Data Science

Transforming Data into Knowledge and Vision
Understanding the Power and Beauty of Data

Linear Regression

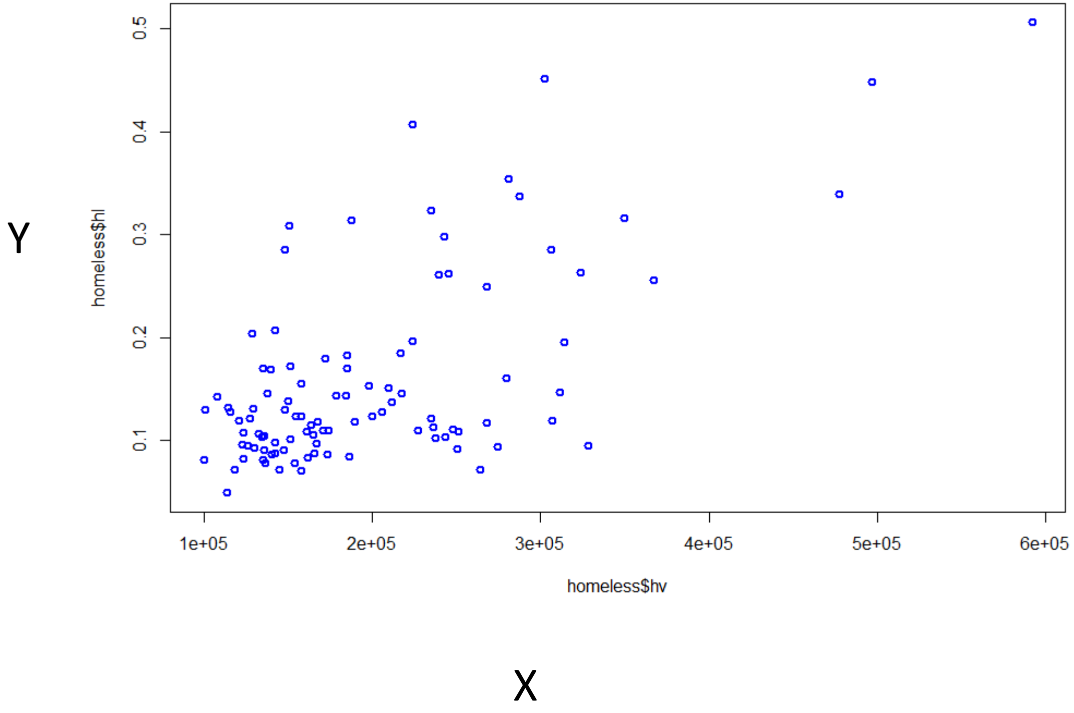
William Yu, PhD

Economist

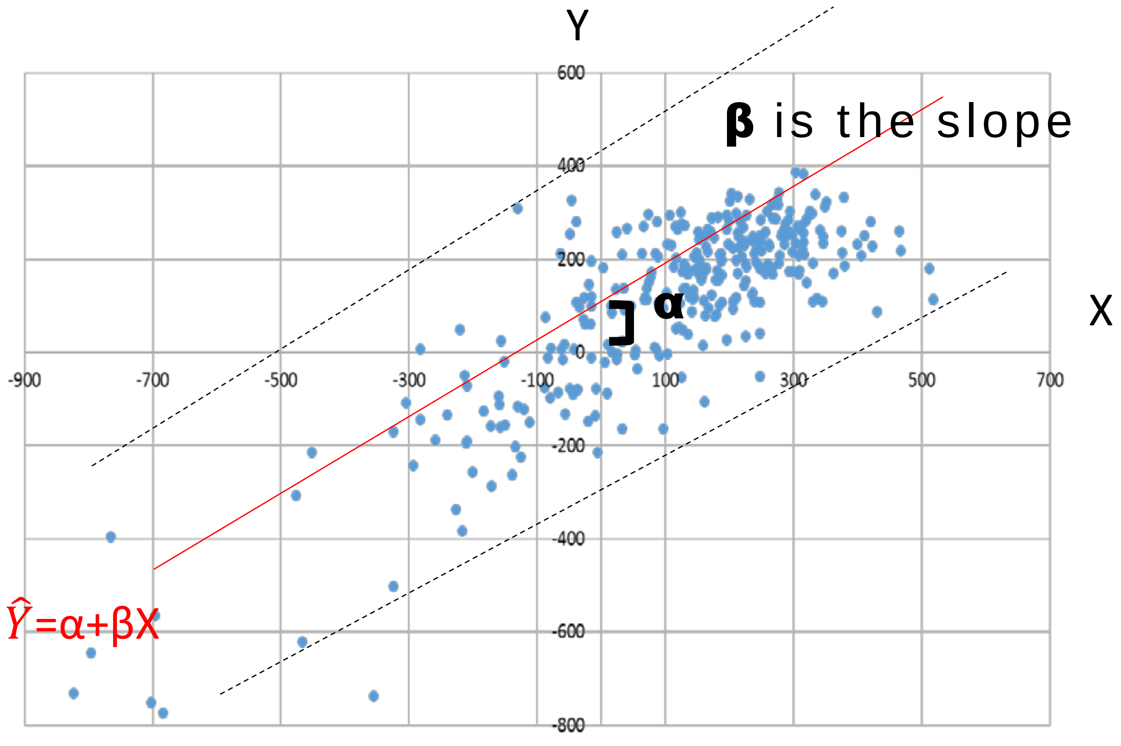
UCLA Anderson Forecast

From correlation to regression

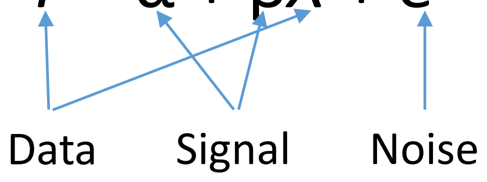
- Correlation between X and Y has no direction.
- But in regression, we *imply* the direction is from X to Y.
- $Y = \alpha + \beta X + e$
- e should be as much as like noise / random number.
- X should be as exogenous as possible.
  - Exogeneity of X means that x is more of a cause than a result.
  - We assume that X causes Y rather than Y causes X.
- X: Independent variable, RHS (right-hand-side) variable, regressors, explanatory variable, predictor
- Y: Dependent variable, LHS variable



Introduction to the linear regression



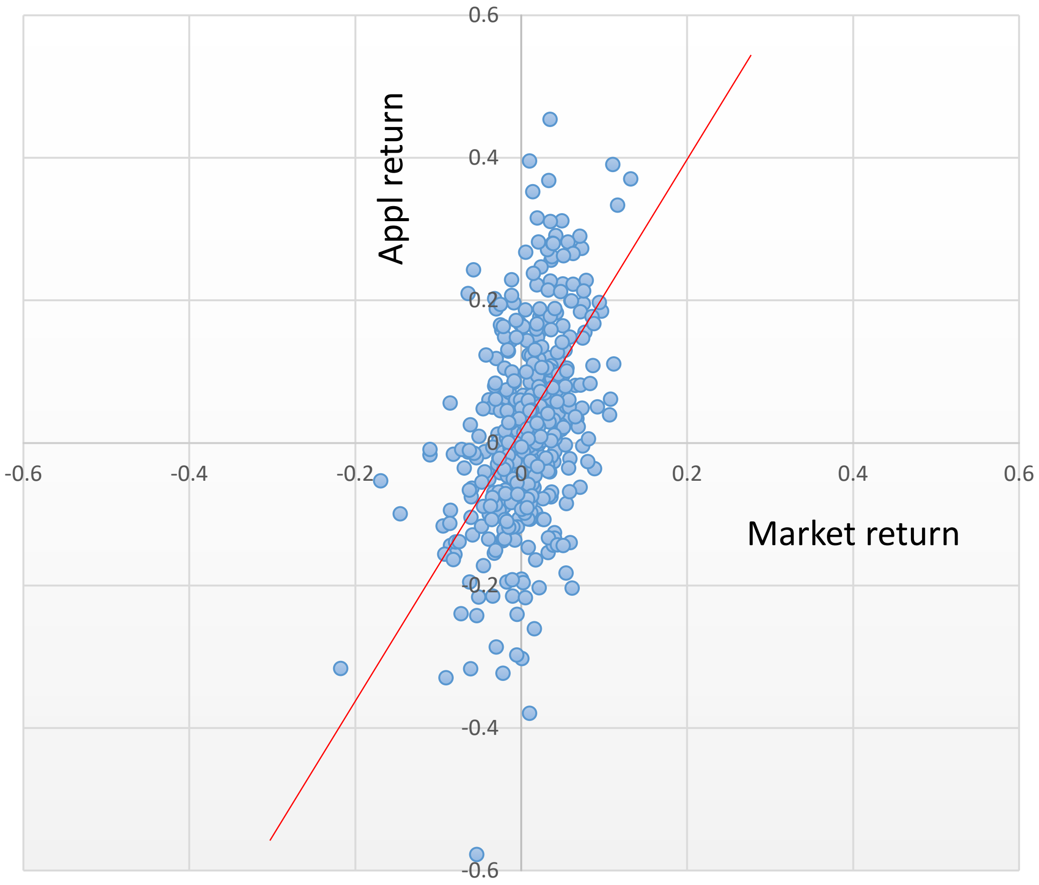
- Linear regression – Ordinary Least Square (OLS) method
- To find(fit) a linear line which minimizes the distance (error, noise) of data points to the fitted line
- Y = α + βX + e



Data Signal Noise

$$Min \sum (e)^2 = Min \sum (Y - \hat{Y})^2$$

Introduction to the linear regression



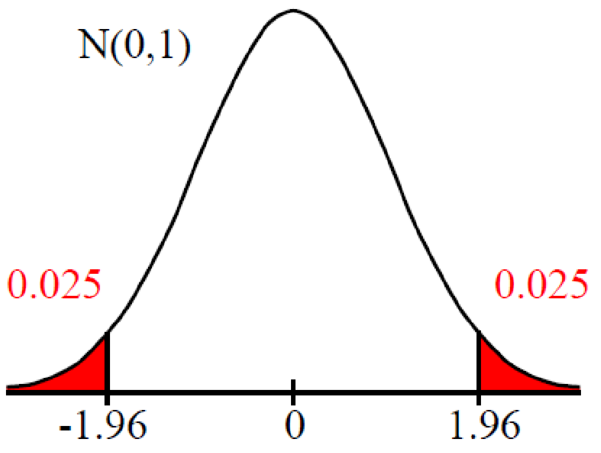
- $Y = \alpha + \beta X + e$
- Y = apple monthly stock return, X = S&P 500 return, 1981m01 to 2016m01
- The red line is the regression line (fit line: $\hat{Y} = \alpha + \beta X$)

	Coefficient	Standard Error	t Stat	P-Value
α	0.01	0.01	1.96	0.05
β	1.40	0.14	10.16	0.00
Adjusted R Square	0.2	Observations		421

- In fact, this regression is similar to the most important theory/pattern in finance, which is called CAPM.

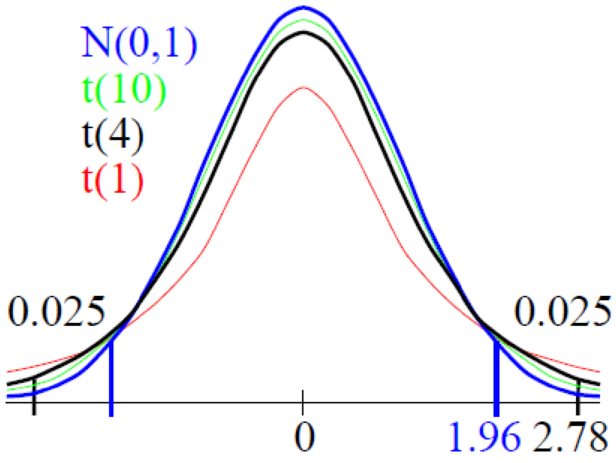
Why t test is so important?

- Standard normal distribution
- When population S.D. (σ) is known



$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

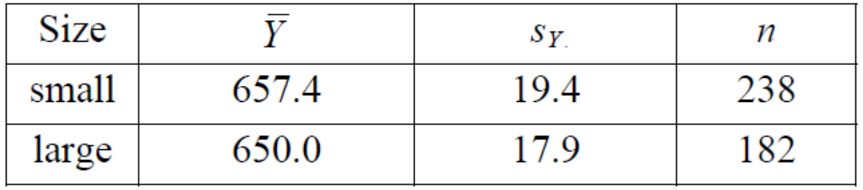
- Student t distribution
- When population S.D. (σ) is unknown



$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{\bar{X} - \mu_0}{\sqrt{\frac{1}{n-1}\Sigma_{k=1}^{n}(X_k - \bar{X})^2}/\sqrt{n}}$$

Two important statistics terms

- Standard errors (S.E.)
  - Standard deviation of an estimator
  - The smaller the S.E., the more precise of the estimator
- T-statistics
  - t = $\frac{\beta - 0}{S.E.(\beta)}$
  - Difference-in-means test:



Size	$\bar{Y}$	$s_{Y.}$	n
small	657.4	19.4	238
large	650.0	17.9	182

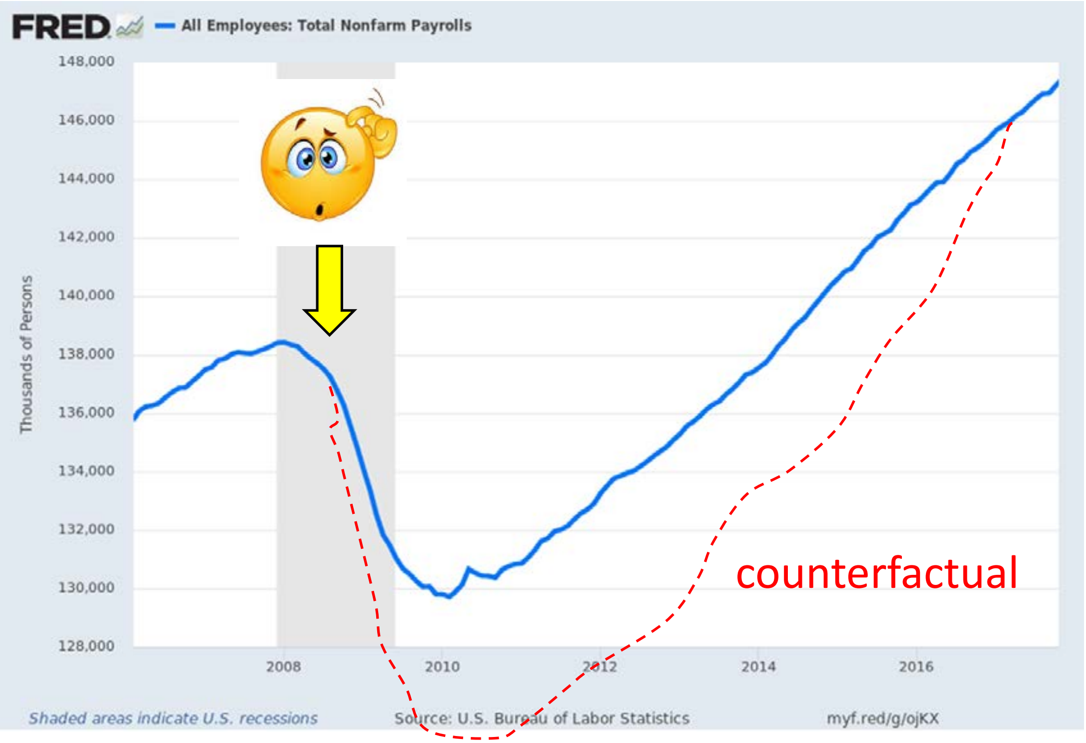
$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)}$$

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{657.4 - 650.0}{\sqrt{\frac{19.4^2}{238} + \frac{17.9^2}{182}}} = \frac{7.4}{1.83} = 4.05$$

|t| > 1.96, so reject (at the 5% significance level) the null hypothesis that the two means are the same.

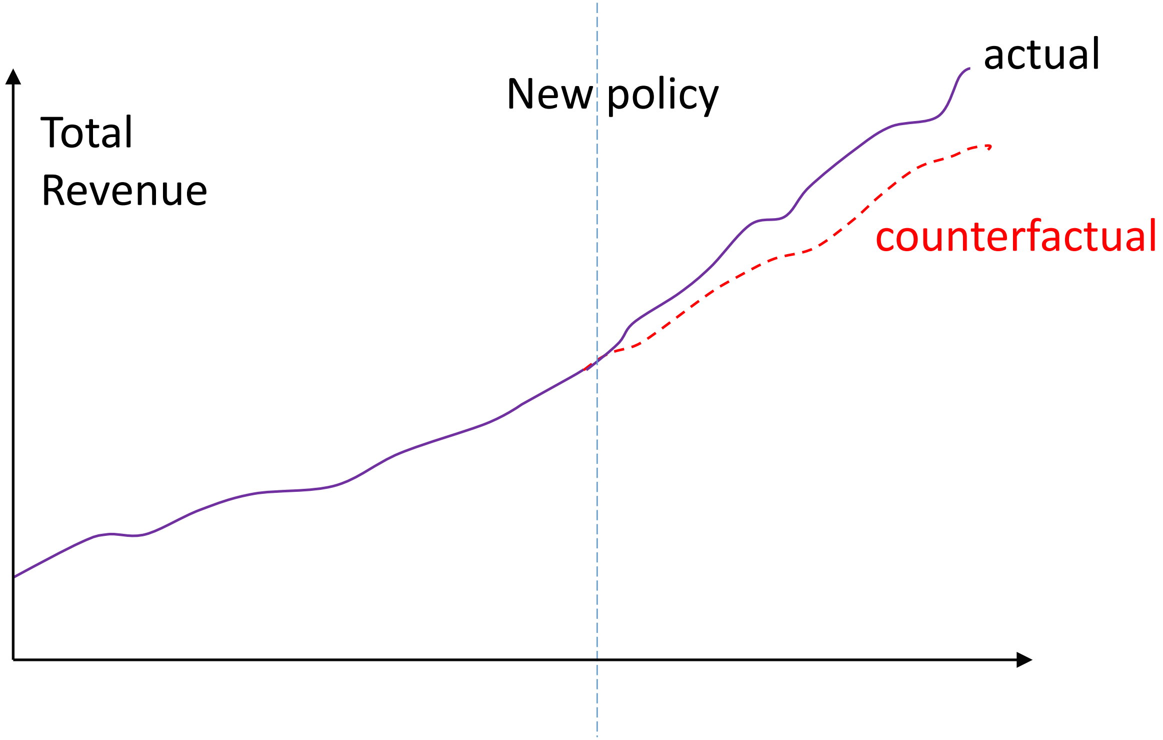
A/B test, policy effectiveness, and counterfactuals

Q1. Did federal gov’t stimulus in 2008/2009 help our economy?



Q2. Did new pricing policy increase total revenue of Warner Brother movies in streaming?

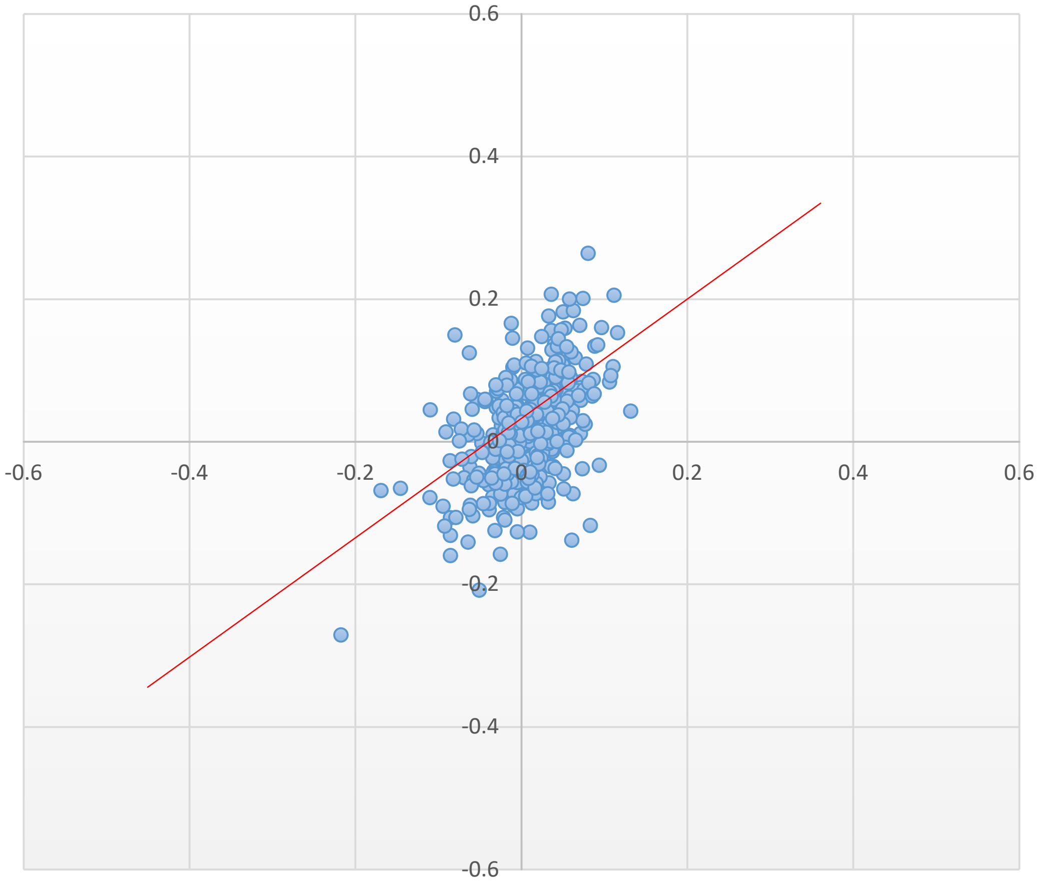
Note: You need to predict the counterfactual first.



Review of statistics and regression

- $Y = \alpha + \beta X + e$
    - Data: X,Y. Signal: α and β. Noise: e
- Statistically Significant
    - To tell if the parameters: α and β are significantly different from zero.
    - If coefficient is large, standard error (SE) is small, |t-stat| is large than 2 (tstat<-2 or >2), p-value is small (usually less than 0.05), then we say the parameter is statistically significant (different from 0).
    - T stat = coefficient / SE
    - SE is the standard deviation (SD) of an estimator/coefficient.
    - Statistical significance ≠ economic significance.

Now replace Y of Apple return with Walmart (WMT) return



	Coefficient	Standard Error	t Stat	P-Value
α	0.01	0.00	3.58	0.00
β	0.81	0.07	12.0	0.00
Adjusted R Square	0.25	Observations		421

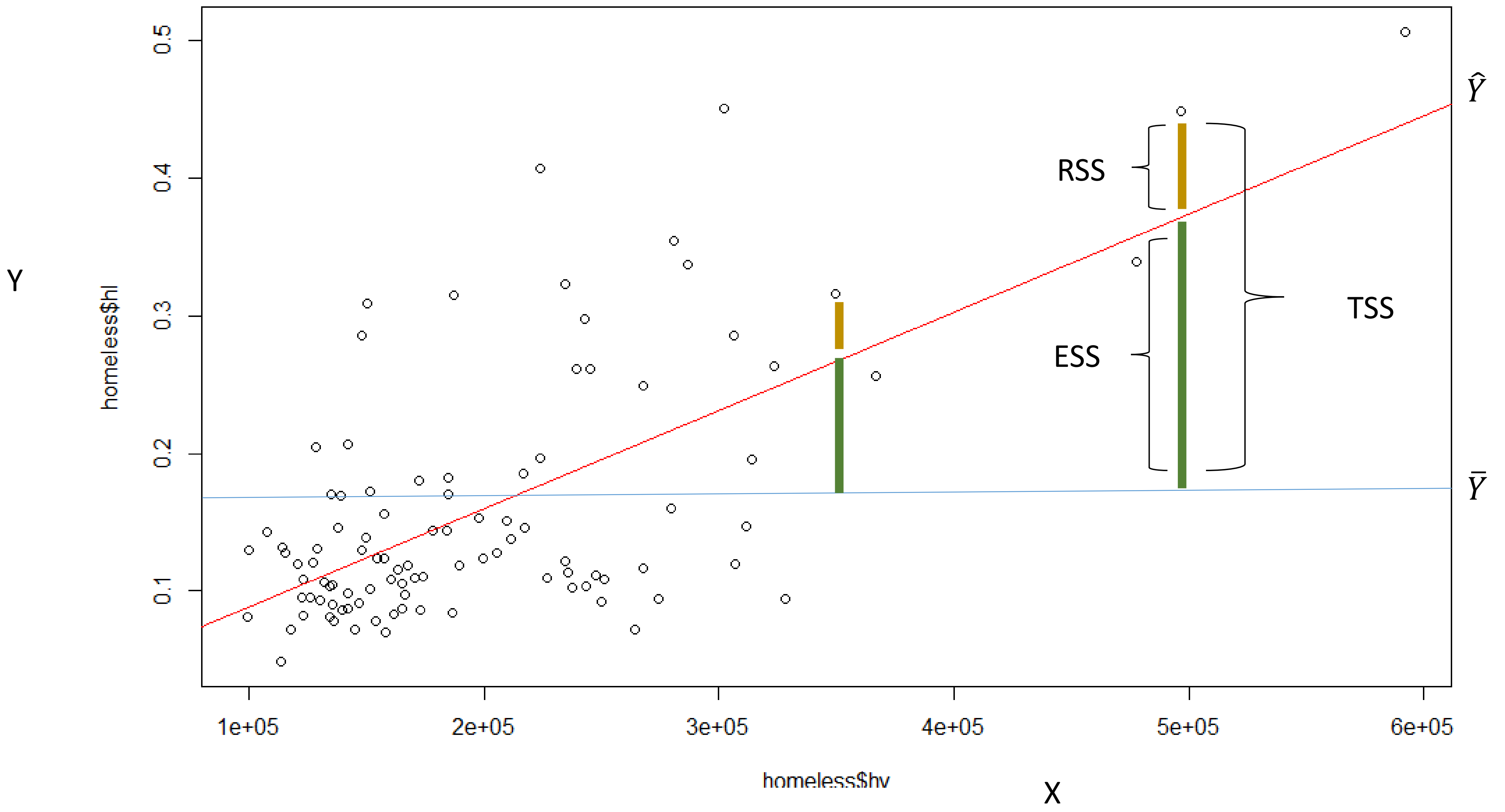
Variation of x (independent variables, right-hand-side variables) could explain 25% of the variation of y (dependent variable, left-hand-side variable)

Can the past/current stock market returns predict the future returns?

- CAPM regression tells us that the current market return can somewhat concurrently show us the individual stock return. But it is not the prediction. We cannot make money out of it.
- $Y = \alpha + \beta X + e$
- Can the current stock return tell us the future ones?
- $X_t = \alpha + \beta X_{t-1} + e_t$
  - Regression, finding pattern/parameter
- $X_{t+1} = \alpha + \beta X_t + e_{t+1}$
  - Prediction, using pattern and current data (X_t) to forecast the future (X_{t+1})

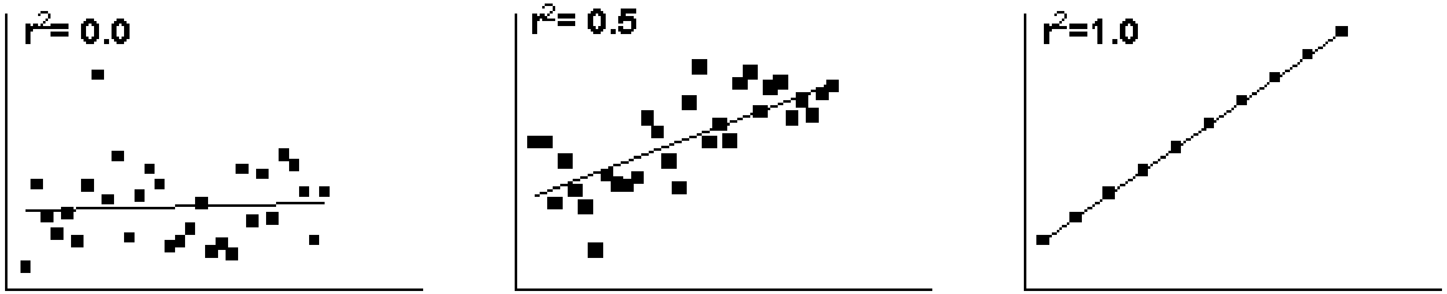
	Coefficient	Standard Error	t Stat	P-Value
α	0.01	0.00	3.34	0.00
β	0.05	0.05	0.95	0.34
Adjusted R Square	0	Observations		420

Statistically Insignificant

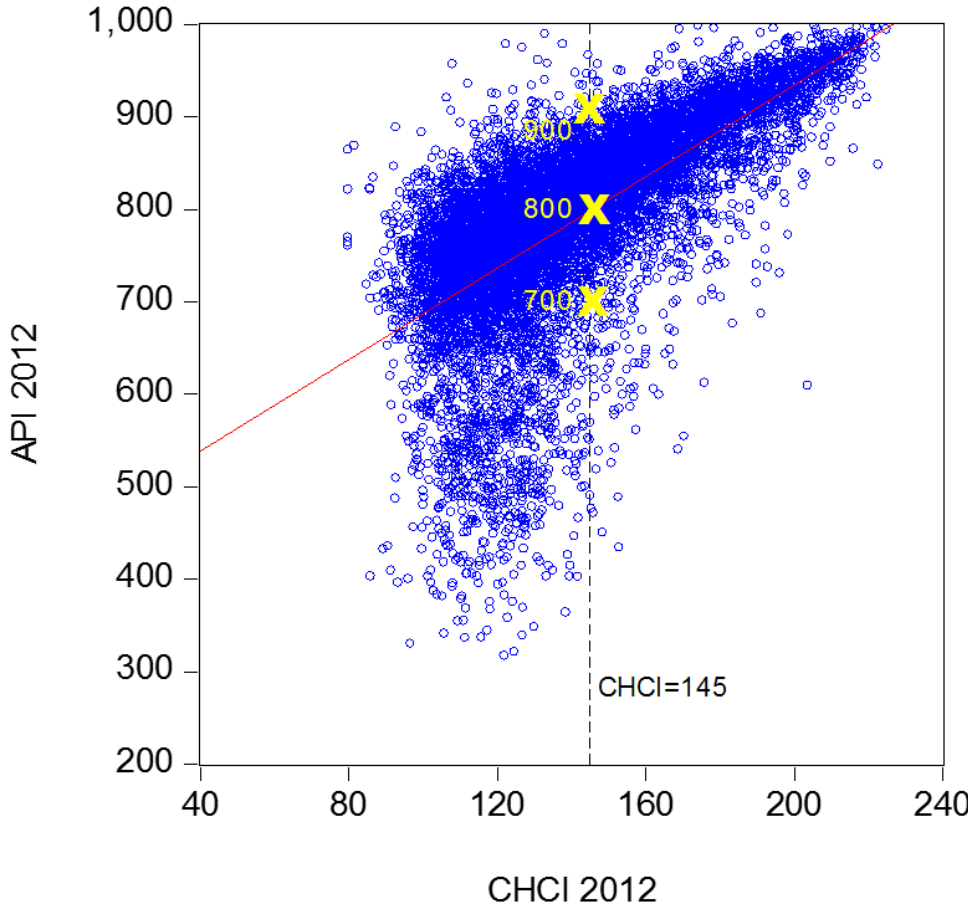

Y
homeless$hl
0.1
0.2
0.3
0.4
0.5
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
homeless$hv
X
RSS
ESS
TSS
$\hat{Y}$
$\bar{Y}$

Percentage of variability of Y explained by the model

- *R* Squared
  - A goodness of in-sample fit
  - Explained by model (signal) $\sum(\hat{Y}-\overline{Y})^2$= ESS Explained variation
  - Unexplained (noise) $\sum(Y-\hat{Y})^2$ = RSS/SSR Residuals sum squared
  - Total $\sum(Y-\overline{Y})^2$= TSS Total variation
  - R squared = 1 - $\frac{RSS}{TSS}$
  - Adj. R squared = 1 - $\frac{T-1}{T-k}\frac{RSS}{TSS}$
  - SER = $\sqrt{\frac{RSS}{T-k}}$
  - SER: stand error of the regression



A multivariable regression



Data: http://www.anderson.ucla.edu/centers/ucla-anderson-forecast/projects-and-partnerships/city-human-capital-index

- $Y = \alpha + \beta X + e$
- Y= API, Academic Performance Index
- X=CHCI, City Human Capital Index
- API = 433 + 2.5*CHCI

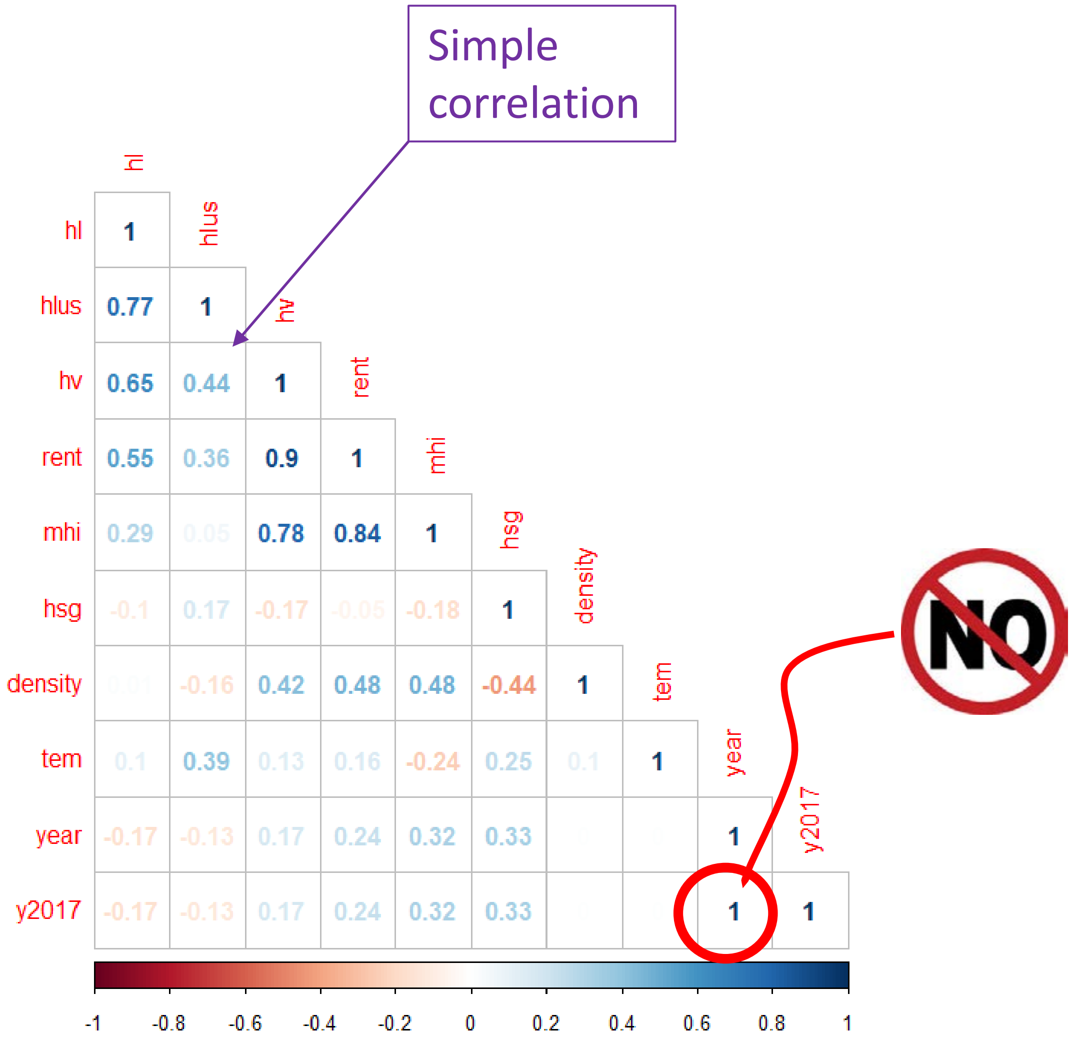
  (t-stat) (70) (66) Adj R2: 0.46
- Can we add more RHS variables?
- Yes. $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e$
- β_1 means that holding X2 and X3 constant (controlling X2 and X3), how change of X1 predicts Y.

Simple correlation

Partial correlation

Multicollinearity

- A state of very high intercorrelations among the independent variables. Reasons:
  - By inaccurate use of dummy variables
  - By inclusion of a variable which is computed from other variables in the data
  - From repetition of the same kind of variable
  - It generally occurs when the variables are highly correlated to each other
- Remember that the meaning of multivariate regression: after controlling other variables, how change of x1 impact y.
- Result: (1) one coefficient would be dropped. (2) error messages show up.



Simple correlation

	hl	hlus	hv	rent	mhi	hsg	density	tem	year	y2017
hl	1									
hlus	0.77	1								
hv	0.65	0.44	1							
rent	0.55	0.36	0.9	1						
mhi	0.29	0.05	0.78	0.84	1					
hsg	-0.1	0.17	-0.17	-0.05	-0.18	1				
density	0.01	-0.16	0.42	0.48	0.48	-0.44	1			
tem	0.1	0.39	0.13	0.16	-0.24	0.25	0.1	1		
year	-0.17	-0.13	0.17	0.24	0.32	0.33			1	
y2017	-0.17	-0.13	0.17	0.24	0.32	0.33			1	1

Variance inflation factor (VIF)

- VIF $(\hat{\beta}_j)$ = $\frac{1}{1-R^2_{X_j|X_{-j}}}$
- Where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.
- If $R^2_{X_j|X_{-j}}$ is close to one, then collinearity is present and so the VIF will be large.
- If VIF is above 5 or 10, it means that the regression model might have multicollinearity problem.

Extensions to the linear model

- Y$= \alpha + \beta_1 X_1$ + $\beta_2 X_2$ + e
- Predictors have synergy or interaction effect
- Removal the additive assumption of predictors
- Y$= \alpha + \beta_1 X_1$ + $\beta_2 X_2$ + $\beta_3 X_1 X_2$+ e
- Y$= \alpha + \beta_1 X_1$ + $(\beta_2$+$\beta_3 X_1) X_2$ + e
- Non-linear model
- Y$= \alpha + \beta_1\ X + \beta_2\ X^2 + \beta_3\ X^3 + \beta_4\ X^4$ +…+ e