Weekly Review 12

Last week we introduced the simple linear regression and discussed statistical inferences for the unknown slope and unknown intercept, and then we developed a measure of the strength of the linear relationship between x and y, viz., the coefficient of determination R^2, which is defined to be

$$R^2 := \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2},$$

and we showed that $0 \leq R^2 \leq 1$.

Recall that for $x_1, \ldots, x_n$, the sample variance is $\sum(x_i - \overline{x})^2/(n-1)$, which is a measure of variation of x. The purpose of dividing the sum of squares by $n-1$ is kind of averaging, and so the sum $\sum(x_i - \overline{x})^2$ itself can be interpreted as the total variation of x.

Now, investigating the expression of the coefficient of determination closely, we can see that the denominator is the total variation of y, whilst the numerator is the total variation of $\hat{y}$. However, the variation in $\hat{y}_i$ does NOT come from randomness but comes purely from the variation in x_i; different x_i-values give different but deterministic $\hat{y}_i$-values ("deterministic" means that we will get the **same** $\hat{y}_i$-value no matter how many times we repeat the **same** x_i-value; the random error is not involved in $\hat{y}_i$), and this variation of $\hat{y}$ can be "explained" entirely by the regression line.

That is to say, R^2 actually represents the proportion of the total variation of y that can be ascribed to the regression line, which specifies a linear relationship between x and y. That is, R^2 is a measure of the strength of the linear relationship between x and y. The higher the value of R^2, the better the regression line can explain the variation in the observed y. E.g. if $R^2 = 90\%$, then we say that the regression line can explain 90% of the total variation of y_i, and if $R^2 = 100\%$, then all y_i lie on the regression line and so the regression line can explain 100% of the variation in y_i. If $R^2 = 0$, it means $\hat{y}_i = \overline{y}$ for all i, i.e. the regression line is a flat line (of zero slope) and so $\hat{y}_i$ is a constant. Then no variation in y_i can be explained by this flat line because a flat line has no variation in the y-value.

If we take the square root of R^2 and choose the sign so that it is the same sign as that of the slope of the regression line, we obtain the *coefficient of correlation*, or *correlation coefficient*, which (after some laborious algebraic work that has not been shown) can be expressed as

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}.$$

Since the coefficient of correlation is typically denoted by r, the coefficient of determination sometimes is denoted by r^2.

The coefficient of correlation must lie between -1 and 1. If it is positive, then when x increases, y increases. If it is negative, then when x increases, y decreases. If it is zero, then x and y have no linear (i.e. straight line) relation. When it is $+1$ or -1, all data points are lying on the straight line, i.e. there is no error. The larger the absolute value of r, the stronger the correlation between x and y. However, note that $r = 0$ does not imply independence; it only implies that there is no linear relation between x and y. Also note that high correlation does not imply cause–effect relationship.

If we consider both X and Y are random variables, then we have a population coefficient of correlation, denoted by the Greek letter ρ (rho). Suppose that both X and Y have normal distributions, then to test the null hypothesis that $\rho = 0$, we use the test statistic:

$$t = r\sqrt{\frac{n-2}{1-r^2}},$$

which, under the null hypothesis, follows the t-distribution with degrees of freedom $n - 2$. This is the first time that we have the t-distribution without having s^2 in the formula, but in fact in the formula of r we actually have sum of squares. (In fact s_e^2 is hidden somewhere in $(1 - r^2)$; a few algebraic steps can show that $(1 - r^2)/(n - 2) = s_e^2/SS_{yy}$.) Note that this test statistic is only applicable when we test the null hypothesis of zero correlation. (For your interest: To test whether $\rho = \rho_0$ for $\rho_0 \neq 0$ or to construct confidence intervals for ρ, we have to consider the *Fisher Z-transformation*, which is not discussed this semester. Make a Google search if you are interested.)

The last chapter is devoted to the so-called *nonparametric tests* (also called *distribution-free tests*), which are tests that do not require the assumption that the population has a normal distribution. That is, a randomly chosen datum from the population does not follow the normal distribution. (I did not say that the test statistic does not have the normal distribution. As a matter of fact, for the tests to be discussed in this chapter, when the sample size is large, some test statistics follow approximately the normal distribution and one test statistic follows approximately the χ^2-distribution.)

The first nonparametric test we discussed, the *sign test*, is a nonparametric version of the one-sample t-test; it is used to test the null hypothesis that a population median is equal to a particular value.

The sign test considers whether data are above or below the hypothesised median. If a datum is above the hypothesised median, it is a "+"; if it is below, then it is replaced by a "−"; if it is the same as the hypothesised median, then discard it. The we count the number of + signs. Suppose we have n signs in total (that is to say, n is the sample size minus the number of data that have been discarded) and the number of + signs is x.

In order to make a decision (reject or not reject the null hypothesis), we are going to calculate the p-value. The decision rule is that we reject the null hypothesis whenever the p-value is less than or equal to α. The calculation of p-value, as we have discussed, depends on the alternative hypothesis in the following way.

- If the alternative is "$>$", then the larger the value of x, the more likely the alternative is true. The p-value is therefore equal to $\Pr(X \geq x)$, where $X \sim \mathrm{B}(n, 0.5)$.
- If the alternative is "$<$", then the p-value is $\Pr(X \leq x)$.
- If the alternative is "$\neq$", the p-value is equal to the smaller one of $2 \times \Pr(X \geq x)$ and $2 \times \Pr(X \leq x)$.

Remind that when n is large, we can approximate the binomial distribution by the normal distribution.

Since we are able to calculate the p-value for any n easily, there is no particular advantage for us to use the critical value approach illustrated in the textbook, and so we stick to the p-value approach for the sign test in this course.

The sign test is a nonparametric analogue to the one-sample t-test. We have another one, which is known as the *Wilcoxon signed-rank test*. That is to say, the signed-rank test also tests the null hypothesis that a population median is equal to a particular value. The idea is to consider not only the signs but also the magnitudes of the differences between the data and the hypothesised median. The procedure is as follows.

1. Calculate all differences between the data and the hypothesised median,
2. Rank the differences according to the absolute values (i.e. according to only the magnitudes but not the signs).
3. **We discard the data that are equal to the hypothesised median.**
   (And the same as in sign test, $n =$ sample size $-$ number of data discarded.)
4. **If two or more absolute values of the differences are equal, we assign each one the mean of the ranks which they jointly occupy.**
5. Let T^+ be the sum of the ranks of all positive differences and T^- be the sum of the ranks of all negative differences and $T = \min(T^+, T^-)$.

Using the same argument as the sign test above, we can see that if the alternative

- is "$<$", the smaller the value of T^+, the more likely the alternative is true;
- is "$>$", the smaller the value of T^-, the more likely the alternative is true;
- is "$\neq$", the smaller the value of T, the more likely the alternative is true.

(Note that the textbook uses T to denote the test statistic in all the three different cases above. But I believe our notations will be clearer.)

Therefore, what we need to decide is: how small is too small. The answer is: it is too small when it is less than or equal to the critical value given in Table IX on page 693. If you find a dash instead of a number in the table, it means that no matter what you observe, your sample size is so small that we will never reject the null hypothesis at that significance level. (It is worthwhile to note that the value 0 in the table can remind you that this table tells you how small is too small, because we will never ask whether 0 is too large or not, as 0 is the smallest possible value for this test statistic.)

If n is large, we may approximate the distribution of T^+ by the normal distribution with mean $\mu_{T^+} = n(n+1)/4$ and variance $\sigma^2_{T^+} = n(n+1)(2n+1)/24$. The large sample test can be based on just T^+ (just like in the sign test, it suffices to consider only the number of positive differences):

$$z = \frac{T^+ - \mu_{T^+}}{\sigma_{T^+}}.$$

If the alternative is "$<$", reject the null hypothesis if $z \leq -z_\alpha$; if the alternative is "$>$", reject the null hypothesis if $z \geq z_\alpha$; if the alternative is "$\neq$", reject the null hypothesis if $|z| \geq z_{\alpha/2}$.

That is to say, our decision rule is:

H_0	H_A	(Critical region) Reject H_0 at the α significance level if
$\tilde{\mu} = \tilde{\mu}_0$ or $\tilde{\mu} \geq \tilde{\mu}_0$	$\tilde{\mu} < \tilde{\mu}_0$	$T^+ \leq$ one-tailed critical value ($z \leq -z_\alpha$ if n is large),
$\tilde{\mu} = \tilde{\mu}_0$ or $\tilde{\mu} \leq \tilde{\mu}_0$	$\tilde{\mu} > \tilde{\mu}_0$	$T^- \leq$ one-tailed critical value ($z \geq z_\alpha$ if n is large),
$\tilde{\mu} = \tilde{\mu}_0$	$\tilde{\mu} \neq \tilde{\mu}_0$	$T \leq$ two-tailed critical value ($\lvert z\rvert \geq z_{\alpha/2}$ if n is large).

The sign test and the signed-rank test are one-sample tests, **which are applicable to the paired-sample situation**. (How? For a paired-sample say $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, we take pairwise differences $d_i = x_i - y_i$ and then apply one-sample tests to $\{d_1, \ldots, d_n\}$.)

The *Wilcoxon rank-sum test* is the nonparametric version of two-independent-sample t-test and is used to test the null hypothesis that two populations have the same median (or mean, because this test requires the assumption that **the two distributions have the same shape**, i.e. if the medians are the same, the two distributions are identical, and consequently the means are also the same). The procedure is as follows.

1. We have two samples.

2. **The sample with smaller sample size is labelled sample 1.** If the two samples are of equal size, then choose arbitrarily one of them to be sample 1.

3. We pool them together and then rank them from the smallest to the largest. (The same as in the signed-rank test, for tied observations, we **assign each of the tied observation the mean of the ranks they jointly occupy**, but unlike in the signed-rank test, **negative numbers are treated as negative numbers and so are smaller than positive numbers**.)

4. Let T be the sum of the ranks of the data in sample 1.

Similar to the signed-rank test,

- if the alternative is $\tilde{\mu}_1 < \tilde{\mu}_2$, then the smaller the value of T, the more likely the alternative is true;
- if the alternative is $\tilde{\mu}_1 > \tilde{\mu}_2$, then the larger the value of T, the more likely the alternative is true;
- if the alternative is $\tilde{\mu}_1 \neq \tilde{\mu}_2$, then the smaller or the larger the value of T, the more likely the alternative is true.

How small is too small and how large is too large? It is too small if the test statistics is less than T_L and is too large if greater than T_U, where T_L and T_U are given in Table X on page 694. Again, if n_1 and n_2 are large, we can approximate the distribution of T by the normal distribution:

$$z = \frac{T - \mu_T}{\sigma_T},$$

where the mean and the variance can be found on page 657. Thus, we have

H_0	H_A	(Critical region) Reject H_0 at the α significance level if
$\tilde{\mu}_1 = \tilde{\mu}_2$ or $\tilde{\mu}_1 \geq \tilde{\mu}_2$	$\tilde{\mu}_1 < \tilde{\mu}_2$	$T \leq$ one-tailed T_L ($z \leq -z_\alpha$ if n is large),
$\tilde{\mu}_1 = \tilde{\mu}_2$ or $\tilde{\mu}_1 \leq \tilde{\mu}_2$	$\tilde{\mu}_1 > \tilde{\mu}_2$	$T \geq$ one-tailed T_U ($z \geq z_\alpha$ if n is large),
$\tilde{\mu}_1 = \tilde{\mu}_2$	$\tilde{\mu}_1 \neq \tilde{\mu}_2$	$T \leq$ two-tailed T_L or $T \geq$ two-tailed T_U ($\lvert z \rvert \geq z_{\alpha/2}$ if n is large).

For your interest: The rank-sum test has another formulation. Suppose that the sum of the ranks in sample i is denoted by T_i, $i = 1$ and 2. Then consider

$$U_i = T_i - \frac{n_i(n_i + 1)}{2}.$$

If the null hypothesis is true, then the distributions of U_1 and U_2 are the same and are known, where the critical values are tabulated in another table (not given in the textbook). Under this formulation the rank-sum test is often known as the *Mann–Whitney U-test*.

After one-sample and two-sample cases, we then discussed k independent samples, taken from k populations whose distributions are not normal. The *H-test*, also known as the *Kruskal–Wallis test*, is the nonparametric version of the ANOVA and is used to test whether $\tilde{\mu}_1 = \tilde{\mu}_2 = \cdots = \tilde{\mu}_k$, under the assumption that the k distributions have the same shape. (This assumption is analogous to the equal variance assumption in ANOVA.) Of course, when the distributions are of the same shape, if the medians are the same, then the means are also the same. The testing procedure is as follows.

1. We pool the data together and rank them (in the same manner as in the rank-sum test).
2. Let R_i be the sum of the ranks of data from sample i.
3. Calculate
$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1),$$
where $n = n_1 + n_2 + \cdots + n_k$.

As I explained in my lecture, $\sum_{i=1}^{k} R_i^2$ is a measure of the variation in $R_1, \ldots, R_k$, and we divide each R_i^2 by n_i in order to take the different sample sizes into consideration, so that the larger the value $\sum_{i=1}^{k} \frac{R_i^2}{n_i}$, the bigger the variation in the sample-size adjusted R_i, meaning that the medians are probably not the same. Why do we multiply $\sum_{i=1}^{k} \frac{R_i^2}{n_i}$ by $\frac{12}{n(n+1)}$ and then deduct $3(n+1)$ from it? The reason is purely mathematical: we do not know the distribution of $\sum_{i=1}^{k} \frac{R_i^2}{n_i}$ but mathematicians worked out an approximation for the distribution of H. (In fact, this formula is originally for the case that all data are distinct and so all ranks are distinct, but it also works well unless there are too many tied observations; if you want to see the mathematics for getting $\frac{12}{n(n+1)}$ and $3(n+1)$, make a Google search.) Now, we still have the fact that the larger the value of H is, the more likely the alternative is true, and when each n_i is at least five, H has approximately a χ^2-distribution with degrees of freedom $k - 1$. (We have k terms but loose one degree of freedom because the sum of R_i must be equal to $1 + \ldots + n$.) Thus, we reject the null hypothesis if $H \geq \chi^2_\alpha$. The same as ANOVA, it is a one-sided test, even though we have a two-sided alternative.

Next week we will finish the last chapter by a nonparametric test for zero correlation and a nonparametric test for testing whether the order of the data is random. The latter is a new question that we have not discussed before. Then we will conclude the course by a very brief review of what we have learnt.

Cheers,
Heng Peng