Week 9 - Evaluation of Regression Models

Introduction:

Welcome to this unit on methods of evaluating regression models. Regression models are essential in mathematical modeling and data analysis, and their performance needs to be assessed rigorously. Without proper evaluation, researchers may misinterpret the relationships between variables. They might erroneously conclude that certain predictors have a significant impact on the outcome when they do not or vice versa. Inadequately evaluated regression models may produce inaccurate predictions. If the model does not capture the underlying relationships between variables effectively, its predictive performance may be poor, leading to unreliable forecasts or estimates. In this unit, we will explore various evaluation methods and metrics used to assess the quality and accuracy of regression models.

Importance of Model Evaluation:

Model evaluation is critical to determine how well a regression model fits the data and makes predictions. It helps in selecting the best model among competing models and guides model refinement. Accurate model evaluation is essential for making informed decisions in various fields, including economics, engineering, and the sciences.

Common Evaluation Metrics:

Several metrics are commonly used to evaluate regression models:

1. Mean Squared Error (MSE):

MSE measures the average squared difference between predicted and actual values. It penalizes larger errors more than smaller ones, making it sensitive to outliers. The formula for MSE is:

$$MSE = \frac{\sum(y_i - \bar{y})^2}{n}$$

where y_i is the actual value, $\bar{y}$ is the mean of actual values, and n is the number of data points.

2. Root Mean Squared Error (RMSE):

RMSE is the square root of MSE and is in the same units as the dependent variable. It provides a more interpretable measure of error.

$$RMSE = \sqrt{MSE}$$

3. Mean Absolute Error (MAE):

MAE measures the average absolute difference between predicted and actual values. It is less sensitive to outliers compared to MSE.

$$MAE = \frac{\sum |y_i - \bar{y}|}{n}$$

where ŷ is the predicted value.

4. R-squared (R^2):

R^2 represents the proportion of variance in the dependent variable explained by the model. It ranges from 0 to 1, where a higher value indicates a better fit. R^2 = 1 - (SSE / SST),

where SSE is the sum of squared errors and SST is the total sum of squares.

5. Adjusted R-squared (Adjusted R^2):

Adjusted R^2 adjusts R^2 for the number of predictors in the model. It penalizes the addition of unnecessary variables.

Adjusted R^2 = 1 - [(1 - R^2) * ((n - 1) / (n - p - 1))],

where n is the number of data points and p is the number of predictors.

Model Validation Techniques:

Summary Statistics:

The summary() function provides a summary of the regression model, including coefficients, standard errors, t-values, and p-values for each predictor. Example:

```
summary(lm_model)
```

Residual Analysis:

Residuals are the differences between the observed values and the values predicted by the model. Plotting residuals against predicted values or predictors can help identify patterns or deviations from assumptions.

Example:

```
plot(lm_model)
```

Diagnostic Plots:

The plot() function can generate diagnostic plots, including scatterplots of residuals against fitted values, histograms of residuals, QQ plots, and more.

Example:

```
plot(lm_model)
```

Coefficient of Determination (R-squared):

R-squared measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model. Higher values indicate better fit.

Example:

```
summary(lm_model)$r.squared
```

Root Mean Squared Error (RMSE) or Mean Absolute Error (MAE):

These metrics quantify the average difference between observed and predicted values. Lower values indicate better predictive accuracy.

Example

```
RMSE: sqrt(mean(residuals(lm_model)^2))
```

Adjusted R-squared:

Adjusted R-squared penalizes the inclusion of unnecessary predictors in the model and provides a more conservative measure of model fit, especially for models with multiple predictors.

Example:

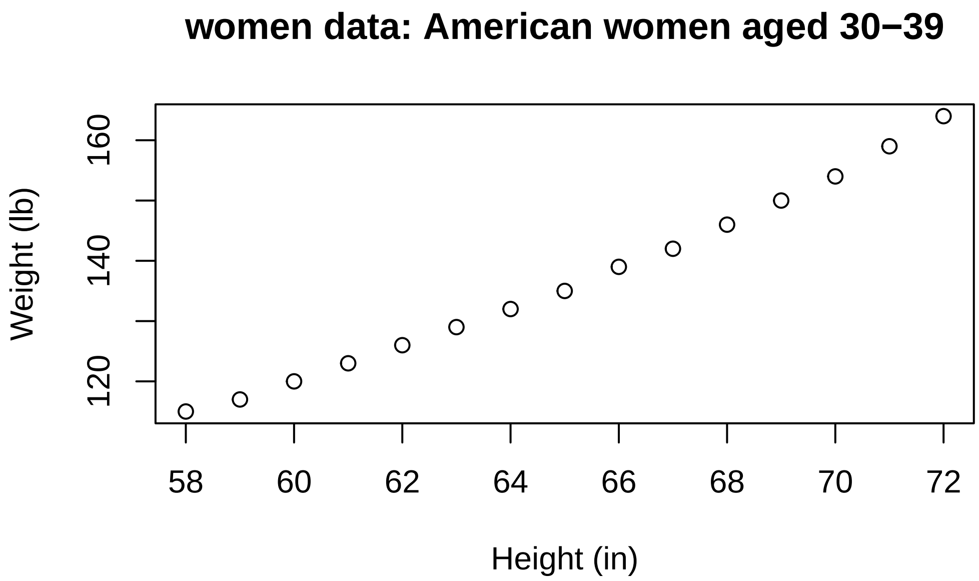
```
summary(lm_model)$adj.r.squared
```

Example

In this example, we will do an exploration of

```
 # You may need to install the datasets library (by using install.packages('datasets'))
library(datasets)
```

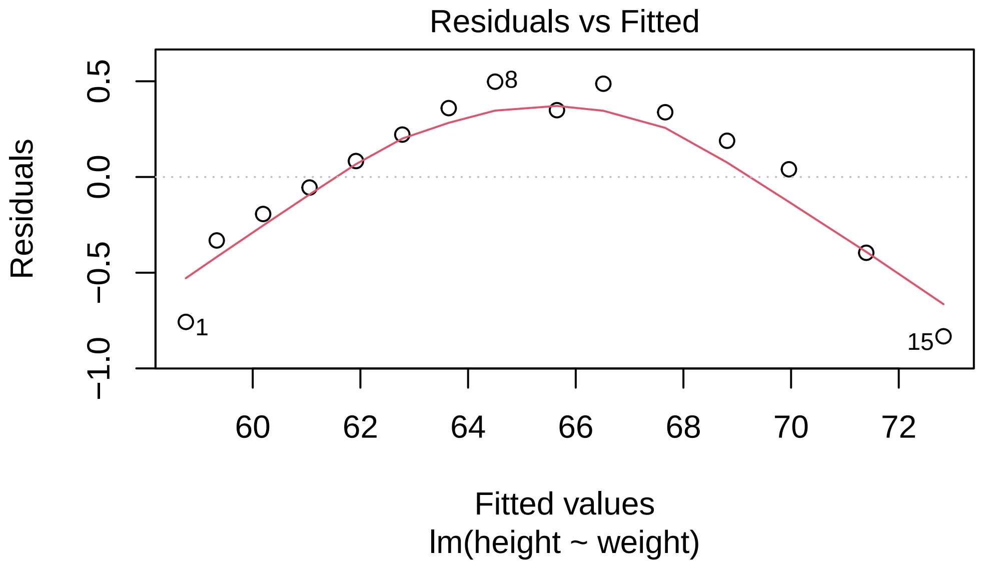
```
plot(women, xlab = "Height (in)", ylab = "Weight (lb)",
     main = "women data: American women aged 30-39")
```

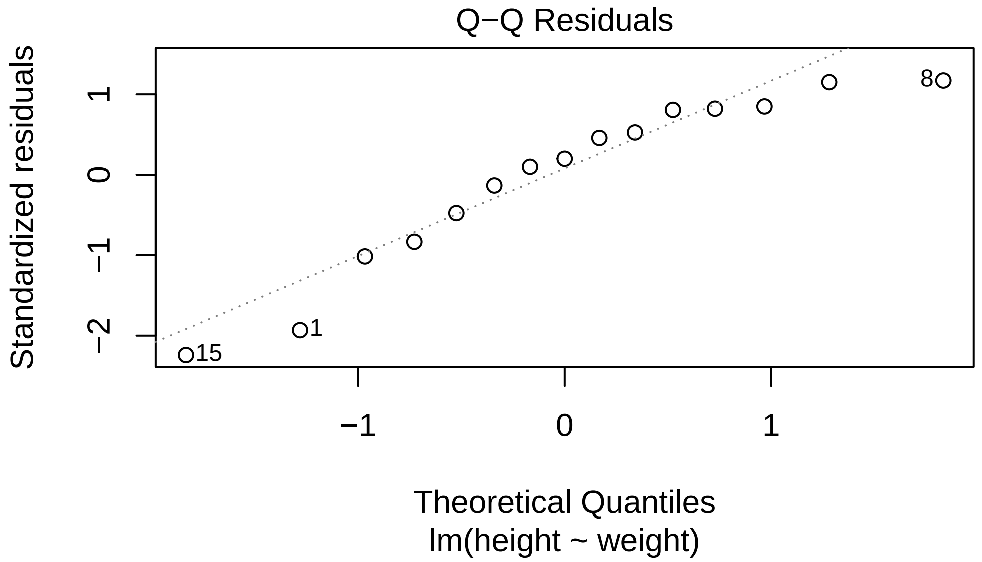


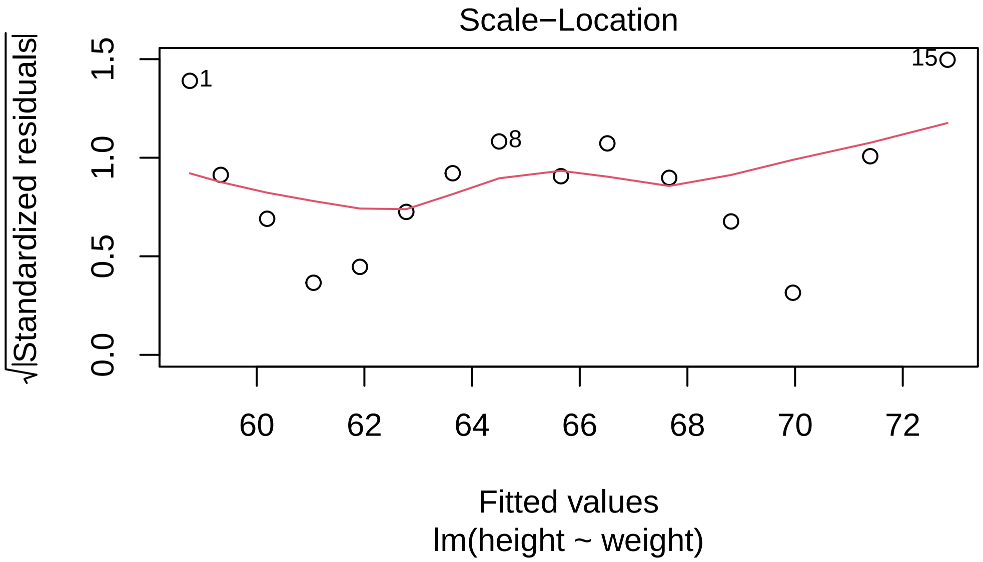
```r
lmHeight = lm(height~weight, data = women) #Create the linear regression
summary(lmHeight) #Review the results
```

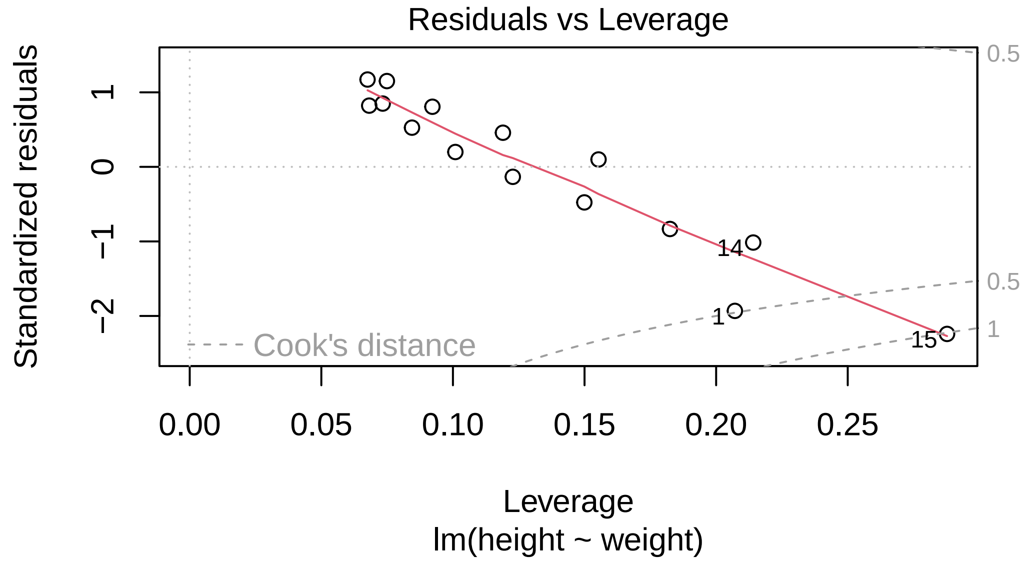
```
Call:
lm(formula = height ~ weight, data = women)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.83233 -0.26249  0.08314  0.34353  0.49790 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 25.723456   1.043746   24.64 2.68e-12 ***
weight       0.287249   0.007588   37.85 1.09e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.44 on 13 degrees of freedom
Multiple R-squared:  0.991, Adjusted R-squared:  0.9903 
F-statistic:  1433 on 1 and 13 DF,  p-value: 1.091e-14
```

```r
plot(lmHeight)
```


Residuals vs Fitted
Residuals
0.5
0.0
−0.5
−1.0
60
62
64
66
68
70
72
Fitted values
lm(height ~ weight)
8
1
15


Q−Q Residuals
Standardized residuals
1
0
−1
−2
−1
0
1
Theoretical Quantiles
lm(height ~ weight)
8
1
15





```
summary(lmHeight)$r.squared
```

```
[1] 0.9910098
```

```r
summary(lmHeight)$adj.r.squared
```

```
[1] 0.9903183
```

Conclusion:

Evaluating regression models is essential to determine their accuracy, reliability, and suitability for a given problem. Metrics like MSE, RMSE, MAE, R^2, and adjusted R^2 provide valuable insights into a model's performance. Model validation techniques and residual analysis help ensure robust and interpretable regression models.