2.1: Introduction to Simple Linear Regression

Dr. Bean - Stat 5100

See **Handout 2.1.1** for information regarding the Toluca power company example.

1 Why Linear Regression?

Linear regression is good for:

- **Inference:** determine if there is a statistically significant linear relationship between two variables, while possibly accounting for the effect of additional variables.
  - Example: after accounting for the effects of square footage and age, are lot size and home sale price significantly linearly related?
- **Prediction:** use variables that are "easy" to measure to predict variables that are harder to measure.
  - Example: Use elevation (easy to measure) to predict annual snow accumulation (hard to measure).

Linear regression only works for variables that share a statistical relationship.

Terminology:

- Y - response variable
- X_i - predictor variables
- ϵ - error (or difference) term
- β_i - model parameters (true values are unknown and are estimated)

Linear Regression focuses on finding appropriate estimates of the model parameters (b_i):

The idea is that we want to select paramater estimates that make the predicted values of Y ($\hat{Y}$) close to the actual values of Y.

2 Ordinary Least Squares (OLS) Regression

If assumptions regarding residuals are satisfied (more in Handout 2.2), then the OLS estimates of the model parameters are "best."

What does it mean to be "best"?

- **unbiased** - given an infinite number of different samples of data, the average of my estimates will be equal to true (and unknown) value of the parameter.
  - In other words, my estimates are "centered" on the truth.
- **minimum variance** - the variation in the estimate from sample to sample is the smallest of all possible estimation methods.

Applications - Toluca Example:

Let X represent the lot size and let Y represent the total work hours. Based on the initial scatterplot, we assume that the relationship between X and Y can be modeled as

$$Y = \beta_0 + \beta_1 X + \epsilon$$

OLS seeks to minimize:

$$Q = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2,$$

which requires us to select estimates b_0 and b_1 that minimize

$$Q = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2 = f(\mathbf{X}).$$

We can use multivariable calculus to find the minimum of Q by finding the critical points, i.e.

$$\nabla Q = \nabla f(\mathbf{X}) = 0.$$

The single critical point that minimizes Q is

$$b_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i \left(X_i - \bar{X}\right)^2}$$
$$b_0 = \bar{Y} - b_1 \bar{X}.$$

Obtain OLS estimates automatically in SAS with:

```
proc reg data=toluca;
  model workhours = lotsize;
  title1 'Simple linear model';
run;
```

Equation Estimates:

$b_0 = 62.37$, $b_1 = 3.57$

Model Equation:

$\hat{Y} = 62.37 + 3.57(lotSize)$

The Critical Assumption

OLS least squares hinges on the assumption that

$$\epsilon \overset{iid}{\sim} N(0, \sigma^2)$$

- **independent:** Knowing the value of one of the model residuals tells you nothing about any of the others.
- **identically distributed:** All of the residuals come from the same distribution.
- **Normal Distribution:** The model residuals follow a normal (bell shaped) distribution.

- **zero mean:** The average of the residuals is zero (unbiased estimates).
- **constant variance:** The spread of the residuals about the line is the same across the range of X and the range of predicted values.

If the assumptions hold, then the simple linear regression can be visualized as in Figure 1.

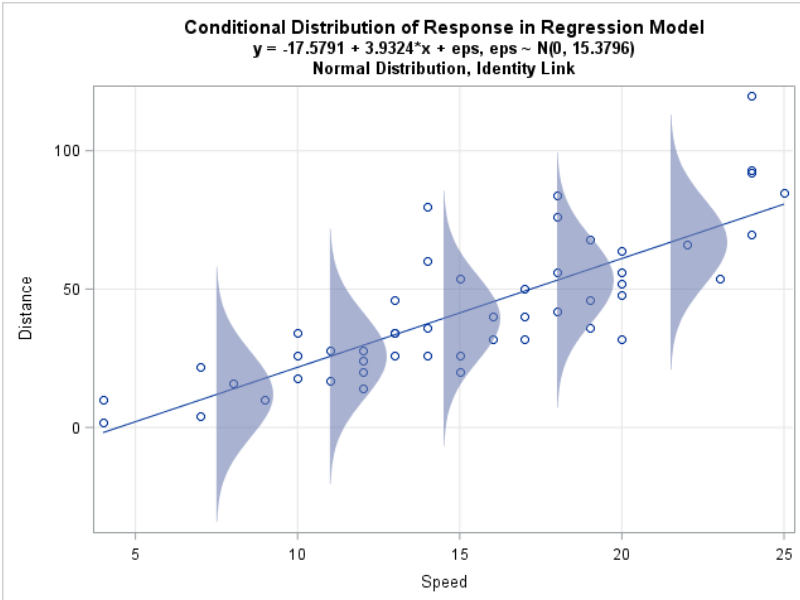

Figure 1: Sample visualization taken from Rick Wicklin on The DO Loop.

In other words, Y follows a normal distribution with a center that is conditional on X.

Estimating σ

Estimating the variance about the regression line:

- Allows us to get a measure of the model fit: lower relative MSE $\rightarrow$ better model.
- All significance tests of model coefficients are based on our estimate of σ.

Estimation of ϵ in Theory

Suppose that $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ were an observed sample from some population. (In practice, ϵ is estimated as the residuals of our OLS model, represented as e_i.)

We could then estimate $\text{Var}(\epsilon)$ as

$$\frac{1}{n-1}\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon})^2$$

.

Note that the variance calculation requires the estimation of $\mu_\epsilon = \bar{\epsilon} = \frac{1}{n}\sum_i \epsilon_i$.

This calculation "constrains" one of the ϵ_i. This means that if we know $\bar{epsilon}$ and $\epsilon_1, \ldots \epsilon_{n-1}$, then we can know ϵ_n.

We call the number of unconstrained observations the "degrees of freedom" (DF).

Every time you estimate a parameter, **you lose one degree of freedom.**

Think of observations as currency. We spend money to estimate things and our degrees of freedom are the leftover cash.

Estimation of ϵ in Practice

Why is it that we can't directly obtain the values of ϵ?

We don't know the true regression line, so we cannot know the true values of epsilon.

We can obtain estimates of the residuals e_i through the regression line:

$$e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i).$$

OLS, by design, makes $\sum_i e_i = 0 \rightarrow \bar{e} = 0$, meaning I don't have to spend any DF to obtain $\bar{e}$.