Chapter 2 Example

Andaleeb Hassan

Computing using user defined function

Forbes's data

User defined function:

```
SLR <- function(x,y) {
  n <- length(x)
  x_bar <- mean(x)
  y_bar <- mean(y)

  SXX <- sum((x-x_bar)**2)
  SXY <- sum((x-x_bar)*(y-y_bar))
  SYY <- sum((y-y_bar)**2)

  beta_1_hat <- SXY/SXX
  beta_0_hat <- y_bar-(x_bar*beta_1_hat)
  RSS <- SYY-(beta_1_hat)^2*SXX
  RMS <- RSS/(n-2)

  se_beta1 <- sqrt(RMS/SXX)
  se_beta0 <- sqrt(RMS*((1/n)+(x_bar^2/SXX)))

   R_2 <- 1-(RSS/SYY)
   cov <- -RMS * (x_bar/SXX)


  return(list(x_bar = x_bar,
              y_bar = y_bar,
              SXX = SXX,
              SXY = SXY,
```

```r
              SYY = SYY,
              beta_1_hat = beta_1_hat,
              beta_0_hat= beta_0_hat,
              RSS = RSS,
              RMS = RMS,
              se_beta1 = se_beta1,
              se_beta0 = se_beta0,
              n = n,
              R_2 = R_2,
              cov = cov
  ))

}

result <- SLR(Forbes$bp,Forbes$lpres)

result[c(6,7)]
```

```
$beta_1_hat
[1] 0.8954937

$beta_0_hat
[1] -42.13778
```

$$\hat{E}(lpres|bp) = -42.138 + 0.895bp$$

```r
result$RSS
```

```
[1] 2.154927
```

```r
result$RMS
```

```
[1] 0.1436618
```

confidence interval for β_0:

```r
result$se_beta0
```

```
[1] 3.340199
```

```r
round(c(result$beta_0_hat-(qt(.95,result$n-2)*result$se_beta0),result$beta_0_hat+(qt(.95,resu
```

```
[1] -47.99 -36.28
```

The 90% confidence interval is $-48.35 \leq \beta_0 \leq -35.93$.

$$H_0 : \beta_0 = -35$$
$$H_1 : \beta_0 \neq -35$$

The test statistic :

$$t = \frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0|x_1, x_2, \cdots, x_n)}$$

```r
t = (result$beta_0_ha+35)/result$se_beta0
min(2*pt(abs(t),result$n-2,lower.tail = FALSE),1)
```

```
[1] 0.04948427
```

Here, for the two sided test the p-value is $p = min\{1, p(t > |t_{cal}|)\}$. The p-value is approximately 0.05, providing evidence against the null hypothesis.

Slope :

$$\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\ se(\hat{\beta}_1|x_1, x_2, \cdots, x_n) \leq \beta_1 \leq \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\ se(\hat{\beta}_1|x_1, x_2, \cdots, x_n)$$

Predicted value and Fitted value

Prediction of value y_* corresponding to x_* which is yet to be observed . With assumption that the fitted model to the observed data can be used to predict for the new case.The point prediction would be:

$$\tilde{y}_* = \hat{\beta}_0 + \hat{\beta}_1\ x_* \qquad\qquad x_* \text{ is a new data point}$$

The true value of y_* is:

$$y_* = \beta_0 + \beta_1 x_* + \epsilon_* \qquad\qquad \epsilon \sim N(0, \sigma^2)$$

The variance of $\tilde{y}_*$ would be the sum of variance due to random error ϵ and variance of the predicted parameters.

$$Var(\tilde{y}_*|x_*) = \sigma^2 + \sigma^2(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX})$$

The standard error of prediction:

$$se(\tilde{y}_*|x_*) = \hat{\sigma}(1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX})^{\frac{1}{2}}$$

For $x_* = 200$, we have, $\tilde{y}_* = 136.961$ and $se(\tilde{y}_*|x_* = 200) = 0.393$

```
c(round(136.961-.393*round(qt(.995,15),3),3),round(136.961+.393*round(qt(.995,15),2),3))
```

```
[1] 135.803 138.120
```

Heights data

```
SLR(Heights$mheight,Heights$dheight)[c(6,7)]
```

```
$beta_1_hat
[1] 0.541747

$beta_0_hat
[1] 29.91744
```

For the model:

$$\widetilde{dheight}_* = E(dheight_*|mheight_*) = \beta_0 + \beta_1 mheight_*$$

Similarly, 95% confidence interval for daughter's height for given mother's height can be calculated as following.

$$\hat{\beta}_0 + \hat{\beta}_1 x_* \pm t_{.025,n-2} \times se(dheight_*|mheight_*)$$

In built function:

```
model <- lm(lpres~bp,Forbes)
```

```
summary(model)[6]
```

```
$sigma
[1] 0.3790275
```

```
broom::tidy(model)%>%
mutate(p.value = scales::pvalue(p.value)) %>%
  kbl(format = "latex", booktabs = T,
      digits = 3, caption = "Estimate of the model") %>%
  kable_styling(latex_options = "hold_position")
```

Table 1: Estimate of the model

term	estimate	std.error	statistic	p.value
(Intercept)	-42.138	3.340	-12.615	<0.001
bp	0.895	0.016	54.431	<0.001

Similarly,

$$\hat{E}(lpres|bp) = -42.138 + 0.895bp$$

The Confidence interval:

```
confint(model,level = .9)
```

```
                    5 %        95 %
(Intercept) -47.9933161 -36.2822424
bp            0.8666529   0.9243344
```

prediction:

```
data <- tibble(bp = c(200))

predict(model, newdata = data, level = .99,interval = "prediction")
```

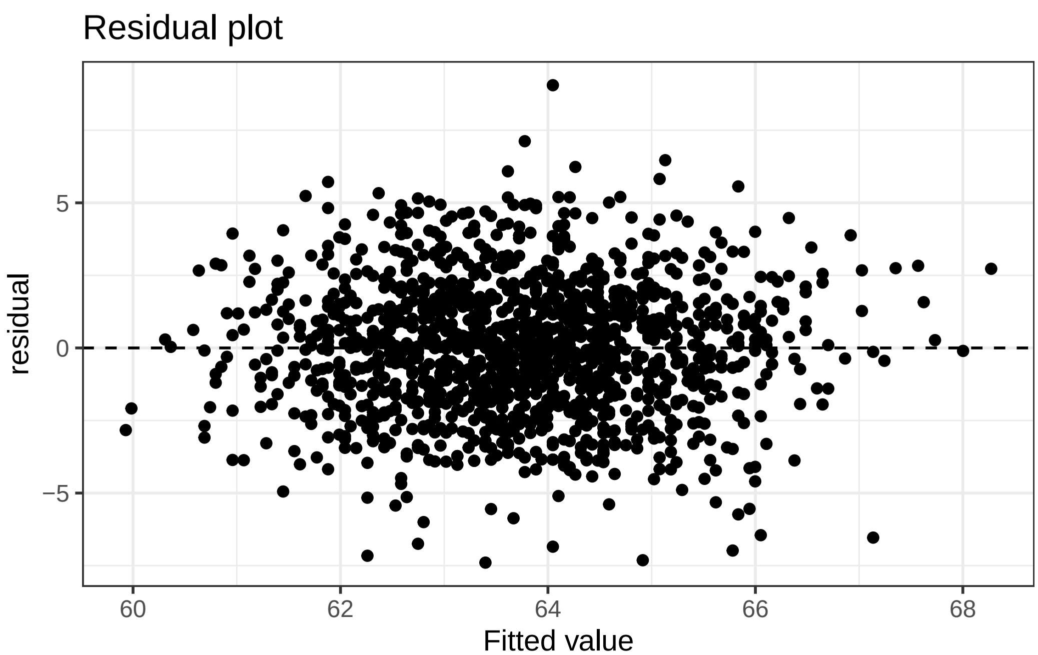
```
      fit      lwr      upr
1 136.961 135.8028 138.1191
```

Residuals

Residual analysis is used for model adequacy checking.

```
model <- lm(dheight~mheight, data = alr4::Heights)


ggplot(mapping = aes(model$fitted.values, model$residuals))+
  geom_point()+
  labs(title = "Residual plot",
       x = "Fitted value", y  = "residual")+
  geom_hline(yintercept = 0, linetype = "dashed")+
  theme_bw()
```



There is no pattern visible in the plot. Thus the assumptions are not violated.

```
model1 <- lm(lpres~bp, data = Forbes)
model2 <- lm(lpres~bp, data = Forbes[-12,])

tidy(model1)%>%
  select(1:2) %>%
  mutate(type = "model1")%>%
  bind_rows(tidy(model2)%>%
```

```r
                select(1:2)%>%
                mutate(type = "model2"))%>%
    pivot_wider(id_cols = term, names_from = type, values_from = estimate) %>%
   kbl(format = "latex", booktabs = TRUE, digits = 3, caption = "Model comparisons") %>%
  kable_styling(latex_options = "hold_position")
```

Table 2: Model comparisons

term	model1	model2
(Intercept)	-42.138	-41.308
bp	0.895	0.891

```r
`model-1` <- lapply(SLR(Forbes$bp,Forbes$lpres),round,digit = 3 )
`model-2` <- lapply(SLR(Forbes[-12,]$bp,Forbes[-12,]$lpres),round,digit = 3)

cbind(`model-1`,`model-2`) %>%
  kbl(format = "latex", booktabs = TRUE, digits = 3, caption = "Model comparisons") %>%
  kable_styling(latex_options = "hold_position")
```

Table 3: Model comparisons

	model-1	model-2
x_bar	202.953	202.85
y_bar	139.605	139.428
SXX	530.782	527.9
SXY	475.312	470.351
SYY	427.794	419.256
beta_1_hat	0.895	0.891
beta_0_hat	-42.138	-41.308
RSS	2.155	0.18
RMS	0.144	0.013
se_beta1	0.016	0.005
se_beta0	3.34	1.001
n	17	16
R_2	0.995	1
cov	-0.055	-0.005