

Lecture 7: Model Assumptions: Assessment and Repair

Part 1 of 2 (Conclusion in Lecture 8)

- Model assumptions for Linear Regression (review)
- What can go wrong
- An Example in which some assumptions are not satisfied
 - House prices in Zip 30062 (A different look!)
- “Repair”
 - Removal (or not) of Outliers
 - Transformations
 - *Polynomial regression is also a possibility. To be discussed later.*
 - Addressing Influential Points

Reading:

- This material is mostly in Chapters 5 & 6 of *Dielman*
 - Some of these chapters involve “multiple regression” ideas, but most of it only uses simple regression. Nevertheless,
- Class Notes may be the most useful source material for now
- As we progress through *Dielman* pay special attention to
 - Section 6.7.2+; Outliers
 - Section 6.4, & 5.2.2-5.2.4; Linearity and curvilinear relations
 - *Chapter 5.2.1, polynomial regression, is also relevant but will be treated later*
 - Section 6.5; Constant variance assumption (Homoscedasticity)
 - Section 6.6; Normality
 - Section 6.7.1; Influential points

Ordinary Linear Model Assumptions (review)

- Properties of errors under ideal model:
 - $\mu_{Y|x} = \beta_0 + \beta_1 x$ for all x .
 - $Y_i = \beta_0 + \beta_1 x_i + e_i$ for all x_i
 - All e_i have the same variance, σ_e^2 [“Homoscedasticity”]
 - The distribution of $e_i|x_i$ is normal. [Also $E(e_i|x_i) = 0$]
 - e_1, \dots, e_n are independent.
- Equivalent description:

For each x_i the corresponding Y_i has a normal distribution with mean $\beta_0 + \beta_1 x_i$, a linear function of x , and constant variance. Also Y_1, \dots, Y_n are independent.

Things to look for in the data (*and what to do about them*)

These things suggest deviations from the assumptions

Order of Inspection

1. Outliers

2. Non-linearity

– A Curved overall pattern as a function of x

3. Heteroscedasticity

– Changes in the *vertical spread* of the data for varying x

4. Non-normality of the residuals

5. Lack of Independence

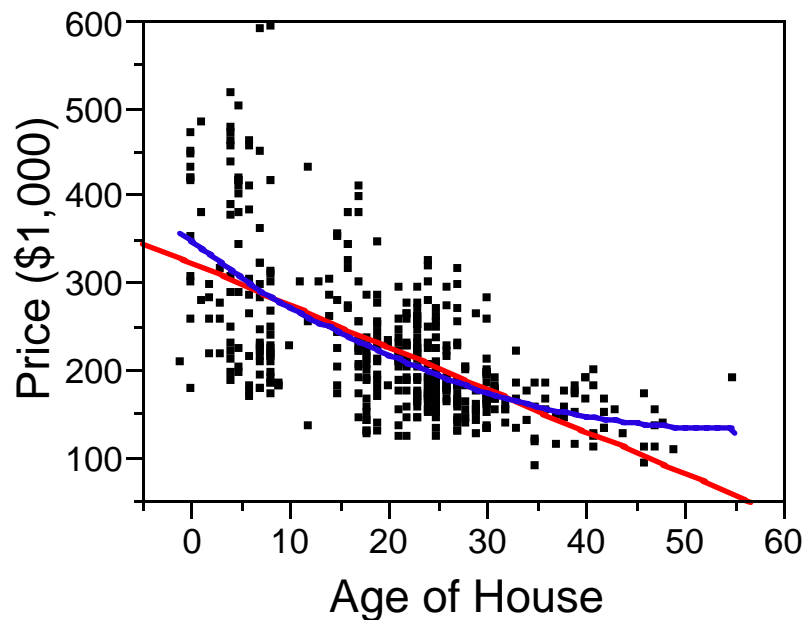
– This is fairly rare except in time-series data, *which will be studied later*

6. Influential Points:

– These are not a “deviation from the assumptions”, but they do require caution in interpreting the analysis

Example: House Prices in Zip 30062 (*cont*)

New perspective: This time we'll look at House Prices (for sales in 2003) as a function of the age of the house.



The red line is the LS line.

The smooth curve (blue) through the data is called a “*smoothing spline*”. The details of its construction are not important (for now). It follows the general pattern of the average of Y at each given x .

1. Outliers

- These are points far from the overall linear pattern for the average of Y given x .
- There are two possible outliers in the data [*Can you find them?*]
- Potential outliers should be examined to see whether they belong in the data.
 - Look for: *Mis-recorded data*, and for
Data that otherwise doesn't belong
- If they don't belong then they should be **Excluded**
 - If they do belong (or if their status is unclear), they can be **retained** in the analysis, **BUT**
- One should later check that any conclusions drawn from the data are not strongly dependent on just a few potential outliers
- This is what we'll do with our two questionable data points

2. Non-linearity

- This is a curved general pattern of the **means of Y given x** .
- Our data follows such a curved, non-linear pattern.
 - This is evident from the curved **spline** on the plot
 - It's pretty easy to see even without that **red** curve
- Non-linearity strongly suggests that the general model **is not valid**.
- Hence, the least squares line, and associated tests and CI^s , are **not** faithful representations of reality
 - *If the plot isn't too non-linear the LS analysis may still be somewhat useful.*
- There are various things that can be done to produce a valid analysis. (*See following slides*)

2. Non-linearity (cont)

Transformations

- Transformations of the variables x and/or Y may be useful.
- Transformations of Y will affect the variance of Y given x .
 - They are best suited for use when heteroscedasticity is also present.
 - In seemingly homoscedastic situations they should only be used with extreme caution to check for heteroscedasticity *after* use of the transformation

2. (cont) Transformations of the x variable

- Look for a transformation $t(x)$ of the x variable so that

$$\mu_{Y|x} = \beta_0 + \beta_1 t(x).$$

- A successful transformation will have $y = \beta_0 + \beta_1 t(x)$ passing nearly through the central pattern of the data
- Common $t(x)$ to try are --

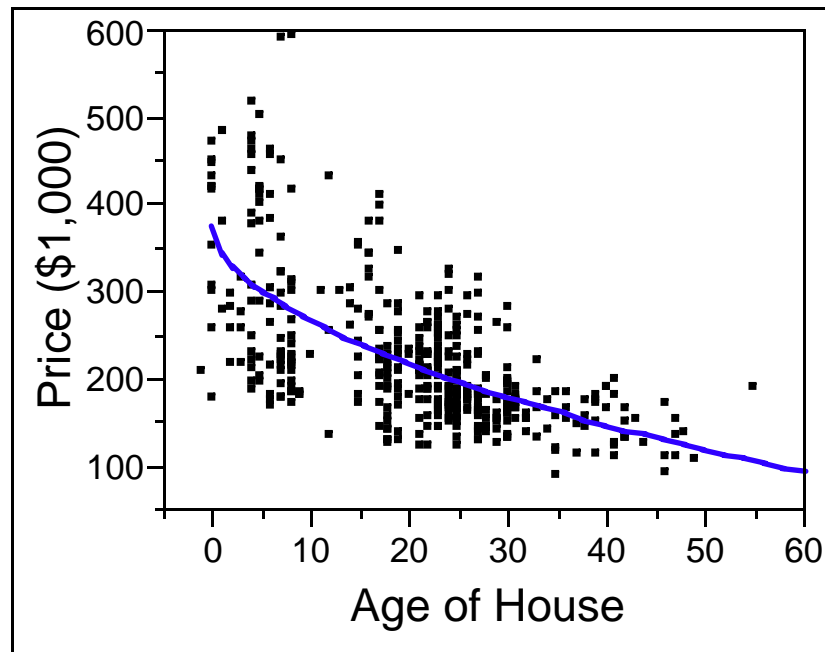
Common Transformations of x

<i>Group A</i>	<i>Group B</i>
$t(x) = \sqrt{x}$	$t(x) = x^2$
$t(x) = \ln(x)$ or $Log_{10}(x)$	$t(x) = e^x$
$t(x) = 1/x$	

- **Group A** transformations correct for situations with “decreasing returns to scale” *ie*, for a one unit increase in x the |change| in the mean of $Y|x$ becomes smaller as x increases.
- **Group B** transformations do the opposite.
- Group A transformations only work if all $x > 0$; if not, try applying them to $x+c$ for suitable c .
- Non-monotone patterns are often best treated by polynomial regression; to be discussed later.
- Our data is of the form suitable for trying a **Group A**.
- The transformation $t(x) = \sqrt{x}$ works reasonably well here.

Here’s the plot showing the **best** curve of form $y = \beta_0 + \beta_1 \sqrt{x}$.

- **Best** of this form is in the sense that β_0, β_1 minimize the SSE for all curves of this form



— Transformed Fit to Sqrt

$$\text{Price } (\$1,000) = 385 - 37.69 \text{ Sqrt}(\text{Age of House})$$

- You can use the corresponding ANOVA tables to compare the value of this fit to that of a straight line:

ANOVA Tables for fit to an ordinary line and to SqRt

Linear Fit: Price (\$1,000) = 323 - 4.857 Age of House

Summary of Fit

RSquare	0.395
Root Mean Square Error	66.0
Observations	438

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1240000	1241676	285.1378
Error	436	1900000	4355	Prob > F
C. Total	437	3140000		<.0001

Transformed Fit to Sqrt: Price (\$1,000) = 385 - 37.69 Sqrt(Age of House)

Summary of Fit

RSquare	0.418
Root Mean Square Error	64.8
Observations	438

Analysis of Variance

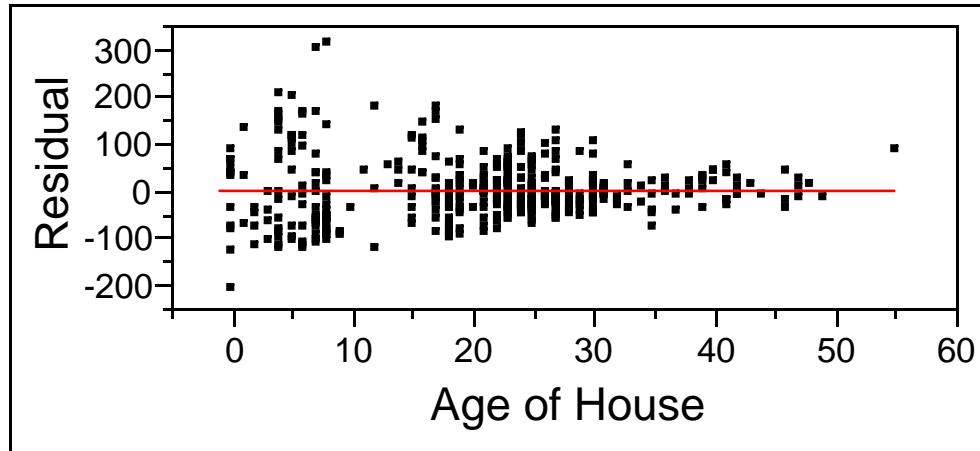
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1310000	1312537	313.0958
Error	436	1830000	4192	Prob > F
C. Total	437	3140000		<.0001

Compare the SSE^s or the R² values to see that the SqRt curve fits somewhat better. {NOTE: *One observation with Age = -1 has been excluded from both analyses.*}

3. Heteroscedasticity

- This refers to situations in which the population model's *error variances for given x* are not equal as x changes
- Look for *vertical spreads* in the scatterplot (or in the plot of residuals) that gradually \nearrow as $x \nearrow$ or that gradually \searrow as $x \nearrow$.
- *Other patterns of vertical spread can occasionally occur and be of interest.*
- We want fairly convincing evidence that the population model is not homoscedastic
 - Do not be overly affected by a few scattered points.
- Look at our **scatterplot**. In spite of the non-linearity confusing the view, you should still see the heteroscedasticity.
- The heteroscedasticity is much more evident in the **residual plot** about the SqRt fit; This is typically the better plot to look at in order to spot heteroscedasticity. Here is that plot --

Residual Plot for the residuals from the SqRt fit



- Such a plot shows the values of the residuals as a function of x .
 - The “Residuals” are the values of $Y_i - \text{Fit}(x_i)$
- You should see that the vertical spread at each x *generally* \searrow as $x \nearrow$. This is heteroscedasticity in the data.

Lecture 8 contains Part 2 of this discussion. It describes how to “fix” heteroscedasticity; and more.