Sales Data of Men's Fashion Stores

Name: Tim Lin

### Introduction

The *Clothing* data set is in {Ecdat} package taking observation of clothing production in Netherlands. This data is very interesting in finding a way to predict future sales based on significant variables. The data set contains 400 numbers of observations and 13 variables. The variables are *tsales* (annual sales in Dutch guilders), *sales* (sales per square meter), *margin* (gross-fit-margin), *nown* (number of owners/managers), *nfull* (number of full-timers), *npart* (number of part-timers), *naux* (number of helpers/temporary workers) *hoursw* (total number of hours worked), *hourspw* (number of hours worked per worker), *inv1* (investment in shop-premises), *inv2* (investment in automation.), *ssize* (sales floor space of the store in m2), and *start* (year start of business). For this data set, we want to find variables that are significant for the annual sales using step function and Mallow’s statistic. The result returns only *sales*, *nfull*, *npart*, *naux*, *hourspw*, and *ssize* variables are significant.

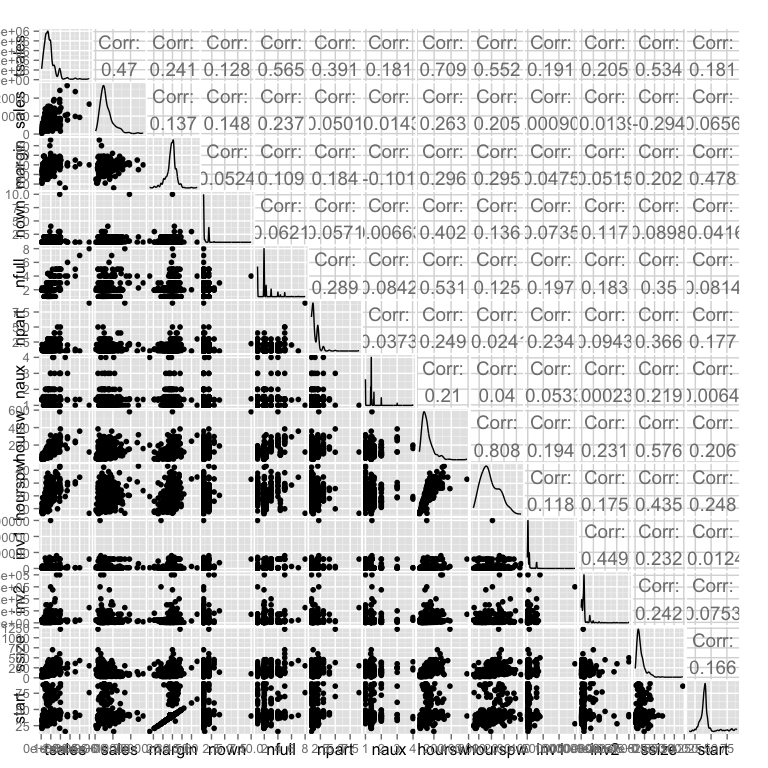
### Data Summary

The summary of the data set:

summary(Clothing)

## tsales sales margin nown   
## Min. : 50000 Min. : 300 Min. :16.00 Min. : 1.000   
## 1st Qu.: 495340 1st Qu.: 3904 1st Qu.:37.00 1st Qu.: 1.000   
## Median : 694227 Median : 5279 Median :39.00 Median : 1.000   
## Mean : 833584 Mean : 6335 Mean :38.77 Mean : 1.284   
## 3rd Qu.: 976817 3rd Qu.: 7740 3rd Qu.:41.00 3rd Qu.: 1.295   
## Max. :5000000 Max. :27000 Max. :66.00 Max. :10.000   
## nfull npart naux hoursw   
## Min. :1.000 Min. :1.000 Min. :1.000 Min. : 32.0   
## 1st Qu.:1.923 1st Qu.:1.283 1st Qu.:1.333 1st Qu.: 80.0   
## Median :1.956 Median :1.283 Median :1.367 Median :104.0   
## Mean :2.069 Mean :1.566 Mean :1.390 Mean :121.1   
## 3rd Qu.:2.066 3rd Qu.:2.000 3rd Qu.:1.367 3rd Qu.:145.2   
## Max. :8.000 Max. :9.000 Max. :4.000 Max. :582.0   
## hourspw inv1 inv2 ssize   
## Min. : 5.708 Min. : 1000 Min. : 350 Min. : 16.0   
## 1st Qu.:13.541 1st Qu.: 20000 1st Qu.: 10000 1st Qu.: 80.0   
## Median :17.745 Median : 22207 Median : 22860 Median : 120.0   
## Mean :18.955 Mean : 58257 Mean : 27829 Mean : 151.1   
## 3rd Qu.:24.303 3rd Qu.: 62269 3rd Qu.: 22860 3rd Qu.: 190.0   
## Max. :43.326 Max. :1500000 Max. :400000 Max. :1214.0   
## start   
## Min. :16.00   
## 1st Qu.:37.00   
## Median :40.00   
## Mean :42.81   
## 3rd Qu.:42.00   
## Max. :90.00

ggpairs(Clothing)



From the summary, we can tell that all the data of only variable *margin* are normal distributed; all other variables are greatly skewed. From the pair plot, we can see that there is high correlation in *tsales* vs. *houesw* and *tsale* vs *nfull*. We can also see there is possible outliners in pair plots.

### 

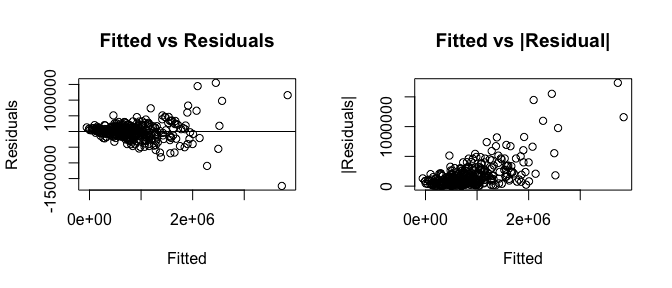
### Data Analysis

First, we use a full model with all variables to diagnosis errors and the fit of the model. Plots of residuals vs fitted values is used to verify constant variance, q-q plot and Shapiro-Wilk test is used to test for normality of errors, and correlation index plot is used to exam independence of errors and confirmed with Durbin-Watson test. Half-normal plots is used to see leverage points, jackknife residuals is used to detect outliners and Cook's statistic is used to find influential points. Then, we refit the data with removed outliners and transformation to find variables that are significant for variable *tsales*.

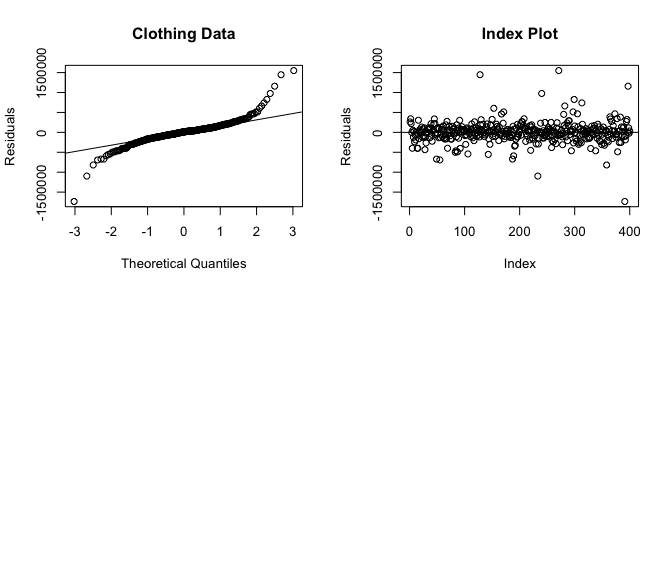
#### Error Diagonistic

fit = lm(tsales~. , Clothing)  
summary(fit)

##   
## Call:  
## lm(formula = tsales ~ ., data = Clothing)  
##   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.245e+06 1.873e+05 -6.649 1.00e-10 \*\*\*  
## sales 7.741e+01 4.762e+00 16.254 < 2e-16 \*\*\*  
## margin -1.314e+03 3.110e+03 -0.422 0.672902   
## nown 5.437e+04 3.504e+04 1.552 0.121488   
## nfull 2.061e+05 2.828e+04 7.288 1.78e-12 \*\*\*  
## npart 1.744e+05 2.568e+04 6.793 4.14e-11 \*\*\*  
## naux 1.577e+05 4.246e+04 3.714 0.000234 \*\*\*  
## hoursw -3.431e+03 1.007e+03 -3.407 0.000725 \*\*\*  
## hourspw 4.141e+04 6.708e+03 6.173 1.69e-09 \*\*\*  
## inv1 -1.700e-01 1.459e-01 -1.165 0.244605   
## inv2 3.398e-01 3.734e-01 0.910 0.363346   
## ssize 2.330e+03 1.960e+02 11.890 < 2e-16 \*\*\*  
## start -1.630e+03 1.195e+03 -1.364 0.173403   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 269200 on 387 degrees of freedom  
## Multiple R-squared: 0.7936, Adjusted R-squared: **0.7872**



The full model have *sales*, *nfull*, *npart*, *naux*,*hoursw*,*hourspw*,and *ssize* as significant variables with = 0.78 .The Fitted vs Residual suggest the variance is heteroscedastic, therefore, transformation of response variable is needed. Box-Cox method is used to determine what is the transformation.



shapiro.test(residuals(fit))

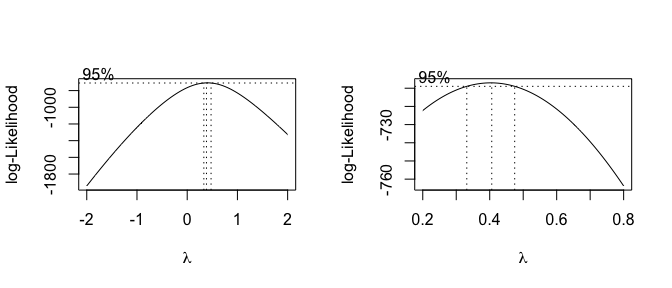
##   
## Shapiro-Wilk normality test  
##   
## data: residuals(fit)  
## W = 0.8553, p-value < 2.2e-16

dwtest(tsales~. , data = Clothing)

##   
## Durbin-Watson test  
##   
## data: tsales ~ .  
## DW = 1.898, p-value = 0.1432  
## alternative hypothesis: true autocorrelation is greater than 0

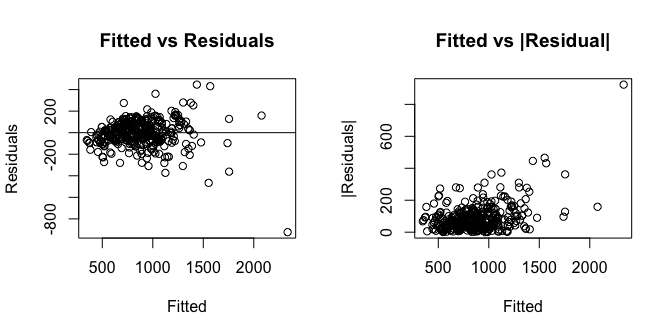
The q-q plot and Shapiro test with small p-value suggest that the error are not normal. From the plot we can tell that the error is long-tailed. The Index plot show that there is no correlated error and the p-value of the Durbin-Watson Test confirms there is no auto-correlation.

#### Transforamtion: Boxcox Plot



The Box-Cox show that the lambda with 95% confidence interval is [ 0.32, 0.48]. We choose a that is close to .48 but a little off the CI for easier interpretation. Therefore, = .5 is used, so square root is used to transform response variable.

fit2 = lm(sqrt(tsales)~. , Clothing)  
summary(fit2)  
## Call:  
## lm(formula = sqrt(tsales) ~ ., data = Clothing)  
##   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.948e+02 8.534e+01 -3.454 0.000612 \*\*\*  
## sales 4.171e-02 2.170e-03 19.223 < 2e-16 \*\*\*  
## margin 2.845e+00 1.417e+00 2.008 0.045368 \*   
## nown 3.738e+01 1.596e+01 2.342 0.019695 \*   
## nfull 6.722e+01 1.289e+01 5.216 2.98e-07 \*\*\*  
## npart 7.103e+01 1.170e+01 6.071 3.04e-09 \*\*\*  
## naux 8.703e+01 1.935e+01 4.498 9.07e-06 \*\*\*  
## hoursw -1.777e+00 4.588e-01 -3.873 0.000126 \*\*\*  
## hourspw 2.233e+01 3.056e+00 7.305 1.59e-12 \*\*\*  
## inv1 1.587e-05 6.647e-05 0.239 0.811483   
## inv2 8.343e-05 1.701e-04 0.490 0.624124   
## ssize 1.311e+00 8.930e-02 14.680 < 2e-16 \*\*\*  
## start -9.843e-01 5.446e-01 -1.807 0.071481 .   
##   
## Residual standard error: 122.6 on 387 degrees of freedom  
## Multiple R-squared: 0.8258, Adjusted R-squared:  **0.8204**



After the transformation, variances are more constant in the Fitted vs residuals plot. *margin* and *nown* variables became significant with =0.824. The model fits better with greater R2 after transformation. Then, we check for outliners and influential points.

#### 

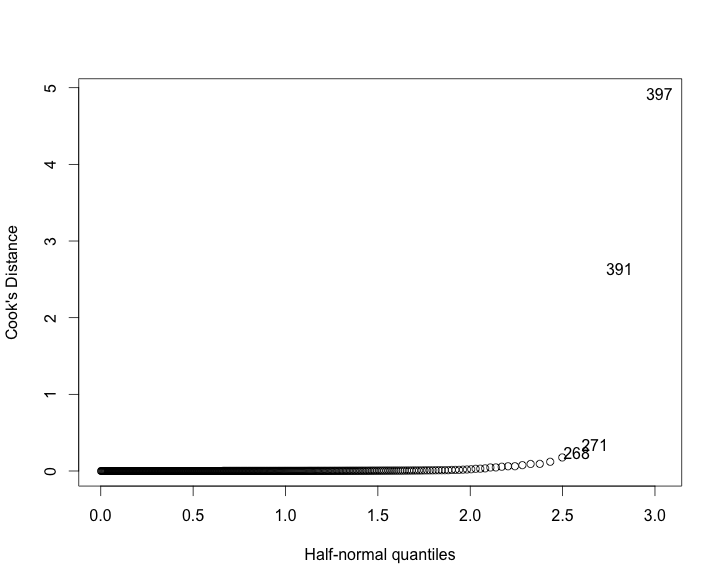
#### Fit Diagnostic

jack = rstudent (fit2)  
qt(.05/(400\*2), 412)

## [1] -3.873002

jack[which(abs(jack) > 3.87)]

## 233 391   
## -4.106789 -10.594099



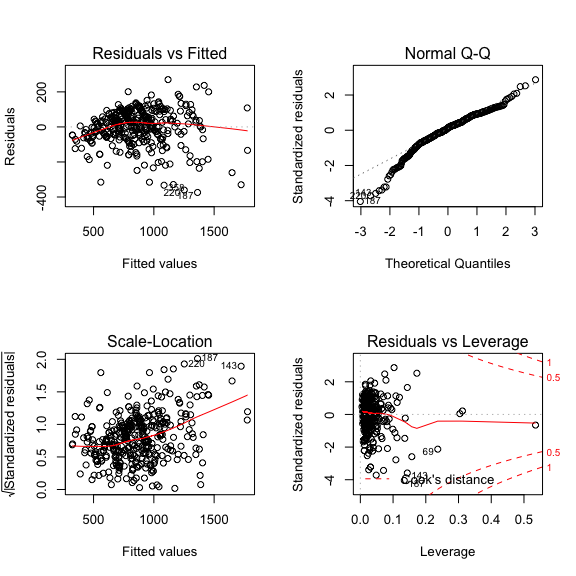
The jackknife value for outliner is anything greater than 3.87, therefore observation 233 and 391 are outliners with jackknife value 4.11 and 10.60 respectively.

The Cook's Distance Plot suggest that observations 397,391,271,268, 233, and 128 are leverage points.

Therefore, we create a new data set (*new.data*) without the outliners and leverage points. Then, we fit the variables with new data to find significant variables.

md = lm(sqrt(tsales) ~., new.data)  
summary(md)

## Call:  
## lm(formula = sqrt(tsales) ~ ., data = new.data)  
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.649e+02 8.787e+01 -1.877 0.06128 .   
## sales 4.635e-02 1.814e-03 25.544 < 2e-16 \*\*\*  
## margin 2.322e+00 1.161e+00 1.999 0.04627 \*   
## nown 1.278e+01 1.495e+01 0.855 0.39290   
## nfull 2.448e+01 1.497e+01 1.636 0.10270   
## npart 4.129e+01 1.458e+01 2.832 0.00487 \*\*   
## naux 7.524e+01 1.770e+01 4.250 2.69e-05 \*\*\*  
## hoursw -5.979e-01 5.662e-01 -1.056 0.29161   
## hourspw 1.131e+01 3.605e+00 3.138 0.00183 \*\*   
## inv1 -2.123e-05 5.585e-05 -0.380 0.70408   
## inv2 -3.060e-05 1.411e-04 -0.217 0.82847   
## ssize 1.851e+00 8.781e-02 21.076 < 2e-16 \*\*\*  
## start -1.301e-01 4.517e-01 -0.288 0.77355   
##   
## Residual standard error: 99.38 on 381 degrees of freedom  
## Multiple R-squared: 0.8682, Adjusted R-squared: 0.864



New model with removed outliners has constant variance, and slightly short-tailed error (which is acceptable), and not leverage points. Variables *sales*, *margin*, *npart*, *naux*, *hourspw*, and *ssize* are significant in this model with = 0.86. This model has better fit after removing the outliners compare to =0.824. Now, we find significant variables using step function and Mallow’s statistic.

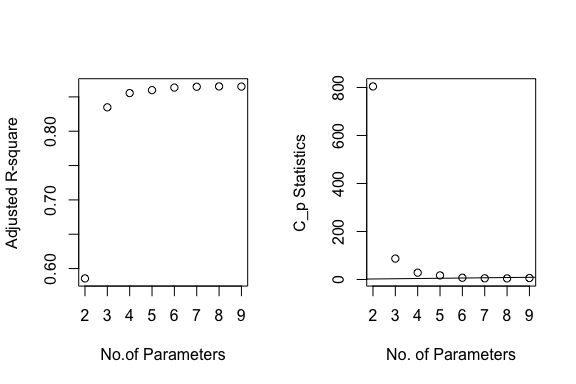
### 

### Variable Selection

all<- regsubsets(sqrt(tsales) ~., new.data)  
(rs<-summary(all))

## sales margin nown nfull npart naux hoursw hourspw inv1 inv2 ssize  
## 1 ( 1 ) " " " " " " " " " " " " "\*" " " " " " " " "   
## 2 ( 1 ) "\*" " " " " " " " " " " " " " " " " " " "\*"   
## 3 ( 1 ) "\*" " " " " " " " " " " "\*" " " " " " " "\*"   
## 4 ( 1 ) "\*" " " " " " " " " "\*" " " "\*" " " " " "\*"   
## 5 ( 1 ) "\*" " " " " " " "\*" "\*" " " "\*" " " " " "\*"   
## 6 ( 1 ) "\*" "\*" " " " " "\*" "\*" " " "\*" " " " " "\*"   
## 7 ( 1 ) "\*" "\*" " " "\*" "\*" "\*" " " "\*" " " " " "\*"   
## 8 ( 1 ) "\*" "\*" " " "\*" "\*" "\*" "\*" "\*" " " " " "\*"   
## start

## 1 ( 1 ) " "   
## 2 ( 1 ) " "   
## 3 ( 1 ) " "   
## 4 ( 1 ) " "   
## 5 ( 1 ) " "   
## 6 ( 1 ) " "   
## 7 ( 1 ) " "   
## 8 ( 1 ) " "



The Mallow’s statistic suggests 7 parameters, with *sales*, *margin*, *npart*, *naux*, *hourspw*, and *ssize* as significant predictors of *tsale.* Now, we want to see if step function give us the same model.

md1 = step(md)

summary(md1)  
## Call:  
## lm(formula = sqrt(tsales) ~ sales + margin + nfull + npart +   
## naux + hourspw + ssize, data = new.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -388.36 -50.30 12.41 66.42 275.68   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -84.630175 44.885882 -1.885 0.06012 .   
## sales 0.046404 0.001786 25.983 < 2e-16 \*\*\*  
## margin 2.135845 1.035810 2.062 0.03988 \*   
## nfull 9.660891 6.129360 1.576 0.11581   
## npart 29.808985 9.701649 3.073 0.00227 \*\*   
## naux 63.507117 13.313022 4.770 2.61e-06 \*\*\*  
## hourspw 7.596472 0.936316 8.113 6.65e-15 \*\*\*  
## ssize 1.839123 0.083963 21.904 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 98.92 on 386 degrees of freedom  
## Multiple R-squared: 0.8677, Adjusted R-squared: 0.8653

The step function with AIC = 3652.91 included *nfull* as a predictor and all other variable are same as the model based on Mallow’s statistic. However, this model have a slightly higier R2 (0.865) Therefore, we want to see the effect of *nfull* on the model.

> md3 = lm(sqrt(tsales) ~ sales + margin + + npart + naux + hourspw + ssize, new.data)

> anova(md3, md1)

Analysis of Variance Table

Model 1: sqrt(tsales) ~ sales + margin + +npart + naux + hourspw + ssize

Model 2: sqrt(tsales) ~ sales + margin + nfull + npart + naux + hourspw + ssize

Res.Df RSS Df Sum of Sq F Pr(>F)

1 387 3801673

2 386 3777362 1 24311 2.4843 0.1158

md3 = lm(sqrt(tsales) ~ sales + margin + + npart + naux + hourspw + ssize, new.data)

Call:

lm(formula = sqrt(tsales) ~ sales + margin + +npart + naux + hourspw + ssize, data = new.data)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -74.296019 44.489516 -1.670 0.09573 .

sales 0.047402 0.001673 28.332 < 2e-16 \*\*\*

margin 2.160118 1.037680 2.082 0.03803 \*

npart 29.717574 9.720062 3.057 0.00239 \*\*

naux 63.852743 13.336718 4.788 2.41e-06 \*\*\*

hourspw 7.328361 0.922499 7.944 2.15e-14 \*\*\*

ssize 1.884713 0.078974 23.865 < 2e-16 \*\*\*

Residual standard error: 99.11 on 387 degrees of freedom

Multiple R-squared: 0.8668, Adjusted R-squared: 0.8648

The Pr(>F) = 0.1158 of ANOVA suggests *nfull* is not significant and with *nfull* remove, from the model the R2  remains almost the same., therefore, *nfull* doesn’t have an effect on the model.

#### 

The Best Model Diagnostic:

The Residual vs Fitted Values plot shows constant variance, the qq plot show errors are normal and residuals vs leverage plot show that there is no outliners. Therefore, this model with variables *sales*, *margin*, *npart*, *naux*, *hourspw*, and *ssize* are the best model with high R2 value (0.8648) and least number of variables.

#### Conclusion

The annual sales in Dutch guilders is only related to sales per square meter,, gross-fit-margin, number of helpers/temporary workers, number of hours worked per worker, sales floor space of the store (in m2). It makes sense because annul sales is highly correlated with total number of hours worked (*hoursw*) which is a measurement of labor, where number of helpers/temporary workers (*npart*), number of hours worked per worker(*naux*) are also a measurement of labor. It is clear that annual sales would be greater if number of labors are greater. Also, for larger sales floor space of the store, it's likely to sell more cloth. For example, compare Macy's to local designer's store in SOHO, Macy's would sell a lot more clothes. Also, with higher gross-fit-margin, proportion of money left over from revenues after accounting for the cost of goods sold, the annual sales would be greater because you more cheaper good in large portion.