Lecture 3

August 28, 2018

Last class

- We studied simple linear regression where there is only one explanatory variable x. We denote the data by $(x_1, y_1), \ldots, (x_n, y_n)$. We assume that these are independent with the relation of y_i to x_i expressed as

$$\mathbb{E}(y_i|x_i) = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n.$$

Last class

- We studied simple linear regression where there is only one explanatory variable x. We denote the data by $(x_1, y_1), \ldots, (x_n, y_n)$. We assume that these are independent with the relation of y_i to x_i expressed as

$$\mathbb{E}(y_i|x_i) = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n.$$

- We can alternatively write the simple linear regression model as

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad \text{where } \mathbb{E}(e_i|x_i) = 0.$$

Last class

- We studied simple linear regression where there is only one explanatory variable x. We denote the data by $(x_1, y_1), \ldots, (x_n, y_n)$. We assume that these are independent with the relation of y_i to x_i expressed as

$$\mathbb{E}(y_i|x_i) = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n.$$

- We can alternatively write the simple linear regression model as

$$y_i = \beta_0 + \beta_1 x_i + e_i, \text{ where } \mathbb{E}(e_i|x_i) = 0.$$

- The parameters β_0 and β_1 are estimated by least squares and this gives the estimates:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2,}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Last class

- One can alternatively write $\hat{\beta}_1$ as $r = s_y/s_x$.

Last class

- One can alternatively write $\hat{\beta}_1$ as $r = s_y/s_x$.
- The OLS regression line is $y = \hat{\beta}_0 + \hat{\beta}_1 x$ and this line can also be written as

$$\frac{y - \bar{y}}{s_y} = r\left(\frac{x - \bar{x}}{s_x}\right).$$

Last class

- One can alternatively write $\hat{\beta}_1$ as $r = s_y/s_x$.
- The OLS regression line is $y = \hat{\beta}_0 + \hat{\beta}_1 x$ and this line can also be written as

$$\frac{y - \bar{y}}{s_y} = r\left(\frac{x - \bar{x}}{s_x}\right).$$

- The regression line always passes through the point $(\bar{x}, \bar{y})$.

Last class

beta1_hat = r(sy/sx)

- One can alternatively write $\hat{\beta}_1$ as $r = s_y/s_x$.
- The OLS regression line is $y = \hat{\beta}_0 + \hat{\beta}_1 x$ and this line can also be written as

$$\frac{y - \bar{y}}{s_y} = r\left(\frac{x - \bar{x}}{s_x}\right).$$

- The regression line always passes through the point $(\bar{x}, \bar{y})$.
- We showed that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased (both conditionally on $X = \{x_1, \ldots, x_n\}$ and unconditionally) estimators of β_1 and β_0 respectively.

Variances of the OLS estimators

- Let us now compute the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$.

Variances of the OLS estimators

- Let us now compute the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Recall the formula for $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Variances of the OLS estimators

- Let us now compute the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Recall the formula for $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- This is a complicated formula to take variance of.

Variances of the OLS estimators

- Let us now compute the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Recall the formula for $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- This is a complicated formula to take variance of.
- If we condition on X, then this is a linear function of $\{y_1, \ldots, y_n\}$ and things are much more tractable. We shall therefore work conditional on X.

Variances of the OLS estimators

- Let us now compute the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Recall the formula for $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

- This is a complicated formula to take variance of.
- If we condition on X, then this is a linear function of $\{y_1, \ldots, y_n\}$ and things are much more tractable. We shall therefore work conditional on X.
- The conditional variance of a random variable Z conditioned on X will be denoted by $\text{var}_X(Z)$.

Variances of the OLS estimators

- In order to go ahead and compute such conditional variances of $\hat{\beta}_0$ and $\hat{\beta}_1$, we need to assume something about how $\{y_1, \ldots, y_n\}$ vary conditional on X.

Variances of the OLS estimators

- In order to go ahead and compute such conditional variances of $\hat{\beta}_0$ and $\hat{\beta}_1$, we need to assume something about how $\{y_1, \ldots, y_n\}$ vary conditional on X.
- The simplest assumption is that: $\text{var}_X(y_i) = \sigma^2$ for each i.

Variances of the OLS estimators

- In order to go ahead and compute such conditional variances of $\hat{\beta}_0$ and $\hat{\beta}_1$, we need to assume something about how $\{y_1, \ldots, y_n\}$ vary conditional on X.
- The simplest assumption is that: $\mathrm{var}_X(y_i) = \sigma^2$ for each i.
- Because we have already assumed that $(x_1, y_1), \ldots, (x_n, y_n)$ are independent, this variance assumption is equivalent to assuming that

$$\mathrm{var}(e_i|x_i) = \mathbb{E}(y_i^2|x_i) = \sigma^2.$$

Variances of the OLS estimators

- In order to go ahead and compute such conditional variances of $\hat{\beta}_0$ and $\hat{\beta}_1$, we need to assume something about how $\{y_1, \ldots, y_n\}$ vary conditional on X.
- The simplest assumption is that: $\mathrm{var}_X(y_i) = \sigma^2$ for each i.
- Because we have already assumed that $(x_1, y_1), \ldots, (x_n, y_n)$ are independent, this variance assumption is equivalent to assuming that

$$\mathrm{var}(e_i|x_i) = \mathbb{E}(y_i^2|x_i) = \sigma^2.$$

- With this assumption, we have

Variances of the OLS estimators

- In order to go ahead and compute such conditional variances of $\hat{\beta}_0$ and $\hat{\beta}_1$, we need to assume something about how $\{y_1, \ldots, y_n\}$ vary conditional on X.
- The simplest assumption is that: $\text{var}_X(y_i) = \sigma^2$ for each i.
- Because we have already assumed that $(x_1, y_1), \ldots, (x_n, y_n)$ are independent, this variance assumption is equivalent to assuming that

$$\text{var}(e_i | x_i) = \mathbb{E}(y_i^2 | x_i) = \sigma^2.$$

- With this assumption, we have

$$\text{var}_X\left(\hat{\beta}_1\right) = \text{var}_X\left(\frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) =$$

Variances of the OLS estimators

- In order to go ahead and compute such conditional variances of $\hat{\beta}_0$ and $\hat{\beta}_1$, we need to assume something about how $\{y_1, \ldots, y_n\}$ vary conditional on X.
- The simplest assumption is that: $\text{var}_X(y_i) = \sigma^2$ for each i.
- Because we have already assumed that $(x_1, y_1), \ldots, (x_n, y_n)$ are independent, this variance assumption is equivalent to assuming that

$$\text{var}(e_i|x_i) = \mathbb{E}(y_i^2|x_i) = \sigma^2.$$

- With this assumption, we have

$$\text{var}_X\left(\hat{\beta}_1\right) = \text{var}_X\left(\frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \text{var}_X(y_i)}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2}$$

Variances of the OLS estimators

- And so
$$\text{var}_X\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Variances of the OLS estimators

- And so

$$\text{var}_X\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Variances of the OLS estimators

- And so

$$\mathrm{var}_X\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

- The formula for $\mathrm{var}_X(\hat{\beta}_0)$ is

$$\mathrm{var}_X\left(\hat{\beta}_0\right) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}$$

  and the proof of this is left as an exercise.

Variances of the OLS estimators

- And so
$$\mathrm{var}_X\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$
- The formula for $\mathrm{var}_X(\hat{\beta}_0)$ is
$$\mathrm{var}_X\left(\hat{\beta}_0\right) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}$$
and the proof of this is left as an exercise.
- Observe that the variance of β_1 increases with σ^2 and decreases with $\sum_{i=1}^n (x_i - \bar{x})^2$. Does this make sense?

Variances of the OLS estimators

- And so
$$\mathrm{var}_X\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$
- The formula for $\mathrm{var}_X(\hat{\beta}_0)$ is
$$\mathrm{var}_X\left(\hat{\beta}_0\right) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}$$
and the proof of this is left as an exercise.
- Observe that the variance of β_1 increases with σ^2 and decreases with $\sum_{i=1}^n (x_i - \bar{x})^2$. Does this make sense?
- In practice, the formula $\mathrm{var}_X(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2$ does not really tell us how good an estimator $\hat{\beta}_1$ is.

Variances of the OLS estimators

- And so
$$\text{var}_X\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$
- The formula for $\text{var}_X(\hat{\beta}_0)$ is
$$\text{var}_X\left(\hat{\beta}_0\right) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}$$
and the proof of this is left as an exercise.
- Observe that the variance of β_1 increases with σ^2 and decreases with $\sum_{i=1}^n (x_i - \bar{x})^2$. Does this make sense?
- In practice, the formula $\text{var}_X(\hat{\beta}_1) = \sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2$ does not really tell us how good an estimator $\hat{\beta}_1$ is.
- This is because σ^2 is unknown.

Variances of the OLS estimators

Example

- Suppose $Z, \ldots, Z_n$ are i.i.d observations with unknown mean μ and unknown variance σ^2. How does one estimate σ^2?

Variances of the OLS estimators
Example

- Suppose $Z, \dots, Z_n$ are i.i.d observations with unknown mean μ and unknown variance σ^2. How does one estimate σ^2? If μ is known, then it is obvious that

$$\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mu)^2$$

is an unbiased estimator of σ^2.

Variances of the OLS estimators
Example

- Suppose $Z, \ldots, Z_n$ are i.i.d observations with unknown mean μ and unknown variance σ^2. How does one estimate σ^2? If μ is known, then it is obvious that

$$\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mu)^2$$

is an unbiased estimator of σ^2.

- Because μ is unknown, a natural fix is to replace μ above by its estimator $\bar{Z} := (Z_1 + \ldots + Z_n)/n$.

Variances of the OLS estimators
Example

- Suppose $Z, \ldots, Z_n$ are i.i.d observations with unknown mean μ and unknown variance σ^2. How does one estimate σ^2? If μ is known, then it is obvious that

$$\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mu)^2$$

is an unbiased estimator of σ^2.

- Because μ is unknown, a natural fix is to replace μ above by its estimator $\bar{Z} := (Z_1 + \ldots + Z_n)/n$. This gives the following estimator for σ^2:

$$\hat{\sigma}_1 := \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})^2$$

Variances of the OLS estimators

- It turns out that this estimator is slightly biased (its bias will be negative which means that it tends to underestimate σ^2).

Variances of the OLS estimators

- It turns out that this estimator is slightly biased (its bias will be negative which means that it tends to underestimate σ^2).
- It is not too hard to prove that

$$\mathbb{E}(\hat{\sigma}_1) := \frac{n-1}{n+1}\sigma^2.$$

Variances of the OLS estimators

- It turns out that this estimator is slightly biased (its bias will be negative which means that it tends to underestimate σ^2).
- It is not too hard to prove that

$$\mathbb{E}(\hat{\sigma}_1) := \frac{n-1}{n+1}\sigma^2.$$

- Therefore a natural way to obtain an unbiased estimator of σ^2 is to take

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Variances of the OLS estimators

- Now let us come to the problem of estimating σ^2 in simple linear regression.

Variances of the OLS estimators

- Now let us come to the problem of estimating σ^2 in simple linear regression. The most natural unbiased estimator is

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Variances of the OLS estimators

- Now let us come to the problem of estimating σ^2 in simple linear regression. The most natural unbiased estimator is

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

- But this is infeasible because β_0 and β_1 are unknown.

Variances of the OLS estimators

- Now let us come to the problem of estimating σ^2 in simple linear regression. The most natural unbiased estimator is

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

- But this is infeasible because β_0 and β_1 are unknown. The natural fix is to replace them with the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. As in the Example above, this results in an estimate of σ^2 that is negatively biased.

Variances of the OLS estimators

- Now let us come to the problem of estimating σ^2 in simple linear regression. The most natural unbiased estimator is

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

- But this is infeasible because β_0 and β_1 are unknown. The natural fix is to replace them with the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. As in the Example above, this results in an estimate of σ^2 that is negatively biased. To correct this, one uses the estimate

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

Variances of the OLS estimators

- Now let us come to the problem of estimating σ^2 in simple linear regression. The most natural unbiased estimator is

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

- But this is infeasible because β_0 and β_1 are unknown. The natural fix is to replace them with the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. As in the Example above, this results in an estimate of σ^2 that is negatively biased. To correct this, one uses the estimate

$$\hat{\sigma}^2 \;=\; \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

- The division by $n-2$ ensures that this estimator is unbiased. Proving that this is unbiased is a little tricky (but not difficult). We will avoid doing this now and later prove it more generally for multiple regression.

- Now for some terminology. We denote $y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ by $\hat{e}_i$ and call it the ith residual.

- Now for some terminology. We denote $y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ by $\hat{e}_i$ and call it the ith residual. The square-root of the estimate of $\hat{\sigma}$ is called the residual standard error i.e.,

$$\hat{\sigma} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i^2}.$$

- Now for some terminology. We denote $y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ by $\hat{e}_i$ and call it the ith residual. The square-root of the estimate of $\hat{\sigma}$ is called the residual standard error i.e.,

$$\hat{\sigma} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i^2}.$$

- The residuals have the following two properties:

- Now for some terminology. We denote $y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ by $\hat{e}_i$ and call it the ith residual. The square-root of the estimate of $\hat{\sigma}$ is called the residual standard error i.e.,

$$\hat{\sigma} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i^2}.$$

- The residuals have the following two properties:
  1. $\sum_{i=1}^{n} \hat{e}_i = 0$.

- Now for some terminology. We denote $y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ by $\hat{e}_i$ and call it the ith residual. The square-root of the estimate of $\hat{\sigma}$ is called the residual standard error i.e.,

$$\hat{\sigma} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i^2}.$$

- The residuals have the following two properties:
  1. $\sum_{i=1}^{n} \hat{e}_i = 0$.
  2. $\sum_{i=1}^{n} \hat{e}_i x_i = 0$.

- Now for some terminology. We denote $y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ by $\hat{e}_i$ and call it the ith residual. The square-root of the estimate of $\hat{\sigma}$ is called the residual standard error i.e.,

$$\hat{\sigma} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} \hat{e}_i^2}.$$

- The residuals have the following two properties:
  1. $\sum_{i=1}^{n} \hat{e}_i = 0$.
  2. $\sum_{i=1}^{n} \hat{e}_i x_i = 0$.
- Because of the above two identities, the n residuals $\hat{e}_1, \dots, \hat{e}_n$ have essentially $n-2$ degrees of freedom.

- Now for some terminology. We denote $y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ by $\hat{e}_i$ and call it the ith residual. The square-root of the estimate of $\hat{\sigma}$ is called the residual standard error i.e.,

$$\hat{\sigma} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} \hat{e}_i^2}.$$

- The residuals have the following two properties:
  1. $\sum_{i=1}^{n} \hat{e}_i = 0$.
  2. $\sum_{i=1}^{n} \hat{e}_i x_i = 0$.
- Because of the above two identities, the n residuals $\hat{e}_1, \dots, \hat{e}_n$ have essentially $n-2$ degrees of freedom.
- Using $\hat{\sigma}$, we can estimate the variance of $\hat{\beta}_1$ by

$$\frac{\hat{\sigma^2}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

- The square root of the above quantity is called the standard error (more precisely estimated standard error) of $\hat{\beta}_1$

$$\text{s.e}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

- The square root of the above quantity is called the standard error (more precisely estimated standard error) of $\hat{\beta}_1$

$$\text{s.e}(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

- Similarly

$$\text{s.e}(\hat{\beta}_0) = \frac{\hat{\sigma}\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{n\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

- The square root of the above quantity is called the standard error (more precisely estimated standard error) of $\hat{\beta}_1$

$$\text{s.e}(\hat{\beta}_1) \; = \; \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

- Similarly

$$\text{s.e}(\hat{\beta}_0) \; = \; \frac{\hat{\sigma}\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{n\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

- There are other topics in simple linear regression such as confidence intervals. We shall do these more generally in multiple linear regression.

Multiple Regression

- There is a response variable y and p explanatory variables $x_1, \ldots, x_p$. The goal is understand the relationship between y and $x_1, \ldots, x_p$.

Multiple Regression

- There is a response variable y and p explanatory variables $x_1, \ldots, x_p$. The goal is understand the relationship between y and $x_1, \ldots, x_p$.
- There are n subjects and data is collected on the variables from these subjects.

Multiple Regression

- There is a response variable y and p explanatory variables $x_1, \ldots, x_p$. The goal is understand the relationship between y and $x_1, \ldots, x_p$.
- There are n subjects and data is collected on the variables from these subjects.
- Data on the response variable is $y_1, \ldots, y_n$. Data on the jth explanatory variable x_j is $x_{1j}, x_{2j}, \ldots, x_{nj}$.

Multiple Regression

- There is a response variable y and p explanatory variables $x_1, \ldots, x_p$. The goal is understand the relationship between y and $x_1, \ldots, x_p$.
- There are n subjects and data is collected on the variables from these subjects.
- Data on the response variable is $y_1, \ldots, y_n$. Data on the jth explanatory variable x_j is $x_{1j}, x_{2j}, \ldots, x_{nj}$.
- The data for the ith subject is $(y_i, x_{i1}, \ldots, x_{ip})$. The linear regression model assumes that these are independent with

$$\mathbb{E}(y_i | x_{i1}, \ldots, x_{ip}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{i1p}, \quad \text{for } i = 1, \ldots, n.$$

Multiple Regression

- There is a response variable y and p explanatory variables $x_1, \ldots, x_p$. The goal is understand the relationship between y and $x_1, \ldots, x_p$.
- There are n subjects and data is collected on the variables from these subjects.
- Data on the response variable is $y_1, \ldots, y_n$. Data on the jth explanatory variable x_j is $x_{1j}, x_{2j}, \ldots, x_{nj}$.
- The data for the ith subject is $(y_i, x_{i1}, \ldots, x_{ip})$. The linear regression model assumes that these are independent with

  $$\mathbb{E}(y_i|x_{i1}, \ldots, x_{ip}) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{i1p}, \quad \text{for } i = 1, \ldots, n.$$

- Equivalently, we can write this as

  $$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + e_i, \quad \text{where } \mathbb{E}(e_i|x_{i1}, \ldots, x_{ip}) = 0.$$

Estimation of $\beta_0, \beta_1, \ldots, \beta_p$

- As before, we use least squares: Estimate $\beta_0, \beta_1, \ldots, \beta_p$ by the minimizers of

$$s(b) = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_{i1} - \ldots b_p x_{ip})^2 \tag{1}$$

over $b_0, \ldots, b_p$.

Estimation of $\beta_0, \beta_1, \ldots, \beta_p$

- As before, we use least squares: Estimate $\beta_0, \beta_1, \ldots, \beta_p$ by the minimizers of

$$s(b) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \ldots b_p x_{ip})^2 \tag{1}$$

  over $b_0, \ldots, b_p$.
- Like in the simple regression case, we can try to take partial derivatives with respect to $b_0, \ldots, b_p$, set them equal to zero and then solve the resulting equations. Doing this directly using the expression (1) is painful. Things are greatly simplified if one uses matrices and vectors.

Estimation of $\beta_0, \beta_1, \ldots, \beta_p$

- As before, we use least squares: Estimate $\beta_0, \beta_1, \ldots, \beta_p$ by the minimizers of

$$s(b) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \ldots b_p x_{ip})^2 \tag{1}$$

  over $b_0, \ldots, b_p$.
- Like in the simple regression case, we can try to take partial derivatives with respect to $b_0, \ldots, b_p$, set them equal to zero and then solve the resulting equations. Doing this directly using the expression (1) is painful. Things are greatly simplified if one uses matrices and vectors.
- Data on the response variable, $y_1, \ldots, y_n$, are represented by the column vector $Y = (y_1, \ldots, y_n)^T$ (the T here stands for transpose).

- Data on the jth explanatory variable x_j is $x_{1j}, x_{2j}, \ldots, x_{nj}$. The data on all the explanatory variables is represented by the $n \times (p+1)$ matrix X whose first column consists of ones, second column has the values $x_{11}, x_{21}, \ldots, x_{n1}$ corresponding to the first explanatory variable, third column has values for the second explanatory variable and so on.

- Data on the jth explanatory variable x_j is $x_{1j}, x_{2j}, \ldots, x_{nj}$. The data on all the explanatory variables is represented by the $n \times (p+1)$ matrix X whose first column consists of ones, second column has the values $x_{11}, x_{21}, \ldots, x_{n1}$ corresponding to the first explanatory variable, third column has values for the second explanatory variable and so on.
- With this notation, the sum of squares in (1) can be rewritten as
$$s(b) = \|Y - Xb\|_2^2.$$
where the norm of the vector x is defined as $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$. Note the equality $\|x\|_2^2 = x^T x$.

- Data on the jth explanatory variable x_j is $x_{1j}, x_{2j}, \ldots, x_{nj}$. The data on all the explanatory variables is represented by the $n \times (p+1)$ matrix X whose first column consists of ones, second column has the values $x_{11}, x_{21}, \ldots, x_{n1}$ corresponding to the first explanatory variable, third column has values for the second explanatory variable and so on.
- With this notation, the sum of squares in (1) can be rewritten as

$$s(b) = \|Y - Xb\|_2^2.$$

where the norm of the vector x is defined as $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$. Note the equality $\|x\|_2^2 = x^T x$.
- Using this, we can write

$$s(b) = (Y - Xb)^T (y - Xb) = Y^T Y - 2b^T Xy + b^T X^T Xb.$$

- This can be minimized via calculus. Take partial derivatives with respect to
  i for $i = 0, 1, \ldots, p$ and equate them to 0. It is easy to check that

$$\nabla S(b) = 2X^T Xb - 2X^T y$$

- This can be minimized via calculus. Take partial derivatives with respect to
  i for $i = 0, 1, \ldots, p$ and equate them to 0. It is easy to check that
$$\nabla S(b) = 2X^T X b - 2X^T y$$
- where
$$\nabla S(b) = \left(\frac{\partial S(b)}{\partial \beta_0}, \ldots, \frac{\partial S(b)}{\partial \beta_p} \right).$$
  denotes the gradient of $S(b)$ with respect to
  $b = (b_1, \ldots, b_p)^T$.

- It follows therefore that the minimizer of $S(b)$ satisfies the equality
$$X^T X b = X^T y.$$

- It follows therefore that the minimizer of $S(b)$ satisfies the equality
$$X^T Xb = X^T y.$$
- This gives p linear equations for the $p + 1$ unknowns $b0, \ldots, b_p$.

- It follows therefore that the minimizer of $S(b)$ satisfies the equality
$$X^T X b = X^T y.$$
- This gives p linear equations for the $p + 1$ unknowns $b0, \ldots, b_p$.
- This important set of equations are called normal equations. Their solution, denoted by $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T$ gives an estimate of β called the least squares estimate.