Homework-2-R-Answers-Problem-3

BIOS507 Spring 2025 | Dr Lukemire | Elizabeth Nemeti Due: February 24 2025 Problem 3.

This example is adapted from "A modern approach to regression with R" by Simon Sheather. The manager of the purchasing department of a large company is interested in developing a regression model to predict the average amount of time it takes to process a given number of invoices. Data were collected over a period of 30 days. For each data point, information was collected on: • The number of invoices processed (Invoices in the dataset) • The number of hours it took to process the set of invoices (Time in the dataset)

```
data_path = "/Users/elizabethnemeti/Documents/GitHub/BIOS507-Coursework/Homeworks/Homework-2/"
data_file <- file.path(data_path, "invoices.txt")
invoices_data <- read.table("invoices.txt", header = TRUE, sep = "") # need to add sep"" so we get indi
```

predictor variable (x) -> number of invoices processed (invoices) response variable (y) -> number of hours it took to process the set of invoices (time)

1. What exploratory analyses should you do using the data? Conduct these and report your findings as well as any supporting figures.

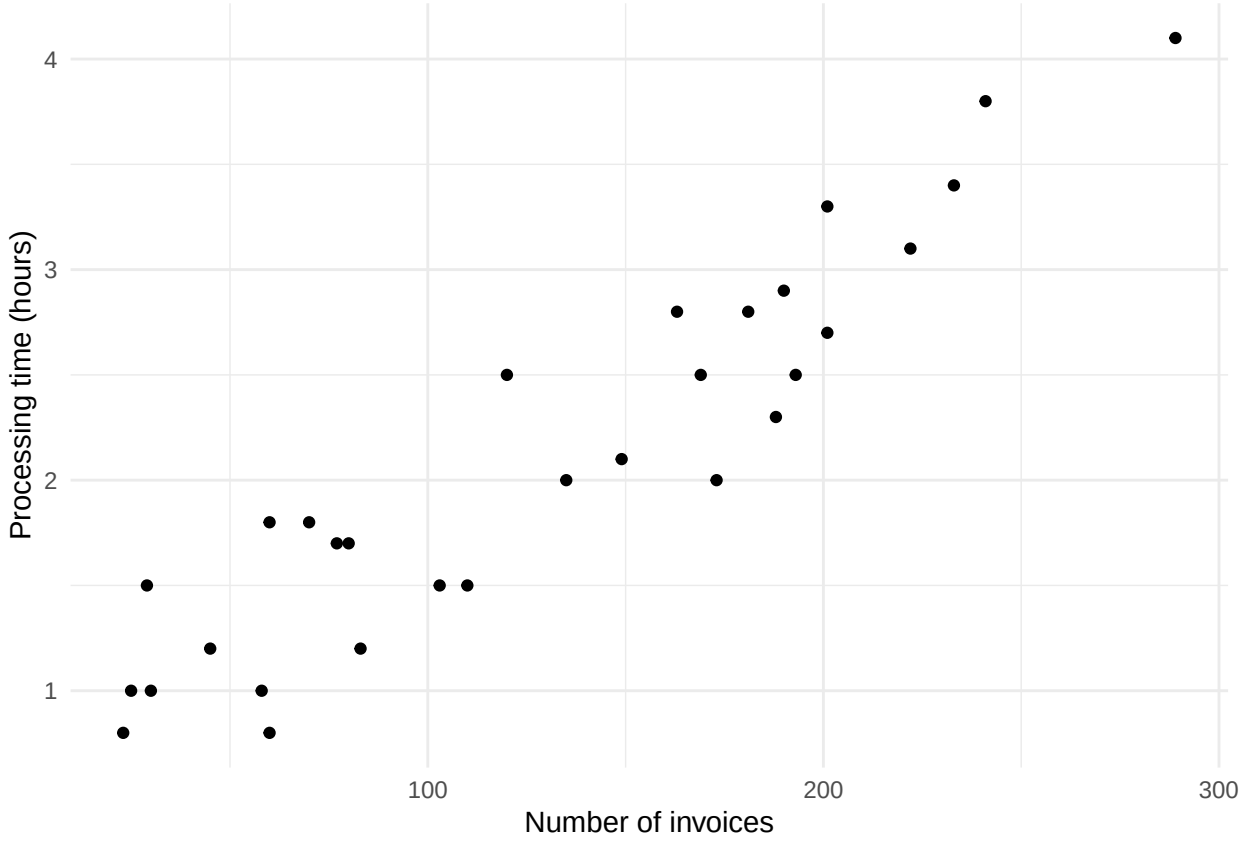
Check everything looks okay with summary():

```
summary(invoices_data)
```

```
##       Day           Invoices          Time
##  Min.   : 1.00   Min.   : 23.0   Min.   :0.800
##  1st Qu.: 8.25   1st Qu.: 62.5   1st Qu.:1.500
##  Median :15.50   Median :127.5   Median :2.000
##  Mean   :15.50   Mean   :130.0   Mean   :2.110
##  3rd Qu.:22.75   3rd Qu.:189.5   3rd Qu.:2.775
##  Max.   :30.00   Max.   :289.0   Max.   :4.100
```

Check whether a linear relationship is an appropriate function via a scatter plot:

```
invoices_data_scatterplot <- ggplot(
  invoices_data,
  aes(x = Invoices, y = Time)) +
  geom_point() +
  theme_minimal() +
  xlab("Number of invoices") +
  ylab("Processing time (hours)")
invoices_data_scatterplot
```



Does a linear relationship appear appropriate? -> Yes

2. Write out the assumed regression model for Y . What are your assumptions about the model error?

processing time (hours) = ß0 + ß1(number of invoices) +

Assumptions about : - the average error should be 0. If it wasn't, that means our model is typically overpredicting/underpredicting. - errors should look like a bell curve and follow a normal distribution - there should be homoscedasticity so the error isn't depending on x - there should not be a pattern in the errors, they need a random spread

3. Fit the model using R. Write out the estimated model.

```
invoices_model <- lm(Time ~ Invoices, data = invoices_data)
summary(invoices_model)
```

```
##
## Call:
## lm(formula = Time ~ Invoices, data = invoices_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59516 -0.27851  0.03485  0.19346  0.53083
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.6417099  0.1222707   5.248 1.41e-05 ***
## Invoices    0.0112916  0.0008184  13.797 5.17e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.3298 on 28 degrees of freedom
## Multiple R-squared:  0.8718, Adjusted R-squared:  0.8672
## F-statistic: 190.4 on 1 and 28 DF,  p-value: 5.175e-14
```

ß0 (intercept) = 1.4615 ß1 (slope) = 0.9231

Estimated model -> E[processing time (hours)] = (1.4615) + (0.9231) * (number of invoices)

4. Fill out the ANOVA table for this analysis.

```
anova(invoices_model)
```

```
## Analysis of Variance Table
##
## Response: Time
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Invoices   1 20.702 20.7020  190.36 5.175e-14 ***
## Residuals 28  3.045  0.1088
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)

Invoices 1 20.702 20.7020 190.36 5.175e-14 Residuals 28 3.045 0.1088 – – Total 29 23.747 – – –

5. Interpret the R2 value for this model.

```
# R2 = SSreg/SStotal
R2 = 20.702/23.747
R2
```

```
## [1] 0.8717733
```

R squared being 0.8717733 means 87.18% of the variation in processing time can be explained by the number of invoices processed, and it is therefore a good predictor. The remaining % is due to other factors not accounted for with this model.

6. Carry out a hypothesis test to test the null hypothesis that the slope is 0. Be sure to write out the , the null and alternative hypothesis, the test statistic, critical value, and your final decision. Interpret the result in the context of the study.

H0: 1 = 0 -> Number of invoices processed does not affect processing time (hours) HA: 1!= 0 -> Number of invoices processed affects processing time (hours) (alpha) = 0.05

Two tailed test, as we're testing if the slope is different from 0 in either direction, not just one. Therefore, (alpha)/2 = 0.025.

The test statistic is the t-statistic.

```
# t = ß1hat/standard error of ß1hat
t = 0.0112916/0.0008184
t
```

```
## [1] 13.79717
```

```
# 13.797 matches our t value (13.797) in the anova table
```

df = 28 (30-2) critical value = +-2.048 (looked it up in a t-table)

```
# to check with R
qt(0.025, df = 28, lower.tail = FALSE)
```

```
## [1] 2.048407
```

To decide on whether to reject the null hypothesis, we need to check if |t| > critical value.

|13.79717| > 2.048407, therefore we reject the null hypothesis.

In the context of the study, this means that number of invoices influences the processing time (hours) significantly. As ß1 is 0.9231, then for each additional invoice processed, processing time increases by 0.9231 hours. The p-value = 5.175e-14, meaning that this result occuring by chance is very low, and we have strong evidence to reject the null hypothesis.

7. Find and interpret a 99% confidence interval for the slope.

Since we're looking for the 99% CI, our (alpha) is 0.01, and as we're doing a two tailed test, a/2=0.005, hence to get our critical value here:

```
qt(0.005, df = 28, lower.tail = FALSE)
```

```
## [1] 2.763262
```

Now, to get our CI:

```
# ß1hat +- critical value * SE(ß1hat)
upper_bound =  0.9231 + 2.763262 * 0.0008184
upper_bound
```

```
## [1] 0.9253615
```

```
lower_bound = 0.9231 - 2.763262 * 0.0008184
lower_bound
```

```
## [1] 0.9208385
```

Our 99% CI is (0.9208385 0.9253615), meaning we are 99% confident that the true increase in processing time for each additional invoice processed, falls between 0.9208385-0.9253615 hours.

8. Find and interpret a 95% confidence interval for the amount of time it would take to process a stack of 160 invoices.

Yhat +- critical value * SE(Yhat)

```
# Yhat = ß0hat + ß1hat * Number of invoices (x)
Yhat = 1.4615 + (0.9231 * 160)
Yhat
```

```
## [1] 149.1575
```

Predicted processing time is 149.1575 hours.

Now, we need to find our confidence interval.

```
invoices_mean = mean(invoices_data$Invoices)
invoices_mean
```

```
## [1] 130.0333
```

```
SS = sum((invoices_data$Invoices - invoices_mean)^2)
SS
```

```
## [1] 162367
```

```
SE = 0.3298 * sqrt((1/30) + ((160 - invoices_mean)^2 / SS))
SE
```

```
## [1] 0.06501664
```

```r
critical_value = qt(0.025, df = 28, lower.tail = FALSE)
critical_value
```

```
## [1] 2.048407
```

```r
upper_bound = Yhat + (critical_value * SE)
upper_bound
```

```
## [1] 149.2907
```

```r
lower_bound = Yhat - (critical_value * SE)
lower_bound
```

```
## [1] 149.0243
```

With this, we are 95% confident that the true mean processing time falls between 149.0243 and 149.2907 hours when processing a stack of 160 invoices.

9. Find and interpret a 95% prediction interval for the amount of time it would take to process a new stack of 160 invoices.

Yhat +- critical value * SE(pred)

Again, first we need the SE.

```r
SE_predicted = 0.3298 * sqrt(1 + (1/30) + ((160 - invoices_mean)^2 / SS))

upper_prediction = Yhat + (critical_value * SE_predicted)
upper_prediction
```

```
## [1] 149.8461
```

```r
lower_prediction = Yhat - (critical_value * SE_predicted)
lower_prediction
```

```
## [1] 148.4689
```

The 95% prediction interval for processing a new stack of 160 invoices is (148.4689, 149.8461).

10. Which interval is wider? Why?

Again, the prediction interval is are wider, because it include uncertainty coming from estimating the mean of the processing time (hours) as well as uncertainty coming from individual observations.