PSTAT126\_Homework2

Yanjie Qi

2019/8/19

## 1.

## (a)

Predictor here is ppgdp and response is fertility.

## (b)

load alr4 package:

library(alr4)

## Loading required package: car

## Loading required package: carData

## Loading required package: effects

## lattice theme set by effectsTheme()  
## See ?effectsTheme for details.

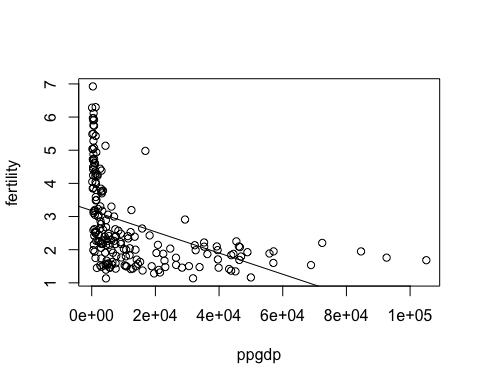
data(UN11)  
attach(UN11)

create two variables from data UN11:

fertility=UN11$fertility  
ppgdp=UN11$ppgdp

draw scattle plot and the linear model:

plot(x=ppgdp,y=fertility)  
abline(lm(fertility~ppgdp))



summary(lm(fertility~ppgdp))

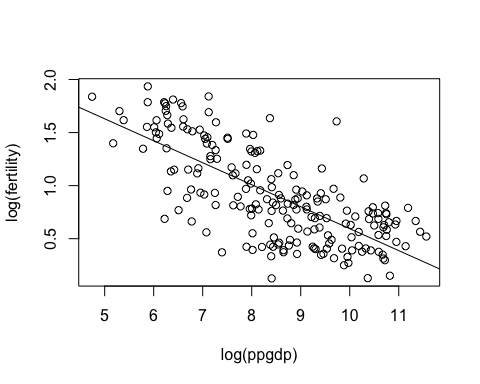
##   
## Call:  
## lm(formula = fertility ~ ppgdp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.9006 -0.8801 -0.3547 0.6749 3.7585   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.178e+00 1.048e-01 30.331 < 2e-16 \*\*\*  
## ppgdp -3.201e-05 4.655e-06 -6.877 7.9e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.206 on 197 degrees of freedom  
## Multiple R-squared: 0.1936, Adjusted R-squared: 0.1895   
## F-statistic: 47.29 on 1 and 197 DF, p-value: 7.903e-11

According to the graph above, The trend is not linear.

## (c)

draw scattle plot and the linear model by using natural log of the variables:

plot(x=log(ppgdp),y=log(fertility))  
abline(lm(log(fertility)~log(ppgdp)))



summary(lm(log(fertility)~log(ppgdp)))

##   
## Call:  
## lm(formula = log(fertility) ~ log(ppgdp))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.79828 -0.21639 0.02669 0.23424 0.95596   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.66551 0.12057 22.11 <2e-16 \*\*\*  
## log(ppgdp) -0.20715 0.01401 -14.79 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3071 on 197 degrees of freedom  
## Multiple R-squared: 0.526, Adjusted R-squared: 0.5236   
## F-statistic: 218.6 on 1 and 197 DF, p-value: < 2.2e-16

The dependence of response variable (fertility) on predictor (ppgdp) is clearly shown on the log model. Therefore, simple linear regression model is plausible for summary of this graph.

## 2.

load faraway Packages:

library(faraway)

##   
## Attaching package: 'faraway'

## The following objects are masked from 'package:alr4':  
##   
## cathedral, pipeline, twins

## The following objects are masked from 'package:car':  
##   
## logit, vif

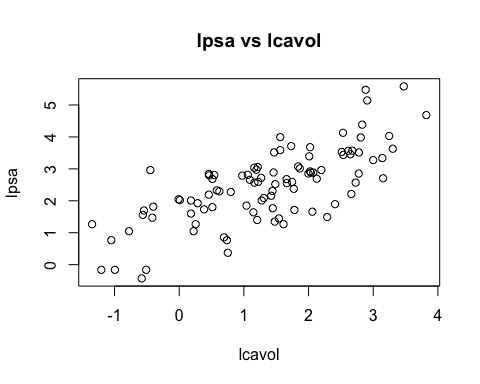
names(prostate)

## [1] "lcavol" "lweight" "age" "lbph" "svi" "lcp" "gleason"  
## [8] "pgg45" "lpsa"

## a)

draw a scatter plot:

plot(prostate$lcavol,prostate$lpsa,xlab="lcavol",ylab="lpsa",main="lpsa vs lcavol")



We can see that there is an overall positive linear relationship between lspa and lcavol; simple linear regression model seems reasonable.

## b)

access data lpsa and lcavol:

y<-prostate$lpsa  
x<-prostate$lcavol

sample means:

xbar<-mean(x)  
ybar<-mean(y)

sum of sqaures:

Sxx<-sum((x-xbar)^2)  
Syy<-sum((y-ybar)^2)  
Sxy<-sum((x-xbar)\*(y-ybar))

estimate the value of slope:

beta1hat<-Sxy/Sxx

estimate the value of intercept:

beta0hat<-ybar-beta1hat\*xbar

Displaying values:

sprintf('The Value of xbar is %.4f',xbar)

## [1] "The Value of xbar is 1.3500"

sprintf('The Value of ybar is %.4f',ybar)

## [1] "The Value of ybar is 2.4784"

sprintf('The Value of Sxx is %.4f', Sxx)

## [1] "The Value of Sxx is 133.3590"

sprintf('The Value of Syy is %.4f', Syy)

## [1] "The Value of Syy is 127.9176"

sprintf('The Value of Sxy is %.4f', Sxy)

## [1] "The Value of Sxy is 95.9278"

sprintf('The estimated value of the beta0 is %.4f',beta0hat)

## [1] "The estimated value of the beta0 is 1.5073"

sprintf('The estimated value of the beta1 is %.4f',beta1hat)

## [1] "The estimated value of the beta1 is 0.7193"

sprintf('The estimated regression line is %.4f+%.4fx',beta0hat,beta1hat)

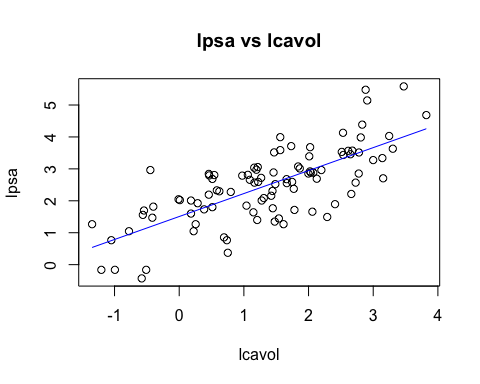
## [1] "The estimated regression line is 1.5073+0.7193x"

calculate the fitted values:

yhat<-beta0hat+beta1hat\*x

draw the fitted line on to the plot from part a):

plot(x,y,xlab="lcavol",ylab="lpsa",main="lpsa vs lcavol")  
lines(sort(x),yhat[order(x)],col="blue")



## c)

get the number of observations:

n<-length(x)

get the sum of square error:

sse<-Syy-beta1hat\*Sxy

get mean square error, which is the estimate of sigma^2:

mse<-sse/(n-2)

estimates of stamdard errors:

sb1<-sqrt(mse/Sxx)  
sb0<-sqrt(mse\*sum(x^2)/(n\*Sxx))  
sprintf('The estimated value of sigma^2 %.4f',mse)

## [1] "The estimated value of sigma^2 0.6202"

sprintf('The standard error of beta1 is %.4f',sb1)

## [1] "The standard error of beta1 is 0.0682"

sprintf('The standard error of beta0 is %.4f',sb0)

## [1] "The standard error of beta0 is 0.1219"

get the covariance:

cov<--mse\*xbar/Sxx  
sprintf('The estimated covariance between beta0&beta1 is %.4f',cov)

## [1] "The estimated covariance between beta0&beta1 is -0.0063"

t-tests for the null hypotheses beta0 = 0:

tb0<-beta0hat/sb0

p-value of beta0 = P(T>tb0)+P(T<-tb0):

pb0<-pt(abs(tb0),df=n-2,lower.tail=FALSE)+ pt(-abs(tb0),df=n-2,lower.tail=TRUE)  
sprintf('The test statistics to test beta0=0 is %.4f, the p-value is %.4f',tb0,pb0)

## [1] "The test statistics to test beta0=0 is 12.3613, the p-value is 0.0000"

t-tests for the null hypotheses beta1 = 0:

tb1<-beta1hat/sb1

p-value of beta1 = P(T>tb1)+P(T<-tb1):

pb1<-pt(abs(tb1),df=n-2,lower.tail=FALSE)+ pt(-abs(tb1),df=n-2,lower.tail=TRUE)  
sprintf('The test statistics to test beta1=0 is %.4f, the p-value is %.4f',tb1,pb1)

## [1] "The test statistics to test beta1=0 is 10.5483, the p-value is 0.0000"

## 3.

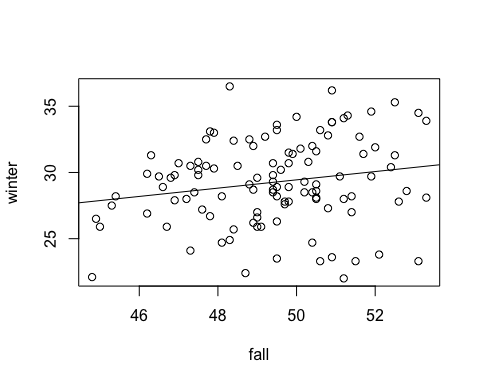
## a)

Loan the data set:

data(ftcollinstemp)  
attach(ftcollinstemp)

draw a scattle plot and the linear model:

plot(x=fall,y=winter)  
abline(lm(winter~fall))

 ## b)

summary(lm(winter~fall))

##   
## Call:  
## lm(formula = winter ~ fall)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.8186 -1.7837 -0.0873 2.1300 7.5896   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.7843 7.5549 1.825 0.0708 .  
## fall 0.3132 0.1528 2.049 0.0428 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.179 on 109 degrees of freedom  
## Multiple R-squared: 0.0371, Adjusted R-squared: 0.02826   
## F-statistic: 4.2 on 1 and 109 DF, p-value: 0.04284

According to Summary, we get p-value is is 0.0428 which is larger than 0.01; therefore, we accept the null hypothesis H0 that the slope is 0 when alpha = 0.01.

## c)

PercofVar<-summary(lm(winter~fall))$r.squared  
PercofVar

## [1] 0.03709854

We could draw the conclusion from above that the percentage of the variability in winter is explained by fall is around 3.7%.

## 4.

## a)

Loan the data set:

data(Heights)  
attach(Heights)

constructe the linear model:

mod <- lm(dheight~mheight)  
summary(mod)

##   
## Call:  
## lm(formula = dheight ~ mheight)  
##   
## 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 \*\*\*  
## mheight 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

Given by summary, we first have Intercept 29.91744 and slope 0.54175; Therefore dheight = 29.91744 + 0.54175 \* mheight. As shown in the summary above, the standard Error of dheight and mheight is 1.62247 and 0.02596, the coefficient of determination is 0.2408, and the estimate of variance is 2.266^2=5.136167.From the coefficient of determination, 24.08% of variance in dheight is explained by mheight.

## b)

confint(mod, 'mheight', level=0.90)

## 5 % 95 %  
## mheight 0.4990166 0.5844774

a 90% confidence interval for β1 is (0.4990166, 0.5844774)

## c)

predict(mod, data.frame(mheight=61), interval="prediction",level=.99)

## fit lwr upr  
## 1 62.964 57.11531 68.8127

From above, we get a 99% confidence interval for daughters whose mother is 61 inches tall is (57.11531, 68.8127)