Statistics 5350/7110
Forecasting

Lecture 6
Regression Modeling and Diagnostics

Professor Stine

Admin Issues

- Questions?
  - Office hours, TA
- Assignments
  - Discussion of solutions
- Learning R
  - *Introduction to Statistical Learning* (with applications in R)
  - On-line at www.statlearning.com
  - Intro to basic R in §2.3
  - Use of R in regression in §3.1-3.4, with examples in §3.6
- Review
  - Multiple regression model
  - Inference, collinearity, prediction
  - Model selection
  - Added comments to Lecture_5.Rmd

Model Selection

- Objective: Select a model that predicts new data as well as possible
  - Model = Set of predictors for chosen response
  - Heuristic: Occam's razor, a preference for simpler models
- Obvious choices
  - R^2: Prefers the model with every predictor that you have available
  - s_w^2: More selective than R^2, but s_w^2 needs an adjustment for predicting new data
- AIC: penalized approach
  - Select the model that has the smallest expected squared prediction error

$$\sum E(\text{prediction error})^2 = (n+k)\,\sigma_w^2$$

k = number estimated parameters in the model

  - Unbiased estimate of this quantity is $(n+k)\,s_w^2 = n(n+k)/(n-k)\,\hat{\sigma}_w^2 \approx n(1+2k/n)\,\hat{\sigma}_w^2$ $\quad \hat{\sigma}_w^2 = SSE/n$
- Measures of complexity produced by adding predictors
  - Unbiased estimator — AIC: Number of predictors x 2
  - Bayesian prior — BIC: Number of predictors x log of sample size
  - Risk inflation — RIC: Number of predictors x log of searched features

Today's Topics

- Regression model
  - Review with eye toward problems and assumptions
  - Emphasis on residual diagnostics produced by R
  - Example Illustrates other types of predictors: categorical variables, interactions

- Key regression diagnostic plots
  - Scatterplot of residuals vs fitted values
  - Normal quantile plot of residuals
  - Abs value of standardized residuals vs fitted values
  - Standardized residuals vs leverage
  - Partial regression plots

- Leverage and influence
  - Some points are more important than others

Plotting a linear model in R generates the first 4 of these plots for a least squares regression.

Example of Leverage

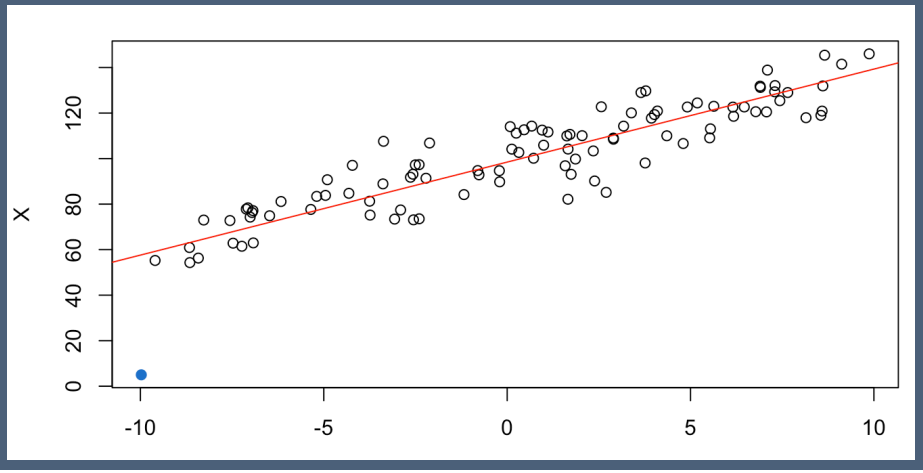
Remember:
Textbook notation uses
X for the response

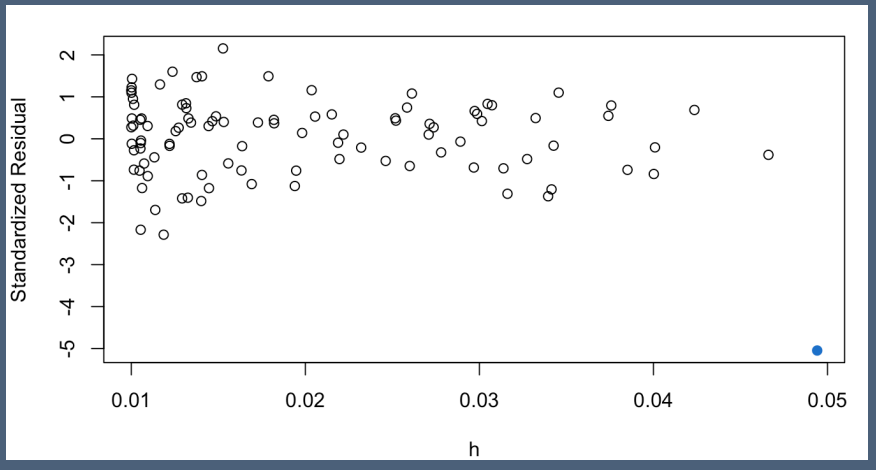
- Leverage in regression
  - Influence = measures how much regression coefficients change if observation is deleted
  - Leverage = h_i is sensitivity of regression to an observation derived from its predictor values $0 < h_i \leq 1$
  - Leverage controls variance at each residual: bigger leverage, smaller variance
  - Standardized residual ideally ≈ N(0,1)

$$\frac{e_i}{\sqrt{1-h_i}\, s_w}$$

where e_i = residual, h_i = leverage and s_w = residual SD

- Example
  - Leveraged outlier on left side at smallest value of predictor.





Leverage and Influence

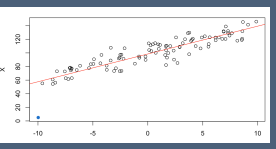
- Leverage
  - Controls how values at each observation influence overall fitted regression
  - In simple regression, leverage measures distance from average of predictor $h_i = \frac{1}{n} + \frac{(Z_i - \bar{Z})^2}{SS_Z}$
  - Variance of a residual depends on its leverage: $\text{Var}(\hat{w}_i) = (1 - h_i)\,\sigma_w^2$
  - Outlier: Observations with atypical values of explanatory variables have high leverage
- Helpful plot
  - Graph of the leverages on the x-axis with standardized residuals

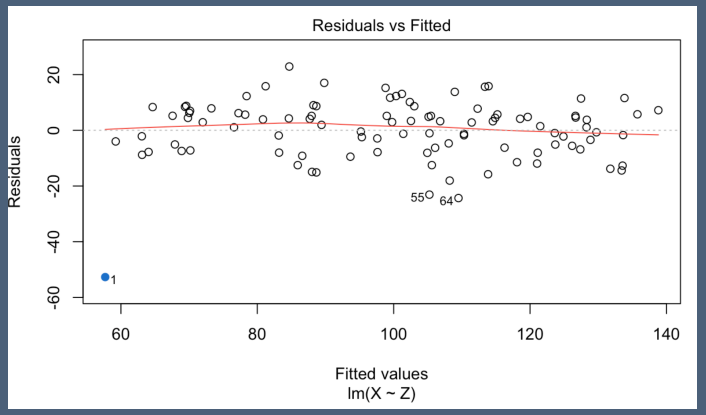
$$\frac{\hat{w}_i}{\sqrt{(1 - h_i)\,s_w^2}}$$

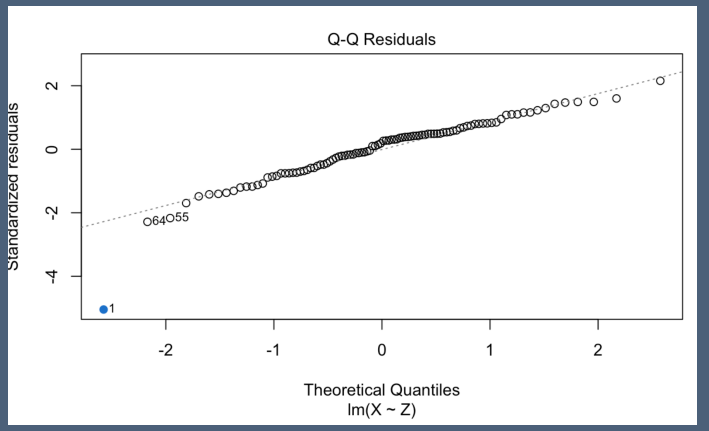
    on y-axis.
  - High leverage (unusual explanatory values) + large residual implies big influence on the fitted model.

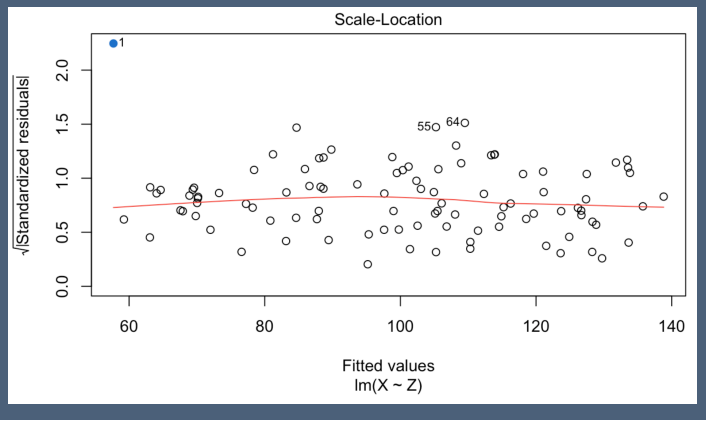
Regression Diagnostic Plots

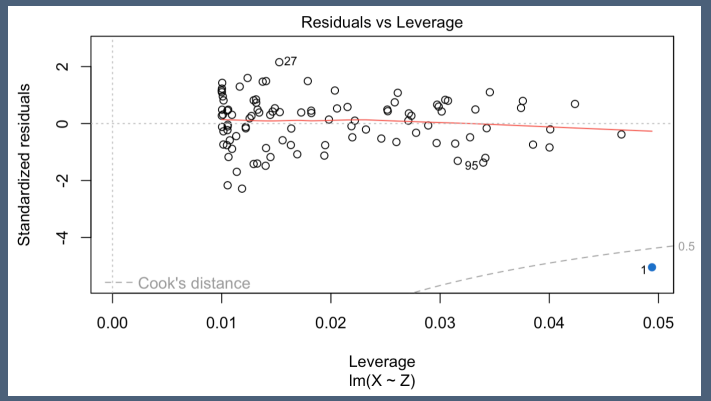
- Standard residual plots for a regression in R
  - Visible leveraged outlier







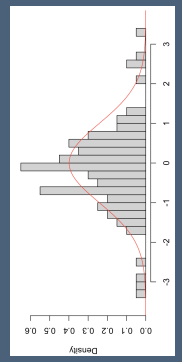


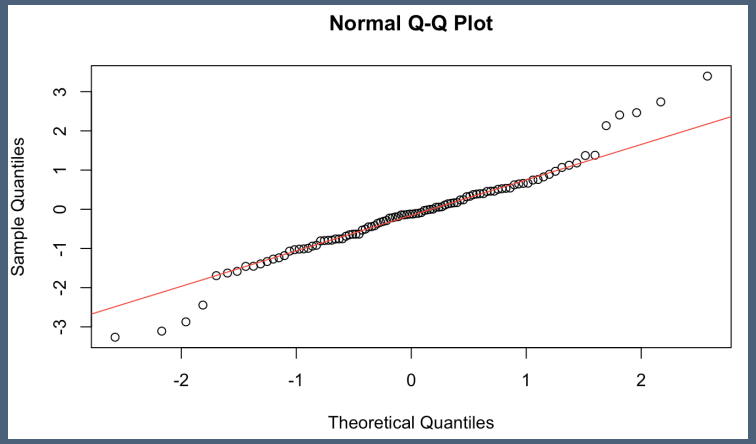


influential outlier because leveraged with large residual

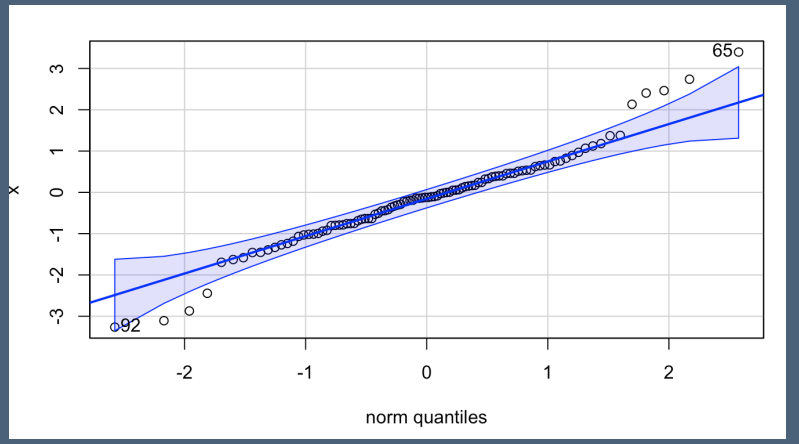
Intervals for the Normal Quantile Plot

- If residuals appear "simple", combine them in histogram
  - Simple: as if a sample from a common distribution (equal mean, equal variance, independent — iid)
- Normal quantile plot
  - Histogram isn't directly helpful since most interesting deviations from normality are in tails
- Example
  - Data are a sample from a t-distribution with 3 d.f.
  - QQ plot is volatile in tails, so intervals help judge significance





Version with confidence bounds (from car package)



Example: Multiple Regression

- Review of model
  - Model for a time series X_t with q predictors $Z_1, Z_2, \ldots, Z_q$
  - Equation

$$X_t = \beta_0 + \beta_1 Z_{t,1} + \cdots + \beta_q Z_{t,q} + w_t$$

  - White noise errors: mean zero, common variance σ_w^2
- Is the model well-specified?
  - Does the model make substantive sense?
  - Does the model have the relevant explanatory variables?
- Diagnostic issues
  - Do the model errors appear to be independent? (at least uncorrelated!)
  - If require prediction intervals, are the model errors normally distributed with common variance?
  - Are any observations unusually leveraged/influential?
  - Do the model effects appear linear?

Choice of symbols matches those in the textbook

Overview of Example Data

- Home prices
  - Prices of 958 detached homes sold in US Midwest
- Explanatory features
  - Have 28 possible features from which to construct possible explanatory variables
  - Mostly self-explanatory from name

    AG = above ground, SF = square feet, Unf = unfinished
  - Several are categorical (e.g. Noise, Lot.Config)

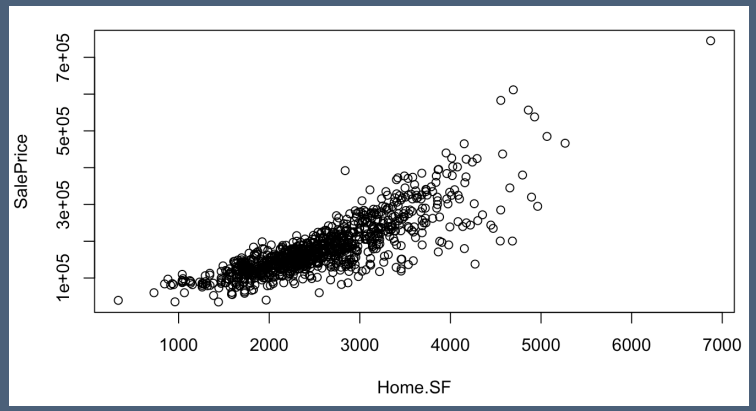
"SalePrice"	"Forclosure.Sale"	"Overall.Condition"	"Age"
"LotArea"	"Lot.Shape"	"Lot.Config"	"Noise"
"Single.Family"	"Remodeled"	"Bsmt.Fin.SF"	"BsmtUnfSF"
"Total.Bsmt.SF"	"CentralAir"	"Elec.Breaker"	"Floors"
"Home.SF"	"Total.SF..AG."	"X1stFlrSF"	"X2ndFlrSF"
"Bathrooms"	"Bsmt.Bathrm"	"Bedrooms..AG."	"Total.Rooms..AG."
"Outdoor.SF"	"WoodDeckSF"	"GarageCars"	"GarageFinish"
"Fireplaces"			

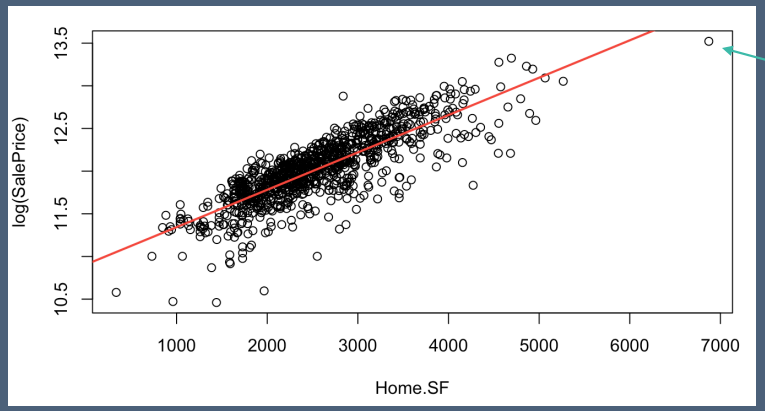
- Process
  - No need to dwell on residuals problems until have a “reasonably good” regression model
  - Start with simple model
    One predictor, price as a function of square footage (slope is nicely interpretable)
  - Add other predictors, Illustrating model-selection criteria
- Residual diagnostics

Initial Model

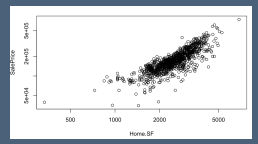
- Transform the response
  - Response will be price on a log scale
- Interpretation
  - Variation on log scale implies percentage change
  - 0.04% increase in price per added square foot, or 4% per added 100 s.f.

```
Coefficients:
(Intercept)      Home.SF
  1.091e+01    4.379e-04
```



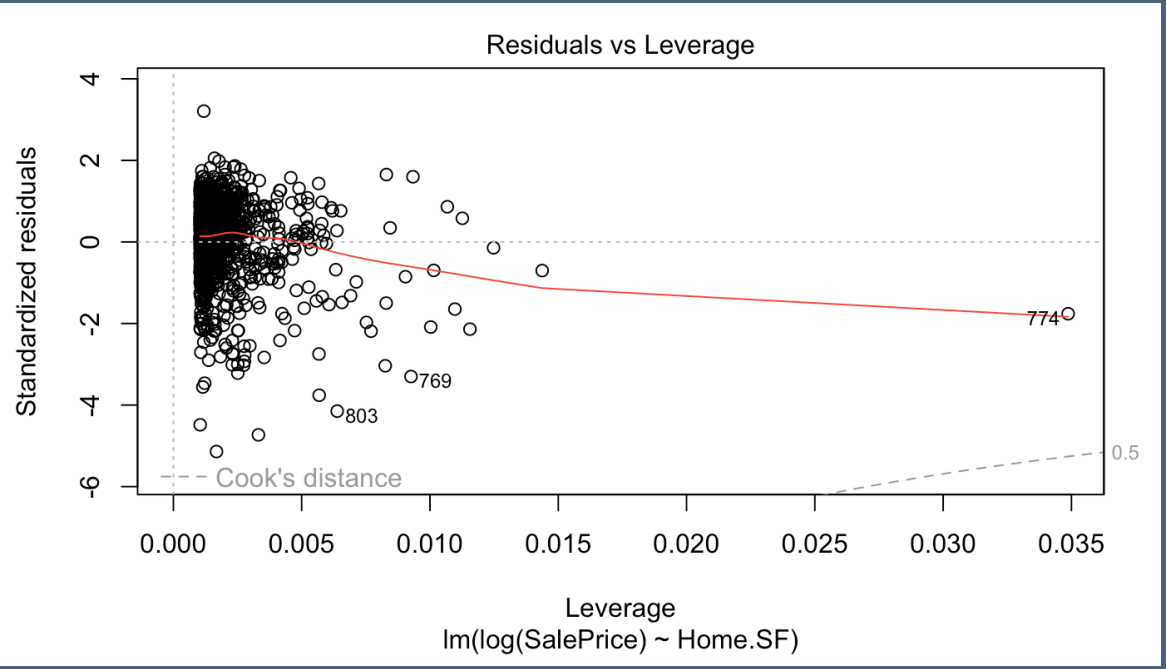


log-log is too much



Leverage Analysis

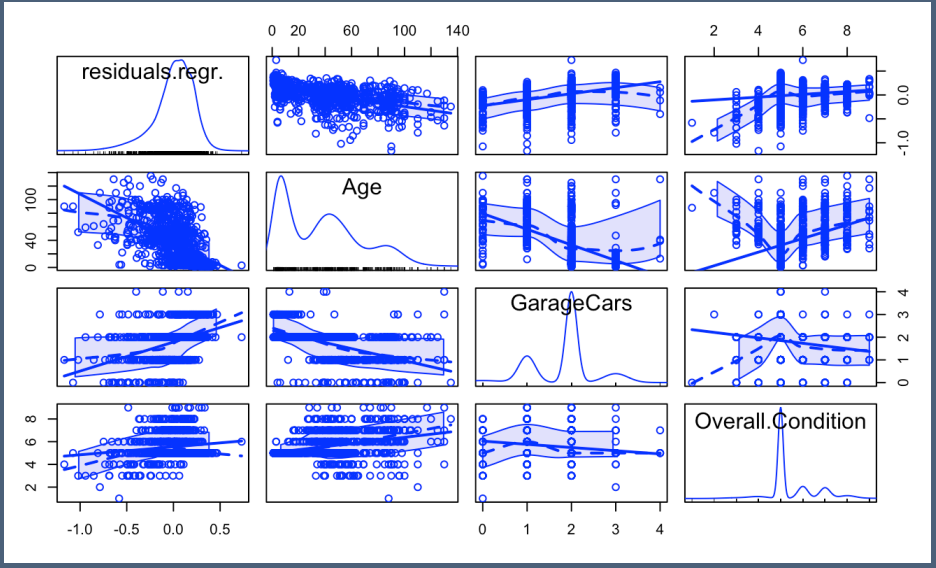
- Intuitive for the simple regression
  - Previously noted outlier is highly leveraged, but not very influential
  - Located about where you'd expect based on other home prices



Premature to be concerned whether data are normally distributed... we haven't finished building the model...

Graphical Analysis

- Scatterplot matrix shows pairwise associations
  - Scatterplot version of a correlation matrix
  - As a “response”, residuals show remaining variation to be explained
- Smooth curves
  - Loess curves indicate nonlinear association
  - Some collinearity, e.g. between Age and GarageCars.



What other features might be relevant?

Multiple Regression Model

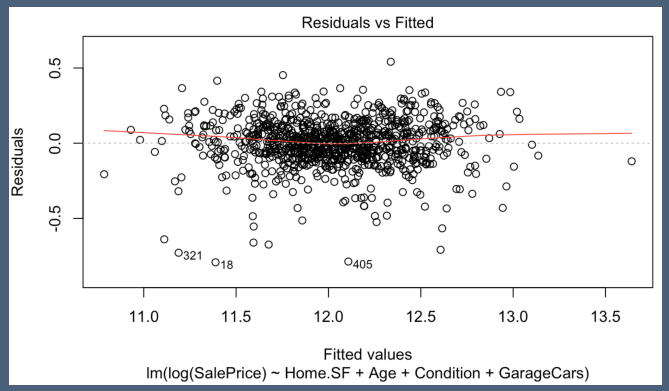
- Four predictors
  - Square footage, plus Age, Garage, and Condition
- Comparison to simple regression
  - Simple regression captures 68% of variation, this model increases to 84%
  - Added coefficients significant
  - Coefficient of Home.SF is smaller (0.00044 down to 0.00033). Why?

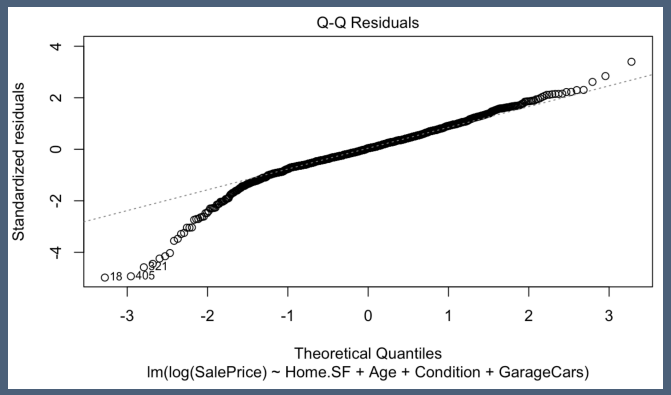
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.073e+01  3.511e-02  305.65   <2e-16 ***
Home.SF      3.291e-04  8.269e-06   39.80   <2e-16 ***
Condition    7.902e-02  4.983e-03   15.86   <2e-16 ***
Age         -4.573e-03  2.167e-04  -21.11   <2e-16 ***
GarageCars   1.058e-01  9.665e-03   10.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1598 on 953 degrees of freedom
Multiple R-squared:  0.8432,    Adjusted R-squared:  0.8426
F-statistic:  1282 on 4 and 953 DF,  p-value: < 2.2e-16
```

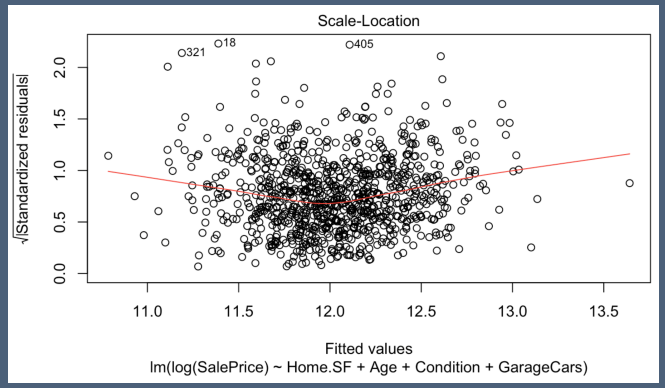
Significantly more predictive model than the prior simple regression?

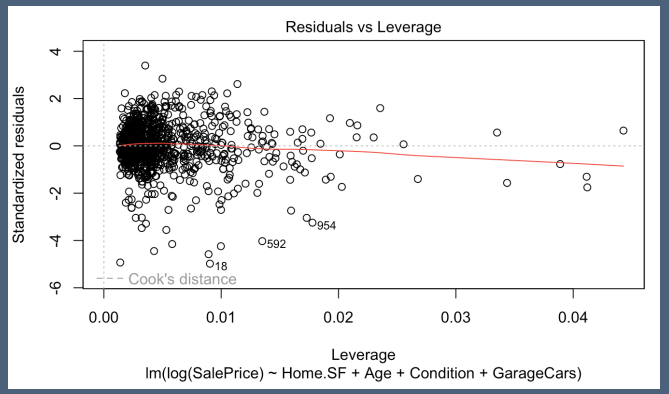
Model Diagnostics

- Checking the model assumptions…
  - Independent? Equal variance? Normally distributed?









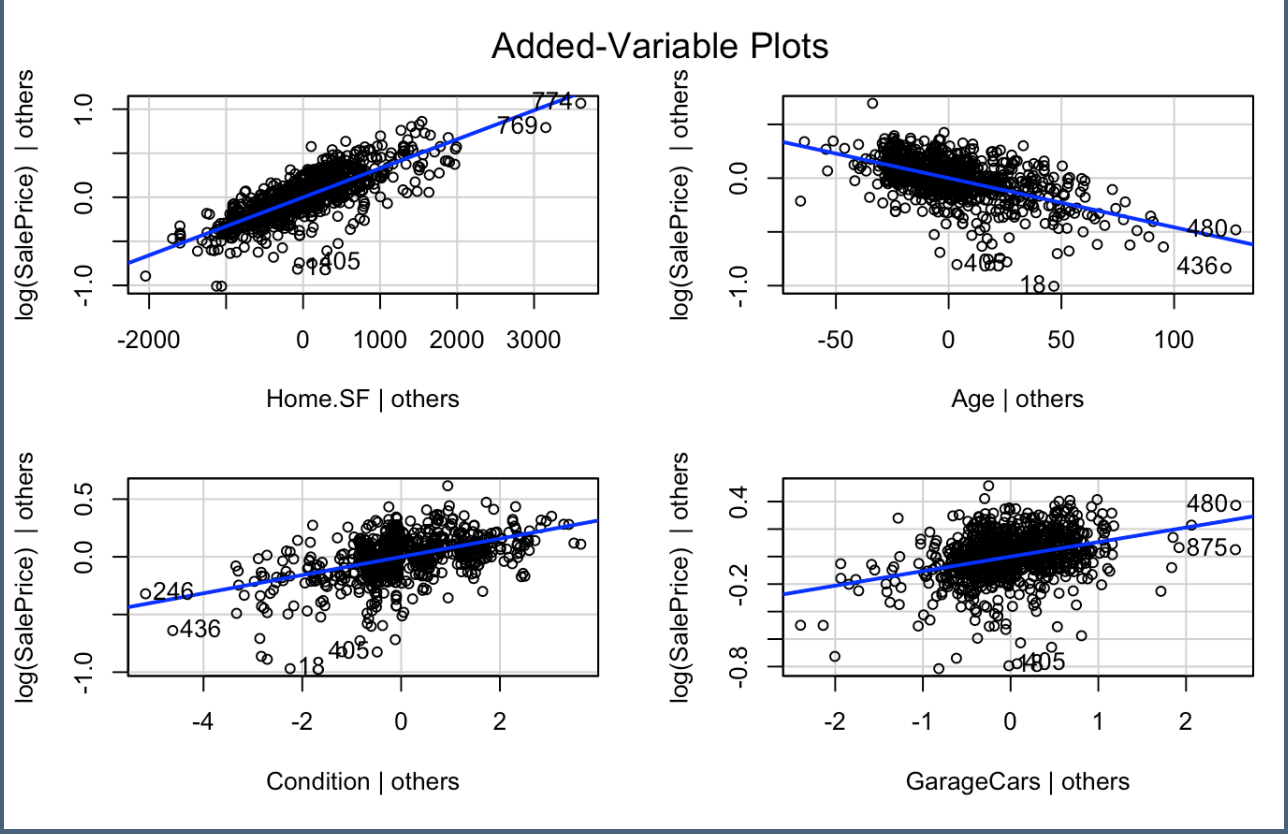
Consequences of deviations from the assumptions?

Partial Regression Plot

- Construction of multiple regression slope for one predictor (say Z_1)
  - Regress response on other predictors, save the residual $\hat{W}_1 = X - a_0 - a_2 Z_2 - \cdots - a_q Z_q$
  - Regress one predictor on other predictors, save the residual $\hat{Z}_1 = Z_1 - b_0 - b_2 Z_2 - \cdots - b_q Z_q$
  - Regress residuals of the response on residuals of the predictor, $\hat{W}_1$ on $\hat{Z}_1$
  - The slope in this simple regression is the coefficient of Z_1 in the regression of X on $Z_1, Z_2, \ldots, Z_q$
  - Also, the t-statistic in the simple regression is almost the t-statistic in the multiple regression. The difference lies in how the degrees of freedom are computed when estimating the error variance.
- Partial regression plot
  - Aliases for this plot: Leverage plot, added variable plot
  - Scatterplot of $\hat{W}_1$ on $\hat{Z}_1$
  - Use: "Simple regression view of a multiple regression slope"
  - Typically will not reveal non-linearity

Partial Regression Plots

- Partial regression plots from multiple regression
  - Some outliers and leveraged cases, but nothing dramatic



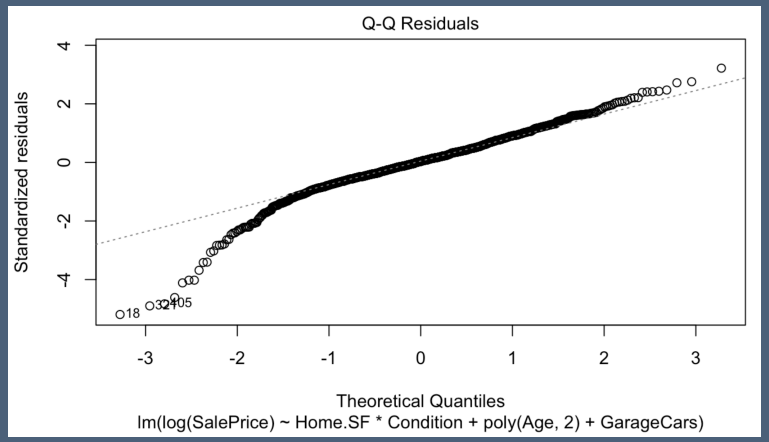
Next Steps

- Identify other explanatory variables
  - Automated search procedures such as stepwise regression, lasso
  - Role of model selection criteria
  - AIC and BIC prefer the model shown to the right
- Manufacture more possible predictors
  - Nonlinear predictors (e.g. logs or polynomials)
  - Interactions: synergy size and condition is not significant in this model.
- Categorical predictors
  - None used so far

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.059e+01  7.940e-02 133.393  < 2e-16 ***
Home.SF            3.187e-04  3.151e-05  10.116  < 2e-16 ***
Condition          7.927e-02  1.395e-02   5.684 1.75e-08 ***
poly(Age, 2)1     -4.522e+00  2.056e-01 -21.995  < 2e-16 ***
poly(Age, 2)2      9.514e-01  1.665e-01   5.715 1.47e-08 ***
GarageCars         9.392e-02  9.739e-03   9.644  < 2e-16 ***
Home.SF:Condition  8.235e-07  5.591e-06   0.147    0.883
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1572 on 951 degrees of freedom
Multiple R-squared:  0.8485,    Adjusted R-squared:  0.8475
F-statistic: 887.4 on 6 and 951 DF,  p-value: < 2.2e-16
```

Ready to Predict?

- Concerns
  - Residuals not normal, model over-predicts more than it under-predicts
  - Evidence of lack of constant variance for unusually small/large properties
  - Consequently might over-predict by a lot more than you'd expect.
  - You can see the problem in the original plot of the simple regression of log price on home size or in the plot of residuals on fitted for either multiple regression.



- How to fix this?
  - Do we have the right response: The original log transformation might not be the best choice.
  - Perhaps we should model cost per square foot instead?
  - Is there another explanatory variable that can explain these pricing errors?

What's Next?

- Check out Chapter 3 of ISL for more discussion of regression in R

- Next time... back to time series
  - Special considerations when using time series in regression models