Lab Week 7

STAT5003

Dr. Justin Wishart

Semester 2, 2022

Contents

Preparation and assumed knowledge

- Viewed the missing data content in Module 7.
- Downloaded the `skin-cancer.csv` file available on Canvas.

Aims

- Investigate the impact of missing data on modelling.

This week, a simple dataset will be explored with and without missing data and its effect on modelling.

1 Simple regression with and without missingness

Consider this older data set that shows the mortality rates (number of deaths per 10 million people) skin cancer for some US states against the latitude of the state. The data is available in `skin-cancer.csv` on Canvas.

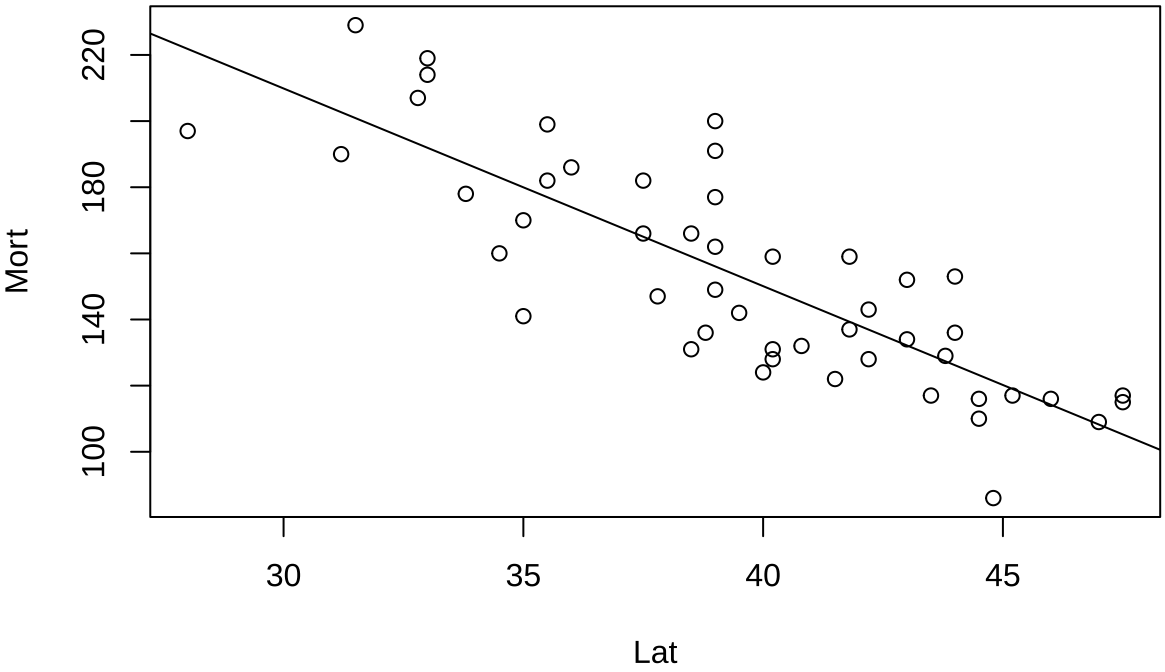
```r
full.data <- read.csv("skin-cancer.csv", header = TRUE)
```

1.1 Standard Simple regression without missingness

Conduct a simple linear regression with the Mortality rates as the response and Latitude as the predictor. Plot the data along with the regression and interpret the regression output.

Solution

```r
full.lm <- lm(Mort ~ Lat, data = full.data)
plot(Mort ~ Lat, data = full.data)
abline(full.lm)
```



```
summary(full.lm)
```

```
##
## Call:
## lm(formula = Mort ~ Lat, data = full.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.972 -13.185   0.972  12.006  43.938
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 389.1894    23.8123   16.34  < 2e-16 ***
## Lat          -5.9776     0.5984   -9.99 3.31e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.12 on 47 degrees of freedom
## Multiple R-squared:  0.6798, Adjusted R-squared:  0.673
## F-statistic:  99.8 on 1 and 47 DF,  p-value: 3.309e-13
```

1.2 Simulate missingness in the latitudes

Simulate half the values in the Latitude feature to be missing at random values. (i.e. replace half the elements with missing `NA` values). Note the assignment of missing values to be is random but not completely at random. That is, the chance that Latitude is missing should depend on Mortality. A good convenient function for this purpose is available in `mice::ampute`. The return object of this function will contain the data with missingness in the element `amp`. E.g. if you assigned the output of the `mice::ampute` to an object called `output`, the data with missing values will be available in `output$amp`. We'll call this the amputed data.

Solution

```
set.seed(5003)
library(mice)
```

```
available.data = ampute(full.data,
                        prop = 0.5,
                        patterns = data.frame(Mort = 1, Lat = 0))
```

1.3 Simple regression on the missing data as the response

Conduct a simple linear regression now on the *amputed* data but instead regressing Latitude on Mortality. That is, do a regression where Latitude is the response and Mortality is the predictor but using the amputed data where some Latitude cases are missing. (Note that `R` will automatically remove missing values from the model when using `lm`)

Solution

```
simple.missing <- lm(Lat ~ Mort, data = available.data$amp)
summary(simple.missing)
```

```
##
## Call:
## lm(formula = Lat ~ Mort, data = available.data$amp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8554 -1.6638  0.2069  1.6368  4.3635
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 55.17688    3.09569  17.824  3.7e-14 ***
## Mort        -0.10157    0.02132  -4.764 0.000105 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.543 on 21 degrees of freedom
##   (26 observations deleted due to missingness)
## Multiple R-squared:  0.5194, Adjusted R-squared:  0.4965
## F-statistic: 22.69 on 1 and 21 DF,  p-value: 0.0001049
```

1.4 Original with missingness complete-case regression

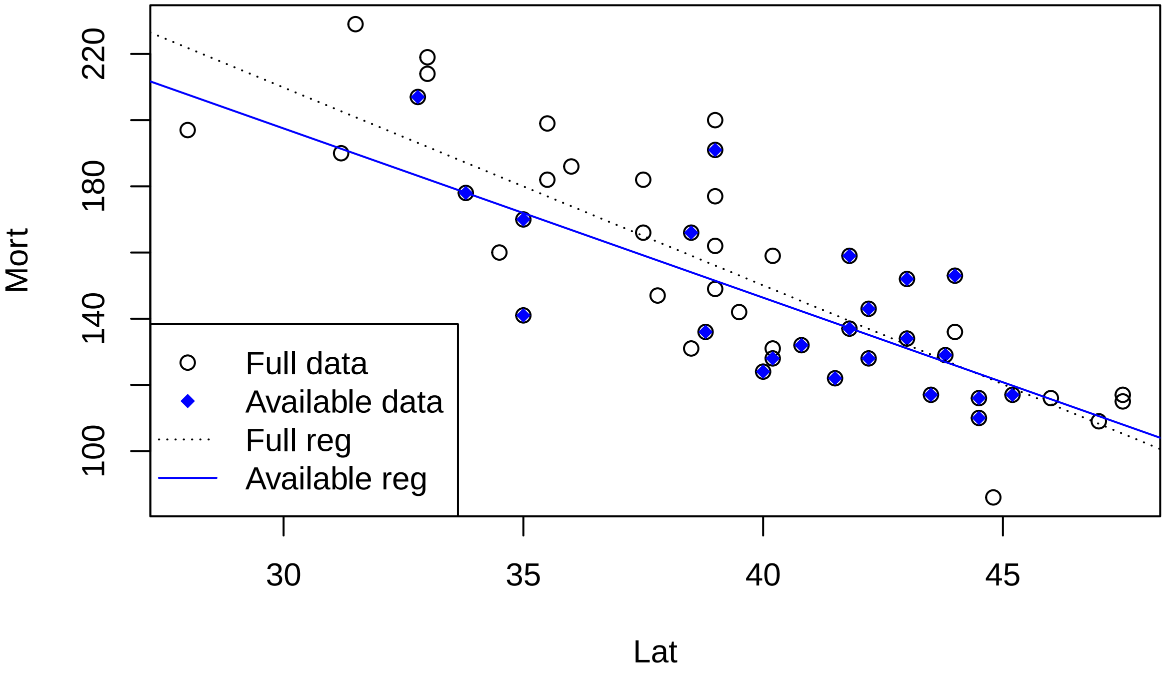
Conduct a complete-case regression on the amputed data with Mortality as the response and Latitude as the predictor and show that it is not consistent with the original regression using the full data.

Solution

```
avaiable.lm <- lm(Mort ~ Lat, data = available.data$amp)
plot(Mort ~ Lat, data = full.data, col = "black")
points(Mort ~ Lat, data = available.data$amp, pch = 18, col = "blue")
abline(full.lm, lty = 'dotted', col = "black")
abline(avaiable.lm, col = "blue")
summary(avaiable.lm)
```

```
##
## Call:
## lm(formula = Mort ~ Lat, data = available.data$amp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.923 -12.402  -1.923   9.934  39.530
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  350.893     43.794   8.012 8.03e-08 ***
## Lat           -5.113      1.073  -4.764 0.000105 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.05 on 21 degrees of freedom
##   (26 observations deleted due to missingness)
## Multiple R-squared:  0.5194, Adjusted R-squared:  0.4965
## F-statistic: 22.69 on 1 and 21 DF,  p-value: 0.0001049
```

```
legend("bottomleft", legend = c("Full data", "Available data", "Full reg", "Available reg"), col = rep(
```



The differences are slight but can see that the intercept and slope estimates are off with a smaller intercept on the model computed with missing data and the slope is also less steep at -5 instead of -5.9.

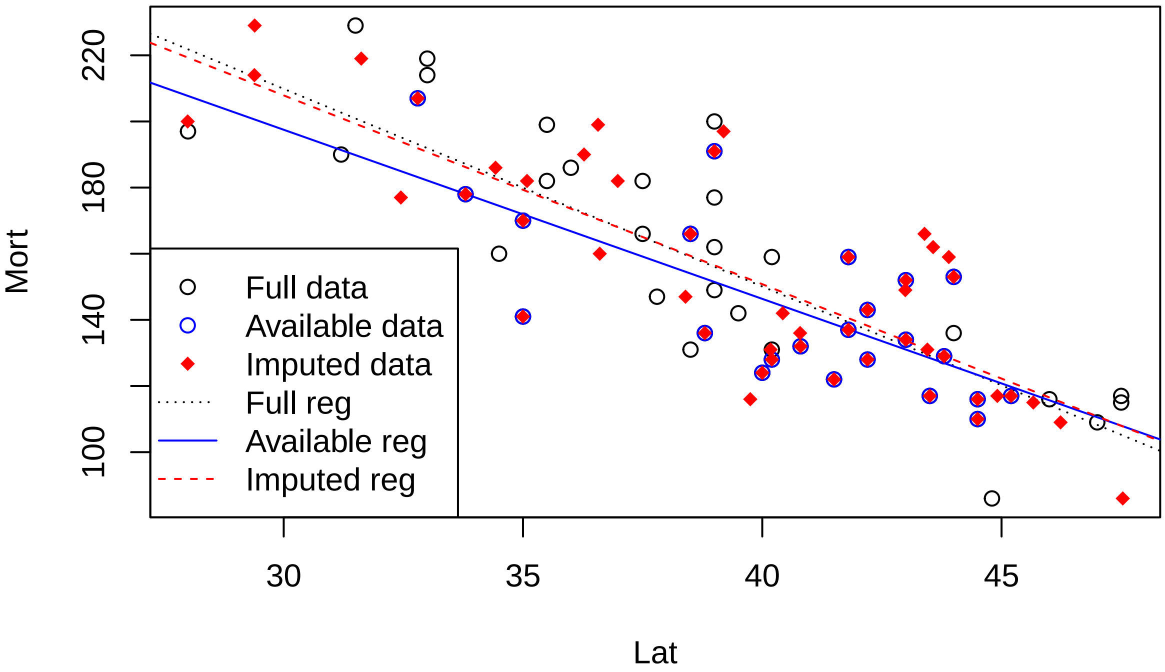
1.5 Random imputation

Randomly impute the missing Latitude observations. That is, predict the missing Latitudes using the output of section 1.3 using random imputation. That is, do the prediction from the linear model in 1.3 but randomly simluate some noise variables using the estimated variability from 1.3 too. Use these random imputed values to fill in the missing values in the amputed data. Call this new data.frame the imputed data. Then conduct the regression with Mortality as the response and Latitude as the predictor on the imputed dataset and compare your results with regression that used the the original complete data with no missingness. Compare the results of the imputed model to the full model without missingness.

Solution

```
missing.lats <- available.data$amp %>%
  filter(is.na(Lat)) %>%
  select(Mort)
```

```r
simple.preds <- predict(simple.missing, newdata = missing.lats)
rand.preds <- simple.preds + rnorm(length(simple.preds), sd = sigma(simple.missing))
imputed.data <- available.data$amp
imputed.data[["Lat"]][is.na(imputed.data[["Lat"]])] <- rand.preds
imputed.lm <- lm(Mort ~ Lat, data = imputed.data)
full.range <- lapply(full.data, range)
imputed.range <- lapply(imputed.data, range)
yrange <- range(c(full.range$Mort, imputed.range$Mort))
xrange <- range(c(full.range$Lat, imputed.range$Lat))
plot(Mort ~ Lat, data = full.data, ylim = yrange, xlim = xrange)
abline(full.lm, lty = "dotted")
points(Mort ~ Lat, data = available.data$amp, col = "blue", pch = 21)
abline(avaiable.lm, col = "blue")
points(Mort ~ Lat, data = imputed.data, col = "red", pch = 18)
abline(imputed.lm, lty = "dashed", col = "red")
legend("bottomleft", legend = c("Full data", "Available data", "Imputed data", "Full reg", "Available r
       col = rep(c("black", "blue", "red"), 2),
       pch = c(21, 21, 18, rep(NA, 3)),
       lty = c(rep(NA, 3), c("dotted", "solid", "dashed")))
```



```r
summary(imputed.lm)
```

```
##
## Call:
## lm(formula = Mort ~ Lat, data = imputed.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.329 -13.807  -3.511  15.276  41.598
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 379.1282    23.4159  16.191  < 2e-16 ***
## Lat          -5.7085     0.5866  -9.731 7.66e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.46 on 47 degrees of freedom
## Multiple R-squared:  0.6683, Adjusted R-squared:  0.6612
## F-statistic: 94.69 on 1 and 47 DF,  p-value: 7.662e-13
```

The data imputation has corrected the bias that was found in the regression estimates before and resulted in estimates that are much closer to the estimates on the full data as evidenced in the plot where the lines are very close.