library(palmerpenguins)
library(moderndive)

## Full Model – ALL Explanatory Variables Included

lm(body\_mass\_g ~ . , data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.877 0.873 79631. 282. 287. 255. 0 9 333

|  |  |
| --- | --- |
|  | **Starting adjusted** $R^{2}$Note the adjusted $R^{2}$ value you are starting with, as this will influence your decisions. |

## Variable Selection

Now, starting with our full model, we will use backwards selection to decide what variable(s) should be removed from the model.

You can only delete a variable if it **increases** adjusted $R^{2}$.

|  |  |
| --- | --- |
|  | Note we are not saying how much adjusted $R^{2}$ needs to be increased, simply that it must be bigger. |

## Models Deleting One Explanatory Variable

lm(body\_mass\_g ~ . -year, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.875 0.872 80659. 284. 288. 284. 0 8 333

lm(body\_mass\_g ~ . -sex, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.853 0.849 94951. 308. 312. 235. 0 8 333

lm(body\_mass\_g ~ . -flipper\_length\_mm, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.863 0.86 88256. 297. 301. 256. 0 8 333

lm(body\_mass\_g ~ . -bill\_depth\_mm, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.873 0.87 81908. 286. 290. 279. 0 8 333

lm(body\_mass\_g ~ . -bill\_length\_mm, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.874 0.871 81384. 285. 289. 281. 0 8 333

lm(body\_mass\_g ~ . -island, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.876 0.874 79869. 283. 286. 329. 0 7 333

lm(body\_mass\_g ~ . -species, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.847 0.843 99062. 315. 319. 257. 0 7 333

|  |  |
| --- | --- |
|  | **Which variable should be deleted?**Based on the adjusted $R^{2}$ values, which variable should be deleted from the model.*Remember:* You are looking for the model that has a higher adjusted $R^{2}$ than what you started with! |

## Models Deleting Two Explanatory Variables

lm(body\_mass\_g ~ . - island - year, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.875 0.873 80828. 284. 287. 380. 0 6 333

lm(body\_mass\_g ~ . - island - sex, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.852 0.849 95580. 309. 312. 313. 0 6 333

lm(body\_mass\_g ~ . - island - flipper\_length\_mm, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.863 0.861 88278. 297. 300. 344. 0 6 333

lm(body\_mass\_g ~ . - island - bill\_depth\_mm, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.873 0.871 82154. 287. 290. 373. 0 6 333

lm(body\_mass\_g ~ . - island - bill\_length\_mm, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.874 0.871 81636. 286. 289. 376. 0 6 333

lm(body\_mass\_g ~ . - island - species, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.831 0.829 108979. 330. 333. 323. 0 5 333

## Models Deleting Three Explanatory Variables

lm(body\_mass\_g ~ . - island - bill\_depth\_mm - year, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.871 0.869 83702. 289. 292. 440. 0 5 333

lm(body\_mass\_g ~ . - island - bill\_depth\_mm - sex, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.832 0.829 108823. 330. 333. 323. 0 5 333

lm(body\_mass\_g ~ . - island - bill\_depth\_mm - flipper\_length\_mm, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.856 0.853 93353. 306. 308. 387. 0 5 333

lm(body\_mass\_g ~ . - island - bill\_depth\_mm - species, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.818 0.815 117911. 343. 346. 368. 0 4 333

lm(body\_mass\_g ~ . - island - bill\_length\_mm - bill\_depth\_mm, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.869 0.867 84623. 291. 294. 434. 0 5 333

## Models Deleting Four Explanatory Variables

lm(body\_mass\_g ~ . - island - bill\_depth\_mm - year - sex, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.824 0.822 113574. 337. 340. 385. 0 4 333

lm(body\_mass\_g ~ . - island - bill\_depth\_mm - year - flipper\_length\_mm, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.856 0.854 93385. 306. 308. 486. 0 4 333

lm(body\_mass\_g ~ . - island - bill\_depth\_mm - year - bill\_length\_mm, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.867 0.865 86047. 293. 296. 534. 0 4 333

lm(body\_mass\_g ~ . - island - bill\_depth\_mm - year - species, data = penguins) %>%
 get\_regression\_summaries()

# A tibble: 1 × 9
 r\_squared adj\_r\_squared mse rmse sigma statistic p\_value df nobs
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 0.807 0.805 125076. 354. 356. 457. 0 3 333

|  |  |
| --- | --- |
|  | **Which variable should be deleted?**Based on the adjusted $R^{2}$ values, which variable should be deleted from the model.*Remember:* You are looking for the model that has a higher adjusted $R^{2}$ than what you started with! |

## New Rules

Now, choose the simplest model that is within 1% of the best adjusted $R^{2}$ you obtained.

|  |  |
| --- | --- |
|  | Note, when I say “simplest model” I mean the model with the fewest variables included. |

|  |  |
| --- | --- |
|  | **What is your best model?**What variables are included in the final model you chose? |