STAT2008 Tutorial Week 5

Tutorial 2 Question 1 and 2

Yunxi (Lucy) Hu

Research School of Finance, Actuarial Studies and Statistics
The Australian National University

- Wattle - STAT2008 - Tutorial 2 and 2014 Assignment 1 - Download "auscars.csv", "prostate.csv", "teengamb.csv"
- Set up your working directory - RStudio - Session - Set Working Directory - Choose Directory - Choose the folder that you download "auscars.csv" before
- Run the R command

Recap

- Hypothesis
  1. Overall Hypothesis: ***anova(name.lm)***
  2. Individual Hypothesis: ***summary(name.lm)***
  3. Linear relation between X and Y: ***cor.test(x,y)***
- Three equations
  1. Population regression function:

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i$$

  2. Model (Observation):

$$Y_i = \beta_0 + \beta_1 X_i + \xi_i = E(Y_i|X_i) + \xi_i$$

  3. Fitted line:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = b_0 + b_1 X_i$$

- Residuals versus errors
  1. Residuals: $ei = Y_i - \hat{Y}_i$
  2. Error: $\xi_i = Y_i - E(Y_i|X_i)$

Overall Hypothesis

- Let k: Number of X type (SLR: k=1)
- Let p=k+1 = Number of parameters
- Let n: Number of observations

```
# ANOVA table (for MLR; SLR: k=1)
#
# Source      Df            SS    MS=SSi/dfi        F(TS)      Pr(>F)
# Model       k             SSR   MSR=SSR/k         MSR/MSE    p(TS)
# Residuals   n-k-1=n-p     SSE   MSE=SSE/(n-p)
# Total       n-1           SST
```

Individual Hypothesis

- By default, the Test Statistic and P value from summary table can only be used for testing Ho: $\beta_0 = 0$ or $\beta_1 = 0$

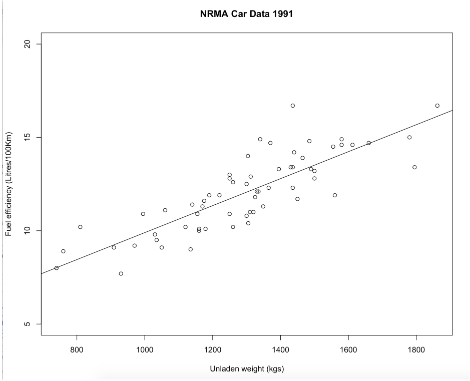
```
# Summary for indidual hypothesis (B0, B1)
#
#             Estimate        SE      t(TS)         Pr(>|t|)
# (Intercept) b0=ybar-b1*xbar SE(b0)  (b0-0)/SE(b0) p(Bo)
# Slope(X)    b1=Sxy/Sxx      SE(b1)  (b1-0)/SE(b1) p(B1)
```

Question 1(b)

```
> # 1. Now fit the requested model using lm():
> auscars.lm <- lm(L.100k ~ Weight)
> auscars.lm
Call:
lm(formula = L.100k ~ Weight)
Coefficients:
(Intercept)       Weight
2.670858      0.007227

> # 2. b0=? b1=? SE(b0)=? SE(b1)=?
> summary(auscars.lm)
Call:
lm(formula = L.100k ~ Weight)

Residuals:
Min       1Q   Median       3Q      Max
-2.2441  -0.7913  -0.0689   0.6378   3.6505

Coefficients:
            Estimate   Std. Error  t value  Pr(>|t|)
(Intercept) 2.6708585  0.8246468   3.239    0.00196 **
Weight      0.0072275  0.0006259   11.548   < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.162 on 60 degrees of freedom
Multiple R-squared:  0.6897, Adjusted R-squared:  0.6845

F-statistic: 133.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

Question 1(b)

```
> # 3. Plot:
> # To generate a scatterplot & the limits on the y-axis range [5,20]
> plot(Weight, L.100k, ylim=c(5,20), xlab="Unladen weight (kgs)",
+ ylab="Fuel efficiency (Litres/100Km)", main="NRMA Car Data 1991")
>
> # To fit a regression line
> abline(coef(auscars.lm))
```



Question 1(c)

```
> # Q1 (c)
> # 1. Plot has a strong association between Weight and L.100k.
>
> # 2. Relationship b.w X and Y
> # Method 1: Individual hypothesis for (B1)
>   summary(auscars.lm)$coef
               Estimate    Std. Error    t value    Pr(>|t|)
(Intercept) 2.670858483   0.824646798   3.238791   1.958279e-03
Weight      0.007227456   0.000625853  11.548169   6.952666e-17
>   qt(0.975,length(L.100k)-2)
[1] 2.000298
```

- For Individual Hypothesis

```
 # Summary for indidual hypothesis (B0, B1)
 #
 #              Estimate         SE      t(TS)          Pr(>|t|)
 # (Intercept) b0=ybar-b1*xbar SE(b0)  (b0-0)/SE(b0) p(Bo)
 # Slope(X)    b1=Sxy/Sxx      SE(b1)  (b1-0)/SE(b1) p(B1)
```

Question 1(c)

```
> # Method 2: Overall hypothesis (anova)
>   anova(auscars.lm)
 Analysis of Variance Table

 Response: L.100k
           Df  Sum Sq   Mean Sq F value    Pr(>F)
 Weight     1  180.031  180.03  133.36   < 2.2e-16 ***
 Residuals 60  80.998   1.35
---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>   qf(0.95,1,length(L.100k)-length(auscars.lm$coef))
 [1] 4.001191
```

- For Overall Hypothesis

```
# Let k: # X type (SLR: k=1), p=k+1 = # parameters, n = # observations
#
# ANOVA table (for MLR; SLR: k=1)
#
# Source     Df           SS   MS=SSi/dfi      F(TS)     Pr(>F)
# Model      k            SSR  MSR=SSR/k       MSR/MSE   p(TS)
# Residuals  n-k-1=n-p    SSE  MSE=SSE/(n-p)
# Total      n-1          SST
```

Question 1(d)

```
> # Q1 (d) R^2=?
> # 1. R^2 = SSR/(SSR+SSE)
> anova(auscars.lm)
Analysis of Variance Table

Response: L.100k
          Df  Sum Sq Mean Sq F value    Pr(>F)    
Weight     1 180.031  180.03  133.36 < 2.2e-16 ***
Residuals 60  80.998    1.35                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> # 2. From Summary:
> summary(auscars.lm)$r.squared
[1] 0.6896983
>
> # 3. r^2=R^2 (SLR):
> cor(Weight,L.100k)^2
[1] 0.6896983
>
> # 4. Interpretation:
> # % of variation of the response explained by the model.
```

Question 1(e)

- $\hat{\beta}_0 = b_0 \sim t(\beta_0, SE(b_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}})$
- $CI(\beta_0) : \hat{\beta}_0 \pm t_{n-2,\alpha/2} SE(\beta_0)$

```
> # Q1 (e)
> # 1. Calculate CI(B0)
> # Generate b0 and SE(b0)
> coef(auscars.lm)
(Intercept)      Weight
2.670858483 0.007227456
> b0 <- coef(auscars.lm)[1]
> b0
(Intercept)
2.670858
>
> summary(auscars.lm)$coef
               Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 2.670858483 0.824646798  3.238791 1.958279e-03
Weight      0.007227456 0.000625853 11.548169 6.952666e-17
> SEb0 <- summary(auscars.lm)$coef[1,2]
> SEb0
[1] 0.8246468
```

Question 1(e)

- $\hat{\beta_0} = b_0 \sim t(\beta_0, SE(b_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}})$
- $CI(\beta_0) : \hat{\beta_0} \pm t_{n-2,\alpha/2} SE(\beta_0)$

```
> # The df for residual:
> auscars.lm$df
 [1] 60
> auscars.lm$df.residual
 [1] 60
>
> # The Critical value:
> qt(0.025,auscars.lm$df)
 [1] -2.000298
> qt(0.975,auscars.lm$df)
 [1] 2.000298
>
> # CI:
> c(b0 + qt(0.025,auscars.lm$df)*SEb0, b0 + qt(0.975,auscars.lm$df)*
  SEb0)
(Intercept) (Intercept)
1.021319    4.320398
```

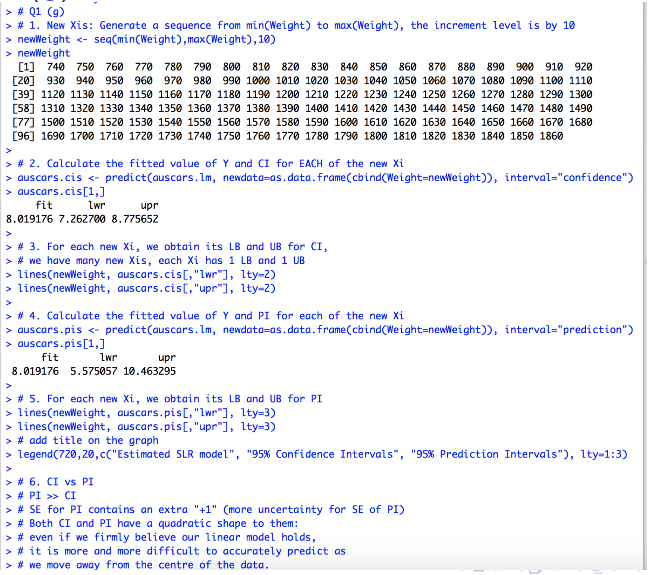
- Interpretation for β_0 and $CI(\beta_0)$?

Question 1(f) CI=? PI=?

- CI
  - $\mu(Y_i|X_0) \sim t(b_0 + b_1 X_0, SE(\mu(Y_i|X_0)) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}})$
  - $CI(\mu(Y_i|X_0)) : (b_0 + b_1 X_0) \pm t_{n-2,\alpha/2} SE(\mu(Y_i|X_0))$
  - CI: In repeated sampling, there is 1-α chance AVERAGE value of Y lies in CI
- PI
  - $Y_i|X_0 \sim t(b_0 + b_1 X_0, SE(Y_i|X_0) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}})$
  - $PI(Y_i|X_0) : (b_0 + b_1 X_0) \pm t_{n-2,\alpha/2} SE(Y_i|X_0)$
  - PI: In repeated sampling, there is 1-α chance SPECIFIC value of Y lies in PI

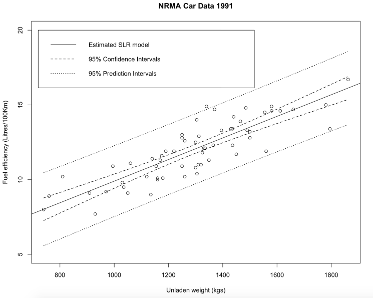
```
> # Define new X0
> newWeight <- 1800
>
> # 1. 95%CI for the EXPECTED value of L.100k when Weight (Xi)= 1800:
> predict(auscars.lm, newdata=as.data.frame(cbind(Weight=newWeight)),
  interval="confidence")
      fit      lwr      upr
1 15.68028 14.98412 16.37644
>
> # 2. 95%PI for the SINGLE value of L.100k when Weight (Xi)= 1800:
> predict(auscars.lm, newdata=as.data.frame(cbind(Weight=newWeight)),
  interval="prediction")
      fit      lwr      upr
1 15.68028 13.25415 18.10641
```

Question 1(g)



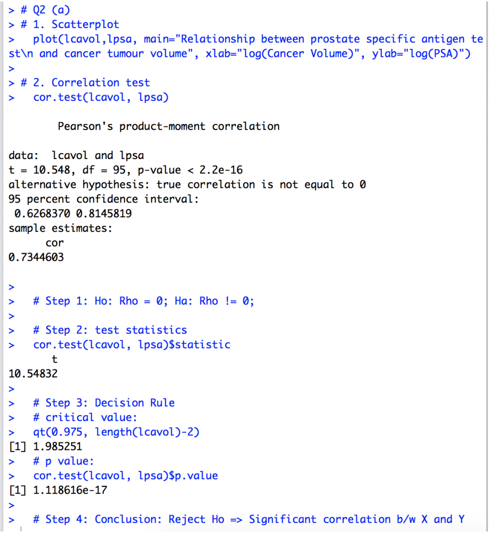
```
> # Q1 (g)
> # 1. New Xis: Generate a sequence from min(Weight) to max(Weight), the increment level is by 10
> newWeight <- seq(min(Weight),max(Weight),10)
> newWeight
  [1]  740  750  760  770  780  790  800  810  820  830  840  850  860  870  880  890  900  910  920
 [20]  930  940  950  960  970  980  990 1000 1010 1020 1030 1040 1050 1060 1070 1080 1090 1100 1110
 [39] 1120 1130 1140 1150 1160 1170 1180 1190 1200 1210 1220 1230 1240 1250 1260 1270 1280 1290 1300
 [58] 1310 1320 1330 1340 1350 1360 1370 1380 1390 1400 1410 1420 1430 1440 1450 1460 1470 1480 1490
 [77] 1500 1510 1520 1530 1540 1550 1560 1570 1580 1590 1600 1610 1620 1630 1640 1650 1660 1670 1680
 [96] 1690 1700 1710 1720 1730 1740 1750 1760 1770 1780 1790 1800 1810 1820 1830 1840 1850 1860
>
> # 2. Calculate the fitted value of Y and CI for EACH of the new Xi
> auscars.cis <- predict(auscars.lm, newdata=as.data.frame(cbind(Weight=newWeight)), interval="confidence")
> auscars.cis[1,]
     fit      lwr      upr
8.019176 7.262700 8.775652
>
> # 3. For each new Xi, we obtain its LB and UB for CI,
> # we have many new Xis, each Xi has 1 LB and 1 UB
> lines(newWeight, auscars.cis[,"lwr"], lty=2)
> lines(newWeight, auscars.cis[,"upr"], lty=2)
>
> # 4. Calculate the fitted value of Y and PI for each of the new Xi
> auscars.pis <- predict(auscars.lm, newdata=as.data.frame(cbind(Weight=newWeight)), interval="prediction")
> auscars.pis[1,]
      fit       lwr       upr
 8.019176  5.575057 10.463295
>
> # 5. For each new Xi, we obtain its LB and UB for PI
> lines(newWeight, auscars.pis[,"lwr"], lty=3)
> lines(newWeight, auscars.pis[,"upr"], lty=3)
> # add title on the graph
> legend(720,20,c("Estimated SLR model", "95% Confidence Intervals", "95% Prediction Intervals"), lty=1:3)
>
> # 6. CI vs PI
> # PI >> CI
> # SE for PI contains an extra "+1" (more uncertainty for SE of PI)
> # Both CI and PI have a quadratic shape to them:
> # even if we firmly believe our linear model holds,
> # it is more and more difficult to accurately predict as
> # we move away from the centre of the data.
```

Question 1(g)

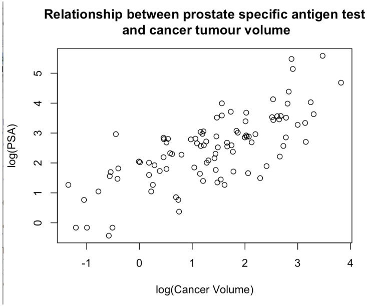


- PI > CI since SE(PI) > SE(CI)
- CI
  - $\mu(Y_i|X_0) \sim t(b_0 + b_1 X_0, SE(\mu(Y_i|X_0)) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}})$
  - $CI(\mu(Y_i|X_0)) : (b_0 + b_1 X_0) \pm t_{n-2,\alpha/2} SE(\mu(Y_i|X_0))$
- PI
  - $Y_i|X_0 \sim t(b_0 + b_1 X_0, SE(Y_i|X_0) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}})$
  - $PI(Y_i|X_0) : (b_0 + b_1 X_0) \pm t_{n-2,\alpha/2} SE(Y_i|X_0)$

Question 2(a)



```
> # Q2 (a)
> # 1. Scatterplot
>   plot(lcavol,lpsa, main="Relationship between prostate specific antigen te
st\n and cancer tumour volume", xlab="log(Cancer Volume)", ylab="log(PSA)")
>
> # 2. Correlation test
>   cor.test(lcavol, lpsa)

        Pearson's product-moment correlation

data:  lcavol and lpsa
t = 10.548, df = 95, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6268370 0.8145819
sample estimates:
      cor
0.7344603


>
>   # Step 1: Ho: Rho = 0; Ha: Rho != 0;
>
>   # Step 2: test statistics
>   cor.test(lcavol, lpsa)$statistic
       t
10.54832
>
>   # Step 3: Decision Rule
>   # critical value:
>   qt(0.975, length(lcavol)-2)
[1] 1.985251
>   # p value:
>   cor.test(lcavol, lpsa)$p.value
[1] 1.118616e-17
>
>   # Step 4: Conclusion: Reject Ho => Significant correlation b/w X and Y
```



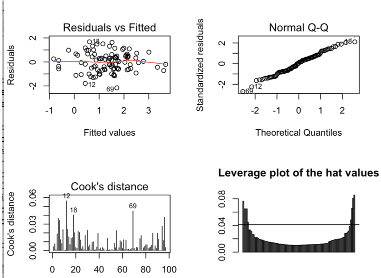
Relationship between prostate specific antigen test and cancer tumour volume

Question 2(b)

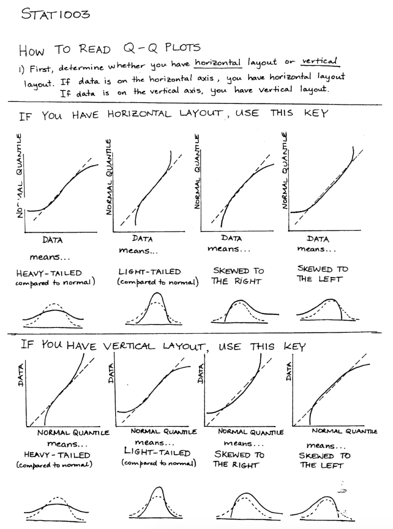
- 3 Assumptions for $\xi_i \sim$ iid $N(0, \sigma^2)$
  1. Independent
  2. Constant variance
  3. Normally distributed
- Residual plots
  - Residuals versus fitted => Check for (1) and (2)
  - Normality/QQ plot => Check for (3)
- Unusual Observation
  1. Potential Outlier
  2. Potential Influential Points

  - Cooks' Distance => Check for both (Mixed Effect)
  - Leverage Plot => Check for the Potential Influential Points
  - Residual versus Fitted => Check for both
  - ...

```
> # Q2(b)
> # 1. Fit a SLR for lcavol(=logY), lpsa (=logX).
>   prostate.lm <— lm(lcavol ~ lpsa)
>   prostate.lm$coef
(Intercept)          lpsa
−0.5086        0.7499
> # 2. Residual plots
> # Three assumptions for error
>   # (1) independent,(2) constant variance,(3) ND
> # Use residual plots to check
>   plot(prostate.lm, which=1) # Residual versus fitted => (1)(2)
>   plot(prostate.lm, which=2) # Normality plot => (3)
> # Unusual observation
>   plot(prostate.lm, which=4) # Cooks Distance (mix effect)
>   barplot(hat(lpsa), main="Leverage plot of the hat values") #
  Leverage (influential point)
>   abline(h=4/length(lpsa))  # Rule:hi>2p/n=2*2/n (SLR),p=#(b0,b1)
```



Question 2(b) Normality Plot



STAT1003

HOW TO READ Q-Q PLOTS

1) First, determine whether you have horizontal layout or vertical layout. If data is on the horizontal axis, you have horizontal layout. If data is on the vertical axis, you have vertical layout.

IF YOU HAVE HORIZONTAL LAYOUT, USE THIS KEY

DATA means... HEAVY-TAILED (compared to normal)

DATA means... LIGHT-TAILED (compared to normal)

DATA means... SKEWED TO THE RIGHT

DATA means... SKEWED TO THE LEFT

IF YOU HAVE VERTICAL LAYOUT, USE THIS KEY

NORMAL QUANTILE means... HEAVY-TAILED (compared to normal)

NORMAL QUANTILE means... LIGHT-TAILED (compared to normal)

NORMAL QUANTILE means... SKEWED TO THE RIGHT

NORMAL QUANTILE means... SKEWED TO THE LEFT

Question 2(c)

```
> # Q2 (c)
> # overall hypothesis (Q2(c))
>   anova(prostate.lm)
Analysis of Variance Table
Response: lcavol
          Df Sum Sq Mean Sq F value    Pr(>F)
lpsa       1 71.938  71.938  111.27 < 2.2e-16 ***
Residuals 95 61.421   0.647
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> qf(0.95,1,length(lcavol)-length(prostate.lm$coef))
[1] 3.941222
> # individual hypothesis (Q2(d))
>   summary(prostate.lm)$coef
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -0.5085802 0.19419311 -2.61894 1.026687e-02
lpsa         0.7499191 0.07109372 10.54832 1.118616e-17
> # correlation test (Q2(a))
>   cor.test(lcavol, lpsa)
Pearson's product-moment correlation

data:  lcavol and lpsa
t = 10.548, df = 95, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6268370 0.8145819
sample estimates:
 cor
0.7344603
```

Question 2(d)

```
> # Q2 (d)
> # Model [B]: lcavol=B0+B1*lpsa+Error. Error ~ iid N(0, sigma^2)
>
> # 1. b0=? b1=? SE(b0)=? SE(b1)=? Hypo(B0,B1)=?
>   summary(prostate.lm)
 Call:
lm(formula = lcavol ~ lpsa)

 Residuals:
Min          1Q     Median         3Q       Max
-2.15948  -0.59383   0.05034   0.50826   1.67751

 Coefficients:
                  Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  -0.50858     0.19419   -2.619    0.0103 *
lpsa          0.74992     0.07109   10.548    <2e-16 ***
---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Residual standard error: 0.8041 on 95 degrees of freedom
 Multiple R-squared:  0.5394, Adjusted R-squared:  0.5346
F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16

>   qt(0.975,length(lcavol)-2)
 [1] 1.985251
>
> # 2. log transformation for (b0)
>   exp(prostate.lm$coef)
 (Intercept)    lpsa
0.6013488   2.1168288
```

CI and PI

- Model (Observation): $Y_i = \beta_0 + \beta_1 X_i + \xi_i$
- CI
  - $\mu(Y_i|X_0) \sim t(b_0 + b_1 X_0, SE(\mu(Y_i|X_0)) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}})$
  - $CI(\mu(Y_i|X_0)) : (b_0 + b_1 X_0) \pm t_{n-2,\alpha/2} SE(\mu(Y_i|X_0))$
  - CI: In repeated sampling, there is 1-α chance **AVERAGE value of Y** lies in CI
- PI
  - $Y_i|X_0 \sim t(b_0 + b_1 X_0, SE(Y_i|X_0) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}}})$
  - $PI(Y_i|X_0) : (b_0 + b_1 X_0) \pm t_{n-2,\alpha/2} SE(Y_i|X_0)$
  - PI: In repeated sampling, there is 1-α chance **SPECIFIC value of Y** lies in PI
- PI (Wider) $>$ CI

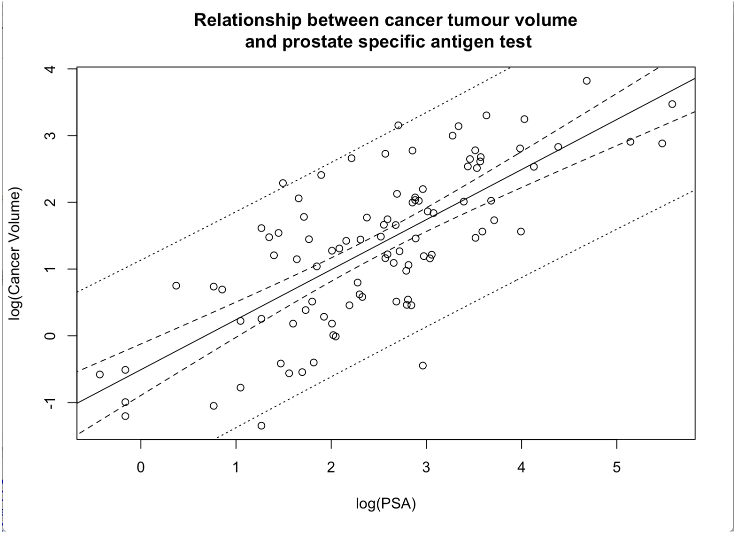
Question 2(e)

```
> # 1. Generate a seqence of Xis
> # check the domain for lpsa
> range(lpsa)
[1] -0.43078  5.58293
> # generate a sequence follows that domain (from -20/20=-1 to 120/
  20=6, the incremental unit is 1/20)
> lpsa.values <- -20:120/20
> lpsa.values
[1] -1.00 -0.95 -0.90 ..
>
> # 2. 95% CI for the mean or expected value of lcavol
> # substitute all 141# lpsa.values to generate CIs
> cintervals <- predict(prostate.lm, newdata=data.frame(lpsa=lpsa.
   values), interval="confidence")
> cintervals[1,]
fit        lwr        upr
-1.2584993 -1.7754976 -0.7415009
>
> # 3. Plot log(psa) versus log(cavol)
> plot(lpsa, lcavol, main="Relationship between cancer tumour volume\n
    and prostate specific antigen test", xlab="log(PSA)", ylab="log(
  Cancer Volume)")
> # Generate the fitted line: lcavol hat = b0 + b1 * lpsa
> abline(prostate.lm$coef)
> # Generate the CI's UB and LB each
> lines(lpsa.values, cintervals[,"lwr"], lty=2)
> lines(lpsa.values, cintervals[,"upr"], lty=2)
```

Question 2(e)

```
> # 4. Comment:
> # based on (a)(hypo for correlation), (c)(hypo for overall),
> # (d)(hypo for b1), tight CI
> # => B1 is significant => when log psa rises ~ log cavol rises
> # => psa rises ~ cavol rises
>
> # A lot of observations lie outside the CIs
> # => a lot of variability around this increasing relationship
> # => so PSA is not necessarily a reliable indicator of tumour size.
>
> # CI for mean value of y, PI for an individual value of y.
> # PI>CI => use PI to check
>
> # 5. PI for 141# of new Xis
> pintervals <- predict(prostate.lm, newdata=data.frame(lpsa=lpsa.
  values), interval="prediction")
> pintervals[1,]
      fit        lwr        upr
-1.2584993 -2.9364232  0.4194247
> lines(lpsa.values, pintervals[,"lwr"], lty=3)
> lines(lpsa.values, pintervals[,"upr"], lty=3)
```

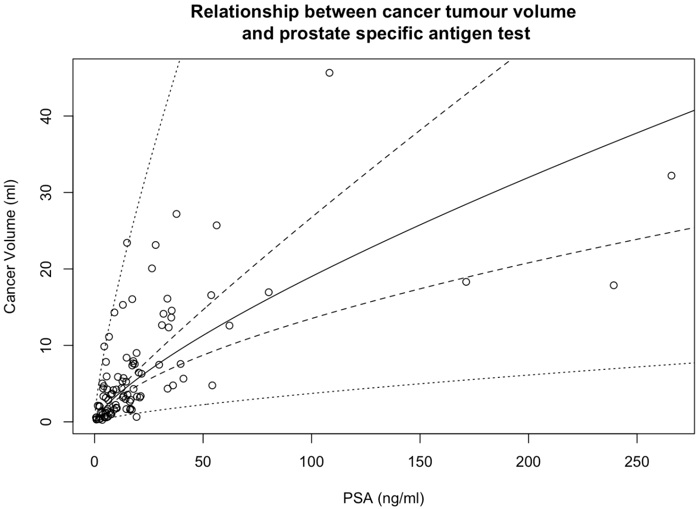
Question 2(e)


Relationship between cancer tumour volume
and prostate specific antigen test
log(Cancer Volume)
-1
0
1
2
3
4
0
1
2
3
4
5
log(PSA)

Question 2(e)

```
> # 6. Transformation
> # Scatterplot: pas versus cavol
> plot(exp(lpsa), exp(lcavol), main="Relationship between cancer
  tumour volume\n and prostate specific antigen test", xlab="PSA (ng/
  ml)", ylab="Cancer Volume (ml)")
> # Generate the fitted line, CI, PI based on Xi = psa (not lpsa) (
  based on exp of 141# of lpsa)
> lines(exp(lpsa.values), exp(cintervals[,"fit"]))
> lines(exp(lpsa.values), exp(cintervals[,"lwr"]), lty=2)
> lines(exp(lpsa.values), exp(cintervals[,"upr"]), lty=2)
> lines(exp(lpsa.values), exp(pintervals[,"lwr"]), lty=3)
> lines(exp(lpsa.values), exp(pintervals[,"upr"]), lty=3)
> legend(164, 4, c("SLR Model on log-log scale", "95% Confidence
  Intervals", "95% Prediction Intervals"), lty=1:3)
```

Question 2(e)


Relationship between cancer tumour volume
and prostate specific antigen test
Cancer Volume (ml)
0
10
20
30
40
PSA (ng/ml)
0
50
100
150
200
250