Homework_3

Emmenta Janneh

2024-02-04

Conceptual Exercises

1. **FALSE**, $E[Yi|Xi] = \theta 0 + \theta 1Xi$ The equation does not explicitly include the noise affecting Yi. $E[Yi|Xi]$, focus on the systematic or average relationship between Yi and Xi, abstracting from the specific realization of the random error term for a given observation. However, the error term is an inheret part of the regression model and is typically included in the broader regression equation: $Yi = \beta 0 + \beta 1Xi + \epsilon i$

2. The key point in the fact that the error terms (ϵi *and* ηi) capture different sources of variance. In the regression of Y on X, the error term (ϵi) accounts for unobserved factors affecting Y, while in the regression of X on Y, the error term(ηi) accounts for unobserved factors affecting X. The asymmetry in the roles of dependent and independent variables contributes to the non-equivalence of the OLS estimates. This is grounded in the idea that the relationship between Y and X may not be symmetric or bidirectional; changes in X may have a different impact on Y compare to changes in Y influencing X.

3. From the graph, Constant variance appears to be violated by the regression.

4. No. The observation of a curved relationship between age and salary, with a peak around 51 years, does not inherently imply a universal trend of salaries increasing until around 51 and then decreasing. The curvature could be influenced by various factors, including industry-specific patterns, generational differences, sample-specific factors, and nonlinear effects of age, which need further analysis for a comprehensive understanding.

University Admissions Data

```
library(tidyverse)
```

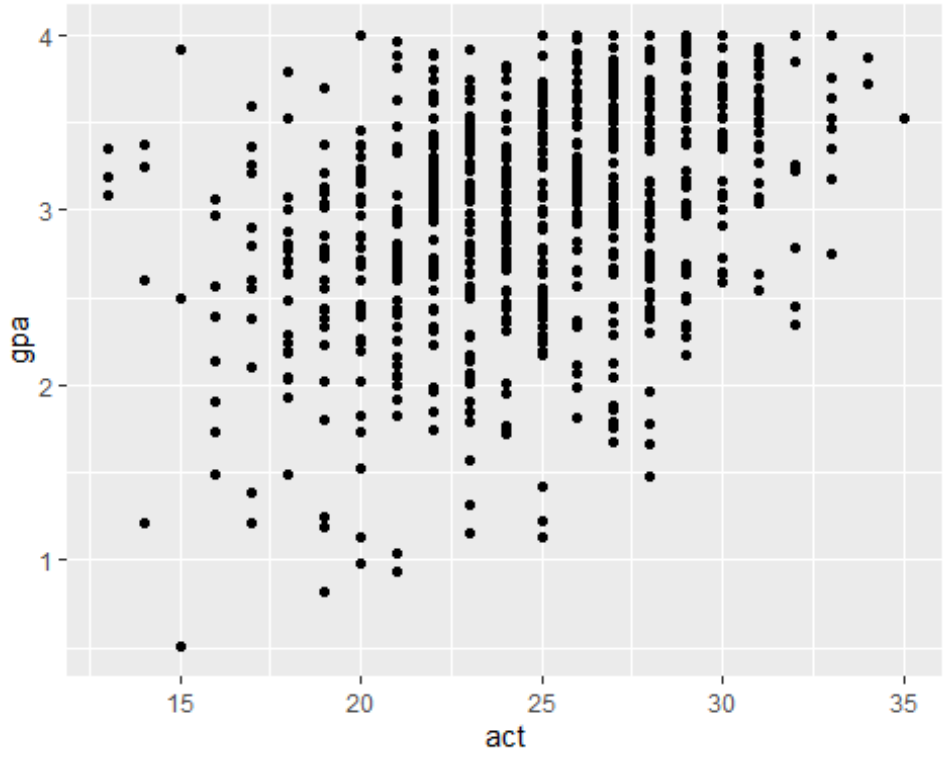
```
## Warning: package 'tidyverse' was built under R version 4.3.2
```

```
## ── 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   3.4.3     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── 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(broom)

## Warning: package 'broom' was built under R version 4.3.2

university <-
read.csv("https://dcgerard.github.io/stat_415_615/data/university.csv")
```

1.

```
ggplot(university, aes(x = act, y = gpa)) +
  geom_point()
```



There is a positive association between ACT score and GPA score.

2. No, a simple linear regression model does not appear appropriately. There seem to be a violation on Constant Variance.

```
lm_university <- lm(gpa ~ act, data = university)
lm_university

##
## Call:
## lm(formula = gpa ~ act, data = university)
##
```

```
## Coefficients:
## (Intercept)          act
##      1.5587       0.0578
```

gpa = 1.5587 + 0.0578(act)

4. **1.5587** is the gpa score of students when act score is 0. For every unit score in act, gpa is higher by **0.0578**

```
newdf <- data.frame(act = c(31, 18, 5, 29))

predict(object = lm_university, newdata = newdf)

##        1        2        3        4
## 3.350518 2.599111 1.847705 3.234917
```

6. Advantages of this approach include simplicity and interpretability, visual comparison and ease of communication.

   The disadvantages with such approach include variability overlooked, samples size impact and potential nonlinear relationships.

   In this scenario, the choice depends on the audience and the specific objectives of the analysis. If simplicity and a broad overview of the relationship between ACT scores and GPAs are paramount, the red triangle point estimates can be effective. However, for a more comprehensive understanding, especially if variability and potential nonlinear relationships are of interest, incorporating additional statistical measures such as confidence intervals or considering advanced modeling techniques like regression analysis may be preferred. A balanced approach that considers both simplicity and statistical rigor would provide a more nuanced interpretation of the data.