Homework-2-R-Answers-Problem-2

BIOS507 Spring 2025 | Dr Lukemire | Elizabeth Nemeti Due: February 24 2025 Problem 2.

On Canvas, you have a data set called solar.txt containing data collected during a solar energy project at Georgia Tech. The data contain several columns, but for now we are going to focus on heat flux (column labeled Y) measured in kilowatts and radial deflection of the deflected rays (column labeled X4) measured in milliradians. The researchers are interested in using the radial deflection to predict the heat flux.

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

predictor variable (x4) -> radical deflection (milliradians) response variable (y) -> heat flux (kilowatts)

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

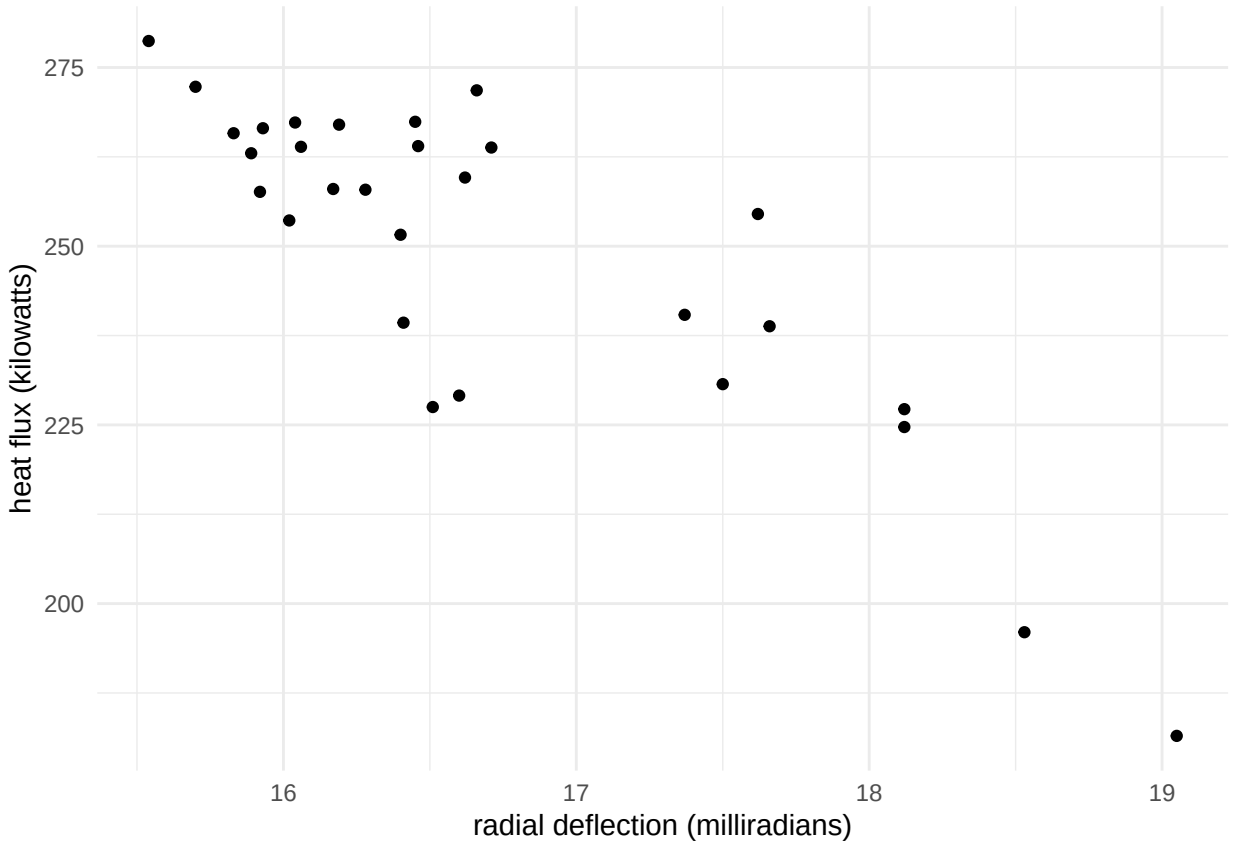
Check everything looks okay with summary():

```
summary(solar_data)
```

```
##        y               x1              x2              x3
##  Min.   :181.5   Min.   :568.5   Min.   :31.08   Min.   :31.84
##  1st Qu.:238.8   1st Qu.:704.0   1st Qu.:34.46   1st Qu.:34.14
##  Median :257.9   Median :756.0   Median :35.35   Median :35.89
##  Mean   :249.6   Mean   :754.5   Mean   :35.10   Mean   :35.53
##  3rd Qu.:265.8   3rd Qu.:801.6   3rd Qu.:35.77   3rd Qu.:36.50
##  Max.   :278.7   Max.   :909.5   Max.   :37.82   Max.   :40.55
##        x4              x5
##  Min.   :15.54   Min.   :10.53
##  1st Qu.:16.04   1st Qu.:11.41
##  Median :16.45   Median :13.10
##  Mean   :16.70   Mean   :13.23
##  3rd Qu.:17.37   3rd Qu.:14.51
##  Max.   :19.05   Max.   :16.73
```

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

```
solar_data_scatterplot <- ggplot(
  solar_data,
  aes(x = x4, y = y)) +
  geom_point() +
  theme_minimal() +
  xlab("radial deflection (milliradians)") +
  ylab("heat flux (kilowatts)")
solar_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?

heat flux (kw) = ß0 + ß1(radial deflection in milliradians) +

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 X4 - 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.

```
solar_model <- lm(y ~ x4, data = solar_data)
summary(solar_model)
```

```
##
## Call:
## lm(formula = y ~ x4, data = solar_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.2487  -4.5029   0.5202   7.9093  24.5080
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  607.103     42.906  14.150 5.24e-14 ***
## x4           -21.402      2.565  -8.343 5.94e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 12.33 on 27 degrees of freedom
## Multiple R-squared:  0.7205, Adjusted R-squared:  0.7102
## F-statistic: 69.61 on 1 and 27 DF,  p-value: 5.935e-09
```

ß0 (intercept) = 607.103 ß1 (slope) = -21.402

Estimated model -> E[heat flux (kw)] = (607.103) + (-21.402) x (radial deflection in milliradians)

4. Fill out the ANOVA table for this analysis.

```
anova(solar_model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x4         1 10578.7   10579  69.609 5.935e-09 ***
## Residuals 27  4103.2     152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

x4 1 10578.7 10579 69.609 5.935e-09 Residuals 27 4103.2 152 – – Total 28 14681.9 – – –

5. Interpret the R2 value for this model.

```
# R2 = SSreg/SStotal
R2 = 10578.7/14681.9
R2
```

```
## [1] 0.7205266
```

R squared being 0.7205266 means 72.05% of the variation in our heat flux variable can be explained by radial deflection, and its therefore a good predictor. The remaining % would have to be explained by other factors beyond radial deflection. Since it's close to 1, our model explains much of the variance.

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 -> Radial deflection does not affect heat flux HA: 1!= 0 -> Radial deflection does affect heat flux = 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 = -21.402/2.565
t
```

```
## [1] -8.34386
```

```
# -8.34386 matches our t value (-8.343) from the summary table
```

df = 27 (29-2) critical value = +-2.052 (looked it up in a t-table)

```
# to check with R
qt(0.025, df = 27, lower.tail = FALSE)
```

```
## [1] 2.051831
```

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

|-8.34386| > 2.051831, therefore we reject the null hypothesis.

In the context of the study, this means that radial deflection influences heat flux significantly. As ß1 is -21.402, then for each milliradian increase in radial deflection, we see a decrease in heat flux by 21.402 kw. The p-value = 5.94e-09, 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 is 0.01, and as we're doing a two tailed test, a(alpha)/2=0.005, hence to get our critical value here:

```
qt(0.005, df = 27, lower.tail = FALSE)
```

```
## [1] 2.770683
```

Now, to get our CI:

```
# ß1hat +- critical value * SE(ß1hat)
upper_bound = -21.402 + 2.770683*2.565
upper_bound
```

```
## [1] -14.2952
```

```
lower_bound = -21.402 - 2.770683*2.565
lower_bound
```

```
## [1] -28.5088
```

Our 99% CI is (-28.5088, -14.2952), meaning we are 99% confident that for each additional milliradian of radial deflection, heat flux decreases by between 14.2952-28.5088 kilowatts.

8. Find and interpret a 95% confidence interval for the mean heat flux when the radial deflection is 16.5 milliradians.

Yhat +- critical value * SE(Yhat)

First, let's get the predicted mean heat flux when the radial deflection is 16.5 milliradians:

```
# Yhat = ß0hat + ß1hat * X4
Yhat = 607.103 + (-21.402 * 16.5)
Yhat
```

```
## [1] 253.97
```

Now, we need to find our confidence interval around our point estimate of 253.97. To do that we first need the standard error of the predicted mean heat flux.

To get the SE we need the mean of X4 and the sum of squared deviations first.

```
x4_mean = mean(solar_data$x4)
x4_mean
```

```
## [1] 16.70207
```

```
SS = sum((solar_data$x4 - x4_mean)^2)
SS
```

```
## [1] 23.09428
```

```
SE = 12.33 * sqrt((1/29) + ((16.5 - x4_mean)^2 / SS))
SE
```

```
## [1] 2.347588
```

```
critical_value = qt(0.025, df = 27, lower.tail = FALSE)
critical_value
```

```
## [1] 2.051831
```

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

```
## [1] 258.7869
```

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

```
## [1] 249.1531
```

With this, we are 95% confident that the mean heat flux falls between 249.1531 and 258.7869 kilowatts when radial deflection is 16.5 milliradians.

9. The lab would like to predict the heat flux when the radial deflection is 16.5 milliradians for a new measurement. Give a 95% prediction interval on the kilowatts.

Yhat +- critical value * SE(pred)

Again, first we need the SE.

```
SE_predicted = 12.33 * sqrt(1 + (1/29) + ((16.5 - x4_mean)^2 / SS))

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

```
## [1] 279.7235
```

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

```
## [1] 228.2165
```

The 95% prediction interval on the kilowatts is (279.7235, 228.2165).

10. Which interval is wider? Why?

Prediction intervals are wider, because they include uncertainty coming from estimating the mean of the heat flux as well as uncertainty coming from individual observations.