Chapter 12. Linear Regression

***Note:** regression line **always** goes through center of data ($\bar{x}$, $\bar{y}$)*

A **regression line** is the *"best fit"* line through a *"scatterplot"* of data.
Regression analysis considers the shape and fit of the regression line.

paired sample data

	A	B
1	**X**	**Y**
2	11.23	13.79
3	8.18	13.94
4	10.38	13.04
5	8.53	17.41
6	14.42	11.99
7	9.75	15.82

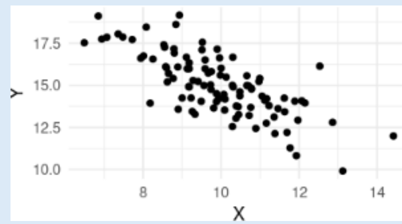
combine

points

D
(x, y)
(11.23, 13.79)
(8.18, 13.94)
(10.38, 13.04)
(8.53, 17.41)
(14.42, 11.99)
(9.75, 15.82)

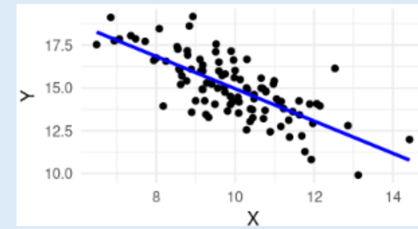
plot

"scatterplot" of data



fit

regression line



Regression Line Vocabulary

- The *independent* variable, plotted along the $\boldsymbol{x}$-axis, is called the **explanatory** or **regressor** variable.
- The *dependent* variable, plotted along the $\boldsymbol{y}$-axis, is called the **response** or **outcome** variable.
- The **regression line equation** is written $\boxed{\hat{\boldsymbol{y}} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\, \boldsymbol{x}}$
- $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are the **regression coefficients**

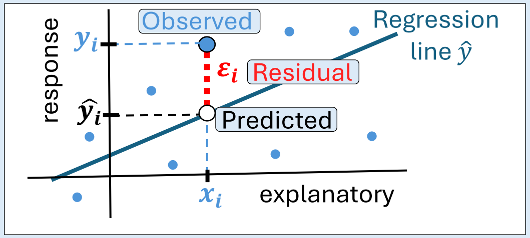
β_0 measures *"lift"*
β_1 measures *"tilt"*

ε_i measures "spread"

  - $\boldsymbol{\beta_0}$ is the regression **intercept**. *("records t-test information" between x and y)*
  - $\boldsymbol{\beta_1}$ is the regression **slope**. It gives the ***expected change*** in y when x increases by 1.
    If $\boldsymbol{\beta_1} = 0$ then $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent! *(if $\beta_1 = 1$ then β_0 is equiv. to paired sample t)*
- Values on the regression line $\hat{\boldsymbol{y}}$ are called **predicted** or **fitted** values. We write $\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\, \boldsymbol{x_i}$
- For each observed data point (x_i, y_i), the distance between observed and predicted values is the **residual error** (or just **residual**) written $\boldsymbol{\varepsilon_i} = \boldsymbol{y_i} - \hat{\boldsymbol{y}}_{\boldsymbol{i}}$ *(residual = observed – expected)*
  - residuals ε_i measure *(vertical)* distance of data points from the regression line *("spread")*
  - regression line minimizes the **sum of squared residuals:** $(\varepsilon_1)^2 + (\varepsilon_2)^2 + (\varepsilon_3)^2 + \cdots$
- The **regression model** is $\boxed{\boldsymbol{y} = (\boldsymbol{\beta_0} + \boldsymbol{\beta_1}\, \boldsymbol{x}) \ + \ \boldsymbol{\varepsilon}}$ where $\boldsymbol{\varepsilon}$ follows a normal distribution with mean $\mu_\varepsilon = 0$ and (constant) standard deviation σ_ε. *(All p-values are computed assuming this)*

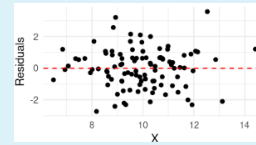
Chapter 12. Regression Analysis

Regression model: $$\boldsymbol{y = \underbrace{(\beta_0 + \beta_1 x)}_{\text{Regression line } \hat{y}} + \underbrace{\varepsilon}_{\text{Residual error}}}$$

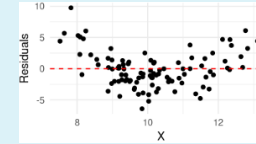


Note: Residuals should be normally distributed, independent of x and y.

No pattern in scatterplot of x vs ε:



good residuals



bad residuals

Hypothesis Tests:

- t-Test for $\beta_0 = 0$ *(If $\beta_1 = 1$ this tests against **equal means**, $\overline{x} = \overline{y}$: same as <u>paired sample</u> t-test!)*
- t-Test for $\beta_1 = 0$ *(Equivalent to **independence test** in <u>single</u> variable <u>linear</u> regression!)*
- F-Test against **independence** of response and explanatory variables.

Note on F Distribution.

- **Chi-Squared** distribution (χ^2) is **sum of squares** of (indep.) normal random variables.

  $$\chi^2_n = (X_1)^2 + (X_2)^2 + \cdots + (X_n)^2 \quad \textit{(n degrees of freedom)}$$

  → *used to compute p-values of variances*

- Fisher's $\boldsymbol{F}$ distribution is **quotient** of two Chi-Squared random variables.

  $$F_{n,\,d} = \frac{\chi^2_n / n}{\chi^2_d / d} \qquad \textit{(n numerator and d denominator degrees of freedom)}$$

  → *used to compute p-values **comparing** variances (if variances are equal then $F = 1$)*

Use this because **difference** of χ^2 is no longer χ^2 !!

To test regression model, use $$F = \frac{\text{variance of } \textbf{predicted} \text{ values } \ \widehat{y_i}}{\text{variance of } \textbf{residuals} \ \ \varepsilon_i}$$

same F in anova

*p-value $p < \alpha$ means that variables are **not independent**, and regression is **meaningful***

Chapter 12. Correlation and Determination Coefficients

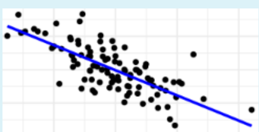
F-test tells if regression line is **meaningful,** but not if it is **useful**.

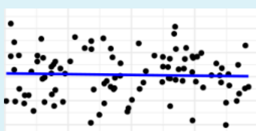
With enough data, even small β_1 values could be **significant**…
but if residual error σ_ε is big, then $\hat{y}_i$ may be far from y_i *(i.e. predictions **not useful**)*

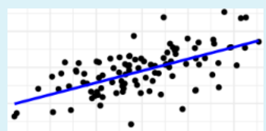
This information is captured by the **correlation** and **determination** coefficients ($\boldsymbol{r}$ and $\boldsymbol{r^2}$)

[Pearson] Correlation

- $-1 \leq r \leq 1$
- $\pm$ sign matches regression slope
- near 0 $\Longrightarrow$ ε **big** compared to β_1
  *(values **far** from regression line)*
- near $\pm 1 \Longrightarrow$ ε **small** compared to β_1
  *(values **close** to regression line)*


negative correlation

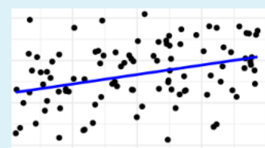

zero correlation

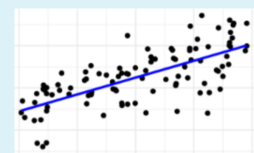

positive correlation

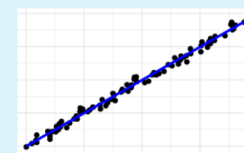
Determination

"How much of y is determined by x?"

- $0 \leq r^2 \leq 1$
- "Percent of variation explained by model"
- $r^2 = \dfrac{\text{variance of } \textbf{predicted} \text{ values } \hat{y}_i}{\text{variance of } \textbf{observed} \text{ values } y_i}$
- $r^2 \approx 0$ means regression line is **not** useful!
- $r^2 \approx 1$ means regression line is **very** useful!


$r^2 \approx 10\%$


$r^2 \approx 50\%$


$r^2 \approx 99\%$

Example regression analysis output

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.3068 -1.7971 -0.1492  0.8094  6.4944

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.0171     1.7780   5.071 0.000963 ***
X             0.5557     0.1588   3.500 0.008084 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.852 on 8 degrees of freedom
Multiple R-squared:  0.6049,	Adjusted R-squared:  0.5555
F-statistic: 12.25 on 1 and 8 DF,  p-value: 0.008084
```

should be symmetric

β_0

β_1

Equal.
want $p < \alpha$

$n-2$

$0 \leq r^2 \leq 1$
want big

Notes:

- Correlation does NOT imply causation.
  - Maybe X causes changes in Y…
  - Maybe Y causes changes in X…
  - Maybe a third, "*lurking* variable" changes both!
- Good correlation does NOT mean data is linear.
  - Maybe quadratic or exponential!
  - You must plot the residuals!
- Must combine multiple measures of goodness…