L-1 Regression

Simple linear Regression

There are n experimental units.
On each unit we measure several variables.
One particular variable is of interest (Y) : response
Other variables $(x_1, \ldots, x_p)$ are predictors.

e.g.

1. A day = unit
   (Y, x_{1i}, x_{2i}) observed for day $i = 1, \ldots, 20$
   - Y = Total precipitation.
   - X_1 = Temp at 10AM
   - X_2 = Relative Humidity at 10AM.

We are asking for a f^n f that helps us predict the "best" value of Y given $(X_1, \ldots, X_p)$.

1. This f is completely unknown.

   Simple: $p = 1$
   Linear: f is a linear f^n, $f(x) = a + bx$.
   For simple linear regression, Y is continuous. One can have $p > 1$. (Multiple Regression.)

2. Drop Linearity: Non-linear Regression.

3. Y can be discrete/categorical data, logistic regression.

Why Regression?

1. To predict the value of Y given $X_1, \ldots, X_p$.

2. To study the impact of a predictor on the response.
   $Y = a + bX$
   What is the change in response for unit change in predictor?

3. Estimator of Parameters?
   Testing Hypothesis e.g. $\beta > 0$.

e.g.

Y = Salary
x_1 = education, x_2 = gender, x_3 = experience
Keeping education and experience fixed, does gender affect Salary? (Testing).
How much does salary change for every year of experience? (Estimation)

Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Model:

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n$

- Linear Model Holds
- Non linear: $y_i = \alpha e^{\beta x_i} + \epsilon_i$

1. ϵ_i are independent of each other
2. ϵ_i are independent of x_i.
   Often x_i are fixed. Otherwise we condition on x_i.

$E[\epsilon_i] = 0 \quad \mathrm{Var}(\epsilon_i) = \sigma^2$ (some #)
ϵ_i are random, so Y_i are also random
$E[Y_i] = \beta_0 + \beta_1 x_i + 0$

Plot of infant mortality (y) on female literacy (x).
Roughly my pattern supports linearity assumption.

Goal:

1. Estimate Parameter $(\beta_0, \beta_1, \sigma^2)$ (fitting).
2. Verify if model assumptions are satisfied. (Diagnostics)
   (Run Diagnostics)
   If assumptions are not satisfied, we modify the model. We fit a model, get from there.
   Keep using this until you get a model that satisfies all assumptions.

First step is to draw a scatter plot and ensure linear model is reasonable at least visually.
Next estimate the β's & σ^2.

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
$Var(\epsilon_i) = \sigma^2$

We want to estimate β_0, β_1.
Given $y_1, \ldots, y_n$ to predict one value of Y, the most common candidate is $\bar{y}$.

$$\arg\min_a \sum_{i=1}^{n} (Y_i - a)^2 = \bar{Y}$$

$\bar{y}$ minimizes the avg Squared distances from all obs.

$$(\hat{\beta_0}, \hat{\beta_1}) = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

Least Squares Estimator

Solution:

$$\hat{\beta_1} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$$

Shift the line around to minimize the sum of squared vertical distance.

1. We do not consider perpendicular distances in regression.
2. Treats x & y asymmetrically.
   If we regress x on y:
   $$x = \alpha_0 + \alpha_1 y + \epsilon$$
   is the model.
   $$\hat{\alpha_1} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (y_i - \bar{y})^2}$$
   $$\hat{\alpha_0} = \bar{x} - \hat{\alpha_1}\bar{y}$$
   This minimizes horizontal distance.
3. Both fitted lines pass through $(\bar{x}, \bar{y})$

$$y - \bar{y} = \hat{\beta_1}(x - \bar{x})$$

$$x - \bar{x} = \hat{\alpha_1}(y - \bar{y})$$

$$\hat{\beta_1}\hat{\alpha_1} = r^2 \quad \text{(r is corr. coeff.)}$$

Regression to the mean

The regression line can be written as:

$$y - \bar{y} = \hat{\beta}_1(x - \bar{x})$$

Standardizing x and y:

$$\frac{y - \bar{y}}{s_y} = r \cdot \frac{x - \bar{x}}{s_x}$$

where r is the correlation coefficient and $|r| \leq 1$.

Galton observed that tall fathers have sons who are not as tall (on average, their height regresses towards the mean). e.g., Scores of 2 tests of the same set of individuals.

$L - 2$

Simple Linear Regression

- **Assumptions**
- **Model:** $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where ϵ_i has mean 0, variance σ^2.
- **Parameters:** $\beta_0, \beta_1, \sigma^2$.
- **Least Squares Criterion**

Goals: Estimate $\beta_0, \beta_1, \sigma^2$.

If ϵ_i are normally distributed, then the least squares estimator is the Maximum Likelihood Estimator (MLE). The least squares criterion is general and does not need a distributional assumption.

Another possible criterion is **Least Absolute Deviation**: $\min \sum |y_i - \beta_0 - \beta_1 x_i|$.

How to estimate σ^2?

The residuals are $\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. An unbiased estimator for σ^2 is:

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

The factor $n - 2$ is a correction factor to make the estimator unbiased. For a simple model $y_i = a + \epsilon_i$, the estimate for variance is $\frac{1}{n-1}\sum(y_i - \bar{y})^2$.

Is x useful in predicting y?

If there was no predictor, the error in predicting y would be measured by the **Total Sum of Squares (SST)**:

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

The **Residual Sum of Squares (SSR)** after fitting the model is:

$$SSR = \sum(y_i - \hat{y}_i)^2 = \sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

If the regression exercise is useful, then $SSR < SST$. The larger the difference $(SST - SSR)$, the more useful(better) is the model.

Decomposition of Variance

Claim A: $SST \geq SSR$.

Let $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$, $S_{xx} = \sum(x_i - \bar{x})^2$, $S_{yy} = \sum(y_i - \bar{y})^2$. Recall $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ and $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$.

$$
\begin{aligned}
SSR &= \sum(y_i - (\bar{y} - \hat{\beta}_1\bar{x}) - \hat{\beta}_1 x_i)^2 = \sum[(y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x})]^2 \\
&= \sum(y_i - \bar{y})^2 - 2\hat{\beta}_1 \sum(x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}_1^{\,2} \sum(x_i - \bar{x})^2 \\
&= S_{yy} - 2\hat{\beta}_1 S_{xy} + \hat{\beta}_1^{\,2} S_{xx} \\
&= S_{yy} - 2\frac{S_{xy}}{S_{xx}} S_{xy} + \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}
\end{aligned}
$$

So, $SSR = SST - \frac{S_{xy}^2}{S_{xx}}$, which implies $SSR \leq SST$.

The amount by which the sum of squares is being reduced due to regression is the **Regression Sum of Squares (SSReg)**.

$$SSReg = SST - SSR = \frac{S_{xy}^2}{S_{xx}}$$

Coefficient of Determination (R^2)

The ratio of SS explained by regression to the total SS of y is denoted by R^2.

$$R^2 = \frac{SSReg}{SST} = \frac{S_{xy}^2/S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = r^2$$

where r is the correlation coefficient between x and y. R^2 is called the coefficient of determination. This is true in general, even with multiple predictors.

Analysis of Variance Table (ANOVA)

Source	**SS**	**df**	**MS (Mean Square)**	**F**
Regression	SSReg	1	$MSReg = \frac{SSReg}{1}$	$\frac{MSReg}{MSR}$
Residual	SSR	n-2	$MSR = \frac{SSR}{n-2}$	
Total	SST	n-1		

- Each SS has chi-square (χ^2) distribution with corresponding degrees of freedom (df).
- Larger values of R^2 are better.
- R^2 measures the proportion of variability of y that can be explained by linear regression on x.
- A larger F-statistic is associated with a larger R^2.
- The ratio of two independent chi-square (χ^2) distributions (divided by their df) gives an F-distribution.
- So we can do a Hypothesis Test (under normality conditions).

Hypothesis Test for Significance of Regression

- H_0: Regression is not significant ($\beta_1 = 0$).

- H_a: Regression is significant ($\beta_1 \neq 0$).

The F-statistic can be written in terms of R^2:

$$F = \frac{MSReg}{MSR} = \frac{SSReg/1}{SSR/(n-2)} = \frac{R^2 \cdot SST}{(1-R^2)SST/(n-2)} = \frac{R^2}{1-R^2}(n-2)$$