Lecture 12
Linear Regression:
Test and Confidence Intervals

Fall 2013

Prof. Yao Xie, yao.xie@isye.gatech.edu

H. Milton Stewart School of Industrial Systems & Engineering

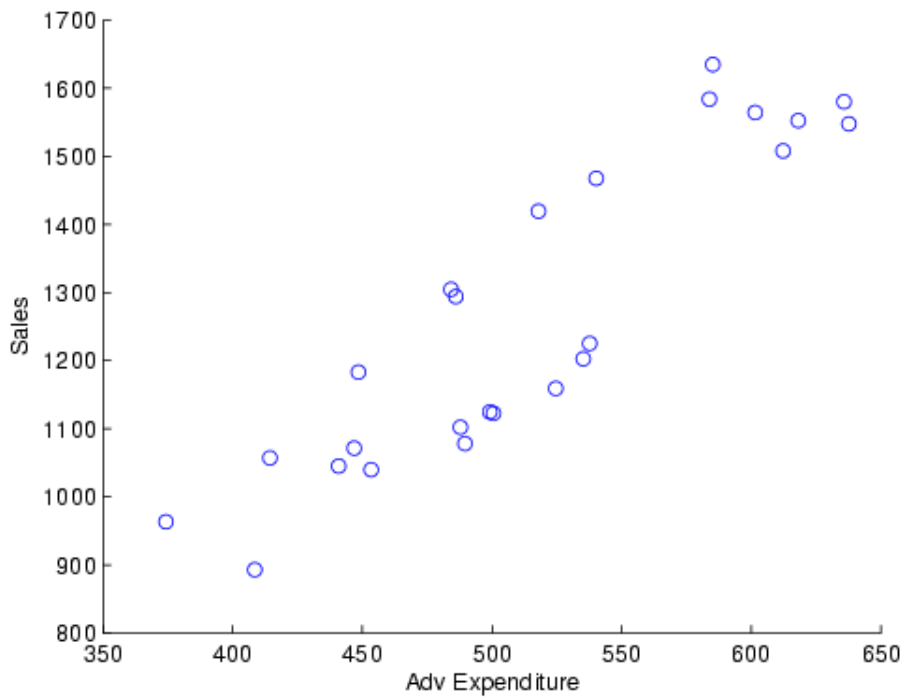
Georgia Tech

Outline

- Properties of $\hat{\beta}_1$ and $\hat{\beta}_0$ as point estimators
- Hypothesis test on slope and intercept
- Confidence intervals of slope and intercept
- Real example: house prices and taxes

Regression analysis

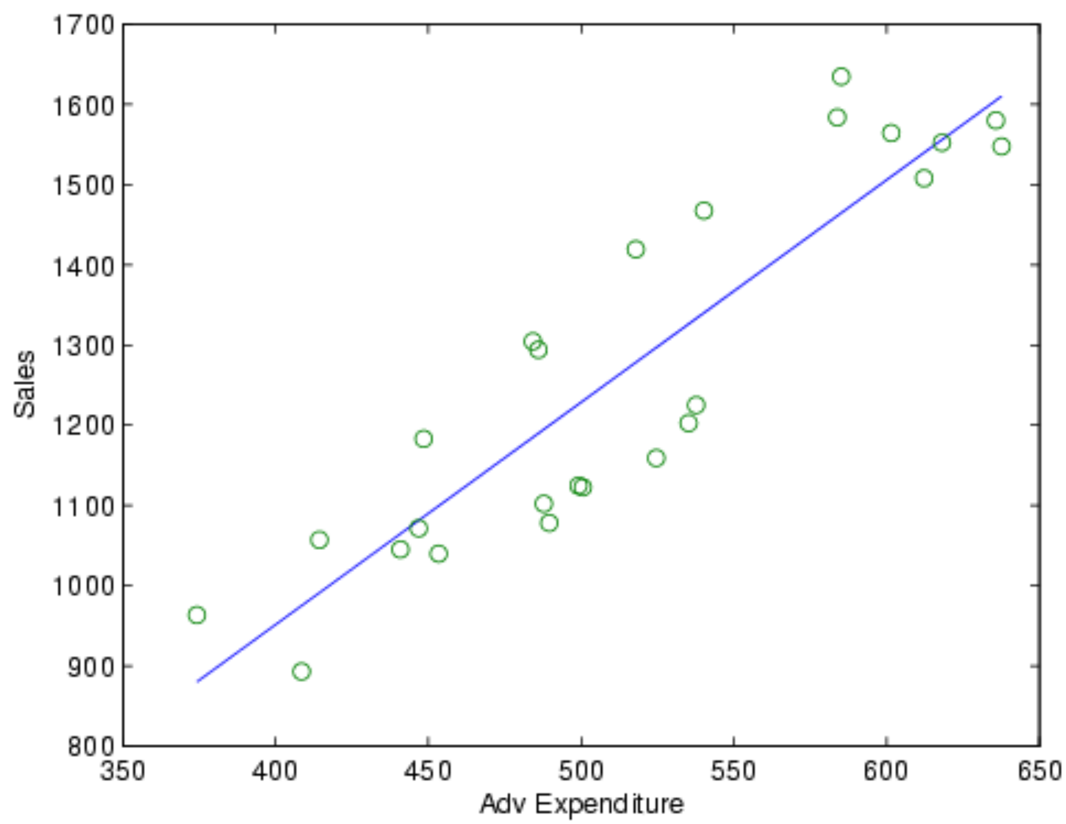
- Step 1: graphical display of data — scatter plot: sales vs. advertisement cost



- calculate correlation

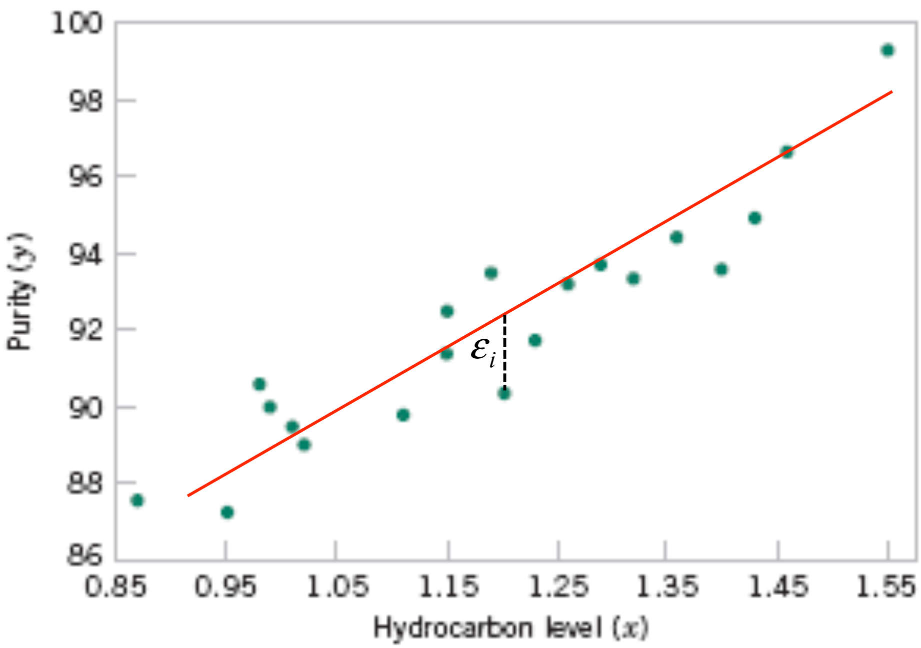
$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad -1 \le \hat{\rho} \le 1$$

- Step 2: find the relationship or association between Sales and Advertisement Cost — Regression



Simple linear regression

Based on the scatter diagram, it is probably reasonable to assume that the mean of the random variable Y is related to X by the following simple linear regression model:



Response

Regressor or Predictor

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1, 2, \cdots, n$$

$$\varepsilon_i \sim \mathrm{N}\left(0, \ \sigma^2\right)$$

Intercept

Slope

Random error

where the slope and intercept of the line are called regression coefficients.

•The case of simple linear regression considers a single regressor or predictor x and a dependent or response variable Y.

Regression coefficients

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \qquad (11\text{-}10)$$

$$S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} \qquad (11\text{-}11)$$

$$\left.\begin{array}{l} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\[2ex] \hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} \end{array}\right\} \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Fitted (estimated) regression model

Caveat: regression relationship are valid only for values of the regressor variable within the range the original data. Be careful with extrapolation.

Estimation of variance

- Using the fitted model, we can estimate value of the response variable for given predictor

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Residuals: $r_i = y_i - \hat{y}_i$
- Our model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1,\ldots,n$, $Var(\varepsilon_i) = \sigma^2$
- Unbiased estimator (MSE: Mean Square Error)

$$\hat{\sigma}^2 = MSE = \frac{\sum_{i=1}^{n} r_i^2}{n-2}$$

Punchline

- the coefficients

$$\hat{\beta}_1 \text{ and } \hat{\beta}_0$$

and both calculated from data, and they are subject to error.

- if the true model is $y = \beta_1 x + \beta_0$, $\hat{\beta}_1$ and $\hat{\beta}_0$ are point estimators for the true coefficients

- we can talk about the ``accuracy'' of $\hat{\beta}_1$ and $\hat{\beta}_0$

Assessing linear regression model

- Test hypothesis about true slope and intercept

$$\beta_1 = ?, \qquad \beta_0 = ?$$

- Construct confidence intervals

$$\beta_1 \in \left[\hat{\beta}_1 - a, \hat{\beta}_1 + a \right] \quad \beta_0 \in \left[\hat{\beta}_0 - b, \hat{\beta}_0 + b \right] \quad \text{with probability } 1 - \alpha$$

- Assume the errors are normally distributed

$$\varepsilon_i \sim \mathrm{N}\left(0, \ \sigma^2 \right)$$

Properties of Regression Estimators

slope parameter β_1	intercept parameter β_0
$E(\hat{\beta}_1) = \beta_1$	$E(\hat{\beta}_0) = \beta_0$
$V(\hat{\beta}_1) = \dfrac{\sigma^2}{S_{xx}}$	$V(\hat{\beta}_0) = \sigma^2 \left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}} \right]$
unbiased estimator	unbiased estimator

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

Standard errors of coefficients

- We can replace σ^2 with its estimator $\hat{\sigma}^2$ …

$$\hat{\sigma}^2 = MSE = \frac{\sum_{i=1}^{n} r_i^2}{n-2}$$

$$r_i = y_i - \hat{y}_i \qquad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Using results from previous page, estimate the

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \qquad \text{and} \qquad se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$$

Hypothesis test in simple linear regression

- we wish to test the hypothesis whether the slope equals a constant $\beta_{1,0}$

$$H_0\colon \beta_1 = \beta_{1,0}$$
$$H_1\colon \beta_1 \neq \beta_{1,0}$$

- e.g. relate **ads** to **sales**, we are interested in study whether or not increase a \$ on ads will increase \$ $\beta_{1,0}$ in sales?
- sale = $\beta_{1,0}$ ads + constant?

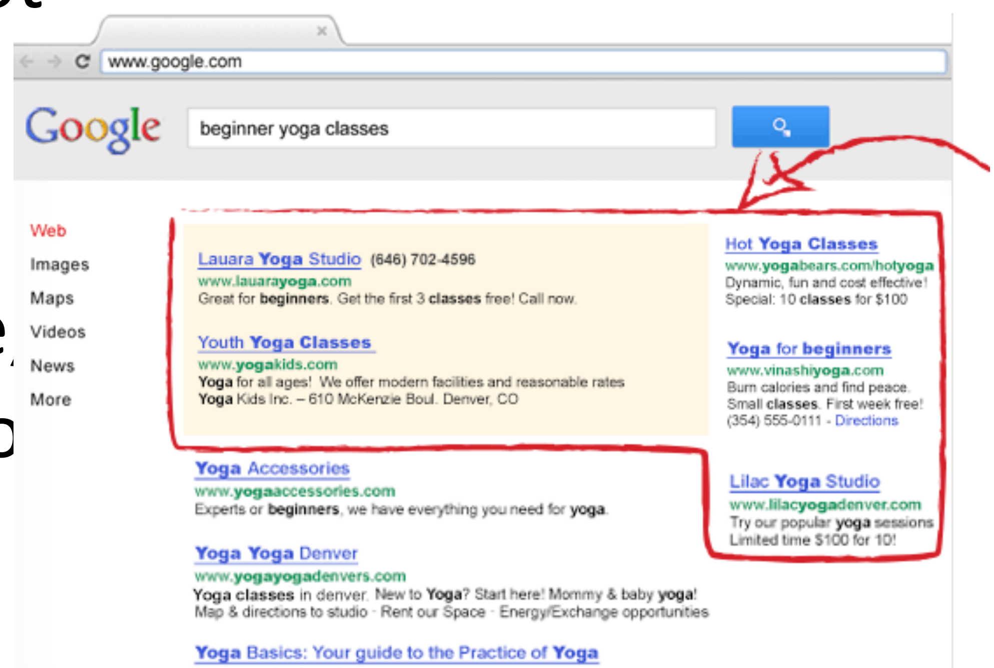
A related and important question…

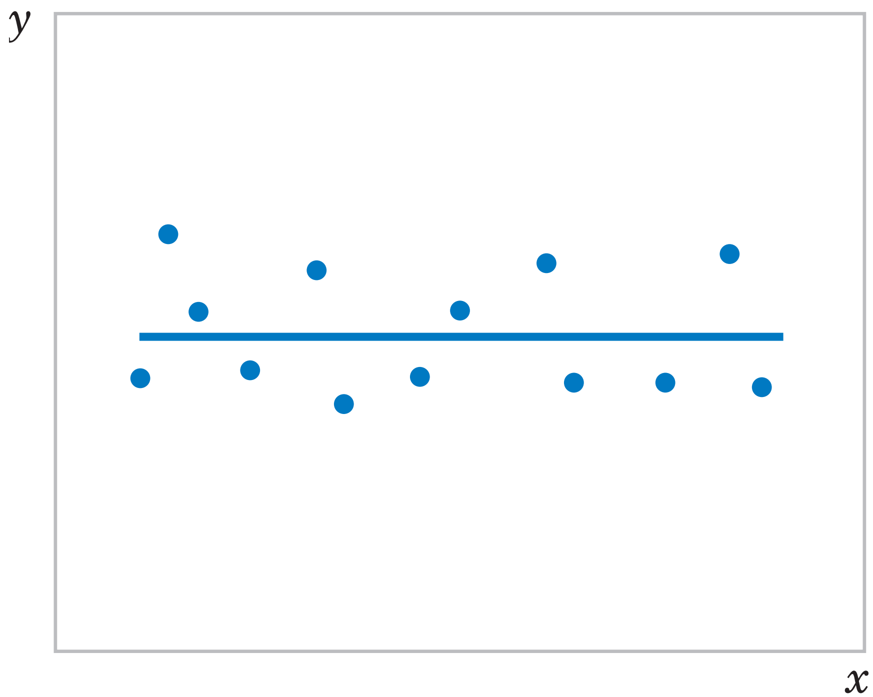
- whether or not the slope is zero?

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

Significance of regression

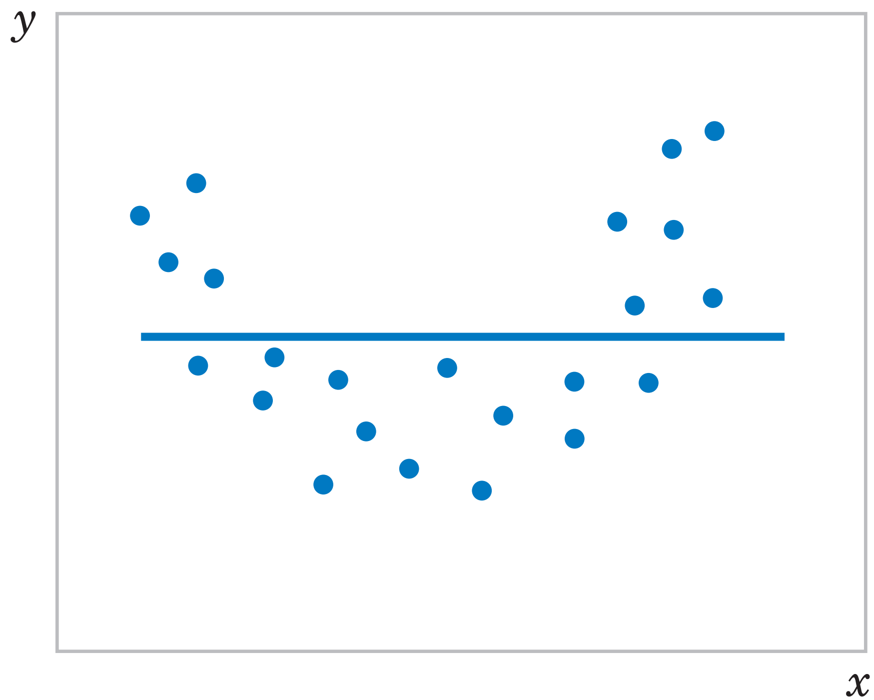
- if $\beta_1 = 0$, that means Y does not depend on X, i.e.,
- Y and X are **independent**
- In the advertisement example, does **ads** increase **sales**? or no effect?





(a)

- H_0 not rejected



(b)

- H_0 rejected

Use t-test for slope

Under H_0

slope parameter β_1
$$E(\hat{\beta}_1) = \beta_{1,0}$$
$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$\hat{\beta}_1 \sim N\left(\beta_{1,0}, \quad \sigma^2 / S_{xx} \right)$$

- Under H_0, test statistic

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$$

~ t distribution with n-2 degree of freedom

- Reject H_0 if

$$|t_0| > t_{\alpha/2, n-2}$$

(two-sided test)

Example: oxygen purity tests of coefficients

- Consider the test

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

$$\hat{\beta}_1 = 14.947 \quad n = 20,$$
$$S_{xx} = 0.68088, \quad \hat{\sigma}^2 = 1.18$$

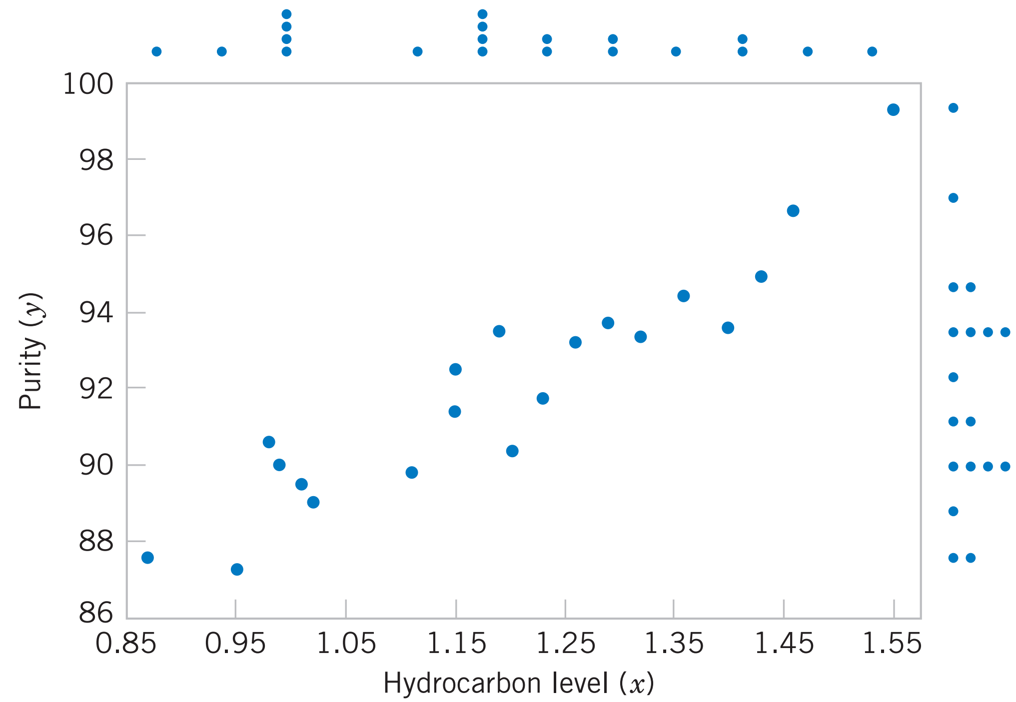


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

- Calculate the test statistic

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{14.947}{\sqrt{1.18/0.68088}} = 11.35$$

- Threshold $t_{\alpha/2,n-2}$ = $t_{0.005,18} = 2.88$
- Reject H0 since $|t_0| > t_{\alpha/2,n-2}$

Use t-test for intercept

- Use a similar form of test

$$H_0\colon \beta_0 = \beta_{0,0}$$
$$H_1\colon \beta_0 \neq \beta_{0,0}$$

- Test statistic $$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}} = \frac{\hat{\beta}_0 - \beta_{0,0}}{se(\hat{\beta}_0)}$$

  Under H_0, T_0 ~ t distribution with n-2 degree of freedom

- Reject H_0 if $|t_0| > t_{\alpha/2, n-2}$

Class activity

Given the regression line:

y = 22.2 + 10.5 x estimated for x = 1,2,3,…,20

1. The estimated slope is:

A. $\hat{\beta}_1 = 22.2$ B. $\hat{\beta}_1 = 10.5$ C. **biased**

2. The predicted value for x*=10 is

A. y*=22.2 B. y*=127.2 **C. y*=32.7**

3. The predicted value for x*=40 is

A. y*=442.2 B. y*=127.2 **C. Cannot extrapolate**

Class activity

1. The estimated slope is significantly different from zero when

A. $\left|\dfrac{\hat{\beta}_1\sqrt{S_{XX}}}{\hat{\sigma}}\right| > t_{\alpha/2,n-2}$ **B.** $\left|\dfrac{\hat{\beta}_1\sqrt{S_{XX}}}{\hat{\sigma}}\right| < t_{\alpha/2,n-2}$ **C.** $\left|\dfrac{\hat{\beta}_1\sqrt{S_{XX}}}{\hat{\sigma}}\right|^2 > F_{\alpha/2,n-1,1}$

2. The estimated intercept is plausibly zero when

A. Its confidence interval contains 0.

B. $\left|\dfrac{\hat{\beta}_0\sqrt{S_{XX}}}{\hat{\sigma}}\right| < t_{\alpha/2,n-2}$ **C.** $\left|\dfrac{\hat{\beta}_0}{\hat{\sigma}\sqrt{1/n+\bar{x}^2/S_{xx}}}\right| > t_{\alpha/2,n-2}$

Confidence interval

- we can obtain confidence interval estimates of slope and intercept
- width of confidence interval is a measure of the overall quality of the regression

true parameter

slope	intercept
$T_0 = \dfrac{\hat{\beta}_1 - \boxed{\beta_{1,0}}}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$	$T_0 = \dfrac{\hat{\beta}_0 - \boxed{\beta_{0,0}}}{\sqrt{\hat{\sigma}^2\left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right]}}$
~ t distribution with n-2 degree of freedom	~ t distribution with n-2 degree of freedom

Confidence intervals

a $100(1-\alpha)\%$ **confidence interval on the slope** β_1

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

a $100(1-\alpha)\%$ **confidence interval on the intercept** β_0

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right]}$$

Example: oxygen purity tests of coefficients

find a 95% confidence interval on the slope $(\alpha = 0.05)$

$$\hat{\beta}_1 = 14.947,\ S_{xx} = 0.68088,\ \text{and}\ \hat{\sigma}^2 = 1.18$$

$$\hat{\beta}_1 - t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$$14.947 - 2.101\sqrt{\frac{1.18}{0.68088}} \leq \beta_1 \leq 14.947 + 2.101\sqrt{\frac{1.18}{0.68088}}$$

$$12.181 \leq \beta_1 \leq 17.713$$

The confidence interval does not include 0, so enough evidence saying there is enough correlation between X and Y.

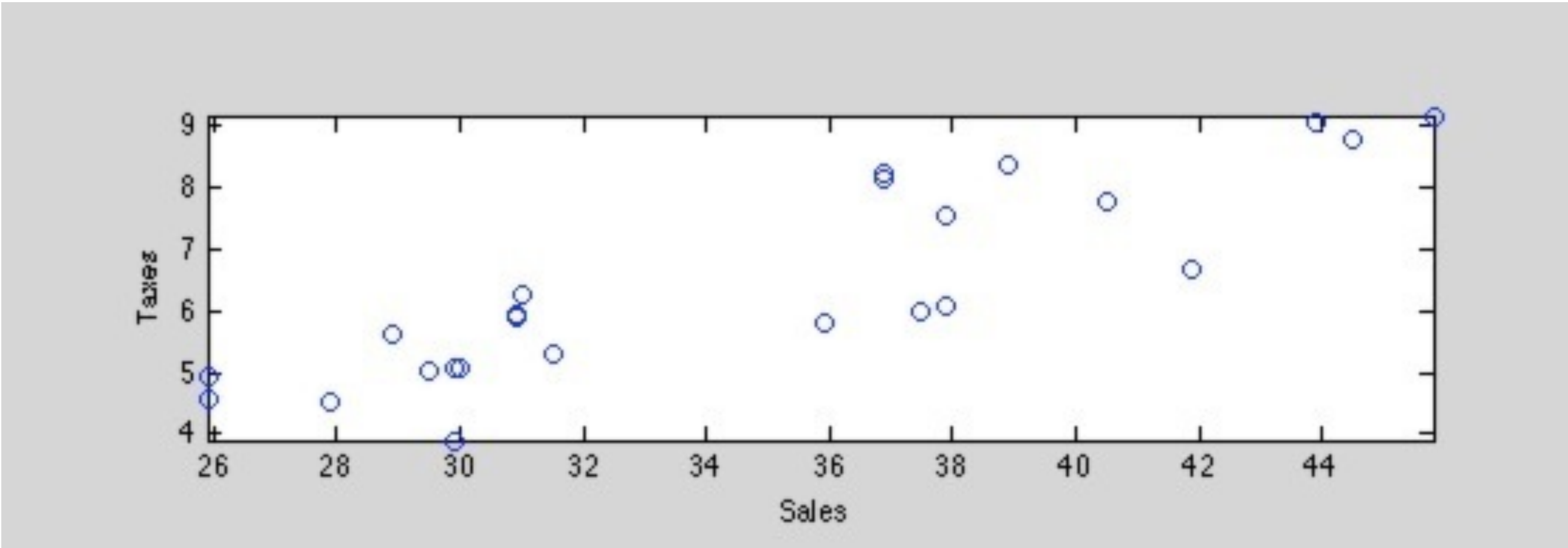
Example: house selling price and annual taxes

Sale Price/1000	Taxes (Local, School), County)/1000	Sale Price/1000	Taxes (Local, School), County)/1000
25.9	4.9176	30.0	5.0500
29.5	5.0208	36.9	8.2464
27.9	4.5429	41.9	6.6969
25.9	4.5573	40.5	7.7841
29.9	5.0597	43.9	9.0384
29.9	3.8910	37.5	5.9894
30.9	5.8980	37.9	7.5422
28.9	5.6039	44.5	8.7951
35.9	5.8282	37.9	6.0831
31.5	5.3003	38.9	8.3607
31.0	6.2712	36.9	8.1400
30.9	5.9592	45.8	9.1416

Independent variable X: SalePrice

Dependent variable Y: Taxes

- qualitative analysis



Calculate correlation

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}} = 0.8760$$

Independent variable Y: SalePrice

Dependent variable X: Taxes

$$n = 24 \qquad \bar{x} = 34.6125 \qquad \bar{y} = 6.4049$$

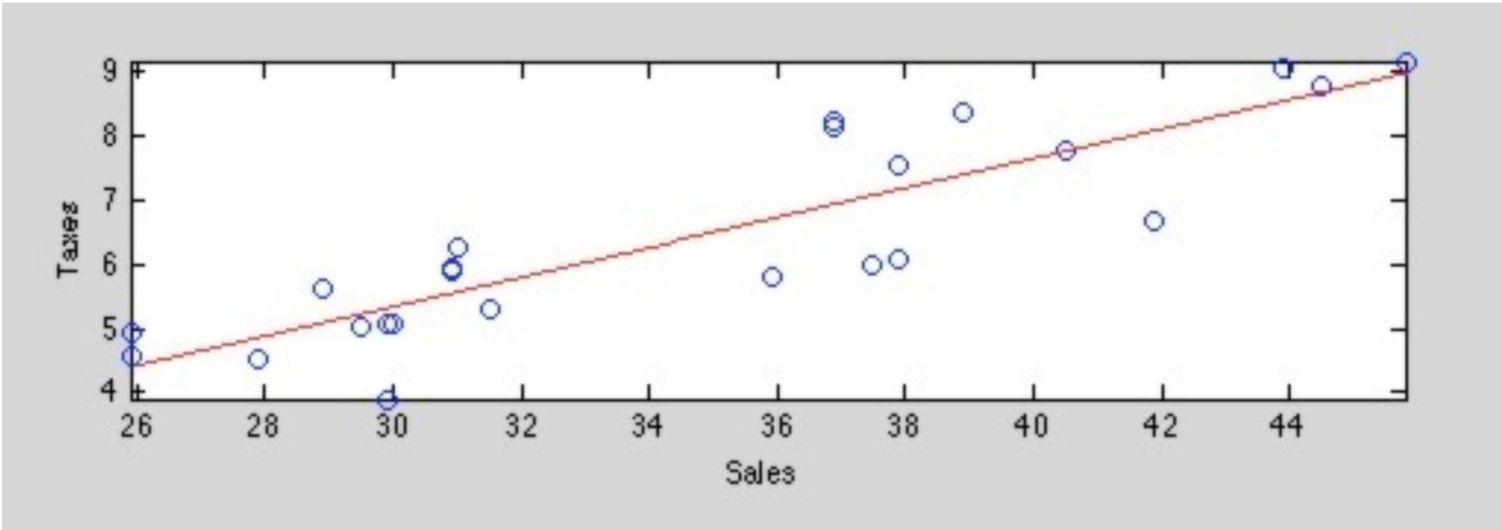
$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = 829.0462$$

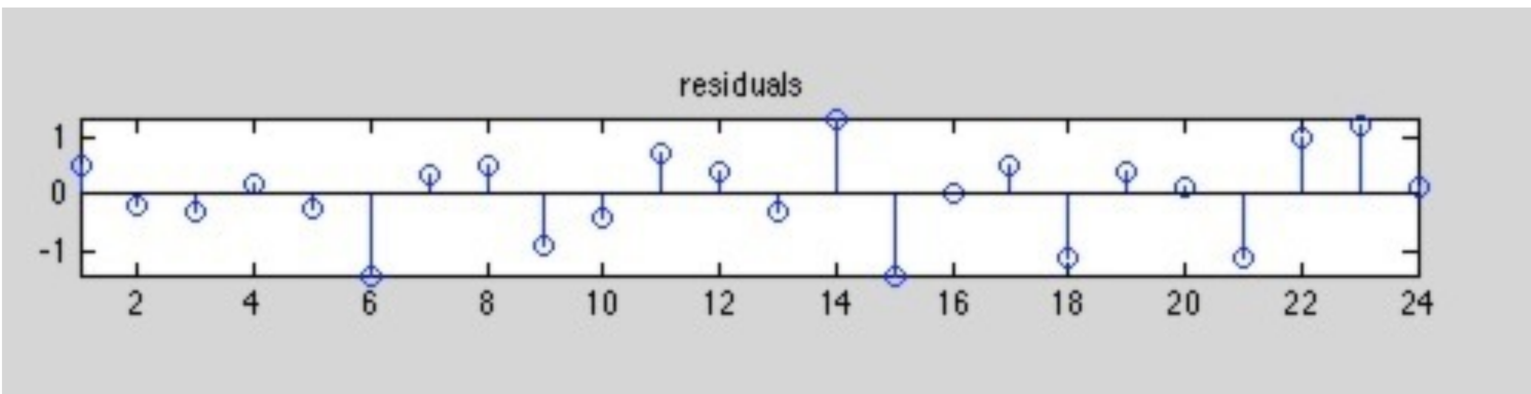
$$S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = 191.3612$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{191.3612}{829.0462} = 0.2308$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 6.4049 - 0.2308 \times 34.6125 = -1.5837$$

Fitted simple linear regression model $\hat{y} = -1.5837 + 0.2308x$





- residuals: $\hat{\sigma}^2 = MSE = \dfrac{\sum_{i=1}^{n} r_i^2}{n-2} = 0.6088$

- standard error of regression coefficients

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \sqrt{\frac{0.6088}{829.0462}} = 0.0271$$

$$se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} = \sqrt{0.6088\left[\frac{1}{24} + \frac{34.6125^2}{829.0462}\right]} = 0.9514$$

- test

Test H_0: $\beta_1 = 0$ using the t-test; use $\alpha = 0.05$

- calculate test statistics

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{0.2308}{0.0271} = 8.5166$$

- threshold

$$t_{\alpha/2, n-2} = t_{0.0025,22} = 3.119$$

- value of test statistic is greater than threshold
- → reject H_0

- construct confidence interval for slope parameter

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

$$t_{\alpha/2,n-2} = t_{0.0025,22} = 3.119$$

$$0.2308 - 3.119 \times 0.0271 \le \beta_1 \le 0.2308 + 3.119 \times 0.0271$$

$$0.14631 \le \beta_1 \le 0.3153$$