Chapter 12. Simple Linear Regression & Analysis

§12.1 Simple Linear Regression Model

Linear regression generalizes the ANOVA ideas from Chapter 10. "Regression is ANOVA with continuous factors."

- ANOVA had discrete factors $f_1, f_2, \ldots$ with corresponding distributions $X^{(f_1)}, X^{(f_2)}, \ldots$
- Regression changes "factor" to continuous variable x with corresponding (continuously changing) distrib. $Y(x)$

Change in Notation from Ch. 10

"Factor" $f_1, f_2, f_3, \ldots$ will now be values of x

$f_1 \longleftrightarrow x = 1$

$f_2 \longleftrightarrow x = 2$

etc...

Now we also have
- $x = 1/2$
- $x = 1/4$
- etc

Random Variable X will now be Y

factor distributions $X^{(f)}$ will be $Y(x)$

$X^{(f_1)} \longleftrightarrow Y(x=1)$

$X^{(f_2)} \longleftrightarrow Y(x=2)$

etc...

Note: The data used for regression is Paired Samples (x_i, y_i) — the same as in §9.3. But, unlike §9.3 we will not think of x_i as a random var. We will consider Y as a random variable & x will be a parameter ("factor") for Y.

Notation: (x_i, Y_i) — x_i: "factor", Y_i: "random var."

Simple Linear Regression Model

Given paired samples (x_i, Y_i) assume that

$$Y \sim \text{Normal}(\beta_0 + \beta_1 x, \sigma)$$

ie $E[Y(x)] = \mu_y^{(x)} = \beta_0 + \beta_1 x$

$\text{Var}[Y(x)] = \sigma^2$

(The distribution for Y is moving with x.)

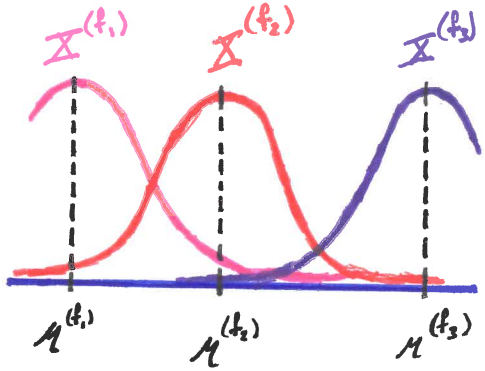
An equivalent way to write this is:

$$Y(x) = (\beta_0 + \beta_1 x) + \varepsilon$$

where $\varepsilon \sim \text{Normal}(0, \sigma)$

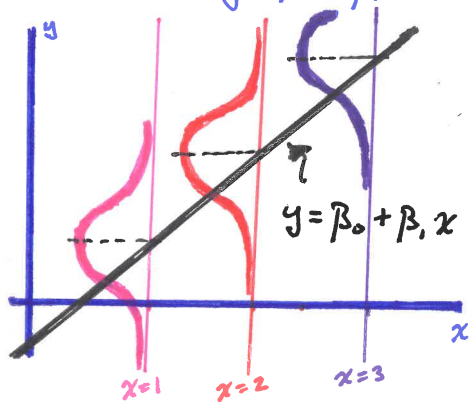
Greek letter "Epsilon" is "Error"

Idea: In Chapter 10 the different factors gave a series of shifted distributions, each with mean $\mu^{(i)}$. Now we have a single distribution which slides as x moves so that the mean follows line $y = \beta_0 + \beta_1 x$.



Shifted Distributions

Means @ $\mu^{(i_1)}, \mu^{(i_2)}, \mu^{(i_3)}$



Sliding Distribution

Means @ $y = \beta_0 + \beta_1 x$

Def: Given sample data $\{(x_i, y_i)\}$ where

$$Y \sim \text{Normal}(\beta_0 + \beta_1 x, \sigma)$$

(or, equivalently, $Y = (\beta_0 + \beta_1 x) + \varepsilon$)

the line of means

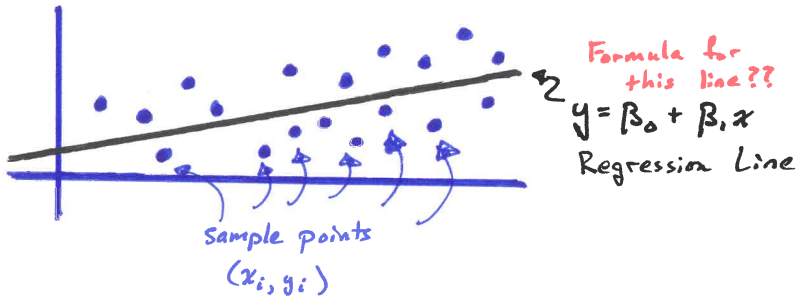
$$y = \beta_0 + \beta_1 x$$

is called the "Regression Line"

§12.2 Estimating Model Parameters

Given paired sample data, we want to figure out formula for regression line (of means)

$$y = \beta_0 + \beta_1 x$$



→ Parameter Estimation!

- Want unbiased estimators $\hat{\beta}_0$ & $\hat{\beta}_1$ for β_0 & β_1

Input: Paired sample data (x_i, y_i)

Output: Expression for $\hat{\beta}_0$, $\hat{\beta}_1$, & Regression Line.

Note: $\hat{\beta}_0$ is easy once we know $\hat{\beta}_1$ because

$$Y(x) = (\beta_0 + \beta_1 x) + \varepsilon$$

$$E[Y(x)] = E[\beta_0 + \beta_1 x + \varepsilon]$$

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} + 0$$

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

→ Use sample means of X & Y

Gauss-Markov Thm: The best (minimum variance) unbiased estimators for β_0 & β_1 are given by using the "least squares best fit" line (from MAT 210).

Brief Reminder of MAT 210:

$$\beta_0 + \beta_1 x = y$$

plug in sample values

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \end{bmatrix}$$

convert to normal equation

$$\begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma (x_i^2) \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \Sigma y_i \\ \Sigma (x_i y_i) \end{bmatrix}$$

invert matrix

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \frac{1}{n\Sigma(x_i^2) - (\Sigma x_i)^2} \begin{bmatrix} \Sigma(x_i^2) & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix} \begin{bmatrix} \Sigma y_i \\ \Sigma (x_i y_i) \end{bmatrix}$$

multiply to get $\hat{\beta}_1$ (ignore $\hat{\beta}_0$ because we already have formula)

$$\hat{\beta}_1 = \frac{n\Sigma(x_i y_i) - (\Sigma x_i)(\Sigma y_i)}{n\Sigma(x_i^2) - (\Sigma x_i)^2}$$

divide top & bottom by n^2

$$= \frac{\frac{1}{n}\Sigma(x_i y_i) - \left(\frac{1}{n}\Sigma x_i\right)\left(\frac{1}{n}\Sigma y_i\right)}{\frac{1}{n}\Sigma(x_i^2) - \left(\frac{1}{n}\Sigma x_i\right)^2}$$

Result:

$$\hat{\beta}_1 = \frac{Cov[X,Y]}{Var[X]} = \frac{S_{xy}}{S_{xx}}$$

where $\begin{cases} S_{xy} = \Sigma(x_i y_i) - \frac{1}{n}(\Sigma x_i)(\Sigma y_i) \\ S_{xx} = \Sigma(x_i^2) - \frac{1}{n}(\Sigma x_i)^2 \end{cases}$

Notation is because S_{xy} & S_{xx} are "sums of squares"

$$S_{xy} = \Sigma(x_i y_i) - \frac{1}{n}(\Sigma x_i)(\Sigma y_i)$$

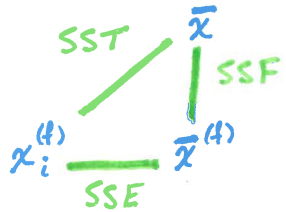
$$= \Sigma (x_i - \bar{x}) \cdot (y_i - \bar{y})$$ ← Divide by $(n-1)$ to get sample covariance.

$$S_{xx} = \Sigma(x_i^2) - \frac{1}{n}(\Sigma x_i)^2$$

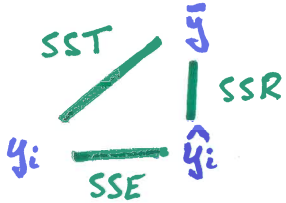
$$= \Sigma (x_i - \bar{x})^2$$ ← Divide by $(n-1)$ to get sample variance

Note: The computation of $\hat{\beta}_1$ used the same kind of stuff that showed up in Ch 10 ANOVA

→ To complete the connection we need to introduce "fitted"/"predicted" values $\hat{y}_i$ ← $\hat{\mu}_Y(x_i)$ (est. mean of Y at x_i)



SS in ANOVA



SS for Regression

Insert Note on Fitted Values $\hat{y}_i$

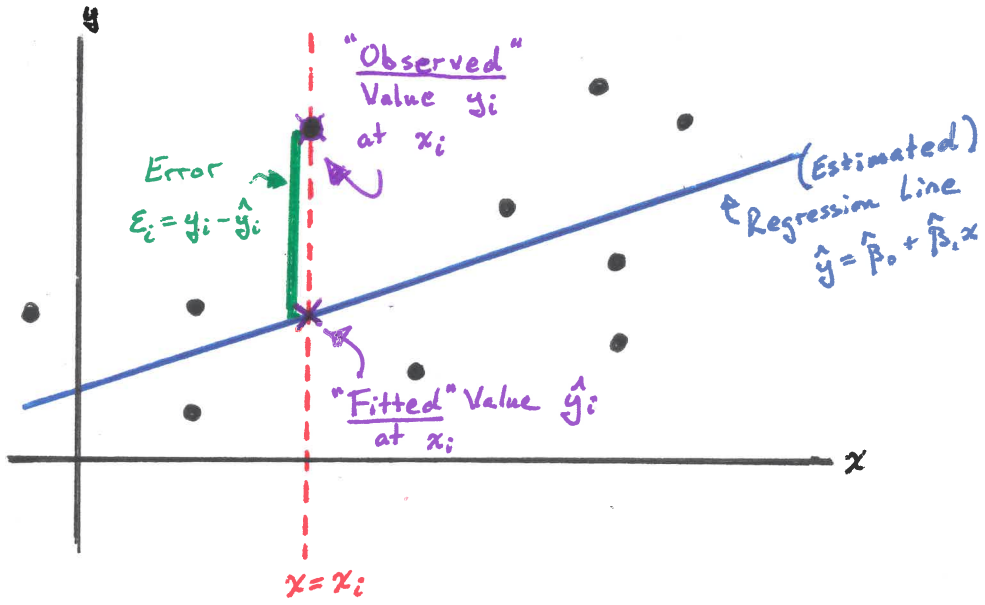
Note: If $Y(x) = (\beta_0 + \beta_1 x) + \varepsilon$

then at each x value

$$\hat{Y}(x) = E[(\hat{\beta}_0 + \hat{\beta}_1 x) + \varepsilon]$$

$$= \beta_0 + \beta_1 x$$

is a "point estimate" for Y.



- As a function, $\hat{Y}(x)$ is the regression line.
- As a value, $\hat{y}_i = \hat{Y}(x_i)$ is the mean of the $Y(x_i)$ distribution.

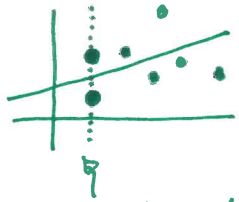
Def: (x_i, y_i) is the "observed" value of Y at x_i.

$(x_i, \hat{y}_i)$ is the "fitted" (or "predicted") value of Y at x_i.

Relating back to chapter 10, if x is a cont. factor, then $\hat{y}_i$ is the factor mean

$$\hat{y}_i = \mu_Y^{(x)} \longleftrightarrow \mu_x^{(i)} = \bar{x}^{(i)}$$

Note: The "observed" vs. "fitted" notation is slightly misleading because it is entirely possible for there to be two different y_i at the same x_i-value ...



Two "observed" y_i at same x_i-value.

Def: Given sample data (x_i, y_i)

- $S_{xx} = \sum (x_i^2) - \frac{1}{n}(\sum x_i)^2 = \sum (x_i - \bar{x})^2$ ← Divide by $n-1$ to get s_x^2 (x sample var)
  (like s_x^2)
- $S_{yy} = \sum (y_i^2) - \frac{1}{n}(\sum y_i)^2 = \sum (y_i - \bar{y})^2$ ← Divide by $n-1$ to get s_y^2 (y sample var)
  (like s_y^2)
- $S_{xy} = \sum (x_i y_i) - \frac{1}{n}(\sum x_i)(\sum y_i) = \sum (x_i - \bar{x})(y_i - \bar{y})$ ← Divide by $n-1$ to get sample covar.
- $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ $\qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
  (like ν_x)
- Fitted (or "predicted") values of y are
  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ ← Expected value of Y at x_i $\left(\mu_Y(x_i) = E[Y(x_i)]\right)$
- Residual (or "error") at x_i is
  $\varepsilon_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ ← This is the same ε from §10.1
- Sum of Squared Error is
  $SSE = \sum (y_i - \hat{y}_i)^2 = \sum (y_i^2) - \hat{\beta}_0 (\sum y_i) - \hat{\beta}_1 \sum (x_i y_i)$ ← unpleasant computation...

Same as for ANOVA we define

- Total Sum of Squares is
  $SST = \sum (y_i - \bar{y})^2 = S_{yy}$!!
- Regression Sum of Squares is
  $SSR = \sum (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 S_{xy}$ ← $\frac{S_{xy}^2}{S_{xx}}$!!

As before, $SST = SSR + SSE$

SSR takes the place of SSF (Factor Sum of Squares)

Note: SSE can also be computed by
$SSE = SST - \hat{\beta}_1 S_{xy} = S_{yy} - \hat{\beta}_1 S_{xy}$ ← $\frac{S_{xx} S_{yy} - S_{xy}^2}{S_{xx}}$!!

Analyze Regression by combining everything into ANOVA table where

	Regression	Error	Total
# deg. of freedom	1	$n-2$	$n-1$

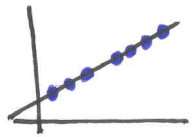
ANOVA Table: $y = \hat{\beta}_0 + \hat{\beta}_1 x$

	df	SS	MS	F-value	p-val
Regression	1	$\hat{\beta}_1 S_{xy}$	$\hat{\beta}_1 S_{xy}$	MSR/MSE	etc...
Error	$n-2$	$S_{yy} - \hat{\beta}_1 S_{xy}$	divide by $n-2$		
Total	$n-1$	S_{yy}	divide by $n-1$		

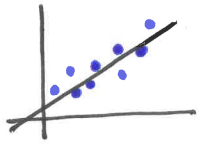
Tests $H_0: \beta_1 = 0$

Note: $R^2 = \dfrac{Cov[X,Y]^2}{Var[X] \cdot Var[Y]}$!!!

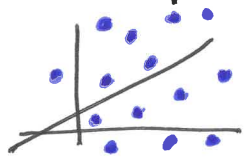
$R^2 = 1 \Rightarrow$ All sample points are on the regression line ($SSE = 0$)



$R^2 \approx 1 \Rightarrow$ Sample points are nearby regression line



$R^2 \approx 0 \Rightarrow$ Sample points are far from regression line



Usually analysis of Regression will also include the "effect size" from §10.2... but it is written using the complement and named something else....

Def: The coefficient of determination for regression is

$$R^2 = 1 - \frac{SSE}{SST}$$

$$= \frac{SSR}{SST}$$

In Chapter 10, SSR was SSF and this was called η^2 "Effect Size"

$$= \frac{S_{xy}^2 / S_{xx}}{S_{yy}}$$

$$= \frac{S_{xy}^2}{S_{xx} S_{yy}}$$ ← (correlation)^2

R^2 tells the proportion of total variance that is coming from the regression line moving the mean $\mu_Y^{(x)}$

R^2 is called a "Goodness of Fit" Statistic

→ low R^2 indicates that (perhaps) linear regression is not appropriate.

(There is also "adjusted R^2" ($\bar{R}^2$) used for non-linear regression)