# 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
```

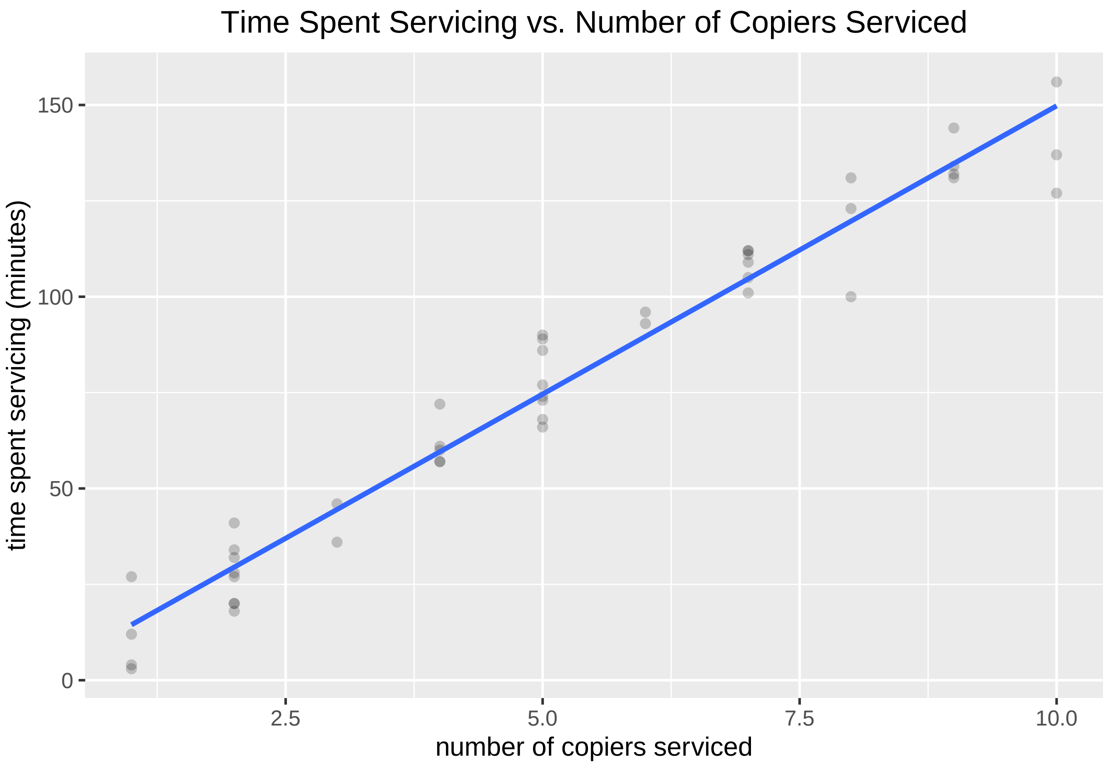
```
library(TomLeversRPackage)
data_set <- times_spent_servicing_and_numbers_of_copiers_serviced
linear_model <- lm(Minutes ~ Serviced, data = data_set)
summarize_linear_model(linear_model)
```

```
##
## 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.

```r
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 / residuals 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 coefficient $R = 0.978$. The adjusted sample linear Pearson correlation coefficient measures the strength of the linear relationship betwen (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 coefficentis 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 erxplain 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 1*b* be interpreted reliably?

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

(d) Obtain the 95 percent confidence interval for the slope $\beta_1$.

```
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 $\beta_1$ is $\left[14.061 \ \frac{min}{1}, 16.009 \ \frac{min}{1}\right]$. Since the confidence interval does not contain 0, we reject the null hypothesis $H_0 : \beta_1 = 0\frac{min}{1}$ 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.

```
predict(
    linear_model,
    data.frame(Serviced = 5),
    level = 0.95,
    interval = "prediction"
)
```

```
##        fit      lwr      upr
## 1 74.59608 56.42133 92.77084
```

An appropriate 95 percent prediction interval that predicts the total service time spent by a service person sent to service 5 copiers is $[56.421 \ min, \ 92.771 \ min]$.

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

Let $\hat{\beta}_0 = -0.5802\ min$ and $\hat{\beta}_1 = 15.0352\ \frac{min}{1}$ 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 $(x_i, y_i) = (Serviced_i, Minutes_i)$ is the difference between the observed time spent servicing $y_i = Serviced_i$ 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 $(x_1, y_1) = (Serviced_1, Minutes_1) = (2, 20\ min)$ in `copier.txt` is the difference between the observed time spent servicing $y_1 = Serviced_1 = 20\ min$ given the observed number of copiers serviced $x_1 = 2$ and the predicted time spent servicing $\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1 = (-0.5802\ min) + \left(15.0352\ \frac{min}{1}\right)(2) = 29.4884\ min$.

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

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

```
mean(linear_model$residuals)
```

```
## [1] -2.612204e-16
```

The average value of all the residuals $E(e_i) = -2.612 \times 10^{-16}\ min \approx 0\ min$. This value is unsurprising and confirms our assumption that errors / residuals between actual and predicted times spent servicing have a mean of 0 $min$.

2. A substance used in biological and medical research is shipped by airfreight to users in cartons of 1000 ampules. The sample consists of 10 shipments. A variables $x = transfer$ represents the number of times a carton was transferred from one aircraft to another during the shipment route. A variable $y = broken$ represents the number of ampules found to be broken upon arrival. A simple linear regression model is fitted using R. The corresponding output from R is shown next, with some values missing.

```
Call:
lm(formula = broken ~ transfer)

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  10.2000     0.6633  ______  ________ ***
transfer      4.0000     0.4690  ______  ________ ***

Residual standard error: 1.483 on 8 degrees of freedom

...

Analysis of Variance Table

Response: broken
          Df Sum Sq Mean Sq F value   Pr(>F)
transfer   1  160.0   160.0  ______ ________ ***
Residuals  8   17.6     2.2
```

The following values are alos provided for you.

$$\bar{x} = 1$$

$$\sum_{i=1}^{n=10} \left[ (x_i - \bar{x})^2 \right] = 10$$

(a) Carry out a hypothesis test to assess if there is a linear relationship between the variables $y = broken$ and $x = transfer$.

The regression mean square

$$MS_R = \frac{SS_R}{df_R} = \frac{160.0}{1} = 160.0$$

Errors / residuals are assumed to be normally and independently distributed with mean 0 and unknown constant variance $\sigma^2$. The residual standard error and estimated standard deviation of errors / residuals $\hat{\sigma} = 1.483\ min$. The residual mean square and estimated variance of errors / residuals

$$MS_{Res} = \hat{\sigma}^2 = (1.483\ min)^2 = \frac{SS_{Res}}{df_{Res}} = \frac{SS_{Res}}{n-p} = \frac{SS_{Res}}{n-2} = \frac{17.6}{8} = 2.2$$

By definition of a test statistic,

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{SE\left(\hat{\beta}_1\right)} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}}$$

follows a $[t_{df_{Res}} = t_{n-p} =]\, t_{n-2}$ distribution if the null hypothesis $H_0 : \beta_1 = \beta_{10}$ is true.

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{SE\left(\hat{\beta}_1\right)} = \frac{\hat{\beta}_1 - 0}{SE\left(\hat{\beta}_1\right)} = \frac{\hat{\beta}_1}{SE\left(\hat{\beta}_1\right)} = \frac{4.0000}{0.4690} = 8.528$$

$$t_0 = \sqrt{{t_0}^2} = \sqrt{\frac{\hat{\beta}_1^2}{SE\left(\hat{\beta}_1\right)^2}} = \sqrt{\frac{\hat{\beta}_1^2}{MS_{Res}/S_{xx}}} = \sqrt{\frac{\hat{\beta}_1^2 S_{xx}}{MS_{Res}}} = \sqrt{\frac{\hat{\beta}_1 S_{xy}}{MS_{Res}}} = \sqrt{F_0}$$

$$t_0 = \sqrt{\frac{MS_R}{MS_{Res}}} = \sqrt{\frac{160.0}{2.2}} = 8.528$$

The quantile of a Student's $t$ probability distribution with $df_{Res} = n - p = n - 2$ degrees of freedom for which the probability that a random test statistic is greater is equal to the significance level $\alpha = 0.05$

$$t_{\alpha,\ df_{Res}} = qt(\alpha,\ df_{Res},\ lower.tail = FALSE) = 1.860$$

The quantile of a Student's $t$ probability distribution with $df_{Res} = n - p = n - 2$ degrees of freedom for which the probability that a random test statistic is greater is equal to half the significance level $\alpha = 0.05$

$$t_{\alpha/2,\ df_{Res}} = qt(\alpha/2,\ df_{Res},\ lower.tail = FALSE) = 2.306$$

```r
significance_level <- 0.05
number_of_observations <- 10
number_of_parameters <- 2
degrees_of_freedom <- number_of_observations - number_of_parameters
qt(significance_level, degrees_of_freedom, lower.tail = FALSE)
```

```
## [1] 1.859548
```

```r
qt(significance_level / 2, degrees_of_freedom, lower.tail = FALSE)
```

```
## [1] 2.306004
```

Given a significance level $\alpha = 0.05$, we test a null hypothesis $H_0 : \beta_1 = 0$ that the slope of a linear model for the number of broken ampules, $y = broken$, vs. the number of transfers, $x = transfer$, is equal to 0. If we have sufficient evidence to reject the null hypothesis, we have sufficient evidence to support an alternate hypothesis $H_1 : \beta_1 \neq 0$ that the slope of the linear model for the number of broken ampules vs. the number of transfers is not equal to 0. Since the alternate hypothesis $H_1$ involves "$\neq$", we have sufficient evidence to reject the null hypothesis if the magnitude $|t_0|$ of the test statistic $t_0$ is greater than the quantile $t_{\alpha/2,\ df_{Res}}$. If the alternative $H_1$ were to involve "$<$" or "$>$", we would have sufficient evidence to reject the null hypothesis if the magnitude $|t_0|$ of the test statistic $t_0$ is greater than $t_{\alpha,\ df_{Res}}$. Since $|t_0| > t_{\alpha/2,\ df_{Res}}$, we reject the null hypothesis. We have sufficient evidence to support the alternate hypothesis.

We have sufficient evidence to reject the null hypothesis if the probability $p$ that a the test statistic for a random sample of shipments is less than $-|t_0|$ or greater than $|t_0|$, assuming the null hypothesis is true, is less than significance level $\alpha$.

```r
p <- pt(8.528, degrees_of_freedom, lower.tail = FALSE) * 2
# 2 if the alternate hypothesis involves "\neq"
# 1 if the alternate hypothesis involves "<" or ">"
p
```

```
## [1] 2.748737e-05
```

Since $p < \alpha$, we reject the null hypothesis. We have sufficient evidence to support the alternate hypothesis.

(b) Calculate a 95 percent confidence interval that estimates the unknown slope of the linear model that fits observations / $(x = transfer, y = broken)$ matched pairs of data for a population of shipments.

The 95 percent confidence interval that estimates the unknown slope of the linear model that fits observations for a population of shipments

$$\left[\hat{\beta}_1 - t_{\alpha/2,\ df_{Res}}\ SE\left(\hat{\beta}_1\right),\ \hat{\beta}_1 + t_{\alpha/2,\ df_{Res}}\ SE\left(\hat{\beta}_1\right)\right]$$

$$[(4.0000) - (2.306)(0.4690),\ (4.0000) + (2.306)(0.4690)]$$

$$[2.918,\ 5.082]$$

(c) A consultant believes the mean number of broken ampules when no transfers are made is different from 9. Conduct an appropriate hypothesis test. State the hypotheses, calculate the test statistic, and write the corresponding conclusion in context, in response to his belief.

By definition of a test statistic,

$$t_0 = \frac{\widehat{E(y|x_0)} - E(y|x_0)}{SE\left(\widehat{E(y|x_0)}\right)} = \frac{\widehat{E(y|x_0)} - E(y|x_0)}{\sqrt{MS_{Res}\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{s_{xx}}\right]}} = \frac{\widehat{E(y|x_0)} - E(y|x_0)}{\sqrt{MS_{Res}\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n=10}\left[(x_i-\bar{x})^2\right]}\right]}}$$

follows a $[t_{df_{Res}} = t_{n-p} =]\, t_{n-2}$ distribution if the null hypothesis $H_0 : E(y|x_0) = \mu$ is true.

$$t_0 = \frac{\widehat{E(y|x_0=0)} - E(y|x_0=0)}{SE\left(\widehat{E(y|x_0=0)}\right)} = \frac{10.2000-9}{0.6633} = \frac{10.2000-9}{\sqrt{2.2\left[\frac{1}{10} + \frac{(0-1)^2}{10}\right]}} = 1.809$$

Given a significance level $\alpha = 0.05$, we test a null hypothesis $H_0 : E(y|x_0 = 0) = 9$ that the mean number of broken ampules when 0 transfers are made is equal to 9. If we have sufficient evidence to reject the null hypothesis, we have sufficient evidence to support an alternate hypothesis $H_1 : E(y|x_0 = 0) \neq 9$ that the mean number of broken ampules when 0 transfers are made is not equal to / different from 9. Since the alternate hypothesis $H_1$ involves "$\neq$", we have sufficient evidence to reject the null hypothesis if the magnitude $|t_0|$ of the test statistic $t_0$ is greater than the quantile $t_{\alpha/2,\ df_{Res}}$. If the alternative $H_1$ were to involve "$<$" or "$>$", we would have sufficient evidence to reject the null hypothesis if the magnitude $|t_0|$ of the test statistic $t_0$ is greater than $t_{\alpha,\ df_{Res}}$. Since $|t_0| < t_{\alpha/2,\ df_{Res}}$, we fail reject the null hypothesis. We have insufficient evidence to support the alternate hypothesis.

We have sufficient evidence to reject the null hypothesis if the probability $p$ that a the test statistic for a random sample of shipments is less than $-|t_0|$ or greater than $|t_0|$, assuming the null hypothesis is true, is less than significance level $\alpha$.

```
p <- pt(1.809, degrees_of_freedom, lower.tail = FALSE) * 2
# 2 if the alternate hypothesis involves "\neq"
# 1 if the alternate hypothesis involves "<" or ">"
p
```

```
## [1] 0.1080558
```

Since $p > \alpha$, we fail to reject the null hypothesis.