Checkpoint: Hypothesis Testing

G10: Xiangtai Hou, Xiangyu Ji, Yaliang Wang, Zhaocheng Yu

Section 1

Hypothesis:

There is a positive correlation between the increasing rate of number of enterprises and overall amount of international import or export.

Design:

We use linear regression to design our hypothesis for this section. The hypothesis is actually a combination of two hypotheses. Let the annual increasing rate of number of enterprises(AIRE) as the independent variable, and the annual international import(AII) and annual international export(AIE) as the dependent variable for each hypothesis. Then we need to test the slope of regression lines:

1. $AII = B_0 + B_1 * AIRE$
2. $AIE = C_0 + C_1 * AIRE$

Therefore, we got the null hypotheses as following:

1. $H_0 : B_1 = 0$
2. $H_0 : C_1 = 0$

And we use one-trial alternative hypotheses as following:

1. $H_a : B_1 > 0$
2. $H_a : C_1 > 0$

As the typical setup, we choose level of significance as 0.05 for both hypotheses.

Data:

The data is queried from our database with the sql as following:

```
SELECT
       t1.year AS year,
       t1.enterprise_inc_rate AS enterprise_inc_rate,
```

```
        t2.import_total AS import,
        t2.export_total AS export
FROM
        stat_enterprise_annu_incr_rate AS t1,
        trading AS t2
WHERE
        t1.year = t2.year
ORDER BY year ASC;
```

The view “stat_enterprise_annu_incr_rate” and other indirect dependent views are given in the file “views.sql” at the root of our repository on github.

Result:

We use Excel to perform these two linear regression t-tests to the data in the year range of [1999, 2012]. Therefore the number of observations is equal to 14 for both tests.

For the hypothesis 1, we get the t-test score and the p-value in the following table.

	Coefficients	Standard Error	t Stat	P-value	Upper 95%
Intercept	1450610.285	120450.9693	12.04315991	2.32312E-08	1713050.402
AIRE	-60296.19833	93393.95595	-0.645611354	0.265341368	143191.7511

We can find out that the coefficient of the AIRE is actually negative which is not the same sign in the alternative hypothesis 1. We also give the boundary value of coefficient for upper-tail alternative hypothesis 1. Obviously, the coefficient of AIRE is not in the rejection region. Therefore we cannot reject the null hypothesis 1.

For the hypothesis 2, we follow the same process and get the table.

	Coefficients	Standard Error	t Stat	P-value	Upper 95%
Intercept	1981355.847	140366.8404	14.11555494	3.88553E-09	2287188.92
AIRE	-109127.9184	108836.1064	-1.002681207	0.167902706	128005.5864

Unfortunately, the coefficient of the AIRE again drops in the negative region and obviously in the rejection region.

For the both hypotheses, we fail to reject the null hypotheses. Since the null hypotheses are the opposite of the hypothesis we want to test. So that, we got that:

There is **no** positive correlation between the increasing rate of number of enterprises and overall amount of international import or export.

Section 2

Hypothesis:

There is no significant correlation between the employment amount of (code 11) and the overall amount of international export of service.

Design:

We still use linear regression to design for this hypothesis. Let the employment amount of agriculture, forestry, fishing and hunting(code 11) as the independent variable, and the overall amount of international export of service as the dependent variable. Then the slope of regression line is:

$$employee\ amount\ of\ code\ 11 = B_0 + B_1 * overall\ amount\ of\ international\ export\ of\ service$$

The null hypothesis is:

$$H_0 : B_1 = 0$$

The alternative hypothesis is:

$$H_a : B_1 \neq 0$$

The level of significance is 0.05.

Data:

The data is queried from our database with the sql as following:

```
SELECT
       t1.year AS year,
       t1.paid_employees AS paid_employees_11,
       t2.export_service AS export_service
FROM
       labors_pattern AS t1,
       trading AS t2
```

```
WHERE
        t1.year = t2.year
        AND t1.naics_code = 11
ORDER BY year ASC;
```

Result:

As same as the first case, we use Excel to perform this linear regression t-test to the data in the range [1998, 2013]. Therefore the number of observations is equal to 16 for the test.

The score and the p-value we get:

	Coefficients	Standard Error	t Stat	P-value
Intercept	2263448.096	230157.4932	9.834344577	1.14841E-07
X Variable	-10.688521	1.337618616	-7.990708914	1.3881E-06

According to the result we have, we cannot accept the null hypothesis. There is a significant correlation between employment amount of agriculture, forestry, fishing and hunting(code 11) and the overall amount of international export of service.

Section 3

Hypothesis:

There is a positive correlation between the increasing rate of number of enterprises and overall amount of international import or export.

Design:

Our analysis on this section based on the slope of the linear regression line. Considering the volume, here we testify the correlation between the sum of IT, Finance and education industries employers and the international export of service. So we are testing the truthfulness of the linear function:

$$international\ service\ export = B_0 + B_1 * sum(ITemployers, finance\ employers, education\ employers)$$

Null hypothesis:

H_0 : The slope of the regression line B_1 is equal to zero.

H_1 : The slope of the regression line B_1 is *not* equal to zero. Using one tail hypothesis, we hypothesised that $B_1 > 0$

The level of significance: 0.05

Test Statistic: T - Test on regression result.

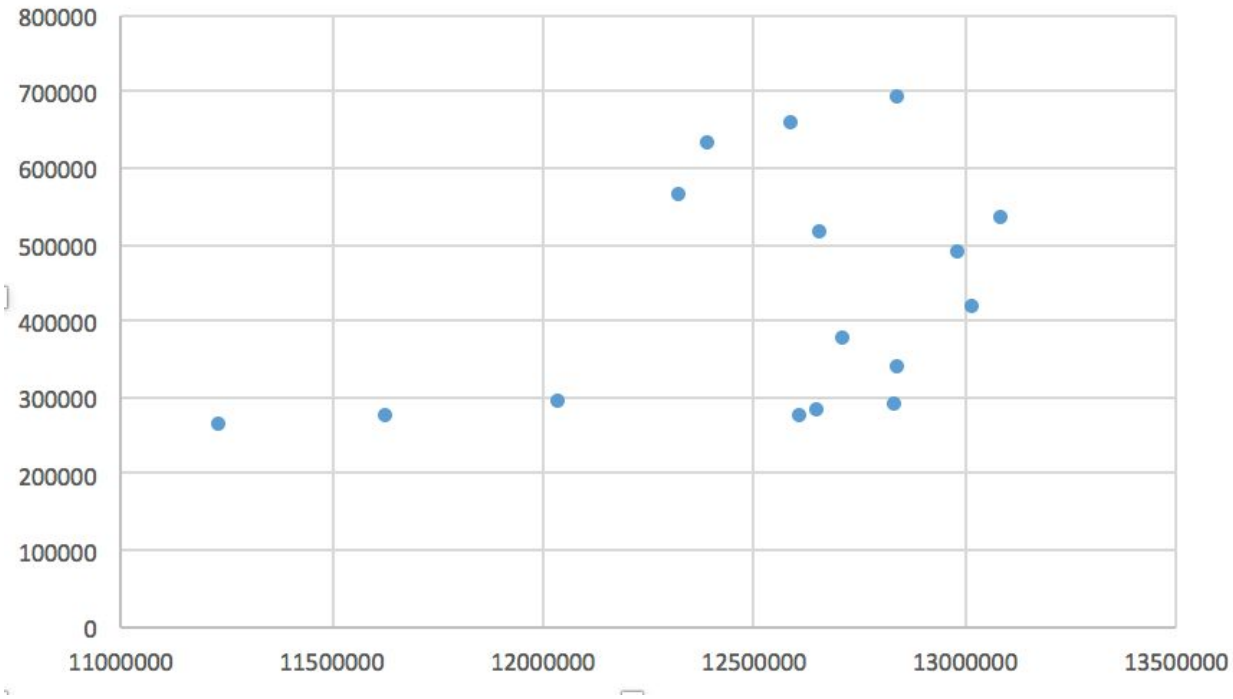
We extract the data as follow SQL function:

```
CREATE OR REPLACE VIEW hypotest_pos_51_52_61_export_service(year,
sum_paid_employees_51_52_61, export_service) AS
SELECT
       year AS year,
       sum(paid_employees) AS sum_paid_employees_51_52_61,
       max(export_service) AS export_service
FROM
       labors_pattern
       NATURAL JOIN trading
WHERE
       naics_code = 51
       OR naics_code = 52
       OR naics_code = 61
GROUP BY year
ORDER BY year ASC;
```

The score and the p-value we get:

	Coefficients	Standard Error	t Stat	P-value
Intercept	-984388.9388	922051.6096	-1.067607202	0.303762098
X Variable	0.112806407	0.073526	1.534238322	0.147255836

As shown on the last table. We may state that the null hypothesis stands and the sum of the paid employers in IT, Finance, and education does not have linear relation with the international service export along the year.



The scatter plot chart we created indicates the same result, the sum of employers and the international service export grows on their own and does not demonstrate any inter-relation.

Section 4

Hypothesis:

There is no significant correlation between increasing rate of number of immigration with international trade of goods and services.

Design:

Linear regression is also our tool here. Our independent variable is increasing immigration rate, and our dependent variable is international balance.Here we pay attention to the the correlation between the international balance(IB) and immigration rate(IR).

$$IB = B_0 + B_1 * IR$$

The null hypothesis is :

$$H_0 : B_1 = 0$$

The alternative hypothesis is :

$$H_a : B_1 > 0$$

The level of the significance is 0.05.

Data:

The data is queried from our database with sql as the following:

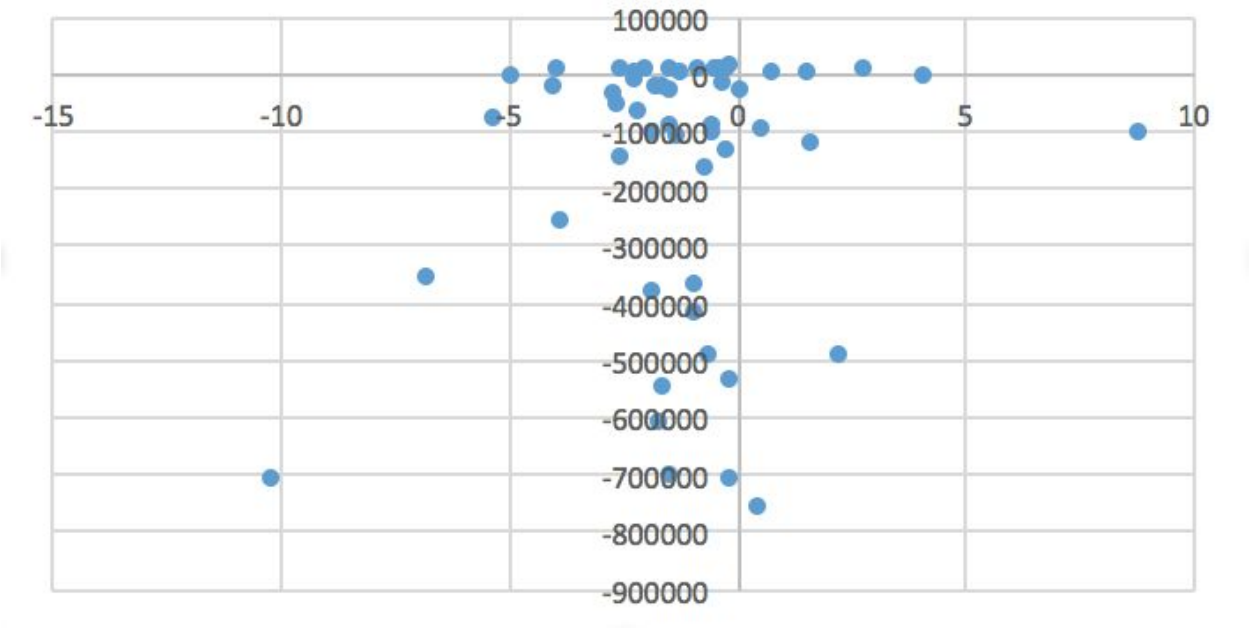
```
SELECT t1.year,
 (t2.variation::numeric - t1.variation::numeric) /      t1.variation::numeric
AS immi_inc_rate,

    t3.balance_total AS balance

FROM
       stat_immi_natu_vari t1,
       stat_immi_natu_vari t2,
        trading t3
WHERE
       t1.year = (t2.year - 1)
       AND t3.year = t2.year
ORDER BY t1.year;
DROP VIEW stat_immi_natu_vari_avg;
DROP VIEW stat_immi_natu_vari;
```

Result:

We use the excel to get the two linear regression and we have already test that the immigration rate has nothing to do with the international trade and services as the following chart shows.



	Coefficients	Standard Error	t Stat	P-value
Intercept	-0.8467	0.456174	-1.8561	0.069222
X variable	1.65E-06	1.59E-06	1.038594	0.303893

As this table show , p-value is larger than our level of significance , which indicates that the null hypothesis is accepted and therefore there is no linear correlation between immigration rate and international base.