Inference in Simple Linear Regression

Keegan Korthauer

Department of Statistics

UW Madison

Recap – Correlation Coefficient

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}}$$

- Measures the strength of **linear** relationship
- Unitless, always between -1 and 1
- Correlation does not imply causation
- If (X,Y) bivariate normal, have CI and HT for population correlation coefficient ρ

Recap –Simple Linear Regression

- The simple linear regression model assumes:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- The least-squares line is:

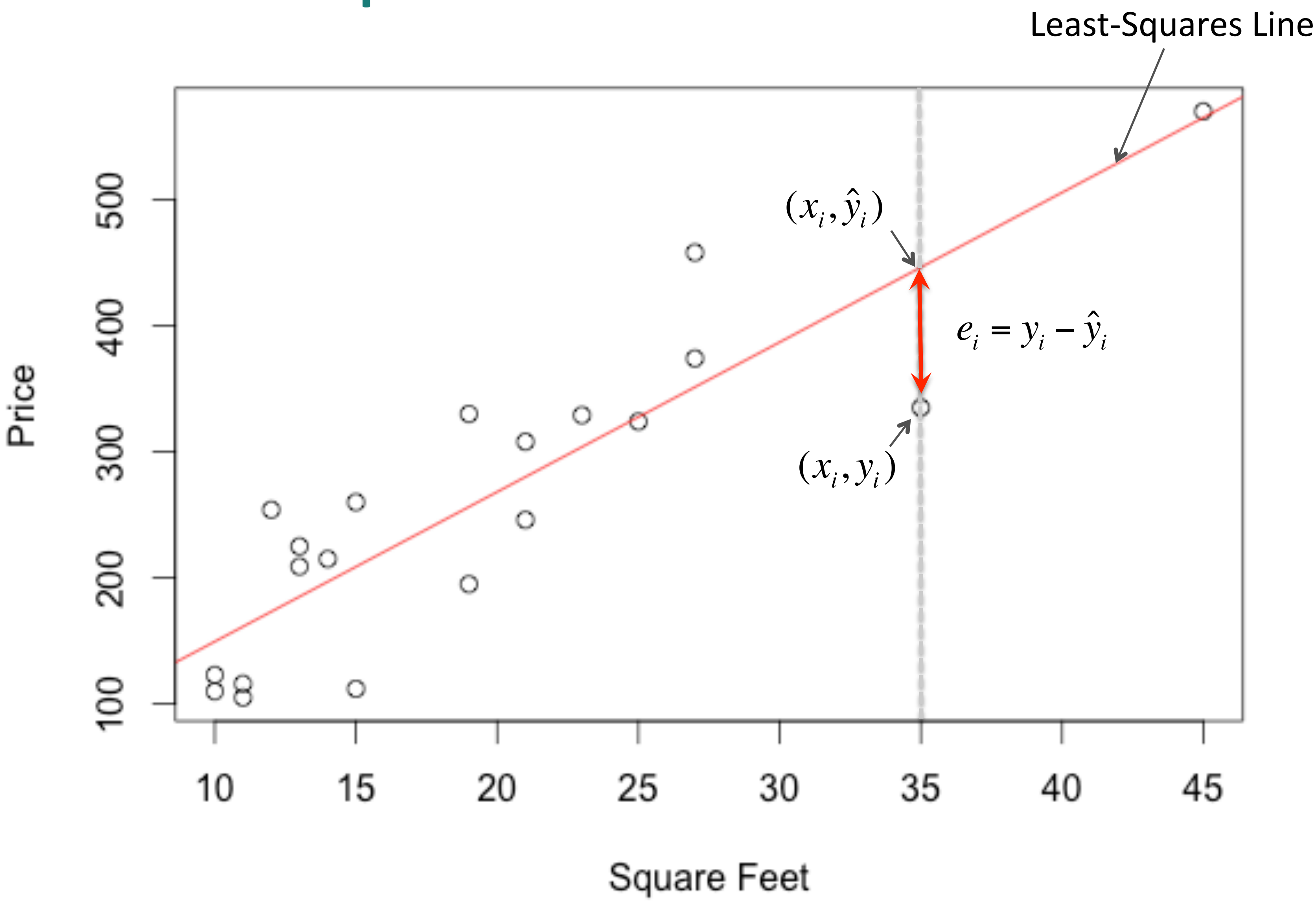
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Where

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}, \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Only applies when relationship is **linear**
- Be wary of extrapolation

Least-Squares Line Minimizes SSE



Sums of Squares

- Error Sum of Squares $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} \hat{y}_i^2$
- Total Sum of Squares $SST = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2$
- Regression Sum of Squares $SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$

Analysis of Variance property: $SST = SSR + SSE$

Coefficient of Determination (Goodness-of-fit measure):

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{SSR}{SST}$$

Example - Finding Sums of Squares

For the housing data example, find SSE, SSR and SST using the following quantities:

$$\bar{x} = 19.3,\ \bar{y} = 259.9,\ n = 20$$

$$\sum_{i=1}^{n} x_i y_i = 119{,}156$$

$$\sum_{i=1}^{n} x_i^2 = 9036,\ \sum_{i=1}^{n} y_i^2 = 1{,}639{,}188,\ \sum_{i=1}^{n} \hat{y}_i^2 = 1{,}574{,}603$$

UNCERTAINTIES AND INFERENCE FOR THE LEAST-SQUARES COEFFICIENTS

ε_i - Error Term in the Simple Linear Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Measurement errors and/or uncontrolled variation in experimental conditions
- Unknown
- Expect to be zero on average

Repeated Experimental Processes

- The errors ε_i will change from experiment to experiment, and so will the estimates of β_0 and β_1
- The errors ε_i create **uncertainty** in the estimates of β_0 and β_1
- Smaller errors are associated with smaller amount of uncertainty in the estimates
  - Likewise larger errors lead to larger uncertainty in the estimates

Assumptions for Errors in Linear Models

1. Errors $\varepsilon_1,\ldots, \varepsilon_n$ are **random** and **independent**. In particular, the magnitude of any error ε_i does not influence the value of the next error ε_{i+1}

2. Errors $\varepsilon_1,\ldots, \varepsilon_n$ all have **mean 0**

3. Errors $\varepsilon_1,\ldots, \varepsilon_n$ all have the **same variance** denoted by σ^2

4. Errors $\varepsilon_1,\ldots, \varepsilon_n$ are **normally distributed**

Violations of Error Assumptions

- If the sample size is large, Assumption 4 (**normality**) is not very important

- Mild violations of Assumption 3 (**constant variance**) are OK, but severe violations are not
  - More on this (how to diagnose, correct) later

Estimation of Error Variance σ^2

- Assumption 3: all errors have variance σ^2

- To estimate uncertainty in estimates of β_0 and β_1, must first estimate σ^2 with s^2:

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$$

- Equivalent formulae:

$$s^2 = \frac{SSE}{n-2} = \frac{SST(1-r^2)}{n-2}$$

Consequences of the Assumptions

- The errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent normal random variables with mean zero and variance σ^2:

$$e_i \sim N(0,\ \sigma^2)$$

- Since $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ the y_i are a linear combination of ε_i so they are also normally distributed:

$$y_i \sim N(\beta_0 + \beta_1 x_i,\ \sigma^2)$$

- We can now calculate the means and standard deviations of the estimates of β_0 and β_1

Uncertainties of Coefficients

Under assumptions 1-4, $\hat{\beta}_0$ and $\hat{\beta}_1$

- are normally distributed
- have mean β_0 and β_1, respectively
- have standard deviations:

$$s_{\hat{\beta}_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \text{ and } s_{\hat{\beta}_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where $s = \hat{\sigma}$

Housing Data Example

Estimate the standard deviations of the regression coefficients using the following quantities:

$$\bar{x} = 19.3,\ \bar{y} = 259.9,\ n = 20$$

$$\sum_{i=1}^{n} x_i y_i = 119{,}156$$

$$\sum_{i=1}^{n} x_i^2 = 9036,\ \sum_{i=1}^{n} y_i^2 = 1{,}639{,}188$$

Ways to Improve Accuracy

- Improving accuracy = decreasing variation
- Increase sample size
- Increase range of x values

Inference on the Coefficients

- Now that we have their mean and standard deviations, we can get CIs/HTs about the true values β_0 and β_1 using the t distribution:

$$\frac{(\hat{\beta}_0 - \beta_0)}{s_{\hat{\beta}_0}} \sim t_{n-2} \text{ and } \frac{(\hat{\beta}_1 - \beta_1)}{s_{\hat{\beta}_1}} \sim t_{n-2}$$

- We can test a hypothesis for β_0 or β_1 using a t-test where the quantities above are the test statistics

Housing Data Example

Someone claims that for every additional 100 square feet, a home will sell for about $10,000 more. To evaluate this claim on our dataset, perform a hypothesis test at the 0.05 level of

$$H_0: \beta_1 = 10 \text{ versus } H_1: \beta_1 \neq 10$$

Confidence Intervals for β_0 and β_1

From the previous results, we can obtain a 100(1-α)% CI for β_0 or β_1 with the following:

$$\hat{\beta}_0 \pm t_{n-2,\ \alpha/2} s_{\hat{\beta}_0}$$

$$\hat{\beta}_1 \pm t_{n-2,\ \alpha/2} s_{\hat{\beta}_1}$$

Housing Data Example

Find 95% confidence intervals for the regression coefficients β_0 and β_1 :

Confidence Interval of Mean Response

What if we want an interval of plausible values for the **mean value of y at a certain value of x**?

A level $100(1-\alpha)\%$ confidence interval for the quantity $\beta_0 + \beta_1 x$ is given by

$$\widehat{\beta}_0 + \widehat{\beta}_1 x \pm t_{n-2,\alpha/2} \cdot s_{\hat{y}} \tag{7.41}$$

where $s_{\hat{y}} = s\sqrt{\dfrac{1}{n} + \dfrac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}$.

Prediction Interval for Future Observations

What if we want an interval of plausible values for **y for a** particular observation **with a certain x value**?

A level $100(1-\alpha)\%$ prediction interval for the quantity $\beta_0 + \beta_1 x$ is given by

$$\widehat{\beta}_0 + \widehat{\beta}_1 x \pm t_{n-2,\alpha/2} \cdot s_{\text{pred}} \tag{7.44}$$

where $s_{\text{pred}} = s\sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$.

Housing Data Example

Find:

- A 95% confidence interval for the mean cost of a home with 2500 square feet

- A 95% prediction interval for a 2500 square foot home that will be put on the market next week

INTERPRETING R OUTPUT

R: Summary of an SLR fit

test statistic and p-value for the test of the null hypothesis that the coefficient is equal to 0

```
> summary(lm(Price~Sqft))
Coefficients:
```

	Estimate		Std. Error		t value	Pr(>|t|)	
(Intercept)	30.731	$\hat{\beta}_0$	31.969	$s_{\hat{\beta}_0}$	0.961	0.349	
Sqft	11.874	$\hat{\beta}_1$	1.504	$s_{\hat{\beta}_1}$	7.895	2.96e-07	***

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

s $\quad$ $n-2$

```
Residual standard error: 59.9 on 18 degrees of freedom
Multiple R-squared: 0.7759,   Adjusted R-squared: 0.7635
F-statistic: 62.33 on 1 and 18 DF,  p-value: 2.957e-07
```

r^2

Next

- Exam 2 handed back Wednesday

- How to Check Assumptions 1-4

- What to do when assumptions are violated
  - Transformation of Data
  - Addressing Outliers and Influential Points