ISE 315: Engineering Statistics

Lecture 18: Simple Linear Regression (Part 2)

Instructor: Mansur M. Arief, PhD

Industrial and Systems Engineering, KFUPM

Office: 22-219 — Email: mansur.arief@kfupm.edu.sa

Based on Montgomery & Runger, Applied Statistics and Probability for Engineers, 6th Ed.

Lecture 18

Simple Linear Regression — Part 2 (Chapter 11, cont'd)

Lecture 18 Outline

- Quick review: fitting the line (from Lecture 15)
- Hypothesis test on the slope β_1 (Sec. 11-4.1)
- ANOVA decomposition for regression (Sec. 11-4.2)
- Coefficient of determination R^2 (Sec. 11-7.2)
- Worked example: production speed & defect rate

Quick Review

What We Did in Lecture 15

Review: Fitting the Regression Line

Model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $\varepsilon_i \sim (0, \sigma^2)$, uncorrelated.

Least squares estimators:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Building blocks:

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}, \quad S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}, \quad S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

Fitted values and residuals:

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

Review: Error Variance $\hat{\sigma}^2$

The **error sum of squares** measures how much scatter remains after fitting the line:

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum (y_i - \hat{y}_i)^2 = S_{yy} - \hat{\beta}_1 \cdot S_{xy}$$

Unbiased estimator of σ^2:

$$\boxed{\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-2}}$$

Why $n-2$? We estimated two parameters ($\hat{\beta}_0$ and $\hat{\beta}_1$), so we lose 2 degrees of freedom. This is the same logic as dividing by $n-1$ when we estimated one parameter ($\bar{x}$).

The Natural Next Question

So far we have a fitted line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

But **fitting a line is always possible** — even if there is no real relationship between x and y!

How do we know if the relationship is **real** or just **noise**?

This is **exactly** the hypothesis testing question you already know from Chapters 9–10.

Connection between Chapter 9 and Chapter 11

Step	**Ch. 9 (One Sample)**	**Ch. 11 (Regression)**
Parameter	μ	β_1
Hypothesized value	μ_0	0 (no relationship)
Estimator	$\bar{X}$	$\hat{\beta}_1$
Standard error	$\sigma/\sqrt{n}$ or $s/\sqrt{n}$	$\sqrt{MS_E/S_{xx}}$
Test statistic	$T_0 = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$	$T_0 = \frac{\hat{\beta}_1}{\sqrt{MS_E/S_{xx}}}$
Distribution	t_{n-1}	t_{n-2}

The structure is **identical**. The only things that change are the parameter, the formula for the standard error, and the degrees of freedom.

Hypothesis Test on the Slope β_1

Section 11-4.1

Question: Is there a statistically significant linear relationship between x and y?

Hypotheses:

$$H_0 : \beta_1 = 0 \quad \text{(no linear relationship)} \qquad H_1 : \beta_1 \neq 0$$

Test statistic:

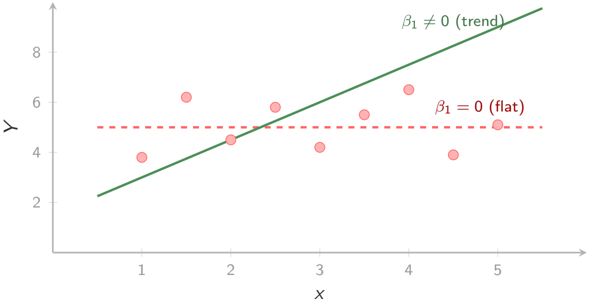
$$\boxed{T_0 = \frac{\hat{\beta}_1}{\sqrt{MS_E / S_{xx}}}}$$

Under H_0, $T_0 \sim t_{n-2}$.

Decision rule (at significance level α):

- Reject H_0 if $|T_0| > t_{\alpha/2,\, n-2}$

What Does "$\beta_1 = 0$" Mean Visually?



If $\beta_1 = 0$, knowing x tells you **nothing** about y. The regression line is flat, and x has no predictive value. Our test asks: is the observed slope far enough from zero to be convincing?

ANOVA for Regression

Decomposing the Total Variability

ANOVA Decomposition

Total = Regression + Error

The total variability in y can be split into two pieces:

$$\underbrace{\sum (y_i - \bar{y})^2}_{SS_T} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{SS_R} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{SS_E}$$

$SS_T = S_{yy}$	Total variability in y
$SS_R = \hat{\beta}_1 \cdot S_{xy}$	Variability **explained** by the regression
$SS_E = S_{yy} - \hat{\beta}_1 \cdot S_{xy}$	Variability **not explained** (residual error)

Key insight: If the model is useful, SS_R should be **large** relative to SS_E. If there is no relationship, most of the variability stays in SS_E.

The ANOVA Table for Simple Linear Regression

Source	**df**	**SS**	**MS**	F_0
Regression	1	SS_R	$MS_R = SS_R/1$	$F_0 = MS_R/MS_E$
Error	$n-2$	SS_E	$MS_E = SS_E/(n-2)$	
Total	$n-1$	SS_T		

Decision rule: Reject H_0: $\beta_1 = 0$ if $F_0 > f_{\alpha,1,n-2}$.

Important fact: For simple linear regression, $F_0 = T_0^2$. The F-test and the t-test give **exactly the same conclusion**. They are two ways to test the same hypothesis.

Coefficient of Determination R^2

Section 11-7.2

Definition:

$$\boxed{R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}}$$

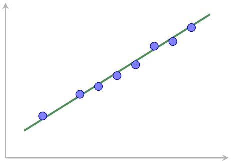
R^2 is the **proportion of total variability** in y that is explained by the linear relationship with x.

- $R^2 = 1$: Perfect fit — all points lie exactly on the line
- $R^2 = 0$: The model explains nothing — x is useless as a predictor
- $0 < R^2 < 1$: Some variability is explained, some is not

For simple linear regression, $R^2 = r^2$, where $r = S_{xy}/\sqrt{S_{xx} \cdot S_{yy}}$ is the **sample correlation coefficient**. The sign of r matches the sign of $\hat{\beta}_1$.

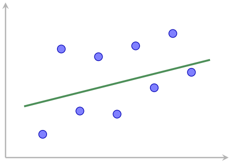
Visualizing R^2: What Does It Look Like?

High R^2 **(**≈ 0.95**)**



Points cluster tightly around the line. The model captures most of the variability.

Low R^2 **(**≈ 0.25**)**



Points are widely scattered. The line captures only a fraction of the variability.

Worked Example

Production Speed and Defect Rate (Continued from Lecture 15)

Example: Recall the Setup

A quality engineer studies how **production speed** (x, units/hr) affects the **defect rate** (y, defects per 100 units).

Observation	**1**	**2**	**3**	**4**	**5**
y (defects/100)	2	5	4	7	7
x (units/hr)	2	4	6	8	10

From Lecture 15, we already computed:

$$S_{xx} = 40, \quad S_{xy} = 24, \quad \bar{x} = 6, \quad \bar{y} = 5$$

$$\hat{\beta}_1 = 0.60, \quad \hat{\beta}_0 = 1.40, \quad \hat{y} = 1.40 + 0.60\,x$$

Example: ANOVA Decomposition

Total sum of squares:

$$SS_T = S_{yy} = 18.0$$

Regression sum of squares:

$$SS_R = \hat{\beta}_1 \cdot S_{xy} = 0.60 \times 24 = 14.4$$

Error sum of squares:

$$SS_E = SS_T - SS_R = 18.0 - 14.4 = 3.6$$

Mean squares: $MS_R = \frac{SS_R}{1} = 14.4, \qquad MS_E = \frac{SS_E}{n-2} = \frac{3.6}{3} = 1.2$

Example: Completed ANOVA Table

Source	**df**	**SS**	**MS**	F_0
Regression	1	14.4	14.4	12.0
Error	3	3.6	1.2	
Total	4	18.0		

Test at $\alpha = 0.05$: From the F-table, $f_{0.05,1,3} = 10.128$.
Since $F_0 = 12.0 > 10.128$, we **reject** H_0.

Conclusion: There is sufficient evidence at $\alpha = 0.05$ that production speed has a statistically significant linear effect on defect rate.

Example: t-Test (Equivalent Approach)

Standard error of $\hat{\beta}_1$:

$$\mathrm{se}(\hat{\beta}_1) = \sqrt{\frac{MS_E}{S_{xx}}} = \sqrt{\frac{1.2}{40}} = \sqrt{0.03} = 0.1732$$

Test statistic:

$$T_0 = \frac{\hat{\beta}_1}{\mathrm{se}(\hat{\beta}_1)} = \frac{0.60}{0.1732} = 3.464$$

Verification: $T_0^2 = (3.464)^2 = 12.0 = F_0$. ✓

Critical value: $t_{0.025,\,3} = 3.182$. Since $|T_0| = 3.464 > 3.182$, we reject H_0 — same conclusion as the F-test.

P-value: Since $t_{0.025,3} = 3.182 < 3.464 < t_{0.01,3} = 4.541$, we have $0.02 <$ P-value < 0.05.

Example: R^2 and Correlation

Coefficient of determination:

$$R^2 = \frac{SS_R}{SS_T} = \frac{14.4}{18.0} = 0.80$$

Interpretation: 80% of the variability in defect rate is explained by the linear relationship with production speed. The remaining 20% is unexplained.

Sample correlation:

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = \frac{24}{\sqrt{40 \times 18}} = \frac{24}{\sqrt{720}} = \frac{24}{26.83} = 0.894$$

Check: $r^2 = (0.894)^2 = 0.80 = R^2$. ✓ And $r > 0$ because $\hat{\beta}_1 > 0$.

Lecture 18 Summary

- The t-test on β_1 asks: is the linear relationship **real**?
  Test statistic: $T_0 = \hat{\beta}_1/\text{se}(\hat{\beta}_1)$ with $n-2$ df.
- The F-test via ANOVA gives the **same answer**: $F_0 = MS_R/MS_E = T_0^2$.
- $SS_T = SS_R + SS_E$. Compute via $SS_R = \hat{\beta}_1 S_{xy}$, $SS_E = S_{yy} - \hat{\beta}_1 S_{xy}$.
- $R^2 = SS_R/SS_T$ measures **how much variability** the model explains.
- The sample correlation $r = S_{xy}/\sqrt{S_{xx} \cdot S_{yy}}$ satisfies $r^2 = R^2$.
- **Next lecture:** Confidence and prediction intervals, residual analysis, and exam strategy.

Reminders

- Office hours: Tuesdays 9–10 AM, Room 22-219 or Zoom
- Continue reading **Chapter 11**, Sections 11-4 and 11-7.2
- Review the ANOVA table structure — it will appear on the exam
- **Next lecture** will complete Chapter 11 with intervals, residual analysis, and a full formula roadmap