PSTAT 126 HW #2

Mujie Wang

04/20/2018

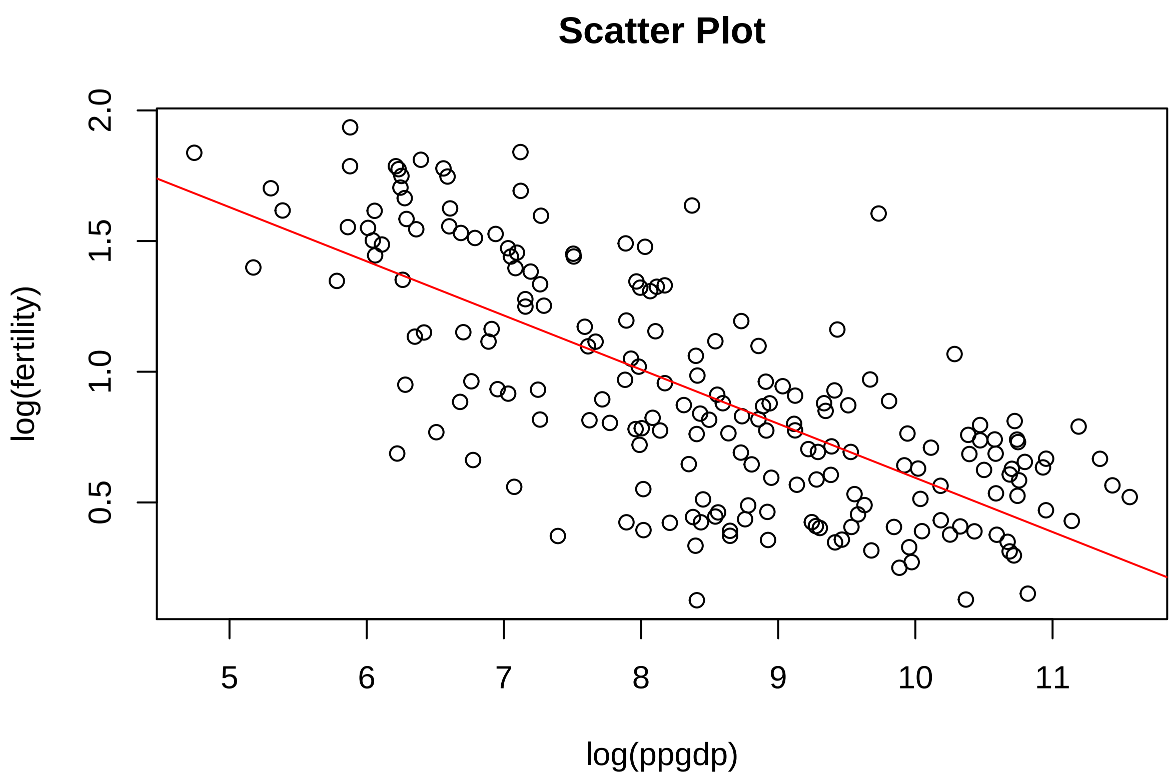
```
#Problem 1
#(a)
library(alr4)
```

```
## Loading required package: car
## Warning: package 'car' was built under R version 3.4.4
## Loading required package: carData
## Warning: package 'carData' was built under R version 3.4.4
## Loading required package: effects
## Warning: package 'effects' was built under R version 3.4.4
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
data(UN11)

attach(UN11)
x = log(ppgdp)
y = log(fertility)

fit <- lm(y~x)
plot(x,y, main = 'Scatter Plot', xlab = 'log(ppgdp)', ylab = 'log(fertility)')
abline(coef(fit), col= 2)
```




```r
#(b)
xbar = mean(x)
ybar = mean(y)
Sxx = sum((x - xbar)^2)
Syy = sum((y - ybar)^2)
Sxy = sum((x - xbar)*(y - ybar))
r = Sxy/sqrt(Sxx*Syy)
b1 = r*sqrt(Syy/Sxx)
b0 = ybar - b1*xbar
yhat = b0 + b1*x
ssto = sum((y - ybar)^2)
sse = sum((y - yhat)^2)
ssr = sum((yhat - ybar)^2)
r2 = ssr/ssto
r2
```

```
## [1] 0.525985
```

```r
#The coefficient of the determination is 0.526. It's a 52.5985% reduced by the predictor log(ppgdp)

#(c)

newdata = data.frame(x = log(43140.9))
fit = lm(y~x)
p = predict(fit,newdata, interval = "predict")

#The 95% prediction interval of log(fertility) is lies between (-0.1554686, 10.64984)
```

```r
#(d)
p[2:3]
```

```
## [1] -0.1554686  1.0649842
```

```r
exp(p[2:3])
```

```
## [1] 0.856014 2.900793
```

```r
# 95% prediction interval of fertility is(0.856014 2.900793)

#Problem 2
#(a)
data(Heights)
x = Heights$mheight
y = Heights$dheight
n = length(y)
xbar = mean(x)
xbar
```

```
## [1] 62.4528
```

```r
ybar = mean(y)
ybar
```

```
## [1] 63.75105
```

```r
Sxx = sum((x - xbar)^2)
Sxx
```

```
## [1] 7620.907
```

```r
Syy = sum((y - ybar)^2)
Syy
```

```
## [1] 9288.616
```

```r
Sxy = sum((x - xbar)*(y - ybar))
Sxy
```

```
## [1] 4128.603
```

```r
b1 = Sxy/Sxx
b1
```

```
## [1] 0.541747
```

```r
b0 = ybar - b1*xbar
b0
```

```
## [1] 29.91744
```

```r
yhat = b0 +b1*x
e = y - yhat
sig2hat = sum(e^2)/(n-2)
sig2hat
```

```
## [1] 5.136167
```

```r
se1= (sig2hat/Sxx)^.5
se1
```

```
## [1] 0.02596069
```

```r
se0 = (sig2hat*(1/n + xbar ^2/Sxx))^.5
se0
```

```
## [1] 1.622469
```

```r
t_b0 = b0/se0
t_b1 = b1/se1
p0 = 1 - pt(t_b0, df = n -2)
p0
```

```
## [1] 0
```

```r
p1 = 1 - pt(t_b1, df = n -2)
p1
```

```
## [1] 0
```

```r
fit = lm(y~x)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.397 -1.529  0.036  1.492  9.053
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.91744    1.62247   18.44   <2e-16 ***
## x            0.54175    0.02596   20.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.266 on 1373 degrees of freedom
## Multiple R-squared:  0.2408, Adjusted R-squared:  0.2402
## F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16
```

```r
#2(b)
ci_b1 = b1 + qt(1-.05/2, df = n - 2) * c(-1, 1) * se1
ci_b1
```

```
## [1] 0.4908201 0.5926739
```

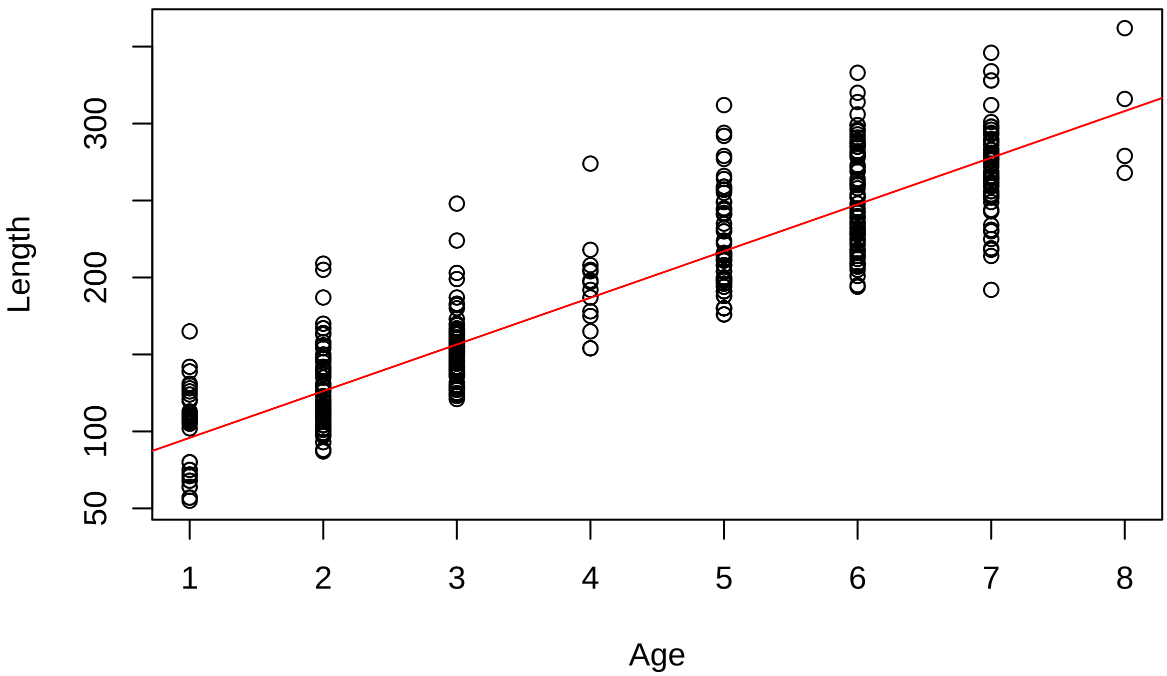
```r
#2(c)
fit = lm(y~x)
ybar = mean(y)
yhat = fitted(fit)
ssr = sum((yhat-ybar)^2)
ssto = sum((y-ybar)^2)
r2 = ssr/ssto
r2
```

```
## [1] 0.2407957
```

```r
#Round 24.08% variation in dheight is explained by the variation in mheight


#Problem 3
#(a)
data(wblake)
x = wblake$Age
y = wblake$Length

fit = lm(y ~ x)
plot(x, y, xlab = 'Age', ylab = 'Length')
abline(coef(fit), col = 2)
```




```r
#3(b)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.794 -19.499  -4.499  16.177  94.853
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  65.5272     3.1974   20.49   <2e-16 ***
## x            30.3239     0.6877   44.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.65 on 437 degrees of freedom
```

```
## Multiple R-squared:  0.8165, Adjusted R-squared:  0.8161
## F-statistic:  1944 on 1 and 437 DF,  p-value: < 2.2e-16
```

```
#3(c)
new = data.frame(x = 4)
ci = predict(fit, new, interval = 'confidence', level = .95)[2:3]
ci
```

```
## [1] 184.1217 189.5237
```

```
n = length(y)
xbar = mean(x)
Sxx = sum((x - xbar)^2)
b0 = coef(fit)[1]
b1 = coef(fit)[2]
yhat = fitted(fit)
e = y - yhat
sig2hat = sum(e^2) / (n-2)

xh = 4
#the point estimate for y at xh = 4
yh = b0 + b1*xh
#the formula for a 95% c.i. for mean response at xh = 4
ci = yh + c(-1, 1) * qt(1-.05/2, df = n - 2) * sqrt(sig2hat) * sqrt(1/n + (xh - xbar)^2/Sxx)
ci
```

```
## [1] 184.1217 189.5237
```

```
#The 95% con dence interval for mean age = 4 is (184.1217, 189.5237)

#3(d)
new = data.frame(x = 9)
ci = predict(fit, new, interval = 'confidence', level = .95)[2:3]
ci
```

```
## [1] 331.4231 345.4612
```

```
#The 95% codence interval for mean age = 9 is between (331.4231, 345.46127).Since the value Age = 9 is
```