H. Guy Williams Case 2: **Housing** April 25, 2006
OPIM303/Hartford

1. Create the descriptive statistics for the data set and provide some analysis of the numbers. Which of the statistics are important in helping to set an asking price for a house in Eastville?

Regression Statistics	
Multiple R	0.908604066
R Square	0.825561349
Adjusted R Square	0.807577983
Standard Error	11.59355988
Observations	108

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	10	61703.81054	6170.381054	45.90694223	2.20682E-32
Residual	97	13037.83117	134.4106306		
Total	107	74741.64171			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-15.2123948	9.817909216	-1.549453602	0.124529355	-34.69822637	4.273436772
sqft	0.037595714	0.003627195	10.36495549	2.18798E-17	0.030396736	0.044794692
beds	4.923746235	1.964688168	2.506120979	0.01387045	1.024384172	8.823108297
baths	-2.911505713	3.023955912	-0.962813546	0.338035206	-8.913220974	3.090209548
heat	-12.90973007	6.100946887	-2.116020728	0.036903588	-25.01842051	-0.801039631
style	2.287745136	1.643716379	1.391812581	0.167162438	-0.974576792	5.550067063
garage	15.75932859	3.824582965	4.120535162	7.96368E-05	8.168590291	23.35006689
basement	9.077212298	3.445422063	2.634571942	0.009807055	2.239003414	15.91542118
age	-1.034169395	0.281336915	-3.675910759	0.000388642	-1.59254528	-0.475793509
fireplace	5.305402605	3.979449564	1.333200112	0.185588818	-2.592703016	13.20350822
school	4.621679542	2.534147074	1.823761371	0.07126744	-0.407900787	9.651259871

Multiple R is the correlation coefficient, the closer it is to one the stronger the relationship between the data. R Squared is more important than R and tells us the percentage of variation in the dependent variable which is explained by the independent variables, the closer to 1 the more the variation in Y is explained by the x variables. R Squared will always increase when a new independent variable model is added to the regression model. Still more important is the Adjusted R Square value. It tells us the percentage of variation in the dependent variable explained by the independent variables *adjusted* for the number of independent variables. Adjusted R Squared does not always increase with the addition of an independent variable but when it does increase it indicates an improvement in the model. Adjusted R Squared can be used to compare two or more multiple regression models including those with different numbers of independent variables. The model with the highest Adjusted R Squared value will give the better solution for the dependent variable and will usually, baring other criteria, be the best model. The model is generally considered good if the values of R Squared and Adjusted R Squared are greater than or equal to 70%.

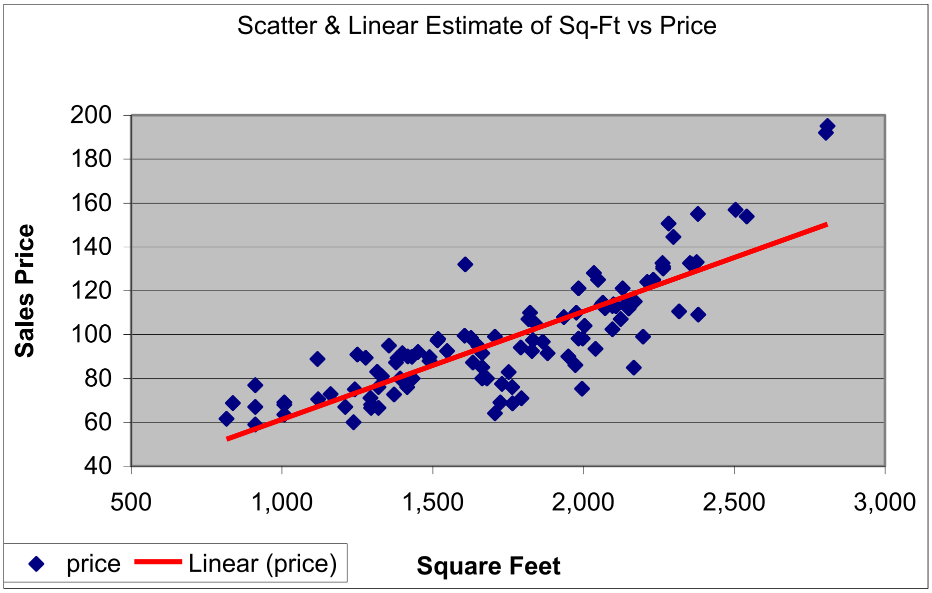
Significance F is a somewhat useful statistic of the model telling us the probability that all model coefficients are equal to zero (Significance F > α) of that some of the coefficients (at least one) are not equal to zero (Significance F < α).

The P-value is a very important statistic of the regression model and tells us the probability that a particular coefficient is equal to zero (P-value > α) or not equal to zero (P-value < α). Specifically, the P-value is the probability of observing the set of data under analysis when the null hypothesis is true.

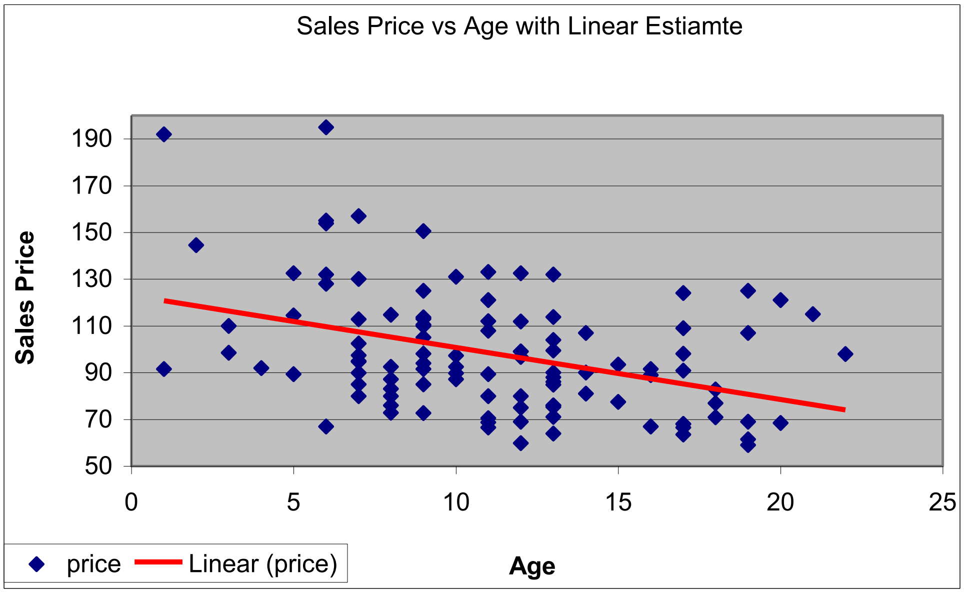
Finally, the confidence interval is a useful statistic telling us the range of possible values, to the specified confidence level, a particular model coefficient may take. A confidence interval which crosses zero indicates there is a possibility the coefficient may actually have the value zero. The corresponding coefficients are usually not used and typically this condition is accompanied by a P-value greater than alpha.

The intercept coefficient is an exception to the rules applied to the P-values and confidence intervals. The intercept coefficient is typically used regardless of it's P-value or confidence interval. The reason is that the logic implied by an intercept of zero typically makes no sense. For instance, a zero square foot house selling for non-zero dollars.

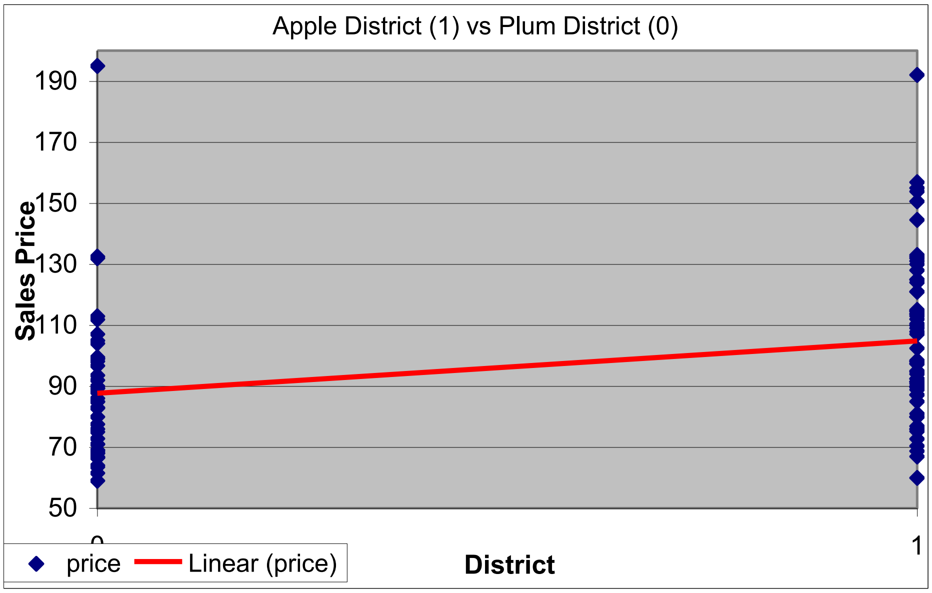
2. Are there simple graphs of the data that would help you determine an asking price for your house? Create a few and describe how you would use them.



It is known that the square foot size of a house is a strong indicator of it's selling price when compared to the selling price of other houses of comparable size. The intercept and slope of the linear estimate of the scatter diagram can be used to solve for a *minimum* selling price of the house under question. It is important to note that adding other indicators to the model will improve the estimate of the true market price.



The scatter and linear estimate of age verses sales price can be used to determine if there is a premium, and approximately how much that premium is, for newer houses verses older houses in the area.



Prospective home buyers may use a linear estimate of selling price verses school district to identify another source of premium.

3. Which model (i.e., which variables are included) predicts the sales price best? Which variables are important, which ones unimportant?

SUMMARY OUTPUT

BEST MODEL

Regression Statistics	
Multiple R	0.9137
R Square	0.8349
Adjusted R Square	0.8233
Standard Error	11.1089
Observations	108

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	62400.9711	8914.4244	72.2361	2.89E-36
Residual	100	12340.6706	123.4067		
Total	107	74741.6417			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	45.4914	17.6331	2.5799	0.0113	10.5079	80.4749
sqft	0.0448	0.0033	13.4721	3.31E-24	0.0382	0.0514
basement	7.8375	3.3027	2.3731	0.0196	1.2850	14.3900
age	-0.9650	0.2549	-3.7865	0.0003	-1.4707	-0.4594
garage	-39.1034	16.3665	-2.3892	0.0188	-71.5740	-6.6328
garage^2	12.2416	3.6920	3.3157	0.0013	4.9168	19.5663
Style1	14.6940	3.3819	4.3448	3.35E-05	7.9843	21.4037
Style2	6.6334	2.7661	2.3981	0.0183	1.1454	12.1213

The best model is given by the equation:

Sale Price =45.4914 + 0.0448*SQFT + 7.8375*Basement - 0.965*AGE - 39.1034*GARAGE + 12.2416*GARAGE^2 + 14.694*STYLE1 + 6.6334*STYLE2

The important variables include square footage, basement, age, garage, style1 and style2. The strongest predictors, based on P-value, are square footage and style1.

Unimportant variables, those removed from the model, include bath, fireplace, and school.

4. Is there a difference in the price in terms of different architectural styles?

Yes. Cape represents the base price with an offset of 45.5, ranch has a premium of 45.5+6.63=52.13, and the two-story has the highest premium at 45.5+14.69=60.19.

5. How does the existence of a basement contribute to selling price?

The coefficient for basement is 7.8375. With all other variables held steady the presents of a basement will result in an increase in sales price of $7, 837.5 (values are in thousands).

6. How does the existence of a fireplace add to the selling price?

Fireplace is a variable which was removed from the model. As such, it has no effect on the sales price value predicted by the model.