STA471 - Homework 6

Richard McCormick

2023-11-09

1. Using least squares procedures, estimate the b's in the model:

$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

```
X1 <- c( 1,4,9,11,3,8,5,10,2,7,6 )
X2 <- c( 8,2,-8,-10,6,-6,0,-12,4,-2,-4 )
Y <- c( 6,8,1,0,5,3,2,-4,10,-3,5 )

data <- data.frame( Y, X1, X2 )
model <- lm( Y ~ X1 + X2 )
summary( model )
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##     -4     -2      1      2      3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.0000     6.0950   2.297   0.0507 .
## X1           -2.0000     1.1984  -1.669   0.1337
## X2           -0.5000     0.5992  -0.834   0.4283
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.915 on 8 degrees of freedom
## Multiple R-squared:  0.6421, Adjusted R-squared:  0.5526
## F-statistic: 7.176 on 2 and 8 DF,  p-value: 0.01641
```

```
print( paste( "b1 = ", coef( model )[2] ) )
```

```
## [1] "b1 =  -2"
```

```
print( paste( "b2 = ", round( coef( model )[3], 3 ) ) )
```

```
## [1] "b2 =  -0.5"
```

2. Write out the analysis of variance table.

```r
anova( model )
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## X1         1 116.082 116.082 13.6567 0.006082 **
## X2         1   5.918   5.918  0.6963 0.428256
## Residuals  8  68.000   8.500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
RSS <- sum( anova( model )[1:2, 2] )
print( paste( "Regression Sum of Squares:", RSS ) )
```

```
## [1] "Regression Sum of Squares: 122"
```

```r
TSS <- sum( anova( model )[,2] )
print( paste( "Total Sum of Squares:", TSS ) )
```

```
## [1] "Total Sum of Squares: 190"
```

3. Using $\alpha = 0.05$, test to determine if the overall regression is statistically significant.

I. Hypothesis
H_0: $\beta_1 = \beta_2 = 0$
H_A: At least one of: β_1, $\beta_2 \neq 0$

```
summary( model )
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##     -4     -2      1      2      3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.0000     6.0950   2.297   0.0507 .
## X1           -2.0000     1.1984  -1.669   0.1337
## X2           -0.5000     0.5992  -0.834   0.4283
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.915 on 8 degrees of freedom
## Multiple R-squared:  0.6421, Adjusted R-squared:  0.5526
## F-statistic: 7.176 on 2 and 8 DF,  p-value: 0.01641
```

II. Test Statistic
Test Statistic: $F = \frac{MS_{reg}}{MS_{resid}}$
Observed Statistic: $F_{obs} = 7.176$, from Summary Table.
p-value = 0.01641, from Summary Table.

III. Conclusion
P-value $= 0.01641 < \alpha = 0.05$.
At the $\alpha = 0.05$ level of significance, the overall regression model is statistically significant. Thus, we **reject** the null hypothesis, and accept the alternative hypothesis.

4. Calculate the square of the multiple correlation coefficient, namely, R2. What portion of the total variation about images is explained by the two variables?

```r
summary( model )
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##     -4     -2      1      2      3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.0000     6.0950   2.297   0.0507 .
## X1           -2.0000     1.1984  -1.669   0.1337
## X2           -0.5000     0.5992  -0.834   0.4283
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.915 on 8 degrees of freedom
## Multiple R-squared:  0.6421, Adjusted R-squared:  0.5526
## F-statistic: 7.176 on 2 and 8 DF,  p-value: 0.01641
```

```r
print( paste( "R-squared value is:", summary( model )$r.squared ) )
```

```
## [1] "R-squared value is: 0.642105263157895"
```

The portion of total variation about images explained by the two variables in this model is 64.21%

5. The inverse of the X'X matrix for this problem is as follows:

$$\begin{bmatrix} 4.3705 & -0.8495 & -0.4086 \\ -0.8495 & 0.1690 & 0.0822 \\ -0.4086 & 0.0822 & 0.0422 \end{bmatrix}$$

```r
xinv_mat <- matrix( c( 4.3705, -0.8495, -0.4086, -0.8495, 0.1690, 0.0822,
                      -0.4086, 0.0822, 0.0422 ), 3, 3 )
```

Using the results of the analysis of variance table with this matrix, calculate estimates of the following:

a. Variance and confidence intervals of b1.

```r
# 1. Find s^2 = RSS / (n - 2)
s.squared = ( deviance(model) / 8 )

# 2. Create variance matrix, using (X'X)^-1
var_mat <- s.squared * xinv_mat

# 3. Get the i-1th element
b1.variance <- var_mat[2, 2]

print( paste( "Variance of b1 =", b1.variance ) )
```

```
## [1] "Variance of b1 = 1.4365"
```

```r
# 95% confidence interval for b1
confint( model, level=0.95 )[2,]
```

```
##      2.5 %     97.5 %
## -4.7636013  0.7636013
```

b. Variance and confidence intervals of b2.

```r
# Get the i-1th element
b2.variance <- var_mat[3,3]

print( paste( "Variance of b2 =", b2.variance ) )
```

```
## [1] "Variance of b2 = 0.3587"
```

```r
# 95% confidence interval for b2
confint( model, level=0.95 )[3,]
```

```
##      2.5 %     97.5 %
## -1.8818007  0.8818007
```

6. How useful is the regression using X_1 alone? What does X_2 contribute, given that X_1 is already in the regression?

```r
writeLines( paste( round( ( anova( model )[1,2] / RSS )*100, 3 ),
            "% of the variation in Y is explained by the regression using X1 alone." ) )
```

```
## 95.149 % of the variation in Y is explained by the regression using X1 alone.
```

```r
model.X1 <- lm( Y ~ X1 + X2 )
anova( model.X1 )
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## X1         1 116.082 116.082 13.6567 0.006082 **
## X2         1   5.918   5.918  0.6963 0.428256
## Residuals  8  68.000   8.500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
RSS.X1 <- sum( anova( model.X1 )[1, 2] )
SS.X1 <- RSS - RSS.X1

writeLines( paste( round( ( SS.X1/RSS )*100, 3 ),
              "% of the variation in Y is explained by the regression using X2,\n",
              "given that X1 is already in the model." ) )
```

```
## 4.851 % of the variation in Y is explained by the regression using X2,
##  given that X1 is already in the model.
```

7. How useful is the regression using X_2 alone? What does X_1 contribute, given that X_2 is already in the regression?

```
writeLines( paste( round( ( anova( lm( Y ~ X2 ) )[1,2] / RSS ) * 100, 3 ),
              "% of the variation in Y is explained by the regression using X2 alone." ) )
```

```
## 80.596 % of the variation in Y is explained by the regression using X2 alone.
```

```
model.X2 <- lm( Y ~ X2 + X1 )
anova( model.X2 )
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## X2         1 98.327  98.327  11.568 0.009344 **
## X1         1 23.673  23.673   2.785 0.133702
## Residuals  8 68.000   8.500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
RSS.X2 <- sum( anova( model.X2 )[1, 2] )
SS.X2 <- RSS - RSS.X2

writeLines( paste( round( ( SS.X2 / RSS ) * 100, 3 ),
              "% of the variation in Y is explained by the regression using X1,\n",
              "given that X2 is already in the model." ) )
```

```
## 19.404 % of the variation in Y is explained by the regression using X1,
##  given that X2 is already in the model.
```

8. What are your conclusions?

Given the model and our variables, it is reasonable to conclude that X_1 alone contributes the most to the explanation of variation in Y in this model, with more than 95% of total variation being explained. The p-value for X_2 is very high, which does not support a conclusion that X_2 contributes much to the model. The model would be more accurate when only using X_1, as X_2 does not contribute much to the overall regression.

a. Fit an appropriate model to the data using $\alpha = 0.05$ and compare the effectiveness of the appropriate model with the full model by adjusted R^2.

```r
fit.model <- lm( Y ~ X1 )
summary( fit.model )
```

```
##
## Call:
## lm(formula = Y ~ X1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.973 -2.082  1.082  2.095  2.946
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.1636     1.8533   4.945 0.000797 ***
## X1           -1.0273     0.2732  -3.759 0.004489 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.866 on 9 degrees of freedom
## Multiple R-squared:  0.611,  Adjusted R-squared:  0.5677
## F-statistic: 14.13 on 1 and 9 DF,  p-value: 0.004489
```

```r
summary( model )
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##     -4     -2      1      2      3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.0000     6.0950   2.297   0.0507 .
## X1           -2.0000     1.1984  -1.669   0.1337
## X2           -0.5000     0.5992  -0.834   0.4283
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.915 on 8 degrees of freedom
## Multiple R-squared:  0.6421, Adjusted R-squared:  0.5526
## F-statistic: 7.176 on 2 and 8 DF,  p-value: 0.01641
```

The adjusted R-squared value is higher when only using X_1 - 0.5677 for the more appropriate model, and 0.5526 for the full model.