Learning Outcomes

1. Use R Studio to inspect data prior to fitting a simple linear model
2. Use R Studio to Fit a simple linear model and display the results of the linear model
3. Use R Studio to inspect the residuals of a simple linear model
4. **Inspecting the data**

The cars dataset comes pre-loaded in R. You can inspect the start of the dataset using the following code:

head(cars)

We are interested in whether we can build a simple regression model to predict Distance (dist) from Speed (speed) by establishing a statistically significant linear relationship.

First, let's check the correlation between the variables

cor(cars$speed, cars$dist)

You can visualise a scatterplot of the dataset using the following code.

plot(cars$speed, cars$dist)

A smoothed line of the points might help to visualise a linear relationship.

scatter.smooth(x=cars$speed, y=cars$dist, main="Dist ~ Speed") # scatterplot

A box plot can help us check for outliers

par(mfrow=c(1, 2)) # divide graph area in 2 columns

boxplot(cars$speed, main="Speed") # box plot for 'speed'

boxplot.stats(cars$speed)$out # display outliers

boxplot(cars$dist, main="Distance") # box plot for 'distance'

boxplot.stats(cars$dist)$out # display outliers

We can also look at the density of the variables.

plot(density(cars$speed), main="Density Plot: Speed") # density plot for 'speed'

plot(density(cars$dist), main="Density Plot: Distance")

par(mfrow=c(1, 1)) # back to 1 plot

1. **Fitting the Model**

We can fit our linear model as follows:

cars.lm <- lm(dist ~ speed, data=cars) # build linear regression model on full data

print(cars.lm)

And now we can inspect the results

summary(cars.lm)

Next, we can visualise our regression line on a scatterplot of our data.

plot(cars$speed, cars$dist)

abline(cars.lm)

1. **Inspecting Residuals**

Next, we can compute the residuals

cars.res <- resid(cars.lm)

We can plot the residuals against the observed values

plot(cars$dist, cars.res, ylab="Residuals", xlab="Distance",

main="Cars Linear Model")

abline(0, 0) # the horizon

We can inspect the residuals density

plot(density(cars.res), main="Density Plot: residuals")

We can also see a number of other plots with a single command

plot(cars.lm) # each plot individually

par(mfrow=c(2,2)) # 2x2 grid of plots

plot(cars.lm)

par(mfrow=c(1, 1)) # back to 1 plot

If we wanted to, we could create the standardised residuals and create a normal probability plot (Q-Q plot) manually.

cars.stdres = rstandard(cars.lm)

qqnorm(cars.stdres,

ylab="Standardized Residuals",

xlab="Normal Scores",

main="Cars dataset")

qqline(cars.stdres)

***What do you think about this model? Is it appropriate for our data?***