Regression: Basics

Professor: Hammou El Barmi
Columbia University

Introduction

- Regression is one of the most widely used of all the statistical methods
- In the univariate case, the data are:
    - one response variable Y
    - p predictor variables $X_1, X_2, \ldots, X_k$.
- One of goals of regression are
    - investigate how Y is related to $X_1, X_2, \ldots, X_p$.
    - estimation of the conditional mean of Y given $X_1, X_2, \ldots, X_p$
    - predict future values of Y when values of $X_1, X_2, \ldots, X_p$ are given
- The multiple regression model relating Y to the predictors $X_1, X_2, \ldots, X_p$ is

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_p X_{i,p} + \epsilon_i$$

  where ϵ_i is called noise, disturbances or errors.
- It is assumed that

$$E(\epsilon_i | X_{i,1}, X_{i,2}, \ldots, X_{i,p}) = 0$$

  as a results

$$E(Y_i | X_{i,1}, X_{i,2}, \ldots, X_{i,p}) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_p X_{i,p}$$

- The parameter β_0 is the intercept and the $\beta_1, \beta_2, \ldots, \beta_p$ are the slopes
- $$\beta_j = \frac{\partial E(Y_i | X_{i,1}, X_{i,2}, \ldots, X_{i,p})}{\partial X_{i,j}}$$
- β_j is the change in the expected value of Y_i when $X_{i,j}$ is increased one unit while holding other predictors fixed.

Assumption

The regression assumptions are

1. Linearity of the conditional expectation (mean)

$$E(Y_i|X_{i,1}, X_{i,2}, \ldots, X_{i,p}) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_p X_{i,p}$$

2. Independent noise (errors): $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent
3. Constant variance: $\mathrm{Var}(\epsilon_i) = \sigma_\epsilon^2$ for all i
4. Gaussian noise: ϵ_i is normally distributed.

Simple Linear Regression (Straight-Line Regression)

- Simple linear regression (straight-line regression) is linear regression with only one predictor variable.
- The model is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- β_0 and β_1 (the regression coefficients) are unknown intercept and slope
- The regression coefficients are estimated by the *method of least squares*
- The least squares estimates are values $\hat{beta}_0$ and $\hat{\beta}_1$ that minimize

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- Using Calculus we get

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\ \hat{\beta}_1 &= \overline{Y} - \hat{\beta}_1 \overline{X} \end{aligned}$$

- It can also be shown that

$$\hat{\beta}_1 = r\frac{s_Y}{s_X}$$

where r is the sample correlation coefficient between X and Y and s_Y and s_Y are the sample standard deviation corresponding to X and Y, respectively.

- The *least squares line* is

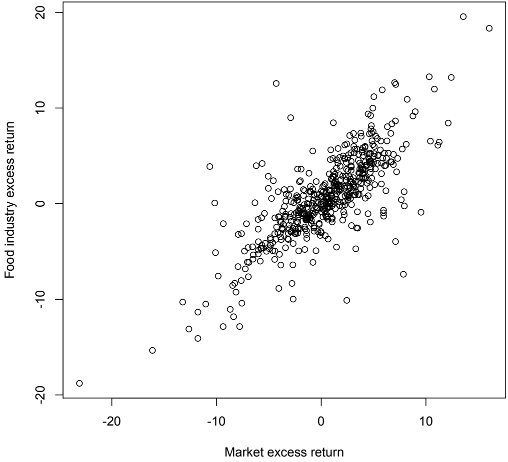
$$\begin{aligned} \hat{Y} &= \hat{\beta}_0 + \hat{\beta}_1 X \\ &= \overline{Y} + \hat{\beta}_1(X - \overline{X}) \\ &= \overline{Y} + \frac{s_{XY}}{s_X^2}(X - \overline{X}) \end{aligned}$$

where

$$s_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

is the sample covariance between X and Y

Straight -Line Regression



Straight -Line Regression

$>$ lm(rfood $\sim$ rmrf) Coefficients:

(Intercept)	rmrf
0.3392	0.7834

> summary(lm(rfood~ rmrf))

Residuals:

Min	1Q	Median	3Q	Max
-13.869	-1.310	-0.194	1.395	15.600

Coefficients:

	Estimate	Std. Error	t -value	$Pr(>\|t\|)$
Intercept	0.33918	0.12756	2.659	0.00808 **
rmrf	0.78342	0.02835	27.631	$< 2e-16$ ***

Multiple R-squared: 0.5976, Adjusted R-squared: 0.5969

F-statistic: 763.5 on 1 and 514 DF, p-value: <2.2e-16

The estimated regression line in this case if

$$\hat{Y} = 0.00339 + 0.78342X$$

Interpretation of the results

- If the excess return of market is 0, the average excess return for the food industry is estimated to be 0.339%
- If the excess return of the market increases of one percent, the average excess return of the food industry will increase by about 0.78342%

Each of the coefficient in the output has three other statistics associated with it:

- Std. Error = standard error. This is the estimated standard deviation of the least squares estimator and tells us the precision of that estimator.
- t-value. This is the t-statistic for testing that the coefficient is 0.
- p-value. This is the p-value for the test of the null hypothesis that the coefficient is 0 versus the alternative that it is not 0. If the p-value is small, then there is evidence that the coefficient is not 0 which means that the predictor has some effect.
- In our example, for the slope, the standard error is equal to 0.02835, the t-value is equal to 27.631 and the p-value is less than 0.0000000000000002. As a results, there is strong evidence that the slope is not 0

- The model is

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \ldots + \beta_p X_{i,p} + \epsilon_{i,p}$$

- The least squares estimates are $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, the solution to

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i,1} - \ldots - \beta_p X_{i,p})^2$$

- The ith fitted value is given by

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i,1} + \ldots + \hat{\beta}_p X_{i,p}.$$

  It is an estimate of $E(Y_i | X_{i,1}, X_{i,2}, \ldots, X_{i,p})$,

- The ith residual is

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

- An unbiased estimate of σ_ϵ^2 is

$$\hat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n - p - 1}$$

- The total variation in Y can be partitioned into two parts:
  - the variation that can be predicted by $X_1, X_2, \ldots, X_p$
  - the variation that can be predicted by $X_1, X_2, \ldots, X_p$
- The total variation in Y is given by

$$\text{total SS} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

- The total variation in Y can be predicted by $X_1, X_2, \ldots, X_p$ is given by

$$\text{regression SS} = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$$

- The variation in Y that cannot be predicted $X_1, X_2, \ldots, X_p$ is given by

$$\text{total SS} = \sum_{i=1}^{n} (Y_i - \hat{Y})^2$$

- R-squared, denoted by R^2 is defined as

$$R^2 = \frac{\text{regression SS}}{\text{total SS}}$$

- R^2 measures the proportion of the variability in Y that can be explained by $X_1, X_2, \ldots, X_p$

Example (continued)

```
anova(lm(rfood rmrf))

Analysis of Variance Table

Response: rfood

            Df Sum Sq Mean Sq F value    Pr(>F)

rmrf         1 6355.7  6355.7  763.49 < 2.2e-16 ***
```

Analysis of Variance, Sums of Squares, R^2

- The degrees of freedom for regression is $p =$ number of predictor variables. For a straight line regression $p = 1$
- The total degrees of freedom is $n - 1$.
- The residual error degrees of freedom is $n - p - 1$.
- The mean sum of squares (MS) for any source is its sum of squared divided by its degrees of freedom.
- The residual MS is an unbiased estimator of σ_ϵ^2.
- The other means sum of squares are used for testing.

- To test $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ against H_a : at least one of them is not zero, we use the $F-$test.
- $$F = \frac{\text{regression MS}}{\text{residual error MS}}$$
- The F-statistic tests null hypothesis that there is no linear relationship between any of the predictors and Y.
- If the p-value corresponding to this test is too small we reject H_0 and conclude that there is a relationship between the response and the predictors.

- R^2 is biased in favor of large models. It always increases by adding more predictors even if they are independent of the response.
- Recall that

$$R^2 = \frac{\text{regression SS}}{\text{total SS}} = 1 - \frac{n^{-1}\text{residual error } SS}{n^{-1}\text{total SS}}$$

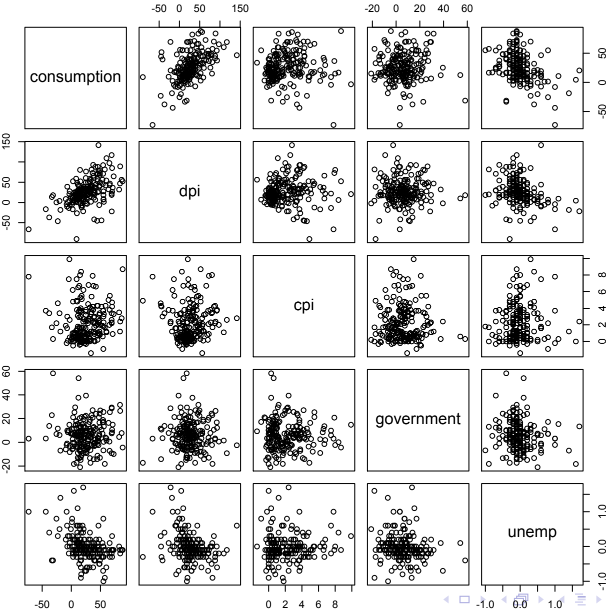
- The bias in R^2 can be removed by using the following adjustment which replaces both occurrences of n by the appropriate degrees of freedom

$$\text{adjusted } R^2 = 1 - \frac{(n-p-1)^{-1}\text{residual error } SS}{(n-1)^{-1}\text{total SS}}$$

- The presence of p is the adjusted R^2 penalizes the criterion for the number of predictor variables, so adjusted R^2 can either increase or decrease when predictor variables are added to the model.
- Adjusted R^2 increases in the added variables decrease the residual sum of squares enough to compensate for the increase in p.
- The adjusted R^2 statistic can be used to select models.

Example:

- The data USMacroG in R's AER package contains quarterly time series on 12 US macroeconomic variables for the period 1950-2000
- The variables we use are:
  - consumption= real consumption expenditure
  - dpi= real disposable personal income
  - cpi= consumer price index
  - government= real government expenditure
  - unemp= unemployment rate
- Goal: predict changes in consumption from changes in the other variables.


consumption
dpi
cpi
government
unemp
-50
0
50
150
-20
20
40
60
2
4
6
8
-1.0
0.0
1.0

```
$>$ summary(fitlm1)

Call:

lm(formula = consumption $~$ dpi + cpi + government + unemp)

Residuals:

    Min      1Q  Median      3Q     Max

-60.626 -12.203  -2.678   9.862  59.283

Coefficients:

              Estimate Std. Error t value Pr(>|t|)

(Intercept)  14.752317   2.520168   5.854 1.97e-08 ***

dpi           0.353044   0.047982   7.358 4.87e-12 ***

cpi           0.726576   0.678754   1.070    0.286

government   -0.002158   0.118142  -0.018    0.985

unemp       -16.304368   3.855214  -4.229 3.58e-05 ***
---
Residual standard error: 20.39 on 198 degrees of freedom

Multiple R-squared:  0.3385,Adjusted R-squared:  0.3252
```

Partial F-Test

- Suppose we have two models I and II and the predictors in model I are a subset of those in model II (model I is submodel of model II)
- A common null hypothesis is H_0 : data generated by model I or equivalently, in model II the slopes of the variables that are not in model I are zero.
- The test of H_0 uses excess regression sum of squares of model II relative to model I:

$$\begin{aligned} SS(II|I) &= \text{regression SS for model II} - \text{regression SS for model I} \\ &= \text{residual SS for model I} - \text{residual SS for model II} \end{aligned}$$

- The degrees of freedom of $SS(II|I) = df_{II|I} = p_{II} - p_I$ where p_{II} and p_I are the numbers of predictors in model II and model I, respectively.
- The partial $F-$ statistics is

$$F = \frac{\mathrm{MS}(II|I)}{\hat{\sigma}_\epsilon^2}$$

where $\hat{\sigma}_\epsilon^2$ is the mean residual sum of squares for model II and

$$\mathrm{MS}(II|I) = \frac{\mathrm{SS}(II|I)}{p_{II} - p_I}$$

- Under the null hypothesis, F has an F-distribution with $df_{II|I}$ and $n - p_{II} - 1$ degrees of freedom and the null hypothesis is rejected if the F statistic exceeds the $\alpha-$upper quantile of this $F-$ distribution.

> fitlm1=lm(consumption dpi+cpi+ governemnt+ unemp)
> summary(fitlm1)
Coefficients:

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
Intercept	14.752317	2.520168	5.854	1.97e-08
dpi	0.353044	0.047982	7.358	4.87e-12
cpi	0.726576	0.678754	1.070	0.286
government	-0.002158	0.118142	-0.018	0.985
unemp	-16.304368	3.855214	-4.229	3.58e-05

> fitlm2=lm(consumption dpi+unemp)
> summary(fitlm2)
Coefficients:

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
Intercept	16.28476	1.91084	8.522	3.79e-15
dpi	0.35567	0.04778	7.444	2.84e-12
unemp	-16.01489	3.79216	-4.223	3.66e-05

> anova(fitlm2, fitlm1)
Analysis of Variance Table
Model 1: consumption dpi + unemp
Model 2: consumption dpi + cpi + government + unemp

	Res.Df	RSS	Df	Sum of Sq	F	$Pr(> F)$
Model 1	200	82767				
Model 2	198	82290	$df_{II\|I} == 2$	$SS(II\|I) = 476.61$	0.5734	0.5645

- When there are many potential predictor variables, often we wish to find a subset of them that provides a parsimonious regression model.
- Model selection means selection of the predictor variables.
- To do so, we can use
  - Adjusted R^2
  - AIC criterion
  $$\text{AIC} = n\log(\hat{\sigma}^2) + 2(1+p)$$
  where $1+p$ is the number of parameters in the model
  - BIC criterion
  $$\text{BIC} = n\log(\hat{\sigma}^2) + (1+p)\log(n)$$
  - C_p criterion
  $$C_p = \frac{SSE(p)}{\hat{\sigma}^2_{\epsilon,M}} - n + 2(p+1)$$
  C_p supposes that there are M predictors, $\hat{\sigma}^2_{\epsilon,M}$ is the estimate of σ^2_ϵ using all of them and $SSE(p)$ is the residual error sum of squares for a model with p predictors ($p \leq M$)
- The smaller values of C_p, AIC and BIC are better.

- If two predictor variables are highly correlated with each other, it becomes difficult to estimate their separate effects on the response
- The effect of high correlation between the predictor variables is that the slope of each variable is very sensitive to whether the other variable is in the model or not. High Collinearity inflates the standard error of the estimates of the slopes and renders them insignificant
- The variance inflation factor (VIF) of a variable tells us how much the squared standard error of its estimated slope is increasing by having the other predictor variables in the model. For example if $VIF = 4$ for some variable, then the variance of its $\hat{\beta}$ is 4 times larger than it would be if the other predictors were deleted.
- Suppose $X_1, X_2, \ldots, X_p$ are the predictors and let R_j^2 be the R^2-value from the regression of X_j on the other predictors. The VIF of X_j is

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

- A value of R_j^2 close to 1 implies large VIF_j. The remedy to collinearity if to reduce the number of predictors.
- if $VIF_j > 10$, there is multicollinearity (some books use 5, this is just a rule of thumb)
- Example

```
> vif(fitlm2)
dpi        unemp
1.095699   1.095699
```