

## Example 4b: Generalized Linear Models and Quantile Regression for Positive Skewed Outcomes (complete syntax, data, and output available for SAS, STATA, and R 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 three types of generalized linear models (all with the same linear predictor) estimated using maximum likelihood: identity link with a normal distribution (typical regression), a log-transformed outcome in a typical regression (which is equivalent to an identity link with a lognormal distribution), and a log link with a gamma distribution. In addition, because this outcome also had several univariate outliers, we will use quantile regression to predict the median home price instead of the mean and to examine predictor slope differences across other percentiles.

For the generalized linear models: In SAS, I am still using GLIMMIX (even though these are not mixed-effects models). Because the corresponding STATA options (using GLM to get conditional distribution fit, also using LGAMMA) do not have denominator degrees of freedom, they were set to "none" in SAS GLIMMIX so that the SAS Wald test results (still labeled as  $t$  or  $F$ ) will match those of STATA (using  $z$  or  $\chi^2$ ). In R, I am using the base function GLM (also using  $z$  or  $\chi^2$ ). For quantile regression: In SAS, I am using QUANTREG. In STATA, I am using SQREG and IQREG, and in R I am using QUANTREG (although I have not yet figured out all the options for obtaining standard errors).

### SAS Data Import and Manipulation:

```
* Location for original files for these models - change this path;
* \\Client\ precedes path in Virtual Desktop outside H drive;
%LET filesave=C:\Dropbox\22_PSQF6270\PSQF6270_Example4b;

* Import Houses XLSX data file;
PROC IMPORT DATAFILE="&filesave.\Houses.xlsx" OUT=work.Example4b DBMS=XLSX REPLACE;
    SHEET="house data"; GETNAMES=YES;
RUN;

* Create predictor variables;
DATA work.Example4b; SET work.Example4b;
* Categories for number of bedrooms;
    bed3v2=.; bed3v4=.;
    IF beds=2 THEN DO; bed3v2=1; bed3v4=0; END;
    IF beds=3 THEN DO; bed3v2=0; bed3v4=0; END;
    IF beds IN(4,5) THEN DO; bed3v2=0; bed3v4=1; END;
* Categories for number of bathrooms;
    bath2v1=.; bath2v3=.;
    IF baths=1 THEN DO; bath2v1=1; bath2v3=0; END;
    IF baths=2 THEN DO; bath2v1=0; bath2v3=0; END;
    IF baths IN(3,4) THEN DO; bath2v1=0; bath2v3=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);
* Label existing and new variables;
    LABEL price= "price: Sale Price in 100,000 units"
           new= "new: House is new construction (0=no, 1=yes)"
           bed3v2= "bed3v2: 2 bedrooms instead of 3 (0=no, 1=yes)"
           bed3v4= "bed3v4: 4 bedrooms instead of 3 (0=no, 1=yes)"
           bath2v1= "bath2v1: 1 bathroom instead of 2 (0=no, 1=yes)"
           bath2v3= "bath2v3: 3 bathrooms instead of 2 (0=no, 1=yes)"
           sqft150= "sqft150: Square Footage per 100 feet (0=150)"
           logprice= "logprice: Natural log of sale price in 100,000 units";
RUN;
```

## STATA Data Import and Manipulation:

```
// Defining global variable for file location to be replaced in code below
// \\Client\ precedes path in Virtual Desktop outside H drive;
global filesave "C:\Dropbox\22_PSQF6270\PSQF6270_Example4b"

// Import Houses XLSX data
import excel "$filesave\Houses.xlsx", firstrow case(preserve) clear

// Categories for number of bedrooms
gen bed3v2=.
gen bed3v4=.
replace bed3v2=1 if beds==2
replace bed3v4=0 if beds==2
replace bed3v2=0 if beds==3
replace bed3v4=0 if beds==3
replace bed3v2=0 if beds==4
replace bed3v4=1 if beds==4
replace bed3v2=0 if beds==5
replace bed3v4=1 if beds==5
// Categories for number of bathrooms
gen bath2v1=.
gen bath2v3=.
replace bath2v1=1 if baths==1
replace bath2v3=0 if baths==1
replace bath2v1=0 if baths==2
replace bath2v3=0 if baths==2
replace bath2v1=0 if baths==3
replace bath2v3=1 if baths==3
replace bath2v1=0 if baths==4
replace bath2v3=1 if baths==4
// Center and rescale size into per 100 square feet (0=1500)
gen sqft150=(size-1500)/100
// Generate quadratic sqft150 for use in some routines
gen sqft150sq=sqft150*sqft150
// Log-transform price for lognormal model
gen logprice=log(price)
// Label existing and new variables
label variable price "price: Sale Price in 100,000 units"
label variable new "new: House is new construction (0=no, 1=yes)"
label variable bed3v2 "bed3v2: 2 bedrooms instead of 3 (0=no, 1=yes)"
label variable bed3v4 "bed3v4: 4 bedrooms instead of 3 (0=no, 1=yes)"
label variable bath2v1 "bath2v1: 1 bathroom instead of 2 (0=no, 1=yes)"
label variable bath2v3 "bath2v3: 3 bathrooms instead of 2 (0=no, 1=yes)"
label variable sqft150 "sqft150: Square Footage per 100 feet (0=150)"
label variable logprice "logprice: Natural log of sale price in 100,000 units"

// Install user-written packages for gamma
search lgamma // install from window
```

## R Data Import and Manipulation:

```
# Define variables for working directory and data name
filesave = "C:\\Dropbox\22_PSQF6270\PSQF6270_Example4b/"
filename = "Houses.xlsx"
setwd(dir=filesave)

# Import Houses XLSX data
Example4b = read_excel(paste0(filesave,filename))
# Convert to data frame without labels to use for analysis
Example4b = as.data.frame(Example4b)

# Categories for number of bedrooms
Example4b$bed3v2=NA; Example4b$bed3v4=NA
Example4b$bed3v2[which(Example4b$beds==2)]=1
Example4b$bed3v4[which(Example4b$beds==2)]=0
Example4b$bed3v2[which(Example4b$beds==3)]=0
```

```

Example4b$bed3v4[which(Example4b$beds==3)]=0
Example4b$bed3v2[which(Example4b$beds==4)]=0
Example4b$bed3v4[which(Example4b$beds==4)]=1
Example4b$bed3v2[which(Example4b$beds==5)]=0
Example4b$bed3v4[which(Example4b$beds==5)]=1
# Categories for number of bathrooms
Example4b$bath2v1=NA; Example4b$bath2v3=NA
Example4b$bath2v1[which(Example4b$baths==1)]=1
Example4b$bath2v3[which(Example4b$baths==1)]=0
Example4b$bath2v1[which(Example4b$baths==2)]=0
Example4b$bath2v3[which(Example4b$baths==2)]=0
Example4b$bath2v1[which(Example4b$baths==3)]=0
Example4b$bath2v3[which(Example4b$baths==3)]=1
Example4b$bath2v1[which(Example4b$baths==4)]=0
Example4b$bath2v3[which(Example4b$baths==4)]=1
# Center and rescale size into per 100 square feet (0=1500)
Example4b$sqft150=(Example4b$size-1500)/100
# Make squared version for use
Example4b$sqftsq=Example4b$sqft150^2
# Log-transform price for lognormal model
Example4b$logprice=log(Example4b$price)

```

## Syntax and SAS Output for Data Description:

```

TITLE "SAS Distribution of Sale Price Outcome";
PROC UNIVARIATE DATA=work.Example4b; VAR price size;
    HISTOGRAM price / MIDPOINTS= 0 TO 600 BY 20 NORMAL(MU=EST SIGMA=EST);
RUN; QUIT;
PROC SGPLOT DATA=work.Example4b;
    VBOX price / DATALABEL=price;
RUN; TITLE;

```

```

TITLE "SAS Descriptive Stats for Example Variables";
PROC MEANS NDEC=1 DATA=work.Example4b; VAR price size; RUN;
PROC FREQ DATA=work.Example4b; TABLE beds baths new; RUN; TITLE;

```

```

display "STATA Distribution of Sale Price Outcome"
summarize price
hist price, percent start(0) width(20)
graph export "$filesave\STATA Price Histogram.png", replace
graph box price
graph export "$filesave\STATA Price Box Plot.png", replace

```

```

display "STATA Descriptive Stats for Example Variables"
summarize price size
tabulate beds
tabulate baths
tabulate new

```

```

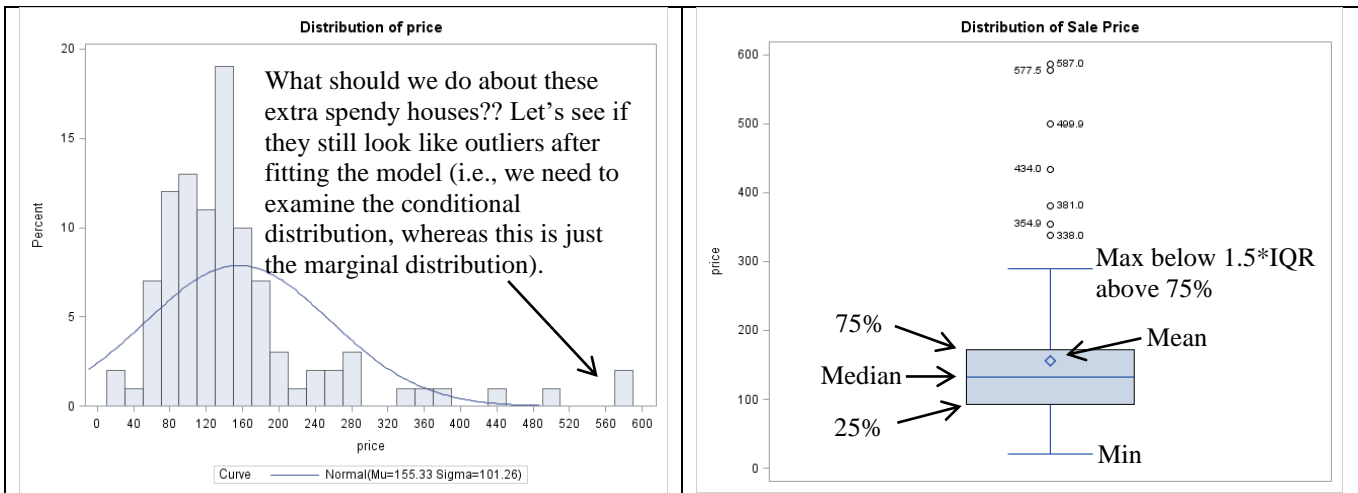
# to save a plot: open a file, create the plot, then close the file
png(file = "R Price Histogram.png") # open file
hist(x=Example4b$price, freq=FALSE,
     ylab="Density", xlab="Sale Price in 100,000 units") # axis labels
dev.off() # close file
png(file = "R Price Boxplot.png") # open file
boxplot(x=Example4b$price)
dev.off() # close file

```

```

print("R Descriptive Stats for Example Variables")
describe(x=Example4b$price); describe(x=Example4b$size)
table(x=Example4b$beds, useNA="ifany")
table(x=Example4b$baths, useNA="ifany")
table(x=Example4b$new, useNA="ifany")

```



Every model we fit in this example will have the same linear predictor so that the reference house is old (i.e., not new construction) and has 3 bedrooms, 2 bedrooms, and 1500 square feet:

$$\hat{y}_i = \beta_0 + \beta_1(New_i) + \beta_2(Bed3v2_i) + \beta_3(Bed3v4_i) + \beta_4(Bath2v1_i) + \beta_5(Bath2v3_i) + \beta_6(SqFt_i - 150) + \beta_7(SqFt_i - 150)^2$$

### 1) Predict Original Price with Identity Link and Normal Conditional Distribution:

$Price_i \sim Normal(\hat{y}_i, \sigma_e^2) \rightarrow$  Regular regression, but using ML estimation of residual variance for comparability

```
TITLE "SAS Predict Price using Identity Link, Normal Distribution";
PROC GLIMMIX DATA=work.Example4b NAMELEN=100 GRADIENT METHOD=MSPL PLOTS=ALL;
  MODEL price = new bed3v2 bed3v4 bath2v1 bath2v3 sqft150 sqft150*sqft150
    / SOLUTION DDFM=NONE LINK=IDENTITY DIST=NORMAL;
  CONTRAST "Multiv Wald test of Model" new 1, bed3v2 1, bed3v4 1,
    bath2v1 1, bath2v3 1, sqft150 1, sqft150*sqft150 1 / CHISQ;
RUN; TITLE;
```

```
display "STATA Predict Price using Identity Link, Normal Distribution"
display "Add vfactor=(N-#fixed effects)/N to match SAS results for SEs"
glm price c.new c.bed3v2 c.bed3v4 c.bath2v1 c.bath2v3 c.sqft150 ///
  c.sqft150#c.sqft150, ml link(identity) family(gaussian) vfactor(0.92)
display "-2LL=" e(11)*-2 // Print -2LL for model
estat ic, n(100) // AIC and BIC do not match SAS and I do not know why
test (c.new=0) (c.bed3v2=0) (c.bed3v4=0) (c.bath2v1=0) (c.bath2v3=0) ///
  (c.sqft150=0) (c.sqft150#c.sqft150=0) // Multiv Wald test of model
```

```
print("R Predict Price using Identity Link, Normal Distribution")
ModelNorm = glm(data=Example4b, family=gaussian(link="identity"), # I(x^2) squares predictor
  formula=price~1+new+bed3v2+bed3v4+bath2v1+bath2v3+sqft150+sqftsq)
# Compute ML residual variance to use as "dispersion"
NewDispersion = sum(ModelNorm$residuals^2) / nrow(Example4b)
summary(ModelNorm, dispersion=NewDispersion)
print("Print ML -2LL, AIC, and BIC that match SAS")
-2*logLik(ModelNorm); AIC(ModelNorm); BIC(ModelNorm)

print("Multiv Wald Test of Model -- matches SAS using REML")
NormR2 = glht(model=ModelNorm, linfct=c("new=0", "bed3v2=0", "bed3v4=0", "bath2v1=0",
  "bath2v3=0", "sqft150=0", "sqftsq=0")) # Couldn't square predictor here
summary(NormR2, test=Chisqtest()) # Joint chi-square test
```

### Partial SAS Output:

```

Fit Statistics
-2 Log Likelihood      1072.96
AIC (smaller is better) 1090.96
BIC (smaller is better) 1114.41
Pearson Chi-Square     267503.1
Pearson Chi-Square / DF 2675.03 → ML residual variance (NOT misfit)
    
```

Effect	Parameter Estimates					
	Estimate	Standard Error	DF	t Value	Pr >  t	Gradient
Intercept	128.14	7.2363	Infty	17.71	<.0001	-263E-17
<b>new</b>	59.5217	18.3575	Infty	3.24	<b>0.0012</b>	-113E-18
bed3v2	14.2148	15.7512	Infty	0.90	0.3668	-17E-17
bed3v4	5.8132	15.7592	Infty	0.37	0.7122	-125E-18
bath2v1	-6.3723	16.2369	Infty	-0.39	0.6947	-184E-18
bath2v3	-14.4904	20.6592	Infty	-0.70	0.4831	1.58E-16
<b>sqft150</b>	10.0297	1.7914	Infty	5.60	<b>&lt;.0001</b>	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 → ML residual variance

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

### Partial STATA Output:

```

Generalized linear models
Optimization      : ML
No. of obs       = 100
Residual df      = 92
Scale parameter  = 2907.643
(1/df) Deviance  = 2907.643
Pearson          = 267503.1219
(1/df) Pearson   = 2907.643 → REML residual variance
Variance function: V(u) = 1
Link function     : g(u) = u
[Gaussian]
[Identity]
AIC              = 10.88959
BIC              = 267079.4
Log likelihood    = -536.4796698
    
```

	price	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
	new	59.52165	18.35751	3.24	0.001	23.54159	95.50172
	bed3v2	14.21484	15.75124	0.90	0.367	-16.65702	45.0867
	bed3v4	5.813162	15.7592	0.37	0.712	-25.07431	36.70063
	bath2v1	-6.372286	16.23691	-0.39	0.695	-38.19605	25.45147
	bath2v3	-14.49037	20.65924	-0.70	0.483	-54.98174	26.00101
	sqft150	10.02966	1.791421	5.60	0.000	6.51854	13.54078
	c.sqft150#c.sqft150	.149102	.0869353	1.72	0.086	-.0212881	.319492
	_cons	128.1352	7.236345	17.71	0.000	113.9523	142.3182

```

. test (c.new=0) (c.bed3v2=0) (c.bed3v4=0) (c.bath2v1=0) (c.bath2v3=0) ///
> (c.sqft150=0) (c.sqft150#c.sqft150=0) // Multiv Wald test of model
chi2( 7) = 279.49
Prob > chi2 = 0.0000
    
```

### Partial R Output:

```

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) 128.135249 7.236345 17.7072 < 0.000000000000000022
new 59.521653 18.357514 3.2424 0.001185
bed3v2 14.214838 15.751239 0.9025 0.366813
bed3v4 5.813161 15.759201 0.3689 0.712222
bath2v1 -6.372286 16.236911 -0.3925 0.694721
bath2v3 -14.490364 20.659244 -0.7014 0.483054
sqft150 10.029661 1.791421 5.5987 0.00000002159
sqftsq 0.149102 0.086935 1.7151 0.086328
    
```

(Dispersion parameter for gaussian family taken to be 2675.0312) → **ML residual variance due to new dispersion**

Null deviance: 1015150 on 99 degrees of freedom  
 Residual deviance: 267503 on 92 degrees of freedom  
 AIC: 1090.96

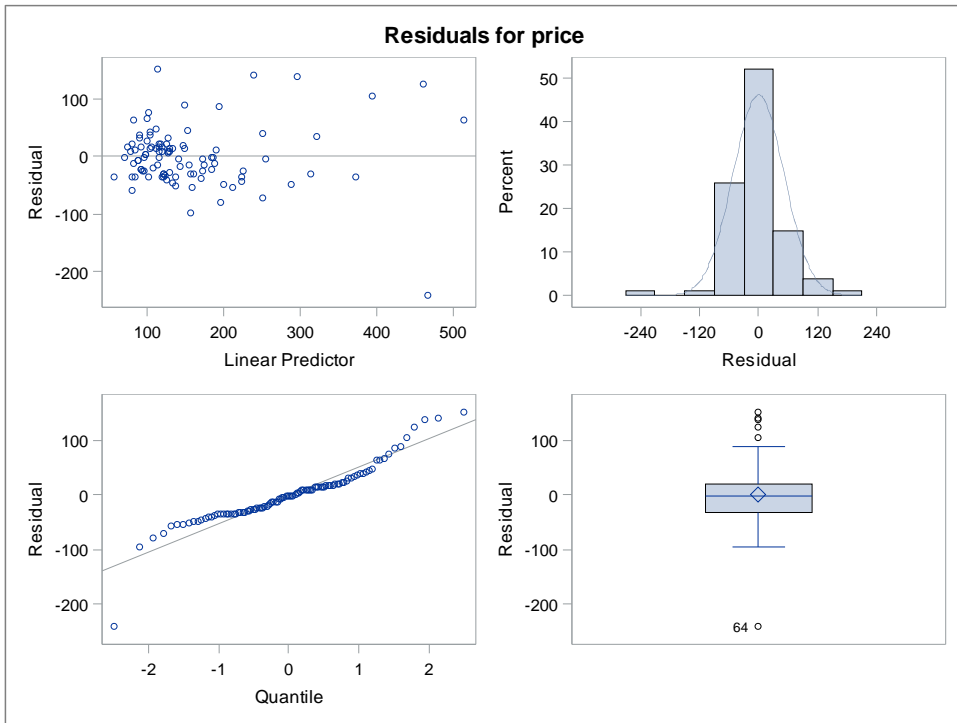
```
[1] "Print ML -2LL, AIC, and BIC that match SAS"
> -2 * logLik(ModelNorm)
'log Lik.' 1072.9593 (df=9)
> AIC(ModelNorm)
[1] 1090.9593
> BIC(ModelNorm)
[1] 1114.4059

[1] "Multiv Wald Test of Model -- matches SAS using REML"
```

General Linear Hypotheses  
 Linear Hypotheses:

	Estimate
new == 0	59.5217
bed3v2 == 0	14.2148
bed3v4 == 0	5.8132
bath2v1 == 0	-6.3723
bath2v3 == 0	-14.4904
sqft150 == 0	10.0297
sqftsq == 0	0.1491

Global Test:  
 Chisq DF Pr(>Chisq)  
 1 257.13 7 8.4006e-52



The conditional distribution still has some outliers and deviates from normal to some extent.  
  
 Let's see if we can do better...

**2a) Predict Log-Transformed Price with Identity Link and Normal Conditional Distribution:**

$LogPrice_i \sim Normal(\hat{y}_i, \sigma_e^2)$  → Regular regression on log-transformed outcome (using ML estimation)

```
TITLE "SAS Predict Log-Transformed Price using Identity Link, Normal Distribution";
PROC GLIMMIX DATA=work.Example4b NAMELEN=100 GRADIENT METHOD=MSPL;
  MODEL logprice = new bed3v2 bed3v4 bath2v1 bath2v3 sqft150 sqft150*sqft150
    / SOLUTION DDFM=NONE LINK=IDENTITY DIST=NORMAL;
  CONTRAST "Multiv Wald test of Model" new 1, bed3v2 1, bed3v4 1,
    bath2v1 1, bath2v3 1, sqft150 1, sqft150*sqft150 1 / CHISQ;
RUN; TITLE;
```

```

display "STATA Predict Log-Transformed Price using Identity Link, Normal Distribution"
glm logprice c.new c.bed3v2 c.bed3v4 c.bath2v1 c.bath2v3 c.sqft150 ///
    c.sqft150#c.sqft150, ml link(identity) family(gaussian) vfactor(0.92)
display "-2LL=" e(11)*-2 // Print -2LL for model
estat ic, n(100) // AIC and BIC do not match SAS and I do not know why
test (c.new=0) (c.bed3v2=0) (c.bed3v4=0) (c.bath2v1=0) (c.bath2v3=0) ///
    (c.sqft150=0) (c.sqft150#c.sqft150=0) // Multiv Wald test of model

print("R Predict Log-Transformed Price using Identity Link, Normal Distribution")
ModelLogNorm = glm(data=Example4b, family=gaussian(link="identity"),
    formula=logprice~1+new+bed3v2+bed3v4+bath2v1+bath2v3+sqft150+sqftsq)
# Compute ML residual variance to use as "dispersion"
NewDispersion = sum(ModelLogNorm$residuals^2) / nrow(Example4b)
summary(ModelLogNorm, dispersion=NewDispersion)
print("Print ML -2LL, AIC, and BIC that match SAS")
-2*logLik(ModelLogNorm); AIC(ModelLogNorm); BIC(ModelLogNorm)

print("Multiv Wald Test of Model -- matches SAS using REML")
LogTNormR2 = glht(model=ModelLogNorm, linfct=c("new=0", "bed3v2=0", "bed3v4=0",
    "bath2v1=0", "bath2v3=0", "sqft150=0", "sqftsq=0"))
summary(LogTNormR2, test=Chisqtest()) # Joint chi-square test

```

## 2b) Predict Price with Identity Link and Lognormal Conditional Distribution:

$Price_i \sim \text{Lognormal}(\hat{y}_i, \sigma_e^2) \rightarrow$  Residuals are expected to follow a lognormal distribution

```

TITLE "SAS Predict Price using Identity Link, Log-Normal Distribution";
PROC GLIMMIX DATA=work.Example4b NAMELEN=100 GRADIENT METHOD=MSPL;
    MODEL price = new bed3v2 bed3v4 bath2v1 bath2v3 sqft150 sqft150*sqft150
        / SOLUTION DDFM=NONE LINK=IDENTITY DIST=LOGNORMAL;
    CONTRAST "Multiv Wald test of Model" new 1, bed3v2 1, bed3v4 1,
        bath2v1 1, bath2v3 1, sqft150 1, sqft150*sqft150 1 / CHISQ;
RUN; TITLE;

// No Stata regression with a lognormal distribution that I could find
# Could not find lognormal conditional distribution in R that was likelihood-equivalent

```

**Partial SAS Output—It is exactly the same either way (so this is one of very few situations in which transforming the outcome is equivalent to fitting a generalized linear model):**

Fit Statistics	
-2 Log Likelihood	61.83
AIC (smaller is better)	79.83
BIC (smaller is better)	103.27
Pearson Chi-Square	10.87
Pearson Chi-Square / DF	0.11 $\rightarrow$ ML residual variance of log(price)

Parameter Estimates						
Effect	Estimate	Standard Error	DF	t Value	Pr >  t	Gradient
Intercept	4.8144	0.04612	Infy	104.39	<.0001	-16E-13
<b>new</b>	0.2392	0.1170	Infy	2.04	<b>0.0409</b>	4.79E-14
bed3v2	0.1540	0.1004	Infy	1.53	0.1251	5.51E-14
bed3v4	0.01298	0.1004	Infy	0.13	0.8972	6.11E-15
bath2v1	-0.1455	0.1035	Infy	-1.41	0.1597	-103E-15
bath2v3	-0.05614	0.1317	Infy	-0.43	0.6698	-256E-16
<b>sqft150</b>	0.07952	0.01142	Infy	6.97	<b>&lt;.0001</b>	2.79E-12
<b>sqft150*sqft150</b>	-0.00126	0.000554	Infy	-2.28	<b>0.0228</b>	-257E-13
Scale	0.1087	0.01537	.	.	.	7.56E-11 $\rightarrow$ ML residual variance

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

**Partial STATA Output:**

```

Generalized linear models          No. of obs    =          100
Optimization      : ML            Residual df   =           92
                                   Scale parameter =   .1180992
                                   (1/df) Deviance =   .1180992
Deviance          = 10.86512647    (1/df) Deviance =   .1180992 → REML residual variance
Pearson          = 10.86512647    (1/df) Pearson =   .1180992
Variance function: V(u) = 1      [Gaussian]
Link function    : g(u) = u      [Identity]
                                   AIC              =   .7782651
                                   BIC              =  -412.8105
Log likelihood   = -30.91325673

```

```

-----
              |           OIM
              |           Coef.  Std. Err.   z    P>|z|    [95% Conf. Interval]
-----+-----
              |
      new     |   .2391817   .1169949   2.04  0.041   .0098759   .4684874
      bed3v2  |   .1539676   .1003847   1.53  0.125  -.0427829   .350718
      bed3v4  |   .0129777   .1004355   0.13  0.897  -.1838722   .2098276
      bath2v1 |  -.1455129   .10348    -1.41  0.160  -.34833    .0573041
      bath2v3 |  -.0561446   .1316641  -0.43  0.670  -.3142016   .2019123
      sqft150 |   .0795194   .011417    6.97  0.000   .0571425   .1018962
c.sqft150#c.sqft150 | -.0012611   .0005541  -2.28  0.023  -.002347  -.0001752
      _cons   |   4.814402   .0461182  104.39 0.000   4.724012   4.904792
-----

```

```

. test (c.new=0) (c.bed3v2=0) (c.bed3v4=0) (c.bath2v1=0) (c.bath2v3=0) ///
>      (c.sqft150=0) (c.sqft150#c.sqft150=0) // Multiv Wald test of model
      chi2( 7) =   187.71
      Prob > chi2 =   0.0000

```

**Partial R Output:**

Coefficients:

```

              Estimate  Std. Error  z value      Pr(>|z|)
(Intercept)  4.81440211  0.04611819  104.3927 < 0.000000000000000022
new          0.23918164  0.11699488   2.0444      0.04092
bed3v2       0.15396753  0.10038474   1.5338      0.12509
bed3v4       0.01297764  0.10043549   0.1292      0.89719
bath2v1     -0.14551293  0.10347999  -1.4062      0.15967
bath2v3     -0.05614470  0.13166411  -0.4264      0.66980
sqft150      0.07951937  0.01141696   6.9650      0.0000000000003284
sqftsq      -0.00126111  0.00055405  -2.2762      0.02284

```

(Dispersion parameter for gaussian family taken to be **0.10865127**) → **ML residual variance**

```

Null deviance: 31.2597 on 99 degrees of freedom
Residual deviance: 10.8651 on 92 degrees of freedom
AIC: 79.8265

```

```

[1] "Print ML -2LL, AIC, and BIC that match SAS"
> -2 * logLik(ModelLogNorm)
'log Lik.' 61.826517 (df=9)
> AIC(ModelLogNorm)
[1] 79.826517
> BIC(ModelLogNorm)
[1] 103.27305

```

```

[1] "Multiv Wald Test of Model -- matches SAS using REML"

```

```

General Linear Hypotheses
Linear Hypotheses:

```

```

              Estimate
new == 0      0.2391816
bed3v2 == 0   0.1539675
bed3v4 == 0   0.0129776
bath2v1 == 0 -0.1455129
bath2v3 == 0 -0.0561447
sqft150 == 0  0.0795194
sqftsq == 0  -0.0012611

```

Global Test:

```

Chisq DF Pr(>Chisq)
1 172.69 7 6.7988e-34

```



3) Predict Price with Log Link and Gamma Conditional Distribution:  $Price_i \sim \text{Gamma}(\mu, \phi)$ , where  $\hat{y}_i = \text{Log}(\mu)$  and  $\phi$  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:

```
TITLE "SAS Predict Price using Log Link, Gamma Distribution";
PROC GLIMMIX DATA=work.Example4b NAMELEN=100 GRADIENT METHOD=MSPL;
  MODEL price = new bed3v2 bed3v4 bath2v1 bath2v3 sqft150 sqft150*sqft150
    / SOLUTION DDFM=NONE LINK=LOG DIST=GAMMA;
  CONTRAST "Multiv Wald test of Model" new 1, bed3v2 1, bed3v4 1,
    bath2v1 1, bath2v3 1, sqft150 1, sqft150*sqft150 1 / CHISQ;
RUN; TITLE;

display "STATA: Price using Log Link, Gamma Distribution"
display "Using LGAMMA that does not allow factor variables or interactions"
display "GLM gives different LL and solution for gamma distribution"
lgamma price new bed3v2 bed3v4 bath2v1 bath2v3 sqft150 sqft150sq
display "-2LL=" e(l1)*-2 // Print -2LL for model
estat ic, n(100) // AIC and BIC match SAS
test (c.new=0) (c.bed3v2=0) (c.bed3v4=0) (c.bath2v1=0) (c.bath2v3=0) ///
  (c.sqft150=0) (c.sqft150#c.sqft150=0) // Multiv Wald test of model

display "STATA LGAMMA: Price using Log Link, Gamma Distribution"
display "Get Incident-Rate Ratios as exp(slope)"
lgamma price new bed3v2 bed3v4 bath2v1 bath2v3 sqft150 sqft150sq, eform

print("R Predict Price using Log Link, Gamma Distribution")
print("SEs and scale parameter are very close to SAS and STATA")
ModelGamma = glm(data=Example4b, family=Gamma(link="log"), # I(x^2) squares predictor
  formula=price~1+new+bed3v2+bed3v4+bath2v1+bath2v3+sqft150+sqftsq)
summary(ModelGamma)
print("Print ML -2LL, AIC, and BIC are very close to those given by SAS")
-2*logLik(ModelGamma); AIC(ModelGamma); BIC(ModelGamma)
print("Pearson Chi-Square / DF Index of Fit matching SAS")
sum(residuals(ModelGamma, type="pearson")^2)/(100) # SAS N

print("Multiv Wald Test of Model -- does not match SAS and STATA")
GammaR2 = glht(model=ModelGamma, linfct=c("new=0","bed3v2=0","bed3v4=0",
  "bath2v1=0","bath2v3=0","sqft150=0","sqftsq=0"))
summary(GammaR2, test=Chisqtest()) # Joint chi-square test
```

## Partial SAS Output:

```
Fit Statistics
-2 Log Likelihood          1034.44
AIC (smaller is better)   1052.44
BIC (smaller is better)   1075.88
Pearson Chi-Square         9.77
Pearson Chi-Square / DF    0.10 → less variance than model expects
```

```
Parameter Estimates
Standard
Effect      Estimate      Error      DF      t Value      Pr > |t|      Gradient
Intercept   4.8550      0.04415   Infy     109.97      <.0001      -2.67E-7
new         0.2047      0.1136    Infy      1.80      0.0715      -0.00001
bed3v2      0.1729      0.1002    Infy      1.72      0.0846      0.000029
bed3v4      0.02188     0.09529   Infy      0.23      0.8184      -9.69E-6
bath2v1     -0.1323     0.09993   Infy     -1.32      0.1855      0.000017
bath2v3     -0.05267    0.1244    Infy     -0.42      0.6720      -4.99E-6
sqft150     0.07520     0.01114   Infy      6.75      <.0001      0.001965
sqft150*sqft150 -0.00100    0.000549   Infy     -1.82      0.0693      -0.02582
Scale       0.1004      0.01397    .          .          .          -2.65E-6 → phi variance multiplier
```

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

**Partial STATA Output:**

```

Log-gamma model
Log likelihood = -517.21898
Number of obs = 100
LR chi2(7) = 117.57
Prob > chi2 = 0.0000
-----+-----
      price |      Coef.   Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+-----
      new |      .204721   .1136043     1.80   0.072   -0.0179394   .4273814
    bed3v2 |     .1728484   .1002319     1.72   0.085   -0.0236026   .3692993
    bed3v4 |     .0218806   .0952913     0.23   0.818   -1.1648869   .2086482
    bath2v1 |    -0.1323233   .0999321    -1.32   0.185   -0.3281866   .06354
    bath2v3 |    -0.0526695   .1244118    -0.42   0.672   -0.2965123   .1911732
    sqft150 |     .0752007   .0111396     6.75   0.000   .0533675   .0970339
    sqft150sq |   -0.0009965   .0005487    -1.82   0.069   -0.0020719   .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  -> log(phi)
-----+-----
      phi |     .1003938   .0139665                .0764346   .1318632  -> phi variance multiplier
-----+-----
. test (new=0) (bed3v2=0) (bed3v4=0) (bath2v1=0) (bath2v3=0) ///
>      (sqft150=0) (sqft150sq) // Multiv Wald test of model
      chi2( 7) = 187.18
      Prob > chi2 = 0.0000

```

**Partial R Output:**

```

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.85495821  0.04559534 106.4793 < 0.000000000000000022
new          0.20472068  0.11566850   1.7699   0.08006
bed3v2       0.17285544  0.09924667   1.7417   0.08491
bed3v4       0.02188128  0.09929685   0.2204   0.82608
bath2v1     -0.13232450  0.10230684  -1.2934   0.19911
bath2v3     -0.05266582  0.13017143  -0.4046   0.68672
sqft150      0.07520161  0.01128753   6.6624   0.000000001942
sqftsq      -0.00099659  0.00054777  -1.8194   0.07211

(Dispersion parameter for Gamma family taken to be 0.10620167) -> slightly different phi variance multiplier

Null deviance: 31.9401 on 99 degrees of freedom
Residual deviance: 10.2072 on 92 degrees of freedom
AIC: 1052.45

```

```

[1] "Print ML -2LL, AIC, and BIC are very close to those given by SAS"
> -2 * logLik(ModelGamma)
'log Lik.' 1034.4521 (df=9)
> AIC(ModelGamma)
[1] 1052.4521
> BIC(ModelGamma)
[1] 1075.8986

[1] "Pearson Chi-Square / DF Index of Fit matching SAS"
> sum(residuals(ModelGamma, type = "pearson")^2)/(100)
[1] 0.09770537

[1] "Multiv Wald Test of Model -- does not match SAS and STATA"

```

```

Global Test:
      Chisq DF Pr(>Chisq)
1 178.37 7 4.2939e-35

```

#### 4) Predict Price Median (50<sup>th</sup> 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 50<sup>th</sup> percentile and the distance from the 25<sup>th</sup> to 75<sup>th</sup> 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.

```
TITLE "SAS Predict Price 50th Percentile (Median) using Quantile Regression";
PROC QUANTREG DATA=work.Example4b NAMELEN=100 CI=RESAMPLING(NREP=500);
  MODEL price = new bed3v2 bed3v4 bath2v1 bath2v3 sqft150 sqft150*sqft150
    / QUANTILE=.50 SEED=8675309; * Jenny is my random seed;
  Model: TEST new bed3v2 bed3v4 bath2v1 bath2v3 sqft150 sqft150*sqft150 / WALD;
RUN; TITLE;

display "STATA Predict Price 50th Percentile (Median) using Quantile Regression"
set seed 8675309 // Set Jenny as seed to get same results each time
sqreg price c.new c.bed3v2 c.bed3v4 c.bath2v1 c.bath2v3 c.sqft150 ///
  c.sqft150#c.sqft150, quantile(.50) reps(500)
test (c.new=0) (c.bed3v2=0) (c.bed3v4=0) (c.bath2v1=0) (c.bath2v3=0) ///
  (c.sqft150=0) (c.sqft150#c.sqft150=0) // Multiv Wald test of model does not match SAS

print("R Predict Price 50th Percentile [Median] using Quantile Regression")
print("Did not figure out how to get same SEs and test statistics as SAS and STATA")
set.seed(8675309) # Jenny is my random seed
ModelQ50 = rq(data=Example4b, tau=.5,
  formula=price~1+new+bed3v2+bed3v4+bath2v1+bath2v3+sqft150+sqftsq)
summary(ModelQ50)
```

#### Partial SAS Output:

Parameter	DF	Parameter Estimates					
		Estimate	Standard Error	95% Confidence Limits		t Value	Pr >  t
Intercept	1	133.0000	7.2909	118.5197	147.4803	18.24	<.0001
new	1	32.1650	24.6156	-16.7236	81.0536	1.31	0.1946
bed3v2	1	1.0778	18.4457	-35.5569	37.7125	0.06	0.9535
bed3v4	1	-28.1157	17.6509	-63.1719	6.9404	-1.59	0.1146
bath2v1	1	-13.7301	15.3765	-44.2691	16.8088	-0.89	0.3742
bath2v3	1	-1.2992	29.5853	-60.0581	57.4596	-0.04	0.9651
sqft150	1	8.6648	2.4979	3.7038	13.6258	3.47	0.0008
sqft150*sqft150	1	0.3827	0.1653	0.0545	0.7110	2.32	0.0228

#### Test Model Results

Test	Statistic	DF	Chi-Square	Pr > ChiSq
Wald	109.8928	7	109.89	<.000

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

<.000 → Translates to F = 109.89/7 = 15.70

#### Partial STATA Output:

Simultaneous quantile regression		Number of obs =		100	
bootstrap(500) SEs		.50 Pseudo R2 =		0.4523	
price	Coef.	Bootstrap Std. Err.	t	P> t	[95% Conf. Interval]
q50					
new	32.16499	29.56973	1.09	0.280	-26.56305 90.89303
bed3v2	1.07779	19.89831	0.05	0.957	-38.44197 40.59755
bed3v4	-28.11573	21.78021	-1.29	0.200	-71.37311 15.14165
bath2v1	-13.73013	14.5145	-0.95	0.347	-42.55717 15.09691
bath2v3	-1.299235	32.61557	-0.04	0.968	-66.07658 63.47811
sqft150	8.664786	2.330797	3.72	0.000	4.035623 13.29395

```

c.sqft150#c.sqft150 | .3827353 .2509158 1.53 0.131 -.1156051 .8810758
      _cons | 133 7.28882 18.25 0.000 118.5238 147.4762
-----
. test (c.new=0) (c.bed3v2=0) (c.bed3v4=0) (c.bath2v1=0) (c.bath2v3=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

```

## Partial R Output:

```

Coefficients:
              coefficients lower bd upper bd
(Intercept) 133.000000    119.479154 139.878004
new          32.164989     3.529067  82.654677
bed3v2       1.077787    -14.270654 32.900320
bed3v4      -28.115733   -44.735514 -2.981709
bath2v1     -13.730133   -35.257264  7.080776
bath2v3     -1.299234   -43.256743 27.989451
sqft150      8.664785     6.543296 13.021328
sqftsq       0.382735    -0.149437  0.491025

```

## 4) Predict Price 25<sup>th</sup> and 75<sup>th</sup> 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 25<sup>th</sup> percentile to ask, “among (relatively) cheap houses, what predicts sale price?” Likewise, we predict the 75<sup>th</sup> 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.

```

TITLE "SAS Predict Price 25th and 75th Percentile using Quantile Regression";
PROC QUANTREG DATA=work.Example4b NAMELEN=100 CI=RESAMPLING(NREP=500);
MODEL price = new bed3v2 bed3v4 bath2v1 bath2v3 sqft150 sqft150*sqft150
          / QUANTILE=.25 .75 SEED=8675309; * Jenny is my random seed;
* Multiv wald test of Model (provided for each quantile);
EachModel: TEST new bed3v2 bed3v4 bath2v1 bath2v3 sqft150 sqft150*sqft150 / WALD;
* Multiv wald test of slope differences between quantiles;
ModelDiff: TEST new bed3v2 bed3v4 bath2v1 bath2v3 sqft150 sqft150*sqft150 / QINTERACT;
newDiff: TEST new / QINTERACT; * How to test single slope diff across quantiles;
RUN; TITLE;

```

```

display "STATA Predict 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.bed3v2 c.bed3v4 c.bath2v1 c.bath2v3 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.bed3v2=0) ([q25]c.bed3v4=0) ([q25]c.bath2v1=0) ///
      ([q25]c.bath2v3=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.bed3v2=0) ([q75]c.bed3v4=0) ([q75]c.bath2v1=0) ///
      ([q75]c.bath2v3=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.bed3v2=[q75]c.bed3v2) ///
      ([q25]c.bed3v4=[q75]c.bed3v4) ([q25]c.bath2v1=[q75]c.bath2v1) ///
      ([q25]c.bath2v3=[q75]c.bath2v3) ([q25]c.sqft150=[q75]c.sqft150) ///
      ([q25]c.sqft150#c.sqft150=[q75]c.sqft150#c.sqft150)
// How to test single slope diff across quantiles
test ([q25]c.new=[q75]c.new)

```

```

display "STATA Predict Price 25-75 Inter-Quantile Regression"
display "Output now directly provides predictor slope differences"
set seed 8675309 // Set Jenny as seed to get same results each time
iqreg price c.new c.bed3v2 c.bed3v4 c.bath2v1 c.bath2v3 c.sqft150 ///
          c.sqft150#c.sqft150, quantile(.25 .75) reps(500)
test (c.new=0) (c.bed3v2=0) (c.bed3v4=0) (c.bath2v1=0) (c.bath2v3=0) ///
      (c.sqft150=0) (c.sqft150#c.sqft150=0) // Multiv Wald test of differences

```

```
print("R Predict Price 25th and 75th Percentile using Quantile Regression")
print("Did not figure out how to get same SEs and test statistics as SAS and STATA")
set.seed(8675309) # Jenny is my random seed
ModelQ2575 = rq(data=Example4b, tau=c(.25,.75),
                formula=price~1+new+bed3v2+bed3v4+bath2v1+bath2v3+sqft150+sqftsq)
summary(ModelQ2575)
```

**Partial SAS Output:**

Parameter Estimates Predicting 25 <sup>th</sup> percentile								
Parameter	DF	Estimate	Standard		95% Confidence		t Value	Pr >  t
			Error	Limits	Limits			
Intercept	1	101.1147	7.2839	86.6482	115.5813	13.88	<.0001	
new	1	45.6732	26.3641	-6.6881	98.0345	1.73	0.0866	
bed3v2	1	4.7000	16.2591	-27.5920	36.9920	0.29	0.7732	
bed3v4	1	-0.2206	18.0406	-36.0508	35.6095	-0.01	0.9903	
bath2v1	1	-0.7478	16.5383	-33.5943	32.0988	-0.05	0.9640	
bath2v3	1	2.3978	39.9465	-76.9394	81.7351	0.06	0.9523	
sqft150	1	9.4049	2.4080	4.6225	14.1874	3.91	0.0002	
sqft150*sqft150	1	0.1069	0.2230	-0.3360	0.5498	0.48	0.6329	

Parameter Estimates Predicting 75 <sup>th</sup> percentile								
Parameter	DF	Estimate	Standard		95% Confidence		t Value	Pr >  t
			Error	Limits	Limits			
Intercept	1	145.7357	7.5581	130.7246	160.7467	19.28	<.0001	
new	1	24.3886	35.5563	-46.2292	95.0065	0.69	0.4945	
bed3v2	1	31.5946	19.8498	-7.8288	71.0179	1.59	0.1149	
bed3v4	1	-31.6868	38.1827	-107.5210	44.1474	-0.83	0.4088	
bath2v1	1	-15.0642	15.3389	-45.5285	15.4001	-0.98	0.3286	
bath2v3	1	-1.2579	38.0627	-76.8537	74.3379	-0.03	0.9737	
sqft150	1	10.8404	3.2413	4.4028	17.2779	3.34	0.0012	
sqft150*sqft150	1	0.3295	0.2020	-0.0718	0.7307	1.63	0.1063	

Test EachModel Results					
Quantile	Level Test	Test Statistic	DF	Chi-Square	Pr > ChiSq
0.25	Wald	65.3371	7	65.34	<.0001 → F= 65.34/7 = 9.33
0.75	Wald	91.5617	7	91.56	<.0001 → F= 91.56/7 = 13.08

Test ModelDiff Results			Test newDiff Results		
Equal Coefficients Across Quantiles			Equal Coefficients Across Quantiles		
Chi-Square	DF	Pr > ChiSq	Chi-Square	DF	Pr > ChiSq
4.4799	7	0.7231	0.3636	1	0.5465

**Partial STATA Output from SQREG—these are the predictor slopes per quantile:**

Simultaneous quantile regression		Number of obs = 100	
bootstrap(500) SEs		.25 Pseudo R2 = 0.3747	.75 Pseudo R2 = 0.5713
price	Coef.	Bootstrap Std. Err.	t P> t  [95% Conf. Interval]
q25			
new	45.67319	23.28024	1.96 0.053 [-.5633818 91.90976]
bed3v2	4.7	16.55032	0.28 0.777 [-28.17036 37.57036]
bed3v4	-.2206333	22.16177	-0.01 0.992 [-44.23583 43.79456]
bath2v1	-.7477557	15.38074	-0.05 0.961 [-31.29524 29.79972]
bath2v3	2.397835	33.72783	0.07 0.943 [-64.58855 69.38422]
sqft150	9.404941	1.757855	5.35 0.000 [5.91369 12.89619]
c.sqft150#c.sqft150	.1068575	.2572658	0.42 0.679 [-.4040946 .6178097]
_cons	101.1147	7.681166	13.16 0.000 [85.85928 116.3702]

```
q75
      new |      24.38865      37.27569      0.65      0.515     -49.64408      98.42139
    bed3v2 |      31.59456      18.9706      1.67      0.099     -6.082685      69.2718
    bed3v4 |     -31.68683      45.05709     -0.70      0.484     -121.1741      57.80045
   bath2v1 |     -15.06422      13.76459     -1.09      0.277     -42.40189      12.27344
   bath2v3 |     -1.257883      43.82958     -0.03      0.977     -88.30722      85.79145
    sqft150 |      10.84037      3.055926      3.55      0.001      4.771039      16.90971
c.sqft150#c.sqft150 |      .3294847      .201842      1.63      0.106     -0.0713909      .7303603
      _cons |      145.7357      5.482533      26.58      0.000      134.8469      156.6244
```

```
. // Multiv Wald test of model at 25th percentile
. test ([q25]c.new=0) ([q25]c.bed3v2=0) ([q25]c.bed3v4=0) ([q25]c.bath2v1=0) ///
> ([q25]c.bath2v3=0) ([q25]c.sqft150=0) ([q25]c.sqft150#c.sqft150=0)
      F( 7, 92) = 12.10
      Prob > F = 0.0000

. // Multiv Wald test of model at 75th percentile
. test ([q75]c.new=0) ([q75]c.bed3v2=0) ([q75]c.bed3v4=0) ([q75]c.bath2v1=0) ///
> ([q75]c.bath2v3=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.bed3v2=[q75]c.bed3v2) ///
> ([q25]c.bed3v4=[q75]c.bed3v4) ([q25]c.bath2v1=[q75]c.bath2v1) ///
> ([q25]c.bath2v3=[q75]c.bath2v3) ([q25]c.sqft150=[q75]c.sqft150) ///
> ([q25]c.sqft150#c.sqft150=[q75]c.sqft150#c.sqft150)
      F( 7, 92) = 0.55
      Prob > F = 0.7918

. // Single-predictor difference across quantiles
. test ([q25]c.new=[q75]c.new)
      F( 1, 92) = 0.37
      Prob > F = 0.5460
```

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

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

```
.75-.25 Interquantile regression
      bootstrap(500) SEs
      Number of obs = 100
      .75 Pseudo R2 = 0.5713
      .25 Pseudo R2 = 0.3747
```

price	Coef.	Bootstrap Std. Err.	t	P> t	[95% Conf. Interval]
new	-21.28454	35.11913	-0.61	0.546	-91.03417 48.46509
bed3v2	26.89456	21.05773	1.28	0.205	-14.92791 68.71703
bed3v4	-31.46619	43.83957	-0.72	0.475	-118.5354 55.60297
bath2v1	-14.31647	16.55987	-0.86	0.390	-47.2058 18.57287
bath2v3	-3.655718	42.55953	-0.09	0.932	-88.18263 80.87119
sqft150	1.435431	2.880917	0.50	0.619	-4.286319 7.157181
c.sqft150#c.sqft150	.2226272	.2837418	0.78	0.435	-.3409085 .7861628
_cons	44.62092	8.548936	5.22	0.000	27.64199 61.59984

```
. test (c.new=0) (c.bed3v2=0) (c.bed3v4=0) (c.bath2v1=0) (c.bath2v3=0) ///
> (c.sqft150=0) (c.sqft150#c.sqft150=0) // Multiv Wald test of differences
      F( 7, 92) = 0.55
      Prob > F = 0.7918
```

**Partial R Output:**

```
tau: [1] 0.25

Coefficients:
      coefficients lower bd upper bd
(Intercept) 101.114737 93.093346 113.687477
new          45.673190 31.445800 92.285814
bed3v2       4.700000 -14.872686 33.256801
bed3v4      -0.220641 -27.352594 19.000892
bath2v1     -0.747755 -18.718106 20.884363
bath2v3      2.397843 -59.449552 37.667577
sqft150      9.404941 6.816233 10.564952
sqftsq       0.106858 -0.258119 0.405855
```

tau: [1] 0.75

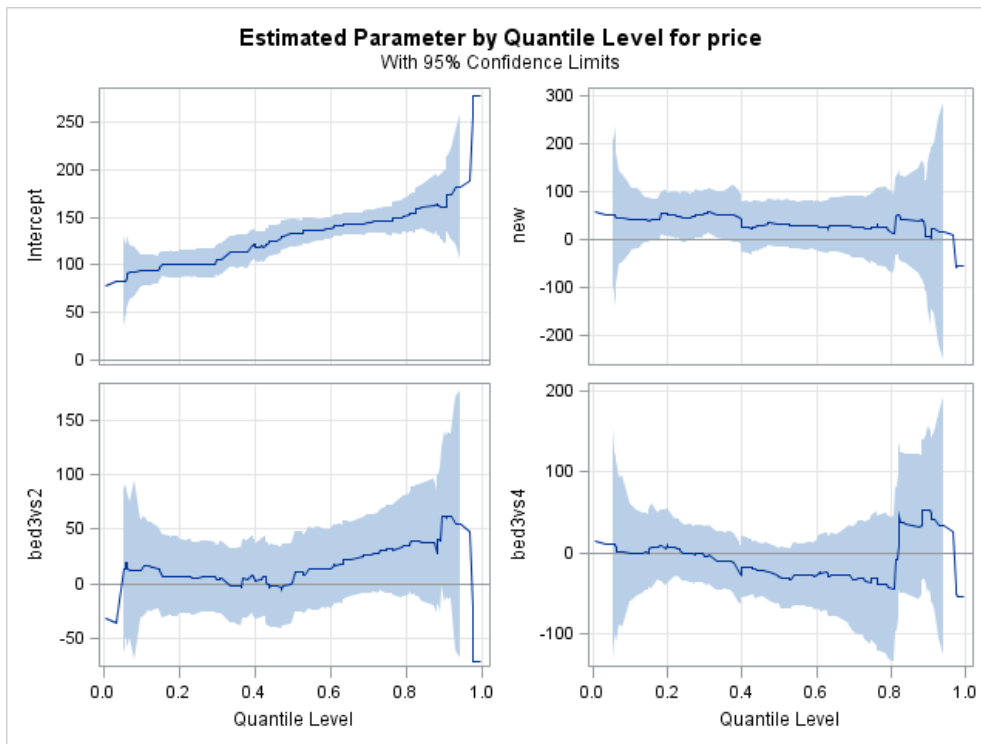
Coefficients:

	coefficients	lower bd	upper bd
(Intercept)	145.735654	141.481847	157.961905
new	24.388649	-0.554536	92.452481
bed3v2	31.594557	4.661877	49.800555
bed3v4	-31.686826	-55.983707	78.477871
bath2v1	-15.064223	-28.281428	3.033738
bath2v3	-1.257882	-47.710414	107.001254
sqft150	10.840372	7.669831	16.773869
sqftsq	0.329485	0.124996	0.816528

**5) Predict Price All Percentiles using Quantile Regression (couldn't find this in STATA or R):**

```
TITLE "SAS Predict Price at All Percentiles using Quantile Regression";
PROC QUANTREG DATA=work.Example4b NAMELEN=100 CI=RESAMPLING(NREP=500);
  MODEL price = new bed3v2 bed3v4 bath2v1 bath2v3 sqft150 sqft150*sqft150
    / QUANTILE=PROCESS PLOT=QUANTPLOT SEED=8675309;
RUN; TITLE;
```

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

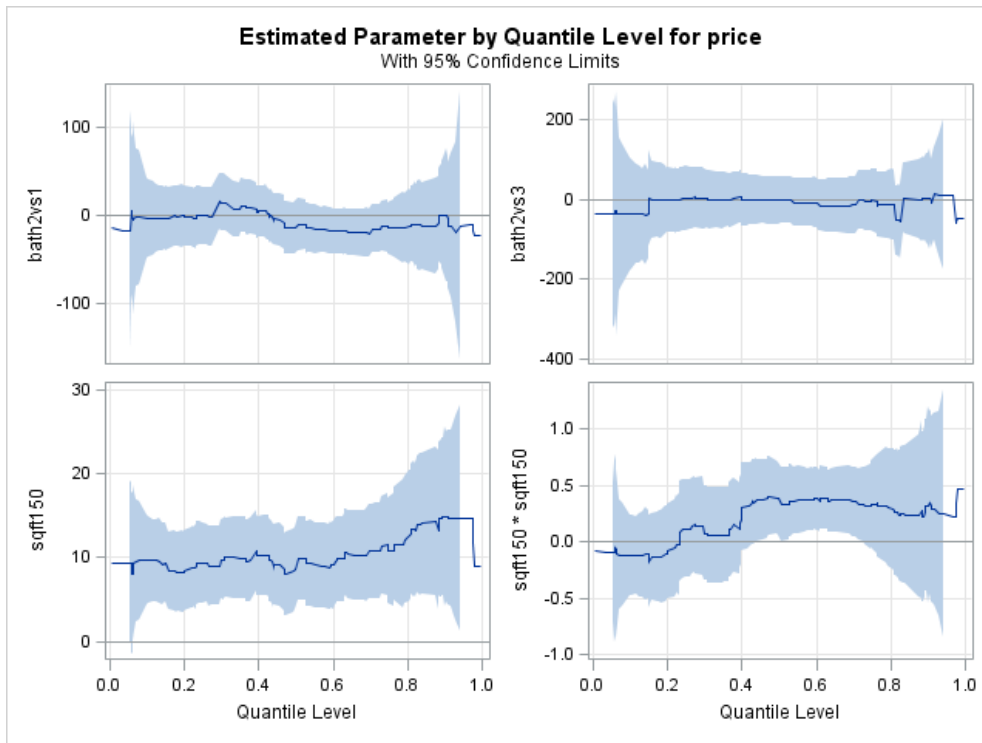


Top left: The intercept increases across percentiles (called “quantiles”) as expected.

Top right: The slope for new construction stays just north of 0 until the 40<sup>th</sup> 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 results 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), linear 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  $\chi^2/DF$  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, the residuals of a model specifying a normal conditional distribution indicated a lack of fit and several outliers.

We then examined two alternative models that were better suited for positively skewed residuals. First, we predicted home sale prices using a lognormal conditional distribution for the residuals, which appeared to result in underdispersion (Pearson  $\chi^2/DF = 0.11$ ). 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  $\chi^2/DF = 0.10$ ). However, the effect of being new construction and the quadratic effect of house size were then nonsignificant ( $p$ 's  $\approx .07$ ).

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 50<sup>th</sup> 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;  $p = .023$ ). We repeated this analysis to predict the 25<sup>th</sup> and 75<sup>th</sup> percentiles to examine potential differences in prediction for relatively inexpensive or relatively expensive houses, respectively. At the 25<sup>th</sup> percentile, there was a marginally significant positive effect of new construction (Est = 45.67,  $p = .087$ ), 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,  $p = .633$ ). At the 75<sup>th</sup> percentile, there was a nonsignificant effect of new construction (Est = 24.29,  $p = .495$ ), a significant linear effect of house size at 1500 square feet (Est = 10.84 per 100 square feet;  $p = .00$ ), but no significant quadratic effect of house size (Est = 0.33,  $p = .106$ ). Finally, Figure 1 provides the results in examining prediction at 144 distinct values ranging from the 0.004<sup>th</sup> to 99.6<sup>th</sup> 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 20<sup>th</sup> to the 40<sup>th</sup> 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 20<sup>th</sup> percentile, to nonsignificantly positive until the 40<sup>th</sup> percentile, to significantly positive until the 70<sup>th</sup> 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.