Lab 5 – Regression Solutions

2024-09-25

```
library(ggplot2)
library(dplyr)

## Less uggo plots
theme_set(theme_bw())

## Copy this to turn off scientific notation (also uggo)
options(scipen = 9999)

## College data
college <- read.csv("https://collinn.github.io/data/college2019.csv")
```

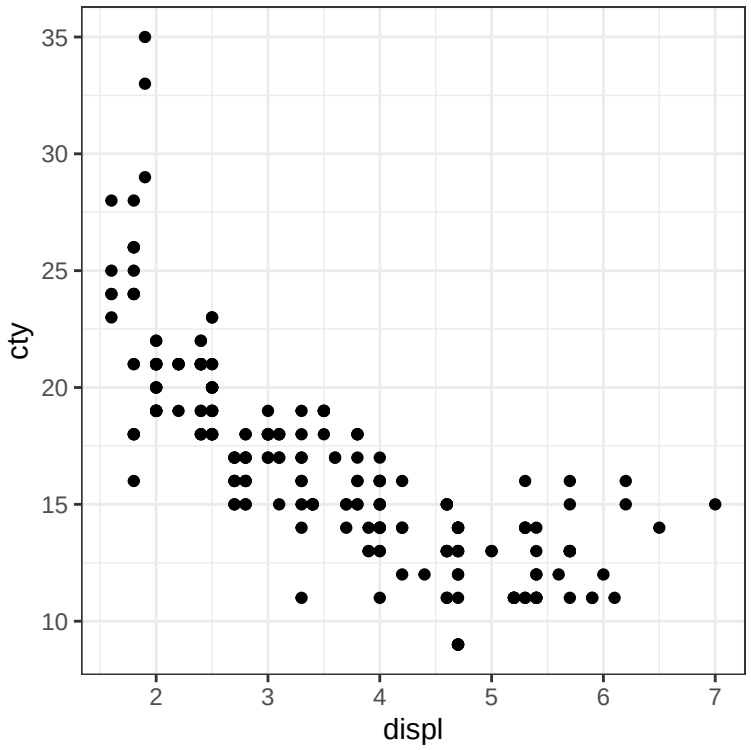
Question 0 Recreate the following expression in latex:

$$\overline{\beta} = x^2 - \hat{9}_9$$

Latex expression:

```
\overline{\beta} = x^2 - \hat{9}_9
```

Question 1: This problem uses the `mpg` dataset with the variables `displ` and `cty`:



1. Consider the scatterplot shown above. Is it more appropriate to use Pearson's or Spearman's rank correlation? Explain
2. Find and report both Pearson's and Spearman's correlation.

```
## More appropriate to use spearman
with(mpg, cor(cty, displ, method = "pearson"))
```

```
## [1] -0.79852
```

```
with(mpg, cor(cty, displ, method = "spearman"))
```

```
## [1] -0.8809
```

Question 2: This question uses the `college` dataset from above

1. Fit a simple linear regression model that predicts `FourYearComp_Males` using the variable `FourYearComp_Females`.
2. Using the fitted coefficients, write out the *fitted regression equation* for your model (you do not need to use hat notation, you can just write text)
3. Next fit a linear model and provide a one-sentence interpretation of the slope coefficient in this model: how does the slope describe the relationship between these variables?
4. Is the intercept meaningful here? Explain.

```
fit <- lm(FourYearComp_Males ~ FourYearComp_Females, college)
coef(fit)
```

```
##          (Intercept) FourYearComp_Females
##            -0.053514             0.966520
```

$$\widehat{\text{Male Completion}} = -0.053 + 0.966 \times \text{Female Completion}$$

Slope – every additional percent in female graduation is 0.96 percent increase in male graduation

Intercept – Not meaningful as it refers to instance in which school has 0% female graduation rate

---

Question 3: This question involves a penguins dataset, including measurements for 344 penguins including their species, size, and sex.

```
penguins <- read.csv("https://collinn.github.io/data/penguins.csv")

## Remove missing values
penguins <- filter(penguins, sex %in% c("male", "female"))
```

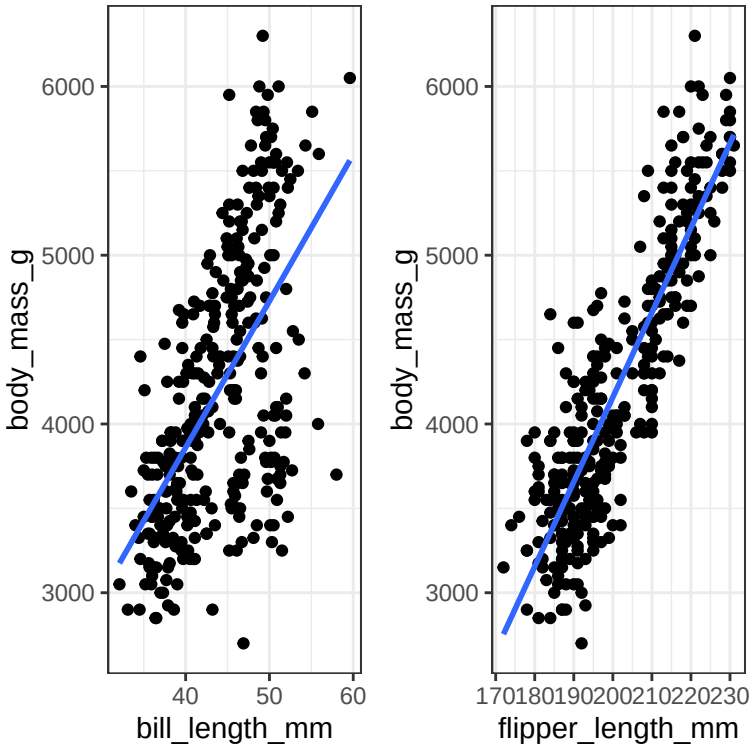
Part A Using the data above, create two different ggplots, both having body mass as the outcome variable, but with plot one using bill length as the predictor, the other using flipper length. For each plot, also include a visual representation of the linear model. Based on these, which predictor seems to have the least amount of unexplained variability? (Note: there are some missing values so you will get warnings – this is totally fine)

Part B: Create linear models based on each of the plots above. From the summaries, what is the R^2 value for each model? Is this consistent with what you found in Part A? Explain

```
p1 <- ggplot(penguins, aes(bill_length_mm, body_mass_g)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE)

p2 <- ggplot(penguins, aes(flipper_length_mm, body_mass_g)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE)

gridExtra::grid.arrange(p1, p2, nrow = 1)
```



```r
(lm(body_mass_g ~ bill_length_mm, penguins) %>% summary())["r.squared"]
```

```
## $r.squared
## [1] 0.34745
```

```r
(lm(body_mass_g ~ flipper_length_mm, penguins) %>% summary())["r.squared"]
```

```
## $r.squared
## [1] 0.76209
```

Looks like using flipper is better by both plot and R^2

Question 4: First, with the `mpg` dataset, use the appropriate dplyr functions to find the average city mpg (`cty`) for each of the categories in `drv`. Then use the `lm()` function to fit a linear model, again with `cty` as the outcome variable and with `drv` as the predictor. What becomes the *reference* variable in this model? Which drive-train appears to have the best city mpg? Which has the worst?

```r
lm(cty ~ drv, mpg)
```

```
##
## Call:
## lm(formula = cty ~ drv, data = mpg)
##
## Coefficients:
## (Intercept)         drvf         drvr
##       14.33         5.64        -0.25
```

```r
# 4wd is 14.33 mpg (reference var)
# fwd is 14.33 + 5.64 = 19.97 mpg (best)
# rwd is 14.33 - 0.25 = 14.08 mpg (worst)
```

Question 5 For this problem, we will again by using the `penguin` dataset included above.

- **Part A** Create a linear model using the bill length (mm) as our response variable and including both sex and flipper length (mm) as explanatory variables. Interpret the intercept and slope. Is the intercept meaningful?
- **Part B** Recreate the model from Part A, this time *also* including species as an explanatory variable. Interpret the intercept in this model and explain why it has changed.
- **Part C** Using your dplyr functions, find the average flipper length for each of the species in the dataset.
- **Part D** Subset your penguin data to *only* include Chinstrap and Gentoo penguins. Using your subset data, recreate the same model from Part B. Why has the indicator variable for Gentoo changed so drastically?
- **Part E** Using your linear model from Part D, find the predicted bill length (mm) of a male Gentoo penguin with a flipper length of 208mm

```r
## Part A, reference var is female. Length of female bill when flipper is 0
# not meaningful. 0.235 tells us bill length increase by that for every mm increase in flipper
lm(bill_length_mm ~ sex + flipper_length_mm, penguins) %>% coef()
```

```
##       (Intercept)           sexmale flipper_length_mm
##          -4.46691           2.07271           0.23593
```

```r
## Part B - Intercept now female and Adelie species, also when flipper is 0
# changed because reference category includes two separate things
lm(bill_length_mm ~ species + sex + flipper_length_mm, penguins) %>% coef()
```

```
##       (Intercept)  speciesChinstrap     speciesGentoo           sexmale
##           18.1785            9.4338            5.9740            3.0041
## flipper_length_mm
##            0.1007
```

```r
## Part C
dd <- filter(penguins, species %in% c("Gentoo", "Chinstrap"))
group_by(dd, species) %>% summarize(mean(bill_length_mm, na.rm = TRUE))
```

```
## # A tibble: 2 x 2
##   species   `mean(bill_length_mm, na.rm = TRUE)`
##   <chr>                                   <dbl>
## 1 Chinstrap                                48.8
## 2 Gentoo                                   47.6
```

```r
## Part D - Intercept has new reference category that is larger than gentoo, so negative
lm(bill_length_mm ~ species + sex + flipper_length_mm, dd) %>% coef()
```

```
##       (Intercept)     speciesGentoo           sexmale flipper_length_mm
##           20.4901           -4.2407            2.9529            0.1372
```

```r
## Part E -- predict
df <- data.frame(species = "Gentoo", sex = "male",
                 flipper_length_mm = 208)

fit <- lm(bill_length_mm ~ species + sex + flipper_length_mm, dd)
predict(fit, df) # guess is 47.74 mm long
```

```
##     1
## 47.74
```