14.2-14.3: Hypothesis Tests on Regression

- HW7 due next Tues
- Please download: 16-Banks.xls

1 Nov 2011
BUSI275
Dr. Sean Ho

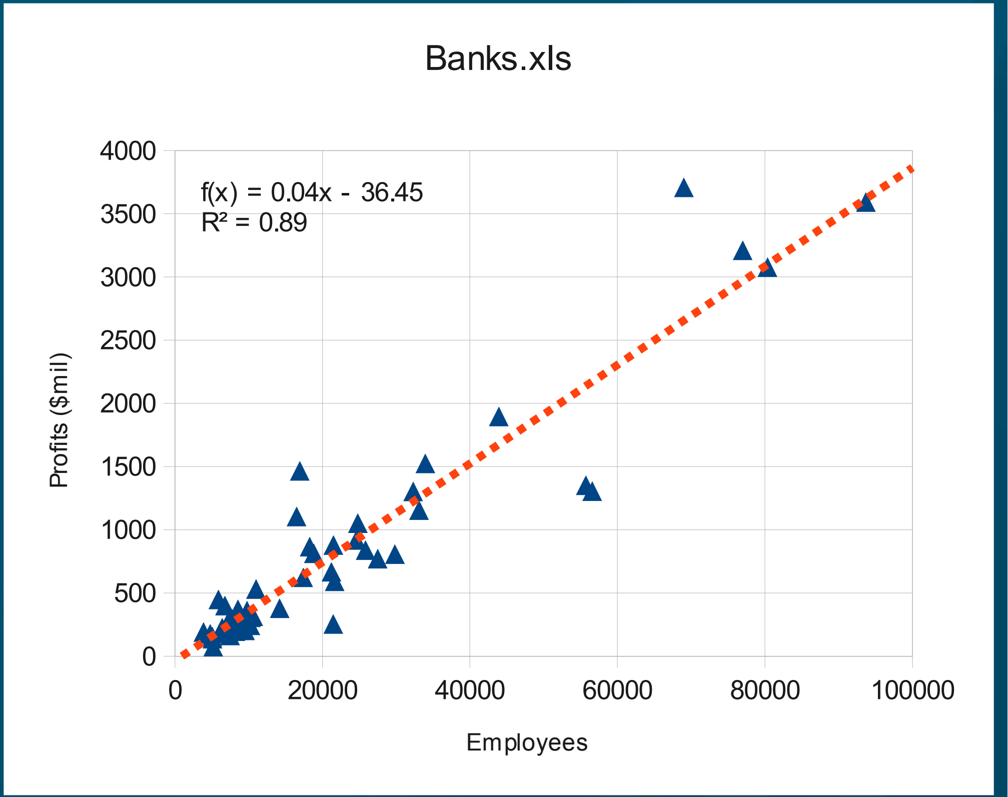
Trinity Western University

Outline for today

- Review of regression model
- Decomposition of variance in regression
  - Model vs. Residual
- F-test on the overall regression model
  - Comparison with t-test on correlation
- T-test on slopes b_i
- Confidence intervals on predicted values

Applying the regression model

- Example: 16-Banks.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*)



Predictions using the model

- 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

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)

F test on overall model (R^2)

- Follow the pattern from the regular SD:

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

	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 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)
    - Bank.xls: 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:
  - $F = t^2$,
    - Banks.xls: 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

T-test on slopes

- In a model with multiple predictors, there will be multiple slopes (b_1, b_2, …)
- A t-test can be run on each to see 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	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

- HW7 (ch10,14): due Tue 8 Nov
- 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.)