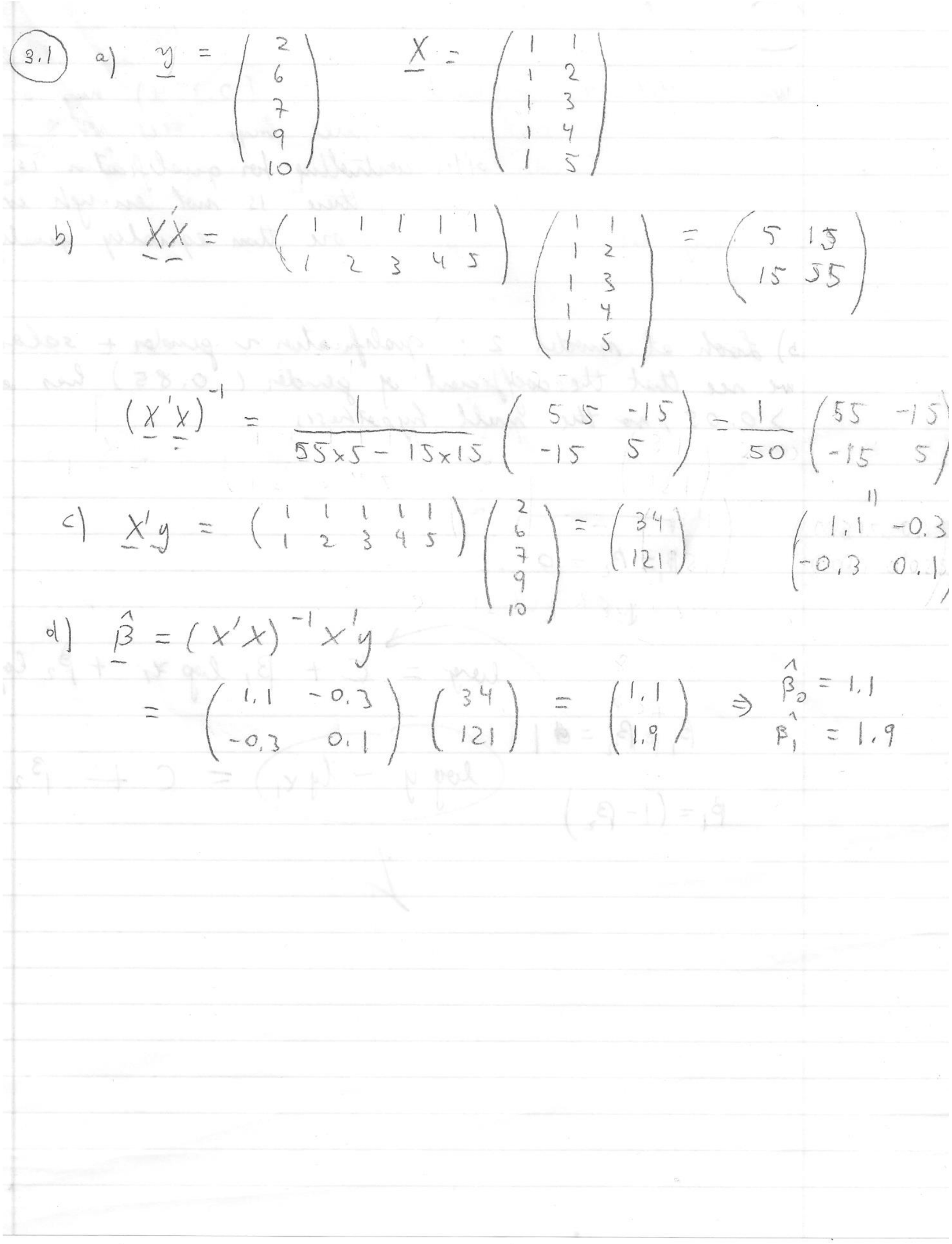
Problem 1 – 3.1



3.1 a) $\underline{y} = \begin{pmatrix}2\\6\\7\\9\\10\end{pmatrix}$ $\underline{X} = \begin{pmatrix}1&1\\1&2\\1&3\\1&4\\1&5\end{pmatrix}$

b) $\underset{\sim}{X}'\underset{\sim}{X} = \begin{pmatrix}1&1&1&1&1\\1&2&3&4&5\end{pmatrix}\begin{pmatrix}1&1\\1&2\\1&3\\1&4\\1&5\end{pmatrix} = \begin{pmatrix}5&15\\15&55\end{pmatrix}$

$(\underline{X}'\underline{X})^{-1} = \frac{1}{55\times5 - 15\times15}\begin{pmatrix}55&-15\\-15&5\end{pmatrix} = \frac{1}{50}\begin{pmatrix}55&-15\\-15&5\end{pmatrix}$

$\parallel$

$\begin{pmatrix}1.1&-0.3\\-0.3&0.1\end{pmatrix}$

c) $\underline{X}'y = \begin{pmatrix}1&1&1&1&1\\1&2&3&4&5\end{pmatrix}\begin{pmatrix}2\\6\\7\\9\\10\end{pmatrix} = \begin{pmatrix}34\\121\end{pmatrix}$

d) $\underline{\hat{\beta}} = (X'X)^{-1}X'y$

$= \begin{pmatrix}1.1&-0.3\\-0.3&0.1\end{pmatrix}\begin{pmatrix}34\\121\end{pmatrix} = \begin{pmatrix}1.1\\1.9\end{pmatrix} \Rightarrow \hat{\beta}_0 = 1.1$

$\hat{\beta}_1 = 1.9$

Problem 2 – 3.2

Part a.

```
> x1 = c(-1,-1,1,1)
> x2 = c(1,-1,1,-1)
> X = cbind(rep(1,4),x1,x2,x1*x2)
> colnames(X)[4]="x1x2"
> y = c(110,120,130,150)
>
> ## Part a)
>
> X
       x1 x2 x1x2
[1,] 1 -1  1   -1
[2,] 1 -1 -1    1
[3,] 1  1  1    1
[4,] 1  1 -1   -1
```

Part b.

```
> # X'X
> t(X)%*%X
       x1 x2 x1x2
     4  0  0    0
x1   0  4  0    0
x2   0  0  4    0
x1x2 0  0  0    4
>
> # inv(X'X)
> solve(t(X)%*%X)
            x1   x2 x1x2
     0.25 0.00 0.00 0.00
x1   0.00 0.25 0.00 0.00
x2   0.00 0.00 0.25 0.00
x1x2 0.00 0.00 0.00 0.25
>
> # beta
> solve(t(X)%*%X)%*%t(X)%*%y
      [,1]
     127.5
x1    12.5
x2    -7.5
x1x2  -2.5
```

The inverse is easy to calculate because X'X is a diagonal matrix as a result of orthogonality, so the inverse is a diagonal matrix with the inverse (e.g 1/4 in this case) of its corresponding entries.

Part c.

Reference: http://www.ats.ucla.edu/stat/mult_pkg/faq/general/effect.htm

The equation is y = 127.5 + 12.5x1 -7.5x2 -2.5x1x2

The intercept represents the overall mean of BP. The overall mean is 127.5

```
> # check intercept
> mean(y)
[1] 127.5
```

With (-1,0,+1) coding, the coefficients represent the distance between the factor levels and the overall mean. Thus, the coefficient b1 of x1 represents the effect of x1 (age) when x2 = 0 (gender). Since x2 = 0 at the mean of the two categories of x2, b1 is a main effect. It's the effect of x1 at the mean value of x2. More specifically, the coefficients of each of the effect variables is equal to the difference between the mean of the group coded 1 and the grand mean.

So in this case, 12.5 represents the deviation of the BP mean of old patient (x1=1) from the overall BP mean. The mean for old patient is 12.5 higher than that of the overall mean (or mean of young patient is -12.5 lower than that of the overall mean). It can also be interpreted as half the average difference between old and young.

```
>
> # check beta 1
> mean(y[x1==1])-mean(y)
[1] 12.5
> mean(y)-mean(y[x1==-1])
[1] 12.5
> mean(y[x1==1]-y[x1==-1])/2
[1] 12.5
>
>
```

The coefficient b2 of x2 represents the effect of x2 (gender) when x1 = 0 (age). Since x1 = 0 at the mean of the two categories of x2, b2 is a main effect. It's the effect of x2 at the mean value of x1. More specifically, the coefficients of each of the effect variables is equal to the difference between the mean of the group coded 1 and the grand mean. So in this case, -7.5 represents the deviation of the BP mean of female from the overall BP mean. The mean for female patient is 7.5 lower than that of the overall mean (or mean of male patients 7.5 higher than that of the overall mean). It can also be interpreted as half the average difference between female and male.

```
>
> # check beta 2
> mean(y[x2==1])-mean(y)
[1] -7.5
> mean(y)-mean(y[x2==-1])
[1] -7.5
> mean(y[x2==1]-y[x2==-1])/2
[1] -7.5
>
```

The equation is y = 127.5 + 12.5x1 -7.5x2 -2.5x1x2

The coefficient of x1*x2 captures the interaction between age and gender (like a cross product). One can think of the interaction coefficient as an additional effect of age (or gender) depending on gender (age).For example, from the equation, when x1=1,x2=1 (old, female), the deviation from the overall mean is now 12.5-7.5-2.5 = 2.5, where the interaction term had an effect of -2.5 in the deviation. But changing the gender of the old patient to male (x1=1,x2=-1), the deviation from the overall mean is now 12.5+7.5+2.5 = 22.5, where the interaction term had an opposite effect of +2.5 in the deviation. More specifically, the coefficient represents the average difference of the differences between old-young within female and old-young within male (or difference of the differences between female-male within old and female-male within young)

```
> # check beta 3
> ((y[x1==1&x2==1]-y[x1==-1&x2==1])/2-(y[x1==1&x2==-1]-y[x1==-1&x2==-1])/2)/2
[1] -2.5
```

beta(x1) > 0 implies that old patients tend to have a higher BP on average than do young on average

beta(x2) <0 implies that female patients tend to have a lower BP on average than do male

beta(x3) <0 implies a negative interaction between age and gender, when older (high level), female (high level) has the effect of reducing the BP, while male (low level) has the effect of increasing the BP.

Part d.

```
> # df of error
> n = length(y)
> p = 3
> df = n - (p+1)
> df
[1] 0
```

Part e.

Assumption: there might be typo in the book. Assume n = # patients per group, so 4*n total patients. Thus, n=1 and 4*1 patients (and not n =4 as it says in first part)

If the sample size is 4n, then the X'X matrix will be a diagonal matrix (4 by 4) with its entries equal to (4n) and its inverse will be a diagonal matrix with its entries equal to 1/(4n).So the current matrix for n=1 will be multiplied by n, and the current inverse matrix for n = 1 will be multiplied by 1/n for the case of n>0.

The X'y vector will be equal to a column vector with 4 entries, the first one equal to the sum of y's, the second equal to the sum of BP from old patients minus the sum of BP from young patients, the third equal to the sum of BP from female minus the sum of BP from male patients, and the fourth one the difference of the differences between old-young within female and old-young within male (or difference of the differences between female-male within old and female-male within young)

The error df will be 4n-(p+1) = 4n-(3+1) = 4n-4, where n = # patients per group

```
>
> # check for n = 2
>
> x1 = append(x1,x1)
> x2 = append(x2,x2)
> X = cbind(rep(1,8),x1,x2,x1*x2)
> y = c(110,120,130,150,110,120,130,150)
>
> # X
> X
       x1 x2
[1,] 1 -1  1 -1
[2,] 1 -1 -1  1
[3,] 1  1  1  1
[4,] 1  1 -1 -1
[5,] 1 -1  1 -1
[6,] 1 -1 -1  1
[7,] 1  1  1  1
[8,] 1  1 -1 -1
```

```
>
> # X'X
> t(X)%*%X
     x1 x2
   8  0  0 0
x1 0  8  0 0
x2 0  0  8 0
   0  0  0 8
>
> # inv(X'X)
> solve(t(X)%*%X)
            x1    x2
   0.125 0.000 0.000 0.000
x1 0.000 0.125 0.000 0.000
x2 0.000 0.000 0.125 0.000
   0.000 0.000 0.000 0.125
>
> # X'y
> t(X)%*%y
   [,1]
   1020
x1  100
x2  -60
    -20
>
> sum(y)
[1] 1020
> sum(y[x1==1])-sum(y[x1==-1])
[1] 100
> sum(y[x2==1])-sum(y[x2==-1])
[1] -60
> (sum(y[x1==1&x2==1])-sum(y[x1==-1&x2==1]))-(sum(y[x1==1&x2==-1])-sum(y[x1==
-1&x2==-1]))
[1] -20
> (sum(y[x1==1&x2==1])-sum(y[x1==1&x2==-1]))-(sum(y[x1==-1&x2==1])-sum(y[x1==
-1&x2==-1]))
[1] -20
>
> # beta
> solve(t(X)%*%X)%*%t(X)%*%y
    [,1]
   127.5
x1  12.5
x2  -7.5
    -2.5
>
> n = length(y)
> p = 3
> df = n - (p+1)
> df
[1] 4
```

Problem 3 – 3.3

```
> # Import data
> filename = "Cobb-Douglas+production+function+data.csv"
> mydata = read.csv(filename,header = T)
```

Part a.

```
> mydata= log(mydata)
> mydata = mydata[,2:4]
> colnames(mydata)=c("log_capital","log_labor","log_output")
> fit = lm(log_output~log_capital+log_labor,mydata)
> summary(fit)

Call:
lm(formula = log_output ~ log_capital + log_labor, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7604 -0.2665 -0.0694  0.1926  3.7975 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.71146    0.09671  -17.70   <2e-16 ***
log_capital  0.20757    0.01719   12.08   <2e-16 ***
log_labor    0.71485    0.02314   30.89   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4781 on 566 degrees of freedom
Multiple R-squared: 0.8378,	Adjusted R-squared: 0.8373 
F-statistic:  1462 on 2 and 566 DF,  p-value: < 2.2e-16
```

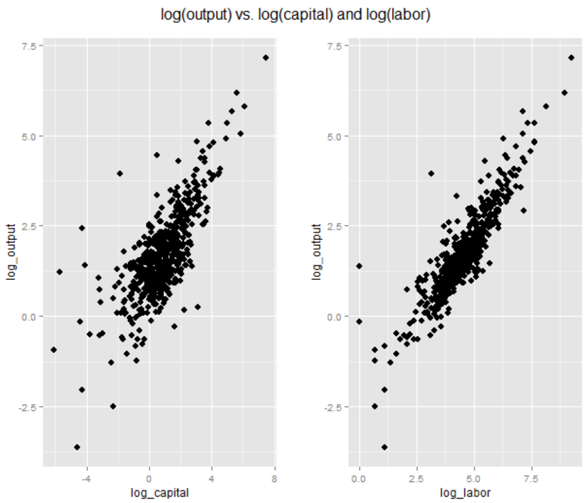
$$y = -1.71146 + x_1{}^{0.20757} x_2{}^{0.71485}$$

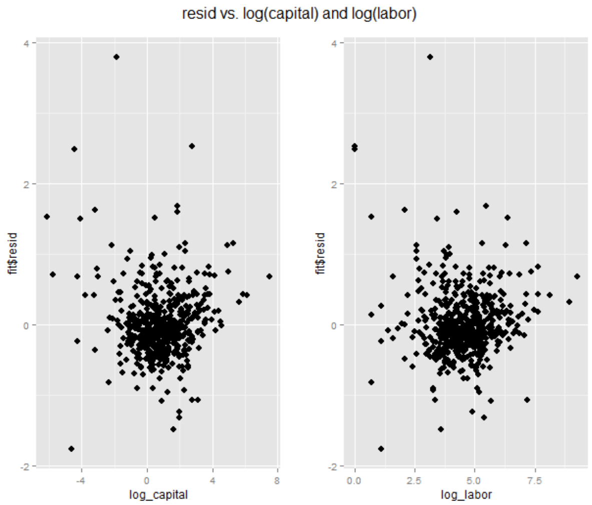
The positive coefficients show that increasing labor or capital leads to an increase in the firm's output. Looking at the p-values for the coefficients, all of them are <0.05, indicating that are significantly different than zero. The F-statistic all shows that jointly all coefficients are significantly different than zero. The R^2 is 0.83, which is a decent value for the data in which the model is accounting for 83% of the variation in the response. The correlation coefficient sqrt(0.83)=0.91 shows also a strong linear association between the response variable and predictors.

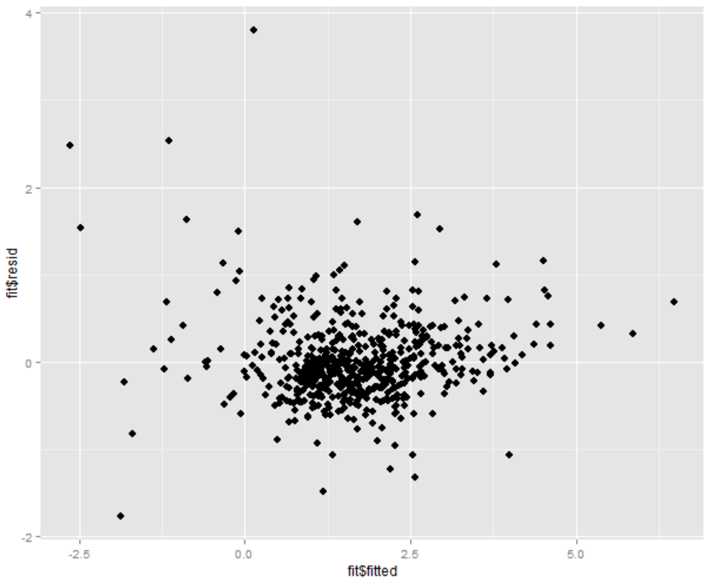
Plotting log(output) versus log(capital) and log(labor) show that a strong linear relationship between log(output) and log(capital), and log(output) and log(labor). The residual plots also show no pattern, indicating that the model assumptions seemed to be satisfied and fit seems to be good.

```
> 
> plot1 = 
+   ggplot(mydata,aes(x=log_capital, y = log_output)) +
+   geom_point(size = 3)
> 
> plot2 = 
+   ggplot(mydata,aes(x=log_labor, y = log_output)) +
+   geom_point(size = 3)
> 
> grid.arrange(plot1,plot2,ncol=2, main = "log(output) vs. log(capital) and l
og(labor)")
```

```
>
>
> plot1 =
+   ggplot(mydata,aes(x=log_capital, y = fit$resid)) +
+   geom_point(size = 3)
>
> plot2 =
+   ggplot(mydata,aes(x=log_labor, y = fit$resid)) +
+   geom_point(size = 3)
>
> grid.arrange(plot1,plot2,ncol=2, main = "resid vs. log(capital) and log(lab
or)")
>
> #plot(fit)
>
> ggplot(mydata,aes(x=fit$fitted, y = fit$resid)) +
+   geom_point(size = 3)
>
```


log(output) vs. log(capital) and log(labor)


resid vs. log(capital) and log(labor)



Part b.

The predicted output is a follows:

```
> newdata = log(data.frame(log_capital=500,log_labor=200))
> exp(predict(fit, newdata,interval="predict"))
       fit      lwr      upr
1 28.96226 11.16622 75.12055
```

Part c.

beta 1 + beta2 =1 is one more constraint that makes the full model more restricted, so use F test to compare the full model and reduced model to test the null hypothesis that beta 1 + beta2 =1. Start by substituting beta1 = 1- beta2 in full model: log(y) = b0 + (1-b2)*log(capital) + b2*log(labor) -> log(y)-log(capital) = b0 + b2(log(labor)-log(capital)). Fit a regression model and compare to full model.

```
>
> mydata$log_output_capital = mydata$log_output-mydata$log_capital
> mydata$log_labor_capital = mydata$log_labor-mydata$log_capital
> head(mydata)
  log_capital log_labor log_output log_output_capital log_labor_capital
1   0.9580326  5.214936   2.224706          1.2666730          4.256903
2   0.2800809  4.510860   1.298640          1.0185592          4.230779
3   3.0952921  6.054439   3.359733          0.2644412          2.959147
4   2.3737750  4.276666   1.416979         -0.9567958          1.902891
5   0.1495957  3.828641   1.061308          0.9117125          3.679046
6   5.6232833  8.941415   6.174079          0.5507952          3.318131
>
> fit_reduced = lm(log_output_capital~log_labor_capital ,mydata)
> summary(fit_reduced)

Call:
lm(formula = log_output_capital ~ log_labor_capital, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.4824 -0.2625 -0.0601  0.1848  3.9127 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -2.03310    0.06641  -30.61   <2e-16 ***
log_labor_capital  0.78511    0.01740   45.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4861 on 567 degrees of freedom
Multiple R-squared: 0.7822,	Adjusted R-squared: 0.7818 
F-statistic:  2036 on 1 and 567 DF,  p-value: < 2.2e-16

>
> anova(fit)
Analysis of Variance Table

Response: log_output
             Df Sum Sq Mean Sq F value    Pr(>F)    
log_capital   1 450.23  450.23 1969.96 < 2.2e-16 ***
log_labor     1 218.08  218.08  954.19 < 2.2e-16 ***
Residuals   566 129.36    0.23                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> df_full = anova(fit)["Residuals","Df"]
> RSS_full = anova(fit)["Residuals","Sum Sq"]
```

```
>
> anova(fit_reduced)
Analysis of Variance Table

Response: log_output_capital
                   Df Sum Sq Mean Sq F value    Pr(>F)
log_labor_capital   1 481.17  481.17  2035.9 < 2.2e-16 ***
Residuals         567 134.01    0.24
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> df_red = anova(fit_reduced)["Residuals","Df"]
> RSS_red = anova(fit_reduced)["Residuals","Sum Sq"]
>
> F_stat = ((RSS_red-RSS_full)/(df_red-df_full))/(RSS_full/(df_full))
> F_crit = qf(.95,df1=df_red-df_full,df2=df_full)
> F_stat
[1] 20.33514
> F_crit
[1] 3.857941
> F_stat > F_crit
[1] TRUE
> pf(F_stat,df1=df_red-df_full,df2=df_full,lower.tail = FALSE)
[1] 7.904722e-06
>
```

F statistic > Fcritical (or p-value < 0.05). Therefore reject Null that beta1 + beta2 = 1, so reject constant returns to scale for this data. Or one can use built in function in R in package car:

```
> library(car)
> linearHypothesis(fit, "log_labor + log_capital = 1")
Linear hypothesis test

Hypothesis:
log_capital  + log_labor = 1

Model 1: restricted model
Model 2: log_output ~ log_capital + log_labor

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    567 134.01
2    566 129.36  1    4.6476 20.335 7.905e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Problem 4 – 3.12

Part a)

Looking at model 1 (salary ~ gender + qualification), we see that the coefficient of gender (0.224) is positive indicating that salary for men tend to be higher than that of female for equal qualification. However, the p-value > 0.05 for gender, so the null hypothesis that the effect of gender after controlling for qualification is zero cannot be rejected at a 0.05 level. In other words the effect of gender on salary is not even close to being significant. Thus, there is not enough evidence to say that men are paid more than equally qualified women.

Part b)

Looking at model 2 (qualification ~ gender + salary), we see that the coefficient of gender (0.85) is positive, suggesting that qualification for men tend to be higher than that for female for equal salary. Therefore, the sign of the coefficient shows the opposite of the proposition that men are less qualified than equally paid women. Also since p-value > 0.05, the null hypothesis that the effect of gender after controlling for salary is zero cannot be rejected at a 0.05 level (although this is borderline rejection level). Assuming a 0.10 level, the null will be rejected indicating that there is a gender effect showing men are actually more qualified than equally paid women. Thus, there is not enough evidence to say that men are less qualified than equally paid women.

Part c)

From the coefficient standpoint, the results show inconsistencies because model 1 says that salary for men tend to be higher than that of female for equal qualification (discrimination against women), and model 2 says qualification for men tend to be higher than that for female for equal salary (discrimination against men). However, from a p-value point of view, the models do not show inconsistencies because the effect of gender in qualification and in salary after controlling for the others is not significant at a 0.05 level.

Part d)

Assuming the defense lawyer is for the company, the lawyer can use either model to make an argument. Model 1 seems to be more relevant to answer the question at hand, which is the discrimination against women, meaning for equal qualification, men tend to be paid more than women. Model 1 shows that this effect is very far from being significant, so there is not enough evidence to show this effect is not zero. Thus, model 1 might be preferred, but the lawyer might also use Model 2 since in some sense it shows the opposite of discrimination against women, it might actually show a discrimination against men because qualification for men tend to be higher than that for female for equal salary. Using a significance level of 0.10 might even make the lawyer’s argument more convincing since it would say that this effect is significant.

Problem 5 – 3.13

Part a)

Let $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$,

where x_1= gender, x_2= education, x_3= experience, x_4= months, y=beginning salaries

- $H_0: \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
- "The null hypothesis claims that there is no significant correlation at all. That is, all of the coefficients are zero and none of the variables belong in the model".
- $H_A: at\ least\ one\ of\ \beta_0, \beta_1, \beta_2, \beta_3\ or\ \beta_4\ is\ diffrent\ than\ zero$
- "The alternative hypothesis is not that every variable belongs in the model but that at least one of the variables belongs in the model".
- Test used: F-test, where F = MSR/MSE = (SSR/p)/(SSE/(n-(p+1))) = 22.98
- Conclusion: Since F = 22.98 > 2.48, reject null hypothesis that all coefficients are zero and conclude at least one is significantly different that zero. From this F-test, conclude overall fit might be okay.

```
> F_stat = (23665352/4)/(22657938/88)
> F_stat
[1] 22.97816
>
> F_crit = qf(.95,df1=4,df2=88)
> F_crit
[1] 2.475277
```

Part b)

- The coefficient of experience is 1.2690. Since it is positive, there is a positive linear positive relationship between salary and experience after accounting for the effect of the variables gender, education and months. In addition, the coefficient (this relationship) is significant (e.g. different than zero) as the tests below show.
- For t-test:
  - $H_0: \beta_3 = 0$
  - $H_A: \beta_3 \neq 0$
  - Note that the test depends on what variables are in the model and in this case all the rest of the variables are in the model.
- F-test can also be used by comparing a full model against the reduced model without the coefficient of experience. The null and alternative are the same as t-test (also can say that the null says the reduced model is better, while the alternative says the full model is better). Both tests are equivalent.
- Test used: t-test, t = coefficient/se = 2.16, or p-value = 0.034
- Conclusion: p-value < 0.05, reject null hypothesis that the effect of experience coefficients is zero and conclude is significantly different than zero.

Part c)

Salary for a man with 12 years of education, 10 months of experience, and 15 months with the company:

$$\widehat{y^*} = 3526.4 + 722.5 \times 1 + 90.02 \times 12 + 1.2690 \times 10 + 23.406 \times 15 = 5692.92$$

```
>
> beta = c(3526.4,722.5,90.02,1.2690,23.406)
> x = c(1,1,12,10,15)
> beta%*%x
        [,1]
[1,] 5692.92
```

Part d)

Average salary for men with 12 years of education, 10 months of experience, and 15 months with the company:

$$\widehat{\mu^*} = E[\widehat{y^*}] = 3526.4 + 722.5 \times 1 + 90.02 \times 12 + 1.2690 \times 10 + 23.406 \times 15 = 5692.92$$

Part e)

Average salary for women with 12 years of education, 10 months of experience, and 15 months with the company:

$$\widehat{\mu^*} = E[\widehat{y^*}] = 3526.4 + 722.5 \times 0 + 90.02 \times 12 + 1.2690 \times 10 + 23.406 \times 15 = 4970.42$$

```
> beta = c(3526.4,722.5,90.02,1.2690,23.406)
> x = c(1,0,12,10,15)
> beta%*%x
        [,1]
[1,] 4970.42
```

Problem 6 – 3.15

```
> # Import data
> filename = "P088.txt"
> mydata = read.table(filename,header = T)
```

Part a)

- For t-test:
  - $H_0 : \beta_{female} = 0$
  - $H_A : \beta_{female} \neq 0$
  - Note that the test depends on what variables are in the model and in this case all the rest of the variables are in the model.
- F-test can also be used by comparing a full model against the reduced model without the coefficient of female. The null and alternative are the same as t-test (also can say that the null says the reduced model is better, while the alternative says the full model is better). Both tests are equivalent.
- Test used: t-test, t = coefficient/se = -0.189, or p-value = 0.85071
- Conclusion: p-value > 0.05, cannot reject null hypothesis that the effect of female coefficients (after accounting for the effect of the variables) is zero and conclude is not significantly different than zero. Thus, the variable female is not needed in the full model as the reduced model without female is preferred over the full model

```
> fit = lm(Sales~Age + HS + Income + Black + Female + Price,mydata)
> summary(fit)

Call:
lm(formula = Sales ~ Age + HS + Income + Black + Female + Price,
    data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.398 -12.388  -5.367   6.270 133.213 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) 103.34485  245.60719   0.421  0.67597   
Age           4.52045    3.21977   1.404  0.16735   
HS           -0.06159    0.81468  -0.076  0.94008   
Income        0.01895    0.01022   1.855  0.07036 . 
Black         0.35754    0.48722   0.734  0.46695   
Female       -1.05286    5.56101  -0.189  0.85071   
Price        -3.25492    1.03141  -3.156  0.00289 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.17 on 44 degrees of freedom
Multiple R-squared: 0.3208,	Adjusted R-squared: 0.2282 
F-statistic: 3.464 on 6 and 44 DF,  p-value: 0.006857

> 
> summary(fit)$coeff["Female","Pr(>|t|)"]
[1] 0.8507058
```

Part b)

- Use F-test:
  - $H_0 : \beta_{female} = \beta_{HS} = 0$
  - $H_A : at\ least\ one\ of\ \beta_{female}\ or\ \beta_{HS}\ is\ not\ zero$
  - Note that the test depends on what variables are in the model and in this case all the rest of the variables are in the model.
  - F-test compares a full model against the reduced model without the coefficient of female and HS. The null is that the reduced model is better, while the alternative says the full model is better.
- Test used: F-test, F=((RSS_reduced-RSS_full)/(df_reduced-df_full))/(RSS_full/df_full)= 3492, or p-value = 0.9789
- Conclusion: p-value > 0.05, cannot reject null hypothesis that the effect of female coefficients and HS (after accounting for the effect of the variables) is zero and conclude they are not significantly different than zero. Thus, the variable female and HS are not needed at the same time in the full model as the reduced model without female and HS is preferred over the full model.

```
> fit_reduced = lm(Sales~Age  + Income + Black + Price,mydata)
> summary(fit_reduced)

Call:
lm(formula = Sales ~ Age + Income + Black + Price, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-46.784 -11.810  -5.380   5.758 132.789 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) 55.329580  62.395293   0.887   0.3798   
Age          4.191538   2.195535   1.909   0.0625 . 
Income       0.018892   0.006882   2.745   0.0086 **
Black        0.334162   0.312098   1.071   0.2899   
Price       -3.239941   0.998778  -3.244   0.0022 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.57 on 46 degrees of freedom
Multiple R-squared: 0.3202,	Adjusted R-squared: 0.2611 
F-statistic: 5.416 on 4 and 46 DF,  p-value: 0.001168

> 
> anova(fit)
Analysis of Variance Table

Response: Sales
          Df Sum Sq Mean Sq F value   Pr(>F)   
Age        1   2640  2639.5  3.3253 0.075019 . 
HS         1    412   412.3  0.5195 0.474883   
Income     1   3939  3939.1  4.9625 0.031063 * 
Black      1   1587  1587.1  1.9994 0.164393   
Female     1     16    16.1  0.0203 0.887307   
Price      1   7905  7905.3  9.9591 0.002886 **
Residuals 44  34926   793.8                    
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> df_full = anova(fit)["Residuals","Df"]
> RSS_full = anova(fit)["Residuals","Sum Sq"]
> 
> anova(fit_reduced)
Analysis of Variance Table

Response: Sales
          Df Sum Sq Mean Sq F value   Pr(>F)   
Age        1   2640  2639.5  3.4731 0.068766 . 
Income     1   3952  3952.1  5.2001 0.027267 * 
Black      1   1877  1876.7  2.4694 0.122936   
Price      1   7997  7997.4 10.5229 0.002199 **
Residuals 46  34960   760.0                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> df_red = anova(fit_reduced)["Residuals","Df"]
> RSS_red = anova(fit_reduced)["Residuals","Sum Sq"]
> 
> F_stat = ((RSS_red-RSS_full)/(df_red-df_full))/(RSS_full/(df_full))
> F_crit = qf(.95,df1=df_red-df_full,df2=df_full)
> F_stat
[1] 0.02128984
> F_crit
[1] 3.209278
> F_stat > F_crit
[1] FALSE
> pf(F_stat,df1=df_red-df_full,df2=df_full,lower.tail = FALSE)
[1] 0.9789453
> 
> anova(fit,fit_reduced)
Analysis of Variance Table

Model 1: Sales ~ Age + HS + Income + Black + Female + Price
Model 2: Sales ~ Age + Income + Black + Price
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     44 34926                           
2     46 34960 -2   -33.799 0.0213 0.9789
> 
```

Part c)

95% Confidence interval for coefficient income:

```
> confint(fit, level=0.95)["Income",]
       2.5 %       97.5 % 
-0.001642517  0.039535423 
```

Part d)

Fit the full model without income and find the R^2. Calculations show that 26.8 % of the variation in sales can be accounted for when income is removed from the full model.

```
> fit = lm(Sales~Age + HS  + Black + Female + Price ,mydata)
> summary(fit)

Call:
lm(formula = Sales ~ Age + HS + Black + Female + Price, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-37.414 -16.454  -5.746   8.541 133.319 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 162.3245   250.0537   0.649  0.51954   
Age           7.3073     2.9238   2.499  0.01616 * 
HS            0.9717     0.6103   1.592  0.11836   
Black         0.8447     0.4213   2.005  0.05101 . 
Female       -3.7815     5.5063  -0.687  0.49576   
Price        -2.8603     1.0362  -2.760  0.00832 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.93 on 45 degrees of freedom
Multiple R-squared: 0.2678,	Adjusted R-squared: 0.1864 
F-statistic: 3.291 on 5 and 45 DF,  p-value: 0.01287

> summary(fit)$r.squared
[1] 0.2677526
```

Part e)

Fit model with only Age, income and price and find the R^2. Calculations show that 30.3 % of the variation in sales can be accounted for by age, income and price.

```
> fit = lm(Sales~Age +Income + Price ,mydata)
> summary(fit)

Call:
lm(formula = Sales ~ Age + Income + Price, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-50.430 -13.853  -4.962   6.691 128.947 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) 64.248227  61.933008   1.037  0.30487   
Age          4.155909   2.198699   1.890  0.06491 . 
Income       0.019281   0.006883   2.801  0.00737 **
Price       -3.399234   0.989172  -3.436  0.00124 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.61 on 47 degrees of freedom
Multiple R-squared: 0.3032,	Adjusted R-squared: 0.2588 
F-statistic: 6.818 on 3 and 47 DF,  p-value: 0.0006565

> summary(fit)$r.squared
[1] 0.3032434
```

Part f)

Fit model with only income and find the R^2. Calculations show that 10.6 % of the variation in sales can be accounted for by income.

```
> 
> fit = lm(Sales~Income ,mydata)
> summary(fit)

Call:
lm(formula = Sales ~ Income, data = mydata)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-54.550 -15.772  -6.517   4.491 144.628 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept) 55.362454  27.743082   1.996   0.0516 .
Income       0.017583   0.007283   2.414   0.0195 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.63 on 49 degrees of freedom
Multiple R-squared: 0.1063,	Adjusted R-squared: 0.08808 
F-statistic: 5.829 on 1 and 49 DF,  p-value: 0.01954

> summary(fit)$r.squared
[1] 0.1063203
```

R-code

```
#### Homework 2
#### Text-Book Problems: 3.12, 3.13, 3.15
#### My Book Problems: 3.1, 3.2, 3.3

#### Setup ##############################################

# Install packages if needed
# install.packages("ggplot2")
# install.packages("grid")
# install.packages("gridExtra")
# install.packages("XLConnect")
# install.packages("corrplot")
# install.packages("Hmisc")
# install.packages("car")

# Load packages
library(ggplot2)
library(grid)
library(gridExtra)
library(XLConnect)
library(corrplot)
library(Hmisc)
library(car)


# My PC
main = "C:/Users/Steven/Documents/Academics/3_Graduate School/2014-2015 ~ NU/"

# Aginity
#main = "\\\\nas1/labuser169"

course = "MSIA_401_Statistical Methods for Data Mining"
datafolder = "Data"
setwd(file.path(main,course, datafolder))

#### Problem 1 - 3.1 ##############################################

# Verify

x = c(1,2,3,4,5)
y = c(2,6,7,9,10)


X = cbind(rep(1,5),x)
t(X)%*%X
solve(t(X)%*%X)
```

```
t(X)%*%y

solve(t(X)%*%X)%*%t(X)%*%y

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

#### Problem 2 - 3.2 ###############################################

x1 = c(-1,-1,1,1)
x2 = c(1,-1,1,-1)
X = cbind(rep(1,4),x1,x2,x1*x2)
colnames(X)[4]="x1x2"
y = c(110,120,130,150)

## Part a)

X

## Part b)

# X'X
t(X)%*%X

# inv(X'X)
solve(t(X)%*%X)

# beta
solve(t(X)%*%X)%*%t(X)%*%y

# check
fit = lm(y~ x1 + x2 +x1*x2)
summary(fit)

# The inverse is easy to calculate because X'X is a diagonal matrix as a result of
# orthogonality, so the inverse is a diagonal matrix with the inverse (e.g 1/4 in this case) of
# its corresponding entries.

## Part c)

# Reference: http://www.ats.ucla.edu/stat/mult_pkg/faq/general/effect.htm

# The equation is y = 127.5 + 12.5x1 -7.5x2 -2.5x1x2
# The intercept represents the overall mean of BP. Ther overall mean is 127.5

# check intercept
mean(y)
```

```
# With (-1,0,+1) coding, the coefficients represent the distance between the factor levels and the overall mean. Thus, the
# coefficient b1 of x1 represents the effect of x1 (age) when x2 = 0 (gender).  Since x2 = 0 at the mean of the two
# categories of x2, b1 is a main effect.  It's the effect of x1 at the mean value of x2. More specifically, the coefficients
# of each of the effect variables is equal to the difference between the mean of the group coded 1 and the grand mean.
# So in this case, 12.5  represents the deviation of the BP mean of old patient (x1=1) from the overall BP mean.
# The mean for old patient is 12.5 higher than that of the overall mean (or mean of young patient is -12.5 lower than that of the overall mean)
# It can also be interpreted as half the average difference between old and young.

# check beta 1
mean(y[x1==1])-mean(y)
mean(y)-mean(y[x1==-1])
mean(y[x1==1]-y[x1==-1])/2


# The coefficient b2 of x2 represents the effect of x2 (gender) when x1 = 0 (age). Since x1 = 0 at the mean of the two
# categories of x2, b2 is a main effect.  It's the effect of x2 at the mean value of x1. More specifically, the coefficients
# of each of the effect variables is equal to the difference between the mean of the group coded 1 and the grand mean.
# So in this case, -7.5  represents the deviation of the BP mean of female from the overall BP mean.
# The mean for female patient is 7.5 lower than that of the overall mean (or mean of male patient
# is 7.5 higher than that of the overall mean). It can also be interpreted as half the average difference between female and male.

# check beta 2
mean(y[x2==1])-mean(y)
mean(y)-mean(y[x2==-1])
mean(y[x2==1]-y[x2==-1])/2

# The equation is y = 127.5 + 12.5x1 -7.5x2 -2.5x1x2

# The coefficient of x1*x2 captures the interaction beetween age and gender (like a crossproduct). One can think of
# the interaction coefficient as an additional effect of age (or gender) depending on gender (age).
# For example, from the equation, when x1=1,x2=1 (old,female), the deviation from the overall mean
# is now 12.5-7.5-2.5 = 2.5, where the intercation term had an effect of -2.5 in the deviation.
# But changing the gender of the old patient to male (x1=1,x2=-1), the
# deviation from the overall mean is now 12.5+7.5+2.5 = 22.5. where the intercation term had an
# an oppositive effect of +2.5 in the deviation.
# More specifically, the coefficient represents the average difference of the differenes between old-
young within female
# and old-young within male (or difference of the differenes between female-male within old
# and female-male within young)

# It is also half the differene between the age effect for males and the age effect for females.

# check beta 3
((y[x1==1&x2==1]-y[x1==-1&x2==1])/2-(y[x1==1&x2==-1]-y[x1==-1&x2==-1])/2)/2

# beta(x1) > 0 implies that old patients tend to hava a higher BP on average than do young on average
# beta(x2) <0 implies that female patients tend to have a lower BP on average than do male
# beta(x3) <0 implies a negative interaction between age and gender, when older (highve level), female
(high level) has the effect
# of reducing the BP, while male (low level) has the effect of increasing the BP.


## Part d)

# df of error
n = length(y)
p = 3
df = n - (p+1)
df

## Part e)

# Assumption: ther might be typo in the book. Assume n = # patients per group, so 4*n total patients
# Thus, n=1 and 4*1 patients (and not n =4 as it says in first part)

# If the sample size is 4n, then the X'X matrix will be a diagonal matrix with its entries
# equal to (4n) and its inverse will be a diagonal matrix with its entries equal to 1/(4n).
# So the current matrix for n=1 will be multiplied by n, and the current inverse matrix for n = 1
# will be multiplied by 1/n for the case of n>0.

# The X'y vector  will be equal to a coulmn vector with 4 entries, the first one equal to
# the sum of y's, the second equal to the the sum of BP from old patients minus
# the sum of BP from young patients, the thrid equal ot the sum of BP from female
# minus the sum of BP from male patients, and the fourth one the difference of the differenes
# between old-young within female and old-young within male (or difference of the differenes
# between female-male within old and female-male within young)

# The error df will be 4n-(p+1) = 4n-(3+1) = 4n-4, where n = # patients per group

# check

x1 = append(x1,x1)
x2 = append(x2,x2)
```

```
X = cbind(rep(1,8),x1,x2,x1*x2)
y = c(110,120,130,150,110,120,130,150)

# X
X

# X'X
t(X)%*%X

# inv(X'X)
solve(t(X)%*%X)

# X'y
t(X)%*%y

sum(y)
sum(y[x1==1])-sum(y[x1==-1])
sum(y[x2==1])-sum(y[x2==-1])
(sum(y[x1==1&x2==1])-sum(y[x1==-1&x2==1]))-(sum(y[x1==1&x2==-1])-sum(y[x1==-1&x2==-1]))
(sum(y[x1==1&x2==1])-sum(y[x1==1&x2==-1]))-(sum(y[x1==-1&x2==1])-sum(y[x1==-1&x2==-1]))

# beta
solve(t(X)%*%X)%*%t(X)%*%y

n = length(y)
p = 3
df = n - (p+1)
df


#### Problem 3 - 3.3 ###############################################

# Import data
filename = "Cobb-Douglas+production+function+data.csv"
mydata = read.csv(filename,header = T)

# Look at data
names(mydata)
head(mydata)
nrow(mydata)
summary(mydata)

## Part a)
mydata= log(mydata)
mydata = mydata[,2:4]
colnames(mydata)=c("log_capital","log_labor","log_output")
fit = lm(log_output~log_capital+log_labor,mydata)
summary(fit)
```

```
plot1 =
  ggplot(mydata,aes(x=log_capital, y = log_output)) +
  geom_point(size = 3)

plot2 =
  ggplot(mydata,aes(x=log_labor, y = log_output)) +
  geom_point(size = 3)

grid.arrange(plot1,plot2,ncol=2, main = "log(output) vs. log(capital) and log(labor)")


plot1 =
  ggplot(mydata,aes(x=log_capital, y = fit$resid)) +
  geom_point(size = 3)

plot2 =
  ggplot(mydata,aes(x=log_labor, y = fit$resid)) +
  geom_point(size = 3)

grid.arrange(plot1,plot2,ncol=2, main = "resid vs. log(capital) and log(labor)")

#plot(fit)

ggplot(mydata,aes(x=fit$fitted, y = fit$resid)) +
  geom_point(size = 3)

# Plotting log(output) versus log(capital) and log(labor) show that a strong linear relationship
# between log(output) and log(capital), and log(output) and log(labor)

# Looking at the p-values for the coefficients, all of them are <0.05, indicating that are signinficantly
# different than zero. The F-statistic all shows that jointly all coefficients are significantly different
# than zero. The R^2 is 0.83, which is a decent value for the data in which the model is accounting
# 83% of the variation in the response. The correlation coefficeint sqrt(0.83)=0.91 shows also a strong
# linear association between the response variable and predictors.

# The residual plots also show no pattern, indicating that the model assumptions seemed to be satisfied
# and fit seems to be good.

# The positive coefficients show that increasing labor or capital leads to an increase in the firm's output

## Part b)

newdata = log(data.frame(log_capital=500,log_labor=200))
exp(predict(fit, newdata,interval="predict"))

# Alternatively, keep the data original, fit log in lm, and for input for predict use original (non-log
transformed)
```

```
# but still need to exponentiate result. Note cannot use mydata$capital, etc for predict
# filename = "Cobb-Douglas+production+function+data.csv"
# mydata = read.csv(filename,header = T)
#
# fit = lm(log(output)~log(capital)+log(labor),mydata)
# summary(fit)
# newdata = data.frame(capital=500,labor=200)
# exp(predict(fit, newdata,interval="predict"))


## Part c)
# beta 1 + beta2 =1 is one more constraint that makes the full model more restricted
# so use F test to compare the full model and reduced model to test the null hypothesis that beta 1 +
beta2 =1
# Start by substituting beta1 = 1- beta2 in full model: log(y) = b0 + (1-b2)*log(capital) + b2*log(labor) ->
# log(y)-log(capital) = b0 + b2(log(labor)-log(capital)). Fit a regression model and compare to full model.

mydata$log_output_capital = mydata$log_output-mydata$log_capital
mydata$log_labor_capital = mydata$log_labor-mydata$log_capital
head(mydata)

fit_reduced = lm(log_output_capital~log_labor_capital ,mydata)
summary(fit_reduced)

anova(fit)
df_full = anova(fit)["Residuals","Df"]
RSS_full = anova(fit)["Residuals","Sum Sq"]

anova(fit_reduced)
df_red = anova(fit_reduced)["Residuals","Df"]
RSS_red = anova(fit_reduced)["Residuals","Sum Sq"]

F_stat = ((RSS_red-RSS_full)/(df_red-df_full))/(RSS_full/(df_full))
F_crit = qf(.95,df1=df_red-df_full,df2=df_full)
F_stat
F_crit
F_stat > F_crit
pf(F_stat,df1=df_red-df_full,df2=df_full,lower.tail = FALSE)

# F statistic > Fcritical (or p-value < 0.05).
# Threrefore reject Null that beta1 + beta2 = 1, so ject constatn returns to scale for this data

# Or one can use built in function in R in package car:
library(car)
linearHypothesis(fit, "log_labor + log_capital = 1")


#### Problem 4 - 3.13 ################################################
```

```
F_stat = (23665352/4)/(22657938/88)
F_stat

F_crit = qf(.95,df1=4,df2=88)
F_crit

beta = c(3526.4,722.5,90.02,1.2690,23.406)
x = c(1,1,12,10,15)
beta%*%x

beta = c(3526.4,722.5,90.02,1.2690,23.406)
x = c(1,0,12,10,15)
beta%*%x

#### Problem 4 - 3.15 #############################################

# Import data
filename = "P088.txt"
mydata = read.table(filename,header = T)

# Look at data
names(mydata)
head(mydata)
nrow(mydata)
summary(mydata)

## Part a

fit = lm(Sales~Age + HS + Income + Black + Female + Price,mydata)
summary(fit)

summary(fit)$coeff["Female","Pr(>|t|)"]

# p-value = 0.85 > 0.05, do not reject null hypothesis that the coefficient
# for females is zero. So female might not be needed in the model

## Part b

fit_reduced = lm(Sales~Age  + Income + Black + Price,mydata)
summary(fit_reduced)

anova(fit)
df_full = anova(fit)["Residuals","Df"]
RSS_full = anova(fit)["Residuals","Sum Sq"]

anova(fit_reduced)
df_red = anova(fit_reduced)["Residuals","Df"]
```

```
RSS_red = anova(fit_reduced)["Residuals","Sum Sq"]

F_stat = ((RSS_red-RSS_full)/(df_red-df_full))/(RSS_full/(df_full))
F_crit = qf(.95,df1=df_red-df_full,df2=df_full)
F_stat
F_crit
F_stat > F_crit
pf(F_stat,df1=df_red-df_full,df2=df_full,lower.tail = FALSE)

anova(fit,fit_reduced)

# p-value > 0.05, cannot reject null hypothesis that both coefficients
# HS and Female are equal to zero. So variables might not be needed in model.


## Part c
confint(fit, level=0.95)["Income",]

## Part d
fit = lm(Sales~Age + HS  + Black + Female + Price ,mydata)
summary(fit)
summary(fit)$r.squared

## Part e
fit = lm(Sales~Age +Income + Price ,mydata)
summary(fit)
summary(fit)$r.squared

## Part f

fit = lm(Sales~Income ,mydata)
summary(fit)
summary(fit)$r.squared
```