Simple Linear Regression (SLR)

- The SLR Model
- Fitting the SLR Model (Least Squares)
- Inferences on the Parameters
- Inferences on the Mean Response
- Prediction Intervals
- Analysis of Variance (ANOVA)
- Residual Analysis

The SLR Model

- A common problem in engineering is to study the relationship between two variables.
  - Let Y be the dependent variable (response).
  - Let x be the independent variable (regressor).
- The SLR model for $i = 1, \ldots, n$ observations

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

  - Parameters: β_1 is the slope, β_0 is the y-intercept.
  - Random Error: ε_i has mean 0 and variance σ^2

$$\Rightarrow \; Y_i \text{ is a r.v.}$$

SLR Examples

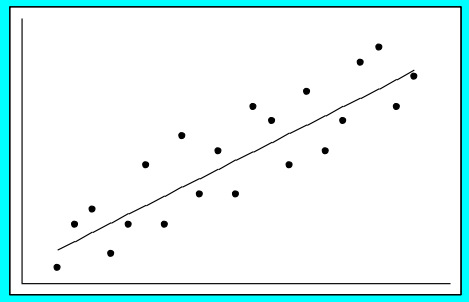
- Goal: Model Y as a function of x.

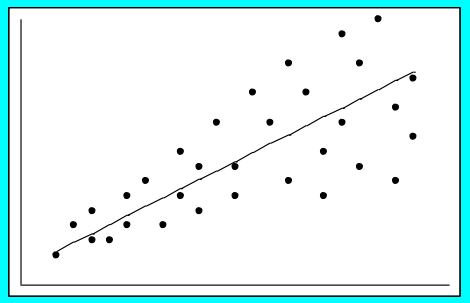
***x* (independent var.)**	***Y* (dependent var.)**
H.S. GPA, SAT score	College GPA
# Cigarettes smoked/day	Blood pressure
Ad expenditure, Price	Sales
Temperature, Reaction time	Yield of a chemical
Height of a person	Weight of a person

- Data is paired: (x_i , y_i)
- See handout: Widget example.

Features of the SLR Model

- Mean and variance of the response Y_i
  - $E[Y_i] = \beta_0 + \beta_1 x_i$ = true line
  - $V(Y_i) = \sigma^2$ = same for all observations





- Covariance between responses Y_i and Y_j
  - $\text{Cov}(Y_i, Y_j) = \text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$
  - Uncorrelated $\Rightarrow$ The outcome of one response does not affect the outcome of another.

Fitting the SLR Model

- The parameters β_0 and β_1 are unknown

  $\Rightarrow$ The true line $\beta_0 + \beta_1 x$ is unknown.

- Point estimators
  - b_1 estimates β_1
  - b_0 estimates β_0
- The fitted regression line is:

$$\hat{Y} = b_0 + b_1 x$$

  - $\hat{y}$ estimates the true line $\beta_0 + \beta_1 x$

Least Squares (LS) Estimation

- Fitted Values: $\hat{y}_i = b_0 + b_1 x_i$
- Residuals: $e_i = y_i - \hat{y}_i$
- Minimize the squared residuals:

$$\min SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

  - Take derivatives w.r.t. b_0 and b_1
  - Set derivatives equal to 0
  - Solve 2 equations for b_0 and b_1

Parameter Point Estimates

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1 \bar{x}$$

- See handout: widget example.

Error Variance Estimation

- Notation

$$S_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$S_{YY} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

$$S_{XY} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

- Rewrite LS estimate for β_1

$$b_1 = S_{XY} / S_{XX}$$

Error Variance Estimation

- Rewrite error sum of squares

$$SSE = \sum_{i=1}^{n} e_i^2 = S_{YY} - b_1 S_{XY}$$

- Define mean square for error

$$MSE = \frac{SSE}{n-2}$$

- Then MSE is an unbiased estimator for σ^2
  - s^2 = MSE

Error Variance Estimation

- Widget example

$$\sum x_i = 500, \sum y_i = 1100 \Rightarrow \bar{x} = 50, \bar{y} = 110$$

$$\sum x_i^2 = 28{,}400, \sum y_i^2 = 134{,}660$$

$$\sum x_i y_i = 61{,}800$$

- Calculate $S_{XX} = 3400$, $S_{YY} = 13{,}660$, $S_{XY} = 6800$

$$MSE = \frac{13{,}660 - (2)(6800)}{10 - 2} = \frac{60}{8} = 7.5$$

Inferences on the Parameters

- See handout for C.I.'s, H.T.'s, and Prediction.
- Widget example: Standard errors for b_1 and b_0

$$s.e.\{b_1\} = \sqrt{\frac{MSE}{S_{XX}}} = \sqrt{\frac{7.5}{3400}} \cong 0.047$$

$$s.e.\{b_0\} = \sqrt{MSE\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right]} = \sqrt{7.5\left[\frac{1}{10} + \frac{50^2}{3400}\right]}$$

$$\cong 2.503$$

Inferences: Slope

- Widget example: Estimate the slope β_1 using a two-sided 95% C.I.

$$t_{\alpha/2,\, n-2} = t_{.025,8} = 2.306$$

$$b_1 \pm t_{.025,8}\, s.e.\{b_1\} = 2 \pm (2.306)(0.047)$$

$$= (1.89, 2.11)$$

We are 95% confident that on average person-hours increases by between 1.89 and 2.11 for each unit increase in lot size.

Inferences: Slope

- Widget example: Test if the slope is significant at the 0.05 level.

H_0: $\beta_1 = 0$ vs. H_1: $\beta_1 \neq 0$ (Form A)

- Using a test statistic:

$$t^* = \frac{b_1 - 0}{s.e.\{b_1\}} = \frac{2}{0.047} \cong 42.6$$

Since $|42.6| > t_{.025,\,8} = 2.306$, we reject H_0

- Using the 95% C.I. (1.89, 2.11):

Since 0 is <u>not</u> in the C.I., we reject H_0

Inferences: Slope

- Widget example: Estimate the slope β_1 using a 95% upperbound C.I.

$$t_{\alpha,\, n-2} = t_{.05,8} = 1.860$$

$$\left(-\infty,\ b_1 + t_{.05,8}\, s.e.\{b_1\}\right) = \left(-\infty,\ 2 + (1.860)(0.047)\right)$$

$$= (-\infty, 2.09)$$

We are 95% confident that on average person-hours increases by 2.09 or less for each unit increase in lot size.

Inferences: Slope

- Widget example: Test if the slope is less than 2.1 at the 0.05 level.

H_0: $\beta_1 = 2.1$ vs. H_1: $\beta_1 < 2.1$ (Form B)

- Using a test statistic:

$$t^* = \frac{b_1 - 2.1}{s.e.\{b_1\}} = \frac{2 - 2.1}{0.047} \cong -2.129$$

Since $-2.129 < -t_{.05,\,8} = -1.860$, we reject H_0

- Using the 95% C.I. $(-\infty, 2.09)$:

Since 2.1 is <u>not</u> in the C.I., we reject H_0

Inferences: *Y*-intercept

- Widget example: Estimate the *y*-intercept β_0 using a 95% lowerbound C.I.

$$t_{\alpha,\, n-2} = t_{.05,8} = 1.860$$

$$\left(b_0 - t_{.05,8}\, s.e.\{b_0\},\, \infty\right) = \left(10 - (1.860)(2.503),\, \infty\right)$$

$$= (7.28,\, \infty)$$

Note: In this example, β_0 is not meaningful because the range of the independent variable (lot size) does not include $x = 0$.

Inferences: *Y*-intercept

- Widget example: Test if the *y*-intercept is greater than 7 at the 0.05 level.

$H_0: \beta_0 = 7$ vs. $H_1: \beta_0 > 7$ (Form C)

- Using a test statistic:

$$t^* = \frac{b_0 - 7}{s.e.\{b_0\}} = \frac{10 - 7}{2.503} \cong 1.199$$

Since $1.199 < t_{.05,\,8} = 1.860$, we fail to reject H_0

- Using the 95% C.I. (7.28, ∞):

Since 7 is in the C.I., we fail to reject H_0

Inferences on the Mean Response

- See handout for C.I.'s, H.T.'s, and Prediction.
- Mean response $E[Y] = \beta_0 + \beta_1 x$ (true line) at specific $x = x_0$
- Widget example: Estimate the mean response when (a) lot size is 55, (b) lot size is 80.
  - Point estimate: $\hat{y} = b_0 + b_1 x_0 = 10 + 2 x_0$

$$\hat{y} \mid_{x = 55} = 10 + 2(55) = 120$$

$$\hat{y} \mid_{x = 80} = 10 + 2(80) = 170$$

Inferences on the Mean Response

- Widget example: Standard errors for $\hat{Y}$ when (a) lot size is 55, (b) lot size is 80.

$$s.e.\{\hat{Y}|_{x=x_0}\} = \sqrt{MSE\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}\right]}$$

$$s.e.\{\hat{Y}|_{x=55}\} = \sqrt{7.5\left[\frac{1}{10} + \frac{(55-50)^2}{3400}\right]} \cong 0.897$$

$$s.e.\{\hat{Y}|_{x=80}\} = \sqrt{7.5\left[\frac{1}{10} + \frac{(80-50)^2}{3400}\right]} \cong 1.654$$

Inferences on the Mean Response

- Widget example: Use two-sided 95% C.I.'s to estimate the mean response when (a) lot size is 55, (b) lot size is 80

$$\hat{y}|_{x=55} \pm t_{.025,8}\, s.e.\{\hat{Y}|_{x=55}\} = (118.3, 121.7) \rightarrow \text{width} = 3.4$$

$$\hat{y}|_{x=80} \pm t_{.025,8}\, s.e.\{\hat{Y}|_{x=80}\} = (166.9, 173.1) \rightarrow \text{width} = 6.2$$

We are 95% confident that the mean person-hours required for a lot size of 55 is between 118.3 and 121.7.

Prediction Intervals

- See handout for C.I.'s, H.T.'s, and Prediction.
- Predict a new observation of Y at $x = x_0$ using a prediction interval (P.I.).
- Widget example:
  - Prediction of person-hours when lot size is 55 with a single value is still $\hat{y} \mid_{x = 55} = 120$.
- Prediction error includes two components
  - Variance of predicted value $\hat{Y}$.
  - Variance of random error ε.

Prediction Intervals

- Widget example: Prediction error for a new observation Y when lot size is 55.

$$p.e.\{Y|_{x=x_0}\} = \sqrt{MSE\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{XX}}\right]}$$

$$= \sqrt{MSE+\left(s.e.\{\hat{Y}|_{x=x_0}\}\right)^2}$$

$$p.e.\{Y|_{x=55}\} = \sqrt{7.5+(0.897)^2} \cong 2.882$$

Prediction Intervals

- Widget example: Use two-sided 95% P.I. to predict a new observation of the response when lot size is 55.

$$\hat{y}\,|_{x=55} \pm t_{.025,8}\ p.e.\{Y\,|_{x=55}\} = (114.6, 125.4)$$

We are 95% confident that a new observation of person-hours for a lot size of 55 will lie between 114.6 and 125.4.

Note: The P.I. for a new response observation is always wider than the C.I. for the mean response.

Analysis of Variance (ANOVA)

- Total Sum of Squares
  - Variability in the response observations y_i (ignoring the independent variable x_i)
  - $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = S_{YY}$
- Regression Sum of Squares
  - Variability explained by the model that includes independent variable x_i
  - $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = b_1 S_{XY} = (b_1)^2 S_{XX}$

Analysis of Variance (ANOVA)

- Error Sum of Squares
  - Variability that remains unexplained by the model
  - $SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = S_{YY} - b_1 S_{XY}$
- Decomposition of Total Sum of Squares
  - See handout with graphical illustration of ANOVA.
  - $(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$
  - $SST = SSR + SSE$

Analysis of Variance (ANOVA)

- Degrees of Freedom (*DF*)
  - *SST*: $DF = n - 1$ (same as sample variance)
  - *SSR*: $DF = 1$ (one independent variable)
  - *SSE*: $DF = DF_{SST} - DF_{SSR} = n - 2$
- Mean Squares: $MS = SS/DF$
  - $MSR = SSR/1 = SSR$
  - $MSE = SSE/(n-2) = s^2$
- *F*-test: $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$
  - *Reject* H_0 if $F^* = \dfrac{MSR}{MSE} > f_{\alpha,1,n-2}$

ANOVA Table

Source	DF	SS	MS	F
Regression	1	SSR	MSR	F*
Error	$n - 2$	SSE	MSE	
Total	$n - 1$	SST		

- Coefficient of Determination:
  - Fraction of variability in the response observations that is explained by the model with *x*.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

ANOVA Table

- Widget example

Source	DF	SS	MS	F
Regression	1	13,600	13,600	1813
Error	8	60	7.5	
Total	9	13,660		

- Since $F^* = 1813 > f_{.01,\,1,\,8} = 11.26$, we conclude the regression is significant (Reject H_0).
- $R^2 = 0.9956 \Rightarrow$ 99.56% of the variability in person-hours is explained by the model with independent variable lot size.

Residual Analysis

- Verify the SLR model assumptions.
- A linear model is reasonable
  - Check plot of e_i vs. $\hat{y}_i$ for curvature.
- The residuals have the same variance
  - Check plot of e_i vs. $\hat{y}_i$ for funnel shape.
- The residuals follow a normal distribution
  - Check histogram or boxplot of the residuals.
  - Check normal probability plot.
- Example: Muscle Mass handout

Transformations

- If a linear model is not reasonable
  - Possibly transform x, e.g., x^2.
- If the residuals do not have constant variance and do not follow a normal distribution
  - Use a variance stabilizing transformation on y.
  - Square root $\sqrt{y}$, $\log(y)$, inverse $\frac{1}{y}$.
- If the residuals do not have constant variance and do follow a normal distribution
  - Use weighted least squares.