CSCI 3022 Intro to Data Science
Regression nbs

Today we do 3 notebooks: nb20, nb21, and an unnumbered notebook about *prediction*. (see "Predicts" on course schedule).

Broad strokes: For a linear model, we have:

- A list of assumptions + 3 about error
- A few important statistics we calculate from the data: $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$, R^2.

  intercept slope error of each observation overall "goodness of fit"
- Today: *hypothesis tests* based on those values and their standard errors.

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

   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

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

Simple Linear Regression Model

The β estimators in the model are:

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

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. The *residuals* are the differences between the observed and fitted y values: $\hat{\varepsilon_i} = r_i = \hat{e_I} = Y_i - \hat{Y_i} = Y_i - \hat{\beta}_0 + \hat{\beta}_1 X_i$

Error of a model

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. $\sum_i \left(\underbrace{\hat{Y} - Y_i}_{\text{residuals}}\right)^2$.
2. SST: Sum of squares, total: total deviation in Y. Looks like $Var[Y]$. $\sum_i \left(Y_i - \bar{Y}\right)^2$
3. SSR: Sum of squares of regression line: the amount of variability tied to the model. $\sum_i \left(\hat{Y}_i - \bar{Y}\right)^2$

The coefficient of determination is: $R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$ This coefficient is a number between 0 and 1 and is the *proportion of observed y variation explained by the model*.

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. $\hat{\beta}_1 =$

Our estimate of the variance of the model is like a measure for an average of this summed errors SSE:
$\hat{\sigma^2} = \frac{SSE}{n-2}$.

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

Our estimate of the variance of the model is like a measure for an average of this summed errors SSE:
$\hat{\sigma^2} = \frac{SSE}{n-2}$.

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

Our estimate of the variance of the model is like a measure for an average of this summed errors SSE:
$\hat{\sigma^2} = \frac{SSE}{n-2}$.

$\hat{\beta}_0$ & $\hat{\beta}_1$ depend on data, so they're random

Inferences about Parameters

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

Distributions:

$$E[\hat{\beta}_0] = \beta_0$$

$$E[\hat{\beta}_1] = \beta_1$$

but

$$Var[\hat{\beta}_0]$$

$$Var[\hat{\beta}_1]$$

$$= Var\left[\frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}\right]$$

include ϵ_i are random

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.

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

1) Normal, each datum's noise $\epsilon \sim N$

Distributions:

2) Unbiased

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

each has its own error, denominator $(X_i - \bar{X})^2$

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

... 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:

estimate σ with $\hat{\sigma}$

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

These lead to CIs of

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

Table : Summary table of model

β_0/const	value	std. error
β_1/slope	value	std. error.

$\hat{\sigma}^2$ too!

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:

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

These lead to CIs of

estimate ± critical · std error

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

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 $s = \frac{SSE}{n-2}$

guess - "baseline"

baseline = 0 =>

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

std error

H_0: assumption that slope = 0 => X & Y unrelated.

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 σ^2. This is also known as a *forecast*.

Idea: If we want to **guess** the *average* $y = \beta_0 + \beta_1 x$, (for a specified x) we have to combine our uncertainties for the βs. If we want to describe *all* the y's for a single value of x, we also would need to include the uncertainty $s^2 \approx \sigma^2$ that accompanies ε.
See: nb accompanying lecture: `SLR Prediction and Confidence`

The usual inference:

The most common inference for linear regression is to answer the question "Does x affect y?" This is a hypothesis test asking about the value of the *slope* of the regression line. We have a CI for this of

$$\beta_1 \pm t_{\alpha/2,n-2} \cdot s.e.(\hat{\beta}_i))$$

where

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

The corresponding hypothesis test is $t = \frac{\hat{\beta}_i}{s.e.(\hat{\beta}_i)}$ with $n = 2$ degrees of freedom.

Two big things to notice

1. The error grows as σ grows: noisy/random data is harder to estimate.
2. The denominator *looks a lot* like the "standard deviation of x." We get more **confident** in our estimates if the predictor variable locations are spread out!

Daily Recap

Today we learned

1. Regression Inference!

Moving forward:

- nb day Friday

Next time in lecture:

- More Regression! More predictor!