SB1 Assessed Practical 1

Candidate number: 1035161

November 20, 2020

1 Data

The data set gives the times run by the fastest male and female athlete for some hill races. It consists of 68 observations and has 4 explanatory variables. Two of them are categorical (the sex of the athlete, and the category of the race, which can be 'short', 'medium' or 'long'); the remaining two are continuous variables, notably the distance of the race (in km) and the total gain in height (in m).

From figure (1a) we can see that variables "time" and "climb" show a decent correlation. From figure (1b) it is possible to observe a stronger relation between time and distance. At a first glance the pattern seems to be slightly convex, which could be physically interpreted as the fact that athletes get tired and so have a lower average speed in longer races. In addition, it can be clearly seen races of length less than 10km are in the short category, races between 10km and 20km are medium, and those longer than 20km are in the long category (this suggests that the variable 'category' might actually be superfluous as the information is already contained by 'distance').

Another noteworthy feature that can be seen in (1b) is that the time difference between the female and male time for a fixed race seems to grow as distance increases. Such relation, as can be seen in figure (2a), looks approximately linear; this leads us to think that there might be a multiplicative interaction between sex and distance. Finally, in (2b) climb and distance can be seen to be correlated.

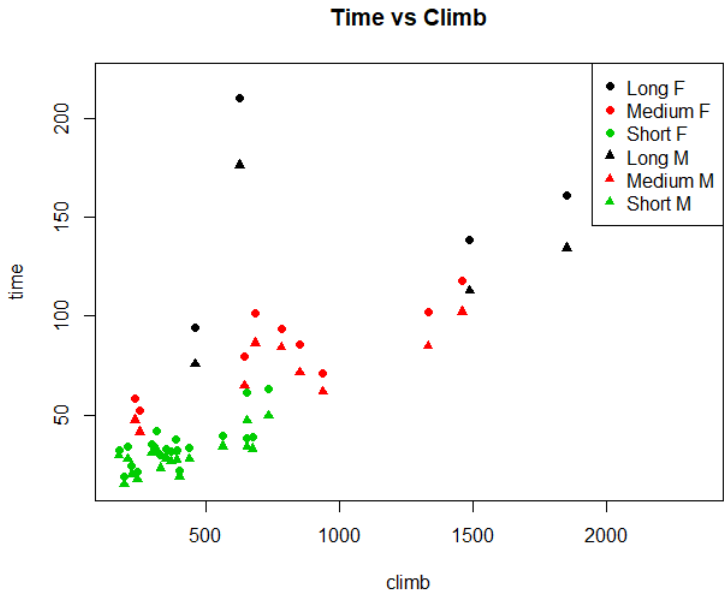

Figure (1a) Exploratory analysis of the data

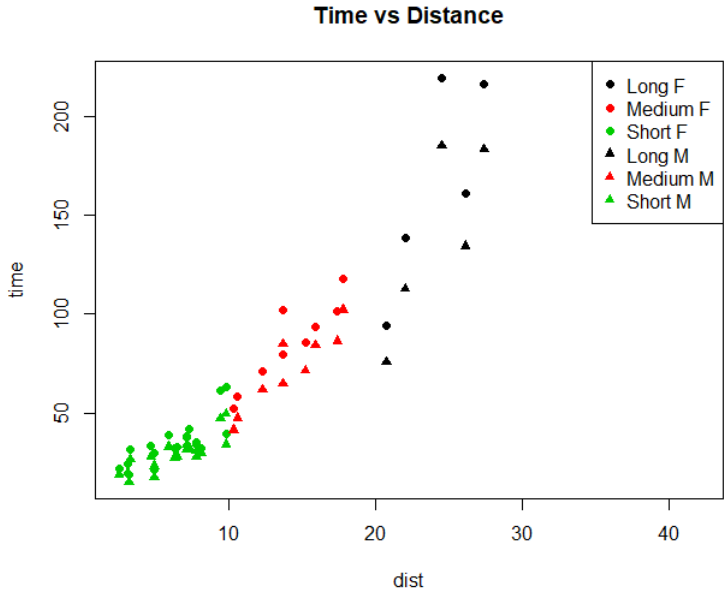

Figure (1b) Exploratory analysis of the data

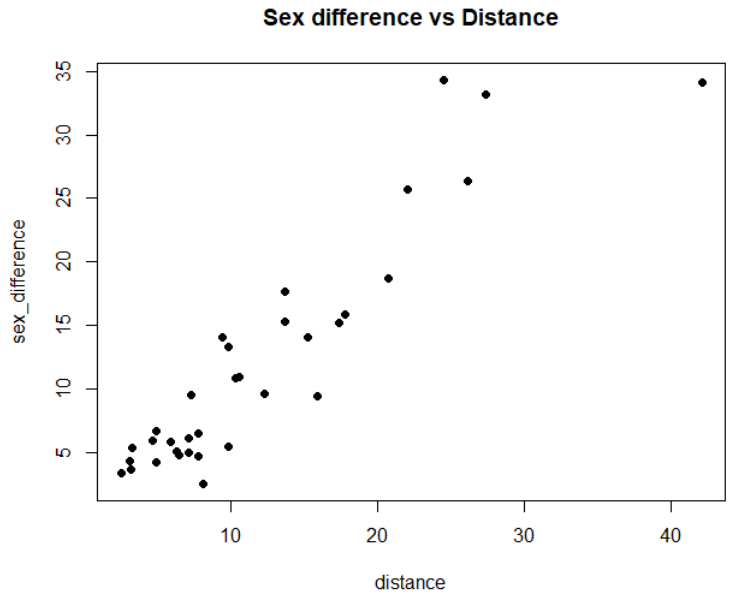

Figure (2a) Difference between female and male times for each of the 34 races

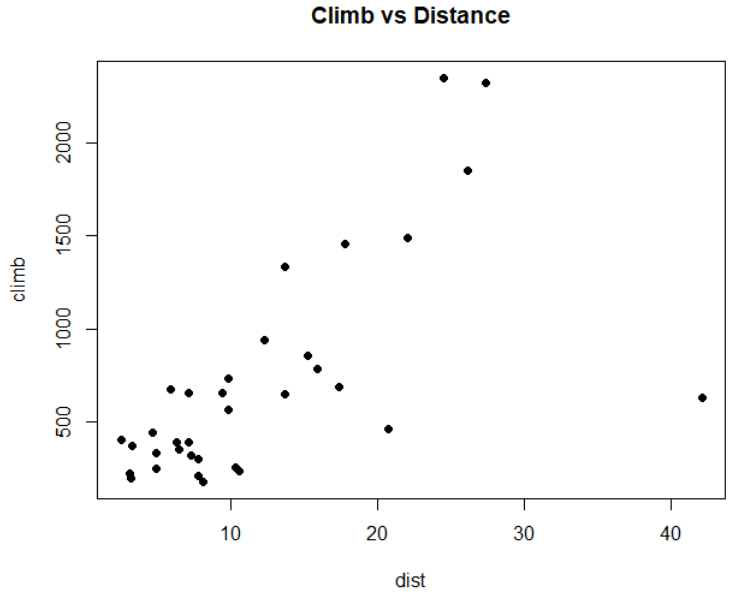

Figure (2b) Exploratory analysis of the data

2 Modelling

We start with a simple linear model with no variable interactions (including all four explanatory variables):

$$E(time) = \beta_0 + \beta_1 dist + \beta_2 climb + \beta 3 I(sex = female)$$

$$+\beta_4 I(category = medium) + \beta_5 I(category = long)$$

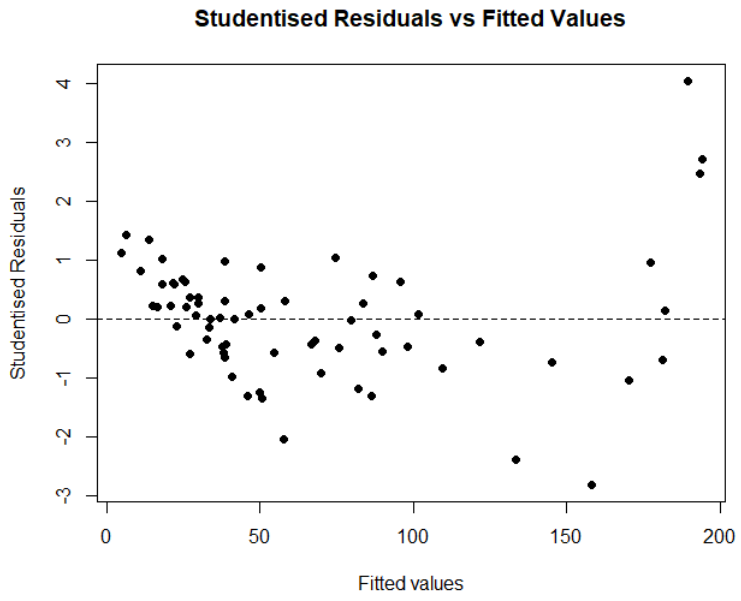

Figure (3) Residuals plot for initial regression

From (3) it can be seen that the variance of the studentized residuals grows with the fitted values, contradicting the assumption of independence of e and ŷ.

From a physical perspective, we could model a race Y of distance n as the sum of n independent races X_i of length 1, thus the variance of the response for an observation would be proportional to its distance value.

$$\text{(a)} \qquad Var(Y) = Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} var(Xi) = n * Var(X_1)$$

However, if we reflect more deeply about the nature of our observations, the independence assumption does not look realistic. It is reasonable to assume

that there are factors affecting the racetime that are not incorporated in our dataset (for example weather conditions, type of surface, ability of the athletes). Such factors will be the same for each of the X_i and hence will lead to a higher variance than under the independence assumption. We could think of a race Y of distance n as $Y = n * X$ where X is a race of distance 1. In this case the variance of the response for an observation would be proportional to the square of its distance.

$$\text{(b)} \qquad Var(Y) = Var(nX) = n^2 Var(X)$$

In figure (4a) we can see that the residuals plot under assumption (a) still exhibits a funnel pattern, meanwhile the more natural interpretation (b) produces a reasonable diagnostic plot (4b)

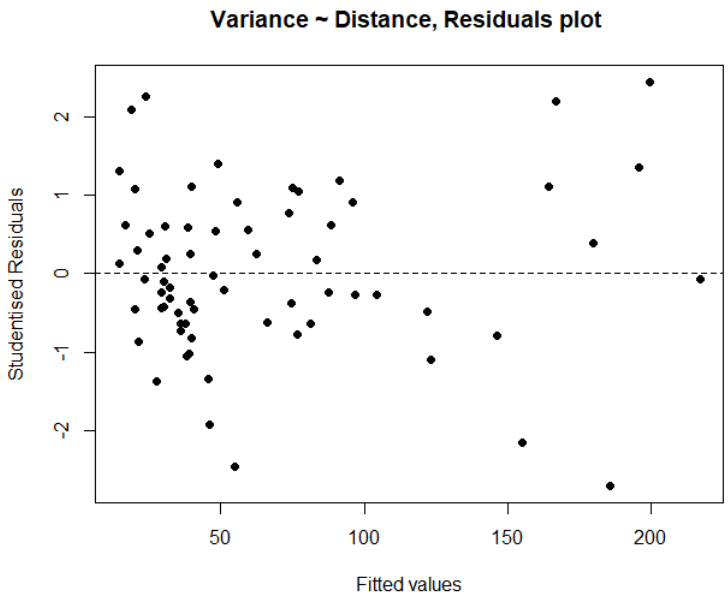

Figure (4a) Residuals plot for variance $\propto$ distance

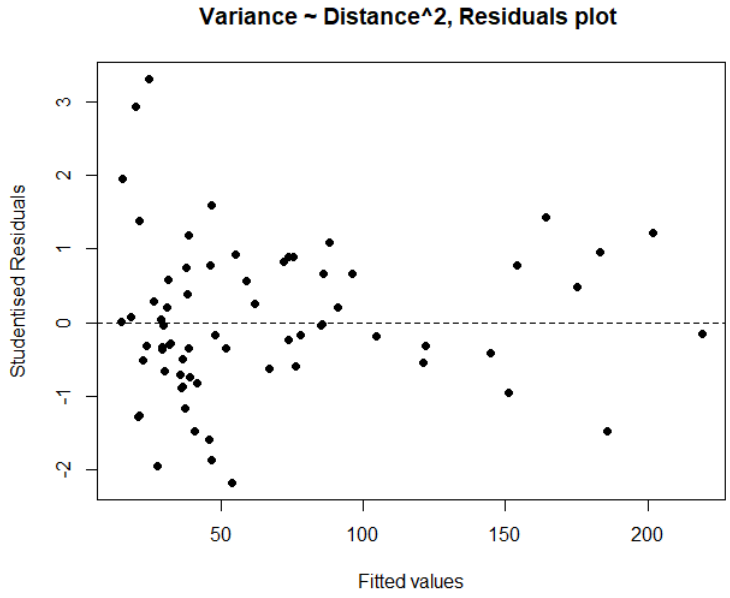

Figure (4b) Residuals plot for variance $\propto distance^2$

By the considerations made in the exploratory data analysis in section 1, we initially exclude *category* from our initial model and also *intercept* , as it does not make physical sense to have a positive predicted time for a race of distance zero. We also add interactions between the remaining 3 explanatory variables (see (2a)). So the new weighted regression model will be:

$$ModelA:$$

$$time = \beta_1 dist + \beta_2 climb + \beta_3 I(sex = female) + \beta_4 dist \cdot climb +$$
$$\beta_5 dist \cdot I(sex = female) + \beta_6 climb \cdot I(sex = female)$$
$$+\beta_7 dist \cdot climb \cdot I(sex = female) + \epsilon$$
$$where \quad \epsilon \sim N(0, dist^2 \sigma^2)$$

Now, we want to test our physical intuition H_0 : $modelA$ against H_1 : $modelA + intercept$. The F test gives us a p-value of 0.052 which is above the standard significance threshold of 0.05. Hence we have no significative evidence against H0.

```
Model 1:  time = dist * climb * sex - 1
Model 2:  time = dist * climb * sex
 Res.Df RSS Df Sum of Sq F Pr(>F)
1 61 31.356
2 60 29.424 1 1.9325 3.9407 0.05171
```

Similarly we want to test the hypothesis that *category* is not significant. $H_0 : modelA, H_1 : modelA + I(category = medium) + I(category = long)$. The p-value for the F test is slightly below 0.05. However, as our main aim is interpretability, we decide that this in not sufficient evidence against H_0.

```
Model 1:  time = dist * climb * sex - 1
Model 2:  time = medium + long + dist * climb * sex - 1
Res.Df RSS Df Sum of Sq F Pr(>F)
1 61 31.356
2 59 27.972 2 3.3845 3.5694 0.03441 *
```

Now that our initial intuitions have been confirmed, we carry out an ANOVA for model A to see if the model can be further simplified. It can be seen from the table below that we can drop $climb \cdot I(sex = female)$ and $dist \cdot climb \cdot I(sex = female)$.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dist	1	2067.18	2067.18	4021.4746	< 2.2e-16 ***
climb	1	67.48	67.48	131.2807	< 2.2e-16 ***
sex	1	8.70	8.70	16.9303	0.0001182 ***
dist:climb	1	14.03	14.03	27.2897	2.241e-06 ***
dist:sex	1	2.17	2.17	4.2187	0.0442737 *
climb:sex	1	0.00	0.00	0.0060	0.9384119 .
dist:climb:sex	1	1.92	1.92	3.7282	0.0581514 .
Residuals	61	31.36	0.51		

Therefore our new simplified model is:

$$ModelB: \quad time = \beta_1 dist + \beta_2 climb + \beta_3 I(sex = female) + \beta_4 dist \cdot climb$$
$$+\beta_5 dist \cdot I(sex = female) + \epsilon$$
$$where \quad \epsilon \sim N(0, dist^2 \sigma^2)$$

The diagnostic plots for Model B (5) look acceptable. From (6) we can see that there are several points with high leverage but only two observations have Cook's distance significantly higher than the approximate treshold 8/(n-2p). Such observations (indices 14, 48) correspond to female and male records

for the exact same race. This is gives additional support to the decision of deletion. Note also that the race is in the short category (we have plenty of observations is such category), therefore deleting should not cause much trouble.

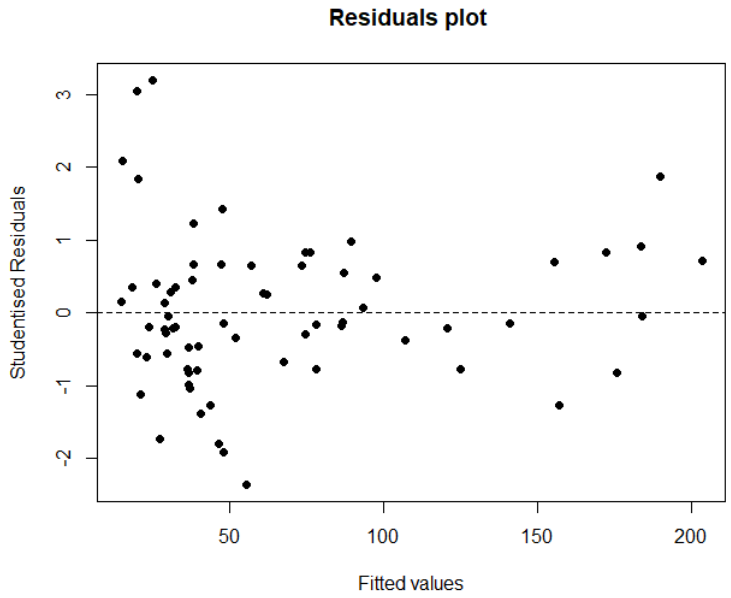

Figure (5a) Studentised residuals plot for Model B

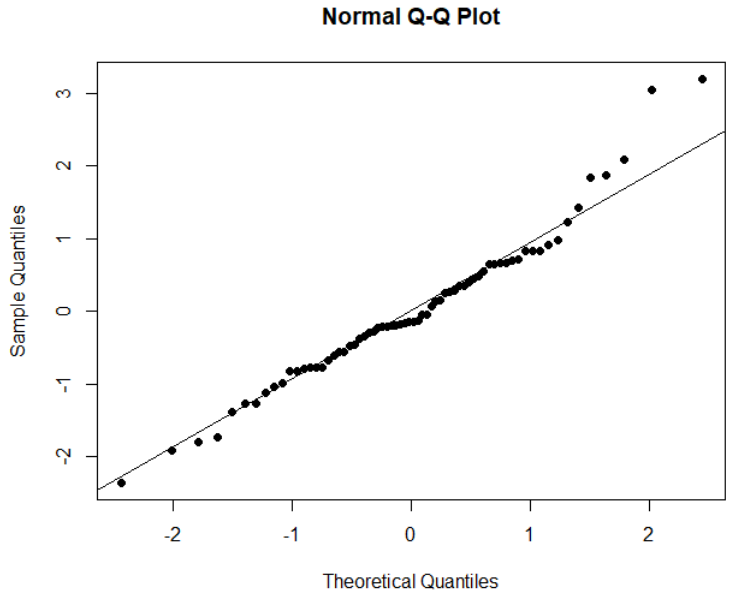

Figure (5b) qq plot for Model B

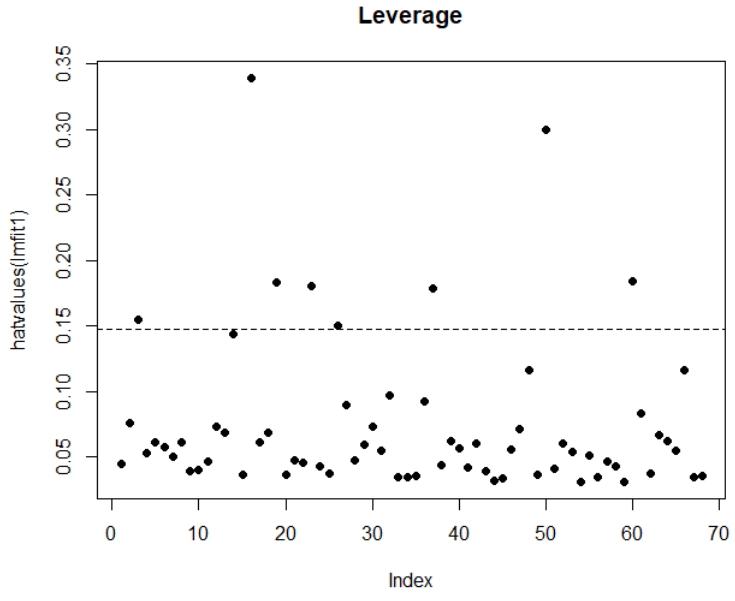

Figure (6a) Leverage plot for Model B

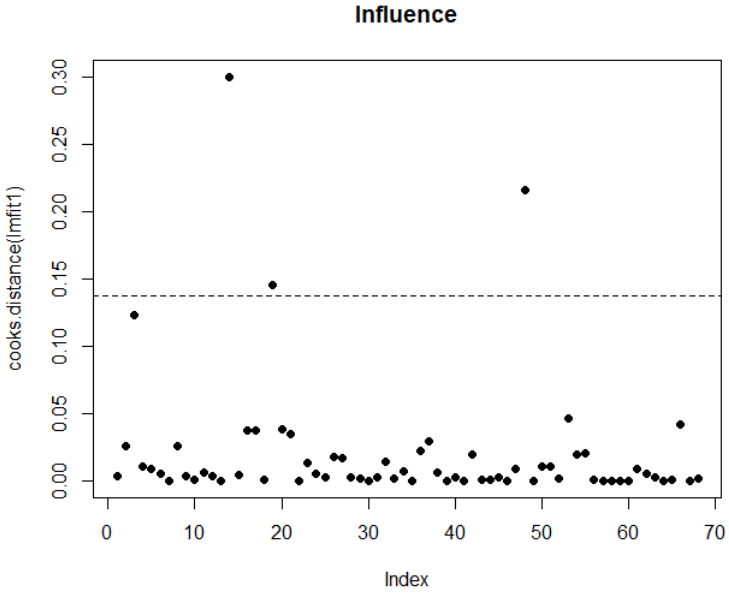

Figure (6b) Cook's distances plot for Model B

After having deleted the two observations, keeping in mind that we aim to have a model easy to interpret, we choose to use BIC to carry out model selection (we are penalizing complex models more than in AIC). Figure (7) together with the table below give us that the best model according to BIC is:

$$ModelC: \qquad time = \beta_1 dist + + \beta_2 dist \cdot climb + \beta_3 dist \cdot I(sex = female) + \epsilon$$

$$where \quad \epsilon \sim N(0, dist^2\sigma^2)$$

	dist	climb	sex	dist:climb	dist:sex
2	TRUE	FALSE	FALSE	TRUE	FALSE
3	TRUE	FALSE	FALSE	TRUE	TRUE
4	TRUE	FALSE	TRUE	TRUE	TRUE
5	TRUE	TRUE	TRUE	TRUE	TRUE

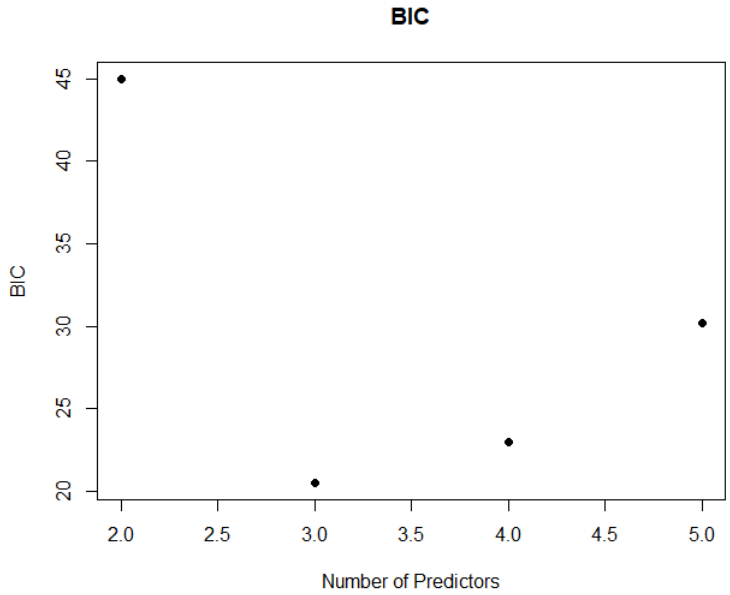

Figure (7) BIC for Model B

The outlier analysis gives us just a point slighty above the 8/(n-2p) threshold which we can easily keep (note that there a good number of points with Cook's distance above 0.15, hence keeping the point is not as problematic as in the case where all the other points have low distances). Therefore Model C will be our final model.

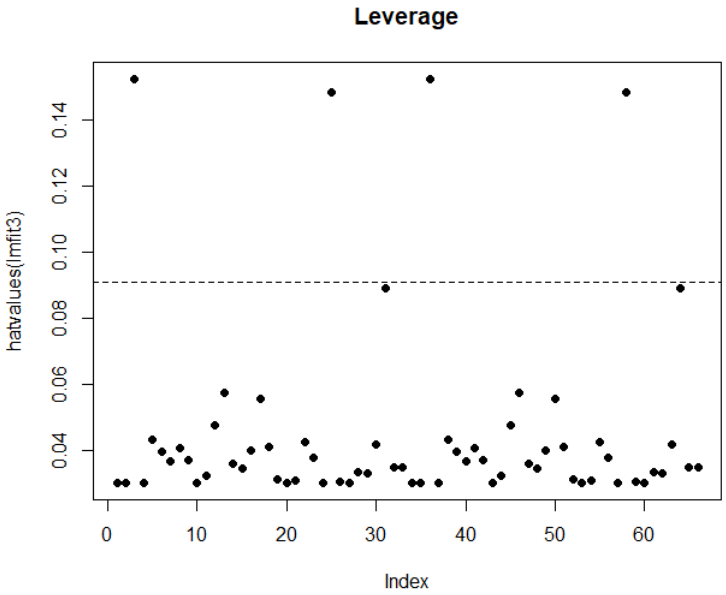

Figure (8a) Leverage plot for Model C

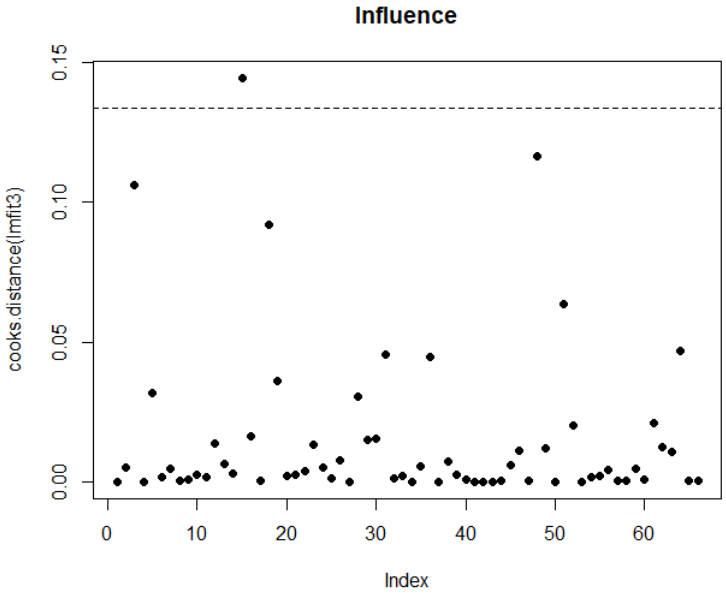

Figure (68b) Cook's distances plot for Model C

3 Interpretation

From the table below we can see that, according to our final linear model, the expected increase in time for a 1km increase in distance when climb is zero and sex = male is 4.089 minutes; the expected increase in time for a 1m increase in climb when distance in fixed is $0.00116 \cdot dist$ minutes, the difference between the expected time of a female and male for a race of distance d is $0.996 \cdot d$ minutes.

```
Coefficients:
             Estimate   Std.  Error  t value      Pr(>|t|)
 dist        4.0888166  0.2191774    18.655    < 2e-16 ***
 dist:climb  0.0011596  0.0002021    5.737    2.96e-07 ***
 dist:sex    0.9958370  0.2332513    4.269    6.73e-05 ***
```

Consider the following observations and let $X = (x_1^T, x_2^T, x_3^T, x_4^T)$ be the corresponding data matrix, where $x_i = (dist_i, climb_i, sex_i)^T$

name	distance	climb	sex
RaceA	5	330	F
RaceB	2	200	F
RaceC	8	2000	F
RaceD	50	3000	F

The predicted time for each observation i (i = 1, 2, 3,4) is :

$$\hat{y}_i = \hat{\beta}^T M x_i \qquad where \quad \hat{\beta} = (4.089, 0.00116, 0.995)^T and \quad M = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

$Hence \quad \hat{y} = (\hat{y}_1, \hat{y}_2, \hat{y}_3, \hat{y}_4)^T = (27.34, 10.64, 59.23, 428.2)^T$
Let y' and X' be the data and design matrix for the weighted regression, that is:

$$y' = diag(1/dist_1, ..., 1/dist_4)y \quad and \quad X = diag(1/dist_1, ..., 1/dist_4)X$$

Similarly $\quad s' = \sqrt{RSS'/66} = 1.103$

Therefore as we have constant variance in this case (n = 66, p = 3) , a $1 - \alpha$ prediction interval for y_i' is:

$$(y_i'^{low}, y_i'^{high}) = (hat_y \pm t_{63}(\alpha/2)s'\sqrt{1 + x_i'^T (X'^T X')^{-1} x_i'})$$

Hence by multiplying tha above by $diag(1/dist_1, ..., 1/dist_4)$, we get that a $1 - \alpha$ prediction interval for y is:

$$(y_{low}, y_{high}) = diag(1/dist_1, ..., 1/dist_4)(y'_{low}, y'_{high})$$

The predicted times and endpoints of the 95%-prediction intervals for X will be:

y_{low}	$\hat{y}$	y_{high}
16.03	27.34	38.64
5.432	10.61	15.83
36.92	59.23	81.53
317.0	428.2	539.4

4 Conclusions

After having seen that a standard linear model resulted in uneven variances of the residuals, we used a weighted regression approach with weights proportional to the inverse of the squared distances. Such model made sense physically and was further simplified to reach a final model with only three parameters. The model indicated that there was a multiplicative relation between distance and climb, that the expected difference int times for males and females was proportional to the distance. In addition, the extreme interpretability of the model allowed us to formulate sensible prediction intervals even for observations whose distances were outside of the range the model was trained on.

5 Appendix

```
racetimes <- read.csv("http://www.stats.ox.ac.uk/~laws/SB1/data/racetimes.csv",
                      stringsAsFactors = TRUE)
#INITIAL EXPLORATION
#WE SEE THAT FUNCTION LOOKS CONVEX, MAKES SENSE
plot(time ~ dist, data = racetimes, main = "Time vs Distance",
     pch = as.numeric(racetimes$sex) + 15,
     col = as.numeric(racetimes$category))
legend("topright", c( "Long F", "Medium F", "Short F",
"Long M", "Medium M",
     "Short M") , col = rep(c(1,2, 3), 2) , pch = c(16,16,16,17,17,17))

#SOME DECENT CORRELATION, NOT AS STRONG AS WITH DIST
plot(time ~ climb, data = racetimes, main = "Time vs Climb",
     pch = as.numeric(racetimes$sex) + 15,
     col = as.numeric(racetimes$category))
legend("topright", c( "Long F", "Medium F", "Short F", "Long M",
      "Medium M", "Short M") , col = rep(c(1,2, 3), 2) ,
      pch = c(16,16,16,17,17,17))
plot( climb ~ dist , data = racetimes,
      main = "Climb vs Distance", pch = 16)

#INDICATES MULTIPLICATIVE RELATION/EFFECT
sex_difference = rep(c(0), 34)
for (i in 1:35){
  sex_difference[i] = - racetimes$time[34+i] + racetimes$time[i]
}
distance = racetimes$dist[1:35]
plot(sex_difference ~ distance, main = "Sex difference vs Distance", pch = 16 )

#PROOF WITH A LINEAR MODEL WITH NO INTERACTIONS
#WE SEE RESIDUALS HAVE UNEVEN VARIANCE
lm1 = lm(time ~ ., data = racetimes)
plot(fitted(lm1),rstudent(lm1),xlab="Fitted values",pch=16, main =
       "Studentised Residuals vs Fitted Values", ylab = "Studentised Residuals")
abline(0,0,lty = 2)

#physical condiseration EITHER DIST OR DIST^2 ARE PROP TO VARIANCE,
^2 case better and even more physical sense(weather example)
w = racetimes$dist
z = lm(time ~ dist*climb*sex -1, data = racetimes, weight = (1/w))
x = lm(time ~ dist*climb*sex -1, data = racetimes, weight = (1/w)^2)
```

```
plot(fitted(z),rstudent(z),xlab="Fitted values",pch=16, main =
       "Variance ~ Distance, Residuals plot ", ylab = "Studentised Residuals")
abline(0,0,lty = 2)
plot(fitted(x),rstudent(x),xlab="Fitted values",pch=16,
     main = "Variance ~ Distance^2, Residuals plot ", ylab =
      "Studentised Residuals")
abline(0,0,lty = 2)

#chosen our weight function, and our starting natural model,
#test hp of no intercept with F test
y = lm(time ~ dist*climb*sex, data = racetimes, weight = 1/w^2)
anova(x,y)
#VALUE 0.5, TAKING INTO CONSIDERATION PHYSICAL

#cATEGORY INFORMATION IS INCLUDED IN DIST TEST NO CATEGORY WITH F TEST
z =  lm(time ~ medium + long + dist*climb*sex -1, data = racetimes,
weight = (1/w)^2)
anova(x,z) #AROUND THE 5% THRESHOLD 3.5%,

#NOW BACKWARDS SELECTION
anova(x)
#WE SEE CLIMB:SEX CAN BE EXCLUDED, SAME FOR TRIPLE TERM
lmfit1 = lm(time ~ dist + climb + sex + dist:climb + dist:sex -1,
 data = racetimes, weight = (1/w)^2)
#4 POINTS (RACES 14,20,48,53) HAVE ABS(ST RES) > 2 SO SHOW MISFIT AND
#WE WILL NEED TO INVESTIGATE AS THE COULD POTENTIALLY BE OUTLIERS
plot(fitted(lmfit1),rstudent(lmfit1),xlab="Fitted values",pch=16,
main = "Residuals plot ",  ylab = "Studentised Residuals")
abline(0,0,lty = 2)
qqnorm(rstudent(lmfit1), pch = 16) #comment briefly saying it looks ok
qqline(rstudent(lmfit1))
which(abs(rstudent(lmfit1)) >2)


#OUTLIER ANALYSIS
n= 68
p = 5
plot(hatvalues(lmfit1), ylabel = "Leverage Component",
 main = "Leverage", pch = 16)
abline(2*p/n, 0, lty = 2)
plot(cooks.distance(lmfit1), ylabel = "Cook's Distance",
main = "Influence", pch = 16)
abline(8/(n - 2*p), 0, lty = 2)
```

```
which(abs(cooks.distance(lmfit1)) >0.15)
#14,48 small and from the same race nice!


#DELETE OUTLIERS
new = racetimes[-c(14,48),]
ww = w[-c(14,48)]
lmfit2 = lm(time ~ dist + climb + sex + dist:climb + dist:sex -1,
            data = new, weight = (1/ww)^2)

#DO BIC and aic, choose bic as we want to favour interpretability
#aggiungi tavola dei which
library(leaps)
b <- regsubsets(time ~ dist + climb + sex + dist:climb + dist:sex -1,
                data = new, weight = (1/ww)^2)
rs <- summary(b)
rs$which
AIC <- 66*log(rs$rss/66) + c(2:5)*2
plot(AIC ~ I(2:5), main = "AIC",ylab = "AIC",
 xlab = "Number of Predictors", pch = 16)
BIC <- 66*log(rs$rss/66) + c(2:5)*2*log(66)
plot(BIC ~ I(2:5), main = "BIC", ylab = "BIC", xlab =
       "Number of Predictors", pch = 16)
       #WE SEE BEST IS DIST + D:C + D:S AS EXPECTED!

#NEW OUTLIER ANALYSIS
lmfit3 = lm(time ~ dist + dist:climb + dist:sex -1,
data = new, weight = (1/ww)^2)
n = 66
p = 3
plot(hatvalues(lmfit3),  main = "Leverage", ylabel =
"Leverage Component", pch = 16)
abline(2*p/n, 0, lty = 2)
plot(cooks.distance(lmfit3),main = "Influence",
     ylabel = "Cook's Distance",  pch = 16)
abline(8/(n - 2*p), 0, lty = 2)
which(abs(cooks.distance(lmfit3)) >0.13)
#only one outlier slightly above the threshold which we decide to keep


#BETA HAT FOR LMFIT3
#MATRIX FOR THE 4 NEW OBSERVATIONS
coeff = c(4.088816624, 0.001159558, 0.995837017)
```

```
mat = matrix (c(5,2,8,50,330,200,2000,3000,1,1,1,1), ncol = 3, nrow= 4)

#PREDICTED TIMES FOR THE 4 OBS USING MODEL3
out = numeric(4)
for (i in 1:4){
  out[i] = (mat[i,1]*coeff[1] + mat[i,2]*mat[i,1]*coeff[2]
            + mat[i,3]*mat[i][1]*coeff[3])

}

#FITTED VALUES Y PRIME, WEIGTHED DATA MATRIX ETC
out_p = out / mat[,1]
mat_p = diag(1/mat[,1]) %*% mat
RSS_p = sum(((1/ww)^2)*residuals(lmfit1)^2)
S_p = sqrt(RSS_p/66)
X = (data.matrix(new)[,c(3,4,5)])
X_p = diag(1/X[,1]) %*% X
XtX_p_inv = solve(t(X_p)%*%X_p)

#SQRT OF 1 MINUS ETC
value = numeric(4)
for (i in 1:4){
  value[i] = S_p * sqrt(1 + t(mat_p[i,]) %*% XtX_p_inv %*% (mat_p[i,]))
}


#these are the intervals for y prime
interval_low = out_p - qt(0.975, df = 63) *value
interval_high =out_p + qt(0.975, df = 63) *value

#PREDICTION INTERVALS AND PREDICTED VALUES
i_l = mat[,1] * interval_low
i_h = mat[,1] * interval_high
out
i_l
i_h
```