

Regression II

Professor: Hammou El Barmi
Columbia University

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi} + \epsilon_i$$

Here

- β_0 is the mean of Y when all the X s are equal to zero
- β_i is the change in the mean of Y when we increase X_i by one while holding all the other X s fixed

In matrix formulation,

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

where

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{p1} \\ 1 & x_{12} & x_{22} & \dots & x_{p2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{pn} \end{pmatrix}$$

Example: antique grandfather clocks

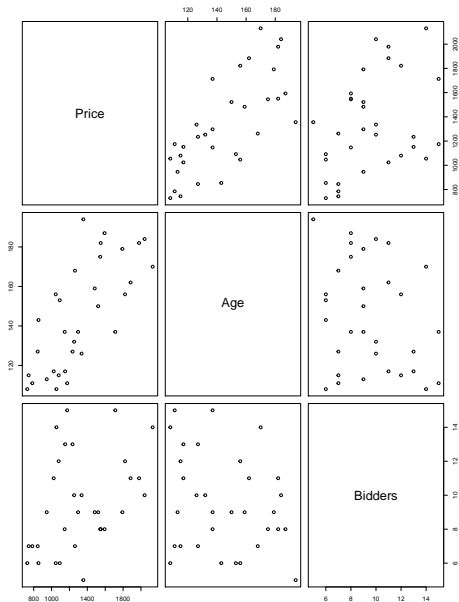
- The data give the selling price at auction of 32 antique grandfather clocks. Also recorded is the age of the clock and the number of people who made a bid.
- The variables are
 - ① Age : Age of the clock (years)
 - ② Bidders: Number of individuals participating in the bidding
 - ③ Price: Selling price (pounds sterling)

	Age	Bidders	Price
1	127	13	1235
2	115	12	1080
3	127	7	845
4	150	9	1522
5	156	6	1047
6	182	11	1979
7	156	12	1822
8	132	10	1253
9	137	9	1297

Example: antique grandfather clocks

10	113	9	946
11	137	15	1713
12	117	11	1024
13	137	8	1147
14	153	6	1092
15	117	13	1152
16	126	10	1336
17	170	14	2131
18	182	8	1550
19	162	11	1884
20	184	10	2041
21	143	6	854
22	159	9	1483
23	108	14	1055
24	175	8	1545
25	108	6	729
26	179	9	1792
27	111	15	1175
28	187	8	1593
29	111	7	785
30	115	7	744
31	194	5	1356
32	168	7	1262

Example (Clocks continued)



To estimate β we minimize

$$\sum_{i=1}^n \epsilon_i^2 = \epsilon^T \epsilon = (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta)$$

The solution is

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Under the assumption we made before

$$\mathbf{b} \sim N(\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1})$$

This implies in particular that

- $E(\mathbf{b}) = \beta$, that is, \mathbf{b} is an unbiased estimator of β .
- $\text{Var}(\mathbf{b}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$ and is estimated by

$$\text{Var}(\mathbf{b}) = \text{MSE} (\mathbf{X}^T \mathbf{X})^{-1}$$

- An estimate of the variance of β_i is

$$SE(b_i) = \text{MSE} (\mathbf{X}^T \mathbf{X})_{ii}^{-1}$$

where $(\mathbf{X}^T \mathbf{X})_{ii}^{-1}$ is the i th diagonal element of $(\mathbf{X}^T \mathbf{X})^{-1}$.

```
> attach(data)
> fit<-lm(Price~ Age+Bidders)
> fit
```

Call:

```
lm(formula = Price ~ Age + Bidders)
```

Coefficients:

(Intercept)	Age	Bidders
-1336.72	12.74	85.82

The regression equation is

$$\widehat{Price} = -1336.72 + 12.74Age + 85.82Bidders$$

- A $100(1 - \alpha)\%$ confidence interval for β_i is

$$b_i \pm t_{n-p-1}(\alpha/2)SE(b_i)$$

The interpretation of this confidence interval is: We are $100(1 - \alpha)\%$ confident that when we increase X_i by one unit while holding all the other X s fixed, the average, Y changes by an amount in this interval.

```
> confint(fit)
```

	2.5 %	97.5 %
(Intercept)	-1691.27514	-982.16896
Age	10.89062	14.58177
Bidders	68.00986	103.62040

We are 95% that when we increase age by one year while holding the number of bidders fixed, on average the price goes by an amount between 10.89 and 12.58 pounds sterling.

To test $H_0 : \beta_i = \beta_{i0}$ against $H_a : \beta_i \neq \beta_{i0}$, the test statistic is

$$t = \frac{b_i - \beta_{i0}}{SE(b_i)}$$

and we reject H_0 if

$$|t| > t_{n-p-1}(\alpha/2) \quad \text{or if} \quad p\text{-value} < \alpha$$

```
> summary(fit)
lm(formula = Price ~ Age + Bidders)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1336.7221	173.3561	-7.711	1.67e-08 ***
Age	12.7362	0.9024	14.114	1.60e-14 ***
Bidders	85.8151	8.7058	9.857	9.14e-11 ***

Residual standard error: 133.1 on 29 degrees of freedom

Multiple R-squared: 0.8927, Adjusted R-squared: 0.8853

F-statistic: 120.7 on 2 and 29 DF, p-value: 8.769e-15

p-values very small we reject $H_0 : \beta_i = 0$ against $H_a : \beta_i \neq 0$

- The ANOVA table is given by

Source	df	SS	MS	F
Model	p	SSR	$MSR=SSR/p$	MSR/MSE
Error	n-p-1	SSE	$MSE=SSE/(n-p-1)$	
Total	n-1	SST		

- The coefficient of determination is

$$R^2 = \frac{SSR}{SST}$$

- Adjusted

$$R_{adj}^2 = 1 - \frac{n-1}{n-p-1} \frac{SSE}{SST}$$

can be used for model selection

- MSE is an estimate of σ^2

- To test $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$ against H_a : at least one of these β s is not zero, we reject H_0 if $F > F(1 - \alpha, p, n - p - 1)$ or if $p - value < \alpha$.
- The test statistic is given by

$$F = \frac{(SSE_R - SSE_F)/(df_R - df_f)}{SSE_F/df_F}$$

and we reject H_0 if

$$F > F(1 - \alpha, df_R - df_F, df_F)$$

or if $p - value < \alpha$.

- In the example, to test $H_0 : \beta_1 = \beta_2 = 0$ against H_a : at least one of them is not equal to zero, we

F-statistic: 120.7 on 2 and 29 DF, p-value: 8.769e-15

Since the p-value is very small we reject H_0

- Suppose we want to test $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0, k < p$ against $H_a : \text{Not } H_0$.
- In this case we have two models:
 - a reduced model (the model in which $\beta_1 = \beta_2 = \dots = \beta_k = 0$) and
 - a full model in which we have all the β s

- we have up to this point assumed that $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, 2, \dots, n$, ϵ_i s are iid $N(0, \sigma^2)$ and made inference about β_0 and β_1
- The goal of the lack of fit test is to test a specific type of regression function fits the data.
- The lack fit test assumes that the observations for a given x are
 - ① independent of each other
 - ② normally distributed
 - ③ the distribution of y given x have the same variance σ^2 .
- We want to test

$$H_0 : \mu_x = \beta_0 + \beta_1 x$$

$$H_1 : \mu_x \neq \beta_0 + \beta_1 x$$

- To carry out a lack of fit test requires repeat observations at one or more x levels

data

x	y	mean under H_0	mean under H_a
x_1	$y_{11}, y_{12}, \dots, y_{1n_1}$	$\beta_0 + \beta_1 x_1$	μ_{x_1}
x_2	$y_{21}, y_{22}, \dots, y_{2n_2}$	$\beta_0 + \beta_1 x_2$	μ_{x_2}
\vdots	\vdots	\vdots	\vdots
x_c	$y_{c1}, y_{c2}, \dots, y_{cn_c}$	$\beta_0 + \beta_1 x_c$	μ_{x_c}

- Under H_a , the model is $y_{ij} = \mu_i + \epsilon_{ij}, i = 1, 2, \dots, c, j = 1, 2, \dots, n_i$
- The estimate of μ_i is $\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$
- $SSE_R = \sum_{i=1}^c \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2$ where $\hat{y}_{ij} = b_0 + b_1 x_i$. H0对应的是reduced model
- $SSE_F = \sum_{i=1}^c \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ H1 对应的是full model
- The partial F-test is

$$F = \frac{SSE_R - SSE_F}{df_R - df_F} \div \frac{SSE_F}{df_F}$$

- Reject H_0 is $F > F(1 - \alpha, df_R - df_F, df_F)$ or if $p - value < \alpha$
- $df_F = \sum_{i=1}^c n_i - c$ and $df_R = \sum_{i=1}^c n_i - 2$
- The difference $SSE_R - SSE_F$ is called the lack of fit sum of squares and is denoted by SSLF
- The test statistic is sometimes expressed as

$$F = \frac{MSLF}{MSE}$$

Lack of fit test

x	y
0.01	127.6
0.48	124.0
0.71	110.8
0.95	103.9
1.19	101.5
0.01	130.1
0.48	122.0
1.44	92.3
0.71	113.1
1.96	83.7
0.01	128.0
1.44	91.4
1.96	86.2

```
> Reduced <- lm(y ~ x, data=corrosion)
```

```
> summary(Reduced)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	129.79	1.40	92.5	< 2e-16
x	-24.02	1.28	-18.8	1.1e-09

Residual standard error: 3.06 on 11 degrees of freedom

Multiple R-Squared: 0.97, Adjusted R-squared: 0.967

F-statistic: 352 on 1 and 11 degrees of freedom, p-value: 1.06e-09

```
>Full <- lm(y~factor(x))
> summary(lm(Full))
Call:
lm(formula = y ~ factor(x))
```

Residuals:

Min	1Q	Median	3Q	Max
-1.2500	-0.9667	0.0000	1.0000	1.5333

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	128.567	0.809	158.914	4.19e-12	***
factor(x)0.48	-5.567	1.279	-4.352	0.00481	**
factor(x)0.71	-16.617	1.279	-12.990	1.28e-05	***
factor(x)0.95	-24.667	1.618	-15.245	5.03e-06	***
factor(x)1.19	-27.067	1.618	-16.728	2.91e-06	***
factor(x)1.44	-36.717	1.279	-28.703	1.18e-07	***
factor(x)1.96	-43.617	1.279	-34.097	4.24e-08	***

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

Residual standard error: 1.401 on 6 degrees of freedom

Multiple R-squared: 0.9965, Adjusted R-squared: 0.9931

F-statistic: 287.3 on 6 and 6 DF, p-value: 4.152e-07

To test for Lack of fit, we use

```
> anova(Reduced,Full)
```

Analysis of Variance Table

Model 1: y ~ x

Model 2: y ~ factor(x)

	Res.Df	Res.Sum Sq	Df	Sum Sq	F value	Pr(>F)
1	11	102.9				
2	6	11.8	5	91.1	9.28	0.0086

$$SSE_R = 102.9, SSE_F = 11.8, df_R = 11, df_F = 6.$$

$$F = \frac{102.9 - 11.8}{11 - 6} \div \frac{11.8}{6} = \frac{91.1}{5} \div \frac{11.8}{6} = 9.28.$$

The p-value is 0.0086. Reject $H_0 : \mu_x = \beta_0 + \beta_1 x$.

Regression with qualitative variables

- Y = volume of sales in July of some electronic store
- x = number of households in the location
- Location of the store = $\begin{cases} \text{Mall} \\ \text{Downtown} \\ \text{Street} \end{cases}$

number of household	location	sales
161	street	157.27
99	street	93.28
135	street	136.81
120	street	123.79
164	street	153.51
221	mall	241.74
179	mall	201.54
204	mall	206.71
214	mall	229.78
101	mall	135.22
231	downtown	224.71
206	downtown	195.29
248	downtown	242.16
107	downtown	115.21
205	downtown	197.82

```
> fit
```

```
Call:
```

```
lm(formula = sales ~ nhousehold + factor(location))
```

```
Coefficients:
```

(Intercept)	nhousehold	factor(location)mall
21.8415	0.8686	21.5100
factor(location)street		
-6.8638		

```
> summary(fit)
```

Call:

```
lm(formula = sales ~ nhousehold + factor(location))
```

Residuals:

Min	1Q	Median	3Q	Max
-13.834	-2.999	2.225	4.357	6.431

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	21.84147	8.55848	2.552	0.026898	*
nhousehold	0.86859	0.04049	21.452	2.52e-10	***
factor(location)mall	21.50998	4.06509	5.291	0.000256	***
factor(location)street	-6.86378	4.77048	-1.439	0.178047	

Residual standard error: 6.349 on 11 degrees of freedom

Multiple R-squared: 0.9868, Adjusted R-squared: 0.9833

F-statistic: 275.1 on 3 and 11 DF, p-value: 1.268e-10

```
> confint(fit)
```

	2.5 %	97.5 %
(Intercept)	3.0043933	40.6785468
nhousehold	0.7794707	0.9577061
factor(location)mall	12.5627722	30.4571864
factor(location)street	-17.3635248	3.6359712

- Multicollinearity: it exists when the explanatory variables are linearly dependent.
- We use the variance inflation factor (VIF) to check whether or not multicollinearity exists
- The VIF for variable x_j is

$$VIF = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination when x_j is regressed on the other x_i s

- As a percentage, R_j^2 is the percentage variability in x_j explained by the other x_i s
- It turns out that $SE(b_j) \propto \sqrt{VIF}$, so when R_j^2 is high, the $SE(b_j)$ is also large and that leads to failing to reject $H_0 : \beta_j = 0$

Outliers and Influential Points

- An outlier is a data point whose response y does not follow the general trend of the rest of the data.
- A data point has high leverage if it has "extreme" predictor x values. With a single predictor, an extreme x value is simply one that is particularly high or low. With multiple predictors, extreme x values may be particularly high or low for one or more predictors, or may be "unusual" combinations of predictor values (e.g., with two predictors that are positively correlated, an unusual combination of predictor values might be a high value of one predictor paired with a low value of the other predictor).
- A data point is influential if it unduly influences any part of a regression analysis, such as the predicted responses, the estimated slope coefficients, or the hypothesis test results.
- Outliers and high leverage data points have the potential to be influential, but we generally have to investigate further to determine whether or not they are actually influential
- One advantage of the case in which we have only one predictor is that we can look at simple scatter plots in order to identify any outliers and influential data points.

- The hat matrix is

$$H = X(X^T X)^{-1} X^T$$

Note that $\hat{y} = H^T y$ so that

$$\begin{aligned}\hat{y}_1 &= h_{11}y_1 + h_{12}y_2 + \dots + h_{1n}y_n \\ \hat{y}_2 &= h_{21}y_1 + h_{22}y_2 + \dots + h_{2n}y_n \\ &\vdots \\ \hat{y}_n &= h_{n1}y_1 + h_{n2}y_2 + \dots + h_{nn}y_n\end{aligned}$$

- h_{ii} is the i th element of the diagonal of H . It measures the distance of the x values of the i th case from the center of the experimental region (ignores the response)
- The leverage h_{ii} , quantifies the influence that the observed response y_i has on its predicted value \hat{y}_i . That is, if h_{ii} is small, then the observed response y_i plays only a small role in the value of the predicted response \hat{y}_i .
- On the other hand, if h_{ii} is large, then the observed response y_i plays a large role in the value of the predicted response \hat{y}_i . It's for this reason that the h_{ii} are called the leverages.
- In simple linear regression

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Some important properties of the leverages:

- 1 The leverage h_{ii} is a measure of the distance between the x value for the i th data point and the mean of the x values for all n data points.
- 2 The leverage h_{ii} is a number between 0 and 1, inclusive.
- 3 The sum of the h_{ii} equals p , the number of parameters (regression coefficients including the intercept).

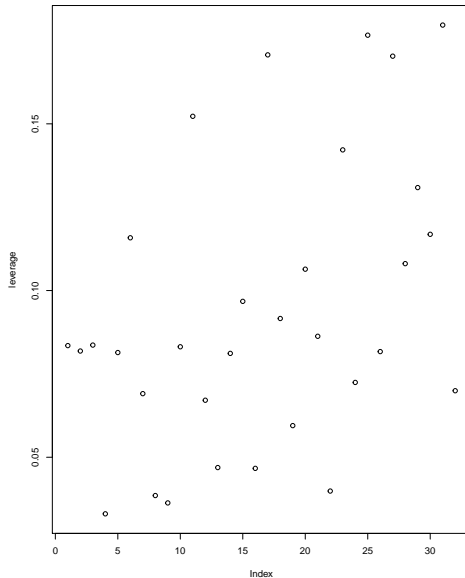
The first bullet indicates that the leverage h_{ii} quantifies how far away the i th x value is from the rest of the x values. If the i th x value is far away, the leverage h_{ii} will be large; and otherwise not. if the x is far away then the corresponding y will play an important influence in predicted value of y

Identifying data points whose **x values are extreme**

- The great thing about leverages is that they can help us identify x values that are extreme and therefore potentially influential on our regression analysis.
- How? All we need to do is determine when a leverage value should be considered large. A common rule is to flag any observation whose leverage value, h_{ii} , satisfies

$$h_{ii} > \frac{2p}{n}$$

Example (Clocks continued)

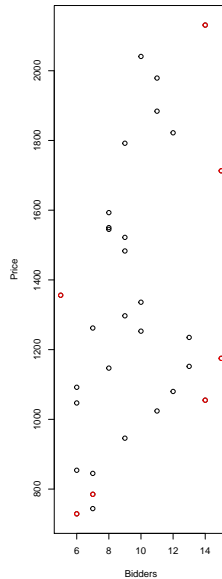
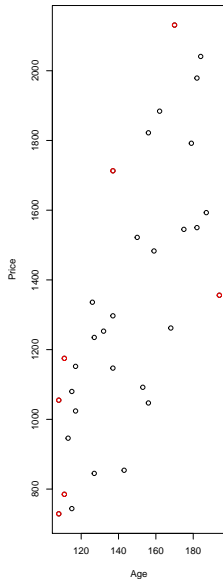


Example (Clocks continued)

```
> data[leverage>2*2/32,]
```

	Age	Bidders	Price
11	137	15	1713
17	170	14	2131
23	108	14	1055
25	108	6	729
27	111	15	1175
29	111	7	785
31	194	5	1356

Example (Clocks continued)



Identifying Outliers (Unusual y Values)

- Residuals: The i th residual is defined as $e_i = y_i - \hat{y}_i, i = 1, 2, \dots, n$.
- Studentized residuals (or internally studentized residuals) are defined for each observation, $i = 1, \dots, n$ as an ordinary residual divided by an estimate of its standard deviation:

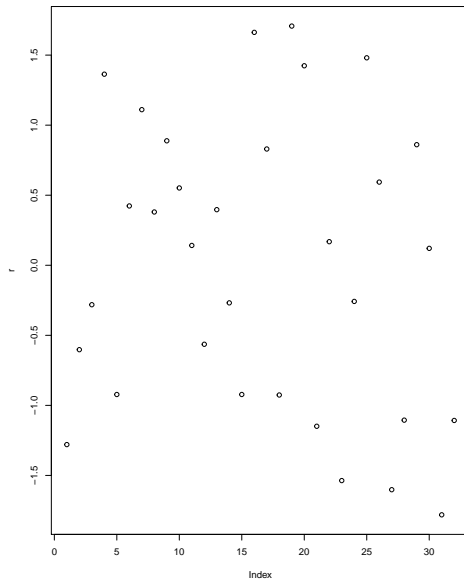
$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

e_i 的均值为0, 方差为mse
(1-leverage)

- An observation with an internally studentized residual that is larger than 3 (in absolute value) is generally deemed an outlier. [Sometimes, the term "outlier" is reserved for an observation with an externally studentized residual that is larger than 3 in absolute value?we consider externally studentized residuals in the next section.]

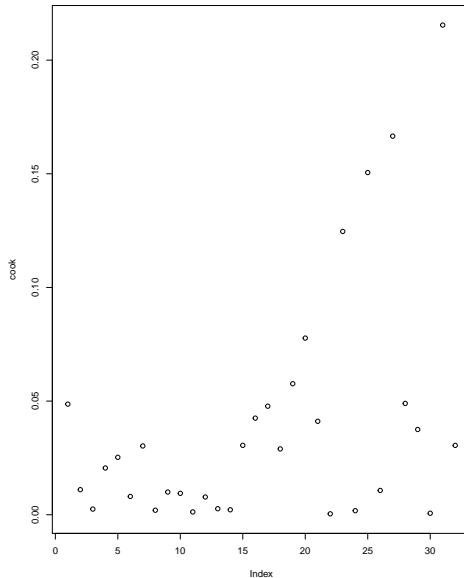
```
> r = rstudent(fit)
> data[abs(r)>3,]
[1] Age      Bidders Price
<0 rows> (or 0-length row.names)
> plot(r)
```

Example (Clocks continued)



- An influential point is one if removed from the data would significantly change the fit.
- An influential point may either be an outlier or have large leverage, or both, but it will tend to have at least one of those properties.
- Cook's distance is a commonly used influence measure that combines these two properties. It can be expressed as
- Typically, points with cook's distance greater than 1 are classified as being influential.
- We can compute the Cook's distance using the following commands
`cook = cooks.distance(fit)`

Example (Clocks continued)



- A Normal probability plot of the residuals can be used to check the normality assumption.
- Here each residual is plotted against its expected value under normality. To make normal probability plots , as well as a histogram, type:

```
> qqnorm(fit$res)  
> qqline(fit$res)  
> hist(fit$res)
```

Example (Clocks continued)

