MAT3375_Assignment-1

Rahul Atre

2023-09-15

Exercise 2.1

Table B.1 gives data concerning the performance of the 26 National Football League teams in 1976. It is suspected that the number of yards gained rushing by opponents (x_8) has an effect on the number of games won by a team (y).

a. Fit a simple linear regression model relating games won y to yards gained rushing by opponents x_8.

Ans: For this example, we will use the method of least squares to obtain a fitted regression model.

First, we must import and load the data from table.b1 using an R package containing this textbook data sets. We can do this by running the following command on the console

```
install.packages("MPV", repos = "http://cran.us.r-project.org")
library(MPV)
```

Now, let x_yards, x_8, rep. the # of yards gained rushing by opponents, and y_won, y rep. the number of games won by a particular team

```
data(table.b1)
x_yards = table.b1$x8
y_won = table.b1$y
```

The desired simple regression model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ According to methods of least squares, $b_0 = \overline{Y} - b_1\overline{X}$ and $b_1 = \frac{\sum(X_i - \overline{X})Y_i}{\sum(X_i - \overline{X})^2}$

We can calculate the parameters b_0 and b_1 using the lm() function to fit the linear model:

```
lin_model = lm(y_won ~ x_yards)
lin_model
```

```
##
## Call:
## lm(formula = y_won ~ x_yards)
##
## Coefficients:
## (Intercept)      x_yards
##   21.788251    -0.007025
```

From the above function call, we obtain $b_0 = 21.788$ and b_1 = -0.007 (rounded to 3 decimal places).

Therefore, a simple linear model relating games won to yards gained rushing by opponents is $\hat{Y} = 21.788 - 0.007X$.

b. Construct the analysis-of-variance table and test for significance of regression.

Ans: The analysis-of-variance (ANOVA) table can be generated using R:

```
anova(lin_model)
```

```
## Analysis of Variance Table
##
## Response: y_won
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x_yards    1 178.09 178.092  31.103 7.381e-06 ***
## Residuals 26 148.87   5.726
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As for testing the significance of regression, the anova table indicates how to test the null hypothesis:

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

As we know, the null hypothesis is that the slope of the line is equal to 0. However, if we take a look at the F and p-value, we can see that the F is 31.103 which is quite large, and the p-value is less than 0.001, which is very small. Therefore, we can reject H_0 and conclude that the slope is not equal to 0. In larger context, this means that there is a correlation between yards gained rushing by opponents and the number of games won by a team.

c. Find a 95% CI on the slope.

Ans: Since we want the CI on the slope β_1, we can use the formula $b_1 \pm t_{1-\alpha/2,n-2} * s(b_1)$. We know that $n = 28$, $b_1 = -0.007$, $\alpha = 0.95$. We can generate the CI using R:

```
confint(lin_model, level = 0.95)
```

```
##                    2.5 %       97.5 %
## (Intercept) 16.246064040 27.330437725
## x_yards     -0.009614347 -0.004435854
```

Therefore, from the above function call, we are 95% confident that the slope is in the interval [-0.0096, -0.0044].

d. What percent of the total variability in y is explained by this model?

Ans: The coefficient of determination R^2 can be used to find the percentage of total variability in y from the formula

$$\frac{\sum(\hat{Y} - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2}$$

. Using R to generate this:

```r
summary(lin_model)
```

```
##
## Call:
## lm(formula = y_won ~ x_yards)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.804 -1.591 -0.647  2.032  4.580
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.788251   2.696233   8.081 1.46e-08 ***
## x_yards     -0.007025   0.001260  -5.577 7.38e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.393 on 26 degrees of freedom
## Multiple R-squared:  0.5447, Adjusted R-squared:  0.5272
## F-statistic:  31.1 on 1 and 26 DF,  p-value: 7.381e-06
```

From the above function call, $R^2 = 0.5447 = 54.47\%$. Therefore, about 54.47% of the total variability in y is explained by this model.

e. Find a 95% CI on the mean number of games won if opponents yards rushing is limited to 2000 yards.

Ans: In order to find this, we need to set $X = 2000$ to get a 95% confidence interval, and use the point estimate $\hat{Y} = b_0 + b_1 X$. So, the CI is $\hat{Y} \pm t_{1-\alpha/2,n-2} * s(\hat{Y})$. Generating the CI using R:

```r
new.dat = data.frame(x_yards = 2000)
predict(lin_model, newdata = new.dat, interval = "confidence")
```

```
##       fit      lwr      upr
## 1 7.73805 6.765753 8.710348
```

Therefore, a 95% CI on the mean number of games won if opponents yards rushing is limited to 2000 yards is [6.7658, 8.7103].

Exercise 2.10

The weight and systolic blood pressure of 26 randomly selected males in the age group 25-30 are shown below. Assume that weight and blood pressure (BP) are jointly normally distributed.

a. Find a regression line relating systolic blood pressure to weight.

Ans: For this example, we will use the method of least squares to obtain the estimates for β_0 and β_1.

Let x_weight rep. the weight of a male. Let y_systolicBP represent the systolic blood pressure of a male.

We must first load the data from the table. Then, we can use the lm() function to find the regression line in R:

```r
data(p2.10)
x_weight = p2.10$weight
y_systolicBP = p2.10$sysbp

lin_model = lm(y_systolicBP ~ x_weight)
lin_model
```

```
##
## Call:
## lm(formula = y_systolicBP ~ x_weight)
##
## Coefficients:
## (Intercept)     x_weight
##     69.1044       0.4194
```

From the above function call, we obtain $b_0 = 69.1044$ and $b_1 = 0.4194$.

Therefore, a regression line relating systolic blood pressure to weight is $\hat{Y} = 69.1044 + 0.4194X$.

b. Estimate the correlation coefficient.

Ans: The sample correlation coefficient is defined as the following:

$$\frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}}$$

. Using R to generate this:

```r
cor(x_weight, y_systolicBP)
```

```
## [1] 0.7734903
```

From the above function call, we obtain $r = 0.7735$. Therefore, the estimated correlation coefficient is 0.7735.

c. Test the hypothesis that $\rho = 0$.

Ans: We have the following hypothesis for ρ:

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

The following test statistic is given for ρ under H_0:

$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim Student_{n-2}$. Using R to generate this:

```r
cor.test(x_weight, y_systolicBP)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x_weight and y_systolicBP
## t = 5.9786, df = 24, p-value = 3.591e-06
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5513214 0.8932215
## sample estimates:
##       cor
## 0.7734903
```

We can see that the t-statistic is 5.9786 which is quite small, and the p-value is less than 0.001, which is very small. Therefore, we can reject H_0 and conclude that the correlation coefficient is not equal to 0. In larger context, this means that there is evidence to suggest a correlation between the systolic blood pressure and weight.

d. Test the hypothesis that $\rho = 0.6$.

Ans: We have the following hypothesis for ρ:

$H_0 : \rho_0 = 0.6$

$H_1 : \rho_0 \neq 0.6$

The test-statistic for $H_0 : \rho = \rho_0$ can be computed as shown below. We know that $r = 0.7735$, $\rho_0 = 0.6$, $n = 26$:

$Z = (arctanh(r) - arctanh(\rho_0))(n-3)^{1/2}$

$Z = (arctanh(0.7735) - arctanh(0.6))(26-3)^{1/2}$ $Z = (1.02898 - 0.693147)(23)^{1/2}$ $Z = 1.61061$

If we let $\alpha = 0.5$, then the rejection for H_0 is if $|Z_0| > Z_{\alpha/2} = 1.96$. Since that is not true, we do not reject H_0, as there is not enough evidence to suggest that $\rho \neq 0.6$, H_1, is true.

e. Find a 95% CI for ρ.

Letting $\alpha = 0.5$, the formula for a 95% CI of ρ is given as follows:

$[tanh(arctanh(r) - \frac{Z_{\alpha/2}}{\sqrt{n-3}}), tanh(arctanh(r) + \frac{Z_{\alpha/2}}{\sqrt{n-3}})]$. Using R to generate this:

```
cor.test(x_weight, y_systolicBP)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x_weight and y_systolicBP
## t = 5.9786, df = 24, p-value = 3.591e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5513214 0.8932215
## sample estimates:
##       cor
## 0.7734903
```

From the above function call, we get the 95% CI interval for ρ to be [0.5513, 0.8932].

Exercise 2.22

Consider the methanol oxidation data in Table B.20. The chemist believes that the ratio of inlet oxygen to the inlet methanol controls the conversion process. Perform a through analysis of these data. Do the data support the chemist's belief?

Ans: First, let's import the data from table.b20 in the textbook data set.

Let x_OxyMethRatio, x_5 rep. the # of ratio of inlet oxygen to inlet methanol, and y_convProcess, y rep. the percent conversion.

```
data(table.b20)
x_OxyMethRatio = table.b20$x5
y_convProcess = table.b20$y
```

Using the lm() and summary function to obtain a summary of the data:

```
lin_model = lm(y_convProcess ~ x_OxyMethRatio)
summary(lin_model)
```

```
##
## Call:
## lm(formula = y_convProcess ~ x_OxyMethRatio)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30.29 -24.15 -16.76  29.42  63.20
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       21.25      20.99   1.013    0.326
## x_OxyMethRatio     7.80      16.78   0.465    0.648
##
## Residual standard error: 35.76 on 16 degrees of freedom
## Multiple R-squared:  0.01333,    Adjusted R-squared:  -0.04834
## F-statistic: 0.2161 on 1 and 16 DF,  p-value: 0.6483
```

From the above function call, we obtain $\hat{Y} = 21.25 + 7.80x_5$.

Testing the null hypothesis, $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$, we obtain F to be 0.2161 and p-value as 0.6483. Also, $R^2 = 0.0133$, meaning that about 1.3% of the total variability in Y is explained by the model, which is not much. Since the F and p-value is small, we fail to reject H_0.

Therefore, this implies that there is not enough evidence to suggest the ratio of inlet oxygen to inlet methanol controls the conversion process. The data does not support the chemist's beliefs.