Section 3.7 Case Study: Predicting in Retail Clothing

Load needed packages.

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

EXAMPLE 3.21 Predicting customer spending for a clothing retailer

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

```
data("Clothing")
str(Clothing)
```

```
## 'data.frame':    60 obs. of  8 variables:
##  $ ID      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Amount  : int  0 0 0 30 33 35 35 39 40 45 ...
##  $ Recency : int  22 30 24 6 12 48 5 2 24 3 ...
##  $ Freq12  : int  0 0 0 3 1 0 5 5 0 6 ...
##  $ Dollar12: int  0 0 0 140 50 0 450 245 0 403 ...
##  $ Freq24  : int  3 0 1 4 1 0 6 12 1 8 ...
##  $ Dollar24: int  400 0 250 225 50 0 415 661 225 1138 ...
##  $ Card    : int  0 0 0 0 0 0 0 1 0 0 ...
```

TABLE 3.5 First few cases of the **Clothing** data

```
head(Clothing)
```

```
##   ID Amount Recency Freq12 Dollar12 Freq24 Dollar24 Card
## 1  1      0      22      0        0      3      400    0
## 2  2      0      30      0        0      0        0    0
## 3  3      0      24      0        0      1      250    0
## 4  4     30       6      3      140      4      225    0
## 5  5     33      12      1       50      1       50    0
## 6  6     35      48      0        0      0        0    0
```

Look at the maximum value for Amount

```
favstats(~Amount, data=Clothing)
```

```
##  min Q1 median  Q3     max     mean       sd  n missing
##    0 50     70 100 1506000 25201.07 194410.5 60       0
```

Remove the unusual observations

```
CleanClothing=subset(Clothing, Amount!=0 & Amount<1000000)
str(CleanClothing)
```

```
## 'data.frame':    56 obs. of  8 variables:
##  $ ID      : int  4 5 6 7 8 9 10 11 12 13 ...
##  $ Amount  : int  30 33 35 35 39 40 45 48 50 50 ...
##  $ Recency : int  6 12 48 5 2 24 3 6 12 5 ...
##  $ Freq12  : int  3 1 0 5 5 0 6 3 1 2 ...
##  $ Dollar12: int  140 50 0 450 245 0 403 155 42 100 ...
##  $ Freq24  : int  4 1 0 6 12 1 8 4 7 8 ...
##  $ Dollar24: int  225 50 0 415 661 225 1138 262 290 700 ...
##  $ Card    : int  0 0 0 0 1 0 0 0 0 1 ...
```

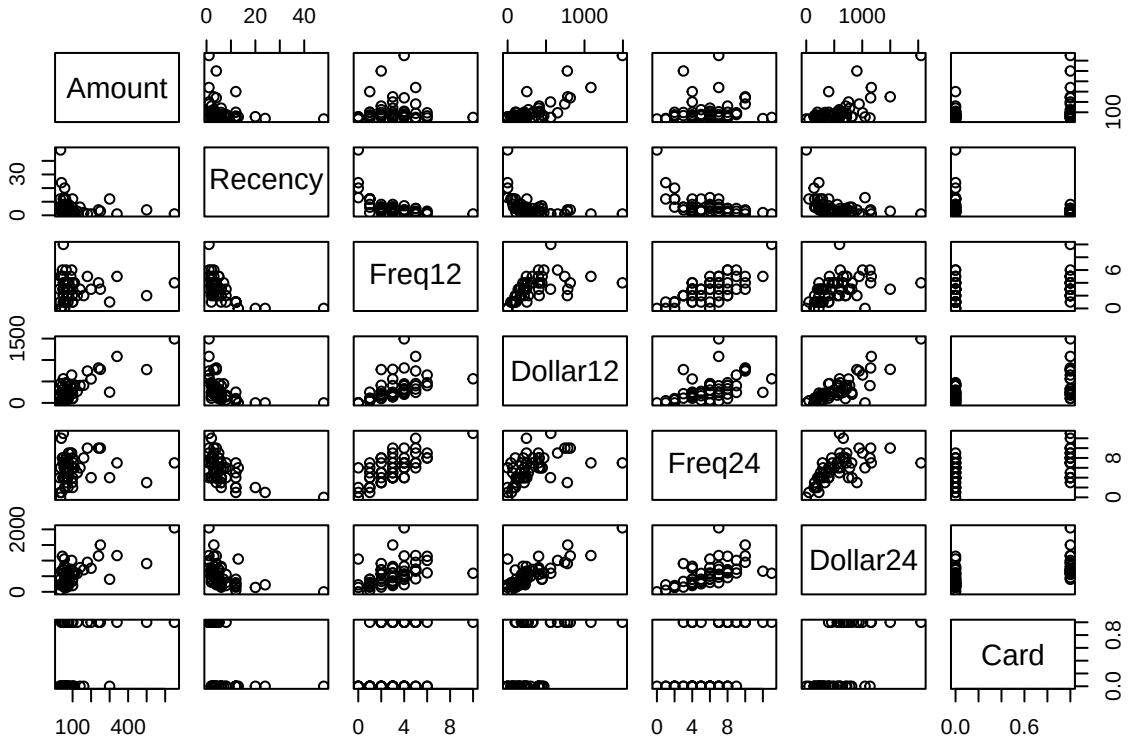
```
favstats(CleanClothing$Amount)
```

```
##  min Q1 median  Q3 max     mean       sd  n missing
##   30 50     71 100 650 108.2857 112.1884 56       0
```

EXAMPLE 3.21 CHOOSE

FIGURE 3.33 Matrix of scatterplots for variables in **Clothing**

```
pairs(CleanClothing[,2:8])
```



Create a matrix of correlation coefficients

```
round(cor(CleanClothing[,2:7]), digit=3)
```

```
##          Amount Recency Freq12 Dollar12 Freq24 Dollar24
## Amount    1.000  -0.221  0.052    0.804  0.102    0.677
## Recency  -0.221   1.000 -0.584   -0.454 -0.549   -0.432
## Freq12    0.052  -0.584  1.000    0.556  0.710    0.421
## Dollar12  0.804  -0.454  0.556    1.000  0.485    0.827
## Freq24    0.102  -0.549  0.710    0.485  1.000    0.596
## Dollar24  0.677  -0.432  0.421    0.827  0.596    1.000
```

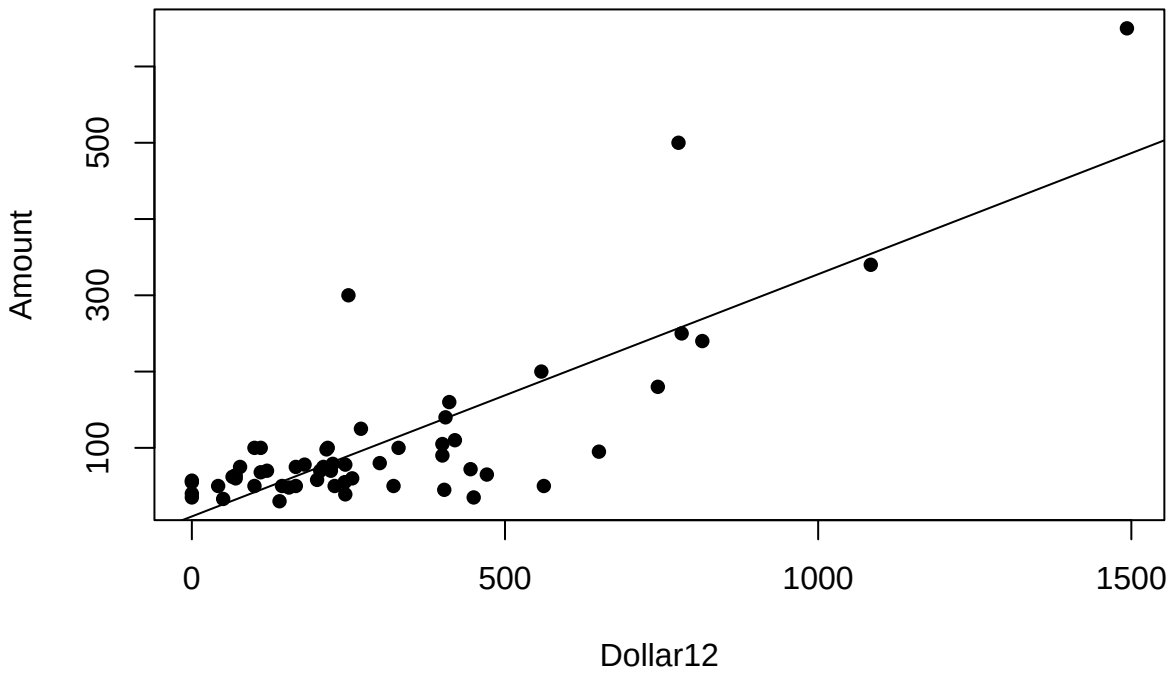
EXAMPLE 3.21 FIT and ASSESS

```
modelSLRDollar12=lm(Amount~Dollar12, data=CleanClothing)
summary(modelSLRDollar12)
```

```
##
## Call:
## lm(formula = Amount ~ Dollar12, data = CleanClothing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -138.54  -31.55   -3.85   25.34  243.18
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.0756    13.3783   0.753    0.455
## Dollar12      0.3176     0.0320   9.925 8.93e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67.37 on 54 degrees of freedom
## Multiple R-squared:  0.6459, Adjusted R-squared:  0.6393
## F-statistic:  98.5 on 1 and 54 DF,  p-value: 8.929e-14
```

FIGURE 3.34 Regression of Amount on Dollar12

```
plot(Amount~Dollar12, data=CleanClothing, pch=16)
abline(modelSLRDollar12)
```



EXAMPLE 3.21 FIT a multiple regression model

```
clothingmodel2=lm(Amount~Dollar12+Dollar24+Recency, data=CleanClothing)
summary(clothingmodel2)
```

```
##
## Call:
## lm(formula = Amount ~ Dollar12 + Dollar24 + Recency, data = CleanClothing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -126.522  -24.098    0.247   23.652  237.852
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -23.05236   21.59290  -1.068   0.2906
## Dollar12      0.32724    0.05678   5.764 4.53e-07 ***
## Dollar24      0.02151    0.04202   0.512   0.6110
## Recency       2.86718    1.37573   2.084   0.0421 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.91 on 52 degrees of freedom
## Multiple R-squared:  0.6736, Adjusted R-squared:  0.6548
## F-statistic: 35.78 on 3 and 52 DF,  p-value: 1.097e-12
```

EXAMPLE 3.21 FIT a 6-predictor model

```
Clothingmodel6=lm(Amount~Recency+Freq12+Dollar12+Freq24+Dollar24+Card,data=CleanClothing)
summary(Clothingmodel6)
```

```
##
## Call:
## lm(formula = Amount ~ Recency + Freq12 + Dollar12 + Freq24 +
##     Dollar24 + Card, data = CleanClothing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.799 -12.218  -3.334   7.299 156.822
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 104.251935  19.834341   5.256 3.20e-06 ***
## Recency      -1.345963   0.971053  -1.386    0.172
## Freq12      -32.353539   5.187870  -6.236 1.01e-07 ***
## Dollar12      0.429683   0.041325  10.398 5.43e-14 ***
## Freq24       -5.173593   3.619661  -1.429    0.159
## Dollar24      0.001756   0.031850   0.055    0.956
## Card         14.624409  14.575770   1.003    0.321
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.83 on 49 degrees of freedom
## Multiple R-squared:  0.882,  Adjusted R-squared:  0.8675
## F-statistic: 61.02 on 6 and 49 DF,  p-value: < 2.2e-16
```

EXAMPLE 3.21 FIT an 11-predictor model with quadratic terms

```
Clothingmodel11=lm(Amount~Recency+I(Recency^2)+Freq12+I(Freq12^2)+Dollar12+I(Dollar12^2)
        +Freq24+I(Freq24^2)+Dollar24+I(Dollar24^2)+Card,data=CleanClothing)
summary(Clothingmodel11)
```

```
##
## Call:
## lm(formula = Amount ~ Recency + I(Recency^2) + Freq12 + I(Freq12^2) +
##     Dollar12 + I(Dollar12^2) + Freq24 + I(Freq24^2) + Dollar24 +
##     I(Dollar24^2) + Card, data = CleanClothing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.514 -21.663  -0.463  11.707 154.280
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.418e+02  4.474e+01   3.170  0.00277 **
## Recency       -2.877e+00  3.218e+00  -0.894  0.37606
## I(Recency^2)   9.237e-03  5.977e-02   0.155  0.87788
## Freq12        -4.801e+01  1.365e+01  -3.518  0.00102 **
## I(Freq12^2)    1.870e+00  1.407e+00   1.330  0.19052
```

```
## Dollar12       3.662e-01  1.101e-01   3.326  0.00179 **
## I(Dollar12^2)  6.993e-05  9.937e-05   0.704  0.48533
## Freq24        -7.014e+00  1.267e+01  -0.554  0.58270
## I(Freq24^2)    1.015e-01  9.522e-01   0.107  0.91558
## Dollar24       4.649e-02  9.078e-02   0.512  0.61110
## I(Dollar24^2) -3.045e-05  5.598e-05  -0.544  0.58926
## Card           7.956e+00  1.747e+01   0.455  0.65106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.09 on 44 degrees of freedom
## Multiple R-squared:  0.8927, Adjusted R-squared:  0.8659
## F-statistic: 33.28 on 11 and 44 DF,  p-value: < 2.2e-16
```

EXAMPLE 3.21 CHOOSE (again)

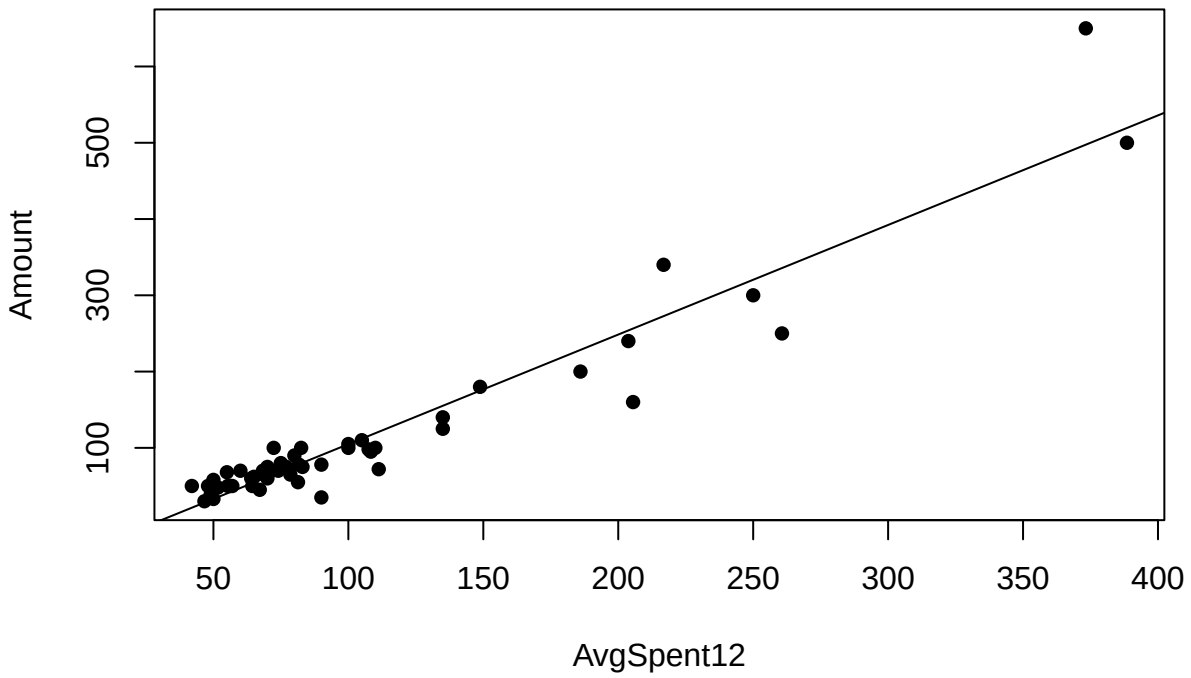
Create the average amount spent, after removing the four cases with no spending

```
Clothing3=subset(CleanClothing,Freq12>0)
Clothing3$AvgSpent12=Clothing3$Dollar12/Clothing3$Freq12
str(Clothing3)
```

```
## 'data.frame':    52 obs. of  9 variables:
##  $ ID        : int  4 5 7 8 10 11 12 13 14 15 ...
##  $ Amount    : int  30 33 35 39 45 48 50 50 50 50 ...
##  $ Recency   : int  6 12 5 2 3 6 12 5 8 1 ...
##  $ Freq12    : int  3 1 5 5 6 3 1 2 3 10 ...
##  $ Dollar12  : int  140 50 450 245 403 155 42 100 144 562 ...
##  $ Freq24    : int  4 1 6 12 8 4 7 8 4 13 ...
##  $ Dollar24  : int  225 50 415 661 1138 262 290 700 202 595 ...
##  $ Card      : int  0 0 0 1 0 0 0 1 0 1 ...
##  $ AvgSpent12: num  46.7 50 90 49 67.2 ...
```

FIGURE 3.35 Amount versus AvgSpent12 with regression line

```
ClothingmodelA12=lm(Amount~AvgSpent12, data=Clothing3)
plot(Amount~AvgSpent12, pch=16, data=Clothing3)
abline(ClothingmodelA12)
```



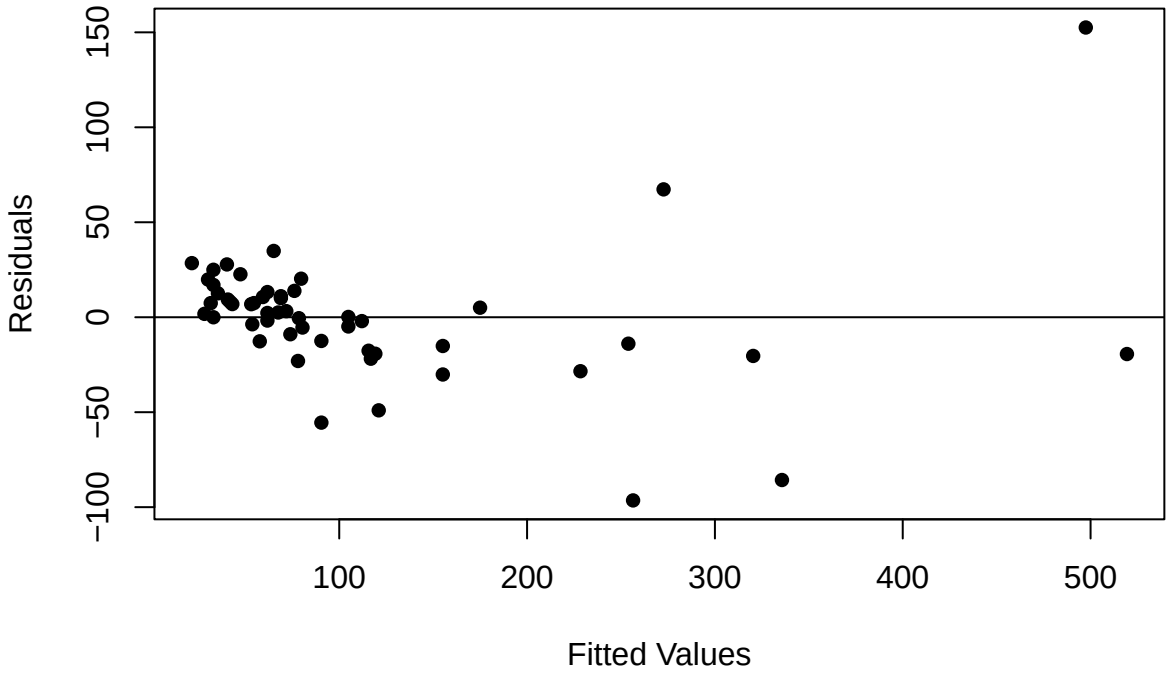
EXAMPLE 3.21 FIT simple linear model with AvgSpent12

```
summary(ClothingmodelA12)
```

```
##
## Call:
## lm(formula = Amount ~ AvgSpent12, data = Clothing3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -96.439 -14.230   2.011  11.446 152.536
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -38.8254     8.3438  -4.653 2.43e-05 ***
## AvgSpent12    1.4368     0.0642  22.380  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.02 on 50 degrees of freedom
## Multiple R-squared:  0.9092, Adjusted R-squared:  0.9074
## F-statistic: 500.9 on 1 and 50 DF,  p-value: < 2.2e-16
```

FIGURE 3.36 Residuals versus fits for the regression of Amount on AvgSpent12

```
plot(ClothingmodelA12$residuals~ClothingmodelA12$fitted,xlab="Fitted Values",ylab="Residuals",pch=16)
abline(h=0)
```



EXAMPLE 3.21 FIT a quadratic model

We use the I( ) notation to add tgeh square term without needing to create a new variable

```
ClothingmodelA12Quad=lm(Amount~AvgSpent12+I(AvgSpent12^2), data=Clothing3)
summary(ClothingmodelA12Quad)
```

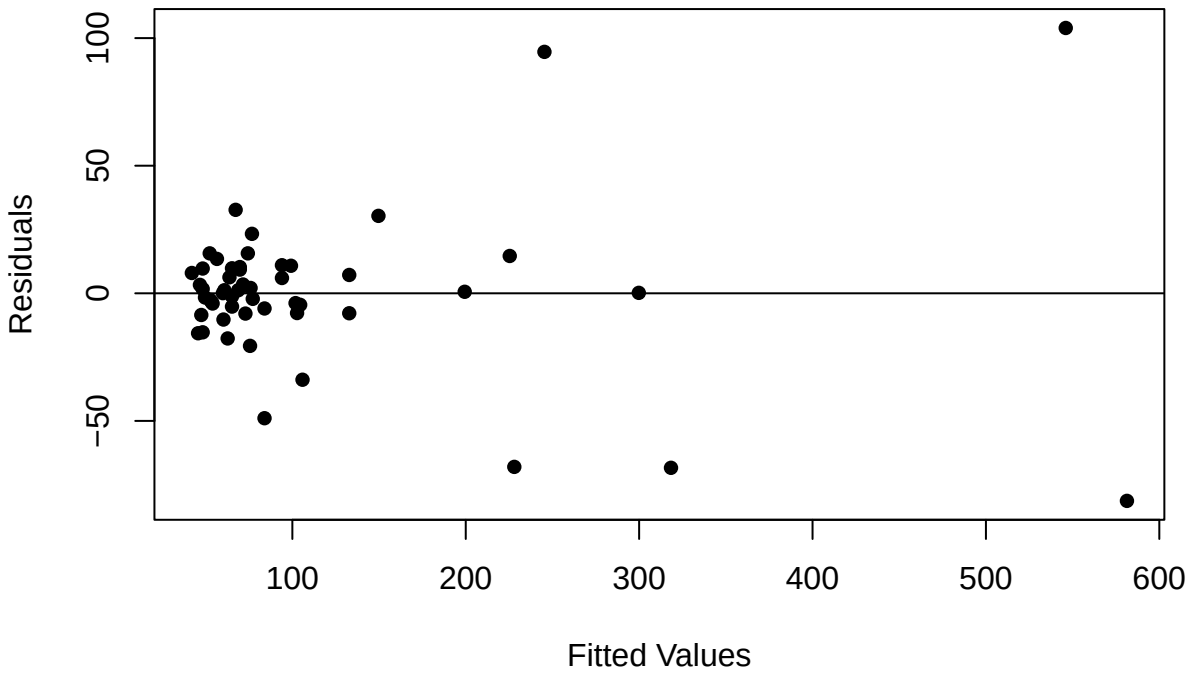
```
##
## Call:
## lm(formula = Amount ~ AvgSpent12 + I(AvgSpent12^2), data = Clothing3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -81.332  -7.752   0.389   9.734 103.968
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.402e+01  1.457e+01   0.963  0.34046
## AvgSpent12      5.709e-01  2.145e-01   2.661  0.01050 *
## I(AvgSpent12^2) 2.289e-03  5.477e-04   4.180  0.00012 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 30.37 on 49 degrees of freedom
## Multiple R-squared:  0.9331, Adjusted R-squared:  0.9304
## F-statistic: 341.7 on 2 and 49 DF,  p-value: < 2.2e-16
```

FIGURE 3.37 Residual plots for quadratic model to predict Amount based on AvgSpent12

(a) Residuals versus fits

```
plot(ClothingmodelA12Quad$residuals~ClothingmodelA12Quad$fitted,xlab="Fitted Values",ylab="Residuals",p
abline(h=0)
```



(b) Normal quantile plot

```
qqnorm(ClothingmodelA12Quad$residuals, xlab="Normal Quantiles", ylab="Residuals",main="", pch=16)
qqline(ClothingmodelA12Quad$residuals)
```

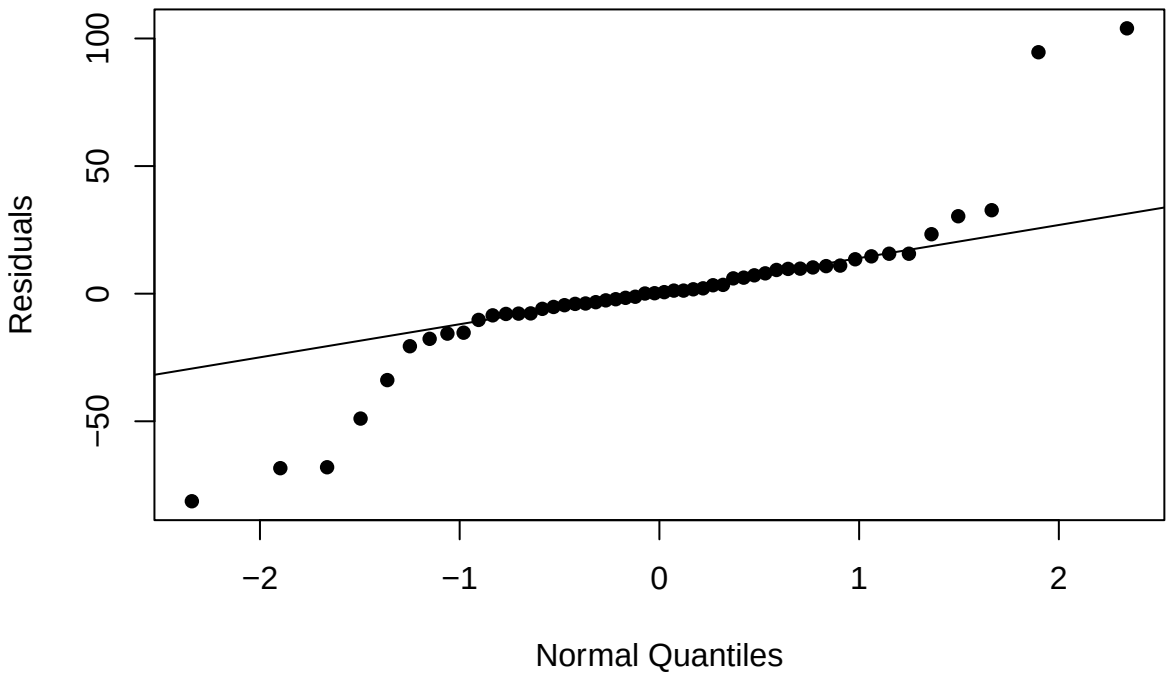
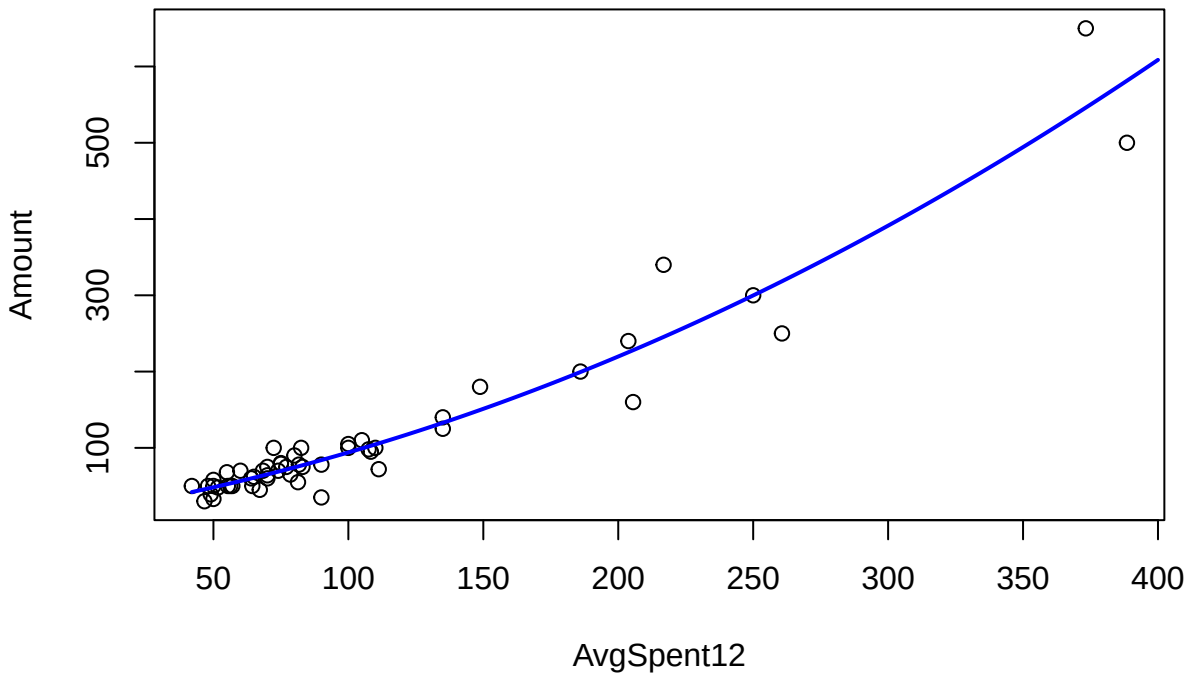


FIGURE 3.38 Quadratic regression fit of Amount on AvgSpent12

```
newx = data.frame(AvgSpent12=seq(42, 400, 0.1))
predictedamt <- predict(ClothingmodelA12Quad,newdata=newx)
plot(Amount~AvgSpent12, data=Clothing3)
lines(predictedamt~newx$AvgSpent12, col="blue", lwd=2)
```


Amount
100
300
500
50
100
150
200
250
300
350
400
AvgSpent12