STA 371G: Statistics and Modeling

Simple Linear Regression: Covariance and Correlation, Goodness of Fit

Mingyuan Zhou
McCombs School of Business
The University of Texas at Austin

http://mingyuanzhou.github.io/teaching

Sample Mean and Sample Variance

- Sample Mean: measure of centrality

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

- Sample Variance: measure of spread

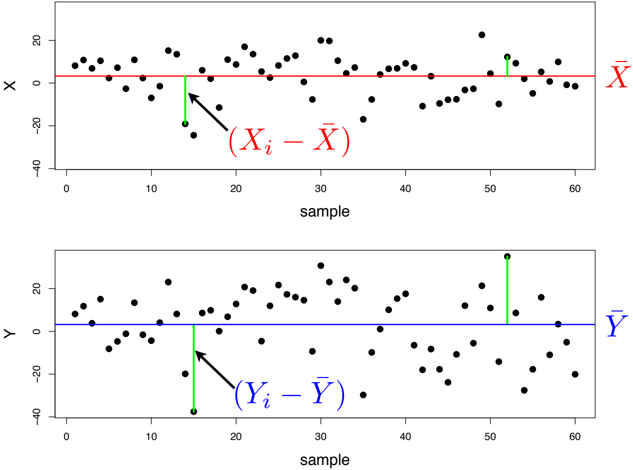
$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

- Sample Standard Deviation:

$$s_y = \sqrt{s_y^2}$$

Example

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

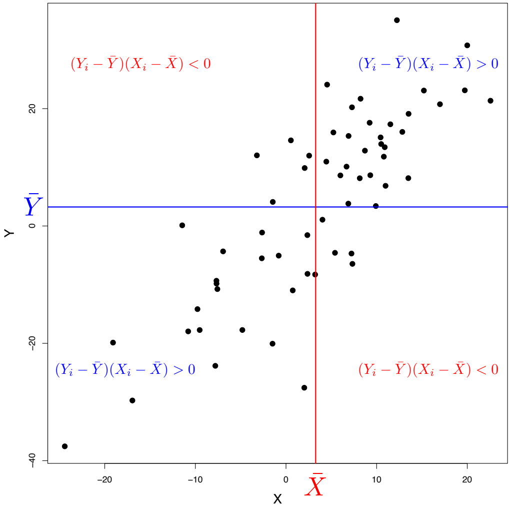


$s_x = 9.7 \qquad s_y = 16.0$

Sample Covariance

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

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



- $s_y = 15.98$, $s_x = 9.7$
- $Cov(X, Y) = 125.9$

How do we interpret that?

Sample Correlation

Correlation is the standardized covariance:

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

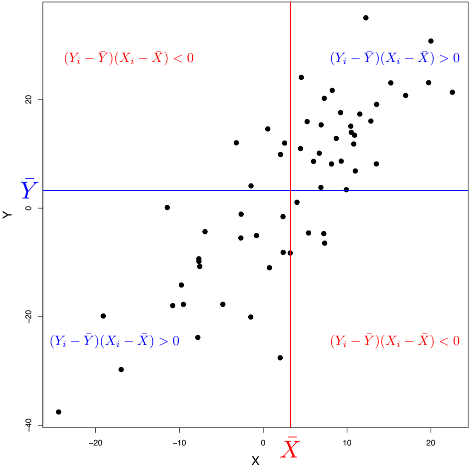
The correlation is scale invariant and the units of measurement don't matter: It is always true that $-1 \leq \mathrm{corr}(X, Y) \leq 1$.

$$\mathrm{corr}(aX, bY) = \frac{\mathrm{cov}(aX, bY)}{s_{ax} s_{by}} = \frac{ab\ \mathrm{cov}(X, Y)}{|a||b|\ s_x s_y} = \mathrm{sign}(ab)\mathrm{corr}(X, Y)$$

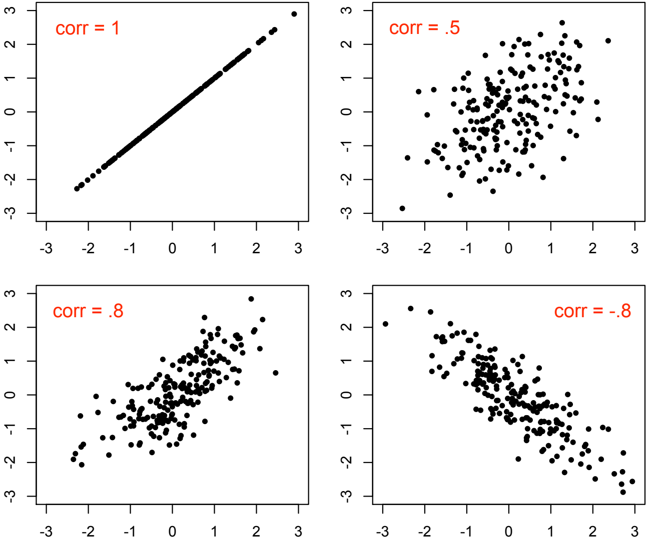
This gives the direction (- or $+$) and strength $(0 \to 1)$ of the linear relationship between X and Y.

Sample Correlation

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



Sample Correlation

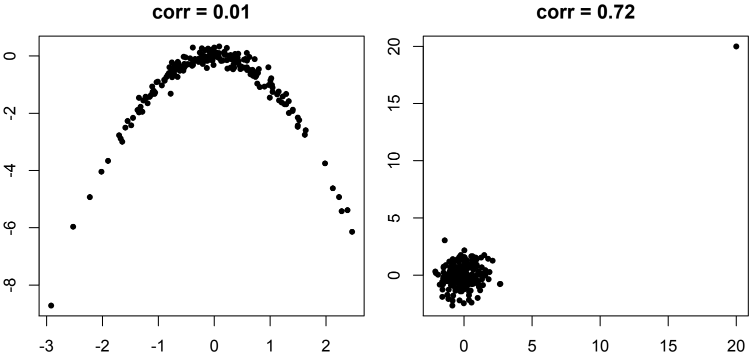


Sample Correlation

Only measures linear relationships:

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

Example: $\text{corr}(X, Y){=}0$ if $Y = -X^2,\ X \sim \mathcal{N}(0, 1)$.



Also be careful with influential observations. Excel Break: correl, stdev,...

Back to Least Squares

1. Intercept:

$$b_0 = \bar{y} - b_1 \bar{x} \Rightarrow \bar{y} = b_0 + b_1 \bar{x}$$

   - The point $(\bar{x}, \bar{y})$ is on the regression line!
   - Least squares finds the point of means and rotate the line through that point until getting the "right" slope

2. Slope:

$$\begin{aligned} b_1 = corr(X, Y) \times \frac{s_y}{s_x} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{Cov(X, Y)}{var(X)} \end{aligned}$$

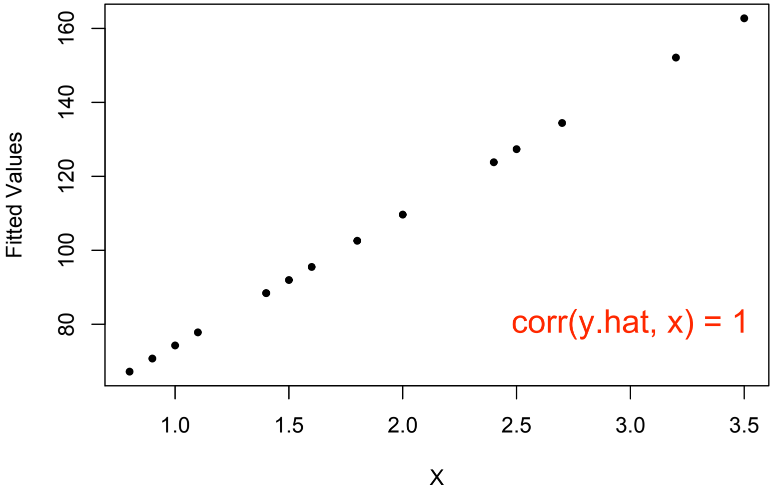
   - So, the right slope is the *correlation coefficient* times a *scaling factor* that ensures the proper units for b_1

More on Least Squares

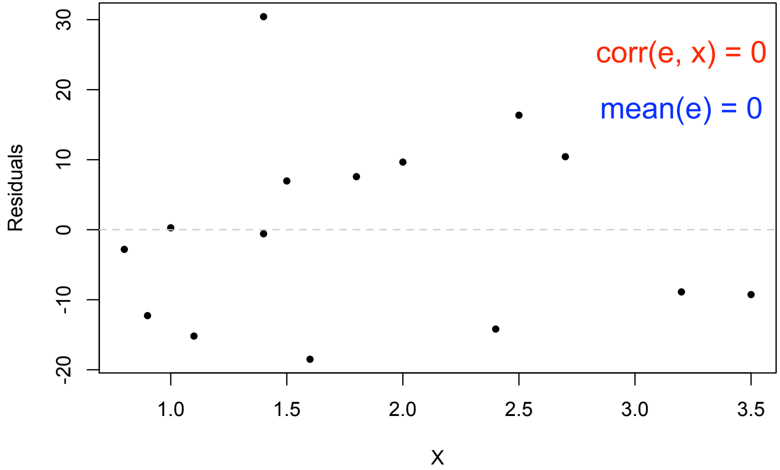
From now on, terms "fitted values" ($\hat{y}_i$) and "residuals" (e_i) refer to those obtained from the least squares line.

The fitted values and residuals have some special properties. Lets look at the housing data analysis to figure out what these properties are...

The Fitted Values and X

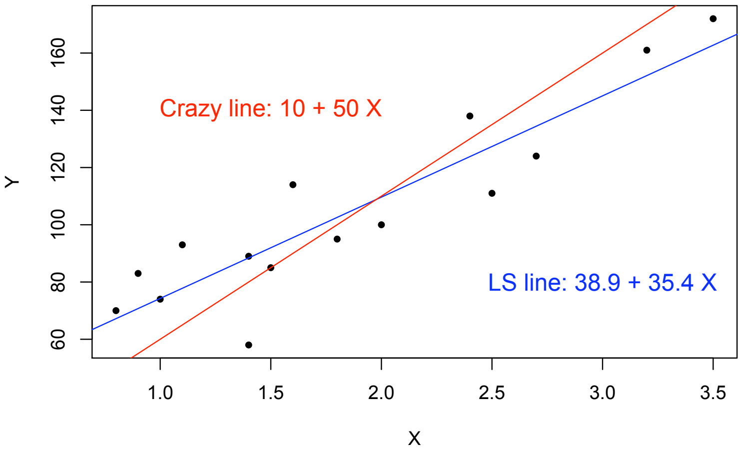


The Residuals and X



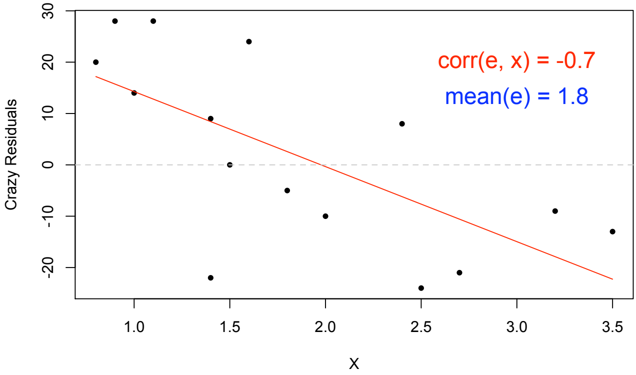
Why?

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!

Fitted Values and Residuals

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 "*Xness*" 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; $\text{corr}(X, e) = 0$.
- $(\bar{X}, \bar{Y})$ is on the regression line.
- $\bar{e} = \frac{\sum_{i=1}^{n} e_i}{n} = 0$.
- $\text{corr}(\hat{Y}, e)$=?

Decomposing the Variance

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

Remember that $Y = \hat{Y} + e$; $\hat{Y}$ and e are uncorrelated, i.e. $\mathrm{corr}(\hat{Y}, e) = 0$; and $\bar{e} = 0$.

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

Decomposing the Variance

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

SSR: Variation in Y explained by the regression line.

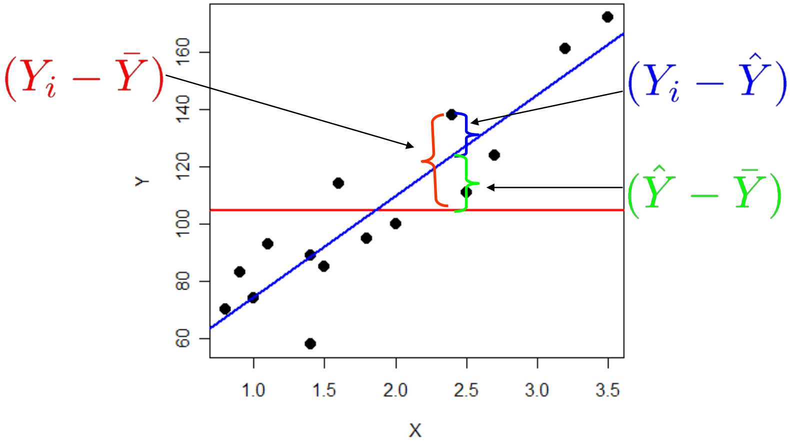
SSE: Variation in Y that is left unexplained.

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

Be careful of similar acronyms; e.g. SSR for "residual" SS.

Decomposing the Variance

$$
\begin{aligned}
(Y_i - \bar{Y}) &= \hat{Y}_i + e_i - \bar{Y} \\
&= (\hat{Y}_i - \bar{Y}) + e_i
\end{aligned}
$$



A Goodness of Fit Measure: R^2

The coefficient of determination, denoted by R^2, measures goodness of fit:

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

- $0 < R^2 < 1$.
- The closer R^2 is to 1, the better the fit.

A Goodness of Fit Measure: R^2 (Optional)

An interesting fact: $R^2 = r_{xy}^2$ (i.e., R^2 is squared correlation).

$$
\begin{aligned}
R^2 &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \frac{\sum_{i=1}^n (b_0 + b_1 X_i - b_0 - b_1 \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \frac{b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{b_1^2 s_x^2}{s_y^2} = r_{xy}^2
\end{aligned}
$$

No surprise: the higher the sample correlation between X and Y, the better you are doing in your regression.

Back to the House Data

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.909209967
R Square	0.826662764
Adjusted R Square	0.81332913
Standard Error	14.13839732
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	12393.10771	12393.10771	61.99831126	2.65987E-06
Residual	13	2598.625623	199.8942787		
Total	14	14991.73333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	38.88468274	9.09390389	4.275906499	0.000902712	19.23849785	58.53086763	19.23849785	58.53086763
X Variable 1	35.38596255	4.494082942	7.873900638	2.65987E-06	25.67708664	45.09483846	25.67708664	45.09483846

SSR

SST

SSE

$$R^2 = \frac{SSR}{SST} = 0.83 = \frac{12393}{12393 + 2599}$$

```
> HouseFit=lm(formula=Price~Size)
> summary(HouseFit)

Call:
lm(formula = Price ~ Size)

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 

> anova(HouseFit)
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                      
```

$$R^2 = \frac{SSR}{SST} = 0.83 = \frac{12393}{12393 + 2599}$$

```
> HouseFit=lm(formula=Price~Size)
> summary(HouseFit)

Call:
lm(formula = Price ~ Size)

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 

> anova(HouseFit)
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                     
```

$$R^2 = \frac{SSR}{SST} = 0.83 = \frac{12393}{12393 + 2599}$$

Interpretation of R^2

- R^2 measures the proportion of variation in Y explained by X.
- R^2 measures the reduction of SSE

  from

  Running regression without predictors X

  to

  Running regression with predictors X

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