PART - 1 of the GPA problem:

```
gpa = read.table("D:\\ASU Stuff\\SEM-1\\STP 530\\120 students.txt")
Y = gpa$V1
X = gpa$V2
head(gpa)
```

```
##      V1 V2
## 1 3.897 21
## 2 3.885 14
## 3 3.778 28
## 4 2.540 22
## 5 3.028 21
## 6 3.865 31
```

```
lm.gpa = lm(Y~X)
lm.gpa
```

```
## 
## Call:
## lm(formula = Y ~ X)
## 
## Coefficients:
## (Intercept)            X  
##     2.11405      0.03883
```

(a)

```
predict.lm(lm.gpa, newdata = data.frame(X=28), interval="confidence", 
level=0.95)
```

```
##        fit      lwr      upr
## 1 3.201209 3.061384 3.341033
```

Thus, we can see that the mean freshmen GPA for students having ACT=28 is 3.2012 and the 95% confidence interval is (3.0613,3.34103).

(b)

```
predict.lm(lm.gpa, newdata = data.frame(X=28), interval="prediction", 
level=0.95)
```

```
##        fit      lwr      upr
## 1 3.201209 1.959355 4.443063
```

Her freshman GPA should be 3.2012 and the 95% prediction interval is (1.9593,4.4430).

(c)

Yes, the prediction interval in part(b) is wider than confidence interval in part(a) as we are predicting the distribution of only a single point and not the mean.

(d)

This seems to be saying about the variation around the line again i.e. the mean and not the new point. So, this is basically same as part(a).

PART - 2 of the GPA problem:

(a)

```
anova(lm.gpa)

## Analysis of Variance Table
##
## Response: Y
##            Df Sum Sq Mean Sq F value   Pr(>F)
## X           1  3.588  3.5878  9.2402 0.002917 **
## Residuals 118 45.818  0.3883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b)

$s^2 + b_1^2 a(X_i - \overline{x})^2$ is estimated by Mean Square due to regression (MSR) and s^2 is estimated by Mean Square due to errors (MSE). When $\beta_1 = 0$, MSR and MSE estimate the same quantity.

(c)

$H_0 : \beta = 0; H_1 : \beta \neq 0$ From the table, it is evident that the value of F*= MSR/MSE = 9.24 and F(0.99,1,118) = 6.855

For the α risk,

Decision Rule: If $F^2 \leq F$ (0.99,1,118), conclude H_0 or else conclude H_1.

Here, $F^* > F$ and thus we will conclude H_1. Therefore, $\beta_1 \leq 0$.

(d)

SSR = 3.588 is the absolute magnitudeof the reduction in the variation of Y when X is introduced into the regression model.

The relative reduction is 3.588/49.406 = 7.262%.

The name of the latter measure is **coefficient of determination**.

(e)

$r = +\sqrt{0.07262} = +0.2695$

(f)

R^2 has the more clear-cut interpretation as it accounts for the percentage of variation in Y explained by X.

PART - 3 Muscle mass problem:

```
musclemass = read.table("D:\\ASU Stuff\\SEM-1\\STP 530\\60 students.txt")
head(musclemass)
```

```
##    V1 V2
## 1 106 43
## 2 106 41
## 3  97 47
## 4 113 46
## 5  96 45
## 6 119 41
```

```
X1 = musclemass$V2
Y1 = musclemass$V1
lm.musclemass = lm(Y1~X1)
lm.musclemass
```

```
##
## Call:
## lm(formula = Y1 ~ X1)
##
## Coefficients:
## (Intercept)           X1
##      156.35        -1.19
```

(a)

```
predict.lm(lm.musclemass, newdata = data.frame(X1=60), interval="confidence",
level=0.95)

##        fit      lwr      upr
## 1 84.94683 82.83471 87.05895
```

Thus, it can be clearly seen that the 95 percent confidence interval for the mean muscle for women aged 60 is (82.83471,87.05895).

(b)

```
predict.lm(lm.musclemass, newdata = data.frame(X1=60), interval="prediction",
level=0.95)

##        fit      lwr     upr
## 1 84.94683 68.45067 101.443
```

The 95% prediction interval is (68.45067,101.443) and no, the interval is not very precisebecause it's much wider than confidence interval in part(a) as we are predicting the distribution of only a single point and not the mean.

(c)

This seems to be saying about the variation around the line again i.e. the mean and not the new point. So, this is basically same as part(a).

PART - 4 Continuation of muscle mass problem:

(a)

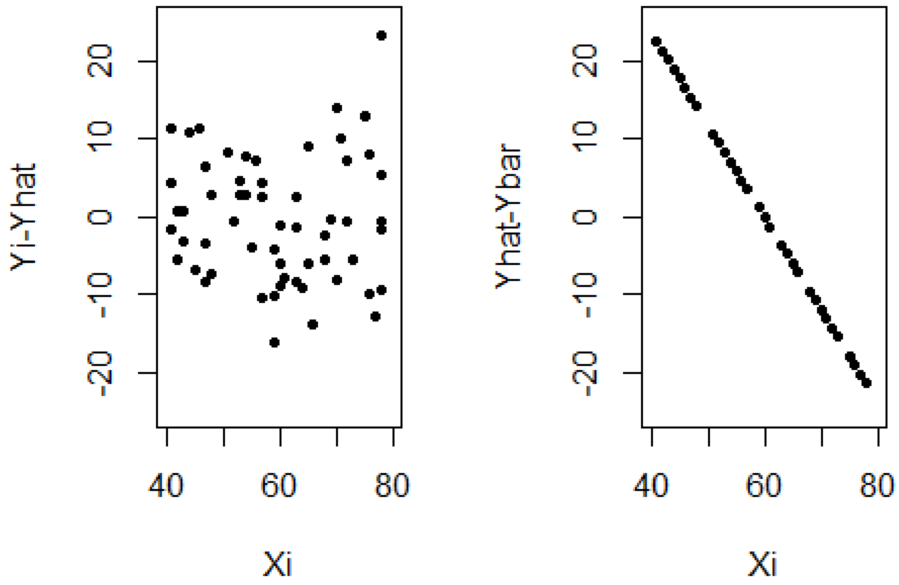
```
lm.musclemass

##
## Call:
## lm(formula = Y1 ~ X1)
##
## Coefficients:
## (Intercept)           X1
##      156.35        -1.19

Y1.mean = mean(Y1)
Y1hat=156.35-1.19*X1

par(mfrow=c(1,2))
plot(X1,Y1-Y1hat, xlab="Xi", ylab="Yi-Yhat", xlim=c(40,80), ylim=c(-
25,25),pch=20)

plot(X1,Y1hat-Y1.mean, xlab="Xi", ylab="Yhat-Ybar", xlim = c(40,80),ylim =
c(-25,25),pch=20)
```



From the graphs, it is apparent that the residuals $Y_i - \hat{Y}_i$ are much smaller than the deviations $Y_i - \overline{y}$ of the predictors from the mean. Thus the SSM should be much larger than the SSE, and we expect R^2 =SSM/SST to be close to 1.

(b)

```
anova(lm.musclemass)

## Analysis of Variance Table
##
## Response: Y1
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1 11627.5 11627.5  174.06 < 2.2e-16 ***
## Residuals 58  3874.4    66.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(c)

$H_0{:}\,\beta = 0$; $H_1{:}\,\beta \neq 0$

Value of $F^* = MSR/MSE$ = 174.06

If $F^* > F(0.95; 1{,}58)$, reject H_0 ; If $F^* < F(0.95; 1{,}58)$, accept H_0.

Here, $F^* = 174.06 > F(0.95; 1{,}58)$ and thus we reject H_0 i.e. $\beta = 0$ and conclude that there is a linear relationship between muscle mass and age.

(d)

$R^2 = SSR/SST = 0.7501$ and thus 1-R^2=0.2499

Therefore, the proportion of the total variation that remains "unexplained" when age is introduced is 24.99% and 75% of the variation is explained, the proportion of unexplained variation is relatively small.

(e)

Because of the negative relationship between mass and age,

r = -$\sqrt{R^2}$= -0.866

PART-5 Grade point average problem

(a)

It would be more reasonable if we consider the X_i as constants because actually director of admissions can decide the minimum ACT score to enter the university. Had the university been selecting students independent of the ACT scores, then we can argue that X_i is a random variable.

(b)

If the X_i were considered to be random variables, there would be modifications in interpretation of confidence coefficients and specified risks of error.

PART - 6 Water Flow:

(a)

install.packages("psychometric")

```
library(psychometric)

## Loading required package: multilevel

## Loading required package: nlme

## Loading required package: MASS

CIr(0.83,147,level = 0.99)

## [1] 0.7502313 0.8859530
```

The 99% confidence interval for the coefficient of correlation is (0.7502,0.8859).

(b)

For the confidence interval of ρ_{12}^2, we just square the lower and upper limits of ρ_{12}.

Hence, lower limit = $(0.7502)^2$ = 0.5628 & Upper limit = $(0.8859)^2$ = 0.7848

Thus, the 99% confidence interval for the ρ_{12}^2 is (0.5628,0.7848).

PART - 7 Muscle mass problem continuation:

(a)

```
cor(X1,Y1,method = "pearson")

## [1] -0.866064
```

Thus, the value of pearson product-moment correlation coefficient r_{12} is -0.86606.

(b)

$H_0 : \rho_{12} = 0$, $H_a : \rho_{12} \neq 0$

Here, $t^* = (-0.86606\sqrt{58})/\sqrt{1-(-0.86606)^2}$ = -13.193.

Also, t(0.975;58) = 2.00172.

If $|t^*| \leq$ 2.00172, conclude H_0, otherwise conclude H_a.

Finally, 13.193 > 2.00172 and thus we conclude H_a.

(c)

```
cor(X1,Y1,method = "spearman")

## [1] -0.8657217
```

Thus, the Spearman rank correlation coefficient is r_s is -0.8657.

(d)

H_0: There is no association between X and Y.

H_a: There is an association between X and Y.

$t^* = (-0.8657\sqrt{58})/\sqrt{1-(-0.8657)^2}$ = -13.1723 and t(0.975;58) = 2.00172.

If $|t^*| \leq$ 2.00172, conclude H_0 or else conclude H_1.

Fianlly, 13.1723 > 2.00172 and therefore we conclude H_a.

(e)

It can be seen from both pearson as well as spearman test that the value of pearson and spearman's coefficient is almost same. There is not much of a difference in the value of test statistic and we have concluded the alternate hypothesis in both the cases.