36-401, Chapter 2: Simple Linear Regression

Zach Branson, Fall 2025

The Simple Linear Regression Model

We'll consider the most basic regression model: the **simple linear regression model**. This model involves a single covariate X and outcome Y.

This model is called "simple" only because it uses just one covariate. Nonetheless, many interesting nuances arise even under this seemingly "simple" regression model. Importantly, this model will serve as a building block for more complex models throughout the course.

As before, we assume we've collected observations $(X_1, Y_1), (X_2, Y_2), \dots (X_n, Y_n)$. The simple linear regression model assumes that

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad i = 1, 2, \dots, n$$

where β_0 is the **intercept** and β_1 is the **slope** for the regression line. Meanwhile, ϵ_i is random noise around the regression line.

Importantly, the above equation communicates a *conditional* relationship: $Y_i | X_i$. Thus, we will consider quantities like $\mathbb{E}[Y_i | X_i]$, $\text{Var}(Y_i | X_i)$, and the conditional distribution $Y_i | X_i$. Often, texts will say the X_i are "fixed" in a linear regression model; but really, they are fixed by conditioning on X_i.

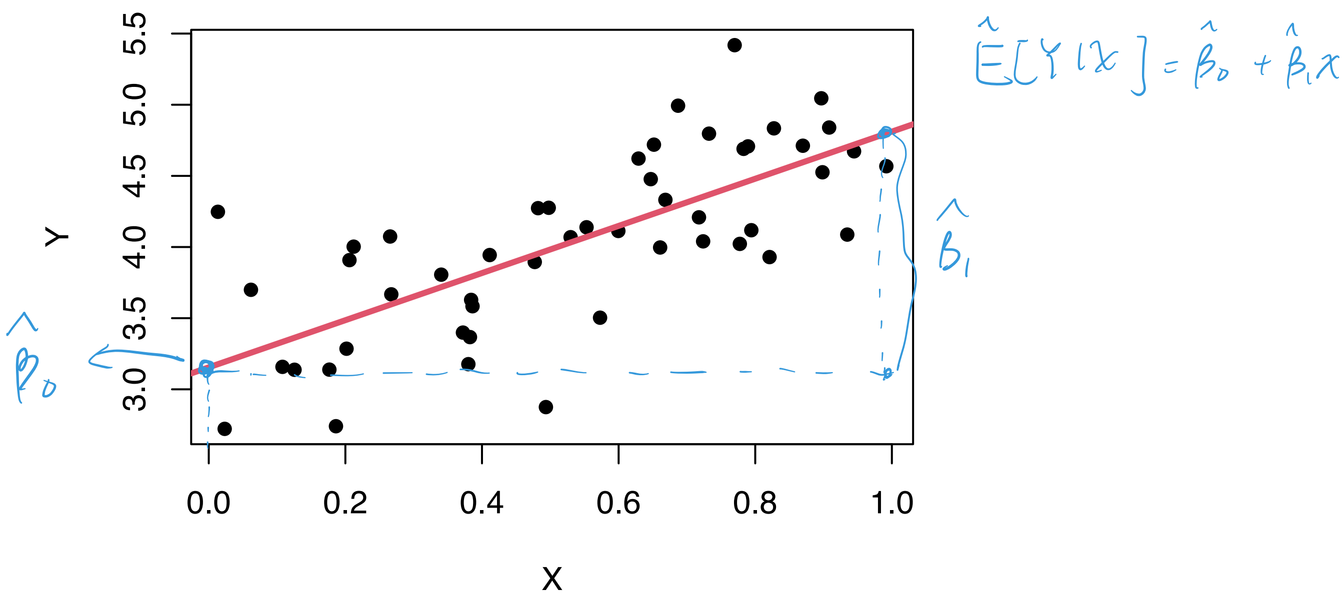

Figure 0.1: Synthetic scatter plot with $n = 50$ and a fitted regression line, $Y = 3.15 + 1.65X$.

Linear regression can be applied to any dataset $(X_1, Y_1), \ldots, (X_n, Y_n)$. However, to characterize the behavior of the linear regression model and conduct inference, we will need assumptions about the ϵ_i to justify our conclusions. Because we are conditioning on X_i and β_0, β_1 are fixed unknown parameters, ϵ_i is the only random variable in the linear regression model.

The base assumptions associated with simple linear regression are:

1. **Mean-Zero Noise**: $\mathbb{E}(\epsilon_i \mid X_i) = 0$ for all i
2. **Constant Variance** ("homoskedasticity") $\mathrm{Var}(\epsilon_i \mid X_i) = \sigma^2$ for all i
3. **Uncorrelated Noise**: The ϵ_i are uncorrelated, i.e., $\mathrm{Cov}(\epsilon_i, \epsilon_j \mid X_i) = 0$ for all $i \neq j$.

Additional assumptions, **if justified**, lead to stronger results. For example, as we'll discuss, one often assumes that the ϵ_i are normally distributed.

Parameter Estimation

The simple linear regression model has three fixed, unknown **parameters**: β_0, β_1, and σ^2. We'll have to estimate these parameters using the data $(X_1, Y_1), \ldots, (X_n, Y_n)$.

We'll first consider maximum likelihood estimation (MLE), which requires a likelihood function. In general, given iid random variables $(Z_1, \ldots, Z_n)$ with parameter(s) θ, the *likelihood function* is

$f(z_i)$ = pdf of z_i

$$L(\theta) = \prod_{i=1}^{n} f(z_i)$$

(involves θ)

Thus, in order to define a likelihood function within the context of linear regression, we'll have to assume $Y_i|X_i$ follows a distribution. To do this, we'll assume $\epsilon_i \mid X_i \overset{iid}{\sim} N(0, \sigma^2)$.

Exercise 0.1. Describe the maximum likelihood approach to parameter estimation within the context of simple linear regression.

Note: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where $\varepsilon_i | X_i \sim N(0, \sigma^2)$

$\Rightarrow Y_i | X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f(y_i | X_i = x_i)$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(y_i - \underbrace{\mu(x_i)}_{=\beta_0+\beta_1 x_i})^2}{2\sigma^2}\right)$$

"proportional to" $\propto \frac{1}{\sigma^n} \exp\left(-\sum_{i=1}^{n} \frac{(y_i - \mu(x_i))^2}{2\sigma^2}\right)$

Equivalently, can maximize the log-likelihood: $\ell(\beta_0, \beta_1, \sigma^2) = -n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu(x_i))^2$

Then maximize (take derivatives, set to zero)

The standard, classic approach to estimating the β parameters is to use **least squares**, which doesn't require distributional assumptions.

Least squares finds the β_0 and β_1 that minimize the following criterion:

$$Q_{\mathrm{LS}}(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

Exercise 0.2. Show that, for simple linear regression, the least squares estimators are:

$$\widehat{\beta}_1 = \frac{\sum_i \left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right)}{\sum_i \left(X_i - \overline{X}\right)^2} = \frac{s_{xy}}{s_{xx}}$$

and

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

To minimize Q_{LS}, take derivatives wrt. β_0 and β_1, set equal to zero and solve.

$$\frac{\partial Q_{LS}}{\partial \beta_0} = 2 \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_i)) = 0$$

$$\Rightarrow \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_i)) = 0$$

$$n\beta_0 = \sum_{i=1}^{n} (Y_i - \beta_1 x_i)$$

$$\Rightarrow \beta_0 = \overline{Y} - \beta_1 \overline{x}$$

Do the same for β_1, solve for β_1.

Technically, must check these are minima (second derivatives are positive)

Additional Basic Definitions

Once $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are determined, the **fitted values** are:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i, \quad i = 1, 2, \ldots, n$$

and the **residuals** are

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i, \quad i = 1, 2, \ldots, n.$$

The quantity

$$\text{RSS} = \sum_{i=1}^{n} \widehat{\epsilon}_i^2$$

is called both the **residual sum of squares (RSS)** and the **sum of squared errors (SSE)**.

Exercise 0.3. Give practical interpretations of the fitted values and the residuals versus ϵ.

Why Least Squares?

Why minimize the sum of squared errors and not another quantity? For example, we could minimize the sum of absolute errors:

$$Q_{\text{L1}}(\beta_0, \beta_1) = \sum_{i=1}^{n} |Y_i - (\beta_0 + \beta_1 X_i)| \,.$$

Minimizing Q_{L1} is indeed possible; it is referred to as **L1 Regression**.

There are several reasons why people have focused on least squares:

1. **Computational Efficiency:** It is straightforward to derive the least squares estimators of β_0 and β_1, as well as properties of them (e.g., means and variances).
2. **The Gauss-Markov Theorem:** Under the three basic assumptions given above, the least squares estimators of β_0 and β_1 are unbiased and have minimum variance among all unbiased estimators.
3. **Least squares gives the MLE** when the ϵ_i are independent and normally distributed.
4. The regression function $\mathbb{E}(Y \mid X)$ is the **mean-squared optimal predictor** of Y.

Exercise 0.4. Recall that our regression model is $Y = f(X) + \epsilon$, where $f(X)$ is the regression function. Show that $f(X) = \mathbb{E}[Y \mid X]$ minimizes MSE.

Exercise 0.5. Assuming $f(X) = \mathbb{E}[Y \mid X]$, what is the optimal linear predictor? How is RSS related to MSE?

Estimating σ^2

The variance σ^2 comes from a distributional assumption on the residuals ϵ_i. If $\epsilon_i \mid X_i \overset{iid}{\sim} N(0, \sigma^2)$, then the MLE for σ^2 is

$$\widehat{\sigma}^2_{\text{MLE}} = \left(\frac{1}{n}\right) \sum_{i=1}^{n} \widehat{\epsilon}_i^2 = \text{RSS} \Big/ n.$$

This estimator is biased. Meanwhile, this adjusted estimator is unbiased:

$$\widehat{\sigma}^2 = \left(\frac{n}{n-2}\right) \widehat{\sigma}^2_{\text{MLE}} = \text{RSS} \Big/ (n-2).$$

This is the estimator we will typically use, and what is reported by R and other software packages. R calls $\hat{\sigma}$ the "residual standard error."

Exercise 0.6. What's the intuition for dividing by $(n-2)$?

Linear Regression in R

In R, we use the `lm()` function to fit linear models using least squares. It takes two key arguments:

- A *model formula*, which is special syntax for specifying the outcome and covariates.
- A *data frame* providing the observed data, which must contain columns whose names match the terms in the model formula.

Model formulas place the outcome to the left of ~ and covariates to the right, separated by + signs.

Formulas can contain transformations and some useful functions:

- `mpg ~ disp` fits a model with `mpg` as the outcome and `disp` as the covariate.
- `mpg ~ log(disp)` log-transforms `disp`.
- `mpg ~ I(disp^2)` takes the square of `disp`. `I()` is necessary because the ^ operator has a specific meaning in formulas, so `I()` tells R to ignore this and evaluate it as-is.
- `mpg ~ disp - 1` removes the intercept from the model.

The `lm()` function returns a fit object containing the data, fit parameters, estimates, and various other useful information.

Example 0.1. Let's return to the Bureau of Economic Analysis (BEA) data example. We'll again consider per-capital GMP ('pcgmp') as the outcome and use population ('pop') as the covariate. We'll log-transform population, because our initial EDA suggested the relationship is linear after log-transformation.

```r
bea <- read.csv("data/bea-2006.csv")
bea_fit <- lm(pcgmp ~ log(pop), data = bea)
summary(bea_fit)
```

```
##
## Call:
## lm(formula = pcgmp ~ log(pop), data = bea)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21572  -4765  -1016   3686  40207
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -23306.2     4957.1  -4.702 3.67e-06 ***
## log(pop)      4449.8      390.9  11.383  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7929 on 364 degrees of freedom
## Multiple R-squared:  0.2625,Adjusted R-squared:  0.2605
## F-statistic: 129.6 on 1 and 364 DF,  p-value: < 2.2e-16
```

To get the estimates in code, we can use the `coef()` (or `coefficients()`) function, which returns a named vector of estimates:

```r
coef(bea_fit)
```

```
## (Intercept)    log(pop)
##  -23306.199    4449.758
```

```r
coef(bea_fit)["log(pop)"]
```

```
## log(pop)
## 4449.758
```