Data Analytics for Business

Regression Diagnostics

Sridhar Narasimhan, Ph.D
Professor
Co-Director, Business Analytics Center
Scheller College of Business

Visual Exploration Before Doing Regression

GTx

Lessons in this Module

A. Visual Exploration Before Doing Regression
B. Anscombe's Quartet
C. Assumptions of Linear Models
D. Common Problems and Fixes in Fitting Linear Regression, Part 1
E. Common Problems and Fixes in Fitting Linear Regression, Part 2

GTx

Some Useful Graphs Before Regression

- Univariate graphs
- Two-dimensional graphs
- 3d plots
- Correlation Matrix plots

GTx

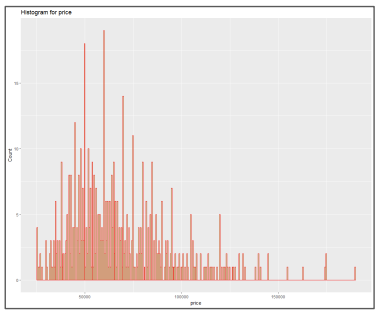
The First 10 Records in Housing

Housing Dataset in the *Ecdat* package in R

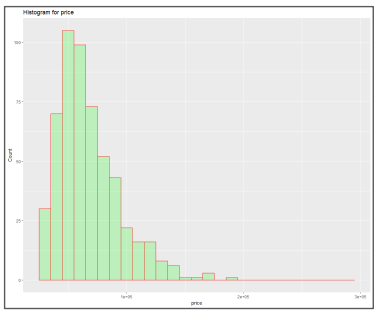
price	lotsize	bedrooms	bathrms	stories	driveway	recroom	fullbase	gashw	airco	garagepl	prefarea
42000	5850	3	1	2	yes	no	yes	no	no	1	no
38500	4000	2	1	1	yes	no	no	no	no	0	no
49500	3060	3	1	1	yes	no	no	no	no	0	no
60500	6650	3	1	2	yes	yes	no	no	no	0	no
61000	6360	2	1	1	yes	no	no	no	no	0	no
66000	4160	3	1	1	yes	yes	yes	no	yes	0	no
66000	3880	3	2	2	yes	no	yes	no	no	2	no
69000	4160	3	1	3	yes	no	no	no	no	0	no
83800	4800	3	1	1	yes	yes	yes	no	no	0	no
88500	5500	3	2	4	yes	yes	no	no	yes	1	no

GTx

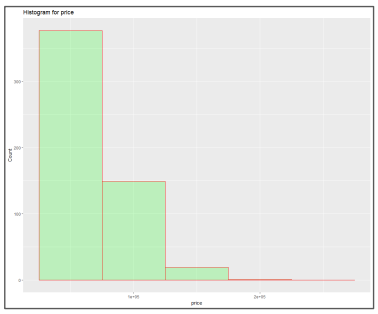
Univariate - Histograms



Bin size = $500



Bin size = $10,000

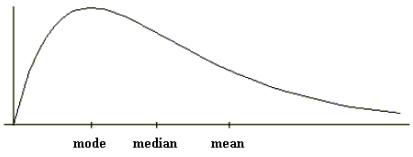


Bin size = $50,000

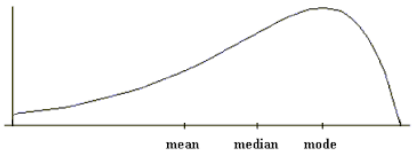
- Histograms of house prices with different bin sizes
- You can get an idea if the data is normally distributed or is skewed
- You need to be careful about the bin size (the form depends on bin size)

GTx

If a Variable is Skewed



Right skewed



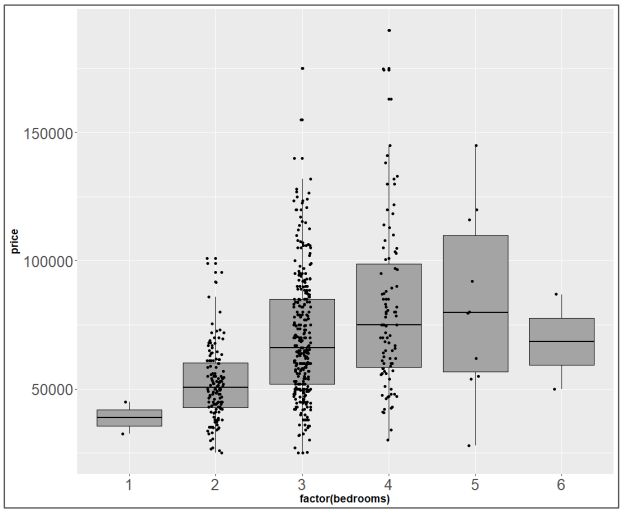
Left skewed

- You may need to use Log(variable) in your model (add 1 to the variable in case 0 values are present)
- Usually, we use natural logs since they are easier to interpret
- Logarithm of a large number is a small number, i.e., taking log is one way to do a variance reducing transformation

GTx

Univariate - Boxplots

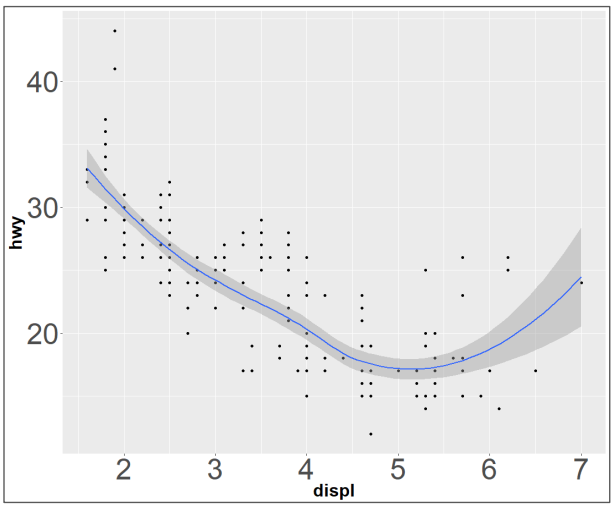
- The central box is drawn between the 25th and 75th percentile of the data
- The line in the box represents the median value
- Can be used to detect skewness (e.g., boxplot for houses with 3 bedrooms which is right skewed)
- Can be used to show outliers (e.g. boxplot for 4 bedrooms which has some very high priced houses)



GTx

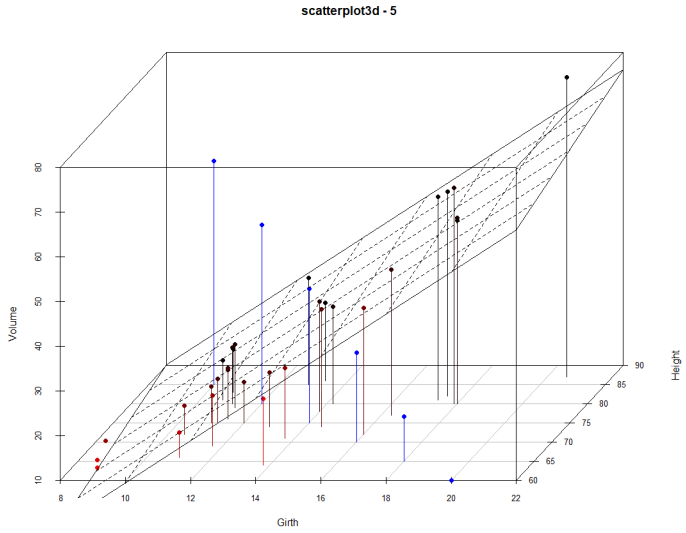
Bivariate Data - Scatterplots

- The scatterplot is probably the most helpful type of graph for displaying two variables
- It is helpful (sometimes) to add a function, like the loess smoothing function in this plot
- Scatterplots with two discrete variables are difficult to use



GTx

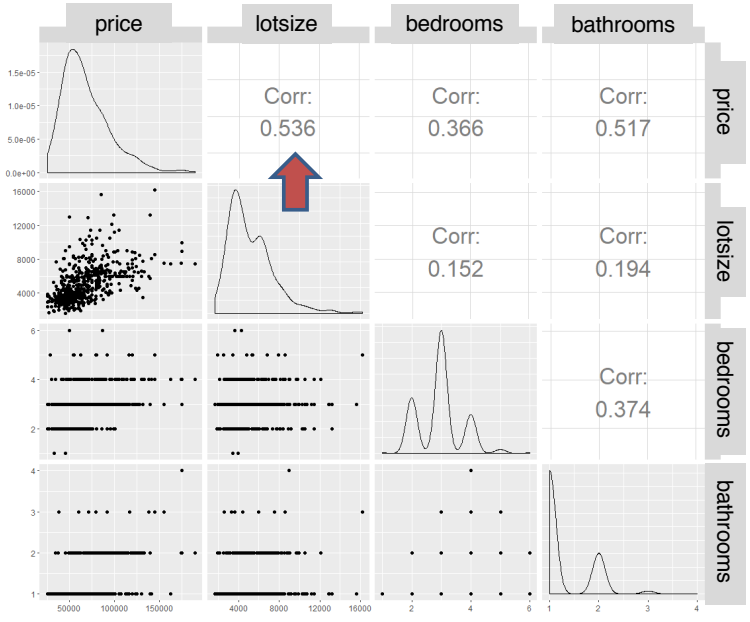
Scatterplot3d



Example 5 in scatterplot3.pdf with linear model, lm(Volume ~ Girth + Height)

GTx

Correlation Matrix



GTx

Data Analytics for Business
Regression Diagnostics

Sridhar Narasimhan, Ph.D
Professor
Co-Director, Business Analytics Center
Scheller College of Business

Anscombe's Quartet

GTx

Anscombe's Quartet (1973).

Francis Anscombe's main idea in this paper is that graphs are very valuable for checking modeling assumptions.

He wanted to counter the impression among statisticians that:

1. "numerical calculations are exact, but graphs are rough;
2. for any particular kind of statistical data there is just one set of calculations constituting a correct statistical analysis;
3. performing intricate calculations is virtuous, whereas actually looking at the data is cheating."

Anscombe, F. J. (1973). "Graphs in Statistical Analysis," American Statistician. 27 (1): 17–21.

GTx

Anscombe's Quartet: Four Regressions

$$y_1 = a_0 + a_1*x_1 \; ; \; y_2 = b_0 + b_1*x_2 \; ; \; y_3 = c_0 + c_1*x_3 \; ; \; y_4 = d_0 + d_1*x_4$$

x1	x2	x3	x4	y1	y2	y3	y4
10	10	10	8	8.04	9.14	7.46	6.58
8	8	8	8	6.95	8.14	6.77	5.76
13	13	13	8	7.58	8.74	12.74	7.71
9	9	9	8	8.81	8.77	7.11	8.84
11	11	11	8	8.33	9.26	7.81	8.47
14	14	14	8	9.96	8.1	8.84	7.04
6	6	6	8	7.24	6.13	6.08	5.25
4	4	4	19	4.26	3.1	5.39	12.5
12	12	12	8	10.84	9.13	8.15	5.56
7	7	7	8	4.82	7.26	6.42	7.91
5	5	5	8	5.68	4.74	5.73	6.89

GTx

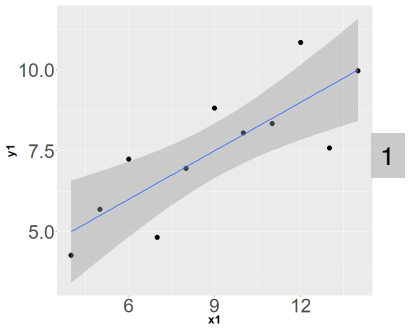
```
lm(formula = y1 ~ x1, data = anscombe)

Residuals:
     Min       1Q   Median       3Q      Max
-1.92127 -0.45577 -0.04136  0.70941  1.83882

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001     1.1247   2.667  0.02573 *
x1            0.5001     0.1179   4.241  0.00217 **
---
Signif. codes:  '**' 0.01 '*' 0.05

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6665,
Adjusted R-squared:  0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

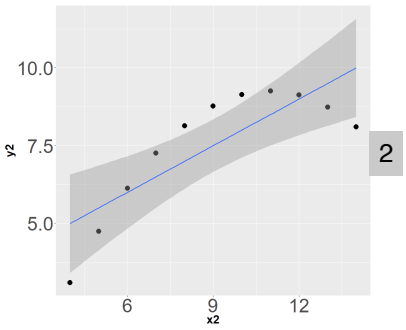
```
lm(formula = y2 ~ x2, data = anscombe)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9009 -0.7609  0.1291  0.9491  1.2691

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.001      1.125   2.667  0.02576 *
x2             0.500      0.118   4.239  0.00218 **
---
Signif. codes:  '**' 0.01 '*' 0.05

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6662,
Adjusted R-squared:  0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179
```

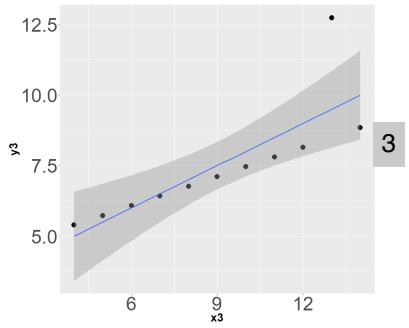
```
lm(formula = y3 ~ x3, data = anscombe)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1586 -0.6146 -0.2303  0.1540  3.2411

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0025     1.1245   2.670  0.02562 *
x3            0.4997     0.1179   4.239  0.00218 **
---
Signif. codes:  '**' 0.01 '*' 0.05

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6663,
Adjusted R-squared:  0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
```

```
lm(formula = y4 ~ x4, data = anscombe)

Residuals:
   Min     1Q Median     3Q    Max
-1.751 -0.831  0.000  0.809  1.839

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0017     1.1239   2.671  0.02559 *
x4            0.4999     0.1178   4.243  0.00216 **
---
Signif. codes:  '**' 0.01 '*' 0.05

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6667,
Adjusted R-squared:  0.6297
F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

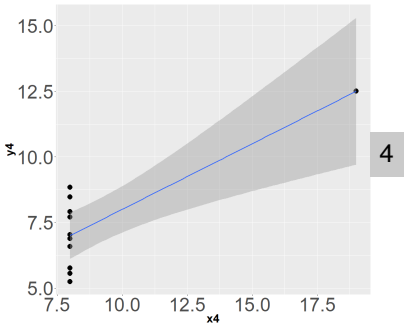
GTx



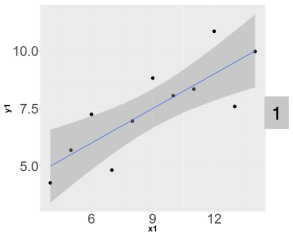


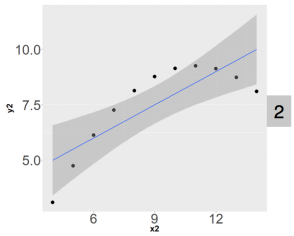


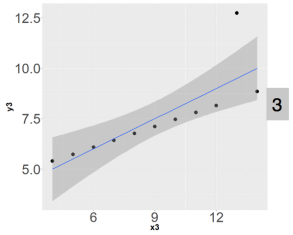
The regression lines are identical. But are there any problems?



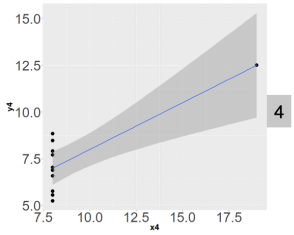
GTx







The regression lines are identical. But are there any problems?



- Regression 1 looks fine.
- Regression 2 requires a nonlinear transformation.
- Regression 3 has a single large (vertical) outlier that has a large residual which is influential on the fitted line.
- Regression 4 has a large (horizontal) outlier (also called a leverage point) which is influential on the fitted line.

GTx

Quiz (True/False)

- Anscombe's main idea in his 1973 paper was that graphs are NOT valuable for checking modeling assumptions.
  Answer: **FALSE**

- For the four models Anscombe fitted for their respective data sets, the parameter estimates are exactly the same and the statistics are identical.
  Answer: **TRUE**

GTx

Data Analytics for Business

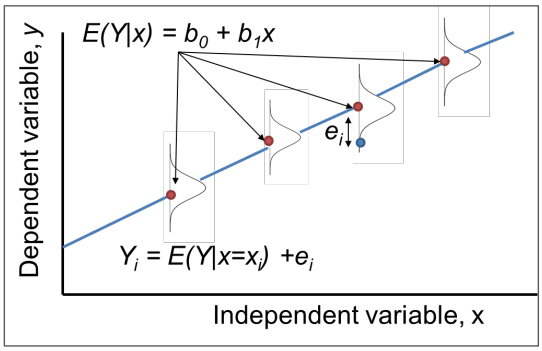
Regression Diagnostics

Sridhar Narasimhan, Ph.D
Professor
Co-Director, Business Analytics Center
Scheller College of Business

Assumptions of Linear Models

GTx

Assumptions of Linear Regression



- **Linearity assumption**: E(y) = $b_0 + b_1x$, i.e., the expected value of Y at each value of X approximates to a straight line
- **Assumption about errors**: The error terms e_i are independently and identically distributed (iid) normal random variables, each with mean zero and constant variance σ^2 (homoscedasticity)
- **Assumptions about predictors**: In multiple regression, the predictor variables are assumed to be linearly independent of one another

GTx

Data Analytics for Business

Regression Diagnostics

Sridhar Narasimhan, Ph.D
Professor
Co-Director, Business Analytics Center
Scheller College of Business

Common Problems and Fixes in Fitting Linear Regression, Part 1

GTx

Most Common Problems in Fitting Linear Regression

1. Non-linearity of the response-predictor relationships
2. Correlation of error terms
3. Non-constant variance of error terms
4. Outliers
5. High-leverage points
6. Collinearity

James, Gareth, et al. *An introduction to statistical learning: with applications in R* (Section 3.3.3). Springer, 2017.

GTx

1. Is the Relationship Nonlinear?

- Check the scatter plots of **Y** vs. each **X** variable. Linear?
- Another plot to use is the residuals plot vs. fitted values plot (especially useful in multiple regression)
  - We want to see no patterns

GTx

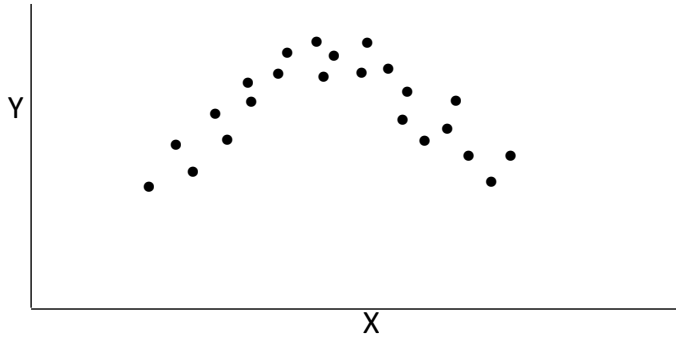
1. Is the Relationship Nonlinear?

- If there are concerns, then
  - Can you model non-linear relationship with higher order terms (e.g., square)?
  - Use variance reducing transformation (such as log) that will give a better linear fit
  - Are there outliers or certain sections of the observations that seem to drive the non-linearity?
  - Is there any important variable that you left out from your model (e.g., age or gender)?
  - Or, maybe, was there systematic bias when collecting data, hence redesign data collection
- Checking residuals helps discover useful insights about your model and data

GTx

Scatterplot of the Response Variable Versus the Explanatory Variable

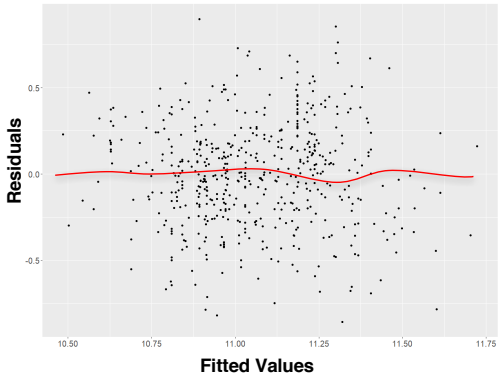
- This is useful to do before fitting a model.
- You can plot Y vs. X to identify any patterns, For example, the scatter plot below suggests using X^2 rather than X as a predictor.



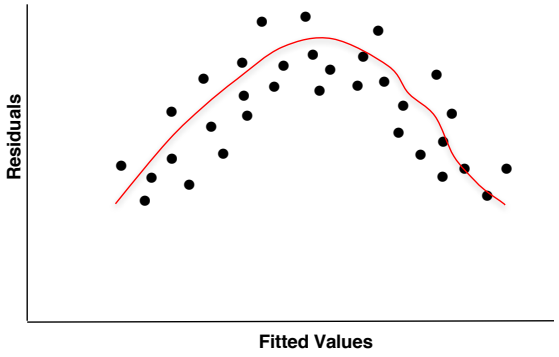
GTx

Residuals vs. Fitted Values Plot

Case 1



Case 2



This plot should be examined after a model is fitted.

GTx

2. Correlation of Error Terms

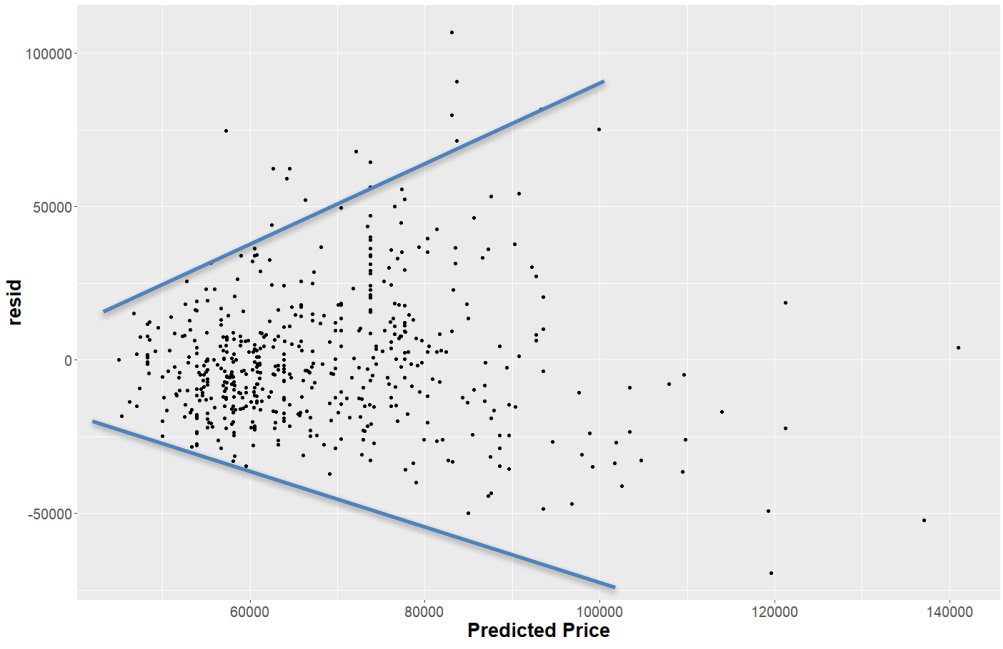
- An important assumption is that error terms $e_1, e_2, \ldots, e_n$ are uncorrelated. If they aren't, then we have autocorrelation.
- So knowing the value of e_i should not have any influence on the magnitude or size of e_{i+1}.
- This property is used to estimate the standard errors of the parameters of the model.
- If there is correlation in the error terms:
  - The estimated standard errors will underestimate the true standard errors.
  - Confidence and prediction intervals will be narrower than they should be and p values will be lower than they should be.
  - We may have sense of confidence in the model that is not warranted.
- The Durbin-Watson test is used to detect autocorrelations in a linear model

GTx

3. Heteroskedasticity (Non-constant Error Variance)

- The assumption is that the spread of the responses around the straight line is the same at all levels of the explanatory variable (i.e., we have constant variance or homoscedasticity).
- You may have non-constant error present (e.g., the errors increase in size with the fitted values). You can detect this with the residuals vs. fitted values plot.
- If non-constant error is present, then Hypotheses tests and Confidence Intervals can be misleading.
- If there is Heteroscedasticity, then transformation of the Y variable may be called for.
- Example: ln(Y), or 1/Y, etc.

GTx



GTx

Quiz (True/False)

- If the scatterplot of Y vs. X shows a nonlinear pattern, then we should not change our linear regression model.
  Answer: **FALSE**
- Autocorrelation is the correlation between each of the e_i variables.
  Answer: **TRUE**
- Heteroskedasticity means having constant Error Variance.
  Answer: **FALSE**

GTx

Data Analytics for Business

Regression Diagnostics

Sridhar Narasimhan, Ph.D
Professor
Co-Director, Business Analytics Center
Scheller College of Business

Common Problems and Fixes in Fitting Linear Regression, Part 2

GTx

4. Outliers

- An outlier is a point that has a y_i value that is far from its predicted value, $\hat{y}_i$
- One way to visualize outliers is to plot residuals (or, better yet, standardized residuals) against predicted values of *y.*
- Outliers could occur because of incorrect data recording or because the phenomenon could very well be non-linear.
- Do not assume that an outlier observation should be removed. It may signal a model deficiency (for e.g., a missing predictor).

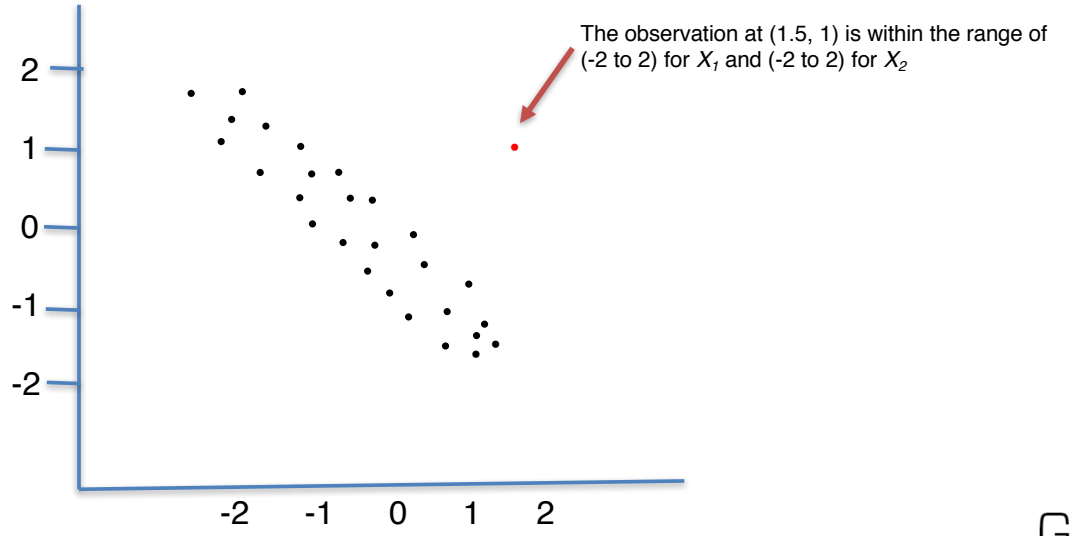
GTx

5. High Leverage Points

- In Anscombe's quartet, Regression 4 has a large (horizontal) outlier (also called a leverage point), which is influential on the fitted line.
- In simple regression, look for observations that have a predictor value outside the normal range of observations.
- A point has high leverage if its deletion (by itself or with 2 or 3 other points) causes noticeable changes in the model.
- With many predictors in a model, one could have an observation that is within the range of each predictor's value but still be unusual.

GTx

. High Leverage Points - Example



GTx

Cook's Distance

- One statistic to identify influential points is the ***Cook's Distance*** C_i that measures the difference between the regression coefficients obtained
  (a) from the full data and
  (b) from deleting observation *i*
- A rule of thumb is to identify points with $C_i > 1$ as highly influential.

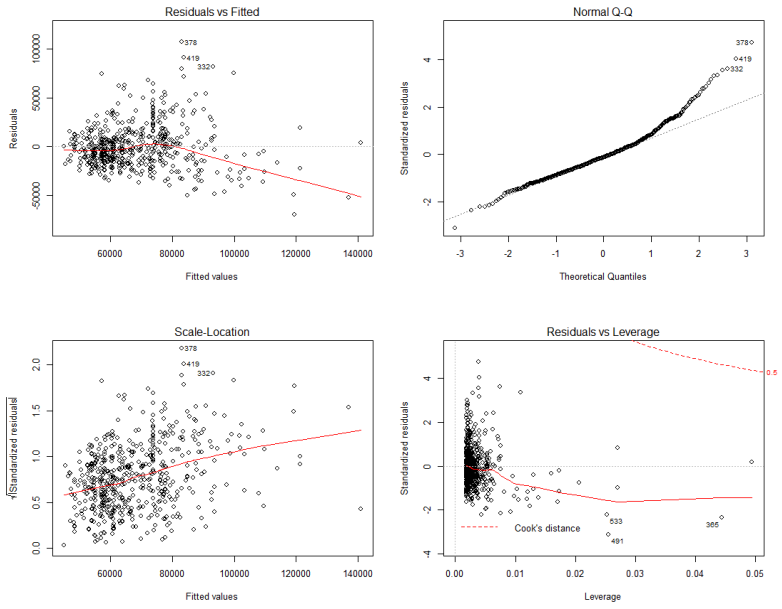
GTx

plot(model) in R

- The plot function in R, when applied to a linear model, provides four plots that are:
  - Residual vs. Fitted (check if residuals have non-linear patterns)
  - Normal Q-Q (check if residuals are normally distributed)
  - Scale-Location (check if √ (standardized residuals) are spread equally along the range of fitted values)
  - Residuals vs Leverage (to find influential points, if any, with $C_i > 1$)

GTx

a.lm <- lm(formula = price ~ lotsize , data = Housing)
plot(a.lm)



GTx

6. (Multi)Collinearity

- Run these two linear regression models in R and then compare their respective coefficient of cylinders:
  - Reg1 <- lm(formula = mpg ~ cylinders, data = Auto)
  - Reg2 <- lm(formula = mpg ~ cylinders + displacement + weight, data = Auto)

- Reg1
  cylinders -3.5581 0.1457 -24.43 <2e-16 ***

- Reg2
  cylinders -0.2678 0.4131 -0.648 0.517

- In Reg2, cylinder's coefficient is no longer statistically significant!

GTx

What Could Be the Reason?

- Such a change in a parameter estimate could indicate the presence of multicollinearity in Reg2.
- Multicollinearity: two or more of the explanatory variables are more or less linearly related.
- To detect multicollinearity, one approach is to use **Variance Inflation Factors (VIF).**
- Regress predictor variable X_j against all other predictor (X) variables. Name the resulting R^2 as R_j^2
- Define $VIF_j = 1/(1-R_j^2)$, $j = 1,2,\ldots, p$
- If X_j has a strong linear relationship to other X variables, then R_j^2 is close to 1, and VIF_j will be large.
- Values of VIF > 5 signify presence of multicollinearity (rule of thumb).

GTx

VIF

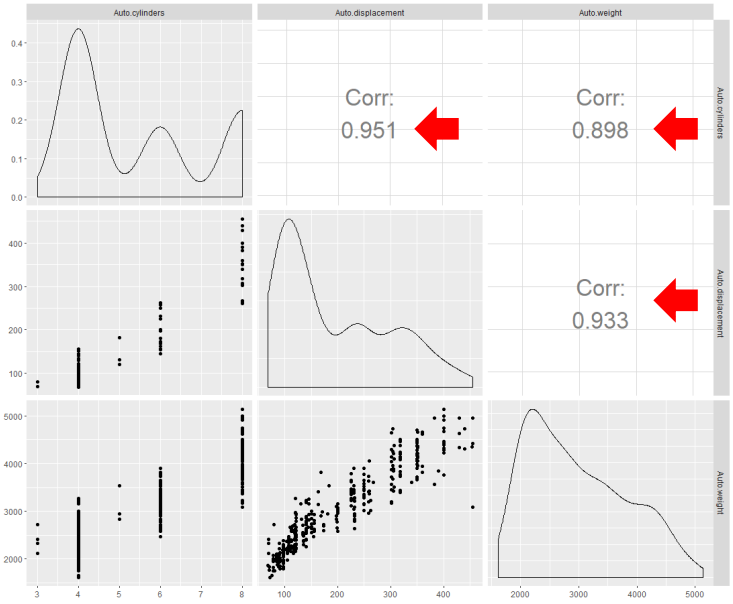
- We use the vif function on the predictors of the Reg2 model and obtain

```
vif(Reg2)
     cylinders      displacement      weight
     10.515508      15.786455         7.788716
```

- These high VIF values indicate Multicollinearity problems

GTx

Correlation Matrix



GTx

Consequences of Multicollinearity

- Multicollinearity: the explanatory variables are highly correlated
  - If $VIF_j = 1/(1-R_j^2) > 5$, multicollinearity present…
- Consequences of multicollinearity
  - OLS estimated parameters may have large variances and covariances, thus making precise estimation difficult
  - The confidence intervals of the estimated parameters tend to be bigger, hence we may not be able to reject H_0 (the null hypothesis, $b_i = 0$)
  - Regression coefficients have the wrong sign, or
  - Regression coefficients are not significantly different from 0 although R^2 is high
  - Adding an explanatory variable changes other variables' coefficients

GTx

Consequences of Multicollinearity

- Solution?
  - Pick one variable if two measure the same "thing"
  - Use Principal Components Analysis or Factor Analysis to create more useful variable(s)

GTx

Recap of this Module

A. Visual Exploration Before Doing Regression
B. Anscombe's Quartet
C. Assumptions of Linear Models
D. Common Problems and Fixes in Fitting Linear Regression, Part 1
E. Common Problems and Fixes in Fitting Linear Regression, Part 2

GTx