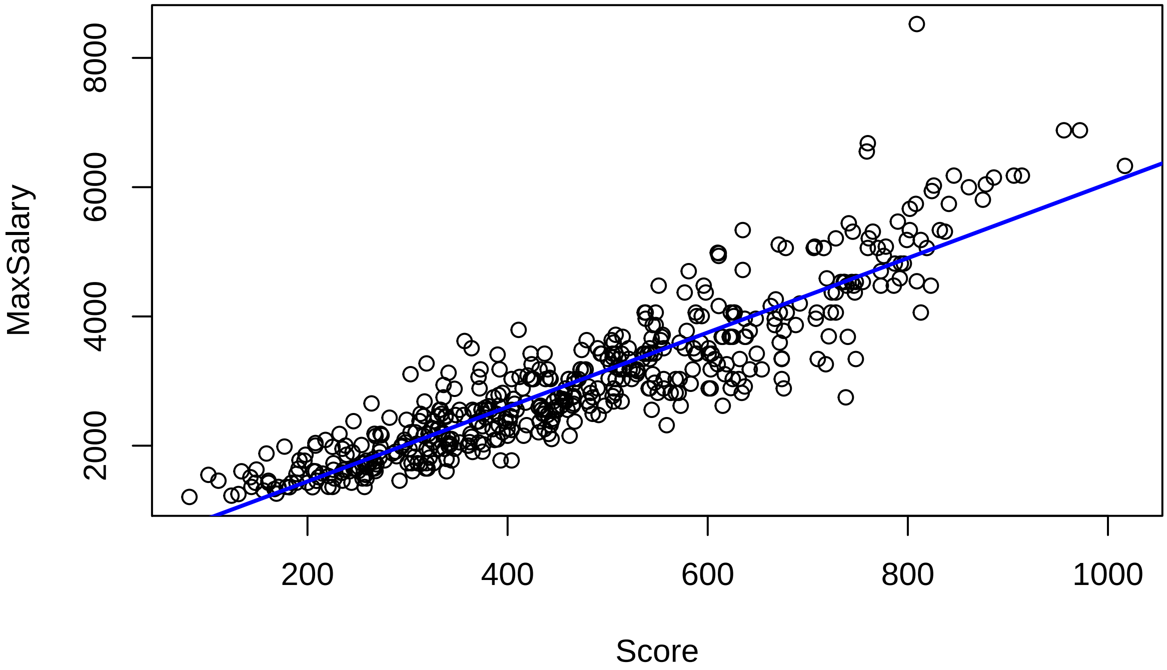
S631 HW7

Shibi He

```
library(alr4)
```

ALR 5.9.1 Scatterplot

```
plot(MaxSalary~Score, data=salarygov)
m1 <- lm(MaxSalary~Score, salarygov)
abline(m1, col="blue", lwd=2)
```



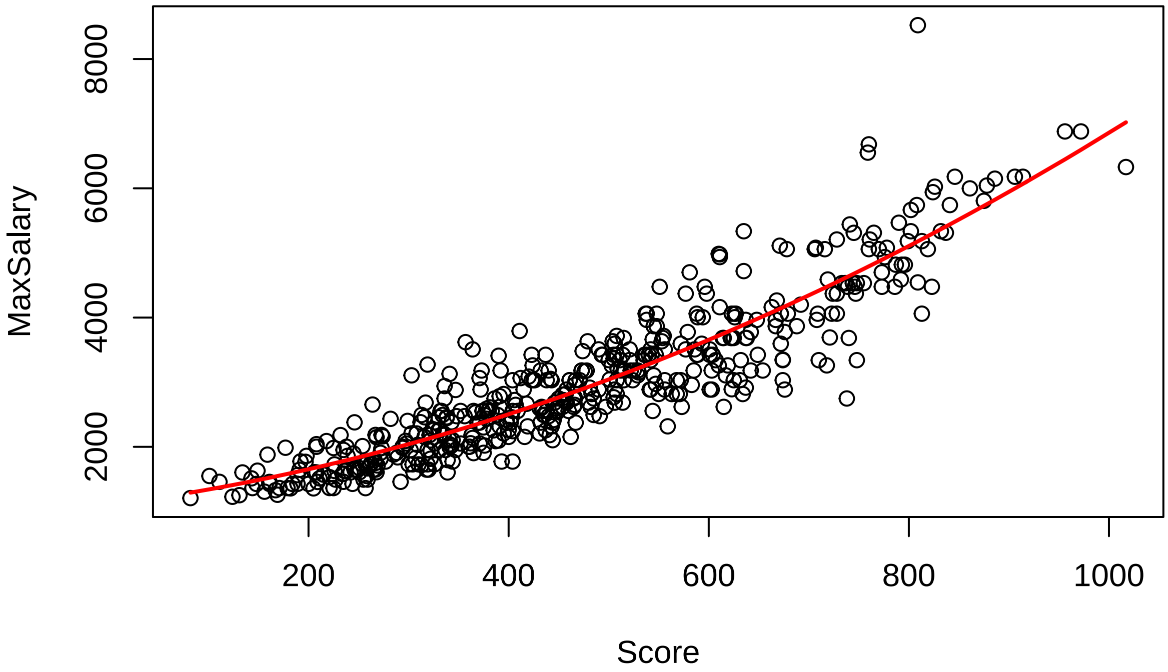
The scatterplot shows that the relationship between Score and MaxSalary seems not to be linear. Many data points are off the fitted linear regression line when the value of Score is high (800-1000). Moreover, the variation in MaxSalary seems to get larger and larger as Score increases.

ALR 5.9.2 Fit a quadratic polynomial model

```
m2 <- lm(MaxSalary ~ Score + I(Score^2), data = salarygov)
plot(MaxSalary~Score, data=salarygov)

#sort the data
newdata = data.frame(Score=salarygov$Score, MaxSalary=fitted(m2))
newdata = newdata[order(newdata$Score), ]

lines(newdata, col="red", lwd=2)
```



The quadratic polynomial model seems to describe the data better. It not only shows the positive relationship between Score and MaxSalary, but also shows that when Score reaches high value, such as 800-1000, the increase in Maxsalary becomes even faster.

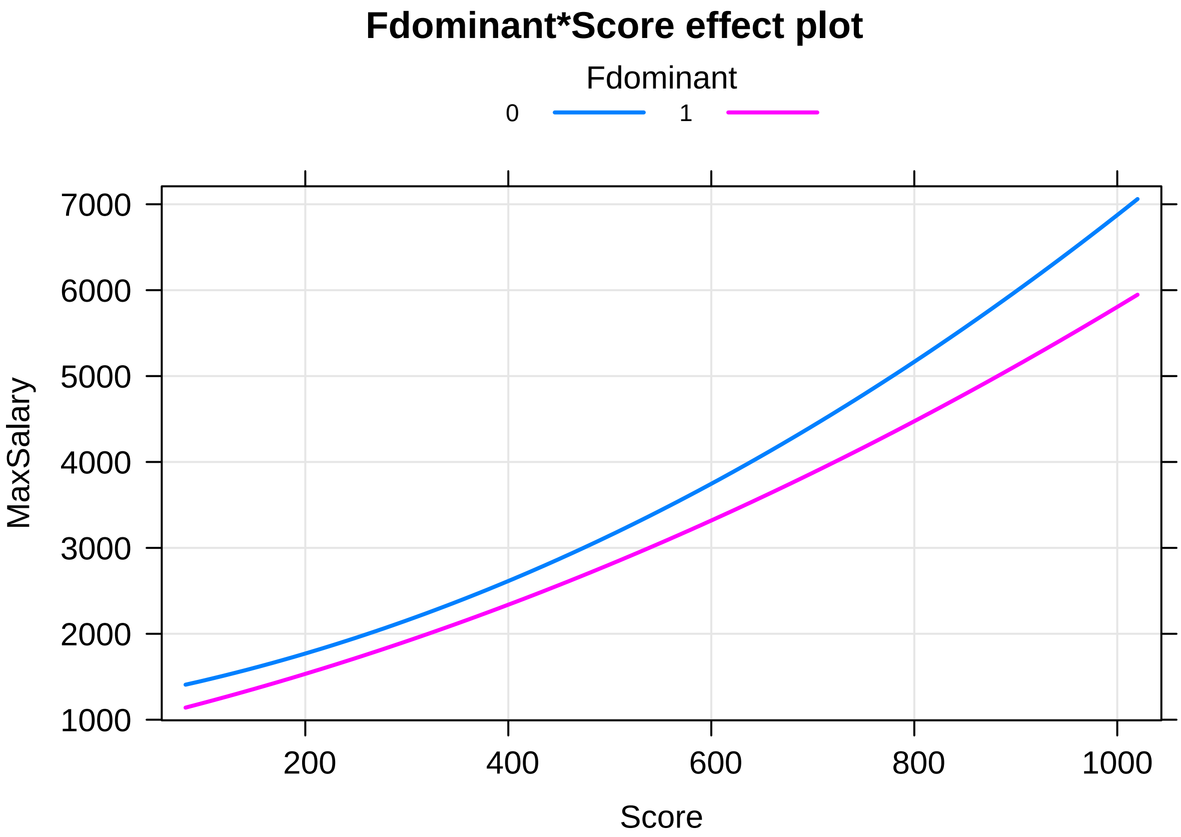
ALR 5.9.3 Are female-dominated positions compensated at lower level?

```
salarygov$percent=(salarygov$NW)/(salarygov$NE)
salarygov$Fdominant = ifelse(salarygov$percent>=0.7, 1, 0)
salarygov$Fdominant <- factor(salarygov$Fdominant)

m3 <- lm(MaxSalary ~ Score*Fdominant + I(Score^2)*Fdominant, salarygov)
summary(m3)$coefficients
```

```
##                            Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)            1.214508e+03 1.662064e+02  7.3072279 1.117979e-12
## Score                  2.059163e+00 6.679168e-01  3.0829636 2.165281e-03
## Fdominant1            -3.108937e+02 2.826799e+02 -1.0998082 2.719568e-01
## I(Score^2)             3.600655e-03 6.235937e-04  5.7740392 1.377106e-08
## Score:Fdominant1       6.526317e-01 1.322561e+00  0.4934604 6.219089e-01
## Fdominant1:I(Score^2) -1.411238e-03 1.446508e-03 -0.9756171 3.297366e-01
```

```
plot(Effect(c("Fdominant", "Score"), m3, xlevels=list(Score = 100)),
     rug=FALSE, grid=TRUE, multiline=TRUE)
```



The effect plot shows that for each Score values, the MaxSalary for female dominated job class (i.e. Fdominant=1) is always lower than the other job class(i.e. Fdominant=0). That means the female dominated positions are compensated at a lower level, adjusting for Score, than other positions.

ALR 6.11 Test for interaction and obtain confidence interval

```
summary(m3)$coefficients
```

```
##                           Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)           1.214508e+03 1.662064e+02  7.3072279 1.117979e-12
## Score                 2.059163e+00 6.679168e-01  3.0829636 2.165281e-03
## Fdominant1           -3.108937e+02 2.826799e+02 -1.0998082 2.719568e-01
## I(Score^2)            3.600655e-03 6.235937e-04  5.7740392 1.377106e-08
## Score:Fdominant1      6.526317e-01 1.322561e+00  0.4934604 6.219089e-01
## Fdominant1:I(Score^2) -1.411238e-03 1.446508e-03 -0.9756171 3.297366e-01
```

The t test results suggest that the interaction terms “Score:Fdominant1” and “Fdominant1:I(Score^2)” do not have statistically significant effects on the expected MaxSalary. Next, I run typeII anova analysis to see if it is adequate to add the interaction terms into a model that already contains the main effects.

```
Anova(m3)
```

```
## Anova Table (Type II tests)
##
## Response: MaxSalary
##                         Sum Sq  Df F value    Pr(>F)
## Score                  3181468   1 14.9047 0.0001283 ***
## Fdominant             10297214   1 48.2409 1.207e-11 ***
## I(Score^2)             7513932   1 35.2016 5.632e-09 ***
## Score:Fdominant          51977   1  0.2435 0.6219089
## Fdominant:I(Score^2)    203172   1  0.9518 0.3297366
## Residuals            104379097 489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results of Type II anova test show that the pvalue for the interaction “Score:Fdominant” and "I(Score^2)*Fdominant" is 0.62 and 0.33, respectively, which suggest that adding the interactions to a model that already contains the main effects is not statistically significant. In other words, there is no sufficient evidence to fit the effects of “Score” separately for female-dominated jobs and all other jobs.

Modify the model:

```
m4 <- lm(MaxSalary ~ Fdominant + Score + I(Score^2), salarygov)
confint(m4)[2,]
```

```
##     2.5 %    97.5 %
## -412.7859 -230.2718
```

The difference in the expected MaxSalary between female-dominated jobs and all other jobs is captured by the estimated coefficient of “Fdominant”. The 95% confidence interval for “Fdominant” is (-412.8 -230.3), suggesting that we are 95% confident that the expected MaxSalary for female-dominated jobs are $230.3 to $412.8 lower than all other jobs.

ALR 7.4.1 WLS

If we change the unit of analysis to employees, we are basically working with the grouped data. In other words, each job class is considered to be a group with different sample size. This suggests using WLS with the number of employees in each job class being the weights. These weightes are reasonable as we should pay more attention to job classes that have a large number of employees.

ALR 7.4.2 Repeat previous problem using WLS

```
m3.W <- lm(MaxSalary ~ Score*Fdominant + I(Score^2)*Fdominant, salarygov, weights = NE)
summary(m3.W)$coefficient
```

```
##                          Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)            1.043658e+03 1.171942e+02  8.9053699 1.047466e-17
## Score                  2.659847e+00 5.363054e-01  4.9595757 9.762940e-07
## Fdominant1            -2.822898e+02 1.713191e+02 -1.6477425 1.000479e-01
## I(Score^2)             3.382556e-03 5.906812e-04  5.7265342 1.791677e-08
## Score:Fdominant1       8.113175e-01 9.424294e-01  0.8608788 3.897267e-01
## Fdominant1:I(Score^2) -2.314039e-03 1.271612e-03 -1.8197681 6.940559e-02
```

Anova test:

```
Anova(m3.W)
```

```
## Anova Table (Type II tests)
##
## Response: MaxSalary
##                          Sum Sq  Df F value    Pr(>F)
## Score                  55017929   1 43.9195 9.057e-11 ***
## Fdominant             110723478   1 88.3880 < 2.2e-16 ***
## I(Score^2)             38059464   1 30.3820 5.751e-08 ***
## Score:Fdominant         928390   1  0.7411   0.38973
## Fdominant:I(Score^2)   4148380   1  3.3116   0.06941 .
## Residuals            612569431 489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again, the t tests results show that the two interaction terms are statistically insignificant and the type II anova test suggests that adding the interaction terms "Score:Fdominant" and "Fdominant:I(Score^2)" to a model that already contains the main effects is not statistically significant (pvalue=0.39 and 0.07, respectively).

Modify the model:

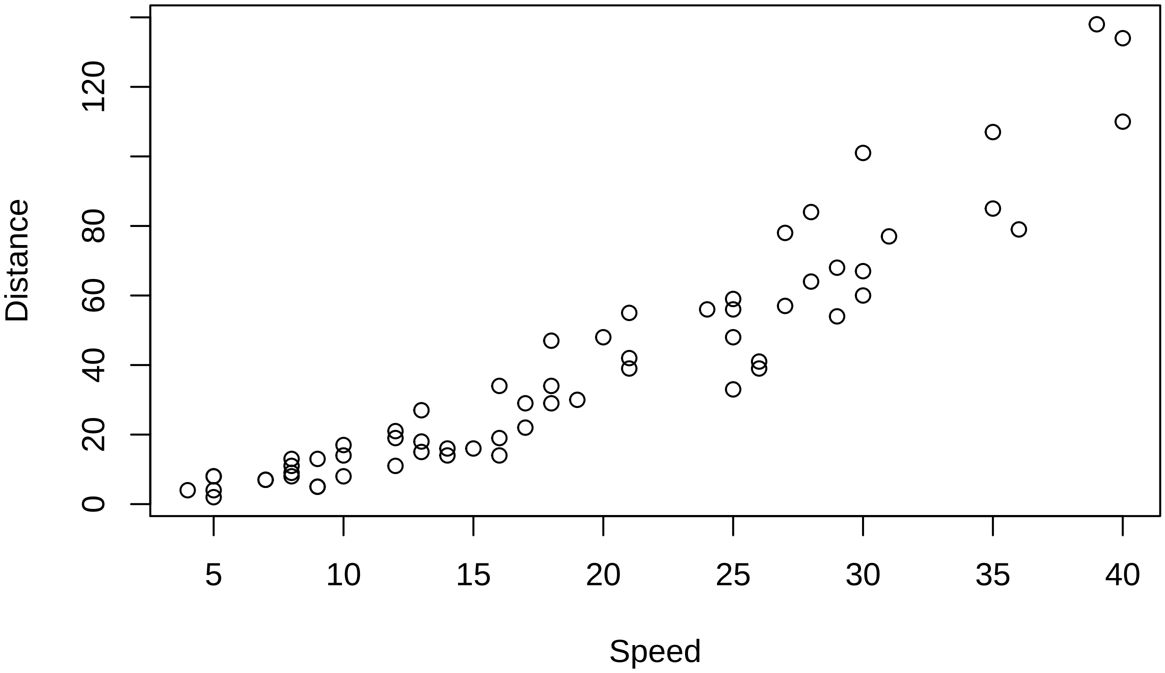
```
m4.W <- lm(MaxSalary ~ Fdominant + Score + I(Score^2), salarygov, weights = NE)
confint(m4.W)[2,]
```

```
##      2.5 %    97.5 %
## -367.7936 -239.3725
```

Using the WLS model, the 95% confidence interval for the difference between female-dominated job classes and all other job classes changes to (-367.8 -239.4).

ALR 7.6.1 Scatterplot

```
plot(Distance ~ Speed, stopping)
```



The relationship between Speed and Distance looks like a part of a parabola. Moreover, as Speed increases, the variation in Dsitance gets larger and larger. So this graph supports fitting a quadratic regression model.

ALR 7.6.2 Fit a quadratic model

Assume constant variance:

```
m1 <- lm(Distance ~ Speed + I(Speed^2), data=stopping)
summary(m1)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 1.58036341 5.10266295 0.3097135 7.578701e-01
## Speed       0.41606845 0.55641125 0.7477714 4.575682e-01
## I(Speed^2)  0.06555584 0.01302574 5.0327930 4.834919e-06
```

Compute score test for nonconstant variance:

```
# Variance depends on the fitted values
Z1=with(stopping,ncvTest(m1))

# Variance depends on Speed
Z2=with(stopping,ncvTest(m1, ~ Speed))

# Variance depends on Speed and Speed^2
Z3=with(stopping,ncvTest(m1, ~ Speed + I(Speed^2)))

# Make a table
table1=rbind(with(Z1,c(Df,ChiSquare,p)),with(Z2,c(Df,ChiSquare,p)),with(Z3,c(Df,ChiSquare,p)))
row.names(table1)=c("Fitted values","Speed","Speed, Speed^2")
colnames(table1)=c("df","Test statistic","p-Value")

table1
```

```
##                df Test statistic      p-Value
## Fitted values   1       22.97013 1.645386e-06
## Speed           1       23.39216 1.321162e-06
## Speed, Speed^2  2       23.46559 8.026245e-06
```

Is adding Speed^2 helpful?

```
1-pchisq(Z3$ChiSquare-Z2$ChiSquare,Z3$Df-Z2$Df)
```

```
## [1] 0.7864035
```

Comparing Z3 with Z2 gives a p-value of 0.7864, suggesting that the simpler Z with only Speed is adequate. Therefore, adding Speed^2 is not helpful.

ALR 7.6.3 Refit the quadratic model

```
m1.W <- lm(Distance ~ Speed + I(Speed^2), data=stopping, weights = 1/Speed)
compareCoefs(m1, m1.W)
```

```
## Calls:
## 1: lm(formula = Distance ~ Speed + I(Speed^2), data = stopping)
## 2: lm(formula = Distance ~ Speed + I(Speed^2), data = stopping,
##   weights = 1/Speed)
##
##             Model 1 Model 2
## (Intercept)    1.58    1.33
## SE             5.10    3.10
##
```

```
## Speed         0.416   0.448
## SE            0.556   0.421
##
## I(Speed^2)   0.0656  0.0648
## SE           0.0130  0.0112
##
```

Model 1 has constant variance and in Model 2, the variance is weighted by Speed. The estimated coefficients are similar, but the standard errors with weighted variance are less than the standard errors in the unweighted case.

ALR 7.6.4 Sandwich estimator of the variance

```
# variance-covariance matrix for betahat
hccm(m1, type="hc3")
```

```
##             (Intercept)       Speed    I(Speed^2)
## (Intercept) 18.45413036 -2.65217062  0.0690953828
## Speed       -2.65217062  0.39729977 -0.0106319865
## I(Speed^2)   0.06909538 -0.01063199  0.0002974929
```

Compare the standard errors

```
cbind("Unweighted SE"=sqrt(diag(vcov(m1))),
      "Weighted SE" = summary(m1.W)$coefficient[,2],
      "HC3 SE"=sqrt(diag(hccm(m1, type="hc3"))))
```

```
##             Unweighted SE Weighted SE     HC3 SE
## (Intercept)    5.10266295  3.09898391 4.29582709
## Speed          0.55641125  0.42064971 0.63031720
## I(Speed^2)     0.01302574  0.01121552 0.01724798
```

All HC3 standard errors are greater than the standard errors in the weighted case. Compared to the standard errors in the unweighted case, the HC3 standard error of intercept is lower while the standard errors of Speed and Speed^2 are higher.