# Principles of Statistics

Xuming He<br>
xmhe@umich.edu

## Lecture 4

## 1 Multiple Tests on the Same Hypothesis

In the brain image example, we examine different areas of the brain, and we have a different hypothesis for each test. **In some applications we may have multiple tests on the same hypothesis.**

**Example**: To study whether eating breakfast is an important factor for living longer, multiple research teams are at work. They may use different sampling plans, obtain different samples, and conduct different statistical tests. Each study has a p-value. For example, two teams get independent p-values $p_1 = 0.06$ and $p_2 = 0.10$.

**Question**: is there a good way to combine the information?

**Remember**: The p-value under the null distribution is a random variable, $U(0, 1)$.

**Discussion**: To combine the p-values we may consider

- $\max\{p_1, p_2\}, or \min\{p_1, p_2\}$.
- Recall that for the uniform random variable $U$, we have $-2\log(U) \sim \chi^2_2$, and $\chi^2$ distribution has the nice additivity property so $-2\log p_1 - 2\log p_2 \sim \chi^2_4$ under the null hypothesis. Then, we have the **"combined p-value"**:

$$P(\chi^2_4 > \text{current value of} - 2\log p_1 - 2\log p_2) = P(\chi^2_4 > 10.23) = 0.037.$$

## 2 Hypothesis Testing When $n$ is Very Large

**Discussion**: Suppose we are testing if the population mean $\theta = 0$. When we increase the sample size $n$ to infinity, the power of the test at any alternative, that is, the probability of rejecting the null under the alternative, goes to 1. In practice $\theta$ is often not exactly 0, but very small, say, 0.01, then, we would reject the null hypothesis eventually. So a slight difference from zero can be detected by increasing the sample size.

Statistical significance from a test when $n$ is very large might not mean very much. In those cases, estimates or confidence intervals might tell a better story. We should look at the practical significance of the difference from zero.

## 3 Modeling

Models are essential for statistical analysis. There are two types of models,

- theory-based models.
- convenience or data-driven models.

Given the sample, many models can be used. We should rule out the poor ones:

- those that are not convenient, too difficult to use.
- those that are not compatible with the data.

### 3.1 Linear Models

**Example**: consider we have two variables: the sales of ice creams ($Y_i$), and the temperature ($X_i$) on a given day. To examine if $Y$ depends on $X$, a linear model is a convenient.

**A Linear Model**: $Y_i = \alpha + \beta X_i + \epsilon_i$, and here $\alpha$ and $\beta$ are unknown parameters, and $\epsilon_i$ are random variables. We often assume $\epsilon_i \sim N(0, \sigma^2)$ with some parameter $\sigma^2$.

**Assumptions in Linear Models**:

- Linear trend. The expectation of $Y$ given $X$ is a linear function of $X$.
- Independence.
- The distribution of $\epsilon_i$ is normal, with a common variance that does not depend on $X$.

These assumptions are chosen for convenience. Some of them are hard to check.

**Checking the Assumptions:** Take the residuals

$$e_i = Y_i - (\hat{\alpha} + \hat{\beta} X_i),$$

where $\hat{\alpha}, \hat{\beta}$ are the least squares estimates. Then we plot the residuals to see if there is any trend.

- We can have the scatter plot of $e_i \sim X_i$.
  - This scatter plot is useful for checking if the variance changes with the explanatory variable. We would expect that the residuals scatter around the horizontal line at 0 with no apparent pattern, like in Figure 1(A).
  - In Figure 1(B), there is a quadratic trend, so the simple linear model is not appropriate.
  - In Figure 1(C), the variance is increasing with $X$, which violates the assumption that the variance is a constant for different $X$.
- We can also check the scatterplot of $e_i$ versus the fitted values $\hat{Y_i} = \hat{\alpha} + \hat{\beta} X_i$. This is equivalent to the previous one when we only have one $X$, but useful when we have multiple $X$'s. So we expect the same thing here: the residuals scatter around the horizontal line around 0 with no

apparent pattern. For example, in Figure 1(D), there is one point far away from the others, which is a potential outlier that needs to be investigated.

**Principle**: after fitting any statistical model, before jumping into conclusions, we should check the residual plot to see if any assumption is violated, and to check if there are any potential outliers. Outliers might be errors, and sometimes they represent something special.

**Question**: can we also use the scatter plot of $e_i \sim Y_i$ to check the assumptions?

**Answer**: no. Reason: We can show that $e_i$ and $X_i$ or $\hat{Y}_i$ are uncorrelated if the linear model is true. But $e_i$ and $Y_i$ are correlated even if the model is true.

**More discussion**: we can not check all the assumptions. For the residual plot, we just need a quick look to see if there is anything obviously "wrong", that is, any major issue. We should not be too demanding, that is, we should not try to look very hard for minor patterns.

**Message from the article "negative height"**: a very useful statistical model is not necessarily a correct model in every aspect.

## 3.2 The Airline Data

Please refer to the pdf file (`http://www.xuminghe.com/Airline-example.pdf`).

**Principle**: We usually start from a simple model, then check the assumptions, detect potential issues, and then update the model accordingly.

## 3.3 Goals of Modeling

**Example**: for the doctors to create the right vaccine against a virus, it is critical to identify the important factors or genes. For this case, we aim to find the right model to do statistical analysis to identify important factors or genes.

There are two main goals of statistical modeling:

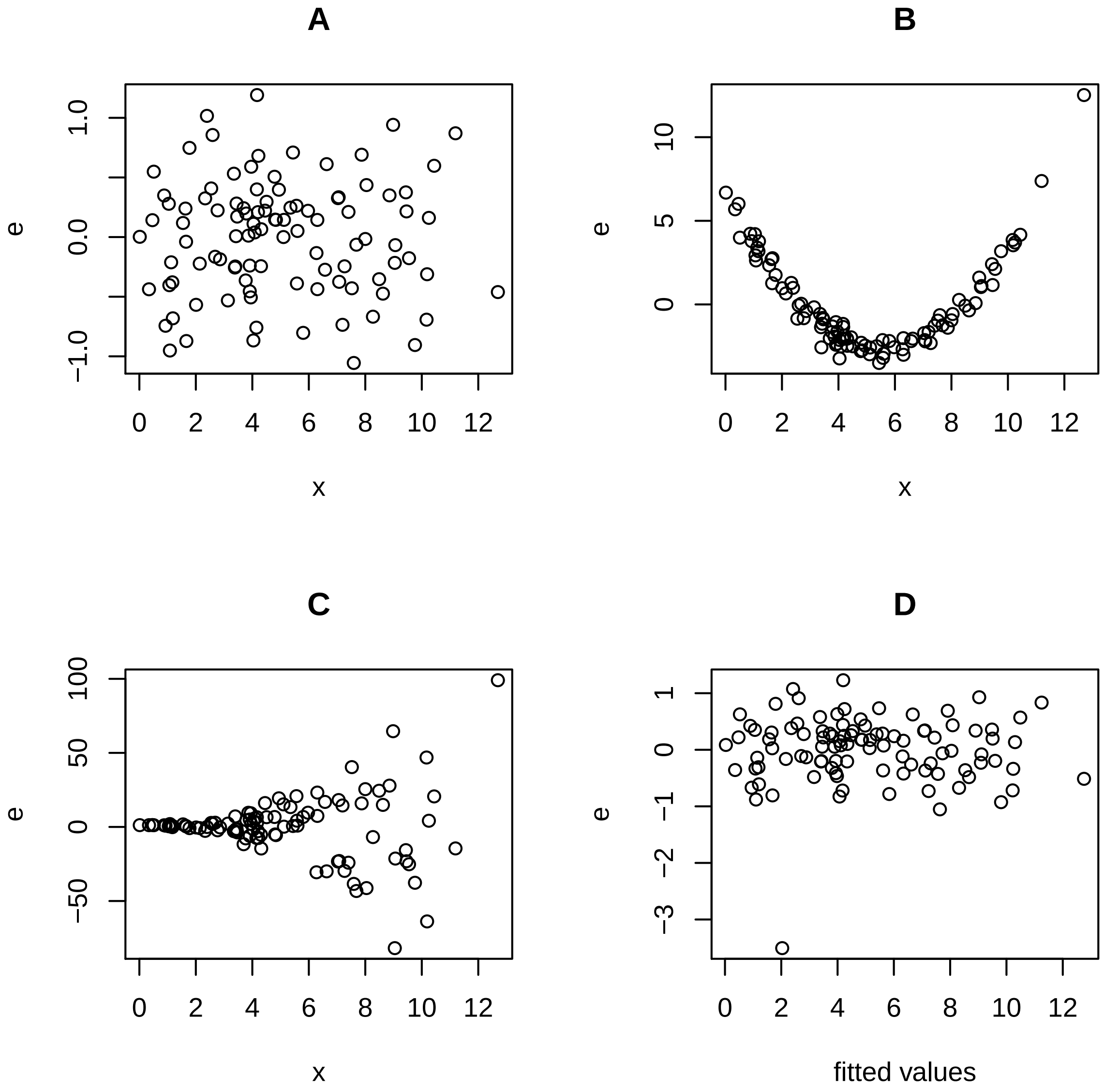

Figure 1: Residual Plots.

- Model the relationship.

- Forecasting, prediction as in e the airline example.

**Remark**: these are two different problems. To model the relationship, we need aim for a right model. But for prediction, we do not necessarily need the right model. The true model does not necessarily gives the best prediction. The reason is that we need to estimate the parameters based on a finite sample, and the estimation may be poorer for more complicated models.

**Question**: if the true model is

$$Y = \alpha + \beta_1 X_i + \beta_2 X_2 + \epsilon. \tag{1}$$

From the three possible models in the following, would the true model (M1) do always better than the other two?

$$\text{M1}\ :\ \hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2;$$

$$\text{M2}\ :\ \hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1;$$

$$\text{M3}\ :\ \hat{Y} = \hat{\alpha} + \hat{\beta}_2 X_2.$$

**Discussion**: the answer is no! The reasons are the parameters are unknown. With fewer number of parameters, we could have better estimation. How good the prediction is depends on two factors:

- How close the model is to the true model.

- How well you can estimate the parameters.

If we have $\text{Var}(\epsilon) = \sigma^2$, and if $\beta_2/\sigma$ is small, then Model M2 might give better prediction than M1.

**Principle of parsimony**: We favor simper models over larger models.