A1

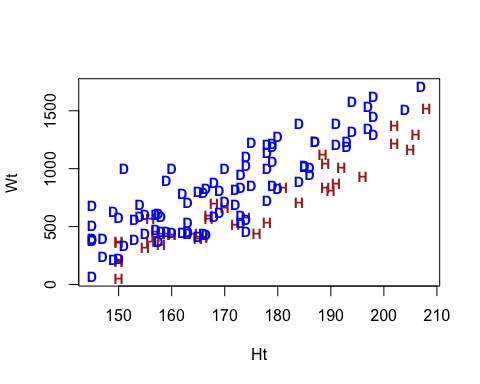
Kerui Du

2021-08-12

## Question 1

### 1a

library(s20x)  
PPW.df <- read.csv('data/PPW.csv')  
plot(Wt~Ht, type="n", data=PPW.df)   
text(Wt~Ht, labels=substr(How,1,1),  
 col=ifelse(How=="Drown","blue", "brown"),  
 data=PPW.df, cex=0.9, font=2)

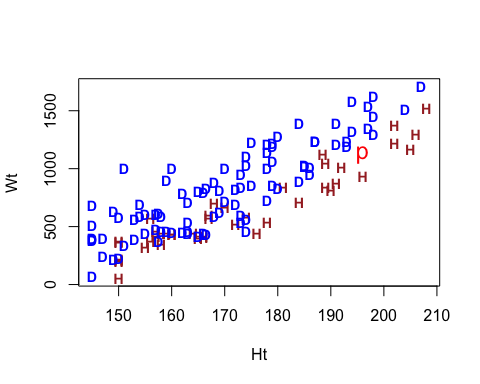


By looking at the scatter plot, there is a positive linear relationship between the person’s height and the weight of the person’s lungs. As height increases, the lungs’ weight increases as well, which means taller people tend to have heavier lungs. In the meantime, trend lines for the manner of death***(drown or hung)*** are both positive and seems to parallel**(we can test this by fitting an interaction term later)**; in that case, it means that the height effect appears the same regardless of the manner of death.

However, it is evident that people who are the same height are more likely to die from drowning than hanging. No wonder that people who died from drowning will have larger lungs. The reason is that it has water on them.

### 1b

plot(Wt~Ht, type="n", data=PPW.df)   
text(Wt~Ht, labels=substr(How,1,1),  
 col=ifelse(How=="Drown","blue", "brown"),  
 data=PPW.df, cex=0.9, font=2)  
text(x = 196, y = 1125, labels = 'p', cex = 1.5, col = 'red')

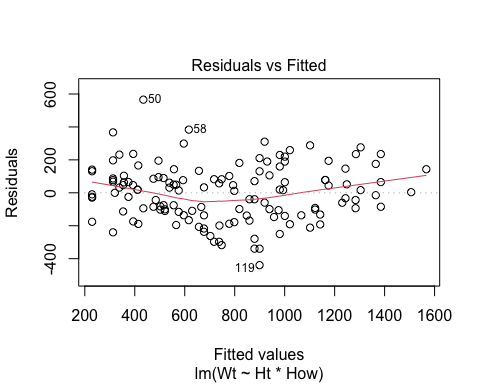


Fitting a point with height(195cm) and lungs weight(1125g). It is pretty hard to distinguish which group this point belongs to at first glance, as it lies in the middle of two different manners of death.

This point relates to prosecution and defense counsels’ is that as a prosecution role, they will emphasize that point P died from drowning even though it is relatively more minor than the same height. On the contrary, defense counsel will insist that point p died by hanging since this point is consistent with the trend line of hung.

### 1c

inter <- lm(Wt~Ht\*How, data = PPW.df)  
plot(inter, which = 1)



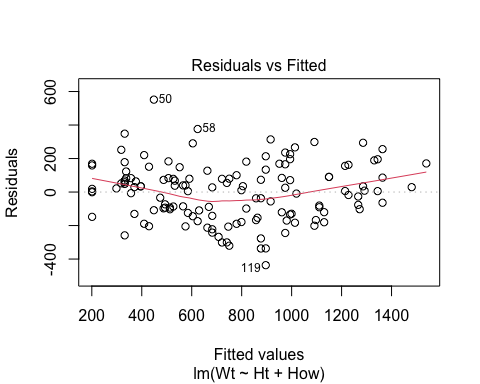
anova(inter)

## Analysis of Variance Table  
##   
## Response: Wt  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Ht 1 13287302 13287302 450.6024 < 2.2e-16 \*\*\*  
## How 1 1510744 1510744 51.2327 5.388e-11 \*\*\*  
## Ht:How 1 34566 34566 1.1722 0.2809   
## Residuals 130 3833422 29488   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

summary(inter)

##   
## Call:  
## lm(formula = Wt ~ Ht \* How, data = PPW.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -439.46 -112.64 15.09 100.18 565.72   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2619.759 193.812 -13.517 <2e-16 \*\*\*  
## Ht 20.225 1.136 17.812 <2e-16 \*\*\*  
## HowHung 118.211 321.213 0.368 0.713   
## Ht:HowHung -2.021 1.866 -1.083 0.281   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 171.7 on 130 degrees of freedom  
## Multiple R-squared: 0.7946, Adjusted R-squared: 0.7899   
## F-statistic: 167.7 on 3 and 130 DF, p-value: < 2.2e-16

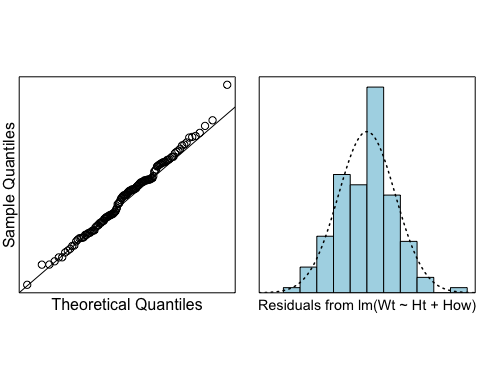
no\_inter <- lm(Wt~Ht+How, data = PPW.df)  
plot(no\_inter, which = 1)



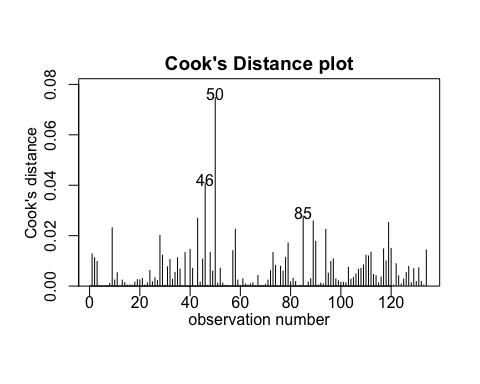
anova(no\_inter)

## Analysis of Variance Table  
##   
## Response: Wt  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Ht 1 13287302 13287302 450.011 < 2.2e-16 \*\*\*  
## How 1 1510744 1510744 51.166 5.382e-11 \*\*\*  
## Residuals 131 3867989 29527   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

normcheck(no\_inter)



cooks20x(no\_inter)



summary(no\_inter)

##   
## Call:  
## lm(formula = Wt ~ Ht + How, data = PPW.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -436.43 -123.78 14.92 90.36 551.55   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2492.6416 154.3037 -16.154 < 2e-16 \*\*\*  
## Ht 19.4774 0.9018 21.599 < 2e-16 \*\*\*  
## HowHung -227.8527 31.8541 -7.153 5.38e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 171.8 on 131 degrees of freedom  
## Multiple R-squared: 0.7928, Adjusted R-squared: 0.7896   
## F-statistic: 250.6 on 2 and 131 DF, p-value: < 2.2e-16

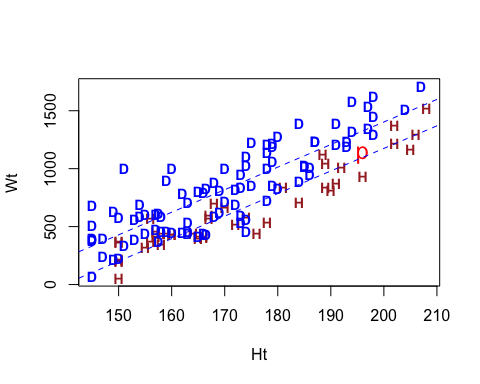
Method and Assumption checks: To begin with a linear model fitted the model with explanatory variables height, how, and their interaction. However, the interaction term was not significant(p-value = 0.2809). Therefore, the model was refitted with the interaction term removed.

For assumption check, there was a curvature in the residual plot, which means the quadratic term can be implemented, whereas most of the data were reasonably constant, which was still able to stick to linear. Most of the data points were along with the line in QQ plot and histogram was normally distributed; also, there were no points of undue influence. Hence, all model assumptions were satisfied.

Our final model is Wti = β0 + β1 × Hti + β2 × HowHungi + εi, where HowHungi set to one if person i died by hanging, otherwise it’s zero, and εi∼iid N(0,σ2).

### 1d

n <- nrow(PPW.df)  
new.df <- PPW.df[c(1:n, n+1), ]  
new.df[n + 1, c('Ht', 'Wt')] = c(196, 1125)  
  
no\_inter\_new <- lm(Wt~Ht+How, data = new.df)  
c <- coef(no\_inter\_new)  
plot(Wt~Ht, type="n", data=PPW.df)   
text(Wt~Ht, labels=substr(How,1,1),  
 col=ifelse(How=="Drown","blue", "brown"),  
 data=PPW.df, cex=0.9, font=2)  
text(x = 196, y = 1125, labels = 'p', cex = 1.5, col = 'red')  
abline(c[1], c[2], col = 'blue', lty = 2)  
abline(c[1] + c[3], c[2], col = 'blue', lty = 2)



It is clear that two lines are going up and parallel.

As a role of the prosecution, point P likely is one of the people who died from drowning. The reason is that point P is a data that belongs to drown but with relatively large variability.

As a role of defence counsel, they may have strong evidence against prosecution as point P is perfectly matched on the line(for hung). Hence, they could declare that point P died by hanging.

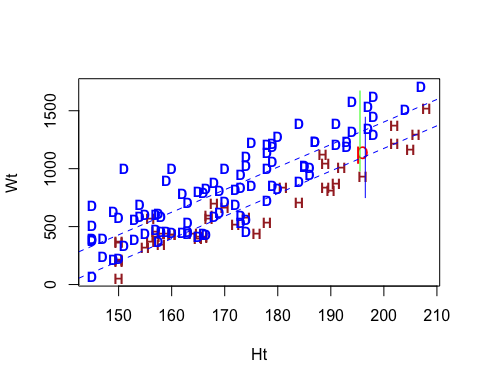
### 1e

preds.df <- data.frame(How = c('Drown', 'Hung'), Ht = 196)  
(preds <- predict(no\_inter\_new, preds.df, interval = 'prediction'))

## fit lwr upr  
## 1 1324.934 979.9984 1669.869  
## 2 1097.081 750.5768 1443.585

### 1f

plot(Wt~Ht, type="n", data=PPW.df)   
text(Wt~Ht, labels=substr(How,1,1),  
 col=ifelse(How=="Drown","blue", "brown"),  
 data=PPW.df, cex=0.9, font=2)  
text(x = 196, y = 1125, labels = 'p', cex = 1.5, col = 'red')  
abline(c[1], c[2], col = 'blue', lty = 2)  
abline(c[1] + c[3], c[2], col = 'blue', lty = 2)  
segments(196-.5,preds[1,2], 196-.5, preds[1,3], col="green")  
segments(196+.5,preds[2,2], 196+.5, preds[2,3], col="blue")



### 1g

For a person whose height is 196cm and weight 1125g died from drowning, we predict the weight of this person’s lungs is between 980.0 to 1670.0(gram), as a role of the prosecution, it is an acceptable value as this point is somewhere in the prediction interval***(green line. However)***, if this person died by hanging, we predict the weight of lungs is between 750.6 to 1443.6(gram), as a role of defence counsels, it is can be an evidence since this point is in the middle of the prediction interval***(blue line)***.

### 1h

glm\_fit <- glm(Wt~Ht+How, data = new.df, family = 'gaussian')  
RSS <- sum(resid(glm\_fit)^2); RSS

## [1] 3867989

R\_squared <- (glm\_fit$null.deviance - RSS) / glm\_fit$null.deviance; R\_squared

## [1] 0.7927793

## Question 2

### 2a

PPW.df$yesno <- with(PPW.df, ifelse(How=='Hung', 1, 0))  
bin <- glm(yesno~Wt+Ht, data = PPW.df, family = 'binomial'); bin

##   
## Call: glm(formula = yesno ~ Wt + Ht, family = "binomial", data = PPW.df)  
##   
## Coefficients:  
## (Intercept) Wt Ht   
## -25.465397 -0.008898 0.181604   
##   
## Degrees of Freedom: 133 Total (i.e. Null); 131 Residual  
## Null Deviance: 168.2   
## Residual Deviance: 120.7 AIC: 126.7

### 2b

pred.interval <- predict(bin, data.frame(Ht = 196, Wt = 1125), se.fit = T)  
lower <- pred.interval$fit-1.96\*pred.interval$se.fit  
upper <- pred.interval$fit+1.96\*pred.interval$se.fit  
ci <- cbind(lower, upper)  
round(exp(ci) / (1+exp(ci)), digits = 3)

## lower upper  
## 1 0.325 0.725

### 2c

We predict that the probability of PPW with height is 196cm and weight 1125g, which died by hanging is between 32.5% to 72.5% on average.

As a role of the prosecution, the confident interval of probability for point P is partly greater than 50%. It is not plausible to prove that PPW died by hanging.

As a role of defence counsel, more than half of the confident interval of probability is over 50%. It is reasonable to trust that PPW died by hanging.

### 2d

Since the confident interval of the probability that PPW died by hanging is between 32.5% to 72.5%, that is, it is somehow strong evidence against that PPW died from drowning, meanwhile, we can back to the plot we were refitted the point P and obverse that P is probably died by hanging, in this case, we possibly assume an accused is innocent.