S631 Assignment6

Shibi He

1a. One-factor model

```r
robey <- read.table("Robey.txt", header=TRUE)

m1 <- lm(tfr ~ region, data=robey)
summary(m1)$coefficients
```

```
##                    Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)       5.855556  0.2674094 21.897344 5.722778e-26
## regionAsia       -2.315556  0.4474615 -5.174871 4.877102e-06
## regionLatin.Amer -1.805556  0.3898128 -4.631853 2.989023e-05
## regionNear.East  -1.055556  0.5348188 -1.973670 5.444346e-02
```

The expected total fertility rate in Africa is about 5.86 children per woman. The expected total fertility rate in Near.East is 1.06 children per woman lower than Africa. The p-value is 5.444346e-02 (>0.05), suggesting that the difference in total fertility rates between Africa and Near.East is not statistically significant.

```r
# change the order of the factor levels to compare Asia vs Latin.Amer
robey$region2 <- factor(robey$region, levels = c("Asia", "Latin.Amer", "Africa", "Near.East"))

m1.2 <- lm(tfr ~ region2, data=robey)
summary(m1.2)$coefficients
```

```
##                    Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)       3.540000  0.3587673 9.867119 6.217703e-13
## region2Latin.Amer 0.510000  0.4573404 1.115143 2.705813e-01
## region2Africa     2.315556  0.4474615 5.174871 4.877102e-06
## region2Near.East  1.260000  0.5858646 2.150667 3.679087e-02
```

The expected total fertility rate in Asia is about 3.54 children per woman. The expected total fertility rate in Latin America is 0.51 children per woman higher than Asia. The p-value is 2.705813e-01(>0.05), suggesting that this difference is not statistically significant.

1b. Explain coefficients and write out the mean functions

In m1, the intercept $\hat{\beta}_0 = 5.86$, meaning that the expected total fertitity rate in Africa is 5.86 children per woman. The estimated coefficients indicate the differences in expected total fertility rate between other regions and Africa. Specifically, compared to Africa, the expected total fertility rate is 2.32 children per woman lower in Asia, 1.81 lower in Latin America , and 1.06 lower in Near.East.

The mean functions for each region:

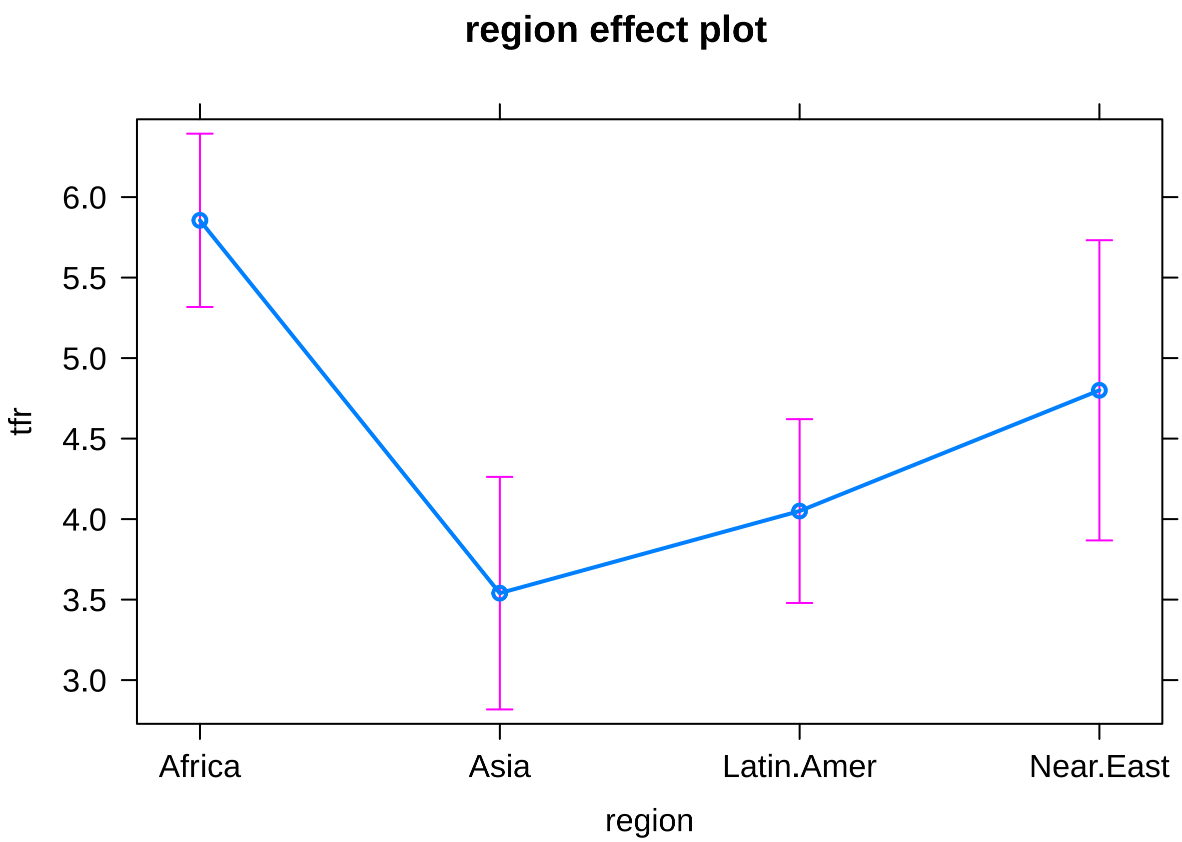
$$\hat{E}(tfr|region = Africa) = \hat{\beta}_0 = 5.86$$
$$\hat{E}(tfr|region = Asia) = \hat{\beta}_0 + \hat{\beta}_{01} = 5.86 - 2.32 = 3.54$$
$$\hat{E}(tfr|region = Latin.Amer) = \hat{\beta}_0 + \hat{\beta}_{02} = 5.86 - 1.81 = 4.05$$
$$\hat{E}(tfr|region = Near.East) = \hat{\beta}_0 + \hat{\beta}_{03} = 5.86 - 1.06 = 4.8$$

1c. Describe the effects plot

```r
library(alr4)
```

```
## Loading required package: car

## Loading required package: carData

## Loading required package: effects

## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

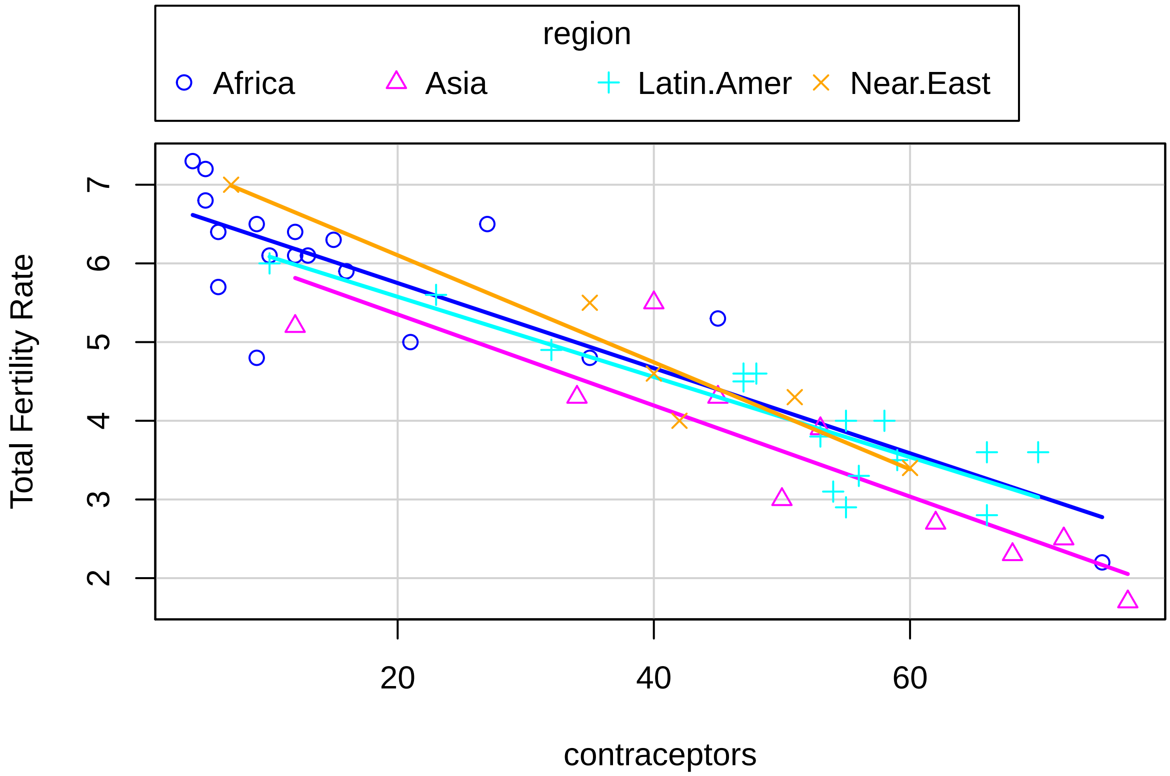
```r
plot(Effect(c("region"), m1))
```



The effect plot shows the fitted values of total fertility rate for different regions and their 95% confidence intervals. Africa has the highest expected total fertility rate of 5.86, and Asia has the lowest expected total fertility rate of 3.54. Latin America and Near.East have moderate total fertility rate.

1d. Scatterplot with both factor and continuous regressor

```
scatterplot(tfr ~ contraceptors|region, data=robey,
            smooth=FALSE, boxplots=FALSE,
            ylab="Total Fertility Rate")
```



The scatter plot shows a negative relationship between contraceptors and total fertility rate. As the percent of contraceptors among married women of childbearing age increases, the total fertility rate decreases. The OLS lines of different regions seem to be parallel to each other, so it's unneccesary to consider different slopes. The intercepts of each OLS line are also very similar, so it's unnecessary to consider different intercepts for different regions.

1e. Model with interaction terms

```
m2 <- lm(tfr~region * contraceptors, data=robey)
summary(m2)$coefficients
```

```
##                                 Estimate  Std. Error    t value
## (Intercept)                  6.832351489 0.194089833 35.2020061
## regionAsia                  -0.322374645 0.563627009 -0.5719645
## regionLatin.Amer            -0.237355799 0.520947681 -0.4556231
## regionNear.East              0.631732515 0.632999103  0.9979991
## contraceptors               -0.054099467 0.007718105 -7.0094238
## regionAsia:contraceptors    -0.003794818 0.012388831 -0.3063096
## regionLatin.Amer:contraceptors  0.003135849 0.012043918  0.2603678
## regionNear.East:contraceptors -0.013919699 0.016140620 -0.8624018
##                                 Pr(>|t|)
## (Intercept)                 8.376212e-33
```

```
## regionAsia                    5.703948e-01
## regionLatin.Amer              6.510080e-01
## regionNear.East               3.239952e-01
## contraceptors                 1.409121e-08
## regionAsia:contraceptors      7.608822e-01
## regionLatin.Amer:contraceptors 7.958523e-01
## regionNear.East:contraceptors  3.933641e-01
```

The results suggest that the expected total fertility rate in Africa is 6.83 children per woman. Compared to Africa, the expected total fertility rates are 0.32, 0.23, 0.63 lower in Asia, Latin America, and Near.East, respectively. In Africa, as the percent of women using contraception increases by 1 percent, the expected total fertility rate decreases by 0.054 children per woman. In Asia, as the percent of women using contraception increases by 1 percent, the expected total fertility rate decreases by $0.054 + 0.004 = 0.058$ children per women. In Latin America, as the percent of women using contraception increases by 1 percent, the expected total fertility rate decreases by $0.054 + 0.003 = 0.057$ children per woman. At last, in near East and North Africa, as the percent of woman using contraception increases by 1 percent, the expected total fertility rate decreases by $0.054 + 0.014 = 0.068$ children per woman. Only the intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ are statistically significant.

1f. ANOVA analysis

```r
Anova(m2)
```

```
## Anova Table (Type II tests)
##
## Response: tfr
##                       Sum Sq Df  F value    Pr(>F)
## region                 1.677  3   1.7018    0.1812
## contraceptors         45.045  1 137.1158 8.226e-15 ***
## region:contraceptors   0.365  3   0.3706    0.7746
## Residuals             13.798 42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

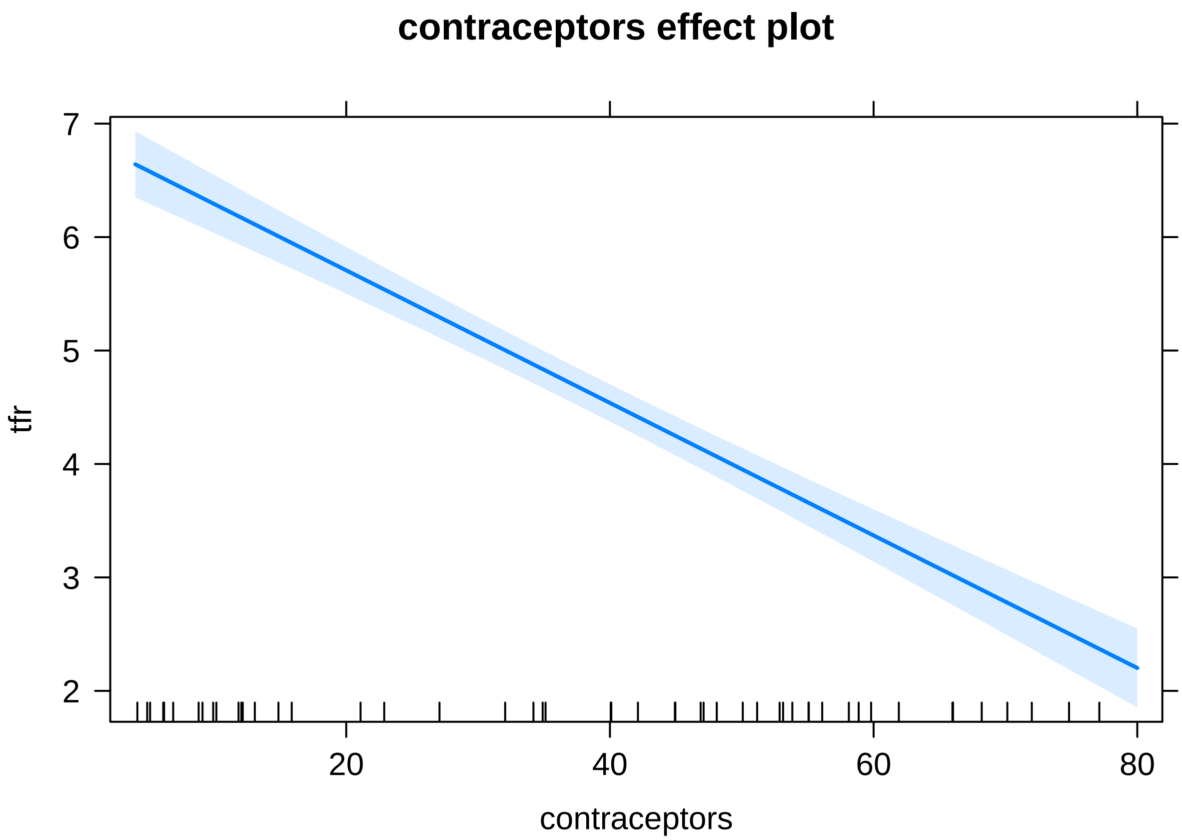
The results of F tests suggest that adding the factor "region" to a model that already contains the regressor "contraceptors" is not statistically significant (p=0.1812). Adding the continuous regressor "contraceptors" to a model that already contains "region" is statistically significant (p=8.226e-15). Adding the interaction terms to a model that already contains the main effects is not statistically significant (p=0.7746). Therefore, I would like to fit a model that only contains the continuous regressor "contraceptors".

1g. Final model: with "contraceptors" only

```r
m3 <- lm(tfr ~  contraceptors, data=robey)
summary(m3)$coefficients
```

```
##                  Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)    6.87508547 0.156860101  43.82941 2.249066e-40
## contraceptors -0.05841574 0.003583935 -16.29933 3.373361e-21
```

```r
plot(Effect("contraceptors", m3))
```



The effect plot shows that as the contraceptors increases, the fitted value of total fertility rate decreases. For example, when the percent of women using contraception increases from 40% to 60%, the fitted total fertility rate decreases approximately from 4.5 to 3.3 children per woman.

1h. Prediction interval

```
mean <- mean(robey$contraceptors)
predict(m3, newdata=data.frame(contraceptors=mean), interval="prediction", level=0.95)
```

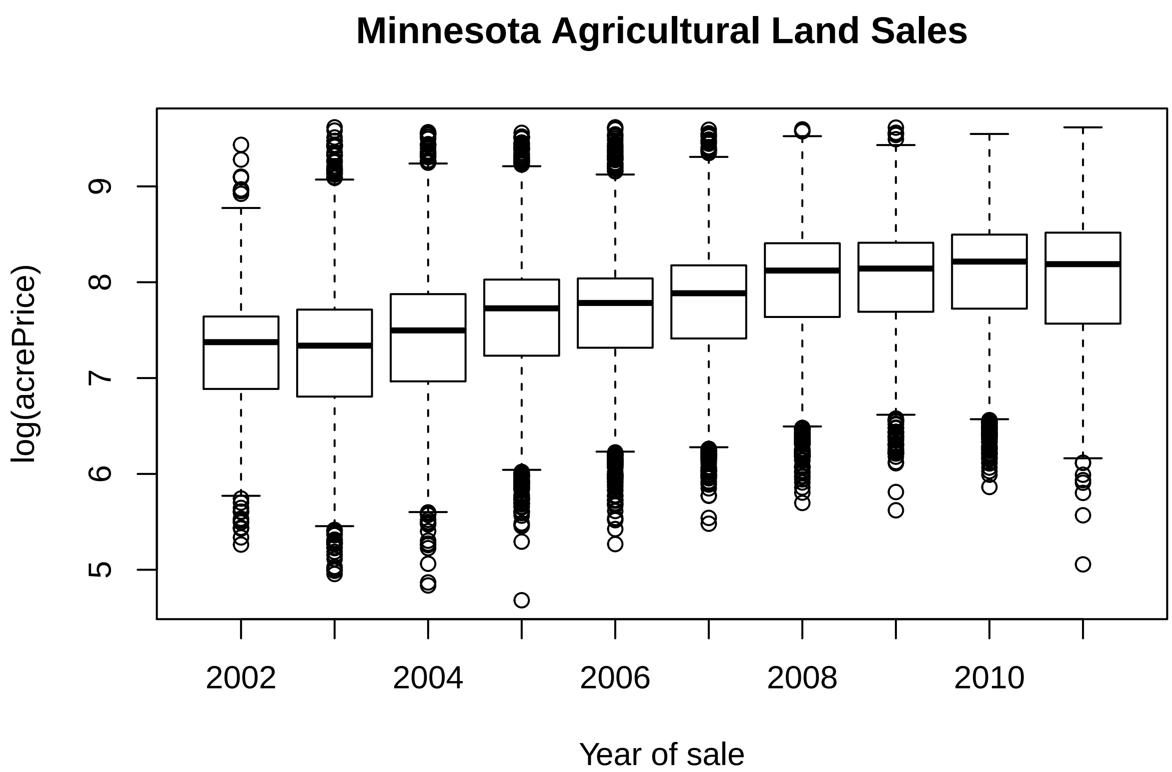
```
##     fit      lwr      upr
## 1 4.688 3.521472 5.854528
```

We are 95% confident that the total fertility rate for a new observation with contraceptors equal to its mean is between 3.521472 and 5.854528 children per woman.

2. ALR 5.4

5.4.1 Boxplot

```
data(MinnLand)
boxplot(log(acrePrice)~year,data=MinnLand, main="Minnesota Agricultural Land Sales",
   xlab="Year of sale", ylab="log(acrePrice)")
```



The boxplots show the log(acrePrice) increases gradually from 2002 to 2011. The pattern in US housing sales prices is not apparently repeated in Minnesota farm sales.

5.4.2 Fit a model with factor "year"

```r
MinnLand$Year <- factor(MinnLand$year)

m1 <- lm(log(acrePrice) ~ Year, MinnLand)
summary(m1)$coefficients
```

```
##                 Estimate Std. Error      t value      Pr(>|t|)
## (Intercept)  7.271748933 0.02847815 255.34487381  0.000000e+00
## Year2003    -0.001550347 0.03206466  -0.04835064  9.614373e-01
## Year2004     0.147944782 0.03154923   4.68933132  2.760361e-06
## Year2005     0.360260534 0.03176125  11.34277022  1.007601e-29
## Year2006     0.393919809 0.03195144  12.32870254  8.664472e-35
## Year2007     0.476822645 0.03186219  14.96515594  2.427602e-50
## Year2008     0.683637098 0.03162006  21.62036379 2.073529e-102
## Year2009     0.714069568 0.03355031  21.28354299  2.431666e-99
## Year2010     0.757331738 0.03260067  23.23056030 1.036926e-117
## Year2011     0.720709867 0.03526532  20.43678853  7.935632e-92
```

The estimated intercept $\hat{\beta}_0 = 7.27$, meaning that the expected log(acrePrice) in 2002 is 7.27. The estimated slopes indicate how much the log(acrePrice) changes in the other years compared to 2002. The t tests can tell us whether these changes in log(acrePrice) are statistically significant.

For example, compared to 2002, the expected log(acrePrice) decreased by 0.0016 in year 2003. This decrease is insignificant, as the pvalue=9.614373e-01 (>0.05). The expected log(acrePrice) increased by 0.1479 in year 2004, compared to 2002. This increase is statistically significant as pvalue=2.760361e-06 (<0.05). The log(acrePrice) increased in each year from 2004 to 2011 and these increases are all statistically significant.

5.4.3 Fit a model with factor "year", omitting intercept

```
m2 <- lm(log(acrePrice) ~ -1+Year, MinnLand)
summary(m2)$coefficients
```

```
##          Estimate Std. Error  t value Pr(>|t|)
## Year2002 7.271749 0.02847815 255.3449        0
## Year2003 7.270199 0.01473559 493.3769        0
## Year2004 7.419694 0.01357751 546.4692        0
## Year2005 7.632009 0.01406314 542.6959        0
## Year2006 7.665669 0.01448757 529.1204        0
## Year2007 7.748572 0.01428966 542.2504        0
## Year2008 7.955386 0.01374129 578.9403        0
## Year2009 7.985819 0.01773750 450.2224        0
## Year2010 8.029081 0.01586816 505.9868        0
## Year2011 7.992459 0.02079995 384.2538        0
```

```
# compute mean of log(acrePrice) for each year and standard errors of sample means
library(dplyr)
MinnLand %>%
  group_by(Year) %>%
  summarise(mean=mean(log(acrePrice)), se=sd(log(acrePrice))/sqrt(length(log(acrePrice))))
```

```
## # A tibble: 10 x 3
##    Year   mean     se
##    <fct> <dbl>  <dbl>
##  1 2002   7.27 0.0267
##  2 2003   7.27 0.0165
##  3 2004   7.42 0.0147
##  4 2005   7.63 0.0147
##  5 2006   7.67 0.0141
##  6 2007   7.75 0.0136
##  7 2008   7.96 0.0126
##  8 2009   7.99 0.0160
##  9 2010   8.03 0.0149
## 10 2011   7.99 0.0215
```

The estimated coefficients are indeed the same as the means of log(acrePrice) for each year. However, the standard errors of the regression coefficients are different from the standard errors of sample means. This is because the standard errors of the regression coefficients are computed as the square root of the diagnol elments of $\hat{\sigma}^2(X^TX)^{-1}$, where $\hat{\sigma}^2$ is the OLS estimator of σ^2, i.e. $\hat{\sigma}^2 = \frac{RSS}{n-10}$. While the standard errors of sample means are computed as $\frac{s_j}{\sqrt{n_j}}$, where s_j is simply the sample standard deviation of log(acrePrice) in the jth year.

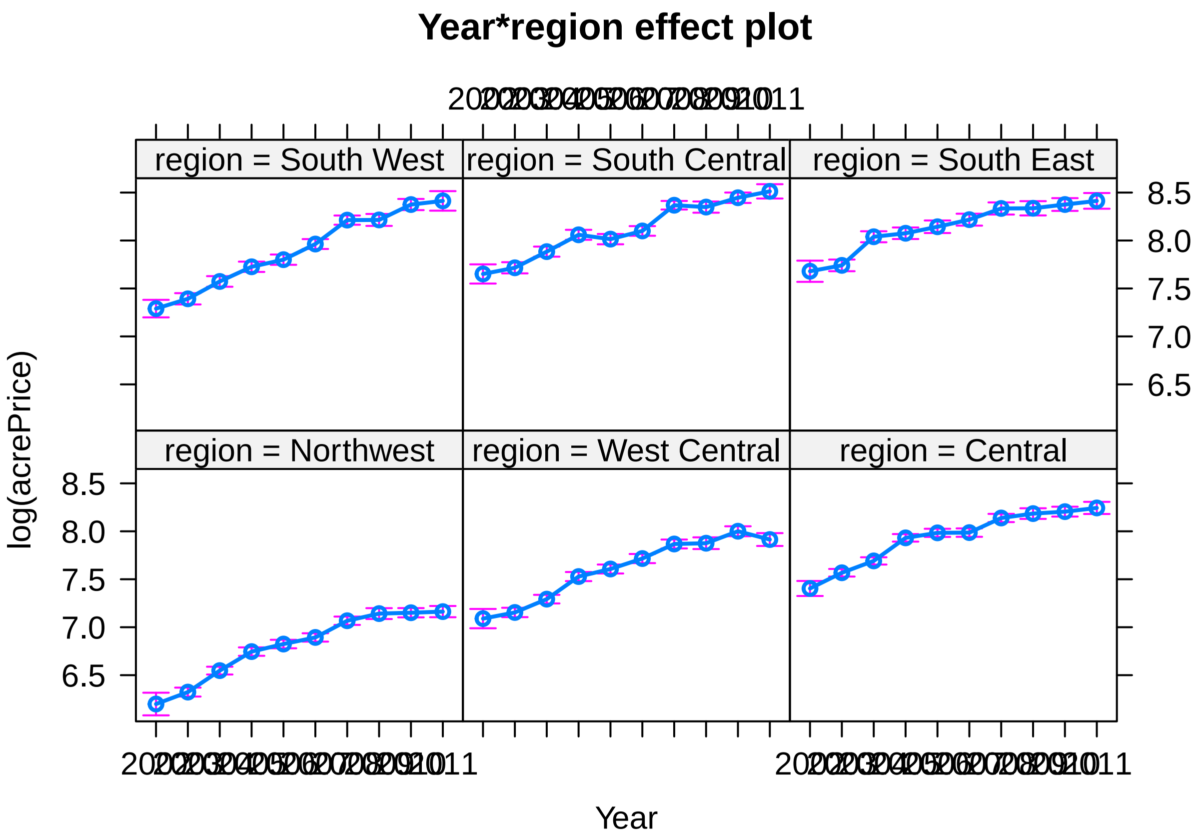
ALR 5.10

5.10.1 Explain the difference

The difference between these two models is that the first model only contains the main effects, while the second model also contains the interactions between "year" and "region".

5.10.2 Fit the models

```r
ma <- lm(log(acrePrice) ~ Year + region, MinnLand)
mb <- lm(log(acrePrice) ~ Year*region, MinnLand)

plot(allEffects(mb))
```



The effects plots show the fitted value of log(acrePrice) for all possible combinations of year and region. Specifically, in a particular year, the fitted log(acrePrice) is different in different regions. Also, the fitted log(acrePrice) in a specific region is different in each year. Generally speaking, the log(acrePrice) consistently increased from 2002 to 2011 in all six regions.

ALR 6.14

6.14.1 Fit a model with "year" being a continuous regressor

```r
mA <- lm(log(acrePrice) ~ year, MinnLand)
summary(mA)$coefficients
```

```
##                 Estimate  Std. Error   t value Pr(>|t|)
## (Intercept) -193.875966 3.983712887 -48.66715        0
## year           0.100464 0.001985464  50.59973        0
```

The model indicates that, on average, the price per acre increases by 10.52% every year (exp(0.1)-1=0.1052).

6.14.2 Fit a model with factor "fyear"

```r
MinnLand$fyear <- factor(MinnLand$year, label=1:10)
mB <- lm(log(acrePrice) ~ 1+fyear, MinnLand)
summary(mB)$coefficients
```

```
##                 Estimate Std. Error      t value      Pr(>|t|)
## (Intercept)  7.271748933 0.02847815 255.34487381  0.000000e+00
## fyear2      -0.001550347 0.03206466  -0.04835064  9.614373e-01
## fyear3       0.147944782 0.03154923   4.68933132  2.760361e-06
## fyear4       0.360260534 0.03176125  11.34277022  1.007601e-29
## fyear5       0.393919809 0.03195144  12.32870254  8.664472e-35
## fyear6       0.476822645 0.03186219  14.96515594  2.427602e-50
## fyear7       0.683637098 0.03162006  21.62036379 2.073529e-102
## fyear8       0.714069568 0.03355031  21.28354299  2.431666e-99
## fyear9       0.757331738 0.03260067  23.23056030 1.036926e-117
## fyear10      0.720709867 0.03526532  20.43678853  7.935632e-92
```

Model B shows that the expected log(acrePrice) in 2002 is 7.27. The estimated coefficients suggest how much the expected log(acrePrice) changes in each year compared to the year 2002. Generally speaking, the increases in the log(acrePrice) gets larger and larger every year, implying the log(acrePrice) increases every year.

6.14.3

In model A, if year=2002, the expected log(acrePrice) = -193.875966 + 0.100464 * 2002 $\approx$ 7.25, which is very close to the intercept in model B. Model A is a special case of model B, assuming there is a constant increase in log(acrePrice) in each year since 2002.

6.14.4 Lack-of-fit test

```r
anova(mA, mB)
```

```
## Analysis of Variance Table
##
## Model 1: log(acrePrice) ~ year
## Model 2: log(acrePrice) ~ 1 + fyear
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1  18698 8666.9
## 2  18690 8579.2  8    87.686 23.878 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result suggest that pvalue<0.05, therefore, we have evidence to reject the null hypethesis that model A is adequate. In other words, model A does not provide an adequate description of the change in log(acrePrice) over time.