Problem 3.4

Part a.

```
> fit1 = lm(log_salary ~ YrsEm + PriorYr + Education + Super + Female +
+                        Advertising + Engineering + Sales, data =mydata)
> summary(fit1)

Call:
lm(formula = log_salary ~ YrsEm + PriorYr + Education + Super +
    Female + Advertising + Engineering + Sales, data = mydata)

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
Female       0.0230683  0.0142917   1.614 0.115002
Advertising -0.0387774  0.0249146  -1.556 0.128124
Engineering -0.0057292  0.0197703  -0.290 0.773597
Sales       -0.0937783  0.0225745  -4.154 0.000185 ***
---
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

> # This fitted equation matches the one given in the book
```

Part b.

```
> fit2 = lm(log_salary ~ YrsEm + PriorYr + Education + Super + Male +
+             Advertising + Engineering + Marketing, data =mydata)
> summary(fit2)

Call:
lm(formula = log_salary ~ YrsEm + PriorYr + Education + Super +
    Male + Advertising + Engineering + Marketing, data = mydata)

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 ***
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
Male        -0.0230683  0.0142917  -1.614 0.115002
Advertising  0.0550009  0.0230111   2.390 0.022045 *
Engineering  0.0880491  0.0180562   4.876 2.07e-05 ***
Marketing    0.0937783  0.0225745   4.154 0.000185 ***
---
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 new coefficients for Male, Advertising, Engineering and Marketing
> # are highlighted above after using Female and Sales as base categories
> # As expected for engineering: 0.0937783-0.0057292 = 0.0880491
> # As expected for advertising: 0.0937783 -0.0387774   = 0.0550009
> # As expected for marketing: -(-0.0937783)= 0.0937783
> # And male: (-0.0230683)= -0.0230683
```

Part c.

```
> # For model in part a), the coefficient of engineering is the effect
> # of engineering on the salary compared to marketing after accounting
> # for other predictors. So the difference in salary between engineering
> # and marketing is not significant because p-value = 0.774 > 0.05.

> # For model in part b), the coefficient of engineering is the effect
> # of engineering on the salary compared to sales after accounting
> # for other predictors. So the difference in salary between engineering
> # and sales is significant because p-value <0.001.

> # If the coefficient of a dummy variable is nonsignificant, it tells
> # you that the effect of the dummy variable on the response compared
> # to the base category after accounting for other predictors is not signifi
cant.
> # Then you can combine nonsignificant category to the base category.
```

Part d.

```
> fit3 = lm(log_salary ~ YrsEm  + Education  +
+             Advertising + Engineering + Sales, data =mydata)
> summary(fit3)

Call:
lm(formula = log_salary ~ YrsEm + Education + Advertising + Engineering +
    Sales, data = mydata)

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 ***
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 ***
---
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

>
> anova(fit1,fit3)
Analysis of Variance Table
```

```
Model 1: log_salary ~ YrsEm + PriorYr + Education + Super + Female + Advertis
ing +
    Engineering + Sales
Model 2: log_salary ~ YrsEm + Education + Advertising + Engineering +
    Sales
  Res.Df      RSS Df  Sum of Sq      F Pr(>F)
1     37 0.077830
2     40 0.087479 -3 -0.0096496 1.5291 0.2231

> # The F-test (p-value > 0.05) to compare the full model vs. reduced model s
hows that the null hypothesis that PriorYr, Super and Female are zero cannot
be rejected, so conclude that the coefficients are not significantly differen
t than zero. Thus, the reduced model is preferred (the variables in the reduc
ed model adequately explain the variation as the variables in the full model)
>
> # Fitting the new model the coefficients of YrsEm, Education and Sales rema
in significant (p-value < 0.05), and the coefficients of Advertising (p-value
=.157) and Engineering (p-value = 0.901) remain insignificant at 0.05 level.
> # The R^2 remains about the same at 0.86. This tells us that Female, Super
> # and PriorYr were not contributing much to explaining the variation in sal
ary. Thus, the model after dropping the variables seems to be preferred.
>
> # Because Advertising and Engineering are insignificant compared to
> # the base (Marketing), then it might be a good idea to combine
> # Advertising, Engineering and Marketing into one "Other" category.
> # So you would need only 1 dummy variables representing Sales vs. "Other".
>
```

Problem 3.5

Note: β is a scalar, so can write: $\underline{y} = \beta \underline{X} + \underline{\varepsilon}$

3.5 $y_i = \beta x_i + \varepsilon_i$

$\Downarrow$

$\underline{y} = \underline{X}\,\underline{\beta} + \underline{\varepsilon}$ where $\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}$, $\underline{\beta} = [\beta]_{1\times 1}$, $\underline{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}_{n\times 1}$

Using eq (3.7): $\hat{\underline{\beta}} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{y}$

$$\hat{\underline{\beta}} = \left([x_1 \cdots x_n] \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \right)^{-1} [x_1 \cdots x_n] \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$= \left[\left(\sum_{i=1}^{n} x_i^2 \right)^{-1} \left(\sum_{i=1}^{n} x_i y_i \right) \right]$$

$$\Rightarrow \hat{\underline{\beta}} = \left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \right] \Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Problem 3.14

3.14 FM : salary ~ gender + educ + exp + months

RM : salary ~ educ

H_0 : Reduced model is adequate

H_1 : full model is adequate

H_0 : $\beta_{gender} = \beta_{exp} = \beta_{months} = 0$

H_1 : at least one of them is $\neq 0$

$$F = \frac{(RSS_{reduced} - RSS_{full}) / (df_{reduced} - df_{full})}{RSS_{full} / df_{full}}$$

$$F = \frac{(38460756 - 22657938) / (91 - 88)}{22657938 / 88}$$

$F = 20.45858$

$F_{(3, 88, 0.05)} = 2.708$

Since $F > F_{(3, 88, 0.05)}$, reject H_0 that the coefficients of gender, experience and months are zero and conclude that not all these coefficients can be taken as zero. The full model is preferred over the reduced model (The variable education does not adequately explain the variation as the full set of four variables).

Problem 5.6

Part a.

Correlation coefficient between price and horsepower:

```
> # r = sqrt(SSR/SST) = sqrt(SSR/(SSR + SSE))
> SSR = 4604.7
> SSE = 1604.44
>
> r = sqrt(SSR/(SSR+SSE))
> r
[1] 0.8611622
```

Part b.

```
> # The estimated price of an American car with a 100 hp engine is:
(answers in thousands)

> # Using Model 1 (American,Japanese,German, and Others are treated the same)
> -6.107 + 0.169*100
[1] 10.793
>
> # Using Model 2
> -4.117 + 0.174*100 -3.162*1
[1] 10.121
>
> # Using Model 3
> -10.882 + 0.237*100 +2.076*1 - 0.052*100*1
[1] 9.694
```

Part c.

```
> # Using Model 2
> # With respect to "Others" ( base category), Japan (-3.818) has the lowest
price compared to Germany (0.311) and USA (-3.162) because the coefficient is
the most negative. The negative coefficient means that the price of Japan is
lower than "Others". In addition, the p-value = 0.0061 shows the difference
between Japan and "Others" is significant.
> # Thus, least expensive car is from Japan.
>
> # Model 3
> # Cannot hold horsepower constant because there is an interaction
> # between country and horsepower, so the least expensive car depends
> # on the horsepower (we cannot estimate the country effect independent of H
P)
>
> # For example, if HP = 100
>
> # Others
> -10.882 + 0.237*100 + 2.076*0 + 4.755*0 + 11.774*0 -0.052*0*100 - 0.077*0*1
00  - 0.095*0*100
[1] 12.818
> # USA
> -10.882 + 0.237*100 + 2.076*1 + 4.755*0 + 11.774*0 -0.052*1*100  - 0.077*0*
100  - 0.095*0*100
[1] 9.694
> # Japan
> -10.882 + 0.237*100 + 2.076*0 + 4.755*1 + 11.774*0 -0.052*0*100  - 0.077*1*
100  - 0.095*0*100
[1] 9.873
```

```
> # Germany
> -10.882 + 0.237*100 + 2.076*0 + 4.755*0 + 11.774*1 -0.052*0*100  - 0.077*0*
100  - 0.095*1*100
[1] 15.092

> # So least expensive is USA, followed by  Japan, "Others", and Germany
>
> # For example, if HP = 1000
>
> # Others
> -10.882 + 0.237*1000 + 2.076*0 + 4.755*0 + 11.774*0 -0.052*0*100 - 0.077*0*
1000  - 0.095*0*1000
[1] 226.118
> # USA
> -10.882 + 0.237*1000 + 2.076*1 + 4.755*0 + 11.774*0 -0.052*1*100  - 0.077*0
*1000  - 0.095*0*1000
[1] 222.994
> # Japan
> -10.882 + 0.237*1000 + 2.076*0 + 4.755*1 + 11.774*0 -0.052*0*100  - 0.077*1
*1000  - 0.095*0*1000
[1] 153.873
> # Germany
> -10.882 + 0.237*1000 + 2.076*0 + 4.755*0 + 11.774*1 -0.052*0*100  - 0.077*0
*1000  - 0.095*1*1000
[1] 142.892
> # So least expensive is Germany, followed by Japan, USA and "Others"
```

Part d.

```
> # F-test
> # Compare model 3 vs model 2
> # Ho: Reduced model (2) is adequate
> # H1: Full model (3) is adequate
>
> # Alternatively,
> # Ho: b_HP*b_USA = b_HP*b_Japan = b_HP*b_Germany = 0
> # H1: At least one of b_HP*b_USA, b_HP*b_Japan or b_HP*b_Germany is differe
nt than zero
>
> # F = ((RSS_reduced - RSS_full)/(df_reduced-df_full))/(RSS_full/df_full)
>
> ((1390.31-1319.85)/(85-82))/(1319.85/82)
[1] 1.459186
> qf(.05,df1=3,df2=82,lower.tail = FALSE)
[1] 2.715937
>
> # Since F_stat = 1.459 < F_crit = 2.716,  the null hypothesis that
> # the interaction terms are zero cannot be rejected, so conclude that the c
oefficients are not significantly different than zero. Thus, the reduced mode
l is preferred  (the variables in the reduced model adequately explain the va
riation as the variables in the full model). Conclusion: there is not a signi
ficant interaction between country and HP.

# Note that we could also test individual coefficients for the interaction te
rms, which in this case all p-values (0.2204,0.0631 and 0.1560) are insignifc
ant, suggesting that there is not a significant interaction between any of th
e countries and HP. This conclusion is consistent with the above.
```

Part e.

```
> # F-test
> # Compare model 2 vs model 1
> # Ho: Reduced model (1) is adequate
```

```
> # H1: Full model (2) is adequate
>
> # Alternatively,
> # Ho: b_USA = b_Japan = b_Germany = 0
> # H1: At least one of b_USA, b_Japan or b_Germany is different than zero
>
> # F = ((RSS_reduced - RSS_full)/(df_reduced-df_full))/(RSS_full/df_full)
>
> ((1604.44-1390.31)/(88-85))/(1390.31/85)
[1] 4.363787
> qf(.05,df1=3,df2=85,lower.tail = FALSE)
[1] 2.711921
>
> # Since F_stat = 4.364 > F_crit = 2.712,  reject the null hypothesis that
> # the coefficients of USA, Japan and Germany are zero, so conclude that at
least one of the coefficients are  significantly different than zero. Thus, t
he full model is preferred  (the variables in the reduced model do not adequa
tely explain the variation as the variables in the full model). Conclusion: g
iven HP, Country is a significant predictor of car price

# Note that we could also test individual coefficients for the coefficient, w
hich in this case USA and Japan are significant (p-values of 0.0216 and 0.006
1), while Germany is insignificant (p-value = 0.8682).So USA and Japan vs. Ot
hers are significant predictors of car price, while Germany is not. This conc
lusion is consistent with the above, which says country as a whole is a signf
icant predictor of car price.
```

Part f.

```
> # Because the p-value > 0.05 for Germany in model 2, the null hypothesis th
at this coefficient is zero cannot be rejected, so conclude that is not signi
ficantly different than zero. Thus, there is not a significant difference of
prices between German and Others car, so it is recommended to add Germans to
the "Other" category.
```

Part g.

```
> # Ho: b_USA = b_Japan
> # H1: b_USA diff b_Japan
>
> # One could fit a model by replacing USA = Japan (called reduced model) and
compared it to the full model (model 2) by using a F-test that will reject or
not the null hypothesis. The statistic would be:
> # F = ((RSS_reduced - RSS_full)/(df_reduced-df_full))/(RSS_full/df_full)
> # Compared to the F_critical with df1 = df_reduced-df_full, df2= df_full at
0.05 level

>
> # Alternatively,
you can use a t-test:  t=(b_USA-b_Japan)/se(b_USA-b_Japan) where
se(b_USA-b_Japan)= sqrt(var(b_USA)+ var(b_Japan) - 2cov(b_USA, b_Japan))
> # and df = n-(p+1) = 85.
> # If |t| > |t_crit (df,0.05/2)|, then reject null, otw cannot reject.
```

Problem 5.9

Part a.

$\hat{V} = \hat{\beta}_0 + \hat{\beta}_1 I + \hat{\beta}_2 D + \hat{\beta}_3 W + \hat{\beta}_4 (G \cdot I) + \hat{\beta}_5 P + \hat{\beta}_6 N$

$\hat{V} = 0.511 - 0.020I + 0.055D + 0.013W + 0.0097(G \cdot I) - 0.00072P - 0.0052N$

So the equation for each possible values of D is (only intercept changes):

D=0:

- $\hat{V} = \hat{\beta}_0 + \hat{\beta}_1 I + \hat{\beta}_3 W + \hat{\beta}_4 (G \cdot I) + \hat{\beta}_5 P + \hat{\beta}_6 N$
- $\hat{V} = 0.511 - 0.020I + 0.013W + 0.0097(G \cdot I) - 0.00072P - 0.0052N$

D=-1:

- $\hat{V} = (\hat{\beta}_0 - \hat{\beta}_2) + \hat{\beta}_1 I + \hat{\beta}_3 W + \hat{\beta}_4 (G \cdot I) + \hat{\beta}_5 P + \hat{\beta}_6 N$
- $\hat{V} = 0.457 - 0.020I + 0.013W + 0.0097(G \cdot I) - 0.00072P - 0.0052N$

D=1:

- $\hat{V} = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 I + \hat{\beta}_3 W + \hat{\beta}_4 (G \cdot I) + \hat{\beta}_5 P + \hat{\beta}_6 N$
- $\hat{V} = 0.566 - 0.020I + 0.013W + 0.0097(G \cdot I) - 0.00072P - 0.0052N$

The coefficient of D says that the democratic share of the two-part presidential vote (V) is on average 0.055 higher if there is a democratic incumbent running for election (D=1) and 0.055 lower if a republican incumbent is running for election (D=-1) compare to any other case ("otherwise") after accounting for the other predictors. The p-value < 0.05, so D is a significant predictor of V.

```
> fit = lm(V~I + D + W + G:I + P + N, data = mydata)
> summary(fit)

Call:
lm(formula = V ~ I + D + W + G:I + P + N, data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.041742 -0.021066 -0.003611  0.011760  0.087914 

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    
P           -0.0007224  0.0040046  -0.180   0.8594    
N           -0.0051822  0.0038083  -1.361   0.1951    
I:G          0.0096901  0.0017712   5.471 8.24e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04113 on 14 degrees of freedom
Multiple R-squared: 0.7898,	Adjusted R-squared: 0.6998 
F-statistic: 8.769 on 6 and 14 DF,  p-value: 0.0004347
```

Part b.

For coefficient I: P – value = 0.2539 > 0.05, so do not reject null hypothesis that the coefficient is equal to zero, concluding that the coefficient is insignificant (not significantly different than zero) and thus should not be kept in the model. However, it is usually not good practice to include the interaction without the main effect (although G is not in the model but I*G is significant).

Part c.

For coefficient IG: P – value < 0.05, so reject null hypothesis that the coefficient is equal to zero, concluding that the coefficient is significant (significantly different than zero) and thus should be kept in the model. Note that G is not in the model; it is usually not good practice to include the interaction without the main effect.

Part d.

It makes sense that the effect on V of both the incumbent (I) and the running incumbent (D) should depend on the absolute value of the grow rate of the GDP deflator in the first 15 quarters of the administration (P) and number of quarters in the first 15 quarters of the administration in which the growth rate of real per capita GDP is greater than 3.2% (N). In addition, it also seems that the effect of the incumbent (I) might depend also on the running incumbent (D). Thus, the interactions D*I, P*I,N*I, P*D and N*D were added to the model. The strategy was to start with this augmented model and remove insignificant terms until all terms in the model are significant and also looking at the adjusted R^2. The highlighted coefficient is the one chosen at each to be removed.

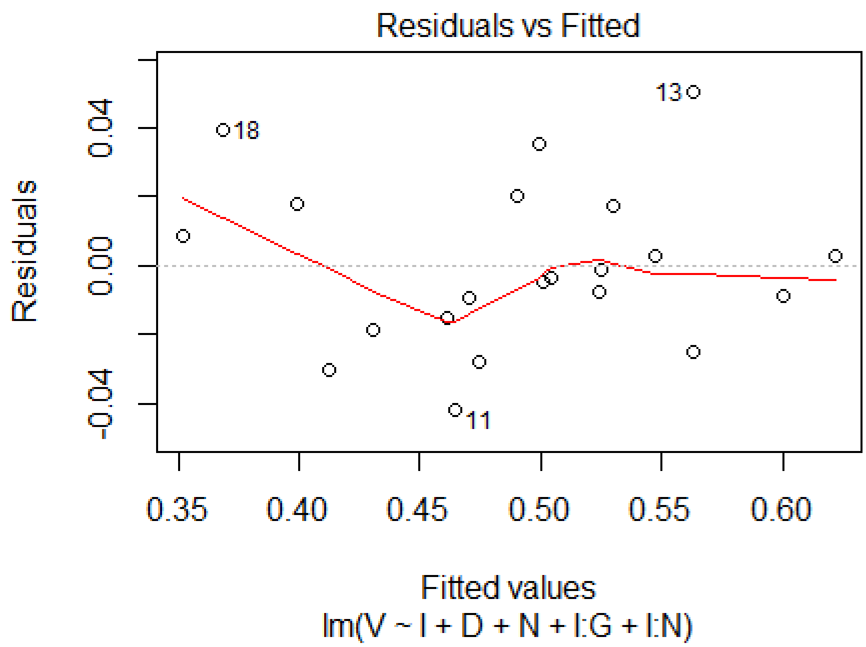
```
> fit1 = lm(V~I + D + W + G:I + P + N + D*I + P*I + N*I + N*D + P*D, data = m
ydata)
> summary(fit1)

Call:
lm(formula = V ~ I + D + W + G:I + P + N + D * I + P * I + N *
    I + N * D + P * D, data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max
-0.041523 -0.010187 -0.000110  0.004481  0.040420

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.5047031  0.0251870  20.038 8.93e-09 ***
I           -0.1382583  0.1228743  -1.125 0.289615
D            0.1592460  0.1222170   1.303 0.224931
W            0.0131441  0.0372922   0.352 0.732605
P            0.0006909  0.0032864   0.210 0.838170
N           -0.0077645  0.0027556  -2.818 0.020123 *
I:G          0.0076111  0.0014521   5.242 0.000534 ***
I:D          0.0123803  0.0185991   0.666 0.522334
I:P         -0.0003627  0.0047500  -0.076 0.940810
I:N          0.0204953  0.0183417   1.117 0.292757
D:N         -0.0117903  0.0188354  -0.626 0.546882
D:P         -0.0075884  0.0047859  -1.586 0.147295
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02831 on 9 degrees of freedom
```

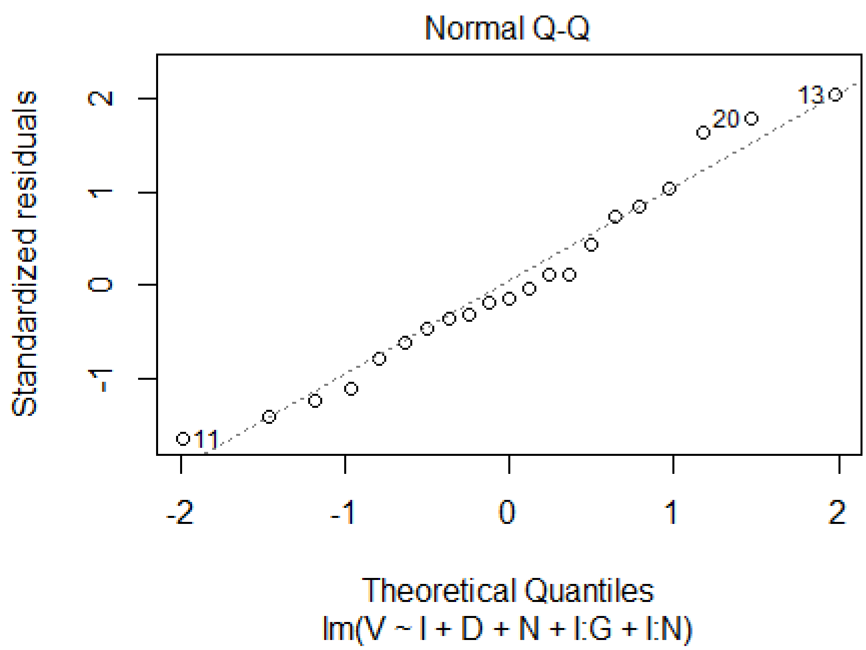
```
Multiple R-squared: 0.936,     Adjusted R-squared: 0.8577
F-statistic: 11.96 on 11 and 9 DF,  p-value: 0.0004421

>
> fit = update(fit1,.~.-I:P)
> summary(fit)

Call:
lm(formula = V ~ I + D + W + P + N + I:G + I:D + I:N + D:N +
    D:P, data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max
-0.041294 -0.010217  0.000393  0.004139  0.040375

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.5042544  0.0232424  21.695 9.67e-10 ***
I           -0.1459078  0.0675086  -2.161 0.055984 .
D            0.1662555  0.0765616   2.172 0.055028 .
W            0.0124784  0.0344093   0.363 0.724416
P            0.0005862  0.0028345   0.207 0.840298
N           -0.0077358  0.0025906  -2.986 0.013665 *
I:G          0.0076343  0.0013474   5.666 0.000208 ***
I:D          0.0131850  0.0145430   0.907 0.385932
I:N          0.0215451  0.0115200   1.870 0.090973 .
D:N         -0.0128411  0.0122039  -1.052 0.317463
D:P         -0.0078045  0.0036629  -2.131 0.058952 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02687 on 10 degrees of freedom
Multiple R-squared: 0.9359,    Adjusted R-squared: 0.8719
F-statistic: 14.61 on 10 and 10 DF,  p-value: 0.0001092

>
> fit = update(fit,.~.-P)
> summary(fit)

Call:
lm(formula = V ~ I + D + W + N + I:G + I:D + I:N + D:N + D:P,
    data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max
-0.041709 -0.010279  0.000817  0.005387  0.039745

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.507141   0.017760  28.555 1.14e-11 ***
I           -0.148974   0.062930  -2.367  0.03733 *
D            0.170314   0.070710   2.409  0.03470 *
W            0.017612   0.022773   0.773  0.45560
N           -0.007898   0.002359  -3.347  0.00651 **
I:G          0.007570   0.001253   6.043 8.39e-05 ***
I:D          0.013501   0.013819   0.977  0.34956
I:N          0.022148   0.010649   2.080  0.06171 .
D:N         -0.013618   0.011095  -1.227  0.24532
D:P         -0.007920   0.003459  -2.290  0.04281 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02567 on 11 degrees of freedom
Multiple R-squared: 0.9357,    Adjusted R-squared: 0.883
```

```
F-statistic: 17.78 on 9 and 11 DF,  p-value: 2.49e-05

>
> fit = update(fit,.~.-W)
> summary(fit)

Call:
lm(formula = V ~ I + D + N + I:G + I:D + I:N + D:N + D:P, data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.042841 -0.011105 -0.001859  0.009731  0.037601 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.507644   0.017448  29.094 1.69e-12 ***
I           -0.135147   0.059318  -2.278  0.04180 *  
D            0.149559   0.064314   2.325  0.03839 *  
N           -0.007455   0.002250  -3.313  0.00619 ** 
I:G          0.007412   0.001215   6.100 5.33e-05 ***
I:D          0.012704   0.013548   0.938  0.36687    
I:N          0.020243   0.010185   1.988  0.07017 .  
D:N         -0.011186   0.010461  -1.069  0.30597    
D:P         -0.006628   0.002978  -2.226  0.04596 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02524 on 12 degrees of freedom
Multiple R-squared: 0.9322,	Adjusted R-squared: 0.887 
F-statistic: 20.62 on 8 and 12 DF,  p-value: 6.835e-06

>
> fit = update(fit,.~.-I:D)
> summary(fit)

Call:
lm(formula = V ~ I + D + N + I:G + I:N + D:N + D:P, data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.036095 -0.010634 -0.007661  0.004968  0.040631 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.517233   0.014072  36.756 1.59e-14 ***
I           -0.117184   0.055878  -2.097  0.05610 .  
D            0.128749   0.060084   2.143  0.05163 .  
N           -0.007678   0.002227  -3.447  0.00433 ** 
I:G          0.007903   0.001092   7.238 6.57e-06 ***
I:N          0.017262   0.009631   1.792  0.09637 .  
D:N         -0.007953   0.009830  -0.809  0.43305    
D:P         -0.006313   0.002945  -2.144  0.05155 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02512 on 13 degrees of freedom
Multiple R-squared: 0.9272,	Adjusted R-squared: 0.888 
F-statistic: 23.65 on 7 and 13 DF,  p-value: 1.991e-06

>
> fit = update(fit,.~.-D:N)
> summary(fit)

Call:
```

```
lm(formula = V ~ I + D + N + I:G + I:N + D:P, data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.036681 -0.012554 -0.003105  0.017478  0.036941 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.519745   0.013555  38.344 1.40e-15 ***
I           -0.073750   0.015306  -4.819 0.000273 ***
D            0.082320   0.017572   4.685 0.000351 ***
N           -0.008047   0.002153  -3.737 0.002209 ** 
I:G          0.008086   0.001055   7.666 2.24e-06 ***
I:N          0.009667   0.002123   4.554 0.000451 ***
D:P         -0.006132   0.002900  -2.114 0.052924 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02481 on 14 degrees of freedom
Multiple R-squared: 0.9235,	Adjusted R-squared: 0.8908 
F-statistic: 28.18 on 6 and 14 DF,  p-value: 4.789e-07

> 
> fit = update(fit,.~.-D:P)
> summary(fit)

Call:
lm(formula = V ~ I + D + N + I:G + I:N, data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.042175 -0.015409 -0.003698  0.017072  0.050259 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.517630   0.015000  34.509 1.05e-15 ***
I           -0.073678   0.016984  -4.338 0.000585 ***
D            0.054566   0.012962   4.210 0.000758 ***
N           -0.007578   0.002377  -3.189 0.006104 ** 
I:G          0.008862   0.001097   8.077 7.65e-07 ***
I:N          0.009562   0.002355   4.060 0.001026 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02753 on 15 degrees of freedom
Multiple R-squared: 0.8991,	Adjusted R-squared: 0.8655 
F-statistic: 26.74 on 5 and 15 DF,  p-value: 5.709e-07
```

Note that this model, compared to the one in the part (a), has a higher adjusted R-squared (0.8655 vs. 0.6698) with the same number of parameters. Further residual analysis should be done to check the fit and validity of the model. The residual plot shows a random behavior and the qq-plot shows the normal assumption is not violated. So the model seems to be adequate in terms of prediction and valid since assumptions seem to hold.





Alternatively, one could use the step function in R as shown next. The model is the same as found by removing insignificant terms step by step, but with the addition of P and D:P. Note that P is insignificant but was included in the model because the interaction D:P is significant (although borderline significant). The model found using the R function has slightly a higher adjusted R-squared.

```
> step(fit1, direction = "backward")
Start:  AIC=-143.5
V ~ I + D + W + G:I + P + N + D * I + P * I + N * I + N * D +
    P * D

       Df Sum of Sq       RSS     AIC
- I:P   1 0.0000047 0.0072187 -145.49
- W     1 0.0000996 0.0073136 -145.21
- D:N   1 0.0003141 0.0075281 -144.61
- I:D   1 0.0003551 0.0075691 -144.49
<none>              0.0072140 -143.50
- I:N   1 0.0010008 0.0082148 -142.77
- D:P   1 0.0020152 0.0092292 -140.33
- I:G   1 0.0220220 0.0292360 -116.11

Step:  AIC=-145.49
V ~ I + D + W + P + N + I:G + I:D + I:N + D:N + D:P

       Df Sum of Sq       RSS     AIC
- W     1 0.0000949 0.0073136 -147.21
- I:D   1 0.0005933 0.0078120 -145.83
<none>              0.0072187 -145.49
- D:N   1 0.0007992 0.0080179 -145.28
- I:N   1 0.0025249 0.0097436 -141.19
- D:P   1 0.0032772 0.0104958 -139.63
- I:G   1 0.0231727 0.0303914 -117.30

Step:  AIC=-147.21
V ~ I + D + P + N + I:G + I:D + I:N + D:N + D:P

       Df Sum of Sq       RSS     AIC
- I:D   1 0.0005433 0.0078569 -147.71
- D:N   1 0.0007082 0.0080218 -147.27
<none>              0.0073136 -147.21
- I:N   1 0.0024888 0.0098024 -143.06
- D:P   1 0.0034790 0.0107926 -141.04
- I:G   1 0.0234180 0.0307316 -119.07

Step:  AIC=-147.71
V ~ I + D + P + N + I:G + I:N + D:N + D:P

       Df Sum of Sq      RSS     AIC
- D:N   1  0.000402 0.008259 -148.66
<none>              0.007857 -147.71
- I:N   1  0.002012 0.009869 -144.92
- D:P   1  0.003235 0.011092 -142.47
- I:G   1  0.032143 0.040000 -115.53

Step:  AIC=-148.66
V ~ I + D + P + N + I:G + I:N + D:P

       Df Sum of Sq      RSS     AIC
<none>              0.008259 -148.66
- D:P   1  0.003093 0.011352 -143.98
- I:N   1  0.012855 0.021114 -130.95
- I:G   1  0.035058 0.043317 -115.86

Call:
lm(formula = V ~ I + D + P + N + I:G + I:N + D:P, data = mydata)

Coefficients:
(Intercept)            I            D            P            N          I:G
I:N          D:P
```

```
   0.512410    -0.076900     0.087976     0.001382    -0.007960     0.008336
0.009705    -0.006755
```

```
> summary(lm(formula = V ~ I + D + P + N + I:G + I:N + D:P, data = mydata))

Call:
lm(formula = V ~ I + D + P + N + I:G + I:N + D:P, data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.035023 -0.014044 -0.007658  0.011799  0.039904 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.512410   0.016885  30.347 1.87e-13 ***
I           -0.076900   0.016106  -4.775 0.000363 ***
D            0.087976   0.019377   4.540 0.000555 ***
P            0.001382   0.001841   0.751 0.466138    
N           -0.007960   0.002191  -3.634 0.003030 ** 
I:G          0.008336   0.001122   7.429 4.98e-06 ***
I:N          0.009705   0.002157   4.498 0.000599 ***
D:P         -0.006755   0.003061  -2.207 0.045937 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0252 on 13 degrees of freedom
Multiple R-squared: 0.9267,    Adjusted R-squared: 0.8873 
F-statistic: 23.48 on 7 and 13 DF,  p-value: 2.077e-06
```

Problem 5.10

Part a.

$\hat{V} = \hat{\beta}_0 + \hat{\beta}_1 I + \hat{\alpha}_1 D_1 + \hat{\alpha}_2 D_2 + \hat{\beta}_3 W + \hat{\beta}_4 (G \cdot I) + \hat{\beta}_5 P + \hat{\beta}_6 N$

$\hat{V} = 0.505 - 0.021I + 0.063D_1 - 0.047D_2 + 0.012W + 0.0094(G \cdot I) - 0.00072P - 0.0051N$

D2=0, D1=0:

- $\hat{V} = (\hat{\beta}_0) + \hat{\beta}_1 I + \hat{\beta}_3 W + \hat{\beta}_4 (G \cdot I) + \hat{\beta}_5 P + \hat{\beta}_6 N$
- $\hat{V} = 0.505 - 0.021I + 0.012W + 0.0094(G \cdot I) - 0.00072P - 0.0051N$

D1=1, D2=0:

- $\hat{V} = (\hat{\beta}_0 + \hat{\alpha}_1) + \hat{\beta}_1 I + \hat{\beta}_3 W + \hat{\beta}_4 (G \cdot I) + \hat{\beta}_5 P + \hat{\beta}_6 N$
- $\hat{V} = 0.569 - 0.021I + 0.012W + 0.0094(G \cdot I) - 0.00072P - 0.0051N$

D2=1, D1=0:

- $\hat{V} = (\hat{\beta}_0 + \hat{\alpha}_2) + \hat{\beta}_1 I + \hat{\beta}_3 W + \hat{\beta}_4 (G \cdot I) + \hat{\beta}_5 P + \hat{\beta}_6 N$
- $\hat{V} = 0.459 - 0.021I + 0.012W + 0.0094(G \cdot I) - 0.00072P - 0.0051N$

The coefficient of D1 says that the democratic share of the two-part presidential vote (V) is on average 0.063 higher if there is a democratic incumbent running for election, and D2 says it is 0.047 lower if a republican incumbent is running for election compare to any other case ("otherwise") after accounting for the other predictors.

```
> fit = lm(V~I + D1 + D2 + W + G:I + P + N, data = mydata)
> summary(fit)

Call:
lm(formula = V ~ I + D1 + D2 + W + G:I + P + N, data = mydata)

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)  0.5054760  0.0364190  13.879 3.58e-09 ***
I           -0.0205982  0.0174858  -1.178 0.259912    
D1           0.0633485  0.0312177   2.029 0.063423 .  
D2          -0.0469714  0.0291912  -1.609 0.131600    
W            0.0123948  0.0436938   0.284 0.781127    
P           -0.0006963  0.0041333  -0.168 0.868808    
N           -0.0051083  0.0039349  -1.298 0.216773    
I:G          0.0094222  0.0019580   4.812 0.000339 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04245 on 13 degrees of freedom
Multiple R-squared: 0.7922,	Adjusted R-squared: 0.6802 
F-statistic: 7.078 on 7 and 13 DF,  p-value: 0.001307
```

Part b.

Substitute $\alpha_1 = -\alpha_2$

$\longrightarrow V = \beta_0 + \beta_1 I + \alpha_1 D_1 + \alpha_2 D_2 + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \varepsilon$

$\longrightarrow \hat{V} = \beta_0 + \beta_1 I + \alpha_1 (D_1 - D_2) + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \varepsilon$

Now compared to

$V = \beta_0 + \beta_1 I + \beta_2 D + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \varepsilon$

Then $D_1 - D_2 = D$ and $\alpha_1 = \beta_2$

- D1=1, D2 = 0 → D=1
- D1=0, D2= 1 → D=-1
- D1=0, D2=0 → D=0

Assuming $\alpha_1 = -\alpha_2$, therefore the second model (with D) can be obtained as a special case of the first model (with D1 and D2).

Part c.

The model was fitted using D1 and D2 as dummy variable, which have coefficients of 0.063 and -0.047. A statistical test should be done to test if the assumption is valid.

Ho: $\alpha_1 + \alpha_2 = 0$, H1: $\alpha_1 + \alpha_2 \neq 0$

Conduct an F-test with the full model and the reduced model by adding the constraint of Ho. Thus:

Full: $V = \beta_0 + \beta_1 I + \alpha_1 D_1 + \alpha_2 D_2 + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \varepsilon$

Reduced: $V = \beta_0 + \beta_1 I + \beta_2 D + \beta_3 W + \beta_4 (G \cdot I) + \beta_5 P + \beta_6 N + \varepsilon$

Because p-value > 0.709, then the null hypothesis cannot be rejected, indicating that there is not enough evidence to say the assumption $\alpha_1 = -\alpha_2$ does not hold. Thus, the data does not seem to violate the assumption $\alpha_1 = -\alpha_2$.

```
> mydata$D1 = (mydata$D==1)*1
> mydata$D2 = (mydata$D==-1)*1
>
> fit = lm(V~ I + D1 + D2 + W + G:I + P + N, data = mydata)
> summary(fit)

Call:
lm(formula = V ~ I + D1 + D2 + W + G:I + P + N, data = mydata)

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)  0.5054760  0.0364190  13.879 3.58e-09 ***
I           -0.0205982  0.0174858  -1.178 0.259912
D1           0.0633485  0.0312177   2.029 0.063423 .
D2          -0.0469714  0.0291912  -1.609 0.131600
W            0.0123948  0.0436938   0.284 0.781127
P           -0.0006963  0.0041333  -0.168 0.868808
N           -0.0051083  0.0039349  -1.298 0.216773
I:G          0.0094222  0.0019580   4.812 0.000339 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04245 on 13 degrees of freedom
Multiple R-squared: 0.7922,    Adjusted R-squared: 0.6802
F-statistic: 7.078 on 7 and 13 DF,  p-value: 0.001307


> linearHypothesis(fit,"D1 + D2 =0 ")
Linear hypothesis test

Hypothesis:
D1  + D2 = 0

Model 1: restricted model
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


> anova(fit0,fit)
Analysis of Variance Table

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
```

R-code

```
#### Homework 3
#### From the text-book: Exercise 3.14, 5.6, 5.9, 5.10
#### From my book: Problem 3.4 and  Problem 3.5

#### (The missing reference (??) in the exercise is to
#### Problem 2.5 from Chapter 2.In problem 3.4(d) of my
#### book, "YrsEm" should be replaced by "PriorYr".
#### Marketing should be Purchase in Table 3.15.

#### 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))

#### 3.4 #################################################

# Import data
filename = "Employee+Salaries.csv"
mydata = read.csv(filename,header = T)
```

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

# Change column name for log salary

colnames(mydata)[colnames(mydata)=="log.Salary."]="log_salary"

### Part a

fit1 = lm(log_salary ~ YrsEm + PriorYr + Education + Super + Female +
              Advertising + Engineering + Sales, data =mydata)
summary(fit1)

# This fitted equation matches the one given in the book

### Part b

fit2 = lm(log_salary ~ YrsEm + PriorYr + Education + Super + Male +
       Advertising + Engineering + Marketing, data =mydata)
summary(fit2)

# The new coefficiens for Male, Advertisting, Engineering and Marketing
# are highlighted after using Female and Sales as base categories
# As expected for engineering: 0.0937783-0.0057292 = 0.0880491
# And male: (-0.0230683)= -0.0230683

### Part c

# For model in part a), the coefficient of engineering is the effect
# of engineering on the salary compared to marketing after accounting
# for other predictors. So the difference in salary between engineering
# and marketing is not significant.
# For model in part b), the coefficient of engineering is the effect
# of engineering on the salary compared to sales after accounting
# for other predictors. So the difference in salary between engineering
# and sales is significant.

# average difference between group coded 1 and group coded 0

mean(mydata$log_salary[mydata$Engineering==1])-mean(mydata$log_salary[mydata$Marketing==1])
mean(mydata$log_salary[mydata$Advertising==1])-mean(mydata$log_salary[mydata$Marketing==1])
mean(mydata$log_salary[mydata$Sales==1])-mean(mydata$log_salary[mydata$Marketing==1])
mean(mydata$log_salary[mydata$Female==1])-mean(mydata$log_salary[mydata$Male==1])

# If the coefficient of a dummy variable is nonsignificant, it tells
```

```
# you that the effect of the dummary variable on the response compared
# to the base category after accounting for other predictors is not significant.
# Then you can combine nonsignificant to the base category.

### Part d

fit3 = lm(log_salary ~ YrsEm  + Education  +
        Advertising + Engineering + Sales, data =mydata)
summary(fit3)

anova(fit1,fit3)
# THe F-test to compare the full model vs. reduced model shows that the null
# hypothesis that PriorYr, Super and Female are zero cannot be rejected, so conclude
# that the coefficients are not significantly different than zero. Thus,
# the reduced model is preferred (the variables in the reduced model adequately
# explain the variation as the variables in the full model)

# Fitting the new model the coefficients of YrsEm, Education and Sales remain
# significant, and the coefficients of Advertising and Engineering
# remain insignificant at 0.05 level.
# The R^2 remains about the same at 0.86. This tells us that Female, Super
# and PriorYr were not contributing much to explaining the variation in salary.
# Thus, the model after dropping the variables seems to be preferred.

# Because Advertisting and Engineering are insignificant compared to
# the base (Marketing), then it might be a good idea to combine
# Advertising, Engineering and Marketing into one "Other" category.
# So you would need only 1 dummy variables representing Sales vs. "Other".
# would represent Sales vs.


#### 3.14 ##############################################

((38460756-22657938)/(91-88))/(22657938/88)
qf(.05,df1=3,df2=88,lower.tail = FALSE)

#### 5.6 ####

## Part a

# r = sqrt(SSR/SST) = sqrt(SSR/(SSR + SSE))
SSR = 4604.7
SSE = 1604.44

r = sqrt(SSR/(SSR+SSE))
r

## Part b
```

```
# The estimated price of an American car with a 100 hp engine is: (answers in thousands)

# Using Model 1
-6.107 + 0.169*100

# Using Model 2
-4.117 + 0.174*100 -3.162*1

# Using Model 3
-10.882 + 0.237*100 +2.076*1 - 0.052*100*1

## Part c

# Model 2

# With respect to "Others", Japan has the the lowest price compared to Germany and USA
# because the coefficient is the most negative.The negative coefficient means
# that the price of Japan is lower than "Others"
# Thus, least expensive car is from Japan.

# Model 3

# Cannot hold horsepower constant because there is an interaction
# between country and horsepower, so the least expensive car depends
# on the horsepower (we cannot estimate the country effect independent of HP)

# For example, if HP = 100

# Others
-10.882 + 0.237*100 + 2.076*0 + 4.755*0 + 11.774*0 -0.052*0*100 - 0.077*0*100  - 0.095*0*100
# USA
-10.882 + 0.237*100 + 2.076*1 + 4.755*0 + 11.774*0 -0.052*1*100  - 0.077*0*100  - 0.095*0*100
# Japan
-10.882 + 0.237*100 + 2.076*0 + 4.755*1 + 11.774*0 -0.052*0*100  - 0.077*1*100  - 0.095*0*100
# Germany
-10.882 + 0.237*100 + 2.076*0 + 4.755*0 + 11.774*1 -0.052*0*100  - 0.077*0*100  - 0.095*1*100
# So least expensive is USA, followed by  Japan, "Others", and Germany

# For example, if HP = 1000

# Others
-10.882 + 0.237*1000 + 2.076*0 + 4.755*0 + 11.774*0 -0.052*0*100 - 0.077*0*1000  - 0.095*0*1000
# USA
-10.882 + 0.237*1000 + 2.076*1 + 4.755*0 + 11.774*0 -0.052*1*100  - 0.077*0*1000  - 0.095*0*1000
# Japan
-10.882 + 0.237*1000 + 2.076*0 + 4.755*1 + 11.774*0 -0.052*0*100  - 0.077*1*1000  - 0.095*0*1000
# Germany
```

```
-10.882 + 0.237*1000 + 2.076*0 + 4.755*0 + 11.774*1 -0.052*0*100  - 0.077*0*1000  - 0.095*1*1000
# So least expensive is Germany, followed by Japan, USA and "Others"


## Part d

# Compare model 3 vs model 2
# Ho: Reduced model (2) is adequate
# H1: Full model (3) is adequate

# Alternatively,
# Ho: HP*USA = HP*Japan = HP*Germany = 0
# H1: At least one of HP*USA, HP*Japan or HP*Germany is different than zero

# F = ((RSS_reduced - RSS_full)/(df_reduced-df_full))/(RSS_full/df_full)

((1390.31-1319.85)/(85-82))/(1319.85/82)
qf(.05,df1=3,df2=82,lower.tail = FALSE)

# Since F_stat = 1.459 < F_crit = 2.716,  the null hypothesis that
# the interaction terms are zero cannot be rejected, so conclude that the cooefficients
# are not significantly different than zero. Thus, the reduced model is preferred
# (the variables in the reduced model adequately explain the variation as the
# variables in the full model). Conclusion: there is not a significant interaction
# between country and HP.

## Part e

# Compare model 2 vs model 1
# Ho: Reduced model (1) is adequate
# H1: Full model (2) is adequate

# Alternatively,
# Ho: USA = Japan = Germany = 0
# H1: At least one of USA, Japan or Germany is different than zero

# F = ((RSS_reduced - RSS_full)/(df_reduced-df_full))/(RSS_full/df_full)

((1604.44-1390.31)/(88-85))/(1390.31/85)
qf(.05,df1=3,df2=85,lower.tail = FALSE)

# Since F_stat = 4.364 > F_crit = 2.712,  reject the null hypothesis that
# the interaction terms are zero, so conclude that the cooefficients
# are  significantly different than zero. Thus, the full model is preferred
# (the variables in the reduced model do not adequately explain the variation as the
# variables in the full model). Conclusion: given HP, Country is a a signficant
# predictor of car price
```

```
## Part f

# Because the p-value > 0.05 for Germany in model 2, the null hypothesis that this
# coefficient is zero cannot be rejected, so conclude that is not significantly
# different than zero. Thus, there is not a signficant difference of prices
# between German and Others car, so it is recommended to add Germans to the
# "Other" category.

## Part g

# Ho: USA = Japan
# H1: USA diff Japan

# One could fit a model by replacing USA = Japan (called reduced model) and compared
# it to the full model (model 2) by using a F-test that will reject or not the
# null hyphothesis.

# Alternatively, you can use a t-test:  t=(USA-Japan)/se(USA-Japan)
# where  se (b1 - b2)= sqrt(var b1 + var b2 - 2cov (b1,b2))
# and df = n-(p+1) = 85


#### 5.9 ##############################################

# Import data
filename = "P160.txt"
mydata = read.table(filename,header = T)

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

## Part a

fit0 = lm(V~I + D + W + G:I + P + N, data = mydata)
summary(fit0)


## Part d

fit1 = lm(V~I + D + W + G:I + P + N + D*I + P*I + N*I + N*D + P*D, data = mydata)
summary(fit1)

fit = update(fit1,.~.-I:P)
summary(fit)
```

```
fit = update(fit,.~.-P)
summary(fit)

fit = update(fit,.~.-W)
summary(fit)

fit = update(fit,.~.-I:D)
summary(fit)

fit = update(fit,.~.-D:N)
summary(fit)

fit = update(fit,.~.-D:P)
summary(fit)


step(fit1, direction = "backward")
summary(lm(formula = V ~ I + D + P + N + I:G + I:N + D:P, data = mydata))


#### 5.10 ###############################################

# Import data
filename = "P160.txt"
mydata = read.table(filename,header = T)

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

mydata$D1 = (mydata$D==1)*1
mydata$D2 = (mydata$D==-1)*1

fit = lm(V~ I + D1 + D2 + W + G:I + P + N, data = mydata)
summary(fit)
```