Textbook:

3.12 qualification – quali

Model 1: salary = β 0 + β 1*quali + β 2 *gender + error1

Model 2: quali = β ’0 + β ’1*gender + β ’2 *salary + error2

(a) YES. Since the coefficient of gender is positive and gender equals to 1 if the employee is male. Therefore, comparing to a female employee, a male employee has more salary based on same level of qualification, which is β 2*gender.

(b) NO. Since the coefficient of gender is positive and gender equals to 1 if the employee is male. Therefore, comparing to a female employee, a male employee has higher qualification level based on same amount of salary, which is β ’1*gender.

(c) YES. If the result of Model 1 tells us that female employees are paid less than male employees based on a same level of qualification. So, we can tell salary and qualification have a positive relationship. The result of Model 2 tells, with paying same amount of salary, women are less qualified than men, which is inconsistent to positive relationship between salary and qualification.

(d) Model 2. If we do a 95% confidence interval hypothesis test and let H0 be the coefficient of Gender is 0. P-value of Gender in Model 1 is 0.6379, which is much larger than 5%. Therefore, data shows a strong evidence that H0 is failed to reject. So, coefficient of gender is zero in Model 1 and there is no gender discrimination. In model 2, by same methodology, p-value of gender is very close to 5%, so that data shows coefficient of gender in model 2 is not 0. Gender is considered in model 2 instead of model 1.

3.13 the multiple regression model is

salary = β 0 + β 1*gender+ β 2*education+ β 3*experience + β 4*months + error

salary^ = 3526.4 + 722.5*gender+90.02*education + 1.269*experience+23.406*months

(a) H0: β 1= β 2 = β 3= β 4 =0; Ha: at least one of β i $\neq$ 0, i=1,2,3,4

Overall F-test: F= MSR/MSE~ $F_{p-1,n-(p+1)}$; n-(p+1)=88, p=4

F = 22.98 ~ $F_{4,88}$ = 2.475277

Since F statistic is larger than $F_{3,84}$, so we reject the null hypothesis and conclude that the model has a good fit for observations.

(b) H0: β 3 $\leq$0; Ha: β 3>0

t statistics of β 3= (β ^3 - β 3)/SE(β ^3) = (1.2690-0)/0.5877=2.159265 ~ $t_{0.95,-88}$= -1.645

t statistics is larger than $t_{0.95,-88}$, so we reject H0 and conclude that β 3>0. Therefore, experience has a positive relationship with salary.

(c) Given gender = 1, education = 12, months = 15 and experience = 10, salary^ = 5692.92.

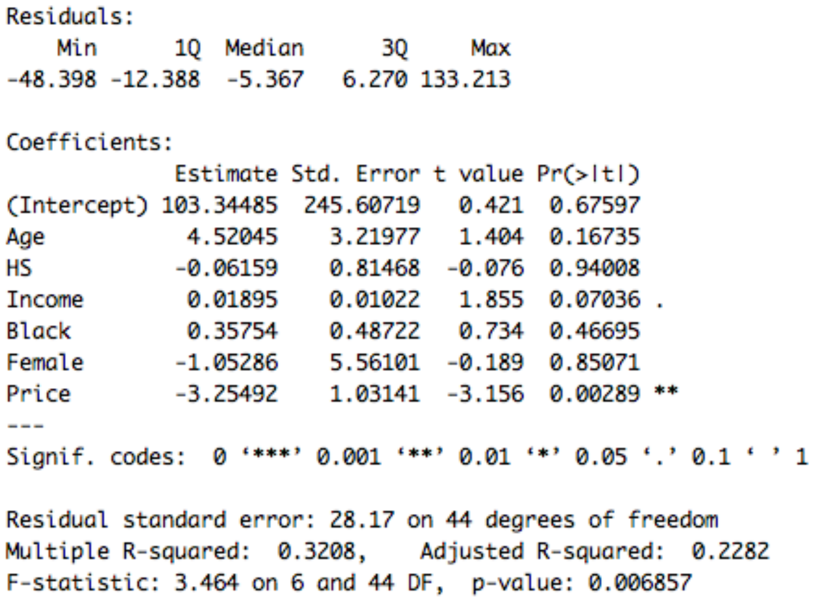
(d) Given gender = 1, education = 12, months = 15 and experience = 10, salary^ = 5692.92.

(e) Given gender = 0, education = 12, months = 15 and experience = 10, salary^ = 4970.42

3.15 The multiple linear regression model is

sales = β 0 + β 1*age + β 2*HS+ β 3*income + β 4*black + β 5*female+ β 6*price + error

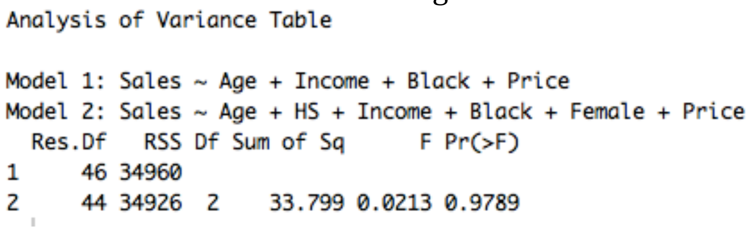
Run regression and get:



```
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
```

(a) H0: β 5 = 0; Ha: β 5 $\neq$ 0
P-value of female is 0.85071, which is larger than 0.05. So, we cannot reject null hypothesis and conclude that female is not needed in the regression equation relating Sales.

(b) H0: β 5 = β 2 = 0; Ha: at least one of β i $\neq$ 0, i= 5,2
Full model: sales = β 0 + β 1*age + β 2*HS+ β 3*income + β 4*black + β 5*female+ β 6*price + error
Reduced model: sales = β 0 + β 1*age + β 3*income + β 4*black + β 6*price + error
After run anova for both model get:



```
Analysis of Variance Table

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

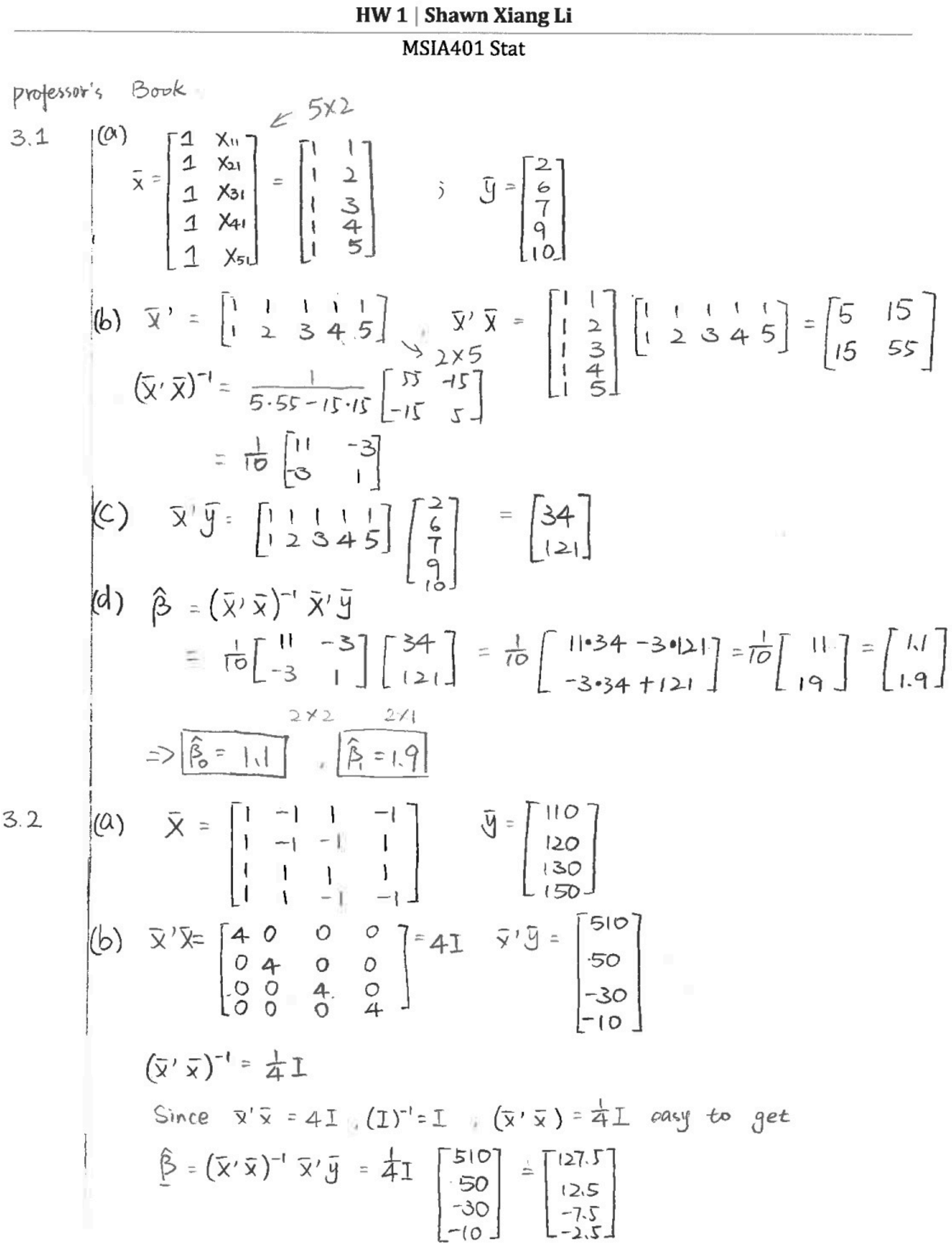
the P-value of reduced model is 0.9789, which is much larger than 0.05. So we fail to reject null hypothesis. We conclude that there is no need to add HS and Female into the model.

(c) 95% CI for the true regression coefficient of the variable Income is [0.005038974, 0.03274515] from reduced model since we have proved there is no need to add HS and Female into the model.

(d) After removed Income from the multiple regression model and ran the regression, we can get R^2 = 0.2678. It means 20% data fit the model, which is not good. As having Income in the model, the R^2 was 0.3208. Since R^2 will increase when adding more variables, and vise versa. Dropping Income variable could be the reason that R^2 decreases.

(e) Under the new model, R^2 = 0.3032, which is larger than R^2 in (d). It tells that data of HS, Black, and Female do not fit the model since R^2 increases after dropping them from the regression model.

(f) R^2 = 0.1063. It tells 10.63% data of Income fit the regression model. Since 10.63% is very low, only using Income to predict Sales is almost impossible.



HW 1 | Shawn Xiang Li

MSIA401 Stat

professor's Book

3.1 (a)

$$\bar{x} = \begin{bmatrix} 1 & X_{11} \\ 1 & X_{21} \\ 1 & X_{31} \\ 1 & X_{41} \\ 1 & X_{51} \end{bmatrix} \overset{\swarrow 5\times2}{=} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} ; \quad \bar{y} = \begin{bmatrix} 2 \\ 6 \\ 7 \\ 9 \\ 10 \end{bmatrix}$$

(b)

$$\bar{x}' = \underbrace{\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}}_{\searrow 2\times5} \qquad \bar{x}'\bar{x} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix} = \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix}$$

$$(\bar{x}'\bar{x})^{-1} = \frac{1}{5\cdot55 - 15\cdot15} \begin{bmatrix} 55 & -15 \\ -15 & 5 \end{bmatrix}$$

$$= \frac{1}{10} \begin{bmatrix} 11 & -3 \\ -3 & 1 \end{bmatrix}$$

(c)

$$\bar{x}'\bar{y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 2 \\ 6 \\ 7 \\ 9 \\ 10 \end{bmatrix} = \begin{bmatrix} 34 \\ 121 \end{bmatrix}$$

(d)

$$\hat{\beta} = (\bar{x}'\bar{x})^{-1} \bar{x}'\bar{y}$$

$$= \frac{1}{10} \underbrace{\begin{bmatrix} 11 & -3 \\ -3 & 1 \end{bmatrix}}_{2\times2} \underbrace{\begin{bmatrix} 34 \\ 121 \end{bmatrix}}_{2\times1} = \frac{1}{10} \begin{bmatrix} 11\cdot34 - 3\cdot121 \\ -3\cdot34 + 121 \end{bmatrix} = \frac{1}{10} \begin{bmatrix} 11 \\ 19 \end{bmatrix} = \begin{bmatrix} 1.1 \\ 1.9 \end{bmatrix}$$

$$\Rightarrow \boxed{\hat{\beta}_0 = 1.1} , \boxed{\hat{\beta}_1 = 1.9}$$

3.2 (a)

$$\bar{X} = \begin{bmatrix} 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \end{bmatrix} \qquad \bar{y} = \begin{bmatrix} 110 \\ 120 \\ 130 \\ 150 \end{bmatrix}$$

(b)

$$\bar{x}'\bar{x} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix} = 4I \qquad \bar{x}'\bar{y} = \begin{bmatrix} 510 \\ 50 \\ -30 \\ -10 \end{bmatrix}$$

$$(\bar{x}'\bar{x})^{-1} = \frac{1}{4}I$$

Since $\bar{x}'\bar{x} = 4I$, $(I)^{-1} = I$, $(\bar{x}'\bar{x}) = \frac{1}{4}I$ easy to get

$$\hat{\beta} = (\bar{x}'\bar{x})^{-1} \bar{x}'\bar{y} = \frac{1}{4}I \begin{bmatrix} 510 \\ 50 \\ -30 \\ -10 \end{bmatrix} = \begin{bmatrix} 127.5 \\ 12.5 \\ -7.5 \\ -2.5 \end{bmatrix}$$

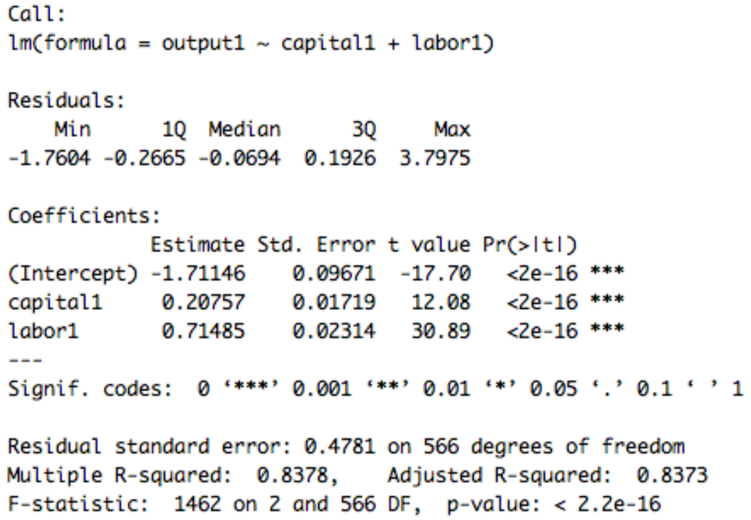
Professor's book

3.2 (c)Based on estimator β hat we get from(b), we can write the model as
y= 127.5+12.5*x1-7.5*x2-2.5*x1*x2+error. Each estimated regression coefficient tells:
1. Hold other predictive variable fixed, if the patient is young, BP will estimately change (12.5+2.5*x2) units; if the patient is old, BP will estimately change (12.5-2.5*x2) units.
2. Hold other predictive variable fixed, if the patient is female, BP will estimately change (-7.5-2.5X1) units; if the patient is male, BP will estimately change (7.5+2.5*x1) units.

(d) error d.f = n-(p+1) = 0
(e) error d.f = 4n-(p+1) =12

matrix $\mathbf{X'X} = 16I_{4by4}$; $(X'X)^{-1} = (1/16)I_{4by4}$; $X'y = \begin{bmatrix} 2040 \\ 200 \\ -120 \\ -40 \end{bmatrix}$

3.3 After log-transformation, model becomes $\ln Y = \ln \beta 0 + \beta 1 * \ln X1 + \beta 2 * \ln X2 + \ln(\text{error})$
After run regression, we get

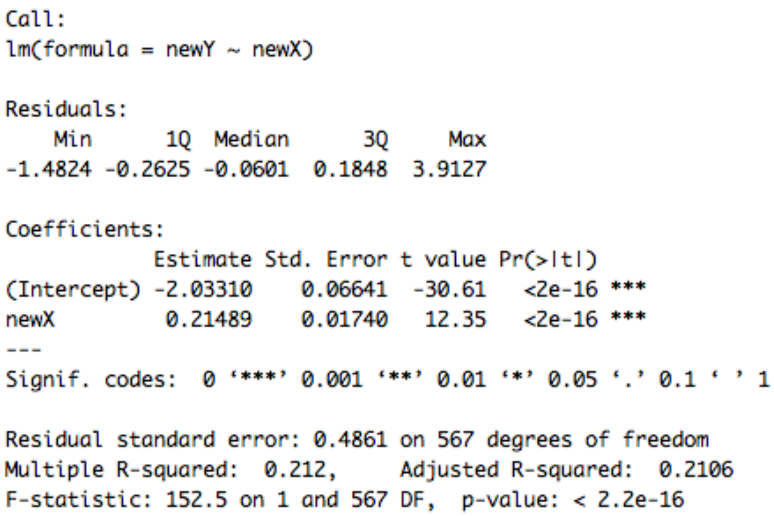


```
Call:
lm(formula = output1 ~ capital1 + labor1)

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 ***
capital1     0.20757    0.01719   12.08   <2e-16 ***
labor1       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
```

(a) β1 & β2's P-value are all less than 0.05, so it shows that capital and labor are all significant predictive variable and are indeed added into the model. Since b1=0.20757 & b2=0.71485 are capital and labor elasticities. We can know that, holding labor fixed, if capital increase 1%, the output will estimatelly increase 0.20757%. Holding labor fixed, if labor increase 1%, the output will estimatelly increase 0.71485%.

(b) lnY=-1.71146+ 0.20757*lnX1+0.71485* lnX2. Given X1=500, X2=200, so we can get ln(Y) = 3.366008, so Y=e^3.366008 = 28.96268776. So the output of the firm is 28.96268776 millions of euros.

(c) Since β1+ β2=1, so we can write β2=1- β1 and put into the model and we get:
lnY – lnX2 = β0+ β1(lnX1-lnX2) + error. So we restructure the data and run the regression.



```
Call:
lm(formula = newY ~ newX)

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 ***
newX         0.21489    0.01740   12.35   <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:  0.212,	Adjusted R-squared:  0.2106 
F-statistic: 152.5 on 1 and 567 DF,  p-value: < 2.2e-16
```

H0: $\beta 1 = 0$; Ha: $\beta 1 \neq 0$
We can see P-value of newX, which is (lnX1-lnX2), is smaller than 0.05. So we reject null hypothesis. So, we conclude (lnX1-lnX2) had significant relationship with response variable and $\beta 1+ \beta 2=1$.