8 October 6 In-class Work

Tuesday, October 6, 2020 11:05 AM

Simple Linear regression: relationship between two variables
- explain y in terms of x or how y varies w/ x

Example
- How sales vary w/ advertising
- Variation of crop yield w/ fertilizer input
- Salary vs years of education

Issues
- other factors need to be considered
- define functional form
- how to capture ceteris paribus (all factors equal) relation between x and y

X	Y
Independent var	dependent var
Explanatory var	Explained var
Control var	Response var
Predictor var	Predicted var
regressor	regressand

Functional Form of Linear regression

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

↓ Dependent ↓ Independent

Investigate the mean/average of response variable

Sales = $\beta_0 + \beta_1$(advertising) + ε

↳ if we spend x on advertising, how much do we get in sales (y)

↳ ε = other factors

Yield = $\beta_0 + \beta_1$(Fertilizer) + ε

↳ other factors such as land quality, rainfall, etc

β_1 = slope → measures change in Y for one unit change in X when all other factors are held constant

β_0 = intercept → value of Y when X = 0

Assumptions
1) Y/X is normally distributed
↳ for given X, Y is normally distributed
2) Average value of ε in the population is zero
3) ε and X are not correlated

Correlation varies between -1 and +1
↳ 0 = no correlation

ρ (rho) = correlation for population

r = correlation for sample

$r_{xy} = \Sigma(x-\bar{x})(y-\bar{y}) / \sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}$

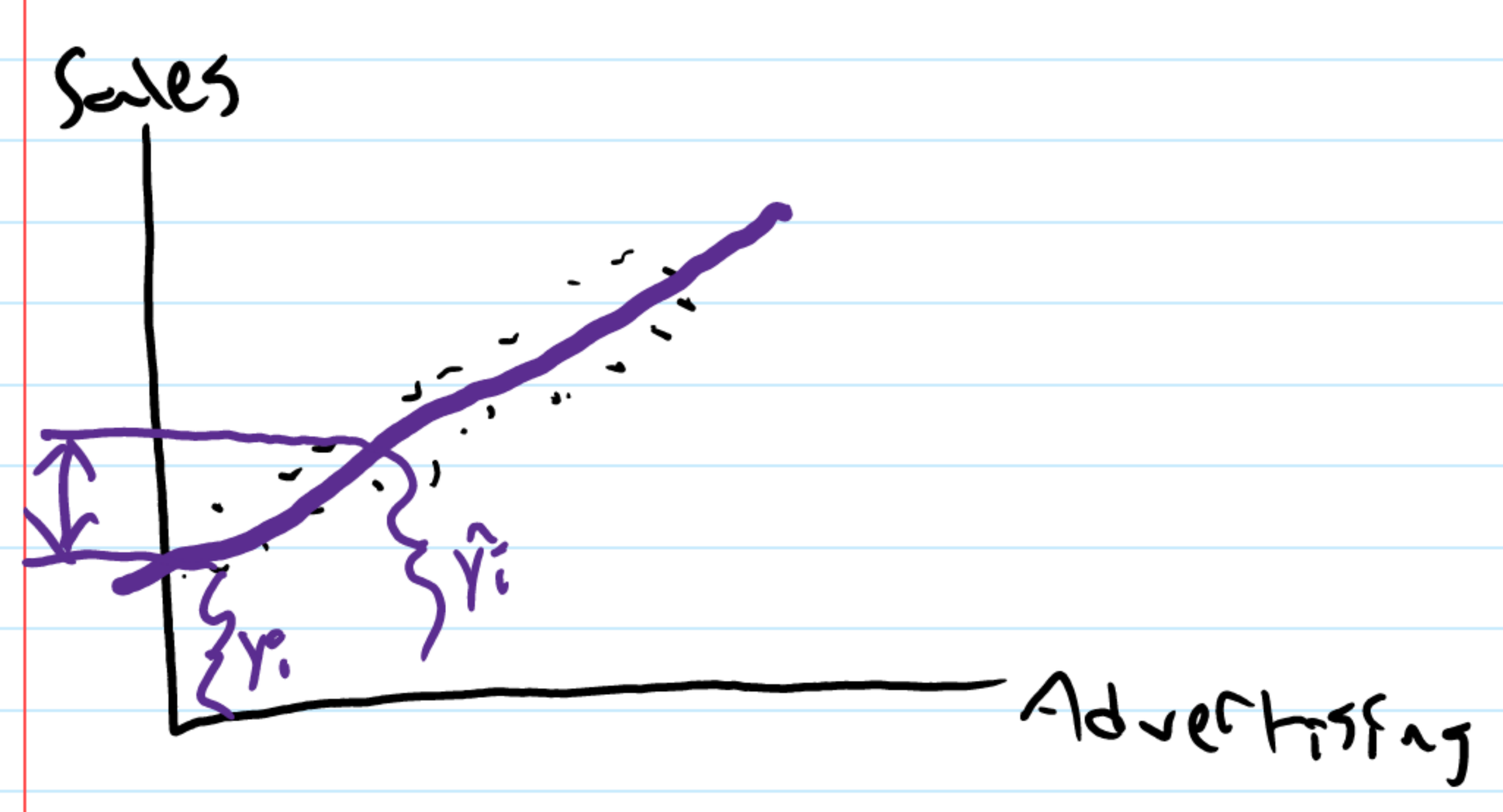
$E(\varepsilon|x) = E(\varepsilon) = 0$ → zero conditional mean assumption
↳ conditional distribution of ε given x
↳ Average of ε does not depend on x
defines value of β_0

example

$Y = 1.5 + .5X$

Y = College GPA X = high school GPA

4 GPA in high school has a 3.5 College GPA



$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x$

error $= Y_i - \hat{Y}_i$
$= Y_i - (\hat{\beta}_0 - \hat{\beta}_1 x)$

$\hat{Y}_i = \beta_0 + \beta_i x_i$

error $e_i = y_i - \hat{y}_i$
$= y_i - (\beta_0 + \beta_i x_i)$

Find B_0 and B_1 such that the square of error is minimum

$\min \Sigma e_i^2 = \min \Sigma [y_i - (\beta_0 + \beta_1 x_i)]^2$

$\frac{d}{d\beta_0} = 0 \quad \frac{d}{d\beta_1} = 0$

$\hat{\beta}_1$ (or b_1) $= \Sigma(x_i - \bar{x})(y_i - \bar{y}) / \Sigma(x_i - \bar{x})^2$

$\bar{X}$ and $\bar{Y}$ are means of X and Y

$\hat{\beta}_0 = b_0 = \bar{y} - b_1\bar{x}$ (or) $\bar{y} - \hat{\beta}_1\bar{x}$

$\hat{\beta}_1 = \partial y / \partial x$

Example → ordinary least square estimates (OLS estimate)

X = advertising Y = sales

X = {5, 2, 3, 1, 4} Y = {50, 15, 25, 10, 40}

$\bar{X} = 3$ $\bar{Y} = 28$

$X_i - \bar{X} = \{2, -1, 0, -2, 1\}$ $Y_i - \bar{Y} = \{22, -13, -3, -18, 12\}$

$(X_i - \bar{X})^2 = \{4, 1, 0, 4, 1\} = 10$ $(x_i - \bar{x})(y_i - \bar{y}) = \{44, 13, 0, 36, 12\} = 105$

$\hat{\beta}_1 = b_1 = 105/10$ → slope = 10.5

$b_0 = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 28 - (10.5 \cdot 3) = -3.5$

If you don't advertise, expect -3.5 in sales

Best fit regression is $Y = -3.5 + 10.5x$

Algebraic Properties of OLS estimates

1) Sum of residuals = 0
residuals = observed - fitted
$= \sum_{i=1}^{n} y - (\beta_0 + \beta_1 x_i)$
$= 0$

2) Covariance of x and Y residuals is 0
$\sum_{i=1}^{n} x_i e_i = 0$

3) $(\bar{x}, \bar{y})$ always fall on estimated regression line

4) Covariance between fitted values and residuals is zero
$e_i = y_i - \hat{y}_i$

Sum of Squares total

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

Total variation from the mean to be explained
measure of total sample variation

Sum of Squares residual

$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$

total variation explained by X

Sum of Square error

$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$

total variation left unexplained

SST = SSR + SSE

Goodness of Fit

how well X explains Y

R^2 = fraction of sample variation in Y explained by X

$R^2 = SSR/SST = 1 - SSE/SST$

Quick Review After Midterm

Determines relationship between X and Y