Homework6

Rohan Sadale

1 March 2016

```
library(alr4)
```

```
## Warning: package 'alr4' was built under R version 3.2.3
```

```
## Loading required package: car
```

```
## Warning: package 'car' was built under R version 3.2.3
```

```
## Loading required package: effects
```

```
## Warning: package 'effects' was built under R version 3.2.3
```

```
##
## Attaching package: 'effects'
```

```
## The following object is masked from 'package:car':
##
##     Prestige
```

5.2 Equation 5.6 is -

$$E(lifeExpF|log(ppgpd) = x, group = j) = \eta_{0j} + \eta_{1j}x$$

Equation 5.7 is

$$E(lifeExpF|log(ppgpd) = x, group) = \beta_0 + \beta_{02}U_2 + \beta_{03}U_3 + \beta_1 x + \beta_{12}U_2 x + \beta_{13}U_3 x$$

- Now in equation 5.6 and 5.7, if we set group = 1 then $U_2 = 0$ and $U_3 = 0$ and

$$E(lifeExpF|log(ppgpd) = x, group = 1) = \eta_{01} + \eta_{11}x$$
$$E(lifeExpF|log(ppgpd) = x, group = 1) = \beta_0 + \beta_1 x$$

Thus we can see that , $\eta_{01} = \beta_0$ and $\eta_{11} = \beta_1$

- Now in equation 5.6 and 5.7, if we set group = 2 then $U_3 = 0$ and

$$E(lifeExpF|log(ppgpd) = x, group = 2) = \eta_{02} + \eta_{12}x$$
$$E(lifeExpF|log(ppgpd) = x, group = 2) = \beta_0 + \beta_{02}U_2 + \beta_1 x + \beta_{12}U_2 x$$

Thus we can see that, $\eta_{02} = \beta_0 + \beta_{02}U_2$ and $\eta_{12} = \beta_1 + \beta_{12}U_2$

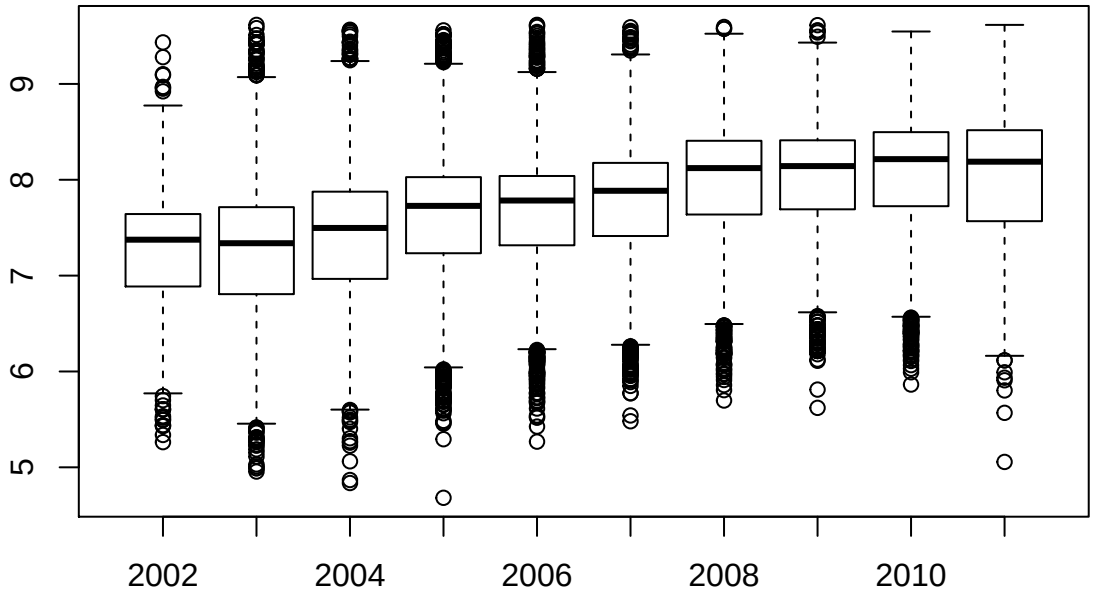
- Similarly, in equation 5.6 and 5.7, if we set group = 3 then $U_2 = 0$ and

$$E(lifeExpF|log(ppgpd) = x, group = 2) = \eta_{03} + \eta_{13}x$$
$$E(lifeExpF|log(ppgpd) = x, group = 2) = \beta_0 + \beta_{03}U_3 + \beta_1 x + \beta_{13}U_3 x$$

Thus we can see that, $\eta_{03} = \beta_0 + \beta_{03}U_3$ and $\eta_{13} = \beta_1 + \beta_{13}U_3$

5.4

```
boxplot(log(acrePrice) ~ year, MinnLand)
```



5.4.1
From the box-plot we can see that -

- The farm sales in each year seems increasing steadily.
- The pattern that housing prices were increasing in US from 2002-2006 and then decreasing from 2007 is not observed in this data. The trend for MinnLand seems increasing from 2002-2007 and from 2008 onwards the rate of increase is slightly less as compared to previous.
- Also we can see that the there is no relationship between interquantile ranges in each successive years.

```r
MinnLand$fyear = as.factor(MinnLand$year)
m1 = lm(log(acrePrice)~fyear, data=MinnLand)
summary(m1)
```

5.4.2

```
##
## Call:
## lm(formula = log(acrePrice) ~ fyear, data = MinnLand)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9499 -0.3785  0.1301  0.4354  2.3456
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.27175    0.02848 255.345  < 2e-16 ***
## fyear2003   -0.00155    0.03207  -0.048    0.961
## fyear2004    0.14794    0.03155   4.689 2.76e-06 ***
## fyear2005    0.36026    0.03176  11.343  < 2e-16 ***
## fyear2006    0.39392    0.03195  12.329  < 2e-16 ***
## fyear2007    0.47682    0.03186  14.965  < 2e-16 ***
## fyear2008    0.68364    0.03162  21.620  < 2e-16 ***
## fyear2009    0.71407    0.03355  21.284  < 2e-16 ***
## fyear2010    0.75733    0.03260  23.231  < 2e-16 ***
## fyear2011    0.72071    0.03526  20.437  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6775 on 18690 degrees of freedom
## Multiple R-squared:  0.1293, Adjusted R-squared:  0.1289
## F-statistic: 308.5 on 9 and 18690 DF,  p-value: < 2.2e-16
```

From the summary above we can say that,

- The intercept represents year 2002.
- All the estimates of slope from the summary (fyear2003 to fyear2011) are actually the difference between estimates of year 2002 and respective year.
- From t-statistics, we can interpret that the change in log(acrePrice) from 2002 to 2003 is not significant while the change in log(acrePrice) from 2002 to other years seem significant.

```r
MinnLand$fyear = as.factor(MinnLand$year)
m1 = lm(log(acrePrice)~fyear-1, data=MinnLand)

with(MinnLand, tapply(log(acrePrice), fyear, mean))
```

5.4.3

```
##     2002     2003     2004     2005     2006     2007     2008     2009
## 7.271749 7.270199 7.419694 7.632009 7.665669 7.748572 7.955386 7.985819
##     2010     2011
## 8.029081 7.992459
```

```
coefficients(m1)
```

```
## fyear2002 fyear2003 fyear2004 fyear2005 fyear2006 fyear2007 fyear2008
##  7.271749  7.270199  7.419694  7.632009  7.665669  7.748572  7.955386
## fyear2009 fyear2010 fyear2011
##  7.985819  8.029081  7.992459
```

From above summary we can see that parameter estimates of model without intercept are equal to means of log(acrePrice) for each year.

```
with(MinnLand, tapply(log(acrePrice), fyear, function(x) sd(x)/sqrt(length(x))))
```

```
##       2002       2003       2004       2005       2006       2007
## 0.02669100 0.01649410 0.01474056 0.01470867 0.01414931 0.01355063
##       2008       2009       2010       2011
## 0.01259338 0.01600373 0.01489691 0.02152664
```

```
summary(m1)
```

```
##
## Call:
## lm(formula = log(acrePrice) ~ fyear - 1, data = MinnLand)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9499 -0.3785  0.1301  0.4354  2.3456
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## fyear2002  7.27175    0.02848   255.3   <2e-16 ***
## fyear2003  7.27020    0.01474   493.4   <2e-16 ***
## fyear2004  7.41969    0.01358   546.5   <2e-16 ***
## fyear2005  7.63201    0.01406   542.7   <2e-16 ***
## fyear2006  7.66567    0.01449   529.1   <2e-16 ***
## fyear2007  7.74857    0.01429   542.2   <2e-16 ***
## fyear2008  7.95539    0.01374   578.9   <2e-16 ***
## fyear2009  7.98582    0.01774   450.2   <2e-16 ***
## fyear2010  8.02908    0.01587   506.0   <2e-16 ***
## fyear2011  7.99246    0.02080   384.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6775 on 18690 degrees of freedom
## Multiple R-squared:  0.9923, Adjusted R-squared:  0.9923
## F-statistic: 2.417e+05 on 10 and 18690 DF,  p-value: < 2.2e-16
```

The difference is because while fitting regression model we consider constant variance for all y values where as with the above tapply function we would consider variance across years

5.8

```
m1 <- lm(Y ~ X1 + X2 + I(X1^2) + I(X2^2) + X1:X2, cakes)
summary(m1)
```

5.8.1

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + I(X1^2) + I(X2^2) + X1:X2, data = cakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4912 -0.3080  0.0200  0.2658  0.5454
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.204e+03  2.416e+02  -9.125 1.67e-05 ***
## X1           2.592e+01  4.659e+00   5.563 0.000533 ***
## X2           9.918e+00  1.167e+00   8.502 2.81e-05 ***
## I(X1^2)     -1.569e-01  3.945e-02  -3.977 0.004079 **
## I(X2^2)     -1.195e-02  1.578e-03  -7.574 6.46e-05 ***
## X1:X2       -4.163e-02  1.072e-02  -3.883 0.004654 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4288 on 8 degrees of freedom
## Multiple R-squared:  0.9487, Adjusted R-squared:  0.9167
## F-statistic:  29.6 on 5 and 8 DF,  p-value: 5.864e-05
```

From the summary we can see that significance levels for quadratic and interaction terms are less than 0.005.

```
m1 <- lm(Y ~ as.factor(block) + X1 + X2 + I(X1^2) + I(X2^2) + X1:X2, cakes)
summary(m1)
```

5.8.2

```
##
## Call:
## lm(formula = Y ~ as.factor(block) + X1 + X2 + I(X1^2) + I(X2^2) +
##     X1:X2, data = cakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4525 -0.3046  0.0200  0.2924  0.4883
##
## Coefficients:
```

```
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -2.205e+03  2.542e+02  -8.672 5.43e-05 ***
## as.factor(block)1  1.143e-01  2.412e-01   0.474 0.650014
## X1                 2.592e+01  4.903e+00   5.287 0.001140 **
## X2                 9.918e+00  1.228e+00   8.080 8.56e-05 ***
## I(X1^2)           -1.569e-01  4.151e-02  -3.779 0.006898 **
## I(X2^2)           -1.195e-02  1.660e-03  -7.197 0.000178 ***
## X1:X2             -4.163e-02  1.128e-02  -3.690 0.007754 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4512 on 7 degrees of freedom
## Multiple R-squared:  0.9503, Adjusted R-squared:  0.9077
## F-statistic: 22.31 on 6 and 7 DF,  p-value: 0.0003129
```

```
m1 <- lm(Y ~ as.factor(block) + X1 + X2 + I(X1^2) + I(X2^2) + X1:X2 - 1, cakes)
summary(m1)
```

```
##
## Call:
## lm(formula = Y ~ as.factor(block) + X1 + X2 + I(X1^2) + I(X2^2) +
##     X1:X2 - 1, data = cakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4525 -0.3046  0.0200  0.2924  0.4883
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## as.factor(block)0 -2.205e+03  2.542e+02  -8.672 5.43e-05 ***
## as.factor(block)1 -2.204e+03  2.542e+02  -8.671 5.43e-05 ***
## X1                 2.592e+01  4.903e+00   5.287 0.001140 **
## X2                 9.918e+00  1.228e+00   8.080 8.56e-05 ***
## I(X1^2)           -1.569e-01  4.151e-02  -3.779 0.006898 **
## I(X2^2)           -1.195e-02  1.660e-03  -7.197 0.000178 ***
## X1:X2             -4.163e-02  1.128e-02  -3.690 0.007754 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4512 on 7 degrees of freedom
## Multiple R-squared:  0.998,  Adjusted R-squared:  0.996
## F-statistic: 504.5 on 7 and 7 DF,  p-value: 6.391e-09
```

From the summary we can see that each block has same effect on the response.

5.10

5.10.1

- In model (a) the change in log(acrePrice) depends only on year and is independent of region. Whereas in model(b), the change is log(acrePrice) depends on both year and region.

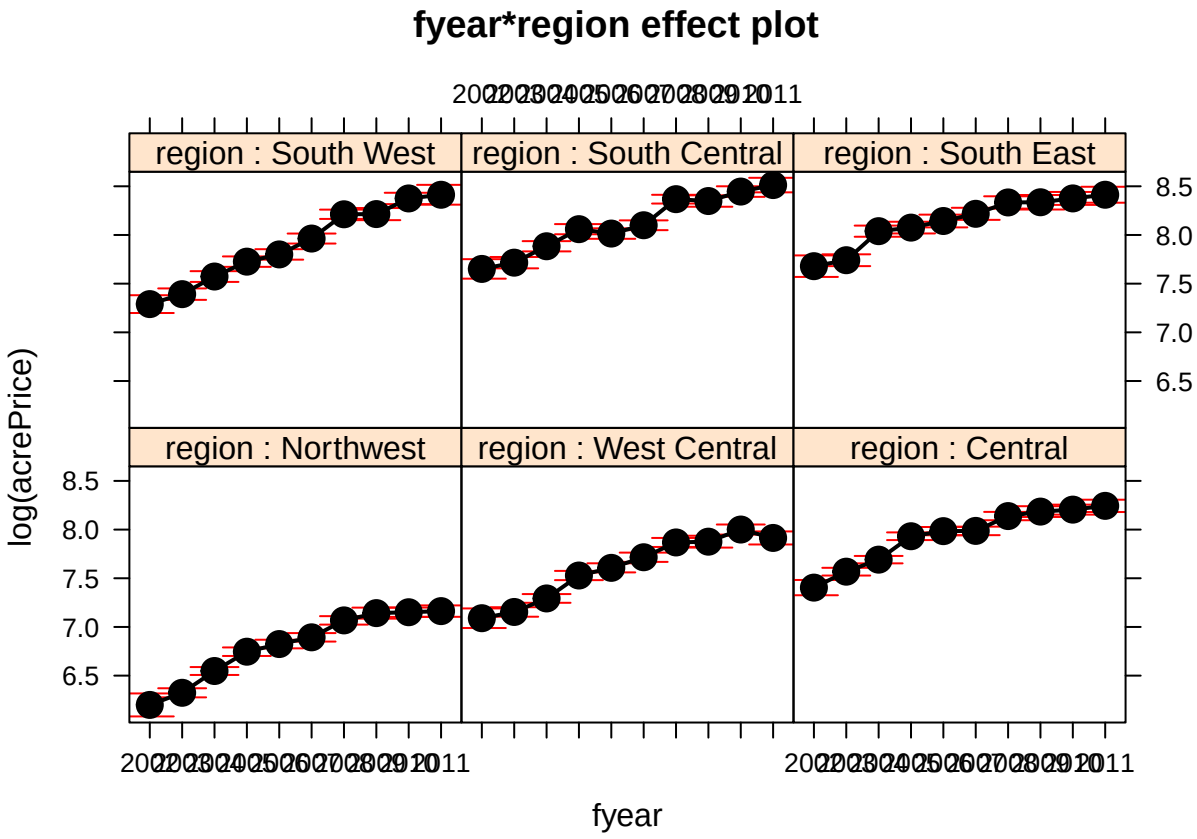
```r
m1 <- lm(log(acrePrice) ~ fyear + region + fyear*region, MinnLand)
#m1 <- lm(log(acrePrice) ~ region*fyear, MinnLand)
summary(m1)
```

5.10.2

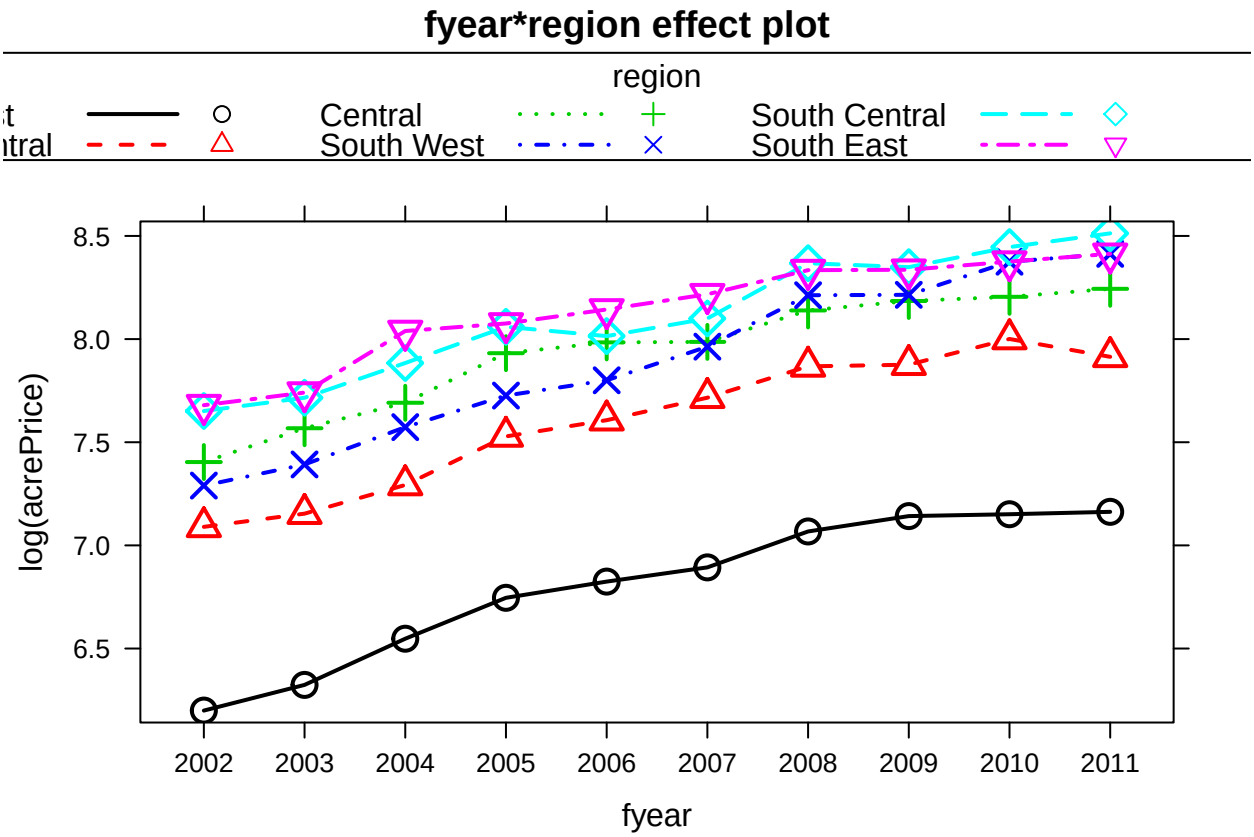
```
##
## Call:
## lm(formula = log(acrePrice) ~ fyear + region + fyear * region,
##     data = MinnLand)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73006 -0.27521  0.01157  0.25561  2.64607
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     6.19911    0.06022 102.939  < 2e-16 ***
## fyear2003                       0.12445    0.06476   1.922  0.05467 .
## fyear2004                       0.34837    0.06367   5.471 4.52e-08 ***
## fyear2005                       0.54665    0.06425   8.508  < 2e-16 ***
## fyear2006                       0.62531    0.06432   9.722  < 2e-16 ***
## fyear2007                       0.69422    0.06419  10.815  < 2e-16 ***
## fyear2008                       0.86828    0.06417  13.532  < 2e-16 ***
## fyear2009                       0.94283    0.06679  14.115  < 2e-16 ***
## fyear2010                       0.95188    0.06505  14.634  < 2e-16 ***
## fyear2011                       0.96351    0.06723  14.333  < 2e-16 ***
## regionWest Central              0.89062    0.07915  11.253  < 2e-16 ***
## regionCentral                   1.20484    0.07230  16.664  < 2e-16 ***
## regionSouth West                1.09079    0.07613  14.328  < 2e-16 ***
## regionSouth Central             1.45223    0.07896  18.392  < 2e-16 ***
## regionSouth East                1.48043    0.08250  17.945  < 2e-16 ***
## fyear2003:regionWest Central   -0.06041    0.08631  -0.700  0.48400
## fyear2004:regionWest Central   -0.14535    0.08493  -1.711  0.08703 .
## fyear2005:regionWest Central   -0.10822    0.08573  -1.262  0.20685
## fyear2006:regionWest Central   -0.10811    0.08568  -1.262  0.20706
## fyear2007:regionWest Central   -0.06810    0.08572  -0.794  0.42693
## fyear2008:regionWest Central   -0.09024    0.08551  -1.055  0.29126
## fyear2009:regionWest Central   -0.15673    0.08981  -1.745  0.08099 .
## fyear2010:regionWest Central   -0.04117    0.08698  -0.473  0.63601
## fyear2011:regionWest Central   -0.13921    0.09123  -1.526  0.12706
## fyear2003:regionCentral         0.03938    0.07877   0.500  0.61715
## fyear2004:regionCentral        -0.06105    0.07766  -0.786  0.43179
## fyear2005:regionCentral        -0.01894    0.07830  -0.242  0.80887
## fyear2006:regionCentral        -0.04535    0.07878  -0.576  0.56486
## fyear2007:regionCentral        -0.11180    0.07888  -1.417  0.15636
## fyear2008:regionCentral        -0.13345    0.07872  -1.695  0.09004 .
## fyear2009:regionCentral        -0.16203    0.08284  -1.956  0.05049 .
## fyear2010:regionCentral        -0.15092    0.08074  -1.869  0.06160 .
## fyear2011:regionCentral        -0.12382    0.08462  -1.463  0.14342
## fyear2003:regionSouth West     -0.02205    0.08522  -0.259  0.79580
## fyear2004:regionSouth West     -0.06516    0.08377  -0.778  0.43671
```

```
## fyear2005:regionSouth West    -0.11040    0.08394  -1.315  0.18842
## fyear2006:regionSouth West    -0.11492    0.08399  -1.368  0.17127
## fyear2007:regionSouth West    -0.02079    0.08352  -0.249  0.80339
## fyear2008:regionSouth West     0.05438    0.08296   0.656  0.51215
## fyear2009:regionSouth West    -0.01820    0.08741  -0.208  0.83508
## fyear2010:regionSouth West     0.13339    0.08534   1.563  0.11805
## fyear2011:regionSouth West     0.15965    0.09689   1.648  0.09942 .
## fyear2003:regionSouth Central -0.06014    0.08766  -0.686  0.49271
## fyear2004:regionSouth Central -0.11564    0.08592  -1.346  0.17838
## fyear2005:regionSouth Central -0.13846    0.08626  -1.605  0.10848
## fyear2006:regionSouth Central -0.26222    0.08656  -3.029  0.00246 **
## fyear2007:regionSouth Central -0.24577    0.08595  -2.859  0.00425 **
## fyear2008:regionSouth Central -0.15184    0.08513  -1.784  0.07451 .
## fyear2009:regionSouth Central -0.24561    0.08923  -2.753  0.00592 **
## fyear2010:regionSouth Central -0.15730    0.08721  -1.804  0.07129 .
## fyear2011:regionSouth Central -0.10230    0.09262  -1.105  0.26938
## fyear2003:regionSouth East    -0.06330    0.09128  -0.693  0.48806
## fyear2004:regionSouth East     0.01118    0.08993   0.124  0.90108
## fyear2005:regionSouth East    -0.15057    0.09083  -1.658  0.09741 .
## fyear2006:regionSouth East    -0.16132    0.09183  -1.757  0.07897 .
## fyear2007:regionSouth East    -0.15648    0.09113  -1.717  0.08598 .
## fyear2008:regionSouth East    -0.21373    0.09131  -2.341  0.01926 *
## fyear2009:regionSouth East    -0.28636    0.09530  -3.005  0.00266 **
## fyear2010:regionSouth East    -0.25598    0.09252  -2.767  0.00567 **
## fyear2011:regionSouth East    -0.22985    0.09718  -2.365  0.01803 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4818 on 18640 degrees of freedom
## Multiple R-squared:  0.5609, Adjusted R-squared:  0.5596
## F-statistic: 403.6 on 59 and 18640 DF,  p-value: < 2.2e-16
```

```
plot(allEffects(m1))
```




```
plot(allEffects(m1), multiline = TRUE)
```



From the effects plot we can see that prices in Northwest regions always lie lower that other region. The prices in all regions are increasing. Also we can see the interaction is present.

5.11

```
m1 <- lm(log(acrePrice) ~ fyear + region + fyear*region + financing, MinnLand)
confint(m1, level = 0.95)["financingseller_financed",]
```

5.11.1

```
##       2.5 %      97.5 %
## -0.11465872 -0.07087851
```

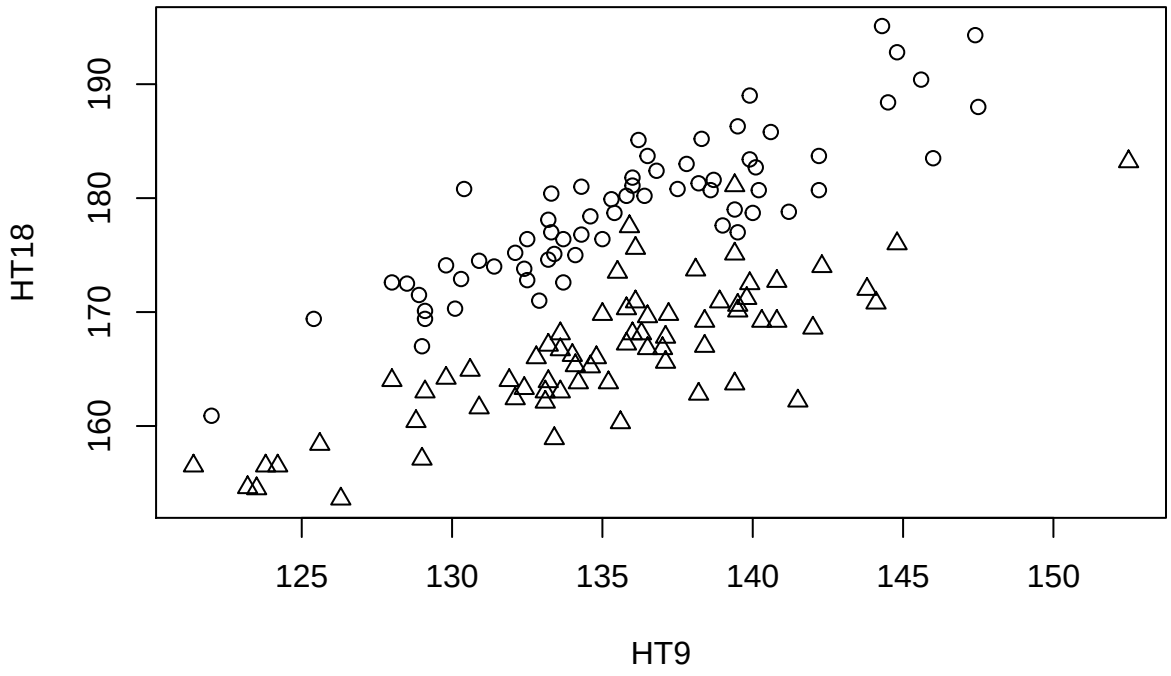
The negative sign suggest that seller_financed sales are lower than title_transfer sales. The seller financed sales estimated between 11% lower and 7% lower.

5.11.2

- This is observational data and we can infer only correlation from it. But the first statement implies causation which is not supported by observational data.
- The second statement seems true for the data but doesn't support strongly for all cases.

5.14

```
plot(HT18 ~ HT9, pch=ifelse(Sex=="0",1,2), BGSall)
```



5.14.1
The appropriate mean functions for boys and girls if taken seperately seem parallel to each other.

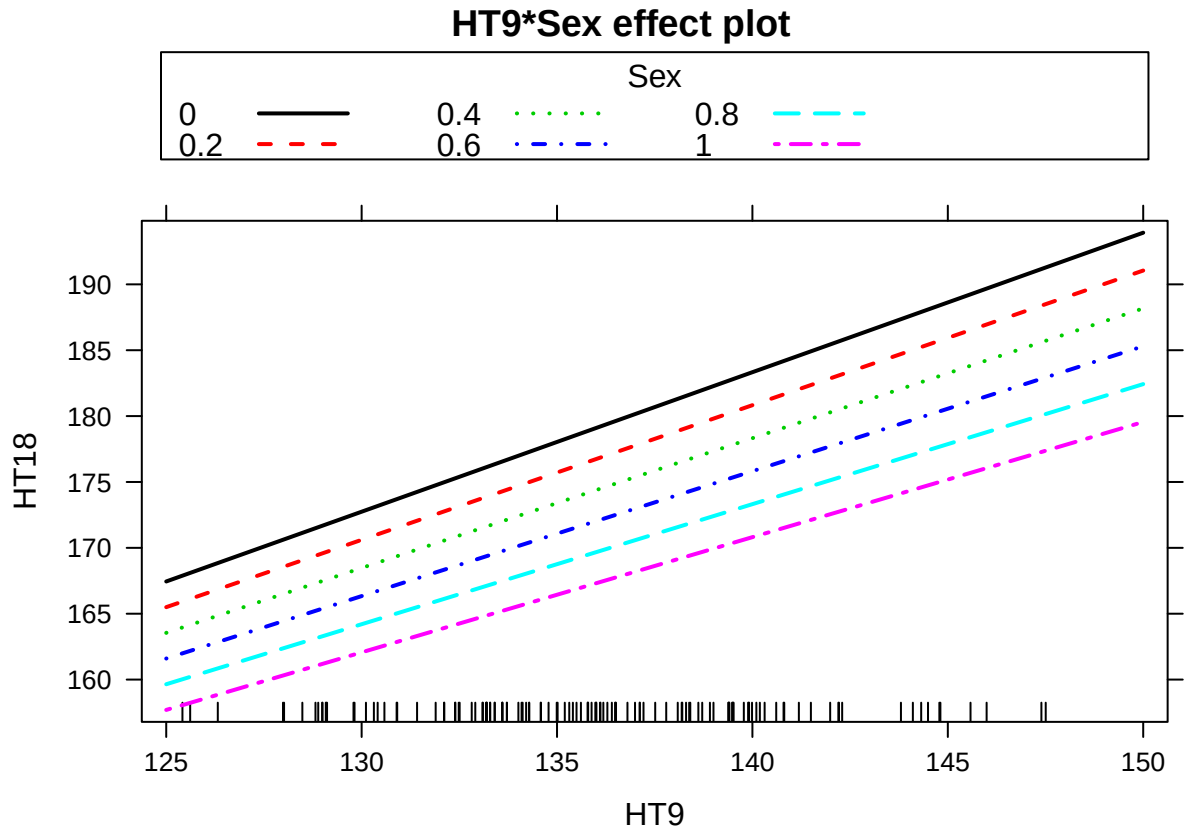
```
m1 <- lm(HT18~HT9 + Sex + HT9:Sex, data=BGSall)
summary(m1)
```

5.14.2

```
##
## Call:
## lm(formula = HT18 ~ HT9 + Sex + HT9:Sex, data = BGSall)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9224 -1.9453 -0.0081  1.7906 10.8136
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.07880   10.67406   3.286   0.0013 **
## HT9          1.05895    0.07849  13.492   <2e-16 ***
## Sex         13.32748   14.54695   0.916   0.3612
## HT9:Sex     -0.18463    0.10725  -1.722   0.0875 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.407 on 132 degrees of freedom
## Multiple R-squared:  0.8549, Adjusted R-squared:  0.8516
## F-statistic: 259.2 on 3 and 132 DF,  p-value: < 2.2e-16
```

```
plot(allEffects(m1), multiline = TRUE)
```




```
confint(m1, level = 0.95)["Sex",]
```

```
##     2.5 %    97.5 %
## -15.44783  42.10279
```

```
m1 <- lm(HT18~HT9 + Sex, data=BGSall)
confint(m1, level = 0.95)["Sex",]
```

```
##     2.5 %    97.5 %
## -12.86355 -10.52813
```

From the t-statistics of the interaction we can see that it is close to zero. Thus we can say they support parallel regression model.