STAT22000 Autumn 2013 Lecture 27

Yibi Huang

December 2, 2013

10.1 Simple Linear Regression

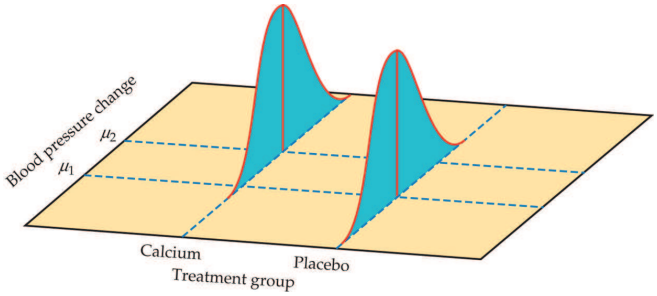
Inference – Where We Are Now

- So far we have discussed inference for one- and two- sample quantitative variables (e.g. t-tests, z-tests)
- Now consider inference where we have a single quantitative response variable (Y) and a single quantitative explanatory (predictor) variable (X)
- In Chapter 2, we covered descriptive tools (scatterplots, least squares regression and correlation)
- Now consider inference for regression

From Two Sample Problems to Many-Sample Problems

Recall the statistical model for comparing the mean responses to two treatments:

Example: randomized clinical trial comparing calcium versus placebo for reducing blood pressure



Statistical Model for Linear Regression (1)

An explanatory variable X can divide the whole population into many **sub-populations**, one for each values of x.

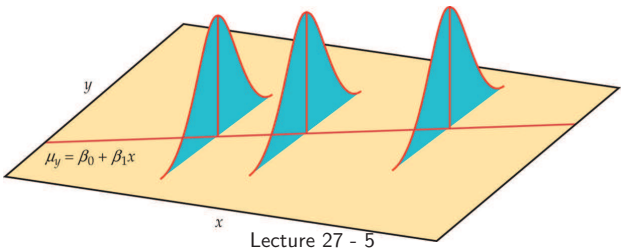
- E.g., say X = father's height, Y = son's height.
  Father-son pairs in which the father is 69 inches tall is a sub-population.
  Father-son pairs in which the father is 72 inches tall is another sub-population.

In different sub-populations, the response variable Y can have different means $\mu_Y(x)$, different SDs $\sigma_Y(x)$, and different distributions.

Statistical Model for Linear Regression (2)

The simple linear regression (SLR) model assumes the following:

1. The means of Y is a linear function of X, i.e.,
$$\mu_Y(x) = \beta_0 + \beta_1 x$$
2. The SD of Y does not change with x, i.e.,
$$\sigma_Y(x) = \sigma \text{ for every } x$$
3. (Optional) Within each subpopulation, the distribution of Y is normal.



Statistical Model for Linear Regression (3)

Equivalently, the SLR model asserts the values of X and Y for individuals in a population are related as follows

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

in which

- the values of β_0 and β_1 are fixed for all individuals, and
- the value of ε, called the **error** or the **noise**, will vary from individual to individual. The distribution of ε's for individuals in a subpopulation follows a normal distribution
$$\varepsilon \sim N(0, \sigma)$$

In the model, the line $y = \beta_0 + \beta_1 x$ is called the **population regression line**.

- β_0 is called the **intercept**
- β_1 is called the **slope**, which describes the change in the mean response (Y) for a single unit increase in X

Data for a Simple Linear Regression Model

Suppose we have a SRS of n individuals from a population. From individual i we observe the response y_i and the explanatory variable x_i:

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$$

The SLR model states that

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Recall in Chapter 2, the least square line of the data above is

$$y = b_0 + b_1 x$$

in which

$$b_1 = r\frac{s_y}{s_x} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}, \quad b_0 = \overline{y} - b_1\overline{x}$$

We can use b_1 to estimate β_1 and b_0 to estimate β_0.

Caution: Sample v.s. Population

Note the population regression line

$$y = \beta_0 + \beta_1 x$$

is *different* from the least square regression line

$$y = b_0 + b_1 x$$

we learned in Chapter 2.

- The latter is merely the least square line for a sample, while the former is the least square line for the entire population.
- The values of b_0 and b_1 will change from sample to sample.

$$b_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}, \quad b_0 = \overline{y} - b_1\overline{x}$$

- We are interested in the population intercept and slope β_0 and β_1, NOT the sample counterparts b_0 and b_1

Estimate of σ^2

To estimate σ, SD of the errors ε_i, recall the **residuals** are the difference between the observed y_i and the predicted y_i:

$$e_i = y_i - \widehat{y}_i = y_i - (b_0 + b_1 x_i)$$

The sample variance of the residuals e_i is used to estimate σ^2:

$$s_e^2 = \frac{\sum e_i^2}{n-2}$$

$s_e = \sqrt{s_e^2}$ is used as the **regression standard error** and it has $n-2$ **degrees of freedom**.

Why $n-2$ degrees of freedom?

- We lose two degrees of freedom because we estimate two parameters, β_0 and β_1.

How Close Is b_1 to β_1?

Recall the slope of the least square line is

$$b_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

Under the SLR model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, replacing y_i in the formula above by $\beta_0 + \beta_1 x_i + \varepsilon_i$, we can show after some algebra that

$$b_1 = \beta_1 + \frac{\sum_i (x_i - \overline{x})\varepsilon_i}{\sum_i (x_i - \overline{x})^2}$$

From the above, one can get the mean, the SD, and the **sampling distribution** of b_1.

- $\mathbb{E}(b_1) = \beta_1$ ..... (b_1 is an **unbiased** estimate of β_1)
- $\mathrm{SD}(b_1) = ?$ ..... (See the next slide)

The Standard Error of b_1

One can show that

$$SD(b_1) = \frac{\sigma}{\sqrt{\sum (x_i - \overline{x})^2}} = \frac{\sigma}{s_x\sqrt{n-1}},$$

where $s_x = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}$ is the sample SD of x_i's.

How to reduce the SD of b_1 (and making b_1 closer to β_1):

- increase the sample size n
- increase the sample SD of x_i's $\Rightarrow$ Better to increase the range of x_i'

As σ is unknown, we estimate it with s_e. The estimated SD of b_1 is called the **standard error (SE)** of b_1

$$SE(b_1) = \frac{s_e}{\sqrt{\sum (x_i - \overline{x})^2}}$$

Confidence Intervals for β_1

The **sampling distribution** of b_1 is normal

$$b_1 \sim N\left(\beta_1, \frac{\sigma}{\sqrt{\sum (x_i - \overline{x})^2}}\right) \quad \Rightarrow \quad z = \frac{b_1 - \beta_1}{\sigma/\sqrt{\sum (x_i - \overline{x})^2}} \sim N(0,1)$$

if

- either the errors ε_i are i.i.d. $N(0, \sigma)$
- or the errors ε_i are independent and the sample size n is large

As σ is unknown, if replaced with s_e, the t-statistic below has a t-distribution with $n-2$ degrees of freedom

$$T = \frac{b_1 - \beta_1}{s_e/\sqrt{\sum (x_i - \overline{x})^2}} = \frac{b_1 - \beta_1}{SE(b_1)} \sim t_{n-2},$$

Hence, the $(1-\alpha)$ **confidence interval for** β_1 is given as

$$b_1 \pm t^* SE(b_1)$$

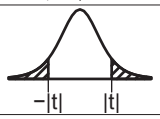
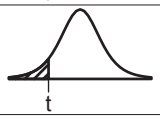
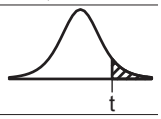
where t^* is the critical value for the $t_{(n-2)}$ distribution at confidence level $1-\alpha$.

Tests for β_1

To **test the hypothesis** $H_0 : \beta_1 = a$, we use the t-statistic

$$t = \frac{b_1 - a}{SE(b_1)} \sim t_{n-2}$$

The p-value can be computed using the t-table based on the alternative hypothesis:

H_a	$\beta_1 \neq a$	$\beta_1 < a$	$\beta_1 > a$
P-value			

Observe that testing $H_0 : \beta_1 = 0$ is equivalent to testing whether x is useful in predicting y linearly.

- This is different from using the correlation coefficient r. It is possible that r is small but β_1 is significantly different from 0.

Inference for the Intercept β_0

Though **the population intercept β_0 is rarely of interest**, all the results for the population slope β_1 have their counterparts for β_0.

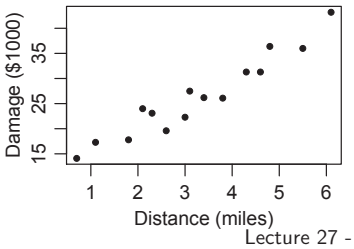
- $b_0 = \beta_0 + \bar{\varepsilon} - \frac{\sum_i \bar{x}(x_i - \bar{x})\varepsilon_i}{\sum_i (x_i - \bar{x})^2}$
- $\mathbb{E}(b_0) = \beta_0$ ..... (b_0 is an **unbiased** estimate of β_0)
- $\mathrm{SD}(b_0) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$
- $\mathrm{SE}(b_0) = s_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$
- The sampling distribution of b_0 (when n is large) is

$$b_0 \sim N\left(\beta_0, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}\right)$$

- $(1-\alpha)$ C.I. for β_0: $b_0 \pm t^* SE(b_0)$
- The test statistic for $H_0 : \beta_0 = a$ is $t = \dfrac{b_0 - a}{SE(b_0)} \sim t_{n-2}$ and the P-value can be computed similarly as for β_1

Example: Fire Damage and Distance to Fire Station

Suppose a fire insurance company wants to relate the amount of fire damage in major residential fires to the distance between the burning house and the nearest fire station. The study is to be conducted in a large suburb of a major city; a sample of 15 recent fires in this suburb is selected. The amount of damage and the distance between the fire and the nearest fire station are recorded in each fire.



Obs.	Distance (mile)	Damage ($1000)
1	0.7	14.1
2	1.1	17.3
3	1.8	17.8
4	2.1	24.0
5	2.3	23.1
6	2.6	19.6
7	3.0	22.3
8	3.1	27.5
9	3.4	26.2
10	3.8	26.1
11	4.3	31.3
12	4.6	31.3
13	4.8	36.4
14	5.5	36.0
15	6.1	43.2


```
> fire = read.table("fire.txt",header=T)
> lm1  = lm(damage ~ dist, data = fire)
> summary(lm1)

Call:
lm(formula = damage ~ dist, data = fire)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4682 -1.4705 -0.1311  1.7915  3.3915

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.2779     1.4203   7.237 6.59e-06 ***
dist          4.9193     0.3927  12.525 1.25e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.316 on 13 degrees of freedom
Multiple R-squared: 0.9235,     Adjusted R-squared: 0.9176
F-statistic: 156.9 on 1 and 13 DF,  p-value: 1.248e-08
```


How to Read R Outputs for Regression? (1)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.2779     1.4203   7.237 6.59e-06 ***
dist          4.9193     0.3927  12.525 1.25e-08 ***
```

- The column "`Estimate`" gives the least square estimate for the intercept b_0 and the slope b_1: $b_0 = 10.2779$, $b_1 = 4.9193$
- The column "`Std. Error`" gives $\mathrm{SE}(b_0)$ and $\mathrm{SE}(b_1)$:

$$SE(b_0) = 1.4203, \quad SE(b_1) = 0.3927$$

So a 95% confidence interval for β_1 is

$$b_1 \pm t^* SE(b_1) = 4.9193 \pm 2.160 \times 0.3927 \approx 4.9193 \pm 0.8482$$

df = 13	···	1.350	1.771	**2.160**	2.282	2.650	3.012	···
	···	80%	90%	**95%**	96%	98%	99%	···
	Confidence level C							

Conclusion: We have 95% confidence that every extra mile from the nearest fire station increases the amount of damage by 4.9193 ± 0.8482 thousands of dollars.


```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.2779     1.4203   7.237 6.59e-06 ***
dist          4.9193     0.3927  12.525 1.25e-08 ***
```

The two columns "`t value`" and "`Pr(>|t|)`" give the t-statistics and two-sided P-values of testing

$$H_0 : \beta_0 = 0 \quad \text{and} \quad H_0 : \beta_1 = 0$$

- Note t-values $b_i/SE(b_i)$ are simply the ratio of the numbers in the "`Estimate`" column and the numbers in the "`Std. Error`" column, e.g.,

$$7.237 = \frac{10.2779}{1.4203}, \quad 12.525 = \frac{4.9193}{0.3927}$$

- Testing $\mathrm{H}_0 : \beta_1 = 0$ is equivalent to testing whether the damage is (linearly) related with the distance to the nearest fire station. The small P-value 1.25×10^{-8} asserts that the distance to the nearest fire indeed has an effect on the amount of damage.

Example: Test for the Slope β_1

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.2779     1.4203   7.237 6.59e-06 ***
dist          4.9193     0.3927  12.525 1.25e-08 ***
```

To test other values of β_1, e.g, $H_0 : \beta_1 = 4$ v.s. $H_1 : \beta_1 > 4$, the t-statistic is

$$t = \frac{b_1 - 3}{SE(b_1)} = \frac{4.9193 - 4}{0.3927} = 2.3409$$

Looking at the t-table for the row with df $= 13$, the one-sided P-value is between 0.01 and 0.02.

	Upper-tail probability p							
df	$\cdots$	0.10	0.05	0.025	**0.02**	**0.01**	0.005	$\cdots$
13	$\cdots$	1.350	1.771	2.160	**2.282**	**2.650**	3.012	$\cdots$

Conclusion: At 5% level, the extra amount of damage for every extra mile from the nearest fire station is significantly higher than \$4000.

How to Read R Outputs for Regression? (2)

```
Residual standard error: 2.316 on 13 degrees of freedom
Multiple R-squared: 0.9235,     Adjusted R-squared: 0.9176
F-statistic: 156.9 on 1 and 13 DF,  p-value: 1.248e-08
```

- `Residual standard error: 2.316 on 13 degrees of freedom`
  This gives the estimate s_e of σ, which is 2.316. As there are 15 points, the degrees of freedom are thus $15 - 2 = 13$.
- `Multiple R-squared` gives r^2, the square of the correlation coefficient.
- `Adjusted R-squared`: Ignore this.
- `F-statistic: 156.9 on 1 and 13 DF, p-value: 1.248e-08`
  See Section 10.2.

Two Similar Problems

- Q_1 Given a specific value of X, say x^*, **estimate** the mean of Y $\mu_y(x^*)$.
- Q_1 Given a specific value of X, say x^*, **predict** Y.

Q_1 is to estimate $\mu_y(x^*) = \beta_0 + \beta_1 x^*$, which is a **parameter**, while Q_2 is to predict $Y = \beta_0 + \beta_1 x^* + \varepsilon = \mu_y(x^*) + \varepsilon$, which is a **random variable**.

For Q_1, $\mu_y(x^*)$ is estimated by

$$\widehat{\mu}_y(x^*) = b_0 + b_1 x^*,$$

while for Q_2, Y is predicted by

$$\widehat{Y} = \widehat{\mu}_y(x^*) + \widehat{\varepsilon} = b_0 + b_1 x^* + 0$$

Though Q_1 and Q_2 end up with identical estimation or prediction, we are less certain (larger standard error) about prediction because of the error term ε.

Confidence Intervals and Prediction Intervals

A $100(1-\alpha)\%$ **confidence interval** for $\mu_y(x^*)$ is given by

$$b_0 + b_1 x^* \pm t^* s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

A $100(1-\alpha)\%$ **prediction interval** for Y is given by

$$b_0 + b_1 x^* \pm t^* s_e \sqrt{\mathbf{1} + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

In both intervals, the t^* is the critical value for the $t_{(n-2)}$ distribution at confidence level $1 - \alpha$.,

- The prediction interval is wider than the confidence interval
- The closer x^* to $\bar{x}$, the narrower the interval.
  This is valid for both the confidence interval, and the prediction interval

Example: Fire Damage and Distance to Fire Station (1)

For houses 2 miles from the nearest fire station, what is the 95% CI for **mean** the amount of damage?

Answer. Recall the fitted regression line is:

$$\text{damage} = 10.28 + 4.92\,\text{distance}$$

1. The estimate is $\hat{\mu}_Y = b_0 + b_1 x^* = 10.28 + 4.92 \times 2 = 20.12$.
2. The critical value t^* is $t_{13,.025} = 2.16$.
3. $s_e = 2.316$, $\bar{x} = 3.28$, $\sum (x_i - \bar{x})^2 = 34.78$
4. The 95% confidence interval for μ_y when $x^* = 2$ is

$$b_0 + b_1 x^* \pm t^* s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$
$$= 20.12 \pm 2.16 \times 2.316 \sqrt{\frac{1}{15} + \frac{(2 - 3.28)^2}{34.78}}$$
$$= 20.12 \pm 1.69 = (18.43,\ 21.80)$$

Example: Fire Damage and Distance to Fire Station (2)

For a house located 2 miles from the nearest fire station, what is the 95% prediction interval for the amount of damage if the house is burned?

Answer.

1. The prediction is still $b_0 + b_1 x^* = 10.28 + 4.92 \times 2 = 20.12$.
2. The critical value t^* is $t_{13,.025} = 2.16$.
3. $s_e = 2.316$, $\bar{x} = 3.28$, $\sum (x_i - \bar{x})^2 = 34.78$
4. The 95% prediction interval for Y when $x^* = 2$ is

$$b_0 + b_1 x^* \pm t^* s_e \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$
$$= 20.12 \pm 2.16 \times 2.316 \sqrt{1 + \frac{1}{15} + \frac{(2 - 3.28)^2}{34.78}}$$
$$= 20.12 \pm 5.28 = (14.84,\ 25.40)$$

Interpretations of the Prediction and Confidence Intervals

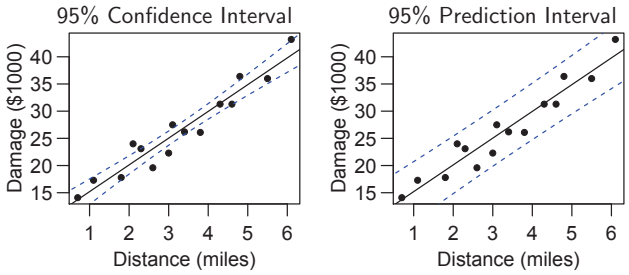
The **prediction interval** (14.84, 25.40) is for the fire damage for a *single* house 2 miles from the nearest fire station

- If one has a house 2 miles from the nearest fire station, once burned, we are 95% confident that the fire damage is between \$14,840 and \$25,400.

The **confidence interval** (18.43, 21.80) is for the **mean** fire damage for all houses 2 miles from the nearest fire station.

- For all houses 2 miles from the nearest fire station, if burned, we are 95% confident that the mean fire damage per house is between \$18,430 and \$21,800.

Example: Fire Damage and Distance to Fire Station



- Observe the 95% prediction interval encloses (almost) all the observations, but the 95% confidence interval does not