3.3

3.3 $F = \frac{MSH_0}{MSE} = \frac{SSH_0}{rMSE} = \frac{SSE_0 - SSE}{rMSE} = \frac{SSE_0 - SSE}{r \cdot \frac{SSE}{n-(p+1)}}$

$= \frac{SSE_0 - SSE}{(p-q) \cdot \frac{SSE}{n-(p+1)}}$

$R^2 = 1 - \frac{SSE}{SST} \quad \therefore SSE = SST(1-R^2)$

$\therefore F = \frac{SST(1-R_q^2) - SST(1-R_p^2)}{(p-q) \cdot \frac{SST(1-R_p^2)}{n-(p+1)}} = \frac{(R_p^2 - R_q^2)(n-(p+1))}{(p-q)(1-R_p^2)}$

when $n=26$. $q=3$. $p=5$. $R_p^2=0.9$. $R_q^2=0.8$

$F = \frac{(0.9-0.8)(26-6)}{(5-3)(1-0.9)} = 10$

$\alpha = 1\%$.

$\therefore F \sim F_{2,20,1\%} = 5.85 \qquad F > 5.85$

$\therefore$ reject H_0

which means the increase in R^2 from partial model to full model is statistically significant at 1% level.

3.10

3.10 a) $X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix}$ $\quad y = \begin{bmatrix} 2 \\ 6 \\ 7 \\ 9 \\ 10 \end{bmatrix}$

b). $X' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}$

$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} = \begin{pmatrix} 5 & 15 \\ 15 & 55 \end{pmatrix}$

$\therefore (X'X)^{-1} = \begin{pmatrix} 5 & 15 \\ 15 & 55 \end{pmatrix}^{-1} = \frac{1}{5 \times 55 - 15 \times 15} \begin{bmatrix} 55 & -15 \\ -15 & 5 \end{bmatrix} = \begin{pmatrix} \frac{11}{10} & -\frac{3}{10} \\ -\frac{3}{10} & \frac{1}{10} \end{pmatrix}$

$(X'X)^{-1}(X'X) = \begin{pmatrix} \frac{11}{10} & -\frac{3}{10} \\ -\frac{3}{10} & \frac{1}{10} \end{pmatrix} \begin{pmatrix} 5 & 15 \\ 15 & 55 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

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

3.11

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

$$= \begin{pmatrix} \frac{11}{10} & -\frac{3}{10} \\ -\frac{3}{10} & \frac{1}{10} \end{pmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix} \begin{pmatrix} 2 \\ 6 \\ 7 \\ 9 \\ 10 \end{pmatrix} = \begin{pmatrix} \frac{11}{10} \\ \frac{19}{10} \end{pmatrix}$$

$\therefore \hat{\beta}_0 = 1.1, \quad \hat{\beta}_1 = 1.9, \quad \hat{y} = 1.1 + 1.9x$

3.11 $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

$$E(y) = \begin{cases} \beta_0 - \beta_1 - \beta_2 + \beta_3 = 40 \\ \beta_0 + \beta_1 - \beta_2 - \beta_3 = 45 \\ \beta_0 - \beta_1 + \beta_2 - \beta_3 = 50 \\ \beta_0 + \beta_1 + \beta_2 + \beta_3 = 65 \end{cases} \Rightarrow \begin{cases} \beta_1 = 5 \\ \beta_2 = 7.5 \\ \beta_3 = 2.5 \\ \beta_0 = 50 \end{cases}$$

β_0 means the unbiased standard salary is \$50.

β_1 means the degree that gender influences salary is \$5.

female for -5 and male for +5. A total difference of \$10 in gender

β_2 means the degree that race influences salary is \$7.5.

non-white for -7.5 and white for +7.5. A total difference of \$15 in gender

β_3 means the degree that gender and race influence salary for \$2.5.

non-white, female and white, male for +2.5. the other situations for -2.5

A total difference of \$5.

Moreover, the interaction between gender and race exists. Male and White have separate positive effects on the salary but if both are offered together then the total effect on the salary can be greater than their sum.

3.12

a)

```
cobb = read.csv("Desktop/Predictive Analytics/Assignment/HW 2/Cobb-Douglas.csv")
row(cobb)
fit1 = lm(log(cobb$output)~log(cobb$capital) + log(cobb$labor))
summary(fit1)
```

The result is:

```
Call:
lm(formula = log(cobb$output) ~ log(cobb$capital) + log(cobb$labor))

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(cobb$capital)  0.20757    0.01719   12.08   <2e-16 ***
log(cobb$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
```

b)

```
var <- vcov(fit1)
var_b3 = var[2,2] + var[3,3] + 2 * var[2,3]
t = (0.20757 + 0.71485 - 1) / sqrt(var_b3)#use estimated b1 and b2 as x bar, mu = 1
t
t > qt(0.025,566) #according to regression result, 566 is degree of freedom
```

The result is:

```
> var <- vcov(fit1)
> var_b3 = var[2,2] + var[3,3] + 2 * var[2,3]
> t = (0.20757 + 0.71485 - 1) / sqrt(var_b3)#use estimated b1 and b2 as x bar, mu = 1
> t
[1] -4.509281
> t > qt(0.025,566) #according to regression result, 566 is degree of freedom
[1] FALSE
> 
```

t < lower confidence interval limit at 95% level, therefore, we reject H0: b1 + b2 = 1, which means: based on current sample, there is not enough evidence to prove that labor and capital count for all productivity.

c)

Fit new full model and partial model:

```
cl <- log(cobb$capital) - log(cobb$labor)
fit2 = lm((log(cobb$output)-log(cobb$labor)) ~ cl+log(cobb$labor))
summary(fit2)

fit3 = lm((log(cobb$output)-log(cobb$labor)) ~ cl)
summary(fit3)
f = (0.2393-0.212) * 566 / ((2-1) * (1-0.2393))
# Rp^2 and Rq^2 are r-squared of full model and partial model
# n - (p+1) is the d.f. of the full model
f
abs(f) < qf(0.95,1,566)
```

The result is:

```
Call:
lm(formula = (log(cobb$output) - log(cobb$labor)) ~ cl)

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 ***
cl           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

> f = (0.2393-0.212) * 566 / ((2-1) * (1-0.2393))
> # Rp^2 and Rq^2 are r-squared of full model and partial model
> # n - (p+1) is the d.f. of the full model
> f
[1] 20.31261
> abs(f) < qf(0.95,1,566)
[1] FALSE
> 
```

Since f is larger than upper limit of confidence interval, we reject the hypothesis that b3 = 0. The conclusion is the same as b).

3.15

a)

```r
salary <- read.csv("Desktop/Predictive Analytics/Assignment/HW 2/salaries.csv")
salary$Gender <- relevel(salary$Gender,ref = "Male")
salary$Dept <- relevel(salary$Dept,ref = "Purchase")
fit_salary = lm(log10(salary$Salary)~salary$YrsEm+salary$PriorYr+salary$Education+salary$Super
               +salary$Dept+salary$Gender)
summary(fit_salary)
```

The result is:

```
Call:
lm(formula = log10(salary$Salary) ~ salary$YrsEm + salary$PriorYr +
    salary$Education + salary$Super + salary$Dept + salary$Gender)

Residuals:
      Min        1Q    Median        3Q       Max
-0.089659 -0.024036 -0.004498  0.028587  0.089410

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.4287934  0.0213399 207.535  < 2e-16 ***
salary$YrsEm        0.0074788  0.0011931   6.269 2.72e-07 ***
salary$PriorYr      0.0016839  0.0019568   0.861 0.395039
salary$Education    0.0170345  0.0033360   5.106 1.02e-05 ***
salary$Super        0.0003901  0.0008056   0.484 0.631115
salary$DeptAdvertse -0.0387774  0.0249146  -1.556 0.128124
salary$DeptEngineer -0.0057292  0.0197703  -0.290 0.773597
salary$DeptSales    -0.0937783  0.0225745  -4.154 0.000185 ***
salary$GenderFemale  0.0230683  0.0142917   1.614 0.115002
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04586 on 37 degrees of freedom
Multiple R-squared:  0.8634,    Adjusted R-squared:  0.8338
F-statistic: 29.22 on 8 and 37 DF,  p-value: 9.629e-14
```

The regression result matches the equation given in the question.

b)

```r
salary$Gender <- relevel(salary$Gender,ref = "Female")
salary$Dept <- relevel(salary$Dept,ref = "Sales")
fit_salary1 = lm(log10(salary$Salary)~salary$YrsEm+salary$PriorYr+salary$Education+salary$Super
                +salary$Dept+salary$Gender)
summary(fit_salary1)
```

The result is:

```
Call:
lm(formula = log10(salary$Salary) ~ salary$YrsEm + salary$PriorYr +
    salary$Education + salary$Super + salary$Dept + salary$Gender)

Residuals:
      Min        1Q    Median        3Q       Max
-0.089659 -0.024036 -0.004498  0.028587  0.089410

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.3580834  0.0248414 175.436  < 2e-16 ***
salary$YrsEm         0.0074788  0.0011931   6.269 2.72e-07 ***
salary$PriorYr       0.0016839  0.0019568   0.861 0.395039
salary$Education     0.0170345  0.0033360   5.106 1.02e-05 ***
salary$Super         0.0003901  0.0008056   0.484 0.631115
salary$DeptPurchase  0.0937783  0.0225745   4.154 0.000185 ***
salary$DeptAdvertse  0.0550009  0.0230111   2.390 0.022045 *
salary$DeptEngineer  0.0880491  0.0180562   4.876 2.07e-05 ***
salary$GenderMale   -0.0230683  0.0142917  -1.614 0.115002
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04586 on 37 degrees of freedom
Multiple R-squared:  0.8634,    Adjusted R-squared:  0.8338
F-statistic: 29.22 on 8 and 37 DF,  p-value: 9.629e-14
```

According to the new regression result, coefficients of variables are:

Variable	Coefficient
Male	-0.023
Purchase	0.094
Advertise	0.055
Engineer	0.088

c)

Different p-values means under different reference categories, the same apartment has different accuracies to predict the dependent variable. If the regression reference is Purchase, then department of Engineer is highly non-significant, and cannot be used to predict the salary. However, if the reference is Sales, then based on this category, department of Engineer is highly significant, thus is an accurate dimension to predict the salary.

d)

```
salary$Dept <- relevel(salary$Dept,ref = "Purchase")
fit_salary2 = lm(log10(salary$Salary)~salary$YrsEm+salary$Education+salary$Dept)
summary(fit_salary2)
```

The regression result is:

```
Call:
lm(formula = log10(salary$Salary) ~ salary$YrsEm + salary$Education +
    salary$Dept)

Residuals:
      Min        1Q    Median        3Q       Max
-0.114193 -0.028068 -0.002002  0.033938  0.081774

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.439005   0.019804 224.142  < 2e-16 ***
salary$YrsEm         0.007660   0.001208   6.341 1.57e-07 ***
salary$Education     0.018371   0.003124   5.881 6.95e-07 ***
salary$DeptAdvertse -0.036488   0.025311  -1.442 0.157208
salary$DeptEngineer -0.002507   0.020037  -0.125 0.901046
salary$DeptSales    -0.087593   0.022740  -3.852 0.000414 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04677 on 40 degrees of freedom
Multiple R-squared:  0.8464,    Adjusted R-squared:  0.8272
F-statistic: 44.09 on 5 and 40 DF,  p-value: 3.099e-15
```

From the regression result, it can be seen that after dropping the PriorYr, Super, and Gender independent variables, p-values of YrsEm and Education are small, indicating that they are correlated to the salary. For one more year employed in the company, you can get $0.007 more in your salary. For one more year educated after high school, you can get $0.18 more in your salary.

Implementing extra SS method:

```
f = (0.8634 - 0.8464) * 37 / ((8 - 5) * (1 - 0.8634))
# Rp and Rq are R-squared in a) and d), n - (p+1) is the d.f in a)
#p and q are independent variables in a) and c)
f
abs(f) < qf(0.95,3,37)
```

The result is:

```
> f = (0.8634 - 0.8464) * 37 / ((8 - 5) * (1 - 0.8634))
> # Rp and Rq are R-squared in a) and d), n - (p+1) is the d.f in a)
> #p and q are independent variables in a) and c)
> f
[1] 1.534895
> abs(f) < qf(0.95,3,37)
[1] TRUE
>
```

It can be seen that we accept the hypothesis that coefficients of gender, PriorYr, and Super are 0.