Correlation and Simple Linear Regression

Mana Azizsoltani
HOA 730
Spring 2026

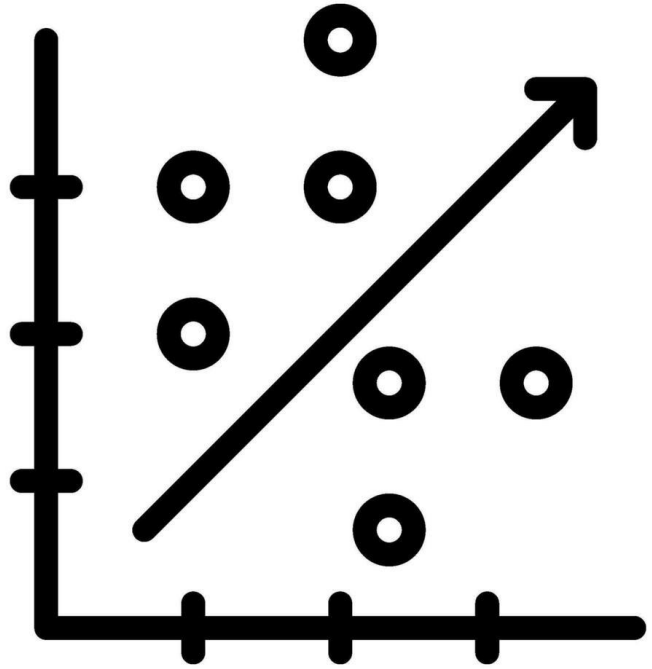
Objective for this lecture

- Understand ways of measuring the relationship between two variables
- Interpreting different relationships between two variables
- Predict one variable using another variable

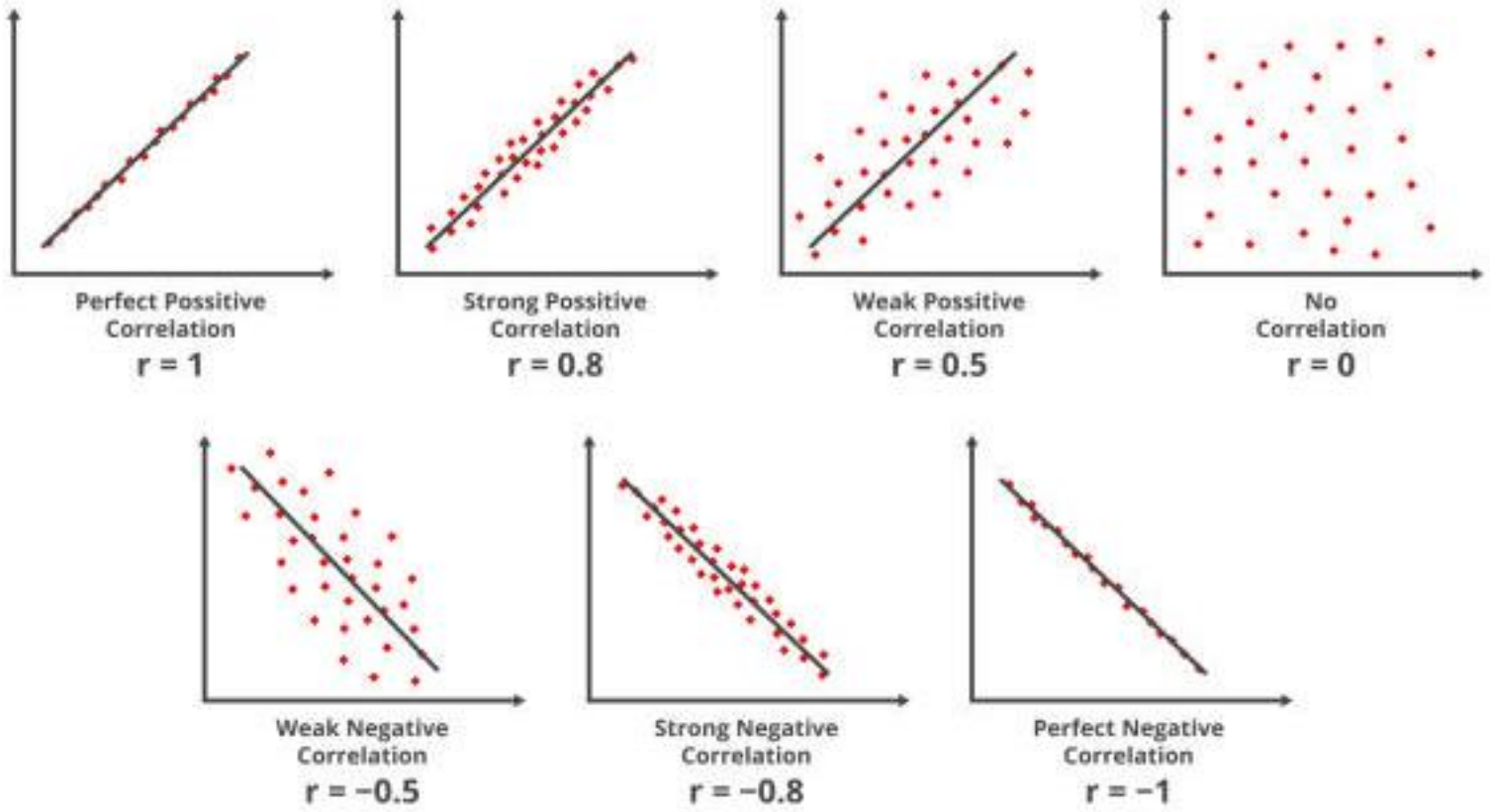
Section 1
Correlation

Correlation: Relationship between variables

- We can only compute correlation between two numeric variables
- Correlation measures the association between two variables – does **not** establish causation
- Examples:
  - Do bigger hotels charge more?
  - Do better reviews lead to a higher prices?
  - Does distance from the city center matter?



Correlation: Definition

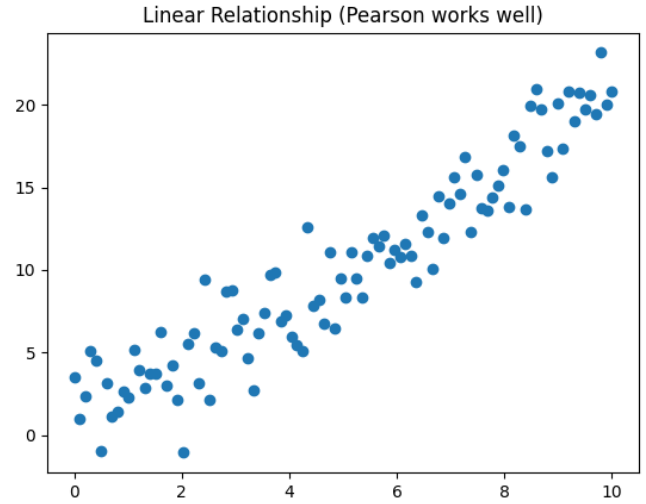


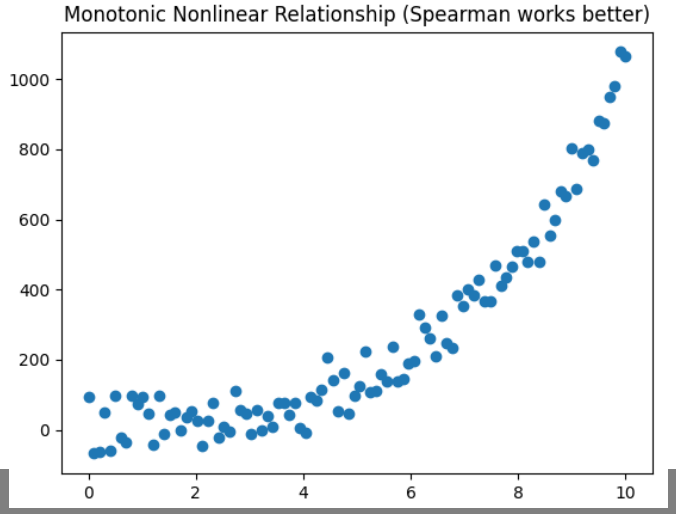
Correlation measures the **strength** and **direction** of a relationship between two variables

r = correlation coefficient

- 0 to 1: how strong
- + / - : directionality

Correlation: Measurement





Ways to measure correlation:

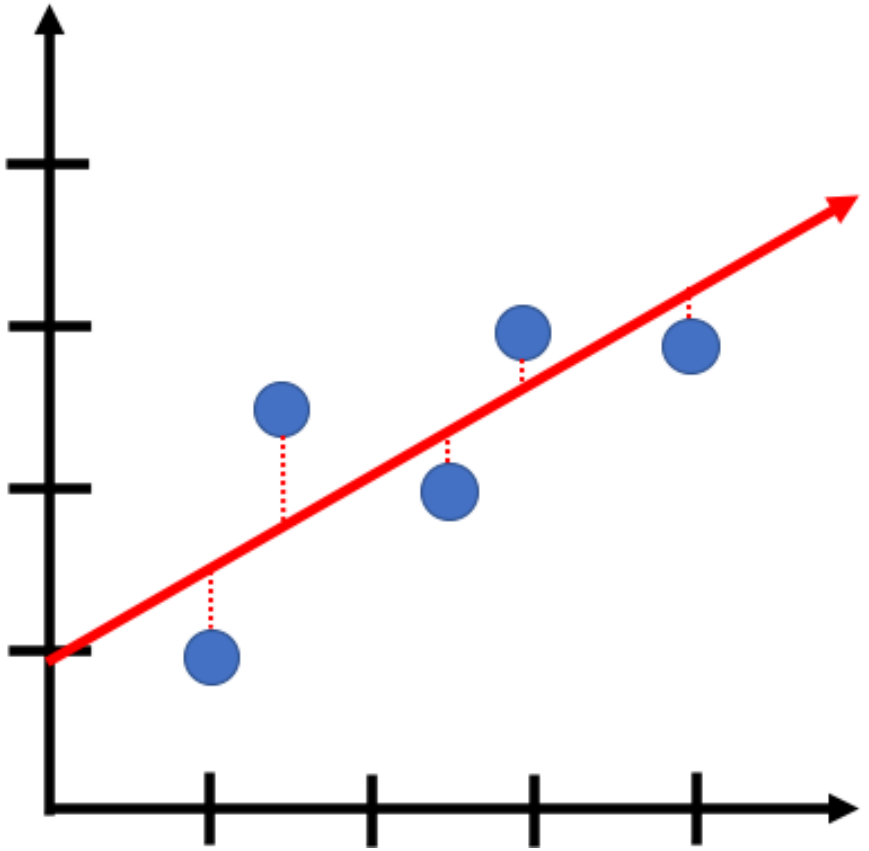
- Pearson (linear): is there a *linear* association between variable 1 and variable 2
- Spearman (monotonic): do things move together in the same order, even if not linear?

Section 2
Simple Linear Regression

Simple linear regression

All equivalent goals:

- Find the line of best fit
- Predict one variable using another
- Understand the impact of one variable on another
- Understand what drives variable Y



What is the line of best fit?

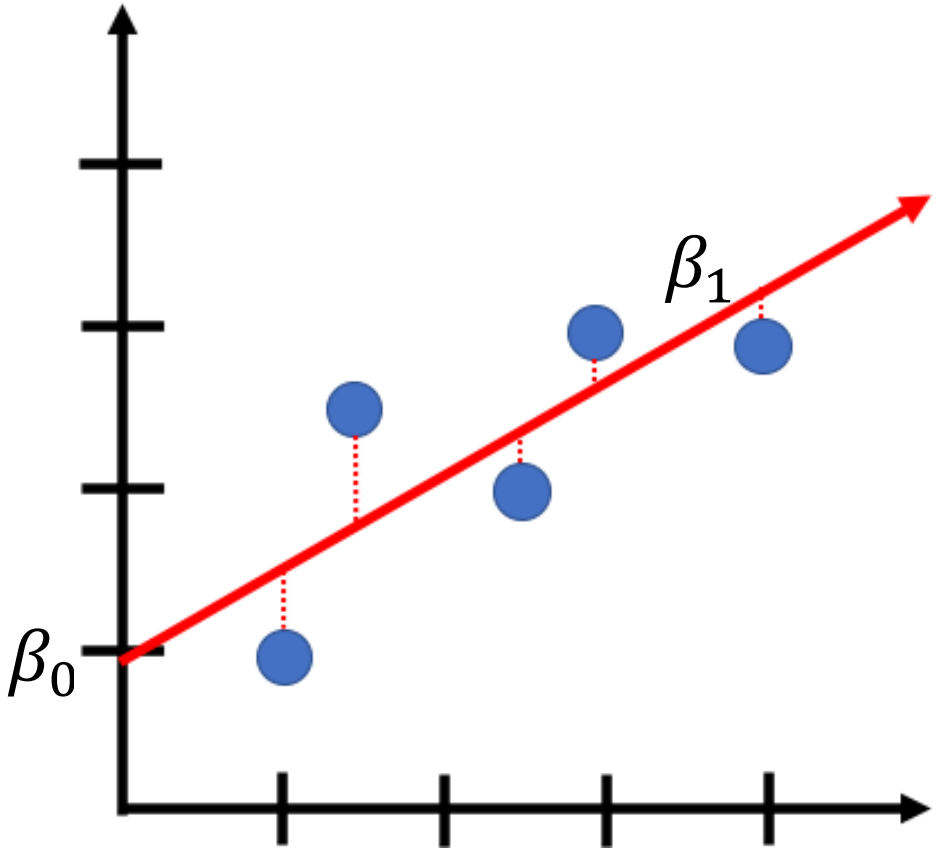
$$Y = \beta_0 + \beta_1 X$$

Y = dependent variable

X = independent variable

β_0 = Y-intercept

β_1 = slope



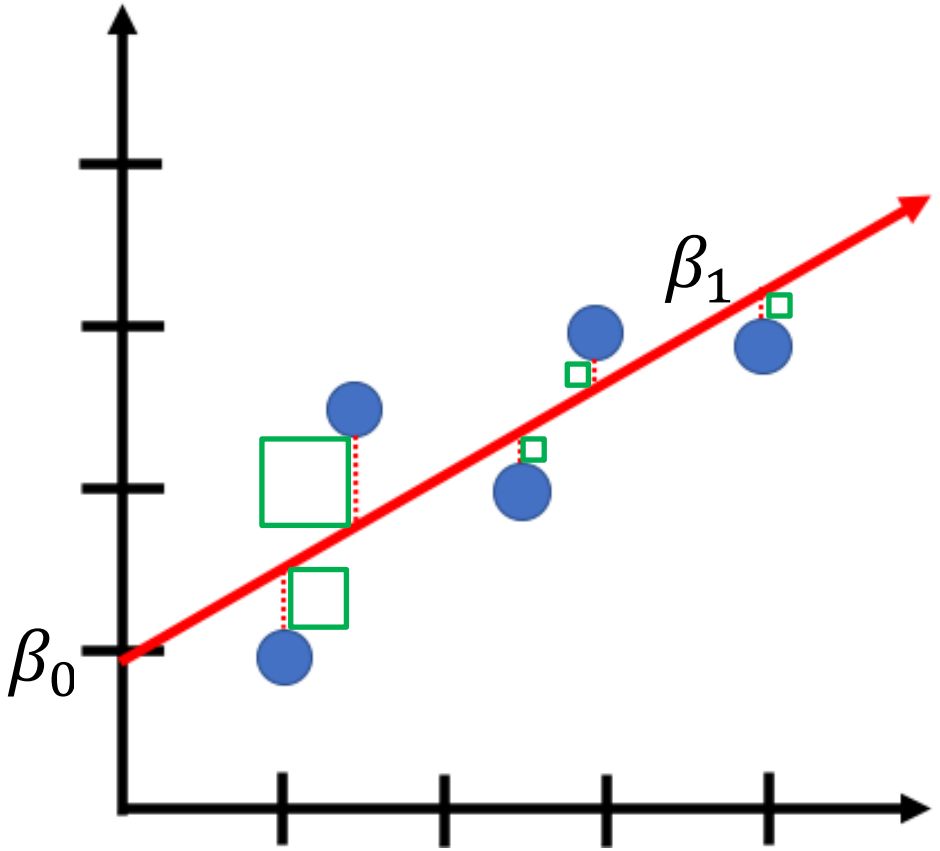
HOA 730	Algebra 1
$Y = \beta_0 + \beta_1 X$	y = mx + b
β_0	b (y-intercept)
β_1	m = slope

UNLV

How do we choose the line?

$$Y = \beta_0 + \beta_1 X$$

- There are many possible lines we can draw
- Regression picks the line that minimizes (mathematically) the total squared error
- Error = actual value – predicted value (on the line)



How do we interpret the line?

$$Y = \beta_0 + \beta_1 X$$

Model Component	Interpretation	Example
β_0	Predicted value of Y when X = 0. Represents the baseline	• Predicted ADR when review score is equal to 0
β_1	Change in Y for a 1-unit increase in X	• Change in ADR for a 1-unit increase in review score

UNLV

Regression Output

Y = ADR and X = distance to city center

Residuals = errors (distance from point to line)

Regression Coefficients (Betas)

Overall fit metrics. How does line fit to data?

```
> lm_dist <- lm(adr_usd ~ dist_center_mi, data = hotels)
> summary(lm_dist)

Call:
lm(formula = adr_usd ~ dist_center_mi, data = hotels)

Residuals:
     Min       1Q   Median       3Q      Max
-106.382  -23.368    0.934   22.825  117.912

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    241.1935     1.2217  197.42   <2e-16 ***
dist_center_mi  -9.8143     0.1816  -54.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.01 on 2998 degrees of freedom
Multiple R-squared:  0.4935,    Adjusted R-squared:  0.4933
F-statistic:  2921 on 1 and 2998 DF,  p-value: < 2.2e-16
```

UNLV

Regression Coefficients

```
> lm_dist <- lm(adr_usd ~ dist_center_mi, data = hotels)
> summary(lm_dist)

Call:
lm(formula = adr_usd ~ dist_center_mi, data = hotels)

Residuals:
     Min       1Q   Median       3Q      Max
-106.382  -23.368    0.934   22.825  117.912

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    241.1935     1.2217  197.42   <2e-16 ***
dist_center_mi  -9.8143     0.1816  -54.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.01 on 2998 degrees of freedom
Multiple R-squared:  0.4935,    Adjusted R-squared:  0.4933
F-statistic:  2921 on 1 and 2998 DF,  p-value: < 2.2e-16
```

Interpretations:

- **(intercept):** when the hotel is 0 miles from the city center, the ADR is $241
- **dist_center_mi:** for every 1 mile we get away from the city center, the ADR goes down by $9.81

UNLV

Regression Coefficients

```
> lm_dist <- lm(adr_usd ~ dist_center_mi, data = hotels)
> summary(lm_dist)

Call:
lm(formula = adr_usd ~ dist_center_mi, data = hotels)

Residuals:
     Min       1Q   Median       3Q      Max
-106.382  -23.368    0.934   22.825  117.912

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    241.1935     1.2217  197.42   <2e-16 ***
dist_center_mi  -9.8143     0.1816  -54.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.01 on 2998 degrees of freedom
Multiple R-squared:  0.4935,    Adjusted R-squared:  0.4933
F-statistic:  2921 on 1 and 2998 DF,  p-value: < 2.2e-16
```

“the variable is significant”

- Same test as we were doing before
  - H0: $\beta_1 = 0$
  - Ha: $\beta_1 \neq 0$
- p-value < .05

- **We don’t worry about the intercept, but dist_center_mi < .05 → there is an effect**

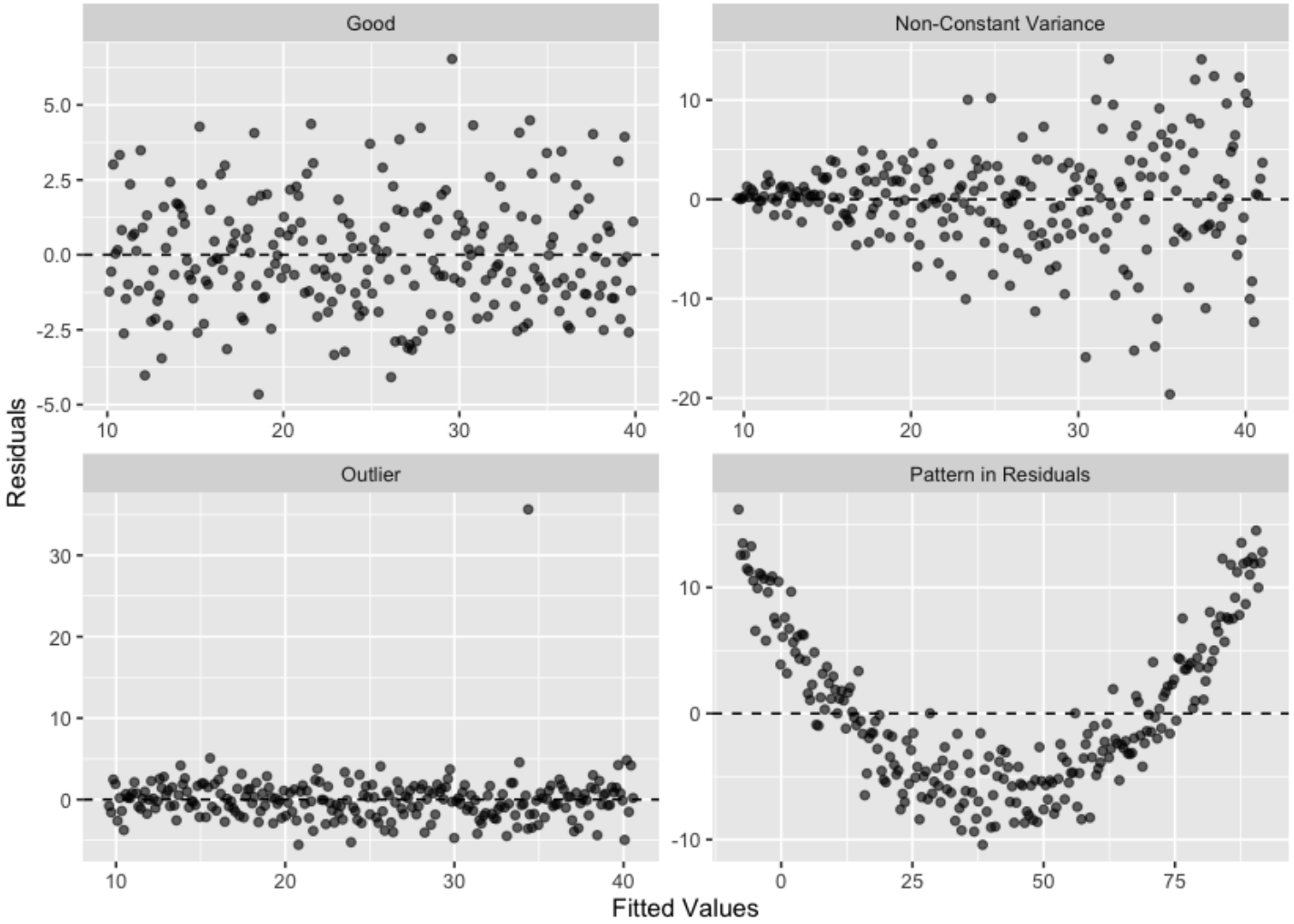
R^2

```
> lm_dist <- lm(adr_usd ~ dist_center_mi, data = hotels)
> summary(lm_dist)

Call:
lm(formula = adr_usd ~ dist_center_mi, data = hotels)

Residuals:
     Min       1Q   Median       3Q      Max
-106.382  -23.368    0.934   22.825  117.912

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    241.1935     1.2217  197.42   <2e-16 ***
dist_center_mi  -9.8143     0.1816  -54.05   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.01 on 2998 degrees of freedom
Multiple R-squared:  0.4935,    Adjusted R-squared:  0.4933
F-statistic:  2921 on 1 and 2998 DF,  p-value: < 2.2e-16
```

- R^2 represents the percent of variation that is explained by the model
- The higher the R^2 the better the model
- r = correlation coefficient. So if you square the correlation you get the R^2

UNLV

Assumptions



- The relationship is roughly linear
- The errors (residuals) have no clear pattern
- The spread of errors (residuals) is consistent
- No extreme outliers are driving the results