Math 11: Discussion 10

Contents

10.1 Upcoming assignments

Assignments	Chapters	Deadlines
Homework	Ch. 21	Wed. 06/03
Quiz	Ch. 21	Thu. 06/04
Homework	Ch. 23	Sat. 06/06
Quiz	Ch. 23	Sun. 06/07
Homework	Ch. 22	Wed. 06/10
Quiz	Ch. 22	Wed. 06/10
Lab 8		Fri. 05/05

Key concepts (not exhaustive):

1. *Regression inference*
2. *Goodness-of-fit* test: **Not on Finals**
3. *Independence* test: **Not on Finals**

10.2 Regression Inference

10.2.1 Testing for Regression Coefficient

Recall the setting of building *linear regression model* which was introduced at the beginning of the quarter. Suppose we have some data points

$$\{(x_i, y_i)\} \quad \text{for } i = 1, 2, \dots, n$$

We modeled this as

$$y_i \;=\; \beta_0 + \beta_1 x_i + \epsilon_i$$

The "best" estimates (in terms of least squares) are

$$\boxed{\hat\beta_0 \;=\; \bar y - \hat\beta_1 \bar x; \qquad \hat\beta_1 \;=\; r \cdot \frac{s_y}{s_x} = \frac{\sum_{i=1}^n (x_i - \bar x)(y_i - \bar y)}{\sum_{i=1}^n (x_i - \bar x)^2}}$$

Motivation: Suppose we have now built the model (ie. calculating $\hat\beta_0$ and $\hat\beta_1$). We want to take a step back and investigate if there was indeed a relationship between x and y. In particular

$$\hat\beta_1 \;\neq\; 0 \quad \Longrightarrow \quad x \text{ and } y \text{ were indeed linearly related}$$

We formulate this as a hypothesis test. For example, we want to test

$$H_0 : \; \hat\beta_1 = 0 \qquad \text{vs.} \qquad H_A : \; \hat\beta_1 \neq 0$$

or H_A other one-sided inequality, depending on the context.

We first define some quantities:

$$s^2 \;=\; \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat y_i)^2$$

$$s_x^2 \;=\; \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar x)^2$$

$$SE\left(\hat\beta_1\right) \;=\; \frac{s}{s_x \sqrt{n-1}}$$

where $SE(\hat\beta_1)$ is the standard error of our estimate $\hat\beta_1$. The appropriate test statistics is

$$T \;=\; \frac{\hat\beta_1 - \beta_1^{(0)}}{SE\left(\hat\beta_1\right)} \;\sim\; T_{n-2}$$

where $\beta_1^{(0)}$ is the current belief for $\hat\beta_1$, here $\beta_1^{(0)} = 0$.

10.2.2 Confidence Interval vs. Prediction Interval

Now, given a new value of x, suppose we have already calculated the predicted new value of $\hat{y}$. Given $(1-\alpha)\%$, we can find the appropriate critical t^* value. We then have:

1. *Confidence Interval.* the Margin of Error is:

$$ME_{CI} = t^* \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

which gives the Confidence Interval:

$$(1-\alpha)\% \text{ CI} = (\hat{\mu}_y \pm ME_{CI})$$

2. *Prediction Interval.* the Margin of Error is:

$$ME_{PI} = t^* \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

which gives the Prediction Interval:

$$(1-\alpha)\% \text{ PI} = (\hat{y} \pm ME_{PI})$$

Remark. The only difference between the ME for Confidence interval and for Prediction interval is that the ME for Prediction interval has a +1 term in the square root, which accounts for the error of any particular observation.

Example 1. *We recall an example from some weeks ago where we modeled the grades as a linear function of the number of hours spent on studying.*

Hours	...	8.5	8.6	8.9	9.1	9.2	...
Grades	...	60	60	68	76	80	...

For your convenience, let H be the number of hours spent, and G the grade, we have:

$$\bar{H} = 10.93; \quad s_H = 2.22; \qquad \bar{G} = 80.8; \quad s_G = 12.74;$$

$$\sum_{i=1}^{n}(H_i - \bar{H})^2 = 93.86; \qquad n = 20$$

And the fitted model has

$$\hat{\beta}_0 \ = \ 25.21; \qquad \hat{\beta}_1 \ = \ 5.09; \qquad s^2 \ = \ 36.27$$

which gives the model of

$$\hat{G} \ = \ 25.21 + 5.09(H)$$

Let us now consider these questions:

(a) *Build a 95% CI for $\hat{\beta}_1$.*

(b) *Assume a confidence level of .05. Test for if $\beta_1 = 0$ or not.*

(c) *Suppose a student studies for 10 hours a week. Build a 95% CI and 95% PI for the predicted grade.*

(d) *Interpret the intervals in part (c).*

Solution:

(a) Since our sample size is 20, $\hat{\beta}_1/SE(\hat{\beta}_1) \sim T_{n-2}$, which has df of 18, At $\alpha = .05$, the critical value is $t^* = 2.1$. The 95% CI is

$$\begin{aligned}
95\%\ CI &= \left(\hat{\beta}_1 \pm t^* SE\left(\hat{\beta}_1\right)\right) \\
&= \left(5.09 \pm 2.1 \cdot \frac{\sqrt{36.27}}{2.22\sqrt{19}}\right) = \big(5.09 \pm 2.1 \cdot (.62)\big) = (5.09 \pm 1.3)
\end{aligned}$$

(b) We first write out the hypothesis:

$$H_0: \ \beta_1 = 0; \qquad H_A: \ \beta_1 \neq 0$$

We now calculate the test statistics:

$$T = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{5.09}{.62} = 8.21$$

We observe that

$$T = 8.21 \notin (5.09 \pm 1.3) \implies \textit{Reject } H_0$$

(c) Given the new value $H = 10$, we have the predicted score:

$$\hat{G} = 25.21 + 5.09(10) = 76.11$$

At $\alpha = .05$, we obtain from t-distribution with df $= 18$ the critical value of $t^* = 2.1$:

- Confidence interval:

$$\begin{aligned}
95\%\ CI &= \left(\hat{\mu}_G \pm t^* s\sqrt{\frac{1}{n} + \frac{(H-\bar{H})^2}{\sum(H_i - \bar{H})^2}}\right) \\
&= \left(76.11 \pm (2.1)(6.02)\sqrt{\frac{1}{20} + \frac{(10-10.93)^2}{93.86}}\right) = (76.11 \pm 3.07)
\end{aligned}$$

- Prediction interval:

$$\begin{aligned}
95\%\ PI &= \left(\hat{G} \pm t^* s\sqrt{1 + \frac{1}{n} + \frac{(H-\bar{H})^2}{\sum(H_i - \bar{H})^2}}\right) \\
&= \left(76.11 \pm (2.1)(6.02)\sqrt{1 + \frac{1}{20} + \frac{(10-10.93)^2}{93.86}}\right) = (76.11 \pm 13.01)
\end{aligned}$$

(d)
- Confidence interval: We are 95% confident that the expected (average) grade given 10 hours of study is between 73.04 and 79.18.
- Prediction interval: We are 95% confident that the predicted grade given 10 hours of study is between 63.1 and 89.12.

10.3 Chi-squared tests

Given observations and expectations, the test statistics is

$$D \;=\; \sum \frac{(Obs - Exp)^2}{Exp}$$

10.3.1 Goodness-of-fit test

Very straightforward where the expected count for group i is

$$\text{Exp}_i = np_i$$

where p_i is the default probability of being in group i. The df is $df = k - 1$, where k is the number of groups.

Example 2. *Let us consider the demographic of student in a Math 11 class. It is commonly believed that among them,* 50% *are first year,* 30% *are second year,* 15% *are third year, and* 5% *are forth year and beyond.*

Suppose the current Math 11 has 240 *students, with the number of students from* 1^{st} *year to* 4^{th} *year in this order as* $132, 75, 27, 6$.

Assume a confidence level of .05*. Determine if this class fits the current belief.*

Solution:

We first write the hypothesis:

$$H_0: \text{ the distribution indeed follows the } 50 - 30 - 15 - 5 \text{ rule} \qquad H_A: \textit{otherwise}$$

For the χ^2 test, to calculate the test statistics, it is easiest to do it in table form:

Group	Year 1	Year 2	Year 3	Year 4
Observation	132	75	27	6
Expected	120	72	36	12
(Obs - Exp)^2	12^2	3^2	9^2	6^2
$\frac{(Obs-Exp)^2}{Exp}$	1.2	.125	2.25	3

Hence the test statistics is

$$D \;=\; \sum \frac{(Obs - Exp)^2}{Exp} \;=\; 6.575$$

Given $n = 4$, we are working with χ^2 distribution with df = 3. Since our χ^2-table does not return p value, we have to instead look-up the threshold χ^2_α. Given $\alpha = .05$, we have $\chi^2_\alpha \;=\; 7.815$. We observe that

$$D \;=\; 6.575 \;<\; 7.815 \;=\; \chi^2_\alpha \quad \Longrightarrow \quad \text{Fail to Reject } H_0$$

We conclude that at the confidence level of $\alpha = .05$ there appears evidence that the demographic of this Math 11 class follows the $50 - 30 - 15 - 5$ rule.

10.3.2 Independence test

In this case, it is a little more involved to calculate the expected counts. Suppose our data is presented with I groups and J categories, put in a form of a I-row, J-column table, then the expected count of group i and category j is

$$\text{Exp}_{i,j} = \frac{R_i C_j}{n}$$

where n is the total sample size, R_i the total number of row i and C_j is the total of column j. The df is $df = (I-1)(J-1)$, ie. the product of $\#(rows) - 1$ and $\#(columns) - 1$.

Example 3. *Ever thought about Med school? (FYI, there is a very chance that your classmates are pre-Med.)*

Let us consider Humanities students and see if the decision to go on to Med school is independent of their majors. Suppose we have carried out a survey and obtained the following:

To Med school?	*History*	*Geograph*	*Literature*
Yes	8	5	7
No	22	35	23

Assume a confidence level of 5%*, determine if it is indeed true that the decision is independent of their majors.*

Solution:
We first write the hypothesis:

$$H_0: \text{ the decision to go to Med school is indeed independent of their majors} \qquad H_A: \textit{otherwise}$$

For the *Independence* test, it is highly recommended that you calculate the total sample size, row and column totals first. The total is $n = 100$.

Different from the previous example, since the data is 2-dimensional, we can't combine *Observations* and *Expectations* on the same table. As such, we calculate the expectation table separately:

To Med school?	History	Geograph	Literature
Yes	6	8	6
No	24	32	24

Hence the test statistics is

$$D = \sum \frac{(Obs - Exp)^2}{Exp} = \frac{2^2}{6} + \frac{2^2}{24} + \cdots + \frac{1^2}{6} + \frac{1^2}{24} = 2.45$$

To calculate df, we note that $I = 2, J = 3$, which gives $df = (2-1)(3-1) = 2$. Given $\alpha = 5\%, \chi^2_\alpha = 5.99$. We observe that

$$D = 2.45 < 5.99 = \chi^2_\alpha \implies \text{Fail to Reject } H_0$$

We conclude that at the confidence level of $\alpha = .05$ there appears evidence that the decision to go on to Med school among the History, Geography and Literature majors is independent of the majors.