uwa6xv_M03_HW

Alanna Hazlett

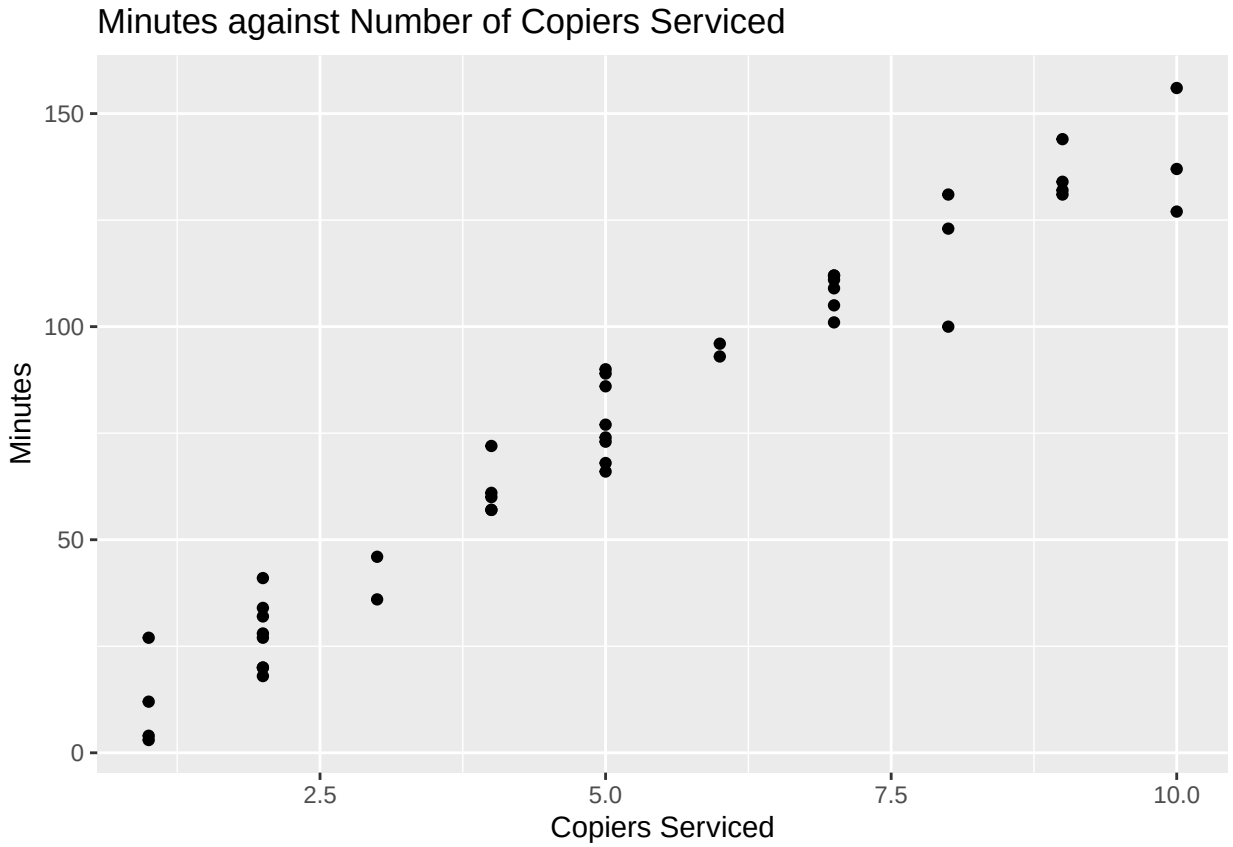
2024-02-14

We will use the dataset "copier.txt" for this question. The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, Serviced is the number of copiers serviced and Minutes is the total number of minutes spent by the service person.

Problem 1

(a)
The predictor variable is "Serviced", the number of copiers serviced. The response variable is "Minutes", the total number of minutes spent servicing the copiers.
(b)
Produce a scatterplot of the two variables. How would you describe the relationship between the number of copiers serviced and the time spent by the service person?

```
copier<-read.table(file="copier.txt", header = TRUE)
ggplot(copier,aes(x=Serviced,y=Minutes))+
  geom_point()+
  labs(x="Copiers Serviced",y="Minutes",title="Minutes against Number of Copiers Serviced")
```



The relationship between number of copiers serviced and minutes spent servicing the copiers appears to be a positive linear relationship. As the number of copiers increase, so does the number of minutes spent servicing them.

(c)

What is the correlation between the total time spent by the service person and the number of copiers serviced? Interpret this correlation contextually.

```
cor(copier$Serviced,copier$Minutes)
```

```
## [1] 0.978517
```

The correlation is a positive value, which indicates that as Serviced increases, so does Minutes. It is a strong association, in that the value is near 1, meaning that the number of copiers serviced highly affects the number of minutes servicing the copiers.

(d)

Can the correlation found in part 1c be interpreted reliably? Briefly explain.

The correlation of 0.97 can be interpreted reliably, because there does not appear to be any curves or distinct outliers in our scatterplot that would alter our correlation value.

(e)

Use the lm() function to fit a linear regression for the two variables. Where are the values of beta1, beta0, R2, and sigma2 for this linear regression?

```
result<-lm(Minutes~Serviced,data=copier)
summary(result)
```

```
##
## Call:
```

```
## lm(formula = Minutes ~ Serviced, data = copier)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7723  -3.7371   0.3334   6.3334  15.4039
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5802     2.8039  -0.207    0.837
## Serviced     15.0352     0.4831  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 43 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

$$\hat{\beta}_1 = 15.0352 \textit{ this isfound in Estimate column and } (Intercept) \textit{ row}$$

$$\hat{\beta}_0 = -0.5802 \textit{ this is found in Estimate column and Serviced row}$$

$$R^2 = 0.9575 \textit{ this is the value associated with Multiple } R - squared$$

$$\hat{\sigma}^2 \textit{is estimated by } s^2,$$

$$s^2 = (8.914)^2 = 79.459396$$

$$s \textit{ is the value associated with Residual standard error}$$

(f)
Interpret the values of beta1, beta0 contextually. Does the value of beta0 make sense in this context?

$$\hat{\beta}_1$$

is the slope, as Serviced increases by one copier, the minutes increases by 15.0352 minutes, on average.

$$\hat{\beta}_0$$

is the estimated intercept. Based on our model the number of minutes it would take to service zero copiers is -0.5802 minutes. This is not possible as we can not measure time negatively.

(g)
Use the anova() function to produce the ANOVA table for this linear regression. What is the value of the ANOVA F statistic? What null and alternative hypotheses are being tested here? What is a relevant conclusion based on this ANOVA F statistic?

```
anova.tab<-anova(result)
anova.tab
```

```
## Analysis of Variance Table
##
## Response: Minutes
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Serviced   1  76960   76960  968.66 < 2.2e-16 ***
## Residuals 43   3416      79
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F statistic is:

$$F = \frac{MS_R}{MS_{res}} = 968.66$$

Our hypotheses are:

$$H_0 : \hat{\beta}_1 = 0$$
$$H_a : \hat{\beta}_1 \neq 0$$

The critical value is:

```
qf(1-0.05, 1, 45-2)
```

```
## [1] 4.067047
```

Our F statistic of 968.66 is larger than our critical value of 4.06 (alternatively our p-value of 2.2X10^-16 is smaller than 0.05), so we reject our null hypothesis of

$$\hat{\beta}_1 = 0$$

our data supports the alternative hypothesis of the slope being different from 0, which indicates a linear association between our variables of Serviced and Minutes for the copiers.

Problem 2

(You may only use R as a simple calculator or to find p-values or critical values) Suppose that for n = 6 students, we want to predict their scores on the second quiz using scores from the first quiz. The estimated regression line is y = 20 + 0.8x

(e)

```
qf(1-0.05,1,4)
```

```
## [1] 7.708647
```

```
(1-pf(25.333,1,4)) * 2
```

```
## [1] 0.01463408
```

Alanna Hazlett
2/14/24

MØ3 HW

2.) Suppose that for n = 6 students, we want to predict their scores on the second quiz using scores from the first quiz. The estimated regression line is $\hat{y} = 20 + 0.8x$,

a.) For each individual observation calculate its predicted score on the second quiz $\hat{y}_i$ and the residual e_i

	col 1	col 2	col 3	col 4	col 5	col 6
x_i	70	75	80	80	85	90
y_i	75	82	80	86	90	91
$\hat{y}_i$	76	80	84	84	88	92
e_i	-1	2	-4	2	2	-1

$e_i = y_i - \hat{y}_i$

Col 1: $\hat{y}_i = 20 + 0.8(70) = 76$ $\quad e_i = 75 - 76 = -1$

Col 2: $\hat{y}_i = 20 + 0.8(75) = 80$ $\quad e_i = 82 - 80 = 2$

Col 3: $\hat{y}_i = 20 + 0.8(80) = 84$ $\quad e_i = 80 - 84 = -4$

Col 4: $\hat{y}_i = 20 + 0.8(80) = 84$ $\quad e_i = 86 - 84 = 2$

Col 5: $\hat{y}_i = 20 + 0.8(85) = 88$ $\quad e_i = 90 - 88 = 2$

Col 6: $\hat{y}_i = 20 + 0.8(90) = 92$ $\quad e_i = 91 - 92 = -1$

b.) Complete ANOVA table for this dataset,

	DF	SS	MS	F-Stat	p-value
Regression	1	190	190	25.3	0.0099
Residual	4	30	7.5	***	***
Total	5	220	***	***	***

$\bar{y} = (75 + 82 + 80 + 86 + 90 + 91) \div 6 = 84$

$$SS_R = \sum_{i=1}^{6} (\hat{y}_i - \bar{y})^2 = (75-84)^2 + (82-84)^2 + (80-84)^2 + (86-84)^2 + (90-84)^2 + (91-84)^2$$

$$= 81 + 4 + 16 + 4 + 36 + 49$$

$$\boxed{SS_R = 190}$$

$$SS_{res} = \sum_{i=1}^{6} (y_i - \hat{y}_i)^2 = (75-76)^2 + (82-80)^2 + (80-84)^2 + (86-84)^2 + (90-88)^2 + (91-92)^2$$

$$= 1 + 4 + 16 + 4 + 4 + 1$$

$$\boxed{SS_{res} = 30}$$

$$SS_T = SS_R + SS_{res}$$

$$= 190 + 30$$

$$\boxed{SS_T = 220}$$

$MS_R = SS_R / df_R = 190/1 = \boxed{190 = MS_R}$

$MS_{res} = SS_{res} / df_{res} = 30/4 = \boxed{7.5 = MS_{res}}$

$F\text{ statistic} = MS_R / MS_{res} = 190/7.5 = \boxed{25.\overline{3} = F\text{ statistic}}$

c.) Calculate the sample estimate of the variance σ^2 for the regression model.

$s^2 = MS_{res} = \boxed{7.5 = s^2}$

d.) What is value of R^2?

$$R^2 = \frac{SS_R}{SS_T} = \frac{190}{220} = \boxed{0.8636 = R^2}$$

This is a strong/good fit to our model.

e.) Carry out ANOVA F test. What is an appropriate conclusion?

$H_0: \beta_1 = 0$ $\quad H_a: \beta_1 \neq 0$

F statistic = 25,3 $\quad$ critical value = qf(1-0.05, 1, 4)
= 7.7086647

F statistic > critical value,
so we reject the null hypothesis. Our data supports the alternative hypothesis of the slope being different from zero. This indicates that there is a linear association between our variables.

- alternatively our p-value = (1-pf(25,333, 1, 4)) x2
= 0.01463

p-value < 0.05, so we reject the null hypothesis. Our data supports the alternative hypothesis.

Alanna Hazlett
2/15/24

MD3 HW Continued

3.) $SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ $\quad e_i = y_i - \hat{y}_i$ $\quad y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$

$\hat{\beta}_0$ = intercept

$\hat{\beta}_1$ = slope

$SS_{res} = \sum_{i=1}^{n} (e_i)^2$ $\quad e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

$$\frac{\partial SS_{res}}{\partial \hat{\beta}_0} = \frac{\partial}{\partial \hat{\beta}_0}\left[\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\right]$$

$$= \sum_{i=1}^{n}\left[\frac{\partial}{\partial \hat{\beta}_0}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\right]$$

$$= \sum_{i=1}^{n} 2 \cdot -1 (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$\frac{\partial SS_{res}}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$0 = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$0 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i$$

$$\sum_{i=1}^{n} \hat{\beta}_0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$\frac{\partial SS_{res}}{\partial \hat{\beta}_1} = \frac{\partial}{\partial \hat{\beta}_1}\left[\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\right] = \sum_{i=1}^{n}\left[\frac{\partial}{\partial \hat{\beta}_1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\right]$$

$$= \sum_{i=1}^{n} 2 \cdot -1 \cdot x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$0 = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \qquad 0 = \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$0 = \sum_{i=1}^{n} (y_i x_i - \hat{\beta}_0 x_i - \hat{\beta}_1 (x_i^2))$$

$$= \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \hat{\beta}_0 x_i - \sum_{i=1}^{n} \hat{\beta}_1 (x_i^2)$$

$$= \sum_{i=1}^{n} y_i x_i - \left(\hat{\beta}_0 \sum_{i=1}^{n} x_i\right) - \left(\hat{\beta}_1 \sum_{i=1}^{n} (x_i)^2\right)$$

$$\hat{\beta}_1 \sum_{i=1}^{n} (x_i)^2 = \sum_{i=1}^{n} y_i x_i - \left(\hat{\beta}_0 \sum_{i=1}^{n} x_i\right)$$

$$\hat{\beta}_1 \sum_{i=1}^{n} (x_i)^2 = \sum_{i=1}^{n} y_i x_i - \left[\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) \sum_{i=1}^{n} x_i\right]$$

$$\hat{\beta}_1 \sum_{i=1}^{n} (x_i)^2 = \left(\sum_{i=1}^{n} y_i x_i\right) - \bar{y}\bar{x} + \hat{\beta}_1 (\bar{x}^2)$$

$$\hat{\beta}_1 \sum_{i=1}^{n} (x_i)^2 - \hat{\beta}_1 (\bar{x}^2) = \left(\sum_{i=1}^{n} y_i x_i\right) - \bar{y}\bar{x}$$

$$\hat{\beta}_1 \left[\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2\right] = \left(\sum_{i=1}^{n} y_i x_i\right) - \bar{y}\bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - \bar{x}^2} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$