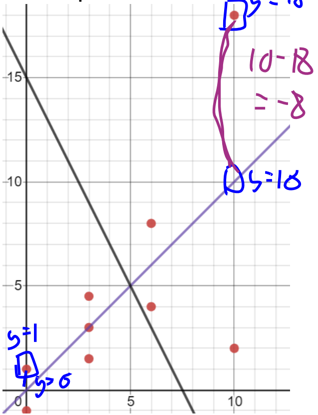
CSCI 3022-002 Intro to Data Science
Regression Inference

1) Linearity
2) Independence
3) Variance of errors changing?
4) Normality

Consider the graph below. Do either of the candidate "best fit" lines violate the 4 big assumptions?



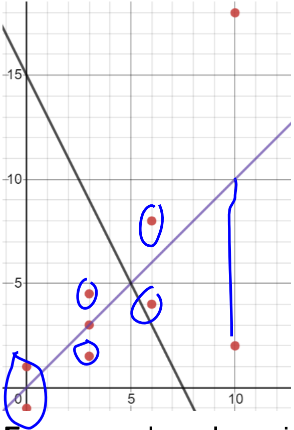
1) ✓ ✓ ∅

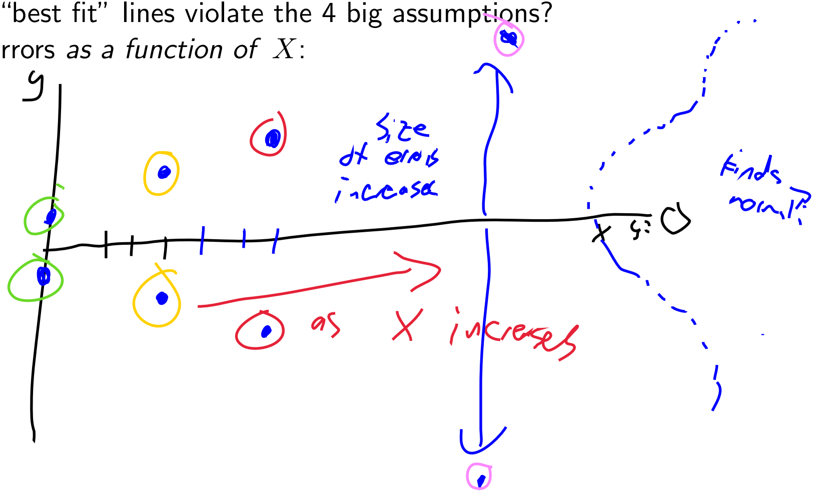
Create new data:

Estimated line − data

Do either of the candidate "best fit" lines violate the 4 big assumptions?

One thing to do: plot the errors *as a function of* X:

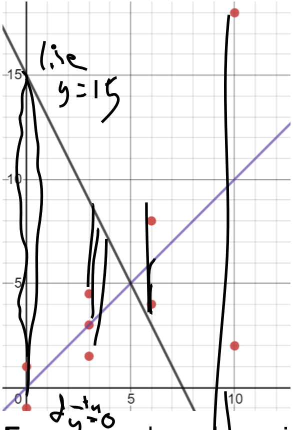


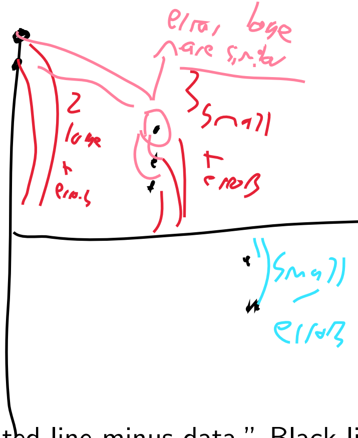


Errors are the values given by "estimated line minus data."

Do either of the candidate "best fit" lines violate the 4 big assumptions?

One thing to do: plot the errors *as a function of* X:





want the points to look random and uncorrelated, but Center at y=0

Not right line/model

Errors are the values given by "estimated line minus data." Black line the errors clump and move up/down as X moves left-right.

Blue line the errors increase in *magnitude* as X goes right.

Announcements and Reminders

- Short HW for next week.
- NB day Friday.

Where we at?

Definition: *Simple Linear Regression* (SLR)
The *Simple Linear Regression* model is a model of the form

With 3 assumptions on ε:

Where we at?

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:

Where we at?

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

Where we at?

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

Where we at?

Definition: *Simple Linear Regression* (SLR)

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

1. Linear
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
With 3 assumptions on ε:
2. error doesn't clump up
$$\text{Cov}[\varepsilon_i, \varepsilon_j] = 0 \qquad \forall i, j$$
Independence of errors
3. size of error doesn't ↑ or ↓ with X ↑ or ↓
$$Var(\varepsilon_i) = \sigma^2 \qquad \forall i$$
Homoskedasticity of errors
4. Normality
$$\varepsilon_i \sim N(0, 1)$$

Simple Linear Regression Model

The β estimators in the model are: estimates for

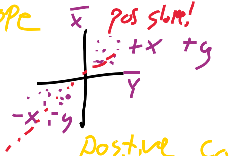
$\hat\beta_0$: intercept : Y-value when $x=0$

$(\bar X, \bar Y)$ is on our line

$\hat\beta_1$: slope

1. $\hat\beta_0 = \bar Y - \hat\beta_1 \bar X$
2. $\hat\beta_1 = \dfrac{Cov[X,Y]}{\sum_{i=1}^{n}\left(X_i - \bar X\right)^2} = \dfrac{\sum_{i=1}^{n}\left(X_i - \bar X\right)\left(Y_i - \bar Y\right)}{\sum_{i=1}^{n}\left(X_i - \bar X\right)^2}$ $= Cov[X,Y]$

recall: $Var[x] = \dfrac{\sum (x_i - \bar x)^2}{n-1}$



positive covariance

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. Has variance σ^2.
- β: the regression coefficients.
- r: the *residuals* or observed errors. Used to estimate σ^2.

opening examples

$y = \boxed{\beta_0} + \boxed{\beta_1} x + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$

$\hat{\beta}_0 \qquad \hat{\beta}_1$

Estimating SLR Parameters

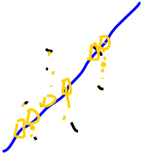
Definitions:

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

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

points on that 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:

data

r_i: diff between line at $x = x_i$ and data point (x_i, y_i)

$r_i = \hat{Y}_i - \underbrace{y_i}_{\text{data value.}}$

Opening plot. (x_i, r_i)

Residuals are estimates of the true error. Why?

std dev.

$\hat{Y}_i$ on the line

Estimating SLR Parameters

Definitions:

1. The *fitted (or predicted) values* $\underline{\hat{Y}_i}$ are obtained by plugging in $\underline{\hat{X}_i}$ 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:

$$\hat{\varepsilon}_i = \underline{r_i} = \underline{\hat{e_I}} = Y_i - \hat{Y}_i = Y_i - \underbrace{\hat{\beta}_0 + \hat{\beta}_1 X_i}_{\text{line}}$$

data

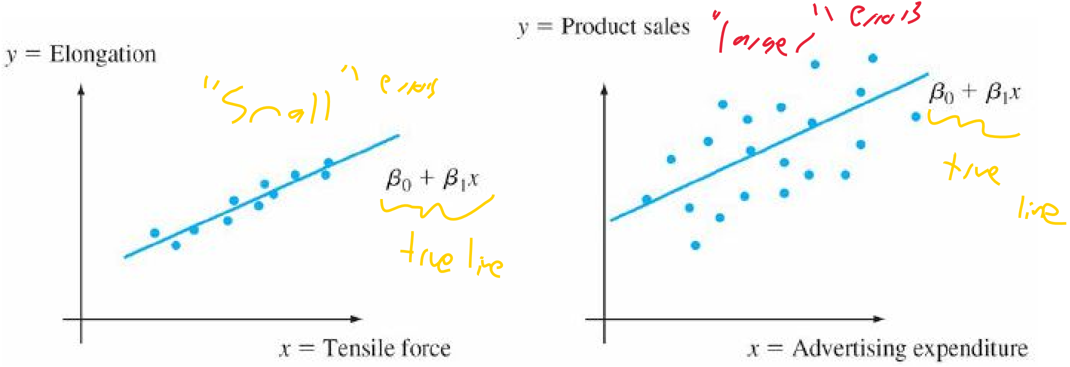
Residuals are estimates of the true error. Why?
We don't have the true values of β_0, β_1, so when we estimate them we get variance and error in our estimates.

Estimating SLR Parameters: σ^2

estimate - reality

$Y_i - Y_i$

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



Estimating SLR Parameters: σ^2

Recall: line we choose minimizes $\sum(\text{line}-\text{data})^2$

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:

$$\text{SSE}= \sum\left(\hat{Y}_i - Y_i\right)^2 = \sum\left(Y_i - \hat{Y}_i\right)^2$$

$$= \sum\left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right)^2$$

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:

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

✓ total error, n observations

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

Recall

$$\text{Var}[Y] = \frac{\sum (Y_i - \bar{Y})^2}{n-1}$$

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:

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

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

$X = [1]$

$Y = [1, 3]$

$Z = [-1, 1, 5]$

center of 1 ?

no spread

center of 2

spread ~ 1 ?

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:

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

first, find $\bar{Y}$

$$Var=\frac{\sum(Y_i-\bar{Y})^2}{n-\boxed{1}}$$

"lost" a degree of freedom

second thing we estimate

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?

   No

2. Can you draw a parabola through them?

   Yes, only 1!

3. Can you draw a cubic function through them?

   $$ax^3 + \underbrace{bx^2 + cx + d}_{\text{includes parabola}}$$

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 $(\underbrace{\hat{\beta}_0, \hat{\beta}_1})$ before we got to σ, which "costs" us two degrees of freedom.

More than just $\bar{X}$: $n-1 \rightarrow n-2$

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.

$E[\hat{\sigma}^2] = \sigma^2$

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

estimate "does ok" for finding σ^2

The Coefficient of Determination

The residual sum of squares SSR can be interpreted as a measure of how much variation in y is left unexplained by the model—that is, how much cannot be attributed to a linear relationship. In the first plot, $SSE = 0$, and there is no unexplained variation, whereas unexplained variation is small for second, and large for the third plot.

Picturing Sums of Squares

The goodness-of-fit of a regressive model is often decomposed into three components based on squared deviations. These are:

1. SSE: Sum of squared errors: (vertical) distances from the regression line to the data values.

   dist line - data

2. SST: Sum of squares, total: total deviation in Y. Looks like $Var[Y]$.

   dist $y - \bar{y}$ baseline: how much y moves

3. SSR: Sum of squares of regression line: the amount of variability tied to the model.

   dist line $- \bar{y}$

Picturing Sums of Squares

The goodness-of-fit of a regressive model is often decomposed into three components based on squared deviations. These are:

1. SSE: Sum of squared errors: (vertical) distances from the regression line to the data values.

$$\sum_i \left(\underbrace{\hat{Y}}_{\text{est}} - \underbrace{Y_i}_{\text{data}}\right)^2$$

2. SST: Sum of squares, total: total deviation in Y. Looks like $Var[Y]$.

$$\sum_i \left(\underbrace{Y_i}_{\text{data}} - \underbrace{\bar{Y}}_{\text{avg}}\right)^2$$

3. SSR: Sum of squares of regression line: the amount of variability tied to the model.

$$\sum_i \left(\underbrace{\hat{Y}_i}_{\text{est}} - \underbrace{\bar{Y}}_{\text{avg}}\right)^2$$

Picturing Sums of Squares

Data (x_i, y_i)

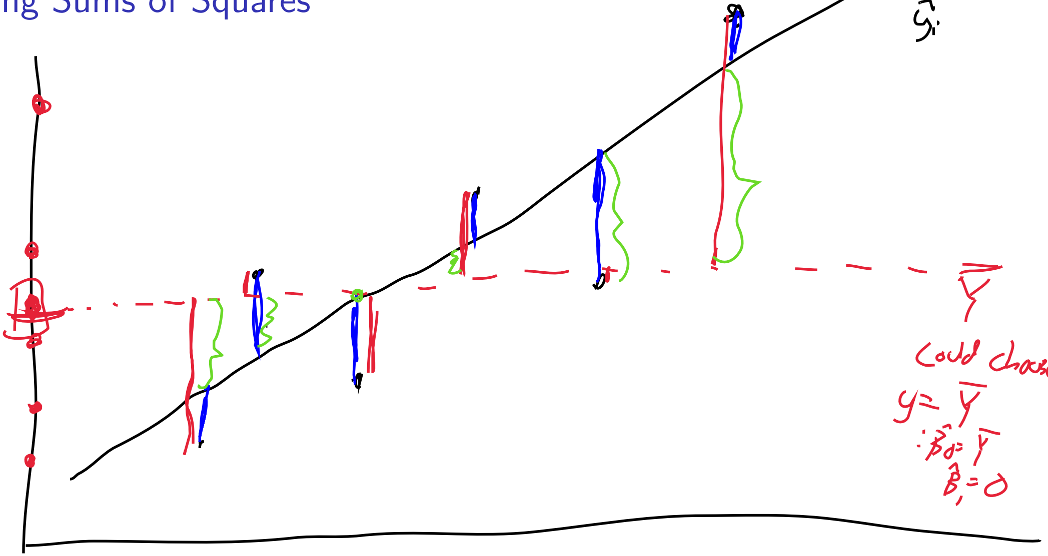
Line: $\hat{y}_i$

SSE

SST:

Find $\bar{Y}$

SSR



Could chose $y = \bar{Y}$

$\therefore \hat{\beta}_0 = \bar{Y}$

$\hat{\beta}_1 = 0$

The Coefficient of Determination

SSE

The sum of squared deviations about the least squares line is smaller than the sum of squared deviations about any other line, i.e. $SSE < SST$ unless the horizontal line itself is the least squares line.

error of estimated line

versus Y movement

The ratio SSE/SST is the proportion of total variation that cannot be explained by the simple linear regression model. The coefficient of determination is:

This coefficient is a number between 0 and 1 and is the *proportion of observed y variation explained by the model*.

The Coefficient of Determination

The sum of squared deviations about the least squares line is smaller than the sum of squared deviations about any other line, i.e. $SSE < SST$ unless the horizontal line itself is the least squares line.

The ratio SSE/SST is the proportion of total variation that cannot be explained by the simple linear regression model. The coefficient of determination is:

$$R^2 = \boxed{1 - \frac{SSE}{SST}} = \frac{SSR}{SST}$$

This coefficient is a number between 0 and 1 and is the *proportion of observed y variation explained by the model*.

The Coefficient of Determination

Again, R^2 is the proportion of observed y variation explained by the model.

The higher the value of R^2, the more successful is the simple linear regression model in explaining y variation, assuming the linear model is correct.

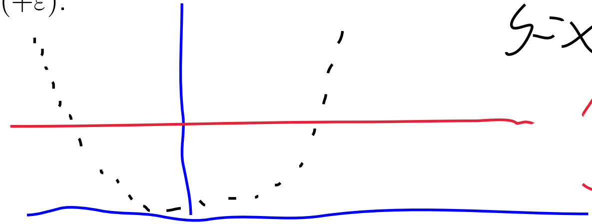
The Coefficient of Determination

Again, R^2 is the proportion of observed y variation explained by the model.

The higher the value of R^2, the more successful is the simple linear regression model in explaining y variation, assuming the linear model is correct.

Crucially, R^2 is a measure of *linear* dependence between X and Y. If $R^2 = 0$, X and Y may still be related! Ex: $Y = X^2(+\varepsilon)$.

$y = x^2$



related, but not linear

Inferences about Parameters

The parameters in SLR have distributions. From these distributions, we can conduct hypothesis tests (e.g., __________), compute confidence intervals, etc.

Distributions:

Inferences about Parameters

The parameters in SLR have distributions. From these distributions, we can conduct hypothesis tests (e.g., $\underline{H_0 : \beta_1 = 0}$), compute confidence intervals, etc.

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}; \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}$

Distributions:

Inferences about Parameters

The parameters in SLR have distributions. From these distributions, we can conduct hypothesis tests (e.g., $\underline{H_0 : \beta_1 = 0}$), compute confidence intervals, etc.

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}; \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

Distributions:

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{X}^2}{(X_i - \bar{X})^2}\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{(X_i - \bar{X})^2}\right)$$

... but of course, we don't know σ^2, so we estimate with $SSE/(n-2)$.

Inferences about Parameters

Confidence Intervals: The CIs for regression are two-sided, and because $\varepsilon \sim N(0, \sigma^2)$, we may use t statistics. Since we have written down the variances of the βs, we can also write down their standard errors:

$$s.e.(\hat{\beta}_0) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\left(X_i - \bar{X}\right)^2}}; \qquad s.e.(\hat{\beta}_1) = \sigma\sqrt{\frac{1}{\left(X_i - \bar{X}\right)^2}}$$

These lead to CIs of

where we replace σ with the estimate $s = \frac{SSE}{n-2}$

Inferences about Parameters

Confidence Intervals: The CIs for regression are two-sided, and because $\varepsilon \sim N(0, \sigma^2)$, we may use t statistics. Since we have written down the variances of the βs, we can also write down their standard errors:

$$s.e.(\hat{\beta}_0) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\left(X_i - \bar{X}\right)^2}}; \qquad s.e.(\hat{\beta}_1) = \sigma\sqrt{\frac{1}{\left(X_i - \bar{X}\right)^2}}$$

These lead to CIs of

$$\beta_i \in (\hat{\beta}_i \pm t_{\alpha/2, n-2} \cdot s.e.(\hat{\beta}_i))$$

where we replace σ with the estimate $s = \frac{SSE}{n-2}$

Inferences about Parameters

Confidence Intervals: The CIs for regression are two-sided, and because $\varepsilon \sim N(0, \sigma^2)$, we may use t statistics. Since we have written down the variances of the βs, we can also write down their standard errors:

$$s.e.(\hat{\beta}_0) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\left(X_i - \bar{X}\right)^2}}; \qquad s.e.(\hat{\beta}_1) = \sigma\sqrt{\frac{1}{\left(X_i - \bar{X}\right)^2}}$$

These lead to CIs of

$$\beta_i \in (\hat{\beta}_i \pm t_{\alpha/2, n-2} \cdot s.e.(\hat{\beta}_i))$$

where we replace σ with the estimate $s = \frac{SSE}{n-2}$

Tests then result from comparing $t = \frac{\hat{\beta}_i}{s.e.(\hat{\beta}_i)}$ to the corresponding critical t values for a one or two-tailed test.

Inferences about Y

There are more types on confidence intervals we may care about!

1. Last slide was how to perform inference on the **parameters** of the *line* β. We also might care about inference on values of Y!
2. A **confidence band** is how sure we are about the mean of Y at specific values of X, or $E[Y|X]$.
3. A **prediction band** is how we estimate the distribution of new Y observations at specific values of X. It's the same as the confidence band, but also includes our estimate for ε.

See: nb accompanying lecture: `SLR Inference`

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 $\frac{\sum\left(X_i - \bar{X}\right)^2}{n}$, which is a different denominator than our usual s^2.

Daily Recap

Today we learned

1. Regression Inference!

Moving forward:

- nb day Friday

Next time in lecture:

- More Regression! More predictor!