# LECTURE 13

REGRESSION ANALYSIS

# SIMPLE LINEAR REGRESSION

- In simple regression our objective is to study the relationship between two variables X and Y. It is the process of estimating a functional relationship between X and Y.
- The function is a mathematical relationship enabling us to predict what values of one variable (Y) correspond to given values of another variable (X).

  Y: is referred to as the dependent variable, the response variable or the predicted variable.

  X: is referred to as the independent variable, the explanatory variable or the predictor variable.

- Another way to study relationship between two variables is **correlation.** It involves measuring the direction and the strength of the **linear** relationship.

# SIMPLE LINEAR REGRESSION MODEL

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

Where

$b_0$= y-intercept

$b_1$= slope of the line

$\varepsilon$ = error variable

This model is

**Simple**: only one X

**Linear in the parameters**: No parameter appears as exponent or is multiplied or divided by another parameter

**Linear in the predictor variable (X)**: X appears only in the first power.

# Examples

- Multiple Linear Regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

- Polynomial Linear Regression:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X^2_i + \varepsilon_i$$

- Linear Regression:

$$\log_{10}(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 \exp(X_i) + \varepsilon_i$$

- Nonlinear Regression:

$$Y_i = \beta_0 / (1 + \beta_1 \exp(\beta_2 X_i)) + \varepsilon_i$$

# THE ERROR TERM

$$\varepsilon = Y - f(X)$$

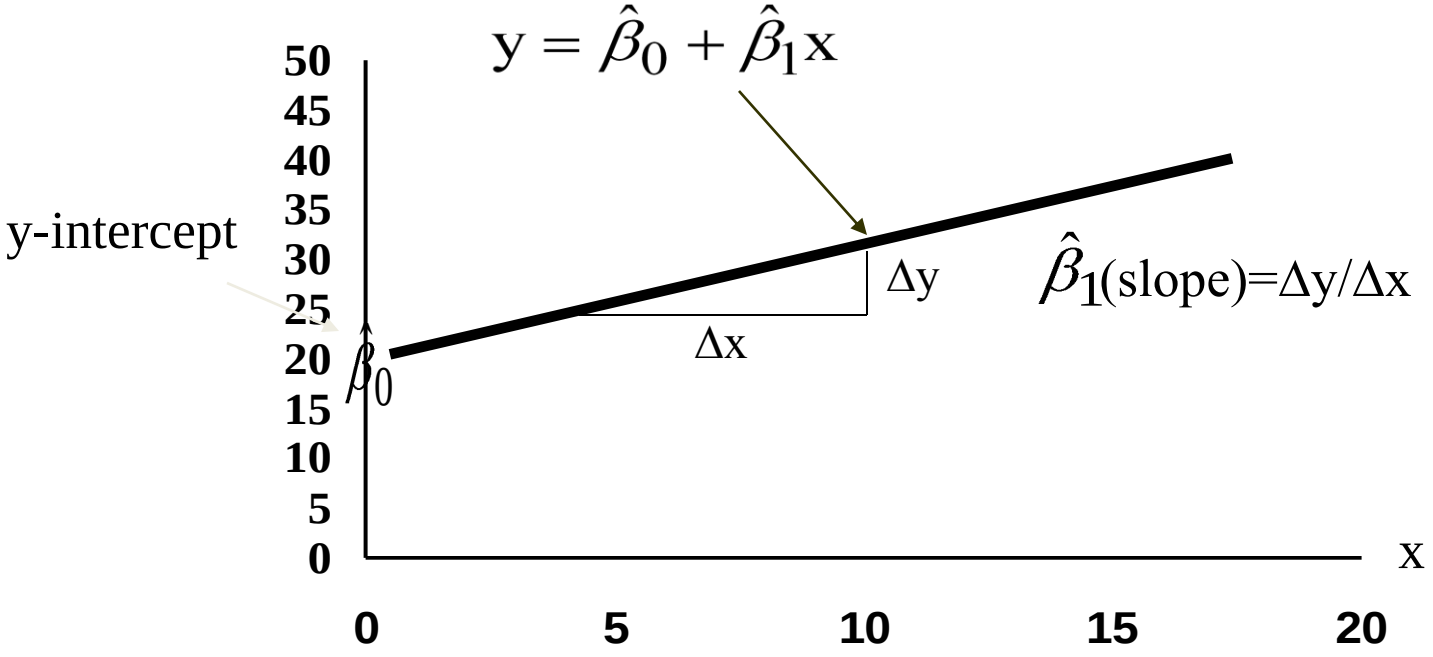
or, emphasizing that f(X) depends on unknown parameters.

$$Y = f(X \mid \beta_0, \beta_1) + \varepsilon$$

What if we don't know the functional form of the relationship?

- Look at a scatter plot of the data for suggestions. The scatter plot shows that the points are not on a line, and so, in addition to the relationship.
- Hypothesize about the nature of the underlying process. Often the hypothesized processes will suggest a functional form.
- We assume that the errors are normal, mutually independent, and have variance $\sigma^2$.

# DETERMINISTIC COMPONENT OF MODEL



# PARAMETER ESTIMATION

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\text{for } i = 1, 2, \ldots, n$$

Objective: Minimize the difference between the observation and its prediction according to the line.

$$\varepsilon_i = y_i - \hat{y}_i$$
$$= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$\hat{y}_i = \text{predicted y value when x} = x_i$$

We want the line which is best for all points. This is done by finding the values of $b_0$ and $b_1$ which minimizes some sum of errors. There are a number of ways of doing this. Consider these two:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} |\varepsilon_i|$$

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \varepsilon_i^{\,2} \longleftarrow$$

Sum of squared residuals

The method of least squares produces estimates with statistical properties (e.g. sampling distributions) which are easier to determine.

$\hat{\beta}_0 \quad \hat{\beta}_1$ Referred to as least squares estimates.

# NORMAL EQUATIONS

Calculus is used to find the least squares estimates.

$$E(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial E}{\partial \beta_0} = 0$$

$$\frac{\partial E}{\partial \beta_1} = 0$$

Solve this system of two equations in two unknowns.

Note: The parameter estimates will be functions of the data, hence they will be statistics.

# Sums of Squares

Let:

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

$$= (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2$$

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

Sums of squares of x.

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

$$= (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2$$

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

Sums of squares of y.

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

$$= (x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_i - \bar{y})$$

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

Sums of cross products of x and y.

# PARAMETER ESTIMATES

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$$

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

# TESTING FOR A STATISTICALLY SIGNIFICANT REGRESSION

$H_o$: There is no relationship between Y and X.

$H_A$: There is a relationship between Y and X.

Which of two competing models is more appropriate?

$$\text{Linear Model}: \quad Y = \beta_0 + \beta_1 X + \varepsilon$$

$$\text{Mean Model}: \quad Y = \mu + \varepsilon$$

We look at the sums of squares of the prediction errors for the two models and decide if that for the linear model is **significantly smaller** than that for the mean model.

# SUMS OF SQUARES ABOUT THE MEAN (TSS)

**Sum of squares about the mean**: sum of the prediction errors for the null (mean model) hypothesis.

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

TSS is actually a measure of the variance of the responses.

# RESIDUAL SUMS OF SQUARES

**Sum of squares for error**: sum of the prediction errors for the alternative (linear regression model) hypothesis.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

SSE measures the variance of the residuals, the part of the response variation that is not explained by the model.

# REGRESSION SUMS OF SQUARES

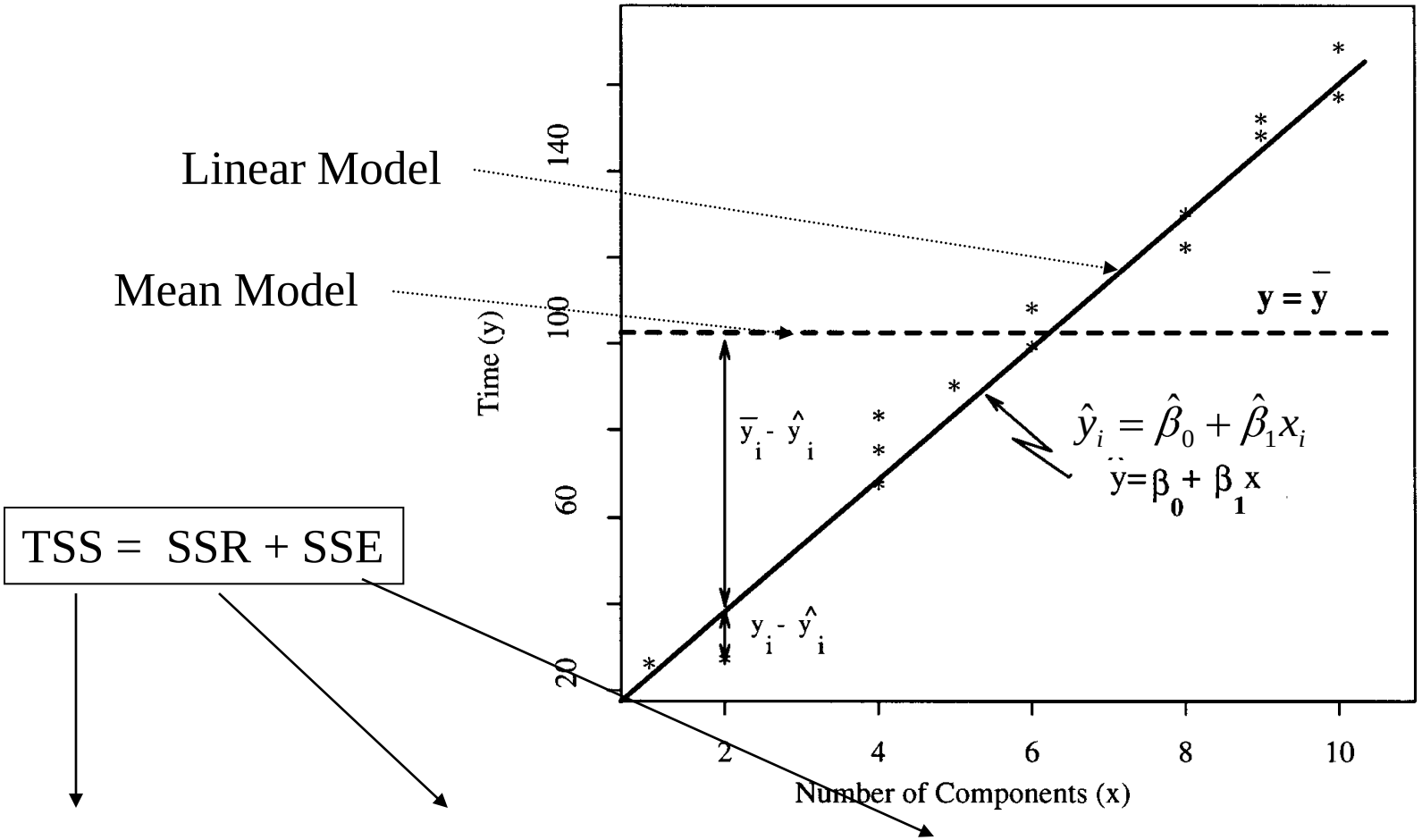
**Sum of squares due to the regression**: difference between TSS and SSE, i.e. SSR = TSS – SSE.

$$SSR = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (\bar{y} - \hat{y}_i)^2$$

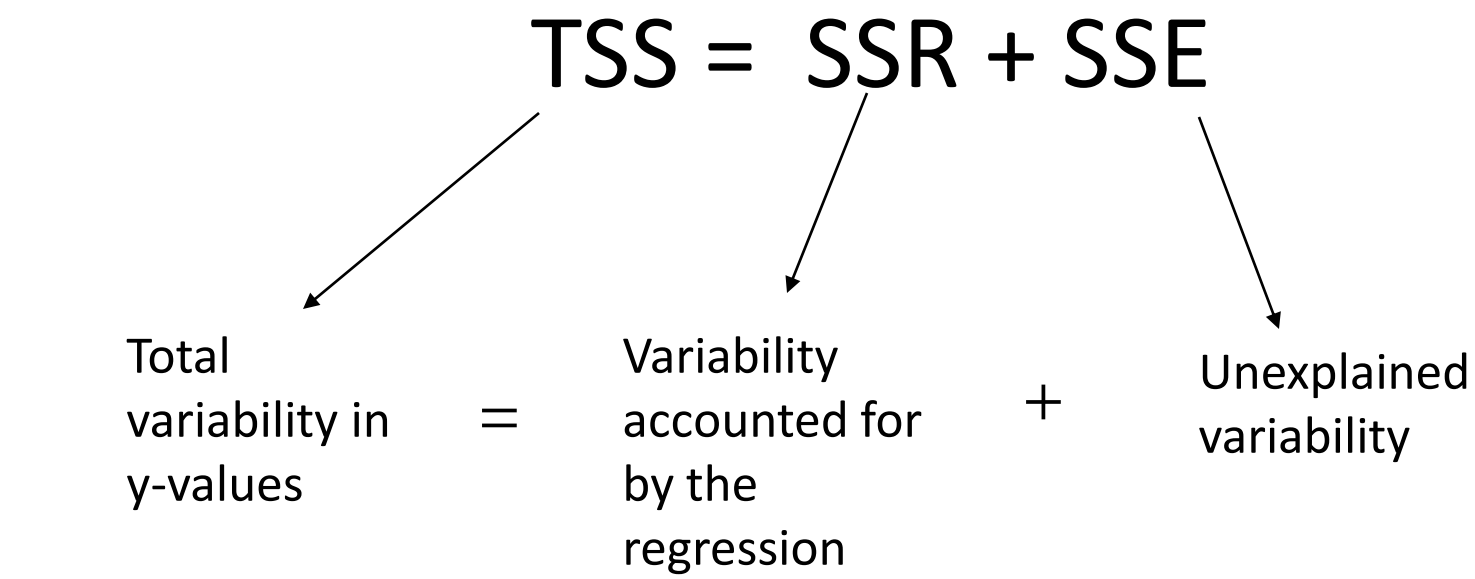
SSR measures how much variability in the response is explained by the regression.

# GRAPHICAL VIEW




TSS = SSR + SSE

Total variability in y-values = Variability accounted for by the regression + Unexplained variability




**regression model fits well**

Then SSR approaches TSS and SSE gets small.

**regression model adds little**

Then SSR approaches 0 and SSE approaches TSS.

# MEAN SQUARE TERMS

Mean Square Total

$$\begin{aligned}\hat{\sigma}_{\mathrm{T}}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 \\ &= \frac{\mathrm{TSS}}{n-1} \\ &= \mathrm{MST}\end{aligned}$$

$$\begin{aligned}\hat{\sigma}_{\mathrm{R}}^2 &= \sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \\ &= \frac{\mathrm{SSR}}{1} \\ &= \mathrm{MSR}\end{aligned}$$

Regression Mean Square:

Residual Mean Square

$$\begin{aligned}\hat{\sigma}_{\varepsilon}^2 &= \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \\ &= \frac{\mathrm{SSE}}{n-2} \\ &= \mathrm{MSE}\end{aligned}$$

# F TEST FOR SIGNIFICANT REGRESSION

Both MSE and MSR measure the same underlying variance quantity under the assumption that the null (mean) model holds.

$$\sigma_R^2 \approx \sigma_\varepsilon^2$$

Under the alternative hypothesis, the MSR should be much greater than the MSE.

$$\sigma_R^2 > \sigma_\varepsilon^2$$

Placing this in the context of a test of variance.

$$F = \frac{\sigma_R^2}{\sigma_\varepsilon^2} = \frac{\text{MSR}}{\text{MSE}}$$

Test Statistic

F should be near 1 if the regression is *not significant*, i.e. $H_0$: mean model holds.

$H_0$: No significant regression fit.
$H_A$: X is a significant predictor of Y.

Test Statistic: $$F = \frac{MSR}{MSE}$$

Reject $H_0$ if: $$F > F_{1,n-2,\alpha}$$

Where $\alpha$ is the probability of a type I error.

# ASSUMPTIONS

1. $e_1, e_2, \ldots e_n$ are independent of each other.
2. The $e_i$ are normally distributed with mean zero and have common variance $s^2$.

**How do we check these assumptions?**

I. Appropriate graphs.
II. Correlations (more later).
III. Formal goodness of fit tests.

# ANALYSIS OF VARIANCE TABLE

We summarize the computations of this test in a table.

| Source | Sums of Squares SSQ | Degrees of Freedom DF | Mean Squares MS | F |
|---|---|---|---|---|
| Regression | SSR | 1 | MSR | $F = \frac{MSR}{MSE}$ |
| Error | SSE | n-2 | MSE | |
| Total | SSM | n-1 | | |

TSS

# Parameter Standard Error Estimates

Under the assumptions for regression inference, the least squares estimates themselves are random variables.

1. $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n$ are independent of each other.
2. The $\varepsilon_i$ are normally distributed with mean zero and have common variance $\sigma^2$.

Using some more calculus and mathematical statistics we can determine the distributions for these parameters.

$$\hat{\beta}_0 \mapsto \mathrm{N}\left(\beta_0, \sigma^2 \frac{\sum x_i^{\,2}}{nS_{XX}}\right) \qquad \hat{\beta}_1 \mapsto \mathrm{N}\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right)$$

# $R^2$ AND $R^2$ ADJUSTED

$$R^2 = 1 - \frac{SSE}{SS_y} \qquad R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SS_y}$$

- $R^2$ measures the degree of linear association between X and Y.
- So, an $R^2$ close to 0 does not necessarily indicate that X and Y are unrelated (relation can be nonlinear)
- Also, a high $R^2$ does not necessarily indicate that the estimated regression line is a good fit.
- As more and more X's are added to the model, $R^2$ always increases. $R^2$adj accounts for the number of parameters in the model.

# TESTING THE SLOPE

- Are X and Y linearly related?

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

- Test Statistic:

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \qquad \text{where} \quad s_{\hat{\beta}_1} = \frac{s_\varepsilon}{\sqrt{SS_x}}$$

- The Rejection Region:

  Reject $H_0$ if $t < -t_{\alpha/2,n-2}$ or $t > t_{\alpha/2,n-2}$.

- If we are testing that high x values lead to high y values, $H_A: \beta_1 > 0$.
- Then, the rejection region is $t > t_{\alpha,n-2}$.

- If we are testing that high x values lead to low y values or low x values lead to high y values, $H_A: \beta_1 < 0$.
- Then, the rejection region is $t < -t_{\alpha,n-2}$.

# PREDICTION AND CONFIDENCE INTERVALS

- Prediction Interval of y for $x=x_g$: The confidence interval for predicting the particular value of y for a given x

$$\hat{y} \pm t_{\alpha/2,n-2} s_e \sqrt{1+\frac{1}{n}+\frac{(x_g-\bar{x})^2}{SS_x}}$$

- Confidence Interval of $E(y|x=x_g)$: The confidence interval for estimating the expected value of y for a given x

$$\hat{y} \pm t_{\alpha/2,n-2} s_e \sqrt{\frac{1}{n}+\frac{(x_g-\bar{x})^2}{SS_x}}$$

# Example

- An educational economist wants to establish the relationship between an individual's income and education. He takes a random sample of 10 individuals and asks for their income ( in $1000s) and education ( in years). The results are shown below. Find the least squares regression line.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Education** | 11 | 12 | 11 | 15 | 8 | 10 | 11 | 12 | 17 | 11 |
| **Income** | 25 | 33 | 22 | 41 | 18 | 28 | 32 | 24 | 53 | 26 |

First Step:

$$\sum x_i = 118$$

$$\sum x_i^2 = 1450$$

$$\sum y_i = 302$$

$$\sum y_i^2 = 10072$$

$$\sum x_i y_i = 3779$$

## Sum of Squares:

$$SS_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 3779 - \frac{(118)(302)}{10} = 215.4$$

$$SS_x = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 1450 - \frac{(118)^2}{10} = 57.6$$

Therefore,

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x} = \frac{215.4}{57.6} = 3.74$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{302}{10} - 3.74\frac{118}{10} = -13.93$$

# The Least Squares Regression Line

- The least squares regression line is

$$\hat{y} = -13.93 + 3.74x$$

- Interpretation of coefficients:

*The sample slope $\hat{\beta}_1 = 3.74$ tells us that on average for each additional year of education, an individual's income rises by $3.74 thousand.

- The y-intercept is $\hat{\beta}_0 = -13.93$ . This value is the expected (or average) income for an individual who has 0 education level (which is meaningless here)

# Example

- In baseball, the fans are always interested in determining which factors lead to successful teams. The table below lists the team batting average and the team winning percentage for the 14 league teams at the end of a recent season.

| Team-B-A | Winning% |
|---|---|
| 0.254 | 0.414 |
| 0.269 | 0.519 |
| 0.255 | 0.500 |
| 0.262 | 0.537 |
| 0.254 | 0.352 |
| 0.247 | 0.519 |
| 0.264 | 0.506 |
| 0.271 | 0.512 |
| 0.280 | 0.586 |
| 0.256 | 0.438 |
| 0.248 | 0.519 |
| 0.255 | 0.512 |
| 0.270 | 0.525 |
| 0.257 | 0.562 |

y = winning % and
x = team batting average

# a) LS Regression Line

$$\sum x_i = 3.642, \sum x_i^2 = 0.949$$

$$\sum y_i = 7.001, \sum y_i^2 = 3.549$$

$$\sum x_i y_i = 1.824562$$

$$SS_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 1.824562 - \frac{(3.642)(7.001)}{14} = 0.0033$$

$$SS_x = \sum x_i^2 - \frac{\sum x_i^{\,2}}{n} = 0.948622 - \frac{(3.642)^2}{14} = 0.00118$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x} = \frac{0.003302}{0.001182} = 0.7941$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 0.5 - (0.7941)0.26 = 0.2935$$

- The least squares regression line is

$$\hat{y} = 0.2935 + 0.7941x$$

- The meaning $\hat{\beta}_1 = 0.7941$ is for each additional batting average of the team, the winning percentage increases by an average of 79.41%.

b) Standard Error of Estimate

$$SSE = S_{yy} - \left(\frac{S_{xy}^2}{S_{xx}}\right) = \left(\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}\right) - \left(\frac{S_{xy}^2}{S_{xx}}\right)$$

$$= (3.548785 - \frac{7.001^2}{14}) - \frac{0.003302^2}{0.00182} = 0.03856$$

So, $$s_\varepsilon^2 = \frac{SSE}{n-2} = \frac{0.03856}{14-2} = 0.00321 \text{ and } s_\varepsilon = \sqrt{s_\varepsilon^2} = 0.0567$$

- Since $s_\varepsilon$=0.0567 is small, we would conclude that “s” is relatively small, indicating that the regression line fits the data quite well.

c) Do the data provide sufficient evidence at the 5% significance level to conclude that higher team batting average lead to higher winning percentage?

$$H_0: \quad \beta_1 = 0$$
$$H_A: \quad \beta_1 > 0$$

$$\text{Test statistic: } t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} = 1.69 \qquad \text{(p-value=.058)}$$

**Conclusion:** Do not reject $H_0$ at $\alpha = 0.05$. The higher team batting average does not lead to higher winning percentage.

d)Coefficient of Determination

$$R^2 = \frac{SS_{xy}^2}{SS_x - SS_y} = 1 - \frac{SSE}{SS_y} = 1 - \frac{0.03856}{0.04778} = 0.1925$$

The 19.25% of the variation in the winning percentage can be explained by the batting average.

e) Predict with 90% confidence the winning percentage of a team whose batting average is 0.275.

$$\hat{y} = 0.2935 + 0.7941(0.275) = 0.5119$$

$$\hat{y} \pm t_{\alpha/2,n-2} s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{x})^2}{SS_x}} =$$

$$0.5119 \pm (1.782)(0.0567)\sqrt{1 + \frac{1}{14} + \frac{(0.275 - 0.2601)^2}{0.001182}}$$

$$0.5119 \pm 0.1134$$

90% PI for y: $(0.3985, 0.6253)$

•The prediction is that the winning percentage of the team will fall between 39.85% and 62.53%.