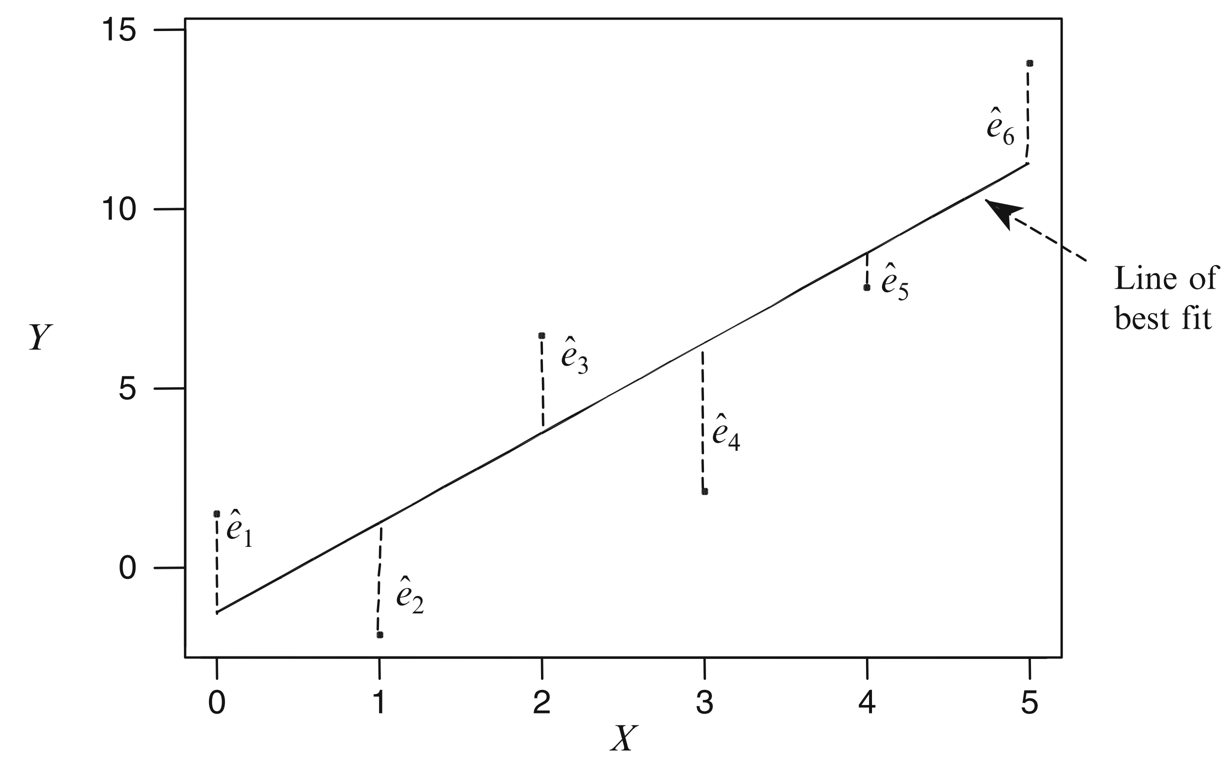
Simple linear regression

MATH 5398
Department of mathematics, UTA

Set-up

- Pairs of observations: (x_i, y_i) for $i = 1, \ldots, n$.
- Y-variable
  - Dependent or response variable
- X-variable
  - Explanatory or predictor variable
  - Its value can sometimes be chosen by a researcher.
- The regression of a random variable *Y* on a random variable *X* is $E(Y|X=x) = g(x)$, which can be any function.

- The regression is linear if $E(Y|X=x)=\beta_0+\beta_1 x$
- β_0 (intercept) and β_1 (slope): unknown regression coefficients
- A line of best fit is chosen by minimizing the residual sum of square (RSS).



$$\mathrm{RSS}=\sum_{i=1}^{n}\hat{e}_i^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2$$

$$(\hat{\beta}_0,\hat{\beta}_1)=\underset{b_0,\,b_1}{\arg\min}\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2$$

- Taking partial derivatives, and we obtain the normal equations.

$$\frac{\partial \mathrm{RSS}}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0$$

$$\frac{\partial \mathrm{RSS}}{\partial b_1} = -2\sum_{i=1}^{n}x_i(y_i - b_0 - b_1 x_i) = 0$$

$$\boxed{\begin{aligned} \sum_{i=1}^{n} y_i &= b_0 n + b_1 \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i &= b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 . \end{aligned}}$$

- Solving the normal equations gives the least square (LS) estimates:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{SXY}{SXX}$$

- LS estimates can be written using the sample correlation coefficient

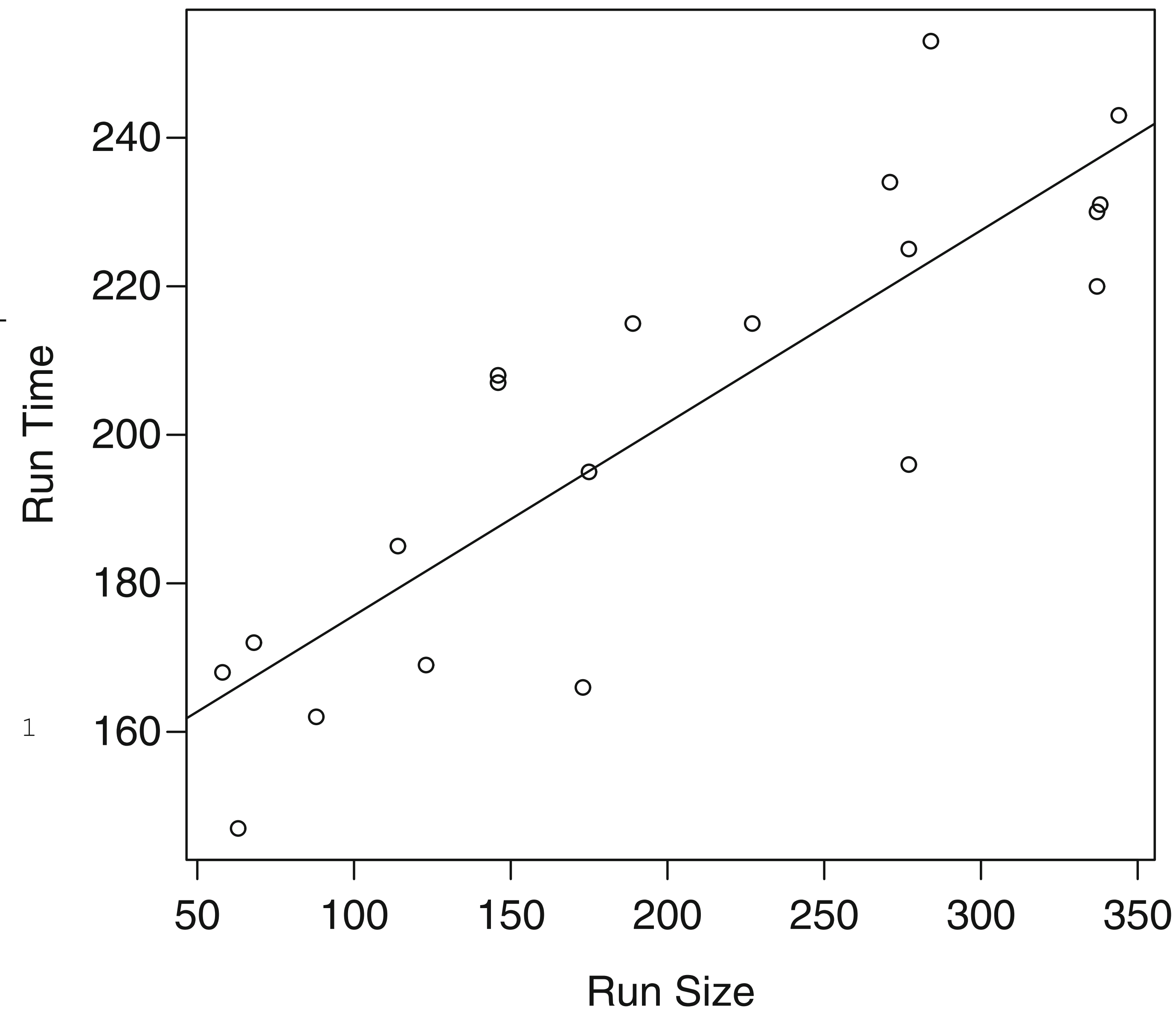
$$r_{XY} = \frac{\frac{1}{n-1} SXY}{SD_X SD_Y} = \frac{SXY}{\sqrt{SXX \cdot SYY}}$$

$$SD_X = \tfrac{1}{n-1}\sqrt{SXX};\ SD_Y = \tfrac{1}{n-1}\sqrt{SYY}$$

$$\hat{\beta}_1 = \frac{SXY}{SXX} = r_{XY}\,\frac{SD_Y}{SD_X}$$

Table 2.1 Production data (production.txt)

Case	Run time	Run size	Case	Run time	Run size
1	195	175	11	220	337
2	215	189	12	168	58
3	243	344	13	207	146
4	162	88	14	225	277
5	185	114	15	169	123
6	231	338	16	215	227
7	234	271	17	147	63
8	166	173	18	230	337
9	253	284	19	208	146
10	196	277	20	172	68




```
Coefficients:
              Estimate Std. Error t value  Pr(>|t|)
(Intercept)   149.74770    8.32815   17.98  6.00e-13 ***
RunSize         0.25924    0.03714    6.98  1.61e-06 ***
---
Signif. codes:0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.25 on 18 degrees of freedom
Multiple R-Squared: 0.7302,      Adjusted R-squared: 0.7152
F-statistic: 48.72 on 1 and 18 DF, p-value: 1.615e-06
```

Estimating variance

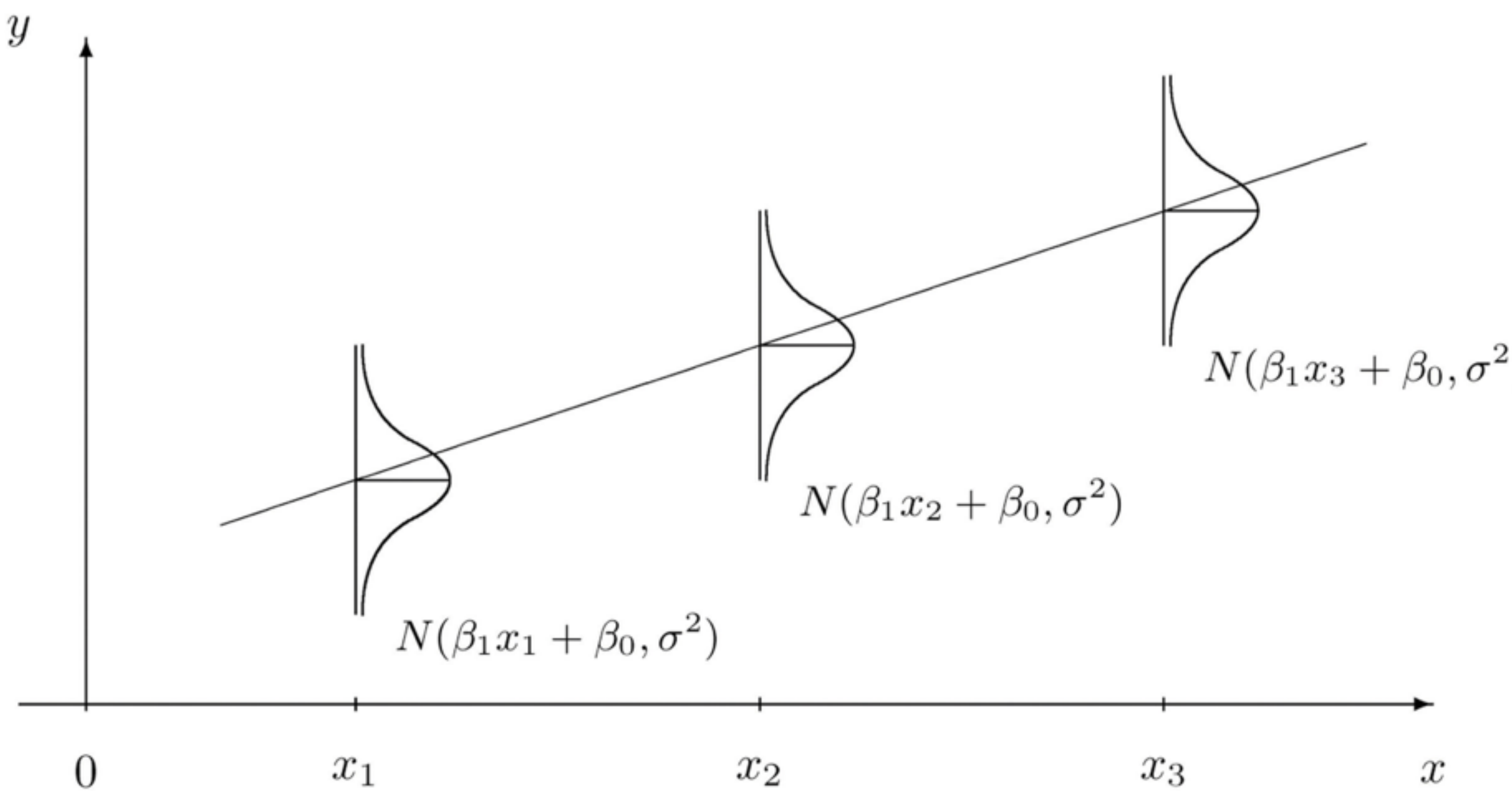
- Linear regression model $Y_i = \beta_0 + \beta_1 x_i + e_i$
  - The last term is random error (mean 0 and variance σ^2)

$$e_i = Y_i - (\beta_0 + \beta_1 x_i) = Y_i - \text{unknown regression line at } x_i.$$

Residual $$\hat{e}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = Y_i - \text{estimated regression line at } x_i.$$

- The sum of residuals is zero. Why?
- The unbiased estimate of σ^2: $S^2 = \frac{\text{RSS}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i^2$
  - Divisor (n-2) is related to the sample size and the number of coefficients estimated.

Simple linear regression concept



Association/Correlation vs Causation

- Causal interpretation: The statement "X causes Y" means that changing the value of X will change the distribution of Y.
    - When X causes Y, X and Y will be associated but the reverse is not true.
- Association interpretation: change in the value of X is associated with changes in the value of Y
    - Association does not necessarily imply causation.
- If the data are from a randomized study, then the causal interpretation is correct.
- If the data are from a observational study, then the causal interpretation is NOT correct.

Interpretation of LS estimates

Interpret the estimate, b_0, only if there are data near zero and setting the explanatory variable to zero makes scientific sense. The meaning of b_0 is the estimate of the mean outcome when $x = 0$, and should always be stated in terms of the actual variables of the study. The p-value for the intercept should be interpreted (with respect to retaining or rejecting $H_0 : \beta_0 = 0$) only if both the equality and the inequality of the mean outcome to zero when the explanatory variable is zero are scientifically plausible.

The interpretation of b_1 is the change (increase or decrease depending on the sign) in the average outcome when the explanatory variable increases by one unit. This should always be stated in terms of the actual variables of the study. Retention of the null hypothesis $H_0 : \beta_1 = 0$ indicates no evidence that a change in x is associated with (or causes for a randomized experiment) a change in y. Rejection indicates that changes in x *cause* changes in y (assuming a randomized experiment).

Simple linear regression assumptions

1. Unbiasedness, linearity
2. Independent errors
3. Constant variance (homoscedasticity)
4. Gaussian errors

1. Y is related to x by the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + e_i \ (i = 1, \ldots, n)$, i.e., $\mathrm{E}(Y \mid X = x_i) = \beta_0 + \beta_1 x_i$
2. The errors $e_1, e_2, \ldots, e_n$ are independent of each other
3. The errors $e_1, e_2, \ldots, e_n$ have a common variance σ^2
4. The errors are normally distributed with a mean of 0 and variance σ^2, that is, $e \mid X \sim N(0, \sigma^2)$

- We should check if these four assumptions hold to make inferences on the linear regression model.
- We can assume X's are non-random.

Distribution of the slope

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

$$\hat{\beta}_1=\sum_{i=1}^{n}c_iy_i \text{ where } c_i=\frac{x_i-\bar{x}}{SXX}$$

- LS estimates are unbiased

$$E(\hat{\beta}_1 \mid X)=\sum_{i}^{n}c_iE(y_i \mid X)=\sum_{i}^{n}c_i(\beta_0+\beta_1x_i)=\beta_1\sum_{i}^{n}c_ix_i=\beta_1$$

$$Var(\hat{\beta}_1 \mid X)=\sum_{i}^{n}c_i^2Var(y_i \mid X)=\sigma^2\sum_{i}^{n}c_i^2=\frac{\sigma^2}{SXX}$$

- It can be shown that

$$\hat{\beta}_1 \mid X \sim N\left(\beta_1,\frac{\sigma^2}{SXX}\right)$$

- If a,b,c,d are constants and X and Y are random variables, then $Cov(aX+b, cY+d) = ac\,Cov(X,Y)$

$$\begin{aligned}
&Cov(\bar{y}, \hat{\beta}_1 \mid X) \\
&= Cov(\beta_0 + \beta_1 \bar{x} + \bar{e}, \sum_i^n c_i(\beta_0 + \beta_1 x_i + e_i) \mid X) \\
&= Cov(\tfrac{1}{n}\sum_i^n e_i, \sum_i^n c_i e_i \mid X) \\
&= \frac{1}{n}\sum_i^n \sum_j^n c_i \cdot Cov(e_i, e_j \mid X)
\end{aligned}$$

- By the assumptions, errors are independent

$$Cov(e_i, e_j \mid X) = \begin{cases} \sigma^2 & i = j \\ 0 & i \neq j \end{cases}$$

- Mean of y and LS estimate of the slope is uncorrelated.

$$Cov(\bar{y}, \hat{\beta}_1 \mid X) = \frac{\sum_i^n (x_i - \bar{x})}{n \cdot SXX}\sigma^2$$

Distribution of the intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- LS estimates are unbiased

$$E(\hat{\beta}_0 \mid X) = E(\bar{y} \mid X) - E(\hat{\beta}_1 \mid X)\bar{x} = E(\beta_0 + \beta_1 \bar{x} + \bar{e} \mid X) - \beta_1 \bar{x} = \beta_0$$

$$Var(\hat{\beta}_0 \mid X) = Var(\bar{y} - \hat{\beta}_1 \bar{x} \mid X) = \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{SXX}$$

- It can be shown that

$$\hat{\beta}_0 \mid X \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)\right)$$

- Covariance between LS estimates

$$\begin{aligned}
Cov(\hat{\beta}_0, \hat{\beta}_1 \mid X) &= Cov(\overline{y}, \hat{\beta}_1 \mid X) - \overline{x} Cov(\hat{\beta}_1, \hat{\beta}_1 \mid X) \\
&= 0 - \overline{x} Var(\hat{\beta}_1) \\
&= \frac{-\overline{x}}{SXX} \sigma^2
\end{aligned}$$

- The variances and covariance for the LS estimators all depend on the unknown σ^2. One needs to estimate this if we want to compute *standard errors.*

- The variances of LS estimates decrease as the distribution of X becomes more spread out.

  - Thus, in a designed experiment greater precision is achieved using a wider range of X values.

Inference about the slope

$$\hat{\beta}_1 \mid X \sim N\left(\beta_1, \frac{\sigma^2}{SXX}\right)$$

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{SXX}} \sim N(0,1)$$

$$T = \frac{\hat{\beta}_1 - \beta_1}{S / \sqrt{SXX}} = \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} \Rightarrow T = \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} \sim t_{n-2}$$

degrees of freedom = sample size – number of mean parameters estimated.

Confidence interval of slope

$$T = \frac{\hat{\beta}_1 - \beta_1}{\mathrm{se}(\hat{\beta}_1)} \sim t_{n-2}$$

- (1-α)x100% confidence interval for slope

$$(\hat{\beta}_1 - t(\alpha/2, n-2)\,\mathrm{se}(\hat{\beta}_1), \hat{\beta}_1 + t(\alpha/2, n-2)\,\mathrm{se}(\hat{\beta}_1))$$

- Example

```
Coefficients:
               Estimate Std. Error t value  Pr(>|t|)
(Intercept)   149.74770    8.32815   17.98  6.00e-13 ***
RunSize         0.25924    0.03714    6.98  1.61e-06 ***
---
Signif. codes:0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.25 on 18 degrees of freedom
Multiple R-Squared: 0.7302,      Adjusted R-squared: 0.7152
F-statistic: 48.72 on 1 and 18 DF, p-value: 1.615e-06
```

Table of t-distribution

Testing hypothesis on slope

$$H_0 : \beta_1 = \beta_1^0 \quad \Longrightarrow \quad T = \frac{\hat{\beta}_1 - \beta_1^0}{\mathrm{se}(\hat{\beta}_1)} \sim t_{n-2} \text{ when } H_0 \text{ is true.}$$

- Consider testing hypotheses

$$H_0 : \beta_1 = 0 \text{ vs } H_A : \beta_1 \neq 0$$

$$T = \frac{\hat{\beta}_1}{se(\beta_1)} \sim t_{n-2} \text{ when } H_0 \text{ is true.}$$

- At significance level α, reject the null if

$$|T| > t(\alpha/2, n-2) \quad \Longleftrightarrow \quad \text{p-value} = 2 \cdot P(t_{n-2} > |T|) < \alpha$$

- What if you fail to reject the null?

- Example: production data

Inferences about the intercept

$$\hat{\beta}_0 \mid X \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{SXX}\right)\right)$$

$$Z = \frac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{1/n + \overline{x}^2/SXX}} \sim N(0,1) \quad \Rightarrow \quad T = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{1/n + \overline{x}^2/SXX}} = \frac{\hat{\beta}_0 - \beta_0}{\text{se}(\hat{\beta}_0)} \sim t_{n-2}$$

- (1-α)x100% confidence interval for slope

$$(\hat{\beta}_0 - t(\alpha/2, n-2)\,\text{se}(\hat{\beta}_0), \hat{\beta}_0 + t(\alpha/2, n-2)\text{se}(\hat{\beta}_0))$$

- Example: production data

Inferences about the Intercept

$$H_0 : \beta_0 = \beta_0^0 \quad \Longrightarrow \quad T = \frac{\hat{\beta}_1 - \beta_1^0}{\mathrm{se}(\hat{\beta}_1)} \sim t_{n-2} \text{ when } H_0 \text{ is true.}$$

- Consider testing hypotheses

$$H_0 : \beta_0 = 0 \text{ vs } H_A : \beta_0 \neq 0$$

$$T = \frac{\hat{\beta}_0}{se(\beta_0)} \sim t_{n-2} \text{ when } H_0 \text{ is true.}$$

- Example: production data

Distribution of the population regression line

- Let x* denote a certain X-value. The population regression line is $E(Y \mid X = x^*) = \beta_0 + \beta_1 x^*$

- The estimator of this unknown conditional expectation is

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

$$E(\hat{y}^* \mid X = x^*) = E(\hat{\beta}_0 + \hat{\beta}_1 x^* \mid X = x^*) = \beta_0 + \beta_1 x^*$$

$$Var(\hat{y}^* \mid X = x^*) = Var(\hat{\beta}_0 + \hat{\beta}_1 x^* \mid X = x^*) = \frac{\sigma^2}{n} + \frac{\sigma^2 (x^* - \overline{x})^2}{SXX}$$

- It can be shown that

$$\hat{y}^* = \hat{y} \mid X = x^* \sim N\left(\beta_0 + \beta_1 x^*, \sigma^2\left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX}\right)\right)$$

Confidence interval for the population regression line

$$\hat{y}^* = \hat{y} \mid X = x^* \sim N\left(\beta_0 + \beta_1 x^*, \sigma^2\left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX}\right)\right)$$

$$Z = \frac{\hat{y}^* - (\beta_0 + \beta_1 x^*)}{\sigma\sqrt{(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX})}} \sim N(0,1) \quad \Rightarrow \quad T = \frac{\hat{y}^* - (\beta_0 + \beta_1 x^*)}{S\sqrt{(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX})}} \sim t_{n-2}$$

- A 100(1 – α)% confidence interval for the population regression line (mean response) at $X = X^*$

$$\hat{y}^* \pm t(\alpha/2, n-2) S\sqrt{(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX})}$$

Estimation vs Prediction

- Estimation: to guess a function of parameters (non-random quantities)

- Prediction: to guess a function of actual data points (random quantities)

- Confidence interval: interval estimation of a function of parameters

- Prediction interval: interval predictions of a function of the actual data points

Distribution of predicted value of Y

$$Y^* = \beta_0 + \beta_1 x^* + e^*$$

$$\begin{aligned}\mathrm{E}(Y^* - \hat{y}^*) &= \mathrm{E}(Y - \hat{y} \mid X = x^*) \\ &= \mathrm{E}(Y \mid X = x^*) - \mathrm{E}(\hat{\beta}_0 + \hat{\beta}_1 x \mid X = x^*) \\ &= 0\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(Y^* - \hat{y}^*) &= \mathrm{Var}(Y - \hat{y} \mid X = x^*) \\ &= \mathrm{Var}(Y \mid X = x^*) + \mathrm{Var}(\hat{y} \mid X = x^*) - 2\mathrm{Cov}(Y, \hat{y} \mid X = x^*) \\ &= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX}\right] - 0\end{aligned}$$

$$Y^* - \hat{y}^* \sim N\left(0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX}\right]\right) \Rightarrow T = \frac{Y^* - \hat{y}^*}{S\sqrt{(1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{SXX})}} \sim t_{n-2}$$

Deviation between $\mathrm{E}(Y \mid X = x^*)$ and *its estimate* plus the random fluctuation in e_i

Prediction interval

- A 100(1–α)% prediction interval for *Y**, the value of *Y* at *X* = *x**

$$\hat{y}^* \pm t(\alpha/2, n-2)S\sqrt{(1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{SXX})}$$

- The prediction interval is wider than the confidence interval because the uncertainty of prediction is the uncertainty of LS estimates plus the uncertainty of the random error.

- The two intervals have the same center.

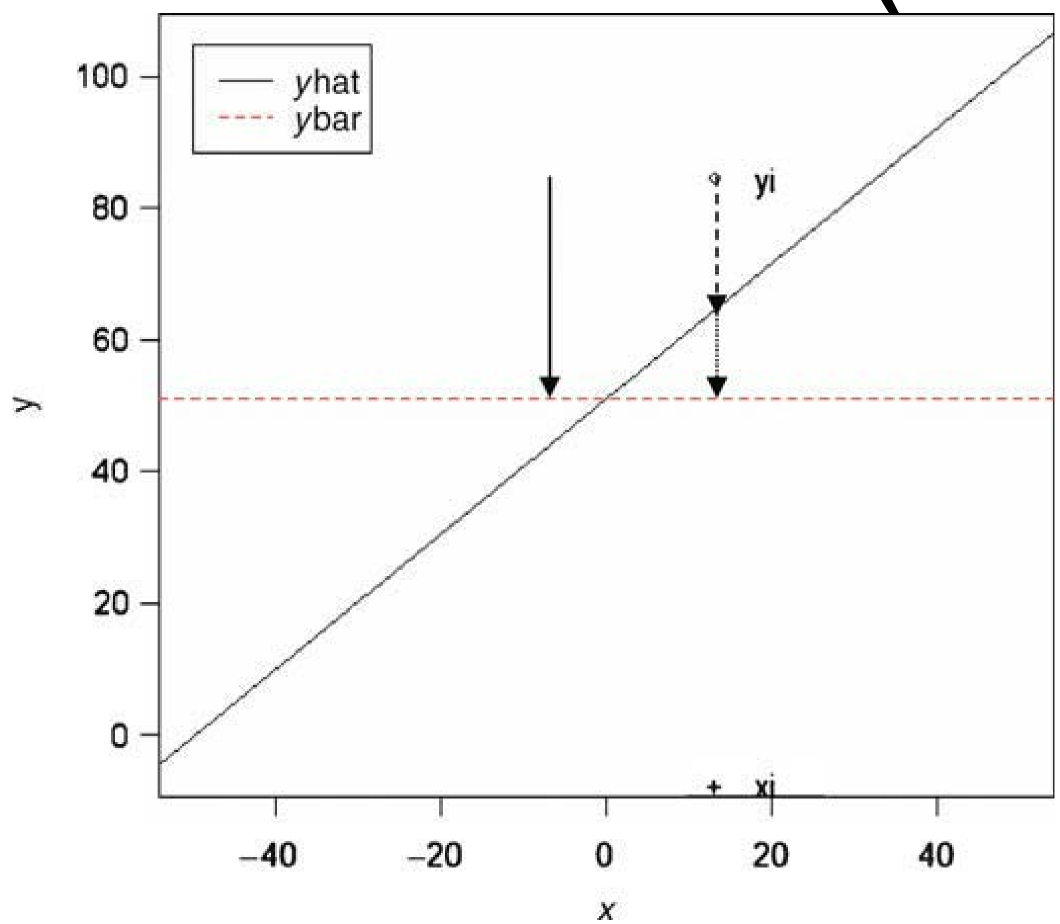
- The width of these intervals decreases if

    - x* gets close to the mean of X

    - n or α increases

    - RSS decreases

Analysis of variance (ANOVA)



$$y_i - \bar{y} = \left(y_i - \hat{y}_i\right) + \left(\hat{y}_i - \bar{y}\right)$$

normal equations

$$\sum_i^n (y_i - \bar{y})^2 = \sum_i^n (\hat{y}_i - \bar{y})^2 + \sum_i^n (y_i - \hat{y})^2$$

SST	= SSreg	+ RSS
Total sample variability	= Variability explained by the model	+ Unexplained (or error) variability

$$\text{SST} = SYY = \sum_i^n (y_i - \bar{y})^2 \qquad \text{SSreg} = \sum_i^n (\hat{y}_i - \bar{y})^2 \qquad \text{RSS} = \sum_i^n (y_i - \hat{y}_i)^2$$

Source of variation	Degrees of freedom (df)	Sum of squares (SS)	Mean square (MS)	F
Regression	1	SSreg	SSreg/1	$F = \dfrac{\text{SSreg}/1}{\text{RSS}/(n-2)}$
Residual	$n - 2$	RSS	RSS/($n - 2$)	
Total	$n - 1$	SST		

- Consider t-statistic for testing

$$H_0 : \beta_1 = 0 \text{ against } H_A : \beta_1 \neq 0$$

$$T = \frac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)} \sim t_{n-2} \quad \xrightarrow{F=T^2} \quad F = \frac{\text{SSreg}/1}{\text{RSS}/(n-2)} \sim F_{1,n-2}$$

- Goodness-of-fit: Coefficient of determination (R^2)
  - The proportion of the total sample variability in the Y's explained by the regression model
  - $$R^2 = \frac{\text{SSreg}}{\text{SST}} = 1 - \frac{\text{RSS}}{\text{SST}} \quad \Longrightarrow$$

R^2 = squared sample correlation coefficient of y and $\hat{y}$

$$R^2 = r_{XY}^2 = \frac{\left(\sum_i^n (y_i - \bar{y})(\hat{y}_i - \bar{y}_i)\right)^2}{\sum_i^n (y_i - \bar{y})^2 \sum_i^n (\hat{y}_i - \bar{y}_i)^2}$$

- It is left as an exercise to show that

$$RSS = \sum\{(y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x})\}^2 = SYY - \hat{\beta}_1^2 SXX$$

- Example:

```
Analysis of Variance Table
Response: RunTime
             Df  Sum Sq  Mean Sq  F value    Pr(>F)
RunSize       1 12868.4  12868.4   48.717 1.615e-06  ***
Residuals    18  4754.6    264.1
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

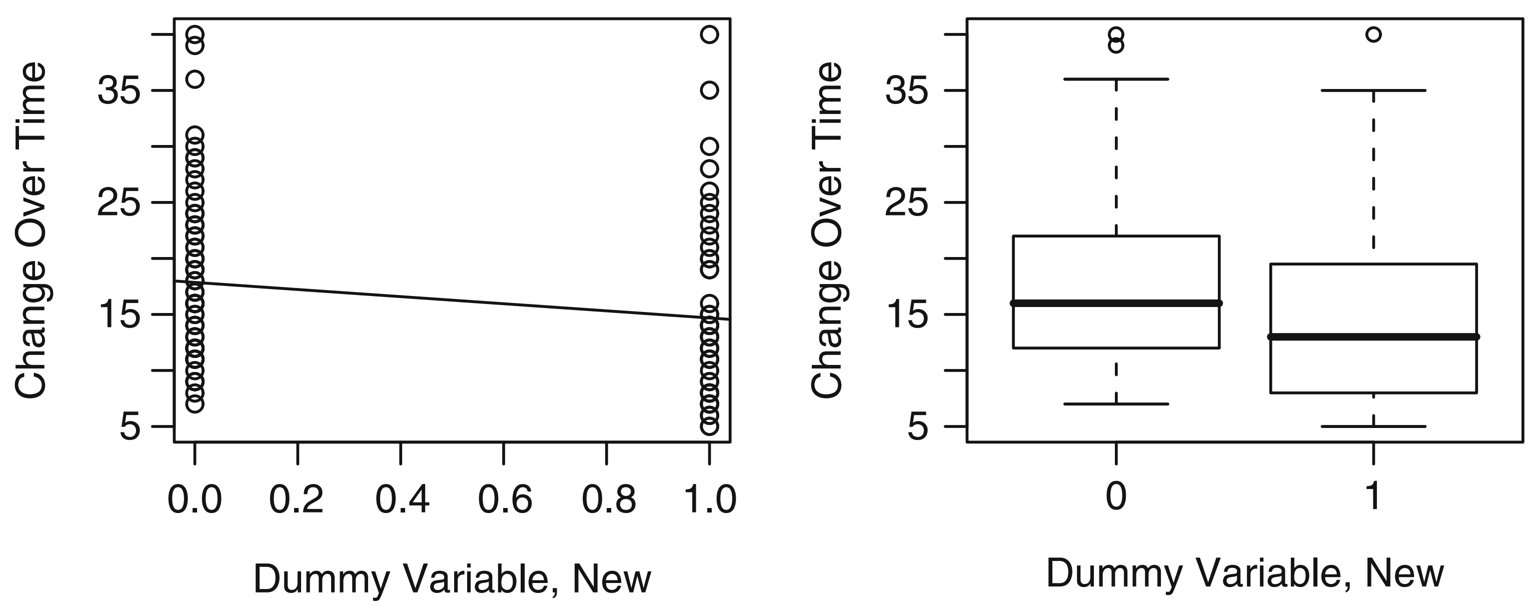
  - Calculate SST and its DF
  - Calculate R^2
  - Verify $F=T^2$

Dummy variable regression

- So far, we have considered a quantitative predictor.
- Now Consider a predictor is categorical with two values (e.g., gender)

Table 2.2 Change-over time data (changeover_times.txt)

Method	Y, Change-over time	X, New
Existing	19	0
Existing	24	0
Existing	39	0
.	.	.
New	14	1
New	40	1
New	35	1



- (Change-over time data) A large food processing center that needs to be able to switch from one type of package to another quickly to react to changes in order patterns.

One-sided alternative

```
Coefficients:
            Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) 17.8611      0.8905    20.058     <2e-16   ***
New         -3.1736      1.4080    -2.254     0.0260   *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 7.556 on 118 degrees of freedom
Multiple R-Squared: 0.04128, Adjusted R-squared: 0.03315
F-statistic: 5.081 on 1 and 118 DF, p-value: 0.02604
```

- Consider testing $H_0 : \beta_1 = 0$ against $H_A : \beta_1 < 0$

$$T = \frac{\hat{\beta}_1 - 0}{\mathrm{se}(\hat{\beta}_1)} \sim t_{n-2} \text{ when } H_0 \text{ is true.}$$

- Reject H_0 at significance level α if -2.254 < -t(α, n-2)

$$p - value = P(T < -2.254 \text{ when } H_0 \text{ is true}) = \frac{0.026}{2} = 0.013$$

- Mean change-over time of the new method

$$17.8611 + (-3.1736) \times 1 = 14.6875 = 14.7 \text{ minutes}$$

- Mean change-over time of the existing method

$$17.8611 + (-3.1736) \times 0 = 17.8611 = 17.9 \text{ minutes}$$

- A 95% confidence interval for the reduction in mean change-over time due to the new method

$$(\hat{\beta}_1 - t(\alpha/2, n-2)\mathrm{se}(\hat{\beta}_1), \hat{\beta}_1 + t(\alpha/2, n-2)\mathrm{se}(\hat{\beta}_1))$$