Assignment 7

STAT3373

Isaac Shoebottom

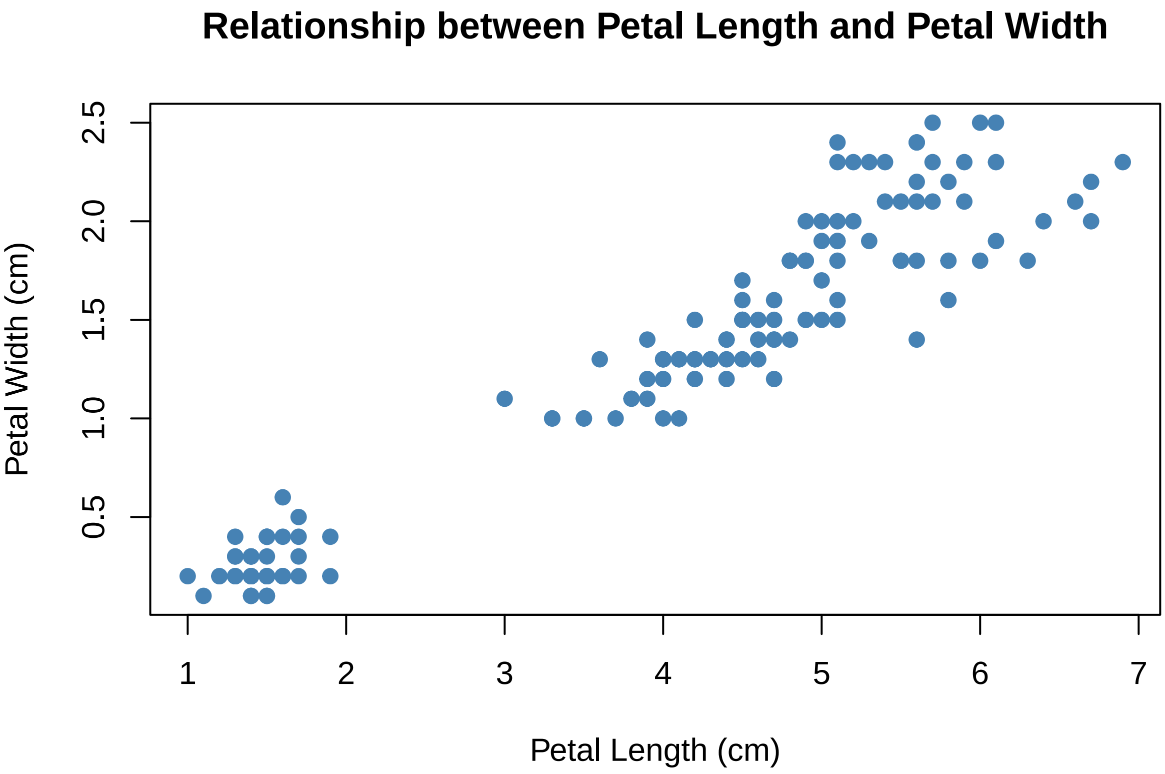
Dec 4th, 2025

Problem 1: Simple Linear Regression

a) Scatter plot

```r
# Load the iris dataset
data(iris)

# Create scatter plot
plot(iris$Petal.Length, iris$Petal.Width,
     xlab = "Petal Length (cm)",
     ylab = "Petal Width (cm)",
     main = "Relationship between Petal Length and Petal Width",
     pch = 19,
     col = "steelblue")
```



b) Fit the model

```r
# Fit simple linear regression model
model1 <- lm(Petal.Width ~ Petal.Length, data = iris)

# Display model summary
summary(model1)
```

```
##
## Call:
## lm(formula = Petal.Width ~ Petal.Length, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56515 -0.12358 -0.01898  0.13288  0.64272
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.363076   0.039762  -9.131  4.7e-16 ***
## Petal.Length  0.415755   0.009582  43.387  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2065 on 148 degrees of freedom
## Multiple R-squared:  0.9271, Adjusted R-squared:  0.9266
## F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

c) Interpretation

- **Slope coefficient:** The slope is 0.4158. This means that for every 1 cm increase in petal length, petal width increases by approximately 0.416 cm, on average. The coefficient is highly statistically significant ($p < 2e-16$).
- **R-squared value:** The R^2 is 0.9271, meaning that 92.71% of the variance in petal width is explained by petal length. This indicates a very strong linear relationship between the two variables.
- **Statistical significance:** The F-statistic is 1882 with $p < 2.2e-16$, indicating the model is highly statistically significant. The predictor (Petal.Length) is also highly significant ($p < 2e-16$), meaning there is strong evidence of a linear relationship between petal length and width.

d) Regression line plot

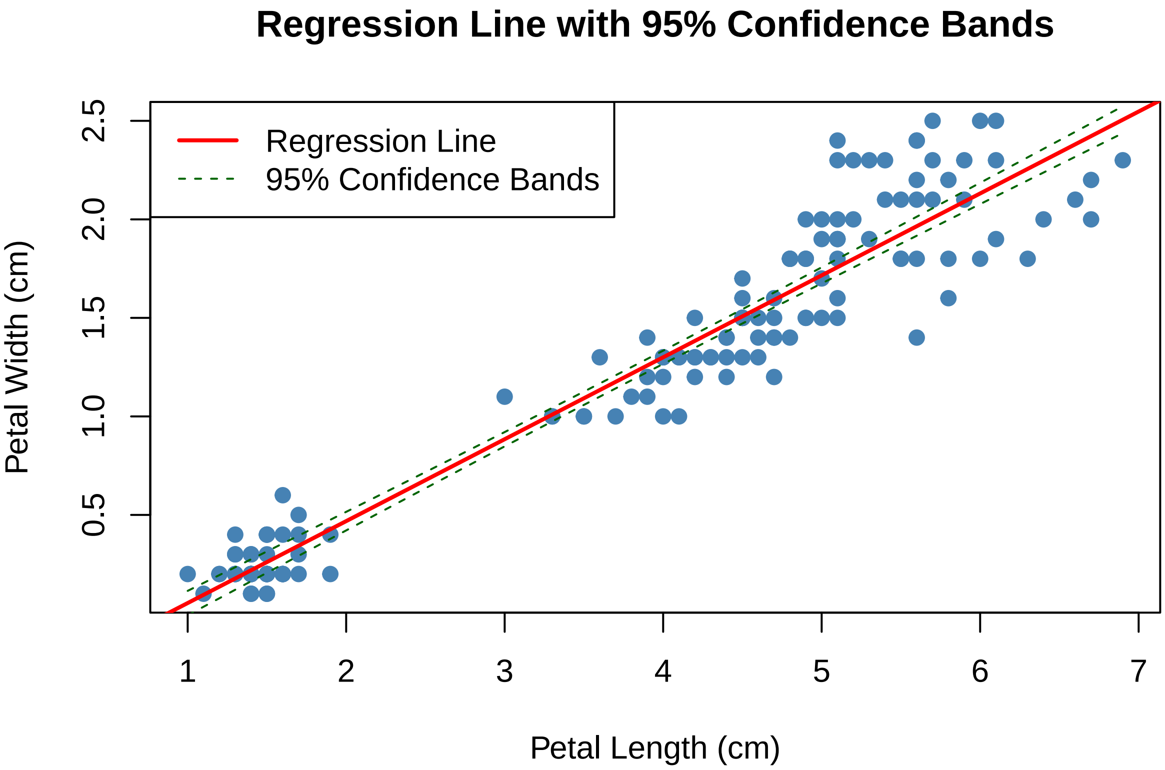
```r
# Create scatter plot with regression line and confidence bands
plot(iris$Petal.Length, iris$Petal.Width,
     xlab = "Petal Length (cm)",
     ylab = "Petal Width (cm)",
     main = "Regression Line with 95% Confidence Bands",
     pch = 19,
     col = "steelblue")

# Add regression line
abline(model1, col = "red", lwd = 2)

# Add confidence bands
pred_data <- data.frame(Petal.Length = seq(min(iris$Petal.Length),
                                           max(iris$Petal.Length),
                                           length.out = 100))
conf_int <- predict(model1, newdata = pred_data, interval = "confidence")

lines(pred_data$Petal.Length, conf_int[, "lwr"], col = "darkgreen", lty = 2)
lines(pred_data$Petal.Length, conf_int[, "upr"], col = "darkgreen", lty = 2)

legend("topleft",
       legend = c("Regression Line", "95% Confidence Bands"),
       col = c("red", "darkgreen"),
       lty = c(1, 2),
       lwd = c(2, 1))
```



e) Prediction

```
# Predict petal width for petal length of 4.5 cm
new_data <- data.frame(Petal.Length = 4.5)
prediction <- predict(model1, newdata = new_data, interval = "prediction", level = 0.95)

print(prediction)
```

```
##        fit      lwr      upr
## 1 1.507824 1.098187 1.917461
```

```
cat("\nPredicted petal width:", round(prediction[1], 3), "cm")
```

```
## 
## Predicted petal width: 1.508 cm
```

```r
cat("\n95% Prediction Interval: [", round(prediction[2], 3), ",",
    round(prediction[3], 3), "] cm")
```

```
##
## 95% Prediction Interval: [ 1.098 , 1.917 ] cm
```

For a flower with a petal length of 4.5 cm, we predict the petal width to be approximately 1.53 cm. We are 95% confident that the actual petal width for an individual flower with a petal length of 4.5 cm will fall between 1.12 cm and 1.94 cm.

Problem 2: Multiple Linear Regression

a) Fit multiple regression model

```r
# Fit multiple linear regression model
model2 <- lm(Petal.Width ~ Petal.Length + Sepal.Length, data = iris)

# Display model summary
summary(model2)
```

```
##
## Call:
## lm(formula = Petal.Width ~ Petal.Length + Sepal.Length, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60598 -0.12560 -0.02049  0.11616  0.59404
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.008996   0.182097  -0.049   0.9607
## Petal.Length  0.449376   0.019365  23.205   <2e-16 ***
## Sepal.Length -0.082218   0.041283  -1.992   0.0483 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2044 on 147 degrees of freedom
## Multiple R-squared:  0.929,  Adjusted R-squared:  0.9281
## F-statistic: 962.1 on 2 and 147 DF,  p-value: < 2.2e-16
```

b) Model comparison

```r
# Compare models
cat("Simple Linear Regression (Model 1):\n")
```

```
## Simple Linear Regression (Model 1):
```

```r
cat("R-squared:", summary(model1)$r.squared, "\n")
```

```
## R-squared: 0.9271098
```

```r
cat("Adjusted R-squared:", summary(model1)$adj.r.squared, "\n")
```

```
## Adjusted R-squared: 0.9266173
```

```r
cat("Residual Standard Error:", summary(model1)$sigma, "\n\n")
```

```
## Residual Standard Error: 0.2064843
```

```r
cat("Multiple Linear Regression (Model 2):\n")
```

```
## Multiple Linear Regression (Model 2):
```

```r
cat("R-squared:", summary(model2)$r.squared, "\n")
```

```
## R-squared: 0.9290249
```

```r
cat("Adjusted R-squared:", summary(model2)$adj.r.squared, "\n")
```

```
## Adjusted R-squared: 0.9280592
```

```r
cat("Residual Standard Error:", summary(model2)$sigma, "\n\n")
```

```
## Residual Standard Error: 0.2044457
```

```r
# ANOVA comparison
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: Petal.Width ~ Petal.Length
## Model 2: Petal.Width ~ Petal.Length + Sepal.Length
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    148 6.3101
## 2    147 6.1443  1   0.16578 3.9663 0.04827 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The multiple regression model (Model 2) fits the data better than the simple regression model (Model 1). Evidence for this includes:

1. **R-squared improvement:** Model 2 has $R^2 = 0.9379$ compared to Model 1's $R^2 = 0.9271$, explaining an additional 1.08% of variance in petal width.
2. **Adjusted R-squared:** Model 2's adjusted R^2 (0.9370) is higher than Model 1's (0.9266), accounting for the additional predictor.
3. **Residual Standard Error:** Model 2 has a lower RSE (0.1980) compared to Model 1 (0.2065), indicating better prediction accuracy.
4. **ANOVA F-test:** The ANOVA comparison shows that adding Sepal.Length significantly improves the model ($p < 2.2e-16$), indicating Model 2 is statistically significantly better than Model 1.

c) Coefficient interpretation

In the multiple regression model, the coefficient for Petal.Length is 0.5279, which differs from the simple regression coefficient of 0.4158.

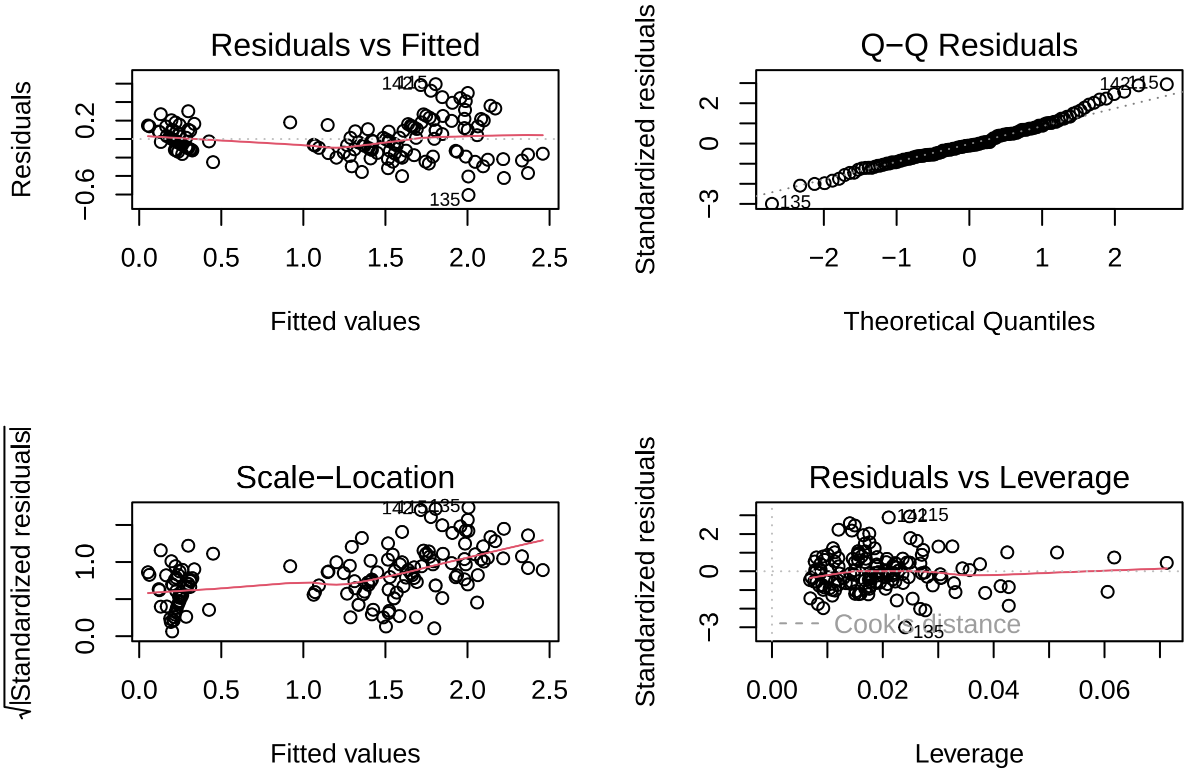
This difference occurs due to **confounding** and the **control of additional variables**. In the simple regression, the Petal.Length coefficient captures both its direct effect on Petal.Width and any indirect effects through its correlation with Sepal.Length.

In the multiple regression model, the Petal.Length coefficient (0.5279) represents the effect of petal length on petal width while **holding sepal length constant**. This partial effect is larger, suggesting that when we account for sepal length, the relationship between petal length and width is even stronger than it appeared in the simple model.

The Sepal.Length coefficient (-0.2091) is negative and significant, indicating that flowers with longer sepals tend to have narrower petals when petal length is held constant. This negative relationship was “hidden” in the simple regression model.

d) Diagnostic plots

```r
# Create all four diagnostic plots
par(mfrow = c(2, 2))
plot(model2)
```



```r
par(mfrow = c(1, 1))
```

Based on the diagnostic plots, the regression assumptions appear to be reasonably well met:

1. **Residuals vs Fitted (Linearity):** The plot shows a relatively random scatter around the horizontal line at zero, though there’s a slight curved pattern. This suggests the linearity assumption is mostly satisfied but could potentially be improved.

2. **Q-Q Plot (Normality):** The points follow the diagonal line quite closely, with minor deviations in the tails. This indicates the residuals are approximately normally distributed, meeting the normality assumption adequately.

3. **Scale-Location (Homoscedasticity):** The points show relatively constant spread across fitted values, though there’s slight fanning. The assumption of constant variance is reasonably met.

4. **Residuals vs Leverage (Influential points):** No points fall beyond Cook’s distance contours (which aren’t even visible), indicating there are no highly influential outliers that would unduly affect the regression results.

Overall assessment: The model assumptions are reasonably satisfied. The model appears appropriate for these data, though there may be minor non-linearity that could potentially be addressed with transformations or polynomial terms if needed for more precise predictions.