Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

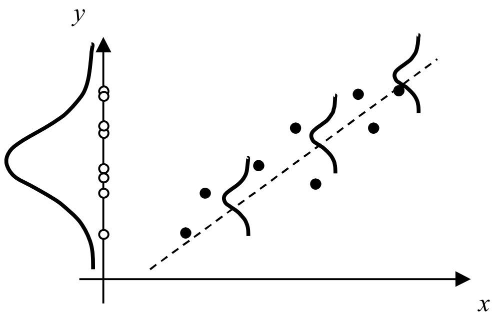
457.212 Statistics for Civil & Environmental Engineers
In-Class Material: Class 25
Regression Analysis (A&T: 8.2-8.4, 8.7)

Given: Sample data set $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$

Question: The functional relation between two random variables X and Y? $Y = f(X)$

→ "Regression" Analysis

1. **Regression & Conditional Mean**



(a) Marginal and conditional standard deviation of Y:

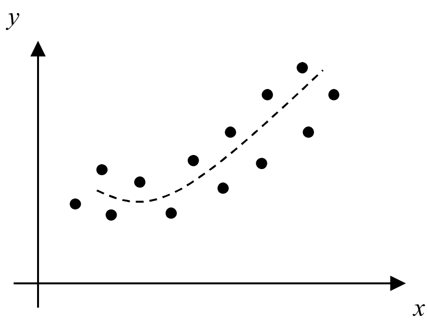
σ_Y $\sigma_{Y|x}$

(b) Marginal and conditional mean of Y:

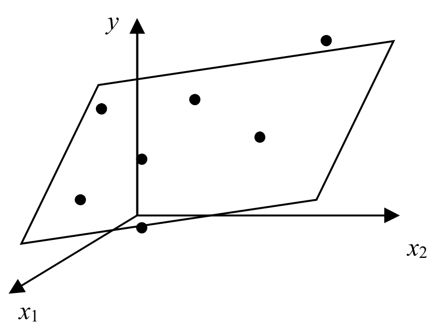
$\mathrm{E}[Y] =$ constant.
$\mathrm{E}[Y \mid x] = f(x)$

→ Conditional mean predicts the outcome of Y more accurately (i.e. smaller variation).
→ Regression analysis aims at finding the functional relationship for the conditional mean to describe the hidden relation between X and Y.

(c) Linear vs. nonlinear regression

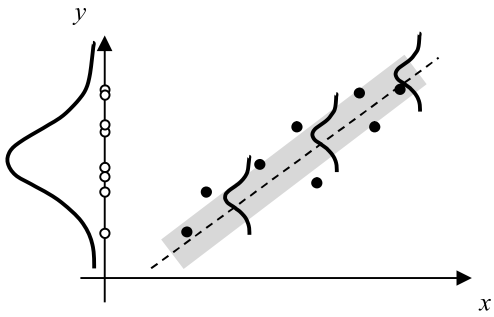


(d) Single vs. multiple regression



Instructor: Junho Song
junhosong@snu.ac.kr

2. Single Linear Regression with Constant (Conditional) Variance



(a) Assumption: the conditional mean is a linear function of x and the conditional variance is constant, i.e.

$\mathrm{E}[Y \mid x] = \alpha + \beta x$ and $\sigma^2_{Y|x} =$ const.

"Linear regression of Y on X"

(b) Estimation of α and β

"Best" estimates on α and β: $\hat{\alpha}$ and $\hat{\beta}$ ~ the values minimizing the sum of squared errors between the prediction by the linear relationship ($y_i' = \alpha + \beta x_i$) and the given data point y_i (least square estimators)

Sum of Squared Errors (SSE):

$$\begin{aligned}\Delta^2 &= \sum_{i=1}^{n} (y_i - y_i')^2 \\ &= \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2\end{aligned}$$

Note: The same weight is given to each data point because the conditional variance is assumed to be constant.

Find α and β that minimize SSE → Solve the following equations for α and β:

$$\frac{\partial \Delta^2}{\partial \alpha} = 2\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)(-1) =$$

$$\frac{\partial \Delta^2}{\partial \beta} = 2\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)(-x_i) =$$

Instructor: Junho Song
junhosong@snu.ac.kr

As a result,

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

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x}$$

Need: $\sum x_i y_i$, $\sum x_i$, $\sum y_i$ and $\sum x_i^2$

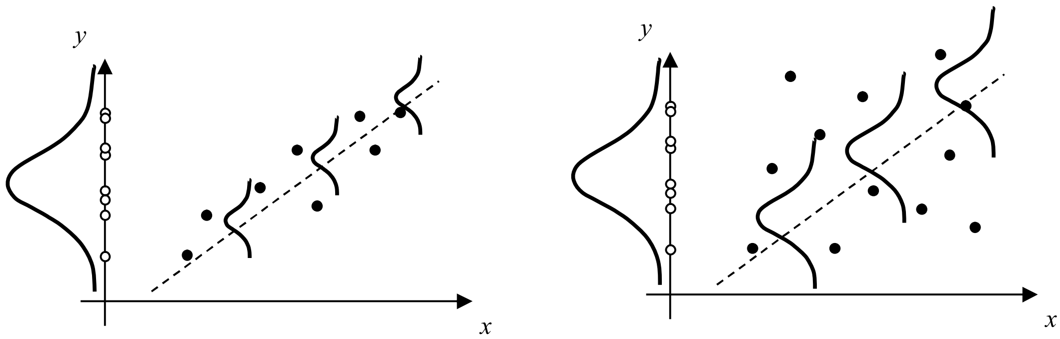
(c) $\sigma_{Y|x}^2$?

Estimated as

$$s_{Y|x}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - y_i')^2 = \frac{\Delta^2}{n-2}$$

(d) Reduction of variance: from marginal $\sigma_Y^2 (s_Y^2)$ to conditional variance $\sigma_{Y|x}^2 (s_{Y|x}^2)$?

→ A measure of the strength of the linear relationship



$$r^2 = \frac{s_Y^2 - s_{Y|x}^2}{s_Y^2} = 1 - \frac{s_{Y|x}^2}{s_Y^2}$$

$r^2 \cong 0$: No reduction (weak linear relationship)

$r^2 \cong 1$: Large reduction (strong linear relationship)

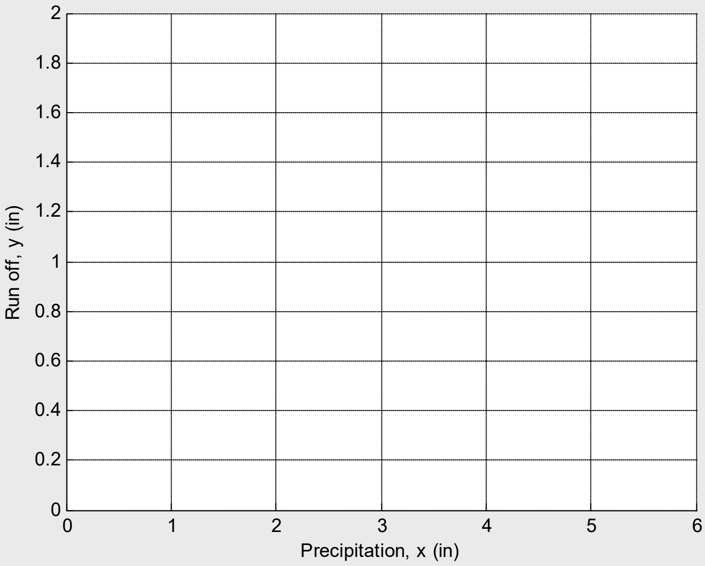
Note: $r^2 \cong \rho_{XY}$ as $n \to \infty$

Instructor: Junho Song
junhosong@snu.ac.kr

Example 1: Regression analysis of Runoff (*Y*) on Precipitation (*X*)

	x_i (in.)	y_i (in)	$x_i y_i$	x_i^2	y_i^2	y_i'	$(y_i - y_i')^2$
	1.01	0.30	0.303	1.0201	0.09		
	2.09	0.95	1.9855	4.3681	0.9025		
	3.57	1.59	5.6763	12.7449	2.5281		
	5.11	1.74	8.8914	26.1121	3.0276		
	2.93	1.12	3.2816	8.5849	1.2544		
Sum							
Avg							

(a) Scatter plot?



(b) Linear regression of Y on X (i.e. Find $\hat{\alpha}$ and $\hat{\beta}$)?

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

$$=$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x} =$$

Thus, $\mathrm{E}[Y \mid x] =$

Show it in the plot.

(c) Estimate on the conditional variance, $s_{Y|x}^2 = \dfrac{\Delta^2}{n-2} =$

(d) Estimate on the marginal variance, $s_Y^2 = \dfrac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) =$

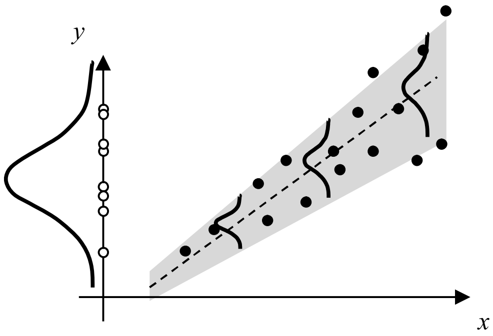
(e) Reduction ratio, $r^2 = 1 - \dfrac{s_{Y|x}^2}{s_Y^2} =$

(f) Suppose the precipitation is 4.0 in.
What is the mean run off?

Probability that the run-off exceeds 2 in.?

Instructor: Junho Song
junhosong@snu.ac.kr

3. Single Linear Regression with Non-constant Variance



(a) Assumption: the conditional mean is a linear function of x and the conditional variance is a function of x, i.e.

$\mathrm{E}[Y\,|\,x] = \alpha + \beta x$

$\sigma_{Y|x} = \sigma \cdot g(x)$ (Thus, $\sigma_{Y|x}^2 = \sigma^2 g^2(x)$)

e.g. $\sigma_{Y|x} = \sigma x$ (linearly increasing)

$\sigma_{Y|x} = \sigma x^2$ (quadratically)

(a) Estimation of α and β

The same as regression with constant variance except that the errors are given non-equal weights.

Sum of **Weighted** Squared Errors (SWSE):

$$\Delta^2 = \sum_{i=1}^{n} w_i' \cdot (y_i - y_i')^2$$
$$= \sum_{i=1}^{n} w_i' \cdot (y_i - \alpha - \beta x_i)^2$$

Note: Give more weights to the data points that require more accurate fitting.

$$w_i' \equiv \frac{1}{\sigma_{Y|x}^2} = \frac{1}{\sigma^2 g^2(x)}$$

Find α and β that minimize SWSE → Solve the following equations for α and β:

$$\frac{\partial \Delta^2}{\partial \alpha} = 0 \text{ and } \frac{\partial \Delta^2}{\partial \beta} = 0$$

Instructor: Junho Song
junhosong@snu.ac.kr

As a result,

$$\hat{\beta} = \frac{(\Sigma w_i)(\Sigma w_i x_i y_i) - (\Sigma w_i y_i)(\Sigma w_i x_i)}{(\Sigma w_i)(\Sigma w_i x_i^2) - (\Sigma w_i x_i)^2}$$

$$\hat{\alpha} = \frac{(\Sigma w_i y_i) - \hat{\beta}(\Sigma w_i x_i)}{\Sigma w_i}$$

where $w_i = \sigma^2 w_i' = \dfrac{1}{g^2(x_i)}$

Need: $\sum w_i x_i y_i$, $\sum w_i x_i$, $\sum w_i y_i$, $\sum w_i x_i^2$ and $\sum w_i$

(c) $\sigma^2_{Y|x}$?

First, an unbiased estimate of σ^2 (not $\sigma^2_{Y|x}$) is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} w_i (y_i - y_i')^2}{n-2}$$

Then, the conditional variance is estimated as

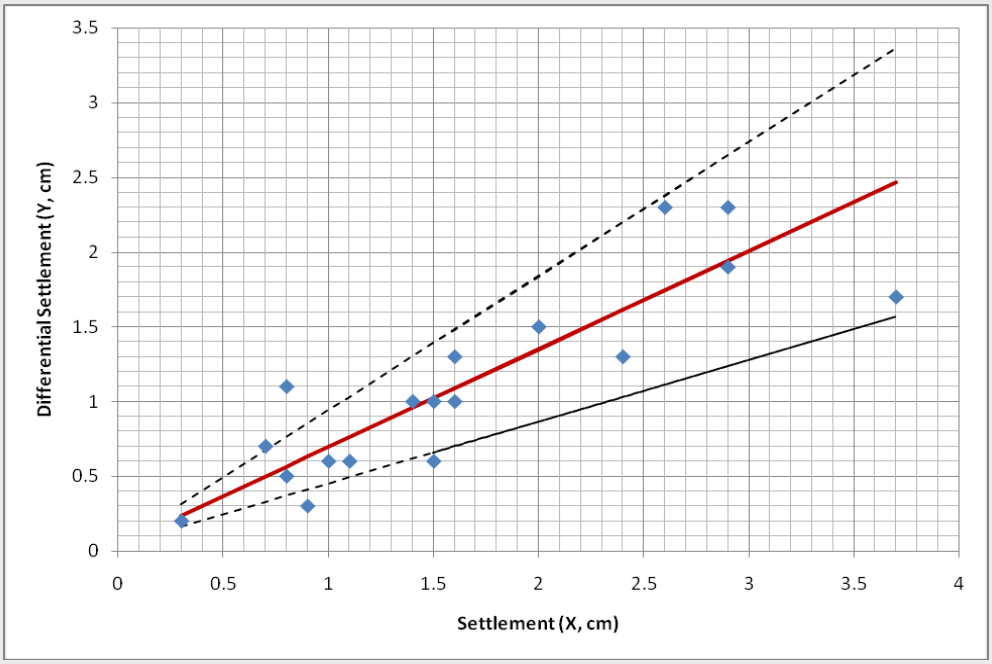
$$s^2_{Y|x} = \hat{\sigma}^2 g^2(x)$$
$$= \frac{\Sigma w_i (y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{n-2} g^2(x)$$

Example 2 (A&T 8.4): 18 Storage tanks. X – maximum settlement (cm), Y – maximum differential settlement (cm)

Assume $\sigma_{Y|x} = \sigma x$ (Linearly increasing)

Thus, $w_i' = \dfrac{1}{\sigma^2 x_i^2}$, $w_i = \sigma^2 w_i' = \dfrac{1}{x_i^2}$

Tank No.	xi	yi	wi	wi*xi	wi*yi	wi*xi*yi	wi*xi^2	yi'	wi*(yi-yi')^2	s_Y|x	yi'+s_Y|x	yi'-s_Y|x
1	0.3	0.2	11.11111	3.333333	2.222222	0.666667	1	0.237608	0.015715398	0.072908	0.310517	0.1647
2	0.7	0.7	2.040816	1.428571	1.428571	1	1	0.499812	0.081786579	0.17012	0.669931	0.329692
3	0.8	0.5	1.5625	1.25	0.78125	0.625	1	0.565362	0.006675366	0.194422	0.759785	0.37094
4	0.8	1.1	1.5625	1.25	1.71875	1.375	1	0.565362	0.446620994	0.194422	0.759785	0.37094
5	0.9	0.3	1.234568	1.111111	0.37037	0.333333	1	0.630913	0.135189509	0.218725	0.849638	0.412188
6	1	0.6	1	1	0.6	0.6	1	0.696464	0.009305291	0.243028	0.939492	0.453436
7	1.1	0.6	0.826446	0.909091	0.495868	0.545455	1	0.762015	0.021693203	0.267331	1.029345	0.494684
8	1.4	1	0.510204	0.714286	0.510204	0.714286	1	0.958667	0.000871635	0.340239	1.298906	0.618428
9	1.5	1	0.444444	0.666667	0.444444	0.666667	1	1.024218	0.000260671	0.364542	1.38876	0.659676
10	1.6	1	0.390625	0.625	0.390625	0.625	1	1.089769	0.003147824	0.388845	1.478613	0.700924
11	1.6	1.3	0.390625	0.625	0.507813	0.8125	1	1.089769	0.017264523	0.388845	1.478613	0.700924
12	2	1.5	0.25	0.5	0.375	0.75	1	1.351972	0.005478075	0.486056	1.838028	0.865916
13	2.4	1.3	0.173611	0.416667	0.225694	0.541667	1	1.614175	0.017136464	0.583267	2.197442	1.030908
14	2.6	2.3	0.147929	0.384615	0.340237	0.884615	1	1.745277	0.045520394	0.631872	2.377149	1.113404
15	2.9	1.9	0.118906	0.344828	0.225922	0.655172	1	1.941929	0.000209044	0.704781	2.64671	1.237148
16	2.9	2.3	0.118906	0.344828	0.273484	0.793103	1	1.941929	0.015245507	0.704781	2.64671	1.237148
17	3.7	1.7	0.073046	0.27027	0.124178	0.459459	1	2.466336	0.042897753	0.899203	3.365539	1.567132
18	1.5	0.6	0.444444	0.666667	0.266667	0.4	1	1.024218	0.079982607	0.364542	1.38876	0.659676
SUM	29.7	19.9	22.40068	15.84093	11.3013	12.44792	18		0.945000837			
beta	0.655508											
alpha	0.040956											
sigma_hat^2	0.059063											
sigma_hat	0.243028											



Instructor: Junho Song
junhosong@snu.ac.kr

4. Multiple Linear Regression

"Linear regression of Y on $X_1, \ldots, X_m$"

(a) Define Δ^2 by assuming $\sigma^2_{Y|x} = \sigma^2$ (constant) or $\sigma^2_{Y|x} = \sigma^2 g^2(x_1,\ldots,x_m)$ (non-constant)

(b) Find

$$\mathrm{E}[Y \mid x_1,\ldots,x_m] = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m$$

(c) Estimate $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m$ by solving

$$\frac{\partial \Delta^2}{\partial \beta_0} = \frac{\partial \Delta^2}{\partial \beta_1} = \cdots = \frac{\partial \Delta^2}{\partial \beta_m} = 0$$

(d) $s^2_{Y|x_1,\ldots,x_m} = \dfrac{\Delta^2}{n-m-1}$

(**Note**: $m = 1$ for single linear regression)

5. Nonlinear Regression & Applications of Regression Analysis (Read A&T 8.6-8.7)

6. Correlation Analysis

(a) (True or theoretical) correlation coefficient

$$\rho_{XY} = \frac{Cov[X,Y]}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[XY] - \mu_X \mu_Y}{\sigma_X \sigma_Y}$$

(b) Unbiased estimator of ρ_{XY}, $\hat{\rho}$

$$\hat{\rho} = \frac{1}{n-1} \frac{\sum_{i=1}^{n} x_i y_i - n\bar{X}\bar{Y}}{s_x s_y}$$

(c) $\hat{\rho}$ and $\hat{\beta}$

$$\hat{\rho} = \hat{\rho}\frac{s_X}{s_X} = \frac{\Sigma(x_i - \bar{X})(y_i - \bar{Y})}{(n-1)s_X^2}\frac{s_X}{s_Y} = \frac{\Sigma(x_i - \bar{X})(y_i - \bar{Y})}{\Sigma(x_i - \bar{X})^2}\frac{s_X}{s_Y} = \hat{\beta}\frac{s_X}{s_Y}$$

(d) $\hat{\rho}^2$ and $r^2 = 1 - s^2_{Y|x} / s^2_Y$

$$\hat{\rho}^2 = 1 - \frac{n-2}{n-1}\frac{s^2_{Y|x}}{s^2_Y}. \qquad \text{As } n \to \infty,\ \hat{\rho}^2 \to 1 - \frac{s^2_{Y|x}}{s^2_Y} = r^2$$