Ch14: Linear Correlation and Regression

- *Please download:* *09-Regression.xls*
- ***HW6** this week*
- ***Projects***

13 Mar 2012
Dr. Sean Ho

busi275.seanho.com

TRINITY WESTERN UNIVERSITY

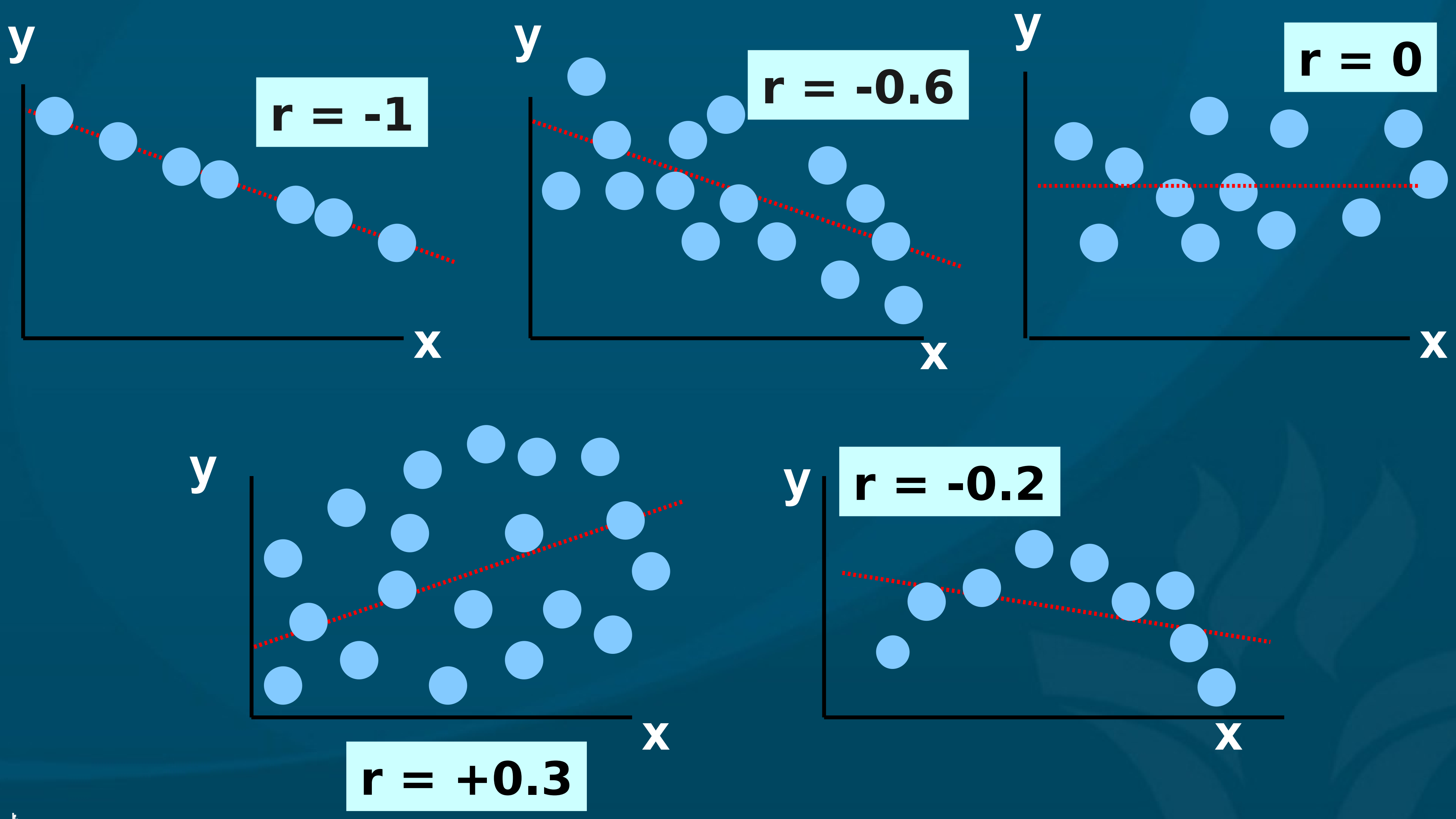
Outline for today

- Correlation
  - r^2 as fraction of variability
  - t-test on correlation
- Prediction using Simple Linear Regression
  - Linear regression model
  - Regression in Excel
  - Analysis of variance: Model vs. Residual
  - The global F-test
  - T-test on slope b_1
  - Confidence intervals on predicted values

Linear correlation

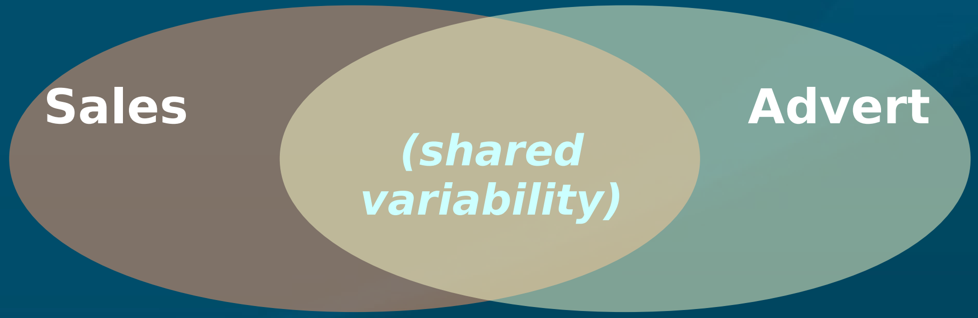
- Correlation measures the strength of a linear relationship between two variables
- Does not determine direction of causation
- Does not imply a direct relationship
  - There might be a mediating variable (e.g., between ice cream and drownings)
- Does not account for non-linear relationships
- The Pearson product-moment correlation coefficient (r) is between -1 and 1
  - Close to -1: inverse relationship
  - Close to 0: no linear relationship
  - Close to +1: positive relationship

Correlation on scatterplots



Correlation is an effect size

- We often want to understand the variance in our outcome variable:
  - e.g., sales: why are they high or low?
- What fraction of the variance in one variable is explained by a linear relationship w/the other?
  - e.g., 50% of the variability in sales is explained by the size of advertising budget



- The effect size is r^2: a fraction from 0% to 100%
  - Also called the coefficient of determination

Correlation: t-test

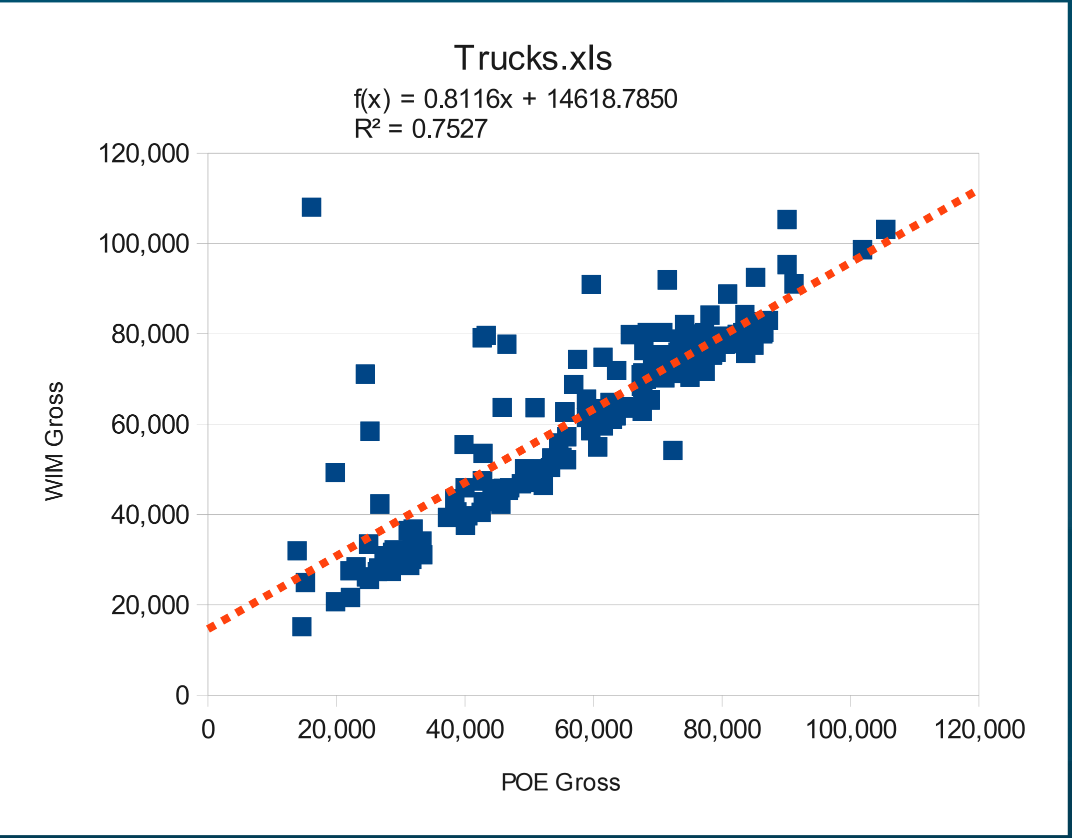
- r is sample correlation (from data)
- ρ is population correlation (want to estimate)
- Hypothesis: H_A: $\rho \neq 0$ (is there a relationship?)
- Standard error: $$SE = \sqrt{\frac{1-r^2}{df}}$$
  - $1 - r^2$ is the variability not explained by the linear relationship
  - df = $n-2$ because we have two sample means
- Test statistic: $t = r / SE$
  - Use TDIST() to get p-value

Correlation: example

- e.g., is there a linear relationship between caffeine intake and time spent in Angry Birds?
  - H_A: $\rho \neq 0$ (i.e., there is a relationship)
  - Data: 8 participants, r = 0.72
- Effect size: $r^2 = 0.72^2 = 51.84\%$
  - About half of variability in AB time is explained by caffeine intake
- Standard error: $SE = \sqrt{(1-0.5184) / 6} \approx 0.2833$
- Test statistic: $t = 0.72 / 0.2833 \approx 2.54$
- P-value: TDIST(2.54, 6, 2) → 4.41%
- At α=0.05, there is a significant relationship

Correlation: Excel

- Example: "Trucks" in 09-Regression.xls
- Scatterplot: POE Gross (G:G), WIM Gross (H:H)
- Correlation: CORREL(*dataX*, *dataY*)
  - Coefficient of determination: r^2
- T-test:
  - Sample r
  - → SE
  - → t-score
  - → p-value



Correl. and χ^2 independence

- Pearson correlation is for two quantitative (continuous) variables
- For ordinal variables, there exists a non-parametric version by Spearman (r_s)
- What about for two categorical variables?
  - χ^2 test of goodness-of-fit (ch13)
  - 2-way contingency tables (pivot tables)
  - Essentially a hypothesis test on independence

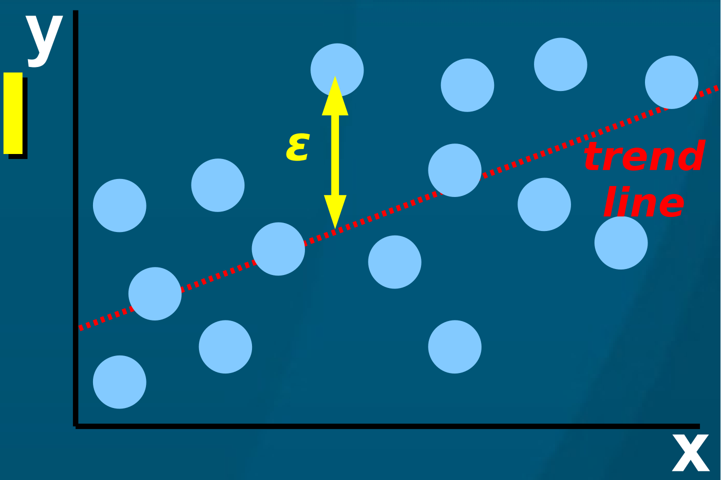
Outline for today

- Correlation
  - r^2 as fraction of variability
  - t-test on correlation
- Prediction using Simple Linear Regression
  - Linear regression model
  - Regression in Excel
  - Analysis of variance: Model vs. Residual
  - The global F-test
  - T-test on slope b_1
  - Confidence intervals on predicted values

Regression: the concept

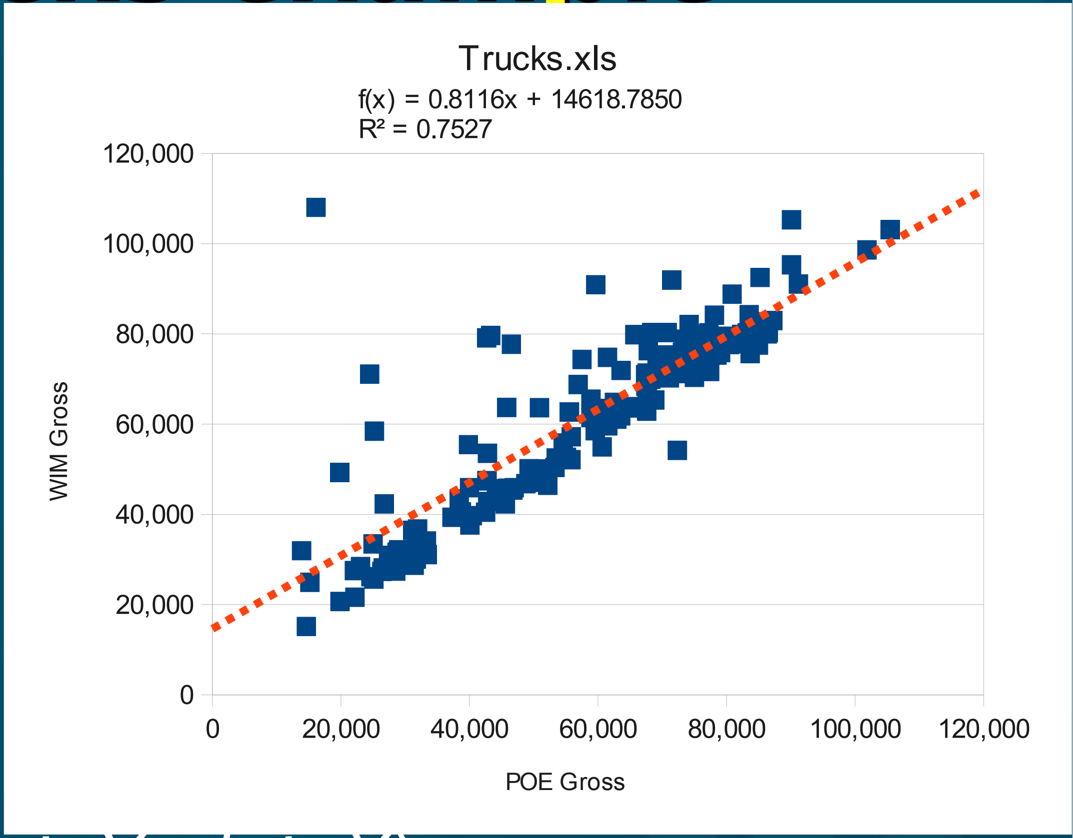
- Regression is about using one or more IVs to predict values in the DV (outcome var)
  - E.g., if we increase advertising budget, will our sales increase?
- The model describes how to predict the DV
  - Input: values for the IV(s). Output: DV value
- Linear regression uses linear functions (lines, planes, etc.) for the models
  - e.g., Sales = 0.5*AdvBudget + 2000
  - Every $1k increase in advertising budget yields 500 additional sales, and
  - With $0 spending, we'll still sell 2000 units

Regression: the model



- The linear model has the form
  - $Y = \beta_0 + \beta_1 X + \varepsilon$
- X is the predictor, Y is the outcome,
  - β_0 (intercept) and β_1 (slope) describe the line of best fit (trend line), and
  - ε represents the residuals: where the trend line doesn't fit the observed data
    - ε = (actual Y) – (predicted Y)
- The residuals average out to 0, and if the model fits the data well, they should be small overall
  - Least-squares fit: minimize SD of residuals

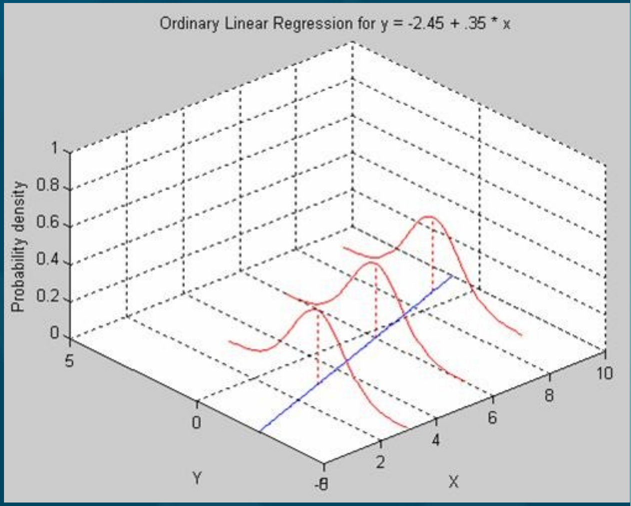
Regression: Trucks example



- Trucks example
- Scatterplot:
  X: POE Gross (G:G)
  Y: WIM Gross (H:H)
- Layout → Trendline
  - Linear, R^2
- Regression model:
  - Slope β_1: SLOPE(*dataY*, *dataX*)
  - Intercept β_0: INTERCEPT(*dataY*, *dataX*)
- SD of the residuals: STEYX(*dataY*, *dataX*)

Regression: assumptions

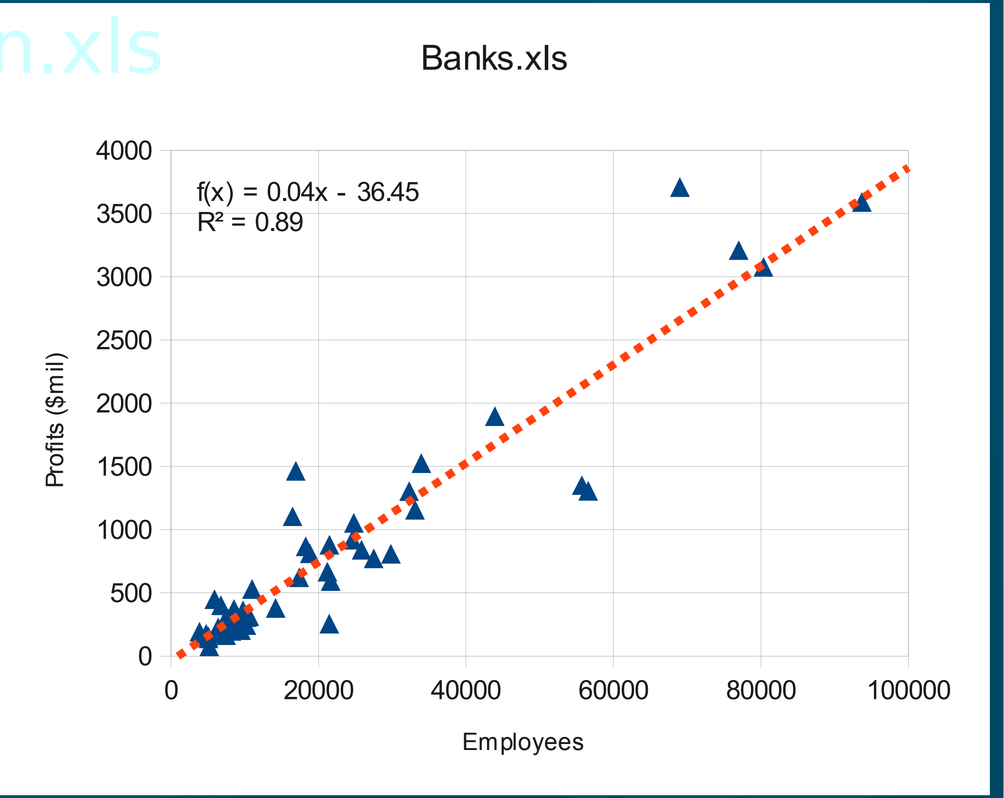
- Both IV and DV must be quantitative
  - (extensions exist for other levels of meas.)
- Independent observations
  - Not repeated-measures or hierarchical
- Normality of residuals
  - DV need not be normal, but residuals do
- Homoscedasticity
  - SD of residuals constant along the line
- These 4 are called: parametricity
  - T-test had similar assumptions



Omid Rouhani

Regression: Banks example

- "Banks" in 09-Regression.xls
- Scatterplot:
  X: Employees (D:D)
  Y: Profit (C:C)
- Layout → Trendline
- Correlation r:
  - CORREL(*datY, datX*)
- Regression model:
  - Intercept b_0: INTERCEPT(*dataY, dataX*)
  - Slope b_1: SLOPE(*dataY, dataX*)
  - SD of residuals (s_ε): STEYX(*dataY, dataX*)



Using regression for prediction

- Assuming that our linear model is correct, we can then predict profits for new companies, given their size (number of employees)
  - Profit ($mil) = 0.039*Employees – 36.45
- e.g., for a company with 1000 employees, our model predicts a profit of $2.558 million
  - This is a point estimate; s_ε adds uncertainty
- Predicted $\hat{Y}$ values: using X values from data
  - Citicorp: $\hat{Y} = 0.039 * 93700 - 36.45 \approx 3618$
- Residuals: (*actual* Y) – (*predicted* Y):
  - $Y - \hat{Y} = 3591 - 3618 = -27.73$ ($mil)
  - Overestimated Citicorp's profit by $27.73 mil

Outline for today

- Correlation
  - r^2 as fraction of variability
  - t-test on correlation
- Prediction using Simple Linear Regression
  - Linear regression model
  - Regression in Excel
  - Analysis of variance: Model vs. Residual
  - The global F-test
  - T-test on slope b_1
  - Confidence intervals on predicted values

Analysis of variance

- In regression, R^2 indicates the fraction of variability in the DV explained by the model
  - If only 1 IV, then $R^2 = r^2$ from correlation
- Total variability in DV: $SS_{tot} = \Sigma(y_i - \bar{y})^2$
  - =VAR(*dataY*) * (COUNT(*dataY*) – 1)
- Explained by model: $SS_{mod} = SS_{tot} * R^2$
- Unexplained (residual): $SS_{res} = SS_{tot} - SS_{mod}$
  - Can also get from $\Sigma(y_i - \hat{y}_i)^2$
- Hence the total variability is decomposed into:
  - $SS_{tot} = SS_{mod} + SS_{res}$
  - (book: SST = SSR + SSE)

Regression: global *F*-test

- Follow the pattern from the regular SD:

$$\sigma = \sqrt{\frac{1}{n-1}\sum\left(x-\bar{x}\right)^2}$$

	Total (on DV)	Model	Residual
SS	$SS_{tot} = \Sigma(y - \bar{y})^2$	$SS_{mod} = \Sigma(\hat{y} - \bar{y})^2$	$SS_{res} = \Sigma(y - \hat{y})^2$
df	n - 1	#vars – 1	n - #vars
MS = SS/df	SS_{tot} / (n-1)	SS_{mod} / 1	SS_{res} / (n–2)
SD = √(MS)	s_Y	-	s_ε (=STEYX)

- The test statistic is $F = MS_{mod} / MS_{res}$
  - Get p-value from FDIST(F, df_{mod}, df_{res})

Calculating the F test

- Key components are the SS_{mod} and SS_{res}
- If we already have R^2, the easiest way is:
  - Find SS_{tot} = VAR(*dataY*) * (*n*-1)
    - e.g., Banks: 38879649 (≈ 39e6)
  - Find $SS_{mod} = SS_{tot} * R^2$
    - e.g., 39e6 * 88.53% ≈ 34e6
  - Find $SS_{res} = SS_{tot} – SS_{mod}$
    - e.g., 39e6 – 34e6 ≈ 5e6
- Otherwise, find SS_{res} using pred ŷ and residuals
- Or, work backwards from $s_ε$ = STEYX(*Y*, *X*)
    - e.g., $SS_{res} = (301)^2 * (n-2)$

F-test on R^2 vs. t-test on r

- If only one predictor, the tests are equivalent:
  - $F = t^2$,
    - e.g., Banks: F ≈ 378, t ≈ 19.4
  - F-dist with $df_{mod} = 1$ is same as t-dist
    - Using same df_{res}
- If multiple IVs, then there are multiple r's
  - Correlation only works on pairs of variables
- F-test is for the overall model with all predictors
  - R^2 indicates fraction of variability in DV explained by the complete model, including all predictors

Outline for today

- Correlation
  - r^2 as fraction of variability
  - t-test on correlation
- Prediction using Simple Linear Regression
  - Linear regression model
  - Regression in Excel
  - Analysis of variance: Model vs. Residual
  - The global F-test
  - T-test on slope b_1
  - Confidence intervals on predicted values

T-test on slopes b_i

- In a model with multiple predictors, there will be multiple slopes (b_1, b_2, …)
- A t-test can be run on each b_i to test if that predictor is significantly correlated with the DV
- Let $SS_X = \Sigma(x - \bar{x})^2$ be for the predictor X:
- Then the standard error for its slope b_1 is
  - $SE(b_1) = s_\varepsilon / \sqrt{SS_X}$
- Obtain t-score and apply a t-dist with df_{res}:
  - =TDIST(b_1 / $SE(b_1)$, df_{res}, *tails*)
- If only 1 IV, the t-score is same as for r

Summary of hypothesis tests

	Correlation	Regression	Slope on X_1
Effect size	r	R^2	b_1
SE	$\sqrt{(1-r^2) / df}$	-	$s_\varepsilon / \sqrt{SS_X}$

df	n - 1	df1 = #var – 1
		df2 = n - #var
			n - #var

Test statistic	$t = r / SE(r)$	$F = MS_{mod} / MS_{res}$	$t = b_1 / SE(b_1)$

- Regression with only 1 IV is same as correlation
  - All tests would then be equivalent

Confidence int. on predictions

- Given a value x for the IV, our model predicts a point estimate $\hat{y}$ for the (single) outcome:
  - $\hat{y} = b_0 + b_1 * x$
- The standard error for this estimate is

$$SE(\hat{y}) = s_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SS_X}}$$

  - Recall that $SS_x = \Sigma(x - \bar{x})^2$
- Confidence interval: $\hat{y} \pm t * SE(\hat{y})$
- When estimating the average outcome, use

$$SE(\hat{y}) = s_\epsilon \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{SS_X}}$$

TODO

- HW6 due Thu
- Projects: be pro-active and self-led
  - If waiting on REB approval:
    generate fake (reasonable) data and
    move forward on analysis, presentation
  - Remember your potential clients:
    what questions would they like answered?
  - Tell a story/narrative in your presentation