STAT 139: STATISTICAL SLEUTHING THROUGH LINEAR MODELS

Lecture 18
Nov 4, 2014

Victoria Liublinska

Odds and Ends

- HW 8 has been posted – start early! Due on Fri, 11/07
  - Q2 (i) part (c) and Q2 part (iii) are identical – just ignore one of them.
- We will review your project proposals and assign one of the teaching staff members to your group.
  - You will hear form us by the end of the week.

Previous lecture: Review

- Calibration: Estimating X that results in $Y = Y_0$
  - Analytical method: $\hat{X} = (Y_0 - \hat{\beta}_0) / \hat{\beta}_1$ and CIs are based on the appropriate estimate of SE.
  - Graphical method.

- Alternative interpretation of the slope estimator,

$$\hat{\beta}_1 = \sum_{i=1}^{n}\left[\omega_i \frac{\left(Y_i - \bar{Y}\right)}{\left(X_i - \bar{X}\right)}\right], \text{ where } \omega_i = \frac{\left(X_i - \bar{X}\right)^2}{\left(\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)} \text{ and } \sum_{i=1}^{n}\omega_i = 1$$

Is Linear Regression Unethical?

http://andrewgelman.com/ Prof. Andrew Gelman's blog.

Is linear regression unethical in that it gives more weight to cases that are far from the average? (July, 8, 2012)

- *If each data point represents a person we are weighting people differently.* Is it ethical?...
- Non-parametric rank-based tests are robust to outliers. *Again, maybe that outlier is the 10th person who dies from an otherwise beneficial medicine. Should we ignore him in assessing the effect of the medicine?*
- The general point […] is that there is no "ethically" neutral method.

Previous lecture: Review

- Correspondence between a pooled two-sample t-test and a simple linear regression with a binary group indicator X.

- Simple Linear Regression: Assumptions and Diagnostics
  - Linearity
  - Independence of errors

Today's overview

- Simple Linear Regression: Assumptions and Diagnostics
  - Linearity
  - Independence of errors, cont.
  - Equal variance of errors
  - Normality of errors
- Interpretation of results after log transformation.
- Sum of Squares decomposition for the linear regression and an R-squared (R^2) statistic.

Reading:

- **Required:** R&S Ch. 8, Ch. 8 R code
- **Supplementary Theory**: A. Sen and M. Srivastava. "Regression Analysis: Theory, Methods, and Applications", Ch 1 Sec. 1.7, Ch. 5, Ch. 6, and Ch. 9.

Simple Linear Regression: Assumptions and Diagnostics

- Independence of errors ε_i. Residuals for any two observations Y_i and Y_j do not "travel together" after taking into account the corresponding X values.

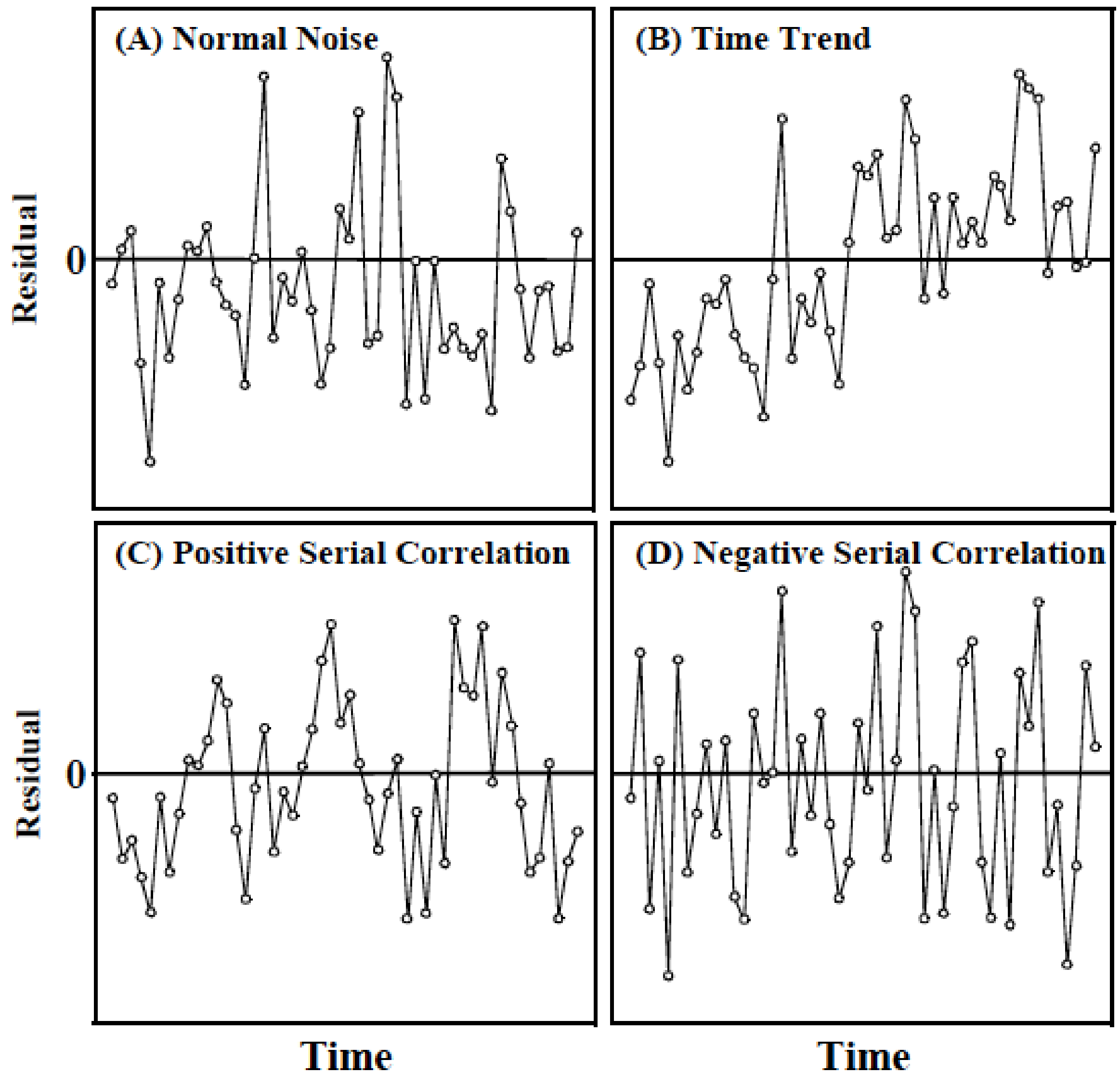
  - Checking: Were all independent predictors included in the model of $\mu(Y|X)$? Examine the design.
    - Plot residuals vs. time/distance, when applicable.
  - If violated:
    - Doesn't lead to bias in $\hat{\beta}_0, \hat{\beta}_1$ but standard errors are affected (tests and CIs can be misleading).

Simple Linear Regression: Assumptions and Diagnostics

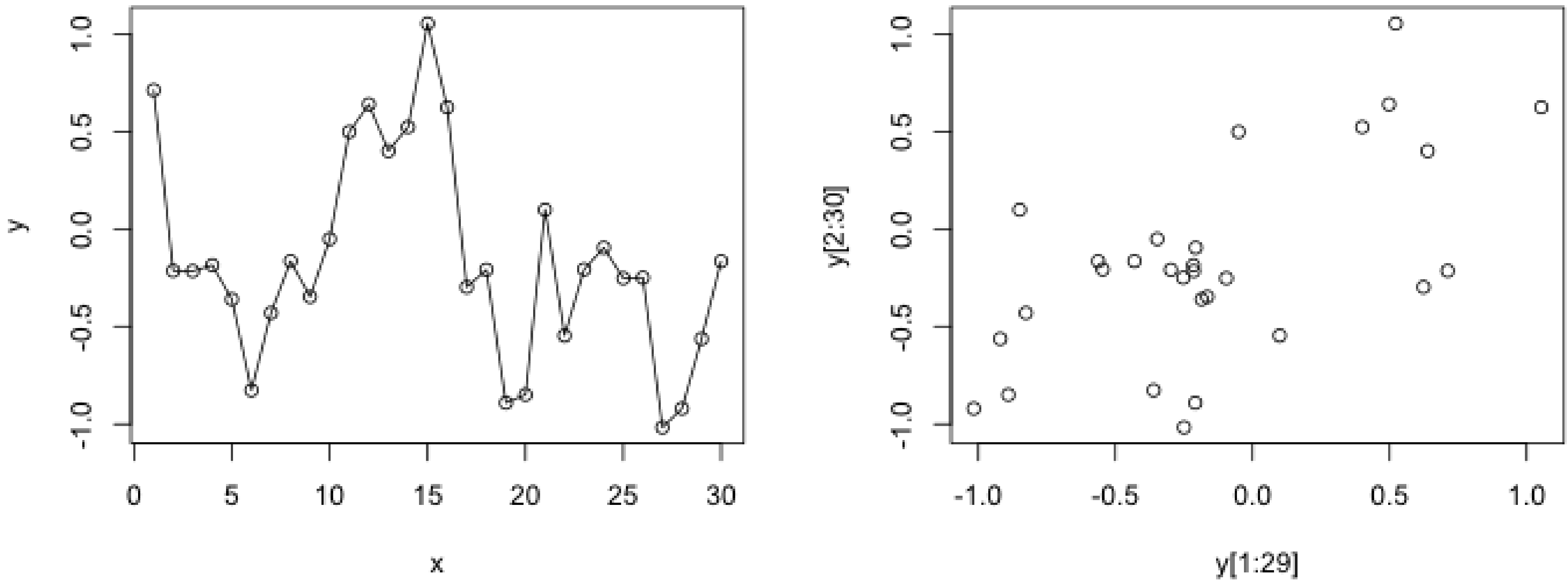
- Independence of errors ε_i. Residuals for any two observations Y_i and Y_j do not "travel together" after taking into account the corresponding X values.

  - Strategies:
    - Add more predictors (Ch. 9), group units in the same cluster.
    - For serial effects see Ch. 15 (models for time series).
    - For cluster effects or repeated observations, consider linear regression with correlated errors, including
      - Multilevel (or random-effect(s)) models (Gelman & Hill, 2007 – on reserve),
      - MANOVA or Repeated Measures ANOVA (Ch. 16).

Possible patterns in plots of residuals versus time order of data collection



Correlation Between Consecutive Observations



Further Notes on Independence

When checking this assumption, consider the following questions:

- Do units interact in any way? (for example, belong to the same household.)
- Is there a spatial (and temporal) proximity? – Study units closer together (in space and time) are more likely to behave similarly than units farther apart.
- Is there a common data-generating source? (for example, repeated measurements on the same subject.)
- Are there clusters where units tend to have similar responses? (for example, weight of cubs in the same litter.)

Further Notes on Independence

- If errors are positively correlated we really have less information than we presume, which results in confidence intervals that are too short and Type I error that is inflated.

Lack of independence between Y_i and Y_j implies that only part of the information about β_0 and β_1 added by Y_j is new; the rest has already been gained from Y_i.

Strategies for dealing with lack of independence:

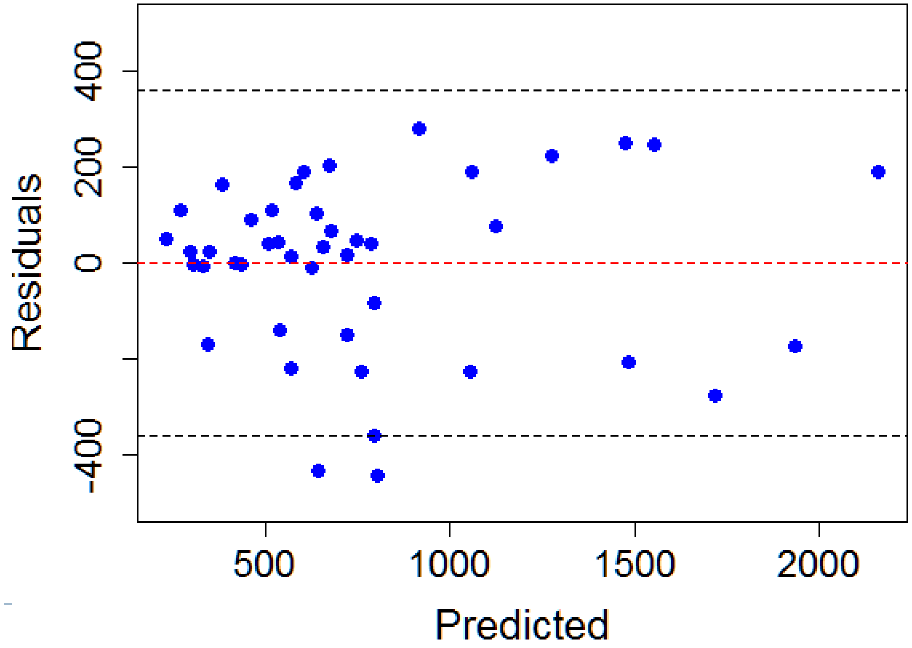
1. Ensure independence by study design;
2. Model dependencies.

Simple Linear Regression: Assumptions and Diagnostics

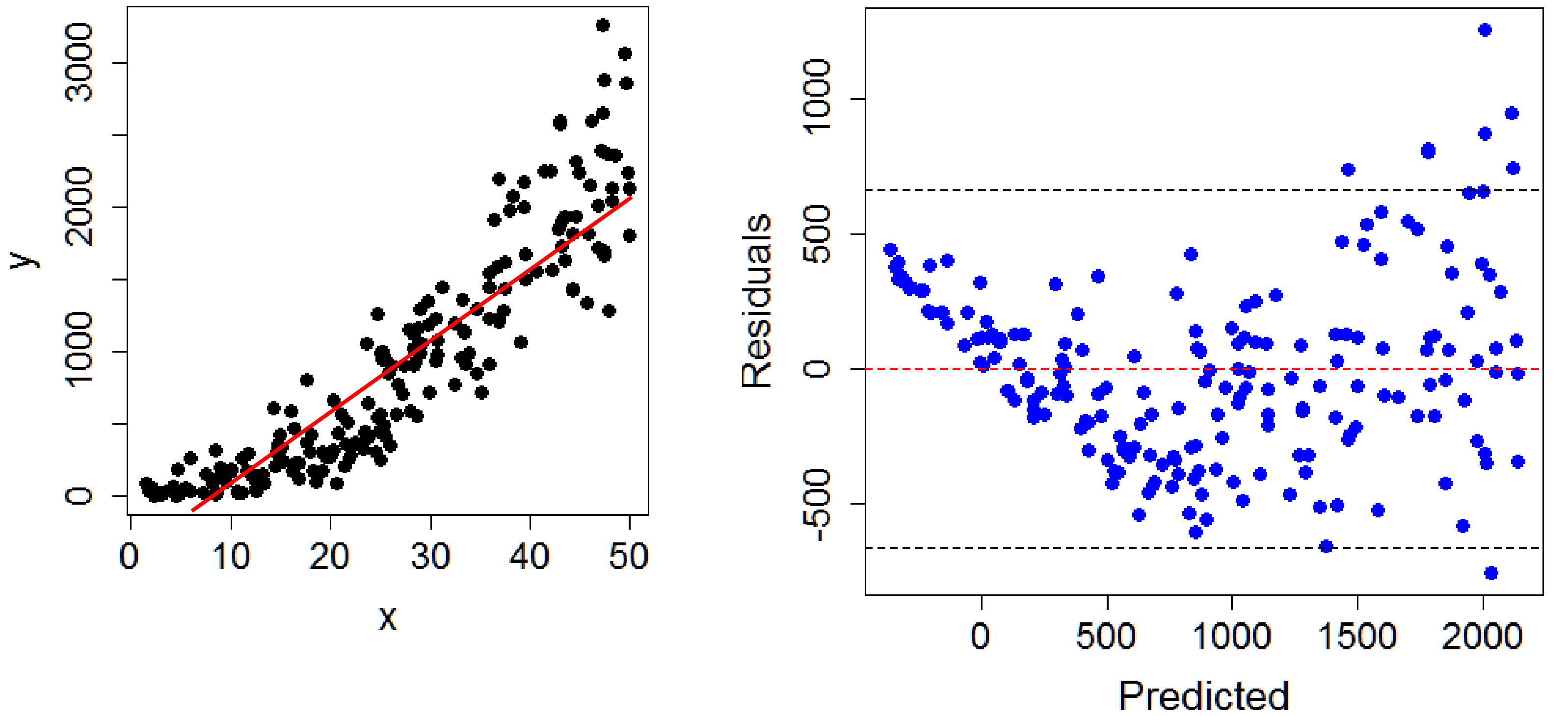
- Equal variance of errors, $\mathrm{Var}(\varepsilon_i) = \sigma^2$.
  - Checking: scatterplot, residual plot.
  - If violated:
    - Doesn't lead to bias in $\hat{\beta}_0, \hat{\beta}_1,$ but standard errors are affected (tests and CIs can be misleading).
    - Prediction is more sensitive.
  - Strategies:
    - Consider transformations ($\log(x)$, $1/x$, x^2, $\log(y)$, $1/y$, etc.).
    - Alternatively, use weighted regression, where each observation is weighted inversely proportional to its variance (R&S Section 11.6.1).

Newton Data: Fitted Values vs. Residuals

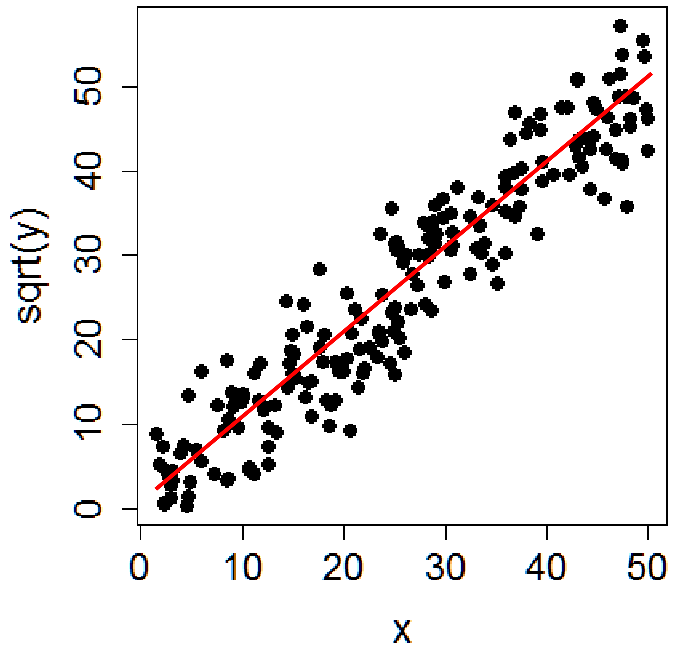
```
regmodel <- lm(Price/1000 ~ Sqft., data = SaleData)
library(arm)
y.hat <- fitted(regmodel)
u <- resid(regmodel)
sigma <- sigma.hat(regmodel)
residual.plot(y.hat, u, sigma, ylim=c(-500,500), main="")
```

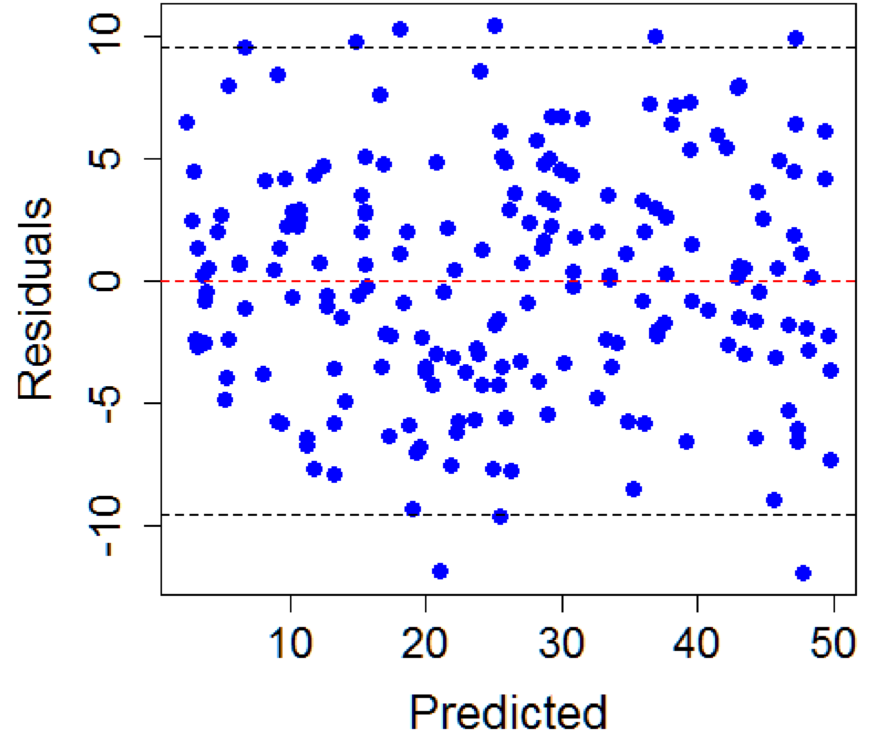


Fitted Values vs. Residuals: Horn- (or Funnel-) Shaped Patten



Fitted Values vs. Residuals: Removal of the Horn-Shaped Patten with a Transformation



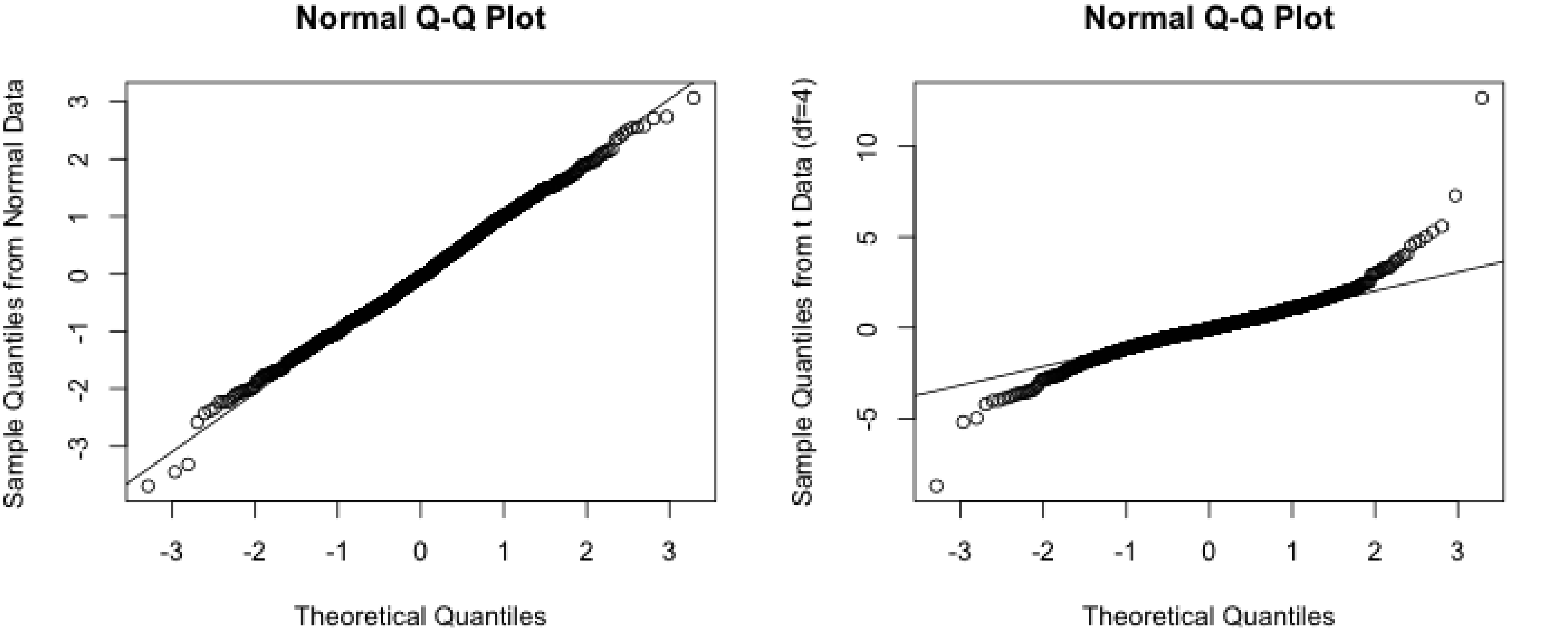


- Horn-shaped pattern in the residual plot suggest response transformation.
- Transformations ordered by their ability to correct unequal spread, from most to least severe one: $1/Y$, $\log(Y)$, $\text{sqrt}(Y)$.

Simple Linear Regression: Assumptions and Diagnostics

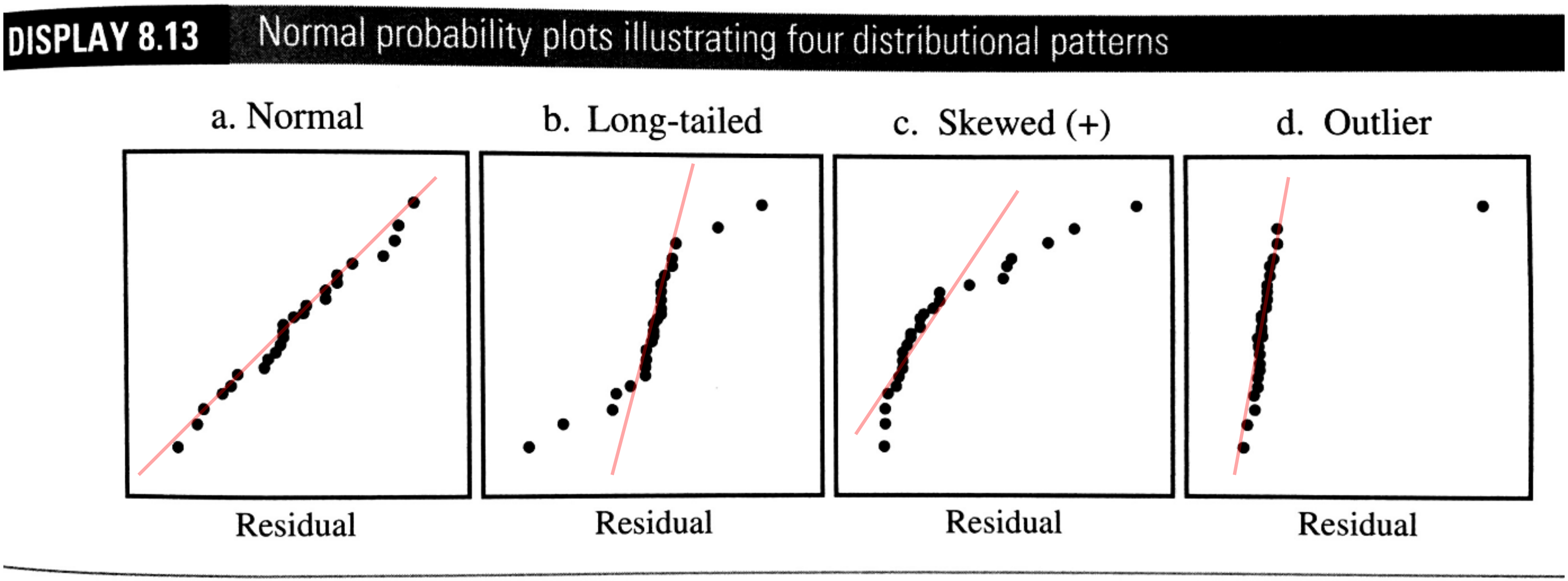
- Normality of errors, $\varepsilon_i \sim N(0, \sigma^2)$.
  - Checking: QQ-plot.
  - If violated:
    - Doesn't lead to bias in $\hat{\beta}_0, \hat{\beta}_1,$ and, due to CLT, standard errors are not affected much (unless residuals are *long-tailed* and there are *outliers*).
    - Prediction is more sensitive, because it is based on the normality of *population distribution* of Y given X.
  - Strategies:
    - Ignore;
    - Alternatively, use regression with *t*-distribution assumption on errors (a form of robust regression).

Normal Probability Plots (QQ-plot)



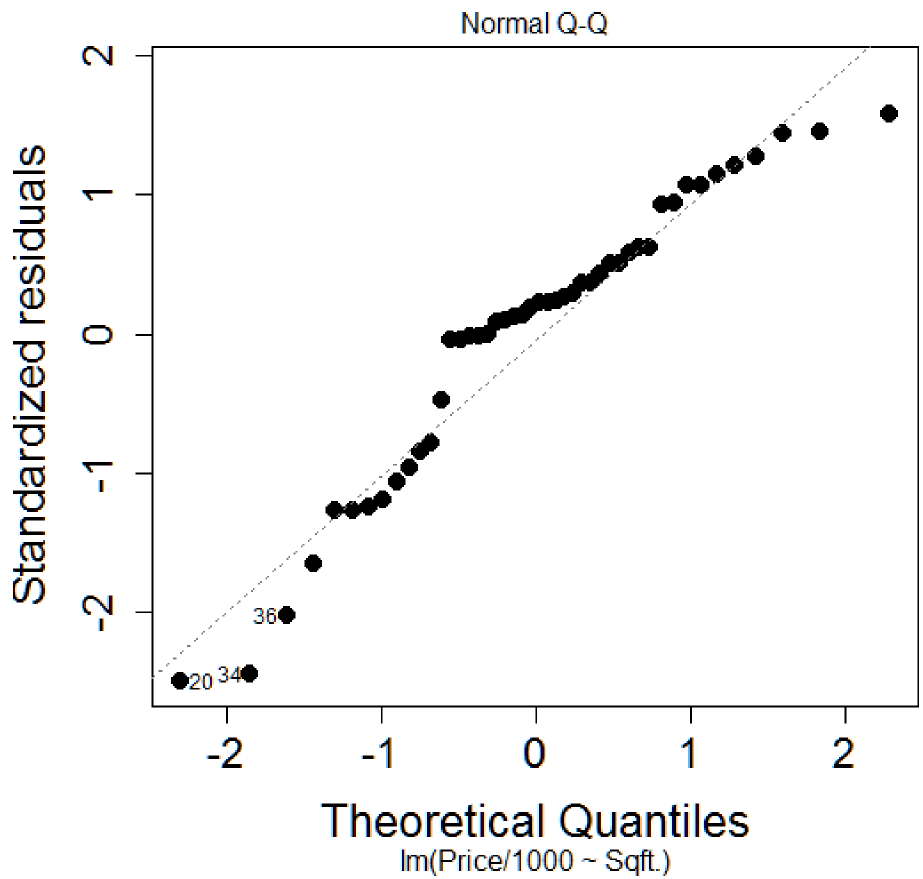
Long-tailed distribution

Normal Probability Plots (QQ-plot)



Newton Data: QQ-plot in R

```
> plot(regmodel, which = 2, pch=19)
```



Interpretation of Results After Log Transformation

Example: Does it pay to advertise?

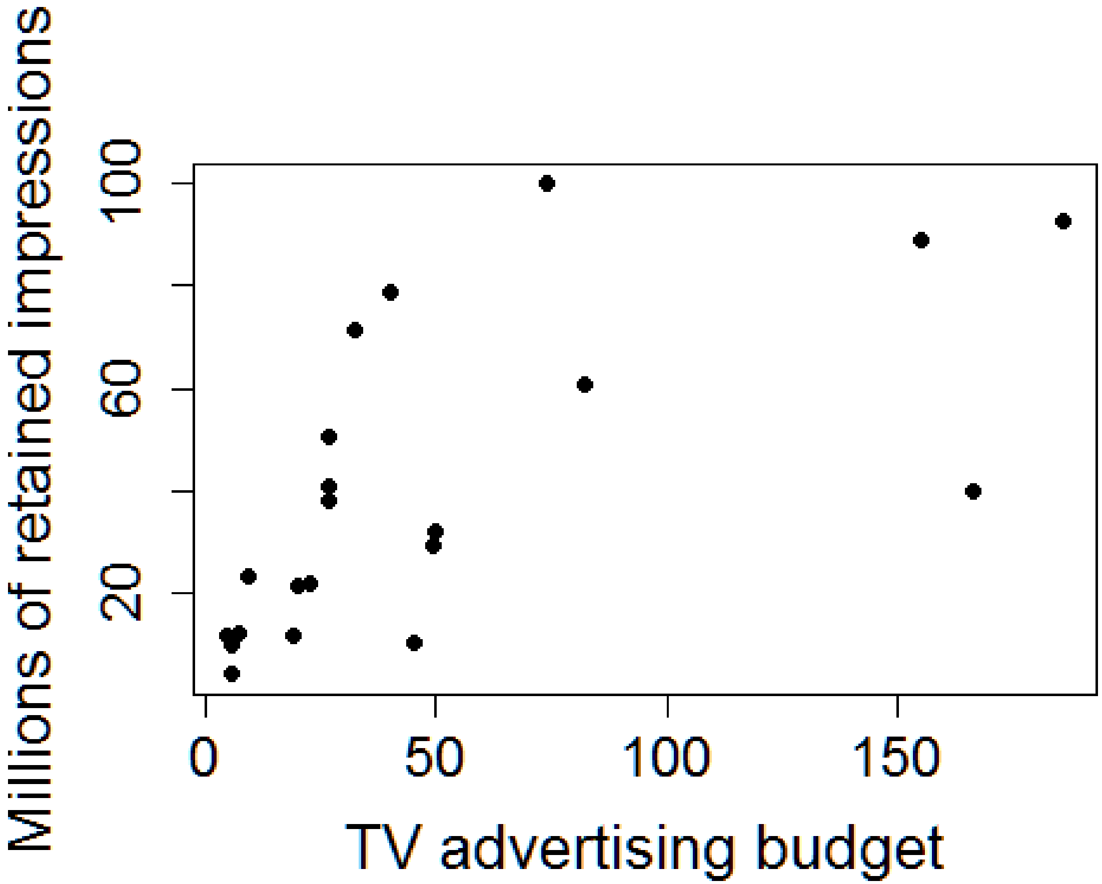
Think about all the commercials during the Super Bowl or the World Series. Does it pay to advertise?

<u>Data:</u> Advertising budget of 21 firm and millions of *impression* retained per week by the users of the products of certain firms. The data are based on a survey of 4,000 adults (Wall Street Journal, 1984).

IMPACT OF ADVERTISING EXPENDITURE

Firm	Impressions, millions	Expenditure, millions of 1983 dollars
1. Miller Lite	32.1	50.1
2. Pepsi	99.6	74.1
3. Stroh's	11.7	19.3
4. Fed'l Express	21.9	22.9
5. Burger King	60.8	82.4

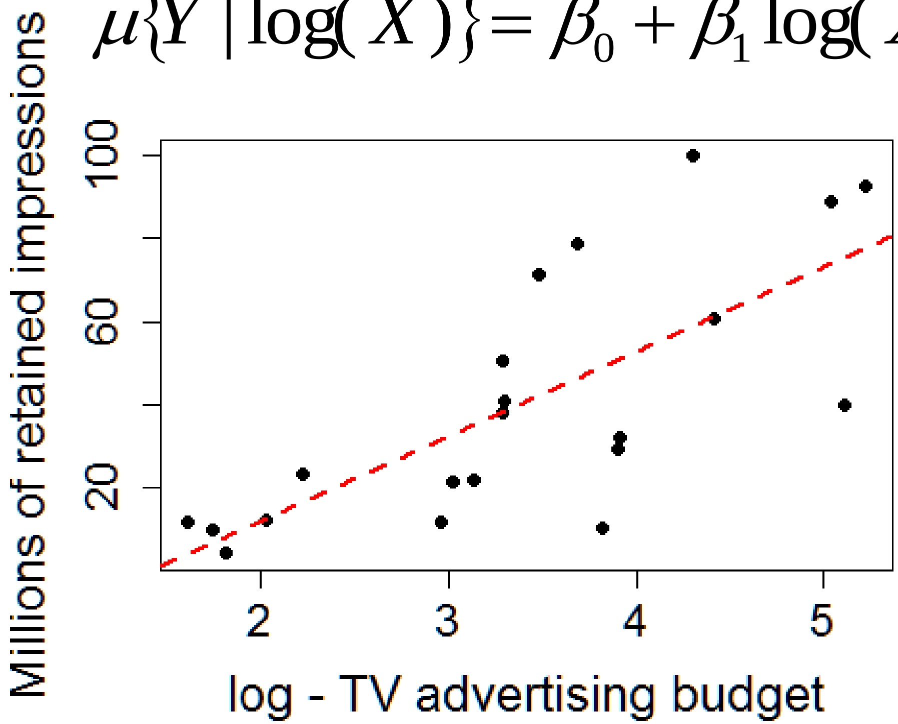
TV Ad Yields (Street Journal, 1984)



What transformation should we use, if any?

Explanatory Variable is Logged

$$\mu\{Y \mid \log(X)\} = \beta_0 + \beta_1 \log(X)$$



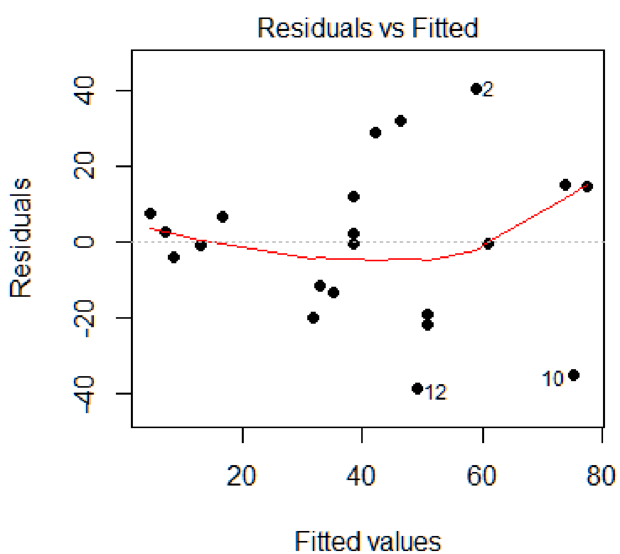
$$\mu\{Y \mid \log(2X)\} - \mu\{Y \mid \log(X)\}$$
$$= \beta_1 \log(2)$$

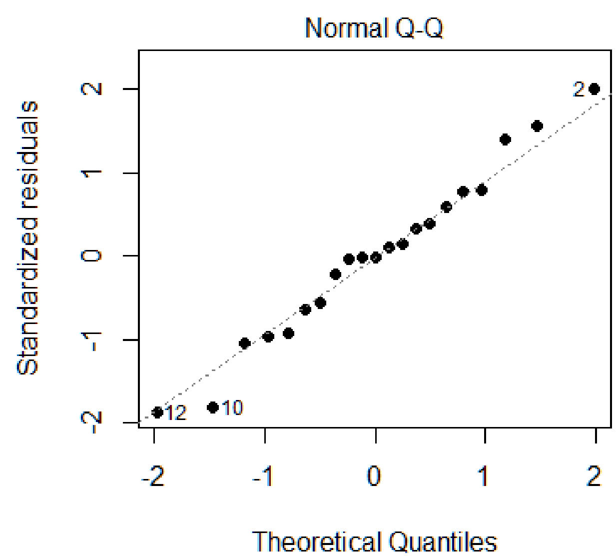
Doubling of X is associated with a $\beta_1 \log(2)$ change in the mean of Y.

$$\hat{\mu}\{\text{Imp} \mid \text{Budget}\} = -28 + 20 \cdot \log(\text{Budget})$$

Explanatory Variable is Logged

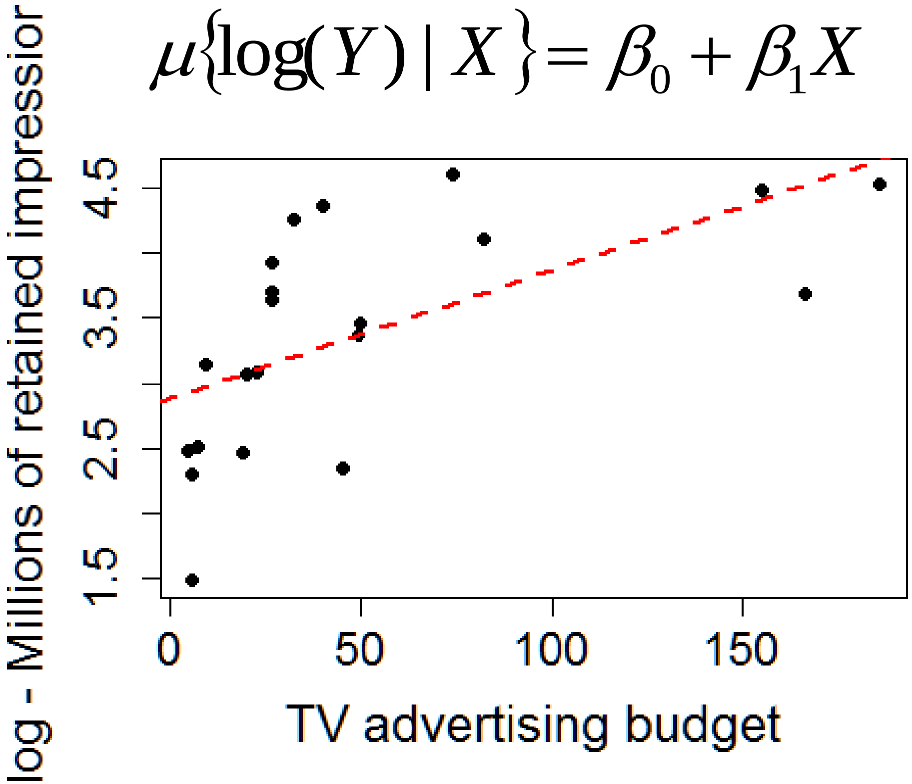
```
lm(formula = MILIMP ~ log(SPEND), data = SaleData)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -28.050     15.441  -1.817 0.085093 .
log(SPEND)    20.180      4.339   4.650 0.000174 ***
```





Response Variable is Logged

$$\mu\{\log(Y) \mid X\} = \beta_0 + \beta_1 X$$



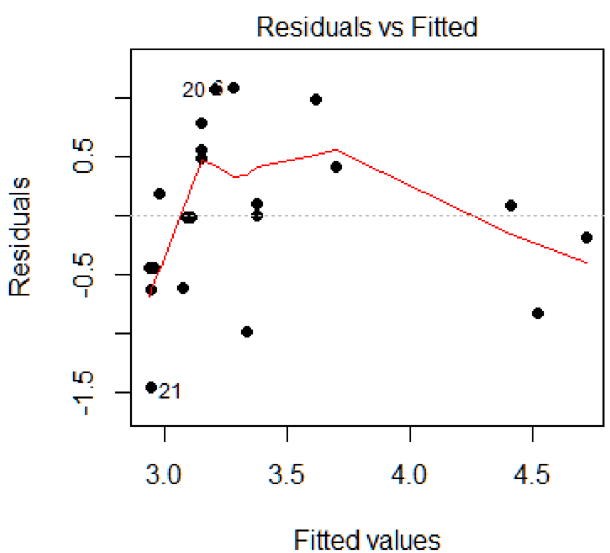
$$\frac{\text{Median}(Y \mid X+1)}{\text{Median}(Y \mid X)} = \exp(\beta_1)$$

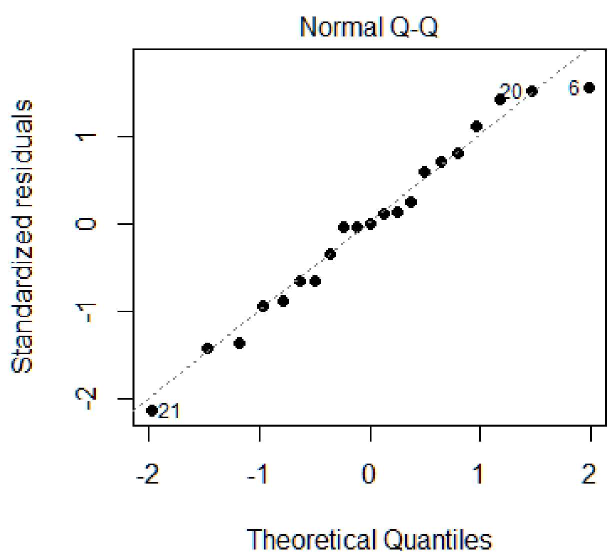
Increase in X of 1 unit is associated with a multiplicative change in Median$\{Y \mid X\}$ *by* exp(β_1).

$$\hat{\mu}\{\log(\text{Imp}) \mid \text{Budget}\} = 2.9 + 0.01 \cdot \text{Budget}$$

Explanatory Variable is Logged

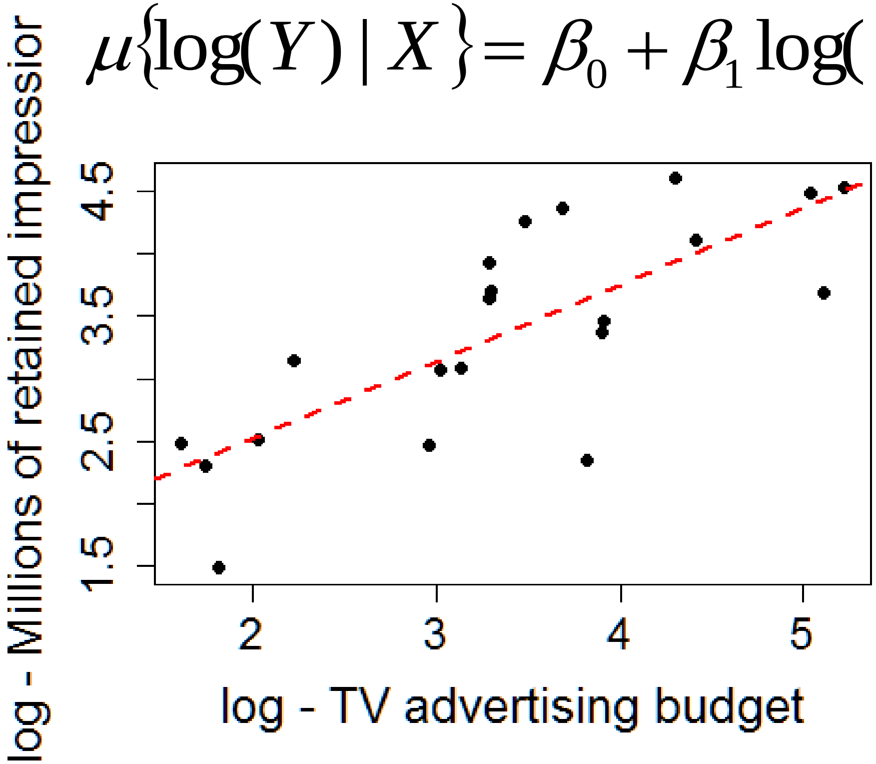
```
lm(formula = log(MILIMP) ~ (SPEND), data = SaleData)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.88592    0.21534  13.401 3.93e-11 ***
SPEND        0.00986    0.00295   3.342  0.00342 **
```





Response and Explanatory Variables are Logged

$$\mu\{\log(Y) \mid X\} = \beta_0 + \beta_1 \log(X)$$



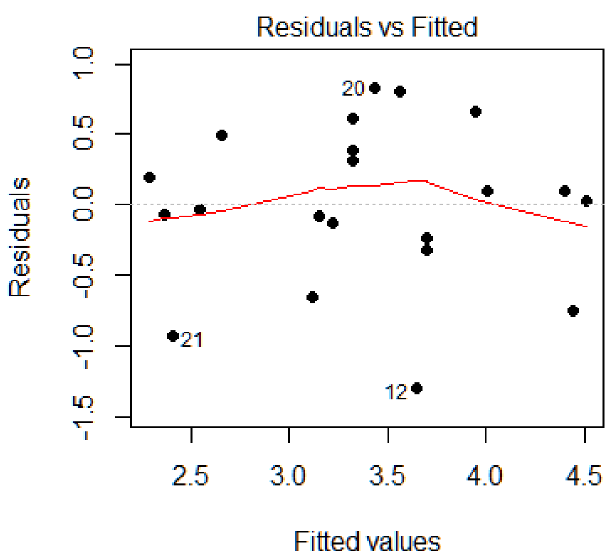
$$\frac{\text{Median}(Y \mid 2X)}{\text{Median}(Y \mid X)} = 2^{\beta_1}$$

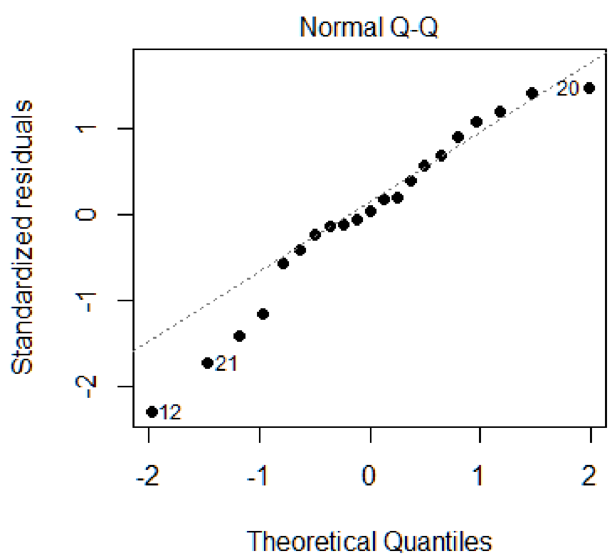
Doubling of X is associated with a multiplicative change in Median$\{Y \mid X\}$ *by* 2^{β_1}*.*

$$\hat{\mu}\{\log(\text{Imp}) \mid \text{Budget}\} = 1.3 + 0.61 \cdot \log(\text{Budget})$$

Explanatory Variable is Logged

```
lm(formula = log(MILIMP) ~ (SPEND), data = SaleData)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2999     0.4236   3.069  0.00632 **
log(SPEND)    0.6135     0.1191   5.153 5.66e-05 ***
```





Transformations and Interpretation

- Other transformations: $\mathrm{sqrt}(Y)$, $1/Y$, etc. – may be more difficult to interpret.
  - If the goal is prediction or confirmation of association, there is, usually, no need to interpret coefficients.

- Randomization of levels of X to study units allows us to make inference about an "effect" of X on Y.

- Without randomization, we can only infer about
  - association between X and Y;
  - X as predictors or risk-factors for Y.

Summary: Exploring Relationship Between X and Y Linear Regression (LR)

1. Identify if units may be considered independent.
   A. Otherwise, model their dependence or redefine units.

2. Review scatterplot for
   A. Non-linearity,
   B. Non-constant spread,
   C. Outliers.

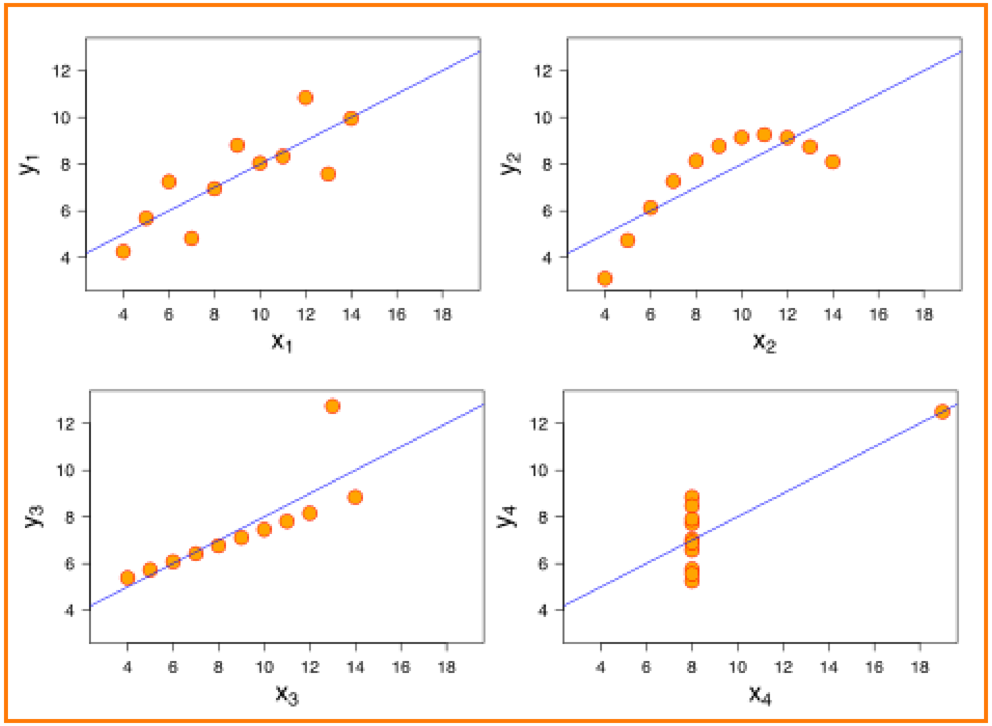
3. Apply transformations to X and/or Y, if necessary.

Summary: Exploring Relationship Between X and Y Linear Regression (LR)

4. (Tentatively) fit the line and examine residual plot for:
   A. Outliers,
   B. Non-constant spread,
   C. Non-normality (especially, if <u>prediction</u> is of interest).
5. If any of the assumptions are violated
   A. try a different transformation;
   B. include more explanatory variables (Multiple LR);
   C. use an alternative version of LR with relaxed assumptions (weighted LR, robust/resistant LR, etc.).
6. Interpret the results.

Anscombe’s Quartet

Francis Anscombe, “Graphs in Statistical Analysis". *American Statistician*, 1973



Read more about Anscombe’s Quartet here.

Anscombe's Quartet

Identical in common summary statistics: mean, variance, (Pearson) correlation, estimated regression line.

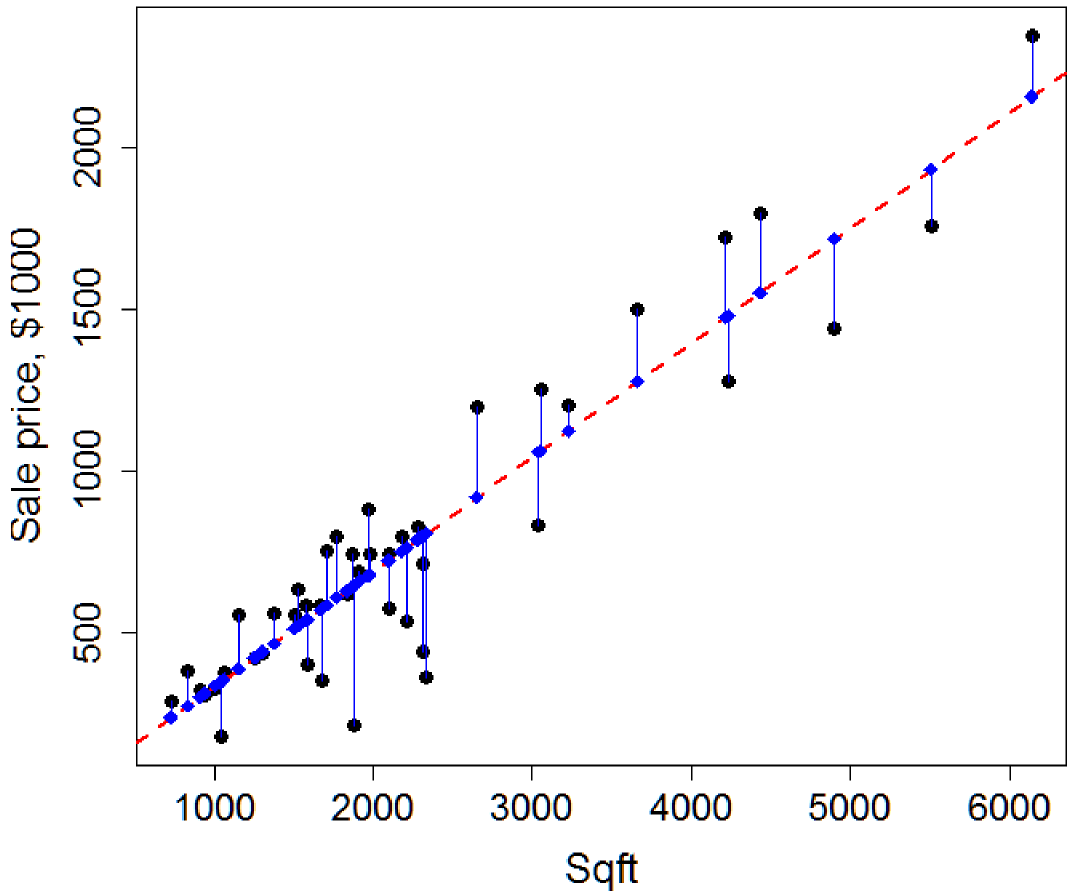
Property	Value
Mean of *x* in each case	9.0
Variance of *x* in each case	11.0
Mean of *y* in each case	7.5
Variance of *y* in each case	4.12
Correlation between *x* and *y* in each case	0.816
Linear regression line in each case	$y = 3 + 0.5x$

Lessons?

Sum of Squares Decomposition for Linear Regression & R-squared Statistic

Residual Sum of Squares for the Regression Model (Full): SSRes

$$E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$$



$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$\text{SSRes} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

$$= \hat{\sigma}^2 (n-2),$$

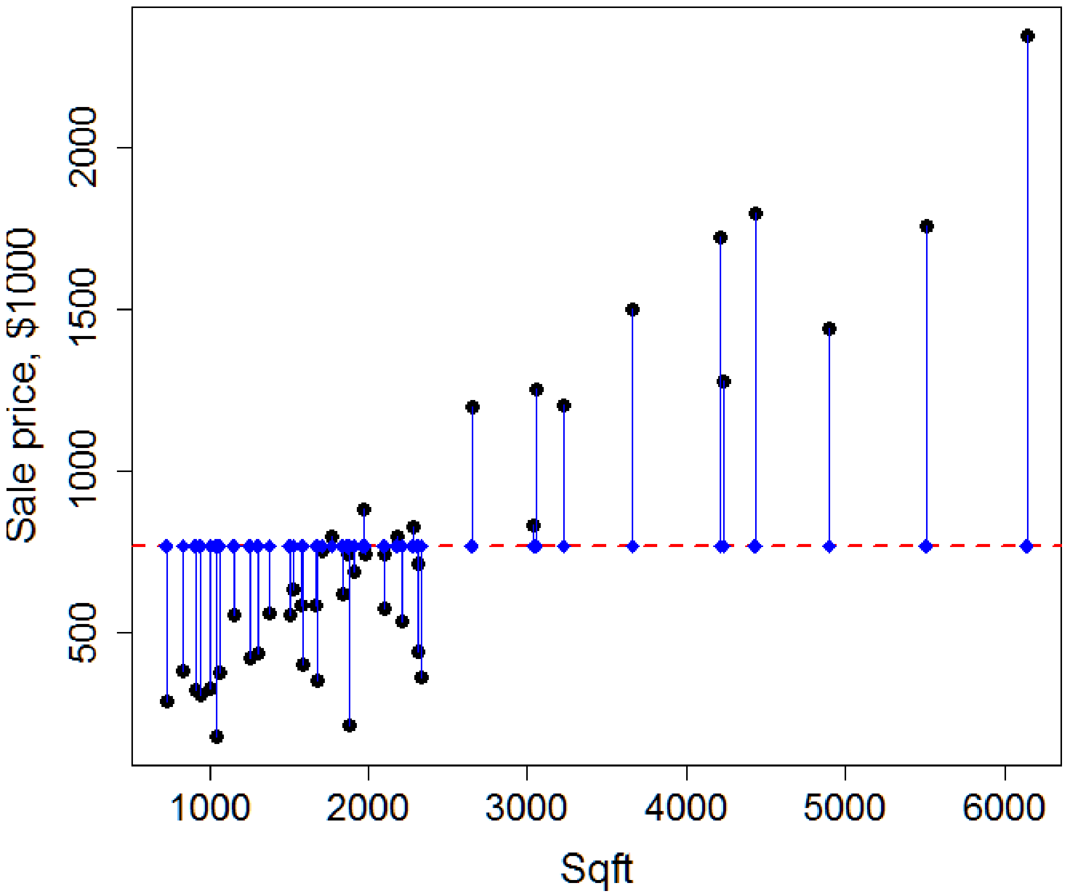
where $\hat{\sigma}^2$ is the estimated residual variance for the regression model.

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-2} \chi^2_{n-2} \Rightarrow$$

$$\Rightarrow \text{SSRes} \sim \sigma^2 \chi^2_{n-2}$$

Residual Sum of Squares for the Equal-Means (Reduced) Model: $SSR_{reduced}$

$$E(Y_i \mid X_i) = \mu$$



$$\hat{Y}_i = \bar{Y}$$

$$SSR_{reduced} = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

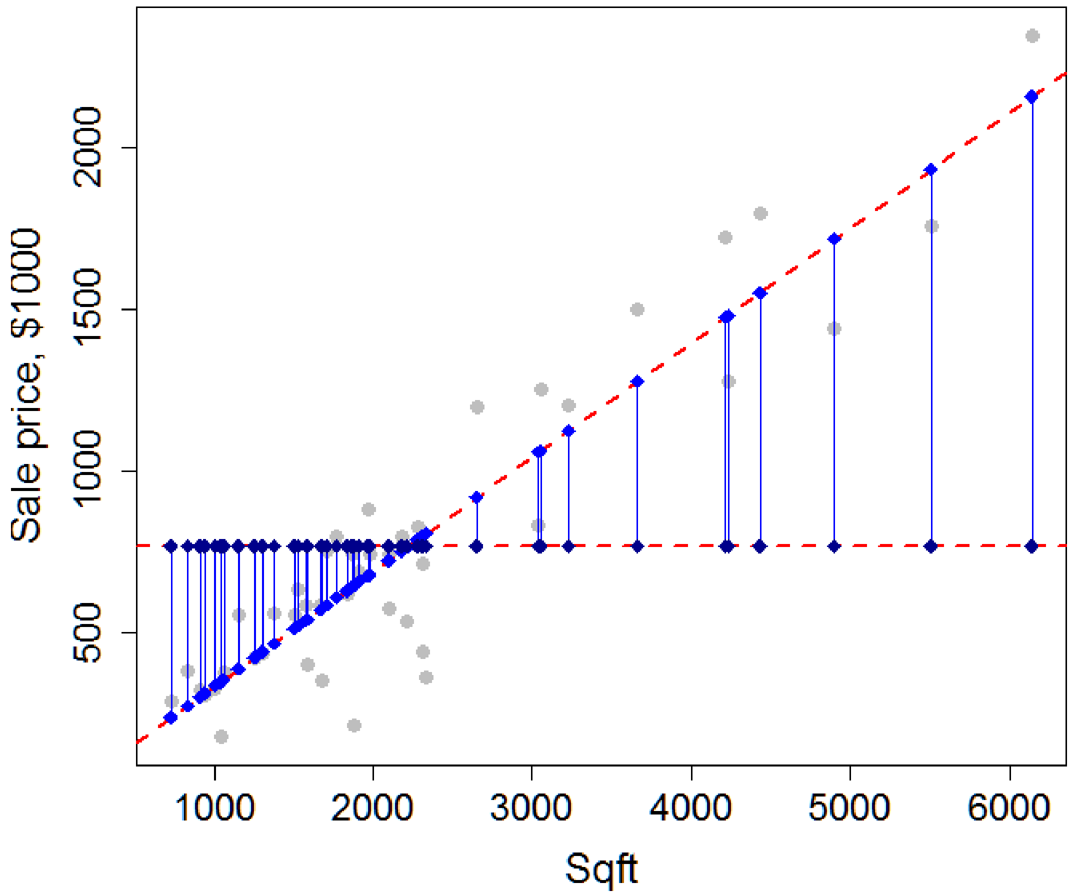
$$= S^2(n-1)$$

where S^2 is the sample variance of the entire sample taken as one group.

$$S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1} \Rightarrow$$

$$\Rightarrow SSR_{reduced} \sim \sigma^2 \chi^2_{n-1}$$

Sum of Squares *Between* Predicted Means and the Overall Mean



$$\text{SSReg} = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$$

Let

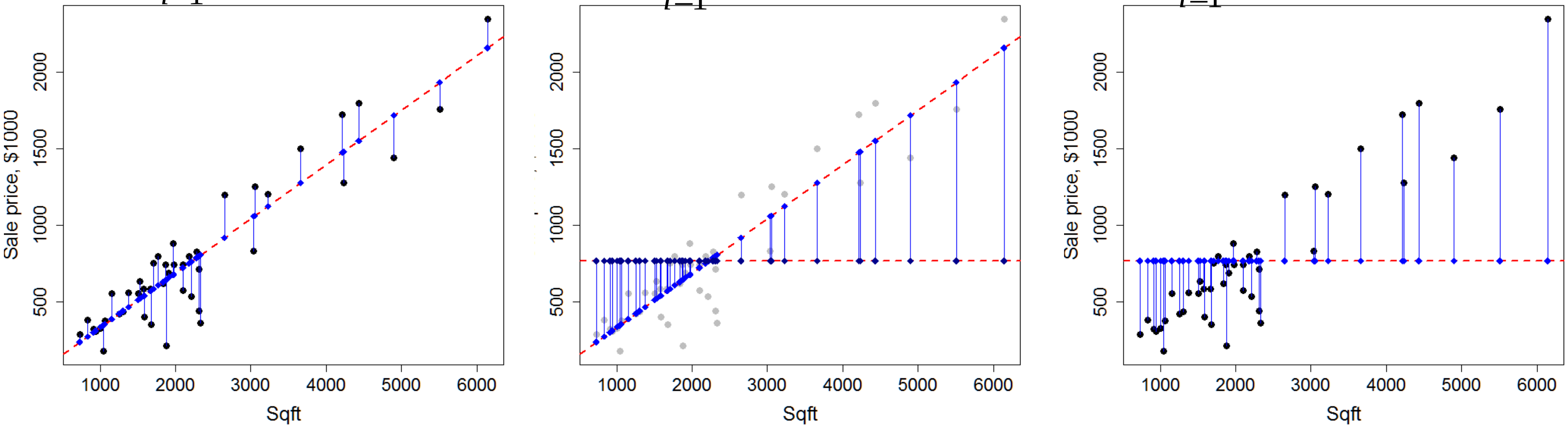
$$H_0 : E(Y_i \mid X_i) = \mu = \beta_0,$$

$$H_a : E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i.$$

Then $\text{SSReg} \overset{H_0}{\sim} \sigma^2 \chi_1^2.$

Sum of Squares Decomposition for the Linear Regression Model

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \qquad \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 \qquad \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$$



SSRes SSReg $\text{SSR}_{\text{reduced}}$

Analogously to the SS decomposition for *I* means for the separate-means model, it can be shown that

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2, \text{ where } \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

R-Squared: The Proportion of Variation Explained

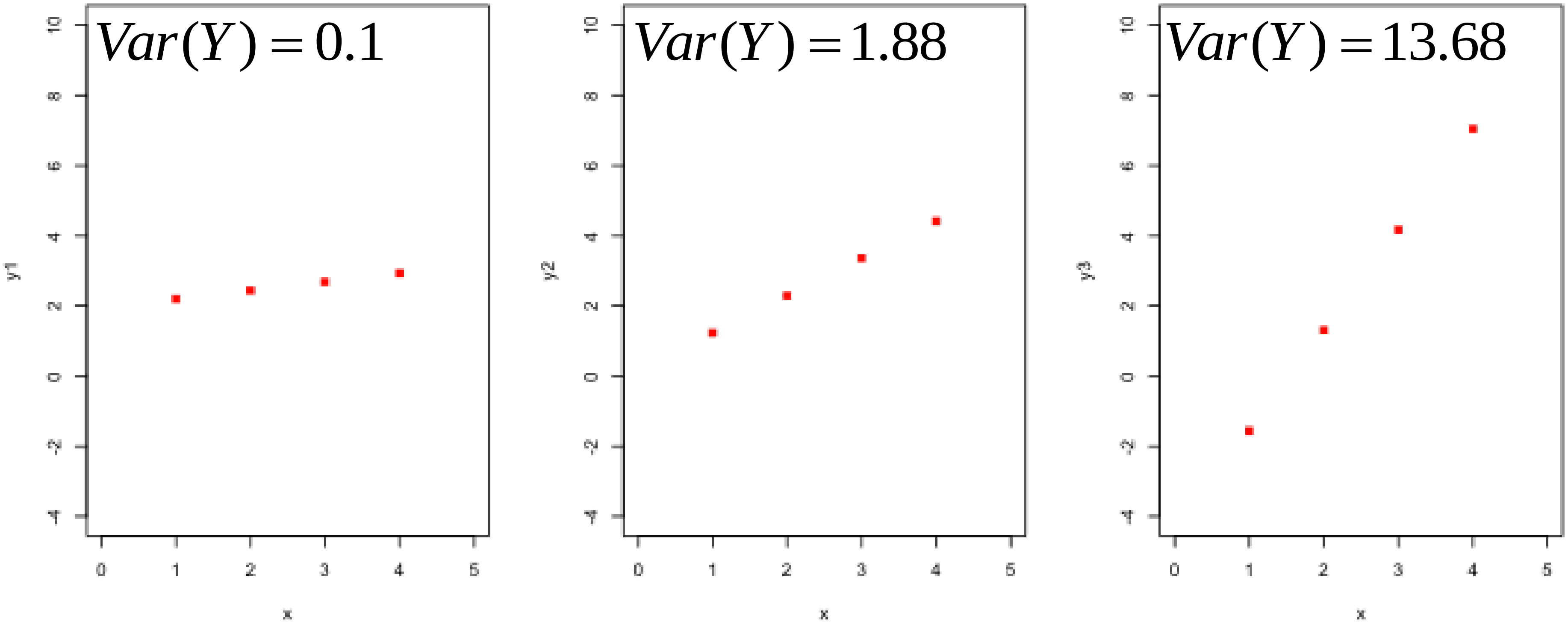
$SSR_{\text{reduced}} = SSRes + SSReg$

R-squared statistic, or a coefficient of determination,

$$R^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 - \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = \frac{SSR_{\text{reduced}} - SSRes}{SSR_{\text{reduced}}} = \frac{SSReg}{SSR_{\text{reduced}}}$$

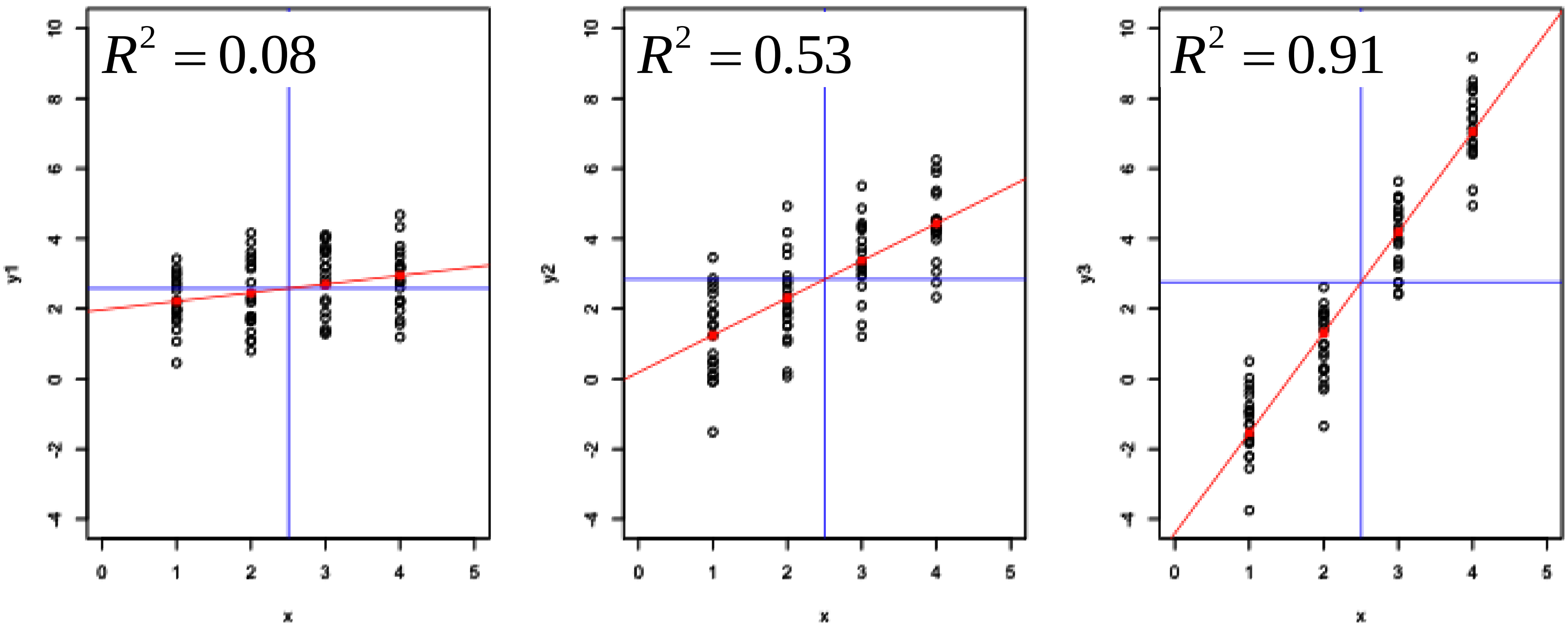
- R^2 is the proportion of variation in the response, Y, explained by the model for the means.
- If the relationship is linear and all other regression assumptions are met, then high R^2 means that X explains Y well.

R-Squared: The Proportion of Variation Explained



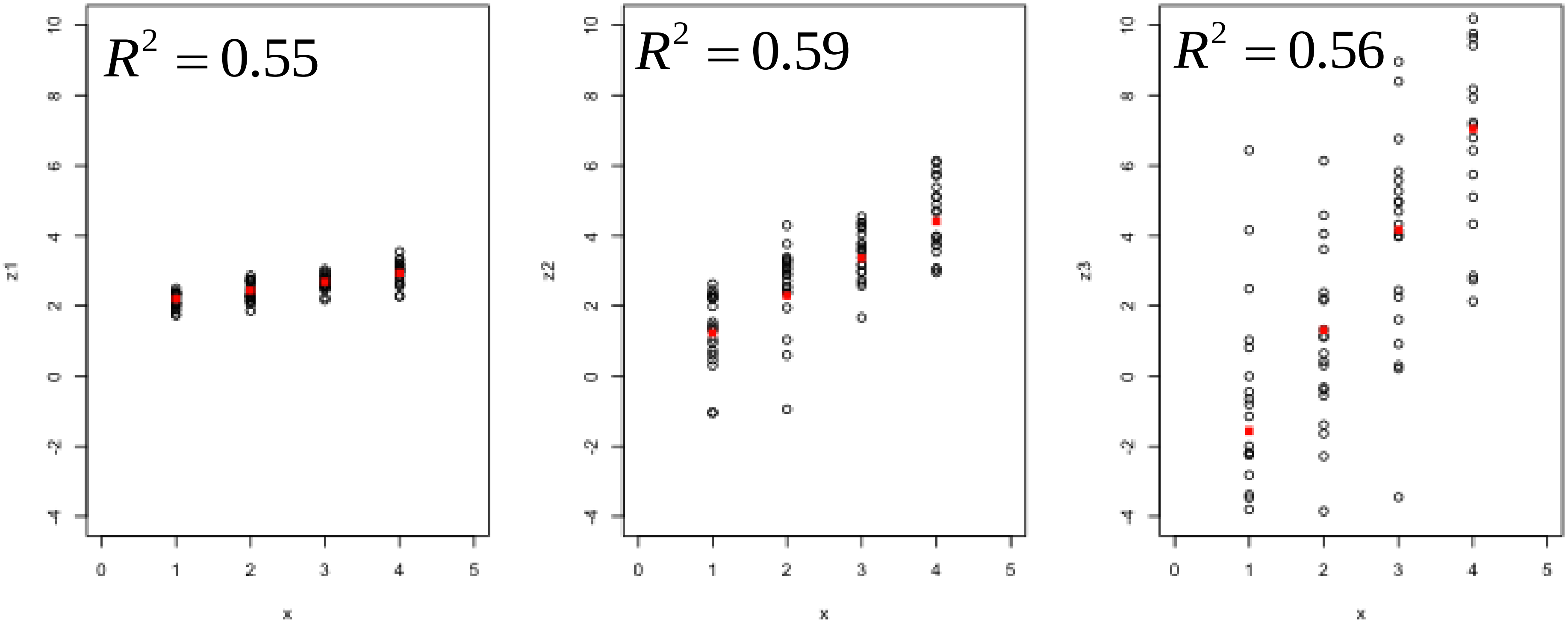
R-Squared: The Proportion of Variation Explained

$$r_{xy} = 0.29 \qquad r_{xy} = 0.72 \qquad r_{xy} = 0.95$$



For Simple Linear Regression: $\boxed{R^2 = r_{xy}^2}$

R-Squared: The Proportion of Variation Explained



R-Squared: The Proportion of Variation Explained

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$$

When does $R^2 = 0$?

When does $R^2 = 1$?

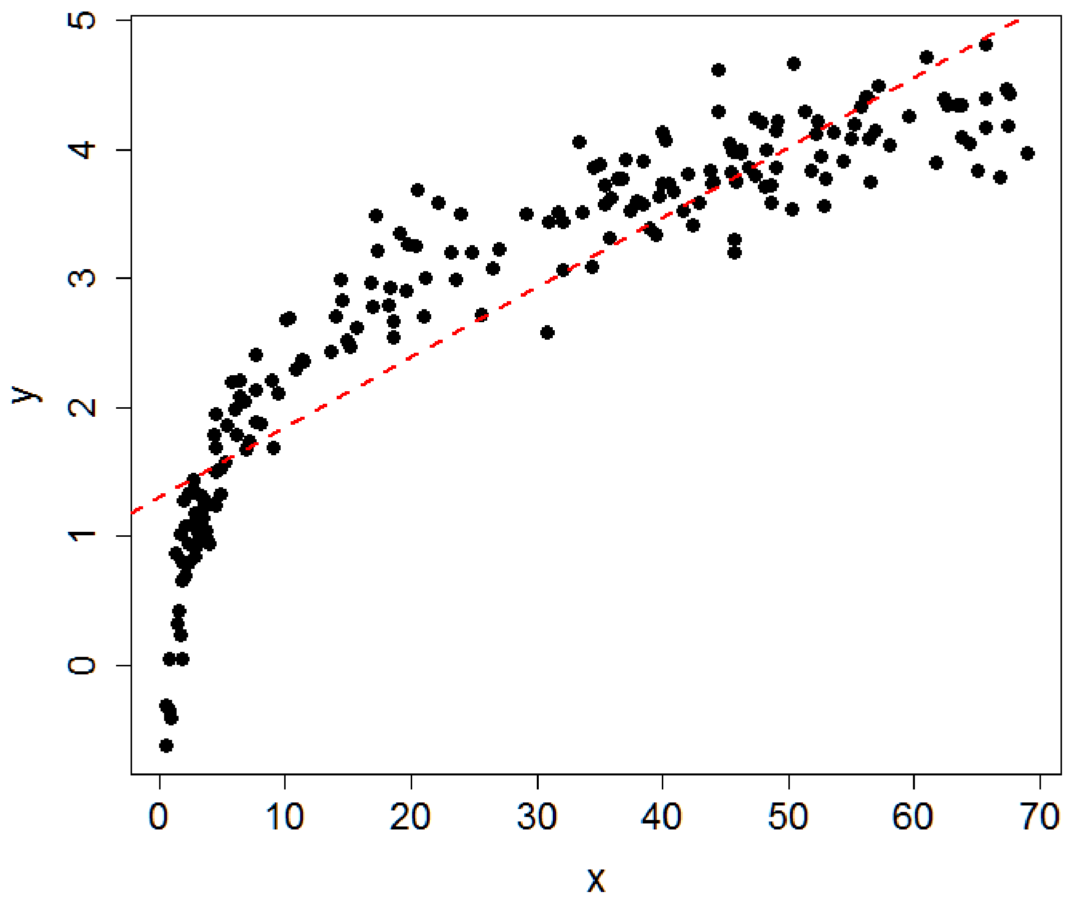
Caution! R^2 may be quite large even when the simple linear regression is inadequate – never use it to assess the adequacy of the straight line model.

- Decreases if we take a subset of a range of *X*.
- Increases as predictors are added to the model.

Later we will learn about *adjusted R-squared*.

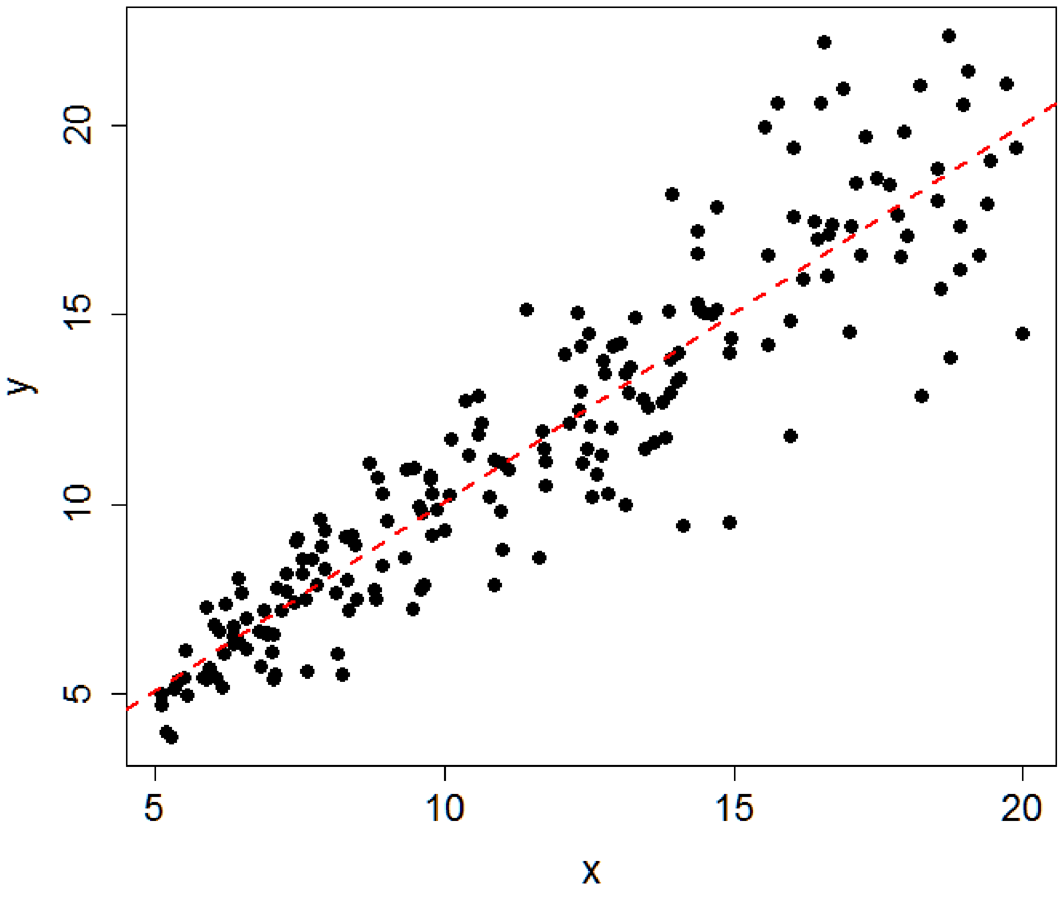
R-Squared: Does Not Indicate the Adequacy of a Straight Line Model

Multiple R-squared: 0.8118



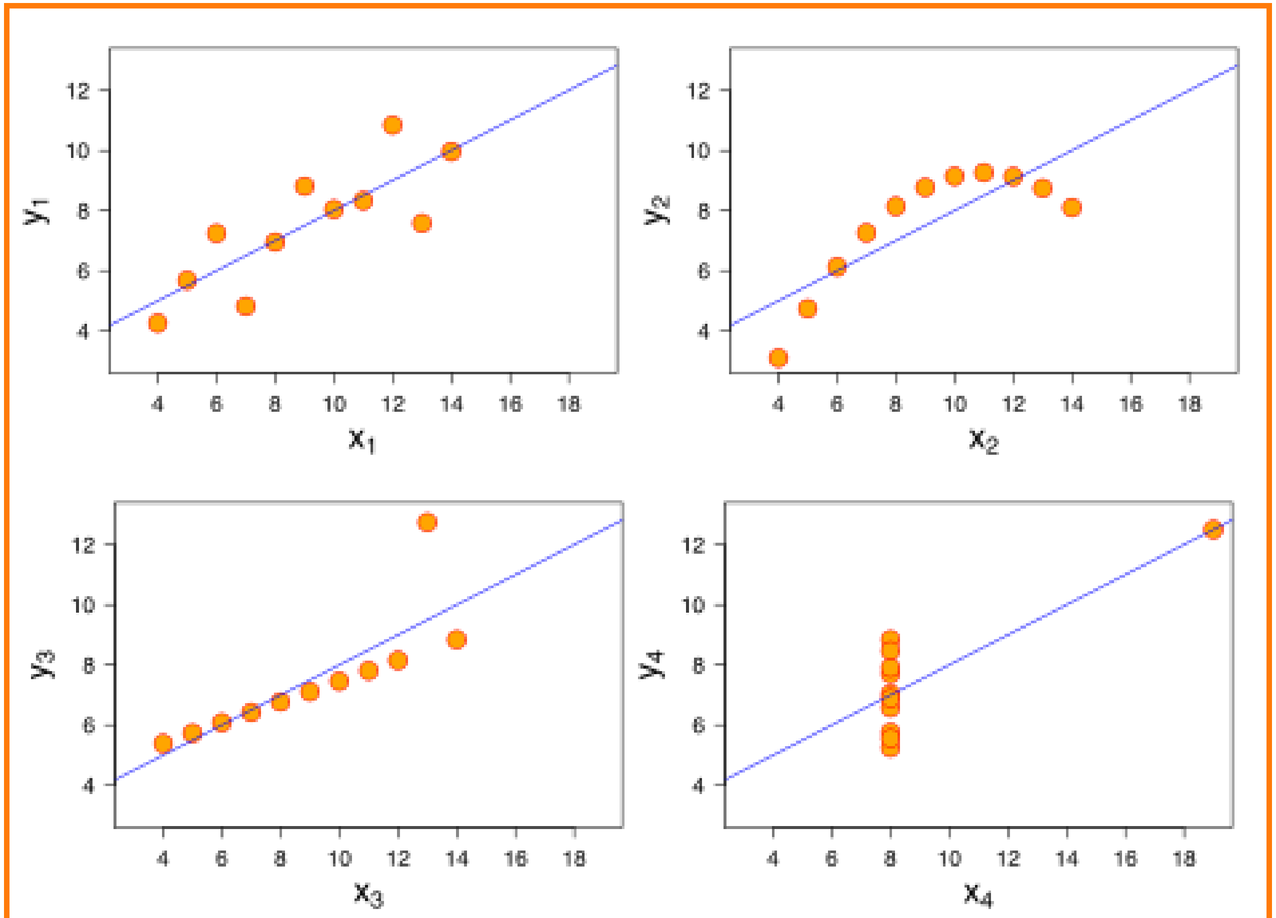
R-Squared: Does Not Indicate the Adequacy of a Straight Line Model

Multiple R-squared: 0.8455



R-Squared: Does Not Indicate the Adequacy of a Straight Line Model

Anscombe's Quartet: R-squared = $r^2 = 0.816^2 = 0.67$



R-Squared: Does Not Indicate the Adequacy of a Straight Line Model

Multiple R-squared: 0.13