week08_homework

Brian Ritz

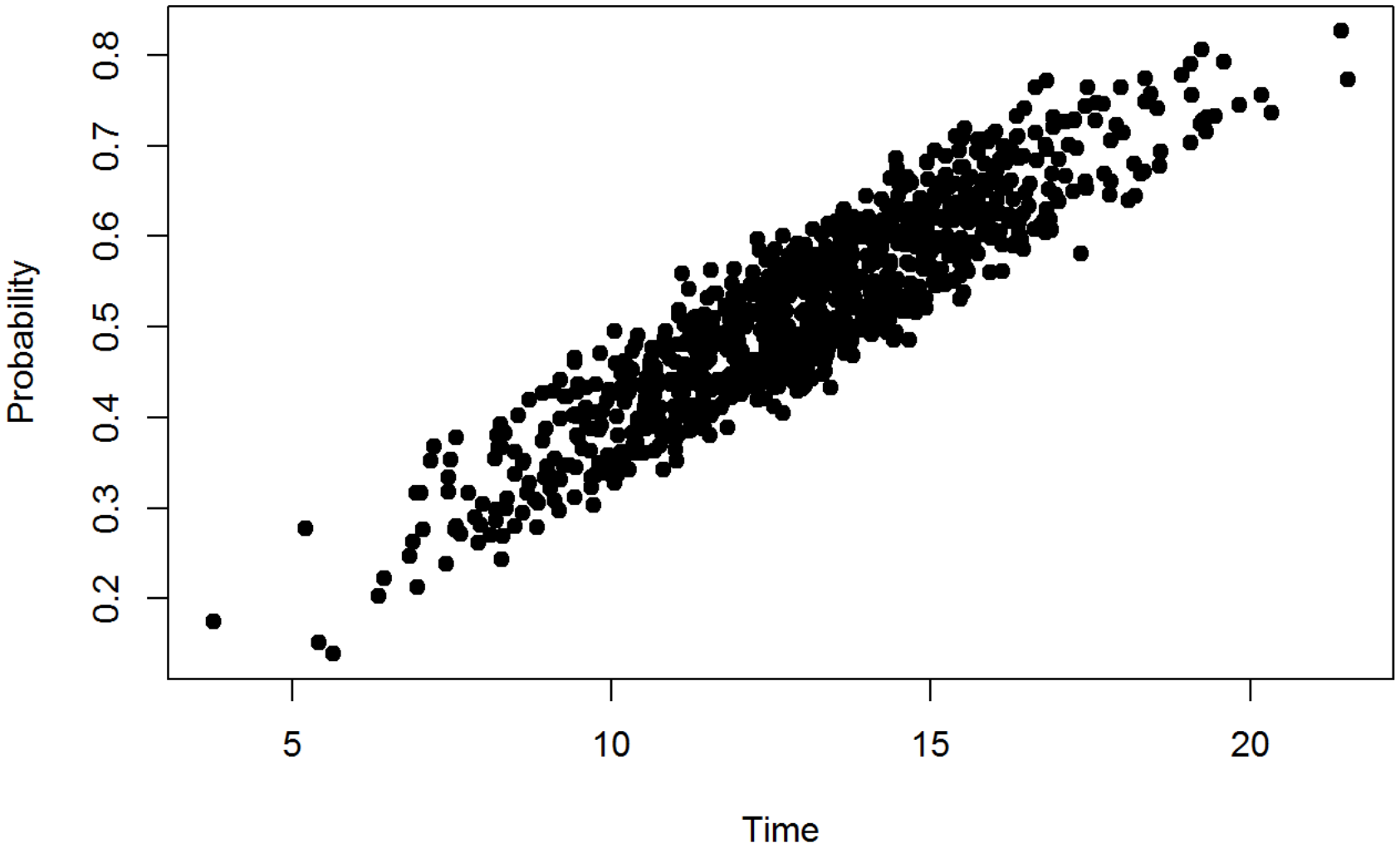
Saturday, March 07, 2015

Look at the data:

```
MarketingData<-read.csv(file="MarketingExperiment.csv",header=TRUE,sep=",")
MarketingData<-as.data.frame(MarketingData)
MarketingData[1:10,]
```

```
##     Time Probability Gender
## 1  13.13      0.4994      M
## 2  15.56      0.5620      M
## 3  11.98      0.4700      F
## 4  15.62      0.6179      M
## 5  16.38      0.6156      M
## 6  16.12      0.5614      M
## 7  10.77      0.4107      F
## 8  14.35      0.5073      M
## 9  12.36      0.5458      F
## 10 16.33      0.5928      M
```

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



Estimate a linear model and check out the output:

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

```
##
## Call:
## lm(formula = Probability ~ Time, data = MarketingData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10764 -0.03842 -0.00142  0.03676  0.11638
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.007165   0.007637    0.94     0.35
## Time        0.039233   0.000576   68.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0458 on 998 degrees of freedom
## Multiple R-squared:  0.823,  Adjusted R-squared:  0.823
## F-statistic: 4.64e+03 on 1 and 998 DF,  p-value: <2e-16
```

```
print(names(MarketingData.EstimatedLinearModel))
```

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

```
print(MarketingData.EstimatedLinearModel$coefficients)
```

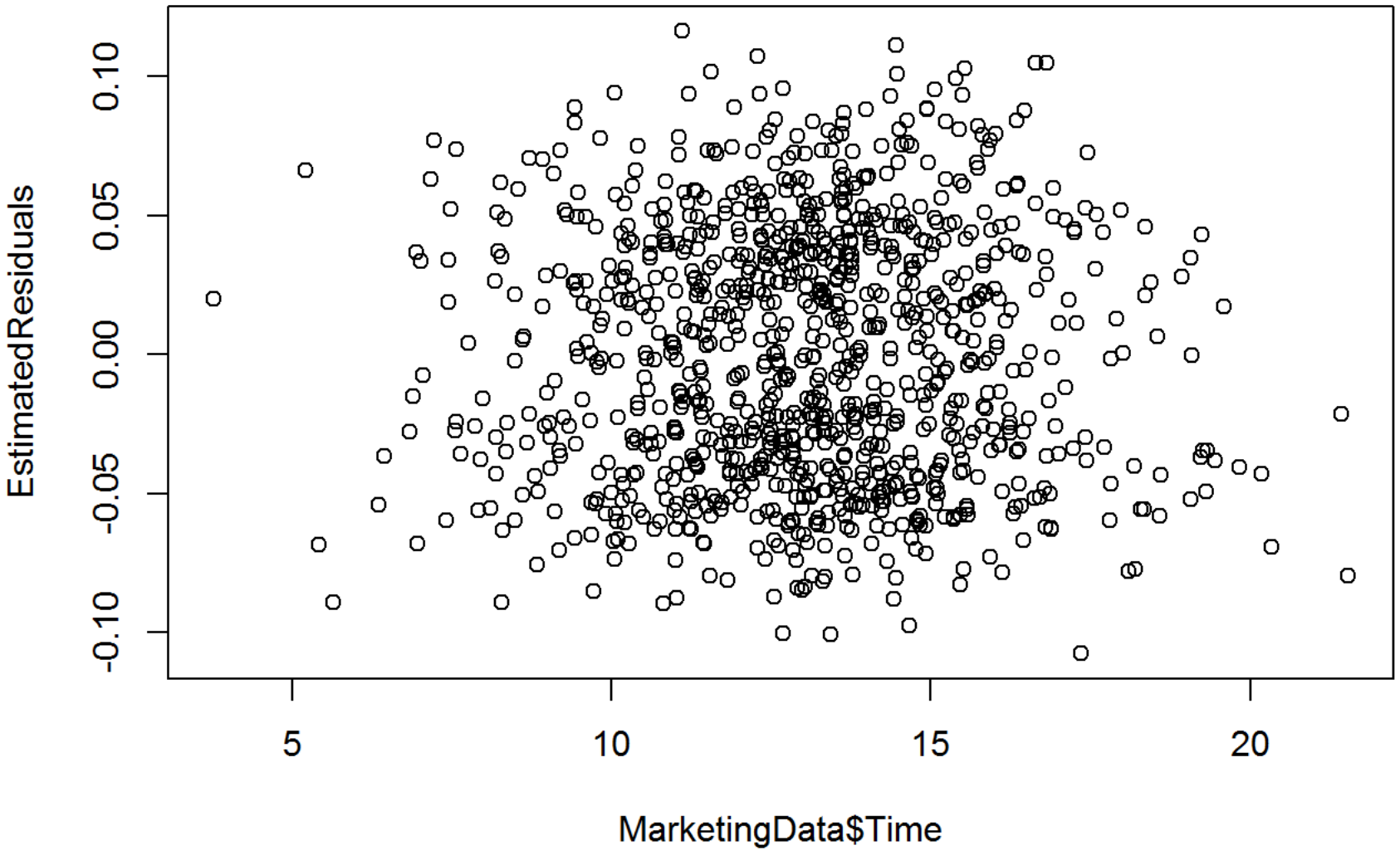
```
## (Intercept)        Time
##    0.007165    0.039233
```

** INTERPRET THE RESULTS OF THE OUTPUT**

Time coefficient is significant and positive. I interpret the coefficient as meaning the following: for a one unit of time increase the probability increases by just under 4 percentage points. The p-value is very small, so we are confident that this coefficient is different from zero

We now check out the residuals of the model:

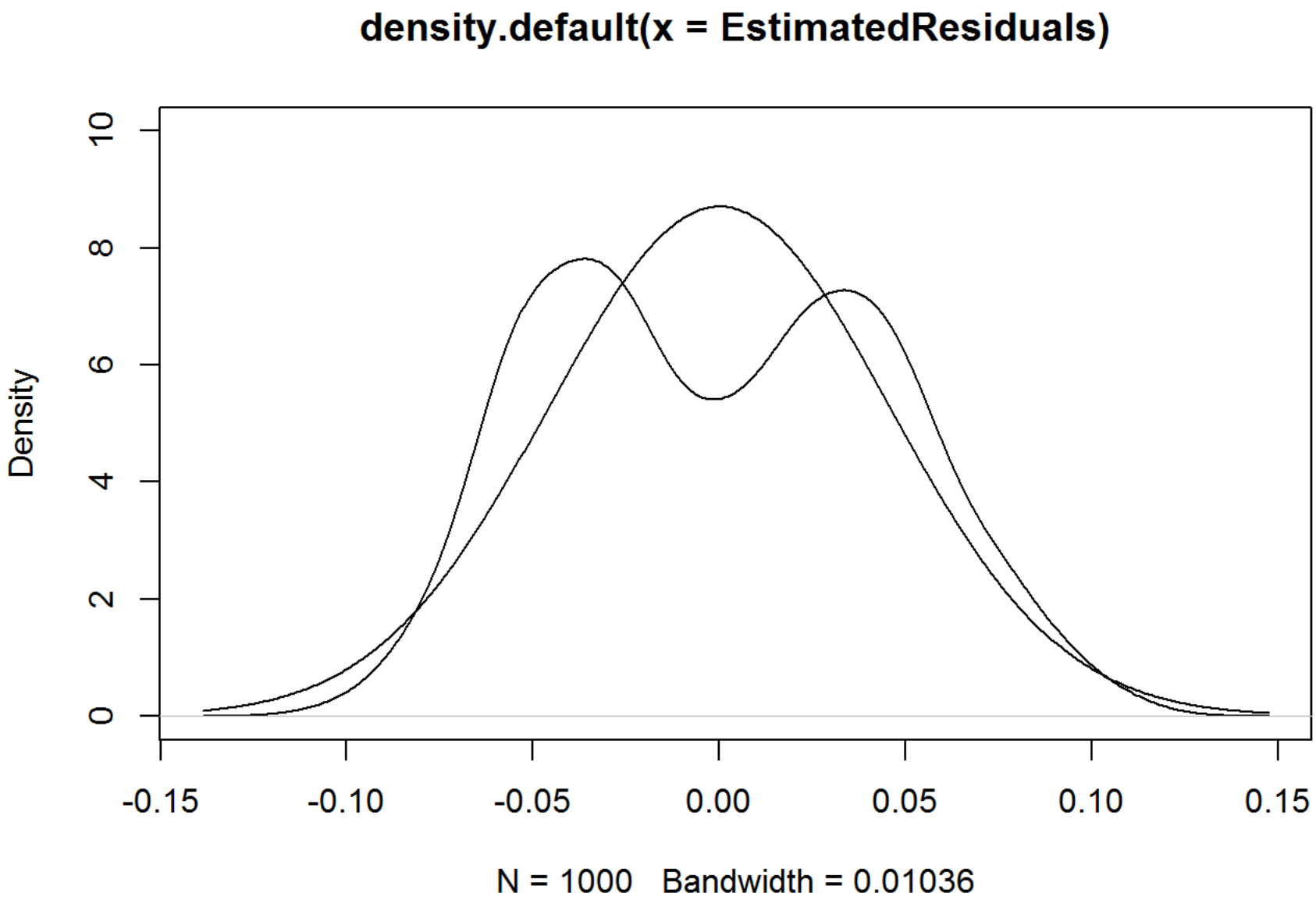
```
EstimatedResiduals<-MarketingData.EstimatedLinearModel$residuals
plot(MarketingData$Time,EstimatedResiduals)
```



It looks like there might be two groups with means just above and below 0, but lets check out the distributions compared the the normal distribution to be sure:

```
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),s
d=sd(EstimatedResiduals)))
```



It looks like there is a “hole” around the mean of the distribution! Perhaps there are actually two groups! It is possible that these two groups are the male/female groups, so lets fit a fixed effect on gender and see if that helps the pattern of the residuals. We will compare the two summaries:

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

```
##
## Call:
## lm(formula = Probability ~ Time + Gender, data = MarketingData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07446 -0.01769 -0.00121  0.01581  0.07604
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.040901   0.004200    9.74   <2e-16 ***
## Time         0.039826   0.000313  127.41   <2e-16 ***
## GenderM     -0.077224   0.001578  -48.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0249 on 997 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.948
## F-statistic: 9.09e+03 on 2 and 997 DF,  p-value: <2e-16
```

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

```
##
## Call:
## lm(formula = Probability ~ Time, data = MarketingData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10764 -0.03842 -0.00142  0.03676  0.11638
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.007165   0.007637    0.94     0.35
## Time        0.039233   0.000576   68.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0458 on 998 degrees of freedom
## Multiple R-squared:  0.823,  Adjusted R-squared:  0.823
## F-statistic: 4.64e+03 on 1 and 998 DF,  p-value: <2e-16
```

The time coefficients are exactly the same, but the genderM coefficient in the gender model is also significant with a coefficient of -.07, meaning that on average being male makes you probability 7 percentage points less.

Now, we will fit a random effects model with lmer() from lme4:

```
library(lme4)
```

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

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

```
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: 4688
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -3.661 -0.657  0.025  0.651  3.367
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Gender   (Intercept) 0.00646  0.0804
##  Residual             6.34199  2.5183
## Number of obs: 1000, groups:  Gender, 2
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  13.0210     0.0979     133
```

WE can see from the summary that variance within the groups is much less than the variance between the groups. We know this because the variance on the gender group is much less than the variance on the residual group.

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

```
##             Estimate Std. Error t value
## (Intercept)    13.02    0.09791     133
```

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

```
## [1] 2.518
```

Now we apply lmer() to fit the model with one predictor Time and one random effect based on Gender:

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Probability ~ Time + (1 | Gender)
##    Data: MarketingData
##
## REML criterion at convergence: -4518
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.9949 -0.7118 -0.0485  0.6354  3.0585
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Gender   (Intercept) 0.002981 0.0546
```

```
##  Residual             0.000618 0.0249
## Number of obs: 1000, groups:  Gender, 2
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 0.002291   0.038826     0.1
## Time        0.039826   0.000313   127.4
##
## Correlation of Fixed Effects:
##      (Intr)
## Time -0.105
```

** COMPARE THE SUMMARIES OF THE RANDOM EFFECT MODEL WITH THE ORIGINAL LINEAR MODEL **

IN the random effects model, the fixed effect on time is very nearly the same as the linear model's coefficient on time. The standard error for the fixed effect time in the random effect model is less than the standard error for the linear model because some of that variance was taken up by the random effect. The residual standard error for hte linear model is higher than the residual standard deviation, meaning that there is less residual error in the random effects model. The t-value of the random effects model also indicates that time is still a significant predictor of probability.