Textbook

3.14 H0: beta(gender) = beta(experience) = beta(months) =0
Ha: at least one of them is not zero
r =3. (test 3 parameters in H0).
$MSH_0=SSH_0/r=(SSE_0-SSE)/r$ = (38460756-22657938)/3 = 5267606. MSE = 257477
And we have F = MSH_0/MSE = 20.45855 ~ F(0.95, 3, 88)= 2.708186.
Since F-statistics is larger than F(0.95,3,88), we reject the H0 and conclude at least one of gender, beta and months is indeed needed to add into the model.

5.6 (a) cor(Y, Horsepower) = 0.8611619. (Y: price)

(b) In the model 3 with interaction terms, all variables except Horsepower becomes insignificant (P-value larger than 0.05). So we should not use model 3 to predict Y. In model 2, Germany variable is highly insignificant since P-value is larger than 0.05. So, we need to exclude Germany first.

After exclude Germany from the category of Country (by substituting "Germany" with "Others") run the regression and get:

```
lm(formula = Y ~ Horsepower + NewCountry, data = Text5.6)

Residuals:
    Min       1Q   Median       3Q      Max
-11.1073  -2.3465  -0.1521   2.1123  12.4388

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -4.244450   1.461595  -2.904  0.00479 **
Horsepower       0.174976   0.009769  17.911  < 2e-16 ***
NewCountryUSA   -3.195623   1.238057  -2.581  0.01172 *
NewCountryJapan -3.856653   1.243542  -3.101  0.00268 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.703 on 78 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.8056,	Adjusted R-squared:  0.7981
F-statistic: 107.8 on 3 and 78 DF,  p-value: < 2.2e-16
```

The new model 2 is: Y^= -4.24445 +0.174976Horsepower -3.195623USA -3.856653Japan
Given USA =1, Japan=0, Germany=0 and Horsepower = 100, LS estimated price = 10.05753 thousands.

(c) Let Horsepower equal to a fixed number x, x>= 0.

The coefficient of Japan is smallest in the new model 2, So, we can get Y^(Japan) is the smallest and Japan car is least expensive car by holding the horsepower fixed.

(d) H0: beta(HP*USA) = 0; Ha: beta(HP*USA) != 0
From output of model 3, we can see p-value of HP*USA is 0.2204, which is larger than 0.05, so we fail to reject H0 and conclude there is not an interaction between horsepower and USA.

H0: beta(HP*Japan) = 0; Ha: beta(HP*Japan) != 0
From output of model 3, we can see p-value of HP*Japan is 0.0631, which is larger than 0.05, so we fail to reject H0 and conclude there is not an interaction between horsepower and Japan.

H0: beta(HP*Germany) = 0; Ha: beta(HP*Germany) != 0
From output of model 3, we can see p-value of HP*Germany is 0.1560, which is larger than 0.05, so we fail to reject H0 and conclude there is not an interaction between horsepower and Germany.
Overall, we can conclude there is an interaction between Country and Horsepower.

(e) Because (d) shows there is not interaction between Country and Horsepower, so all interaction variables can be excluded. Since Germany is not significant in Model 2, so the new Model 2 is good to be used here.

H0: beta(USA) = 0; Ha: beta(USA) != 0
We can see p-value of USA is 0.01172, which is smaller than 0.05, so USA is significant in the model.

H0: beta(Japan) = 0; Ha: beta(Japan) != 0
We can see p-value of Japan is 0.00268, which is smaller than 0.05, so Japan is significant in the model.
USA and Japan are significant and indeed needed included in the model. Therefore, we can conclude Country is an important predictor of the price of a car.

(f) Since Germany is not significant in Model 2 and there is no need include Germany in the mode, we can reduce Germany and let it joins with Other category together.

(g)
H0: $\beta_{USA} = \beta_{Japan}$ (also can be $\beta_{USA} - \beta_{Japan} = 0$) ; Ha: $\beta_{USA} \mathrel{!=} \beta_{Japan}$
T-statistics: T = [($\beta_{USA} - \beta_{Japan}$)-0]/ SE($\beta_{USA} - \beta_{Japan}$) ~ t(n-2,0.025)=t(88,0.025)
And SE($\beta_{USA} - \beta_{Japan}$) =sqrt(Var($\beta_{USA}$)+var($\beta_{Japan}$) – 2*Cov($\beta_{USA}$, β_{Japan}))

5.9 After change the baseline to be "0" for D and set a new variable **Z = G*I**, run the regression:

```
Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.5111627  0.0321992  15.875 2.40e-10 ***
I         -0.0201077  0.0168979  -1.190   0.2539
D           0.0546159  0.0205705   2.655   0.0188 *
W            0.0133905  0.0422639   0.317   0.7560
Z           0.0096901  0.0017712   5.471 8.24e-05 ***
P          -0.0007224  0.0040046  -0.180   0.8594
N          -0.0051822  0.0038083  -1.361   0.1951
```

(a) V^ = 0.5111627 -0.0201077 I+0.0546159 D +0.0133905 W+0.0096901G*I- 0.0007224 *P -0.0051822 *N

The coefficient of D is 0.0546159. If a Democratic incumbent is running for election, D=1, the Democratic share of the two-party presidential vote will increase 0.0633485 units. If a Republic incumbent is running for election, D=-1, the Democratic share of the two-party presidential vote will decrease 0.0633485 units. If Other incumbent is running for election, D=0, the Democratic share of the two-party presidential vote will not change.

V^ = 0.5657786 + -0.0201077 I +0.0133905 W+0.0096901G*I- 0.0007224 *P -0.0051822 *N when D=1

V^ = 0.4565468 + -0.0201077 I +0.0133905 W+0.0096901G*I- 0.0007224 *P -0.0051822 *N when D=-1

V^ = 0.5111627 + -0.0201077 I +0.0133905 W+0.0096901G*I- 0.0007224 *P -0.0051822 *N when D=-1

(b) We don't need to keep I in the above model since the p-value of I is 0.2539, which is larger than 0.05. *I* is not significant and should be excluded from the model.

(c) Yes. G*I 's corresponding P-value is 8.24e-05, which is highly significant. G*I should be included in the model.

(d) After run many different models, the best model I found is below:

```
Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.481837   0.008522  56.543  < 2e-16 ***
D           0.050000   0.014198   3.522 0.002619 **
G*I           0.007626   0.001603   4.757 0.000183 ***
I*P        -0.003705   0.001994  -1.858 0.080534 .
```

V^ = 0.481837 + 0.050000 D+0.007626 G*I -0.003705 I*P

I tried all possible interaction and found out I*P is a significant interaction added into the model. In the best model, all variables are significant under 90% CI test.

5.10 (a)

```
lm(formula = V ~ factor(I) + D1 + D2 + factor(W) + Z + P + N)

Residuals:
     Min       1Q   Median       3Q      Max
-0.044201 -0.022728 -0.002548  0.011671  0.084681

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.5260743  0.0369736  14.228 2.64e-09 ***
factor(I)1  -0.0411965  0.0349716  -1.178 0.259912
D1           0.0633485  0.0312177   2.029 0.063423 .
D2          -0.0469714  0.0291912  -1.609 0.131600
factor(W)1   0.0123948  0.0436938   0.284 0.781127
G*I          0.0094222  0.0019580   4.812 0.000339 ***
P           -0.0006963  0.0041333  -0.168 0.868808
N           -0.0051083  0.0039349  -1.298 0.216773
```

V^ = 0.5260743-0.0411965I+0.0633485D1-0.0469714D2 +0.0123948W+0.0094222G*I-0.0006963*P -0.0051083*N.

A Democratic incumbent is running for election, the Democratic share of the two-party presidential vote will estimately increase 0.0633485 units.

A Republic incumbent is running for election, the Democratic share of the two-party presidential vote will estimately decrease 0.0469714 units.

(b) when $\alpha 1 = - \alpha 2$, V= $\beta 0+ \beta 1$*I+ $\alpha 1$*(D1-D2) + $\beta 3$*W+ $\beta 4$*(G*I) + $\beta 5$*P+ $\beta 6$*N+error. 5.11 is a special case of 5.12 since D2=0 when D1=1, which is category “1” above. D1=0 when D2=1, which is category “-1” above. D1=0, D2=0 is category “0” above. So we can consider (D1-D2) being a variable with 3 categories, which is variable D from 5.9.

(c)when $\alpha 1= - \alpha 2$, we get reduce model under H0: V= $\beta 0+ \beta 1$*I+ $\alpha 1$*(D1-D2) + $\beta 3$*W+ $\beta 4$*(G*I) + $\beta 5$*P+ $\beta 6$*N+error, which is the model from 5.9.
Full model from 5.10.

H0: Reduced model = full model ; Ha : reduced model ≠ full model

Run the ANOVA

```
Model 1: V ~ I + D + W + G:I + P + N
Model 2: V ~ I + D1 + D2 + W + G:I + P + N
  Res.Df      RSS Df  Sum of Sq      F Pr(>F)
1     14 0.023686
2     13 0.023423  1 0.00026227 0.1456  0.709
```

P-value = 0.709 larger than 0.05, so we fail to reject H0 and conclude reduced model = full model, so $\alpha 1 = -\alpha 2$.

Prof's book

3.4 (a) Yes. The fitted equation is checked.

```
lm(formula = log.Salary. ~ YrsEm + PriorYr + Education + Super + 
    factor(Female) + factor(Advertising) + factor(Engineering) + 
    factor(Sales))

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 ***
YrsEm                0.0074788  0.0011931   6.269 2.72e-07 ***
PriorYr              0.0016839  0.0019568   0.861 0.395039    
Education            0.0170345  0.0033360   5.106 1.02e-05 ***
Super                0.0003901  0.0008056   0.484 0.631115    
factor(Female)1      0.0230683  0.0142917   1.614 0.115002    
factor(Advertising)1 -0.0387774  0.0249146  -1.556 0.128124    
factor(Engineering)1 -0.0057292  0.0197703  -0.290 0.773597    
factor(Sales)1      -0.0937783  0.0225745  -4.154 0.000185 ***

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
```

(b)

```
Coefficients:
         Estimate Std. Error t value Pr(>|t|)    
```

```
(Intercept)          4.3580834  0.0248414 175.436  < 2e-16 ***
YrsEm                0.0074788  0.0011931   6.269 2.72e-07 ***
PriorYr              0.0016839  0.0019568   0.861 0.395039    
Education            0.0170345  0.0033360   5.106 1.02e-05 ***
Super                0.0003901  0.0008056   0.484 0.631115    
factor(Male)1       -0.0230683  0.0142917  -1.614 0.115002    
factor(Advertising)1  0.0550009  0.0230111   2.390 0.022045 *  
factor(Engineering)1  0.0880491  0.0180562   4.876 2.07e-05 ***
factor(Marketing)1    0.0937783  0.0225745   4.154 0.000185 ***
```

If we set Female and Sales to be baseline, the new regression is above. So the new coefficient for Male is -0.0230683, for Advertising is 0.0550009, for Engineering is 0.0550009, and for Marketing is 0.0550009.

(c) The coefficients for the other categories represent the differences from that group. Engg is nonsignificant with base categories Male&Marketing, so the influence on log(salary) is not much difference from Male and Marketing's influence on log(salary), so, here, Engg is not significant. However, Engg is significant with base categories Female&Sales, so the influence on log(salary) is much difference from Female and Sale's influence on log(salary), so, here, Engg is significant.

(d) after drop those three variables, run the regression:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.439005   0.019804 224.142  < 2e-16 ***
YrsEm        0.007660   0.001208   6.341 1.57e-07 ***
Education    0.018371   0.003124   5.881 6.95e-07 ***
Advertising -0.036488   0.025311  -1.442 0.157208
Engineering -0.002507   0.020037  -0.125 0.901046
Sales       -0.087593   0.022740  -3.852 0.000414 ***
```

If we do 95% CI test, Advertising and Engineering is not significant since its P-value is larger than 0.05.

3.5 y= $\begin{pmatrix} y1 \\ y2 \\ y3 \\ \vdots \\ yn \end{pmatrix}$, X = $\begin{pmatrix} x1 \\ x2 \\ x3 \\ \vdots \\ xn \end{pmatrix}$, e= $\begin{pmatrix} e1 \\ e2 \\ e3 \\ \vdots \\ en \end{pmatrix}$, $\beta = \beta$. Given 3.7 formula: β ^ = (X'X)$^{-1}$X'y

X'X = $\sum_{i=1}^{n} x_i^2$, (X'X)^-1 = $\frac{1}{\sum_{i=1}^{n} x_i^2}$; X'y = $\sum_{i=1}^{n} x_i y_i$

So β ^ = (X'X)$^{-1}$X'y = $\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$