HW2

Yu-Ting Weng, Mengdi Hao, Elena Li, Minji Park, Sarah Lee

2024-03-31

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 following is the regression result of the full model:

```
##
## Call:
## glm(formula = log(LPRICE) ~ . - AMMORT, data = homes)
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.104e+01  5.165e-02 213.825  < 2e-16 ***
## EAPTBLY         -5.093e-02  1.938e-02  -2.629 0.008584 **
## ECOM1Y          -3.300e-02  1.590e-02  -2.076 0.037926 *
## ECOM2Y          -1.529e-01  3.969e-02  -3.852 0.000118 ***
## EGREENY          4.497e-02  1.157e-02   3.887 0.000102 ***
## EJUNKY          -2.110e-01  4.215e-02  -5.005 5.66e-07 ***
## ELOW1Y           5.433e-02  1.909e-02   2.846 0.004440 **
## ESFDY            8.740e-02  2.443e-02   3.577 0.000348 ***
## ETRANSY         -4.083e-03  2.091e-02  -0.195 0.845201
## EABANY          -1.582e-01  2.972e-02  -5.324 1.03e-07 ***
## HOWHgood         7.428e-02  2.174e-02   3.417 0.000635 ***
## HOWNgood         9.363e-02  1.812e-02   5.169 2.39e-07 ***
## ODORAY          -8.504e-02  2.735e-02  -3.109 0.001883 **
## STRNAY          -8.214e-02  1.327e-02  -6.189 6.20e-10 ***
## ZINC2            4.005e-07  4.690e-08   8.539  < 2e-16 ***
## PER              7.617e-02  5.171e-03  14.731  < 2e-16 ***
## ZADULT          -1.058e-01  8.984e-03 -11.775  < 2e-16 ***
## HHGRADBach       1.285e-01  1.897e-02   6.776 1.28e-11 ***
## HHGRADGrad       1.451e-01  2.143e-02   6.771 1.33e-11 ***
## HHGRADHS Grad   -7.215e-02  1.793e-02  -4.025 5.72e-05 ***
## HHGRADNo HS     -3.115e-01  2.629e-02 -11.852  < 2e-16 ***
## NUNITS           6.239e-04  4.297e-04   1.452 0.146513
## INTW            -7.115e-02  3.652e-03 -19.481  < 2e-16 ***
## METROurban      -3.126e-02  1.498e-02  -2.087 0.036941 *
## STATECO         -3.785e-03  2.440e-02  -0.155 0.876722
```

```
## STATECT         -2.293e-02  2.614e-02  -0.877 0.380363
## STATEGA         -9.580e-02  2.657e-02  -3.606 0.000312 ***
## STATEIL         -3.999e-01  4.829e-02  -8.280  < 2e-16 ***
## STATEIN         -1.772e-01  2.650e-02  -6.688 2.34e-11 ***
## STATELA         -2.718e-01  3.131e-02  -8.682  < 2e-16 ***
## STATEMO         -1.832e-01  2.844e-02  -6.442 1.21e-10 ***
## STATEOH         -1.319e-01  2.783e-02  -4.740 2.16e-06 ***
## STATEOK         -3.260e-01  2.853e-02 -11.424  < 2e-16 ***
## STATEPA         -4.613e-01  2.899e-02 -15.914  < 2e-16 ***
## STATETX         -3.265e-01  2.985e-02 -10.935  < 2e-16 ***
## STATEWA          1.223e-01  2.558e-02   4.779 1.78e-06 ***
## BATHS            1.933e-01  9.972e-03  19.380  < 2e-16 ***
## BEDRMS           3.767e-03  8.353e-03   0.451 0.651996
## MATBUYY          2.973e-01  1.131e-02  26.284  < 2e-16 ***
## DWNPAYprev home  9.728e-02  1.491e-02   6.525 7.00e-11 ***
## VALUE            1.191e-06  4.064e-08  29.301  < 2e-16 ***
## FRSTHOY         -1.170e-01  1.428e-02  -8.196 2.68e-16 ***
## gt20dwnTRUE      2.055e-01  1.265e-02  16.246  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.4552749)
##
##     Null deviance: 13003.4  on 15564  degrees of freedom
## Residual deviance:  7066.8  on 15522  degrees of freedom
## AIC: 31969
##
## Number of Fisher Scoring iterations: 2

## R-squared:  0.4565419
```

From the above regression output, this regression has 42 coefficients in total, excluding the intercept. Its R^2 is 1-7066.8/13003.4 = 0.4565419, indicating that about 45.65% of the variability in the logarithm of the price is explain by the model. Out of the 42 coefficients, 37 are significant under a False Discovery Rate (FDR) of 10%.

```
## p-value at FDR 10%:  0.03792594

## number of significant coefficients at FDR 10%:  37

## insignificant variables:  ETRANSY NUNITS STATECO STATECT BEDRMS

##
## Call:
## glm(formula = log(LPRICE) ~ . - AMMORT - ETRANS - NUNITS - BEDRMS,
##     data = homes)
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.105e+01  5.019e-02 220.212  < 2e-16 ***
## EAPTBLY         -4.851e-02  1.921e-02  -2.525 0.011564 *
## ECOM1Y          -3.251e-02  1.578e-02  -2.061 0.039358 *
## ECOM2Y          -1.544e-01  3.931e-02  -3.929 8.58e-05 ***
## EGREENY          4.461e-02  1.155e-02   3.861 0.000114 ***
## EJUNKY          -2.113e-01  4.215e-02  -5.012 5.44e-07 ***
## ELOW1Y           5.291e-02  1.899e-02   2.786 0.005337 **
## ESFDY            8.705e-02  2.427e-02   3.586 0.000336 ***
```

```
## EABANY          -1.591e-01  2.970e-02  -5.358 8.55e-08 ***
## HOWHgood         7.475e-02  2.172e-02   3.441 0.000580 ***
## HOWNgood         9.382e-02  1.811e-02   5.182 2.22e-07 ***
## ODORAY          -8.543e-02  2.733e-02  -3.125 0.001778 **
## STRNAY          -8.219e-02  1.323e-02  -6.210 5.42e-10 ***
## ZINC2            4.008e-07  4.689e-08   8.546  < 2e-16 ***
## PER              7.658e-02  5.010e-03  15.284  < 2e-16 ***
## ZADULT          -1.058e-01  8.972e-03 -11.789  < 2e-16 ***
## HHGRADBach       1.292e-01  1.896e-02   6.814 9.86e-12 ***
## HHGRADGrad       1.457e-01  2.142e-02   6.803 1.06e-11 ***
## HHGRADHS Grad   -7.226e-02  1.792e-02  -4.032 5.55e-05 ***
## HHGRADNo HS     -3.118e-01  2.628e-02 -11.864  < 2e-16 ***
## INTW            -7.119e-02  3.651e-03 -19.497  < 2e-16 ***
## METROurban      -3.119e-02  1.497e-02  -2.084 0.037208 *
## STATECO         -2.994e-03  2.437e-02  -0.123 0.902247
## STATECT         -2.217e-02  2.612e-02  -0.849 0.395961
## STATEGA         -9.502e-02  2.651e-02  -3.584 0.000339 ***
## STATEIL         -3.998e-01  4.828e-02  -8.280  < 2e-16 ***
## STATEIN         -1.772e-01  2.647e-02  -6.694 2.25e-11 ***
## STATELA         -2.715e-01  3.129e-02  -8.677  < 2e-16 ***
## STATEMO         -1.828e-01  2.842e-02  -6.432 1.30e-10 ***
## STATEOH         -1.295e-01  2.767e-02  -4.681 2.88e-06 ***
## STATEOK         -3.258e-01  2.852e-02 -11.422  < 2e-16 ***
## STATEPA         -4.607e-01  2.894e-02 -15.919  < 2e-16 ***
## STATETX         -3.261e-01  2.984e-02 -10.930  < 2e-16 ***
## STATEWA          1.231e-01  2.557e-02   4.812 1.51e-06 ***
## BATHS            1.947e-01  9.308e-03  20.915  < 2e-16 ***
## MATBUYY          2.973e-01  1.130e-02  26.314  < 2e-16 ***
## DWNPAYprev home  9.759e-02  1.491e-02   6.547 6.04e-11 ***
## VALUE            1.192e-06  4.041e-08  29.505  < 2e-16 ***
## FRSTHOY         -1.174e-01  1.426e-02  -8.231  < 2e-16 ***
## gt20dwnTRUE      2.057e-01  1.264e-02  16.270  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.4552534)
##
##     Null deviance: 13003.4  on 15564  degrees of freedom
## Residual deviance:  7067.8  on 15525  degrees of freedom
## AIC: 31965
##
## Number of Fisher Scoring iterations: 2

## R-squared:  0.4564626
```

After removing the insignificant variables ("ETRANS", "NUNITS", "BEDRMS") and re-estimate the model, the R^2 is 1-7067.8/13003.4 = 0.4564626. There is only a very small decrease in R^2, suggesting that the removed variables contributed little to explaining the variability in log-price.

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)

```
##
## 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
```

In the logistic regression model predicting the likelihood of a buyer putting down more than 20% of the purchase price, significant predictors include the state of Pennsylvania ("STATEPA"), whether the buyer is purchasing their first home ("FRSTHOY"), and the number of bathrooms ("BATHS"). The positive coefficient for "STATEPA" (0.6011) indicates that, all else equal, buyers in Pennsylvania are about 1.824 times more likely to make a larger down payment compared to buyers in other states. This might reflect state-specific market conditions or policies that favor or require larger down payments.

Conversely, "FRSTHOY" has a negative coefficient (-0.37), suggesting first-time home buyers are less likely to make a down payment of over 20%, with the odds being about 30.3% lower than repeat buyers (odds multiplier is 0.697). This could be due to first-time buyers having less accumulated wealth or being more cautious with their initial home investment.

The negative coefficient for "BATHS" (-0.2445) implies that as the number of bathrooms increases, the likelihood of making a larger down payment decreases, with the odds being about 21.7% lower than repeat buyers (odds multiplier is 0.783), possibly reflecting higher overall property prices or buyer preferences for more modest homes when making larger down payments.

```
##
## Call:
## glm(formula = gt20dwn ~ . - AMMORT - LPRICE + BATHS * FRSTHO,
##     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 interaction between "FRSTHO" and "BATHS" indicates that the negative impact of having more bathrooms on the likelihood of a larger down payment is further reduced for first-time buyers, with an odds multiplier of exp(-0.202)=0.817, suggesting different purchasing behaviors or financial strategies among this group.

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 following is a regression on the subset data:

```
##
## Call:
## glm(formula = log(LPRICE) ~ . - AMMORT, data = homes[subset,
##     ])
##
```

```
## 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
```

```
## OOS R-squared:  -0.04904513
```

The model trained on homes valued over $100k resulted in an out-of-sample (OOS) R^2 of -0.049. This negative R^2 suggests that the model performs worse on unseen data than a naive model that predicts the average log price for all observations, indicating potential over-fitting to the training data.

Negative R^2 in this context points to the model's limited generalizability. For GLMs, especially with transformed outcomes like log prices, R^2 may not be the most appropriate measure of model performance. Alternative evaluation metrics, such as AIC, BIC, or cross-validation, should be considered to better assess the model's predictive accuracy and to ensure it captures underlying trends rather than noise.