LNL_HW_week8

Patrick Kelly

Saturday, March 07, 2015

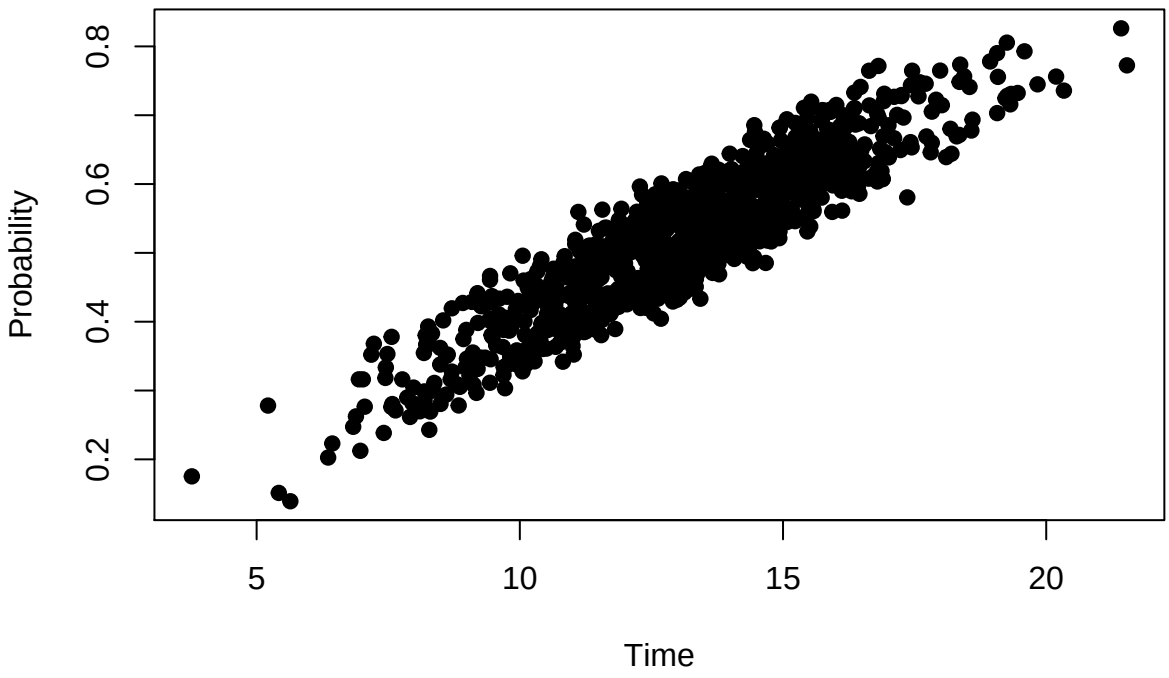
Week 8 Assignment

Look at the sample in the file LinearModelCase1.csv. Below nSample is the length of the sample imported from the file. The first 10 rows and the X-Y plot are:

```
MarketingData<-read.csv(file="C:/Users/Patrick/Documents/R/UChicago/Linear_NonLinear/MarketingExperimen
MarketingData<-as.data.frame(MarketingData)
MarketingData[1:10,]
```

```
##        Time Probability Gender
## 1  13.13286   0.4994391      M
## 2  15.56113   0.5619825      M
## 3  11.98454   0.4700490      F
## 4  15.61916   0.6179313      M
## 5  16.37815   0.6155603      M
## 6  16.12320   0.5613695      M
## 7  10.76662   0.4106634      F
## 8  14.35351   0.5072862      M
## 9  12.36272   0.5457849      F
## 10 16.32724   0.5928357      M
```

```
plot(MarketingData$Time,MarketingData$Probability, type="p",pch=19,xlab="Time",ylab="Probability")
```



Estimate linear model using function lm look at the output of the function

```
MarketingData.EstimatedLinearModel<-lm(Probability~Time,data=MarketingData)
names(MarketingData.EstimatedLinearModel)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```
MarketingData.EstimatedLinearModel$coefficients
```

```
## (Intercept)        Time
## 0.007165356 0.039232604
```

look at the model summary

```
summary(MarketingData.EstimatedLinearModel)
```

```
##
## Call:
## lm(formula = Probability ~ Time, data = MarketingData)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

```
## -0.107643 -0.038418 -0.001423  0.036755  0.116378
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0071654  0.0076371   0.938    0.348
## Time        0.0392326  0.0005758  68.142   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04584 on 998 degrees of freedom
## Multiple R-squared:  0.8231, Adjusted R-squared:  0.8229
## F-statistic:  4643 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
names(summary(MarketingData.EstimatedLinearModel))
```

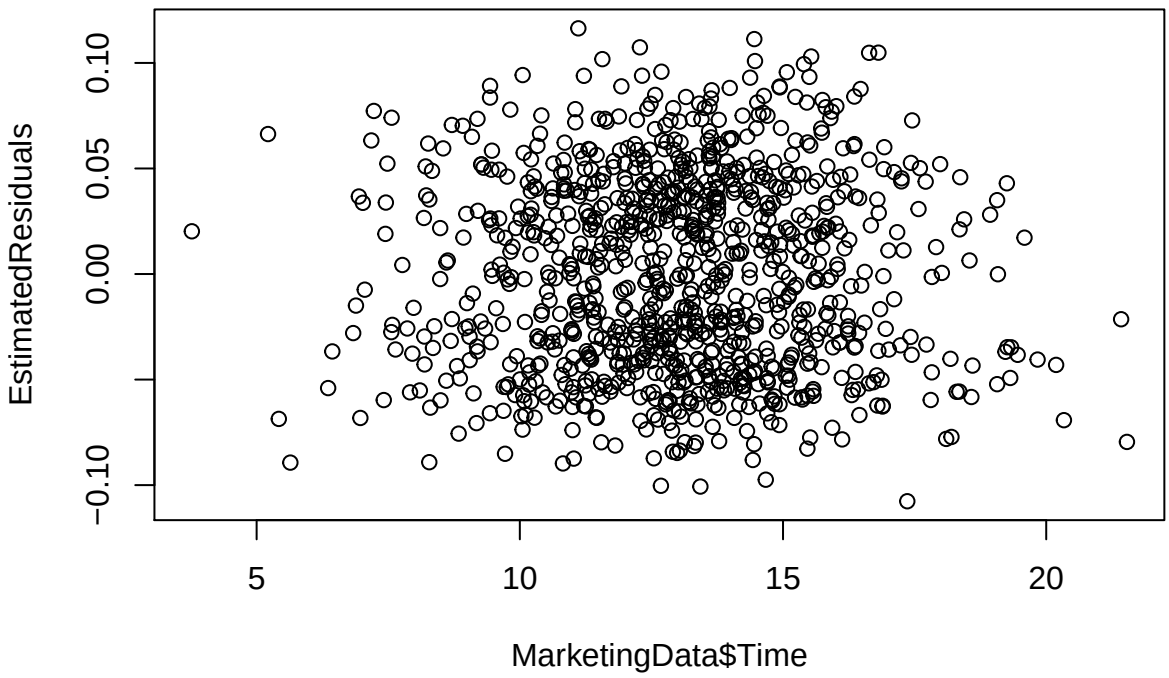
```
##  [1] "call"          "terms"         "residuals"     "coefficients"
##  [5] "aliased"       "sigma"         "df"            "r.squared"
##  [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

Interpret the results in the output.

The model above is generally a good fit - low residual sq error, high multiple r-squared and our independent variable (time) is highly significant. However, through observing the original plot of the data we can see that there may be multiple samples within this data which is definitely worth further exploring, particularly through observing our other independent variable (gender).

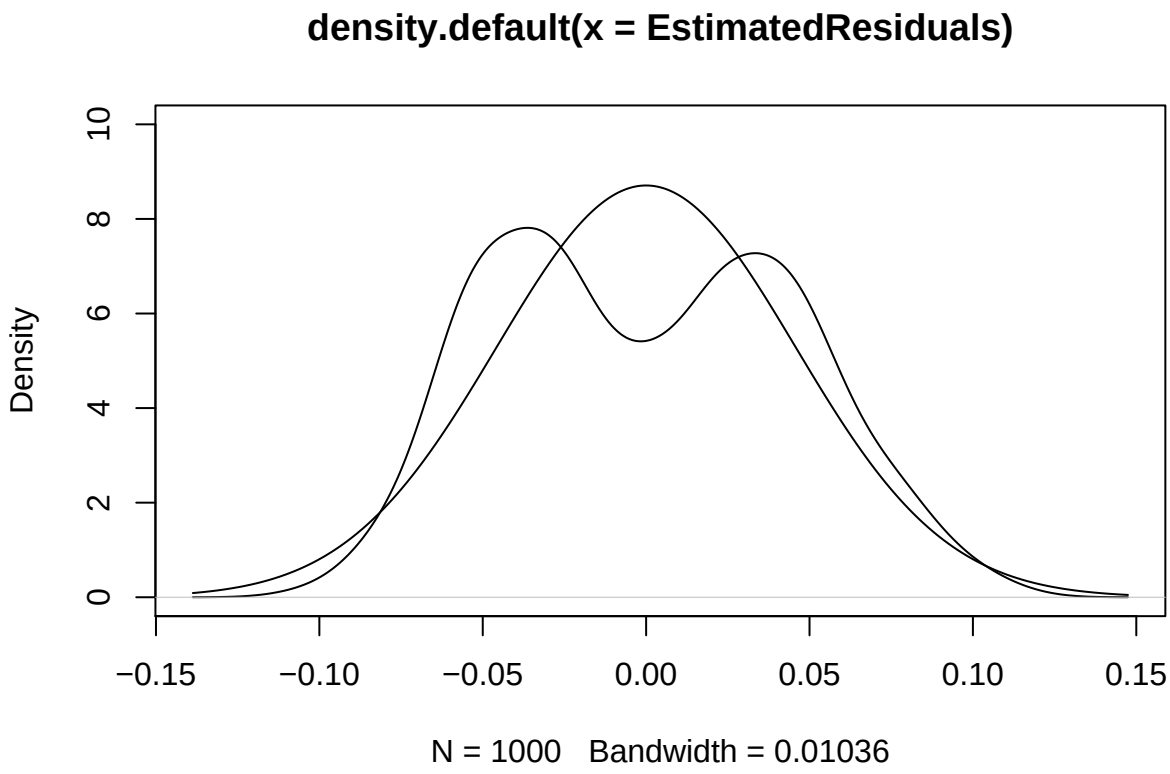
Observe the residuals

```
EstimatedResiduals<-MarketingData.EstimatedLinearModel$residuals
plot(MarketingData$Time,EstimatedResiduals)
```



and their probability density in comparison with the normal density

```
Probability.Density.Residuals<-density(EstimatedResiduals)
plot(Probability.Density.Residuals,ylim=c(0,10))
lines(Probability.Density.Residuals$x,dnorm(Probability.Density.Residuals$x,mean=mean(EstimatedResiduals
```

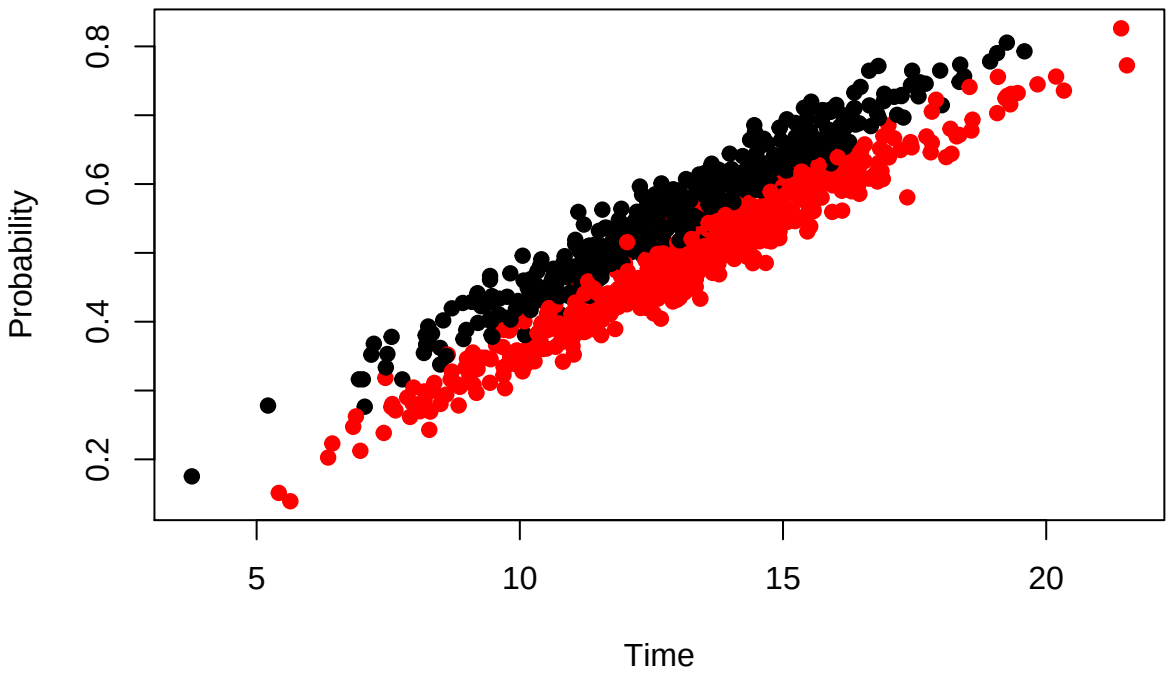


What do you conclude from the analysis of residuals?

The analysis of residuals confirms my suspision mention above that there are two seperate samples within the data. This can be see through the plot of the residuals but can be strongly seen in the probability density plot of the residuals.

This is futher supported when checking the plot of the data again, differing the color by gender.

```
plot(MarketingData$Time,MarketingData$Probability, type="p",pch=19,
     xlab="Time",ylab="Probability",col=MarketingData$Gender)
```



Add the fixed effect based on gender. Compare the two summaries.

```
MarketingData.LinearModel.Gender<-lm(Probability~Time+Gender,data=MarketingData)
summary(MarketingData.LinearModel.Gender)
```

```
##
## Call:
## lm(formula = Probability ~ Time + Gender, data = MarketingData)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.074464 -0.017687 -0.001208  0.015808  0.076044
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0409007  0.0042002   9.738   <2e-16 ***
## Time         0.0398264  0.0003126 127.408   <2e-16 ***
## GenderM     -0.0772236  0.0015783 -48.927   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02487 on 997 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9479
## F-statistic:  9085 on 2 and 997 DF,  p-value: < 2.2e-16
```

```r
summary(MarketingData.EstimatedLinearModel)
```

```
##
## Call:
## lm(formula = Probability ~ Time, data = MarketingData)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.107643 -0.038418 -0.001423  0.036755  0.116378
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0071654  0.0076371   0.938    0.348
## Time        0.0392326  0.0005758  68.142   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04584 on 998 degrees of freedom
## Multiple R-squared:  0.8231, Adjusted R-squared:  0.8229
## F-statistic:  4643 on 1 and 998 DF,  p-value: < 2.2e-16
```

Comparison of the two models above

The model including linear model including both `Time` and `Gender` is a better fit. Both of `Time` and `Gender` are significant, the residual standard error improved, and the multiple r-square improved. Overall it is clear that the second model is a better fit.

Learn how to fit the model using `lmer()` from `lme4`

First use the simplest random effects model with Time as output and Gender as the only random effect.

```r
library(lme4)
```

```
## Loading required package: Matrix
## Loading required package: Rcpp
```

```r
MarketingData.Time.Random.Effect<-lmer(Time~1+(1|Gender),data=MarketingData)
summary(MarketingData.Time.Random.Effect)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Time ~ 1 + (1 | Gender)
##    Data: MarketingData
##
## REML criterion at convergence: 4687.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.6606 -0.6568  0.0249  0.6506  3.3671
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Gender   (Intercept) 0.006464 0.0804
##  Residual             6.341992 2.5183
```

```
## Number of obs: 1000, groups:  Gender, 2
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 13.02104    0.09791     133
```

One way of thinking about the variances returned by the summary of `lmer()` is: residual variance is the variance within the groups and each random effect variance is the variance between the groups.

Look at other fields in `lmer()` object.

```
names(summary(MarketingData.Time.Random.Effect))
```

```
##  [1] "methTitle"    "objClass"     "devcomp"      "isLmer"
##  [5] "useScale"     "logLik"       "family"       "link"
##  [9] "ngrps"        "coefficients" "sigma"        "vcov"
## [13] "varcor"       "AICtab"       "call"         "residuals"
```

```
summary(MarketingData.Time.Random.Effect)$coefficients
```

```
##             Estimate Std. Error  t value
## (Intercept) 13.02104 0.09790661 132.9945
```

```
summary(MarketingData.Time.Random.Effect)$sigma
```

```
## [1] 2.518331
```

```
##summary(MarketingData.Time.Random.Effect)$residuals
```

Now apply `lmer()` to fit the model with one predictor Rime and one random effect based on Gender

```
MarketingData.Probability.Random.Effect <- lmer(Probability ~ Time + (1|Gender), data=MarketingData)

summary(MarketingData.Probability.Random.Effect)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Probability ~ Time + (1 | Gender)
##    Data: MarketingData
##
## REML criterion at convergence: -4517.9
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.99486 -0.71179 -0.04851  0.63536  3.05851
##
## Random effects:
##  Groups   Name        Variance  Std.Dev.
##  Gender   (Intercept) 0.0029805 0.05459
##  Residual             0.0006184 0.02487
## Number of obs: 1000, groups:  Gender, 2
##
```

```
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept) 0.0022910  0.0388256    0.06
## Time        0.0398262  0.0003126  127.41
##
## Correlation of Fixed Effects:
##      (Intr)
## Time -0.105
```

Compare the summaries and residuals of `MarketingData.Probability.Random.Effect` the linear model `Probability~Time`.

In checking the summaries of the two models we see that the residual standard error of the linear model (0.0458) is higher than the residual std dev of the mixed model (0.02487). Furthermore the AIC value for the mixed model is considerably lower than the linear model (-4509.9 < -3323.3). Both models show that Time is important to the model, and in the mixed model the ratio of the Gender variance is greater than the residual variance - supporting its significance in the model as well.

Through plotting the residials of the mixed model (see below) and comparing to the visualizations of the linear model residuals - from earlier in the analysis - we can see that the residuals from the mixed model appear to uniform (a much more desirable outcome).

```
Mixed.Residuals <- summary(MarketingData.Probability.Random.Effect)$residuals

plot(MarketingData$Time,Mixed.Residuals)
```