Lecture 4 Simple Linear Regression II

Donghyeon Yu

Materials in this lecture notes come from two lecture notes written by Jean Kyung Kim (Inha Univ.) and Pratheepa Jeganathan (Stanford Univ.).

Outline

- Inference on simple linear regression model
- Prediction
- Example

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 *estimator of standard error (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

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

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 $|T| \geq t_{\alpha/2}(n-2)$.

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.

```
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 '.' 0.1 ' ' 1

Residual standard error: 0.4038 on 2176 degrees of freedom
Multiple R-squared:  0.1351,	Adjusted R-squared:  0.1347 
F-statistic:   340 on 1 and 2176 DF,  p-value: < 2.2e-16
```

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 $\widehat{\theta} \sim N(\theta, \sigma^2_\theta)$, and the estimator of standard error $\widehat{SE}(\widehat{\theta})$ such that

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

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

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

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

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

Confidence interval for regression parameters

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

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

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

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

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

Confidence interval for regression parameters

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

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

- 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 $\mathbb{E}[Y|x_0] = \beta_0 + \beta_1 x_0$.
  - Let $\mathbb{E}[Y|x_0] = \mu_0$ so $\mu_0 = \beta_0 + \beta_1 x_0$.
  - The best estimator for μ_0 is $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.
- $\mathbb{V}[\hat{\mu}_0] = \mathbb{V}\left[\hat{\beta}_0 + \hat{\beta}_1 x_0\right]$.
- $\widehat{\mathrm{SE}}(\hat{\mu}_0) = \hat{\sigma}\sqrt{\dfrac{1}{n} + \dfrac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}},\ \hat{\sigma}^2 = \dfrac{SSE}{n-2}$.
  - The estimation is much more accurate around $\bar{x}$.
- $\hat{\mu}_0 \sim \mathrm{N}(\mu_0, \mathbb{V}[\hat{\mu}_0])$.

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 $\mathbb{V}\left[Y_0 - \hat{\mu}_0\right] = \mathbb{V}\left[Y_0\right] + \mathbb{V}\left[\hat{\mu}_0\right]$
- $\widehat{\mathrm{SE}}\left(Y_0 - \hat{\mu}_0\right) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}$.

Comparing SE of predicted response and mean response

- $\widehat{\mathrm{SE}}\,(Y_0 - \hat{\mu}_0) > \mathrm{SE}\,(\hat{\mu}_0)$.
  - 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

$$\frac{\hat{\mu}_0 - \mu_0}{\widehat{\mathrm{SE}}\,(\hat{\mu}_0)} \sim t(n-2).$$

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

$$\hat{\mu}_0 \pm t_{\alpha/2}(n-2)\widehat{\mathrm{SE}}\,(\hat{\mu}_0)\,.$$

  - Confidence limits.

Prediction interval

- We can show that

$$\frac{Y_0 - \hat{\mu}_0}{\widehat{\mathrm{SE}}\,(Y_0 - \hat{\mu}_0)} \sim t(n-2).$$

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

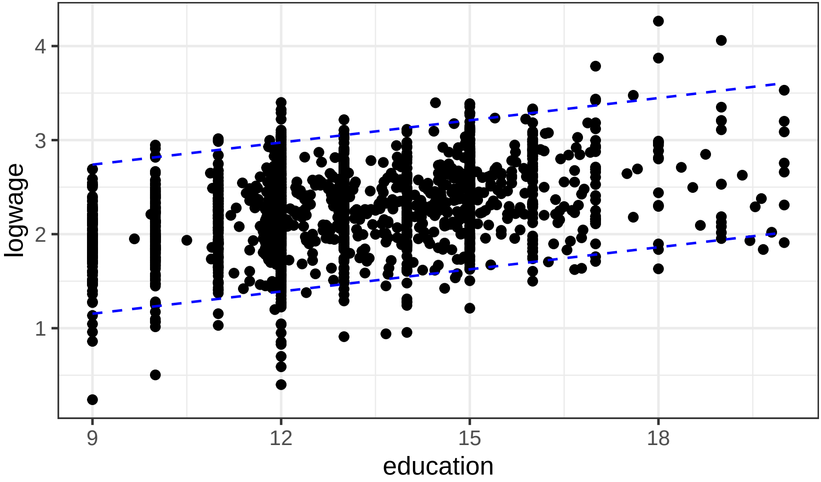
$$\hat{Y}_0 \pm t_{\alpha/2}(n-2)\widehat{\mathrm{SE}}\,(Y_0 - \hat{\mu}_0)\,.$$

  - Prediction limits.

Wages vs. education example

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

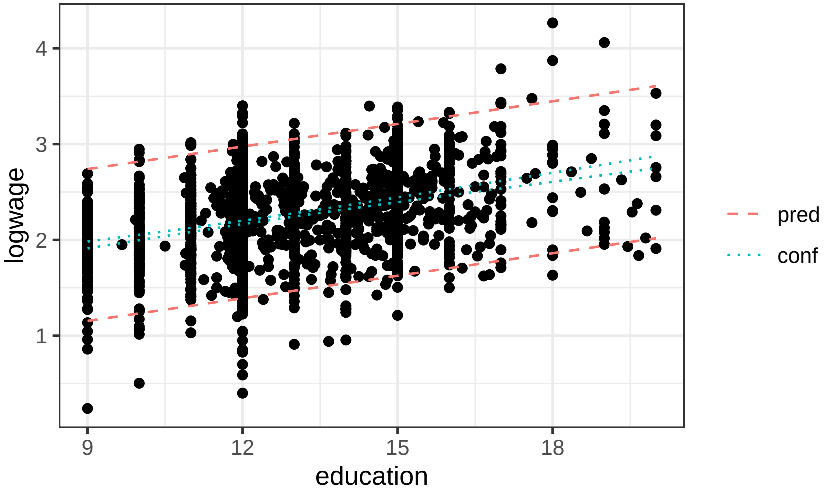
```
fname = './wage.csv'
wages = read.table(fname, sep=',',
        header=TRUE)
wages.lm = lm(logwage ~ education,
  data = wages)
xval = data.frame(education = seq(min(wages$education),
  max(wages$education), length.out = 100))
prediction_bands = predict(wages.lm, xval,
  interval = "prediction")
```



logwage
4
3
2
1
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")
```


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

References for this lecture

- Based on the lecture notes of Pratheepa Jeganathan
- Based on the lecture notes of Jonathan Taylor .