Lecture 4: Simple linear regression

Pratheepa Jeganathan

10/02/2019

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

Outline

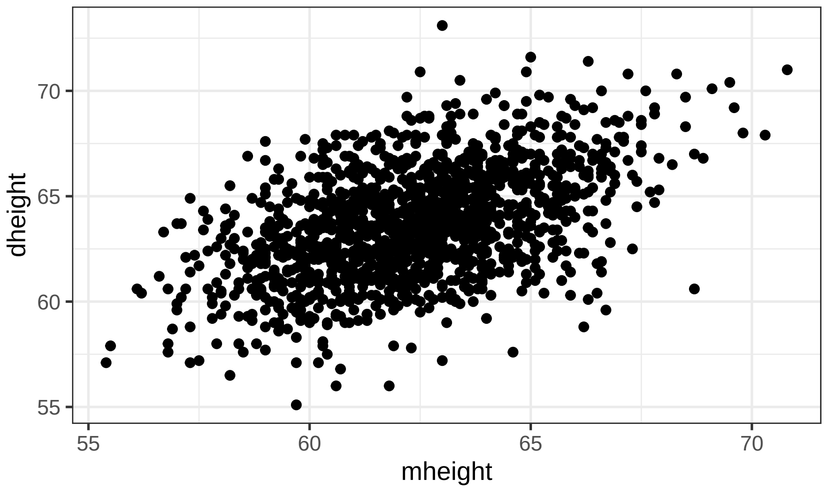
- Correlation
- Introduction to simple linear regression
- Loss functions
- Estimation
- Example

Simple linear regression

- The first type of model, which we will spend a lot of time on, is the *simple linear regression model*.
- One simple way to think of it is via scatter plots.
- Below are heights of mothers and daughters collected by Karl Pearson in the late 19th century.

Scatter plot

```r
library(alr4)
data(Heights)
M = Heights$mheight
D = Heights$dheight
library(ggplot2)
heights_fig = ggplot(Heights,
  aes(mheight, dheight)) +
  geom_point() + theme_bw()
```



dheight
mheight
55
60
65
70
55
60
65
70

Covariance and Correlation

Covariance

- Consider random pairs (X, Y). The strength of the relationship or association between X and Y is of our main interest.
- If X and Y are continuous, the direction of the linear relationship between X and Y can be measured by **covariance**.

$$\text{Cov}(X, Y) = \mathbb{E}(X - \mu_X)(Y - \mu_Y)$$

.

- Given a random sample $(X_1, Y_1), \cdots, (X_n, Y_n)$, the estimator of $\text{Cov}(X, Y)$ is a sample covariance.

$$\hat{\text{Cov}}(X, Y) = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n - 1}$$

Correlation coefficient

- If X and Y are continuous, from random sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ we can use Pearson correlation coefficient or nonparametric Kendall or Spearman statistics to measure the direction and strength of a linear relationship.
- Let X and Y be continuous random variables with mean μ_X, μ_Y and standard deviation σ_X, σ_Y.
- Correlation coefficient is

$$\rho = \frac{\mathbb{E}\left(X - \mu_X\right)\left(Y - \mu_Y\right)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}\left(XY\right) - \mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)}{\sigma_X \sigma_Y}.$$

- If X and Y are independent, $\mathbb{E}\left(XY\right) = \mathbb{E}\left(X\right)\mathbb{E}\left(Y\right)$. Thus, $\rho = 0$, converse is not true.
  - If X and Y are bivariate normal, converse is also true.
- If X and Y are dependent, $\rho \neq 0$.
- Pearson correlation coefficient measures the linear association between X and Y.

Pearson's correlation coefficient

- Sample Pearson's correlation coefficient:

$$\hat{\rho} = r = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}}.$$

- Compute the Pearson's correlations coefficient of heights data

```r
cor(D, M, method = "pearson")
```

```
## [1] 0.4907094
```

- Examine the scatter plot.
- Interpret $\hat{\rho}$.
- $\hat{\rho}$ cannot be used for prediction purposes.

Regression

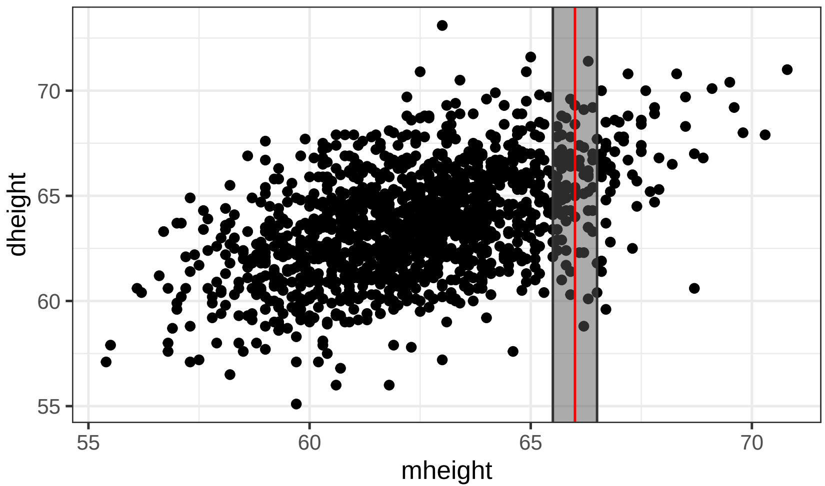
- Ignore mother's height and guessing the daughter's height, we would guess the average height of daughters

```r
mean(D)
```

```
## [1] 63.75105
```

- Can we do better?

- A simple linear regression model fits a line through the above scatter plot in a particular way.
- Specifically, it tries to estimate the height of a new daughter in this population, say D_{new}, whose mother had height M_{new}.
- It does this by considering each slice of the data.
- Here is a slice of the data near $M = 66$, the slice is taken over a window of size 1 inch.



dheight
70
65
60
55
55
60
65
70
mheight

```r
selected_points = (M <= X+.5) & (M >= X-.5)
mean_within_slice = mean(D[selected_points])
mean_within_slice
```

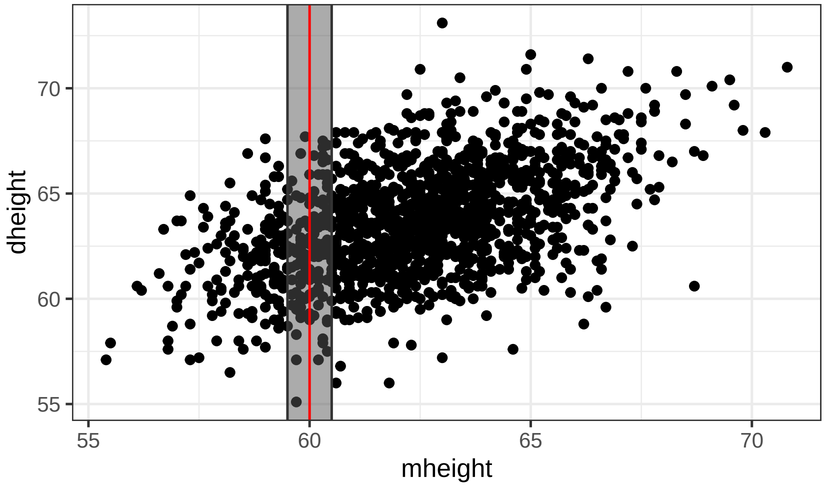
```
## [1] 65.17333
```

- We see that, in our sample, the average height of daughters whose height fell within our slice is about 65.2 inches.

- Of course this height varies by slice. For instance, at 60 inches:

```r
X = 60
selected_points = (M <= X+.5) & (M >= X-.5)
mean_within_slice = mean(D[selected_points])
mean_within_slice
```

```
## [1] 62.42829
```



70
65
60
55
dheight
55
60
65
70
mheight

- The regression model puts a line through this scatter plot in an *optimal* fashion.
- To do this, simple linear regression assumes that the mean in slice M lies on some line

$$\beta_0 + \beta_1 M.$$

- It then chooses (β_0, β_1) based on the data.

```r
parameters.est = lm(D ~ M)$coef
print(parameters.est)
```

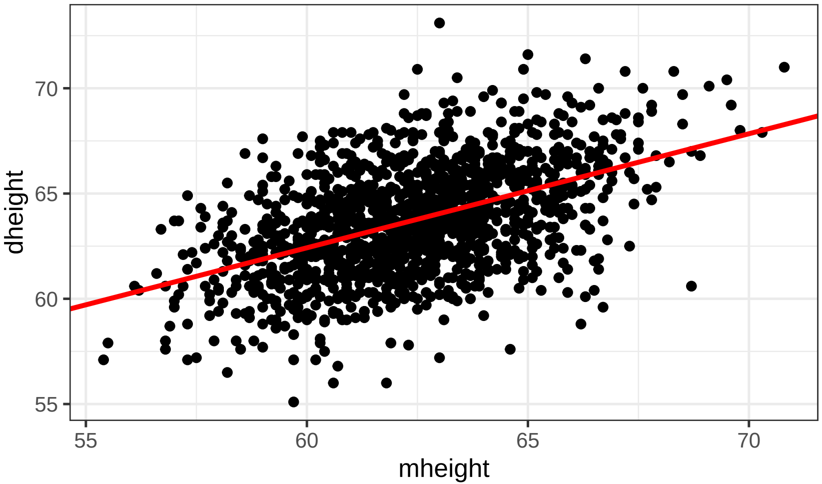
```
## (Intercept)           M
##   29.917437    0.541747
```

```r
intercept = parameters.est[1]; intercept
```

```
## (Intercept)
##    29.91744
```

```r
slope = parameters.est[2]; slope
```

```
##        M
## 0.541747
```


dheight
70
65
60
55
55
60
65
70
mheight

Mathematical formulation

For height of couples data: a mathematical model:

$$\texttt{Daughter} = f(\texttt{Mother}) + \varepsilon,$$

where f gives the average height of the daughter of a mother of height Mother and ε is the random variation within the slice.

Linear regression models

- A *linear* regression model says that the function f is a sum (linear combination) of functions of `Mother`.
- Simple linear regression model:

$$f(\texttt{Mother}) = \beta_0 + \beta_1 \cdot \texttt{Mother}$$

  for some unknown parameter vector (β_0, β_1).
- Could also be a sum (linear combination) of *fixed* functions of `Mother`:

$$f(\texttt{Mother}) = \beta_0 + \beta_1 \cdot \texttt{Mother} + \beta_2 \cdot \texttt{Mother}^2$$

Simple linear regression model

- Let Y_i be the height of the i-th daughter in the sample, X_i be the height of the i-th mother.
- We have a sample of $(X_1, Y_1), \cdots, (X_n, Y_n)$.
- Model:

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{regression equation}} + \underbrace{\varepsilon_i}_{\text{error}},$$

  where ε_i are random error.
  - $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{V}[\epsilon_i] = \sigma^2$
- $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ specifies a *distribution* for the Y's given the X's.
  - i.e. $Y_i | x_i \sim \mathrm{N}(\beta_0 + \beta_1 X_i, \sigma^2)$ is a *statistical model*..

Fitting the model

- We will be using *least squares* regression.
  - This measures the *goodness of fit* of a line by the sum of squared errors, SSE.
- Least squares regression chooses the line that minimizes

$$\mathrm{SSE}(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \cdot X_i)^2.$$

- In principle, we might measure goodness of fit differently by sum of absolute deviation (SAD):

$$\mathrm{SAD}(\beta_0, \beta_1) = \sum_{i=1}^{n} |Y_i - \beta_0 - \beta_1 \cdot X_i|.$$

- For some *loss function* L we might try to minimize

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} L(Y_i - \beta_0 - \beta_1 X_i).$$

Why least squares?

- With least squares, the minimizers have explicit formula
  - not so important with today's computer power – especially when L is convex.
- Resulting formula are *linear* in the outcome Y. This is important for inferential reasons.
  - For only predictive power, this is also not so important.
- If assumptions are correct, then this is *maximum likelihood estimation*.
- Statistical theory tells us the *maximum likelihood estimators (MLEs)* are generally good estimators (consistency, asymptotic normality).

Choice of loss function

- The choice of the function we use to measure goodness of fit, or the *loss* function, has an outcome on what sort of estimates we get out of our procedure.
- For instance, if, instead of fitting a line to a scatter plot, we were estimating a *center* of a distribution, which we denote by μ, then we might consider minimizing several loss functions.

- If we choose the sum of squared errors:

$$\text{SSE}(\mu) = \sum_{i=1}^{n} (Y_i - \mu)^2.$$

  - Then, we know that the minimizer of $\text{SSE}(\mu)$ is the sample mean of Y.
- On the other hand, if we choose the sum of the absolute errors

$$SAD(\mu) = \sum_{i=1}^{n} |Y_i - \mu|.$$

  - Then, the resulting minimizer is the sample median of Y.

- Both of these minimization problems also have *population* versions as well.
- For instance, the population mean minimizes, as a function of μ

$$\mathbb{E}((Y - \mu)^2)$$

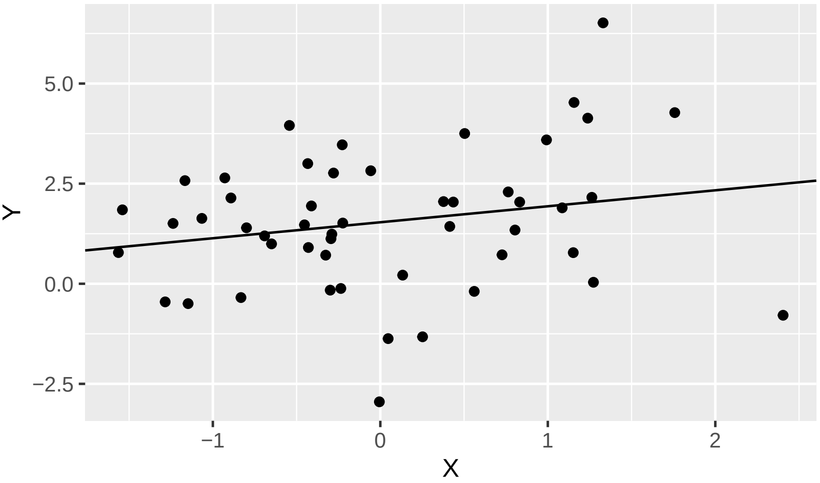
while the population median minimizes

$$\mathbb{E}(|Y - \mu|).$$

Visualizing the loss function

Let's take a random scatter plot of X and Y and view the loss function $L(\beta_0, \beta_1)$.

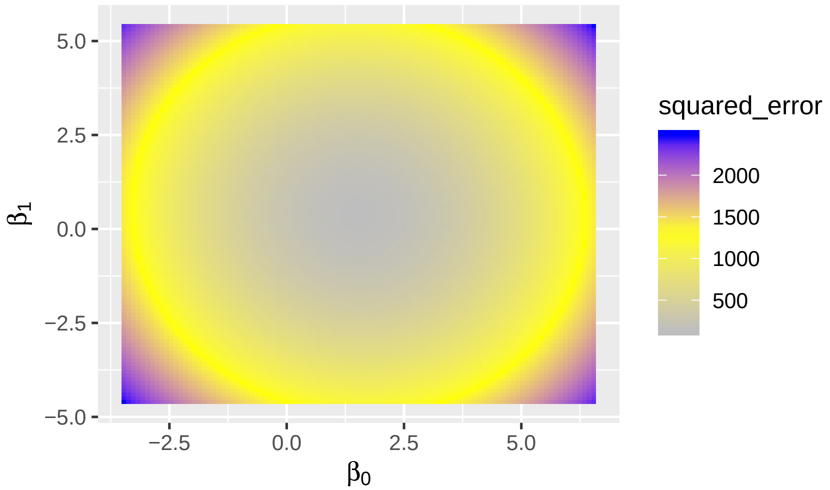
```
X = rnorm(50)
Y = 1.5 + 0.1 * X + rnorm(50) * 2
parameters.est = lm(Y ~ X)$coef
intercept = parameters.est[1]
slope = parameters.est[2]
```



5.0
2.5
0.0
−2.5
Y
−1
0
1
2
X

- Let's plot the *loss* as a function of the parameters.
- Note that the *true* intercept is 1.5 while the *true* slope is 0.1.

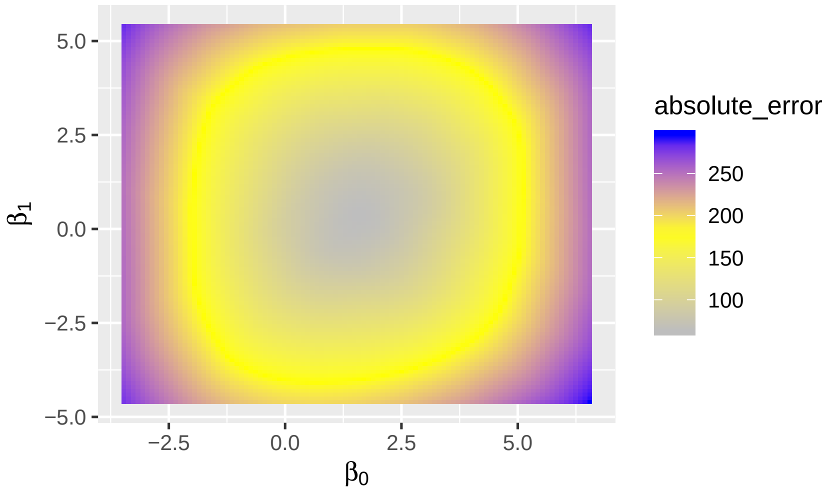
```
grid_intercept = seq(intercept - 5,
  intercept + 5, length = 100)
grid_slope = seq(slope - 5,
  slope + 5, length = 100)
loss_data = expand.grid(intercept_ = grid_intercept,
  slope_=grid_slope)

loss_data$squared_error = numeric(nrow(loss_data))
for (i in 1:nrow(loss_data)) {
    loss_data$squared_error[i] =
      sum((Y - X * loss_data$slope_[i] -
          loss_data$intercept_[i])^2)
}
```



squared_error
2000
1500
1000
500
5.0
2.5
0.0
−2.5
−5.0
−2.5
0.0
2.5
5.0
β_0
β_1

Let's contrast this with the sum of absolute errors.

```
loss_data$absolute_error = numeric(nrow(loss_data))
for (i in 1:nrow(loss_data)) {
    loss_data$absolute_error[i] =
      sum(abs(Y - X * loss_data$slope_[i] -
          loss_data$intercept_[i]))
}
absolute_error_fig = (ggplot(loss_data,
  aes(intercept_, slope_,
    fill = absolute_error)) +
    geom_raster() +
    scale_fill_gradientn(colours = c("gray",
      "yellow", "blue")))
```


absolute_error
250
200
150
100
5.0
2.5
0.0
−2.5
−5.0
−2.5
0.0
2.5
5.0
β_1
β_0

Geometry of least squares

- The following picture depicts the geometry involved in least squares regression.

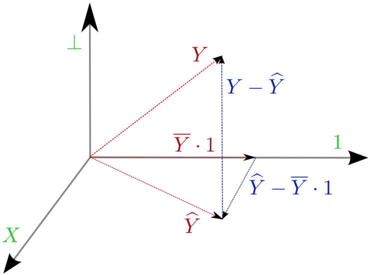


Figure 1: Source - Jonathan Taylor

- It requires some imagination but the picture should be thought as representing vectors in n-dimensional space, I where n is the number of points in the scatter plot.
- In our height data, $n = 1375$. The bottom two axes should be thought of as 2-dimensional, while the axis marked "$\perp$" should be thought of as $(n - 2)$ dimensional, or, 1373 in this case.

```r
dim(Heights)
```

```
## [1] 1375    2
```

Least squares estimators

- There are explicit formula for the least squares estimators, i.e. the minimizers of the error sum of squares.
- For the slope, $\hat{\beta}_1$, it can be shown that

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\widehat{Cov}(X, Y)}{\widehat{Var}(X)}.$$

- Knowing the slope estimate, the intercept estimate can be found easily:

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \cdot \overline{X}.$$

Important lengths

- We can describe an observation as
$$\underbrace{y_i}_{\text{Observed}} = \underbrace{\hat{y}_i}_{\text{Fit}} + \underbrace{(y_i - \hat{y}_i)}_{\text{Deviation from fit}}.$$
- Subtract $\bar{y}$ from both sides
$$\underbrace{y_i - \bar{y}}_{\text{Deviation from mean}} = \underbrace{\hat{y}_i - \bar{y}}_{\text{Deviation due to fit}} + \underbrace{(y_i - \hat{y}_i)}_{\text{Residual}}.$$

- The (squared) lengths of the vectors $\left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right)$, $\left(\bar{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}\right)$, $\left(\boldsymbol{Y} - \bar{\boldsymbol{Y}}\right)$ are important quantities in what follows.

$$
\begin{aligned}
\text{SSE} &= \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2 \\
\text{SSR} &= \sum_{i=1}^{n}(\overline{Y} - \widehat{Y}_i)^2 = \sum_{i=1}^{n}(\overline{Y} - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2 \\
\text{SST} &= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = SSE + SSR \\
R^2 &= \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \widehat{Cor}(\boldsymbol{X}, \boldsymbol{Y})^2.
\end{aligned}
$$

Coefficient of determination

An important summary of the fit is the ratio

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

which measures *how much variability in* Y is explained by X.

Estimate of σ^2

- There is one final quantity needed to estimate all of our parameters in our (statistical) model.
- This is σ^2, the variance of the random variation within each slice (the regression model assumes this variance is constant within each slice).
- The estimate most commonly used is

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \frac{\text{SSE}}{n-2} = \text{MSE}$$

- Above, note the practice of replacing the quantity $\mathrm{SSE}(\hat{\beta}_0, \hat{\beta}_1)$, i.e. the minimum of this function, with just SSE.
- The term MSE above refers to mean squared error: a sum of squares divided by what we call its *degrees of freedom*.
  - The degrees of freedom of *SSE*, the *error sum of squares* is therefore $n - 2$.
  - Remember this $n - 2$ corresponded to $\perp$ in the picture above.

- Using some statistical calculations that we will not dwell on, if our simple linear regression model is correct, then we can see that

$$\frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{\chi^2_{n-2}}{n-2}$$

  where the right hand side denotes a *chi-squared* distribution with $n-2$ degrees of freedom.
- (Note: our estimate of σ^2 *is not* the maximum likelihood estimate.)

Example wages vs. education

- In this example, we'll look at the output of *lm* for the wage data and verify that some of the equations we present for the least squares solutions agree with the output.
- The data was compiled from a study in econometrics Learning about Heterogeneity in Returns to Schooling.

```r
url = 'http://www.stanford.edu/class/stats191/data/wage.cs
wages = read.table(url, sep=',',
  header=TRUE)
print(head(wages))
```

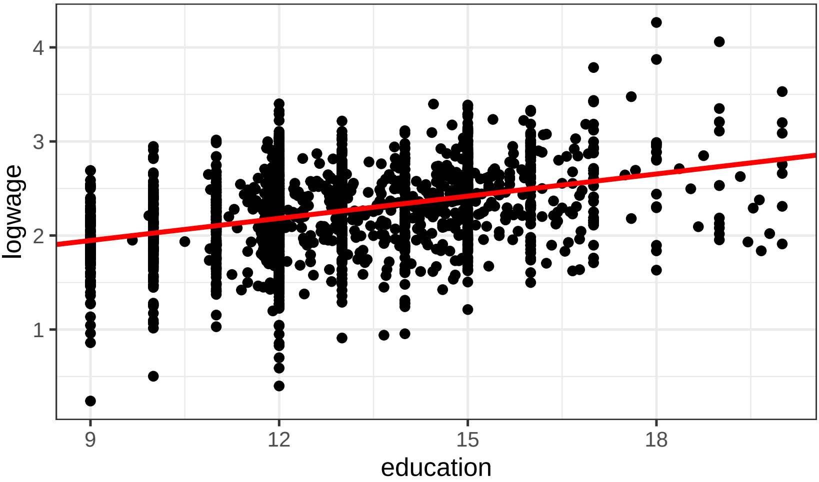
```
##   education  logwage
## 1  16.75000 2.845000
## 2  15.00000 2.446667
## 3  10.00000 1.560000
## 4  12.66667 2.099167
## 5  15.00000 2.490000
## 6  15.00000 2.330833
```

- Let's fit the linear regression model.

```r
wages.lm = lm(logwage ~ education,
  data = wages)
print(wages.lm)
```

```
##
## Call:
## lm(formula = logwage ~ education, data = wages)
##
## Coefficients:
## (Intercept)    education
##      1.2392       0.0786
```

- As in the mother-daughter data, we might want to plot the data and add the regression line.



- Compute the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ using the formula

```r
beta.1.hat = cov(wages$education,
  wages$logwage) / var(wages$education)
beta.0.hat = mean(wages$logwage) -
  beta.1.hat * mean(wages$education)
```

- Compare the above with the `lm` output

```r
print(c(beta.0.hat, beta.1.hat))
```

```
## [1] 1.23919433 0.07859951
```

```r
print(coef(wages.lm))
```

```
## (Intercept)   education
##  1.23919433  0.07859951
```

- Compute $\hat{\sigma}^2$ using the formula

```
sigma.hat = sqrt(sum(resid(wages.lm)^2) /
    wages.lm$df.resid)
c(sigma.hat, sqrt(sum((wages$logwage -
    predict(wages.lm))^2) / wages.lm$df.resid))
```

```
## [1] 0.4037828 0.4037828
```

- The summary from *R* also contains this estimate of σ: (Residual standard error)

```r
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.
##
## Residual standard error: 0.4038 on 2176 degrees of free
## Multiple R-squared:  0.1351, Adjusted R-squared:  0.134
## F-statistic:   340 on 1 and 2176 DF,  p-value: < 2.2e-1
```

References for this lecture

- Based on the lecture notes of Jonathan Taylor .
- Lecture notes of Stats 205

Chatterjee, Samprit, and Ali S Hadi. 2015. *Regression Analysis by Example*. John Wiley & Sons.