

Stat 6021: HW Set 4

Tom Lever

09/22/22

1. You will use the dataset `copier.txt` for this question. This is the same data set that you used in the last homework. 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 by 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.

It is hypothesized that the total time spent by the service person can be predicted using the number of copiers serviced. Fit an appropriate linear regression and answer the following questions.

```
times_spent_servicing_and_numbers_of_copiers_serviced <- read.table(
  "../Module_3--Simple_Linear_Regression/Homework/copier.txt", header = TRUE
)
head(times_spent_servicing_and_numbers_of_copiers_serviced, n = 3)
```

```
##   Minutes Serviced
## 1      20         2
## 2      60         4
## 3      46         3
```

```
library(TomLeversRPackage)
data_set <- times_spent_servicing_and_numbers_of_copiers_serviced
linear_model <- lm(Minutes ~ Serviced, data = data_set)
summarize_linear_model(linear_model)
```

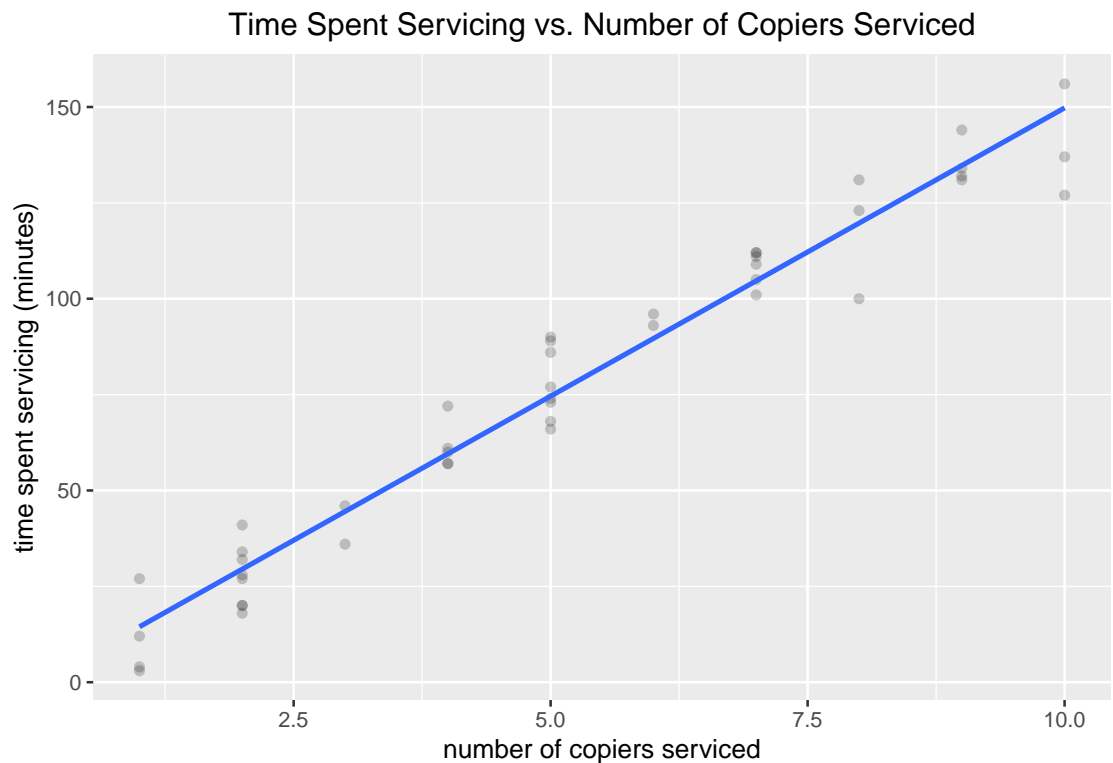
```
##
## Call:
## lm(formula = Minutes ~ Serviced, data = data_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7723  -3.7371   0.3334   6.3334  15.4039
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5802     2.8039  -0.207   0.837
## 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
##
## E(y | x) = B_0 + B_1 * x = -0.5802 + 15.0352 * x
```

```
## Number of observations: 45
## Estimated variance of errors: 79.459396
## Multiple R: 0.978516981701943 Adjusted R: 0.978008179924892
```

- (a) Produce an appropriate scatterplot and comment on the relationship between the total time spent by the service person and the number of copiers serviced.

The response variable is **Minutes**, the total number of minutes spent by the service person. The predictor is **Serviced**, the number of copiers serviced.

```
library(ggplot2)
ggplot(
  times_spent_servicing_and_numbers_of_copiers_serviced,
  aes(x = Serviced, y = Minutes)
) +
  geom_point(alpha = 0.2) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    x = "number of copiers serviced",
    y = "time spent servicing (minutes)",
    title = "Time Spent Servicing vs. Number of Copiers Serviced"
  ) +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 0)
  )
)
```



The relationship between time spent servicing and number of copiers serviced appears linear. A line of best fit has been rendered to aid in this determination. A simple linear regression model appears reasonable for time spent servicing and number of copiers data.

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

We assume that the sample of service calls is simple random. We assume that (Serviced, Minutes) matched pairs of data for the service calls are independent. We assume that errors between actual and predicted times spent servicing are normally and independently distributed with mean 0 and constant variance. The linear model for time spent servicing vs. number of copiers serviced has an adjusted sample linear Pearson coefficient $R = 0.978$. The adjusted sample linear Pearson correlation coefficient measures the strength of the linear relationship between (Serviced, Minutes) matched pairs of data for the sample of service calls. The linear model has a positive linear Pearson correlation. The value of the adjusted sample linear Pearson correlation coefficient lies between -1 and 1 inclusive. Since the scatterplot shows a reasonable linear association, the adjusted sample linear Pearson correlation coefficient is reliable. Since the above probability $p < 2.2 \times 10^{-16}$ is less than a significance level $\alpha = 0.05$, we reject a null hypothesis of a linear regression t test that there is no correlation between time spent servicing and number of copiers serviced, and conclude that there is a correlation between time spent servicing and number of copiers serviced. The linear model has an adjusted sample coefficient of determination of $R^2 = 0.957$. The adjusted sample coefficient of determination is the proportion of variation in time spent servicing that is explained by the linear relationship / number of copiers. The adjusted sample coefficient of determination lies between 0 and 1. Since the adjusted sample coefficient of determination is greater than 0.8, the linear model is precise and good for prediction.

- (c) Can the correlation found in part 1b be interpreted reliably?

As above, since the scatterplot shows a reasonable linear association, the correlation coefficient $R = 0.978$ is reliable.

- (d) Obtain the 95 percent confidence interval for the slope β_1 .

```
confidence_level <- 0.95
confint(linear_model, level = confidence_level)
```

```
##                2.5 %    97.5 %
## (Intercept) -6.234843  5.074529
## Serviced    14.061010 16.009486
```

The 95 percent confidence interval for the slope β_1 is $[14.061 \frac{\text{min}}{1}, 16.009 \frac{\text{min}}{1}]$. Since the confidence interval does not contain 0, we reject the null hypothesis $H_0 : \beta_1 = 0 \frac{\text{min}}{1}$ that the slope of the linear model for time spent servicing vs. number of copiers serviced is 0. We have sufficient evidence to support the alternate hypothesis $H_1 : \beta_1 \neq 0$.

- (e) Suppose a service person is sent to service 5 copiers. Obtain an appropriate 95 percent interval that predicts the total service time spent by the service person.

STUFF

- (f) What is the value of the residual for the first observation? Interpret this value contextually.

Let $\hat{\beta}_0 = -0.5802 \text{ min}$ and $\hat{\beta}_1 = 15.0352 \frac{\text{min}}{1}$ be the estimated intercept and slope of the linear model for time spent servicing vs. number of copiers serviced.

The residual e_i for the i th observation $(x_i, y_i) = (\text{Serviced}_i, \text{Minutes}_i)$ is the difference between the observed time spent servicing $y_i = \text{Serviced}_i$ given the observed number of copiers serviced x_i and the predicted time spent servicing $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Mathematically,

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

The residual e_1 for the first observation $(x_1, y_1) = (\text{Serviced}_1, \text{Minutes}_1) = (2, 20 \text{ min})$ in `copier.txt` is the difference between the observed time spent servicing $y_1 = \text{Serviced}_1 = 20 \text{ min}$

given the observed number of copiers serviced $x_1 = 2$ and the predicted time spent servicing $\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1 = (-0.5802 \text{ min}) + (15.0352 \frac{\text{min}}{1}) (2) = 29.4884 \text{ min}$.

$$e_i = y_i - \hat{y}_i = (20 \text{ min}) - (29.4884 \text{ min}) = -9.4884 \text{ min}$$

- (g) What is the average value of all the residuals? Is this value surprising?

STUFF