Sylvia (Shuyuan) Ding
Shuyuand

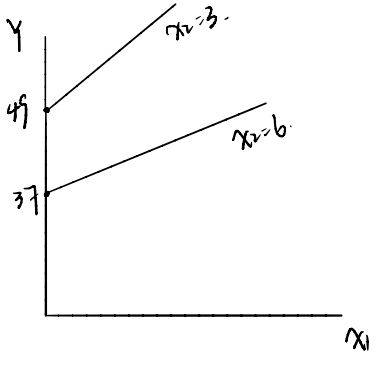
36-401 HW #10.

1. $X_2 = 3$: $E[Y|X_1, X_2 = 3] = 25 + 3 \cdot X_1 + 4 \cdot 3 + 1.5 \cdot X_1 \cdot 3$

$= 25 + 3X_1 + 12 + 4.5X_1 = 37 + 7.5X_1$

$X_2 = 6$: $E[Y|X_1, X_2 = 6] = 25 + 3 \cdot X_1 + 4 \cdot 6 + 1.5 \cdot X_1 \cdot 6$

$= 25 + 3X_1 + 24 + 9X_1 = 49 + 12X_1$



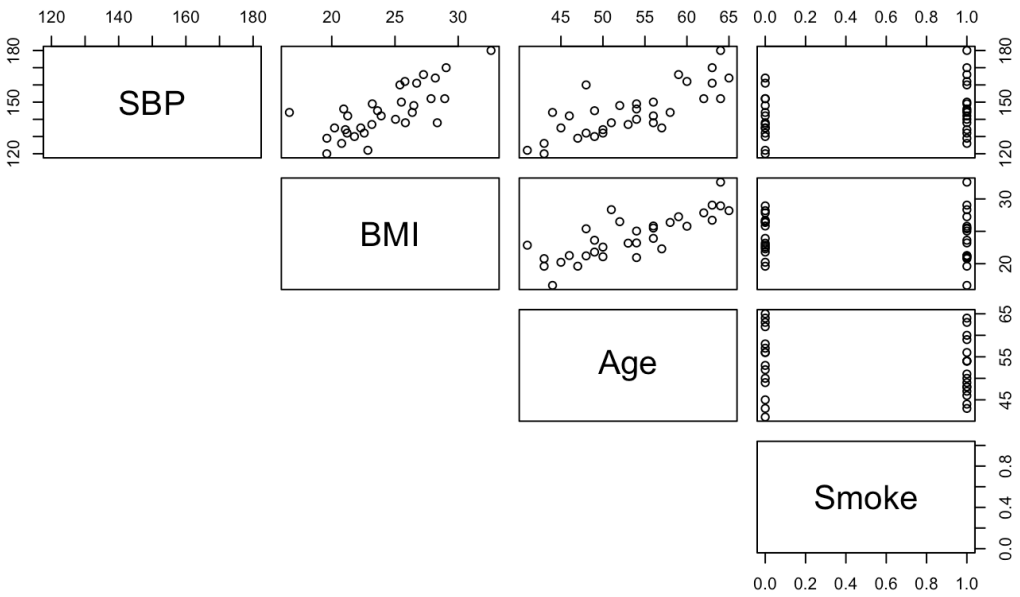
The interaction effect of X1 and X2 here is apparent as the slope of the function changes as X2 changes from 3 to 6. It changes from 7.5 to 12, which means a unit of increase in X1 now has a larger effect on Y after X2 is increase from 3 to 6. The nature of the interaction effect refers to the idea that the effect of one variable depends on the effect of another variable.

36401 HW10

Sylvia (Shuyuan) Ding
Shuyuand

Problem #2

(a)



It seems like SBP has a positive linear relationship with both BMI and Age, and BMI and Age themselves show a positive linear relationship as well. Moreover, from this scatterplot matrix, we think Smoke is a categorical variable as the dataset is described, with 0 represents no smoke and 1 represents smoke. Also for right now it seems like there are relatively equal amount of 0 and 1 under variable Smoke.

(b)

round(cor(sbp), 3)

```
##            ID   SBP    BMI    Age  Smoke
## ID      1.000 0.099  0.165  0.327 -0.112
## SBP     0.099 1.000  0.742  0.775  0.247
## BMI     0.165 0.742  1.000  0.803 -0.071
## Age     0.327 0.775  0.803  1.000 -0.139
## Smoke -0.112 0.247 -0.071 -0.139  1.000
```

round(cor(sbp)[-1,-1], 3)

```
##          SBP    BMI    Age  Smoke
## SBP    1.000  0.742  0.775  0.247
## BMI    0.742  1.000  0.803 -0.071
## Age    0.775  0.803  1.000 -0.139
## Smoke 0.247 -0.071 -0.139  1.000
```

The difference between the two is that the second one eliminates the variable ID, which in this dataset is just the index of the observation, and it should not be taken into consideration of the model. And as we can see, the correlation coefficients make more sense after we remove ID, as now SBP shows 1 to itself, and BMI's coefficient with SBP significantly increased comparing to when we had ID considered as a variable.

(c)

```
## Call:
## lm(formula = SBP ~ Age + BMI + Smoke + Age:Smoke + BMI:Smoke,
##     data = sbp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.046  -5.145  -1.385   5.473  17.185
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  48.6127    17.0154   2.857   0.0083 **
## Age           1.0289     0.5018   2.051   0.0505 .
## BMI           1.4866     1.2987   1.145   0.2628
## Smoke        -0.5374    23.2300  -0.023   0.9817
## Age:Smoke     0.4373     0.7128   0.614   0.5449
## BMI:Smoke    -0.5273     1.5317  -0.344   0.7334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.622 on 26 degrees of freedom
## Multiple R-squared:  0.765,  Adjusted R-squared:  0.7198
## F-statistic: 16.92 on 5 and 26 DF,  p-value: 1.868e-07
```

(i) It seems like all these explanatory variables do not show a statistically significant relationship with the response variable: SBP at alpha level of 0.05, because it seems like all p-values are greater than 0.05.

(ii) The null hypothesis for the global F-test is beta1 = beta2 = beta3 = beta4 = beta5 = 0
Alternative hypothesis is at least one of the betas is not 0.

The F* value for this test is 16.92 and the result is statistically significant because the p-value is only 1.868e-07, which is approximately 0. The R(adj)^2 value is 0.7198. Based on the F-test, we can conclude that we have sufficient evidence to reject null hypothesis and conclude that at least one of the betas in the model is not 0, which means there is a regression relation here. And after adjusting for the number of explanatory variables in the model, the proportion of variability of SBP can be explained by the set of explanatory variables in the model is 71.98%. However, since none of the individual explanatory variable in the model has a statistically significant relationship with the response variable, we suspect there is some correlation between two explanatory variables.

(d) Model with no interaction (lm(SBP~Agew+BMI+Smoke))

```
##
## Call:
## lm(formula = SBP ~ Age + BMI + Smoke, data = sbp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5420  -6.1812  -0.7282   5.2908  15.7050
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45.1032    10.7649   4.190 0.000252 ***
## Age           1.2127     0.3238   3.745 0.000829 ***
## BMI           1.2223     0.6399   1.910 0.066427 .
## Smoke         9.9456     2.6561   3.744 0.000830 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.407 on 28 degrees of freedom
## Multiple R-squared:  0.7609, Adjusted R-squared:  0.7353
## F-statistic: 29.71 on 3 and 28 DF,  p-value: 7.602e-09
```

i) Now is seems like Age and Smoke has statistically significant relationship with SBP because they have p values less than 0.05.

ii) Null hypothesis in this is beta1 = beta2 = beta3 = 0, alternative hypothesis is at least one of the betas is not 0. The F* value for the test is 29.71 and it is statistically significant because we can see that the p-value is 7.602e-09, which is really close to 0. Also the R^2(adjusted) is 0.7353, which means that after adjusting for the number of explanatory variables in the model, the proportion of variability of SBP can be explained by the set of explanatory variables in the model with Age, BMI and Smoke as explanatory variables is 73.53%. Based on this F-test, we can conclude that we have sufficient evidence to reject null hypothesis and conclude that at least one of the betas is not 0, which means some regression relation is present here. Moreover, since the coefficient changes drastically for Smoke, we suspect there is multicollinearity between smoke and another explanatory variable.

(e) The null hypothesis of interest is beta4 = beta 5 = 0 (reduced model without interaction)

```
## Analysis of Variance Table
##
## Model 1: SBP ~ Age + BMI + Smoke + Age:Smoke + BMI:Smoke
## Model 2: SBP ~ Age + BMI + Smoke
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 1510.3
## 2     28 1536.1 -2   -25.805 0.2221 0.8023
```

Here we failed to reject the null hypothesis since the p-value is greater than 0.05, and we can therefore conclude that the reduced model without interaction terms fit better than the full model with interaction terms.

(f)

```
## Call:
## lm(formula = SBP ~ Smoke, data = sbp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.824  -9.056  -2.812  11.200  32.176
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  140.800      3.661  38.454   <2e-16 ***
## Smoke          7.024      5.023   1.398    0.172
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.18 on 30 degrees of freedom
## Multiple R-squared:  0.06117,    Adjusted R-squared:  0.02988
## F-statistic: 1.955 on 1 and 30 DF,  p-value: 0.1723
```

We conclude here that smoke by itself does not has statistically significant linear relationship with the response variable SBP, which makes sense, because it is a categorical variable.

(g)

```
## Analysis of Variance Table
##
## Model 1: SBP ~ Age + BMI + Smoke
## Model 2: SBP ~ Smoke
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     28 1536.1
## 2     30 6032.9 -2   -4496.7 40.982 4.816e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis in this case is beat1 = beta2 = 0. From the summary, we can see that the p-value is less than 0.05, which means that we reject the null hypothesis and conclude that M2 is better fitted than M3, which means the model that includes age and BMI fits better than the model that only includes smoke as explanatory variable as it explains more variability in the response variable-SBP.

(h)

```
## Analysis of Variance Table
##
## Model 1: SBP ~ Age + BMI + Smoke
## Model 2: SBP ~ Age + BMI
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     28 1536.1
## 2     29 2305.4 -1   -769.23 14.021 0.00083 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis of interest is beta3 = 0, and here since the p-value is less than 0.05, we successfully reject the null hypothesis and conclude that the model with explanatory variable Smoke is better than the model without the explanatory variable Smoke as the model with Smoke explains more variability in the response variable-SBP.

3. (a). $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

(b). Commercial: $E[Y|x_1, x_2, x_3] = \beta_0 + \beta_1 x_1 + \beta_2 = (\beta_0 + \beta_2) + \beta_1 x_1$

Mutual Saving: $E[Y|x_1, x_2, x_3] = \beta_0 + \beta_1 x_1 + \beta_3 = (\beta_0 + \beta_3) + \beta_1 x_1$

Saving & Loan: $E[Y|x_1, x_2, x_3] = \beta_0 + \beta_1 x_1 - \beta_2 - \beta_3 = (\beta_0 - \beta_2 - \beta_3) + \beta_1 x_1$.

(c) $\beta_0 + \beta_1 x_1 = \dfrac{(\beta_0 + \beta_2) + \beta_1 x_1 + (\beta_0 + \beta_3) + \beta_1 x_1 + (\beta_0 - \beta_2 - \beta_3) + \beta_1 x_1}{3}$ ← (Sum of profits of all 3 types of bank).

(i). β_2.

It is the profit of the Commercial bank comparing to the average of all three types of banks

(ii). β_3

It is the profit of the Mutual Savings bank comparing to the average of all three types of banks

(iii). $-\beta_2 - \beta_3$

It is the profit of the Savings and Loan bank comparing to the average of all three types of banks

Appendix

```
sbp <- read.csv("SBP.csv")
head(sbp)
pairs(~SBP+BMI+Age+Smoke, data = sbp, lower.panel = NULL)

round(cor(sbp), 3)
round(cor(sbp)[-1,-1], 3)

full.mod <- lm(SBP~Age+BMI+Smoke+Age:Smoke+BMI:Smoke, data = sbp)
summary(full.mod)

model.noint <- lm(SBP~Age+BMI+Smoke, data = sbp)
summary(model.noint)

model.noint <- lm(SBP~Age+BMI+Smoke, data = sbp)
summary(model.noint)

anova(full.mod, model.noint)

mode.smoke <- lm(SBP~Smoke, data = sbp)
summary(mode.smoke)

anova(model.noint, mode.smoke)

M4 <- lm(SBP~Age+BMI, data = sbp)
anova(model.noint, M4)
```