default
Outline

Contents

label=sec:orgb1fb7b8,fragile]Applied Regression Analysis - Homework 1

2 (p33) $Y = 300 + 2X$

This is a functional relation since it is deterministic.

5 (p33) *When asked to state the SLR model, a student wrote it as follows:* $E(Y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$
you agree?\

The model proposed by the student does not calculate the expectation of ϵ. The SLR model should be as follows: $E(Y_i) = \beta_0 + \beta_1 x_i$ since $E(\epsilon_i) = 0$

8 (p33) If another observation is obtained at $X = 45$, the expected value will still remain as $Y = 104$ since the expected value represents the average response for a given predictor. The new Y value may not be 108 however since each observation has its own error term.

11 (p34) /The regression function relating production output by an employee after taking a training program (Y) to the production output before the training program (X) is $E(Y) = 20 + .95X$ where x:[40, 100]. An observer concludes that the training program does not raise production output on average because β_1 is not greater than 1. Comment./

The average production output of an employee after taking the training program is represented by Y_i. β_1 represents the effect of an employee's pre-training-program production output in the model. For all values of X within the model, the average post-training-program production output (Y) exceeds the pre-training-program production output (X) so the observer's interpretation of β_1 is incorrect.

16 (p34) *Evaluate the following satement: "For the least squares method to be fully valid, it is required that the distribution of Y be normal."*

One of the assumptions for the Least Squares method is that all observations are Independent and Identically Distributed (i.i.d). Least Squares does not require a specific distribution thus normality for Y is not a requirement.

18 (p34) /According to (1.17), $\Sigma e_i = 0$ when regression model (1.1) is fitted to a set of n cases by the method of least squares. Is it also true that $\Sigma \epsilon_i = 0$?. Comment./

It is not necessarily true that $\Sigma \epsilon_i = 0$.

Let the residual e_i be defined as $e_i = Y_i - \hat{Y}_i$

Let the response value be defined as $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$.

Let the predicted response value be defined as: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1$
Then $e_i = \beta_0 + \beta_1 X_1 + \epsilon_i - \hat{\beta}_0 - \hat{\beta}_1 X_1$.
Rearranging terms, this becomes
$(\beta_0 - \hat{\beta}_0) + (\beta_1 X_i - \hat{\beta}_1 X_i) + \epsilon_i = e_i$
Summing the residuals gives us the following:
$\sum_1^n (\beta_0 - \hat{\beta}_0) + \sum_1^n (\beta_1 X_i - \hat{\beta}_1 X_i) + \sum_i^n \epsilon_i = \Sigma e_i = 0$
$\rightarrow \sum_i^n \epsilon_i = \sum_i^n [(\hat{\beta}_0 - \beta_0) + (\hat{\beta_1 X_i} - \beta_1 X_i)]$
Thus it is not true that $\Sigma \epsilon_i = 0$
19 (p35) a

```
data <- readxl::read_excel("~/Downloads/GradePointAverage.xlsx")
model <- lm(GPA ~ ACT, data = data)
summary(model)

Call:
lm(formula = GPA ~ ACT, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.74004 -0.33827  0.04062  0.44064  1.22737

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
ACT          0.03883    0.01277   3.040  0.00292 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6231 on 118 degrees of freedom
Multiple R-squared:  0.07262,Adjusted R-squared:  0.06476
F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

Least Squares estimates:

- $\beta_0 = 2.11405$
- $\beta_1 = 0.03883$

Estimated Regression Function: $\hat{Y} = 2.11405 + 0.03883 X_{GPA}$ b

```
data %>%
   ggplot(aes(x = ACT, y = GPA)) +
```

```
geom_point() +
geom_smooth(method = "lm", se = FALSE)
```

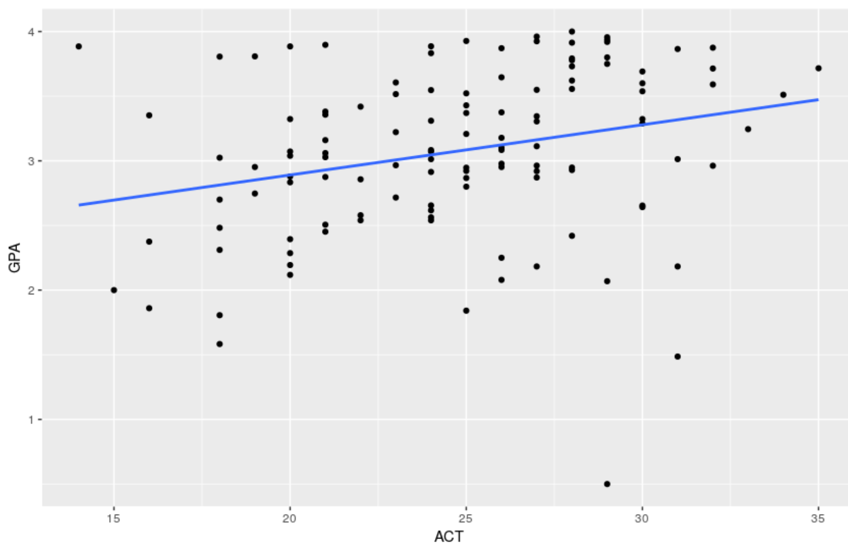

Figure 1: Regression Line

The regression line fits decently. It is far from a perfect fit but it does capture a general positive correlation between ACT Scores and GPA. c

```
predict(model, data.frame(ACT = c(30)))
```

3.278863 d The point estimate for the mean response increases by 0.03883 for each additional point scored on the ACT.

23b (p36) *Estimate σ^2 and σ. In what units is σ expressed?*

```
mean((data$GPA - predict(model))^2)
# [1] 0.3818134
```

$\hat{\sigma^2} = 0.3818134$

$\hat{\sigma} = 0.6179105$

σ is expressed as GPA.

22 (p36) a $\hat{Y} = 168.6 + 2.03438 \times X_{HOURS}$

The estimated line gives a pretty good fit to the data. b

```
predict(model2, data.frame(Hours = c(40)))
# 249.975
```

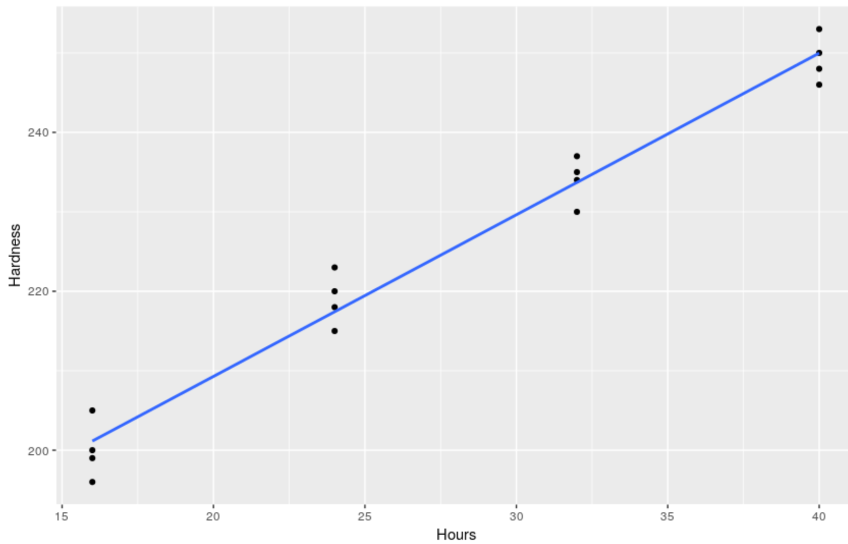

Figure 2: Regression Line

$\hat{Y_{40}} = 249.975$

c The mean Brinell Hardness score increases by 2.03438 per hour of elapsed time.

26b (p36) *Estimate σ^2 and σ. In what units is σ expressed?*

```
mean((plastic$Hardness - predict(model2))^2)
# [1] 9.151563
```

$\hat{\sigma^2} = 9.151563$

$\hat{\sigma} = 83.7511$

σ is expressed as the Brinell Hardness Score.

30 (p37) /What is the implication for the regression function if $\beta_1 = 0$ so that the model is $Y_i = \beta_0 + \epsilon_i$? How would the regression function plot on a graph?/

This type of model is known as an intercept-only model. The model is a constant so it would appear as a straight line on a graph. There are many uses but a primary use is as a baseline for comparing with a model containing parameters. If an intercept-only model is considered a better fit than models with parameters, it means that additional parameters do not help explain the model any more than the intercept.

33 (p37) $Q = \sum_1^n [Y_i - \beta_0]^2$

$\rightarrow \frac{\partial Q}{\partial \beta_0} = -2\sum_1^n [Y_i - \beta_0] = 0$
$\rightarrow -2\sum_1^n Y_i - n\beta_0 = 0$
$\rightarrow -2n\bar{y} - n\beta_0 = 0$
$\rightarrow \beta_0 = -2\bar{y}$