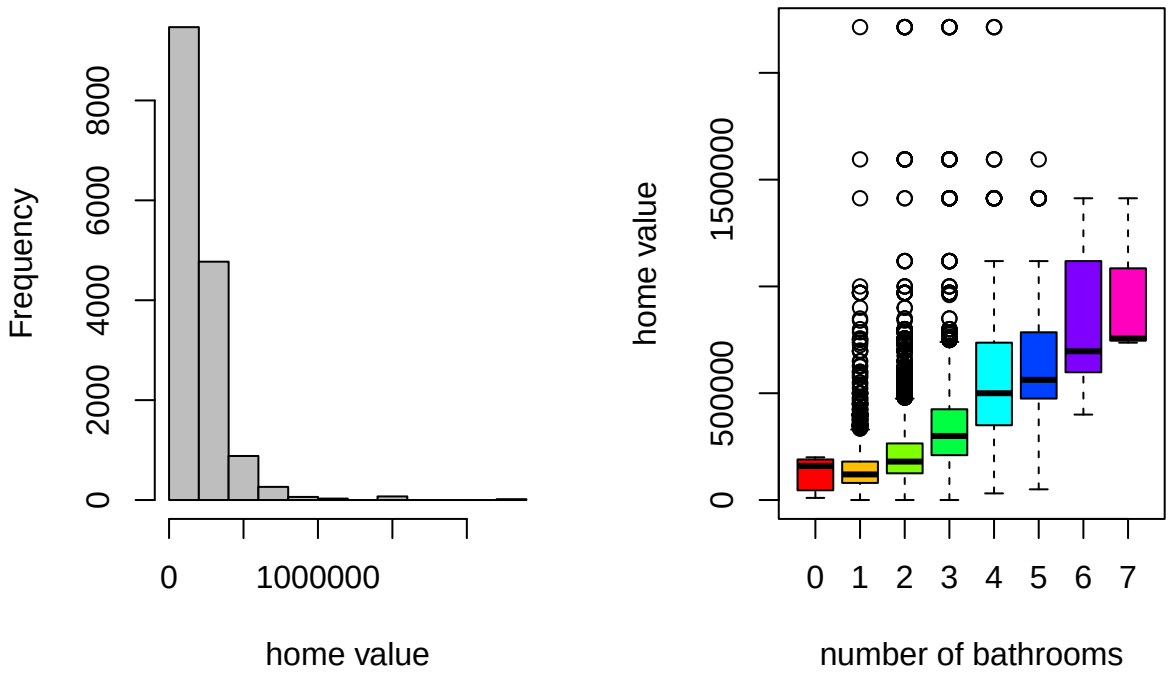
BUS 41201 Homework 2 Assignment

Shihan Ban, Yi Cao, Shri Lekkala, Ningxin Zhang

2 April 2024

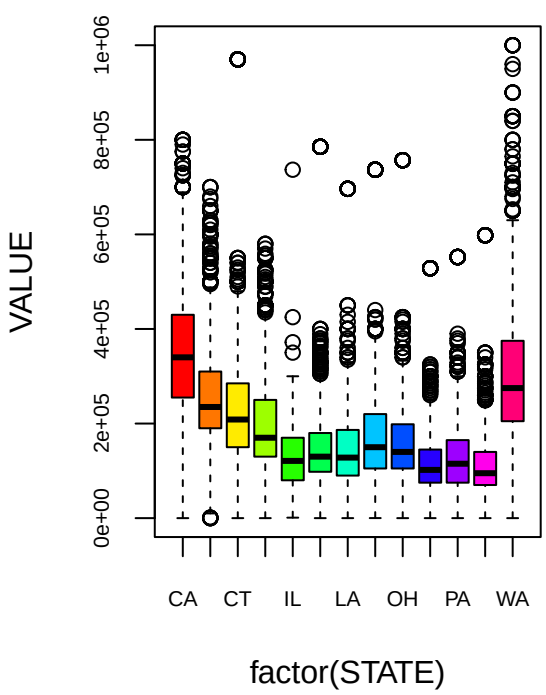
Setup

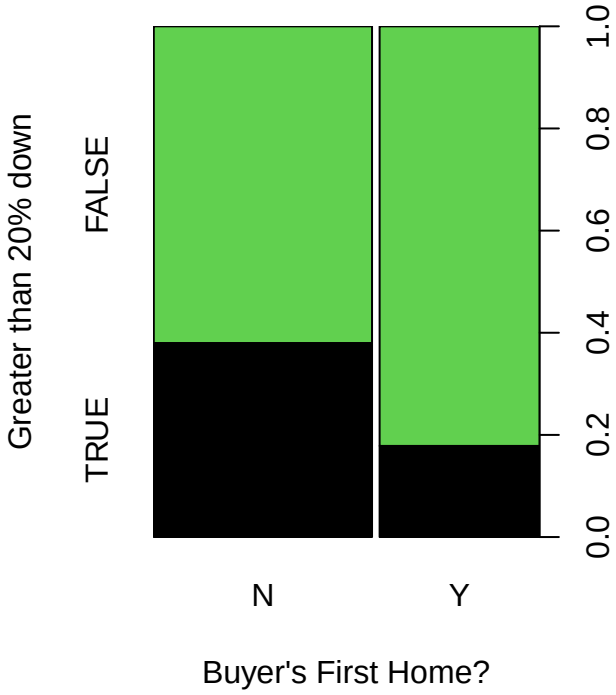
```
library(knitr) # library for markdown output
# Set so that long lines in R will be wrapped:
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=80), tidy=TRUE)

##### ******** Mortgage and Home Sales Data ******** #####

## Read in the data

homes <- read.csv("homes2004.csv")

# conditional vs marginal value


par(mfrow=c(1,2)) # 1 row, 2 columns of plots

hist(homes$VALUE, col="grey", xlab="home value", main="")

plot(VALUE ~ factor(BATHS),
    col=rainbow(8), data=homes[homes$BATHS<8,],
    xlab="number of bathrooms", ylab="home value")
```



```r
# create a var for down payment being greater than 20%
homes$gt20dwn <- factor(0.2<(homes$LPRICE-homes$AMMORT)/homes$LPRICE)
```

```r
# You can try some quick plots.  Do more to build your intuition!

par(mfrow = c(1, 2))
plot(VALUE ~ factor(STATE), data = homes, col = rainbow(nlevels(factor(homes$STATE))),
    ylim = c(0, 10^6), cex.axis = 0.65)
plot(gt20dwn ~ factor(FRSTHO), data = homes, col = c(1, 3), xlab = "Buyer's First Home?",
    ylab = "Greater than 20% down")
```





Question 1

Regress log price onto all variables but mortgage.
What is the R2? How many coefficients are used in this model and how many are significant at 10% FDR?
Re-run regression with only the significant covariates, and compare R2 to the full model. (2 points)

```
# First convert all non-numeric columns in 'homes' to factors
homes = lapply(homes, function(x) if (!is.numeric(x)) factor(x) else x)

# Convert 'homes' back to a data frame, as lapply returns a list
homes = as.data.frame(homes)
```

```
# regress log(PRICE) on everything except AMMORT
pricey <- glm(log(LPRICE) ~ . - AMMORT, data = homes)

# Extract R-squared value the summary
summary_pricey <- summary(pricey)
R2_reduced = 1 - summary_pricey$deviance/summary_pricey$null.deviance
R2_reduced
```

```
## [1] 0.4565419
```

So the R2 score is 0.4565419.

```
# extract pvalues
pvals <- summary(pricey)$coef[-1, 4]
length(pvals)
```

```
## [1] 42
```

So there are 42 coefficients in this model.

```
# Find the p-value cutoff at the 10% FDR level

# To find the p-value cut off we first order the p values
pvals_ordered <- pvals[order(pvals, decreasing = F)]

# Next we use the function fdr_cut function defined in class class to find the
# cutoff at level 0.1
fdr_cut <- function(pvals, q) {
    pvals <- pvals[!is.na(pvals)]
    N <- length(pvals)
    k <- rank(pvals, ties.method = "min")
    alpha <- max(pvals[pvals <= (q * k/N)])
    return(alpha)
}

p_cutoff = fdr_cut(pvals_ordered, q = 0.1)
p_cutoff
```

```
## [1] 0.03792594
```

```r
# Find the number of significant coefficients at this level
sum(pvals < p_cutoff)
```

```
## [1] 36
```

So out of the 42 coefficients, 36 are significant at the 10% FDR level.

```r
# Extract significant coefficients
significant_covariates = names(pvals)[pvals < p_cutoff]
significant_covariates
```

```
##  [1] "EAPTBLY"          "ECOM2Y"           "EGREENY"          "EJUNKY"
##  [5] "ELOW1Y"           "ESFDY"            "EABANY"           "HOWHgood"
##  [9] "HOWNgood"         "ODORAY"           "STRNAY"           "ZINC2"
## [13] "PER"              "ZADULT"           "HHGRADBach"       "HHGRADGrad"
## [17] "HHGRADHS Grad"    "HHGRADNo HS"      "INTW"             "METROurban"
## [21] "STATEGA"          "STATEIL"          "STATEIN"          "STATELA"
## [25] "STATEMO"          "STATEOH"          "STATEOK"          "STATEPA"
## [29] "STATETX"          "STATEWA"          "BATHS"            "MATBUYY"
## [33] "DWNPAYprev home" "VALUE"            "FRSTHOY"          "gt20dwnTRUE"
```

As there are covariates that correspond to factors, we extract only the relevant variable names and use them for our reduced model.

```r
# Get the names of significant variables in the dataset
significant_vars = c("EAPTBL", "ECOM2", "EGREEN", "EJUNK", "ELOW1", "ESFD", "EABAN",
    "HOWH", "HOWN", "ODORA", "STRNA", "ZINC2", "PER", "ZADULT", "HHGRAD", "INTW",
    "METRO", "STATE", "BATHS", "MATBUY", "DWNPAY", "VALUE", "FRSTHO", "gt20dwn")

# Construct the formula for the reduced model
reduced_formula_str = paste("log(LPRICE)", "~", paste(significant_vars, collapse = " + "))

# Rerun the regression with the significant covariates
reduced_model = glm(reduced_formula_str, data = homes)

# Extract R-squared value the summary
summary_reduced_model = summary(reduced_model)
R2_reduced = 1 - summary_reduced_model$deviance/summary_reduced_model$null.deviance
R2_reduced
```

```
## [1] 0.4563139
```

So the R2 score for the reduced model is 0.4563139.

Which is slightly less than the R2 score of the full model (which was 0.4565419), this is expected as our reduced model has fewer covariates than the full model.

Question 2

Fit a regression for whether the buyer had more than 20 percent down (onto everything but AMMORT and LPRICE).

Interpret effects for Pennsylvania state, 1st home buyers and the number of bathrooms.

Add and describe an interaction between 1st home-buyers and the number of baths. (2 points)

```
# - don't forget family='binomial'!  - use +A*B in formula to add A interacting
# with B Fit the logistic regression model excluding AMMORT and LPRICE
down_payment_model <- glm(gt20dwn ~ . - AMMORT - LPRICE + FRSTHO * BATHS, family = "binomial",
    data = homes)

# Summary of the model to interpret coefficients
summary_down_payment_model <- summary(down_payment_model)

# Print the summary to interpret effects
summary_down_payment_model
```

```
##
## Call:
## glm(formula = gt20dwn ~ . - AMMORT - LPRICE + FRSTHO * BATHS,
##     family = "binomial", data = homes)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.378e+00  1.851e-01  -7.444 9.76e-14 ***
## EAPTBLY          1.217e-02  7.020e-02   0.173 0.862337
## ECOM1Y          -1.608e-01  5.806e-02  -2.770 0.005612 **
## ECOM2Y          -3.181e-01  1.598e-01  -1.991 0.046511 *
## EGREENY         -2.305e-03  3.987e-02  -0.058 0.953900
## EJUNKY          -5.332e-03  1.606e-01  -0.033 0.973520
## ELOW1Y           4.950e-02  6.627e-02   0.747 0.455066
## ESFDY           -2.715e-01  8.276e-02  -3.280 0.001036 **
## ETRANSY         -6.147e-02  7.612e-02  -0.808 0.419333
## EABANY          -9.206e-02  1.155e-01  -0.797 0.425505
## HOWHgood        -1.324e-01  7.938e-02  -1.668 0.095245 .
## HOWNgood         1.630e-01  6.728e-02   2.423 0.015399 *
## ODORAY           1.022e-01  9.804e-02   1.043 0.297090
## STRNAY          -9.672e-02  4.736e-02  -2.042 0.041136 *
## ZINC2           -1.479e-07  1.897e-07  -0.780 0.435530
## PER             -1.266e-01  1.859e-02  -6.811 9.67e-12 ***
## ZADULT           2.195e-02  3.193e-02   0.687 0.491817
## HHGRADBach       1.818e-01  6.597e-02   2.755 0.005863 **
## HHGRADGrad       2.770e-01  7.294e-02   3.797 0.000146 ***
## HHGRADHS Grad   -1.967e-02  6.374e-02  -0.309 0.757647
## HHGRADNo HS     -7.767e-02  9.837e-02  -0.790 0.429774
## NUNITS           2.284e-03  1.415e-03   1.613 0.106646
## INTW            -6.421e-02  1.371e-02  -4.684 2.81e-06 ***
## METROurban      -8.407e-02  5.391e-02  -1.560 0.118848
## STATECO         -3.523e-02  8.516e-02  -0.414 0.679103
## STATECT          7.739e-01  8.837e-02   8.758  < 2e-16 ***
## STATEGA         -2.317e-01  9.489e-02  -2.441 0.014636 *
## STATEIL          5.738e-01  1.635e-01   3.509 0.000450 ***
```

```
## STATEIN          2.367e-01  9.369e-02   2.526 0.011534 *
## STATELA          5.893e-01  1.079e-01   5.464 4.66e-08 ***
## STATEMO          5.194e-01  9.749e-02   5.328 9.95e-08 ***
## STATEOH          7.505e-01  9.493e-02   7.906 2.66e-15 ***
## STATEOK          1.174e-01  1.029e-01   1.141 0.253976
## STATEPA          5.816e-01  1.009e-01   5.761 8.34e-09 ***
## STATETX          2.875e-01  1.075e-01   2.675 0.007473 **
## STATEWA          1.535e-01  8.829e-02   1.739 0.082036 .
## BATHS            2.994e-01  3.824e-02   7.829 4.92e-15 ***
## BEDRMS          -2.157e-02  2.913e-02  -0.741 0.458931
## MATBUYY          2.590e-01  3.929e-02   6.592 4.33e-11 ***
## DWNPAYprev home  7.338e-01  4.868e-02  15.073  < 2e-16 ***
## VALUE            1.448e-06  1.458e-07   9.927  < 2e-16 ***
## FRSTHOY         -2.137e-02  1.184e-01  -0.180 0.856799
## BATHS:FRSTHOY   -2.020e-01  6.207e-02  -3.255 0.001135 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 18873  on 15564  degrees of freedom
## Residual deviance: 16958  on 15522  degrees of freedom
## AIC: 17044
##
## Number of Fisher Scoring iterations: 4
```

The negative coefficient for the interaction term (BATHS:FRSTHOY) suggests that for first-time homebuyers, each additional bathroom decreases the log odds of putting down more than 20% by 0.202 compared to buyers who are not purchasing their first home. This could indicate that first-time home buyers are either purchasing less expensive homes with more bathrooms or that the presence of additional bathrooms diminishes their ability or inclination to make larger down payments, perhaps due to the overall higher costs associated with homes that have more bathrooms.

Question 3

Focus only on a subset of homes worth $> 100k$.

```
# this is your training sample
subset_index = which(homes$VALUE > 1e+05)
subset_homes = homes[subset_index, ]
```

Train the full model from Question 1 on this subset.

```
# Train the full model on this subset
full_model_subset = glm(log(LPRICE) ~ . - AMMORT, data = subset_homes)
```

Predict the left-out homes using this model.

```
predicted_log_prices = predict(full_model_subset, newdata = homes[-subset_index,
    ])
```

What is the out-of-sample fit (i.e. R2)?

```
# find the actual out of sample log prices
actual_log_prices = log(homes[-subset_index, ]$LPRICE)

# Use the code ``deviance.R' to compute OOS deviance
source("deviance.R")
OOS_fit = R2(actual_log_prices, predicted_log_prices)
OOS_fit
```

```
## [1] -0.04904513
```

So the out-of-sample fit is -0.04904513.

```
# Null model has just one mean parameter
ybar <- mean(log(homes$LPRICE[-subset_index]))
ybar
```

```
## [1] 10.7779
```

```
D0 <- deviance(y = log(homes$LPRICE[-subset_index]), pred = ybar)
D0
```

```
## [1] 2879.554
```

Explain why you get this value.

A negative R2 model on the out-of-sample data suggests that the model may be overfitting to the data with homes worth >100k.
Thus using the same model for the left-out homes may be unreliable as the model may not necessarily generalize well, as the left-out homes might have different characteristics and patterns compared to the training data.