Lecture 7: Correlation + Regression

→ correlation does NOT imply causation

- two variables may be positively correlated because they're influenced by a third variable
- but here we'll consider the case of two variables

→ scatter plot

- each point represents a pair of x & y values
- we're not connecting the points

→ correlation coefficient

- $$r_{xy} = \frac{cov(x,y)}{\sqrt{var(x)}\cdot\sqrt{var(y)}} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$$

- Note: if $y_i = \alpha x_i \ \forall i$ then $r_{xy} = sign(\alpha)\cdot 1$

→ note that correlation coefficient may be zero if relationship is non-linear

The Gauss-Markov Assumptions

$$y = \alpha + \beta x + \varepsilon$$

1. $E[\varepsilon] = 0$

   expected value of the error term equals zero

2. $E[\varepsilon^2] = \sigma^2 < \infty$

   constant + finite variance of error term

3. $cor(\varepsilon, x) = 0$ ← zero correlation between error + regressors

4. $cor(\varepsilon_i, \varepsilon_j) = 0$ ← zero correlation among error terms

5. $cor(x_A, x_B) = 0$ ← zero correlation among regressors

6. **linear model**

   - a percentage should not be dependent variable because

$$0 < pct < 1$$

Minimizing Sum of Squares

$$y = \alpha + \beta x + \varepsilon$$

$$\varepsilon = y - \alpha - \beta x$$

$$\underset{\alpha,\beta}{\text{Min}} \quad \Sigma \varepsilon^2 = \Sigma (y - \alpha - \beta x)^2$$

$$\frac{\partial \Sigma \varepsilon^2}{\partial \alpha} = -2 \Sigma (y - \alpha - \beta x) = 0$$

necessary condition #1

two implications:

a. $\frac{1}{N} \Sigma \varepsilon = 0 \iff E[\varepsilon] = 0$

b. $\bar{y} = \hat{\alpha} + \hat{\beta} \bar{x} \implies \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$

$$\frac{\partial \Sigma \varepsilon^2}{\partial \beta} = -2 \Sigma (y - \alpha - \beta x) x = 0$$

necessary condition #2

two implications:

a. $\frac{1}{N} \Sigma \varepsilon x = 0 \iff cov(\varepsilon, x) = 0$

note: $cov(\varepsilon, x) = \frac{1}{N} \Sigma \varepsilon x - \bar{\varepsilon} \bar{x}$

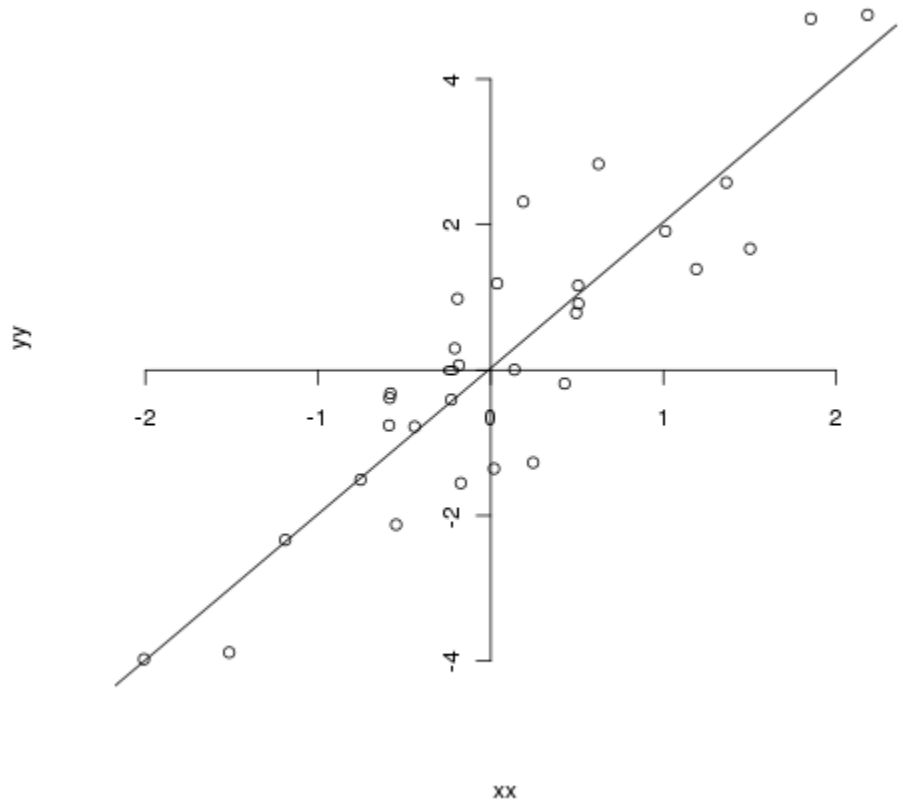
but $\bar{\varepsilon} = 0 \quad \therefore \quad cov(\varepsilon, x) = 0$

b. $$\hat{\beta} = \frac{\frac{1}{N} \Sigma y x - \bar{y} \bar{x}}{\frac{1}{N} \Sigma x^2 - \bar{x}^2} = \frac{\Sigma (y - \bar{y})(x - \bar{x})}{\Sigma (x - \bar{x})^2} = \frac{cov(x, y)}{var(x)}$$

```
R version 2.11.1 (2010-05-31)
Copyright (C) 2010 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> ## quick R script to illustrate correlation and regression
>
> ## randomly generate X and Y with positive relation
> xx <- rnorm(30)
> yy <- 2*xx + rnorm(30)
>
> ## correlation
> cor(xx,yy)
[1] 0.8966973
>
> ## regression
> ols <- lm(yy~xx)
>
> ## plot the relation and
> ## insert regression line into plot
> png( "scatterplot.png" )
> plot( xx ,yy , axes = FALSE )
> axis(1 , pos = 0 )
> axis(2 , pos = 0 )
> abline( a = coef(ols)[1] , b = coef(ols)[2] )
> dev.off()
null device
          1
>
> ## how did we get coefficients?
> ## we minimized the sum of squares
> beta <- seq( 0.5 , 3.5 , 0.01 )
> sqerr <- NA
> for (i in 1:length(beta)) {
+   alfa <- mean(yy) - beta[i]*mean(xx)
+   sqerr[i] <- sum((yy - alfa - beta[i] * xx)^2)
+ }
>
> ## show sum of squares plotted over different possible values of beta
> png( "min-sum-squares.png" )
> plot( beta , sqerr , type = "l" , axes = FALSE )
> axis(1)
> axis(2)
> dev.off()
null device
          1
>
```

```
> ## find beta which minimizes sum of squares
> betahat <- beta[which(sqerr == min(sqerr))]
>
> ## corresponding alpha
> alfahat <- mean(yy) - betahat*mean(xx)
>
> ## compare coefficients with regression output
> alfahat
[1] 0.02580671
> betahat
[1] 2.01
>
> ## summarize regression
> summary(ols)

Call:
lm(formula = yy ~ xx)

Residuals:
     Min       1Q   Median       3Q      Max
-1.79051 -0.72444  0.02788  0.63851  1.91080

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02611    0.17507   0.149    0.883
xx           2.00706    0.18724  10.719 2.03e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.953 on 28 degrees of freedom
Multiple R-squared: 0.8041,     Adjusted R-squared: 0.7971
F-statistic: 114.9 on 1 and 28 DF,  p-value: 2.032e-11

>
> ## the R-squared:
> r2 <- 1 - (var(ols$residuals)/var(yy))
> r2
[1] 0.804066
>
> ## square root of R-squared is equal to the correlation coefficient
> sqrt(r2)
[1] 0.8966973
> cor(xx,yy)
[1] 0.8966973
>
> proc.time()
   user  system elapsed
  0.308   0.040   0.333
```

scatterplot and regression line



sum of squares as a function of “beta