Stat 6021: HW Set 3

Tom Lever

09/15/22

1. 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 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.

   (a) What is the response variable in this analysis? What is te predictor in this analysis?

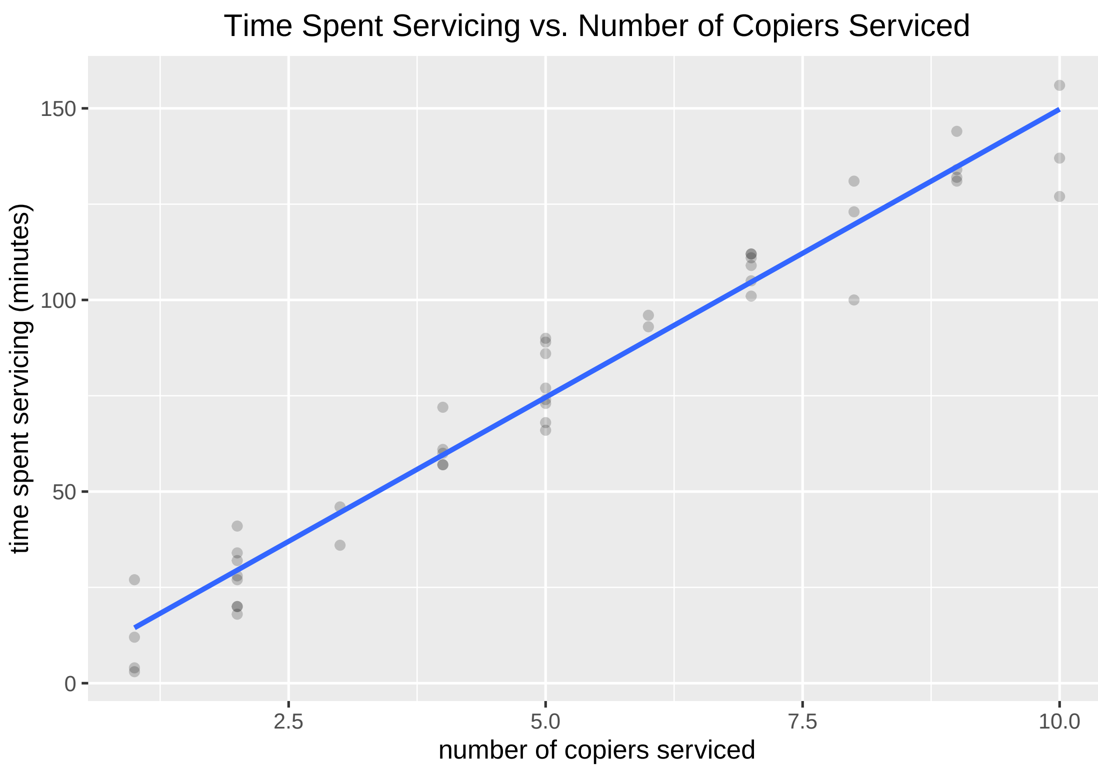
   The response variable is `Minutes`, the total number of minutes spent by the service person. The predictor is `Serviced`, the number of copiers serviced.

   (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?

```r
times_spent_servicing_and_numbers_of_copiers_serviced <-
    read.table("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
```

```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.

(c) Use the `lm()` function to fit a linear regression model for the two variables. Where are the values for $\hat{\beta}_1$, $\hat{\beta}_0$, R^2, and $\hat{\sigma}^2$ for this linear regression?

```
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
```

$\hat{\beta}_1 = 15.035 \frac{min}{1}$ is the cell value for row `Serviced` and column `Estimate` in table `Coefficients` above, and is given in the linear-regression equation.

$\hat{\beta}_0 = -0.580\ min$ is the cell value for row `(Intercept)` and column `Estimate`, and is given in the linear-regression equation.

$R^2 = 0.957$ is the value corresponding to `Adjusted R-squared`.

Errors are assumed to have mean 0 and unknown constant variance σ^2. An estimated variance is the residual mean square $\hat{\sigma}^2$. The residual standard error is $\hat{\sigma}$. The estimated value for the standard deviation of the error terms for the regression model is also $\hat{\sigma}$. $\hat{\sigma} = 8.914\ min$. $\hat{\sigma}^2 = 79.459\ min^2$ is the value corresponding to `Estimated variance of errors`.

(d) Interpret the values of $\hat{\beta}_1$ and $\hat{\beta}_0$ contextually. Does the value of $\hat{\beta}_0$ make sense in the context?

The estimated slope $15.035 \frac{min}{1}$ indicates that for every change in number of copiers serviced of 1, the predicted time of service will increase by $15.035\ min$.

A time of service cannot be negative. An estimated time of service of approximately $0\ min$ makes sense for a number of copiers serviced of 0. An estimated time of service of $-0.580\ min$ makes sense as an intercept / offset / bias that allows the estimated time of service to take on specific values for specific numbers of copiers serviced.

(e) 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?

```
analyze_variance(linear_model)
```

```
## 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
## DFT: 44, SST: 80376
## R2: 0.957499751169503
## Number of observations: 45
```

The value of the ANOVA F statistic is $F_0 = 968.66$. Note that $MS_{Res} \approx 79$. The null hypothesis for an ANOVA F test is that the slope β_1 of a linear model is equal to 0. The alternate hypothesis for an ANOVA F test is that the slope β_1 of a linear model is not equal to 0.

```
test_null_hypothesis_involving_slope(linear_model, 0.05)
```

```
## Since probability 2.2e-16
## is less than significance level 0.05,
## we reject the null hypothesis.
## We have sufficient evidence to support the alternate hypothesis.
```

2. Suppose that for $n = 6$ students, we want to predict the students' scores on a second quiz using scores from a first quiz. The estimated regression line is

$$\hat{y} = 20 + 0.8\ x$$

(a) For each individual observation (x_i, y_i), calculate the corresponding predicted score on the second quiz $\hat{y}_i$ and the residual e_i. You may show your results in the table below.

v_i	1	2	3	4	5	6
x	70	75	80	80	85	90
y	75	82	80	86	90	91
$\hat{y}$	76	80	84	84	88	92
e	-1	2	-4	2	2	-1

```
20 + 0.8 * 70
```

```
## [1] 76
```

```
20 + 0.8 * 75
```

```
## [1] 80
```

```
20 + 0.8 * 80
```

```
## [1] 84
```

```
20 + 0.8 * 85
```

```
## [1] 88
```

```
20 + 0.8 * 90
```

```
## [1] 92
```

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

```
75 - 76
```

```
## [1] -1
```

```
82 - 80
```

```
## [1] 2
```

```
80 - 84
```

```
## [1] -4
```

```
86 - 84
```

```
## [1] 2
```

```
90 - 88
```

```
## [1] 2
```

```
91 - 92
```

```
## [1] -1
```

(b) Complete the ANOVA table below for the above data set. Note that cells with * are typically left blank.

	DF	SS	MS	F0	p
Regression	1	160	160	21.333	0.0099
Residual	4	30	7.5	*	*
Total	5	190	*	*	*

```r
1
```

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

```r
6 - 2
```

```
## [1] 4
```

```r
6 - 1
```

```
## [1] 5
```

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} [y_i]$$

$$SS_R = \sum_{i=1}^{n} \left[(\hat{y}_i - \bar{y})^2 \right]$$

```r
(75 + 82 + 80 + 86 + 90 + 91) / 6
```

```
## [1] 84
```

```r
(76 - 84)^2 + (80 - 84)^2 + (84 - 84)^2 + (84 - 84)^2 + (88 - 84)^2 + (92 - 84)^2
```

```
## [1] 160
```

$$SS_T = \sum_{i=1}^{n} \left[(y_i - \bar{y})^2 \right]$$

```r
(75 - 84)^2 + (82 - 84)^2 + (80 - 84)^2 + (86 - 84)^2 + (90 - 84)^2 + (91 - 84)^2
```

```
## [1] 190
```

$$SS_{Res} = SS_T - SS_R$$

```r
190 - 160
```

```
## [1] 30
```

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

```r
160 / 1
```

```
## [1] 160
```

$$MS_{Res} = \frac{SS_{Res}}{df_{Res}}$$

```r
30 / 4
```

```
## [1] 7.5
```

$$F_0 = \frac{MS_R}{MS_{Res}}$$

$F_{\alpha,\ df_R,\ df_{res}}$: quantile such that the probability of that statistic F_0 is less than this quantile is $1 - \alpha$, and the probability of static F_0 being greater than this quantile is α

```r
160 / 7.5
```

```
## [1] 21.33333
```

```r
qf(1 - 0.05, 1, 4, lower.tail = TRUE)
```

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

```r
qf(0.05, 1, 4, lower.tail = FALSE)
```

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

p: Probability that a random statistic F is greater than statistic F_0

```r
1 - pf(160 / 7.5, 1, 4, lower.tail = TRUE)
```

```
## [1] 0.009889991
```

```r
pf(160 / 7.5, 1, 4, lower.tail = FALSE)
```

```
## [1] 0.009889991
```

(c) Calculate the sample estimate $\hat{\sigma}^2$ of the variance σ^2 for the regression model.

Errors are assumed to have mean 0 and unknown constant variance σ^2. An estimated variance is the residual mean square $\hat{\sigma}^2$. The residual standard error is $\hat{\sigma}$. The estimated value for the standard deviation of the error terms for the regression model is also $\hat{\sigma}$.

$$\hat{\sigma}^2 = MS_{Res} = \frac{SS_{Res}}{n-2}$$

```r
30 / 4
```

```
## [1] 7.5
```

(d) What is the value of R^2 here?

$$R^2 = \frac{SS_R}{SS_T}$$

```r
160 / 190
```

```
## [1] 0.8421053
```

The coefficient of determination R^2 is the proportion of the variation in a student's score on a second quiz that is explained by the linear model of a student's score on a second quiz vs. the student's score on a first quiz / the student's score on a first quiz. The correlation of determination lies between 0 and 1. Since the coefficient of determination is greater than 0.8, the linear model is precise and good for prediction.

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

Since the above statistic $F_0 = 21.333$ is greater than $F_{\alpha,\ DF_R,\ DF_{Res}} = 7.709$, and the probability $p = 0.0099$ is less than a standard significance level $\alpha = 0.05$, we reject the null hypothesis that the slope β_1 is equal to 0 for the linear model of a student's score on a second quiz vs. the student's score on a first quiz. We have sufficient evidence to conclude that the slope β_1 is not equal to 0, and that there is a linear relationship between a student's score on a second quiz and the student's score on a first quiz.

3. The least squares estimators of the simple linear regression model are

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

and

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

These are found by minimizing the sum of squared errors; i.e., by minimizing

$$SS_{Res} = \sum_{i=1}^{n} \left[(y_i - \hat{y}_i)^2\right]$$

Recall that fitted values and residuals from the fitted regression line are defined as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1\ x_i$$

and

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

The partial derivatives of SS_{Res} with respect to the coefficients of the linear model are derived as follows.

$$SS_{Res} = \sum_{i=1}^{n} \left[(y_i - \hat{y}_i)^2\right]$$

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

$$SS_{Res} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1\ x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1\ x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1\ x_n)^2$$

$$\frac{\partial \left[SS_{Res}\right]}{\partial \left[\hat{\beta}_0\right]} = 2\ (y_1 - \hat{\beta}_0 - \hat{\beta}_1\ x_1) \frac{\partial \left[y_1 - \hat{\beta}_0 - \hat{\beta}_1\ x_1\right]}{\partial \left[\hat{\beta}_0\right]} + \ldots$$

$$\frac{\partial \left[SS_{Res}\right]}{\partial \left[\hat{\beta}_0\right]} = -2\ (y_1 - \hat{\beta}_0 - \hat{\beta}_1\ x_1) - 2\ (y_2 - \hat{\beta}_0 - \hat{\beta}_1\ x_2) - \ldots - 2\ (y_n - \hat{\beta}_0 - \hat{\beta}_1\ x_n)$$

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

$$\frac{\partial \left[SS_{Res}\right]}{\partial \left[\hat{\beta}_0\right]} = -2 \sum_{i=1}^{n} \left[y_i - \hat{y}_i\right]$$

$$\frac{\partial \left[SS_{Res}\right]}{\partial \left[\hat{\beta}_1\right]} = 2\ (y_1 - \hat{\beta}_0 - \hat{\beta}_1\ x_1) \frac{\partial \left[y_1 - \hat{\beta}_0 - \hat{\beta}_1\ x_1\right]}{\partial \left[\hat{\beta}_1\right]} + \ldots$$

$$\frac{\partial \left[SS_{Res}\right]}{\partial \left[\hat{\beta}_1\right]} = -2\ (y_1 - \hat{\beta}_0 - \hat{\beta}_1\ x_1)\ x_1 - 2\ (y_2 - \hat{\beta}_0 - \hat{\beta}_1\ x_2)\ x_2 - \ldots - 2\ (y_n - \hat{\beta}_0 - \hat{\beta}_1\ x_n)\ x_n$$

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

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

Using the above equations, show that the following equalities hold. Also, give a one-sentence interpretation of what the equalities mean.

The method of least squares is used to determine $\hat{\beta}_0$ and $\hat{\beta}_1$ of a sample simple linear regression model. We estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the sum of the squares of the differences between the observations y_i and the predicted values $\hat{y}_i = \hat{\beta}_0 - \hat{\beta}_1 \ x_i$ is a minimum. The sum of the squares of the differences between the observations y_i and the predicted values $\hat{y}_i$ is a minimum if the following two conditions hold.

$$\frac{\partial \left[SS_{Res}\right]}{\partial \left[\hat{\beta}_0\right]} = -2 \sum_{i=1}^{n} \left[y_i - \hat{y}_i\right] = 0$$

$$\frac{\partial \left[SS_{Res}\right]}{\partial \left[\hat{\beta}_1\right]} = -2 \sum_{i=1}^{n} \left[(y_i - \hat{y}_i) \ x_i\right] = 0$$

Given the first equation,

$$\sum_{i=1}^{n} \left[y_i - \hat{y}_i\right] = 0$$

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

$$\sum_{i=1}^{n} \left[e_i\right] = 0$$

The sum of the residuals in any sample simple linear regression model that contains an intercept β_0 is always 0.

$$\sum_{i=1}^{n} \left[y_i - \hat{y}_i\right] = 0$$

$$y_1 - \hat{y}_1 + y_2 - \hat{y}_2 + \ldots + y_n - \hat{y}_n = 0$$

$$y_1 + y_2 + \ldots + y_n = \hat{y}_1 + \hat{y}_2 + \ldots + \hat{y}_n$$

$$\sum_{i=1}^{n} \left[y_i\right] = \sum_{i=1}^{n} \left[\hat{y}_i\right]$$

The sum of the observed values y_i equals the sum of the fitted values $\hat{y}_i$.

$$\frac{\partial \left[SS_{Res}\right]}{\partial \left[\hat{\beta}_1\right]} = -2 \sum_{i=1}^{n} \left[(y_i - \hat{y}_i) \ x_i\right] = 0$$

$$\sum_{i=1}^{n} \left[(y_i - \hat{y}_i) \ x_i\right] = 0$$

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

$$\sum_{i=1}^{n} \left[e_i \ x_i\right] = 0$$

The sum of the residuals e_i weighted by the corresponding value of the regressor / predictor variable x_i always equals 0.

$$\sum_{i=1}^{n} [e_i \ x_i] = 0$$

$$\sum_{i=1}^{n} \left[e_i \ \frac{\hat{y}_i - \hat{\beta}_0}{\hat{\beta}_1}\right] = 0$$

$$\frac{1}{\hat{\beta}_1} \sum_{i=1}^{n} \left[e_i \ \left(\hat{y}_i - \hat{\beta}_0\right)\right] = 0$$

$$\sum_{i=1}^{n} \left[e_i \ \left(\hat{y}_i - \hat{\beta}_0\right)\right] = 0$$

$$\sum_{i=1}^{n} \left[e_i \ \hat{y}_i - e_i \ \hat{\beta}_0\right] = 0$$

$$\sum_{i=1}^{n} \left[e_i \ \hat{y}_i - e_i \ \hat{\beta}_0\right] = 0$$

$$\sum_{i=1}^{n} [e_i \ \hat{y}_i] - \sum_{i=1}^{n} \left[e_i \ \hat{\beta}_0\right] = 0$$

$$\sum_{i=1}^{n} [e_i \ \hat{y}_i] - \hat{\beta}_0 \sum_{i=1}^{n} [e_i] = 0$$

$$\sum_{i=1}^{n} [e_i \ \hat{y}_i] - \hat{\beta}_0 \, (0) = 0$$

$$\sum_{i=1}^{n} [e_i \ \hat{y}_i] - 0 = 0$$

$$\sum_{i=1}^{n} [e_i \ \hat{y}_i] = 0$$

The sum of the residuals e_i weighted by the corresponding fitted value $\hat{y}_i$ always equals 0.