HW

Monday, February 17, 2025 1:59 PM

HW Module 5

Michael Puchalski

2025-02-17

1. (R required) For this question, 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 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.

```r
library(tidyverse)
```

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:

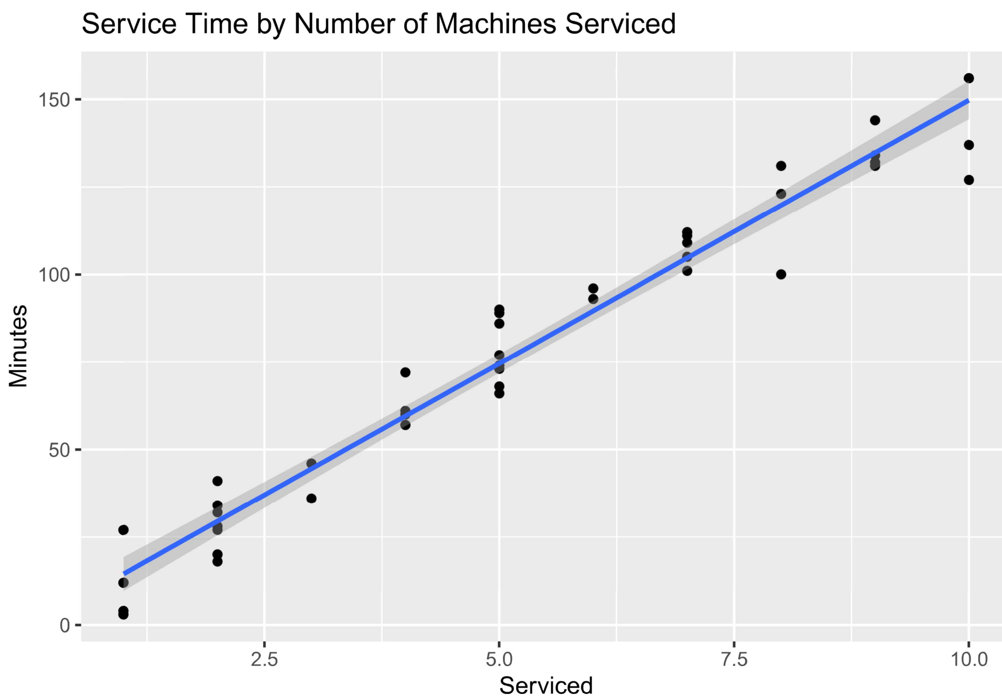
```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
data <- read.table("copier.txt", header=TRUE , sep="")

mylm <- lm(Minutes~Serviced, data=data)

ggplot(data, aes(x=Serviced,y=Minutes))+
  geom_point()+
  geom_smooth(method = "lm", se = TRUE, level=0.95)+
  labs(x="Serviced", y="Minutes", title="Service Time by Number of Machines Serviced")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Shown above is a scatterplot fitted with an appropriate linear regression fitted as well as the 95% confidence interval shaded in grey.

```r
summary(mylm)
```

(a) Obtain the 95% confidence interval for the slope, B1.

```
##
## Call:
## lm(formula = Minutes ~ Serviced, data = data)
##
## 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
```

```r
confint(mylm, level=0.95)
```

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

Here the confidence interval for the slope, indicated in the Serviced row, falls between the lower bound of 14.061 and 16.009 Minutes. This means it is expected that 95% of the time, the average time to fix a copier should be essentially between 14 and 16 minutes with the predicted mean being 15.035 minutes.

```r
new_copiers <- data.frame(Serviced=5)
predict(mylm, new_copiers, level=0.95, interval="confidence")
```

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

```
##        fit      lwr      upr
## 1 74.59608 71.91422 77.27794
```

The predicted time is 74.596 minutes, at just under an hour and 15 with a corresponding confidence interval of (71.914, 77.280)mins to be spent on servicing the 5 copiers.

(c) What is the value of the residual for the first observation? Interpret this value contextually. The value of the residual in the first observation is 8.914 on 43 degrees of freedom. This is a low residual value, meaning the extraneous variance not described by the model is low suggesting the model does a fairly good job modeling the relationship between copiers serviced and total service time.

2. (You may only use R as a simple calculator or to find p-values, critical values, or multipliers) A substance used in biological and medical research is shipped by airfreight to users in cartons of 1000 ampules. The data consist of 10 shipments. The variables are number of times the carton was transferred from one aircraft to another during the shipment route (*transfer*), and the number of ampules found to be broken upon arrival (*broken*). We want to fit a simple linear regression. 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 also provided for you, and may be used for the rest of this question: $\bar{x} = 1$, $\sum_{i=1}^{10}(x_i - \bar{x})^2 = 10$.

(a) Carry out a hypothesis test to assess if there is a linear relationship between the variables of interest.

(b) Calculate a 95% confidence interval that estimates the unknown value of the population slope.

(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 statements, calculate the test statistic, and write the corresponding conclusion in context, in response to his belief).

(d) Calculate a 95% confidence interval for the mean number of broken ampules and a 95% prediction interval for the number of broken ampules when the number of transfers is 2.

(e) What happens to the intervals from the previous part when the number of transfers is 1? (Describe what happens without calculating)

(f) What is the value of the F statistic for the ANOVA table?

(g) Calculate the value of R^2, and interpret this value in context.

(f) What is the value of the F statistic for the ANOVA table?

(g) Calculate the value of R^2, and interpret this value in context.

3. Please remember to complete the Module 1 to 4 Guided Question Set Participation Self- and Peer-Evaluation Questions. Complete via Quizzes on Canvas. Link provided in Canvas in the same place as where you found this PDF.

HW Mod 4

a) $se(\hat{\beta}_1) = 0.4690$ $\hat{\beta}_1 = 4.0000$

$se(\hat{\beta}_0) = 0.6633$ $\hat{\beta}_0 = 10.2000$

$se_{res} = 1.483$ on 8 df $\bar{x} = 1$

$H_0: \beta_1 = 0$ $H_a: \beta_1 \neq 0$ $\sum_{i=1}^{10} (x_i - \bar{x})^2 = 10$?

$$t = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{4.0000}{0.4690} = 8.5288$$

t-test vs t critical value has an issue, unsure where mistake is.

?

$$se(\hat{\beta}_1) = \frac{1.483}{\sqrt{10}} = 0.4690$$

~~[illegible]~~ $q_t(1 - 0.05/2, 8) = 2.3060$

$(1 - pt(8.5288, 8)) \cdot 2$ p-value < 0.05

p-value $\approx 2.747 \times 10^{-5}$

because p-value is less than ~~almost~~ ~~any [illegible] need~~, 0.05, we reject the null and hold that there is a linear relationship.

(b) CI Calculation for Pop true Slope

estimator ± (multiplier × se of estimator)

Meaning $100\cdot(1-\alpha)\%$ CI for β_1 is

$$\hat{\beta}_1 \pm t_{1-\alpha/2;\,n-2}\, se(\hat{\beta}_1) =$$

$$\hat{\beta}_1 \pm t_{1-\alpha/2;\,n-2}\left(\frac{s}{\sqrt{\sum(x_i-\bar{x})^2}}\right)$$

$\hat{\beta}_1 = 4.0000$ $\quad t = 2.3060$

$s = 1.483$ $\quad \sum(x_i-\bar{x})^2 = 10$

$$4.0000 + 2.3060\left(\frac{1.483}{\sqrt{10}}\right)$$

$= 5.0814$

$$4.0000 - 2.3060\left(\frac{1.483}{\sqrt{10}}\right)$$

$= 2.9186$

95% CI of mean slope is (2.9186, 5.0814)

t

(c). Claim: Broken Ampules $\neq 9$

$H_0: \mu = 9$ $\quad H_a: \mu \neq 9$

$t = \dfrac{\hat{\beta}_1 - 9}{se\,\hat{\beta}}$ $\quad \hat{\beta}_1 = 4.0000$

$t = \dfrac{4.0000 - 9}{(s)/\sqrt{10}}$ $\quad \dfrac{-5.0000}{1.483/\sqrt{10}}$

$t = -10.6618$

$p = (1 - pt(-10.6618, 8)) \cdot 2$

$= 1.99999$

≈ 2.0000

pvalue > 0.05 So we fail to reject the null hypothesis regarding the claim that the broken ampules w/ no transfer = 9. This means the Consultants Claim is wrong.

↳ Absolute val of t is also > than Critical Value

(d) 95% CI for Broken ampules
(Confidence Interval for Mean Response)

$$\hat{\mu}_{y|x_0} \pm t_{1-\alpha/2,\, n-2}\, s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

transfers = 2 → Predictor

$\hat{y} = \beta_0 + \beta_1 x$

x_0 = transfer = 2

$\hat{y} = \beta_0 + \beta_1 x$

$\hat{y} = 10.2 + 4.0000 \cdot 2$

$\hat{y} = 18.2000$

~~14.2 ± qt(1-0.05/2, 8) s~~ $\cancel{\sqrt{\frac{1}{10} + \left(\frac{0}{10}\right)}}$

18.2

~~14.2~~ ± qt(1-0.05/2, 8) 0.469 $\sqrt{\frac{1}{10} + \left(\frac{2-1}{10}\right)}$

$\Rightarrow \sqrt{\frac{1}{10} + \frac{1}{10}}$

18.2 ± 0.4837 = ~~18.6837~~

$\boxed{\text{95\% CI} = 17.7163,\ 18.6837}$

$\Rightarrow \sqrt{\frac{2}{10}}$

~~14.2~~ ± 2.3060 (0.469)(0.4472) ⇒ 0.4472

18.2 ± 0.4837 = CI for Mean

(d)

PI for ~~Mean~~ transfer = 2

$18.2000 \pm \underset{2.3060}{(0.469)} \sqrt{1 + \frac{1}{10} + \frac{1}{10}}$

$18.2000 \pm \underset{2.3060}{(0.4969)} (\sqrt{1.2})$

$18.2000 \pm (0.469)(1.095)$

$18.2000 \pm (2.3060)\, 0.5135$

18.2000 ± 1.1843

$$\boxed{\text{95\% Prediction interval of response is } [17.0157, 19.3843]}$$

(e) What happens when transfers = 1 is you have a ~~higher interval~~ smaller interval because ~~a portion~~ a portion of the entire equation is zeroed out by $x_0 - \bar{x} = 0$ in the numerator.

In other words, our Intervals get tighter / more narrow

(f) F-Statistic for the ANOVA

$F = \frac{MSR}{MSE}$ Both given in table

For own note taking purposes

$SSR = 160$

$MSR = \frac{SSR}{df_{reg}}$ $\quad \frac{160}{1}$

$SSE = 17.6$

$MSE = \frac{SSE}{df_{res}}$ $\quad \frac{17.6}{8}$

$$F = \frac{160}{2.2} \approx 72.7273$$

(g) R^2 value

$R^2 = 1 - \frac{SSE}{SST}$ $\quad SST = SSE + SSR$

$R^2 = 1 - \frac{17.6}{177.6}$ $\quad \boxed{R^2 = .9009009}$
≈ 0.9009

This value accounts for how much Variance can be attributed to the number of transfers.

Here $\approx 90.09\%$ of broken ampules variance can be attributed to the number of transfers.