# Problem 1

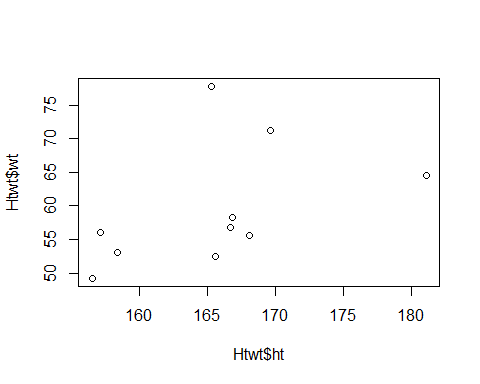
**Height and weight data** (Data file: Htwt) The table below and the the data file give ht = height in centimeters and wt = weight in kilograms for a sample of 18-year-old girls. The interest is in predicting weight from height.

library(alr4)  
data(Htwt)  
  
Htwt

## ht wt  
## 1 169.6 71.2  
## 2 166.8 58.2  
## 3 157.1 56.0  
## 4 181.1 64.5  
## 5 158.4 53.0  
## 6 165.6 52.4  
## 7 166.7 56.8  
## 8 156.5 49.2  
## 9 168.1 55.6  
## 10 165.3 77.8

1. Draw a scatterplot of wt on the vertical axis versus ht on the horizontal axis. On the basis of this plot, does a simple linear model make sense for these data? Why or why not?

plot(Htwt$wt~Htwt$ht)

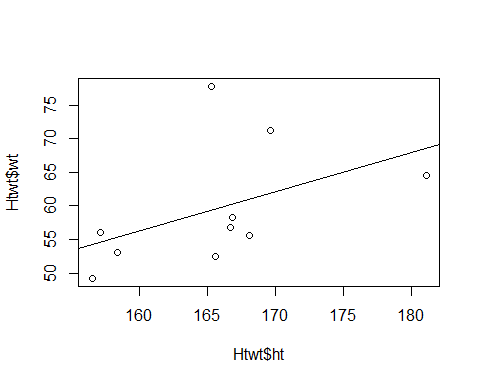


1. Compute estimates of the slope and the intercept for the regression of on . Draw the fitted line on your scatterplot.

lm\_Htwt<-lm(wt~ht, data=Htwt)  
lm\_Htwt

##   
## Call:  
## lm(formula = wt ~ ht, data = Htwt)  
##   
## Coefficients:  
## (Intercept) ht   
## -36.8759 0.5821

plot(Htwt$wt~Htwt$ht)  
abline(lm(Htwt$wt~Htwt$ht))



1. Interpret the parameter estimates and . Obtain the -tests for the hypotheses that and and -values using two-sided tests. What is your conclusion based on the -values.

summary(lm\_Htwt)

##   
## Call:  
## lm(formula = wt ~ ht, data = Htwt)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.1166 -4.7744 -2.8412 0.5696 18.4581   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -36.8759 64.4728 -0.572 0.583  
## ht 0.5821 0.3892 1.496 0.173  
##   
## Residual standard error: 8.456 on 8 degrees of freedom  
## Multiple R-squared: 0.2185, Adjusted R-squared: 0.1208   
## F-statistic: 2.237 on 1 and 8 DF, p-value: 0.1731

1. Obtain and adjusted . What can you say about the them?

summary(lm\_Htwt)$r.squared

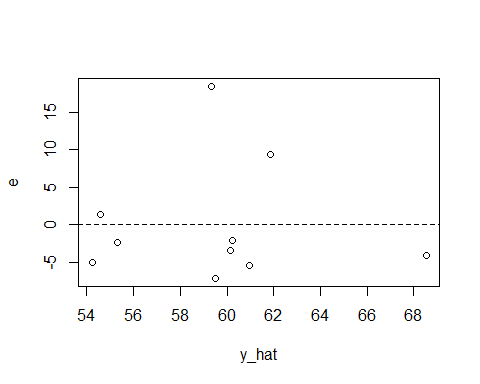
## [1] 0.2185191

summary(lm\_Htwt)$adj.r.squared

## [1] 0.120834

1. Check all the model assumptions for this simple linear regression.

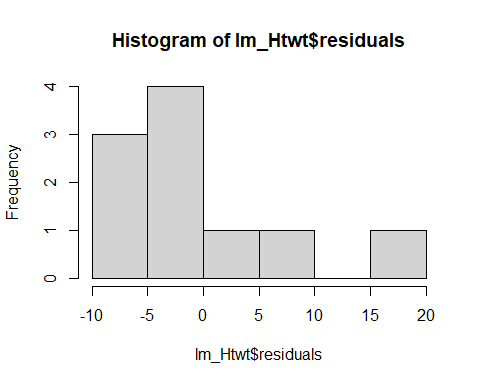
#Linearity and independent of error  
e<-lm\_Htwt$residuals  
y\_hat <- lm\_Htwt$fitted.values  
  
plot(y\_hat, e)  
abline(h=0, lty=2)



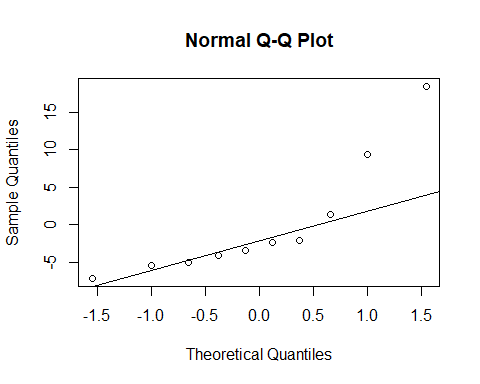
#H\_0: Independence of Error  
#H\_A: The correlation exists  
durbinWatsonTest(lm\_Htwt)

## lag Autocorrelation D-W Statistic p-value  
## 1 -0.05772706 1.366839 0.398  
## Alternative hypothesis: rho != 0

# Normality  
hist(lm\_Htwt$residuals)



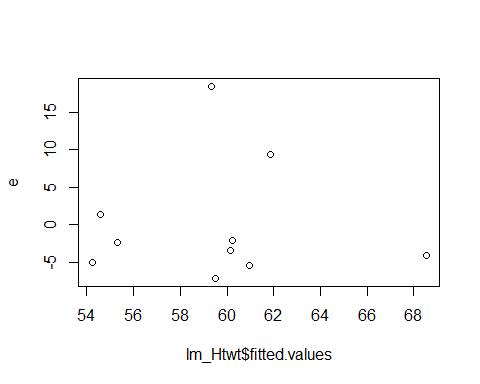
qqnorm(lm\_Htwt$residuals)  
qqline(lm\_Htwt$residuals)



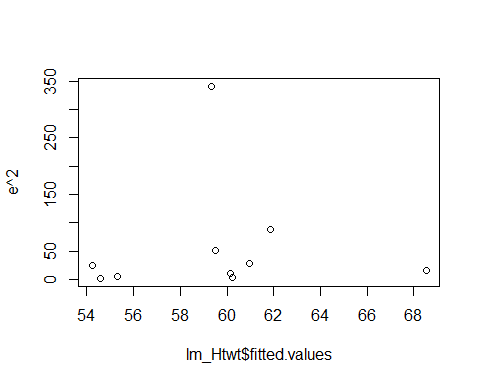
#H\_0: The data is normal  
#H\_A: The data is not normal   
shapiro.test(lm\_Htwt$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: lm\_Htwt$residuals  
## W = 0.78496, p-value = 0.009514

# Homoschedasticity   
plot(lm\_Htwt$fitted.values, e)



plot(lm\_Htwt$fitted.values, e^2)



#H\_0: Homoscedasticity holds  
#H\_A: The variance is not constant  
library(car)  
ncvTest(lm\_Htwt)

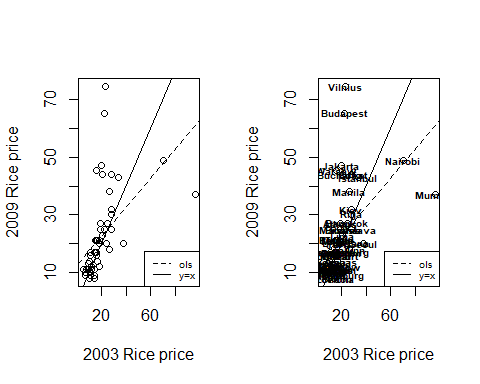
## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 0.03970496, Df = 1, p = 0.84206

# Problem 2

(Data file: UBSprices) The international bank UBS regularly produces a report (UBS, 2009) on prices and earnings in major cities throughout the world. Three of the measures they include are prices of basic commodities, namely 1kg of rice, a 1kg loaf of bread, and the price of a Big Mac hamburger at McDonalds. An interesting feature of the prices they report is that prices are measured in the minutes of labor required for a “typical” worker in that location to earn enough money to purchase the commodity. Using minutes of labor corrects at least in part for currency fluctuations, prevailing wage rates, and local prices. The data file includes measurements for rice, bread, and Big Mac prices from the 2003 and the 2009 reports. The year 2003 was before the major recession hit much of the world around 2006, and the year 2009 may reflect changes in prices due to the recession. The figure below is the plot of y = rice2009 versus x = rice2003, the price of rice in 2009 and 2003, respectively, with the cities corresponding to a few of the points marked.

library(alr4)  
data("UBSprices")

par(mfrow=c(1,2))  
plot(x=UBSprices$rice2003, y=UBSprices$rice2009,  
 xlab="2003 Rice price",   
 ylab="2009 Rice price")  
#identify(x=UBSprices$rice2003, y=UBSprices$rice2009,   
# labels=row.names(UBSprices), n=5)  
abline(lm(rice2009~rice2003, data=UBSprices), lty=2)  
abline(a=0, b=1, lty=1)  
legend("bottomright", legend=c("ols", "y=x"), lty=2:1, cex=0.6)  
  
plot(x=UBSprices$rice2003, y=UBSprices$rice2009,  
 xlab="2003 Rice price",   
 ylab="2009 Rice price")  
text(x=UBSprices$rice2003, y=UBSprices$rice2009,  
 labels=row.names(UBSprices), cex=0.6, font=2)  
abline(lm(rice2009~rice2003, data=UBSprices), lty=2)  
abline(a=0, b=1, lty=1)  
legend("bottomright", legend=c("ols", "y=x"), lty=2:1, cex=0.6)



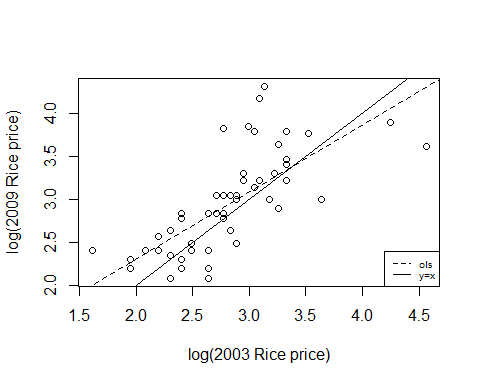
par(mfrow=c(1,1))

1. The line with equation is shown on this plot as the solid line. What is the key difference between points above this line and points below the line?
2. Which city had the largest increase in rice price? Which had the largest decrease in rice price?
3. The ols line is shown on the figure as a dashed line, and evidently . Does this suggest that prices are lower in 2009 than in 2003? Explain your answer.
4. Give two reasons why fitting simple linear regression to the figure in this problem is not likely to be appropriate.

# Problem 3

(Data file: UBSprices) This is a continuation of Problem 2. An alternative representation of the data used in the last problem is to use log scales, as in the following figure:

plot(x=log(UBSprices$rice2003), y=log(UBSprices$rice2009),  
 xlab="log(2003 Rice price)",   
 ylab="log(2009 Rice price)")  
  
abline(lm(log(rice2009)~log(rice2003), data=UBSprices), lty=2)  
abline(a=0, b=1, lty=1)  
legend("bottomright", legend=c("ols", "y=x"), lty=2:1, cex=0.6)



1. Explain why this graph and the graph in Problem 2 suggests that using log-scale is preferable if fitting simple linear regression is desired.
2. Suppose we start with a proposed model

This is a common model in many areas of study. Examples include allometry (Gould, 1966), where x could represent the size of one body characteristic such as total weight and y represents some other body characteristic, such as brain weight, psychophysics (Stevens, 1966), in which x is a physical stimulus and y is a psychological response to it, or in economics, where x could represent inputs and y outputs, where this relationship is often called a Cobb-Douglas production function (Greene, 2003).

If we take the logs of both sides of the last equation, we get

If we approximate , and write , to the extent that the logarithm of the expectation equals the expectation of the logarithm, we have

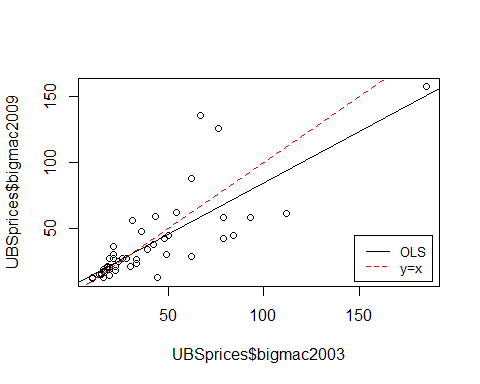
Give an interpretation of and in this setting, assuming .

# Problem 4

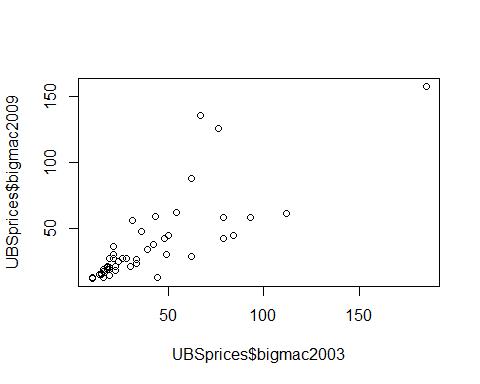
(Data file: UBSprices) This problem continues with the data file UBSprices described in Problem 2.

1. Draw the plot of y=bigmac2009 versus x=bigmac2003, the price of a Big Mac hamburger in 2009 and 2003. On this plot draw (1) the ols fitted line; (2) the line . Identify the most unusual cases and describe why they are unusual.

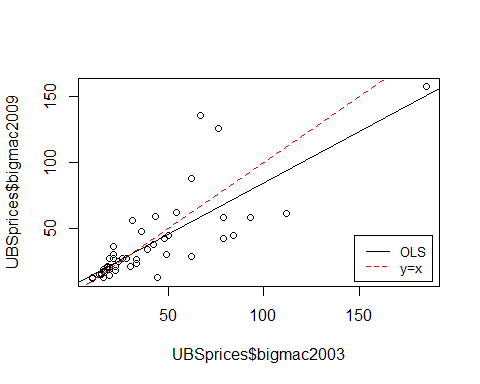
plot(UBSprices$bigmac2009~UBSprices$bigmac2003)  
abline(lm(UBSprices$bigmac2009~UBSprices$bigmac2003))  
abline(0,1, col="red", lty=2)  
legend("bottomright", inset=0.02, legend=c("OLS", "y=x"),col=c("black", "red"), lty=1:2, cex=0.8)



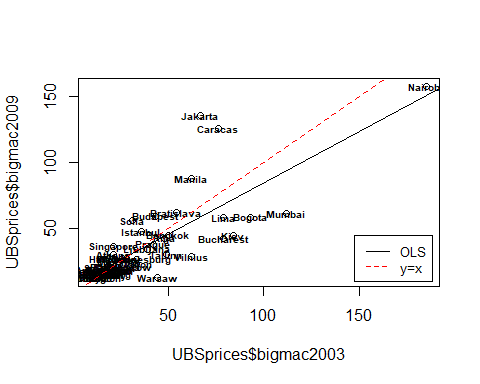
plot(UBSprices$bigmac2009~UBSprices$bigmac2003)



plot(UBSprices$bigmac2009~UBSprices$bigmac2003)  
abline(lm(UBSprices$bigmac2009~UBSprices$bigmac2003))  
abline(0,1, col="red", lty=2)  
legend("bottomright", inset=0.02, legend=c("OLS", "y=x"),  
 col=c("black", "red"), lty=1:2, cex=0.8)

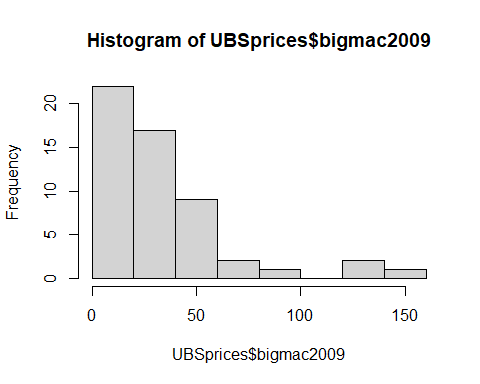


plot(UBSprices$bigmac2009~UBSprices$bigmac2003)  
#identify(x=UBSprices$bigmac2003, y=UBSprices$bigmac2009,   
# labels=row.names(UBSprices), n=3)  
text(x=UBSprices$bigmac2003, y=UBSprices$bigmac2009,  
 labels=row.names(UBSprices), cex=0.6, font=2)  
abline(lm(UBSprices$bigmac2009~UBSprices$bigmac2003))  
abline(0,1, col="red", lty=2)  
legend("bottomright", inset=0.02, legend=c("OLS", "y=x"),  
 col=c("black", "red"), lty=1:2, cex=0.8)

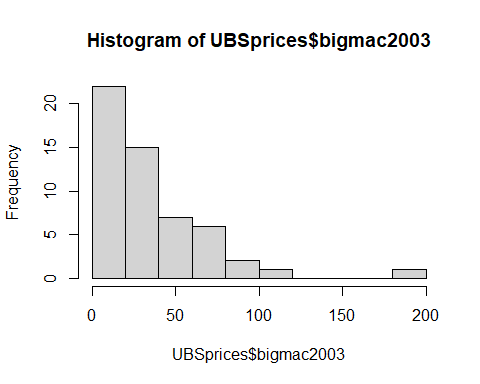


1. Give two reasons why fitting simple linear regression to the figure in this problem is not likely to be appropriate.

hist(UBSprices$bigmac2009)

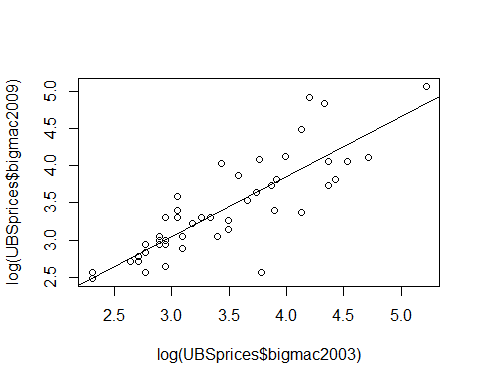


hist(UBSprices$bigmac2003)



1. Plot log(bigmac2009) versus log(bigmac2003) and explain why this graph is more sensibly summarized with a linear regression.

plot(log(UBSprices$bigmac2009)~log(UBSprices$bigmac2003))  
abline(lm(log(UBSprices$bigmac2009)~log(UBSprices$bigmac2003)))



# Problem 5

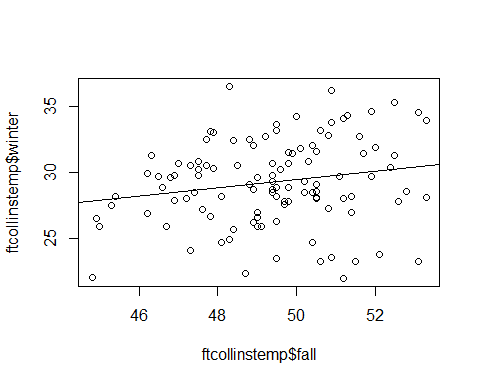
**Ft. Collins temperature data** (Data file: ftcollinstemp) The data file gives the mean temperature in the fall of each year, defined as September 1 to November 30, and the mean temperature in the following winter, defined as December 1 to the end of February in the following calendar year, in degrees Fahrenheit, for Ft. Collins, CO (Colorado Climate Center, 2012). These data cover the time period from 1900 to 2010. The question of interest is: Does the average fall temperature predict the average winter temperature?

library(alr4)  
data("ftcollinstemp")  
head(ftcollinstemp)

## year fall winter  
## 1 1900 50.2 28.5  
## 2 1901 48.1 28.2  
## 3 1902 48.1 24.7  
## 4 1903 49.9 31.4  
## 5 1904 47.8 26.7  
## 6 1905 46.9 29.8

1. Draw a scatterplot of the response versus the predictor, and describe any pattern you might see in the plot.

plot(ftcollinstemp$winter~ftcollinstemp$fall)  
abline(lm(ftcollinstemp$winter~ftcollinstemp$fall))



library(dplyr)

##   
## Attaching package: 'dplyr'

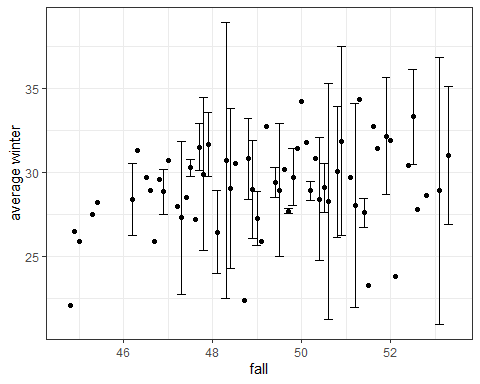
## The following object is masked from 'package:car':  
##   
## recode

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
c.temp<-ftcollinstemp %>%  
 group\_by(fall) %>%   
 summarise\_at(vars(winter), list(mean = mean, sd=sd))  
  
ggplot(c.temp, aes(fall,mean)) +   
 geom\_point() +  
 geom\_errorbar(aes(ymin=mean-sd, ymax=mean+sd), width=0.2,   
 position=position\_dodge(0.05)) +  
 labs(x="fall", y="average winter") +   
 theme\_bw()

## Warning: `position\_dodge()` requires non-overlapping x intervals



1. Use statistical software to fit the regression of the response on the predictor. Add the fitted line to your graph. Test the slope to be 0 against a two-sided alternative, and summarize your results.

lm\_temp <- lm(winter~fall, data=ftcollinstemp)  
summary(lm\_temp)

##   
## Call:  
## lm(formula = winter ~ fall, data = ftcollinstemp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.8186 -1.7837 -0.0873 2.1300 7.5896   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.7843 7.5549 1.825 0.0708 .  
## fall 0.3132 0.1528 2.049 0.0428 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.179 on 109 degrees of freedom  
## Multiple R-squared: 0.0371, Adjusted R-squared: 0.02826   
## F-statistic: 4.2 on 1 and 109 DF, p-value: 0.04284

1. Compute or obtain from your computer output the value of the variability in winter explained by fall and explain what this means.

summary(lm\_temp)$r.squared

## [1] 0.03709854

1. Divide the data into 2 time periods, an early period from 1900 to 1989, and a late period from 1990 to 2010. You can do this using the variable year in the data file. Are the results different in the two time periods?

temp1989 <- filter(ftcollinstemp, year<=1989)  
temp2010 <- filter(ftcollinstemp, year>=1990)  
nrow(temp1989)

## [1] 90

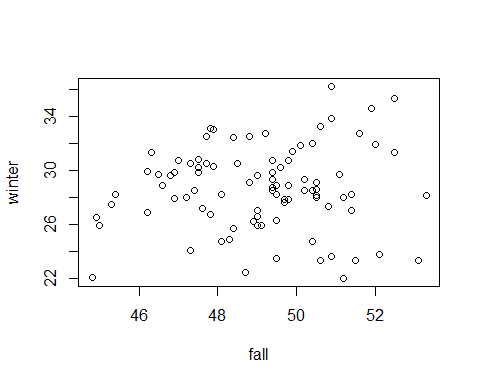
nrow(temp2010)

## [1] 21

nrow(ftcollinstemp)

## [1] 111

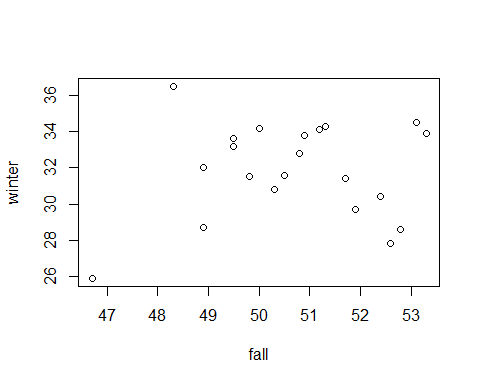
plot(winter~fall, data=temp1989)



lm\_1989<-lm(winter~fall, data=temp1989)  
summary(lm\_1989)

##   
## Call:  
## lm(formula = winter ~ fall, data = temp1989)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.8976 -1.6349 0.0118 2.0079 7.3387   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 22.7079 8.2600 2.749 0.00725 \*\*  
## fall 0.1209 0.1681 0.719 0.47397   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.057 on 88 degrees of freedom  
## Multiple R-squared: 0.005842, Adjusted R-squared: -0.005455   
## F-statistic: 0.5171 on 1 and 88 DF, p-value: 0.474

plot(winter~fall, data=temp2010)



lm\_2010<-lm(winter~fall, data=temp2010)  
summary(lm\_2010)

##   
## Call:  
## lm(formula = winter ~ fall, data = temp2010)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.4174 -1.7097 0.3768 1.8988 4.9602   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 24.8260 17.7973 1.395 0.179  
## fall 0.1390 0.3509 0.396 0.696  
##   
## Residual standard error: 2.699 on 19 degrees of freedom  
## Multiple R-squared: 0.00819, Adjusted R-squared: -0.04401   
## F-statistic: 0.1569 on 1 and 19 DF, p-value: 0.6965