qbs121_hw5_gibran

Gibran Erlangga

2/15/2022

Questions

Choose two online datasets that are suitable for use demonstrating (1) normal linear mixed and (2) binary/poisson mixed models. These can be either longitudinal or simply clustered, but should include covariates as well as cluster indicators. 1. For both the linear and nonlinear analyses, describe and justify the longitudinal/clustered outcomes and covariates and the plan for fitting and interpreting mixed random and fixed effects models. 2. Fit the models using lmer and glmer and provide summary statistics and graphs for summarizing the results and assessment of modeling assumptions.

Dataset Justifications

For normal linear mixed models, I am using the "House Prices in the City of Windsor, Canada" dataset, which contains these following variables: sell = sale price of a house lot = the lot size of a property in square feet bdms = the number of bedrooms fb = the number of full bathrooms sty = the number of stories excluding basement drv = 1 if the house has a driveway rec = 1 if the house has a recreational room ffin = 1 if the house has a full finished basement ghw = 1 if the house uses gas for hot water heating ca = 1 if there is central air conditioning gar = the number of garage places reg = 1 if the house is located in the preferred neighbourhood of the city

For binary/poisson mixed models, I am using "Do Workplace Smoking Bans Reduce Smoking?" dataset.

Normal Linear Mixed Model

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7.9000
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.1.1

## Warning: package 'tibble' was built under R version 4.1.1

## Warning: package 'tidyr' was built under R version 4.1.1
```

```
## Warning: package 'readr' was built under R version 4.1.1

## Warning: package 'stringr' was built under R version 4.1.1

## Warning: package 'forcats' was built under R version 4.1.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
data <- read.csv('HousePrices.csv')
```

The dataset describes house prices in the city of Windsor, Canada (546 rows and 13 columns). The dependent variable is house price, which signifies by the "price" column. The rest of the variables signify all the details about each house presented in the dataset (house size in square feet, number of bedrooms, bathrooms, garages, as well as other factors such as whether or not the house is located in the preferred neighborhood of the city). We can see some sample data from the dataset along with the distribution of the dependent variable below:

```
paste('# of rows/# of columns:', dim(data)[1] ,'/', dim(data)[2])
```

```
## [1] "# of rows/# of columns: 546 / 13"
```

```
print('list of columns: ')
```

```
## [1] "list of columns: "
```

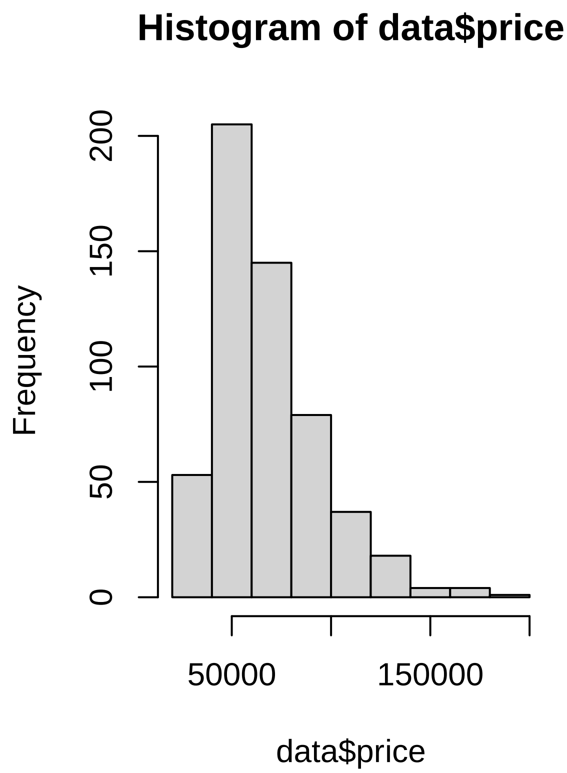
```
names(data)
```

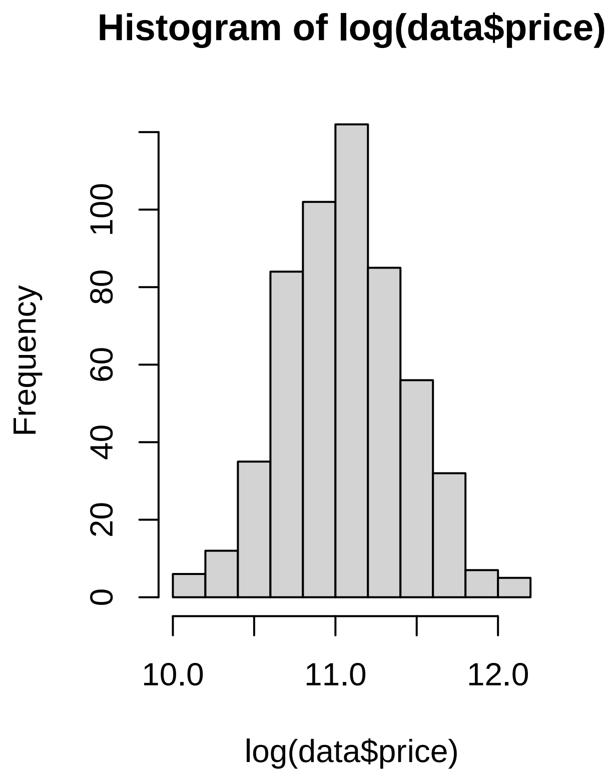
```
##  [1] "X"          "price"      "lotsize"    "bedrooms"   "bathrooms"
##  [6] "stories"    "driveway"   "recreation" "fullbase"   "gasheat"
## [11] "aircon"     "garage"     "prefer"
```

```
# see some sample data
print(head(data, 3))
```

```
##   X price lotsize bedrooms bathrooms stories driveway recreation fullbase
## 1 1 42000    5850        3         1       2      yes         no      yes
## 2 2 38500    4000        2         1       1      yes         no       no
## 3 3 49500    3060        3         1       1      yes         no       no
##   gasheat aircon garage prefer
## 1      no     no      1     no
## 2      no     no      0     no
## 3      no     no      0     no
```

```
# plot dependent and independent variables
par(mfrow=c(1,2))
hist(data$price)
hist(log(data$price))
```

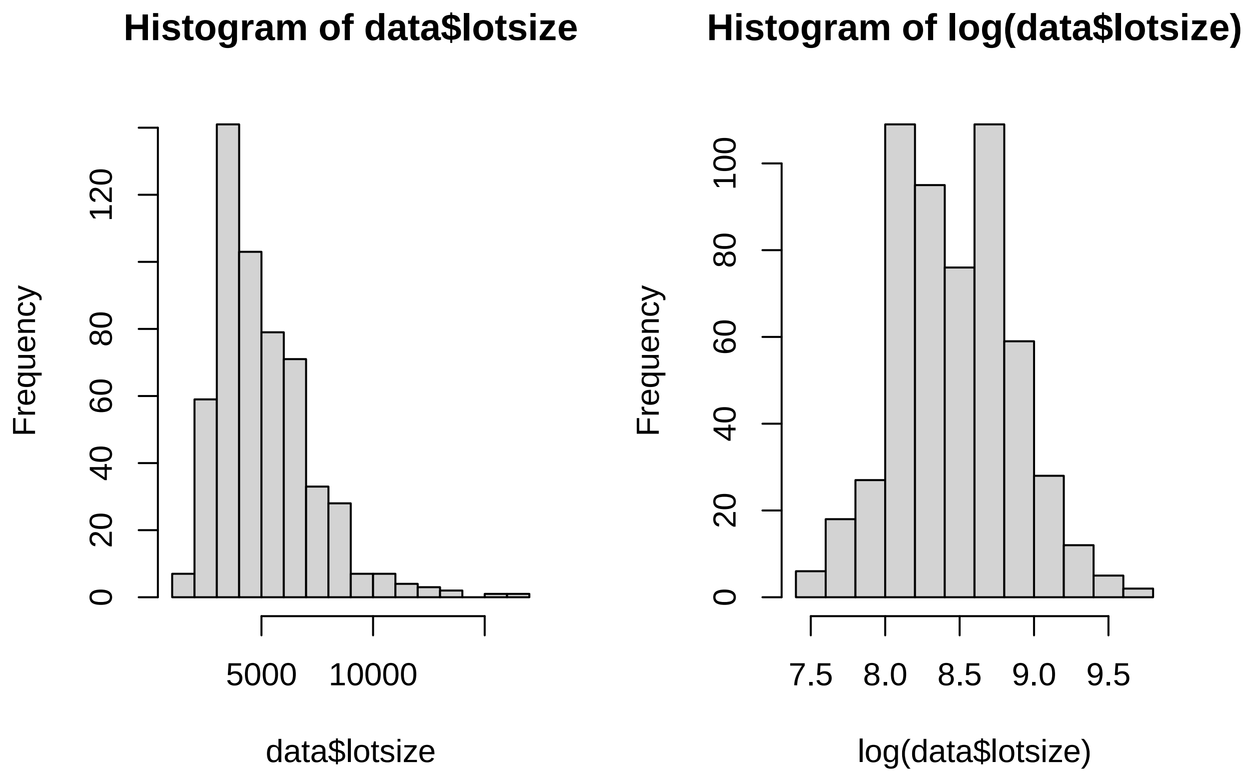




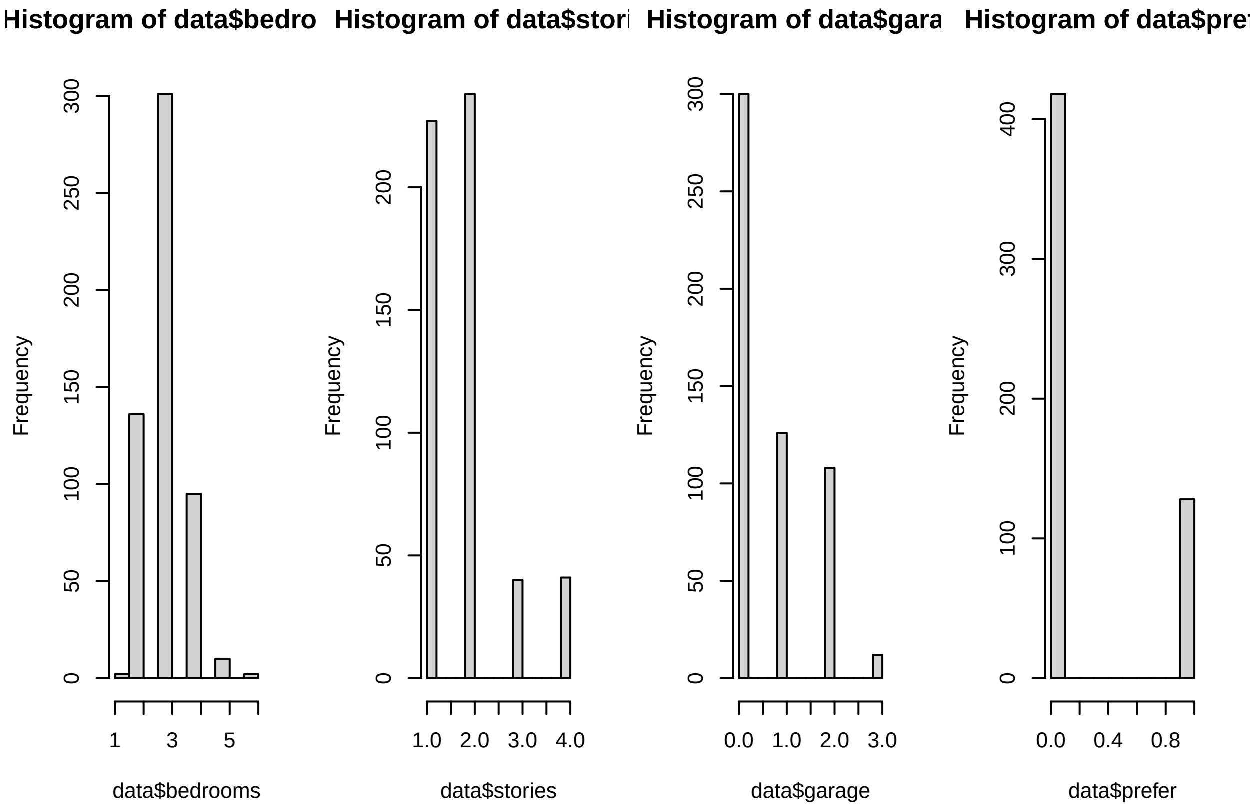
Above figures show the distribution of the dependent variable, the original one (left) and the one after applying a log transformation to the data (right). We can observe that the house price distribution is skewed to the right, meaning that it has a long right tail and the variable mean to the right of the median. To comply with one of the assumptions of linear regression, I applied a log-transform to the house price variable to make it more normally distributed.

I selected a handful of independent variables as potential predictors for the house price. The variables are: lotsize -> lot size of a property in square feet bedrooms -> number of bedrooms stories -> number of stories excluding basement garage -> number of garages prefer -> a flag that shows whether the house located in the preferred neighborhood of the city

```
boolean_convert <- function(data) {
  if (data == "yes") {
    return(1)
  } else {
    return(0)
  }
}

data$prefer <- sapply(data$prefer, boolean_convert)

df <- data %>%
      select('price', 'lotsize', 'bedrooms', 'stories', 'garage', 'prefer')

par(mfrow=c(1,2))
hist(data$lotsize)
hist(log(data$lotsize))
```




```
par(mfrow=c(1, 4))
hist(data$bedrooms)
hist(data$stories)
hist(data$garage)
hist(data$prefer)
```




```r
library(lme4)
```

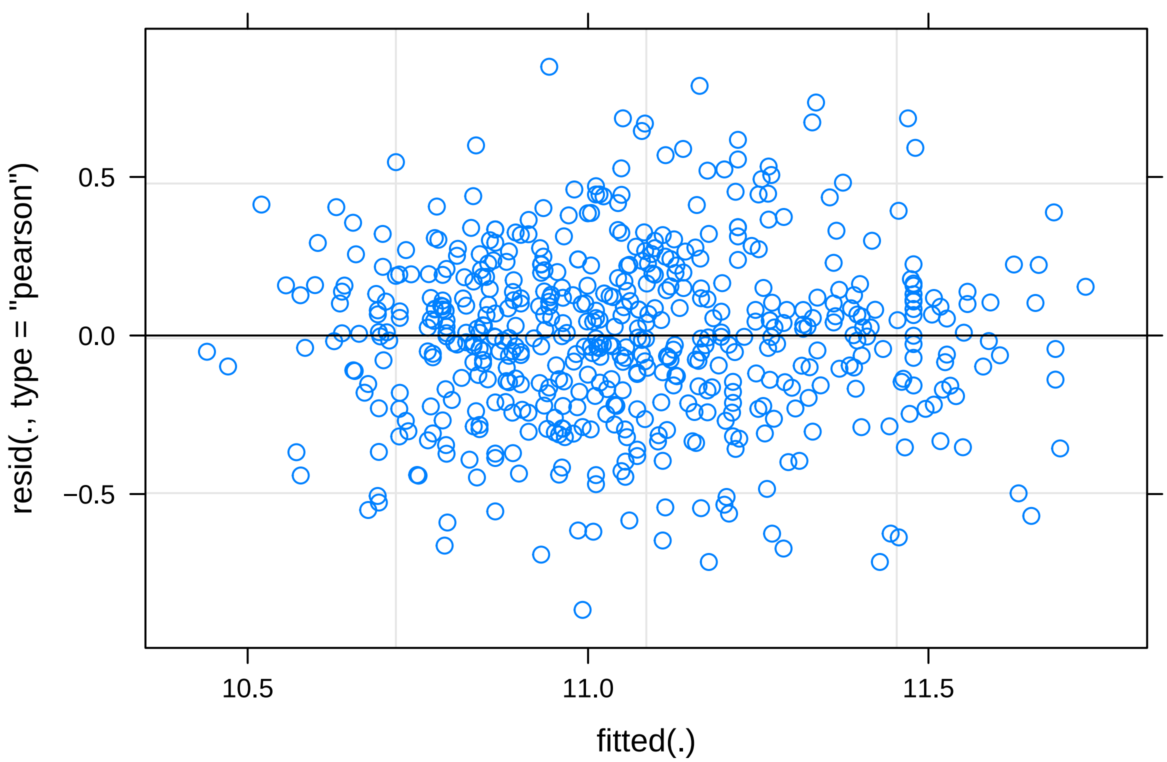
```
## Loading required package: Matrix

## Warning: package 'Matrix' was built under R version 4.1.1

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```r
df$log_price <- log(df$price)

log_price_mixed <- lmer(log_price ~ log(lotsize) + (1 | stories), data = df)
summary(log_price_mixed)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log_price ~ log(lotsize) + (1 | stories)
##    Data: df
##
## REML criterion at convergence: 155
```

```
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1647 -0.6112  0.0263  0.5902  3.0997
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  stories  (Intercept) 0.03527  0.1878
##  Residual             0.07477  0.2734
## Number of obs: 546, groups:  stories, 4
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)   6.80358    0.27618   24.64
## log(lotsize)  0.51422    0.03037   16.93
##
## Correlation of Fixed Effects:
##             (Intr)
## log(lotsiz) -0.939
```

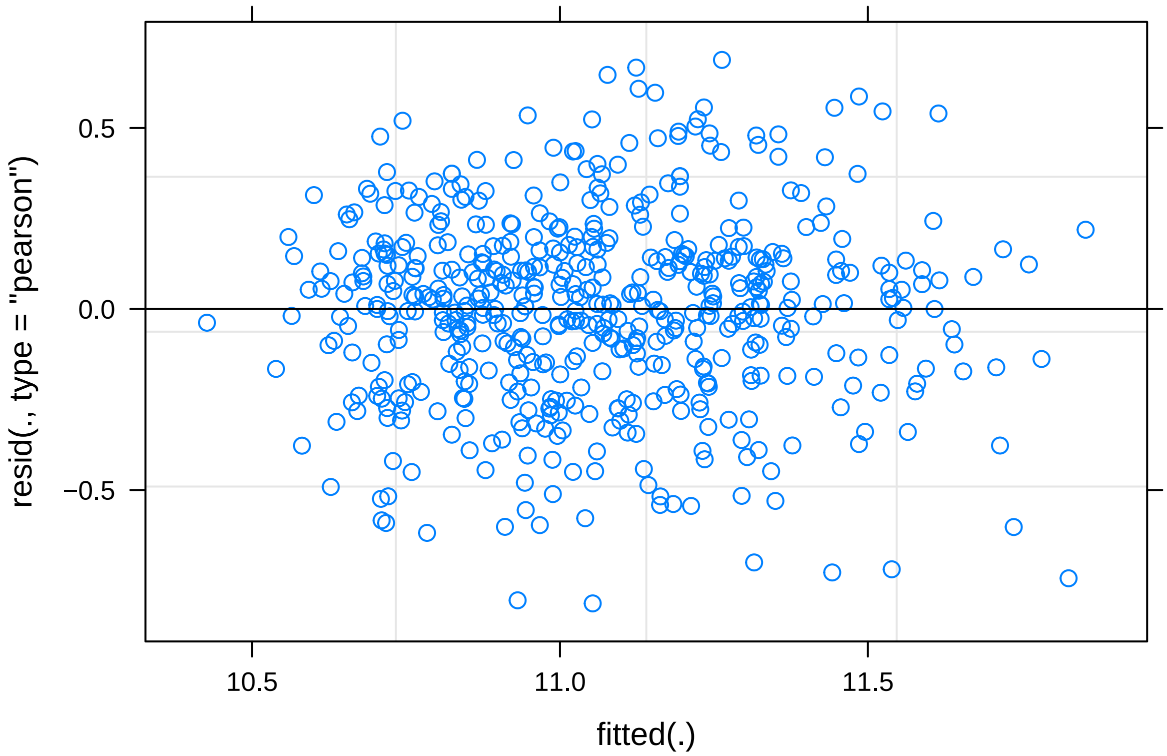
```
plot(log_price_mixed)
```




```
# cluster level covariates
log_price_mixed_cluster <- lmer(log_price ~ log(lotsize) + bedrooms + garage +
                                  (1 | stories), data = df)
summary(log_price_mixed_cluster)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log_price ~ log(lotsize) + bedrooms + garage + (1 | stories)
##    Data: df
##
## REML criterion at convergence: 114.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.11627 -0.61709  0.05699  0.57747  2.63679
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  stories  (Intercept) 0.02673  0.1635
##  Residual             0.06810  0.2610
## Number of obs: 546, groups:  stories, 4
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)   7.21582    0.27370  26.364
## log(lotsize)  0.42842    0.03132  13.677
## bedrooms      0.08631    0.01808   4.773
## garage        0.07351    0.01407   5.226
##
## Correlation of Fixed Effects:
##             (Intr) lg(lt) bedrms
## log(lotsiz) -0.932
## bedrooms    -0.042 -0.162
## garage       0.302 -0.325 -0.105
```

```
plot(log_price_mixed_cluster)
```




```r
# mixed random effects
log_price_mixed_random <- lmer(log_price ~ log(lotsize) +
                               (1 + log(lotsize) | stories), data = df)
```

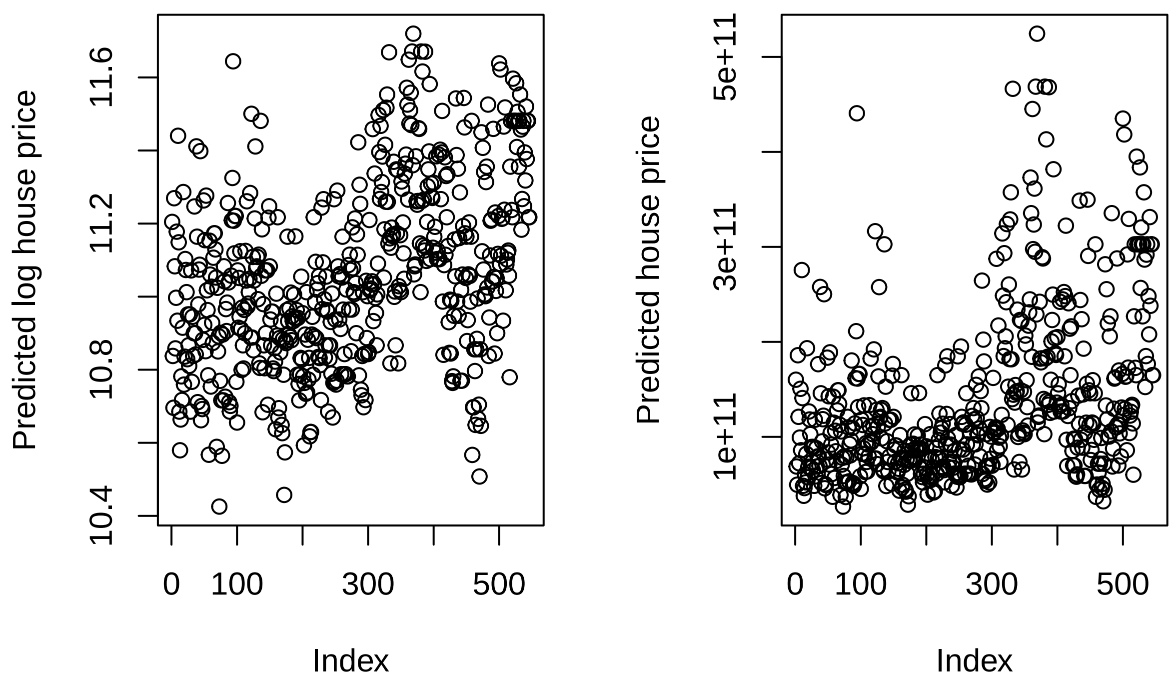
```
## boundary (singular) fit: see ?isSingular
```

```r
summary(log_price_mixed_random)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log_price ~ log(lotsize) + (1 + log(lotsize) | stories)
##    Data: df
##
## REML criterion at convergence: 154.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1694 -0.6077  0.0255  0.5733  3.0940
##
## Random effects:
##  Groups   Name         Variance  Std.Dev. Corr
##  stories  (Intercept)  0.1755055 0.41893
##           log(lotsize) 0.0007101 0.02665  -1.00
##  Residual              0.0747226 0.27335
## Number of obs: 546, groups:  stories, 4
##
```

```
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept)    6.95028    0.33704   20.62
## log(lotsize)  0.49731    0.03355   14.82
##
## Correlation of Fixed Effects:
##              (Intr)
## log(lotsiz) -0.964
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

```
predicted_log_price <- predict(log_price_mixed_random, type="response")

# back-transform from log into original unit
back_transform <- function(data) {
  return(10**data)
}

predicted_price <- sapply(predicted_log_price, back_transform)

par(mfrow=c(1,2))
plot(predicted_log_price, ylab="Predicted log house price")
plot(predicted_price, ylab="Predicted house price")
```



Binary/Poisson Mixed Model

```
data_binary <- read.csv('HousePrices.csv')
```