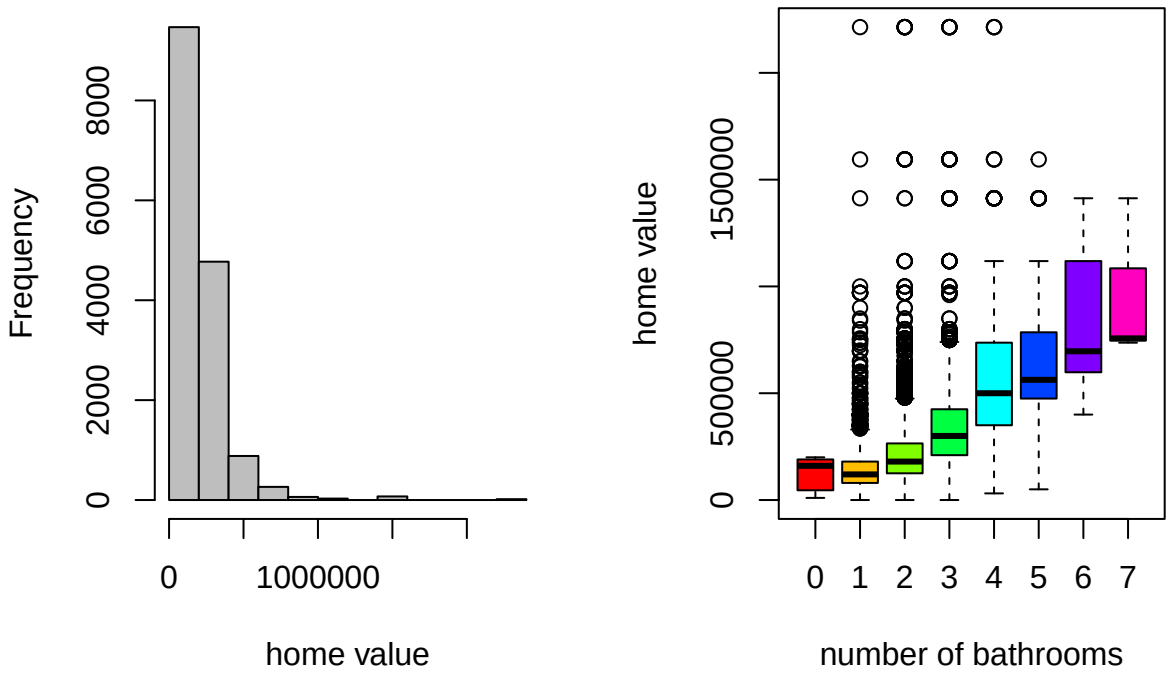
BUS 41201 Homework 2 Assignment

Shihan Ban, Yi Cao, Shri Lekkala, Ningxin Zhang

2 April 2024

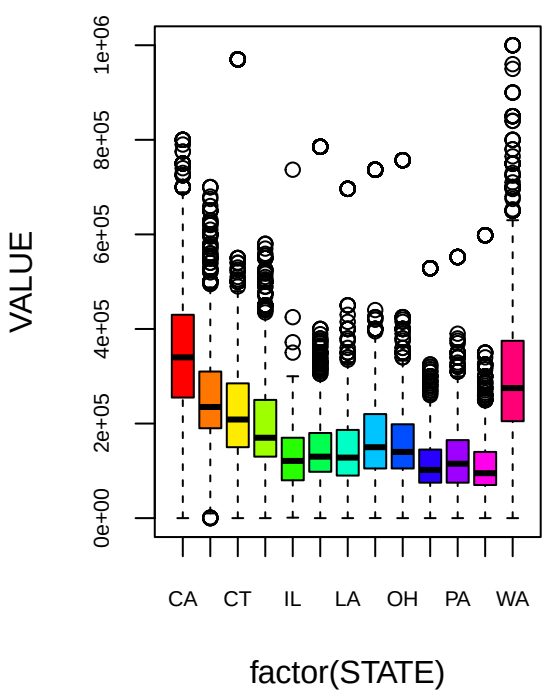
Setup

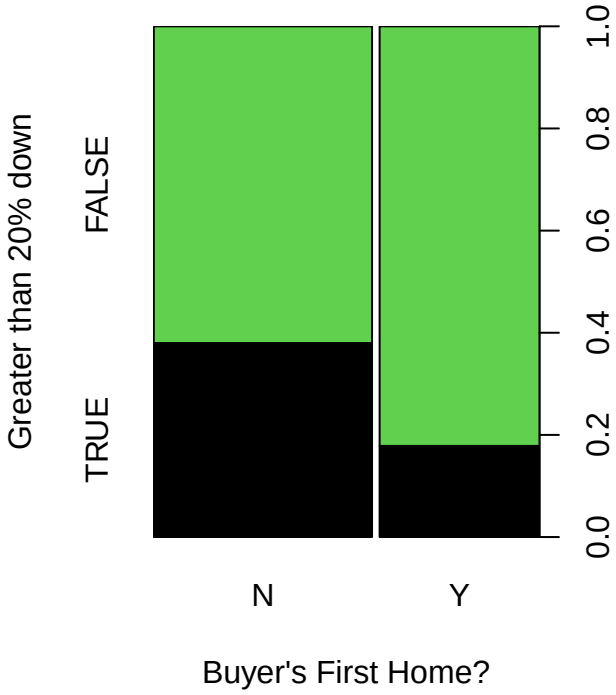
```
library(knitr) # library for markdown output
# Set so that long lines in R will be wrapped:
knitr::opts_chunk$set(tidy.opts=list(width.cutoff=80), tidy=TRUE)

##### ******** 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

```
# 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)
```

Regress log price onto all variables but mortgage.

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

What is the R2?

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

So the R2 score is 0.4565419.

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

How many coefficients are used in this model and how many are significant at 10% FDR?

```
## [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
```

Re-run regression with only the significant covariates, and compare R2 to the full model. (2 points)

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

```
# 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.

```
# 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.

```
# 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
```

```r
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 the logistic regression model excluding AMMORT and LPRICE
down_payment_model = glm(gt20dwn ~ . - AMMORT - LPRICE, family = "binomial", data = homes)
```

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

```
# Summary of the model to interpret coefficients
summary_down_payment_model = summary(down_payment_model)
summary_down_payment_model
```

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

```
##
## Call:
## glm(formula = gt20dwn ~ . - AMMORT - LPRICE, family = "binomial",
##     data = homes)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.293e+00  1.831e-01  -7.065 1.61e-12 ***
## EAPTBLY          1.505e-02  7.025e-02   0.214 0.830424
## ECOM1Y          -1.619e-01  5.809e-02  -2.787 0.005325 **
## ECOM2Y          -3.131e-01  1.600e-01  -1.957 0.050385 .
## EGREENY         -1.569e-03  3.984e-02  -0.039 0.968582
## EJUNKY          -9.697e-03  1.608e-01  -0.060 0.951913
## ELOW1Y           4.635e-02  6.627e-02   0.699 0.484292
## ESFDY           -2.670e-01  8.276e-02  -3.227 0.001252 **
## ETRANSY         -6.270e-02  7.616e-02  -0.823 0.410416
## EABANY          -8.187e-02  1.157e-01  -0.708 0.479137
## HOWHgood        -1.372e-01  7.947e-02  -1.726 0.084398 .
## HOWNgood         1.597e-01  6.730e-02   2.372 0.017669 *
## ODORAY           1.041e-01  9.811e-02   1.061 0.288528
## STRNAY          -9.644e-02  4.737e-02  -2.036 0.041783 *
## ZINC2           -1.277e-07  1.874e-07  -0.682 0.495530
## PER             -1.253e-01  1.855e-02  -6.752 1.46e-11 ***
## ZADULT           1.944e-02  3.188e-02   0.610 0.542024
## HHGRADBach       1.797e-01  6.596e-02   2.725 0.006431 **
## HHGRADGrad       2.729e-01  7.288e-02   3.745 0.000181 ***
## HHGRADHS Grad   -2.064e-02  6.376e-02  -0.324 0.746192
## HHGRADNo HS     -7.246e-02  9.845e-02  -0.736 0.461720
## NUNITS           2.377e-03  1.428e-03   1.664 0.096100 .
## INTW            -6.327e-02  1.372e-02  -4.613 3.98e-06 ***
## METROurban      -8.000e-02  5.389e-02  -1.485 0.137672
## STATECO         -2.513e-02  8.491e-02  -0.296 0.767257
## STATECT          7.870e-01  8.825e-02   8.918  < 2e-16 ***
## STATEGA         -2.223e-01  9.455e-02  -2.351 0.018716 *
```

```
## STATEIL          5.870e-01  1.635e-01   3.590 0.000330 ***
## STATEIN          2.431e-01  9.352e-02   2.599 0.009336 **
## STATELA          5.932e-01  1.077e-01   5.506 3.67e-08 ***
## STATEMO          5.309e-01  9.730e-02   5.456 4.87e-08 ***
## STATEOH          7.642e-01  9.480e-02   8.061 7.59e-16 ***
## STATEOK          1.291e-01  1.027e-01   1.257 0.208850
## STATEPA          6.011e-01  1.007e-01   5.968 2.40e-09 ***
## STATETX          2.935e-01  1.073e-01   2.736 0.006221 **
## STATEWA          1.525e-01  8.819e-02   1.730 0.083717 .
## BATHS            2.445e-01  3.419e-02   7.152 8.57e-13 ***
## BEDRMS          -2.086e-02  2.908e-02  -0.717 0.473120
## MATBUYY          2.587e-01  3.927e-02   6.588 4.45e-11 ***
## DWNPAYprev home  7.417e-01  4.857e-02  15.272  < 2e-16 ***
## VALUE            1.489e-06  1.452e-07  10.256  < 2e-16 ***
## FRSTHOY         -3.700e-01  5.170e-02  -7.156 8.29e-13 ***
## ---
## 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: 16969  on 15523  degrees of freedom
## AIC: 17053
##
## Number of Fisher Scoring iterations: 4
```

```
# Extract the relevant coefficients
coef(summary_down_payment_model)[c("(Intercept)", "STATEPA", "BATHS", "FRSTHOY"),
    ]
```

```
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -1.2934308 0.18308134 -7.064788 1.608619e-12
## STATEPA      0.6010720 0.10071324  5.968152 2.399554e-09
## BATHS        0.2445396 0.03419338  7.151664 8.573193e-13
## FRSTHOY     -0.3699814 0.05170003 -7.156309 8.287814e-13
```

The coefficient of ≈ 0.601 for STATEPA suggests that being in the state of Pennsylvania increases the log-odds of having more than a 20% down payment by approximately 0.601 compared to the baseline.

The coefficient of ≈ 0.245 for BATHS means that each additional bathroom increases the log-odds of having a $>$ 20% down payment by approximately 0.245. By exponentiating the coefficient, $exp(0.245) \approx 1.277$, we can interpret this as meaning that each additional bathroom will need to a 27.7% increase in the odds of having a more than 20% down payment.

Finally the coefficient of $\approx$ -0.370 for FRSTHOY suggests that first home buyers have a lower log odds of having a $>$ 20% down payment than non-first home buyers by about 0.370.

Further, in the summary above, we notice that there are 3 stars (***) for each of the three coefficients above, which suggests that the p values for these covariates are likely to be statistically significant.

```
# - don't forget family='binomial'!  - use +A*B in formula to add A interacting
# with B
```

```r
# Fit the logistic regression model excluding AMMORT and LPRICE, and include an
# interaction term
interaction_model <- glm(gt20dwn ~ . - AMMORT - LPRICE + FRSTHO * BATHS, family = "binomial",
    data = homes)

# Summary of the model to interpret coefficients
summary_interaction_model <- summary(interaction_model)

# Print the summary to interpret effects
summary_interaction_model
```

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

```
##
## 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
```

```
# Extract the relevant coefficients
coef(summary_interaction_model)[c("(Intercept)", "STATEPA", "BATHS", "FRSTHOY", "BATHS:FRSTHOY"),
    ]
```

```
##                  Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)   -1.37788087 0.18509708 -7.4440984 9.760870e-14
## STATEPA        0.58156042 0.10094148  5.7613622 8.343779e-09
## BATHS          0.29940364 0.03824324  7.8289296 4.920411e-15
## FRSTHOY       -0.02136922 0.11842159 -0.1804504 8.567990e-01
## BATHS:FRSTHOY -0.20203156 0.06207417 -3.2546800 1.135202e-03
```

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

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

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

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

Train the full model from Question 1 on this subset.

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

Predict the left-out homes using this model.

```
# Use the code ``deviance.R' to compute OOS deviance
source("deviance.R")

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

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

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

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

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

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

OOS_fit = R2(actual_log_prices, predicted_log_prices)
OOS_fit
```

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

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

Explain why you get this value. Based on the output, we can see the value of r-squared is negative for the out-of-sample data. When the R^2 is negative, it typically indicates that the model's performance is worse than simply using the mean value as the prediction. In this case, it's likely that the model has over-fitted to the initial data, possibly due to the presence of outliers or due to the initial subset having different characteristics compared to the left out data. Additionally, if the model was trained only on homes worth >100k, it may not generalize well to homes worth <100k, leading to poor performance when applied to this subset of data.