Stat 135 Lab 11

GSI: Yutong Wang

Apr 17, 2020

Berkeley
UNIVERSITY OF CALIFORNIA

Source: 四月 @douban

To-Do Today:

1. Bonferroni t-test
2. Kruskal-Wallis Test
3. Summary of Statistical Tests
    a. Parametric tests
    b. Nonparametric tests
4. Linear Regression
    a. Simple linear regression
    b. Linear regression in R

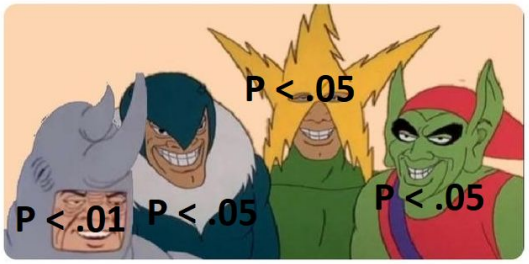
Review: Bonferroni Correction

(from Lab 10)

If k null hypotheses are to be tested, a desired overall type I error rate of at most α can be guaranteed by testing each null hypothesis at level α/k.

Equivalently, if k confidence intervals are each formed to have confidence level $100(1 - \alpha/k)\%$, they all hold simultaneously with confidence level at least $100(1 - \alpha)\%$.

Me and the significant boys



Me and the significant boys after Bonferroni correction

Review: Bonferroni t-test

Bonferroni correction might be too conservative, because as the number of tests increases, it will be harder to reject the null for a t-test. Thus, instead of testing each hypothesis at level α/k, we could do something else.

@ FB: statistical statistics meme

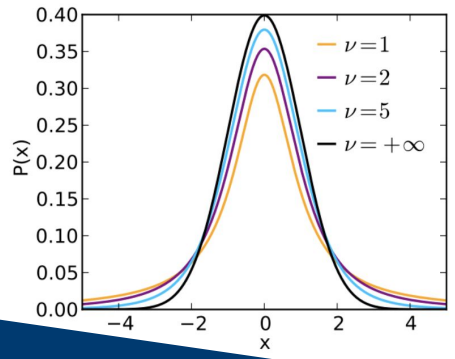
Review: Bonferroni t-test

Intuition: Given the same test statistic (t on x-axis), p-value (area on the right tail) will go smaller as the degree of freedom increases. So, if we can design some corrected t-test with a larger degree of freedom, it will become easier to get a smaller p-value, and thus reject H_0.

$$\text{Bonferroni t-test: } t_{I(J-1)} = \frac{\bar{Y}_{i.} - \bar{Y}_{j.}}{S_p\sqrt{\frac{2}{J}}}$$

$$S_p^2 = \frac{\text{SSW}}{I(J-1)} = \frac{(J-1)S_1^2 + (J-1)S_2^2 + \ldots + (J-1)S_I^2}{I(J-1)}$$

You should compute such Bonferroni t-test for each pair of groups.
In R, use pairwise.t.test(group_A_vector, group_B_vector, p.adjust = "bonferroni") to get a table of p-values of all the pairs.



Review: Kruskal-Wallis test

Non-parametric tests are very helpful whenever you are not so sure about the parametric assumptions in parametric tests.

We have learned Mann-Whitney test to compare the distribution between treatm-ent and control group.

But what if there are more than 2 groups?

@ FB: statistical statistics meme

Review: Kruskal-Wallis test

The Kruskal-Wallis test is a generalization of the Mann-Whitney test.

Assumptions & Notations: The observations are assumed to be independent, but no particular distributional assumption. The observations are pooled together and ranked. Suppose

$$R_{ij} = \text{the rank of } Y_{ij} \text{ in the combined sample}$$

$$\bar{R}_{i\cdot} = \frac{1}{J_i}\sum_{j=1}^{J_i} R_{ij}, \text{ avg rank in the } i\text{th group}$$

$$\bar{R}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J_i} R_{ij} = \frac{N+1}{2}$$

where N is the total number of observations.

Review: Kruskal-Wallis test

Intuition: H_0 states that the probability distributions in different groups/treatment are identical. Thus,

$$\text{SSB} = \sum_{i=1}^{I} J_i(\bar{R}_{i\cdot} - \bar{R}_{\cdot\cdot})^2$$

is a nice metric to measure the difference between $\bar{R}_{i\cdot}$. The larger SSB is, the more likely we would reject H_0.

Test Statistic: Under H_0, the test statistic

$$K = \frac{12}{N(N+1)}\text{SSB}$$

is approximately distributed as a chi-squared random variable with degree of freedom equal to $I - 1$.
Note, it might be easier to compute the test statistic in the other form

$$K = \frac{12}{N(N+1)}\left(\sum_{i=1}^{I} J_i \bar{R}_{i\cdot}^2\right) - 3(N+1)$$

Take-away from Statistical Tests (1)

Given the data and hypothesis, which test should we use?

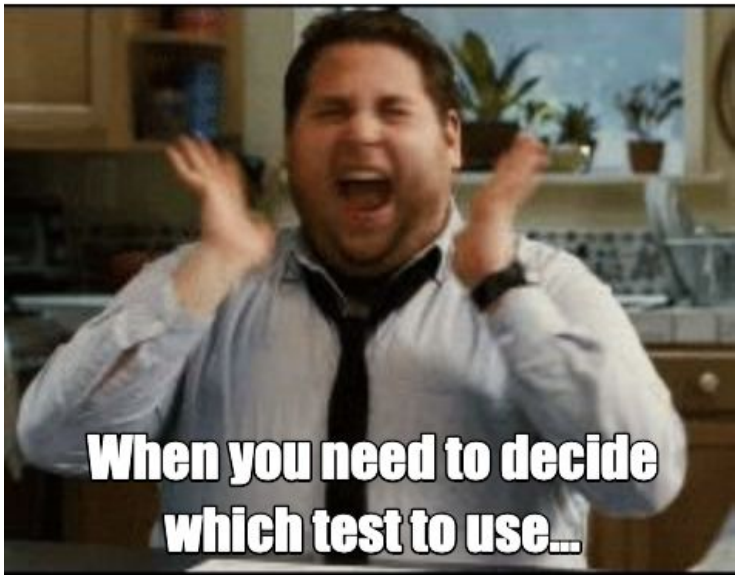
Parametric tests we have learnt:

1. 2-sample t-test (paired/unpaired, equal/unequal var)
2. Goodness of fit chi-squared test
3. Test of homogeneity
4. Test of independence
5. F test (one-way ANOVA)
6. Bonferroni t-test

Nonparametric tests we have learnt:

1. Mann-Whitney Test/Wilcoxon rank sum test
2. Wilcoxon signed rank test
3. Kruskal-Wallis test

Maybe even more!!



Take-away from Statistical Tests (2)

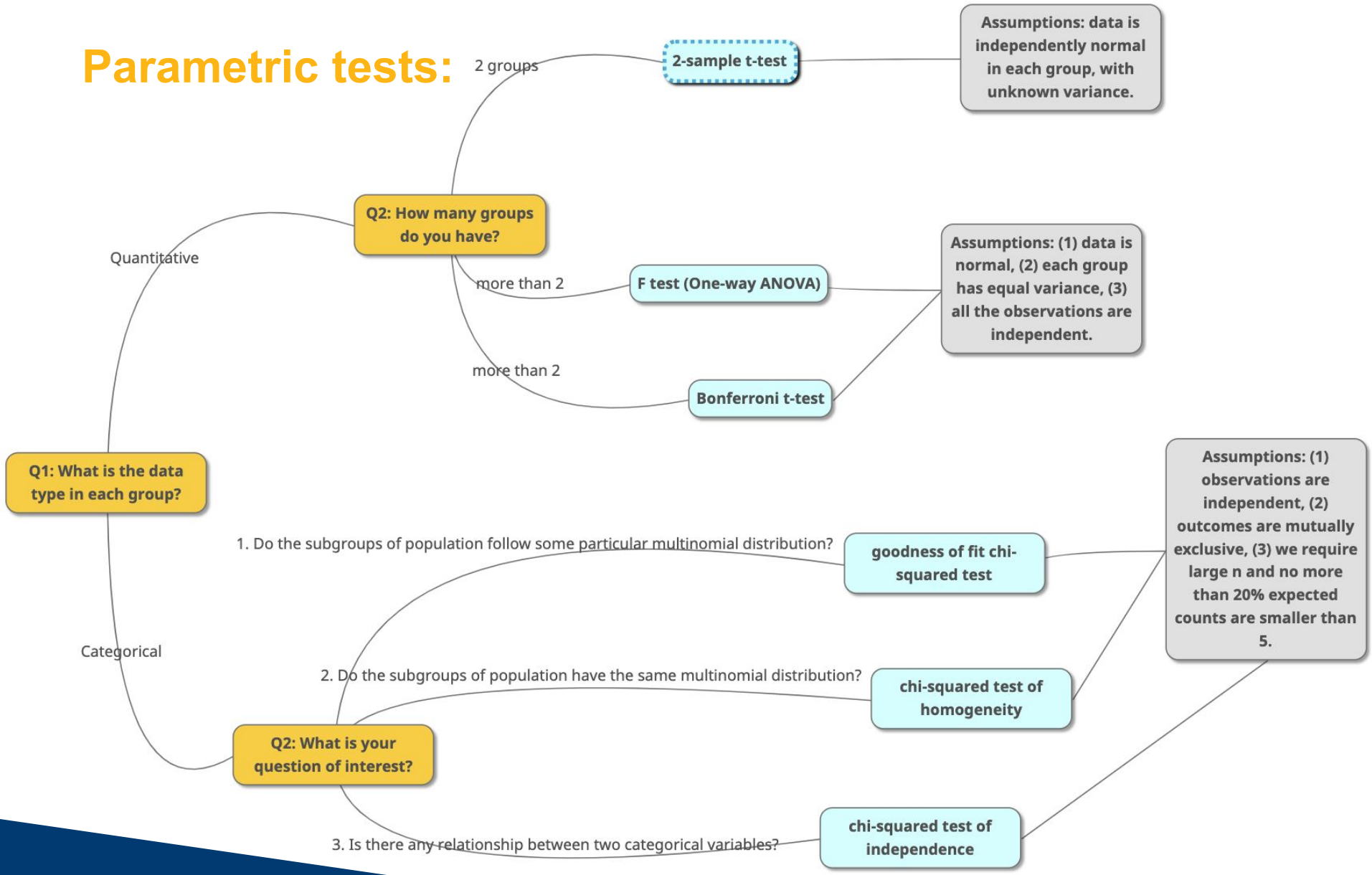
Given the data and hypothesis, which test should we use?

Normally, the question of interest is whether there is any significant difference among the treatments/groups/levels.

Step 1: Check the diagram in the next page and choose the appropriate parametric test accordingly.

Step 2: Make sure to check if the parametric assumptions hold. If not, nonparametric tests would always be happy to help!

Parametric tests:



Take-away from Statistical Tests (4)

Q1: How could I choose between F test and Bonferroni t-test?

F test can only tell you if there is at least group that is significant different, but you do not know which one is different.

Meanwhile, Bonferroni t-test can locate the pair of groups that are significantly different.

Q2: What is the difference between TOH and TOI?

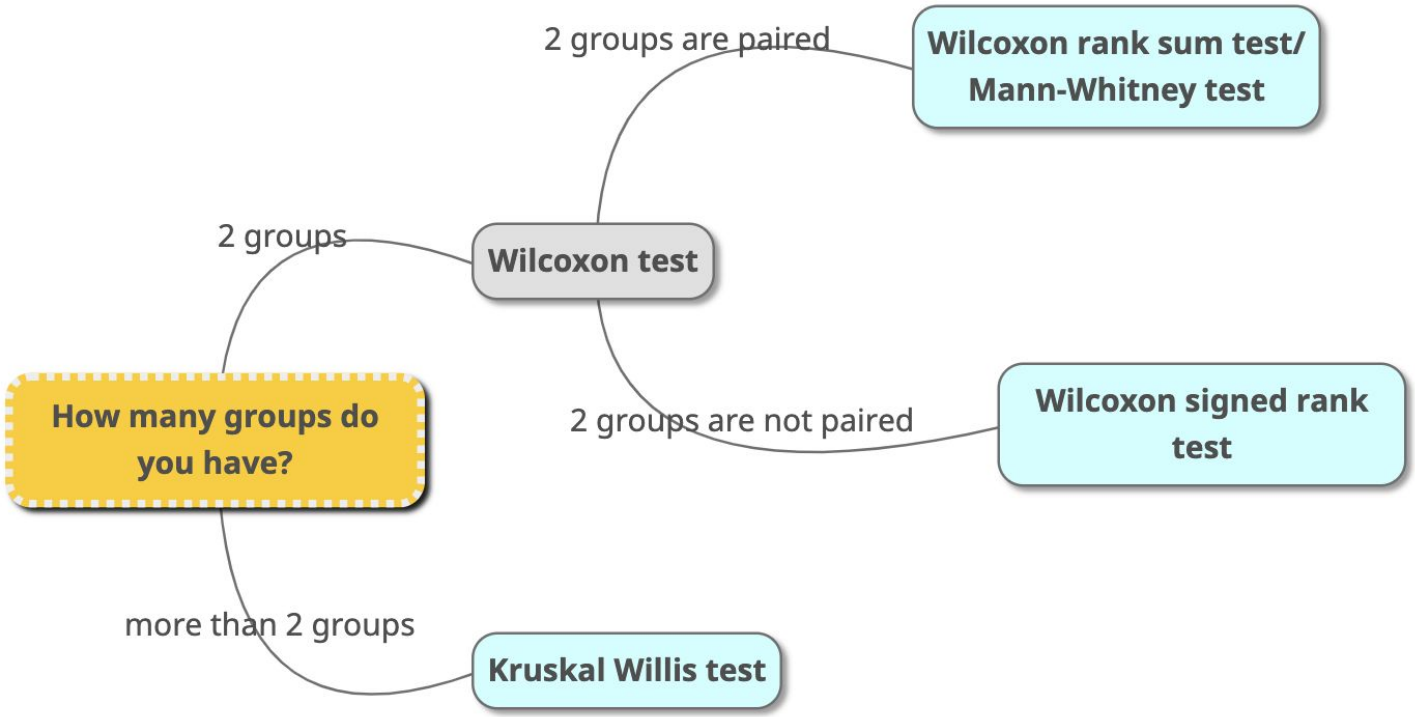
They are very similar to each other, except that they are answering a slightly different question. The testing procedures are exactly the same.

Q3: What if any of the assumption is not met in the test I chose?

Good question! Then we will need to use nonparametric tests in the next page.

Take-away from Statistical Tests (5)

Non-parametric tests:



Take-away from Statistical Tests (6)

How to decide whether the data is paired or not?

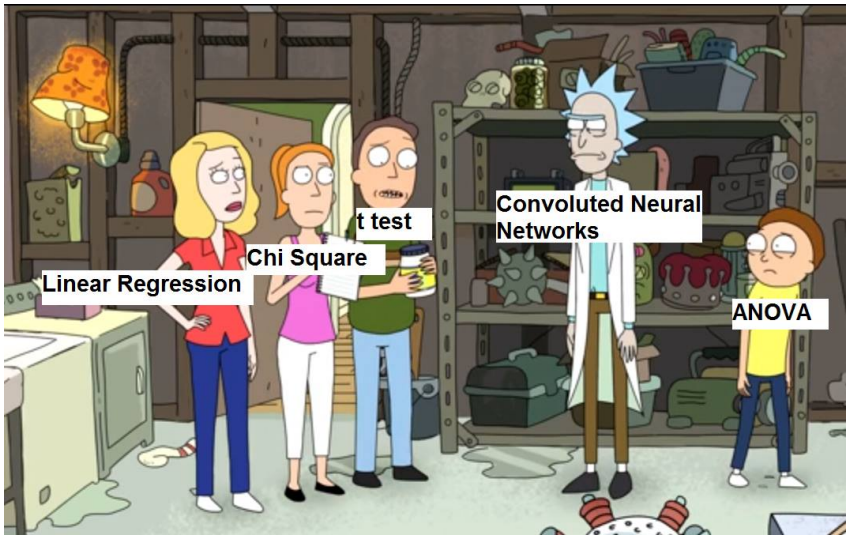
Pairing comes from consideration of what was sampled, instead of whether two groups have the same sample size or not.

The pairs of values are at least somewhat positively dependent, while unpaired values are not dependent. Often the dependence-pairing occurs because they're observations on the same unit (repeated measures), e.g. before-versus-after measure.

Reference: stackexchange by Glen_b -Reinstate Monica

Review: Linear Regression

Now, it is time for us to dive into linear regression, which is the key element in statistical learning. You will be able to predict some outcome given predictors using linear models!



Review: Linear Regression

Terminology

$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$ where $\mathbf{Y} \in \mathbb{R}^n$, $\mathbf{X}$ is the design matrix.
β is an unknown constant vector, and $\mathbf{e}$ is the noise term.

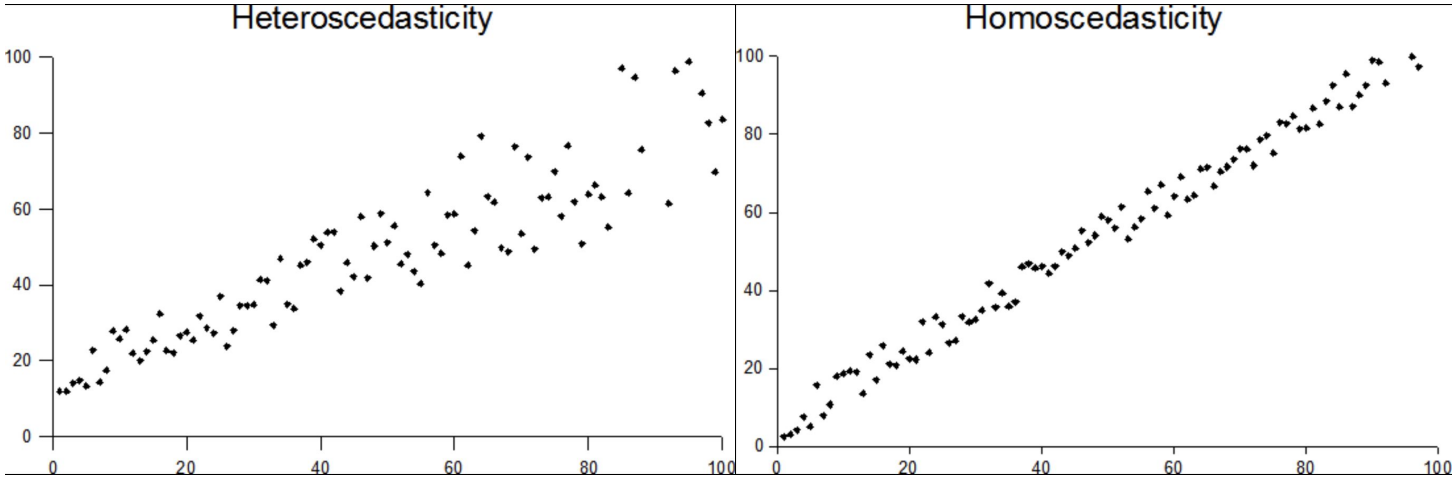
- $n \geq k + 1$, and the **design matrix X** spans a $k + 1$ dimension subspace of $\mathbb{R}^n$.
- $\mathbf{X}$ is of full rank, i.e., the columns of $\mathbf{X}$ are independent.
- $e_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$.
- **Homoscedasticity**: Each y_i is of the same variance σ^2, and independent of $\mathbf{X}$.

$$\begin{bmatrix} y_1 \\ \dots \\ \dots \\ \dots \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{1k} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}_{n\times(k+1)} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{bmatrix}_{(k+1)\times 1} + \begin{bmatrix} e_1 \\ \dots \\ \dots \\ \dots \\ e_n \end{bmatrix}_{n\times 1}$$

Review: Simple Linear Regression

Terminology

Simple linear regression is the case when $k = 1$, i.e., $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \mathbf{e}$.
The **Thyche line** is $\mathbb{E}(\mathbf{y}) = \beta_0 + \beta_1 \mathbf{x}$.
The **regression line** is $\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}$, which is the optimal fitting line through the data.
The **residual error** is $\hat{e}_i = y_i - \hat{y}_i$, which is an estimator of e_i.



Source: Wikipedia

Review: Simple Linear Regression

How to find $\hat{\beta}_0$ and $\hat{\beta}_1$?

The optimal fitting line is found by minimizing the residual sum of squares (RSS) which is

$$\text{RSS} = \|\hat{\mathbf{e}}\|_2^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 = \|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

To minimize RSS, we can

- either use least squares (set the partial derivatives of $\|\mathbf{y} - \mathbf{X}\hat{\beta}\|$ w.r.t. β_0, β_1 equal to 0),
- or apply the orthogonal projection $(\mathbf{X}'_{(k+1)\times n} \cdot (\mathbf{y} - \mathbf{X}\hat{\beta})_{n\times 1} = \mathbf{0}_{(k+1)\times 1})$

$$\Longrightarrow \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Here, we implicitly assumed that $\mathbf{X}'\mathbf{X}$ is invertible in the formulation of $\hat{\beta}$. This is not true in general, and this is the reason why we have to assume that $n \geq k + 1$ and the design matrix $\mathbf{X}$ is full rank in the first place.

See Lemma A in page 566, Rice for why such assumptions are needed for $\mathbf{X}'\mathbf{X}$ to be invertible.

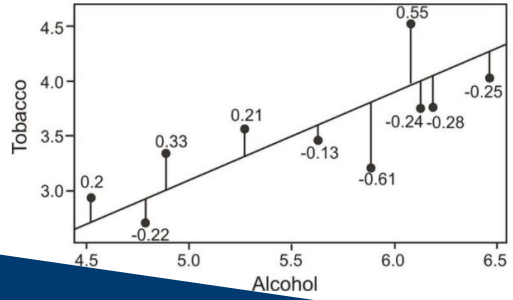


Figure: https://www.ck12.org/statistics/least-squares-regression/lesson/Least-Squares-Regression-ADV-PST/

Berkeley
UNIVERSITY OF CALIFORNIA

Review: Simple Linear Regression

We can also express $\hat{\beta}_0$ and $\hat{\beta}_1$ in terms of the covariance and variance of $\mathbf{X}$ and $\mathbf{Y}$ in the simple linear regression scenario.
When $k = 1$, we have $\mathbf{y} = [y_1, ..., y_n]^T_{n\times 1}, \mathbf{x} = [x_1, ..., x_n]^T_{n\times 1}$. Thus,

$$\mathbb{E}[\mathbf{x}] = \bar{\mathbf{x}}, \mathbb{E}[\mathbf{y}] = \bar{\mathbf{y}}$$

$$s_{xx} = \mathrm{Var}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2, \; s_{yy} = \mathrm{Var}(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{\mathbf{y}})^2$$

$$s_{xy} = \mathrm{Cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})(y_i - \bar{\mathbf{y}})$$

$$\Longrightarrow \hat{\beta}_1 = \frac{\mathrm{Cov}(\mathbf{x}, \mathbf{y})}{\mathrm{Var}(\mathbf{x})}, \; \hat{\beta}_0 = \mathbb{E}(\mathbf{y}) - \hat{\beta}_1\mathbb{E}(\mathbf{x})$$

Q1: Problem 9G in HW: The fitted values.

Problem 9G: The fitted values. The estimated values $\hat{y}$ are often called the $f$itted values because they are obtained by fitting the regression model to the data. Use the fact that $\hat{y}$ is a linear function of x to show that:

a) $\bar{\hat{y}} = \bar{y}$. That is, the average of the fitted values equals the average of the original values.

b) $\sigma^2_{\hat{y}} = r^2\sigma^2_y$. Check that this gives sensible answers when r $= 0$ and when r $= 1$.

Q2: Problem 9H in HW9: The residuals.

Problem 9H: The residuals. For each $i = 1, 2, 3, \ldots, n$, define $\hat{e}_i = \hat{y}_i - y_i$ to be the ith *residual*, that is, the error in the regression estimate of y_i.

a) Show that $\bar{\hat{e}} = 0$.

b) Show that $\sigma_{\hat{e}}^2 = (1 - r^2)\sigma_y^2$. This implies that the larger r gets, the less overall error there is in the regression. Check that the answer is sensible when $r = 0$ and when $r = 1$.

Review: Linear Regression in R

In R, suppose you have two vectors x, y of size n. You may use lm() function to fit a linear regression line. (insert ?lm in R for more details.)

formula: A typical model has the form response $\sim$ terms where response is the (numeric) response vector and terms is a series of terms which specifies a linear predictor for response.

returned values: An object of class "lm" is a list containing at least the following components: coefficients, residuals, and fitted mean values.

omit the intercept: A formula has an implied intercept term. To remove this use either $y \sim x-1$ or $y \sim x+0$.

To compute the estimated regression coefficients, there is another method. You could implement $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ in R via

$$\text{beta\_hat} \leftarrow \text{solve(t(X)\%*\%X) \%*\%t(X)\%*\%y}$$

We can also calculate the **confidence intervals** of the estimated regression coefficients, **plot** the regression lines as well as residual plots, etc. For more details in R, see the demo R file in my course website.

Q3: Linear regression in R (Rice 14.38)

Rice 14.38: The file sapphire lists observed values of Youngs modulus (g) measured at various temperatures (T) for sapphire rods (Ku 1969). Fit a linear relationship $g = \beta_0 + \beta_1 t$ and form confidence intervals for the coefficients. Examine the residuals.
See data on bCourses.

Q4:Simple Linear Regression for Quadratic Functions (Rice 14.9.39)

As part of a nuclear safeguards program, the contents of a tank are routinely measured. The determination of volume is made indirectly by measuring the difference in pressure at the top and at the bottom of the tank. The tank is cylindrical in shape, but its internal geometry is complicated by various pipes and agitator paddles. Without these complications, pressure and volume should have a linear relationship. To calibrate pressure with respect to volume, known quantities (x) of liquid are placed in the tank and pressure readings (y) are taken. The data in the file tank volume are from Knafl et al. (1984). The units of volume are kiloliters and those of pressure are pascals.

a. Plot pressure versus volume. Does the relationship appear linear?

b. Calculate the linear regression of pressure on volume, and plot the residuals versus volume. What does the residual plot show?

c. Try fitting pressure as a quadratic function of volume. What do you think of the fit?

(See data on bCourses)