Mod2\_LinearRegression

Grace Williams

2023-05-24

# Simple Linear Regression and Correlation

#### Libraries and Datasets Needed

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(tidymodels)

## ── Attaching packages ────────────────────────────────────── tidymodels 1.1.0 ──  
## ✔ broom 1.0.4 ✔ rsample 1.1.1  
## ✔ dials 1.2.0 ✔ tune 1.1.1  
## ✔ infer 1.0.4 ✔ workflows 1.1.3  
## ✔ modeldata 1.1.0 ✔ workflowsets 1.0.1  
## ✔ parsnip 1.1.0 ✔ yardstick 1.2.0  
## ✔ recipes 1.0.6   
## ── Conflicts ───────────────────────────────────────── tidymodels\_conflicts() ──  
## ✖ scales::discard() masks purrr::discard()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ recipes::fixed() masks stringr::fixed()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ yardstick::spec() masks readr::spec()  
## ✖ recipes::step() masks stats::step()  
## • Learn how to get started at https://www.tidymodels.org/start/

library(GGally)

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

library(lmtest)

## Loading required package: zoo  
##   
## Attaching package: 'zoo'  
##   
## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

air = airquality

### Question 1: How many rows are in the “air” dataframe?

##**153**

nrow(air)

## [1] 153

### Question 2: How many columns are in the “air” datagrame?

##**6**

ncol(air)

## [1] 6

### Question 3: True/False: There is missing data in “Ozone” variable in the dataframe.

##**TRUE**

summary(air$Ozone)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.00 18.00 31.50 42.13 63.25 168.00 37

is.na(air$Ozone)

## [1] FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE FALSE  
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [25] TRUE TRUE TRUE FALSE FALSE FALSE FALSE TRUE TRUE TRUE TRUE TRUE  
## [37] TRUE FALSE TRUE FALSE FALSE TRUE TRUE FALSE TRUE TRUE FALSE FALSE  
## [49] FALSE FALSE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## [61] TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  
## [73] FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE TRUE  
## [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [97] FALSE FALSE FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE TRUE FALSE  
## [109] FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE TRUE FALSE  
## [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [145] FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE

### Question 4: Which variable is the most likely to be the response (Y) variable?

### A. Ozone

### B. Solar.R

### C. Wind

### D. Temp

### E. Month

### F. Day

##**D. Temp**

### Cleaning up the data.

air2 = air %>% drop\_na()

### Question 5: How many rows in this new (air2) data frame?

##**111**

nrow(air2)

## [1] 111

ncol(air2)

## [1] 6

### Question 6: How many columns remain in this new (air2) data frame?

##**6**

### Use the “ggpairs” function to develop a visualization of the relationships in this dataset and to show correlation values for the combinations of variables. Then use the “ggcorr” function to develop a correlation matrix for the variables. Hint: Use “label = TRUE” in the “ggcorr” function to show the correlation values.

### Question 7: Which variable is the most strongly correlated with the “Ozone” variable?

### A. Solar.R

### B. Wind

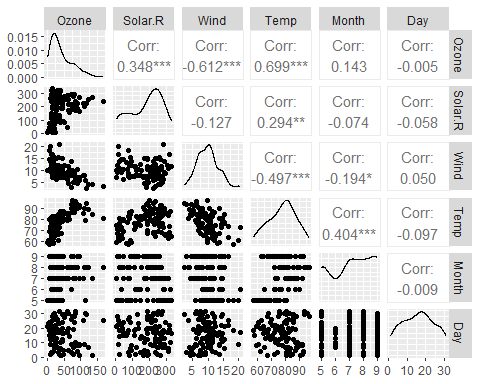
### C. Temp

### D. Month

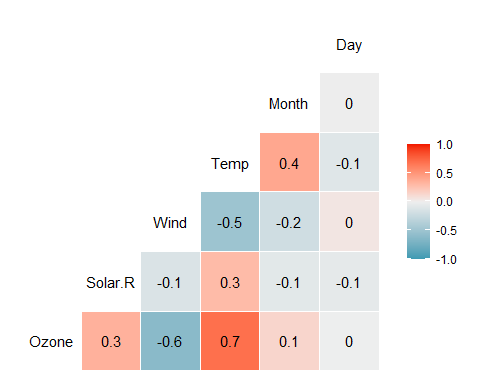
### E. Day

##**C. Temp**

ggpairs(air2)



ggcorr(air2, label = TRUE)

 ### Question 8: Which variable is least strongly correlated with the “Ozone” variable? ### A. Solar.R ### B. Wind ### C. Temp ### D. Month ### E. Day ##**E. Day**

### Question 9: Plot “Temp” (x axis) versus “Ozone (y axis) using the”ggplot” function. Choose an appropriate chart type. Which statement best describes the relationship between “Temp” and “Ozone”?

### A. As Temp increases, Ozone decreases

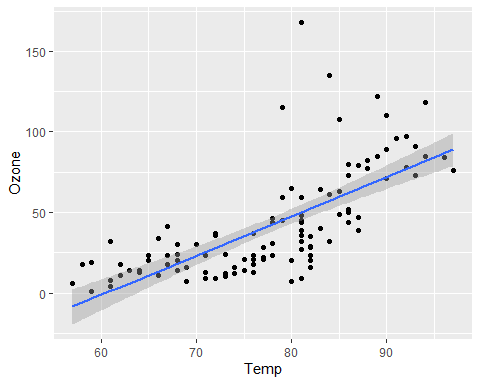
### B. As Temp increases there is no noticeable change in Ozone

### C. As Temp increases, Ozone increase

##**C. As Temp increases, Ozone increases**

ggplot(air2, aes(x = Temp, y = Ozone))+  
 geom\_point()+  
 geom\_smooth(method=lm)

## `geom\_smooth()` using formula = 'y ~ x'



### Use Tidymodels to create a linear regression model using “Temp” to predict “Ozone”. You miss wish to callyour model fit “lm\_fit”.

OT2 = recipe(Ozone ~ Temp, air2)  
OT2

##

## ── Recipe ──────────────────────────────────────────────────────────────────────

##

## ── Inputs

## Number of variables by role

## outcome: 1  
## predictor: 1

lm\_model =   
 linear\_reg() %>%  
 set\_engine("lm")  
lm\_model

## Linear Regression Model Specification (regression)  
##   
## Computational engine: lm

lm\_workflow =  
 workflow() %>%  
 add\_model(lm\_model) %>%  
 add\_recipe(OT2)  
lm\_workflow

## ══ Workflow ════════════════════════════════════════════════════════════════════  
## Preprocessor: Recipe  
## Model: linear\_reg()  
##   
## ── Preprocessor ────────────────────────────────────────────────────────────────  
## 0 Recipe Steps  
##   
## ── Model ───────────────────────────────────────────────────────────────────────  
## Linear Regression Model Specification (regression)  
##   
## Computational engine: lm

lm\_fit = fit(lm\_workflow, air2)  
  
summary(lm\_fit$fit$fit$fit)

##   
## Call:  
## stats::lm(formula = ..y ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -40.922 -17.459 -0.874 10.444 118.078   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -147.6461 18.7553 -7.872 2.76e-12 \*\*\*  
## Temp 2.4391 0.2393 10.192 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 23.92 on 109 degrees of freedom  
## Multiple R-squared: 0.488, Adjusted R-squared: 0.4833   
## F-statistic: 103.9 on 1 and 109 DF, p-value: < 2.2e-16

confint(lm\_fit$fit$fit$fit)

## 2.5 % 97.5 %  
## (Intercept) -184.818372 -110.473773  
## Temp 1.964787 2.913433

testdata = data.frame(Temp = 80)  
predict(lm\_fit, testdata)

## # A tibble: 1 × 1  
## .pred  
## <dbl>  
## 1 47.5

### Question 10: What is the slope of this regression model (to four decimal places)?

##**2.4391**

### Question 11: What is the R-squared value of this model (not Adjusted R-squared) (to three decimal places)?

##**0.488**

### Question 12: Is the “Temp” variables significant in the model?

##**Yes, Temp is significant**

### Use the code below to generate 95% confidence intervals for the coefficients. Note that you may need to change “lm\_fit” to the name of your model fit if you used a different name.

#### confint(lm\_fitfit$fit)

### Question 13: True/False: A 95% confidence interval for the slope coefficient does not contain zero.

##**TRUE**

### Question 14: Using your linear regression model with “Temp” to predict “Ozone”, what is the predicted “Ozone” value when “Temp” is equal to 80 (to two decimal places)?

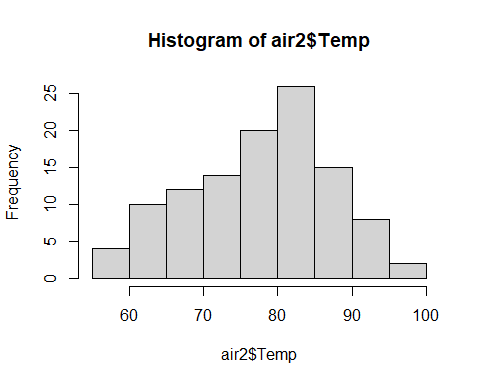
##**47.5**

### Perform appropriate model diagnostics to verify whether or not the model appears to meet the four linear regression model assumptions.

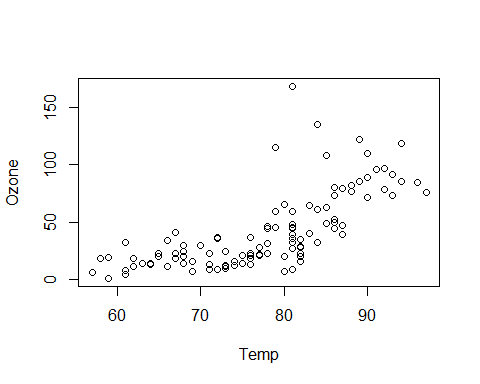
### Question 15: True/False: There is no evidence of non-independent (autocorrelated) residuals.

##**True**

hist(air2$Temp)



plot(Ozone ~ Temp, air2)



cor(air2$Ozone, air2$Temp)

## [1] 0.6985414