Comparison of Least Squares and Least Absolute Deviation Regression

group 6

2025-04-04

Introduction

Regression analysis is a fundamental technique in statistics used to model relationships between variables. This report compares two regression methods:

1. **Least Squares Estimation (LSE)**: Minimizes the sum of squared residuals.
2. **Least Absolute Deviation (LAD)**: Minimizes the sum of absolute residuals (also known as median regression or quantile regression at $\tau = 0.5$).

The dataset used in this analysis contains information on advertising spending (TV) and sales performance.

Data Loading and Preprocessing

```
data_advertise <- read.csv("C:/Users/Aryan Deo/Downloads/advertising_cleaned.csv", header=TRU

biv_data <- data.frame(Sales = data_advertise$Sales, TV = data_advertise$TV)
```

Regression Models

Least Squares Estimation (LSE)

LSE finds the regression line by minimizing the **Sum of Squared Errors (SSE)**:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where y_i is the actual value, and $\hat{y}_i$ is the predicted value.

```r
lse_model <- lm(Sales ~ TV, data = biv_data)
```

Least Absolute Deviation (LAD)

LAD regression (also known as **quantile regression at $\tau = 0.5$**) minimizes the sum of absolute residuals:

$$LAD = \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

This method is **robust to outliers** since it does not square the residuals.

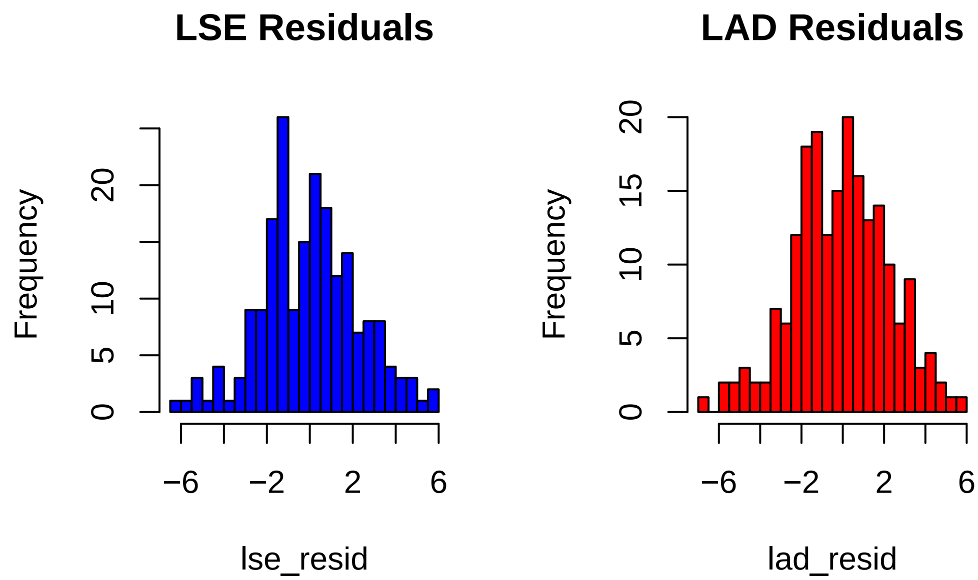
```r
lad_model <- rq(Sales ~ TV, tau = 0.5, data = biv_data)
```

Residual Analysis

Residuals are the differences between actual and predicted values. Examining their distribution helps assess model fit.

```r
# Compute residuals
lse_resid <- resid(lse_model)
lad_resid <- resid(lad_model)

# Plot histograms
par(mfrow = c(1, 2))  # Arrange plots side by side
hist(lse_resid, main = "LSE Residuals", col = "blue", breaks = 20)
hist(lad_resid, main = "LAD Residuals", col = "red", breaks = 20)
```



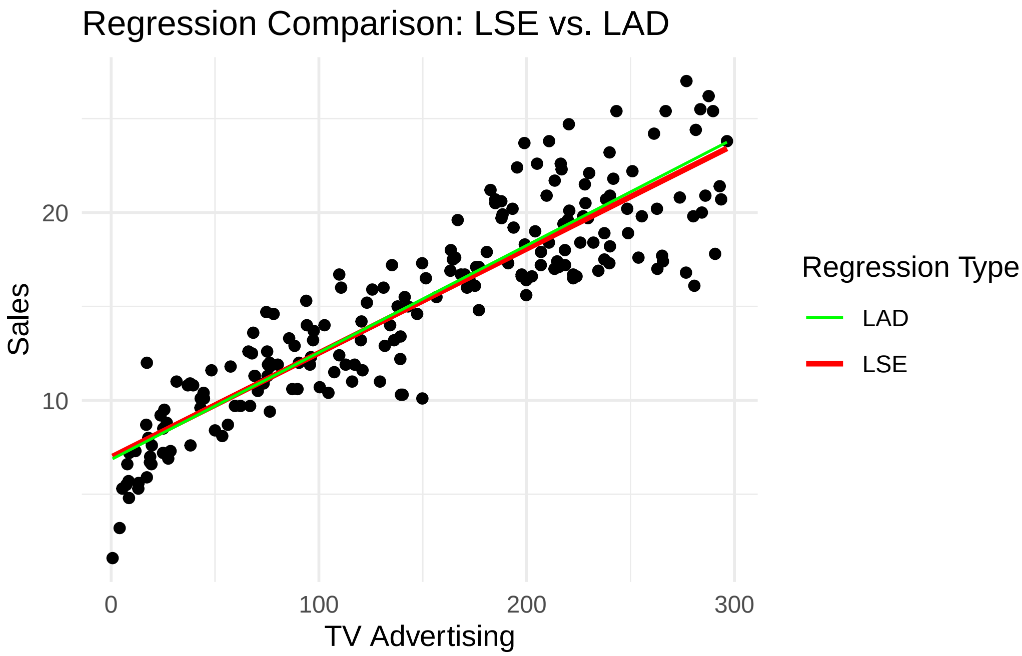
Model Comparison

We visualize both regression models on a scatter plot:

```
# Reset plot layout
par(mfrow = c(1, 1))

ggplot(biv_data, aes(x = TV, y = Sales)) +
  geom_point(color = "black") +
  geom_smooth(method = "lm", aes(color = "LSE"), se = FALSE) +
  stat_quantile(quantiles = 0.5, aes(color = "LAD"), formula = y ~ x) +
  scale_color_manual(values = c("LSE" = "red", "LAD" = "green")) +
  labs(title = "Regression Comparison: LSE vs. LAD",
       x = "TV Advertising",
       y = "Sales",
       color = "Regression Type") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Summary Statistics

```r
summary(lse_model)
```

```

Call:
lm(formula = Sales ~ TV, data = biv_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.4438 -1.4857  0.0218  1.5042  5.6932 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.974821   0.322553   21.62   <2e-16 ***
TV          0.055465   0.001896   29.26   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.296 on 198 degrees of freedom
Multiple R-squared:  0.8122,    Adjusted R-squared:  0.8112 
```

```
F-statistic: 856.2 on 1 and 198 DF,  p-value: < 2.2e-16
```

```r
summary(lad_model)
```

```
Call: rq(formula = Sales ~ TV, tau = 0.5, data = biv_data)

tau: [1] 0.5

Coefficients:
            coefficients lower bd upper bd
(Intercept) 6.84561      6.16219  7.83892
TV          0.05702      0.04901  0.05986
```

Error Metrics

To compare model performance, we calculate: - **Mean Absolute Error (MAE)**: Measures average absolute differences between predicted and actual values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **Root Mean Squared Error (RMSE)**: Penalizes larger errors more heavily.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

```r
# Compute MAE
lse_mae <- mean(abs(lse_resid))
lad_mae <- mean(abs(lad_resid))

# Compute RMSE
lse_rmse <- sqrt(mean(lse_resid^2))
lad_rmse <- sqrt(mean(lad_resid^2))

cat("LSE MAE:", lse_mae, "\nLAD MAE:", lad_mae, "\n")
```

```
LSE MAE: 1.830587
LAD MAE: 1.829467
```

```r
cat("LSE RMSE:", lse_rmse, "\nLAD RMSE:", lad_rmse, "\n")
```

```
LSE RMSE: 2.284238
LAD RMSE: 2.290251
```

Conclusion

- **LSE regression** minimizes squared errors and is sensitive to outliers.
- **LAD regression** minimizes absolute errors and is robust to outliers.
- MAE and RMSE provide insight into model performance, with lower values indicating better fit.

This analysis highlights the importance of choosing an appropriate regression method based on data characteristics.