknowns (β_0, β_1).

Estimating SLR Parameters: Pen and Paper

Goal: find the minimizers of the function $f(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$. Sounds like a Calculus problem!

$$\frac{df}{d\beta_0} = \frac{d}{d\beta_0} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

$$\frac{df}{d\beta_1} = \frac{d}{d\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

For finding the joint maximum/minimum of multiple inputs, we end up with a system of equations: set both equal to zero and find the values that make both equal to zero.

Estimating SLR Parameters: Results

For a model of the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$; $\varepsilon \sim N(0, \sigma^2)$

1. $\hat{\beta}_0 =$

2. $\hat{\beta}_1 =$

What happens if $\beta_0 \approx 0$? If $\beta_1 \approx 0$?

Estimating SLR Parameters: Results

For a model of the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$; $\varepsilon \sim N(0, \sigma^2)$

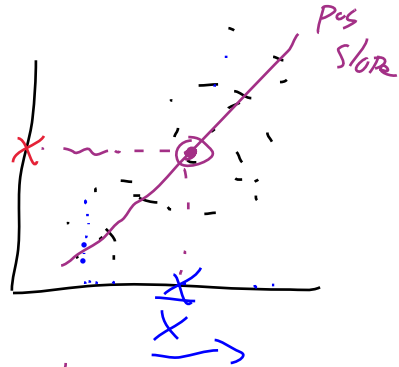
1. $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
2. $\hat{\beta}_1 = \frac{Cov[X,Y]}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

slope! (later)

What happens if $\beta_0 \approx 0$? If $\beta_1 \approx 0$?

Estimating SLR Parameters: Results

For a model of the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$; $\varepsilon \sim N(0, \sigma^2)$



1. $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
2. $\hat{\beta}_1 = \frac{Cov[X,Y]}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

Numerator looks like $Cov(X,Y)$. + if X,Y move together
looks like $Var[X]$.
− if X,Y move apart (as X↑, Y↓)

What happens if $\beta_0 \approx 0$? If $\beta_1 \approx 0$?
One result: the regression line goes through $(0, \beta_0)$. It also goes through $(\bar{X}, \bar{Y})$!

Estimating SLR Parameters

Definitions:

1. The *fitted (or predicted) values* __ are obtained by plugging in __ to the equation of the estimated regression line:

2. The *residuals* are the differences between the observed and fitted y values:

Residuals are estimates of the true error. Why?

Estimating SLR Parameters

Definitions:

1. The *fitted (or predicted) values* $\hat{Y}_i$ are obtained by plugging in $\hat{X}_i$ to the equation of the estimated regression line:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

2. The *residuals* are the differences between the observed and fitted y values:

Residuals are estimates of the true error. Why?

Estimating SLR Parameters

Definitions:

1. The *fitted (or predicted) values* __ are obtained by plugging in __ to the equation of the estimated regression line:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

2. The *residuals* are the differences between the observed and fitted y values:

$$\hat{\varepsilon_i} = r_i = \hat{e_I} = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 + \hat{beta}_1 X_i$$

Residuals are estimates of the true error. Why?

Estimating SLR Parameters

Definitions:

1. The *fitted (or predicted) values* __ are obtained by plugging in __ to the equation of the estimated regression line:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

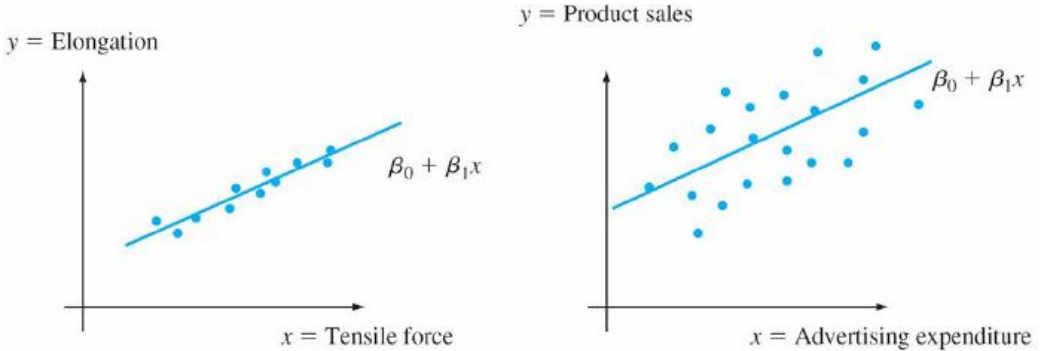
2. The *residuals* are the differences between the observed and fitted y values:

$$\hat{\varepsilon_i} = r_i = \hat{e_I} = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 + \hat{beta_1} X_i$$

Residuals are estimates of the true error. Why?
We don't have the true values of β_0, β_1, so when we estimate them we get variance and error in our estimates.

Estimating SLR Parameters: σ^2

The parameter σ^2 determines the amount of spread about the true regression line. Two separate examples:



Estimating SLR Parameters: σ^2

An estimate of σ^2 will be used in confidence interval formulas and hypothesis testing procedures presented in the next days. Recall that the residual sum of squares or sum of squared errors (SSE) is:

$$\text{SSE} =$$

So, our estimate of the variance of the model is like a measure for an average of this summand:

Estimating SLR Parameters: σ^2

An estimate of σ^2 will be used in confidence interval formulas and hypothesis testing procedures presented in the next days. Recall that the residual sum of squares or sum of squared errors (SSE) is:

$$\text{SSE} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2$$

So, our estimate of the variance of the model is like a measure for an average of this summand:

Estimating SLR Parameters: σ^2

An estimate of σ^2 will be used in confidence interval formulas and hypothesis testing procedures presented in the next days. Recall that the residual sum of squares or sum of squared errors (SSE) is:

$$\text{SSE}=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}\left(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i\right)^2$$

So, our estimate of the variance of the model is like a measure for an average of this summand:

$$\hat{\sigma^2}=\frac{SSE}{n-2}$$

Wait, what? Why the $n-2$??

Estimating SLR Parameters: σ^2

An estimate of σ^2 will be used in confidence interval formulas and hypothesis testing procedures presented in the next days. Recall that the residual sum of squares or sum of squared errors (SSE) is:

$$\text{SSE}=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}\left(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i\right)^2$$

So, our estimate of the variance of the model is like a measure for an average of this summand:

$$\hat{\sigma^2}=\frac{SSE}{n-2}$$

Wait, what? Why the $n-2$??
These are again *degrees of freedom*.

Degrees of Freedom Intuition

Suppose you have 3 (random) points on the XY plane.

1. Can you draw a line through them?

2. Can you draw a parabola through them?

3. Can you draw a cubic function through them?

4. Can you draw a quartic function through them?

Degrees of Freedom Intuition

Suppose you have 3 (random) points on the XY plane.

1. Can you draw a line through them?
   It's very unlikely. In fact, for truly random (normal) points, this result has probability zero!
2. Can you draw a parabola through them?
   It's very unlikely. In fact, for truly random (normal) points, this result has probability zero! Yes, but there's only one such parabola.
3. Can you draw a cubic function through them?
   Yes. Not only that, you could choose *any one* of a, b, c, d in the $ax^3 + bx^2 + cx + d = 0$ and then solve for the others. You have **one degree of freedom**.
4. Can you draw a quartic function through them?
   Yes. Not only that, you could choose *any two* of a, b, c, d, e in the $ax^4 + bx^3 + cx^2 + dx + e = 0$ and then solve for the others. You have **two degrees of freedom**.

Degrees of Freedom

The takeaway?

One property of mathematical estimation: the more you estimate, the more you risk *overfitting*. In this model we've estimated **2** means ($\hat{\beta}_0$,$\hat{\beta}_1$) before we got to σ, which "costs" us two degrees of freedom.

The more we estimate, the less options - degrees of freedom - we get for the remaining terms.

Estimating SLR Parameters: σ^2

Some properties of our estimate:

1. The divisor $n–2$ in is the number of degrees of freedom (df) associated with SSE and $\hat{\sigma}^2$.

2. This is because to obtain $\hat{\sigma}^2$, two parameters must first be estimated, which results in a loss of 2 df.

3. Replacing each y_i in the formula for $\hat{\sigma}^2$ by the r.v. Y_i gives a random variable.

4. It can be shown that the r.v. $\hat{\sigma}^2$ is an unbiased estimator for σ^2.

On Optimization

In any data science technique, there are two important considerations:

1. What are we optimizing? What are solving for and why?
2. How do we solve for that?
   - 2.1 Subsequent data science classes - e.g. "Advanced Data Science," "Machine Learning," etc. - involve a *lot* of algorithmic considerations: memory allocation, flop counts, etc.
   - 2.2 Do we have to approximate, or can we solve for an exact solution?

Estimating SLR Parameters: the MLE

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.

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

What's a likelihood function?

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

We want the function that gives the probability of outcomes as a function of p. In this case, that's:

$$l(p) = P(\text{data GIVEN P}) = P(\text{5 heads and one tails}|p)$$

What's a likelihood function?

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

We want the function that gives the probability of outcomes as a function of p. In this case, that's:

$$l(p) = P(\text{data GIVEN P}) = P(\text{5 heads and one tails}|p)$$

If we know p, this is a binomial, and the function is $l(p) = \binom{6}{5}p^5(1-p) = 6p^5(1-p)$. We want to find the value of p that maximizes this!

What's a likelihood function?

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

We want the function that gives the probability of outcomes as a function of p. In this case, that's:

$$l(p) = P(\text{data GIVEN P}) = P(\text{5 heads and one tails}|p)$$

If we know p, this is a binomial, and the function is $l(p) = \binom{6}{5}p^5(1-p) = 6p^5(1-p)$. We want to find the value of p that maximizes this!

$$\frac{dl}{dp} = 30p^4(1-p) - 6p^5 = p^4(30 - 30p - 6) = p^4(24 - 30p)$$

What's a likelihood function?

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

We want the function that gives the probability of outcomes as a function of p. In this case, that's:

$$l(p) = P(\text{data GIVEN P}) = P(\text{5 heads and one tails}|p)$$

If we know p, this is a binomial, and the function is $l(p) = \binom{6}{5}p^5(1-p) = 6p^5(1-p)$. We want to find the value of p that maximizes this!

$$\frac{dl}{dp} = 30p^4(1-p) - 6p^5 = p^4(30 - 30p - 6) = p^4(24 - 30p)$$

which equals zero at $p = 0$ and $p = 5/6$.

Why care?

For any problem with underlying probability distributions - a pmf or a pmf - we can typically write down a likelihood function, which often reduces our data science problem to a numerical maximization problem.

For other problems, we may instead solve a least-squares or cost minimization problem. In either case, there's some metric by which we're coming up with the *best* solution.

For simple linear regression, they provide the same values of $\hat{\beta}$! This isn't always true. For example, the MLE for σ^2 of a normal data set is $\frac{\sum(X_i - \bar{X})^2}{n}$, which is a different denominator than our usual s^2.

Daily Recap

Today we learned

1. Regression!

Moving forward:

- nb day Friday

Next time in lecture:

- More Regression!