SHC 798 Assignment 1, 2025

Richard Lubega

2025-07-14

SHC 798 Assignment 1, 2025

Part 3: Simple regression

Question 1

```
# The dataset cars
# A SLR to analyse the relationship between speed and stopping distance

cat("\n A SLR between speed and stopping distance \n")
```

```
##
##  A SLR between speed and stopping distance
```

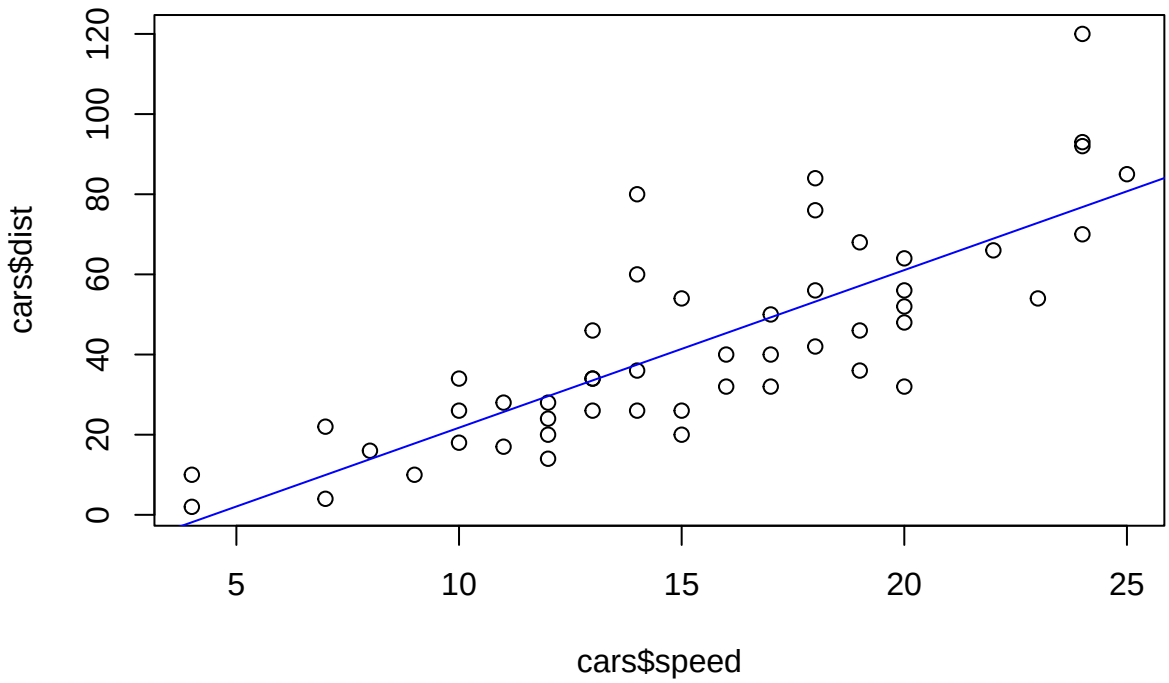
```
lm_s.sd <- lm(dist ~ speed, data = cars)
summary(lm_s.sd)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

```r
cat("\n ===  SLR Model Plot  === \n")
```

```
##
##  ===  SLR Model Plot  ===
```

```r
plot(cars$speed, cars$dist)
abline(lm_s.sd, col = "blue")
```



(a) From the model summary, **Multiple R-squared**: 0.6511, Adjusted R-squared: 0.6438

Thus, **65.11%** of the variation in stopping distance is explained by speed

(b) **Intercept** (-17.5791): This means that for a theoretical speed of 0 mph the predicted stopping distance is -17.5791 feet. This is not practically rational but ensures the regression line fits the data best within the observed speed range. It is not meaningful to extrapolate to speed = 0.

- It's p-value (0.0123) is small and statistically significant at the 5% level, but its practical importance is limited.

Slope (3.9324): For every 1 mph increase in speed, stopping distance increases by about 3.9324 feet. Higher driving speeds require longer stopping distances.

- The p-value (1.49e-12) is much smaller than 0.05 (even at a 1% significance level), so the relationship between speed and stopping distance is statistically significant.We reject the null hypothesis that speed has no effect on stopping distance. Thus, speed has an considerable impact on stopping distance.

```r
# Predicting stopping distance for  speed = 20 mph;  compute a 95% prediction interval.
cat("\n ===  Stopping distance at a speed of 20 mp and the 95% prediction interval ===\n ")
```

(c)

```
##
##  ===  Stopping distance at a speed of 20 mp and the 95% prediction interval ===
##
```

```r
predict(lm_s.sd, newdata = data.frame(speed = 20), interval = "prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 61.06908 29.60309 92.53507
```

```r
cat("\n Evaluating Model Assumptions \n")
```

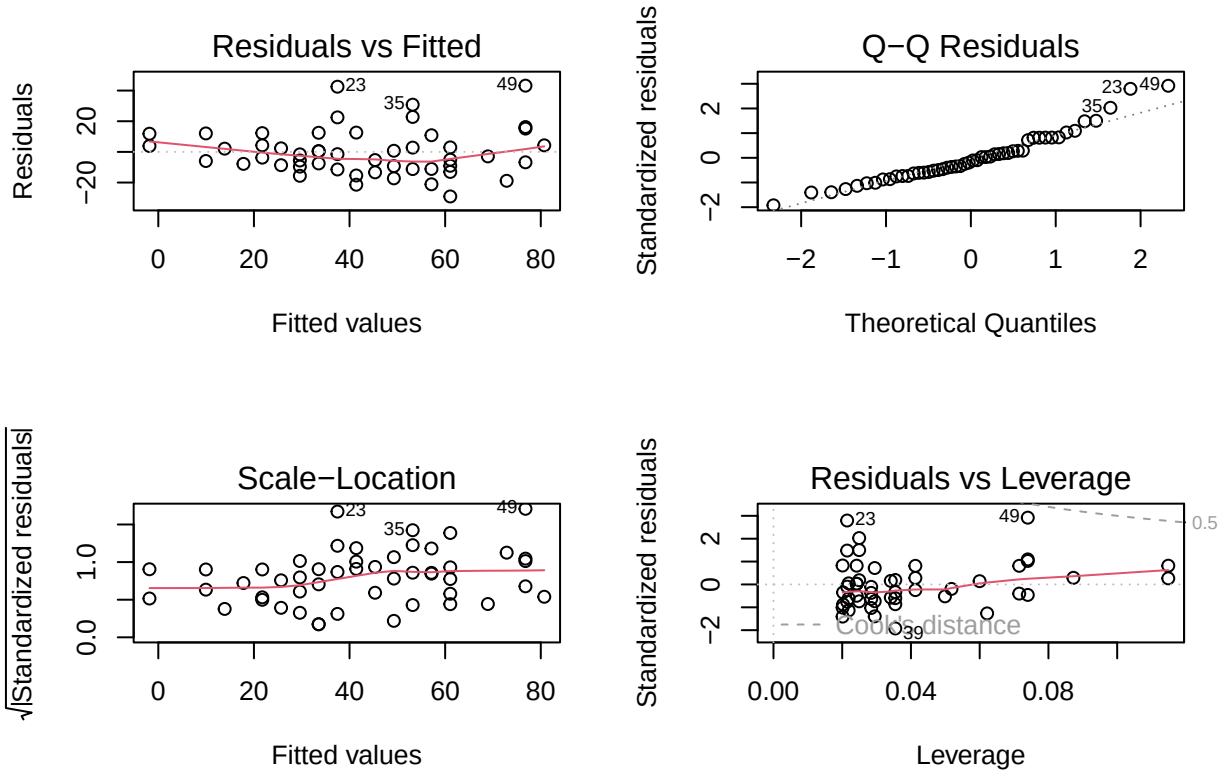
(d)

```
##
##  Evaluating Model Assumptions
```

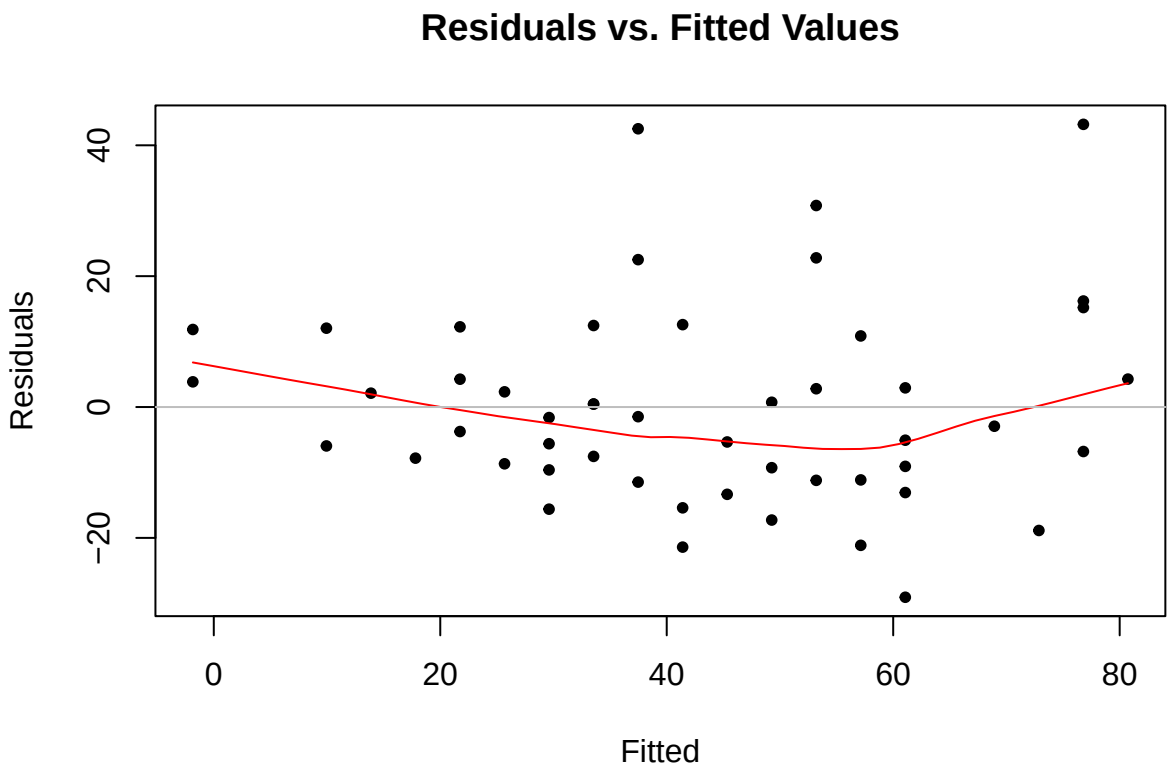
```r
cat("\n ===  Model Diagnostics Plots  === \n")
```

```
##
##  ===  Model Diagnostics Plots  ===
```

```r
# Diagnostics plots
par(mfrow = c(2,2))
plot(lm_s.sd)
```

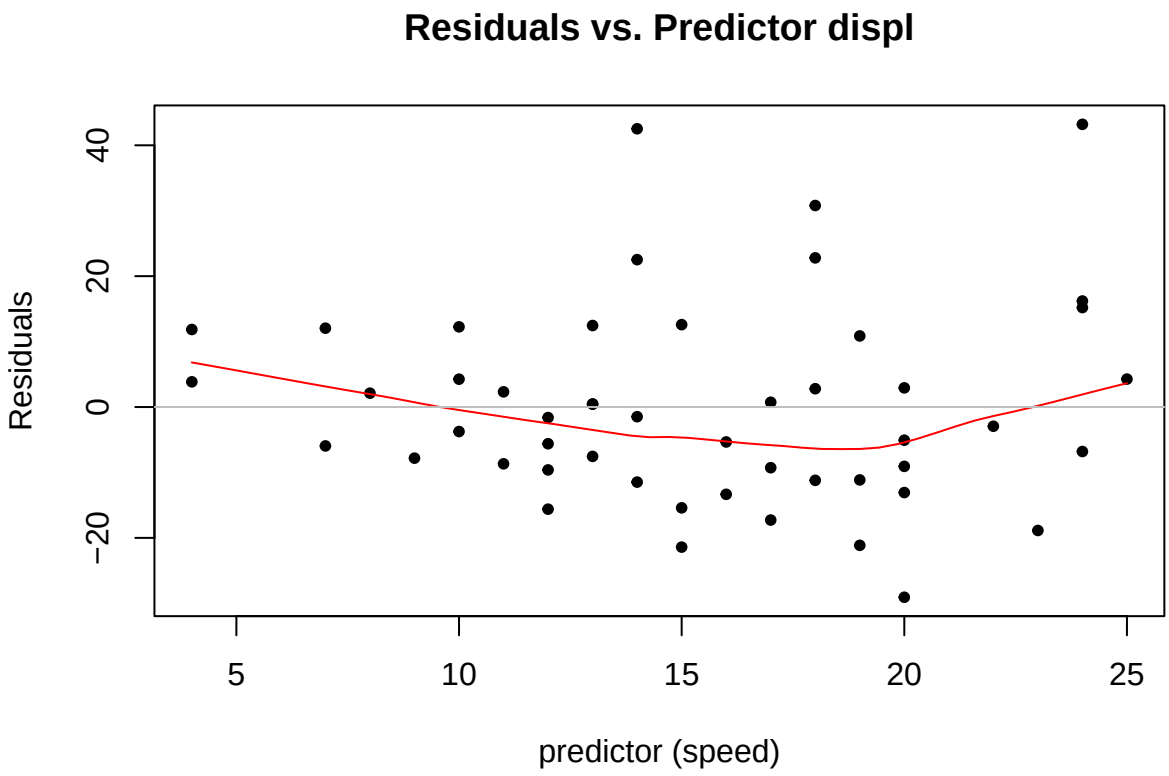



```
par(mfrow = c(1,1))
# Tukey-Anscombe Plot
plot(lm_s.sd$fitted.values, lm_s.sd$residuals, xlab="Fitted", ylab="Residuals", pch=20) +
  title("Residuals vs. Fitted Values") +
  lines(loess.smooth(lm_s.sd$fitted.values, lm_s.sd$residuals),col="red") +
  abline(h=0, col="grey")
```



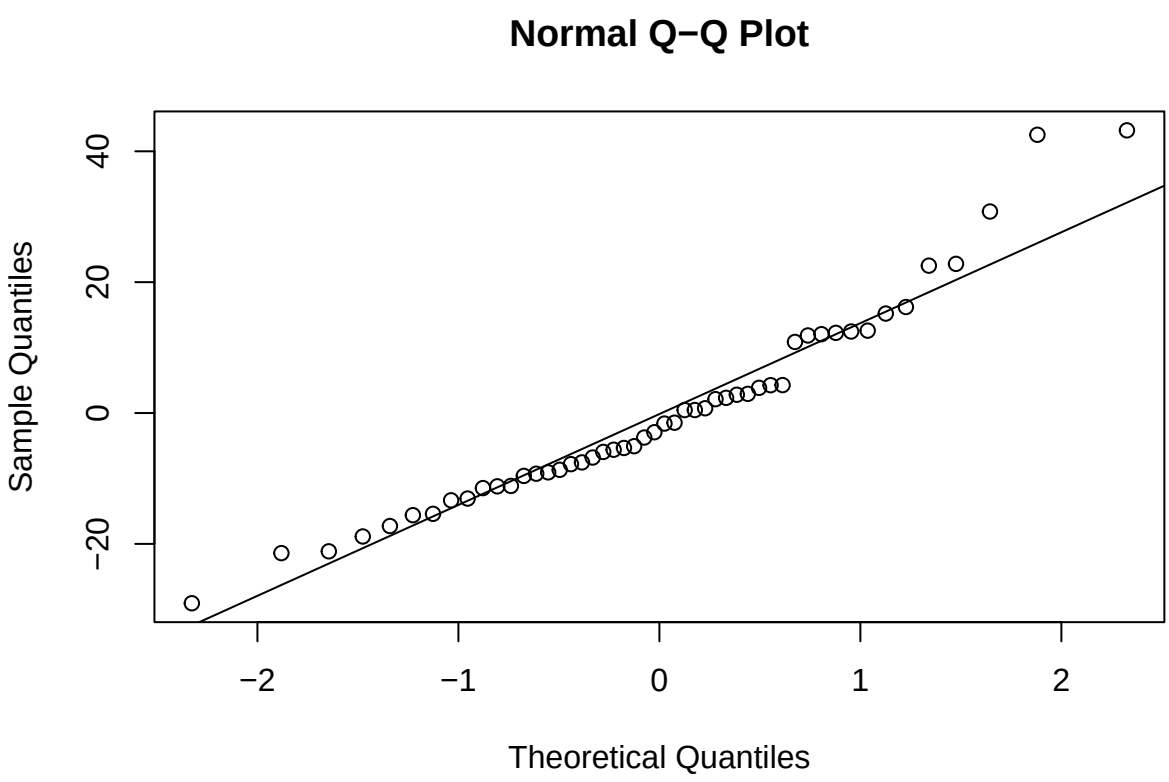

```
## integer(0)
```

```
# Residuals vs. Predictor Plot
plot(cars$speed, lm_s.sd$residuals, xlab="predictor (speed)", ylab="Residuals", pch=20) +
  title("Residuals vs. Predictor displ") +
  lines(loess.smooth(cars$speed, lm_s.sd$residuals),col="red") +
  abline(h=0, col="grey")
```




```
## integer(0)
```

```
# Quantile-Quantile Plot
qqnorm(lm_s.sd$residuals) #Quantile-Quantile Plot
qqline(lm_s.sd$residuals) # adds the diagonal line
```



Model Assumption Evaluation

1. **Linearity** — *From the Tukey-Anscombe Plot (Residuals vs. Fitted)*:

- By inspection, the residuals generally hover around the zero line which suggests that they likely approximate a mean of zero. There is, however, slight curvature (a kink) in the red LOESS smoother line (deviation from the horizontal) which implies mild (misspecified) non-linearity. This is confirmed by the systematic misprediction in the middle (overpredicting) and the extremes (underpredicting). In this case, there is is a clear violation of the linearity ($E[E_i] = 0$) assumption; a straight line is not the correct fit to the data and the model ought to be improved.
- **Transformation**: Add a quadratic term (dist $=\beta_0 + \beta_1$. speed $+\beta_2$. speed^2 + E_i) to fix this and improve the model (as for this pair, the true relationship is quadratic). This constitutes a multiple linear regression problem.

2. **Homoskedasticity** — *From the Scale-Location Plot*:

- The red line is slightly upward-trending, indicating that variance increases with fitted values (minor heteroscedasticity)
- The Tukey-Anscombe plot also seems to indicate that the scatter is not constant for the entire range of speed/fitted values (less scatter for lower values and more scatter for higher values). There is an obvious violation of homoskedasticity.
- **Transformation**: Log-transform on dist (since stopping distance cannot be negative)

3. **Independence**

- Since the data is *not time-dependent*, residual independence is likely satisfied (no autocorrelation expected)
- **Transformation**: None needed

4. **Normality** — *From the Q-Q Plot*:

- The bulk of the residuals (in the central region) are approximately Gaussian distributed. A noticeable deviations (or outliers) at the upper tail indicates right skewness hence departure from normality. The assumption of Gaussian errors is slightly violated by the model due to this moderate non-normality.
- **Transformation**: Log-transform on dist to correct right-skewness (improve normality and heteroskedasticity)

Model Evaluation and Improvements

- Therefore, this model (lm_s.sd = dist ~ speed) has minor assumption violations (non-linearity, heteroscedasticity, non-normality).
- Suggested transformations like the log(dist) ~ speed and a quadratic term could be made and the diagnostics re-checked. The best model is the one with the most stable residuals, best-fulfilled assumptions, and highest adjusted R^2.