Lecture 13:
Inference for Simple Linear Regression
STAT 310, Spring 2023

Review of Key Terms

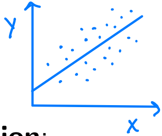
- A **parameter** is a numerical characteristic of the population (fixed number, that is usually unknown).
  - For example, the population mean height μ of all students at CSUEB.
- A **statistic** is a numerical characteristic of the sample (varies depending on sample). The statistic is also referred to as a **point estimate**, since it is our best guess at the value of a population parameter.
  - For example, the sample mean height $\bar{x}$ of $n = 100$ randomly selected CSUEB students.

Review of Key Terms

Statistical inference refers to the process of using data collected from a sample to answer questions about population parameters.

- **Standard error (SE)**: measures the variability of a statistic (or point estimate) from sample to sample.
- **Confidence interval**: a plausible range of values for the population parameter.
- **Hypothesis test**: Do the sample data provide convincing evidence that the population parameter is different than some value?

Inference for Linear Regression



Simple linear regression model for the population:

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

β_0 and β_1 are the population parameters. We can refer to β_0 as the *population intercept* and β_1 as the *population slope*.

Least squares regression line:

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

$\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of the slope and intercept, obtained from a random sample of data.

	Point Estimate	Population Parameter
mean	$\bar{x}$	μ
proportion	$\hat{p}$	p
standard deviation	s	σ
simple linear regression	$\hat{\beta}_0, \hat{\beta}_1$	β_0, β_1

Confidence interval for the population slope β_1:

$$\hat{\beta}_1 \pm t^* SE_{\hat{\beta}_1}$$

- $\hat{\beta}_1$ is the point estimate
- t^* is the t-critical value, which depends on the confidence level and has $n-2$ degrees of freedom
- $SE_{\hat{\beta}_1}$ is the standard error of the slope estimate

Hypothesis test for whether the population slope β_1 is different than zero. We can also interpret this as a hypothesis test for whether there is a linear association between x and y.

$H_0 : \beta_1 = 0$ there is <u>no</u> linear association between x and y

$H_A : \beta_1 \neq 0$ there is a linear association between x and y

Test statistic:

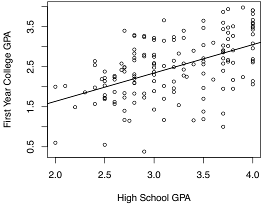
$$t = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}}; \quad \text{df} = n - 2$$

The test statistic is then used to compute the p-value. When using the default significance level ($\alpha = 0.05$), we would reject H_0 when the p-value < 0.05.

Example

Let's go back to the example of using a student's high school GPA (x) to predict their college GPA (y). Shown below is a scatter plot of the data, and the output from fitting this linear regression model in R.

```
Coefficients:                SE                 p-value
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.2168     0.3250   0.667    0.506
hs_gpa         0.7091     0.1003   7.069 5.73e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Equation of least squares line:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 0.22 + 0.709x$$

Example

(a) Do the data provide strong evidence of a linear association between high school GPA and first-year college GPA? State the null and alternative hypothesis, report the test statistic and *p*-value, and state your conclusion.

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

$t = 7.069$

p-value = 5.73e-11

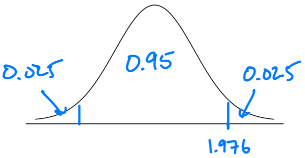
$= 5.73 \cdot 10^{-11} \approx 0$

Since p-value < 0.05, we reject H_0.

The data provide strong evidence of a linear association between high school and first-year college GPA.

Example

(b) Calculate a 95% confidence interval for the slope parameter β_1. Note that there are $n = 150$ students in this data set.



$$t^* = qt(0.975,\ df = \overset{n-2}{148})$$

$$= 1.976$$

$$\hat{\beta}_1 \pm t^* SE \Rightarrow 0.709 \pm 1.976(0.1003)$$

$$\Rightarrow \boxed{(0.511,\ 0.907)}$$

We are 95% confident that the population slope β_1 is between 0.511 and 0.907

Conditions for Simple Linear Regression

- **Linearity**. The data should follow a linear trend.
- **Constant variability**. The variability of the points around the least squares line remains roughly constant.
- **Normality**. The residuals should be approximately normally distributed with mean 0.
- **Independence**. Values of the response variable are independent of each other. This is satisfied when the data come from a random sample.

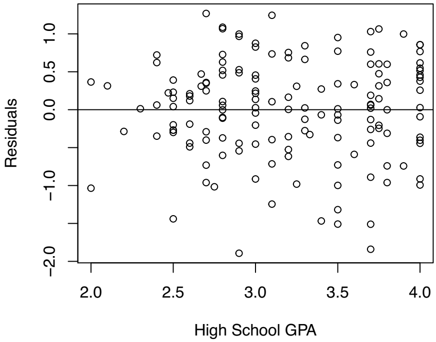
Residual Plots

residual = observed − predicted

- One useful way to check the conditions is to look at a plot of the residuals, $\hat{e}_i = y_i - \hat{y}_i$, versus the predictor, x_i.
- One purpose of residual plots is to identify characteristics or patterns still apparent in the data after fitting the model.
- Residual plots are especially useful for checking the constant variability condition.
- Ideally, the residual plot should show no obvious pattern, and the points are randomly scattered around 0.

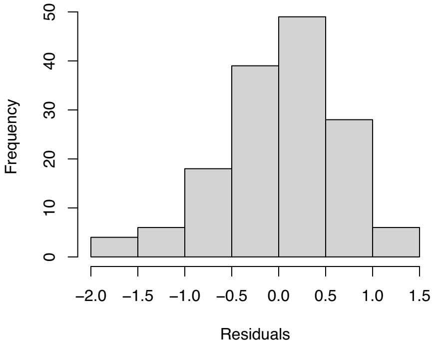
Example: Residual Plot

For the simple linear regression model between high school and first-year college GPA, the points in the residual plot look randomly scattered around the horizontal line at 0. This indicates that the condition of constant variability is met.



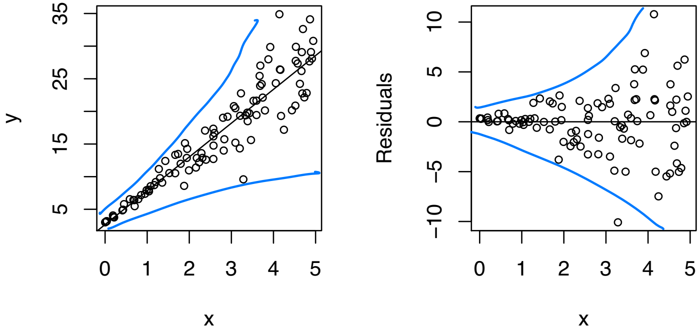
Example: Normality of Residuals

To check whether the the residuals are normally distributed we can make a histogram. The histogram should be symmetric about 0 and have an approximate bell-curve shape. For the GPA example, the residuals appear to have an approximate normal distribution, and there are no outliers.



Example: Nonconstant variability

An example of a violation of the constant variability condition. This residual plot shows a **fan pattern**.



Example: Nonlinearity

An example of a violation of the linearity condition.