Linear Regression

Harshvardhan

2024-11-26

Rising Popularity of Golf

Golf has experienced a notable resurgence in the United States, with participation reaching 45 million Americans. This growth reflects a demographic shift, as the sport becomes more inclusive and appealing to a broader audience. Over the past five years, the number of women and girls playing on courses has increased by 23%, and participation among Asian, Black, and Hispanic golfers has risen by 43%. Additionally, junior golfers have seen a 40% uptick in participation. These trends indicate that golf is becoming younger and more diverse, marking an exciting era for the sport.

But golf courses aren't popular in all states! Some states are more suitable to play golf like tropical state of Florida. On the other hand, some are naturally not the best like Alaska. What can we learn from the number of golf courses?

NOTE: The data used for this demonstration follows the general trend but hasn't been updated in a while.

Loading Libraries and Data

We need to load the data to do some analysis! First of all, we will load required libraries and the data set in CSV format. To make sure all names are in snake_case (my preferred style), I use `janitor` package's `clean_names()` function.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(maps)
```

```
## Warning: package 'maps' was built under R version 4.2.3
```

```
##
## Attaching package: 'maps'
##
## The following object is masked from 'package:purrr':
##
##     map
```

```
theme_set(theme_minimal())
```

```
df = read_csv("Golf Courses.csv") %>%
   janitor::clean_names()
```

```
## Rows: 51 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): State, Growth
## dbl (2): Golf Courses, Location Quotient
## num (2): Population, Poulation Density
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df
```

```
## # A tibble: 51 x 6
##    state      golf_courses location_quotient population growth poulation_density
##    <chr>             <dbl>             <dbl>      <dbl> <chr>              <dbl>
##  1 Florida             884              2.66   22177997 1.06%              414.
##  2 Hawaii               57              2.5      1401709 -0.34%             218.
##  3 South Car~          256              2.1      5342388 1.22%              178.
```

```
##  4 Arizona             168              1.58    7640796 1.60%              67.3
##  5 Nevada               83              1.39    3238601 1.66%              29.5
##  6 North Car~          402              1.35   10807491 0.99%             222.
##  7 Alabama             150              1.14    4949697 0.31%              97.7
##  8 Pennsylva~          551              1.13   12805190 0.01%             286.
##  9 Georgia             279              1.1    10936299 0.98%             190.
## 10 Montana              77              1.09    1093117 0.75%               7.51
## # i 41 more rows
```

Some data cleaning is needed: growth column has percentage sign % at the end that needs to be removed.

```
# Substitute "%" with "" and convert to numeric
df$growth = gsub("%", "", df$growth) %>%
   as.numeric()

df
```

```
## # A tibble: 51 x 6
##    state      golf_courses location_quotient population growth poulation_density
##    <chr>             <dbl>             <dbl>      <dbl>  <dbl>             <dbl>
##  1 Florida             884              2.66   22177997   1.06             414.
##  2 Hawaii               57              2.5     1401709  -0.34             218.
##  3 South Car~          256              2.1     5342388   1.22             178.
##  4 Arizona             168              1.58    7640796   1.6               67.3
##  5 Nevada               83              1.39    3238601   1.66              29.5
##  6 North Car~          402              1.35   10807491   0.99             222.
##  7 Alabama             150              1.14    4949697   0.31              97.7
##  8 Pennsylva~          551              1.13   12805190   0.01             286.
##  9 Georgia             279              1.1    10936299   0.98             190.
## 10 Montana              77              1.09    1093117   0.75               7.51
## # i 41 more rows
```

Does location determine number of golf courses?

One way to answer this question is through creating a state-wise map plot of the number of golf courses in the USA. We will use `maps` package in R for this.

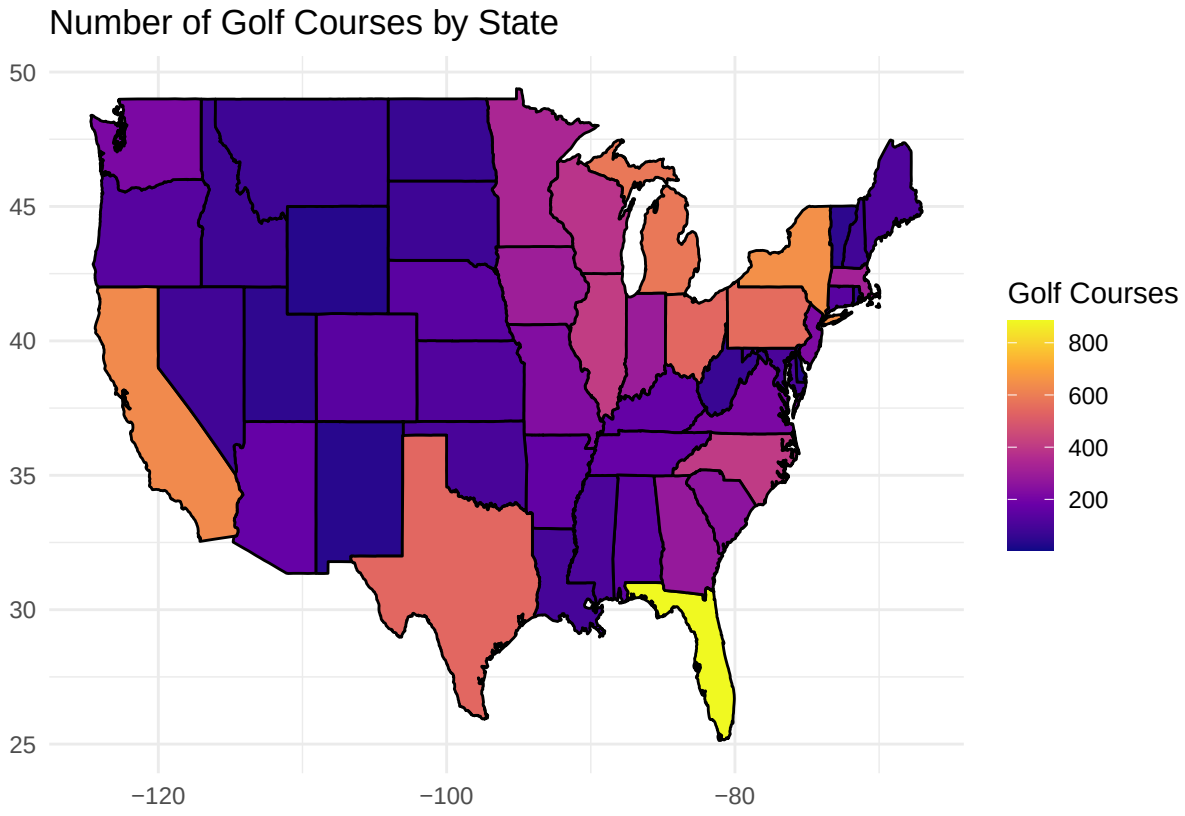
```
# Load US state map data
us_states = map_data("state")

# Convert state names to lowercase to match the map data
df$state = tolower(df$state)

# Merge map data with golf data
map_data = us_states %>%
   left_join(df, by = c("region" = "state"))


map_data %>% tibble() # Print beautifully
```

```
## # A tibble: 15,537 x 11
##     long   lat group order region  subregion golf_courses location_quotient
##    <dbl> <dbl> <dbl> <int> <chr>   <chr>            <dbl>             <dbl>
##  1 -87.5  30.4     1     1 alabama <NA>               150              1.14
##  2 -87.5  30.4     1     2 alabama <NA>               150              1.14
##  3 -87.5  30.4     1     3 alabama <NA>               150              1.14
```

```
##  4 -87.5  30.3     1     4 alabama <NA>                150              1.14
##  5 -87.6  30.3     1     5 alabama <NA>                150              1.14
##  6 -87.6  30.3     1     6 alabama <NA>                150              1.14
##  7 -87.6  30.3     1     7 alabama <NA>                150              1.14
##  8 -87.6  30.3     1     8 alabama <NA>                150              1.14
##  9 -87.7  30.3     1     9 alabama <NA>                150              1.14
## 10 -87.8  30.3     1    10 alabama <NA>                150              1.14
## # i 15,527 more rows
## # i 3 more variables: population <dbl>, growth <dbl>, poulation_density <dbl>
```

Next, we will create the map plot of the number of golf courses by state.

```
ggplot(map_data, aes(x = long, y = lat, group = group, fill = golf_courses)) +
  geom_polygon(color = "black") +
  scale_fill_viridis_c(option = "C", name = "Golf Courses") +
  theme_minimal() +
  labs(
    title = "Number of Golf Courses by State",
    x = "",
    y = ""
  )
```

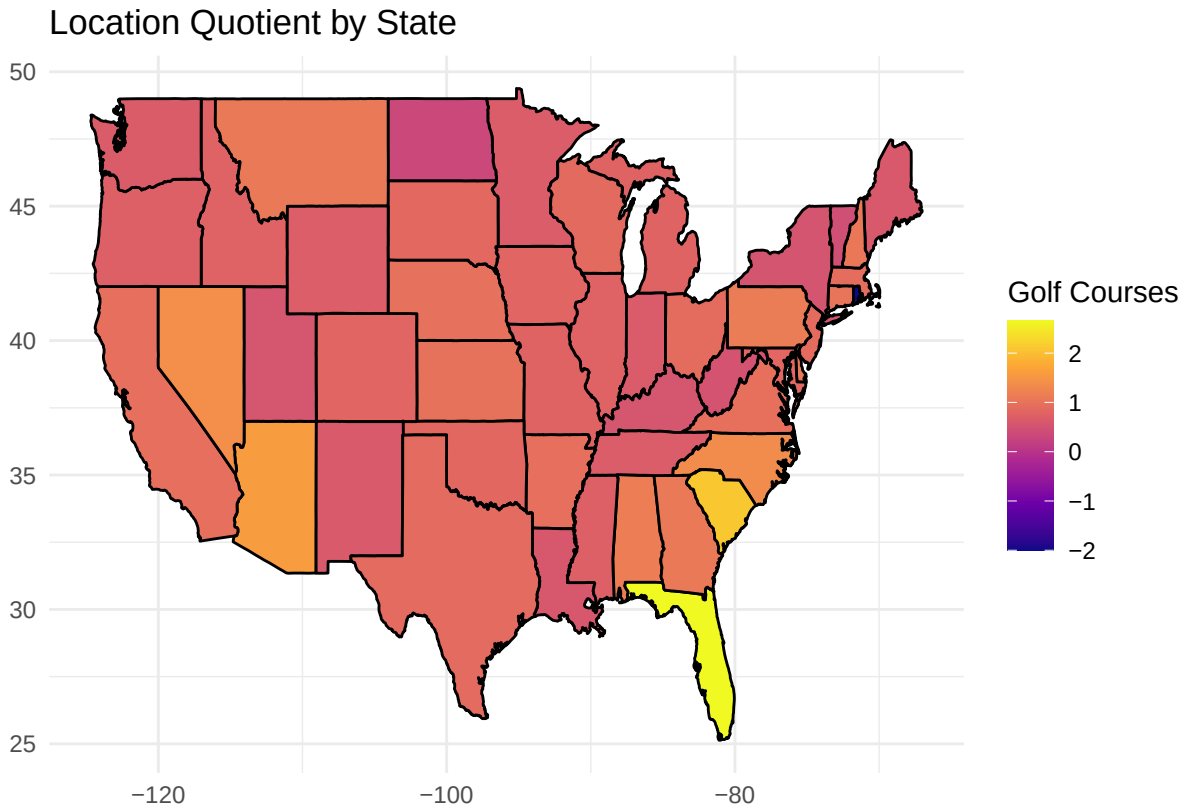


Some locations are more favourable to playing golf due their climate, like Florida. In our data, this is noted through the column `location_quotient`.

Let's look at its distribution.

```
ggplot(map_data, aes(x = long, y = lat, group = group, fill = location_quotient)) +
  geom_polygon(color = "black") +
  scale_fill_viridis_c(option = "C", name = "Golf Courses") +
  theme_minimal() +
```

```
  labs(
    title = "Location Quotient by State",
    x = "",
    y = ""
  )
```

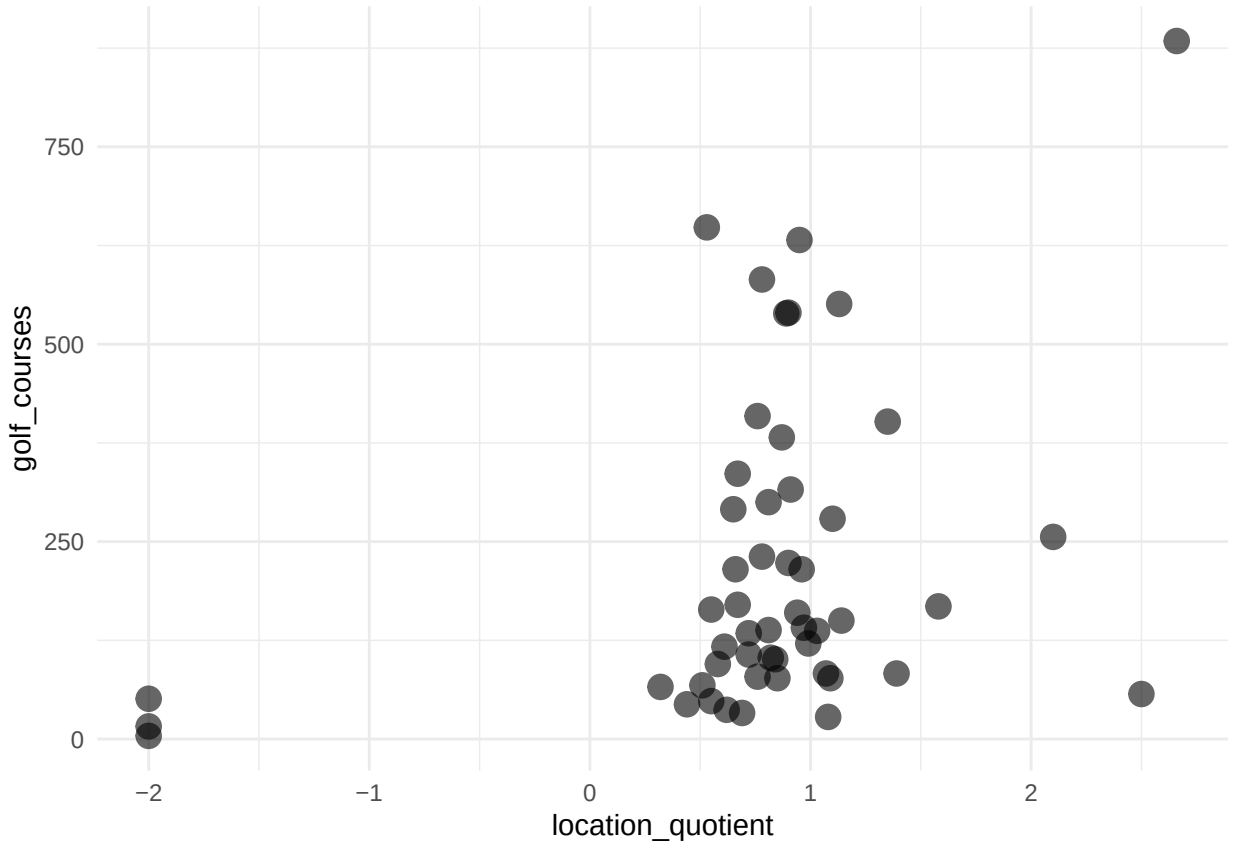


We see that regions around Florida are most favourable, while generally northern states are least favourable.

Scatterplot

To the see the relationship between **Location Quotient** and **Number of Golf Courses** visually, we should use a scatterplot. Recall that a scatterplot shows relationship between two quantitative variables.

```
ggplot(df, aes(x = location_quotient, y = golf_courses)) +
   geom_point(size = 4, alpha = 0.6)
```



If we look at this, it becomes clear that high location quotient helps, but there is a LOT of variation.

Thus the question becomes: is there a ***significant*** relationship between location and number of golf courses? (Food for thought: what do we mean by *significant*?)

Variables

1. **Independent Variable:** Variable(s) that are used to predict or explain changes in another variable. **Location Quotient** is independent variable in our case. It is also known as predictor or explanatory variable.
2. **Dependent Variable:** Variable(s) that we are trying to predict or explain. **Number of Golf Courses** is dependent variable in our case. It is also known as response or outcome variable.

Mathematical Model

Mathematically, we can describe this model with a simple linear equation:

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

What do the components mean?

1. y_i: The dependent variable, in this case, the number of golf courses in state i.
2. x_i: The independent variable (predictor), here the location quotient in state i.
3. β_0: The intercept, which represents the expected value of y (number of golf courses) when x (location quotient) is 0. It is the starting point of the model on the y-axis.
4. β_1: The slope, which measures how much y (number of golf courses) changes, on average, for a one-unit increase in x (location quotient). It captures the strength and direction of the relationship between x and y.

5. ϵ_i: The residual, which is the error term for state i. It represents the difference between the observed value y_i and the predicted value $\hat{y}_i$ from the model. Mathematically:

$$\epsilon_i = y_i - \hat{y}_i.$$

How are β_0 (intercept) and β_1 (slope) calculated?

The values of β_0 and β_1 are determined using **least squares regression**, which minimizes the sum of squared residuals, $\sum \epsilon_i^2$. This ensures the best possible fit for the data.

1. **Slope (β_1)**:
$$\beta_1 = \frac{\text{Cov}(x, y)}{\text{Var}(x)} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}.$$
   - This formula calculates how much y changes for a unit change in x, standardized by the variability in x.
2. **Intercept (β_0)**:
$$\beta_0 = \bar{y} - \beta_1 \bar{x}.$$
   - Here, $\bar{x}$ is the mean of x, and $\bar{y}$ is the mean of y. This centers the line through the data's average values.

What do these values mean in context?

- **Intercept (β_0)**: If a state's location quotient (x) is 0, the model predicts that the state would have β_0 golf courses on average. While it may not be realistic for x to be exactly 0, the intercept gives a baseline for the relationship.
- **Slope (β_1)**: For each one-unit increase in the location quotient, the number of golf courses is expected to increase (or decrease) by β_1, keeping all other predictors constant. A positive slope means more golf courses with higher location quotient, and a negative slope would indicate fewer.
- **Residual (ϵ_i)**: This is the "unexplained" part of the number of golf courses for state i. A smaller residual means the model prediction is closer to the actual data point.

Putting it together

- $y_i = \beta_0 + \beta_1 x_i$ represents the line that best fits the data, with β_0 and β_1 summarizing the relationship between x (location quotient) and y (number of golf courses).
- ϵ_i measures how well the model performs for each state. A large ϵ_i suggests the model may be missing some important factors for that state i.

Simple Linear Regression

Let's use R's `lm()` to fit a simple linear regression model. The model is called "simple" because it only has a single independent variable — Location Quotient.

```
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
```

What to look for?

1. **P-value**:
   - Indicates whether the relationship between a predictor (e.g., `location_quotient`) and the response variable (`golf_courses`) is statistically significant.
   - A small p-value (< 0.05) suggests the predictor has a meaningful impact on the response variable.
2. **R-squared (R^2)**:
   - Represents the proportion of variance in the response variable that is explained by the predictor(s).
   - Values range from 0 to 1:
     - $R^2 = 0$: No relationship.
     - $R^2 = 1$: Perfect fit.
3. **F-statistic**:
   - Tests whether the regression model as a whole provides a better fit than a model with no predictors.
   - A large F-statistic with a low p-value indicates the model explains a significant portion of the variability in the response variable.
4. **Residual Standard Error (RSE)**:
   - Measures the average distance between observed values and predicted values.
   - Smaller values indicate a better fit.
5. **Significance Codes**:
   - Visual shorthand (e.g., `***`, `*`, `.`) to indicate how strong the evidence is against the null hypothesis ($H_0 : \beta = 0$) for each coefficient.

Interpreting the Output:

1. **P-value for Coefficients**:
   - β_1 (location_quotient): p-value = 0.0113 (< 0.05), meaning the predictor is statistically significant.
   - The relationship between location quotient and golf courses is unlikely due to random chance.
2. **R-squared (R^2)**:
   - $R^2 = 0.1238$: The model explains about 12.4% of the variation in `golf_courses`, suggesting a weak relationship.
3. **F-statistic**:
   - $F = 6.923, p = 0.01134$: The model as a whole is statistically significant at the 5% level.
4. **Residual Standard Error**:
   - 186.6: On average, the number of golf courses deviates from the predicted values by about 186.6.
5. **Significance Codes**:
   - `location_quotient` has a single `*`, indicating it is moderately significant ($p < 0.05$).

Multiple Linear Regression

Just like simple linear regression, we can use `lm()` to create a linear model with more than one independent variables. Here is how the outputs look like:

```
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
```

Interpreting the results

1. **Significant Predictors**:
   - `location_quotient`: p-value = 0.0412 (< 0.05), significant. A one-unit increase in location quotient is associated with ~49 additional golf courses, holding other variables constant.
   - `population`: p-value = 1.99×10^{-12} ($p < 0.001$), highly significant. For every additional 100,000 people in a state's population, the number of golf courses increases by ~2.
2. **Non-Significant Predictors**:
   - `growth`: p-value = 0.1503 ($p > 0.05$), not significant. No strong evidence that growth impacts the number of golf courses.
   - `poulation_density`: p-value = 0.6584 ($p > 0.05$), not significant.
3. **Model Fit**:
   - $R^2 = 0.7115$: The model explains ~71.15% of the variance in `golf_courses`.
   - Adjusted $R^2 = 0.6865$: Adjusted for the number of predictors, the model explains ~68.65% of the variance.
4. **Residual Standard Error**:
   - 110.5: On average, predictions deviate by ~110.5 golf courses from actual values.
5. **F-statistic**:
   - $F = 28.37, p = 6.632 \times 10^{-12}$: The model as a whole is highly significant ($p < 0.001$), confirming that at least one predictor contributes meaningfully.

How is this different from SLR?

- The model includes multiple predictors, allowing us to account for several factors simultaneously.
- R^2 is much higher in MLR (~71%) compared to SLR (~12%), suggesting that the additional predictors significantly improve the model's explanatory power.
- Interpretation of coefficients now considers the **effect of each predictor while holding others constant**, a key advantage of MLR.

95% Confidence Intervals

One can estimate 95% confidence intervals using R's `confint()` function. Here is the 95% confidence interval for all estimates. Note that whichever confidence interval includes zero, are the ones that aren't significant.

```
confint(mlr, level = 0.95)
```

```
##                           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 Models and Model Diagnostics

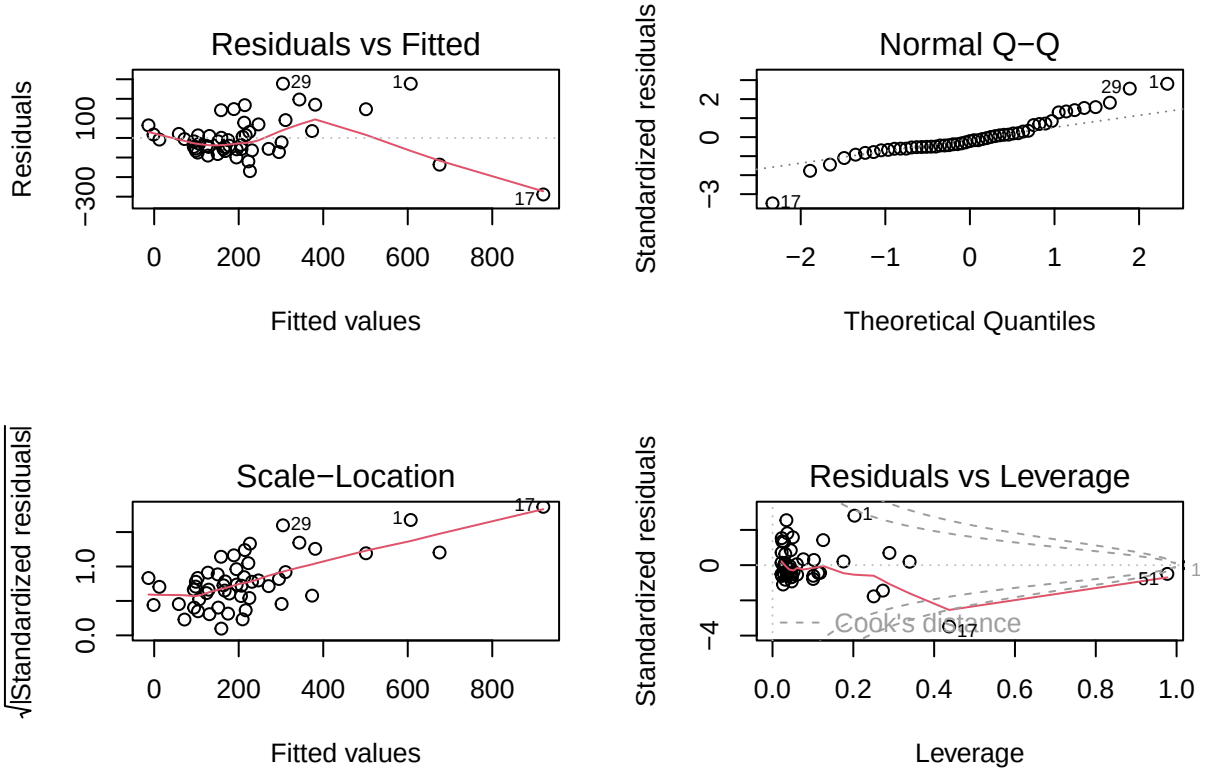
Linear regression relies on several assumptions to ensure valid results.

1. **Linearity** — assumed linear relationship between X and Y
2. **Independence** — residuals (errors) are independent of each other
3. **Normal** distribution of residuals $N(0, \sigma^2)$
4. **Equal variance** across values of X (homoskedasticity)

Let's create the diagnostic plots to check the assumptions of linear regression. Here's how you can do it:

```
# Create diagnostic plots
par(mfrow = c(2, 2))  # arrange plots in a 2x2 grid
plot(mlr)
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



1. Residual vs Fitted Plot

Ideally, residuals should be randomly scattered around the horizontal line at zero. If you observe a clear pattern, such as a curve, it suggests that a non-linear relationship may exist.

In our plot, the residuals seem to be somewhat randomly scattered but some values near the end indicate some extreme (fitted) values, $\hat{Y}$. Thus, it likely violates non-linearity assumption due to some data points (e.g. 1 and 17).

Should we remove those problematic data points? Maybe, but depends on data points.

2. Normality: Normal Q-Q Plot and Shapiro-Wilk Test

This plots residuals against quantiles of standard normal distribution to investigate if the residuals follow a normal distribution.

In our plot, most data points are fine — but 1, 29 and 17 are interesting.

Shapiro-Wilk test can also test if residuals are Normally distributed.

```
shapiro.test(resid(mlr))  # Shapiro-Wilk test
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(mlr)
## W = 0.93467, p-value = 0.007551
```

If $p - value > 0.05$, the residuals follow a normal distribution with 95% confidence.

3. Scale-Location Plot

This plot evaluates the homoskedasticity (equal variance of residuals). In this plot, if the residuals appear to be spread out equally around the line, it suggests that the assumption of equal variance is met. If we see a spread that increases or decreases, it indicates heteroscedasticity.

Again, in our plot, we see some outliers — same observations again!

Breusch-Pagan (BP) test from `lmtest` package in R specifically tests for homoskedasticity.

```
lmtest::bptest(mlr)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mlr
## BP = 26.196, df = 4, p-value = 2.889e-05
```

A $p - value > 0.05$ indicates homoskedasticity.

4. Residuals vs Leverage

This identifies influential observations. Points outside the Cook's distance lines are considered influential observations. If you see points that stand out far from the rest (high leverage), they can significantly affect the model. In your plot, if there are only a few points far from the main cluster, you might have some influential observations to consider.

Again, in our plot, we see some outliers and they are the same observations again!

5. Independence of Residuals

To test the independence of residuals, we try to calculate autocorrelation in the residuals — correlation with itself. (This is a critical concept in time-series analysis which you will learn in another course!)

We do this using Durbin-Watson test from `car` package in R.

```
car::durbinWatsonTest(mlr)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.06291598      1.989209   0.788
##  Alternative hypothesis: rho != 0
```

When $p - value > 0.05$, the residuals are likely independent of each other.

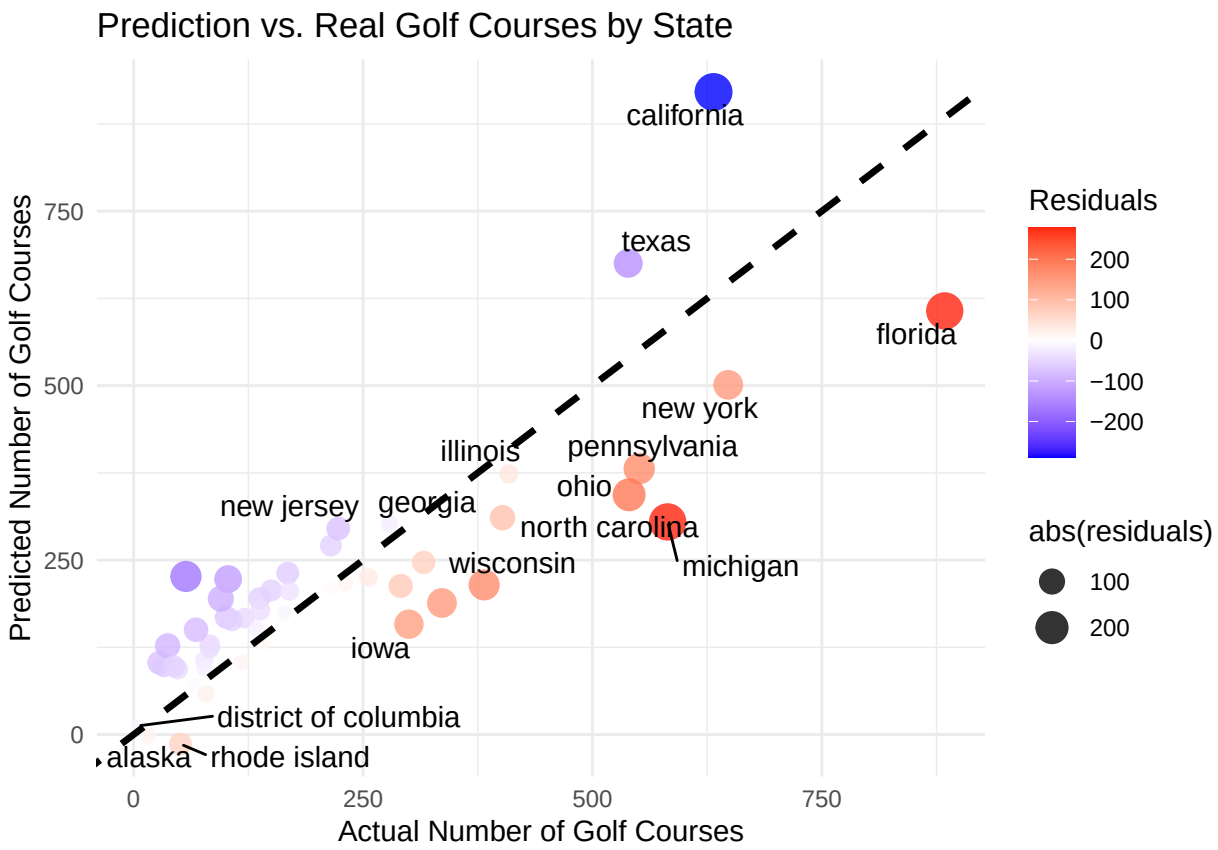
Predictions

Given our model, we can use the data to predict the number of golf courses. If the predicted number of golf courses differs widely from given number of golf courses, this indicates a potential for spreading the love of golf!

```
# Predict Yhat for all states
df$pred_golf_courses = predict(mlr, newdata = df)

# Calculate residuals
df$residuals = df$golf_courses - df$pred_golf_courses

ggplot(df, aes(x = golf_courses, y = pred_golf_courses, label = state)) +
  geom_point(aes(color = residuals, size = abs(residuals)), alpha = 0.8) + # Points for actual vs predi
  geom_abline(slope = 1, intercept = 0, color = "black", linetype = "dashed", size = 1.2) + # 1:1 line
  scale_color_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0, name = "Residuals") +
   ggrepel::geom_text_repel() +
  theme_minimal() +
  labs(
    title = "Prediction vs. Real Golf Courses by State",
    x = "Actual Number of Golf Courses",
    y = "Predicted Number of Golf Courses"
  ) +
  theme(
    legend.position = "right"
  )
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: ggrepel: 35 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



1. **Underestimated States (Red Points):** These states may have unaccounted factors making them favorable for golf course expansion. Investigate local conditions like emerging markets, tourism, or policy changes.
2. **Overestimated States (Blue Points):** These states may have challenges not captured in the data, such as cultural or economic barriers limiting golf course development.
3. **Well-Fitted States (White Points):** These states align well with the model's predictions, suggesting the current factors explain their golf course counts effectively.

Business Insights

1. 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.

2. Strategic Golf Course Development

Insight: Location matters! (`location_quotient` is significant.) States with a higher location quotient (e.g., Florida and Hawaii) are more favorable for golf courses due to climate and topography. These areas already have a strong presence, suggesting opportunities for luxury golf resorts or high-end services.

Action: Target these regions for premium golf-related tourism or memberships, leveraging their natural advantages to attract both domestic and international clientele.

Here are the states with highest location quotient:

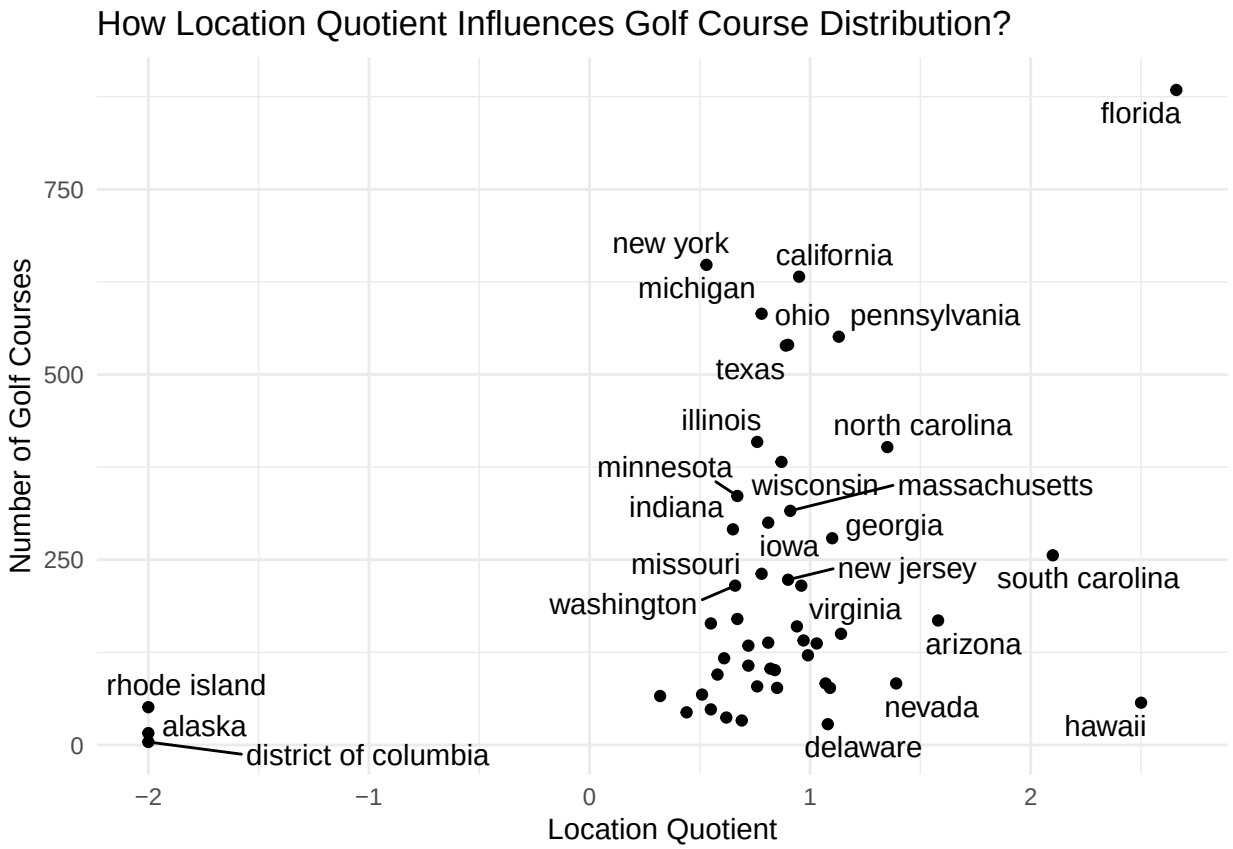
```r
df %>%
   arrange(desc(location_quotient))
```

```
## # A tibble: 51 x 8
##    state      golf_courses location_quotient population growth poulation_density
##    <chr>             <dbl>             <dbl>      <dbl>  <dbl>             <dbl>
##  1 florida             884              2.66   22177997   1.06             414.
##  2 hawaii               57              2.5     1401709  -0.34             218.
##  3 south car~          256              2.1     5342388   1.22             178.
##  4 arizona             168              1.58    7640796   1.6               67.3
##  5 nevada               83              1.39    3238601   1.66              29.5
##  6 north car~          402              1.35   10807491   0.99             222.
##  7 alabama             150              1.14    4949697   0.31              97.7
##  8 pennsylva~          551              1.13   12805190   0.01             286.
##  9 georgia             279              1.1    10936299   0.98             190.
## 10 montana              77              1.09    1093117   0.75               7.51
## # i 41 more rows
## # i 2 more variables: pred_golf_courses <dbl>, residuals <dbl>
```

Can we identify which states are worth investing?

```r
df %>%
   ggplot(aes(y = golf_courses, x = location_quotient, label = state)) +
   geom_point() +
   ggrepel::geom_text_repel() +
   labs(x = "Location Quotient",
        y = "Number of Golf Courses",
        title = "How Location Quotient Influences Golf Course Distribution?")
```

```
## Warning: ggrepel: 24 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



3. Market Potential: Real vs Predicted

Insight: The relationship between predicted and actual values reveals key insights about market potential.

1. **Underestimated states (red points)** are more favorable than the model predicts, suggesting untapped opportunities if the driving factors are sustainable.
2. **Overestimated states (blue points)** are less favorable than expected, indicating potential challenges or barriers to development.
3. **Well-fitted states (white points)** align closely with current predictions, representing stable and predictable markets.

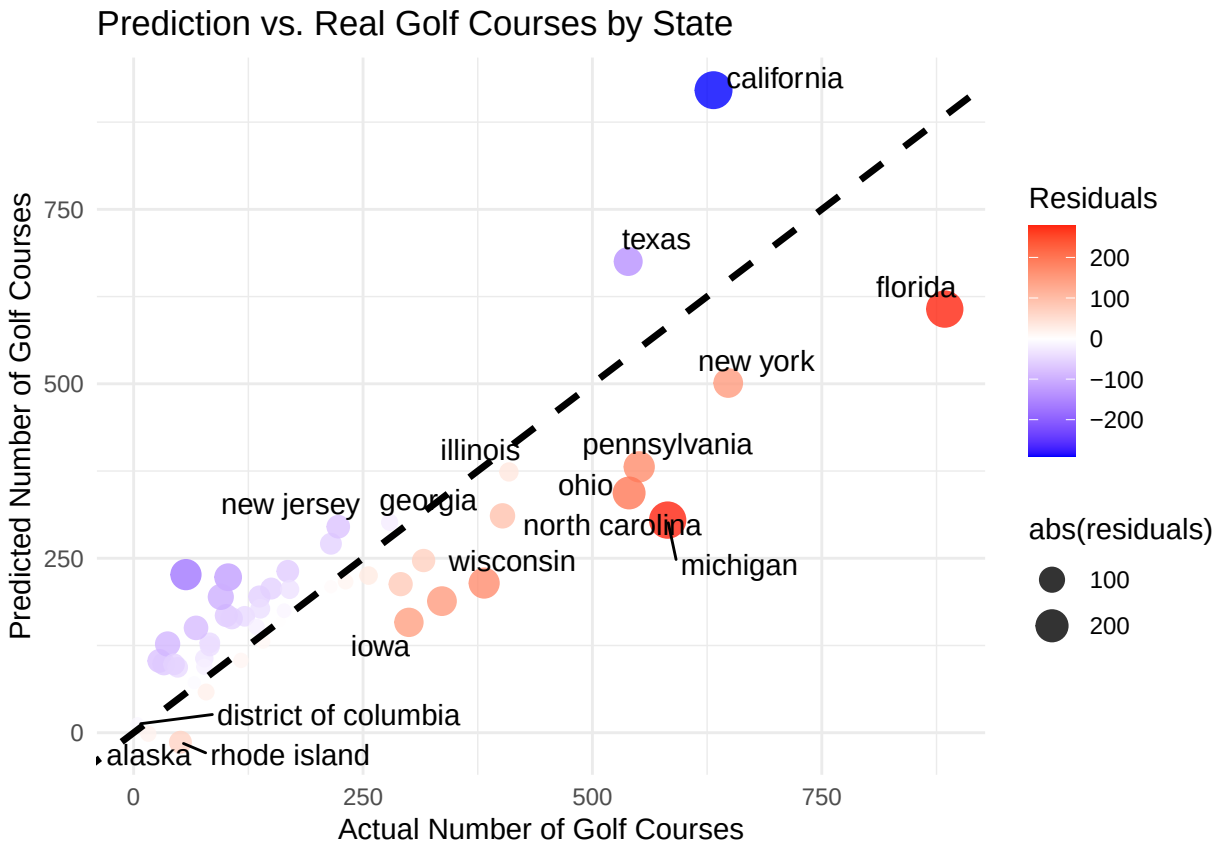
Action: Use residuals to identify actionable targets:

- **Underestimated States (Red Points):** Explore factors contributing to the higher-than-predicted number of golf courses (e.g., tourism, local culture). If these factors are sustainable, these states may represent excellent growth opportunities.
- **Overestimated States (Blue Points):** Investigate barriers to development, such as limited demand or cultural disinterest, and recalibrate strategies accordingly.
- **Well-Fitted States (White Points):** These states meet expectations, suggesting they are stable markets. Focus on maintaining current investments and leveraging existing strengths.

```
ggplot(df, aes(x = golf_courses, y = pred_golf_courses, label = state)) +
  geom_point(aes(color = residuals, size = abs(residuals)), alpha = 0.8) + # Points for actual vs predi
  geom_abline(slope = 1, intercept = 0, color = "black", linetype = "dashed", size = 1.2) + # 1:1 line
  scale_color_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0, name = "Residuals") +
   ggrepel::geom_text_repel() +
  theme_minimal() +
  labs(
```

```
    title = "Prediction vs. Real Golf Courses by State",
    x = "Actual Number of Golf Courses",
    y = "Predicted Number of Golf Courses"
  ) +
  theme(
    legend.position = "right"
  )
```

```
## Warning: ggrepel: 35 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Concluding Remarks

Through the case study of golf courses, we demonstrated how linear regression models relationships between variables and provides actionable insights.

1. **Regression Basics**:
   - Regression models the relationship between predictors (e.g., Location Quotient) and a response variable (e.g., Golf Courses).
   - Metrics like R^2 (71% variability explained) and p-values (e.g., Location Quotient: $p = 0.0412$) assess model fit and predictor significance.
2. **Interpreting Coefficients**:
   - A one-unit increase in **Location Quotient** is associated with ~49 additional golf courses, holding other factors constant.
   - Residuals reveal deviations from the model, highlighting areas for further analysis.

3. **Case Study Insights**:
    - Significant predictors (e.g., Location Quotient, Population) influence golf course distribution.
    - Residuals identify **underestimated states** (growth opportunities) and **overestimated states** (challenges).