Stat 415 Regression: Classwork/Lab 1

Frankie Tyndall

2022-06-28

## Study of Airfreight breakage

A substance used in biological and medical research is shipped by airfreight to users in cartons of 1,000 ampules. The data below, involving 10 shipments, were collected on the number of times the carton was transferred from one aircraft to another over the shipment route (X) and the number of ampules found to be broken upon arrival (Y).Assume that first-order regression model is appropriate.

x = The number of times the carton was transferred from one aircraft to another over the shipment route

y = The number of ampules found to be broken upon arrival

airfreight\_shipments <- tribble(~Xi, ~Yi,  
 1 , 16,  
 0 , 9,  
 2 , 17,  
 0 , 12,  
 3 , 22,  
 1 , 13,  
 0 , 8,  
 1 , 15,  
 2 , 19,  
 0 , 11  
 )   
  
airfreight\_shipments

## # A tibble: 10 × 2  
## Xi Yi  
## <dbl> <dbl>  
## 1 1 16  
## 2 0 9  
## 3 2 17  
## 4 0 12  
## 5 3 22  
## 6 1 13  
## 7 0 8  
## 8 1 15  
## 9 2 19  
## 10 0 11

## Hypothesis for Simple Linear Regression for Airfreight breakage Study

H0: β = 0 null hypothesis is that the theoretical slope is equal to 0 HA: β ≠ 0 alternative hypothesis is that the theoretical slope does not 0

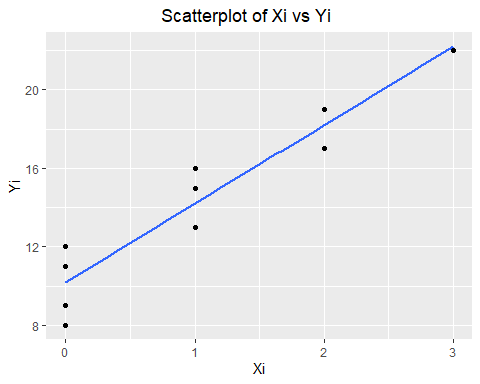
## Obtain the estimated regression function. Plot the estimated regression function and the data (use R to generate the scatter plot and the regression line) Does a linear regression function appear to give a good fit in this case?

as.out = lm(airfreight\_shipments$Yi ~ airfreight\_shipments$Xi)  
summary(as.out)

##   
## Call:  
## lm(formula = airfreight\_shipments$Yi ~ airfreight\_shipments$Xi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.2 -1.2 0.3 0.8 1.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.2000 0.6633 15.377 3.18e-07 \*\*\*  
## airfreight\_shipments$Xi 4.0000 0.4690 8.528 2.75e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.483 on 8 degrees of freedom  
## Multiple R-squared: 0.9009, Adjusted R-squared: 0.8885   
## F-statistic: 72.73 on 1 and 8 DF, p-value: 2.749e-05

qplot(x = Xi, y = Yi, data = airfreight\_shipments, geom = "point") +  
 geom\_smooth(method = lm, se = FALSE) +  
 ggtitle(" Scatterplot of Xi vs Yi")

## `geom\_smooth()` using formula 'y ~ x'



cor(airfreight\_shipments$Yi, airfreight\_shipments$Xi)

## [1] 0.949158

The estimated regression function is: y(hat) = 10.2 + 4Xi. The p value of the predictor variable (Xi) is 2.749e-05 which is below the 0.05 threshold indicating that we would reject the null hypothesis that the slope is equal to 0.The Multiple R-squared value is 0.9009.The correlation coefficient is 0.949158 which indicates a very strong positive correlation. The linear model does appear to give a good fit as a linear relationship does exist between Xi, the independent variable, and Yi is the dependent variable as seen in the scatterplot above.

## Slope and y-intercept

slope: 4

y-intercept: 10.2

The slope indicates that for every increase by 1 transfer (Xi), the number of ampules found to be broken upon arrival (Yi) increases by 4 transfers on average.

## Obtain a point estimate of the expected number of broken ampules when X = 1 transfer is made.

y(hat) = 10.2 + 4Xi y(hat) = 10.2 + 4(1) y(hat) = 14.2

Expected number of broken ampules when X = 1 transfer is made is 14.2

## Estimate the increase in the expected number of ampules broken when there are 2 transfers as compared to 1 transfer.

Estimated increase for 1 transfer = 4Xi = 4(1) = 4/1 = 4 Estimated increase for 2 transfers = 4Xi = 4(2) = 8/1 = 8

The estimated increase for 2 transfers is 8

## Residuals for Airfreight Shipment Breakage of Entire Dataset

residuals(as.out)

## 1 2 3 4 5 6 7 8 9 10   
## 1.8 -1.2 -1.2 1.8 -0.2 -1.2 -2.2 0.8 0.8 0.8

## Find the specific residual for Xi = 3. Show all of your work. Determine if the observed value of 22 is above or below average.

observed\_Yi\_for3 <- airfreight\_shipments %>% filter(Xi == 3)  
observed\_Yi\_for3

## # A tibble: 1 × 2  
## Xi Yi  
## <dbl> <dbl>  
## 1 3 22

y(hat) = 10.2 + 4Xi y(hat) = 10.2 + 4(3) expected value = y(hat) = 22.2

residual = observed - expected = 22 - 22.2 = -0.2

Since the residual is a negative number, the value of 22 is below average

## Verify that your fitted regression line goes through the point (Xbar,Ybar)

Xbar <- mean(airfreight\_shipments$Xi)  
ybar <- mean(airfreight\_shipments$Yi)  
Xbar

## [1] 1

ybar

## [1] 14.2

The value for Xbar is 1 The value for Ybar is 14.2

y(hat) = 10.2 + 4(1) = 14.2

The fitted regression line does goes through the point (Xbar,Ybar)

## Which value in the R summary output table determines if your model is doing a good job explaining the variation in the dependent variable produced by the model. In this case identify this specific proportion of variation.

The Multiple R-squared value determines if your model does a good job at explaining the variation in the Yi variable produced by the model. The R-squared value is 0.9009. This means that 90.09% of the Yi variation can be explained by the regression model with Xi.