WEEK 04 PART 02

INSTRUCTOR: YANAN WU

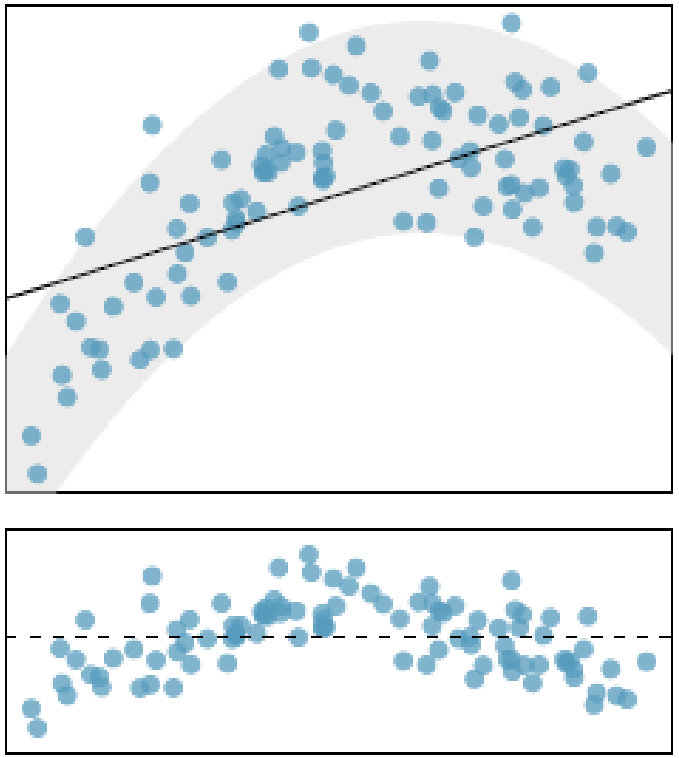
TA: KHADIJA NISAR

SPRING 2025

4.1
KEY ASSUMPTION

1. Linearity: The relationship between the independent variable and dependent variable is **linear**, if there is a nonlinear trend, an advanced regression method should be applied

Non-linearity of the Data

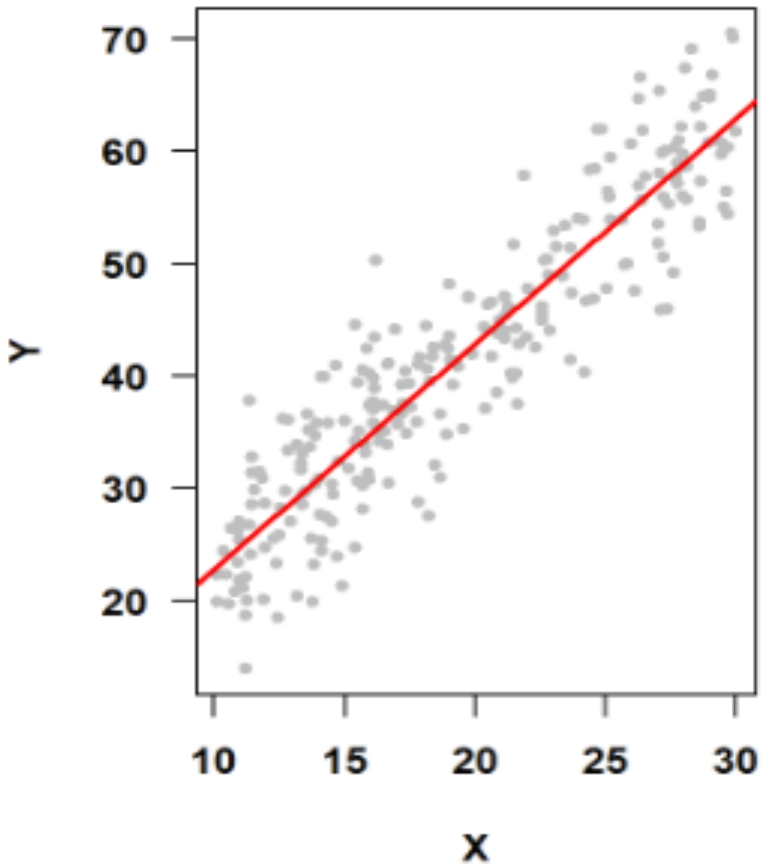


Liner regression line

Residual plot is a useful graphical tool for identifying non-linearity.

2. The error at any level of x_i share an identical distribution, with constant variance

constant variance of Error Terms



non-constant variance of Error Terms

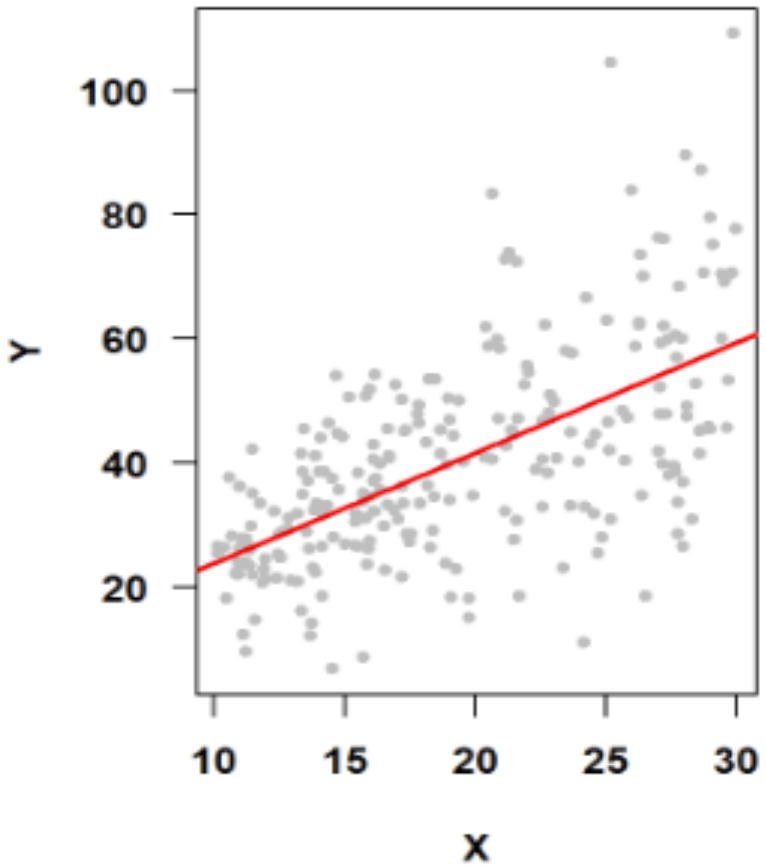
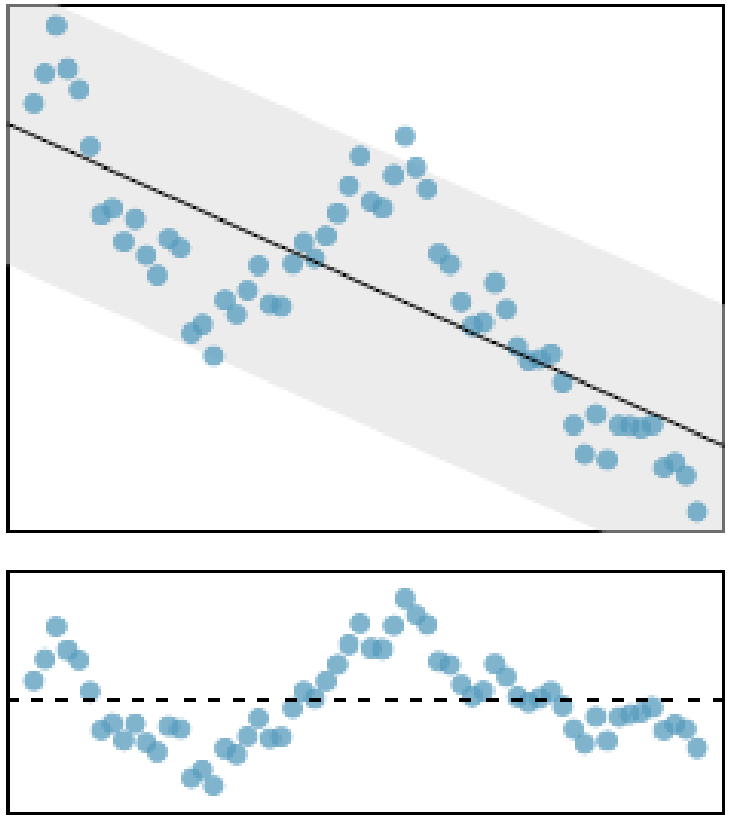


IMAGE SOURCE: HTTPS://WWW.BOOKDOWN.ORG/RWNAHHAS/RMPH/MLR-CONSTANT-VARIANCE.HTML

3. Error are assumed to be independent (uncorrelated) among each other

Example of correlated Error



4. i.i.d Normality of Error

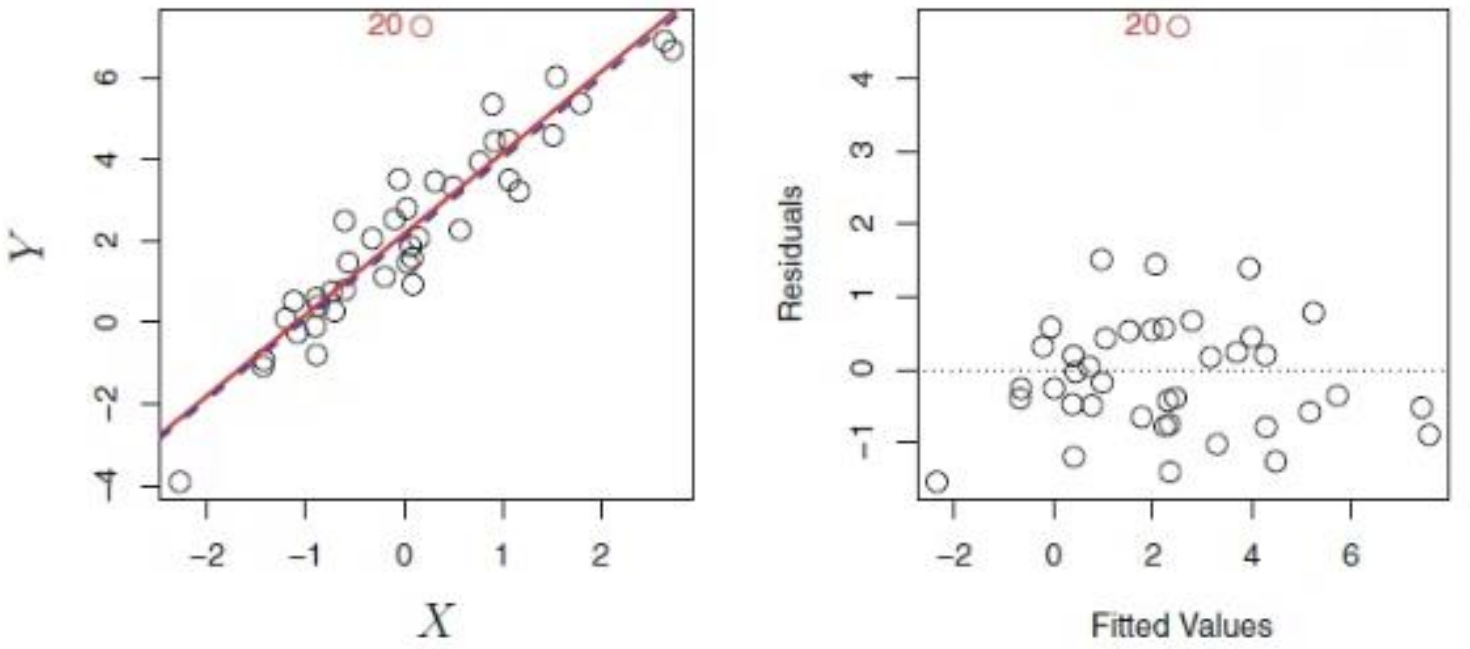
This assumption states that the **disturbances (errors) in a regression model are**:

1) Independently and identically distributed (i.i.d)

2) Normally distributed (i.e.,$\varepsilon_i \sim N(0, \sigma^2)$

This assumption is **important** because it allows for **valid hypothesis testing and confidence intervals**, even when the sample size is **very small**.

4. OUTLIER



If we drop outlier, 20, the RSE decrease from 1.09 to 0.77.
R^2 increase from 0.805 to 0.892.

If we believe the outlier is due to an error in data collection, we can simply remove the observation.

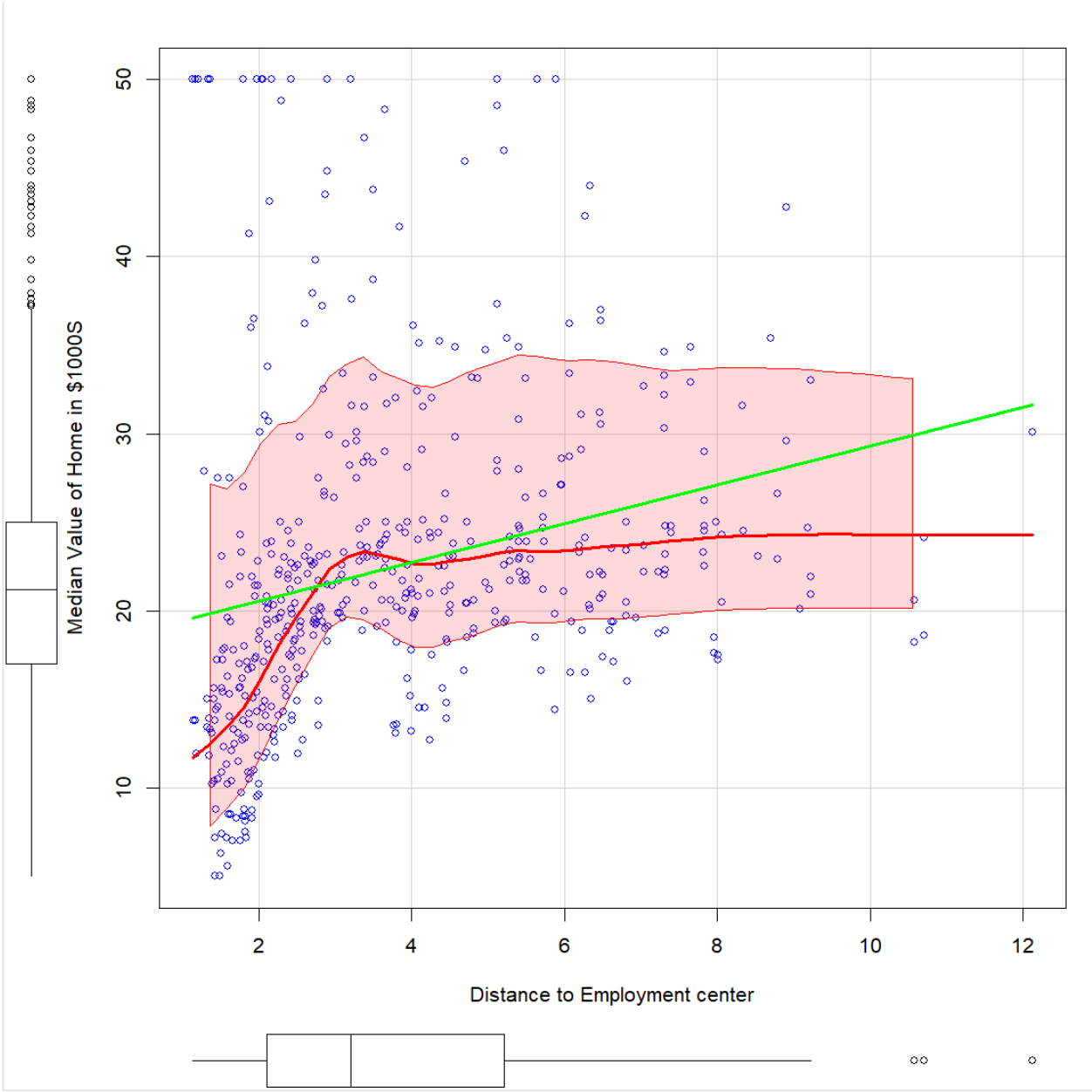
However, care should be taken ,since an outlier may indicate a deficiency in the model, such as missing x variables

Transformation of dependent and independent variable

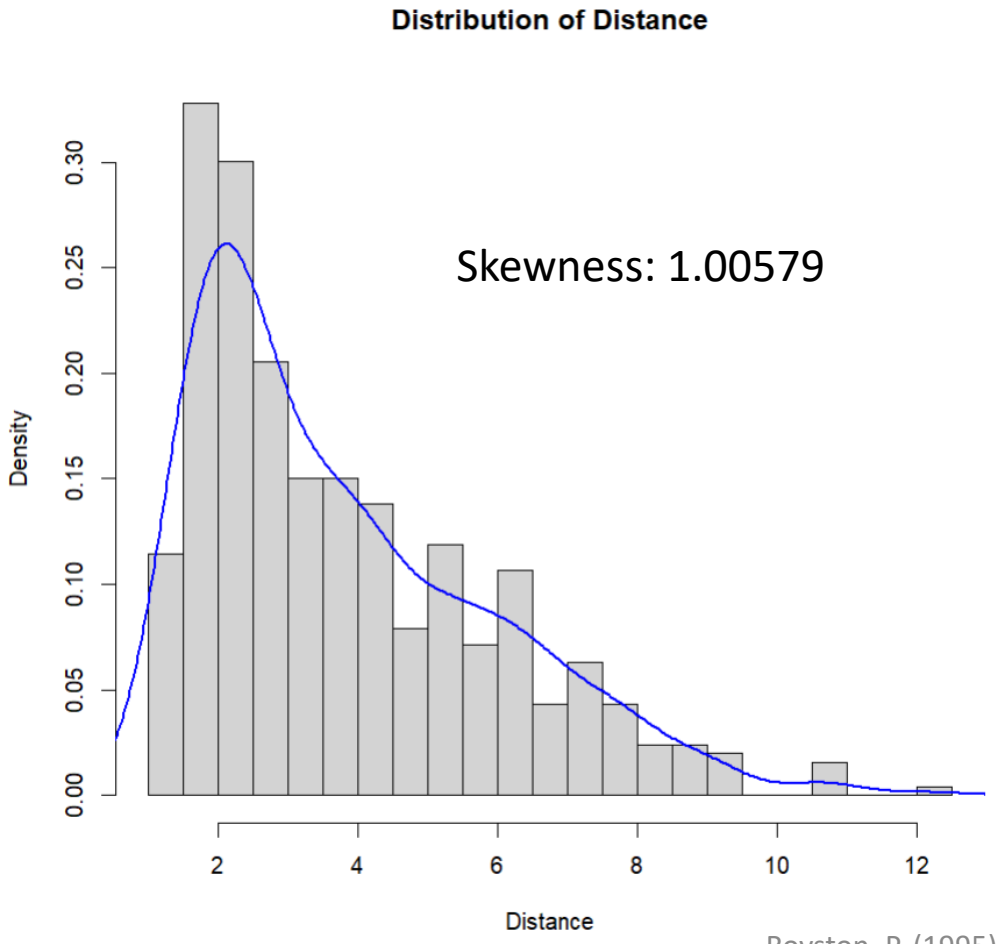
The transformation is needed for the dependent and independent variable is able to:

1) Fix a curvilinear relationship and making relationship linear
2) Make the distribution of error to a normal distribution or close to the normal distribution
3) To stabilize the variability of the regression residuals by a transformation of the dependent variable

Assess the univariate distribution in x_i and y_i variables: scatterplot



Shapiro Test: Assess the normality in univariate distribution



Skewness: 1.00579

Shapiro and Wilk's (1965) W-statistic is well-established and powerful test of departure from normality

H_0: The sample comes from a normally distributed population.

```
> shapiro.test(Boston$dis)

        Shapiro-Wilk normality test

data:  Boston$dis
W = 0.90323, p-value < 2.2e-16
```

Royston, P. (1995). Remark AS R94: A remark on algorithm AS 181: The W-test for normality. Journal of the Royal Statistical Society. Series C (Applied Statistics), 44(4), 547-551.

Find a transformation for the independent and dependent variable

- Scaled Power Transformation (This only works for positive variable ($Z > 0$))

$$\mathrm{T}(\mathrm{Z}, \lambda) = \begin{cases} \dfrac{Z^{\lambda} - 1}{\lambda}, & where\ \lambda \neq 0 \\ \log(Z), & where\ \lambda = 0 \end{cases}$$

- Z can be either an independent variable (X_i) or a dependent variable (Y_i)

- $Y_i^* = T(Y_i, \lambda)$ and $X_i^* = T(X_i\ \lambda)$

- For $\lambda \neq 0$, the scaled power transformations are essentially x^{λ}, because the scaled-power family only subtracts 1 and divides by the constant λ.

- The λ parameter of the transformation can be estimated by the Box-Cox power family

Box-Cox Transformation

The Box-Cox can provide answers to the following questions:

1) Is there a transformation needed to normalize the data?

H_0: There is no transformation needed

H_1: There is a transformation needed

2) What is the optimal value of the transformation parameter?

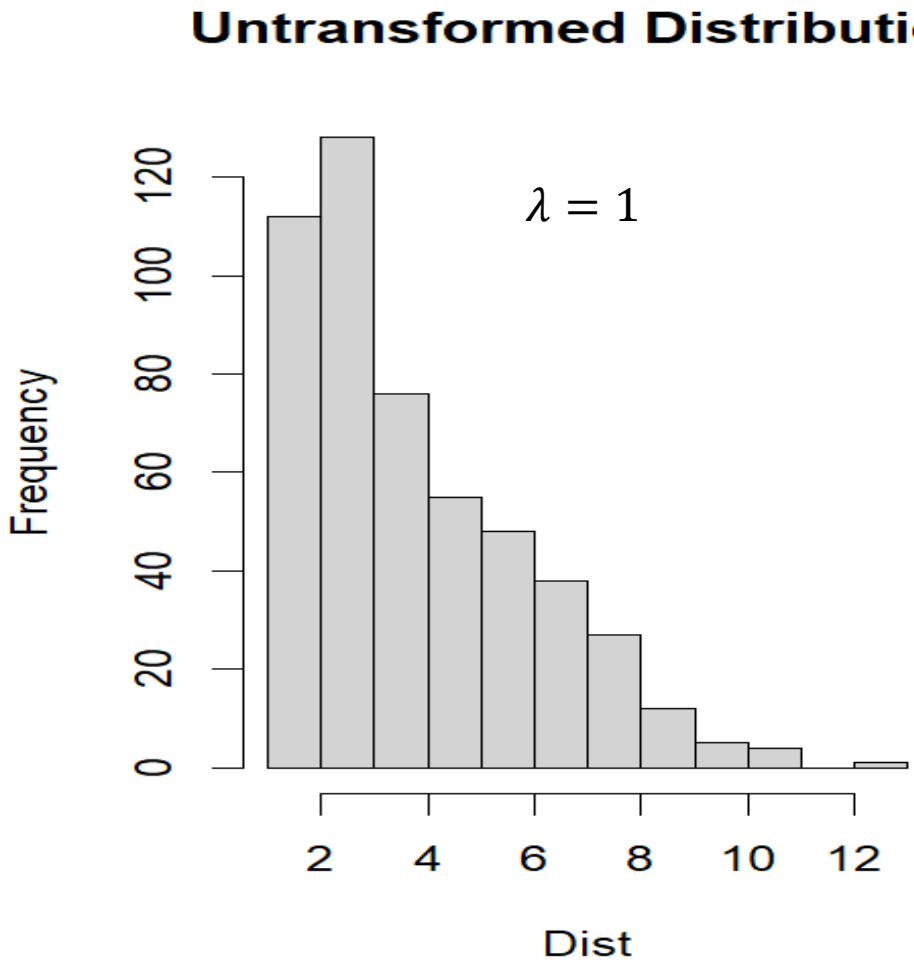
$H_0: \lambda = 0$

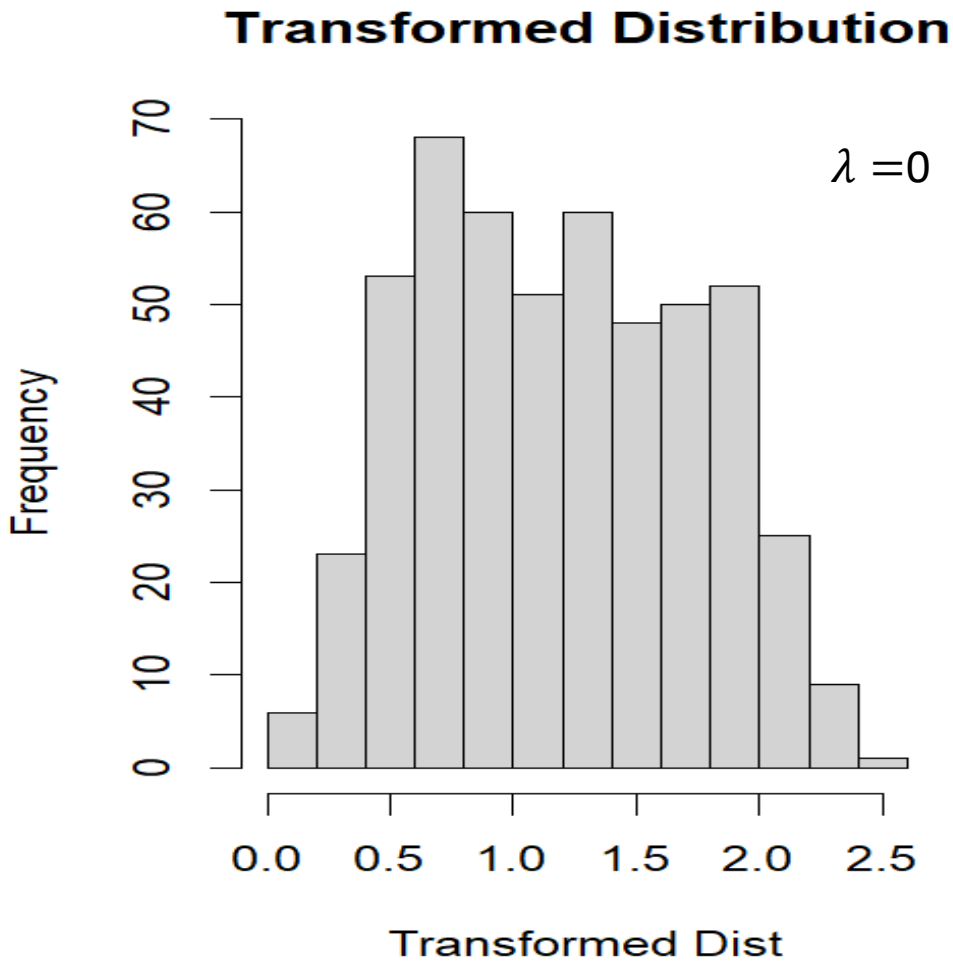
$H_1: \lambda \neq 0$

4. Procedure for Data Transformation in Regression Analysis

1) Apply a suitable transformation to independent variable (X_i^*) to improve the linearity of the model
   - $X_i^* = Z(X_i, \lambda)$
2) Identify an appropriate transformation function for the dependent variable ($Y_i^* = \mathrm{Z}(\mathrm{Y_i}, \lambda)$), so that the regression residuals approximate a normal distribution:
   - $e_i^* = Y_i^* - b_0^* - b_1^* * X_i^*, e \sim N(0, \sigma^2)$
3) Perform regression in the transformed system
   - $\widehat{Y_i^*} = b_0^* - b_1^* * X_i^*$

Impact of λ on Transformation

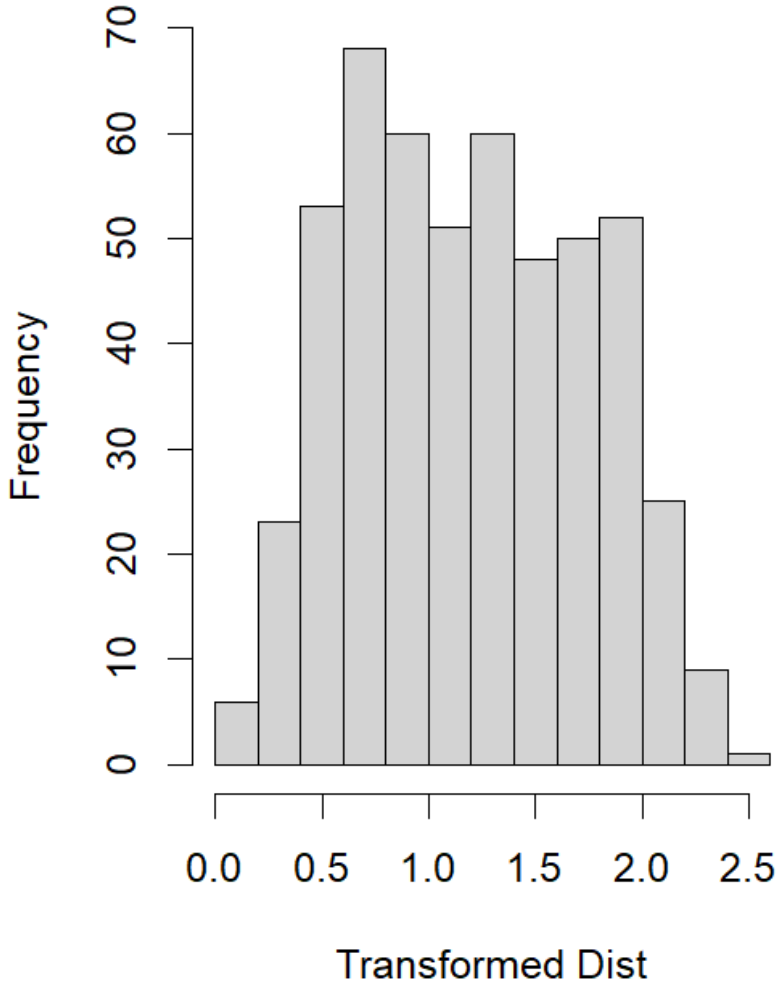




Impact of λ on Transformation

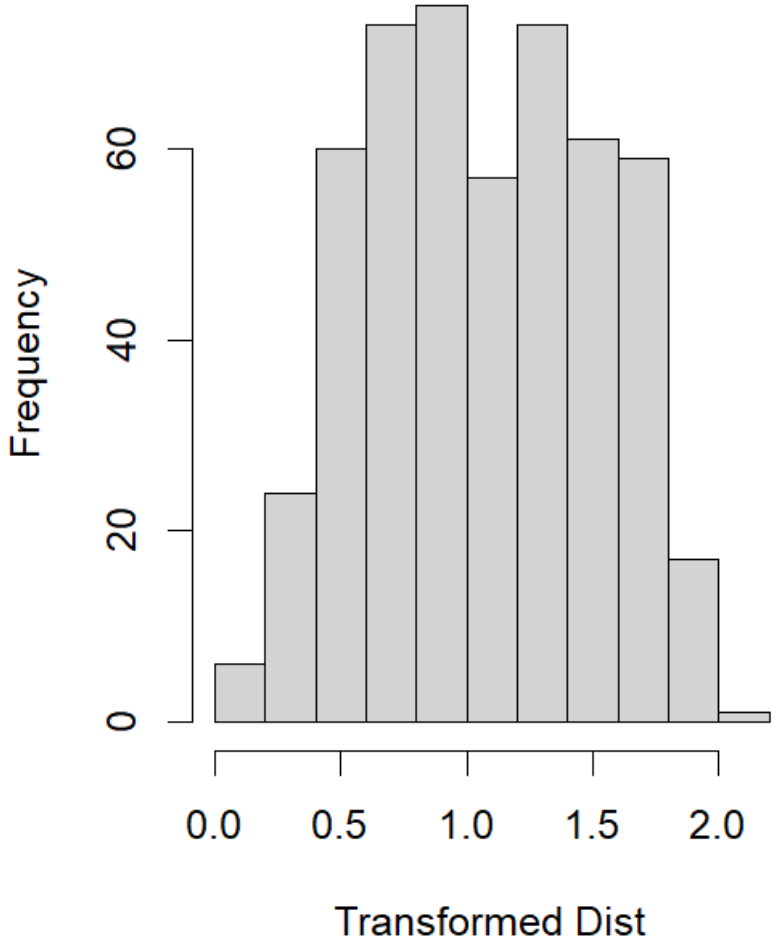
Transformed Distribution in $\log(X_i)$

$\lambda = 0$



Transformed Distribution in $\frac{X_i^{\lambda}-1}{\lambda}$

λ =-0.1556



Elasticity

- A bivariate model in transformed system:

  1) $\log(y_i) = b_0 + b_1 \cdot \log(x_i) + \varepsilon$.

- Exponentiate both side to get:

  1) $y_i = \exp(b_0 + b_1 \log(x_i) + \varepsilon)$

     $= \exp(b_0) * x_i^{b_1} * \exp(\varepsilon)$

- In the exponential model, the estimated regression coefficient b_1 is interpreted as a relative rate of change (i.e., percentage change) at a given value y_i and x_i

Log-log curve

If $b_1 > 1$ implies the steeper-to-right shape of a waxing exponential curve.

If $0 < b_i < 1$, it implies the flatter-to-right shape of a waning exponential.

If $b_i < 0$, it results in down-to-right version of these curves

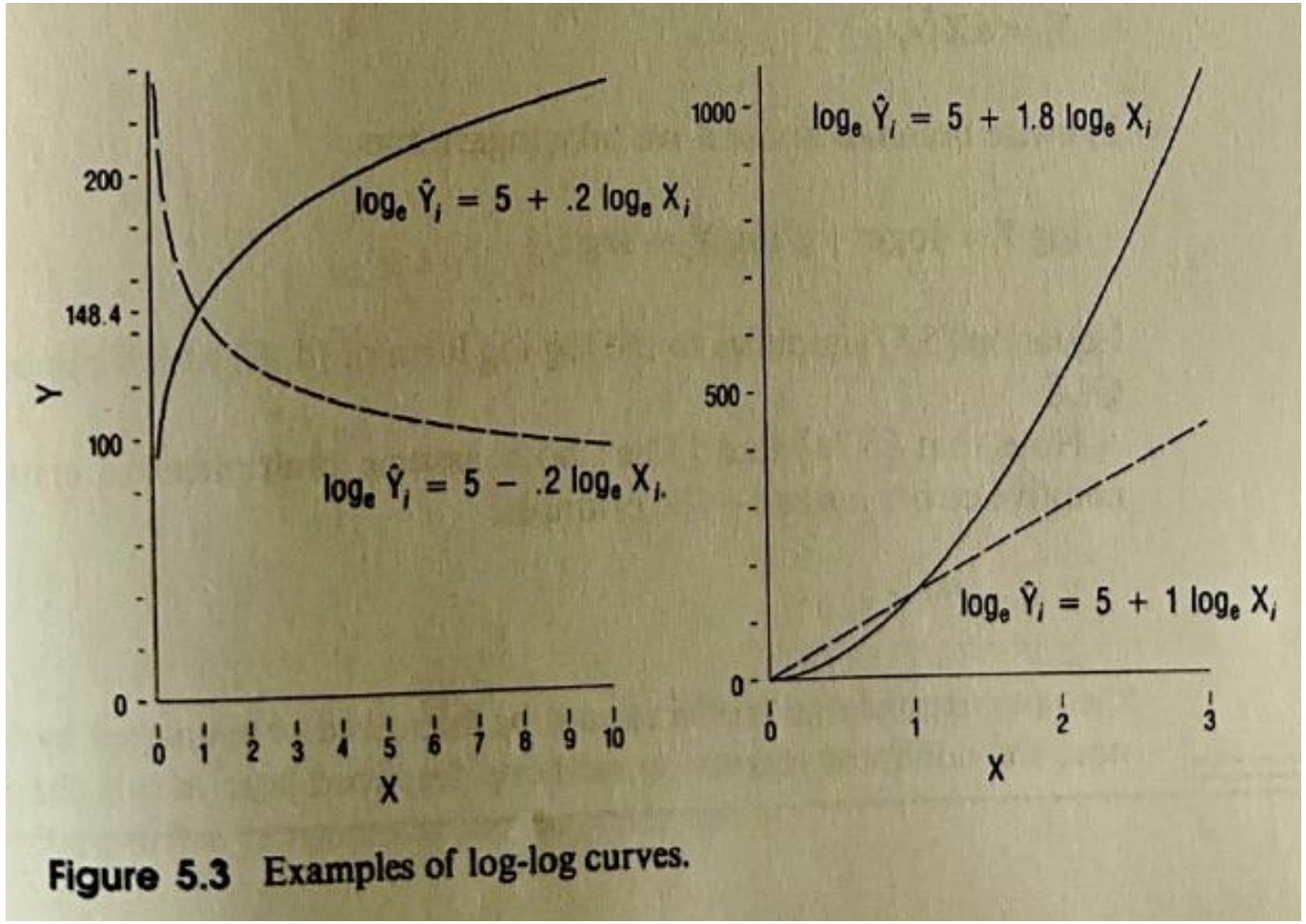


Figure 5.3 Examples of log-log curves.

Image source: Hamilton, L. C. (1992). Regression with Graphics: A Second Course In Applied Statistics. Brooks/Cole.