Statistics Coursework 2

Q1:

We have two populations, the male athletes and female athletes respectively. We assume the sample of heights of each is random, following a normal distribution and that the populations are independent.

Test: H_0: $\sigma_{male}^2=\sigma_{female}^2$ vs. H_1: $\sigma_{male}^2\neq\sigma_{female}^2$

Under the null hypothesis, the test statistic is given by $F=s_{male}^2/s_{female}^2 \sim F_{n(male)-1,n(female)-1}$

Take the significance level to be $\alpha=0.05$

We have that $s_{male}=10.3525$ and $s_{female}=9.08438$ and $n_{male}=6997$, $n_{female}=5755$.

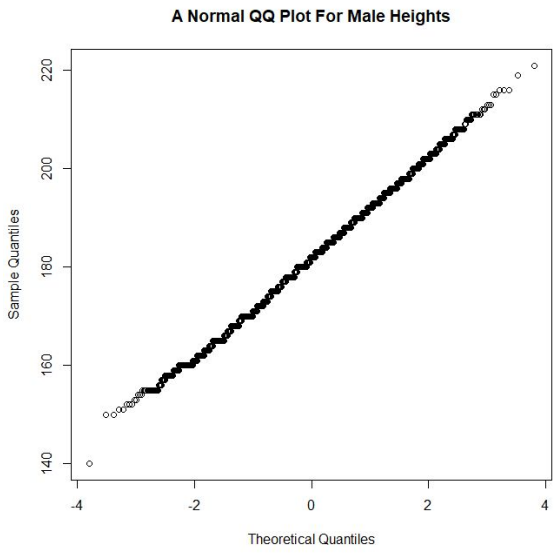
We reject H_0 if $F>F_{6996,5754,0.025}=1.05$ or if $F<F_{6996,5754,0.975}=1/F_{5754,6996,0.025}=0.952$(to 3 s.f).

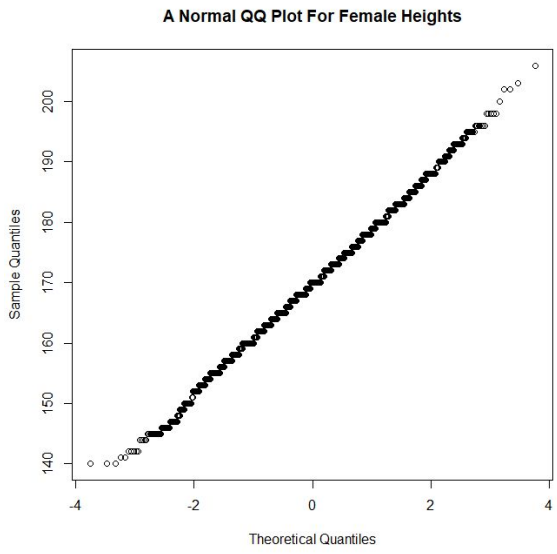
Here $F=10.3525^2/9.08438^2=1.30$(to 3 s.f)$>1.05$

Thus we reject H_0 at the 5% significance level, there is evidence to suggest that σ_{male}^2 is different to the female variance. Since our variance for male athlete height here is greater than the female variance, this supports the greater male variability theory which states that males tend to have a larger variance in traits than females do.

Assumptions: We assumed that the sample of heights were random, however I used all the data available as I was able to do so using R.

We also assume that the heights of each gender follow a normal distribution, to check this assumption is reasonable I plotted a normal QQ plot for each gender, as presented below.





It is quite clear that the heights of each may be modelled accurately by a normal distribution since the QQ plots show that the theoretical quantiles line up closely to the sample quantiles, and the sample sizes of each were huge (central limit theorem).

Lastly we assumed that the two distributions were independent. To test this, I considered the correlation coefficient $r=s_{xy}/s_x s_y$ with the knowledge from the probability module that the correlation coefficient is 0 if and only if the covariance is zero. The distributions being independent implies that the covariance is zero.
s_{xy}=17.59184 (setting x as male heights and y as female heights, and using a sample of 50 from each)
s_x=10.84586
s_y=8.870957
r=0.183(3 s.f.)
This is almost zero, so I took a larger sample from each of 5000 and found that the covariance was -0.9967293, standard deviation for males was 10.39887 and 9.132889 for females. Using the formula above this gives r=-0.0105(3 s.f). This suggests strongly that r=0 and that the two distributions are independent.

Q2:
(a):
Test: H_0:μ=15 vs H_1:μ≠15
We are given that σ=0.25, n=3 and I calculated that $\bar{X}$=15.583.
Take α=0.05.
Reject H_0 if $\left|\frac{x-\mu}{\sigma\div\sqrt{n}}\right| > z_{0.025}$=1.96
The test statistic is $\left|\frac{15.583-15}{0.25\div\sqrt{3}}\right|$=4.04>1.96 thus we reject H_0 at the 5% significance level, meaning that we reject that the true mean parts per billion of lead in drinking water is 15.
For the p-value:
p-value=P($|x| \geq 15.583$ |X~N(15,0.25^2/3))
=P($\left|\frac{x-\mu}{\sigma\div\sqrt{n}}\right| \geq \frac{15.583-\mu}{\sigma\div\sqrt{n}}$ |X~N(15,0.25^2/3))
=P($|z|$ >4.039|z~N(0,1))
=2P(z<-4.039)
=$5.36\text{x}10^{-5}$
Hence since p<0.01 there is strong evidence against the null hypothesis.
(b):
Considering α=0.01:

$\left|\frac{x-15}{0.25\div\sqrt{3}}\right|$ >2.5758=$\mathcal{Z}_{0.005}$

$\frac{x-15}{0.25\div\sqrt{3}}$ >2.5758

x>15+2.5758($\frac{0.25}{\sqrt{3}}$)=15.37

$\frac{x-15}{0.25\div\sqrt{3}}$ <-2.5758

x<15-2.5758($\frac{0.25}{\sqrt{3}}$)=14.63

So H_0 rejected for sample means x>15.37 or x<14.63

(c):

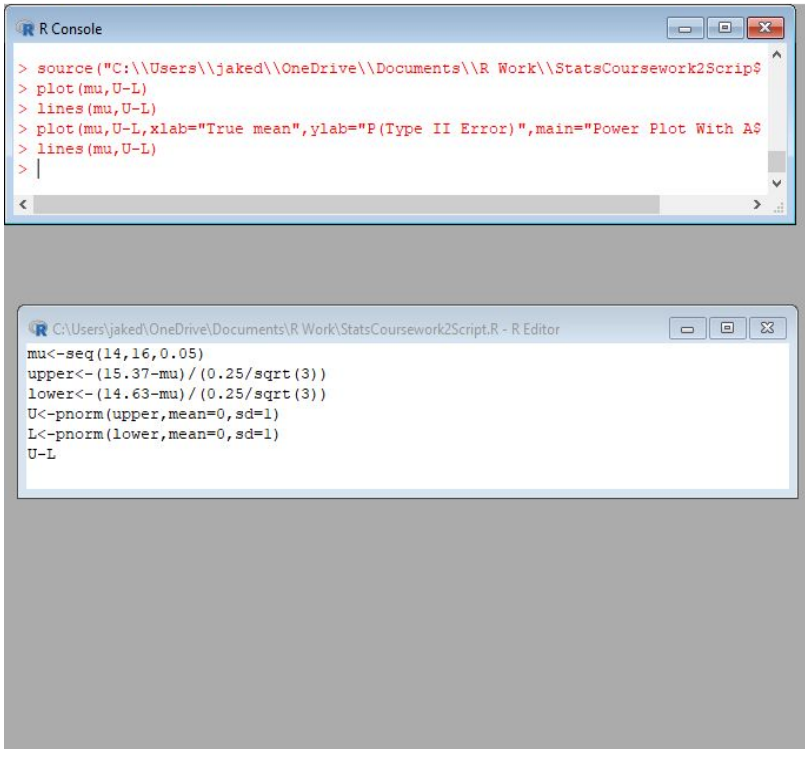
The probability of a type II error is P(accept H_0|H_0 false).

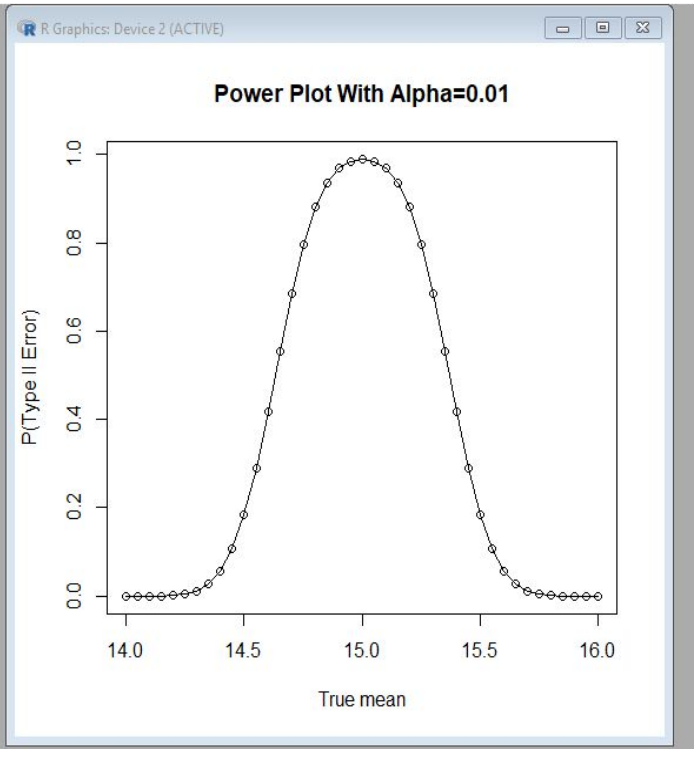
That is, in terms of our problem, letting μ_1 be the true mean, given by P(14.63<x<15.37|$\mu=\mu_1$)

=P($\frac{14.63-\mu}{(\frac{0.25}{\sqrt{3}})}$ <z< $\frac{15.37-\mu}{(\frac{0.25}{\sqrt{3}})}$ |z~N(0,1))

= Φ($\frac{15.37-\mu}{(\frac{0.25}{\sqrt{3}})}$) - Φ($\frac{14.63-\mu}{(\frac{0.25}{\sqrt{3}})}$)

Using values of μ_1 in the range of 14-16, calculating this and plotting the probability of type II error against values of μ_1 produces the power plot below (the R codes used presented alongside in image).





If instead a significance level of α=0.05 was used, the region in which H_0 accepted decreases and so I would expect the power plot to have a narrower peak. This also follows from the converse of the fact that the probability of a type II error increases when the significance level is decreased.

(d):

Power=1-P(Type II Error)=0.8

P(Type II Error)=0.2=P(accept $H_0|\mu_1$=15.3)

We accept H_0 if $\left|\frac{x-15}{\frac{0.25}{\sqrt{n}}}\right|$ <2.5758 which upon rearranging and using the reverse triangle inequality gives us $|x|$ <15+ $\frac{0.64395}{\sqrt{n}}$.

P($|x|$ <15+ $\frac{0.64395}{\sqrt{n}}$ |μ_1=15.3)=P(-15- $\frac{0.64395}{\sqrt{n}}$ <x<15+ $\frac{0.64395}{\sqrt{n}}$ |μ_1=15.3)

=P((-30.3- $\frac{0.64395}{\sqrt{n}}$)/ $\frac{0.25}{\sqrt{n}}$ <z<2.5758-1.2$\sqrt{n}$)=0.2

(where z~N(0,1), we ignore the lower value because its way into the lower tail and so P(z<-30.3)$\approx$0)

$\approx$P(z<2.5758-1.2$\sqrt{n}$)=0.2 (now find that P(z<-0.8416212)=0.2)

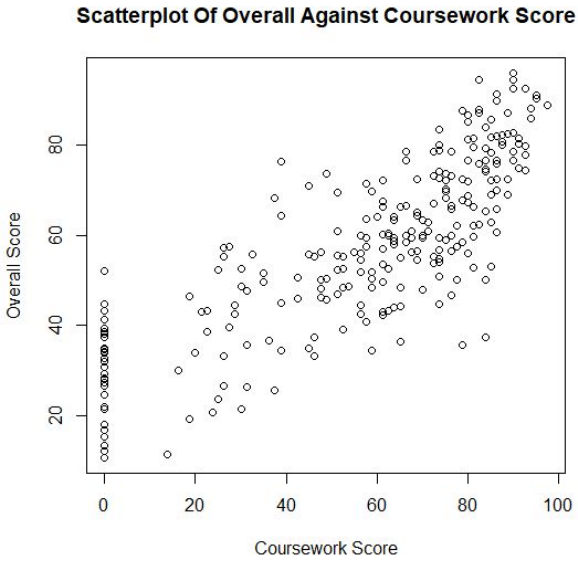
2.5758-1.2$\sqrt{n}$ =-0.8416212 (rearrange for $\sqrt{n}$ and square)

n=8.11 (3 s.f) hence for a power of 80%, n>8.11. Therefore n=9 will achieve this.

Q3

(a):

There is a clear positive correlation between score in coursework and overall, as better performers in coursework had higher overall scores and vice versa. Those who got zero in their coursework had a maximum overall score of below 60.



(b):

	Overall score (y)	Coursework score (x)
Mean	57.47708	57.06181
Variance (s^2)	374.9821	835.2182

And the value of s_{xy} is 457.4966.

(c):

The sample correlation coefficient is $r=s_{xy}/(s_x^2 s_y^2)^{1/2}$

$=457.4966/(374.9821 \times 835.2182)^{½}=0.817$ (to 3s.f.)

Test: H_0: $\rho=0$ vs H_1:$\rho>0$

Assume (x,y) has a bivariate normal distribution. Choose $\alpha=0.001$.

Reject H_0 if $t>t_{269,0.0005}=3.33$(3 s.f.)

$t=0.817(269/(1-0.817^2))^{½}=23.3>3.33$(3 s.f) so we reject H_0, there is strong evidence of a positive correlation between the coursework scores and overall scores, as the test statistic is well into the tail of the t distribution.

(d):

We wish to find a suitable linear regression model for these data, since there is strong evidence of a linear relationship between the coursework and overall scores as shown from the sample correlation coefficient of 0.817, we can model the data with the linear equation $y_i=\beta_0+\beta_1 x_i+\epsilon_i$ for i=1,2,...,n

We first estimate the slope of the line by $\hat{\beta_1} =s_{xy}/s_x^2=457.4966/835.2182=0.548$(3 s.f.).

The intercept of the line is given by

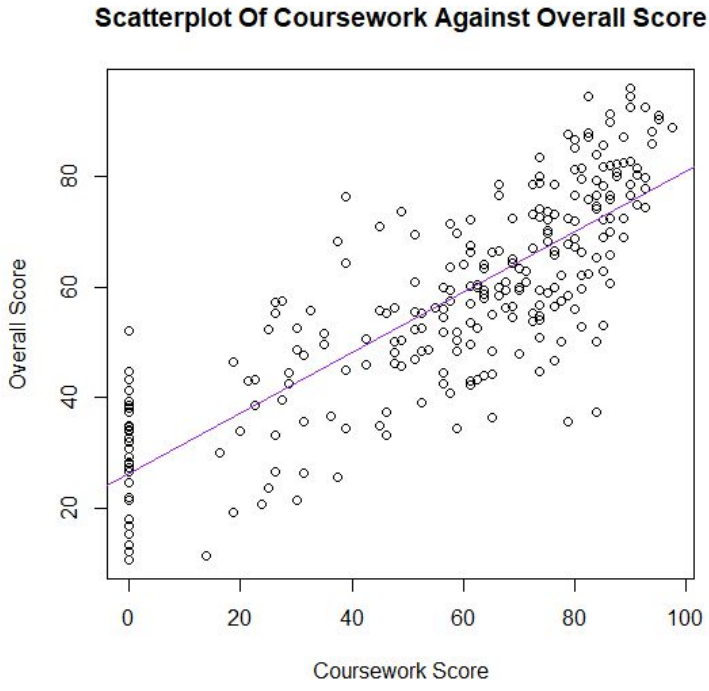
$\hat{\beta_0}=y_{mean}-\hat{\beta_1} x_{mean}=57.47708-(0.548 \times 57.06181)=26.2$(3 s.f)

Thus our estimated model is given by $\hat{y}=26.2+0.548x$, where x is a known x value we wish to use to predict the value of y, this estimate being denoted $\hat{y}$.

This model assumes that the unobservable random errors ϵ_i have zero mean, are uncorrelated and have the same variances.

(e):

As above our regression model is $\hat{y}=26.2+0.548x$. This line is fitted onto the plot as shown below by the purple line.



(f):

The residuals are given by the difference between the y values given in the data and the fitted y values from our regression model:

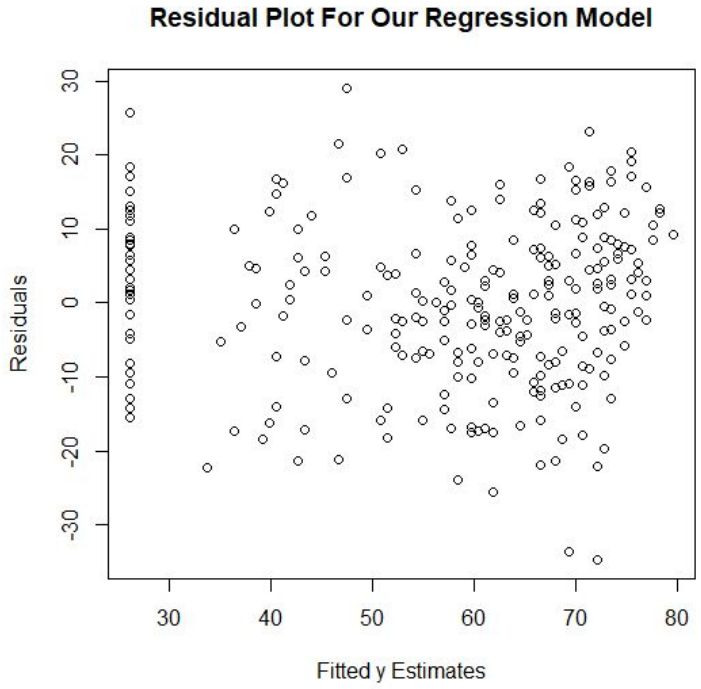
R Console

```
> residuals
  [1] -22.315 -15.530 -14.200 -12.870 -10.870  -9.530  -8.200 -17.395  -4.870 -18.465
 [11]  -4.870 -21.310  -4.200 -16.230  -1.530  -1.530 -21.250 -17.075   1.130   1.130
 [21]   0.470 -14.005   2.130   1.800   3.130  -5.185   4.470   4.470   5.800   5.800
 [31]   5.800 -18.295  -7.335   6.470 -13.015 -23.975   7.800  -3.160   7.970 -15.860
 [41]   8.470   8.800 -25.490  -7.745 -33.605  11.130 -14.295 -34.675  -9.485  11.800
 [51]  13.130  12.470  -1.770  -0.030 -15.800  15.130 -16.880   0.465 -17.515  17.130
 [61] -14.445   4.640 -16.845 -17.280 -16.845   5.075   2.465 -17.055 -17.490  -2.355
 [71] -12.445  18.470 -21.865  -6.060   9.945  -3.660  -7.165 -21.405  -7.365  -9.975
 [81] -13.490  -4.060   4.255 -16.560   6.030  -6.905  -6.470  -2.495 -10.185  -7.975
 [91]  -2.060 -22.015   4.290 -18.500   1.010 -15.865  -6.645  -2.035   6.290  25.800
[101]  12.430  -5.105 -17.805 -19.780  -7.950  -2.470  10.030  -9.455 -12.535  -6.185
[111] -12.100  14.665   0.200   3.705 -10.760  -6.820 -11.865  -2.445   3.940 -14.040
[121]  -7.455  -0.010   4.810  -1.105  11.820   1.295  -7.020  -9.865  16.665 -11.405
[131] -11.405  -2.845 -10.935  16.230  -3.055  -0.210 -11.170  -3.925  -2.385  -8.300
[141]  -5.230  -7.195  -2.385  -1.715   2.895   0.485   1.790  -8.065  -0.620  -4.560
[151]  -2.585 -11.145  -3.690   6.635   0.050 -12.885  -4.325  -2.360  -6.500  -8.475
[161]  -1.230  -9.780  -8.910  -2.325   2.285   2.945   4.920   0.545  16.985   5.790
[171]   1.205  -6.675   1.205  -7.545  -4.475  -2.065   4.510   6.485   1.240  -1.405
[181]   4.075  -2.710  -1.605  21.420   7.815   1.030  -3.780  -1.370  -5.755   2.370
[191]  -3.545   3.030  15.295   2.370  11.355  20.140  -0.450  13.790   3.065   1.960
[201]   8.545   5.030  12.485  -2.415   7.240   5.265   6.135  -0.885   1.985   6.370
[211]  20.835   7.465   2.655  -1.285  -2.390   4.420   2.455  28.985   4.655   3.115
[221]   1.150   6.630  14.075   0.940   5.550  10.595   7.325  12.570  16.075  12.135
[231]   3.150   4.045   2.940  13.465   6.020   8.855   8.855   6.680  11.290   5.375
[241]   7.585   8.020   8.890  10.855   8.455   7.150   7.150  16.805   7.150  11.985
[251]  15.290   8.505  12.890  16.630  15.760  12.245  15.760  18.395  18.395  10.505
[261]  16.420   9.200  16.455  12.070  12.740  17.785  17.150  15.610  23.090  19.150
[271]  20.480
>
```

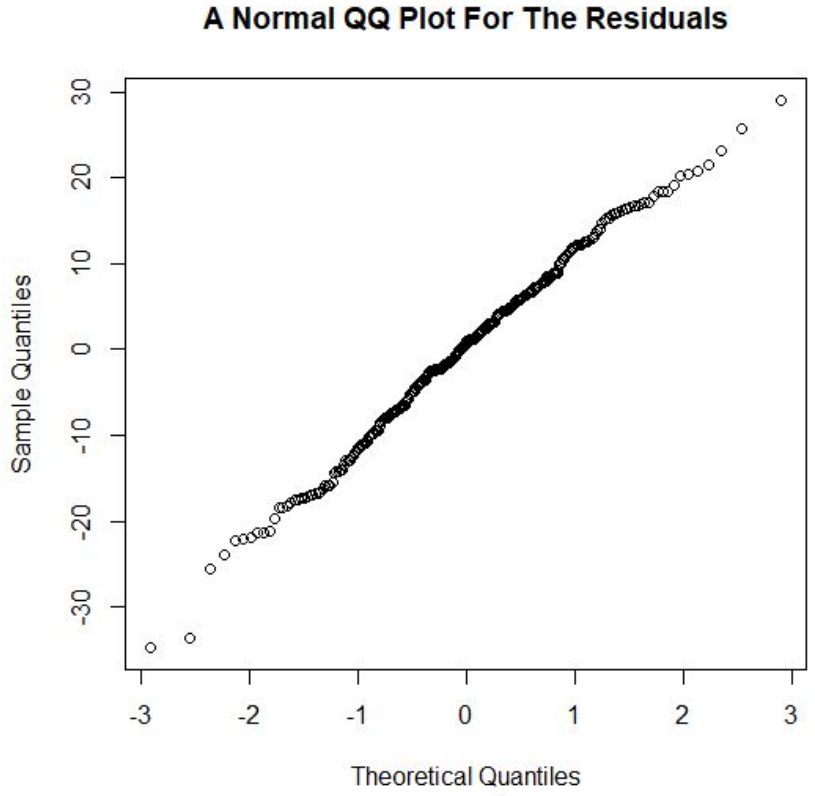
C:\Users\jaked\OneDrive\Documen...

```
x<-coursework
fittedY<-26.2+0.548*x
residuals<-overall-fittedY
```

A residual plot for the data is shown below.



A QQ plot for the residuals is also shown below, which shows a linear relationship between the theoretical and sample quantiles and thus means that we may model the residuals approximately by a normal distribution.



(g):

(i)

Firstly we need to calculate an unbiased estimate for the variance of the errors ϵ_i, given by $\hat{\sigma}^2$:

$\hat{\sigma}^2=(270/269)(374.9821-0.548^2 28.9^2)=124.6$ ($\hat{\sigma}=11.2$ to 3 s.f.)

Now calculate the 95% prediction interval for fitted value when $x_0=70$.

$\hat{\beta}_0+\hat{\beta}_1 x_0 \pm t_{n-2,\alpha/2}\hat{\sigma}(1+1/n+(x_0-x_{mean})^2/(n-1)s_x^2)^{1/2}$

$=26.2+(0.548x70)\pm t_{269,0.975}\hat{\sigma}(1+1/271+(70-57.1)^2/270(28.9^2))^{1/2}$

$=64.56\pm 1.968822x(11.2(1+1/271+(70-57.1)^2/270(28.9^2))^{1/2}$

$=64.56\pm 22.1$

$=(42.5,86.7)$

(ii)

We are given that the individual predicted $\hat{y}$ has a normal distribution. The mean of the distribution will be the fitted value at $x_0=60$. This is given by 26.2+0.548(60)=59.1(3s.f)

from our regression model. The variance is given by $\sigma^{\wedge 2}(1+1/n+(x_0-x_{mean})^2/(n-1)s_x^2)$ which is equal to 125 (3s.f.).
This gives us $y^{\wedge}$~N(59.1,125).
To find the probability that the student passes, need to find probability that y^ is greater than or equal to 40 (the pass score).
$P(y^{\wedge} \geq 40)$=
$P((y^{\wedge}-59.1)/\sqrt{125} \geq (40-59.1)/\sqrt{125})$= (this standardises variable to Z~N(0,1))
$P(z \geq -1.71)=1-P(z<-1.71)=1-0.04363294=0.956$ (3 s.f.).
Therefore a student who scores 60 on coursework has approximately 95.6% chance of passing the module overall.

Appendix:
Q1:
Test:

```
>F<-subset(dat,Sex=="F")
> M<-subset(dat,Sex=="M")
> FemaleHeight<-F[,5]
> MaleHeight<-M[,5]
>sd(na.omit(FemaleHeight))
>sd(na.omit(MaleHeight))
> length(na.omit(FemaleHeight))
> length(na.omit(MaleHeight))
> qf(.975,df1=6996,df2=5754)
> qf(.975,df1=5754,df2=6996)
```

Assumptions:

```
>qqnorm(na.omit(MaleHeight),main=”A Normal QQ Plot For Male Heights”)
>qqnorm(na.omit(FemaleHeight),main=”A Normal QQ Plot For Male Heights”)
>MHeightSample<-sample(na.omit(MaleHeight),50,replace==FALSE, prob=NULL)
>FHeightSample<-sample(na.omit(FemaleHeight),50,replace==FALSE, prob=NULL)
>sd(MHeightSample)
>sd(FHeightSample)
>cov(MHeightSample,FHeightSample)
> MHeightSample<-sample(na.omit(MaleHeights),5000,replace=FALSE,prob=NULL)
#Much larger sample to test the independence
> FHeightSample<-sample(na.omit(FemaleHeights),5000,replace=FALSE,prob=NULL)
> cov(MHeightSample,FHeightSample)
```

```
> sd(MHeightSample)
> sd(FHeightSample)
```

Q2:

```
 >pnorm(-4.039,mean=0,sd=1)
#The code for the power plot is given with the power plot in ©
>qnorm(0.2,mean=0,sd=1) #Code for 2(d)
```

Q3:

(a)

```
> plot(overall~coursework,xlab="Coursework Score",ylab="Overall
Score",main="Scatterplot Of Overall Against Coursework Score")
```

(b)

```
> mean(coursework)
> mean(overall)
> sd(overall)
> sd(coursework)
> cov(overall,coursework)
```

(c)

```
> qt(0.9995,269)
```

(e)

```
>abline(a=26.2,b=0.548,col=”Purple”)
```

(f)

```
>plot(residuals~fittedY, main=”Residual Plot For Our Regression Model”, xlab=”Fitted y
Estimates”,ylab=”Residuals”)
>qqnorm(residuals,main=”A Normal QQ Plot For The Residuals”)
```

(g)(i)

```
>length(residuals)
>qt(0.975,269)
```

(g)(ii)

```
> pnorm(-1.71,mean=0,sd=1)
```