1 Simple Linear Regression: Background and Model

1.1 Regression vs. Correlation

1.1.1 Roles of explanation vs. prediction in science

1.1.2 Response vs. explanatory variables

1.1.3 Regression vs. correlation: relationship betw/ variables

regression: one variable (explanatory) *determines* other (response)

correlation: both variables change together; *covary*

1.1.4 When to use regression or correlation?

regression to measure effect of one variable on other

correlation to measure strength of association between variables

1.2 Other kinds of regression

1.2.1 nonlinear regression

1.2.2 multiple (= multivariate) regression

1.2.3 logistic regression

1.3 Linear Model for Simple Linear Regression

$Y_i = \alpha + \beta X_i + \varepsilon_i$ X_i = explanatory variable

Y_i = response variable

1.4 Fitting Regression Model

1.4.1 Least Squares estimates for *a*, *b*:

minimize squared deviations of data points from regression line;

$\rightarrow$ minimize $\sum_{i=1}^{n}\left[Y_i - (a + bX_i)\right]^2$

1.4.2 Σx^2, Σy^2, Σxy

$\Sigma x^2 = \sum (X_i - \bar{X})^2$ machine formula: $\Sigma x^2 = \sum X_i^2 - \left(\sum \bar{X}^2\right)/n$

$\Sigma y^2 = \sum (Y_i - \bar{Y})^2$ $\Sigma xy = \sum (X_i - \bar{X})(Y_i - \bar{Y})$

1.4.3 Regression coefficient (*b*)

$$b = \frac{\sum xy}{\sum x^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

potential values for *b*: $(-\infty, \infty)$

1.4.4 Y intercept (*a*):

from linear model, $Y_i = \alpha + \beta X_i$ => $\alpha = \bar{Y} - \beta\bar{X}$

estimate: $a = \bar{Y} - b\bar{X}$

need both *a*, *b* to specify unique regression line

1.5 Assumptions

1.5.1 For each *X* value, normal distribution of *Y* values (=> normal distribution of ε's)

1.5.2 Equal variances – e.g., range, SD of *Y* values not ↑ w/ larger *X*

1.5.3 Actual relationship is linear

(1.5.1)-(1.5.3) can be addressed w/ transformations

1.5.4 *Y*'s randomly sampled & independent of each other

1.5.5 *X* measurements w/out error (impossible; => assumption effectively is that *X* error negligible)

1.5.6 Concern about outliers

2 Predicting the Response Variable ($\hat{Y}$)

2.1 Knowing a, b, use regression eqn. to predict Y: $\hat{Y} = a + bX_i$

2.2 Caution: dangerous to extrapolate from regression equation

3 Testing significance of Regression

3.1 Hypotheses: H_0: $\beta = 0$ H_A: $\beta \neq 0$

3.2 Testing H_0 using ANOVA
partition variability: total SS = regression SS + residual SS

3.3 Calculations

3.3.1 Total variability in response variable: deviations of Y_i from mean Y

total SS $= \sum (Y_i - \bar{Y})^2 = \sum y^2$

3.3.2 Variability explained by regression: deviations of predicted $\hat{Y}_i$ from mean Y

$$\text{regression SS} = \sum (\hat{Y}_i - \bar{Y})^2 = \frac{\left(\sum xy\right)^2}{\sum x^2} = b\sum xy$$

3.3.3 Variability <u>not</u> explained by regression

residual SS $= \sum (Y_i - \hat{Y}_i)^2$ = total SS − regression SS

3.3.4 Degrees of freedom

total DF = $n - 1$

regression DF = 1 (always = 1 for simple linear regression)

residual DF = total DF − regression DF = $n - 2$

3.3.5 Mean Squares

MS = SS/DF

3.3.6 F-statistic $F = \dfrac{\text{regression MS}}{\text{residual MS}}$ critical value: $F_{\alpha(1),\nu_1,\nu_2} = F_{\alpha(1),1,n-2}$

3.4 Coefficient of determination, r^2: proportion of total variation in Y explained by regression

$$r^2 = \frac{\text{regression SS}}{\text{total SS}}$$

3.5 Testing H_0 using t-test

3.5.1 Applications: *cannot be tested w/ ANOVA

H_0: $\beta = 0$ H_A: $\beta \neq 0$

H_0: $\beta = \beta_0$ H_A: $\beta \neq \beta_0$ *

H_0: $\beta \geq \beta_0$ H_A: $\beta < \beta_0$ (one tailed tests) *

3.5.2 t-statistic

$$t = \frac{b - \beta_0}{s_b}$$

s_b = standard error of regression coefficient

$$\text{var}(b) = s_b^2 = \frac{s_{Y\cdot X}^2}{\sum x^2} \qquad s_b = \sqrt{\frac{s_{Y\cdot X}^2}{\sum x^2}}$$

were $s_{Y\cdot X}^2$ = residual MS

4 Confidence Intervals in Regression

4.1 General form for confidence intervals: CI = statistic ± (t_α)(SE of statistic)

4.2 Confidence interval for regression coefficient, b

$1-\alpha$ confidence limits: $b \pm t_{\alpha(2),(n-2)} s_b$

5 Interpreting Regression Results

Regression ≅ fitting line to data; quantitative exercise

Does not prove that relationship exists

6 Data Transformations in Regression

6.1 Purpose of transformations: to adjust distribution of data to satisfy assumptions (i.e., normality, equality of variances)

→ not to straighten curved lines

− e.g., log or sq.root transform may straighten points into line, but then other assumptions violated

6.2 Transformations of explanatory data (X) not affect distribution of Y, so can be used to straighten curved line

Caution with transformation of response data (Y):

→ inappropriate transformation will violate assumptions

6.3 Inspection of Residuals

7 Comparing Two Slopes

7.1 Hypotheses: H_0: $\beta_1 = \beta_2$ H_A: $\beta_1 \neq \beta_2$

7.2 Student's t-test:

7.3 test statistic:

$$t = \frac{b_1 - b_2}{s_{b_1-b_2}} \qquad s_{b_1-b_2} = \sqrt{\frac{(s^2_{Y\cdot X})_p}{\left(\sum x^2\right)_1} + \frac{(s^2_{Y\cdot X})_p}{\left(\sum x^2\right)_2}}$$

$$(s^2_{Y\cdot X})_p = \frac{(\text{residual SS})_1 + (\text{residual SS})_2}{(\text{residual DF})_1 + (\text{residual DF})_2}$$

7.4 Critical value (t) has ($n_1 - 2$) + ($n_2 - 2$) degrees of freedom:

$$\nu = n_1 + n_2 - 4$$

7.5 If H_0 not rejected, estimate common regression coefficient:

$$b_c = \frac{\left(\sum xy\right)_1 + \left(\sum xy\right)_2}{\left(\sum x^2\right)_1 + \left(\sum x^2\right)_2}$$

or (w/ more rounding error): $$b_c = \frac{\left(\sum x^2\right)_1 b_1 + \left(\sum x^2\right)_2 b_2}{\left(\sum x^2\right)_1 + \left(\sum x^2\right)_2}$$