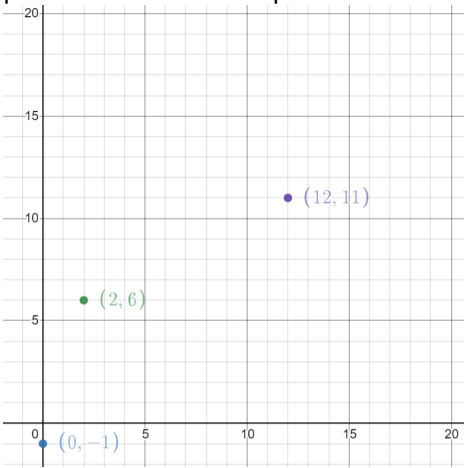
CSCI 3022 Intro to Data Science
Regression

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



We've looked at the following test statistics for hypothesis testing.

1. To compare proportions against a baseline or against each other, we use Z-statistics.

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \textbf{ OR } \frac{(\hat{p_1} - \hat{p_2}) - \Delta_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$$

2. To compare means when the samples are large **or** underlying normal with *known* variances, we also use Z-statistics.

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \textbf{ OR } \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \textbf{ OR } \frac{\left(\bar{X} - \bar{Y}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \textbf{ OR } \frac{\left(\bar{X} - \bar{Y}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}$$

3. To compare means when the samples are small **and** underlying normal, we use t-statistics.

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \textbf{ OR } \frac{\left(\bar{X} - \bar{Y}\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}$$

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 a median

Suppose we want a 90% CI for the variance of a data set. Code to **bootstrap:**

Bootstrapping a median

Suppose we want a 90% CI for the variance of a data set. Code to **bootstrap:**

1. vars=[]
   nsamp=10000
2. for i in range(nsamp):
   newX=np.random.choice(X, size=len(X), replace=True)
   vars.append(np.var(newX, ddof=1))
3. CI= np.percentile(vars, [5,95])

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.
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!

Now what?

On to Regression!

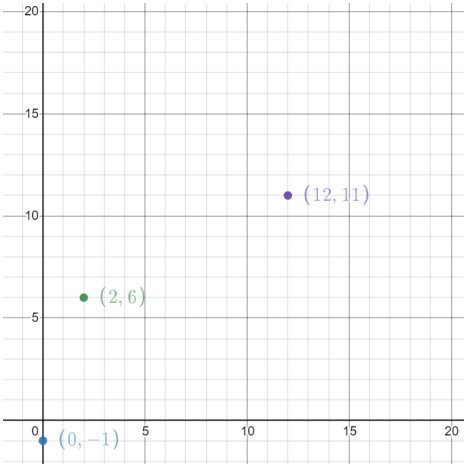
A few things to note: you are **not expected** nor even encouraged to memorize all of these formulas. Instead, you want a few basic vocabulary words and associations:

1. **Normals** are for large sample measures of the *mean* (or proportions). They are (difference)/(standard deviation) formulas.
2. **t's** are for small sample measures of the *mean*. *Assumption:* populations are normal. They are (difference)/(standard deviation) formulas.
3. **Chi-squared** are for measures of the variance. *Assumption:* of a normal. They use (sums of squared deviations) in the formula.
4. **F** are for measures of the variance. *Assumption:* of a normal. They are a *ratio* of two variances/chi-squareds.

With those associations and basic algebraic intuitions, just look up the one you need at any given time!

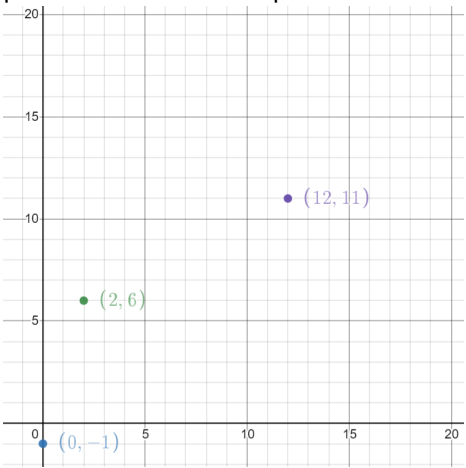
Opening

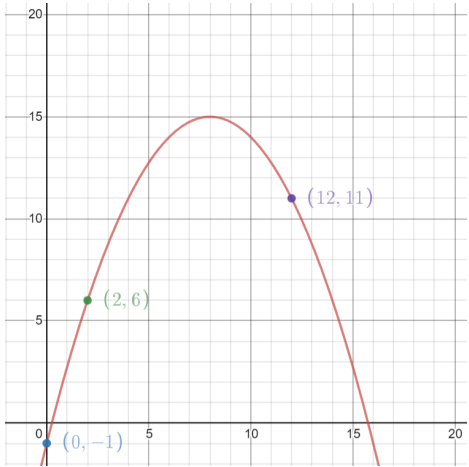
Consider the graph below. Can you draw a straight line that passes through each and every point? What about a parabola?



Opening

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





Mathematical Interpolation

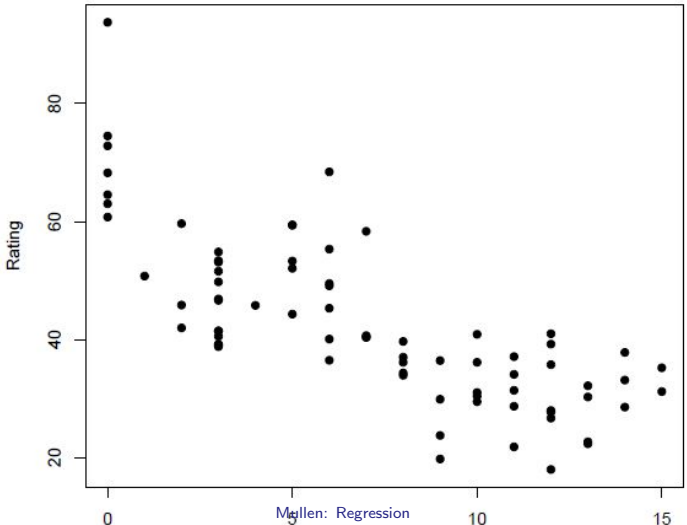
These problems are known as *interpolation*. We actually don't care about them! As statisticians we ask a different problem: how can we *best* draw a line through data values if the process that's generated them is subject to **error**?

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!

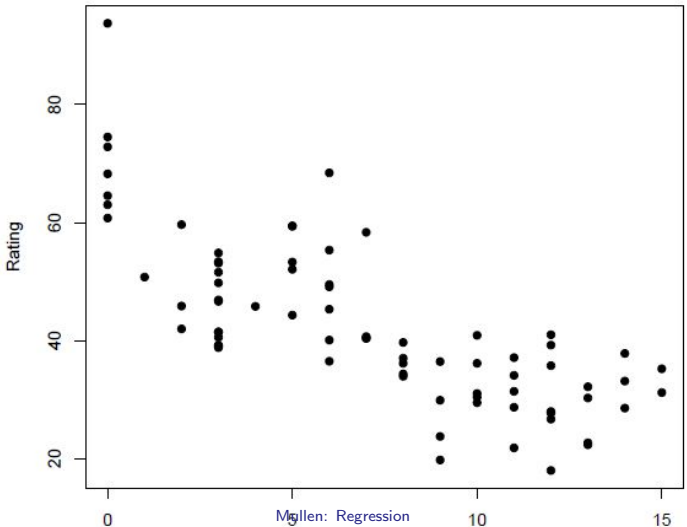
Simple Linear Regression

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

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.

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:

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

Simple Linear Regression Model

Definition: *Simple Linear Regression* (SLR)

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

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:

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

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

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

4. $$\varepsilon_i \sim N(0, 1)$$

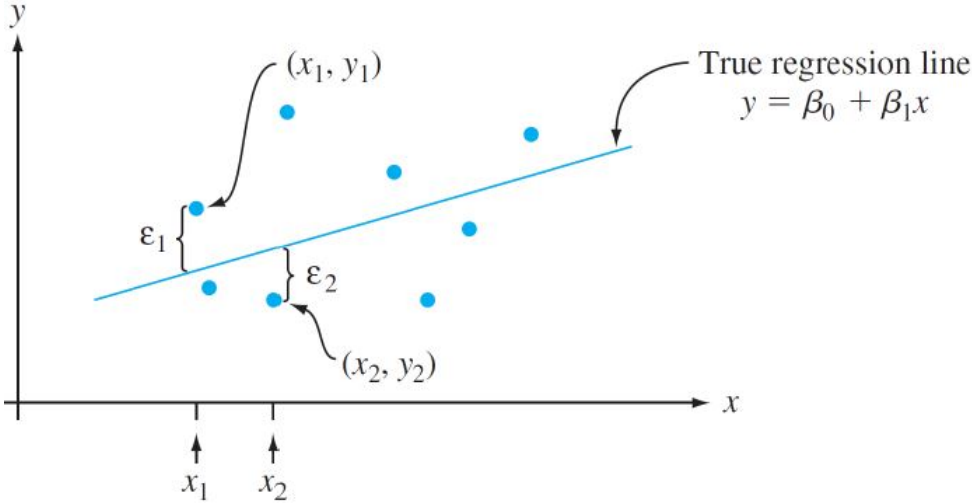
Simple Linear Regression Model

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.

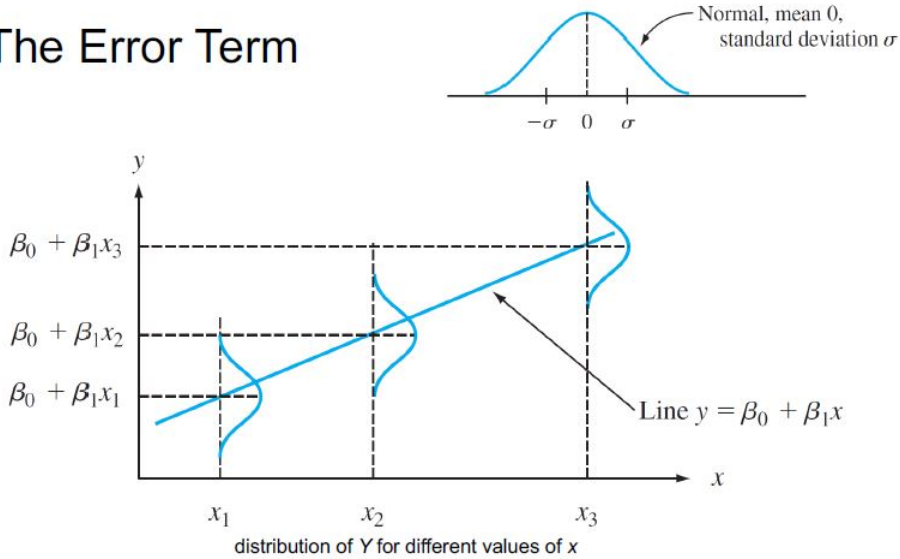
What exactly does ε do?

Simple Linear Regression: ε



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?

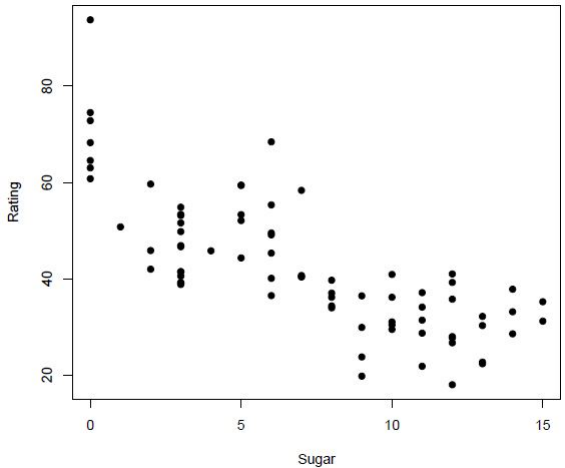
1. **Linearity:** Is a *straight* line actually appropriate?
2. **Independence:** Are there patterns or structures in my data not covered by a straight line?
3. **Homoskedasticity:** Do the sizes of my errors vary depending on where I am in the data set? Larger x, larger y?
4. **Normality:** Do the errors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

Simple Linear Regression: Assumptions

How do we know when the SLR model is appropriate?

Scatter plots

Assumptions

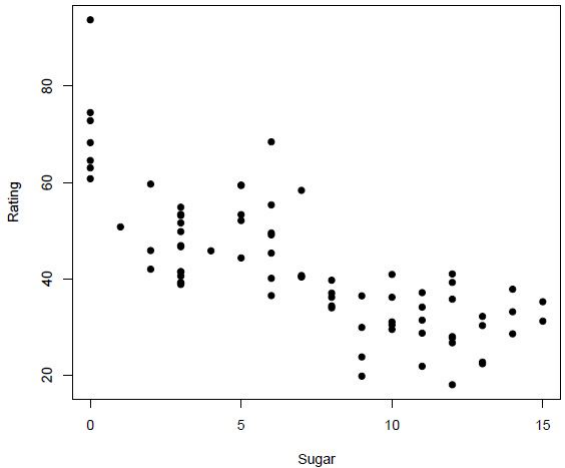


Simple Linear Regression: Assumptions

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



Simple Linear Regression: Interpretations

Interpreting parameters:

- 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":

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

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 $E[Y_i]$:

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 $Var[Y_i]$:

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

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

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

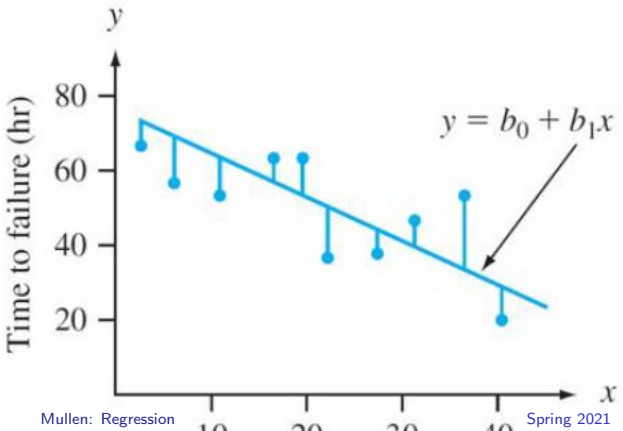
$$= \beta_0 + \beta_1 X_i + \beta_1$$

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

Estimating SLR Parameters

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

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

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!

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

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

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

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

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:

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

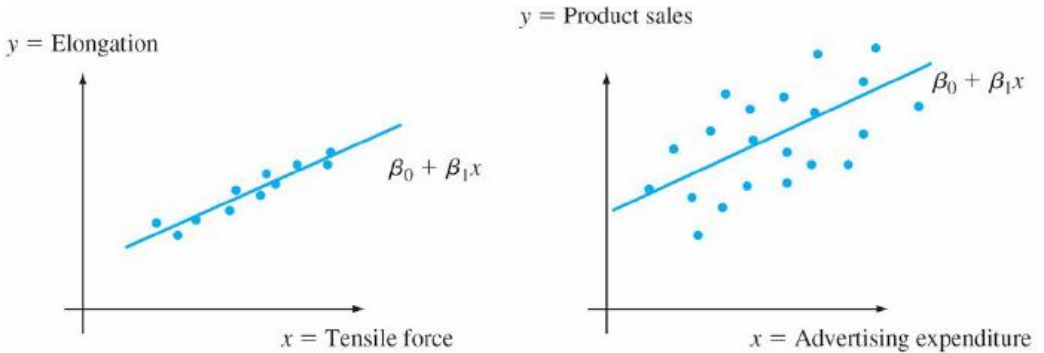
2. The *residuals* are the differences between the observed and fitted y values:

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

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:

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

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:

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

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

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

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:

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

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

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

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?
   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.

Daily Recap

Today we learned

1. Regression!

Moving forward:

- nb day Friday

Next time in lecture:

- More Regression!