Section 3.3 Comparing Two Regression Lines

Load needed packages.

```
library(Stat2Data)
library(mosaic)
library(ggplot2)
```

Create a dataframe for **ButterfliesBc** and look at the structure of the data.

```
data("ButterfliesBc")
str(ButterfliesBc)
```

```
## 'data.frame':    32 obs. of  4 variables:
##  $ Temp   : num  0.9 1.1 1.4 1.6 1.6 1.6 2.3 2.4 2.4 2.8 ...
##  $ Wing   : num  18.1 18.2 18.4 18.1 17.9 17.8 17.8 17.9 17.7 18.3 ...
##  $ Sex    : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Species: Factor w/ 1 level "Bc": 1 1 1 1 1 1 1 1 1 1 ...
```

EXAMPLE 3.9 Butterfly size: females and males

TABLE 3.2 Previous summer temperature and average wing length for male and female butterlies

```
head(ButterfliesBc)
```

```
##   Temp Wing  Sex Species
## 1  0.9 18.1 Male      Bc
## 2  1.1 18.2 Male      Bc
## 3  1.4 18.4 Male      Bc
## 4  1.6 18.1 Male      Bc
## 5  1.6 17.9 Male      Bc
## 6  1.6 17.8 Male      Bc
```

Create subsets of Male and Female butterflies

```
MaleButterflies=subset(ButterfliesBc, Sex=='Male')
FemaleButterflies=subset(ButterfliesBc, Sex=='Female')
```

EXAMPLE 3.9 FIT Separate regression lines for males and females

```
SLRmodelBcM <- lm(Wing ~ Temp, data=MaleButterflies)
SLRmodelBcM
```

```
##
## Call:
```

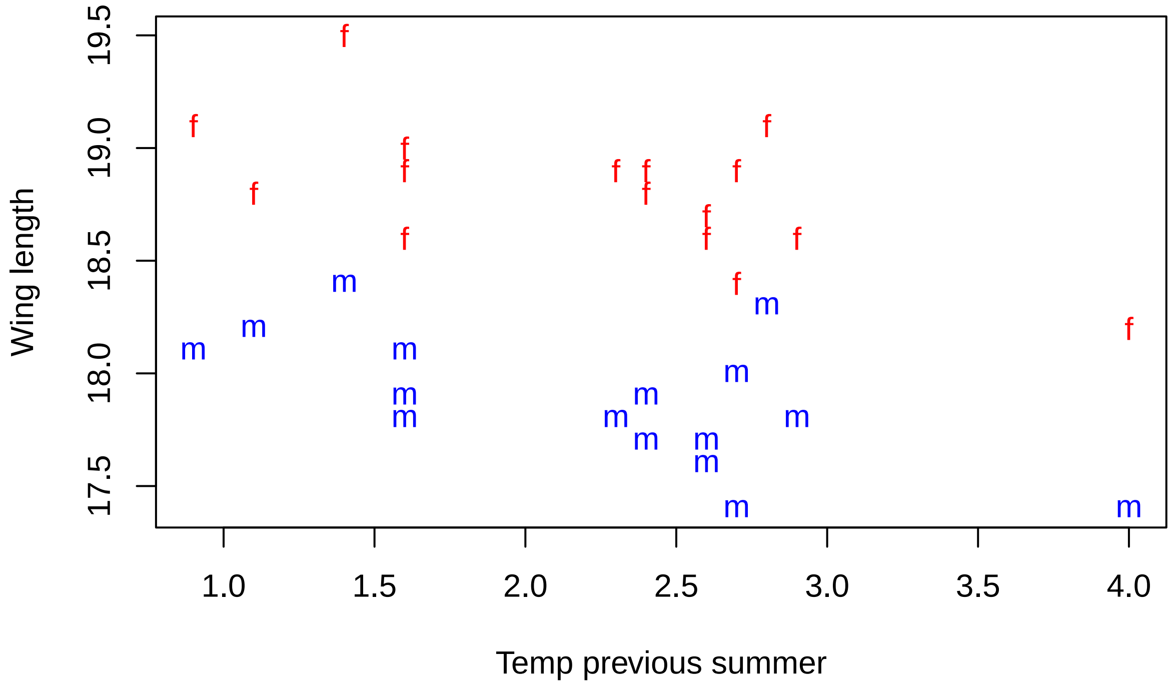
```
## lm(formula = Wing ~ Temp, data = MaleButterflies)
##
## Coefficients:
## (Intercept)         Temp
##     18.3958      -0.2313
```

```
SLRmodelBcF <- lm(Wing ~ Temp, data=FemaleButterflies)
SLRmodelBcF
```

```
##
## Call:
## lm(formula = Wing ~ Temp, data = FemaleButterflies)
##
## Coefficients:
## (Intercept)         Temp
##     19.3439      -0.2388
```

FIGURE 3.5 Scatterplot of wing length versus temperature for male and female butterflies

```
plot(Wing ~ Temp, type="n", data=ButterfliesBc, ylab="Wing length", xlab="Temp previous summer")
points(Wing ~ Temp, pch="m", col="blue", data=MaleButterflies)
points(Wing ~ Temp, pch="f", col="red", data=FemaleButterflies)
```



EXAMPLE 3.9 CHOOSE and FIT

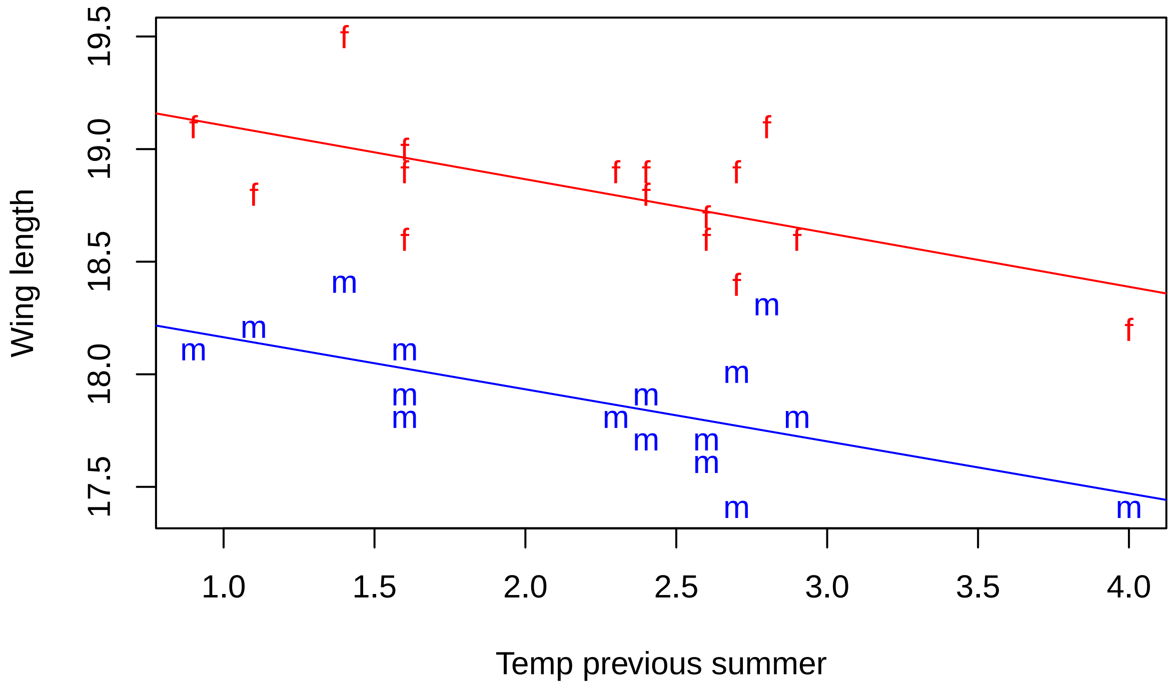
Note the trick below to create an indicator variable (IMale for males) in the ButterfliesBc dataframe.

```
ButterfliesBc$IMale=(ButterfliesBc$Sex=="Male")*1
TwoLinesmodelBc=lm(Wing~Temp+IMale,data=ButterfliesBc)
TwoLinesmodelBc
```

```
##
## Call:
## lm(formula = Wing ~ Temp + IMale, data = ButterfliesBc)
##
## Coefficients:
## (Intercept)         Temp        IMale
##     19.3355      -0.2350      -0.9313
```

FIGURE 3.6 Separate regression lines for male and female butterflies

```
plot(Wing ~ Temp, type="n", data=ButterfliesBc, ylab="Wing length", xlab="Temp previous summer")
points(Wing ~ Temp, pch="m", col="blue", data=filter(ButterfliesBc, Sex=="Male"))
points(Wing ~ Temp, pch="f", col="red", data=filter(ButterfliesBc, Sex=="Female"))
abline(SLRmodelBcM, col='blue')
abline(SLRmodelBcF, col='red')
```



EXAMPLE 3.9 ASSESS

```
summary(TwoLinesmodelBc)
```

```
##
```

```
## Call:
## lm(formula = Wing ~ Temp + IMale, data = ButterfliesBc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36961 -0.13114 -0.03910  0.08433  0.55390
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.33547    0.13372  144.60  < 2e-16 ***
## Temp        -0.23504    0.05391   -4.36  0.00015 ***
## IMale       -0.93125    0.08356  -11.14 5.35e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2364 on 29 degrees of freedom
## Multiple R-squared:  0.8316, Adjusted R-squared:   0.82
## F-statistic:  71.6 on 2 and 29 DF,  p-value: 6.059e-12
```

If we fit a simple linear regression model, ignoring sex of the butterfly.

```
SLRmodelBc <- lm(Wing ~ Temp, data=ButterfliesBc)
summary(SLRmodelBc)
```

```
##
## Call:
## lm(formula = Wing ~ Temp, data = ButterfliesBc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.83523 -0.52935  0.09725  0.44137  0.95922
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.8698     0.2870  65.740   <2e-16 ***
## Temp         -0.2350     0.1218  -1.929   0.0632 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.5341 on 30 degrees of freedom
## Multiple R-squared:  0.1104, Adjusted R-squared:  0.08072
## F-statistic: 3.722 on 1 and 30 DF,  p-value: 0.0632
```

Output from fitting a simple linear regression model for only male butterflies

```
summary(SLRmodelBcM)
```

```
##
## Call:
## lm(formula = Wing ~ Temp, data = MaleButterflies)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max 
## -0.37140 -0.12954 -0.06733  0.07437  0.55173 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 18.39582    0.17829 103.180  < 2e-16 ***
## Temp        -0.23127    0.07567  -3.056  0.00854 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2346 on 14 degrees of freedom
## Multiple R-squared:  0.4002, Adjusted R-squared:  0.3573 
## F-statistic: 9.341 on 1 and 14 DF,  p-value: 0.008543
```

Output from fitting a simple linear regression model for only female butterflies

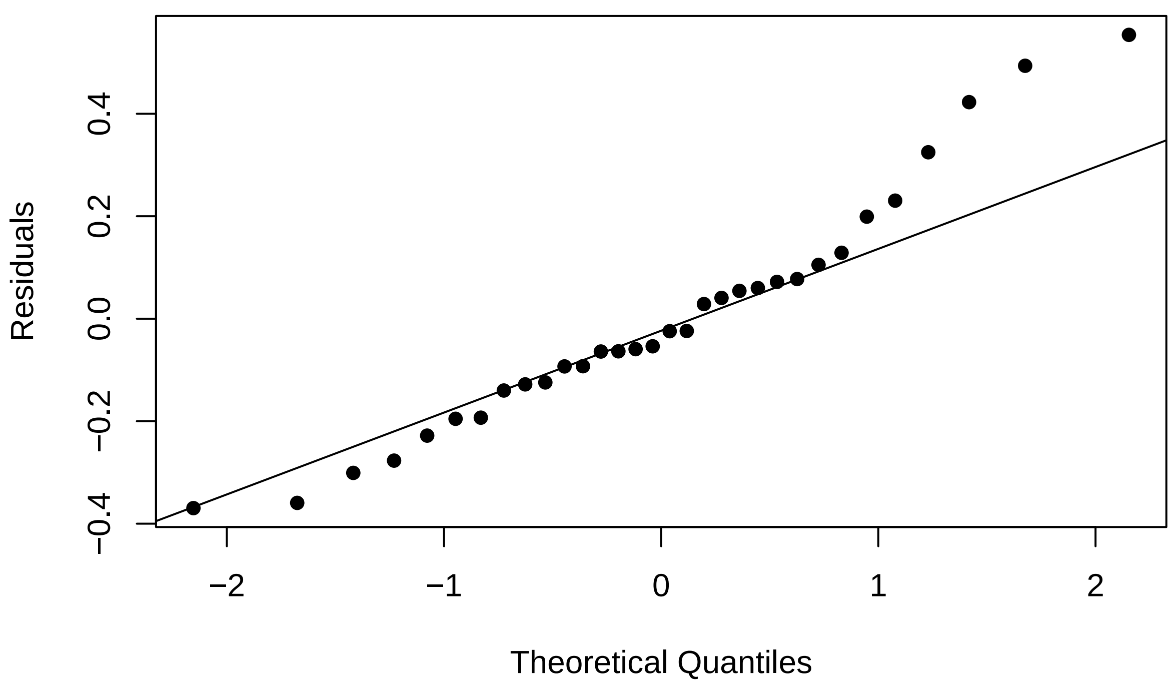
```
summary(SLRmodelBcF)
```

```
## 
## Call:
## lm(formula = Wing ~ Temp, data = FemaleButterflies)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.36176 -0.13936 -0.02594  0.11138  0.49048 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 19.34386    0.18721 103.326  < 2e-16 ***
## Temp        -0.23881    0.07946  -3.006  0.00945 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.2463 on 14 degrees of freedom
## Multiple R-squared:  0.3922, Adjusted R-squared:  0.3488 
## F-statistic: 9.033 on 1 and 14 DF,  p-value: 0.009447
```

FIGURE 3.7 Residual plots for predicting wing length using IMale and Temperature

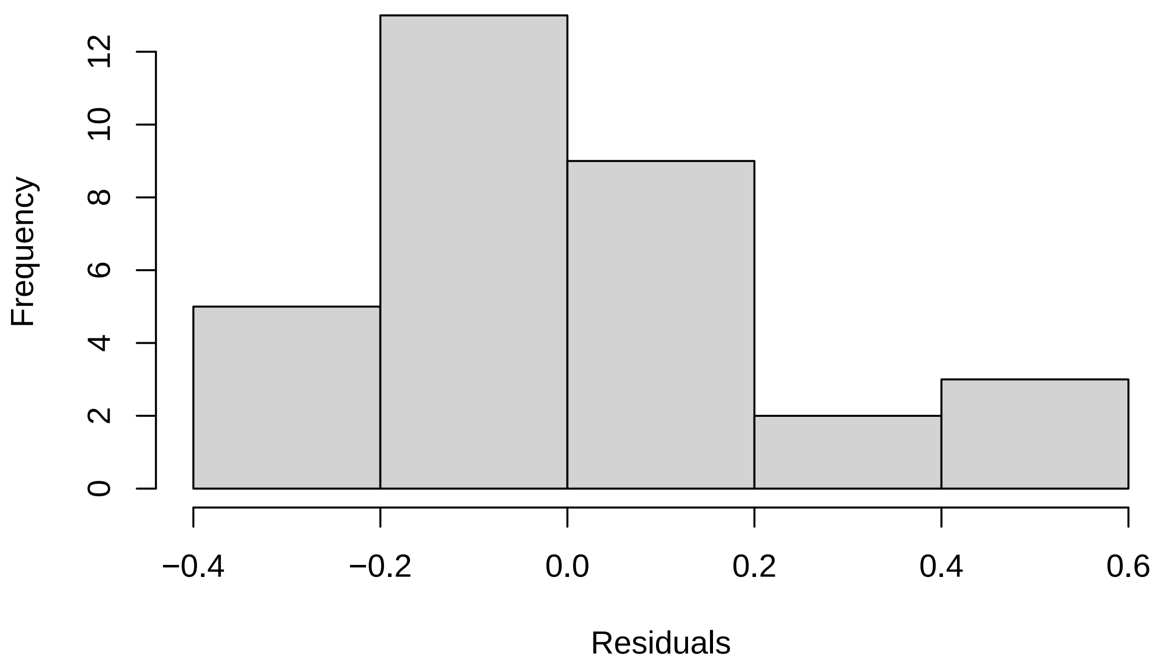
(a) Normal quantile plot of residuals

```
qqnorm(TwoLinesmodelBc$residuals, xlab="Theoretical Quantiles", ylab="Residuals",main="", pch=16)
qqline(TwoLinesmodelBc$residuals)
```



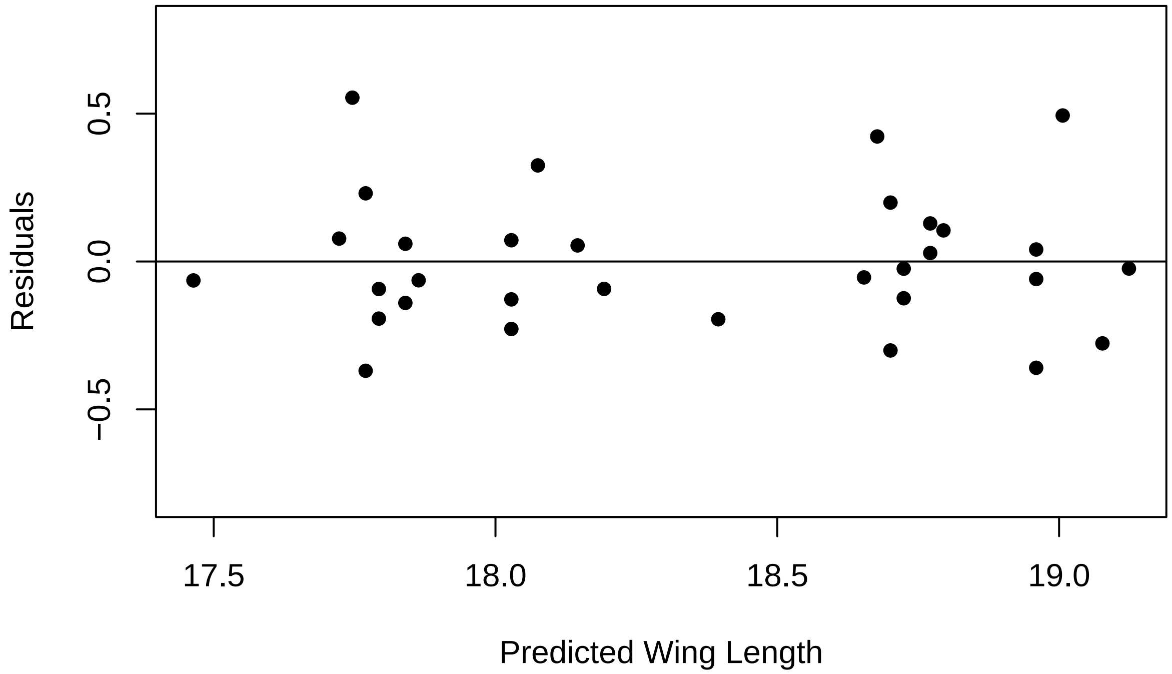
(b) Histogram of residuals

```
hist(TwoLinesmodelBc$residuals, xlab="Residuals", main="")
```



(c) Residuals versus fits

```
plot(TwoLinesmodelBc$residuals~TwoLinesmodelBc$fitted, pch=16, ylab="Residuals",xlab="Predicted Wing Len
abline(h=0)
```



EXAMPLE 3.9 USE

```
confint(TwoLinesmodelBc)
```

```
##                  2.5 %     97.5 %
## (Intercept) 19.0619800 19.6089552
## Temp        -0.3453057 -0.1247776
## IMale       -1.1021593 -0.7603407
```

EXAMPLE 3.10 Growth rates of kids

Create a dataframe for **Kids198** and look at the structure of the data.

```
data("Kids198")
str(Kids198)
```

```
## 'data.frame':    198 obs. of  5 variables:
##  $ Height: num  67.8 63 50.1 55.7 63.2 48.8 63.8 61.3 61.1 54.7 ...
##  $ Weight: int  166 93 54 69 115 52 108 89 118 80 ...
##  $ Age   : int  210 144 119 130 157 102 198 155 199 134 ...
##  $ Sex   : int  0 1 0 1 0 0 1 0 1 0 ...
##  $ Race  : int  1 0 0 0 0 0 0 0 0 0 ...
```

FIGURE 3.8 Weight versus Age by Sex for kids

Note: This version uses the qplot( ) function from the ggplot2 package. This provides another way to show group information in the scatterplot.

```
qplot(Age, Weight, color=as.factor(Sex), data=Kids198)
```

```
## Warning: ‘qplot()‘ was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

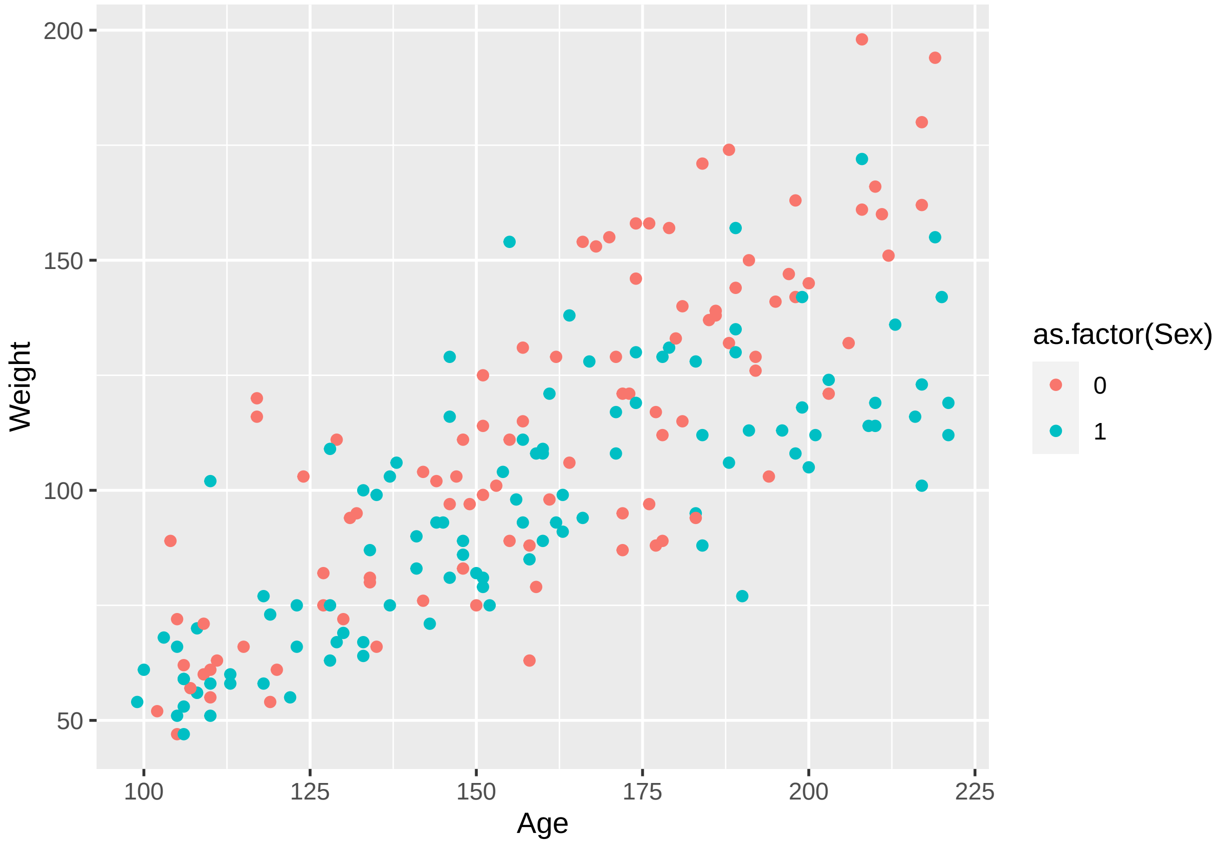


FIGURE 3.9 Separate regressions of Weight versus Age for boys and girls

```
qplot(Age, Weight, facets= ~Sex, color= as.factor(Sex), data=Kids198, geom=c("point", "lm"))
```

```
## Warning: Using the ‘size‘ aesthetic with geom_line was deprecated in ggplot2 3.4.0.
## i Please use the ‘linewidth‘ aesthetic instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

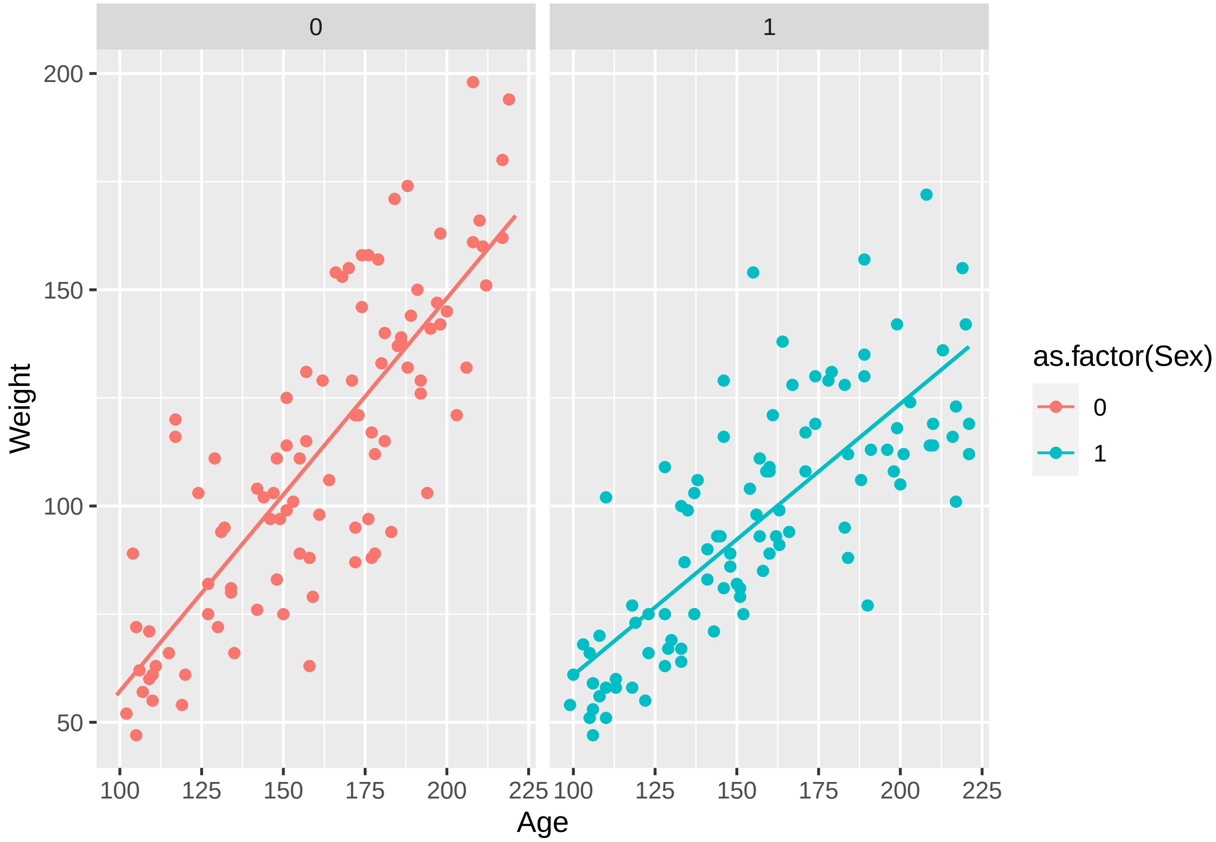
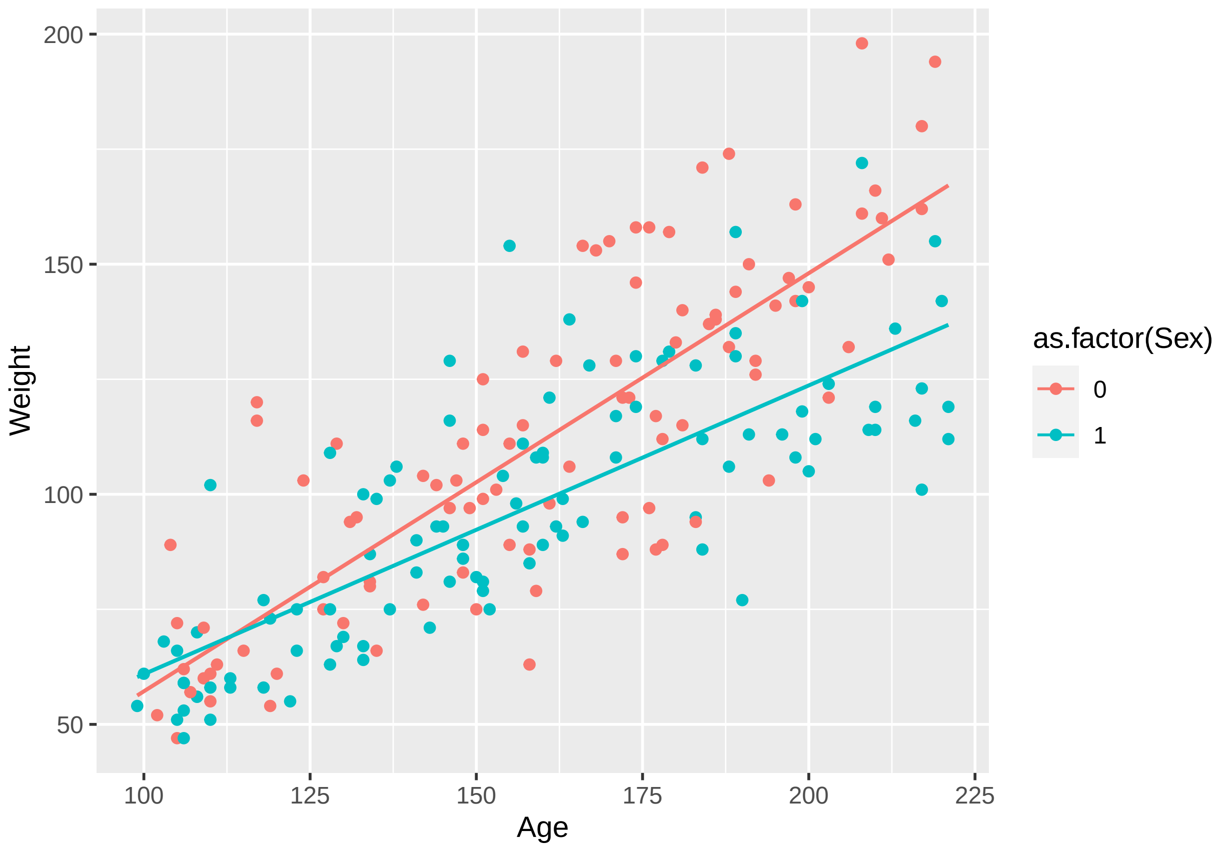


FIGURE 3.10 Compare regression lines by Sex

```
qplot(Age, Weight, color=as.factor(Sex),  data=Kids198, geom=c("point", "lm"))
```



EXAMPLE 3.10 FIT

Note: The variable Sex is already coded the same way as the variable IGirl, so we did not create another variable in the code below. The next chunk fits the multiple regression model using Age, Sex, and the product Age*Sex to predict Weight.

```
regmodelbg=lm(Weight~Age+Sex+Sex*Age, data=Kids198) #model for both boys and girls
summary(regmodelbg)
```

```
##
## Call:
## lm(formula = Weight ~ Age + Sex + Sex * Age, data = Kids198)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.884 -12.055  -2.782  10.185  58.581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -33.69254   10.00727  -3.367 0.000917 ***
## Age           0.90871    0.06106  14.882  < 2e-16 ***
## Sex          31.85057   13.24269   2.405 0.017106 *
## Age:Sex      -0.28122    0.08164  -3.445 0.000700 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 19.19 on 194 degrees of freedom
## Multiple R-squared:  0.6683, Adjusted R-squared:  0.6631
## F-statistic: 130.3 on 3 and 194 DF,  p-value: < 2.2e-16
```

Note: As discussed in the text, we can compute the separate simple regression lines for predicting Weight from Age separately for boys and girls from the multiple regression model above. To help confirm this the next chunk fits the simple regressions for subsets consisting of each sex.

```
lm(Weight~Age,data=subset(Kids198,Sex==0))   #for boys
```

```
##
## Call:
## lm(formula = Weight ~ Age, data = subset(Kids198, Sex == 0))
##
## Coefficients:
## (Intercept)          Age
##    -33.6925       0.9087
```

```
lm(Weight~Age,data=subset(Kids198,Sex==1))   #for girls
```

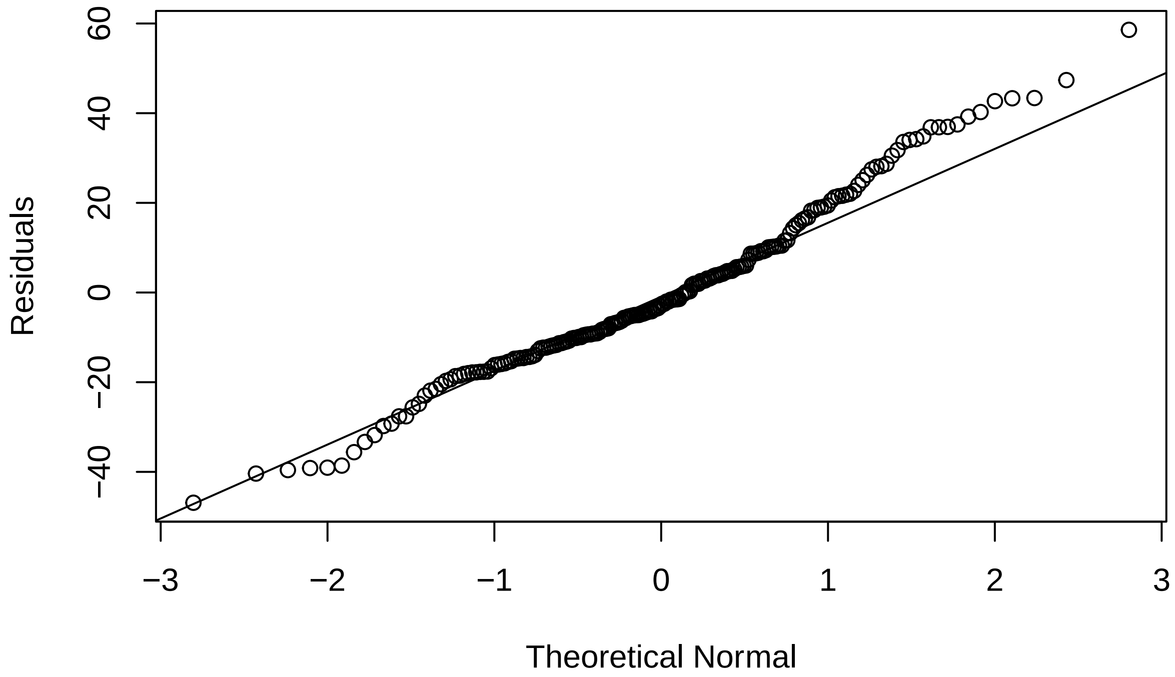
```
##
## Call:
## lm(formula = Weight ~ Age, data = subset(Kids198, Sex == 1))
##
## Coefficients:
## (Intercept)          Age
##     -1.8420       0.6275
```

EXAMPLE 3.10 ASSESS

FIGURE 3.11 Residual plots for multiple regression model for Weight based on Age and Sex

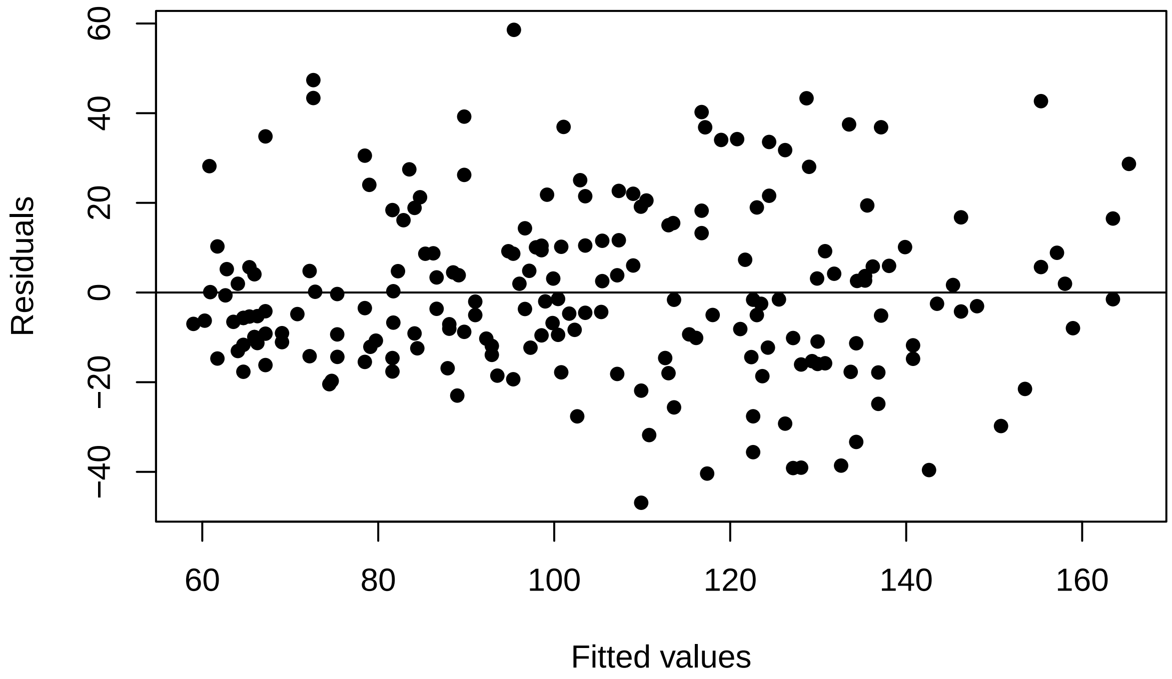
(a) Normal quantile plot (similar to the probability plot shown in the text)

```
qqnorm(regmodelbg$residuals, xlab="Theoretical Normal", ylab="Residuals", main="")
qqline(regmodelbg$residuals)
```



(b) Residuals versus fits

```
plot(regmodelbg$residuals~regmodelbg$fitted,ylab="Residuals",xlab="Fitted values", pch=16)
abline(h=0)
```



EXAMPLE 3.10 USE

```
confint(regmodelbg)
```

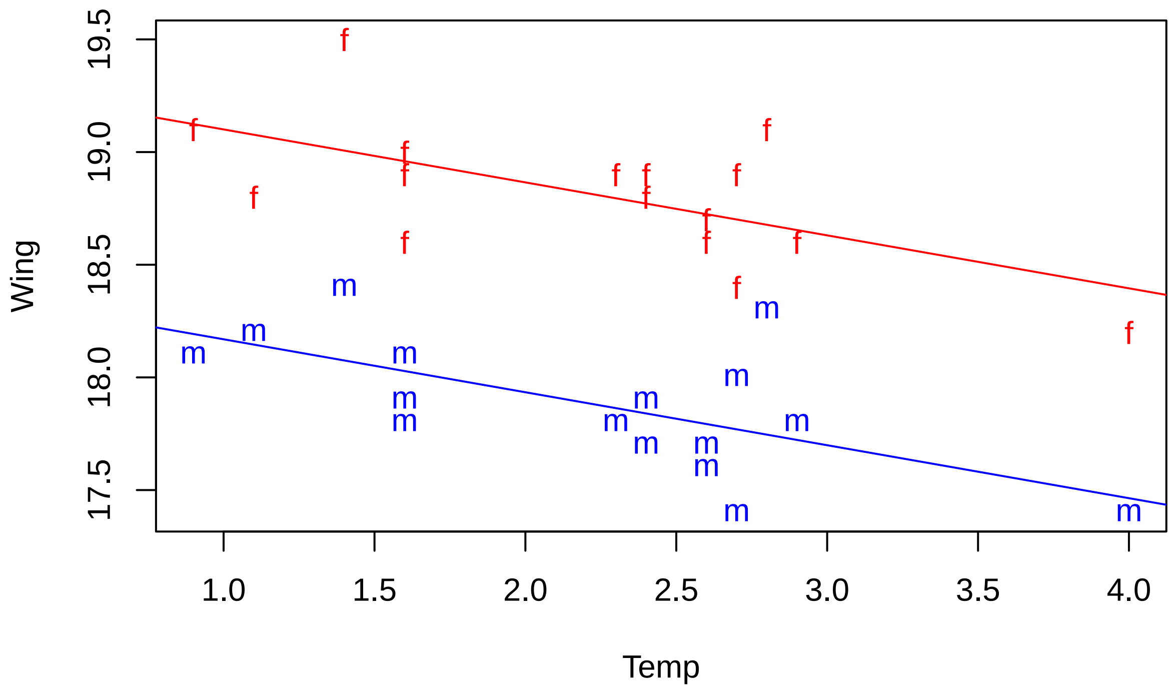
```
##                    2.5 %      97.5 %
## (Intercept) -53.4295458 -13.9555301
## Age           0.7882792   1.0291404
## Sex           5.7324387  57.9686922
## Age:Sex      -0.4422327  -0.1202112
```

Alternate Solutions

Alternate way to get Figure 3.6

This version uses the IMale indicator and ifelse( ) to control plotting symbol and color. It gets the slope and intercepts for the two lines directly from the fitted multiple regression with the indicators and requires no subsetting.

```
plot(Wing~Temp,pch=ifelse(IMale,"m","f"),col=ifelse(IMale,"blue","red"),data=ButterfliesBc)
abline(TwoLinesmodelBc$coeff[1],TwoLinesmodelBc$coeff[2],col="red")
abline(TwoLinesmodelBc$coeff[1]+TwoLinesmodelBc$coeff[3],TwoLinesmodelBc$coeff[2],col="blue")
```



Another way to approach Figures 3.9 and 3.10: create subsets for each sex.

FIGURE 3.9a Boys

```
boys=subset(Kids198,Sex==0)
plot(Weight~Age,data=boys)
abline(lm(Weight~Age,data=boys),col="blue")
```

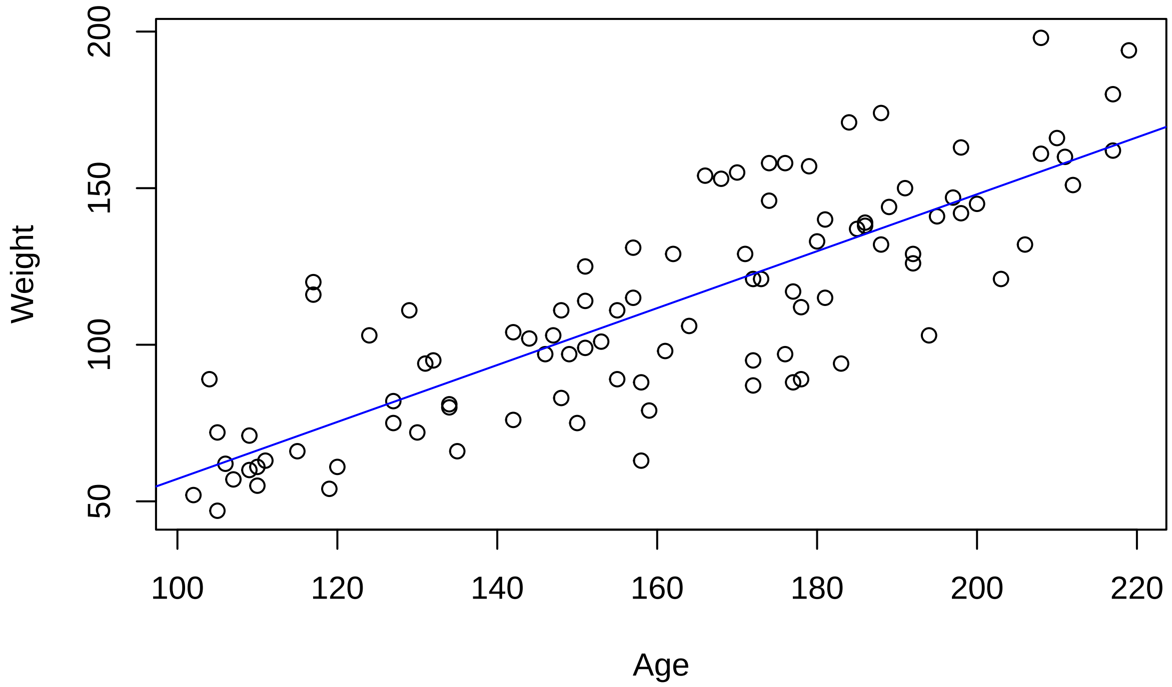


FIGURE 3.9b Girls

```
girls=subset(Kids198,Sex==1)
plot(Weight~Age,data=girls)
abline(lm(Weight~Age,data=girls),col="red")
```

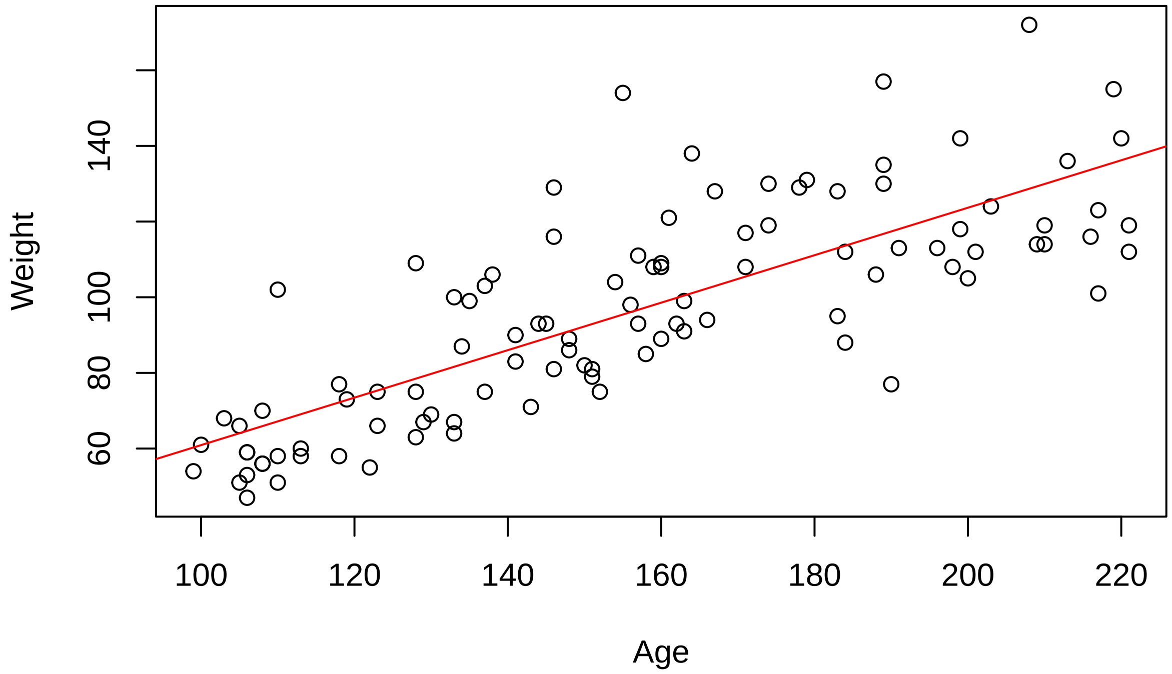
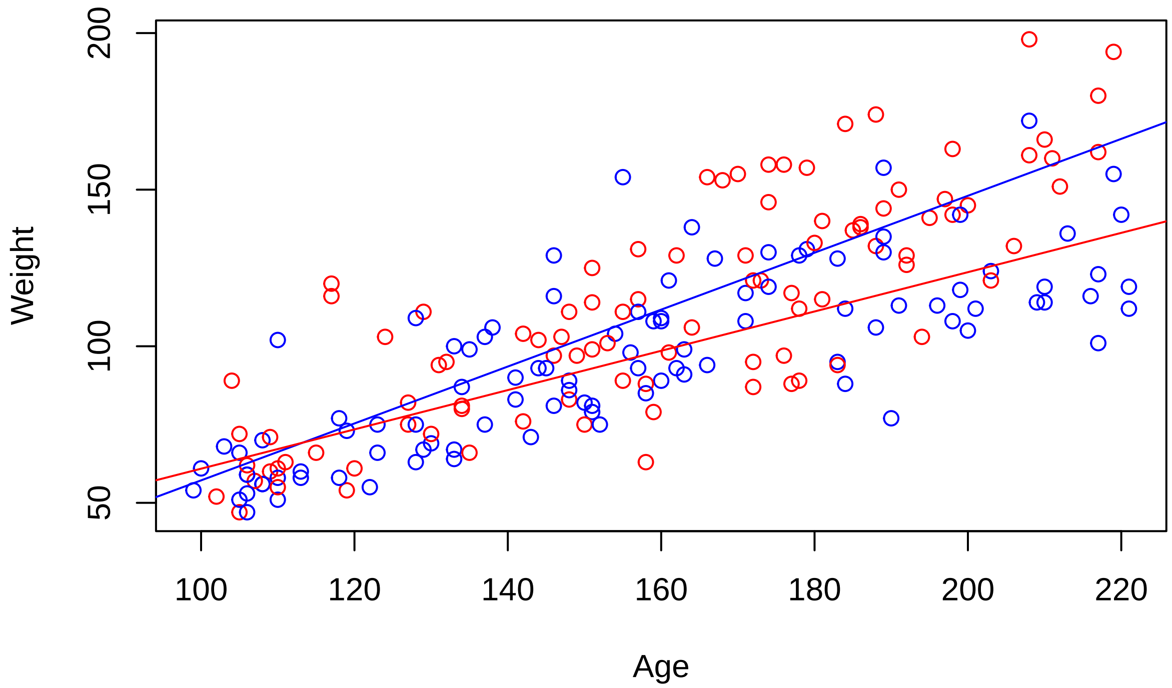


FIGURE 3.10 Compare regression lines by Sex

```
plot(Weight~Age,col=ifelse(Sex,"blue","red"),data=Kids198)
abline(lm(Weight~Age,data=subset(Kids198,Sex==0)),col="blue")
abline(lm(Weight~Age,data=subset(Kids198,Sex==1)),col="red")
```


Weight
200
150
100
50
100
120
140
160
180
200
220
Age