STAT151A - hw5

Xuanpei Ouyang (Esther)

March 15, 2017

(a) Construct dummy regressors for the factor or factors and fit an additive dummy-regression model. Use an incremental F-test to test the null hypothesis that each factor has no effect. Explain what each regression coeffcient means. If there is a single factor in the model, write out the regression equation for each category of the factor.

```
duncan = read.table("~/Desktop/STAT 151A/STAT-151A/hw/hw5/Duncan.txt")
summary(duncan)
```

```
##    type        income         education         prestige    
##  bc  :21   Min.   : 7.00   Min.   :  7.00   Min.   : 3.00  
##  prof:18   1st Qu.:21.00   1st Qu.: 26.00   1st Qu.:16.00  
##  wc  : 6   Median :42.00   Median : 45.00   Median :41.00  
##            Mean   :41.87   Mean   : 52.56   Mean   :47.69  
##            3rd Qu.:64.00   3rd Qu.: 84.00   3rd Qu.:81.00  
##            Max.   :81.00   Max.   :100.00   Max.   :97.00  
```

```
# Train the full model for duncan dataset
duncan_model = lm(prestige~income+education+type, data = duncan)
summary(duncan_model)
```

```
## 
## Call:
## lm(formula = prestige ~ income + education + type, data = duncan)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -14.890  -5.740  -1.754   5.442  28.972 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -0.18503    3.71377  -0.050  0.96051    
## income        0.59755    0.08936   6.687 5.12e-08 ***
## education     0.34532    0.11361   3.040  0.00416 ** 
## typeprof     16.65751    6.99301   2.382  0.02206 *  
## typewc      -14.66113    6.10877  -2.400  0.02114 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 9.744 on 40 degrees of freedom
## Multiple R-squared:  0.9131, Adjusted R-squared:  0.9044 
## F-statistic:   105 on 4 and 40 DF,  p-value: < 2.2e-16
```

```
# Anova table for duncan full model
anova(duncan_model)
```

```
## Analysis of Variance Table
## 
## Response: prestige
##           Df  Sum Sq Mean Sq F value    Pr(>F)    
## income     1 30664.8 30664.8 322.962 < 2.2e-16 ***
## education  1  5516.1  5516.1  58.096 2.590e-09 ***
## type       2  3708.7  1854.4  19.530 1.208e-06 ***
## Residuals 40  3798.0    94.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# show the dummy variable construction
contrasts(duncan$type)
```

```
##      prof wc
## bc      0  0
## prof    1  0
## wc      0  1
```

```
# Train the null model for duncan dataset
duncan_model_null = lm(prestige~income+education, data=duncan)
summary(duncan_model_null)
```

```
##
## Call:
## lm(formula = prestige ~ income + education, data = duncan)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.538  -6.417   0.655   6.605  34.641
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.06466    4.27194  -1.420    0.163
## income       0.59873    0.11967   5.003 1.05e-05 ***
## education    0.54583    0.09825   5.555 1.73e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.37 on 42 degrees of freedom
## Multiple R-squared:  0.8282, Adjusted R-squared:   0.82
## F-statistic: 101.2 on 2 and 42 DF,  p-value: < 2.2e-16
```

Since for the region response variable, I have 3 categories, I use 2 parameters to construct the dummy variables.

	D1	D2
bc	0	0
prof	1	0
wc	0	1

X_1: Income

X_2: Education

D_1: type prof

D_2: type wc

Model: $\hat{Y_i} = -0.18503 + 0.59755X_1 + 0.34532X_2 + 16.65751D_1 - 14.66113D_2$

Blue Collar: $\hat{Y_i} = -0.18503 + 0.59755 X_1 + 0.34532 X_2$

Professional: $\hat{Y_i} = 16.47248 + 0.59755 X_1 + 0.34532 X_2$

White Collar: $\hat{Y_i} = -14.84616 + 0.59755 X_1 + 0.34532 X_2$

To test the partial effects of type of occupation, conduct F test with H_0 is no effect of occupation type on Prestige keeping income and education is kept constant. Here, null hypothesis is the null model without type effect and alternative hypothesis is the full model with type effect.

H_0: $\gamma_1 = \gamma_2 = 0$

H_a: At least one of γ_1, γ_2 is not zero.

$F_0 = \frac{n-k-1}{q} \frac{{R_1}^2 - {R_0}^2}{1-{R_1}^2}$

where $n - k - 1 = 45 - 4 - 1 = 40$

```
dim(duncan)
```

```
## [1] 45  4
```

```
F_0 = ((45-4-1)/2)*((0.9131 - 0.8282)/(1-0.9131))
F_0
```

```
## [1] 19.5397
```

Therefore, $F_0 = \frac{n-k-1}{q} \frac{{R_1}^2 - {R_0}^2}{1-{R_1}^2}$

$= \frac{45-4-1}{2} \frac{0.9131^2 - 0.8282^2}{1-0.9131^2}$

$= 19.5397008$

If the null model is significantly different from the full mode, then $R_1 >> R_0$ and F_0 will be much higher.

F critical value for $df = 2, 40$ is

Since $F_0 = 19.5397008$ is much greater than $F_c = 3.231727$, I reject the null hypothesis about the null model.

And I can conclude that the type also has effects on prestige and our model is: $\hat{Y_i} = -0.18503 + 0.59755 X_1 + 0.34532 X_2 + 16.65751 D_1 - 14.66113 D_2$

$\alpha = -0.1853$ indicates the value of prestige when the income and education are all zero.

$\beta_1 = 0.59755$ means with every unit increase in income (one thousand dollar) and keeping all the other variables constant, the person will have $\beta_0 = 0.59755$ unit increase in prestige.

$\beta_2 = 0.34532$ means with every unit increase in education (years) and keeping all the other variables constant, the person will have $\beta_0 = 0.34532$ unit increase in prestige.

$\gamma_1 = 16.65751$ indicates, for profession occupation, compared to blue collar, the value of prestige will increase when the income and education are all zeros.

$\gamma_2 = -14.66113$ indicates, for white collar occupation, compared to blue collar, the value of prestige will decrease when the income and education are all zeros.

(b) Now include interaction regressors in the model, allowing each quantitative explanatory variable to interact with each factor. Test each interaction by an incremental F-test. If there is a single factor in the model, write out the regression equation for each category of the factor, and confirm that you get the same results (within rounding error) as you obtain by performing the regression on the quantitative explanatory variables separately for each category of the factor.

```r
duncan_interact_model = lm(prestige~income+education+type+education:type+
                             income:type, data = duncan)
summary(duncan_interact_model)
```

```
##
## Call:
## lm(formula = prestige ~ income + education + type + education:type +
##     income:type, data = duncan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.2629  -5.5337  -0.2431   5.1065  22.5198
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -3.95054    6.79402  -0.581   0.5645
## income               0.78341    0.13074   5.992 7.12e-07 ***
## education            0.31962    0.27979   1.142   0.2608
## typeprof            32.00781   14.10923   2.269   0.0294 *
## typewc              -7.04320   20.63835  -0.341   0.7349
## education:typeprof   0.01859    0.31837   0.058   0.9538
## education:typewc     0.10677    0.36216   0.295   0.7698
## income:typeprof     -0.36914    0.20388  -1.811   0.0786 .
## income:typewc       -0.36031    0.25957  -1.388   0.1736
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.647 on 36 degrees of freedom
## Multiple R-squared:  0.9233, Adjusted R-squared:  0.9063
## F-statistic: 54.17 on 8 and 36 DF,  p-value: < 2.2e-16
```

```r
# Anova table for duncan interaction model
anova(duncan_interact_model)
```

```
## Analysis of Variance Table
##
## Response: prestige
##                Df  Sum Sq Mean Sq  F value    Pr(>F)
## income          1 30664.8 30664.8 329.4692 < 2.2e-16 ***
## education       1  5516.1  5516.1  59.2661 4.071e-09 ***
## type            2  3708.7  1854.4  19.9237 1.495e-06 ***
## education:type  2    75.1    37.6   0.4036    0.6709
## income:type     2   372.2   186.1   1.9994    0.1502
## Residuals      36  3350.6    93.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-incremental Table

Source	Null Hypothesis	p-value
Income	$\beta_1 = 0 \Vert \delta_{11} = \delta_{12} = 0$	< 2.2e-16
Education	$\beta_2 = 0 \Vert \delta_{21} = \delta_{22} = 0$	4.071e-09
Type	$\gamma_1 = \gamma_2 = 0 \Vert \delta_{11} = \delta_{12} = \delta_{21} = \delta_{22} = 0$	1.495e-06
Income*Type	$\delta_{11} = \delta_{12} = 0$	0.6709

Source	Null Hypothesis	p-value
Education*Type	$\delta_{21} = \delta_{22} = 0$	0.1502

Since the p-value for interaction terms are all not small enough, the interaction between education and type and the interaction between income and type are both insignificant.

And the final model is the same as the model I obtained from part (a).

```r
prof_duncan = Duncan %>% filter(type == "prof")
wc_duncan = Duncan %>% filter(type == "wc")
bc_duncan = Duncan %>% filter(type == "bc")

prof_model = lm(prestige~income+education, prof_duncan)
summary(prof_model)
```

```
##
## Call:
## lm(formula = prestige ~ income + education, data = prof_duncan)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.338  -5.216  -0.416   5.920  21.833
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.0573    12.9430   2.168   0.0467 *
## income        0.4143     0.1637   2.530   0.0231 *
## education     0.3382     0.1590   2.127   0.0504 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.1 on 15 degrees of freedom
## Multiple R-squared:  0.5478, Adjusted R-squared:  0.4875
## F-statistic: 9.086 on 2 and 15 DF,  p-value: 0.002599
```

```r
wc_model = lm(prestige~income+education, wc_duncan)
summary(wc_model)
```

```
##
## Call:
## lm(formula = prestige ~ income + education, data = wc_duncan)
##
## Residuals:
##      1      2      3      4      5      6
## -2.450  2.341  7.023  1.233 -1.551 -6.596
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.9937    12.1337  -0.906   0.4317
## income        0.4231     0.1396   3.030   0.0563 .
## education     0.4264     0.1432   2.978   0.0587 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.007 on 3 degrees of freedom
## Multiple R-squared:  0.8443, Adjusted R-squared:  0.7405
## F-statistic: 8.136 on 2 and 3 DF,  p-value: 0.06142
```

```
bc_model = lm(prestige~income+education, bc_duncan)
summary(bc_model)
```

```
##
## Call:
## lm(formula = prestige ~ income + education, data = bc_duncan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.2629  -6.3580   0.0742   5.1065  22.5198
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.9505     6.8698  -0.575    0.572
## income        0.7834     0.1322   5.926 1.31e-05 ***
## education     0.3196     0.2829   1.130    0.273
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.755 on 18 degrees of freedom
## Multiple R-squared:  0.7373, Adjusted R-squared:  0.7081
## F-statistic: 25.26 on 2 and 18 DF,  p-value: 5.964e-06
```

prof model: $\hat{Y_i} = 28.0573 + 0.4143X_1 + 0.3382X_2$

wc_model: $\hat{Y_i} = -10.9937 + 0.4231X_1 + 0.4264X_2$

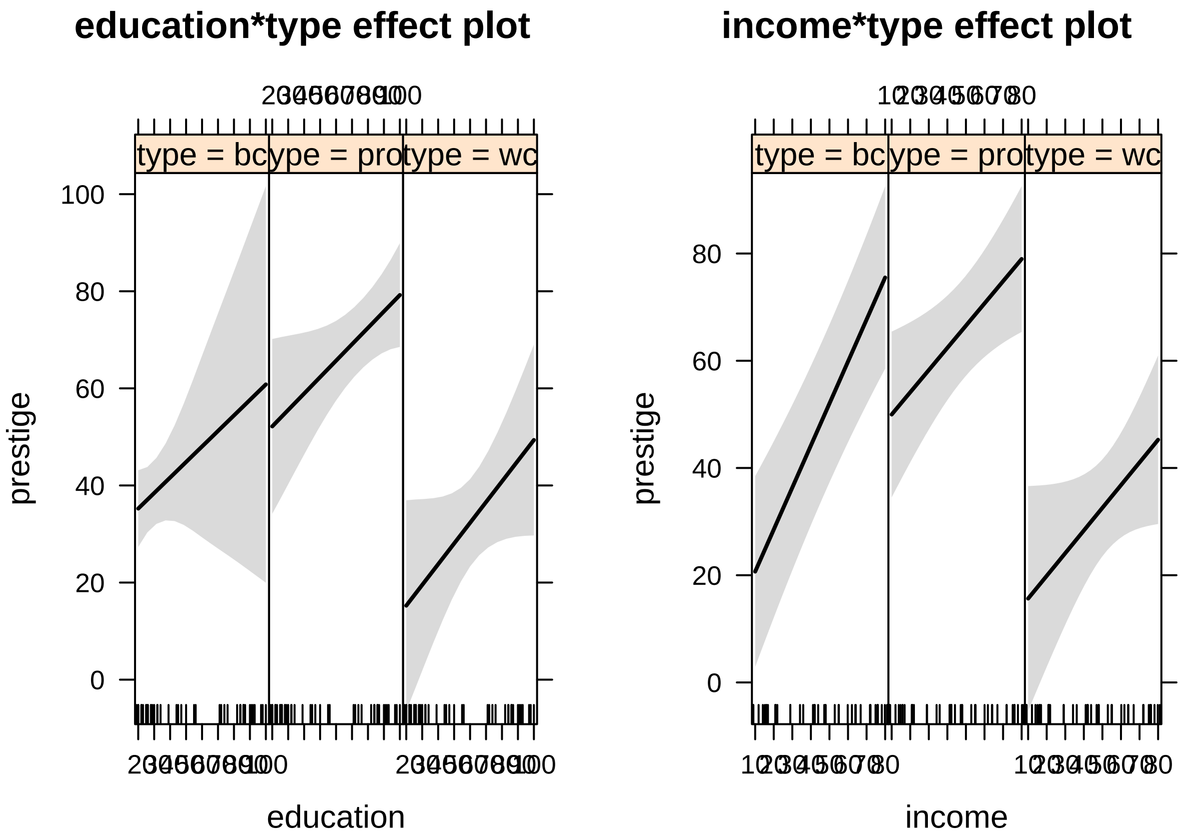
bc_model: $\hat{Y_i} = -3.9505 + 0.7834X_1 + 0.3196X_2$

By comparing the Anova table (F-incremntal table) and the result obtained by performing the regression on the quantitative explanatory variables seperately for each category of the factor (ignore rounding error), I get the same results that the interaction between education and type and the interaction between income and type are both insignificant.

(c) Fit a final model to the data that includes the statistically significant effects. If there are interactions in this model, construct an effect display for each high-order term in the model.

From part b, I can conclude that there is no interaction.

```
plot(allEffects(duncan_interact_model))
```



```
# there is no interaction but here I still plot the effect display for
# high-order terms in the model just for understanding.
summary(duncan_interact_model)
```

```
##
## Call:
## lm(formula = prestige ~ income + education + type + education:type +
##     income:type, data = duncan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.2629  -5.5337  -0.2431   5.1065  22.5198
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -3.95054    6.79402  -0.581   0.5645
## income               0.78341    0.13074   5.992 7.12e-07 ***
## education            0.31962    0.27979   1.142   0.2608
## typeprof            32.00781   14.10923   2.269   0.0294 *
## typewc              -7.04320   20.63835  -0.341   0.7349
## education:typeprof   0.01859    0.31837   0.058   0.9538
## education:typewc     0.10677    0.36216   0.295   0.7698
## income:typeprof     -0.36914    0.20388  -1.811   0.0786 .
## income:typewc       -0.36031    0.25957  -1.388   0.1736
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.647 on 36 degrees of freedom
## Multiple R-squared:  0.9233, Adjusted R-squared:  0.9063
## F-statistic: 54.17 on 8 and 36 DF,  p-value: < 2.2e-16
```

```r
vif(duncan_interact_model)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## income            4.824438  1        2.196460
## education        32.778424  1        5.725244
## type            480.931237  2        4.682963
## education:type 1233.746316  2        5.926612
## income:type     176.244403  2        3.643584
```

In our dataset, I can see there is a discrepency between the model I obtained from full model with interaction and the model I obtained from full model. By checking variance inflation factors for our linear models with interact, I can see that the discrepency might because there is high multicollinearity between education and education:type interaction term.

Therefore, I use stepwise regression with BIC.

```r
n = dim(duncan)[1]
BIC_model = step(duncan_interact_model, Prestige~1 , direction = "backward",
                 k = log(n))
```

```
## Start:  AIC=228.22
## prestige ~ income + education + type + education:type + income:type
##
##                  Df Sum of Sq    RSS    AIC
## - education:type  2     11.56 3362.2 220.76
## - income:type     2    372.17 3722.8 225.35
## <none>                        3350.6 228.22
##
## Step:  AIC=220.76
## prestige ~ income + education + type + income:type
##
##               Df Sum of Sq    RSS    AIC
## - income:type  2    435.75 3798.0 218.63
## <none>                     3362.2 220.76
## - education    1    891.30 4253.5 227.54
##
## Step:  AIC=218.63
## prestige ~ income + education + type
##
##             Df Sum of Sq    RSS    AIC
## <none>                    3798.0 218.63
## - education  1     877.2 4675.2 224.18
## - type       2    3708.7 7506.7 241.68
## - income     1    4246.1 8044.1 248.60
```

The final model I obtained from stepwise regression with BIC is

X_1: income

X_2: education

D_1: type prof

D_2: type wc

Final Model: $\hat{Y}_i = -0.18503 + 0.59755X_1 + 0.34532X_2 + 16.65751D_1 - 14.66113D_2$