

## **Section 2.2: Covariance, Correlation, and Least Squares**

Jared S. Murray

The University of Texas at Austin

McCombs School of Business

Suggested reading: OpenIntro Statistics, Chapter 7.1, 7.2

## A Deeper Look at Least Squares Estimates

Last time we saw that least squares estimates had some special properties:

- ▶ The fitted values  $\hat{Y}$  and  $x$  were **very** dependent
- ▶ The residuals  $Y - \hat{Y}$  and  $x$  had no apparent relationship
- ▶ The residuals  $Y - \hat{Y}$  had a sample mean of zero

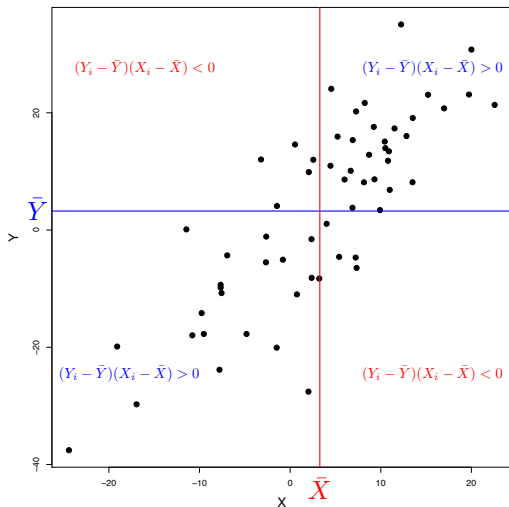
What's going on? And what exactly are the least squares estimates?

We need to review **sample covariance** and **correlation**

# Covariance

Measure the *direction* and *strength* of the linear relationship between  $Y$  and  $X$

$$\text{Cov}(Y, X) = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{n - 1}$$



- ▶  $s_y = 15.98,$   
 $s_x = 9.7$
- ▶  $\text{Cov}(X, Y) = 125.9$

How do we interpret that?

# Correlation

Correlation is the standardized covariance:

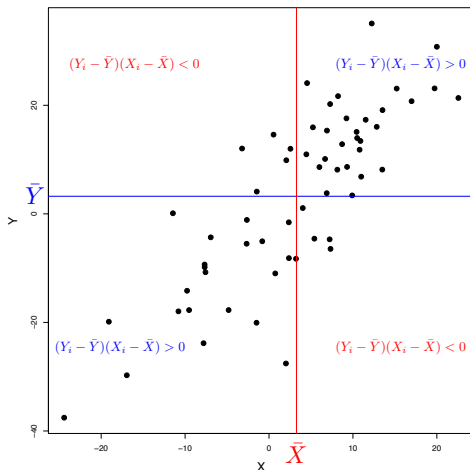
$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{s_x^2 s_y^2}} = \frac{\text{cov}(X, Y)}{s_x s_y}$$

The correlation is scale invariant and the units of measurement don't matter: **It is always true that  $-1 \leq \text{corr}(X, Y) \leq 1$ .**

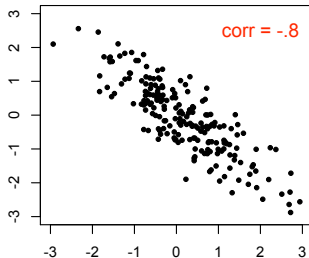
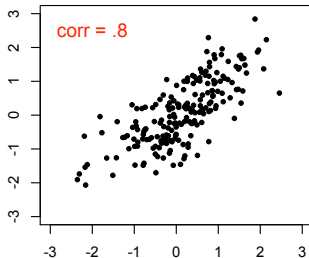
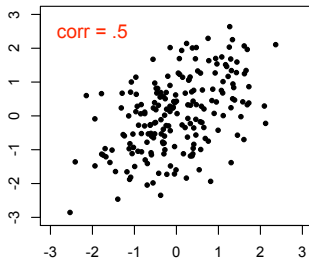
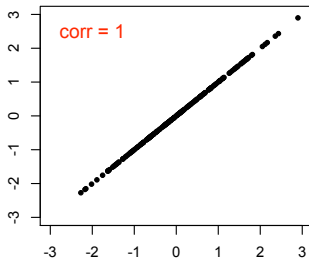
This gives the direction (- or +) and strength ( $0 \rightarrow 1$ ) of the **linear** relationship between  $X$  and  $Y$ .

# Correlation

$$\text{corr}(Y, X) = \frac{\text{cov}(X, Y)}{\sqrt{s_x^2 s_y^2}} = \frac{\text{cov}(X, Y)}{s_x s_y} = \frac{125.9}{15.98 \times 9.7} = 0.812$$



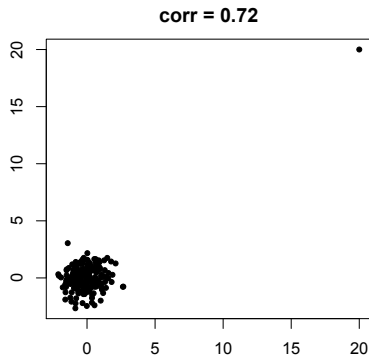
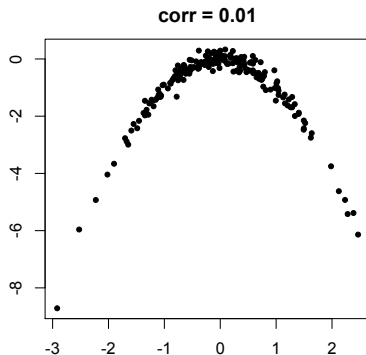
# Correlation



# Correlation

Only measures **linear** relationships:

$\text{corr}(X, Y) = 0$  does not mean the variables are not related!



Also be careful with influential observations...

# The Least Squares Estimates

The values for  $b_0$  and  $b_1$  that minimize the least squares criterion are:

$$b_1 = r_{xy} \times \frac{s_y}{s_x} \quad b_0 = \bar{Y} - b_1 \bar{X}$$

where,

- ▶  $\bar{X}$  and  $\bar{Y}$  are the sample mean of  $X$  and  $Y$
- ▶  $\text{corr}(x, y) = r_{xy}$  is the sample correlation
- ▶  $s_x$  and  $s_y$  are the sample standard deviation of  $X$  and  $Y$

These are the **least squares estimates** of  $\beta_0$  and  $\beta_1$ ; they minimize the sum of squared residuals.



# The Least Squares Estimates

The values for  $b_0$  and  $b_1$  that minimize the least squares criterion are:

$$b_1 = r_{xy} \times \frac{s_y}{s_x} \quad b_0 = \bar{Y} - b_1 \bar{X}$$

How do we interpret these?

- ▶  $b_0$  ensures the line goes through  $(\bar{x}, \bar{y})$
- ▶  $b_1$  scales the correlation to appropriate units by multiplying with  $s_y/s_x$  (what are the units of  $b_1$ ?)

These are the **least squares estimates** of  $\beta_0$  and  $\beta_1$ ; they minimize the sum of squared residuals.

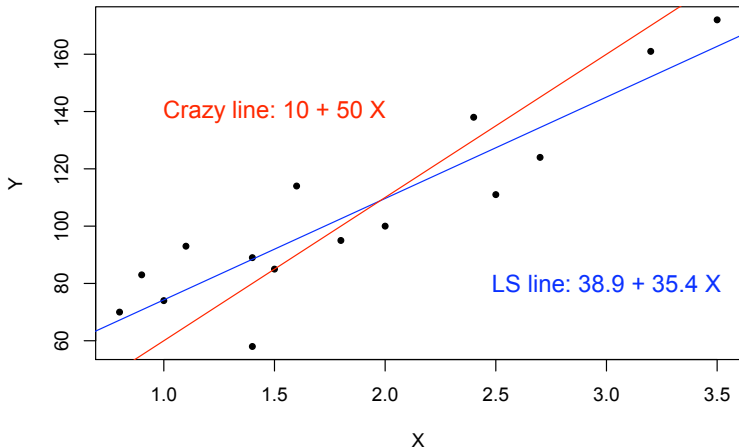
# Properties of Least Squares Estimates

Remember from the housing data, we had:

- ▶  $\text{corr}(\hat{Y}, x) = 1$  (a perfect linear relationship)
- ▶  $\text{corr}(e, x) = 0$  (no linear relationship)
- ▶  $\text{mean}(e) = 0$  (sample average of residuals is zero)

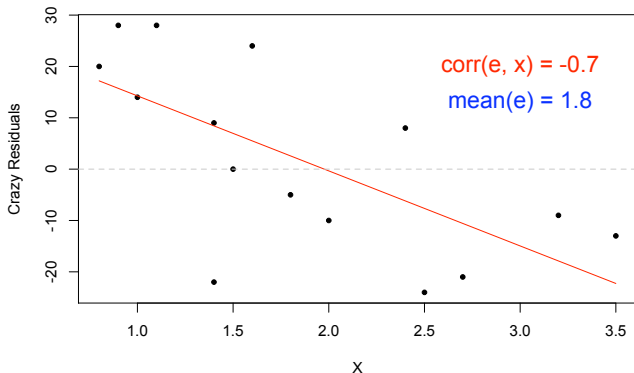
## Why?

What is the intuition for the relationship between  $\hat{Y}$  and  $e$  and  $X$ ? Lets consider some “crazy” alternative line:



## Fitted Values and Residuals

This is a bad fit! We are underestimating the value of small houses and overestimating the value of big houses.



Clearly, we have left some predictive ability on the table!

## Summary: LS is the best we can do!!

As long as the correlation between  $e$  and  $X$  is non-zero, we could always adjust our prediction rule to do better.

We need to exploit all of the predictive power in the  $X$  values and put this into  $\hat{Y}$ , leaving no “ $X$ ness” in the residuals.

In Summary:  $Y = \hat{Y} + e$  where:

- ▶  $\hat{Y}$  is “made from  $X$ ”;  $\text{corr}(X, \hat{Y}) = 1$ .
- ▶  $e$  is unrelated to  $X$ ;  $\text{corr}(X, e) = 0$ .
- ▶ On average, our prediction error is zero:  $\bar{e} = \sum_{i=1}^n e_i = 0$ .

## Decomposing the Variance

How well does the least squares line explain variation in  $Y$ ?

Remember that  $Y = \hat{Y} + e$

Since  $\hat{Y}$  and  $e$  are uncorrelated, i.e.  $\text{corr}(\hat{Y}, e) = 0$ ,

$$\begin{aligned}\text{var}(Y) &= \text{var}(\hat{Y} + e) = \text{var}(\hat{Y}) + \text{var}(e) \\ \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1} &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})^2}{n-1} + \frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n-1}\end{aligned}$$

Given that  $\bar{e} = 0$ , and the sample mean of the fitted values  $\bar{\hat{Y}} = \bar{Y}$  (why?) we get to write:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n e_i^2$$

## Decomposing the Variance

$$\underbrace{\sum_{i=1}^n (Y_i - \bar{Y})^2}_{\substack{\text{Total Sum of} \\ \text{Squares} \\ \text{SST}}} = \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}_{\substack{\text{Regression SS} \\ \text{SSR}}} + \underbrace{\sum_{i=1}^n e_i^2}_{\substack{\text{Error SS} \\ \text{SSE}}}$$

SSR: Variation in  $Y$  explained by the regression line.

SSE: Variation in  $Y$  that is left unexplained.

$\text{SSR} = \text{SST} \Rightarrow$  perfect fit.

*Be careful of similar acronyms; e.g. SSR for “residual” SS.*

## The Coefficient of Determination $R^2$

The **coefficient of determination**, denoted by  $R^2$ , measures how well the fitted values  $\hat{Y}$  follow  $Y$ :

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- ▶  $R^2$  is the proportion of variance in  $Y$  that is “explained” by the regression line (in the mathematical – not scientific – sense!):  $R^2 = 1 - \text{Var}(e)/\text{Var}(Y)$
- ▶  $0 < R^2 < 1$
- ▶ For simple linear regression,  $R^2 = r_{xy}^2$ . Similar caveats to sample correlation apply!



## $R^2$ for the Housing Data

```
summary(fit)

##
## Call:
## lm(formula = Price ~ Size, data = housing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.425  -8.618   0.575  10.766  18.498
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   38.885      9.094   4.276 0.000903 ***
## Size          35.386      4.494   7.874 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.14 on 13 degrees of freedom
## Multiple R-squared:  0.8267, Adjusted R-squared:  0.8133
## F-statistic:    62 on 1 and 13 DF,  p-value: 2.66e-06
```

## $R^2$ for the Housing Data

```
anova(fit)

## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Size       1 12393.1  12393.1   61.998 2.66e-06 ***
## Residuals 13  2598.6    199.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0
```

$$R^2 = \frac{SSR}{SST} = \frac{12393.1}{2598.6 + 12393.1} = 0.8267$$

## Back to Baseball

Three very similar, related ways to look at a simple linear regression... with only one  $X$  variable, life is easy!

	$R^2$	corr	SSE
OBP	0.88	0.94	0.79
SLG	0.76	0.87	1.64
AVG	0.63	0.79	2.49