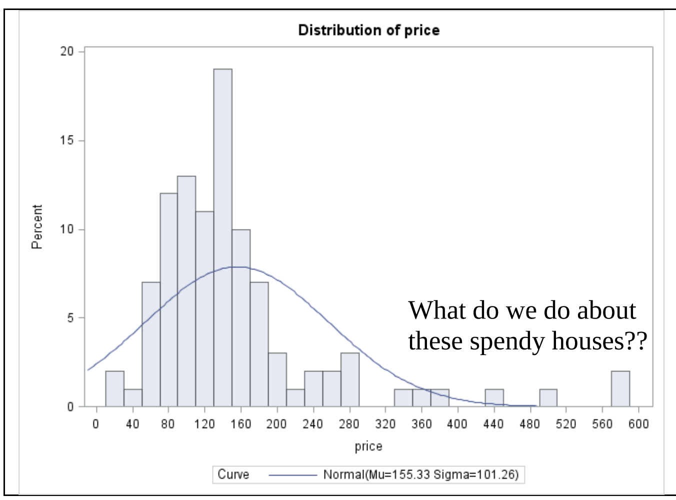
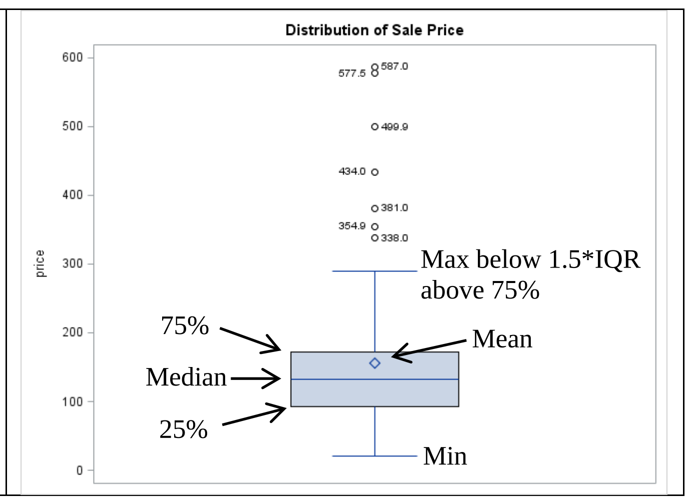
Example 3b: Generalized Linear Models for Positive Skewed Outcomes using SAS and STATA
(complete syntax, data, and output available for SAS and STATA electronically)

The data for this example come from chapter 4 of Agresti (2015) available here: http://users.stat.ufl.edu/~aa/glm/data/ We will be predicting the sale price of 100 homes from four characteristics: whether they are brand new (0=no, 1=yes), square footage in 100s (centered at 1500), number of bedrooms (2, 3, or 4+), and number of bathrooms (1, 2, or 3+). Because this sample's distribution of home sale prices is bounded by 0 and is positively skewed, we will compare four types of generalized linear models estimated using maximum likelihood: identity link with a normal distribution (typical regression), a log-transformed outcome in a typical regression, an identity link with a log-normal distribution, and a log link with a gamma distribution. In addition, because this sample also had several outliers, we will use quantile regression to predict the median home price instead of the mean and to examine predictor effect differences across other percentiles. In SAS GLIMMIX I am not using denominator DF so that the results match those of STATA as closely as possible.

SAS Data Manipulation and Description:

```
* Location for original files for these models – change this path;
%LET filesave= C:\Dropbox\20_PSQF7375_Generalized\PSQF7375_Generalized_Example3b;
LIBNAME filesave "&filesave.";

* Import XLSX data file into SAS;
PROC IMPORT DATAFILE="&filesave.\Houses.xlsx" OUT=work.Example3b DBMS=XLSX REPLACE;
     SHEET="house data"; GETNAMES=YES; RUN;

* Create predictor variables;
DATA work.Example3b; SET work.Example3b;
* Categories for number of bedrooms;
       IF beds=2        THEN DO; bed3vs2=1; bed3vs4=0; END;
  ELSE IF beds=3        THEN DO; bed3vs2=0; bed3vs4=0; END;
  ELSE IF beds IN(4,5) THEN DO; bed3vs2=0; bed3vs4=1; END;
* Categories for number of baths;
       IF baths=1        THEN DO; bath2vs1=1; bath2vs3=0; END;
  ELSE IF baths=2        THEN DO; bath2vs1=0; bath2vs3=0; END;
  ELSE IF baths IN(3,4) THEN DO; bath2vs1=0; bath2vs3=1; END;
* Center and rescale size into per 100 square feet (0=1500); sqft150=(size-1500)/100;
* Log-transform price for demonstration; logprice=LOG(price); RUN;

* Export data to STATA format;
PROC EXPORT DATA=work.Example3b OUTFILE="&filesave.\Example3b.dta" DBMS=STATA REPLACE; RUN;

TITLE "Distribution of Sale Price";
PROC UNIVARIATE DATA=work.Example3b; VAR price size;
     HISTOGRAM price / MIDPOINTS= 0 TO 600 BY 20 NORMAL(MU=EST SIGMA=EST); RUN; QUIT;
PROC SGPLOT DATA=work.Example3b; VBOX price / DATALABEL=price; RUN; TITLE;
```





STATA Data Manipulation and Description:

```
* Import data
use "$filesave\Example3b.dta", clear
* Generate quadratic sqft150 for use in some routines
gen sqft150sq=sqft150*sqft150

* Install lgamma
search lgamma // install from window

display as result "Distribution of Sale Price"
summarize price
hist price, percent start(0) width(20)
graph box price

display as result "Descriptive Stats for Example Variables"
summarize price size
tabulate beds
tabulate baths
tabulate new
```

Every model we fit in this example will have the same linear predictor so that the reference house is old, has 3 bedrooms, 2 bedrooms, and 1500 square feet:

$$\hat{y}_i = \beta_0 + \beta_1(New_i) + \beta_2(Bed3vs2_i) + \beta_3(Bed3vs4_i) + \beta_4(Bath2vs1_i) + \beta_5(Bath2vs3_i) + \beta_6(SqFt_i - 150) + \beta_7(SqFt_i - 150)^2$$

---

1) Two Ways to Predict Original Price Assuming Normal Residuals: $Price_i \sim Normal(\hat{y}_i, \sigma_e^2)$

```
display as result "STATA MIXED: Price using Identity Link, Normal Distribution"
mixed price c.new c.bed3vs2 c.bed3vs4 c.bath2vs1 c.bath2vs3 c.sqft150 ///
          c.sqft150#c.sqft150, ml,
estat ic, n(100),
test (c.new=0) (c.bed3vs2=0) (c.bed3vs4=0) (c.bath2vs1=0) (c.bath2vs3=0) ///
     (c.sqft150=0) (c.sqft150#c.sqft150=0) // Multiv Wald test of model


display as result "STATA GLM: Price using Identity Link, Normal Distribution"
glm price c.new c.bed3vs2 c.bed3vs4 c.bath2vs1 c.bath2vs3 c.sqft150 ///
          c.sqft150#c.sqft150, link(identity) family(gaussian),
estat ic, n(100),
test (c.new=0) (c.bed3vs2=0) (c.bed3vs4=0) (c.bath2vs1=0) (c.bath2vs3=0) ///
     (c.sqft150=0) (c.sqft150#c.sqft150=0) // Multiv Wald test of model


TITLE "SAS MIXED: Price using Identity Link, Normal Distribution";
PROC MIXED DATA=work.Example3b NOCLPRINT NAMELEN=100  METHOD=ML;
     MODEL price = new bed3vs2 bed3vs4 bath2vs1 bath2vs3 sqft150 sqft150*sqft150
                    / SOLUTION;
     CONTRAST "Multiv Wald test of Model" new 1, bed3vs2 1, bed3vs4 1,
               bath2vs1 1, bath2vs3 1, sqft150 1, sqft150*sqft150 1 / CHISQ;
RUN; TITLE;

TITLE "SAS GLIMMIX: Price using Identity Link, Normal Distribution";
PROC GLIMMIX DATA=work.Example3b NOCLPRINT NAMELEN=100 GRADIENT METHOD=MSPL;
     MODEL price = new bed3vs2 bed3vs4 bath2vs1 bath2vs3 sqft150 sqft150*sqft150
                    / SOLUTION DDFM=NONE LINK=IDENTITY DIST=NORMAL;
     CONTRAST "Multiv Wald test of Model" new 1, bed3vs2 1, bed3vs4 1,
               bath2vs1 1, bath2vs3 1, sqft150 1, sqft150*sqft150 1 / CHISQ;
RUN; TITLE;
```

STATA Output from GLM:

```
Generalized linear models                          No. of obs      =        100
Optimization     : ML                              Residual df     =         92
                                                   Scale parameter =   2907.643
Deviance         =  267503.1219                    (1/df) Deviance =   2907.643
Pearson          =  267503.1219                    (1/df) Pearson  =   2907.643 → Um, this is really bad

Variance function: V(u) = 1                        [Gaussian]
Link function    : g(u) = u                        [Identity]
                                                   AIC             =   10.88959
Log likelihood   = -536.4796698                    BIC             =   267079.4
------------------------------------------------------------------------------------
                    |                 OIM
              price |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------------+---------------------------------------------------------------
                new |   59.52165   19.13903     3.11   0.002     22.00984    97.03347
            bed3vs2 |   14.21484    16.4218     0.87   0.387     -17.9713    46.40098
            bed3vs4 |   5.813161    16.4301     0.35   0.723    -26.38925    38.01557
           bath2vs1 |  -6.372286   16.92815    -0.38   0.707    -39.55085    26.80628
           bath2vs3 |  -14.49036   21.53875    -0.67   0.501    -56.70554    27.72481
            sqft150 |   10.02966   1.867685     5.37   0.000     6.369065    13.69026
c.sqft150#c.sqft150 |    .149102   .0906363     1.65   0.100    -.0285419    .3267458
              _cons |   128.1352   7.544411    16.98   0.000     113.3485     142.922
------------------------------------------------------------------------------------
Akaike's information criterion and Bayesian information criterion
-----------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+---------------------------------------------------------------
           . |        100         .  -536.4797       8    1088.959   1109.801
-----------------------------------------------------------------------------
. test (c.new=0) (c.bed3vs2=0) (c.bed3vs4=0) (c.bath2vs1=0) (c.bath2vs3=0) ///
>      (c.sqft150=0) (c.sqft150#c.sqft150=0) // Multiv Wald test of model
           chi2(  7) =  257.13
         Prob > chi2 =    0.0000
```

SAS Output from GLIMMIX:

Fit Statistics	
-2 Log Likelihood	1072.96
AIC (smaller is better)	1090.96
AICC (smaller is better)	1092.96
BIC (smaller is better)	1114.41
CAIC (smaller is better)	1123.41
HQIC (smaller is better)	1100.45
Pearson Chi-Square	267503.1
Pearson Chi-Square / DF	2675.03 → Um, this is really bad (should be 1)

Parameter Estimates

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|	Gradient	
Intercept	128.14	7.2363	Infty	17.71	<.0001	-263E-17	
new	59.5217	18.3575	Infty	3.24	**0.0012**	-113E-18	
bed3vs2	14.2148	15.7512	Infty	0.90	0.3668	-17E-17	
bed3vs4	5.8132	15.7592	Infty	0.37	0.7122	-125E-18	
bath2vs1	-6.3723	16.2369	Infty	-0.39	0.6947	-184E-18	
bath2vs3	-14.4904	20.6592	Infty	-0.70	0.4831	1.58E-16	
sqft150	10.0297	1.7914	Infty	5.60	**<.0001**	2.87E-15	
sqft150*sqft150	0.1491	0.08694	Infty	1.72	0.0863	6.38E-15	
Scale	2675.03	378.31	.	.	.	4.77E-18	→ Residual variance

Contrasts

Label	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
Multiv Wald test of Model	7	Infty	279.49	39.93	<.0001	<.0001

Before interpreting these results, let’s see if we can get better distribution fit… here are two equivalent models:

2) Predict Log-Transformed Price Assuming Normal Residuals: $LogPrice_i \sim Normal(\hat{y}_i, \sigma_e^2)$

```
display as result "STATA: Log-Transformed Price using Identity Link, Normal Distribution"
glm logprice c.new c.bed3vs2 c.bed3vs4 c.bath2vs1 c.bath2vs3 c.sqft150 ///
             c.sqft150#c.sqft150, link(identity) family(gaussian),
estat ic, n(100),
test (c.new=0) (c.bed3vs2=0) (c.bed3vs4=0) (c.bath2vs1=0) (c.bath2vs3=0) ///
     (c.sqft150=0) (c.sqft150#c.sqft150=0) // Multiv Wald test of model

TITLE "SAS: Log-Transformed Price using Identity Link, Normal Distribution";
PROC GLIMMIX DATA=work.Example3b NOCLPRINT NAMELEN=100 GRADIENT METHOD=MSPL;
     MODEL logprice = new bed3vs2 bed3vs4 bath2vs1 bath2vs3 sqft150 sqft150*sqft150
                    / SOLUTION DDFM=NONE LINK=IDENTITY DIST=NORMAL;
     CONTRAST "Multiv Wald test of Model" new 1, bed3vs2 1, bed3vs4 1,
               bath2vs1 1, bath2vs3 1, sqft150 1, sqft150*sqft150 1 / CHISQ; RUN; TITLE;
```

3) Predict Price Assuming Log-Normal Residuals: $Price_i \sim Lognormal(\hat{y}_i, \sigma_e^2)$ (not readily in Stata)

```
TITLE "SAS: Price using Identity Link, Log-Normal Distribution";
PROC GLIMMIX DATA=work.Example3b NOCLPRINT NAMELEN=100 GRADIENT METHOD=MSPL;
     MODEL price = new bed3vs2 bed3vs4 bath2vs1 bath2vs3 sqft150 sqft150*sqft150
                    / SOLUTION DDFM=NONE LINK=IDENTITY DIST=LOGNORMAL;
     CONTRAST "Multiv Wald test of Model" new 1, bed3vs2 1, bed3vs4 1,
               bath2vs1 1, bath2vs3 1, sqft150 1, sqft150*sqft150 1 / CHISQ; RUN; TITLE;
```

STATA Output:

```
Generalized linear models                          No. of obs      =        100
Optimization     : ML                              Residual df     =         92
                                                   Scale parameter =    .1180992
Deviance         =  10.86512691                    (1/df) Deviance =    .1180992
Pearson          =  10.86512691                    (1/df) Pearson  =    .1180992 → Much better!
Variance function: V(u) = 1                        [Gaussian]
Link function    : g(u) = u                        [Identity]
                                                   AIC             =    .7782652
Log likelihood   = -30.91325871                    BIC             =  -412.8105
------------------------------------------------------------------------------------
                    |                 OIM
           logprice |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------------+---------------------------------------------------------------
                new |   .2391816   .1219756     1.96   0.050     .0001139    .4782494
            bed3vs2 |   .1539675   .1046583     1.47   0.141    -.051159     .3590941
            bed3vs4 |   .0129776   .1047112     0.12   0.901    -.1922526    .2182079
           bath2vs1 |  -.1455129   .1078853    -1.35   0.177    -.3569643    .0659385
           bath2vs3 |  -.0561447   .1372693    -0.41   0.683    -.3251876    .2128982
            sqft150 |   .0795194   .011903      6.68   0.000     .0561899    .1028488
c.sqft150#c.sqft150 |  -.0012611   .0005776    -2.18   0.029    -.0023933   -.000129
              _cons |   4.814402   .0480815   100.13   0.000     4.720164    4.90864
                        Note that scale factor is provided up above instead of here...
------------------------------------------------------------------------------------
Akaike's information criterion and Bayesian information criterion
-----------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+---------------------------------------------------------------
           . |        100         .  -30.91326       8    77.82652   98.66788
-----------------------------------------------------------------------------

. test (c.new=0) (c.bed3vs2=0) (c.bed3vs4=0) (c.bath2vs1=0) (c.bath2vs3=0) ///
>      (c.sqft150=0) (c.sqft150#c.sqft150=0) // Multiv Wald test of model
           chi2(  7) =  172.69
         Prob > chi2 =    0.0000
```

SAS's Output is exactly the same either way:

```
           Fit Statistics
-2 Log Likelihood            61.83
AIC  (smaller is better)     79.83
AICC (smaller is better)     81.83
BIC  (smaller is better)    103.27
CAIC (smaller is better)    112.27
```

```
HQIC (smaller is better)        89.32
Pearson Chi-Square              10.87
Pearson Chi-Square / DF          0.11 → Much better!

                               Parameter Estimates
                              Standard
Effect              Estimate     Error      DF    t Value    Pr > |t|    Gradient
Intercept             4.8144   0.04612   Infty     104.39      <.0001     -16E-13
new                   0.2392    0.1170   Infty       2.04      0.0409    4.79E-14
bed3vs2               0.1540    0.1004   Infty       1.53      0.1251    5.51E-14
bed3vs4              0.01298    0.1004   Infty       0.13      0.8972    6.11E-15
bath2vs1             -0.1455    0.1035   Infty      -1.41      0.1597    -103E-15
bath2vs3            -0.05614    0.1317   Infty      -0.43      0.6698    -256E-16
sqft150              0.07952   0.01142   Infty       6.97      <.0001    2.79E-12
sqft150*sqft150     -0.00126  0.000554   Infty      -2.28      0.0228    -257E-13
Scale                 0.1087   0.01537       .          .           .    7.56E-11 → Residual variance

                                          Contrasts
Label                        Num DF   Den DF    Chi-Square     F Value      Pr > ChiSq     Pr > F
Multiv Wald test of Model         7    Infty        187.71       26.82          <.0001     <.0001
```

---

4) Predict Price with Log Link Assuming Gamma Residuals: $Price_i \sim Gamma(\mu, \phi)$, where $\hat{y}_i = Log(\mu)$ and ϕ is a "scale" multiplier of the variance, such that variance = $\mu^2\phi$ (or at least I think that's right).

Stata's GLM does not give the same LL as in SAS for gamma, but here is an "Lgamma" routine that does:

```
display as result "STATA: Price using Log Link, Gamma Distribution"
display as result "Using LGAMMA that does not allow factor variables"
lgamma price new bed3vs2 bed3vs4 bath2vs1 bath2vs3 sqft150 sqft150sq,
estat ic, n(100),
test (new=0) (bed3vs2=0) (bed3vs4=0) (bath2vs1=0) (bath2vs3=0) ///
     (sqft150=0) (sqft150sq) // Multiv Wald test of model

display as result "STATA LGAMMA: Price using Log Link, Gamma Distribution"
display as result "Get Incident-Rate Ratios as exp(slope)"
lgamma price new bed3vs2 bed3vs4 bath2vs1 bath2vs3 sqft150 sqft150sq, eform

TITLE "SAS: Price using Log Link, Gamma Distribution";
PROC GLIMMIX DATA=work.Example3b NOCLPRINT NAMELEN=100 GRADIENT METHOD=MSPL PLOTS=ALL;
     MODEL price = new bed3vs2 bed3vs4 bath2vs1 bath2vs3 sqft150 sqft150*sqft150
                / SOLUTION DDFM=NONE LINK=LOG DIST=GAMMA;
     CONTRAST "Multiv Wald test of Model" new 1, bed3vs2 1, bed3vs4 1,
              bath2vs1 1, bath2vs3 1, sqft150 1, sqft150*sqft150 1 / CHISQ; RUN; TITLE;
```

STATA Output:

```
Log-gamma model                                   Number of obs   =        100
                                                  LR chi2(7)      =     117.57
Log likelihood = -517.21898                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         new |    .204721   .1136043     1.80   0.072    -.0179394    .4273814
     bed3vs2 |   .1728484   .1002319     1.72   0.085    -.0236026    .3692993
     bed3vs4 |   .0218806   .0952913     0.23   0.818    -.1648869    .2086482
    bath2vs1 |  -.1323233   .0999321    -1.32   0.185    -.3281866      .06354
    bath2vs3 |  -.0526695   .1244118    -0.42   0.672    -.2965123    .1911732
     sqft150 |   .0752007   .0111396     6.75   0.000     .0533675    .0970339
   sqft150sq |  -.0009965   .0005487    -1.82   0.069    -.0020719    .0000789
       _cons |   4.854958   .0441468   109.97   0.000     4.768432    4.941484
-------------+----------------------------------------------------------------
     /ln_phi |  -2.298655   .1391173   -16.52   0.000     -2.57132    -2.02599
---------+--------------------------------------------------------------------
         phi |   .1003938   .0139665                      .0764346    .1318632 → scale variance multiplier
------------------------------------------------------------------------------
```

```
Akaike's information criterion and Bayesian information criterion
-----------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+---------------------------------------------------------------
           . |        100  -576.002   -517.219       9    1052.438   1075.884
-----------------------------------------------------------------------------
. test (new=0) (bed3vs2=0) (bed3vs4=0) (bath2vs1=0) (bath2vs3=0) ///
>      (sqft150=0) (sqft150sq) // Multiv Wald test of model
           chi2(  7) =  187.18
         Prob > chi2 =    0.0000
```

SAS Output:

```
           Fit Statistics
-2 Log Likelihood            1034.44
AIC  (smaller is better)     1052.44
AICC (smaller is better)     1054.44
BIC  (smaller is better)     1075.88
CAIC (smaller is better)     1084.88
HQIC (smaller is better)     1061.93
Pearson Chi-Square              9.77
Pearson Chi-Square / DF         0.10 → Still good!

                                Parameter Estimates
                               Standard
Effect               Estimate     Error       DF    t Value     Pr > |t|     Gradient
Intercept              4.8550   0.04415    Infty     109.97       <.0001     -2.67E-7
new                    0.2047    0.1136    Infty       1.80       0.0715     -0.00001
bed3vs2                0.1729    0.1002    Infty       1.72       0.0846     0.000029
bed3vs4               0.02188   0.09529    Infty       0.23       0.8184     -9.69E-6
bath2vs1              -0.1323   0.09993    Infty      -1.32       0.1855     0.000017
bath2vs3             -0.05267    0.1244    Infty      -0.42       0.6720     -4.99E-6
sqft150               0.07520   0.01114    Infty       6.75       <.0001     0.001965
sqft150*sqft150      -0.00100  0.000549    Infty      -1.82       0.0693     -0.02582
Scale                  0.1004   0.01397        .          .            .     -2.65E-6 → phi variance multiplier

                                          Contrasts
                              Num      Den
Label                          DF       DF     Chi-Square     F Value      Pr > ChiSq     Pr > F
Multiv Wald test of Model       7    Infty         187.18       26.74          <.0001     <.0001
```

---

4) Predict Price Median (50th Percentile) instead of Mean using Quantile Regression

Back in intro stat you learned that variables with skewness, outliers, or other kinds of non-normal distributions could be better described using median and interquartile range (i.e., the 50th percentile and the distance from the 25th to 75th percentile) than using the mean and standard deviation. **So why not predict these percentiles instead of the mean using a regression model?** This is the basis of **quantile regression**: the slope estimates are those that minimize a weighted absolute value of the residuals (rather than an unweighted sum of squared residuals as in traditional regression). While the residuals are still assumed to be normal, this is of little consequence because most quantile procedures use some kind of resampling (i.e., bootstrapping in SAS and STATA) to get the standard errors without relying on distributional properties.

```
display as result "STATA: Price 50th Percentile using Quantile Regression"
set seed 8675309 // Set Jenny as seed to get same results each time
sqreg price c.new c.bed3vs2 c.bed3vs4 c.bath2vs1 c.bath2vs3 c.sqft150 ///
              c.sqft150#c.sqft150, quantile(.50),
test (c.new=0) (c.bed3vs2=0) (c.bed3vs4=0) (c.bath2vs1=0) (c.bath2vs3=0) ///
     (c.sqft150=0) (c.sqft150#c.sqft150=0) // Multiv Wald test of model

TITLE "SAS: Price 50th Percentile (Median) using Quantile Regression";
PROC QUANTREG DATA=work.Example3b NAMELEN=100 CI=RESAMPLING(NREP=500);
     MODEL price = new bed3vs2 bed3vs4 bath2vs1 bath2vs3 sqft150 sqft150*sqft150 / QUANTILE=.50;
     Model: TEST new bed3vs2 bed3vs4 bath2vs1 bath2vs3 sqft150 sqft150*sqft150 / WALD; RUN; TITLE;
```

STATA Output:

```
Simultaneous quantile regression                    Number of obs =        100
  bootstrap(500) SEs                                 .50 Pseudo R2 =     0.4523
--------------------------------------------------------------------------------------
                     |              Bootstrap
               price |      Coef.   Std. Err.      t     P>|t|     [95% Conf. Interval]
---------------------+----------------------------------------------------------------
q50              new |   32.16499   29.68706      1.08   0.281    -26.79608    91.12606
             bed3vs2 |   1.077787   19.89456      0.05   0.957    -38.43453    40.59011
             bed3vs4 |  -28.11573   21.71178     -1.29   0.199     -71.2372    15.00574
            bath2vs1 |  -13.73013   14.54949     -0.94   0.348    -42.62668    15.16642
            bath2vs3 |  -1.299234   32.61532     -0.04   0.968    -66.07607     63.4776
             sqft150 |   8.664785   2.330797      3.72   0.000     4.035622    13.29395
c.sqft150#c.sqft150 |   .3827353   .2509158      1.53   0.131    -.1156051    .8810758
               _cons |        133   7.293593     18.24   0.000     118.5143    147.4857
--------------------------------------------------------------------------------------
. test (c.new=0) (c.bed3vs2=0) (c.bed3vs4=0) (c.bath2vs1=0) (c.bath2vs3=0) ///
>       (c.sqft150=0) (c.sqft150#c.sqft150=0) // Multiv Wald test of model
       F(  7,    92) =   10.50 → Note is very different than provided by SAS, but not sure why
            Prob > F =    0.0000
```

SAS Output:

Parameter Estimates Predicting 50th Percentile (Median)

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	133.0000	6.4939	120.1026	145.8974	20.48	<.0001
new	1	32.1650	21.8180	-11.1674	75.4974	1.47	0.1438
bed3vs2	1	1.0778	19.4887	-37.6285	39.7841	0.06	0.9560
bed3vs4	1	-28.1157	18.1543	-64.1716	7.9402	-1.55	0.1249
bath2vs1	1	-13.7301	12.9477	-39.4453	11.9851	-1.06	0.2917
bath2vs3	1	-1.2992	29.3305	-59.5522	56.9538	-0.04	0.9648
sqft150	1	8.6648	2.5004	3.6987	13.6309	3.47	**0.0008**
sqft150*sqft150	1	0.3827	0.1760	0.0332	0.7323	2.17	**0.0322**

Test Model Results

Test	Test Statistic	DF	Chi-Square	Pr > ChiSq
Wald	93.2328	7	93.23	<.0001

→ Translates to F = 93.23/7 = 13.32

For unknown reasons, the multivariate Wald test results differ between SAS and STATA (beyond correcting for F vs. χ^2)

---

4) Predict Price 25th and 75th Percentile using Quantile Regression:

Besides "handling" outliers, another use of quantile regression is to answer research questions about differences at other points of a distribution. Here, we predict the 25th percentile to ask, "among (relatively) cheap houses, what predicts sale price?" Likewise, we predict the 75th percentile to ask, "among (relatively) expensive houses, what predicts sale price?" We can also ask for differences in the predictor effects across these quantiles (e.g., is being a new house more important if the house is expensive than if the house is cheap?), which is analogous to an interaction of the predictor with the quantiles.

```
display as result "STATA: Price 25th and 75th Percentile using Quantile Regression"
set seed 8675309 // Set Jenny as seed to get same results each time
sqreg price c.new c.bed3vs2 c.bed3vs4 c.bath2vs1 c.bath2vs3 c.sqft150 ///
            c.sqft150#c.sqft150, quantile(.25 .75) reps(500),
// Multiv Wald test of model at 25th percentile
test ([q25]c.new=0) ([q25]c.bed3vs2=0) ([q25]c.bed3vs4=0) ([q25]c.bath2vs1=0) ///
     ([q25]c.bath2vs3=0)([q25]c.sqft150=0)([q25]c.sqft150#c.sqft150=0)
// Multiv Wald test of model at 75th percentile
test ([q75]c.new=0) ([q75]c.bed3vs2=0) ([q75]c.bed3vs4=0) ([q75]c.bath2vs1=0) ///
     ([q75]c.bath2vs3=0)([q75]c.sqft150=0)([q75]c.sqft150#c.sqft150=0)
// Multiv Wald test of difference in model between 25th and 75th percentile
test ([q25]c.new=[q75]c.new)([q25]c.bed3vs2=[q75]c.bed3vs2) ///
     ([q25]c.bed3vs4=[q75]c.bed3vs4)([q25]c.bath2vs1=[q75]c.bath2vs1) ///
     ([q25]c.bath2vs3=[q75]c.bath2vs3)([q25]c.sqft150=[q75]c.sqft150) ///
      ([q25]c.sqft150#c.sqft150=[q75]c.sqft150#c.sqft150)
```

```
// Single-predictor difference across quantiles
test ([q25]c.new=[q75]c.new)
display as result "STATA: Price 25-75 Inter-Quantile Regression"
display as result "Model directly predicts predictor slope differences"
set seed 8675309 // Set Jenny as seed to get same results each time
iqreg price c.new c.bed3vs2 c.bed3vs4 c.bath2vs1 c.bath2vs3 c.sqft150 ///
            c.sqft150#c.sqft150, quantile(.25 .75) reps(500)
test (c.new=0) (c.bed3vs2=0) (c.bed3vs4=0) (c.bath2vs1=0) (c.bath2vs3=0) ///
     (c.sqft150=0) (c.sqft150#c.sqft150=0) // Multiv Wald test of differences

TITLE "SAS: Price 50th and 75th Percentile using Quantile Regression";
PROC QUANTREG DATA=work.Example3b NAMELEN=100 CI=RESAMPLING(NREP=500);
     MODEL price = new bed3vs2 bed3vs4 bath2vs1 bath2vs3 sqft150 sqft150*sqft150
                 / QUANTILE=.25 .75;
     EachModel: TEST new bed3vs2 bed3vs4 bath2vs1 bath2vs3 sqft150 sqft150*sqft150 / WALD;
     ModelDiff: TEST new bed3vs2 bed3vs4 bath2vs1 bath2vs3 sqft150 sqft150*sqft150 / QINTERACT;
     newDiff:   TEST new / QINTERACT; * How to test predictor effect across quantiles; RUN; TITLE;
```

STATA Output from SQREG:

```
Simultaneous quantile regression                    Number of obs =        100
  bootstrap(500) SEs                                .25 Pseudo R2 =     0.3747
                                                    .75 Pseudo R2 =     0.5713
--------------------------------------------------------------------------------
                   |              Bootstrap
             price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+------------------------------------------------------------
q25            new |   45.67319    23.32531     1.96   0.053     -.652896    91.99928
           bed3vs2 |        4.7    16.71575     0.28   0.779    -28.49892    37.89892
           bed3vs4 |  -.2206411    21.92028    -0.01   0.992     -43.7562    43.31492
          bath2vs1 |  -.7477554    15.37286    -0.05   0.961    -31.27959    29.78407
          bath2vs3 |   2.397843    33.71776     0.07   0.943    -64.56854    69.36422
           sqft150 |   9.404941    1.757854     5.35   0.000      5.91369    12.89619
c.sqft150#c.sqft150 |   .1068575    .2572658     0.42   0.679    -.4040946    .6178097
             _cons |   101.1147    7.680341    13.17   0.000     85.86092    116.3686
-------------------+------------------------------------------------------------
q75            new |   24.38865    37.27962     0.65   0.515     -49.6519     98.4292
           bed3vs2 |   31.59456    18.98626     1.66   0.100    -6.113803    69.30292
           bed3vs4 |  -31.68683    45.09697    -0.70   0.484    -121.2533    57.87966
          bath2vs1 |  -15.06422    13.74436    -1.10   0.276     -42.3617    12.23326
          bath2vs3 |  -1.257882    43.82478    -0.03   0.977    -88.29768    85.78192
           sqft150 |   10.84037    3.055926     3.55   0.001     4.771038    16.90971
c.sqft150#c.sqft150 |   .3294847     .201842     1.63   0.106     -.071391    .7303603
             _cons |   145.7357    5.484035    26.57   0.000     134.8439    156.6274
--------------------------------------------------------------------------------

. // Multiv Wald test of model at 25th percentile
. test ([q25]c.new=0) ([q25]c.bed3vs2=0) ([q25]c.bed3vs4=0) ([q25]c.bath2vs1=0) ///
>       ([q25]c.bath2vs3=0)([q25]c.sqft150=0)([q25]c.sqft150#c.sqft150=0)
       F(  7,    92) =   12.03
            Prob > F =    0.0000

. // Multiv Wald test of model at 75th percentile
. test ([q75]c.new=0) ([q75]c.bed3vs2=0) ([q75]c.bed3vs4=0) ([q75]c.bath2vs1=0) ///
>       ([q75]c.bath2vs3=0)([q75]c.sqft150=0)([q75]c.sqft150#c.sqft150=0)
       F(  7,    92) =   9.48
            Prob > F =    0.0000

. // Multiv Wald test of difference in model between 25th and 75th percentile
. test ([q25]c.new=[q75]c.new)([q25]c.bed3vs2=[q75]c.bed3vs2) ///
>       ([q25]c.bed3vs4=[q75]c.bed3vs4)([q25]c.bath2vs1=[q75]c.bath2vs1) ///
>       ([q25]c.bath2vs3=[q75]c.bath2vs3)([q25]c.sqft150=[q75]c.sqft150) ///
>           ([q25]c.sqft150#c.sqft150=[q75]c.sqft150#c.sqft150)
       F(  7,    92) =    0.56
            Prob > F =    0.7689

. // Single-predictor difference across quantiles
. test ([q25]c.new=[q75]c.new)
       F(  1,    92) =    0.37
            Prob > F =    0.5470
```

For unknown reasons, the multivariate Wald test results continue to differ between SAS and STATA (beyond correcting for F vs. χ^2)

STATA Output from IQREG—these are the *differences* in predictor slopes across quantiles:

```
.75-.25 Interquantile regression                    Number of obs =        100
  bootstrap(500) SEs                                .75 Pseudo R2 =     0.5713
                                                    .25 Pseudo R2 =     0.3747
-------------------------------------------------------------------------------------
                    |              Bootstrap
              price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------------+----------------------------------------------------------------
                new |  -21.28454     35.214    -0.60   0.547    -91.22259     48.6535
            bed3vs2 |   26.89456   21.00194     1.28   0.204    -14.81711    68.60622
            bed3vs4 |  -31.46618   43.78631    -0.72   0.474    -118.4296    55.49721
           bath2vs1 |  -14.31647   16.65664    -0.86   0.392      -47.398    18.76506
           bath2vs3 |  -3.655725   42.57896    -0.09   0.932    -88.22121    80.90976
            sqft150 |   1.435431   2.880917     0.50   0.619    -4.286319    7.157181
c.sqft150#c.sqft150 |   .2226271   .2837418     0.78   0.435    -.3409085    .7861628
              _cons |   44.62092   8.528189     5.23   0.000      27.6832    61.55864
-------------------------------------------------------------------------------------
test (c.new=0) (c.bed3vs2=0) (c.bed3vs4=0) (c.bath2vs1=0) (c.bath2vs3=0) ///
>      (c.sqft150=0) (c.sqft150#c.sqft150=0) // Multiv Wald test of differences
       F(  7,    92) =    0.56
            Prob > F =    0.7869
```

SAS Output:

Parameter Estimates Predicting 25th percentile

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	101.1147	7.2033	86.8084	115.4211	14.04	<.0001
new	1	45.6732	24.7080	-3.3990	94.7454	1.85	0.0677
bed3vs2	1	4.7000	15.2906	-25.6685	35.0685	0.31	0.7593
bed3vs4	1	-0.2206	18.5831	-37.1283	36.6870	-0.01	0.9906
bath2vs1	1	-0.7478	16.9679	-34.4474	32.9519	-0.04	0.9649
bath2vs3	1	2.3978	40.7497	-78.5345	83.3302	0.06	0.9532
sqft150	1	9.4049	2.3382	4.7611	14.0488	4.02	**0.0001**
sqft150*sqft150	1	0.1069	0.2097	-0.3097	0.5234	0.51	0.6116

Parameter Estimates Predicting 75th percentile

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	145.7357	7.4091	131.0205	160.4508	19.67	<.0001
new	1	24.3886	31.2605	-37.6973	86.4746	0.78	0.4373
bed3vs2	1	31.5946	18.3438	-4.8379	68.0270	1.72	0.0884
bed3vs4	1	-31.6868	40.6147	-112.3511	48.9774	-0.78	0.4373
bath2vs1	1	-15.0642	15.5390	-45.9261	15.7977	-0.97	0.3349
bath2vs3	1	-1.2579	42.7840	-86.2306	83.7149	-0.03	0.9766
sqft150	1	10.8404	3.3255	4.2357	17.4450	3.26	**0.0016**
sqft150*sqft150	1	0.3295	0.2223	-0.1119	0.7709	1.48	0.1416

Test EachModel Results

Quantile Level	Test	Test Statistic	DF	Chi-Square	Pr > ChiSq
0.25	Wald	78.4206	7	78.42	<.0001
0.75	Wald	96.8727	7	96.87	<.0001

Test ModelDiff Results
Equal Coefficients Across Quantiles

Chi-Square	DF	Pr > ChiSq
4.4799	7	0.7231

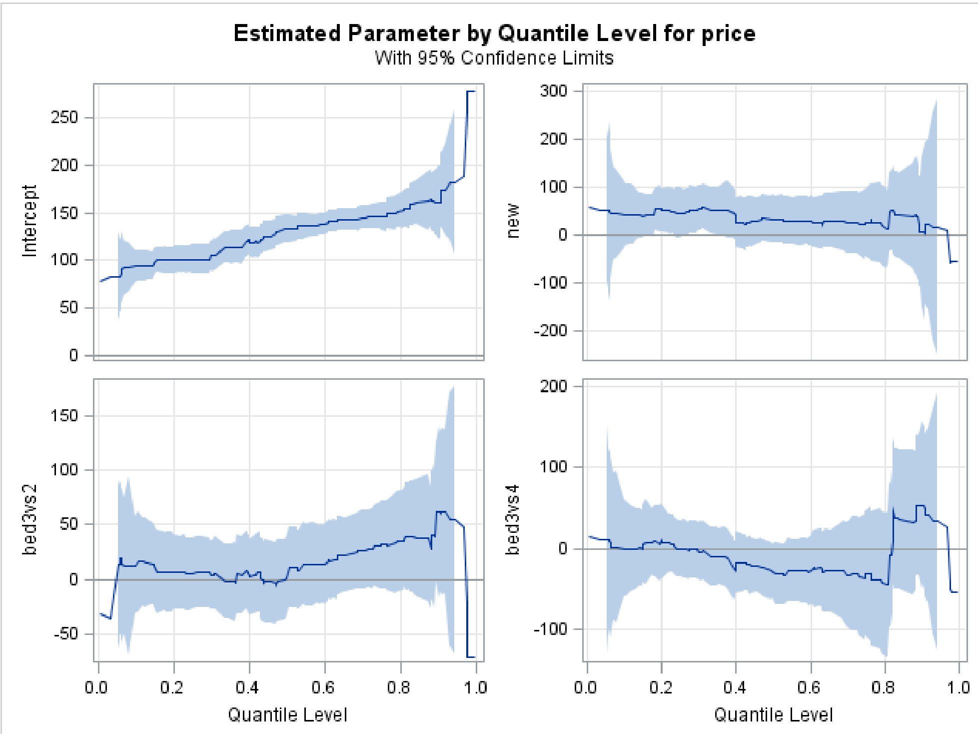
Test newDiff Results
Equal Coefficients Across Quantiles

Chi-Square	DF	Pr > ChiSq
0.3636	1	0.5465

4) Predict Price All Percentiles using Quantile Regression (couldn't find this in STATA):

```
TITLE "SAS: Price All Percentiles using Quantile Regression";
PROC QUANTREG DATA=work.Example3b NAMELEN=100 CI=RESAMPLING(NREP=500);
     MODEL price = new bed3vs2 bed3vs4 bath2vs1 bath2vs3 sqft150 sqft150*sqft150
                       / QUANTILE=PROCESS PLOT=QUANTPLOT SEED=8675309; * Jenny is the random seed;
RUN; TITLE;
```

SAS Output Graphical Summary (lots of voluminous output omitted; is Figure 1 in results section):

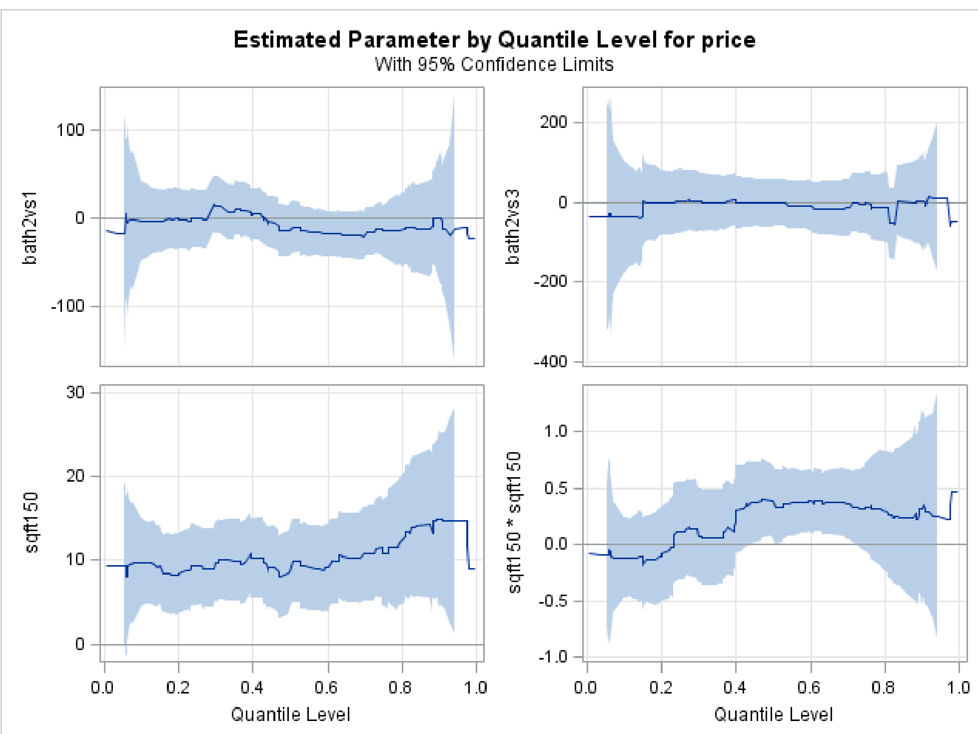


Top left: The intercept increases across percentiles (called "quantiles") as expected.

Top right: The slope for new construction stays just north of 0 until the 40th percentile or so.

Bottom left: The slope for 3 vs 2 bedrooms appears to not be different than 0 through most percentiles, although with an apparent increase in the upper quantiles (with lots of noise).

Bottom right: The slope for 3 vs 4 bedrooms appears to not be different than 0 through most of the percentiles, although with an apparent decrease in the upper percentiles (with lots of noise) until .80 or so, in which it suddenly jumps up to positive (with lots of noise)…?



Top left: The slope for bath 2 vs 1 is 0 with no trend across percentiles.

Top right: The slope for bath 2 vs 3 is 0 with no trend across percentiles.

Bottom left: The slope for the linear effect of square footage (which is the instantaneous slope at 1500 sq ft) is significantly positive across percentiles and looks to grow in strength after .60 or so.

Bottom right: The slope the quadratic effect of square footage is not different than 0 until about .50, at which point it is significantly positive (i.e., an accelerated effect of square footage). Although it stays positive, there is greater noise making it not different than 0 after .70 or so.

Sample Write-up using SAS output:

The present analysis sought to predict the final sale price of 100 homes from four characteristics: whether they were new construction (0=no, 1=yes), liner and quadratic effects of square footage in 100s (centered at 1500), number of bedrooms (2,3, or 4+), and number of bathrooms (1,2, or 3+). Because the observed distribution of home sale prices was positively skewed and contained seven potential outliers, the robustness of the model results to these characteristics was examined using several distinct approaches. All models included the same predictor effects and were estimated using maximum likelihood within SAS GLIMMIX unless otherwise noted. The extent of conditional distribution fit was examined using the Pearson χ^2/DF statistic (in which 1=good fit); all predictor fixed effects were tested univariately using z-distributions without denominator degrees of freedom unless otherwise noted. As expected given the positively skewed distribution of sale prices, a model specifying a normal conditional distribution have severe overdispersion (Pearson χ^2/DF = 2675.03).

We then examined two alternative models that were better suited for positively skewed residuals. First, we predicted home sale prices using a log-normal conditional distribution for the residuals, which appeared to have much better fit but also to result in underdispersion (Pearson χ^2/DF = 0.11). In the lognormal solution, after controlling for the number of bedrooms and bathrooms, new houses sold for significantly more money (0.24 log $1000 units; p < .041), and sale prices were also uniquely predicted by a quadratic function of square footage. More specifically, the sale price increased significantly by 0.08 log $1000 units per 100 additional square feet as evaluated at 1500 square feet (p < .001), but this positive slope of house size became significantly less positive by twice the quadratic coefficient of –0.001 per additional 100 square feet (i.e., the impact of being a bigger house was reduced in bigger houses; p < .023). The number of bedrooms or bathrooms did not have significant unique effects. Second, we fit the same predictive model using a log link function and a gamma conditional distribution, which exhibited a similar level of conditional distribution fit (Pearson χ^2/DF = 0.10). However, the effect of being new construction and the quadratic effect of house size were then nonsignificant (p's ≈ .07).

We then turned to a different modeling approach that would be more robust to outliers—quantile regression, in which one can predict any percentile of the distribution (labeled a "quantile") instead of the mean as in traditional regression. In our quantile regressions, the point estimates for the predictor slopes were found by minimizing a weighted function of the absolute value of the model residuals (in which the weights reflect the chosen percentile). Standard errors were found through 500 bootstrap replications (i.e., in which 500 samples with replacement were generated to capture the empirical sampling distribution of the slope estimates for more valid standard errors). SAS QUANTREG was used to conduct the analyses, and residual denominator degrees of freedom were used to evaluate the significance of the model predictors.

First, in predicting the 50th percentile (i.e., the median home price), no unique predictor effects were significant except square footage, for which significant positive linear and quadratic effects were found. More specifically, the sale price increased by 8.66 $1000 units per 100 additional square feet as evaluated at 1500 square feet (p < .001), and this positive slope of house size became significantly more positive by twice the quadratic coefficient of 0.38 per additional 100 square feet (i.e., the price bonus of being a bigger house was magnified in bigger houses; p < .0322). We repeated this analysis to predict the 25th and 75th percentiles to examine potential differences in prediction for relatively inexpensive or relatively expensive houses, respectively. At the 25th percentile, there was a marginally significant positive effect of new construction (Est = 45.67, p = .067), a significant linear effect of house size at 1500 square feet (Est = 9.40 per 100 square feet; p < .001), but no significant quadratic effect of house size (Est = 0.107, p = .612). At the 75th percentile, there was a nonsignificant effect of new construction (Est = 24.29, p = .437), a significant linear effect of house size at 1500 square feet (Est = 10.84 per 100 square feet; p < .002), but no significant quadratic effect of house size (Est = 0.33, p = .142). Finally, Figure 1 provides the results in examining prediction at 144 distinct values ranging from the 0.004th to 99.6th percentiles, in which the solid line in each image depicts the point estimate for the slope (y-axis) as a function of the percentile (x-axis), and the shading conveys the 95% confidence interval around the slope estimates. The unique effects of number of bedrooms and number of bathrooms did not appear to be significant at any percentile. The effect of new construction appeared marginally significantly positive from approximately the 20th to the 40th percentiles, and nonsignificantly positive otherwise. The linear effect of house size at 1500 square feet was significantly positive at nearly every percentile and appeared to grow in size as home prices increased. The quadratic effect of house size appeared to transition from nonsignificantly negative until the 20th percentile, to nonsignificantly positive until the 40th percentile, to significantly positive until the 70th percentile, after which it remained nonsignificantly positive. Thus, it appears that having a bigger house is even more helpful among midrange houses, but not for inexpensive or very expensive houses.