Lecture 12
Linear Regression: Test and Confidence Intervals
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

- 1

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

-1 \leq \hat{\rho} \leq 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:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, 2, \ldots, n$$

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

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

(11-10)

\[ S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left( \sum_{i=1}^{n} x_i \right)\left( \sum_{i=1}^{n} y_i \right)}{n} \]  

(11-11)

\[
\begin{align*}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\
\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i
\end{align*}
\]

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: \( Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \ i = 1, \ldots, n, \ \text{Var}(\varepsilon_i) = \sigma^2 \)
- 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 \( y = \beta_1 x + \beta_0 \), \( \hat{\beta}_1 \text{ and } \hat{\beta}_0 \) are point estimators for the true coefficients

• we can talk about the “accuracy” of \( \hat{\beta}_1 \text{ and } \hat{\beta}_0 \)
Assessing linear regression model

• Test hypothesis about true slope and intercept

\[ \beta_1 = ?, \quad \beta_0 = ? \]

• 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

\[ \varepsilon_i \sim N \left( 0, \sigma^2 \right) \]
## Properties of Regression Estimators

<table>
<thead>
<tr>
<th>slope parameter $\beta_1$</th>
<th>intercept parameter $\beta_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E(\hat{\beta}_1) = \beta_1$</td>
<td>$E(\hat{\beta}_0) = \beta_0$</td>
</tr>
<tr>
<td>$V(\hat{\beta}<em>1) = \frac{\sigma^2}{S</em>{xx}}$</td>
<td>$V(\hat{\beta}<em>0) = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S</em>{xx}} \right]$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>unbiased estimator</th>
<th>unbiased estimator</th>
</tr>
</thead>
</table>

\[
S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left( \sum_{i=1}^{n} x_i \right)^2}{n}
\]
Standard errors of coefficients

• We can replace $\sigma^2$ with its estimator $\hat{\sigma}^2$ ...

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

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

• Using results from previous page, estimate the

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \quad \text{and} \quad se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left[ \frac{1}{n} + \frac{x^2}{S_{xx}} \right]}$$
Hypothesis test in simple linear regression

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

\[ H_0: \beta_1 = \beta_{1,0} \]
\[ H_1: \beta_1 \neq \beta_{1,0} \]

• 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?
  
  \[ H_0: \beta_1 = 0 \]
  
  \[ H_1: \beta_1 \neq 0 \]

- 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?

Significance of regression
EXAMPLE 11-2 Oxygen Purity Tests of Coefficients

We will test for significance of regression using the model for the oxygen purity data from Example 11-1. The hypotheses are

\[ H_0: \beta = 1 \quad H_1: \beta \neq 1 \]

and we will use \( \alpha = 0.01 \). From Example 11-1 and Table 11-2 we have

\[ b = 14.947, \quad s_e = 1.18, \quad s_x^2 = 0.68088, \quad n = 20 \]

so the \( t \)-statistic in Equation 10-20 becomes

\[ t = \frac{14.947 - 1}{0.68088} = 11.35 \]

Hence, since the reference value of \( t \) is \( t_{0.005, 18} = 2.88 \), the value of the test statistic is very far into the critical region, implying that \( H_0: \beta = 1 \) should be rejected. There is strong evidence to support this claim.

The \( P \)-value for this test is \( P = 0.000 \). This was obtained manually with a calculator.

Table 11-2 presents the Minitab output for this problem. Notice that the \( t \)-statistic value for the slope is computed as 11.35 and that the reported \( P \)-value is \( P = 0.000 \). Minitab also reports the \( t \)-statistic for testing the hypothesis \( H_0: \beta = 0 \). This statistic is computed from Equation 11-22, with \( n = 20 \), \( s_x^2 = 0.68088 \), as

\[ t = \frac{14.947 - 0}{1.18} = 12.59 \]

Clearly, then, the hypothesis that the intercept is zero is rejected.

11-4.2 Analysis of Variance Approach to Test Significance of Regression

A method called the analysis of variance can be used to test for significance of regression. The procedure partitions the total variability in the response variable into meaningful components as the basis for the test. The analysis of variance identity is as follows:

\[
\sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2
\]
Use t-test for slope

Under $H_0$

slope parameter $\beta_1$

$$E(\hat{\beta}_1) = \beta_{1,0}$$

$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

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

- 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
  
  \[ H_0: \beta_1 = 0 \]
  
  \[ H_1: \beta_1 \neq 0 \]
  
  \[ \hat{\beta}_1 = 14.947 \quad n = 20, \]
  
  \[ S_{xx} = 0.68088, \quad \hat{\sigma}^2 = 1.18 \]

- 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 \( H_0 \) since \( |t_0| > t_{\alpha/2,n-2} \)

*Figure 11-1*  Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.
Use t-test for intercept

- Use a similar form of test

\[ H_0: \beta_0 = \beta_{0,0} \]
\[ H_1: \beta_0 \neq \beta_{0,0} \]

- 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 \( H_0 \), \( T_0 \sim t \) distribution with \( n-2 \) degree of freedom

- Reject \( H_0 \) if \( |t_0| > t_{\alpha/2,n-2} \)
Class activity

Given the regression line:
\[ y = 22.2 + 10.5 \times \text{x} \] estimated for \( x = 1, 2, 3, \ldots, 20 \)

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

   \[ \frac{\hat{\beta}_1 \sqrt{S_{XX}}}{\hat{\sigma}} > t_{\alpha/2, n-2} \]

   A. \hspace{1cm} B. \hspace{1cm} C. \hspace{1cm} \frac{\hat{\beta}_1 \sqrt{S_{XX}}}{\hat{\sigma}} < t_{\alpha/2, n-2} \hspace{1cm} \frac{\hat{\beta}_1 \sqrt{S_{XX}}}{\hat{\sigma}} > F_{\alpha/2, n-1,1}^2 \]

2. The estimated intercept is plausibly zero when

   A. Its confidence interval contains 0.

   \[ \frac{\hat{\beta}_0 \sqrt{S_{XX}}}{\hat{\sigma}} < t_{\alpha/2, n-2} \]

   B. \hspace{1cm} C. \hspace{1cm} \frac{\hat{\beta}_0}{\hat{\sigma} \sqrt{1/n + \bar{x}^2/S_{xx}}} > 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

<table>
<thead>
<tr>
<th>slope</th>
<th>intercept</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_0 = \frac{\hat{\beta}<em>1 - \beta</em>{1,0}}{\sqrt{\hat{\sigma}^2/S_{xx}}}$</td>
<td>$T_0 = \frac{\hat{\beta}<em>0 - \beta</em>{0,0}}{\sqrt{\hat{\sigma}^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]}}$</td>
</tr>
<tr>
<td>(\sim \text{t distribution with n-2 degree of freedom} )</td>
<td>(\sim \text{t distribution with n-2 degree of freedom} )</td>
</tr>
</tbody>
</table>
Confidence intervals

a 100(1 − \(\alpha\))% confidence interval on the slope \(\beta_1\)

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

a 100(1 − \(\alpha\))% confidence interval on the intercept \(\beta_0\)

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

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

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

\[ n = 24 \quad \bar{x} = 34.6125 \quad \bar{y} = 6.4049 \]

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

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

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

\[ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 6.4049 - 0.2308 \times 34.6125 = -1.5837 \]
Fitted simple linear regression model \( \hat{y} = -1.5837 + 0.2308x \)

- residuals: \( \hat{\sigma}^2 = MSE = \frac{\sum_{i=1}^{n} r_i^2}{n-2} = 0.6088 \)
• standard error of regression coefficients

\[ se(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2} \frac{1}{S_{xx}} = \sqrt{\frac{0.6088}{829.0462}} = 0.0271 \]

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

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

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

\[
0.2308 - 3.119 \times 0.0271 \leq \beta_1 \leq 0.2308 + 3.119 \times 0.0271
\]

\[
0.14631 \leq \beta_1 \leq 0.3153
\]