Ch14: Correlation and Linear Regression

- **HW6 due today**
- **Please download: 15-Trucks.xls**

27 Oct 2011
BUSI275
Dr. Sean Ho

Trinity Western University

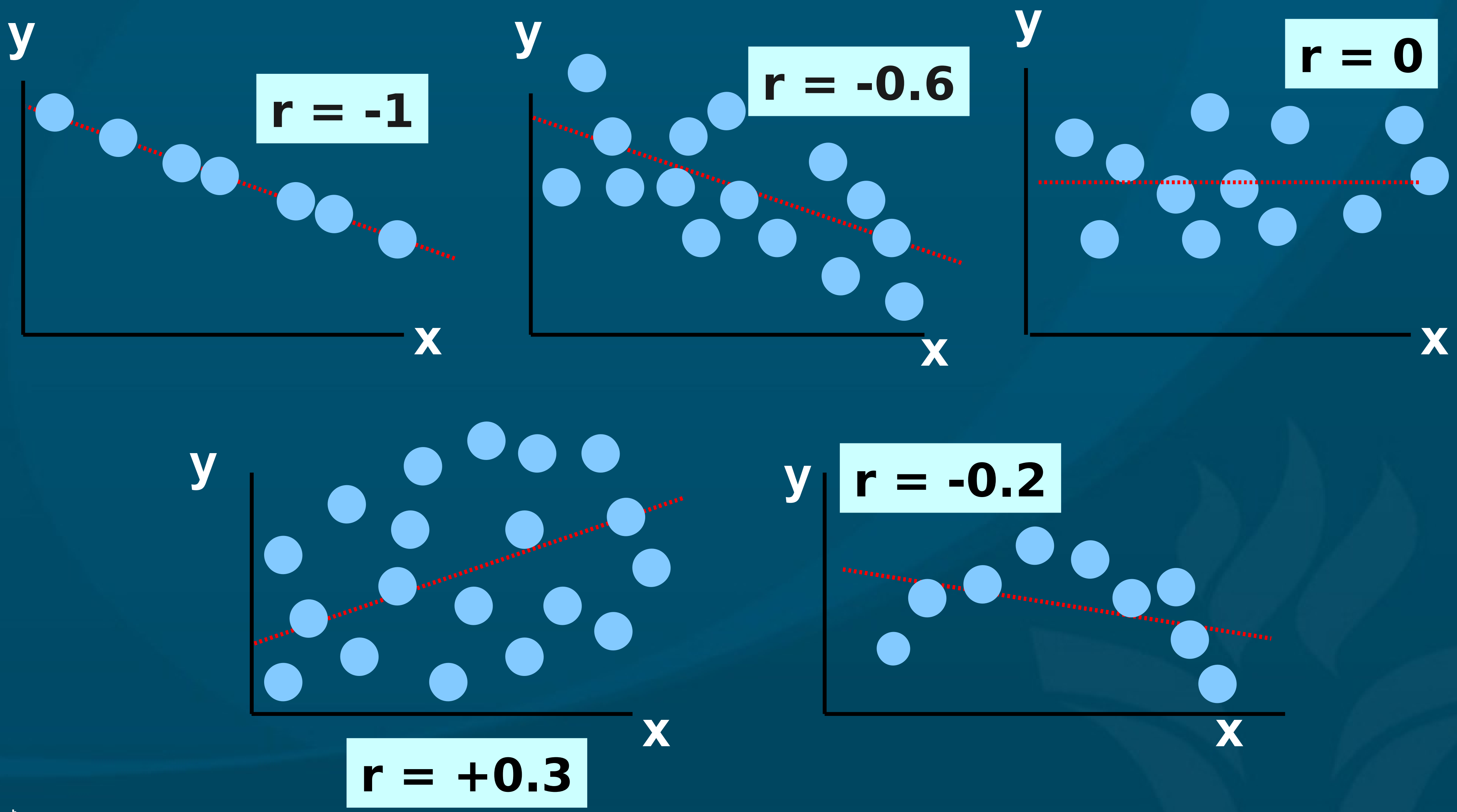
Outline for today

- Correlation
  - Intuition / concept
  - r^2 as fraction of variability
  - t-test on correlation
    - Doing it in Excel
  - Extending to categorical variables
- Intro to Regression
  - Linear regression model
  - In Excel
  - Assumptions

Linear correlation

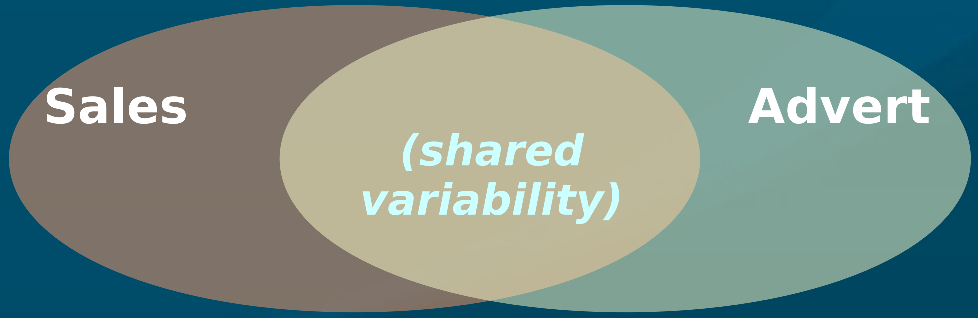
- Correlation measures the strength of a linear relationship between two variables
- Does not determine direction of causation
- Does not imply a direct relationship
  - There might be a mediating variable (e.g., between ice cream and drownings)
- Does not account for non-linear relationships
- The Pearson 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

Scatterplots and correlation



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

t-test on correlation

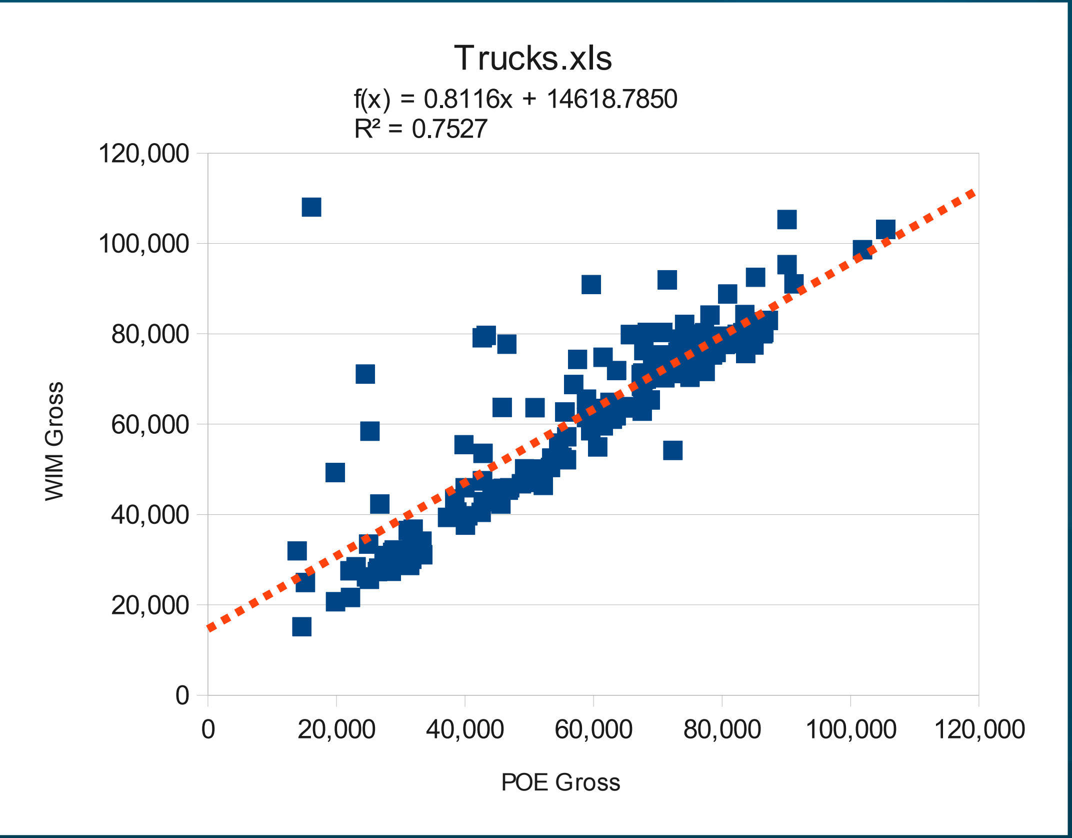
- r is sample correlation (from data)
- ρ is population correlation (want to estimate)
- Hypothesis: H_A: $\rho \neq 0$ (is there a relationship?)
- Standard error: $$SE = \sqrt{\frac{1 - r^2}{df}}$$
  - $1 - r^2$ is the variability not explained by the linear relationship
  - df = n–2 because we have 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 in Excel

- Example in 15-Trucks.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



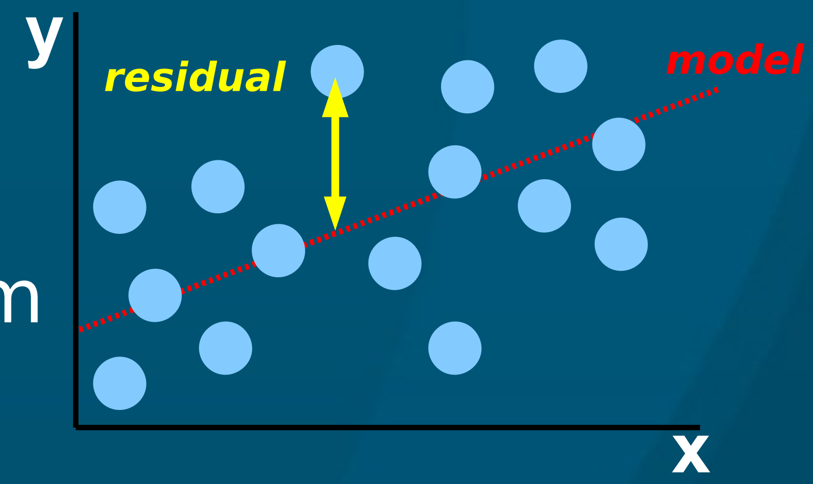
Correl. and χ^2 independence

- 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

Intro to regression

- 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

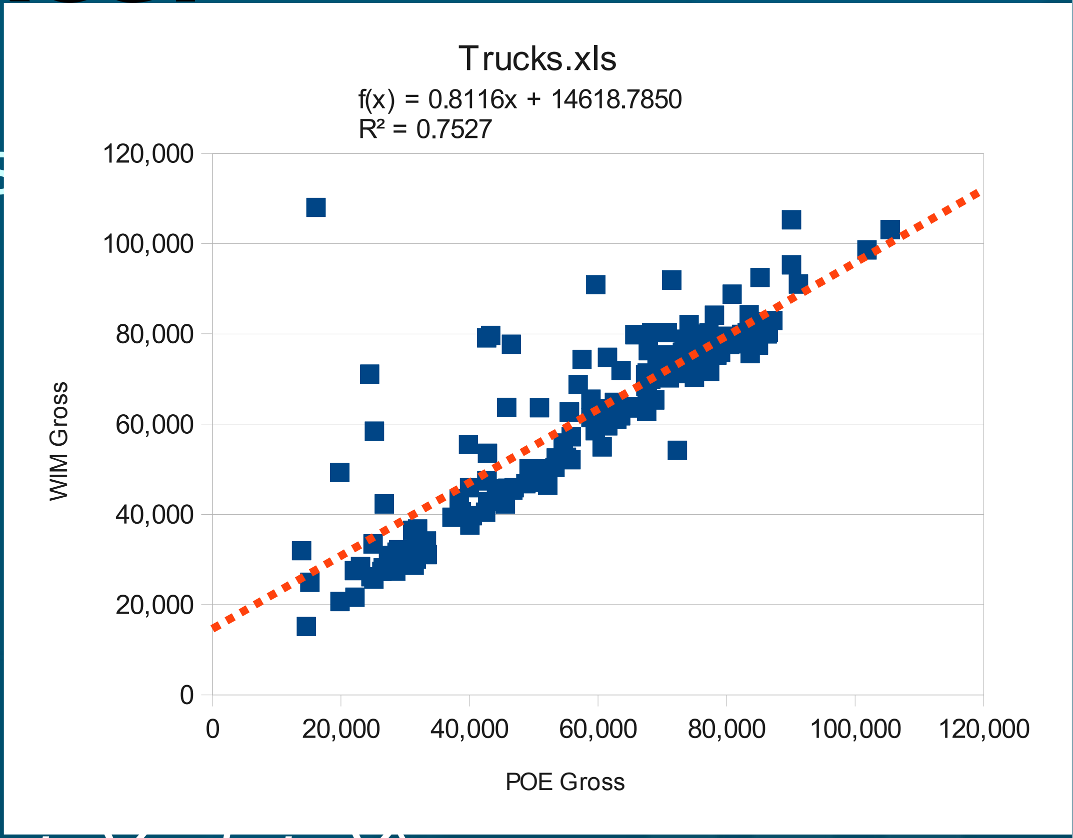
Regression model



- The linear model has the form
  - $Y = \beta_0 + \beta_1 X + \varepsilon$
- Where X is the predictor, Y is the outcome,
  - β_0 (intercept) and β_1 (slope of X) are parameters of the line of best fit,
    - Software can calculate these
  - ε represents the residuals: where the linear model 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: minimize SD of residuals

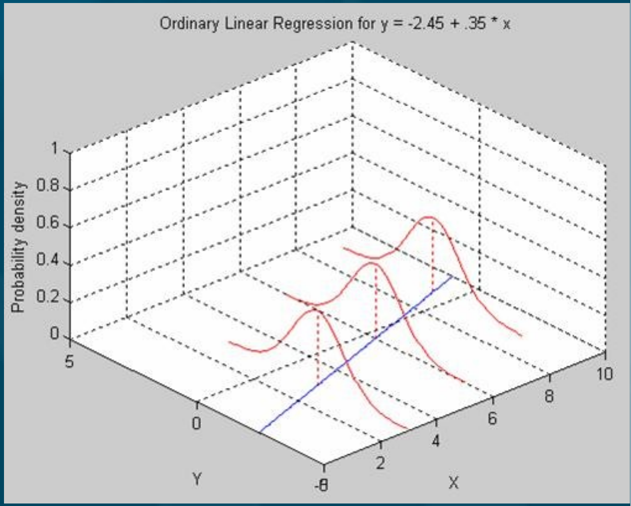
Regression in Excel

- Example: 15-Trucks.xls
- 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*)



Assumptions of regression

- 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

TODO

- HW6 (ch9-10): due Thu 27 Oct
- Projects:
  - Acquire data if you haven't already
    - If waiting for REB: try making up toy data so you can get started on analysis
  - Background research for likely predictors of your outcome variable
  - Read ahead on your chosen method of analysis (regression, time-series, logistic, etc.)