Big Data HW 2

2024-03-31

Group 22:

- Panji Al 'Alam
- Misbah Arshad
- Colley Windya Buwana
- Hazna Faiza
- Sagarika Krishnan

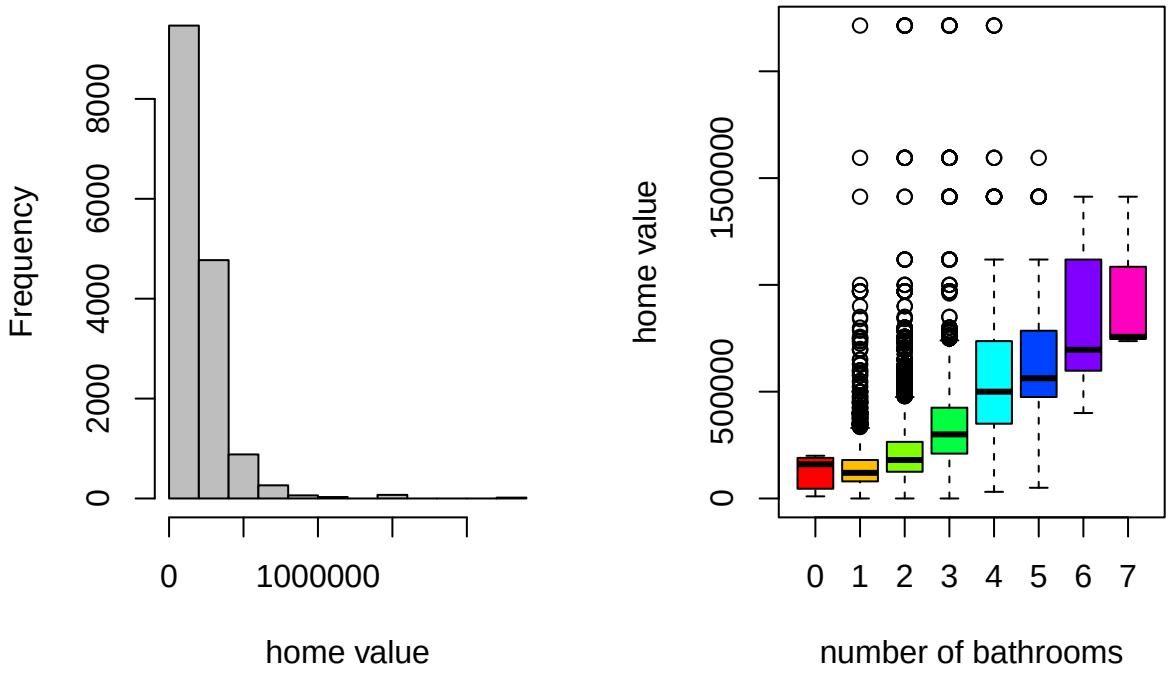
QUESTION 1 Regress log price onto all variables but mortgage. What is the R2? How many coefficients are used in this model and how many are significant at 10% FDR? Re-run regression with only the significant covariates, and compare R2 to the full model. (2 points)

The R2 is 0.447 and there are 41 coefficients used in the model. There are 35 coefficients significant at the FDR of 10% level in this model.

```
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")
```




```
pricey <- glm(log(LPRICE) ~ .-AMMORT, data=homes)
summary(pricey)
```

```
##
## Call:
## glm(formula = log(LPRICE) ~ . - AMMORT, data = homes)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -11.5315   -0.2036    0.0956    0.3492    2.6791
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.108e+01  5.202e-02 213.063  < 2e-16 ***
## EAPTBLY         -5.068e-02  1.954e-02  -2.594 0.009497 **
## ECOM1Y          -3.875e-02  1.603e-02  -2.418 0.015634 *
## ECOM2Y          -1.617e-01  4.002e-02  -4.041 5.35e-05 ***
## EGREENY          4.495e-02  1.167e-02   3.853 0.000117 ***
## EJUNKY          -2.107e-01  4.251e-02  -4.956 7.27e-07 ***
## ELOW1Y           5.584e-02  1.926e-02   2.900 0.003736 **
## ESFDY            7.676e-02  2.463e-02   3.117 0.001832 **
## ETRANSY         -6.172e-03  2.109e-02  -0.293 0.769743
## EABANY          -1.599e-01  2.997e-02  -5.337 9.60e-08 ***
## HOWHgood         6.894e-02  2.192e-02   3.145 0.001664 **
## HOWNgood         9.863e-02  1.827e-02   5.400 6.76e-08 ***
## ODORAY          -8.105e-02  2.758e-02  -2.938 0.003306 **
## STRNAY          -8.550e-02  1.338e-02  -6.389 1.71e-10 ***
## ZINC2            3.962e-07  4.730e-08   8.377  < 2e-16 ***
## PER              7.186e-02  5.208e-03  13.799  < 2e-16 ***
## ZADULT          -1.051e-01  9.060e-03 -11.605  < 2e-16 ***
## HHGRADBach       1.352e-01  1.912e-02   7.072 1.59e-12 ***
## HHGRADGrad       1.561e-01  2.160e-02   7.230 5.06e-13 ***
## HHGRADHS Grad   -7.271e-02  1.808e-02  -4.022 5.79e-05 ***
## HHGRADNo HS     -3.125e-01  2.651e-02 -11.788  < 2e-16 ***
## NUNITS           7.306e-04  4.333e-04   1.686 0.091767 .
## INTW            -7.311e-02  3.681e-03 -19.861  < 2e-16 ***
## METROurban      -3.385e-02  1.511e-02  -2.241 0.025044 *
## STATECO         -4.380e-03  2.460e-02  -0.178 0.858706
## STATECT          8.528e-03  2.629e-02   0.324 0.745628
## STATEGA         -1.030e-01  2.679e-02  -3.844 0.000121 ***
## STATEIL         -3.760e-01  4.868e-02  -7.724 1.20e-14 ***
## STATEIN         -1.668e-01  2.672e-02  -6.243 4.41e-10 ***
## STATELA         -2.491e-01  3.154e-02  -7.899 2.99e-15 ***
## STATEMO         -1.616e-01  2.864e-02  -5.640 1.73e-08 ***
## STATEOH         -1.016e-01  2.800e-02  -3.628 0.000287 ***
## STATEOK         -3.193e-01  2.877e-02 -11.097  < 2e-16 ***
## STATEPA         -4.375e-01  2.920e-02 -14.985  < 2e-16 ***
## STATETX         -3.139e-01  3.010e-02 -10.428  < 2e-16 ***
## STATEWA          1.277e-01  2.580e-02   4.952 7.42e-07 ***
## BATHS            2.027e-01  1.004e-02  20.195  < 2e-16 ***
## BEDRMS           2.878e-03  8.424e-03   0.342 0.732630
## MATBUYY          3.072e-01  1.139e-02  26.969  < 2e-16 ***
## DWNPAYprev home  1.302e-01  1.489e-02   8.745  < 2e-16 ***
## VALUE            1.257e-06  4.078e-08  30.810  < 2e-16 ***
## FRSTHOY         -1.288e-01  1.438e-02  -8.959  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for gaussian family taken to be 0.4629866)
##
##     Null deviance: 13003.4  on 15564  degrees of freedom
## Residual deviance:  7186.9  on 15523  degrees of freedom
## AIC: 32230
##
## Number of Fisher Scoring iterations: 2
```

```
null_deviance <- pricey$null.deviance
resid_deviance <- pricey$deviance

R2 <- 1 - (resid_deviance / null_deviance)
R2
```

```
## [1] 0.447301
```

```
num_coefficients <- length(coefficients(pricey))
print(num_coefficients)
```

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

```
fdr_cut <- function(pvals, q, plotit=FALSE, ...){
  pvals <- pvals[!is.na(pvals)]
  N <- length(pvals)

  k <- rank(pvals, ties.method="min")
  alpha <- max(pvals[ pvals<= (q*k/N) ])

  if(plotit){
    sig <- factor(pvals<=alpha)
    o <- order(pvals)
    plot(pvals[o], col=c("grey60","red")[sig[o]], pch=20, ...,
       ylab="p-values", xlab="tests ordered by p-value", main = paste('FDR =',q))
    lines(1:N, q*(1:N)/N)
  }

  return(alpha)
}

pvals <- summary(pricey)$coef[-1,4]
q <- .1
p <- length(pvals)
cutoff <- fdr_cut(pvals,q)
cutoff
```

```
## [1] 0.02504391
```

```
significant_predictors <- names(pvals)[pvals < cutoff]
length(significant_predictors)
```

```
## [1] 35
```

Re-run the regression with the only significant covariates

```
notsignif_predictors <- names(pvals[pvals >= cutoff])
notsignif_predictors
```

```
## [1] "ETRANSY"    "NUNITS"     "METROurban" "STATECO"    "STATECT"
## [6] "BEDRMS"
```

```
homes_reduced <- homes |>
  select(!c("ETRANS", "NUNITS", "METRO", "BEDRMS"))

pricey_rerun <- glm(log(LPRICE) ~ .-AMMORT, data=homes_reduced)
summary(pricey_rerun)
```

```
##
## Call:
## glm(formula = log(LPRICE) ~ . - AMMORT, data = homes_reduced)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -11.5205   -0.2037    0.0969    0.3508    2.6912
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.109e+01  5.051e-02 219.481  < 2e-16 ***
## EAPTBLY         -5.196e-02  1.928e-02  -2.694 0.007059 **
## ECOM1Y          -4.065e-02  1.587e-02  -2.561 0.010453 *
## ECOM2Y          -1.655e-01  3.963e-02  -4.177 2.97e-05 ***
## EGREENY          4.718e-02  1.159e-02   4.069 4.74e-05 ***
## EJUNKY          -2.131e-01  4.250e-02  -5.015 5.37e-07 ***
## ELOW1Y           5.317e-02  1.914e-02   2.777 0.005490 **
## ESFDY            7.346e-02  2.445e-02   3.005 0.002661 **
## EABANY          -1.653e-01  2.989e-02  -5.531 3.24e-08 ***
## HOWHgood         6.853e-02  2.190e-02   3.129 0.001757 **
## HOWNgood         9.987e-02  1.825e-02   5.471 4.53e-08 ***
## ODORAY          -8.332e-02  2.755e-02  -3.024 0.002501 **
## STRNAY          -8.657e-02  1.334e-02  -6.489 8.90e-11 ***
## ZINC2            3.968e-07  4.730e-08   8.389  < 2e-16 ***
## PER              7.249e-02  5.042e-03  14.376  < 2e-16 ***
## ZADULT          -1.052e-01  9.049e-03 -11.625  < 2e-16 ***
## HHGRADBach       1.348e-01  1.911e-02   7.056 1.78e-12 ***
## HHGRADGrad       1.543e-01  2.156e-02   7.156 8.65e-13 ***
## HHGRADHS Grad   -7.323e-02  1.807e-02  -4.052 5.11e-05 ***
## HHGRADNo HS     -3.154e-01  2.648e-02 -11.909  < 2e-16 ***
## INTW            -7.326e-02  3.680e-03 -19.905  < 2e-16 ***
## STATECO         -3.783e-03  2.458e-02  -0.154 0.877692
## STATECT          1.466e-02  2.617e-02   0.560 0.575203
## STATEGA         -9.912e-02  2.669e-02  -3.713 0.000205 ***
## STATEIL         -3.702e-01  4.860e-02  -7.617 2.75e-14 ***
## STATEIN         -1.745e-01  2.647e-02  -6.592 4.47e-11 ***
## STATELA         -2.513e-01  3.150e-02  -7.978 1.59e-15 ***
## STATEMO         -1.586e-01  2.861e-02  -5.544 3.01e-08 ***
## STATEOH         -9.763e-02  2.784e-02  -3.507 0.000454 ***
## STATEOK         -3.284e-01  2.846e-02 -11.539  < 2e-16 ***
## STATEPA         -4.335e-01  2.911e-02 -14.892  < 2e-16 ***
## STATETX         -3.295e-01  2.922e-02 -11.276  < 2e-16 ***
## STATEWA          1.278e-01  2.579e-02   4.958 7.22e-07 ***
## BATHS            2.058e-01  9.327e-03  22.068  < 2e-16 ***
## MATBUYY          3.067e-01  1.138e-02  26.966  < 2e-16 ***
## DWNPAYprev home  1.315e-01  1.489e-02   8.831  < 2e-16 ***
```

```
## VALUE            1.250e-06  4.042e-08  30.927  < 2e-16 ***
## FRSTHOY         -1.299e-01  1.436e-02  -9.044  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.4631046)
##
##     Null deviance: 13003.4  on 15564  degrees of freedom
## Residual deviance:  7190.6  on 15527  degrees of freedom
## AIC: 32230
##
## Number of Fisher Scoring iterations: 2
```

```
null_deviance_re <- pricey_rerun$null.deviance
resid_deviance_re <- pricey_rerun$deviance

R2 <- 1 - (resid_deviance_re / null_deviance_re)
R2
```

```
## [1] 0.4470176
```

After omitting some variables which are not significant, the r-squared become lower (from 0.447 to 0.456). This might be because there are fewer covariates (tests) applied when we re-run the regression as there are less variation can be explained.

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)

```
homes$gt20dwn <- factor(0.2<(homes$LPRICE-homes$AMMORT)/homes$LPRICE)

pricey_dp <- glm(gt20dwn ~ . - AMMORT - LPRICE, data = homes, family = "binomial")
summary(pricey_dp)
```

```
##
## Call:
## glm(formula = gt20dwn ~ . - AMMORT - LPRICE, family = "binomial",
##     data = homes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4502  -0.8084  -0.5985   1.0693   2.4772
##
## 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
```

For the down payment regression:

```
# STATEPA
(exp(0.6011) - 1) * 100
```

```
## [1] 82.41242
```

- Buying a house in Pennsylvania is associated with 0.6011 higher log odds, or 82.41% higher odds, of having more than 20% down payment for a home compared to the baseline state (California). Holding all else constant. We found statistically significant evidence of the relationship between those variables as the p-value is less than the significance level of alpha = 5%.

```
# FRSTHO
(exp(-0.3700) - 1) * 100
```

```
## [1] -30.92657
```

- Being a first-time home buyer is associated with 0.3700 lower log odds, or 30.93% lower odds, of having more than 20% down payment for a home compared to a buyer not buying her/his first house. Holding other variables constant. Its p-value is very small at the 0.05 significance level, meaning it is highly statistically significant

```
# BATHS
(exp(0.2445) - 1) * 100
```

```
## [1] 27.69827
```

- One additional of bathroom is associated with 0.2445 higher log odds, or 27.70% higher odds, of having more than 20% down payment for a home. Holding all else constant. We found statistically significant evidence of the relationship between those variables as the p-value is less than the significance level of alpha = 5%.

–

Add the interaction term

```
pricey_dp2 <- glm(gt20dwn ~ . - AMMORT - LPRICE + FRSTHO*BATHS, data = homes, family = "binomial")
summary(pricey_dp2)
```

```
##
## Call:
## glm(formula = gt20dwn ~ . - AMMORT - LPRICE + FRSTHO * BATHS,
```

```
##     family = "binomial", data = homes)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4400  -0.8054  -0.5974   1.0654   2.4456
##
## 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
```

For the interaction regression:

```
# BATHS:FRSTHO
(exp(0.2994 - 0.202) - 1) * 100
```

```
## [1] 10.23012
```

- The effect of being a first time homeowner on receiving a 20% down payment decreases by 10.23% with every additional bathroom. This is lower compared to when we did not account for the interaction term. The p-value for the coefficients of BATHS and the interaction term is less than the significance level at 0.05 meaning that it is statistically significant.

QUESTION 3 Focus only on a subset of homes worth >100k. Train the full model from Question 1 on this subset. Predict the left-out homes using this model. What is the out-of-sample fit (i.e. R2)? Explain why you get this value. (1 point)

The out of sample R2 is 0.39. Our model explains around 39% of the out of sample data. Although a "high" R2 is a subjective number based on the context, in this case our model doesn't account for everything that goes into the variance of housing prices. It is lower than our R2 in question 1 which makes sense because the out of sample data is explained less by our model than our in sample data.

```
train_homes <- homes[homes$VALUE > 100000, ]
test_homes <- homes[homes$VALUE <= 100000, ]

deviance <- function(y, pred, family=c("gaussian","binomial")){
    family <- match.arg(family)
    if(family=="gaussian"){
        return( sum( (y-pred)^2 ) )
    }else{
        if(is.factor(y)) y <- as.numeric(y)>1
        return( -2*sum( y*log(pred) + (1-y)*log(1-pred) ) )
    }
}

R2 <- function(y, pred, family=c("gaussian","binomial")){
    fam <- match.arg(family)
    if(fam=="binomial"){
        if(is.factor(y)){ y <- as.numeric(y)>1 }
    }
    dev <- deviance(y, pred, family=fam)
    dev0 <- deviance(y, mean(y), family=fam)
    return(1-dev/dev0)
}

# Train the full model
train_reg <- glm(log(LPRICE) ~ .-AMMORT, data=train_homes)
summary(train_reg)
```

```
##
## Call:
## glm(formula = log(LPRICE) ~ . - AMMORT, data = train_homes)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -11.5896   -0.1823    0.0826    0.3144    2.5242
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.131e+01  5.504e-02 205.422  < 2e-16 ***
## EAPTBLY         -2.872e-02  2.093e-02  -1.372 0.170129
## ECOM1Y          -3.188e-02  1.701e-02  -1.874 0.061002 .
## ECOM2Y          -9.633e-02  4.745e-02  -2.030 0.042369 *
## EGREENY          4.134e-02  1.162e-02   3.556 0.000377 ***
## EJUNKY          -1.228e-01  5.060e-02  -2.427 0.015222 *
## ELOW1Y           9.421e-03  1.928e-02   0.489 0.625147
## ESFDY            2.852e-02  2.675e-02   1.066 0.286381
## ETRANSY         -1.085e-03  2.239e-02  -0.048 0.961344
```

```
## EABANY          -6.315e-02  3.799e-02  -1.662 0.096515 .
## HOWHgood         1.809e-02  2.435e-02   0.743 0.457627
## HOWNgood         5.975e-02  1.992e-02   3.000 0.002709 **
## ODORAY          -8.679e-02  3.000e-02  -2.894 0.003815 **
## STRNAY          -6.705e-02  1.397e-02  -4.800 1.61e-06 ***
## ZINC2            3.392e-07  4.308e-08   7.873 3.77e-15 ***
## PER              8.356e-02  5.237e-03  15.958  < 2e-16 ***
## ZADULT          -1.121e-01  9.212e-03 -12.166  < 2e-16 ***
## HHGRADBach       1.267e-01  1.903e-02   6.658 2.90e-11 ***
## HHGRADGrad       1.431e-01  2.116e-02   6.766 1.39e-11 ***
## HHGRADHS Grad   -3.670e-02  1.860e-02  -1.974 0.048457 *
## HHGRADNo HS     -1.774e-01  2.990e-02  -5.933 3.05e-09 ***
## NUNITS           4.627e-04  4.893e-04   0.946 0.344404
## INTW            -6.720e-02  4.340e-03 -15.482  < 2e-16 ***
## METROurban      -1.392e-02  1.608e-02  -0.866 0.386729
## STATECO          7.515e-03  2.231e-02   0.337 0.736258
## STATECT         -4.112e-02  2.427e-02  -1.694 0.090236 .
## STATEGA         -7.813e-02  2.485e-02  -3.144 0.001671 **
## STATEIL         -1.336e-01  5.412e-02  -2.469 0.013574 *
## STATEIN         -1.338e-01  2.615e-02  -5.119 3.13e-07 ***
## STATELA         -2.053e-01  3.216e-02  -6.382 1.81e-10 ***
## STATEMO         -1.078e-01  2.789e-02  -3.866 0.000111 ***
## STATEOH         -1.026e-01  2.707e-02  -3.792 0.000150 ***
## STATEOK         -1.762e-01  3.171e-02  -5.556 2.82e-08 ***
## STATEPA         -3.124e-01  3.118e-02 -10.020  < 2e-16 ***
## STATETX         -1.458e-01  3.402e-02  -4.287 1.82e-05 ***
## STATEWA          1.203e-01  2.342e-02   5.138 2.82e-07 ***
## BATHS            1.705e-01  9.923e-03  17.182  < 2e-16 ***
## BEDRMS          -1.765e-02  8.483e-03  -2.080 0.037528 *
## MATBUYY          2.988e-01  1.143e-02  26.140  < 2e-16 ***
## DWNPAYprev home  7.793e-02  1.464e-02   5.324 1.04e-07 ***
## VALUE            1.046e-06  3.859e-08  27.112  < 2e-16 ***
## FRSTHOY         -1.091e-01  1.486e-02  -7.345 2.18e-13 ***
## gt20dwnTRUE      2.189e-01  1.261e-02  17.352  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.3668366)
##
##     Null deviance: 7300.4  on 12143  degrees of freedom
## Residual deviance: 4439.1  on 12101  degrees of freedom
## AIC: 22330
##
## Number of Fisher Scoring iterations: 2
```

```
null_deviance_train <- train_reg$null.deviance
resid_deviance_train <- train_reg$deviance

R2_train <- 1 - (resid_deviance_train / null_deviance_train)
R2_train
```

```
## [1] 0.3919407
```

```
# Predict the left-out-homes
prediction <- predict(train_reg, newdata=test_homes, type="response")
```

```
head(prediction)
```

```
##       12       13       17       20       29       31
## 11.29185 11.32100 11.73195 10.97023 11.43371 11.33384
```

```
# Calculate the out-of-sample fit
test_homes$pred_log_price <- predict(train_reg, newdata=test_homes, type="response")
test_homes$pred_price <- exp(test_homes$pred_log_price)

R2_test <- R2(test_homes$LPRICE, test_homes$pred_price, family="gaussian")
print(R2_test)
```

```
## [1] 0.2133504
```

The out of sample R2 is 0.213. Our model explains around 21.33% of the out of sample data. It is lower than our r-squared on the training data (39.19%) which makes sense because the out of sample data is explained less by our model than our in sample data. If the model is too complex and tailored to the nuances of the training data (the higher-valued homes in this context), it may not generalize well to unseen data, particularly if that data (lower-valued homes) has different characteristics. The model may capture the noises in the training data which makes it fit the training data very well, but it does not generalize to the new data.