hw_03

lisa liubovich

2024-02-08

Conceptual Exercises

1.

True. This is the linearity assumption of the simple linear regression model, where it assumed that the relationship between the predictor and response variables can be described by a linear relationship with some random error. If the above equation were to be simplified, its simplified form is B0 +B1Xi, where B0 is the y-intercept and B1 is the slope.

2.

One explanation for why the OLS estimates of a regression of Y on X are not generally the same as OLS estimates of a regression of X on Y is the direction of causality. The directionality of causality might not be the same; that is, changes in X may not affect Y in the same way changes in Y affect X. For example, weight of a car may affect its miles per gallon, but a car's miles per gallon may not affect its weight.

3.

In the plot, the assumption of the linear regression model that seems to be violated is constant variance; the funnel shape of the points indicates heteroskedacity, or varying variances within the distribution of points.

4.

This does not imply that as people age, their salaries will increase until they are about 51, then start to decrease. There are multiple reasons why this could occur, such as market demands, career trajectories retirement plans, and other omitted variables that lead to variation in salary after the age of 51.

Source for help on this section: chatpgt (https://chat.openai.com/c/dc5ff8d3-b601-4621-abe9-e88d109ac2e5)

University Admissions Data

1.

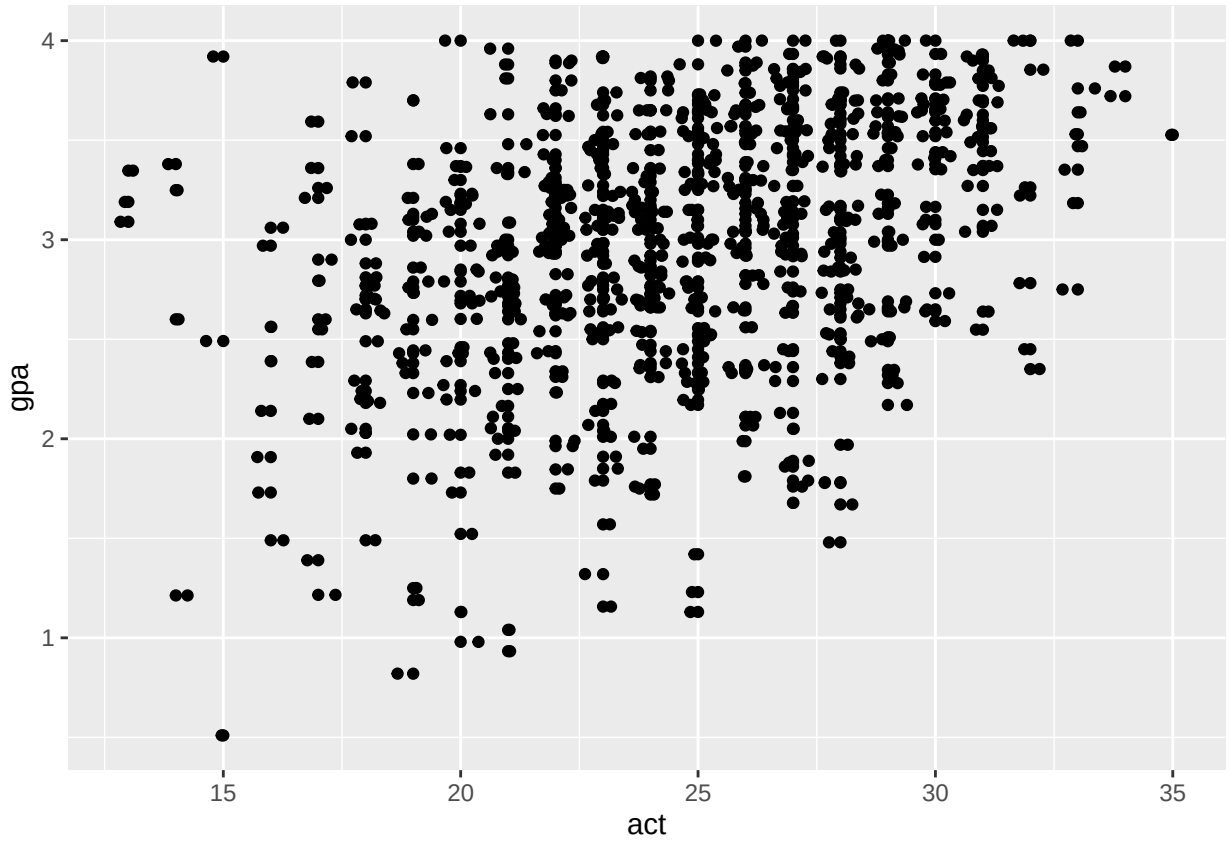
```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
uni <- read_csv("../week_3/university.csv")
```

```
## Rows: 705 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (5): id, gpa, rank, act, year
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ggplot(uni, mapping = aes(x = act, y = gpa)) +
  geom_point() +
  geom_jitter()
```



There appears to be a moderate positive association; it appears that as ACT score is higher, so is freshman GPA.

2.

Based on the plot above, it seems as though a linear relationship could be appropriate because the shape of the plots indicates relatively similar variances across the data.

3.

```
uni_model <- lm(gpa ~ act, uni)
summary(uni_model)
```

```
##
## Call:
## lm(formula = gpa ~ act, data = uni)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.91571 -0.33951  0.09489  0.43068  1.49429
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.55870    0.13802   11.29   <2e-16 ***
## act          0.05780    0.00555   10.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.591 on 703 degrees of freedom
## Multiple R-squared:  0.1337, Adjusted R-squared:  0.1324
## F-statistic: 108.5 on 1 and 703 DF,  p-value: < 2.2e-16
```

estimated regression function: y = 1.55870 + 0.05780x

4.

It is not possible to get an ACT score of 0, the lowest possible score is 1. Therefore, an interpretation of the y-interecept is as follows: If a student scores 1 on the ACT, their predicted GPA is 1.6165. The interpretation for the slope is as follows; for every additional unit in ACT score, the GPA is 0.05780 more on average.

5.

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

```
##   new_act
## 1      31
```

```
## 2      18
## 3       5
## 4      29
```

```
predict_uni <- function(x){
  1.55870 + 0.05780 * x
}

predict_uni(31)
```

```
## [1] 3.3505
```

```
predict_uni(18)
```

```
## [1] 2.5991
```

```
predict_uni(5)
```

```
## [1] 1.8477
```

```
predict_uni(29)
```

```
## [1] 3.2349
```

An ACT score of 5 seems to give an inappropriate prediction, as it seems out of range of our plot above.

6.

Some advantages of this method are simplicity, quick interpretation, and easy comparison. This method is straightforward, easy to implement and understand, and the clear visualization of the relationship between ACT and GPA makes it easy to interpret. Moreover, since the data is summarized to discrete points, it easier to compare the average GPA across different ACT levels. However, this method does come with some disadvantages. First, this method overlooks individual variability within each group. Some students with the same ACT scores might have different GPAs, which can skew the accuracy of the average. There is also possibility for potential misrepresentation. Averaging GPAs across ACT levels may not accurately capture the true relationship between ACT scores and GPA. It may oversimplify the relationship, leading to potential misinterpretations or incorrect conclusions. Finally, this approach is inadequate for predictive modeling because it doesn't provide that accuracy that is needed to build predictive models and understand nuanced relationships between variables. Personally, I would prefer a more comprehensive approach that has more predictive power and less room for misinterpretation.

Received help from chat gpt: https://chat.openai.com/c/070141db-bfc9-41f5-bb2b-1d261a156335