Regression Models

Math 463, Spring 2009, University of Oregon

David A. Levin

April 8, 2009

Recall the model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n\,.$$

The ε_i's are assumed to be uncorrelated with $\mathbb{E}(\varepsilon_i) = 0$ and $\operatorname{Var} \varepsilon_i = \sigma^2$.

Recall the model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n\,.$$

The ε_i's are assumed to be uncorrelated with $\mathbb{E}(\varepsilon_i) = 0$ and $\operatorname{Var} \varepsilon_i = \sigma^2$.
The *least-squares estimates* of β_0 and β_1 are

$$\hat{\beta}_1 = \sum_{i=1}^{n} (x_i - \bar{x}) Y_i$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_0 \bar{x}\,.$$

Recall the model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n\,.$$

The ε_i's are assumed to be uncorrelated with $\mathbb{E}(\varepsilon_i) = 0$ and $\operatorname{Var} \varepsilon_i = \sigma^2$.
The *least-squares estimates* of β_0 and β_1 are

$$\hat{\beta}_1 = \sum_{i=1}^{n} (x_i - \bar{x}) Y_i$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_0 \bar{x}\,.$$

These are the values of b_0 and b_1 which minimize

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2\,.$$

Recall the model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n\,.$$

The ε_i's are assumed to be uncorrelated with $\mathbb{E}(\varepsilon_i) = 0$ and $\operatorname{Var} \varepsilon_i = \sigma^2$.
The *least-squares estimates* of β_0 and β_1 are

$$\hat{\beta}_1 = \sum_{i=1}^{n} (x_i - \bar{x}) Y_i$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_0 \bar{x}\,.$$

These are the values of b_0 and b_1 which minimize

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2\,.$$

When the ε_i's are assumed to Normal, the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are the maximum likelihood estimators.

The random variable

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

(where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$) is an unbiased estimator of σ^2.

The random variable

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

(where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$) is an unbiased estimator of σ^2.
Also under the assumption of Normal errors, the statistics

$$T = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{1/n + (\bar{x})^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$T = \frac{\hat{\beta}_1 - \beta_1}{S / \sum_{i=1}^{n}(x_i - \bar{x})^2}$$

have t-distributions with $n-2$ degree of freedom.

The random variable

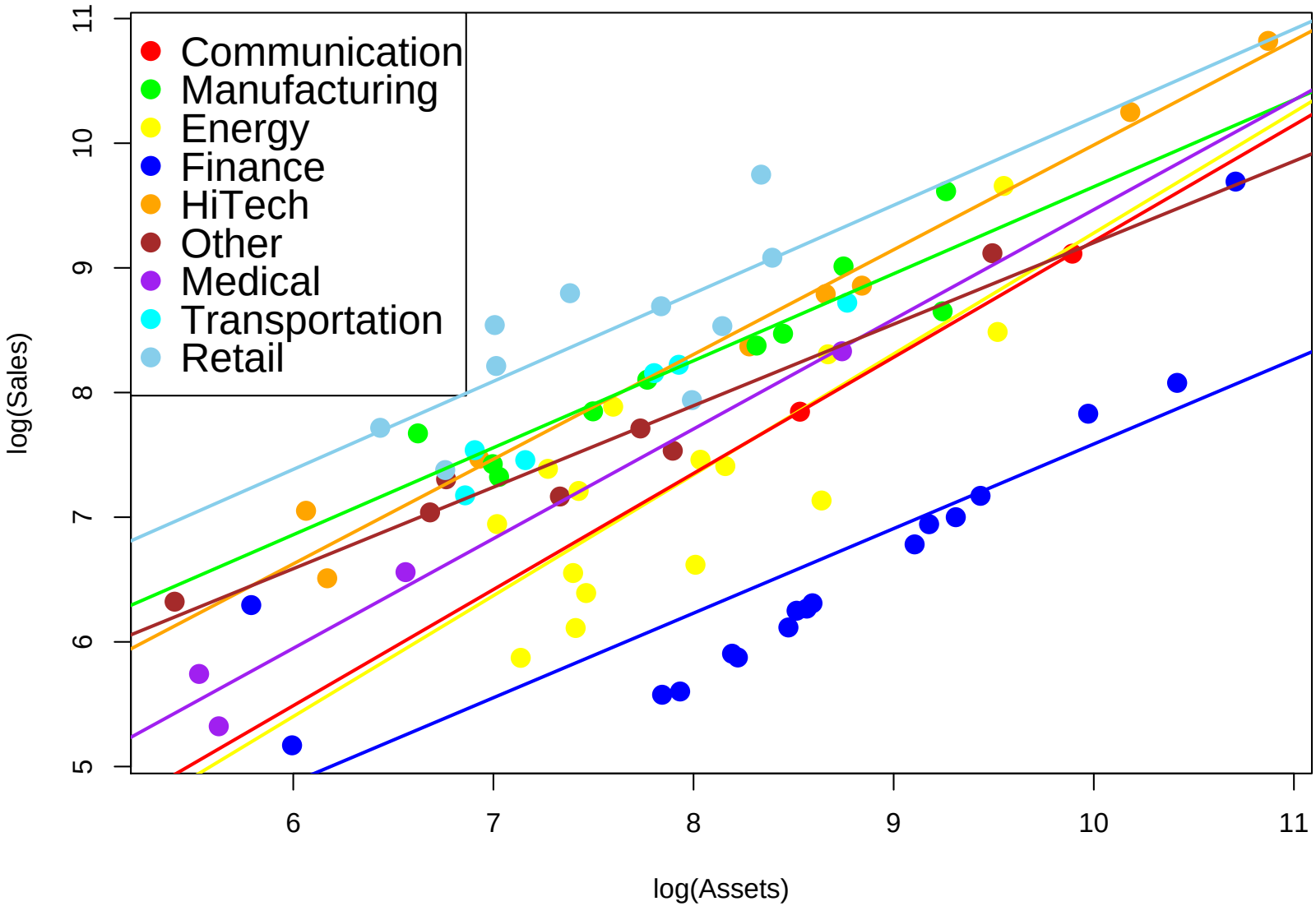
$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

(where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$) is an unbiased estimator of σ^2.
Also under the assumption of Normal errors, the statistics

$$T = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{1/n + (\bar{x})^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$T = \frac{\hat{\beta}_1 - \beta_1}{S / \sum_{i=1}^{n}(x_i - \bar{x})^2}$$

have t-distributions with $n-2$ degree of freedom.
Thus tests and confidence intervals for β_0 and β_1 can be based on these statistics and the t-distribution.


Communication
Manufacturing
Energy
Finance
HiTech
Other
Medical
Transportation
Retail
log(Sales)
log(Assets)
5
6
7
8
9
10
11

Define

$$\hat{Y}_i = \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, \ldots, n\,.$$

to be the *fitted values*. The quantity

$$\hat{\varepsilon}_i = \hat{Y}_i - Y_i$$

is called the *residual*.

Define

$$\hat{Y}_i = \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, \ldots, n\,.$$

to be the *fitted values*. The quantity

$$\hat{\varepsilon}_i = \hat{Y}_i - Y_i$$

is called the *residual*.
The *residual sum of squares* is defined as

$$\mathrm{SS}_{\mathrm{Res}} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2\,.$$

Define

$$\hat{Y}_i = \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, \ldots, n\,.$$

to be the *fitted values*. The quantity

$$\hat{\varepsilon}_i = \hat{Y}_i - Y_i$$

is called the *residual*.
The *residual sum of squares* is defined as

$$\mathrm{SS}_{\mathrm{Res}} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2\,.$$

The *regression sum of squares* is defined as

$$\mathrm{SS}_{\mathrm{Reg}} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2\,.$$

Define

$$\hat{Y}_i = \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, \ldots, n\,.$$

to be the *fitted values*. The quantity

$$\hat{\varepsilon}_i = \hat{Y}_i - Y_i$$

is called the *residual*.
The *residual sum of squares* is defined as

$$\mathrm{SS}_{\mathrm{Res}} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2\,.$$

The *regression sum of squares* is defined as

$$\mathrm{SS}_{\mathrm{Reg}} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2\,.$$

The *total sum of squares* is defined as

$$\mathrm{SS}_{\mathrm{Tot}} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2\,.$$

Fundamental sum-of-squares decomposition:

$$\mathrm{SS}_{\mathrm{Tot}} = \mathrm{SS}_{\mathrm{Reg}} + \mathrm{SS}_{\mathrm{Res}} \, .$$

Fundamental sum-of-squares decomposition:

$$\mathrm{SS}_{\mathrm{Tot}} = \mathrm{SS}_{\mathrm{Reg}} + \mathrm{SS}_{\mathrm{Res}} .$$

We now make the **additional assumption** that the ε_i's are Normal. The random variable $\mathrm{SS}_{\mathrm{Res}}/\sigma^2$ has a chi-squared distribution with $n - 2$ degrees of freedom. If the hypothesis

$$H_0 : \beta_1 = 0$$

is true, all of the sum-of-squares above divided by σ^2 have a chi-squared distribution. The degrees of freedom are summarized in the table below:

sum-of-squares	degrees of freedom
residual	$n - 2$
regression	1
total	$n - 1$

Under $H_0 : \beta_1 = 0$, the statistic

$$F = \frac{\mathrm{SS}_{\mathrm{Reg}}}{\mathrm{SS}_{\mathrm{Res}}/(n-2)} .$$

has a F-distribution with 1 and $n - 2$ degrees of freedom.

Under $H_0 : \beta_1 = 0$, the statistic

$$F = \frac{\mathrm{SS}_{\mathrm{Reg}}}{\mathrm{SS}_{\mathrm{Res}}/(n-2)} .$$

has a F-distribution with 1 and $n-2$ degrees of freedom.
Can base a test of $H_0 : \beta_1 = 0$ on F.

Under $H_0 : \beta_1 = 0$, the statistic

$$F = \frac{\mathrm{SS}_{\mathrm{Reg}}}{\mathrm{SS}_{\mathrm{Res}}/(n-2)} .$$

has a F-distribution with 1 and $n - 2$ degrees of freedom.
Can base a test of $H_0 : \beta_1 = 0$ on F.
This test will give the same result as the t-test.

$$R^2 = \frac{\mathrm{SS}_{\mathrm{Reg}}}{\mathrm{SS}_{\mathrm{Tot}}}$$

The value R_2 is in $[0, 1]$, and is interpreted as *the percent of variation in the response data that is explained by the model*.

- May want to estimate $\mathbb{E}(Y \mid x)$.

- May want to estimate $\mathbb{E}(Y \mid x)$.
- Recall $\mathbb{E}(Y \mid x) = \beta_0 + \beta_1 x$.

- May want to estimate $\mathbb{E}(Y \mid x)$.
- Recall $\mathbb{E}(Y \mid x) = \beta_0 + \beta_1 x$.
- Natural estimator:

$$\widehat{\mathbb{E}(Y \mid x)} = \hat{\beta}_0 + \hat{\beta}_1 x = \hat{Y} \, .$$

- May want to estimate $\mathbb{E}(Y \mid x)$.
- Recall $\mathbb{E}(Y \mid x) = \beta_0 + \beta_1 x$.
- Natural estimator:

$$\widehat{\mathbb{E}(Y \mid x)} = \hat{\beta}_0 + \hat{\beta}_1 x = \hat{Y}\,.$$

- For a confidence interval, we need to know distribution of $\hat{Y}$. It is a linear combination of independent Normal variables, so it has a Normal distribution. Thus confidence intervals can be worked out using the t distribution.