Diagnostic Plots

BIOS 6611

CU Anschutz

Week 9

1. Refresher: Linear Regression Assumptions

2. Types of Diagnostic Plots

3. Assumption Violation Examples

Refresher: Linear Regression Assumptions

Refresher: Linear Regression Assumptions

Recall, once again, the assumptions of a linear model. Two of them we usually infer from scientific knowledge and/or the study design:

- **Existence**: for any fixed value of X, Y is a random variable with a certain probability distribution having finite mean and variance (usually inferred from scientific knowledge or study design)
- **Independence**: The Y-values are statistically independent of one another.

Refresher: Linear Regression Assumptions (cont.)

The other three we can evaluate using diagnostic plots:

- **Linearity**: $E[Y] = \mu_{Y|X}$ is (approximately) a straight line function of X
- **Homoscedasticity**: The variance of Y is the same for any X. That is, for all possible values of X,

$$\sigma^2_{Y|X} = \sigma^2_{Y|X=1} = \sigma^2_{Y|X=2} = \ldots = \sigma^2_{Y|X=x}$$

- **Normality**: For any fixed value of X, $Y|X$ is normally distributed. (Note: does not imply that Y is normally distributed)

Types of Diagnostic Plots

Types of Diagnostic Plots

The following graphs can be used to assess linearity, homoscedasticity, and normality:

- **Y-X Scatter plot**: Plot the dependent (Y) vs independent (X) variable.
  - Very informative for simple linear regression, when we have one predictor. Less useful for visualizing with multiple predictors.
- **Residual Scatter plot**: Plot the residuals (or jackknife residuals) versus predictor(s) to look for patterns.
  - When multiple predictors, can also plot against predicted values.
- **Residual Histogram:** Allows us to visualize the distribution of residuals.
- **Normal Probability Plot (Q-Q plot)**: Plot residuals versus the expected standard deviation from a Normal Distribution.

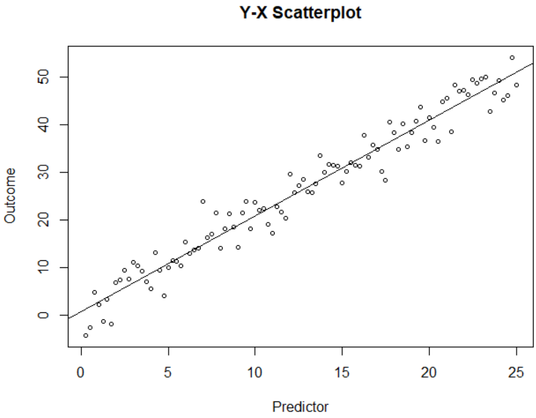
Simulate data

To illustrate what these plots should look like, we will simulate a data set where all the assumptions are satisfied:

```r
set.seed(888)
n <- 100 # number of data points
Y <- rep(NA,n) #initialize outcome vector
X <- seq(0.25,25,by=0.25) # predictor values
B0 <- 1 # pick some value of intercept
B1 <- 2 # pick some value for slope

for (i in 1:n){
  error_i <- rnorm(1,mean=0,sd=3) # Normality and heterosc assumptions
  Y[i] <- B0 + B1*X[i]+error_i # linearity assumption
}

df <- data.frame(outcome=Y, predictor=X)

mod1 <- glm(Y ~ X, data=df) # Fit model
```

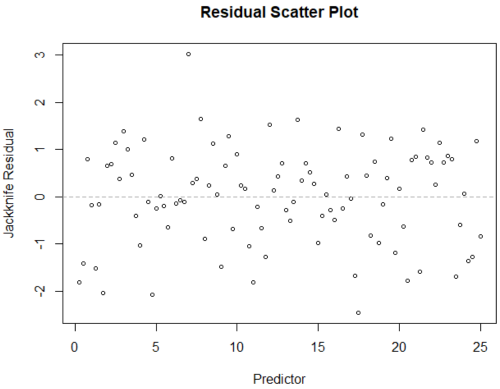
Y-X Scatterplot

If assumptions are met, we will see a linear relationship between X and Y, with points pretty evenly distributed around the fitted regression line.



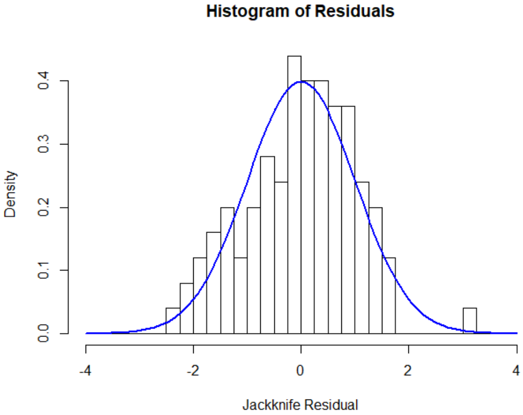
Residual Scatterplot

If assumptions are met, we will see an evenly distributed cloud of points around the x-axis.



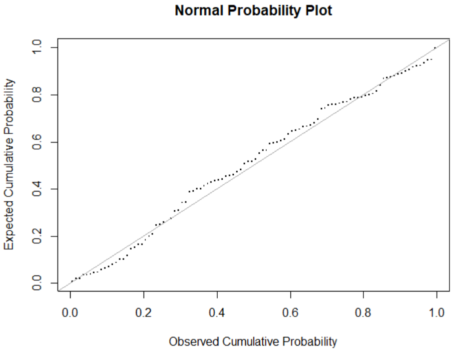
Residual Histogram

If assumptions are met, will look approximately normal.



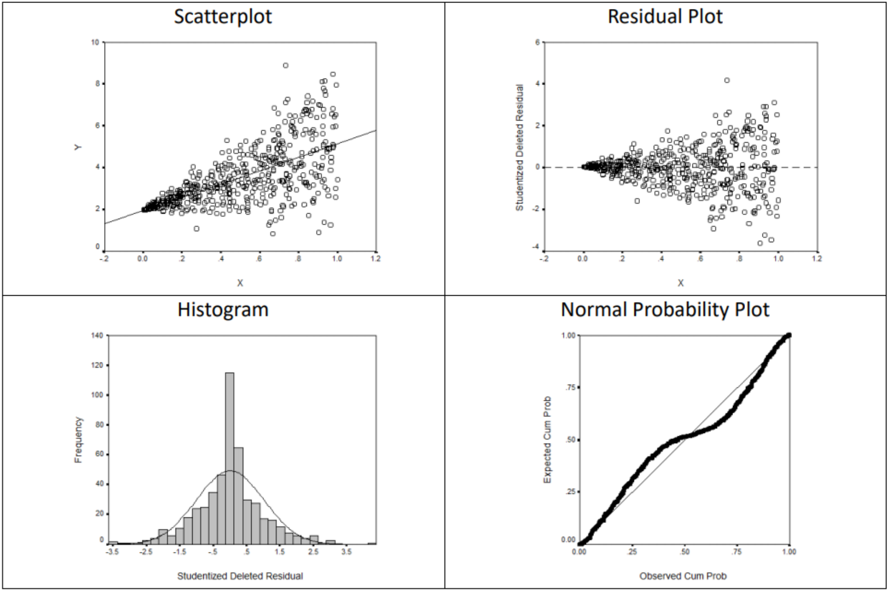
Normal Probability Plot (Q-Q plot)

If assumptions are met, should be a straight line.

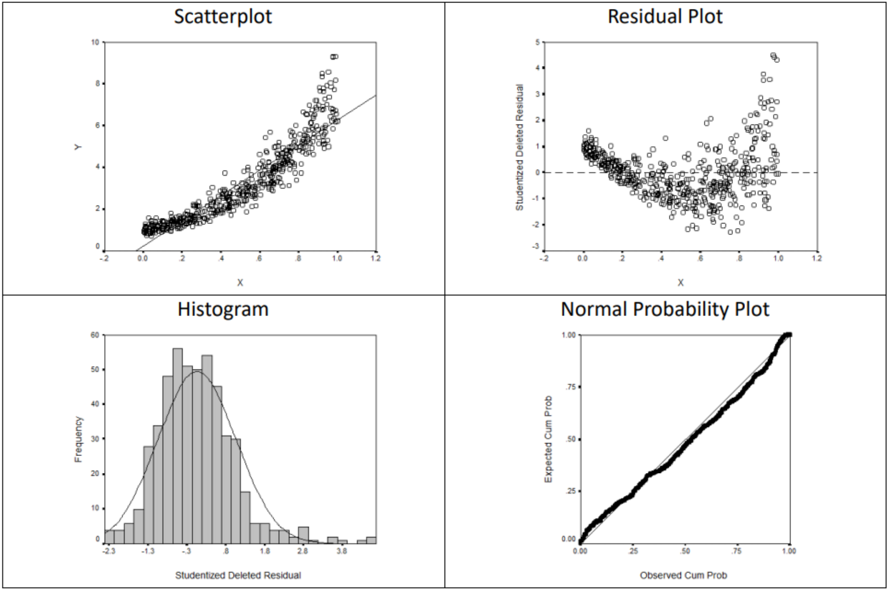


Assumption Violation Examples

Assumption Violated: Heteroscedasticity



Assumption Violated: Linearity



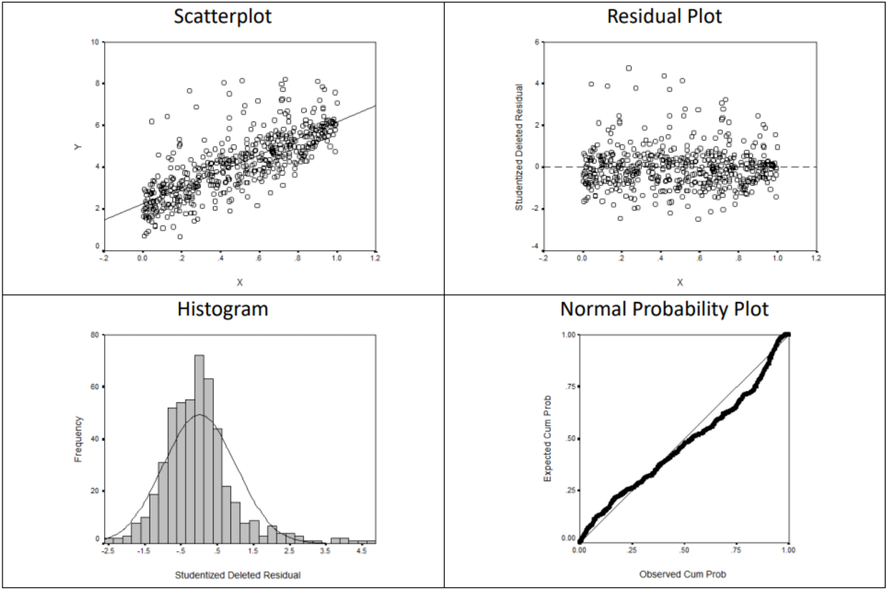
Scatterplot

Residual Plot

Histogram

Normal Probability Plot

Assumption Violated: Normality



Scatterplot

Residual Plot

Histogram

Normal Probability Plot

Summary

Diagnostic plots can be used to assess assumptions of linear models. In next lecture, we will look at transformation we can consider when these plots indicate our assumptions are violated.