$Cov(X,Y) = E\big((X-\mu_X)(Y-\mu_Y)\big)$ $Cov > 0$: *on average* X, Y *above (below) mean together.*

$Cov < 0$: *on average, when* X *is above (below) mean,* Y *is below (above).*

(X,Y): *independent* $\Longrightarrow Cov(X,Y) = 0$ $Cov(X,Y) = 0 \nRightarrow (X,Y)$: *independent*

$Cov\big((a_1X+b_1),(a_2Y+b_2)\big) = a_1a_2Cov(X,Y)$ $Var(aX+bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$

$Corr(X,Y) = \rho_{XY} = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$ $-1 \le \rho \le 1$ (X,Y): *independent* $\Longrightarrow Corr(X,Y) = 0$

$Corr(X,Y) = 0 \nRightarrow (X,Y)$: *independent* $E(X) = \sum_{i=1}^{k} x_i f_i(x_i)$ $[or] \int_{-\infty}^{\infty} xf(x)dx$

$E\big(g(x)\big) = \sum_{i=1}^{k} g(x_i)f_i(x_i)$ $[or]$ $\int_{-\infty}^{\infty} g(x)xf(x)dx$ $E(c) = c$, $E(aX+b) = aE(X) + b$,

$E(X+Y) = E(X) + E(Y)$ $Var(X) = E(X-\mu)^2 = \sigma^2 = E(X^2) - \big(E(X)\big)^2$ $Var(c) = 0$, $Var(aX + b) = a^2Var(X)$ $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$ $\Phi(c) = z - table\ reading\ for\ number\ c$

$P(Z > z) = 1 - \Phi(z)$, $P(Z < -z) = P(Z > z)$, $P(a \le Z \le b) = \Phi(b) - \Phi(a)$

$P(|Z| > c) = P(Z > c) + P(Z < -c) = 2\big(1 - \Phi(c)\big)$ $E(Y|X=x) = \sum_j y_j f_{Y|X}(y_i|x)$ *or* $\int y_i f_{Y|X}(y_i|x)dy$

Simple linear regression model. $y = \beta_0 + \beta_1 x + u$ assuming $E\langle u|x\rangle = 0 \Longrightarrow E\langle y|x\rangle = \beta_0 + \beta_1 x$ $E(u) = 0 \Longrightarrow E(y - \beta_0 - \beta_1 x) = 0$ $Cov(x,u) = E(xu) = 0 \Longrightarrow E\big(x(y - \beta_0 - \beta_1 x)\big) = 0$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(y_i-\bar{y})}{\sum_{i=1}^{n} x_i(x_i-\bar{x})} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x$ $\hat{u}_i = y_i - \hat{y}_i$

$\sum_{i=1}^{n} \hat{u}_i = 0$ $\sum_{i=1}^{n} x_i\hat{u}_i = 0$ *The point* $(\bar{x},\bar{y})$ *is always on the regression line.*

$bias\big(\widehat{\beta_1}\big) = E\big(\widehat{\beta_1}\big) - \beta_1$ $E\big(\widehat{\beta_1}\big) = \beta_1 + E\left(\frac{\sum u_i x_i}{\sum x_i^2}\right) = \beta_1 + \frac{cov(u,x)}{var(x)} = \beta_1 + 0$

$\therefore bias\big(\widehat{\beta_1}\big) = \beta_1 - \beta_1 = 0$ ■

$SST = total\ sum\ of\ squares = total\ variation\ of\ y\ about\ its\ mean$ $\sum(y_i - \bar{y})^2$

$SST_X = total\ variation\ of\ x\ about\ its\ mean$ $\sum(x_i - \bar{x})^2$ $SSR = \sum(y_i - \hat{y}_i)^2$

$SSE = explained\ sum\ of\ squares = predicted\ variation\ of\ y\ about\ \bar{y} = \sum(\hat{y}_i - \bar{y})^2$

$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$ *shows fraction of variation in y explained by x*

$E\big(\hat{\beta}_0\big) = \beta_0$ $E\big(\hat{\beta}_1\big) = \beta_1$ $Var\big(\hat{\beta}_0\big) = \frac{\sigma^2\sum x_i^2}{nSST_x}$ $Var\big(\hat{\beta}_1\big) = \frac{\sigma^2}{SST_x}$

$[SST_x = \sum(x_i - \bar{x})^2]$ *Unbiased estimator of population variance:* $\hat{\sigma}^2 = \frac{1}{n-2}\sum\hat{u}_i^2 = \frac{SSR}{n-2}$

$SER = \hat{\sigma} = \sqrt{\hat{\sigma}^2}$ $SE\big(\hat{\beta}_0\big) = \sqrt{Var\big(\hat{\beta}_0\big)} = \sqrt{\frac{\sigma^2\sum x_i^2}{nSST_x}}$ $SE\big(\hat{\beta}_1\big) = \sqrt{Var\big(\hat{\beta}_1\big)} = \sqrt{\frac{\sigma^2}{SST_x}}$

modify dependent: $cy = c\beta_0 + c\beta_1 x$ *modify independent:* $y = \beta_0 + \left(\frac{\beta_1}{k}\right)(kx)$

OVB: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ *but we estimate* $y = \widetilde{\beta_0} + \widetilde{\beta_1}x_1$ *and* $\widetilde{x_2} = \widetilde{\delta_0} + \widetilde{\delta_1}x_1$ *then* $\widetilde{\beta_1} = \widehat{\beta_1} + \widehat{\beta_2}\widetilde{\delta_1}$ *then* $E\big(\widetilde{\beta_1}\big) = \beta_1 + \beta_2\widetilde{\delta_1}$ Including irrelevant vars maintains unbiasedness, but precision of estimates is bad.

Multicollinearity: correlation among expl vars leads to bias in coefficients

$Var(\widehat{\beta_J}) = \frac{\sigma^2}{SST_j(1-R_j^2)}$ $\quad where\ SST_j = \sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2 \quad and$

$test\ statistic: t = \frac{\widehat{\beta_1} - 0}{SE(\widehat{\beta_1})}$ $\quad null\ hypothesis\ H_0: \beta_1 = 0 \quad alt\ H_A: \beta_1 < 0,\ \beta_1 > 0,\ \beta_1 \neq 0$

$Y - Variables: Quantitative\ (Cardinal\ or\ Ordinal), or\ Qualitative\ (Ordered\ or\ Unordered)$

$Regression: Cardinal\ (OLS), Ordinal\ (try\ to\ make\ cardinal, if\ not, ordered\ logit),$

$Categorical\ 2\ groups\ (logit), Unordered > 2\ groups$

$(multinomial\ logit), Ordered > 2\ groups\ (ordered\ logit)$

$X\ causes\ Y: causality; Y\ causes\ X: reverse\ causality; both: simultaneous\ causality$

$Sources\ of\ Bias: OVB, Measurement\ error, Sample\ Selection\ Bias, Misspecification.$

$Overstated: |\widehat{\beta_1}| > \widehat{\beta_1} \quad Understated: |\widehat{\beta_1}| < \widehat{\beta_1}$

$OVB: Y = X + Z, but\ Z\ omitted. Biased\ when\ Corr(X,Z) \neq 0\ and\ Corr(Y,Z) \neq 0\ (hold\ X\ constant)$

	$\widehat{\beta_1} > 0$		$\widehat{\beta_1} < 0$	
	$Corr(Y,Z) > 0$	$Corr(Y,Z) < 0$	$Corr(Y,Z) > 0$	$Corr(Y,Z) < 0$
$Corr(X,Z) > 0$	Overstated	Understated	Understated	Overstated
$Corr(X,Z) < 0$	Understated	Overstated	Overstated	Understated

$Error\ in\ X: Attenuation\ (slope\ understated), Imprecision\ (large\ std\ errors)$

$Errors\ in\ Y: None\ (slope\ unbiased), Imprecision\ (large\ std\ errors)$

$Coefficient\ interpretation\ with\ quadratic\ X\ (\ldots \beta_1 X + \beta_2 X^2 \ldots):$

$\beta_1 > 0,\ \beta_2 > 0: incr\ at\ incr\ rate \quad \beta_1 > 0,\ \beta_2 < 0: incr\ at\ decr\ rate$

$\beta_1 < 0,\ \beta_2 > 0: decr\ at\ incr\ rate \quad \beta_1 < 0,\ \beta_2 < 0: decr\ at\ decr\ rate$

$lin - lin: y = \beta_0 + \beta_1 x \qquad \beta_1 = unit\ change\ in\ y\ for\ a\ 1\ unit\ change\ in\ x$

$lin - log: y = \beta_0 + \beta_1 \ln x \qquad \frac{\beta_1}{100} = unit\ change\ in\ y\ for\ a\ 1\%\ change\ in\ x$

$log - lin: \ln y = \beta_0 + \beta_1 x \qquad 100\beta_1 = \%\ change\ in\ y\ for\ a\ 1\ unit\ change\ in\ x$

$log - log: \ln y = \beta_0 + \beta_1 \ln x \qquad \beta_1 = \%\ change\ in\ y\ for\ a\ 1\%\ change\ in\ x$

$Difference\ in\ Differences: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 * X_2) + \varepsilon$

	$X_2 = 0$	$X_2 = 1$	Difference
$X_1 = 1$	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$
$X_1 = 0$	β_0	$\beta_0 + \beta_2$	β_2
		Difference in Differences:	β_3

$Good\ instrument: Relevant\ (Corr(X,Z) \neq 0)\ and\ Exogenous\ (Corr(Z,\varepsilon) \neq 0)$

$Pvalue = .02 \Longrightarrow in\ one\ out\ of\ fifty\ samples, we\ would\ observe\ result\ if\ H_0: true.$