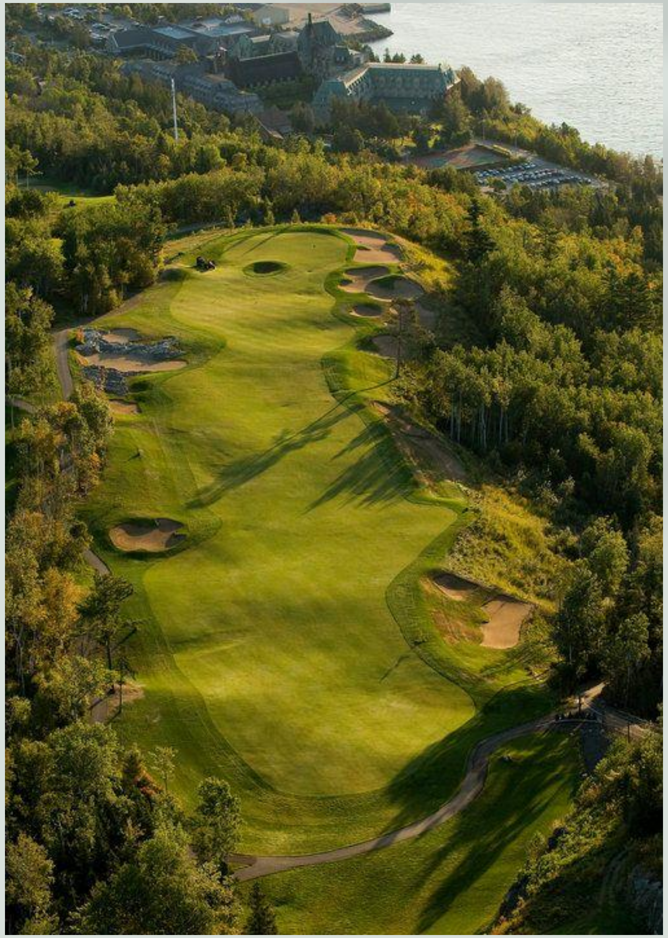
Linear Regression

Harshvardhan

University of Tennessee

Teaching Demo at UMass Amherst | December 2, 2024

Table of Contents



Case Study:
Number of Golf Courses in USA

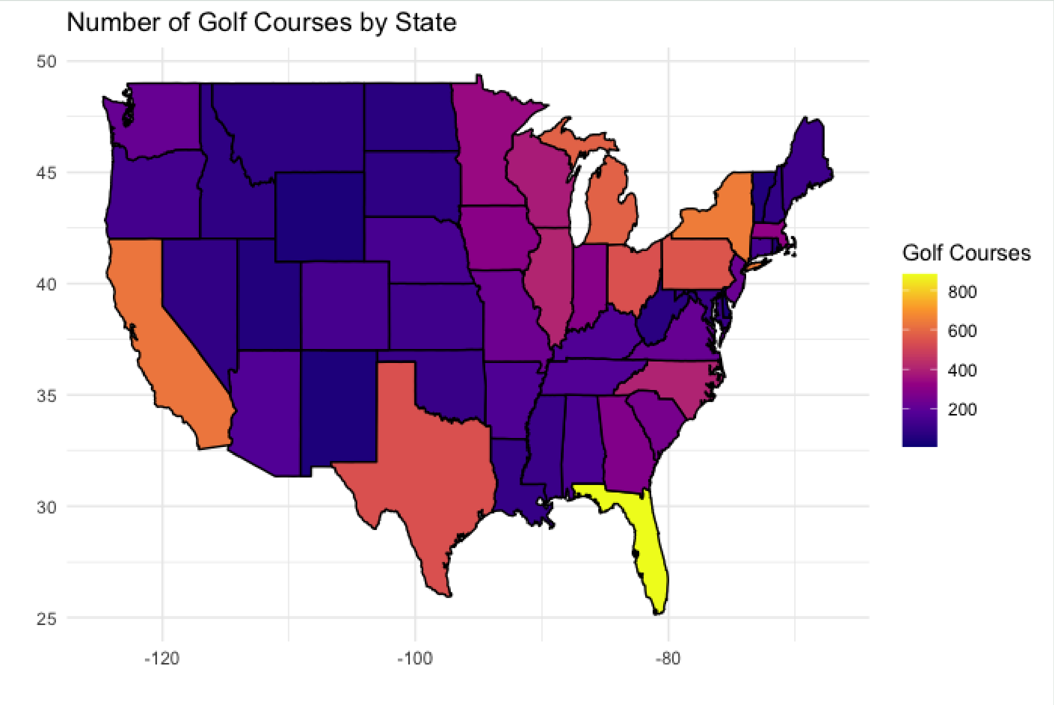
Golf Courses by State: How Many Are There?

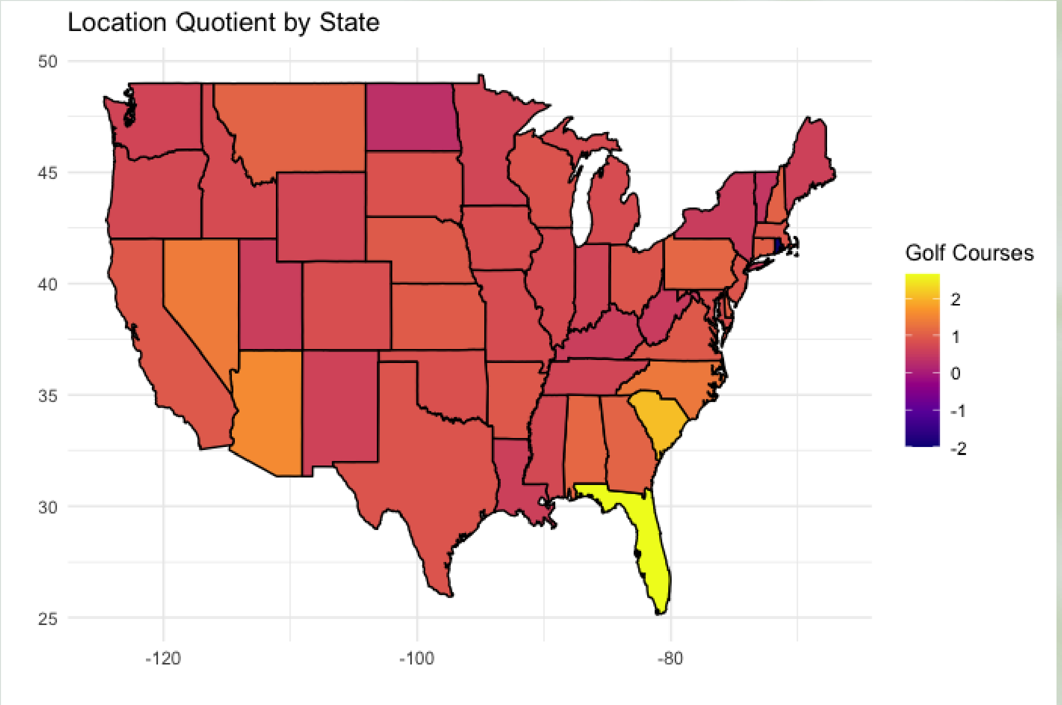
- In 2023, more than 45 million people played golf in US
- What determines how popular is golf in any state?
  - Location
  - Population / Population Density
  - Population Growth
  - Etc.

Fun fact: Florida has the highest number of golf courses in USA, currently 1000+

- golf.png


THE BEST GOLF PARTNERS ARE
THE ONES WHO ARE JUST A
LITTLE BIT WORSE THAN YOU.





Golf Courses by Location

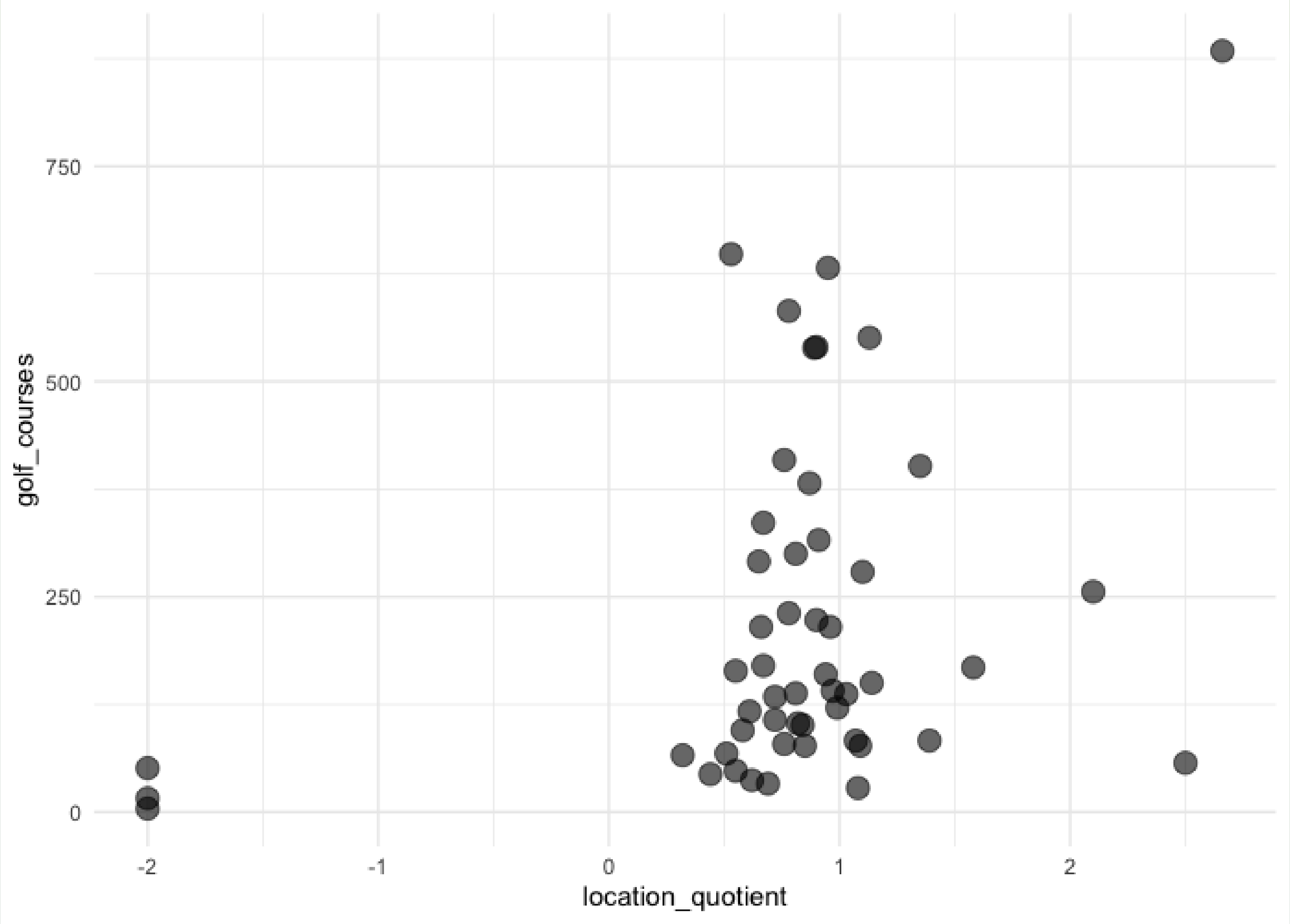
Location Quotient: A numerical index indicating location's favorability to have golf courses

What is Linear Regression?

- Linear regression is a *model* that estimates *linear relationship* between a dependent variable and one or more independent variable(s)
- It is an attempt to find the best fit line between independent and dependent variables
- **Strengths:**
  - Explainability and interpretability
  - Simple, quick and easy
  - Statistical basis for usage and interpretation
- **Weaknesses:**
  - Assumes linear relationship – simple model is too simple
  - Sensitive to outliers
  - Assumptions – we will talk about them
  - **No causation implied, only a sophisticated form of correlation**

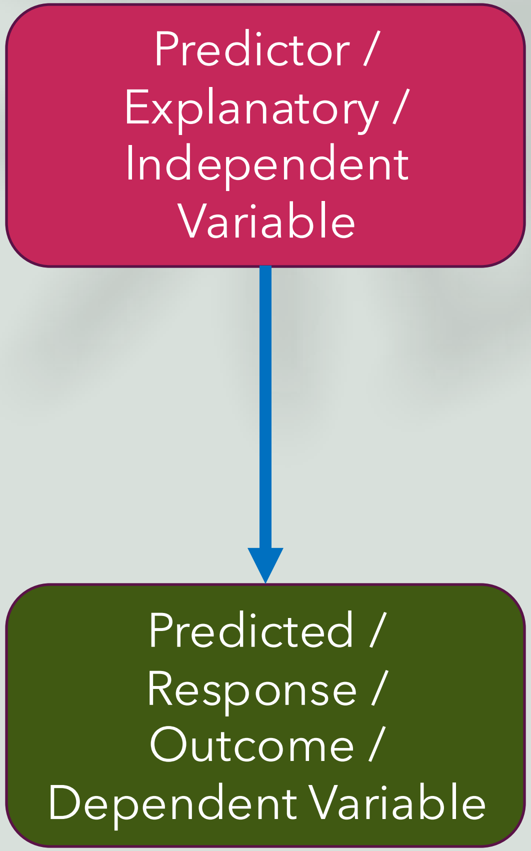
Scatterplot and Correlation

- Note for:
  - Direction
  - Strength
  - Outliers
- Linear regression is square of correlation between Y and X
- In multilinear regression, its squared correlation between $\hat{Y}$ and Y (Why?)
- https://digitalfirst.bfwpub.com/stats_applet/stats_applet_5_correg.html



Key Components

- **Dependent Variable:**
  - Variable to be predicted or explained
  - Also known as **Response** or **Outcome** variables.
  - Usually written as y_i
  - Example: *Number of golf courses*
- **Independent Variable(s):**
  - Variables used to predict or explain dependent variable
  - Also known as **Predictor** or **Explanatory** variables
  - Usually written as x_i
  - Example: *Location Quotient*



Linear Regression Mathematics

Linear Model, Coefficients and Predicted Responses

Mathematics of Linear Regression

- Simple Linear Regression (single predictor)

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Multiple Linear Regression (*p* predictors)

$$y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_p x_i^p + \epsilon_i$$
$$Y = X\beta + \mathrm{E}$$

Calculating Coefficients

Simple Linear Regression

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

- $\beta_1 = {}^{Cov(x,y)}/_{Var(x)}$
  $$= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$
- $\beta_0 = \bar{y} \; - \beta\bar{x}$
- β_1 measures how much y changes w.r.t. x, standardized by variability in x

Multiple Linear Regression

$Y = X\beta + \mathrm{E}$

- $\beta = (X'X)^{-1}X'Y$
- where X is data matrix, Y is response vector

Predicted Response: $\hat{Y}$

- $\hat{Y}$ is the estimated or predicted value of the dependent variable Y based on the estimated linear regression model
- $\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$ – Notice there is no residual term here. Why?
- **Interpretation:** $\hat{y}_i$ is the best guess we have for y_i given information about (x_i, y_i)
- **Residual:** $\hat{\epsilon}_i = y_i - \hat{y}_i$
- **Golf Example:** Number of golf courses given location, etc.

Interpretation, Assumptions and Diagnostics

How reliable are our conclusions?

Significance and Strength of Relationship

```
> fit = lm(golf_courses ~ location_quotient, data = df)
> summary(fit)

Call:
lm(formula = golf_courses ~ location_quotient, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-305.33 -120.55  -72.85   81.32  508.28 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         153.08      35.79   4.278 8.73e-05 ***
location_quotient    83.70      31.81   2.631   0.0113 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 186.6 on 49 degrees of freedom
Multiple R-squared:  0.1238,	Adjusted R-squared:  0.1059 
F-statistic: 6.923 on 1 and 49 DF,  p-value: 0.01134
```

- **P-value** measures significance of a relationship, typical threshold is 0.05

- **R-squared** measures proportion of variance in Y explained by Xs
  - R2 = 0 means no relationship
  - R2 = 1 implies perfect relationship
  - Also called "goodness of fit"

- **Adjusted R-squared** accounts for number of variables

- **F-statistic** tests whether the regression model provides a better fit than a model with no predictors (i.e. simple mean)
  - Higher is better (and will have low p-value)

- **Residual Standard Error** measures average distance between $\widehat{\boldsymbol{Y}}$ and $\boldsymbol{Y}$

Interpretation

```
> fit = lm(golf_courses ~ location_quotient, data = df)
> summary(fit)

Call:
lm(formula = golf_courses ~ location_quotient, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-305.33 -120.55  -72.85   81.32  508.28 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         153.08      35.79   4.278 8.73e-05 ***
location_quotient    83.70      31.81   2.631   0.0113 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 186.6 on 49 degrees of freedom
Multiple R-squared:  0.1238,	Adjusted R-squared:  0.1059 
F-statistic: 6.923 on 1 and 49 DF,  p-value: 0.01134
```

Based on the results:

1. Dependent variable (Y) = ________________
2. Independent variable (X) = _______________
3. Linear model, mathematically:

Interpretation

```
> fit = lm(golf_courses ~ location_quotient, data = df)
> summary(fit)

Call:
lm(formula = golf_courses ~ location_quotient, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-305.33 -120.55  -72.85   81.32  508.28 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)         153.08      35.79   4.278 8.73e-05 ***
location_quotient    83.70      31.81   2.631   0.0113 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 186.6 on 49 degrees of freedom
Multiple R-squared:  0.1238,	Adjusted R-squared:  0.1059 
F-statistic: 6.923 on 1 and 49 DF,  p-value: 0.01134
```

Is regression model significant?

- F-statistic =
- R2 =
- Adj R2 =

R2 Interpretation:

Is the relationship between location and number of golf courses significant?

- P-value =

Slope Interpretation:

Intercept Interpretation:

Multiple Linear Regression

- We can choose to include more than one variables in regression. Let's see an example.

```
> mlr = lm(golf_courses ~ location_quotient + population + growth + poulation_density,
data = df)
> summary(mlr)

Call:
lm(formula = golf_courses ~ location_quotient + population +
    growth + poulation_density, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-288.77  -57.55  -17.66   33.05  277.30

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        6.104e+01  2.738e+01   2.229   0.0307 *
location_quotient  4.875e+01  2.320e+01   2.101   0.0412 *
population         2.061e-05  2.168e-06   9.506 1.99e-12 ***
growth            -4.028e+01  2.754e+01  -1.463   0.1503
poulation_density  4.901e-03  1.101e-02   0.445   0.6584
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 110.5 on 46 degrees of freedom
Multiple R-squared:  0.7115,    Adjusted R-squared:  0.6865
F-statistic: 28.37 on 4 and 46 DF,  p-value: 6.632e-12
```

Interpret R2:

Estimate effect of 'population':

Confidence Intervals of Estimated Coefficients

confint (model) function in R can give us 95% confidence intervals for all intercepts

```
> confint(mlr)
                         2.5 %       97.5 %
(Intercept)       5.919174e+00 1.161585e+02
location_quotient 2.039280e+00 9.545420e+01
population        1.624440e-05 2.497202e-05
growth           -9.571472e+01 1.514525e+01
poulation_density -1.726619e-02 2.706847e-02
```

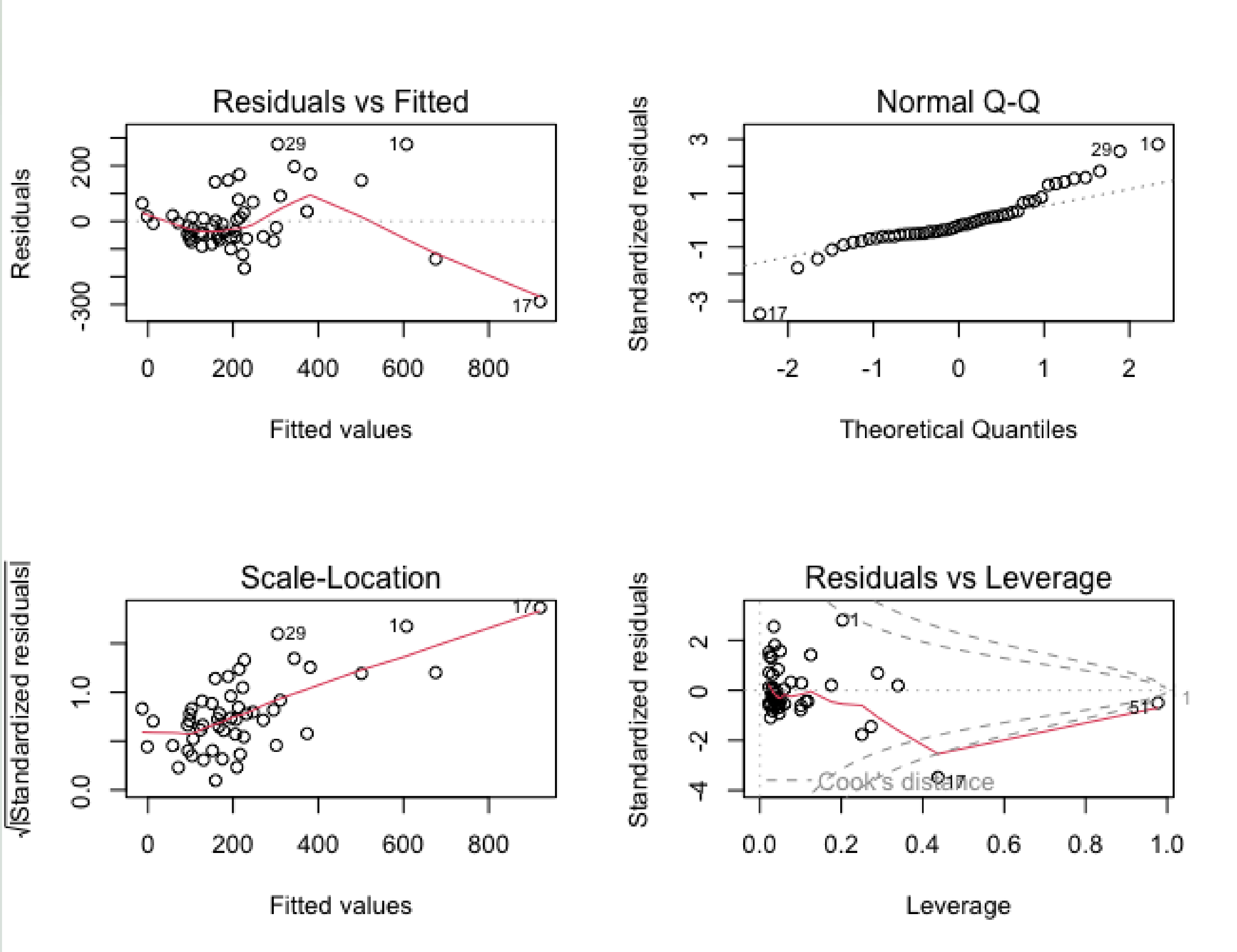
Assumptions in Linear Regression (LINE)

- Like all statistical models, Linear Regression works under certain assumptions:
  1. **L**inearity – assumed linear relationship between X and Y
  2. **I**ndependence – residuals (errors) are independent of each other
  3. **N**ormal distribution of residuals $N(0, \sigma^2)$
  4. **E**qual variance across values of X (homoskedasticity)

Model Diagnostics

- R provides four diagnostic plots:
  1. Residual vs Fitted plot (Linearity assumption)
     - Good if horizontal line shows with no distinct patterns
     - plot(model, which = 1)
  2. Normal Q-Q plot (Normality assumption)
     - Good if residuals follow diagonal dotted line
     - plot(model, which = 2)
  3. Scale-Location plot (Equal variance or Homoskedasticity assumption)
     - Good if horizontal line with equally spread points
     - plot(model, which = 3)
  4. Residuals vs Leverage (Detecting outliers)
     - Good if few points stand out
     - plot(model, which = 4)

Model Diagnostics Case



Mid-class Quiz

- Let’s see how much we understood from today’s class so far
- Visit https://play.blooket.com/ and enter code XXXXXX

Business Insights

Real-World Implications and Actionable Insights

Impact of Location, Population, and Value of Linear Model

- From our MLR results, we found that a one-unit increase in location quotient is associated with ~49 additional golf courses, holding other variables constant.
- For every additional 100,000 people in a state's population, the number of golf courses increases by ~2.
- Location, Population, Population Growth, and Population Density together explain 71% of variability in the number of golf courses by state.

Multiple Linear Regression

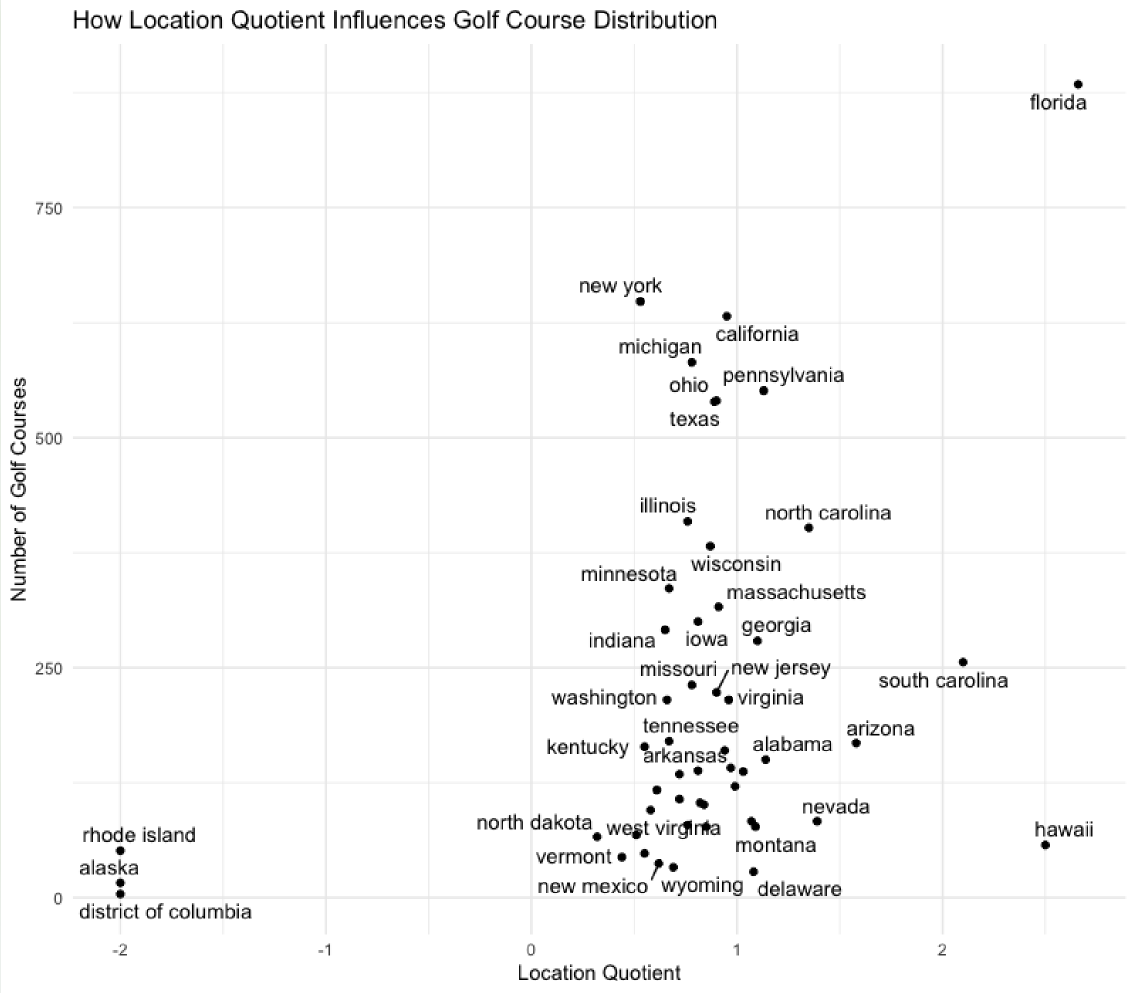
```
> mlr = lm(golf_courses ~ location_quotient + population + growth + poulation_density,
data = df)
> summary(mlr)

Call:
lm(formula = golf_courses ~ location_quotient + population +
    growth + poulation_density, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-288.77  -57.55  -17.66   33.05  277.30

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.104e+01  2.738e+01   2.229   0.0307 *
location_quotient 4.875e+01  2.320e+01   2.101   0.0412 *
population        2.061e-05  2.168e-06   9.506 1.99e-12 ***
growth           -4.028e+01  2.754e+01  -1.463   0.1503
poulation_density 4.901e-03  1.101e-02   0.445   0.6584
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 110.5 on 46 degrees of freedom
Multiple R-squared:  0.7115,	Adjusted R-squared:  0.6865
F-statistic: 28.37 on 4 and 46 DF,  p-value: 6.632e-12
```

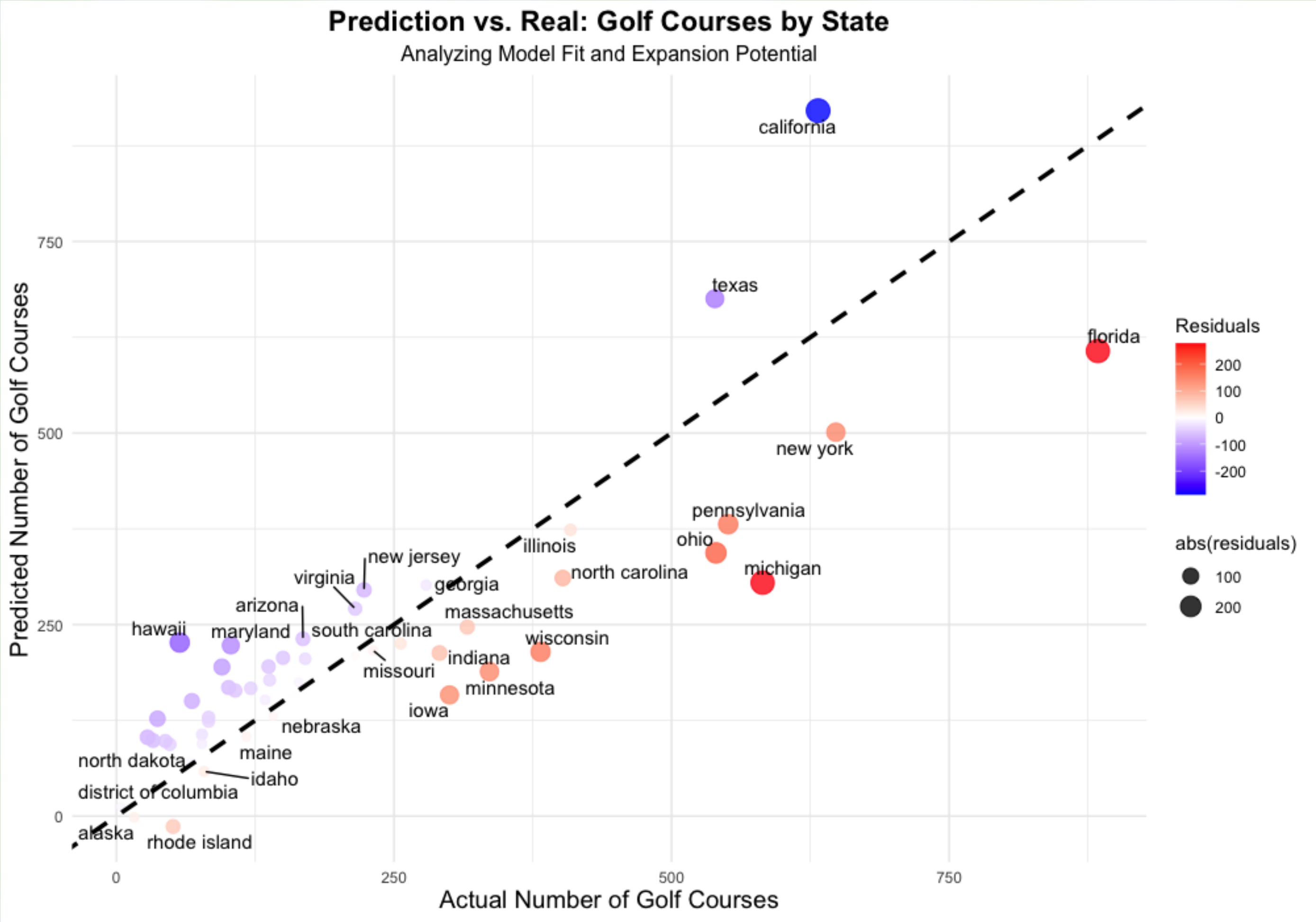


Location Matters!

Strategic Golf Course Development

- **Insight:**
  - Location matters! (location_quotient is significant.)
  - States with a higher Location Quotient (e.g., Florida, Hawaii):
    - More favorable for golf courses due to climate and topography.
    - Strong existing presence of golf courses.
- **Action:**
  - Target high-Location Quotient regions for:
    - Premium golf-related tourism?
  - Leverage natural advantage to attract domestic and international clientele at those locations

Market Potential



- We can see market potential by visualizing the residuals.
  - **Underestimated States** (Red Points)
    - More favourable than expected
  - **Overestimated States** (Blue Points)
    - Less favourable than expected
  - **Well-Fitted States** (White Points)
    - As favourable as expected

Concluding Remarks

What we learned today?

- Linear regression helps us identify relationship and patterns between *independent* and *dependent* variables
- Location quotient and population matter in predicting the number of golf courses in a state
- Model coefficients tell us the impact of an individual variable
- RSquared tells us "goodness of fit" of a model
- Assumptions of LR should be verified before interpretation
- Residuals analysis can tell us more about the data than we think

What to remember?

- Regression's strength is interpretability
- Regression coefficient being significant doesn't imply causality
- Assumption of "linear model" might be too simplistic for pure prediction
- **LINE Assumptions:** Linearity, Independence, Normality, Equal Variance
- **There is more to regression! We will cover additional topics in coming classes**

Next Up…

- **Interaction between Variables**
  - Identifying relationships that have combined/conditional effects, not additive effects
  - Effect of exercise on weight might depend on diet type
- **Ranking Importance of Independent Variables**
  - Scaling X by its mean and standard deviation
- **Linear regression for optimization**
  - https://blog.harsh17.in/using-linear-regression-to-find-optimal-value/
- **Variable selection:** Lasso and ridge regressions


Questions?