

Handout: Multiple Regression.

Math 283, Summer 2011

Least Squares Regression

“true” equation: $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$
where β_0 is “true” intercept, β_i are “true” slopes, and μ_y is the average response to the values of x_1, x_2, \dots, x_p .

Model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i$ where ε_i are independent Normal (mean = 0, st.dev. = σ)

Parameters are $\beta_0, \beta_1, \dots, \beta_p$ and σ .

The estimated equation: $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$, where b_0, b_1, \dots, b_p are the least squares estimates (difficult).

Residual of the observation i : $e_i = y_i - \hat{y}_i$

σ is estimated using $S = \sqrt{S^2}$, where
$$S^2 = \sum e_i^2 / (n-p-1) = MS(\text{Residual Error})$$

ANOVA (ANalysis Of VAriance) table

First, we would see if any x-variables at all are significant. (F-test)
Once we establish that, we can test variables one by one to see which are significant.

F-test

F random variable has 2 kinds of degrees of freedom:

$$\text{Df(Numerator)} = p$$

$$\text{Df(Denominator)} = n - p - 1 \quad (\text{same as } s^2 = \text{“Mean square error”})$$

P-values are given in Table E.

T-statistic:

Is used to test one variable at a time. As before, $t = \text{Estimate} / \text{SE}(\text{Estimate})$

Diagnostics: plot Residuals versus Fits to spot any trends in the data.

Example:

predicting house prices. Sample of $n = 63$ houses.

(Data are available at <http://www.nmt.edu/~olegm/283labs/house.txt>)

Y = selling price, X_1 = Area (sq .ft.) X_2 = # rooms, X_3 = # bedrooms, X_4 = # baths,
 X_5 = age.

The regression equation is
 Price = 2271 + 44.8 Area + 5805 Rooms - 9436 Bedr - 369 age +
 7652 bath

Step 1: does the price depend on any of the variables? Look at the ANOVA table.

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	5	55639266497	11127853299	31.10	?
Residual Error	57	20397723344	357854796		
Total	62	76036989841			

- Note that:
- 1) first two rows add to Total.
 - 2) MS (mean square) = SS (sum of squares) / DF
 - 3) F = MS(Regression) / MS(Error).

High values of F signify that regression variables explain a great deal, compared to random error. Thus, high F will lead to low p-values for the test.

Use Table E (page T-13 and further) for p-value.

Step 2: which X-variables are significant?

Predictor	Coef	SE Coef	T	P
Constant	2271	13531	0.17	0.867
Area	44.815	9.305	4.82	0.000
Rooms	5805	2563	2.26	0.027
Bedrooms	-9436	6038	-1.56	0.124
age	-369.5	122.6	-3.01	0.004
bath	7652	6782	1.13	0.264

S = 18917 R-Sq = 73.2% R-Sq(adj) = 70.8%

Results of individual T-tests are given. As we see, Area, Rooms and Age are significant. (Caution: T-test results depend on other terms in the model!)

How well will price be predicted? _____

How well, in general, the price correlates with the variables given?

“R-sq” is called **Squared Multiple Correlation** (which is r^2 in one-dimensional case) says 73.2% of variation in prices is explained away by the predictors.

Also, note that $R^2 = SS(\text{Regression}) / SS(\text{Total}) = (\text{Book: SSM/SST})$

Step 3: are there any unusual observations?

Unusual Observations

Obs	Area	Price	Fit	SE Fit	Residual	St Resid
10	3600	149000	187323	10771	-38323	-2.46RX
50	2353	210000	130004	5370	79996	4.41R
59	2038	141000	117915	10764	23085	1.48 X
61	1545	139000	90010	3629	48990	2.64R

R denotes an observation with a large standardized residual

X denotes an observation whose X value gives it large influence.

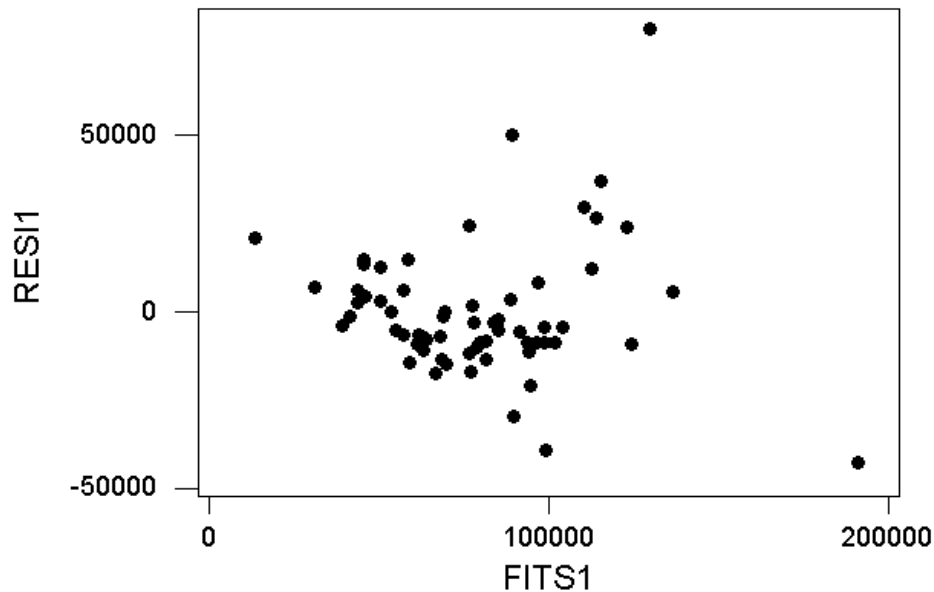
Minitab has a way of pointing out influential observations. The results could drastically change if one or more of these observations are omitted.

It may make sense to re-fit the equation omitting the insignificant variables:

$$\text{Price} = -88 + 42.7 \text{ Area} + 4517 \text{ Rooms} - 434 \text{ age}$$

The estimates changed quite a bit!

Step 4. Look at the residual plot.



Two obvious outliers aside (which observations are they?), there is overall funnel shape. (An ideal residual plot should have no structure.) This means, for example, that errors in prediction are higher for higher-priced houses.

We might want to use some kind of nonlinear regression (discussed next).