UQ, STAT2201, 2017,
Lecture 8 (and part of 9).
Unit 8 – Two Sample Inference.
Unit 9 – Linear Regression.

Unit 8 – Two Sample Inference

Sample $x_1, \ldots, x_{n_1}$ modelled as an i.i.d. sequence of random variables, $X_1, \ldots, X_{n_1}$ and another sample $y_1, \ldots, y_{n_2}$ modelled by an i.i.d. sequence of random variables, $Y_1, \ldots, Y_{n_1}$.

Observations, x_i and y_i (for same i) are not paired. Possible that $n_1 \neq n_2$ (unequal sample sizes).

Model: $X_i \overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2), \qquad Y_i \overset{i.i.d.}{\sim} N(\mu_2, \sigma_2^2).$

Two Variations:

(i) **equal variances:** $\sigma_1^2 = \sigma_2^2 := \sigma^2$.

(ii) **unequal variances:** $\sigma_2^2 \neq \sigma_2^2$.

Focus on **difference in means**,

$$\Delta_\mu := \mu_1 - \mu_2 = E[X_i] - E[Y_i].$$

Ask if

$$\Delta_\mu \ (=, <, >) \ 0$$

i.e. if $\mu_1 \ (=, <, >) \ \mu_2$.

But we can also replace the "0" with other values, e.g. $\mu_1 - \mu_2 = \Delta_0$ for some Δ_0.

A point estimator for Δ_μ is $\overline{X} - \overline{Y}$ (difference in sample means).

The estimate from the data is denoted by $\overline{x} - \overline{y}$ (the difference in the individual sample means), with,

$$\overline{x} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i, \qquad \overline{y} = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i.$$

In the case (ii) of **unequal variances**: Point estimates for σ_1^2 and σ_2^2 are the individual sample variances,

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \overline{x})^2, \qquad s_2^2 = \frac{1}{n_2 - 2} \sum_{i=1}^{n_2} (y_i - \overline{y})^2.$$

In case (i) of **equal variances**, both S_1^2 and S_2^2 estimate σ^2. In this case, a more reliable estimate can be obtained via the **pooled variance estimator**

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

In case (i), under H_0:

$$T = \frac{\overline{X} - \overline{Y} - \Delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \sim \quad t\big(n_1 + n_2 - 2\big).$$

The T test statistic follows a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

In case (ii), under H_0, there is only the approximate distribution,

$$T = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad \sim^{\text{approx}} \quad t(v).$$

where the degrees of freedom are

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_s^2/n_s\right)^2}{n_s - 1}}.$$

If v is not an integer, may round down to the nearest integer (for using a table).

Case (i): two sample T-Tests with equal variance

Model: $X_i \overset{i.i.d.}{\sim} N(\mu_1, \sigma^2), \qquad Y_i \overset{i.i.d.}{\sim} N(\mu_2, \sigma^2).$

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0.$

Test statistic:
$$t = \frac{\bar{x} - \bar{y} - \Delta_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \qquad T = \frac{\bar{X} - \bar{Y} - \Delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Alternative Hypotheses	***P*-value**	**Rejection Criterion for Fixed-Level Tests**
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	$P = 2\left[1 - F_{n_1+n_2-2}\left(\lvert t\rvert\right)\right]$	$t > t_{1-\alpha/2, n_1+n_2-2}$ or $t < t_{\alpha/2, n_1+n_2-2}$
$H_1 : \mu_1 - \mu_2 > \Delta_0$	$P = 1 - F_{n_1+n_2-2}(t)$	$t > t_{1-\alpha, n_1+n_2-2}$
$H_1 : \mu_1 - \mu_2 < \Delta_0$	$P = F_{n_1+n_2-2}(t)$	$t < t_{\alpha, n_1+n_2-2}$

Case (ii): two sample T-Tests with unequal variance

Model: $X_i \overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2), \qquad Y_i \overset{i.i.d.}{\sim} N(\mu_2, \sigma_2^2).$

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0.$

Test statistic:

$$t = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, \qquad T = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

Alternative Hypotheses	***P*-value**	**Rejection Criterion for Fixed-Level Tests**
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	$P = 2\left[1 - F_v\left(\lvert t\rvert\right)\right]$	$t > t_{1-\alpha/2,v}$ or $t < t_{\alpha/2,v}$
$H_1 : \mu_1 - \mu_2 > \Delta_0$	$P = 1 - F_v\left(t\right)$	$t > t_{1-\alpha,v}$
$H_1 : \mu_1 - \mu_2 < \Delta_0$	$P = F_v\left(t\right)$	$t < t_{\alpha,v}$

$1-\alpha$ Confidence Intervals

Case (i) (Equal variances):

$$\bar{x}-\bar{y}-t_{1-\alpha/2,n_1+n_2-2}\ s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \ \leq \ \mu_1-\mu_2 \ \leq \ \bar{x}-\bar{y}+t_{1-\alpha/2,n_1+n_2-2}\ s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Case (ii) (Unequal variances):

$$\bar{x}-\bar{y}-t_{1-\alpha/2,v}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \ \leq \ \mu_1-\mu_2 \ \leq \ \bar{x}-\bar{y}+t_{1-\alpha/2,v}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

Unit 9 – Linear Regression

The collection of statistical tools that are used to model and explore relationships between variables that are related in a nondeterministic manner is called **regression analysis**.

Of key importance is the conditional expectation,

$$E(Y \mid x) = \mu_{Y \mid x} = \beta_0 + \beta_1 x \qquad \text{with} \qquad Y = \beta_0 + \beta_1 x + \epsilon,$$

where x is not random and ϵ is a Normal random variable with $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2$.

Simple Linear Regression is the case where both x and y are scalars, in which case the data is,

$$(x_1, y_1), \ldots, (x_n, y_n).$$

Then given estimates of β_0 and β_1 denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ we have

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad i = 1, 2, \ldots, n,$$

where e_i, are the **residuals** and we can also define the **predicted observation**,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

Ideally it would hold that $y_i = \hat{y}_i$ ($e_i = 0$) and thus **total mean squared error**

$$L := SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2,$$

would be zero.

But in practice, unless $\sigma^2 = 0$ (and all points lie on the same line), we have that $L > 0$.

The standard (classic) way of determining the statistics ($\hat{\beta}_0$, $\hat{\beta}_1$) is by minimisation of L.

The solution, called the **least squares estimators** must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0 \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0 \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Simplifying these two equations yields

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

These are called the **least squares normal equations**. The solution to the normal equations results in the **least squares estimators** $\hat{\beta}_0$ and $\hat{\beta}_1$. Using the sample means, $\overline{x}$ and $\overline{y}$ the estimators are,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad \hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} y_i x_i - \frac{\left(\sum\limits_{i=1}^{n} y_i\right)\left(\sum\limits_{i=1}^{n} x_i\right)}{n}}{\sum\limits_{i=1}^{n} x_i^2 - \frac{\left(\sum\limits_{i=1}^{n} x_i\right)^2}{n}}.$$

The following quantities are also of common use:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x^i\right)\left(\sum_{i=1}^{n} y^i\right)}{n}$$

Hence,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}.$$

Further,

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \qquad SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

The **Analysis of Variance Identity** is

$$\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2+\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$$

or,

$$SS_T=SS_R+SS_E.$$

Also, $SS_R=\hat{\beta}_1 S_{xy}$.

An **Estimator of the Variance**, σ^2 is

$$\hat{\sigma}^2:=MS_E=\frac{SS_E}{n-2}$$

A widely used measure for a regression model is the following ratio of sum of squares, which is often used to judge the adequacy of a regression model:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

$$E\left(\hat{\beta}_0\right) = \beta_0, \qquad V\left(\hat{\beta}_0\right) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right]$$

$$E\left(\hat{\beta}_1\right) = \beta_1, \qquad V\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{XX}}.$$

$$se\left(\hat{\beta}_1\right) = \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}} \quad \text{and} \quad se\left(\hat{\beta}_0\right) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right]}$$

The **Test Statistic for the Slope** is

$$T = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{XX}}}$$

$$H_0 : \beta_1 = \beta_{1,0} \qquad\qquad H_1 : \beta_1 \neq \beta_{1,0}$$

Under H_0 the test statistic T follows a **t - distribution** with "$n-2$ degree of freedom".

An alternative is to use the F statistic as is common in **ANOVA** (Analysis of Variance) – not covered fully in the course.

$$F = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E}.$$

Under H_0 the test statistic F follows an **F - distribution** with "1 degree of freedom in the numerator and $n-2$ degrees of freedom in the denominator".

Analysis of Variance Table for Testing Significance of Regression

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	$\boldsymbol{F_0}$
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	MS_R/MS_E
Error	$SS_E = SS_T - \hat{\beta}_1 S_{xy}$	$n-2$	MS_E	
Total	SS_T	$n-1$		

There are also confidence intervals for β_0 and β_1 as well as prediction intervals for observations. We don't cover these formulas.

To check the regression model assumptions we plot the residuals e_i and check for (i) Normality. (ii) Constant variance. (iii) Independence.

Logistic Regression

Take the response variable, Y_i as a Bernoulli random variable. In this case notice that $E(Y) = P(Y = 1)$.

The **logit response function** has the form

$$E\Big(Y\Big) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp\Big(\beta_0 + \beta_1 x\Big)}.$$

Fitting a logistic regression model to data yields estimates of β_0 and β_1. The following formula is called the **odds**

$$\frac{E\Big(Y\Big)}{1 - E\Big(Y\Big)} = \exp\Big(\beta_0 + \beta_1 x\Big).$$