STA302/STA1001, Week 3

Mark Ebden, 21 September 2017, morning

With grateful acknowledgment to Alison Gibbs and Becky Lin

Today's class

- The Confidence Interval in Linear Regression
- Hypothesis testing on β_0 and β_1
- Reference: Simon Sheather §§2.2, 2.3

Computing Labs with R installed

Robarts has a Computer Lab open whenever the library itself is open:

- `https://mdl.library.utoronto.ca/technology/computer-lab`
- Monday to Friday 8:30 am to 11 pm
- Saturday 9 am - 10 pm
- Sunday 10 am - 10 pm

There are also four IIT (Information & Instructional Technology) labs:

- In Sidney Smith Hall, Carr Hall, and in Ramsay Wright
- Need Help with an IIT lab? Phone: 416-946-HELP (4357)
- Email: `iit@artsci.utoronto.ca`
- Walk-in: Come to Sidney Smith Room 572 (IIT Office), Monday to Friday, 8:45 am - 5:00 pm

More about the IIT Computer Labs

The four are:

- Sidney Smith Hall room 561 (lower level) (49 seats) - 100 St. George Street: 8:45 am to 7 pm
- Carr Hall room 325 (3rd floor) (30 seats) - 100 St. Joseph Street: 8:45 am to 9 pm
- Ramsay Wright room 107 (20 seats) - 25 Harbord Street: 8:45 am to 9 pm
- Ramsay Wright room 109 (24 seats) - 25 Harbord Street: 8:45 am to 9 pm

Before dropping in, click the links at left here to ensure the room hasn't been booked: `http://lab.chass.utoronto.ca/schedules.php`

More about the IIT Computer Labs

Logging in:

- You must use a valid UTORid and password to log in to lab computers
- If you have trouble logging in, please verify your UTORid credentials at `https://www.utorid.utoronto.ca` (click on the "verify" link under the yellow "Problems with your UTORid?" heading). If your UTORid username and password do not work, reset your password on this page.
- For more help, contact the IIT labs, or reach the Information Commons helpdesk at 416-978-HELP (4357) or `help.desk@utoronto.ca`

More about the IIT Computer Labs

Printing:

- Printing is available in the Sidney Smith and Ramsay Wright labs, but not Carr Hall
- You must have a TCard with sufficient value stored on it. A card reader attached to the print release station will debit the print job cost from your TCard at the time of printing

Saving Data:

- Data is not saved on the lab computers
- Back-up your data frequently, and ensure you have an appropriate storage and/or back-up method for your files (e.g. use a USB key or email materials to yourself)

Towards a Confidence Interval

For a chosen value of x^*,

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

Because $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimates,

$$\mathbb{E}(\hat{y}^*) = \beta_0 + \beta_1 x^*$$

And, using our equations from last Thursday,

formula of variance

$$\begin{aligned}
\text{var}(\hat{y}^*) &= \text{var}(\hat{\beta}_0) + \ (x^*)^2 \text{var}(\hat{\beta}_1 x^*) \ + \ 2x^* \text{cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) \\
&= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] + \ \frac{(x^*)^2 \sigma^2}{S_{xx}} \ - \ \frac{2x^* \sigma^2 \bar{x}}{S_{xx}} \\
&= \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]
\end{aligned}$$

sigma^2 is unknown from data, usually have to estimate

Towards a Confidence Interval

Now bringing in our assumption from Tuesday that the errors are normally distributed:

$$\hat{y}^* \sim \mathcal{N}\left(\beta_0 + \beta_1 x^*, \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]\right)$$

Equivalently we can write this as

$$Z = \frac{\hat{y}^* - (\beta_0 + \beta_1 x^*)}{\sigma\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}} \sim \mathcal{N}(0, 1)$$

standardization

Towards a Confidence Interval

We don't generally know σ^2, but can estimate using the mean square error, S^2, as in question 3 from last week. This changes our Z score into a T score:

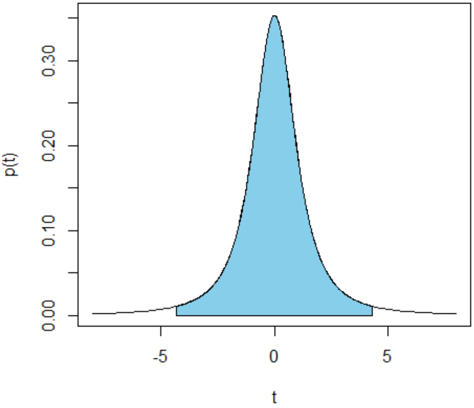
$$T = \frac{\hat{y}^* - (\beta_0 + \beta_1 x^*)}{S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}} \ \sim \ t_{n-2}$$

This distribution tells us that for a given value of x^*:

- Our best estimate for the ordinate, $\hat{y}^*$, is centred on $\beta_0 + \beta_1 x^*$
- Our uncertainty follows a (scaled) t_{n-2} distribution around that point.

A Confidence Interval

What upper- and lower bounds on $\hat{y}^*$ can be expected to encompass the population regression line, i.e. encompass the true $\mathbb{E}(Y^*)$, 95% of the time?



4 points fitting?

The answer is called a 95% confidence interval.

R code to shade a graph

```
c1 = qt(0.025,2) # Left bound of shaded region
c2 = qt(0.975,2)
x0 = 8 # Highest t-score to plot
myseq = seq(c1, c2, 0.01)
cx <- c(c1,myseq,c2) # vector of x-points to outline shaded region
cy <- c(0,dt(myseq,2),0)
curve(dt(x,2),xlim=c(-x0,x0),xlab='t',ylab='p(t)')
polygon(cx,cy,col='skyblue') # connect the dots
```

You don't need to know the ‘curve‘ and ‘polygon‘ commands

A Confidence Interval

Rearranging:

$$\hat{y}^* - (\beta_0 + \beta_1 x^*) \;\sim\; t_{n-2}\; S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

which may seem an unusual way of employing the $\sim$ symbol, but, continuing:

$$\hat{y}^* \sim (\beta_0 + \beta_1 x^*) \;+\; t_{n-2}\; S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

We'll represent the quantile function, $F'(p)$, of the t distribution by $t(1-p, \nu)$, where p is the cumulative probability (between 0 and 1) and ν is the number of degrees of freedom. For our 95% confidence interval, in the lower bound we'll set $p = \alpha/2 = 0.05/2$ and in the upper bound we'll set $p = 1 - \alpha/2 = 0.975$.

A Confidence Interval

Thus we're interested in two cases: $t(\alpha/2, n-2)$ and $t(1-\alpha/2, n-2)$. Equivalently, because the t distribution is symmetric, and because $\alpha = 0.05$, we're interested in $\pm\, t(0.025, n-2)$.

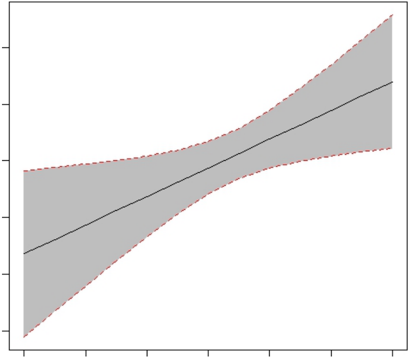
Therefore, the 95% confidence interval is bounded from below by

$$(\beta_0 + \beta_1 x^*) \;\; - \;\; t(0.025, n-2)\; S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

and from above by

$$(\beta_0 + \beta_1 x^*) \;\; + \;\; t(0.025, n-2)\; S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

Plot of Pointwise Confidence intervals



Exercise: Produce this kind of plot for a small data set:

$$\{(2,1),(4,3),(6,4)\}$$

Don't worry about shading, but you should know how to plot the three lines: upper, mean, lower.

What about Confidence Intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$?

Developing on question #3

Our estimator of σ^2 in question #3 from last week, S^2, is the Mean Square Error (MSE).

Our means and variances are expressed in terms of σ, which is unknown, hence the importance of question #3.

For example, the variance of $\hat{\beta}_1$ was found to be

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

However, we use S in place of σ to get:

$$\widehat{\text{var}\left(\hat{\beta}_1\right)} = \frac{S^2}{S_{xx}}$$

Standard error

The square root of this is known as the *standard error* (the estimate of the standard deviation of a parameter) in regression. So,

$$\text{se}\left(\hat{\beta}_1\right) = \sqrt{\frac{S^2}{S_{xx}}}$$

and of course

$$\text{se}\left(\hat{\beta}_0\right) = \sqrt{S^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

You're already used to more simply referring to standard error as the standard deviation of a sampling distribution.

Recap of our guesses about β_1

We've shown how to estimate the mean and variance of $\hat{\beta}_1$.

Then, following the same kind of logic we used in the confidence intervals for $\hat{y}^*$, we can show that:

$$T = \frac{\hat{\beta}_1 - \beta_1}{\mathrm{se}\left(\hat{\beta}_1\right)} \sim t_{n-2}$$

And thus the bounds of the confidence interval are:

$$\hat{\beta}_1 \ \pm \ t(0.025, n-2)\,\mathrm{se}(\hat{\beta}_1)$$

Similarly, for $\hat{\beta}_0$:

$$\hat{\beta}_0 \ \pm \ t(0.025, n-2)\,\mathrm{se}(\hat{\beta}_0)$$

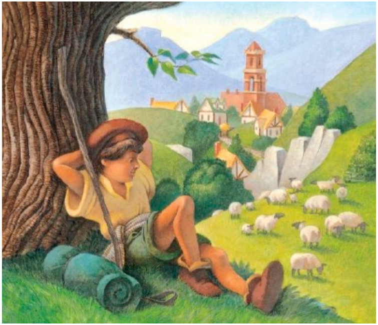
Today's class

- The Confidence Interval in Linear Regression
- **Hypothesis testing on β_0 and β_1**
- Reference: Simon Sheather §§2.2, 2.3

Suppose we want to test whether our random variable β_1 is likely to have a particular mean, β_1^0. For example, perhaps $\beta_1^0 = 0$.

This is an example of the kind of problem on which we can apply a *hypothesis test*.

Statistical hypotheses



The type I error rate is defined as:

$$\begin{aligned}\alpha &= P(\text{type I error})\\ &= P(\text{Reject } H_0 | H_0 \text{ is true})\end{aligned}$$

The type II error rate is defined as:

$$\begin{aligned}\beta &= P(\text{type II error})\\ &= P(\text{Don't reject } H_0 | H_1 \text{ is true})\end{aligned}$$

Decision Theory

Decision	H_0 True	H_0 False
Do not reject H_0	Correct	Type II error
Reject H_0	Type I error	Correct

p-value $= \mathrm{P}(|\text{test stat}| \leq |\text{observed test stat}| \; |H_0 \text{ true})$

$\alpha = \mathrm{P}(\text{type I error } |H_0 \text{ true})$

$\beta = \mathrm{P}(\text{type II error } |H_1 \text{ true})$

$1 - \beta =$ power of test

Statistical hypotheses and power

Power (a.k.a. sensitivity) is defined as:

$$\begin{aligned} \text{power} &= 1 - \beta \\ &= 1 - P(\text{Don't reject } H_0 | H_1 \text{ is true}) \\ &= P(\text{Reject } H_0 | H_1 \text{ is true}) \,. \end{aligned}$$

The probability that a fixed-level α test will reject H_0 when a particular alternative value of the parameter is true is called the *power* of the test to detect that alternative.

The Student's t-test

- You've encountered several statistics which measure central tendency, variability, etc, in an effort to describe/summarize some data
- When a statistic is used in hypothesis testing, it's known as the *test statistic*
- And when this statistic follows a t-distribution under the null hypothesis, our hypothesis test is an example of a ***t*-test**, a.k.a. Student's t-test
- These should usually be two-sided (we prepare for the test statistic's being abnormally high or low) but you do see one-sided tests as well (when the analyst says they have good reason to only check for one or the other of the high/low cases)

Procedure for a t test

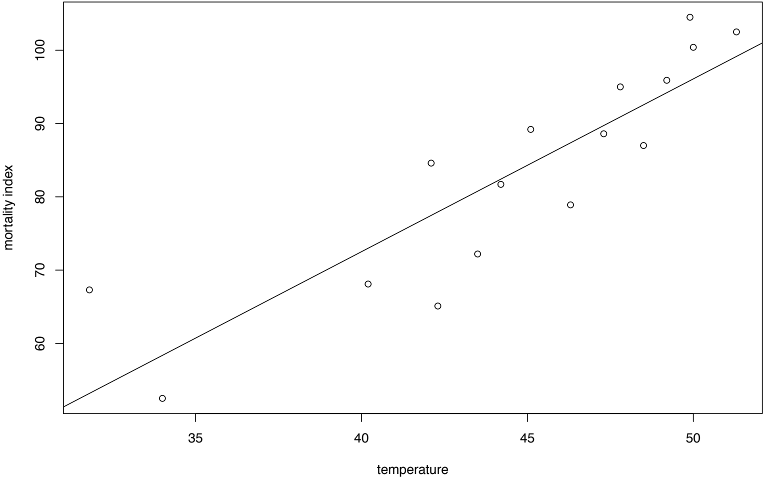
1. Assume the null hypothesis, H_0
2. Calculate your T statistic given H_0
3. How likely is your observed result?

Results of a hypothesis test

Is there any contradiction between H_0 and the observed data?

- The p-value, the probability under the null hypothesis of obtaining a result as extreme or more extreme than the observed result
- A small p-value implies evidence against the null hypothesis
- A large p-value implies no evidence against the null hypothesis
- If the p-value is large does this imply that the null hypothesis is true?
- What does the p-value say about the probability that the null hypothesis is true? Try using Bayes' rule to figure this out

Returning to the temperature/mortality dataset



R has already calculated our p-value

```r
summary(myFit)
```

```
##
## Call:
## lm(formula = M ~ T)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8358  -5.6319   0.4904   4.3981  14.1200
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.7947    15.6719  -1.391    0.186
## T             2.3577     0.3489   6.758  9.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our p-value affects our interpretation

Interpreting b_0 and b_1 when their p-value is low:

- What does the slope mean? For each unit increase in X, Y can be expected to increase by $b_1 X$
- What does the intercept mean? The b_0 has meaning when you are studying very small values of X. It tells you what Y might be when X is around 0

Interpreting b_0 and b_1 when their p-value is high:

- We can say very little in such cases

Extra information: the two-sample t-test

Suppose that there is a clinical trial, in which subjects are randomized to treatments A or B with equal probability. Let μ_A be the mean response in the group receiving drug A and μ_B be the mean response in the group receiving drug B. The null hypothesis is that there is no difference between A and B; the alternative claims there is a clinically meaningful difference between them.

$$H_0 : \mu_A = \mu_B \text{ versus } H_1 : \mu_A \neq \mu_B$$

We want to know if the standard treatment is better than the experimental treatment, or vice versa

The two-sample t-test

Let's assume the patient data are independent random samples from a normal distribution with means μ_A and μ_B but the same variance.

Let's use $\bar{y}_A - \bar{y}_B$ as our test statistic. The distribution is

$$\bar{y}_A - \bar{y}_B \sim \mathcal{N}\left(\mu_A - \mu_B, \sigma^2(1/n_A + 1/n_B)\right).$$

So,

$$\frac{(\bar{y}_A - \bar{y}_b) - \delta_\mu}{\sigma\sqrt{1/n_A + 1/n_B}} \sim \mathcal{N}(0, 1),$$

Next steps

- Try today's plotting exercise
- Try the seven questions at the back of Chapter 2 in Simon Sheather's textbook
- Solutions to HW #1 to be posted very soon – last chance to try them without peaking!
- Next TA office hours: tomorrow morning