Lecture 7: Simple linear regression II

Pratheepa Jeganathan

10/07/2019

Recall

- What is a regression model?
- Descriptive statistics – graphical
- Descriptive statistics – numerical
- Inference about a population mean
- Difference between two population means
- Some tips on R
- Simple linear regression (covariance, correlation, estimation, geometry of least squares)

Outline

- Inference on simple linear regression model
- Example

What do we mean by inference?

- Generally, by inference, we mean "learning something about the relationship between X and Y based on the sample $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$."
- In the simple linear regression model, this often means learning about β_0, β_1.
  - Particular forms of inference are **confidence intervals** or **hypothesis tests**.
- Most of the questions of *inference* in this course can be answered in terms of t-statistics or F-statistics.
- First we will talk about t-statistics, later F-statistics.

Examples of (statistical) hypotheses

- One sample problem: given an independent sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ where $X_i \sim N(\mu, \sigma^2)$, the *null hypothesis* $H_0 : \mu = \mu_0$ says that in fact the population mean is some specified value μ_0.
- Two sample problem: given two independent samples $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$, $\boldsymbol{W} = (W_1, \ldots, W_m)$ where $Z_i \sim N(\mu_1, \sigma^2)$ and $W_i \sim N(\mu_2, \sigma^2)$, the *null hypothesis* $H_0 : \mu_1 = \mu_2$ says that in fact the population means from which the two samples are drawn are identical.

Testing a hypothesis

- We test a null hypothesis, H_0 based on some test statistic T whose distribution is fully known when H_0 is true.
- For example, in the one-sample problem, if $\bar{X}$ is the sample mean of our sample $(X_1, \ldots, X_n)$ and

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

  is the sample variance. Then

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

  has what is called a Student's t distribution with $n-1$ degrees of freedom *when* $H_0 : \mu = \mu_0$ *is true.*
- **When the null hypothesis is not true, it does not have this distribution!**

General form of a (Student's) T statistic

- A t statistic with k degrees of freedom, has a form that becomes easy to recognize after seeing it several times.
- It has two main parts: a numerator and a denominator. The numerator $Z \sim N(0,1)$ while $D \sim \sqrt{\chi^2_k/k}$ that is assumed *independent* of Z.
- The t-statistic has the form

$$T = \frac{Z}{D}.$$

One sample problem revisited

- Above, we used the one sample problem as an example of a t-statistic. Let's be a little more specific.
- Given an independent sample $\boldsymbol{X} = (X_1, \dots, X_n)$ where $X_i \sim N(\mu, \sigma^2)$ we can test $H_0 : \mu = 0$ using a T-statistic.
- We can prove that the random variables

$$\overline{X} \sim N(\mu, \sigma^2/n), \qquad \frac{S_X^2}{\sigma^2} \sim \frac{\chi^2_{n-1}}{n-1}$$

  are independent.
- Therefore, whatever the true μ is

$$\frac{\overline{X} - \mu}{S_X/\sqrt{n}} = \frac{(\overline{X} - \mu)/(\sigma/\sqrt{n})}{S_X/\sigma} \sim t_{n-1}.$$

- Our null hypothesis specifies a particular value for μ, i.e. 0. Therefore, under $H_0 : \mu = 0$ (i.e. assuming that H_0 is true),

$$\overline{X}/(S_X/\sqrt{n}) \sim t_{n-1}.$$

- Another form of the *t*-statistic is

$$T = \frac{\text{estimate of parameter} - \text{true parameter}}{\text{accuracy of the estimate}}.$$

- In more formal terms, we write this as

$$T = \frac{\hat{\theta} - \theta}{SE(\hat{\theta})}.$$

- Note that the denominator is the accuracy of the *estimate* and not the "accuracy" of the true parameter (which is usually assumed fixed, though not for Bayesians).
- The term *SE* or *standard error* will, in this course, usually refer to an estimate of the accuracy of estimator. Therefore, it is the square root of an estimate of the variance of an estimator.

- In our simple linear regression model, a natural (**unobservable**) t-statistic is

$$T = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}.$$

- We've seen how to compute $\hat{\beta}_1$, we never get to see the true β_1, so the only quantity we have anything left to say about is the standard error $SE(\hat{\beta}_1)$.
- How many degrees of freedom would this T have?

Comparison of Student's t to normal distribution

- As the degrees of freedom increases, the population histogram, or density, of the T_k distribution looks more and more like the standard normal distribution usually denoted by $N(0,1)$.

```
rejection_region = function(dens, q_lower,
  q_upper, xval) {
    fig = ggplot(data.frame(x = xval),
      aes(x)) +
      stat_function(fun = dens,
        geom = 'line') +
      stat_function(fun = function(x) {
        ifelse(x > q_upper | x < q_lower,
          dens(x), NA)
        }, geom='area', fill='#CC7777') +
      labs(y='Density', x='T') +
      theme_bw()
}
```

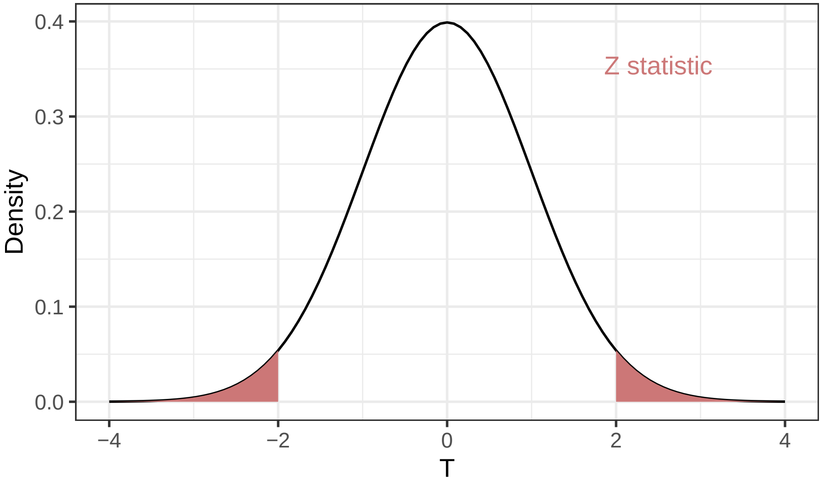
```r
xval = seq(-4, 4, length=101)
q = qnorm(0.975);q
```

```
## [1] 1.959964
```

```r
Z_fig = rejection_region(dnorm,
  -q, q, xval) +
  annotate('text', x = 2.5,
    y = dnorm(2)+0.3,
    label = 'Z statistic',
    color = '#CC7777')
```

- This change in the density has an effect on the *rejection rule* for hypothesis tests based on the T_k distribution.
- For instance, for the standard normal, the 5% rejection rule is to reject if the so-called Z-score is larger than about 2 in absolute value.

```
Z_fig
```



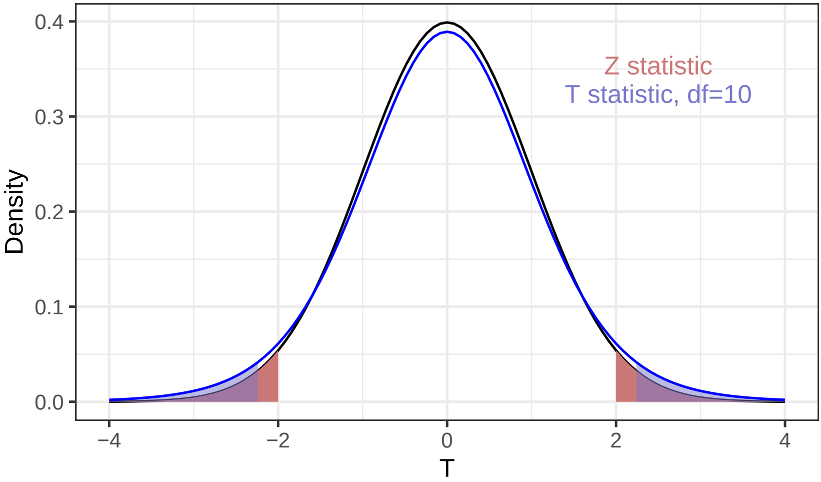
- For the T_{10} distribution, however, this rule must be modified.

```
q10 = qt(0.975, 10); q10
```

```
## [1] 2.228139
```

```r
T_fig = Z_fig +
  stat_function(fun=function(x) {
    ifelse(x > q10 | x < -q10,
      dt(x, 10), NA)
    },
    geom='area',
    fill='#7777CC', alpha=0.5) +
  stat_function(fun=function(x) {
    dt(x, 10)
    },
    color='blue') +
  annotate('text', x=2.5,
    y=dnorm(2)+0.27,
    label='T statistic, df=10',
    color='#7777CC')
```

```
T_fig
```



Confidence interval

- The following are examples of confidence intervals we saw in our review.
  - One sample problem: instead of deciding whether $\mu = 0$, we might want to come up with an (random) interval $[L, U]$ based on the sample $\boldsymbol{X}$ such that the probability the true (nonrandom) μ is contained in $[L, U]$ is at least $1 - \alpha$, i.e. 95%.
  - Two sample problem: find a (random) interval $[L, U]$ based on the samples $\boldsymbol{Z}$ and $\boldsymbol{W}$ such that the probability the true (nonrandom) $\mu_1 - \mu_2$ is contained in $[L, U]$ is at least $1 - \alpha$, i.e. 95%.

Confidence interval for one sample problem

- In the one sample problem, we might be interested in a confidence interval for the unknown μ.
- Given an independent sample $(X_1, \ldots, X_n)$ where $X_i \sim N(\mu, \sigma^2)$ we can construct a $(1 - \alpha) * 100\%$ confidence interval using the numerator and denominator of the t-statistic.

Confidence interval for one sample problem

- Let $q = t_{n-1,(1-\alpha/2)}$

$$
\begin{aligned}
1-\alpha &\le P_\mu\left(-q \le \frac{\mu-\overline{X}}{S_X/\sqrt{n}} \le q\right) \\
&\le P_\mu\left(-q\cdot S_X/\sqrt{n} \le \mu-\overline{X} \le q\cdot S_X/\sqrt{n}\right) \\
&\le P_\mu\left(\overline{X}-q\cdot S_X/\sqrt{n} \le \mu \le \overline{X}+q\cdot S_X/\sqrt{n}\right)
\end{aligned}
$$

- Therefore, the interval $\overline{X} \pm q\cdot S_X/\sqrt{n}$ is a $(1-\alpha)*100\%$ confidence interval for μ.

Inference for β_0 or β_1

- Recall our model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where errors ε_i are independent $N(0, \sigma^2)$.

- In our heights example, we might want to know if there really is a linear association between `Daughter` $= Y$ and `Mother` $= X$.
  - This can be answered with a *hypothesis test* of the null hypothesis $H_0 : \beta_1 = 0$.
  - This assumes the model above is correct, but that $\beta_1 = 0$.
- Alternatively, we might want to have a range of values that we can be fairly certain β_1 lies within.
  - This is a *confidence interval* for β_1.

Setup for inference

- We can show that

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right).$$

- Therefore,

$$\frac{\widehat{\beta}_1 - \beta_1}{\sigma\sqrt{\frac{1}{\sum_{i=1}^n (X_i - \overline{X})^2}}} \sim N(\ 0, 1).$$

- The other quantity we need is the *standard error* or SE of $\hat{\beta}_1$.
  - This is obtained from estimating the variance of $\widehat{\beta}_1$, which, in this case means simply plugging in our estimate of σ, yielding

$$SE(\widehat{\beta}_1) = \widehat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^n (X_i - \overline{X})^2}} \qquad \text{independent of } \widehat{\beta}_1$$

Testing $H_0 : \beta_1 = \beta_1^0$

- Suppose we want to test that β_1 is some pre-specified value, β_1^0 (this is often 0: i.e. is there a linear association)
- Under $H_0 : \beta_1 = \beta_1^0$

$$T = \frac{\widehat{\beta}_1 - \beta_1^0}{\widehat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^n (X_i - \overline{X})^2}}} = \frac{\widehat{\beta}_1 - \beta_1^0}{\frac{\widehat{\sigma}}{\sigma} \cdot \sigma\sqrt{\frac{1}{\sum_{i=1}^n (X_i - \overline{X})^2}}} \sim t_{n-2}.$$

- Reject $H_0 : \beta_1 = \beta_1^0$ if $|T| \geq t_{n-2,1-\alpha/2}$.

Example

Wage example

- Let's perform this test for the wage data.

```
SE.beta.1.hat = (sigma.hat * sqrt(1 /
    sum((wages$education - mean(wages$education))^2)))
Tstat = (beta.1.hat - 0) / SE.beta.1.hat
data.frame(beta.1.hat, SE.beta.1.hat, Tstat)
```

```
##   beta.1.hat SE.beta.1.hat    Tstat
## 1 0.07859951   0.004262471 18.43989
```

Wage example

- Let's look at the output of the `lm` function again.

```
summary(wages.lm)
```

```
##
## Call:
## lm(formula = logwage ~ education, data = wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.78239 -0.25265  0.01636  0.27965  1.61101
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.239194   0.054974   22.54   <2e-16 ***
## education   0.078600   0.004262   18.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
```

Wage example

- We see that R performs this test in the second row of the `Coefficients` table.
- It is clear that wages are correlated with education.

Why reject for large |T|?

- Observing a large $|T|$ is unlikely if $\beta_1 = \beta_1^0$: reasonable to conclude that H_0 is false.
- Common to report p-value:

$$\mathbb{P}(|T_{n-2}| \geq |T_{obs}|) = 2\mathbb{P}(T_{n-2} \geq |T_{obs}|)$$

```
2*(1 - pt(Tstat, wages.lm$df.resid))
```

```
## [1] 0
```

Confidence interval based on Student's t distribution

- Suppose we have a parameter estimate $\widehat{\theta} \sim N(\theta, \sigma_\theta^2)$, and standard error $SE(\widehat{\theta})$ such that

$$\frac{\widehat{\theta} - \theta}{SE(\widehat{\theta})} \sim t_\nu.$$

- We can find a $(1-\alpha) \cdot 100\%$ confidence interval by:

$$\widehat{\theta} \pm SE(\widehat{\theta}) \cdot t_{\nu,1-\alpha/2}.$$

- To prove this, expand the absolute value as we did for the one-sample CI

$$1 - \alpha \leq \mathbb{P}_\theta\left(\left|\frac{\widehat{\theta} - \theta}{SE(\widehat{\theta})}\right| < t_{\nu,1-\alpha/2}\right).$$

Confidence interval for regression parameters

- Applying the above to the parameter β_1 yields a confidence interval of the form

$$\hat{\beta}_1 \pm SE(\hat{\beta}_1) \cdot t_{n-2,1-\alpha/2}.$$

- We will need to compute $SE(\hat{\beta}_1)$. This can be computed using this formula

$$SE(a_0\hat{\beta}_0 + a_1\hat{\beta}_1) = \hat{\sigma}\sqrt{\frac{a_0^2}{n} + \frac{(a_0\overline{X} - a_1)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}$$

with $(a_0, a_1) = (0, 1)$.

Confidence interval for regression parameters

- We also need to find the quantity $t_{n-2,1-\alpha/2}$. This is defined by

$$\mathbb{P}(T_{n-2} \geq t_{n-2,1-\alpha/2}) = \alpha/2.$$

- In *R*, this is computed by the function `qt`.

```
alpha = 0.05
n = nrow(wages); n
```

```
## [1] 2178
```

```
qt(1-0.5*alpha, n-2)
```

```
## [1] 1.961055
```

- Not surprisingly, this is close to that of the normal distribution, which is a Student's t with ∞ for degrees of freedom.

```
qnorm(1 - 0.5*alpha)
```

```
## [1] 1.959964
```

- We will not need to use these explicit formulae all the time, as *R* has some built in functions to compute confidence intervals.

```r
L = beta.1.hat -
  qt(0.975, wages.lm$df.resid) * SE.beta.1.hat
U = beta.1.hat +
  qt(0.975, wages.lm$df.resid) * SE.beta.1.hat
data.frame(L, U)
```

```
##            L          U
## 1 0.07024057 0.08695845
```

```r
confint(wages.lm)
```

```
##                  2.5 %     97.5 %
## (Intercept) 1.13138690 1.34700175
## education   0.07024057 0.08695845
```

Predictions

The estimation of the mean response

- Given $Y = \beta_0 + \beta_1 x + \epsilon$ and the least squares estimators of β_0 and β_1 are $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively.
- For a chosen value x_0, what is the prediction value of the **mean response variable**?
  - We need to estimate $\mathbb{E}\left[Y|x_0\right] = \beta_0 + \beta_1 x_0$.
  - Let $\mathbb{E}\left[Y|x_0\right] = \mu_0$ so $\mu_0 = \beta_0 + \beta_1 x_0$.
  - The best estimator for μ_0 is $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.
- $\mathbb{V}\left[\hat{\mu}_0\right] = \mathbb{V}\left[\hat{\beta}_0 + \hat{\beta}_1 x_0\right]$.
- $\mathrm{SE}\left(\hat{\mu}_0\right) = \hat{\sigma}\sqrt{\dfrac{1}{n} + \dfrac{\left(x_0 - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}$, $\hat{\sigma}^2 = \dfrac{SSE}{n-2}$.
  - The estimation is much more accurate around $\bar{x}$.
- $\hat{\mu}_0 \sim \mathrm{N}\left(\mu_0, \mathbb{V}\left[\hat{\mu}_0\right]\right)$.

Predicting the response of an individual observation

- Given $Y = \beta_0 + \beta_1 x + \epsilon$ and the least squares estimators of β_0 and β_1 are $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively.
- For a chosen value x_0, what is the prediction value of the response variable Y_0? Here Y_0 is a random variable.
  - $Y_0 \sim \mathrm{N}\left(\mathbb{E}\left[Y|x_0\right], \sigma^2\right)$.
  - We took $\mathbb{E}\left[Y|x_0\right] = \mu_0$.
  - The best estimator for Y_0 is $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$
- The predicted response distribution is the predicted distribution of the residuals $Y_0 - \hat{\mu}_0$ at the given point x_0. So the variance is given by $\mathbb{V}\left[Y_0 - \hat{\mu}_0\right] = \mathbb{V}\left[Y_0\right] + \mathbb{V}\left[\hat{\mu}_0\right]$
- $\mathrm{SE}\left(\hat{Y}_0\right) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$.

Comparing SE of predicted response and mean response

- $\mathsf{SE}\left(\hat{Y}_0\right) > \mathsf{SE}\left(\hat{\mu}_0\right)$.
    - Greater uncertainity in predicting one observation than in estimating the mean response.
    - Averaging in the mean response reduces the variability.

Confidence interval for mean response

- We can show that

$$\frac{\hat{\mu}_0 - \mu_0}{\mathsf{SE}\left(\hat{\mu}_0\right)} \sim t_{n-2}.$$

  - $(1-\alpha)\,100\%$ confidence interval for μ_0 is

$$\hat{\mu}_0 \pm t_{n-2,\alpha/2}\mathsf{SE}\left(\hat{\mu}_0\right).$$

  - Confidence limits.

Prediction interval

- We can show that

$$\frac{\hat{Y}_0 - Y_0}{\mathrm{SE}\left(\hat{Y}_0\right)} \sim t_{n-2}.$$

  - $(1-\alpha)\,100\%$ prediction interval for Y_0 is

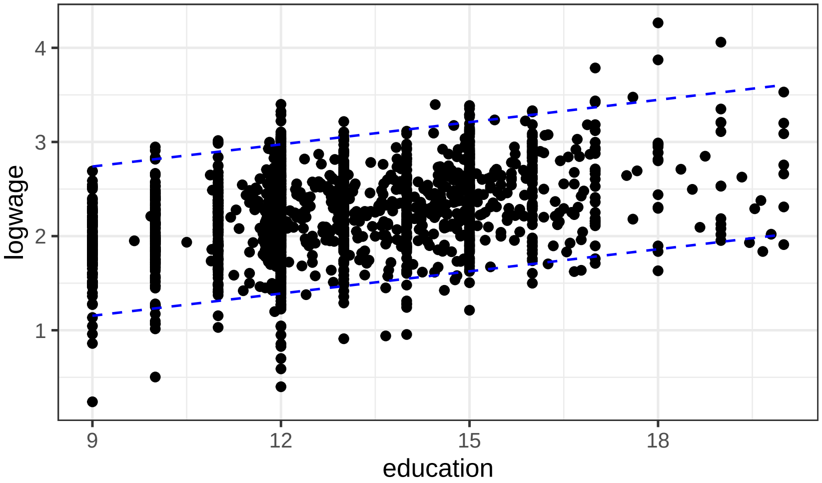
$$\hat{Y}_0 \pm t_{n-2,\alpha/2}\mathrm{SE}\left(\hat{Y}_0\right).$$

  - Prediction limits.

Wages vs. education example

- Construct CI for the mean response for a sequence of x.

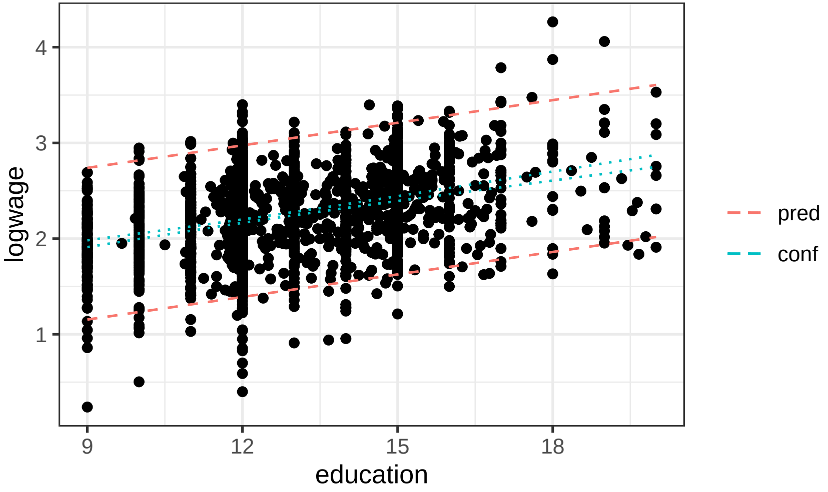
```
url = 'http://www.stanford.edu/class/stats191/data/wage.csv
wages = read.table(url, sep=',',
  header=TRUE)
wages.lm = lm(logwage ~ education,
  data = wages)
xval = data.frame(education = seq(min(wages$education),
  max(wages$education), length.out = 100))
prediction_bands = predict(wages.lm, xval,
  interval = "prediction")
```



4
3
2
1
logwage
9
12
15
18
education

- Construct prediction intervals for the response for a sequence of x.

```
xval = data.frame(education =
    seq(min(wages$education),
      max(wages$education), length.out = 100))
confidence_bands = predict(wages.lm, xval,
  interval = "confidence")
```



logwage
4
3
2
1
9
12
15
18
education
pred
conf

References for this lecture

- Based on the lecture notes of Jonathan Taylor .