2.2: Diagnostics and Remedial Measures

Dr. Bean - Stat 5100

1 Why Diagnostics

Recall that the nice properties of the OLS coefficient estimates relied on the assumption that

$$\epsilon \overset{iid}{\sim} N(0, \sigma^2).$$

Model Assumptions in Linear Regression

1. X and Y share a linear relationship
    - X and Y can be related in a non-linear way, but OLS regression cannot be used in this case
2. model describes all observations
    - no outliers or influential points
3. additional predictor variables are unnecessary
    - there is no additional information to "extract" from ϵ
4. ϵ's follow a normal distribution
    - Crucial for small sample sizes, not so critical for large (> 500) sample sizes due to central limit theorem.
5. ϵ's have constant variance
6. ϵ's are independent (possibly related to item #3)

We check assumptions using **diagnostics** and fix violated assumptions using **remedial measures**. Violations are most apparent in the **error terms** $(\epsilon_1, \ldots, \epsilon_n)$ so we focus on **residuals** $(e_1 \ldots e_n)$.

There are both **graphical** and **numerical** checks of the assumptions regarding residuals, but the graphical assumptions are more informative.

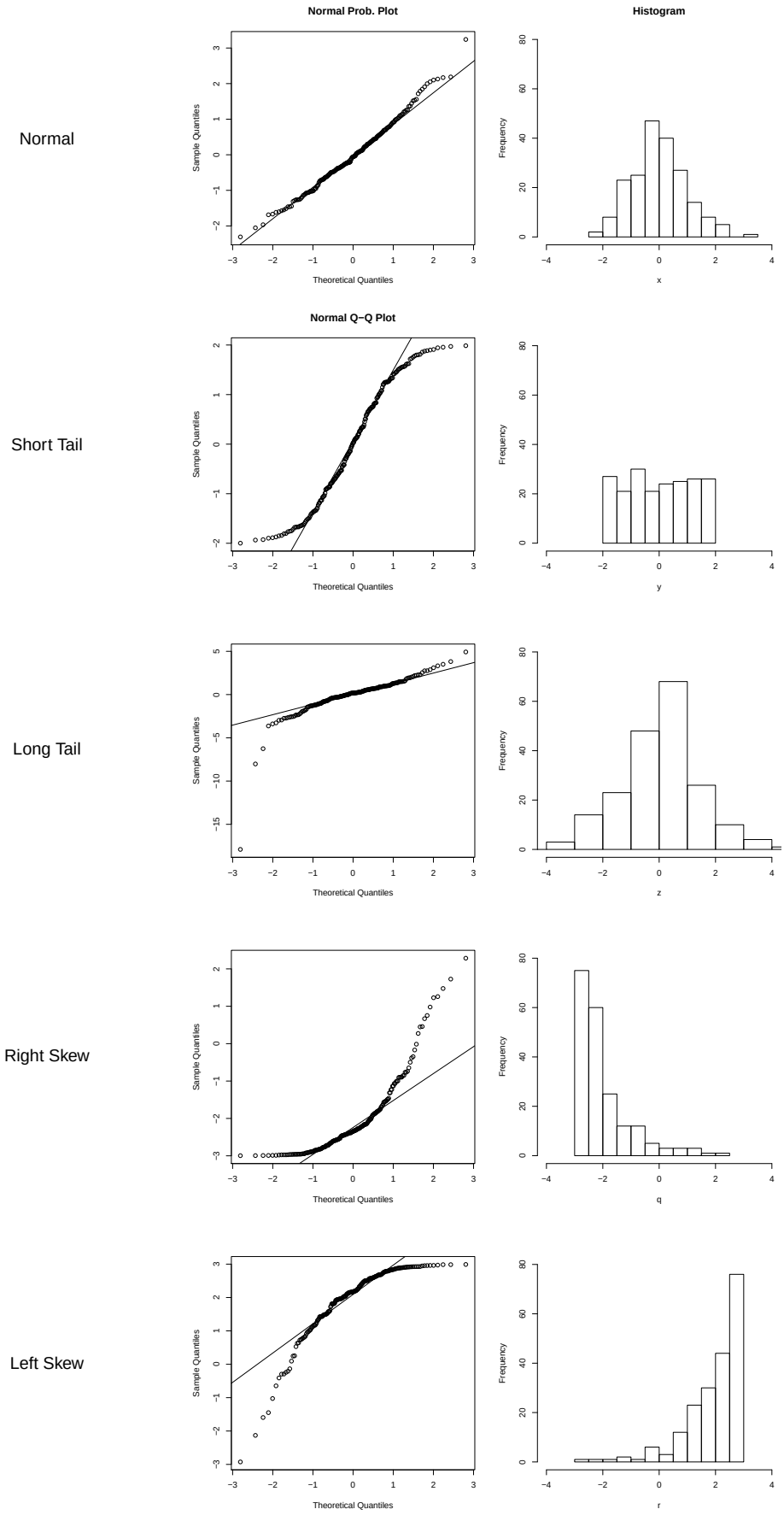


Figure 1: Example scatterplot and histograms for different data distributions.

2 Graphical Diagnostics

- **Boxplot:** quick way to check for the symmetry of residuals.
- **Histogram:** Way to check the shape of a distribution.
  - SAS will overlay a normal curve on the histogram of residuals to help check for normality.
- **Normal Probability Plot:** a qq-plot where the quantiles of the data are compared to the expected quantiles under a normal distribution
  - Expected values under normality have a mean of 0 and a SD = $\sqrt{\text{MSE}}$.
  - See page 111 in textbook for method for details about how to approximate expected observations under normality.
  - If data are approximately normal, the residuals in the Normal Probability Plot should closely follow a straight line.
- **Sequence Plot:** Line plot with residual values on the X axis and observation number on the X-axis.
  - Can "connect" the dots because there is only one Y value for every X value.
  - Look for patterns in the residuals across time/observation number.
    - Patterns would suggest that the residuals are **not independent.**
- **Residual Plot**
  - Plot e vs X or e vs $\hat{Y}$
  - Look for non-linearity and or non-constant variance in these scatterplots.

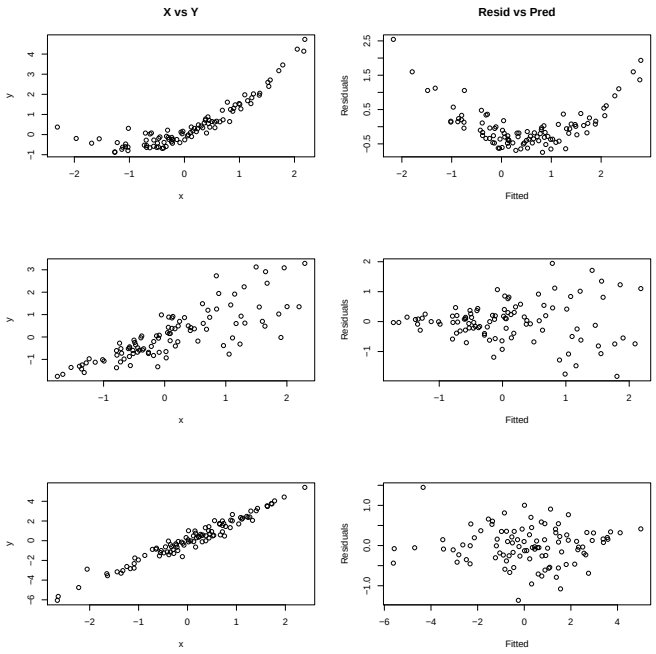


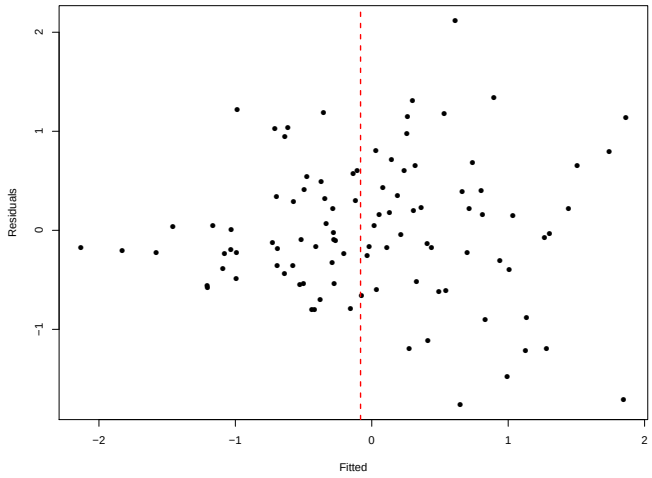
Figure 2: Plots showing 1) non-linearity, 2) non-constant variance, and 3) satisfied assumptions.

3 Numerical Diagnostics

Numerical diagnostics seek to determine if violations of model assumptions are statistically significant.

3.1 Some Numerical Tests

- Broth-Forsythe (BF) test of constant variance:



  - Split the data into two groups based on the median predicted value.
  - Calculate the median absolute deviation (MAD) $d_i = |e_i - \tilde{e}|$, where $\tilde{e}$ is the median within each group (lower and upper).
  - Conduct a two-sample t-test of d's to determine if the average of d_i's within each group are significantly different.
  - **Null Hypothesis: The variance of ϵ is constant.** (Estimated with residuals e).

Toluca Example: BF p-value is .2, which suggests there is not significant evidence of non-constant variance.

- Correlation Test of Normality
  - Calculate correlation between observed e's and the "normal-expected" e's ($e*$). Similar to expected residuals in qqplot.
  - **Null hypothesis: ϵ follows a normal distribution.**
  - If the correlation isn't at least as big as the cirtical value for $\alpha = 0.05$ in Table B.6 for a given sample size n, then reject H_0.
  - **NOTE: The p-values provided in the SAS macro output mean *nothing* for the correlation test of normality.**

Toluca Example: Correlation of $0.992 > 0.96$, so there is not significant evidence of non-normality.

- F-test for lack of fit (test of linearity between X and Y)
  - See textbook 3.7 for details.
  - Requires multiple observations at one or more X-levels. (Hard to do in observational studies or studies with multiple X-variables).
  - Basically, the test compares the regression predictions to the empirical average of Y at X-levels with multiple observations.
  - **Null hypothesis: The regression function is linear.**

Toluca Example: p-value $0.69 > .05$ suggests there is no significant evidence of non-linearity.

TABLE B.6 Critical Values for Coefficient of Correlation between Ordered Residuals and Expected Values under Normality when Distribution of Error Terms Is Normal.

	Level of Significance α				
n	.10	.05	.025	.01	.005
5	.903	.880	.865	.826	.807
6	.910	.888	.866	.838	.820
7	.918	.898	.877	.850	.828
8	.924	.906	.887	.861	.840
9	.930	.912	.894	.871	.854
10	.934	.918	.901	.879	.862
12	.942	.928	.912	.892	.876
14	.948	.935	.923	.905	.890
16	.953	.941	.929	.913	.899
18	.957	.946	.935	.920	.908
20	.960	.951	.940	.926	.916
22	.963	.954	.945	.933	.923
24	.965	.957	.949	.937	.927
26	.967	.960	.952	.941	.932
28	.969	.962	.955	.944	.936
30	.971	.964	.957	.947	.939
40	.977	.972	.966	.959	.953
50	.981	.977	.972	.966	.961
60	.984	.980	.976	.971	.967
70	.986	.983	.979	.975	.971
80	.987	.985	.982	.978	.975
90	.988	.986	.984	.980	.977
100	.989	.987	.985	.982	.979

Source: Reprinted, with permission, from S. W. Looney and T. R. Gulledge, Jr., "Use of the Correlation Coefficient with Normal Probability Plots," *The American Statistician* 39 (1985), pp. 75–79.

4 Remedial Measures

If assumptions are violated, your options are:

- Give up (at least on linear regression).

- Alternatives to OLS like Regression Trees, Quantile Regression etc. (more later in the semester).
- Depending on the X vs Y relationship, try non-linear regression (more later in the semester).
- *For non-normality and heteroskedasticity:* Variable Transformations on X or Y (not e).
  - Sometimes, a combination of transformations on both X and Y variables is needed.
  - NOTE: Removing "outlier" points from a model should be seen as a measure of last resort.

Box-Cox Approach

Great starting point to determine candidate transformations, but no "magical" solution.

- Define new response variable
$$Y' = \begin{cases} \text{sign}(\lambda)Y^\lambda & \lambda \neq 0 \\ \log(Y) & \lambda = 0 \end{cases}$$
(Note that sign(λ) preserves the original ordering of the response variable).
- Consider the theoretical model
$$Y_i^\lambda = \beta_0 + \beta_1 X_i + \epsilon_i$$
- Consider a set of candidate lambda values, use maximum likelihood estimation to determine the "best value."
  - Maximum likelihood estimation: "which value of λ is the most likely, given the data that I have observed?
- **When possible: pick an *interpretable* transformation that is close to the transformation recommended by SAS.**

  Ex: if $\lambda = .009$ is recommended, probably go with $\lambda = 0 \rightarrow \log(\lambda)$

In SAS:

```
proc transreg data=plasma;
 model boxcox(<response variable> / lambda=<lower> to <upper> by <step size>)
       = identity(<explanatory variable>) ...;
run;
```

In R:

```
library(MASS)
boxcox(<response variable> ~ <explanatory variable>, data = <dataframe>)
```

Omitted Predictors

Think of regression as a form of data mining: we want to extract as much *information* as we can from our *data*.

If we failed to extract all the information from the data, this may show up as a trend in the plot the residuals (which SHOULD have a constant mean of 0).

We will discuss more about how to extract *time* related information in data at the end of the semester.

If you apply multiple methods to fix violations of assumptions, make sure to check that the final model actually fixed the violations of assumptions.