Data Analytics for Materials Science

27-737

A.D. (Tony) Rollett, R.A. LeSar (Iowa State Univ.)

Dept. Materials Sci. Eng., Carnegie Mellon University

Multiple Linear Regression

Lecture 5

Revised: 15th Feb. 2021

The background to this lecture can be found in Ch. 4 of the Jobson book, Volume 1.

At the end of this lecture and on Wednesday we offer a tutorial on the use of R for MLR, which will be very useful for the second homework assignment.

The instructor will be Dr. Amit K. Verma

1. Regression residuals are assumed to be normally distributed, as we had in normal linear regression analysis (LR)
2. A linear relationship is assumed between the dependent variable and the independent variables (also as in LR).
3. The residuals all have the same finite variance (homoscedastic) and are approximately rectangular-shaped.
4. The independent variables $(x_1, x_2, x_3, ...)$ are not too highly correlated. In-class question: what do we find numerically when two variables are highly correlated? What is a practical method for checking?

Assumptions in multiple linear regression analysis

Assume a data set with *n* dependent variables (y_i, i=1..n), each depending on *p* independent variables (x_{ik}, i=1..n , k=1..p) in a linear way, i.e.,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i = \mathbf{x}_i^T \boldsymbol{\beta} + e_i$$

(remember that the inner product (dot product) between 2 vectors can be written as $\mathbf{a}^T\mathbf{b} = \mathbf{b}^T\mathbf{a}$). Here,

$\mathbf{x}_i^T = \begin{pmatrix} 1 & x_{i1} & x_{i2} & \cdots & x_{ip} \end{pmatrix}$ and $\boldsymbol{\beta}^T = \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 & \cdots & \beta_p \end{pmatrix}$

We can write this as a matrix equation: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$

in which the matrices are defined on the next page.

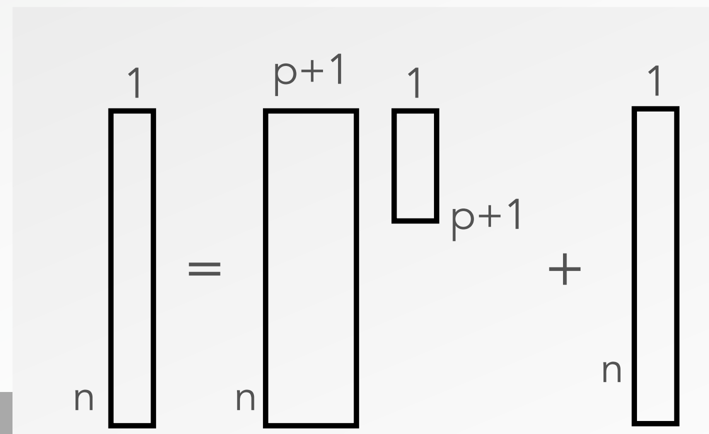
The number of rows (datapoints) in our data matrix is *n* and the number of variables to be used (to explain *y*) is *p* so *n* (number of datapoints) must be greater than *p*.

Constructing the equations

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Dimensions:



Matrix Eq.

As described in Lecture 4, in least squares optimization the residual vector is defined as $\mathbf{X}\boldsymbol{\beta} - \mathbf{Y}$ and the error as the sum of the squares of the residuals, i.e., as the inner (dot) product:

$$L(\boldsymbol{\beta}) = \| \mathbf{X}\boldsymbol{\beta} - \mathbf{Y} \|^2 = (\mathbf{X}\boldsymbol{\beta} - \mathbf{Y})^T(\mathbf{X}\boldsymbol{\beta} - \mathbf{Y})$$

$$= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} - (\mathbf{X}\boldsymbol{\beta})^T\mathbf{Y} + (\mathbf{X}\boldsymbol{\beta})^T\mathbf{X}\boldsymbol{\beta}$$

Remembering that $(\mathbf{X}\boldsymbol{\beta})^T = \boldsymbol{\beta}^T\mathbf{X}^T$, we have

$$L(\boldsymbol{\beta}) = \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

The optimal value for $\boldsymbol{\beta}$ $(\widehat{\boldsymbol{\beta}})$ is the solution of $\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$

Solving for the coefficients 1: least squares

$$L(\boldsymbol{\beta}) = \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

What is $\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$? We need results from *matrix calculus*. (Not discussed here!)

We can show that $\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{Y}^T\mathbf{X} + 2\widehat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X} = 0 \Rightarrow \widehat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X} = \mathbf{Y}^T\mathbf{X}$

Using $\widehat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X} = \mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}}$ and $\mathbf{Y}^T\mathbf{X} = \mathbf{X}^T\mathbf{Y}$, we have $(\mathbf{X}^T\mathbf{X})\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}$

Finally, we have: $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Exactly what we saw in OLS.

Knowing $\mathbf{X}$ and $\mathbf{Y}$, we can find $\widehat{\boldsymbol{\beta}}$.

Solving for the coefficients 2: least squares

Suppose we have data relating fuel consumption (in 2001) to a number of parameters. 50 states plus DC: N = 51.

To install in R:
`library(alr4)`

Drivers	FuelC	Income	Miles	MPC	Pop	Tax	State
3.5599×10^6	2.38251×10^6	23 471.	94 440.	12 737.	3.45159×10^6	18.	Alabama
472 211.	235 400.	30 064.	13 628.	7639.16	457 728.	8.	Alaska
3.55037×10^6	2.42843×10^6	25 578.	55 245.	9411.55	3.90753×10^6	18.	Arizona
1.96188×10^6	1.35817×10^6	22 257.	98 132.	11 268.4	2.07262×10^6	21.7	Arkansas
2.16238×10^7	1.46918×10^7	32 275.	168 771.	8923.89	2.55993×10^7	18.	California
3.28792×10^6	2.04866×10^6	32 949.	85 854.	9722.73	3.32246×10^6	22.	Colorado
2.65037×10^6	1.45828×10^6	40 640.	20 910.	9021.35	2.65145×10^6	25.	Connecticut
564 099.	382 043.	31 255.	5814.	10 891.3	610 269.	23.	Delaware
328 094.	148 769.	37 383.	1534.	6555.94	468 575.	20.	Dist_of_Col
1.27434×10^7	7.47112×10^6	28 145.	117 299.	9531.23	1.27418×10^7	13.6	Florida
5.8338×10^6	4.6937×10^6	27 940.	115 534.	13 248.6	6.25071×10^6	7.5	Georgia
787 820.	404 684.	28 221.	4278.	7108.75	949 184.	16.	Hawaii
896 666.	609 051.	24 180.	46 310.	10 879.4	969 166.	25.	Idaho
7.8095×10^6	5.01522×10^6	32 259.	138 359.	8239.09	9.53033×10^6	19.	Illinois
4.11692×10^6	3.12099×10^6	27 011.	94 038.	12 916.9	4.68239×10^6	15.	Indiana
1.97875×10^6	1.47581×10^6	26 723.	113 437.	10 258.4	2.281×10^6	20.	Iowa
1.8713×10^6	1.23695×10^6	27 816.	134 725.	10 656.7	2.05849×10^6	21.	Kansas
2.75663×10^6	2.08563×10^6	24 294.	78 914.	11 301.7	3.16128×10^6	16.4	Kentucky
2.71821×10^6	2.15144×10^6	23 334.	60 829.	9505.31	3.39485×10^6	20.	Louisiana
942 556.	590 093.	25 623.	22 672.	11 320.	1.01027×10^6	22.	Maine
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Note that there is a wide range in the magnitudes of these quantities, from 7.5 to 2.6 x 10^7 — 6 orders of magnitude.

To avoid mathematical instabilities that can arise from such a large spread in data, we will rescale some of the variable.

Note: *Drivers* is the number of people over 16 with licenses, *Pop* is the total population over 16.

A sample problem (not materials!)

Weisberg, S. (2014). *Applied Linear Regression*, third edition. New York: Wiley.

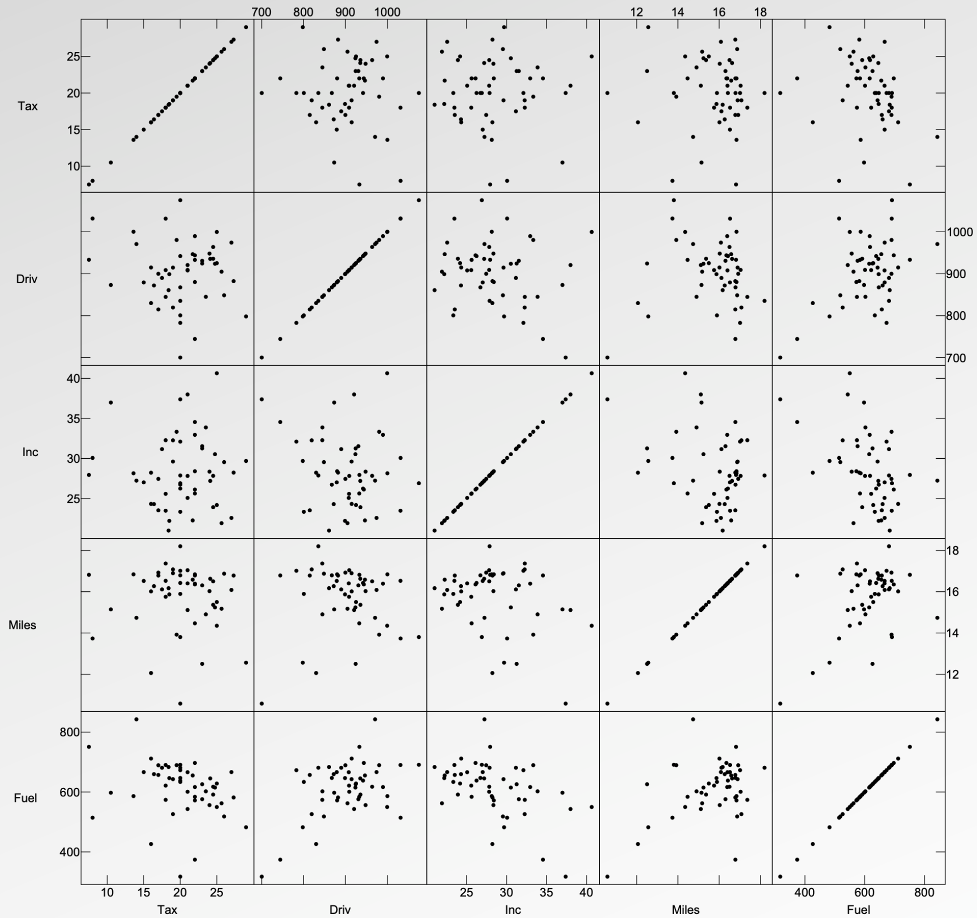
Note that use of scaled data reduces the range of the data to 7.5 to 10^3 — 2 orders of magnitude.

Tax	$\frac{1000\ \text{Drivers}}{\text{Pop}}$	$\frac{\text{Income}}{1000}$	$\frac{\text{Log[Miles]}}{\text{Log[2]}}$	$\frac{1000\ \text{FuelC}}{\text{Pop}}$	State
18.	1031.38	23.471	16.5271	690.264	Alabama
8.	1031.64	30.064	13.7343	514.279	Alaska
18.	908.597	25.578	15.7536	621.475	Arizona
21.7	946.571	22.257	16.5824	655.293	Arkansas
18.	844.703	32.275	17.3647	573.913	California
22.	989.606	32.949	16.3896	616.612	Colorado
25.	999.593	40.64	14.3519	549.993	Connecticut
23.	924.345	31.255	12.5053	626.024	Delaware
20.	700.195	37.383	10.5831	317.492	Dist_of_Col
13.6	1000.12	28.145	16.8398	586.346	Florida
7.5	933.303	27.94	16.818	750.907	Georgia
16.	829.997	28.221	12.0627	426.349	Hawaii
25.	925.193	24.18	15.499	628.428	Idaho
19.	819.437	32.259	17.0781	526.238	Illinois
15.	879.235	27.011	16.521	666.536	Indiana
20.	867.491	26.723	16.7915	647.002	Iowa
21.	909.065	27.816	17.0397	600.902	Kansas
16.4	871.998	24.294	16.268	659.741	Kentucky
20.	800.685	23.334	15.8925	633.735	Louisiana
22.	932.972	25.623	14.4686	584.093	Maine
⋮	⋮	⋮	⋮	⋮	⋮

We can use a number of approaches to examine what correlations there are between **pairs** of variables in this data.

Here we will show 2 types: a scatterplot matrix and a correlation matrix.

A sample problem (not materials!)



A scatterplot matrix (`pairs`, `scaterplotMatrix` in R, e.g.) shows each variable plotted against all the other variables.

What do we see: a large variation in *Fuel*. Some states seem to have more that 1 driver per person over the age of 16. *Fuel* use tends to decrease as *Tax* increases.

However, while useful in providing information about correlation of pairs of data, they do not tell us anything about *joint* relationships between the data.

Scatterplot matrix

Call each type of data, i.e., each variable or "feature" X_i.

We can write down the following statistical measures of each datatype.

$$\overline{X_i} = \frac{1}{N}\sum_{k=1}^{N} X_{ki}, \;\; S_i = \frac{1}{N-1}\sum_{k=1}^{N}\left(X_{ki} - \overline{X_i}\right)^2, \quad \sigma_i = \sqrt{S_i}$$

Variable	Mean	Std Dev	Minimum	Median	Maximum
Tax	20.155	4.5447	7.5	20	29
Drivers	903.68	72.858	700.20	909.07	1075.29
Income	28.404	4.4516	20.993	27.871	40.64
logMiles	15.745	1.4867	10.583	16.268	18.198
Fuel	603.13	88.96	317.49	626.02	842.79

To put all variables on the same scale, we *autoscale* the data, i.e., subtract the mean and normalize by the standard deviation.

Define: $X'_{ki} = \frac{X_{ki} - \overline{X_i}}{\sigma_i}$ for each data entry

with i = the type of data (e.g., *Income*)

We have: $\overline{X'_i} = 0$ and $\sigma_{X'_i} = 1$

Suppresses the wide range of the data and makes comparisons possible and calculations more numerically stable.

The autoscaled data:

Tax	$\frac{1000\,\text{Drivers}}{\text{Pop}}$	$\frac{\text{Income}}{1000}$	$\frac{\text{Log[Miles]}}{\text{Log[2]}}$	$\frac{1000\,\text{FuelC}}{\text{Pop}}$
-0.474153	1.75276	-1.10811	0.525958	0.867082
-2.6745	1.75634	0.372919	-1.35256	-1.11117
-0.474153	0.0675195	-0.634801	0.00564679	0.0938204
0.339975	0.588719	-1.38082	0.563171	0.473964
-0.474153	-0.809447	0.86959	1.08935	-0.440827
0.405986	1.1794	1.02099	0.433464	0.0391491
1.06609	1.31648	2.74867	-0.937138	-0.709714
0.626021	0.283662	0.640461	-2.1792	0.144954
-0.0340838	-2.79287	2.01703	-3.47214	-3.32325
-1.44231	1.32376	-0.0581588	0.736302	-0.301065
-2.78452	0.406611	-0.104209	0.721589	1.54877
-0.914223	-1.01129	-0.0410864	-2.4769	-2.09959
1.06609	0.295309	-0.948842	-0.16555	0.171978
-0.254119	-1.15624	0.865996	0.896538	-0.976744
-1.13426	-0.335484	-0.312897	0.521819	0.600356
-0.0340838	-0.496682	-0.377592	0.703814	0.380765
0.185951	0.0739455	-0.132064	0.87071	-0.137437
-0.826209	-0.434811	-0.923234	0.35167	0.523972
-0.0340838	-1.41361	-1.13888	0.0990846	0.231632
0.405986	0.402067	-0.624692	-0.85863	-0.326396
0.736038	-0.805872	1.22833	-0.566942	-0.121882
0.185951	0.233071	2.15384	-0.426022	-0.78571
-0.254119	0.150374	0.271383	0.772761	0.335452
-0.0340838	-1.65892	0.830503	0.852938	0.672104
-0.386139	-0.588406	-1.66476	0.285351	0.791065
⋮	⋮	⋮	⋮	⋮

Covariance matrix

Calculate the *covariance matrix* (measure of the variance between variables) for the autoscaled data $X'_{ki} = \frac{X_{ki} - \overline{X_i}}{\sigma_i}$ with $\overline{X'_i} = 0$ and $\sigma_{X'_i} = 1$

$$C_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} X'_{ki} X'_{kj} = \frac{1}{N-1} \sum_{k=1}^{N} \frac{\left(X_{ki} - \overline{X_i}\right)\left(X_{kj} - \overline{X_j}\right)}{\sigma'_i \sigma'_j}$$

$$C_{ii} = \frac{1}{N-1} \sum_{k=1}^{N} X'^2_{ki} = \frac{1}{N-1} \sum_{k=1}^{N} \frac{\left(X_{ki} - \overline{X_i}\right)^2}{\sigma'^2_i} = \frac{S_i}{S_i} = 1$$

C is a measure of the correlation between the variables.

Covariance matrix

$\mathbf{C} =$

	Tax	Drivers	Income	logMiles	Fuel
Tax	1	-0.0858	-0.0107	-0.0437	-0.2594
Drivers	-0.0858	1	-0.1760	0.0306	0.4568
Income	-0.0107	-0.1760	1	-0.2959	-0.4644
logMiles	-0.0437	0.0306	-0.2959	1	0.4220
Fuel	-0.2594	0.4568	-0.4644	0.4220	1

This is a more traditional summary of two-variable relationships.

We see what is also apparent in the scatterplot matrix: relatively small correlations between the predictors and *Fuel*, and essentially no correlation between the predictors themselves. In other words, *no single variable explains the variance in the fuel consumption*.

Correlation matrix

We have defined (p = 5, N = 51):

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & x_{14} \\ 1 & x_{21} & x_{22} & x_{23} & x_{24} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & x_{N2} & x_{N3} & x_{N4} \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}$$

$$\Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Please note that numerical instabilities can arise if there are orders of magnitude variations between the data types and you could get incorrect answers.

We will use the matrix equation in this case and will test by comparing to another solution (in the textbook† where you can find this dataset, or in the *alr4* package in R) based on rescaled data.

Matrix form

†Weisberg, S. (2014). *Applied Linear Regression*, third edition. New York: Wiley.

$$\mathbf{M} = \mathbf{X}^T\mathbf{X} = \begin{bmatrix} 51 & 1027.9 & 46087.6 & 1448.6 & 803.003 \\ 1027.9 & 21750. & 927469 & 29185.6 & 16169.7 \\ 46087.6 & 927469. & 4.19137 \times 10^7 & 1.30621 \times 10^6 & 725822. \\ 1448.6 & 29185.6 & 1.30621 \times 10^6 & 42136.6 & 22710.5 \\ 803.003 & 16169.7 & 725822 & 22710.5 & 12753.9 \end{bmatrix}$$

$$\mathbf{M}^{-1} = \begin{bmatrix} 9.02151 & -0.0285205 & -0.00408 & -0.0598114 & -0.193151 \\ -0.0285205 & 0.000978751 & 5.59937 \times 10^{-6} & 0.000042633 & 0.000160232 \\ -0.00408 & 5.59937 \times 10^{-6} & 3.92216 \times 10^{-6} & 0.0000118901 & 5.4018 \times 10^{-6} \\ -0.0598114 & 0.0000426338 & 0.0000118901 & 0.00114276 & 0.00100021 \\ -0.19315 & 0.000160232 & 5.4018 \times 10^{-6} & 0.00100021 & 0.00994784 \end{bmatrix}$$

Note the 7 orders of magnitude difference in the entries in $\mathbf{M}^{-1}$

Solution

Using $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, we find

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 154.193 \\ -4.22798 \\ 0.471871 \\ -6.13533 \\ 18.5453 \end{pmatrix}, \begin{array}{c} \text{(intercept)} \\ \text{Tax} \\ \text{Drivers} \\ \text{Income} \\ \text{logMiles} \end{array}$$

which agrees with solutions based on alternative numerical solutions.

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 Tax + \hat{\beta}_2 Drivers + \hat{\beta}_3 Income + \hat{\beta}_4 logMiles$

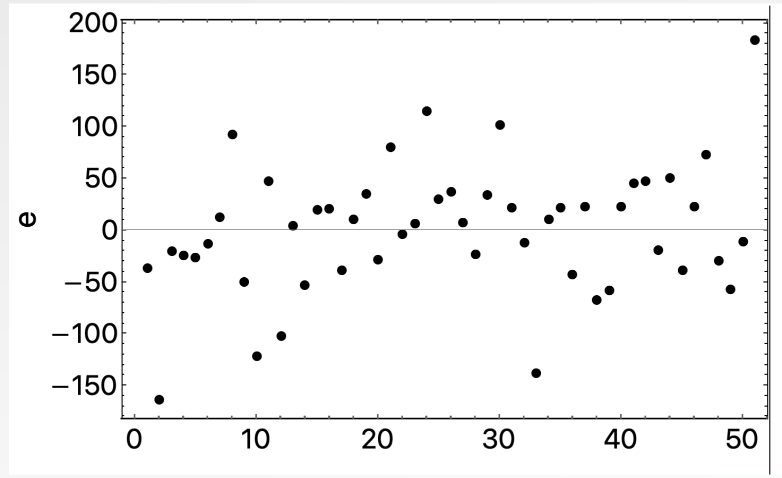
How do we analyze the quality of this solution?

- We define the *residual sum of squares* (*RSS*) [Jobson *SSE* = *error sum of the squares*] to be the value of E, e.g., $RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$, evaluated at the minimum. You can say that this is the residue left over after subtracting the fitted values.

- The variance σ^2 is RSS divided by its *degrees of freedom* (df)

- df = number of cases (data points) minus the number of parameters in the mean function, here $df = N - (1+4) = N - 5$ and the *residual mean square is* $\hat{\sigma}^2 = \frac{RSS}{N-5}$

Residual error over the 51 datapoints

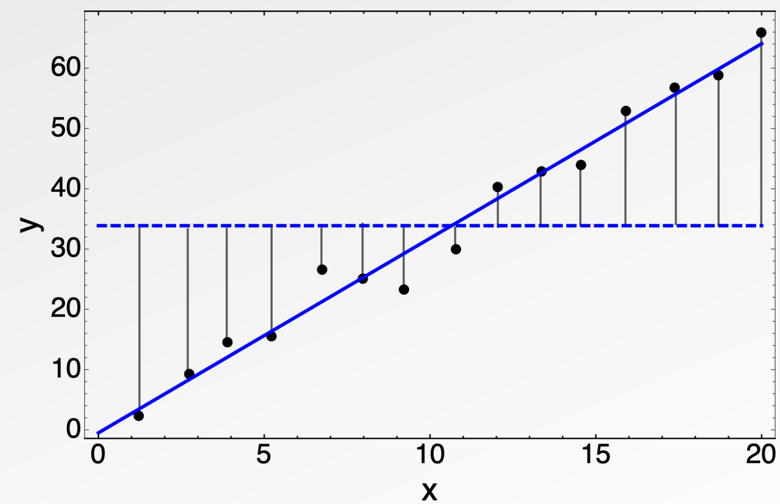


- $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ is often called the *standard error of regression*
- *the smaller RSS, the smaller the residuals*
- *Randomness in the error values is a good thing*

The residual sum of squares (RSS)

- We can also measure the error relative to the mean $\overline{y}$, which is equivalent to using $y(x) = \beta_0$ as our fitting function.
- The OLS estimate would be found by minimizing $\sum_i (y_i - \beta_0)^2$
- The minimum is found with $\hat{\beta}_0 = \overline{y}$ (note difference from value with $\widehat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$)
- The residual sum of squares is [Jobson: *SST = total sum of the squares*]

$$SYY = \sum_{i=1}^{N} (y_i - \overline{y})^2$$



- There is only one parameter, so $df = N - 1$

Another measure of variance

The *sum of squares due to regression*, SS_{reg}, is $SS_{reg} = SYY - RSS$

- the df of SS_{reg} is the df for the mean function $(N-1)$ minus the df for simple regression $(N-(1+p))$, $df = p$. Here, $p = 4$

SS_{reg} is the reduction in the residual sum of square from enlarging the mean function from $\hat{y} = \beta_0 = \bar{y}$ to $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 Tax + \hat{\beta}_2 Drivers + \hat{\beta}_3 Income + \hat{\beta}_4 logMiles$

Consider

$$R^2 = \frac{SS_{reg}}{SYY} = 1 - \frac{RSS}{SYY}$$

R^2 is the *coefficient of multiple regression* and is 1 minus the *fraction* of the data that is unexplained by the variance as described with RSS.

The $\hat{\beta}_i$ are the *coefficients* for the solution, which is generally all that is reported (unfortunately!).

For this dataset, we find:

- $RSS = 193{,}700$
- $\hat{\sigma}^2 = \frac{RSS}{N-5} = 4210.87$
- $SYY = 395{,}694$
- $SSreg = SYY - RSS = 201{,}994.$
- $R^2 = 1 - \frac{RSS}{SYY} = 0.510$

About half of the variance is accounted for using
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 Tax + \hat{\beta}_2 Drivers + \hat{\beta}_3 Income + \hat{\beta}_4 logMiles$

Explore details of how the various quantities are calculated in R from this website:
http://r-statistics.co/Linear-Regression.html

Estimates of variants

When the number of variables (columns-1) is *p* and the number of datapoints (rows) is *n*. In ordinary circumstances we must have more datapoints than variables. If the two are close to each other, however, we need to be careful. If, e.g., *n=p+1* then the fit is exact with no degrees of freedom.

Therefore, we can compute an *adjusted R^2* to account for this:

$$R^2 = 1 - \frac{n-1}{n-p-1}(1-R^2) = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} = 1 - \frac{RSS/(n-p-1)}{SYY/(n-1)}$$

Here, [Jobson] *SSE* is the *error sum of the squares = RSS*
and [Jobson] *SST* is the *total sum of the squares = SYY* (previous slides)

If the two values of R^2 differ by a large amount then check the dimensions of your data matrix because it may signal that you do not have enough data.

Adjusted R^2

We can show that the variances of the coefficients are given by
$Var(\widehat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\mathbf{X}^T\mathbf{X})^{-1}$

The standard errors are $\sqrt{Var(\widehat{\boldsymbol{\beta}})}$, in which the square root is taken of each element.

We find for the coefficients and their standard errors:

$$\begin{array}{c} \text{(intercept)} \\ \text{Tax} \\ \text{Drivers} \\ \text{Income} \\ \text{logMiles} \end{array} \quad \boldsymbol{\beta} = \begin{pmatrix} 154.193 \\ -4.22798 \\ 0.471871 \\ -6.13533 \\ 18.5453 \end{pmatrix} \; stderror(\boldsymbol{\beta}) = \begin{pmatrix} 194.9062 \\ 2.0301 \\ 0.1285 \\ 2.1936 \\ 6.4722 \end{pmatrix} \; t\ value = \frac{\hat{\beta}_k}{stderror(\hat{\beta}_k)} = \begin{pmatrix} 0.791 \\ -2.083 \\ 3.672 \\ -2.797 \\ 2.865 \end{pmatrix}$$

Variance of the coefficients

There are a number of other tests that can be applied to this data.

- F-test
- t-test
- hypothesis testing
- sequential analysis testing
- prediction

All (or most) of these are available in standard statistics packages, such as R.

Other tests

When independent random samples, n_1 and n_2, are taken from two normal distributions with equal variances, the sampling distribution of the ratio of the sample variances, $F = \frac{s_1^2}{s_2^2}$, follows the F distribution:

$F(p, n - p')$, in which $p' = p$ for a function with no intercept and $p' = 1 + p$ if an intercept is included.

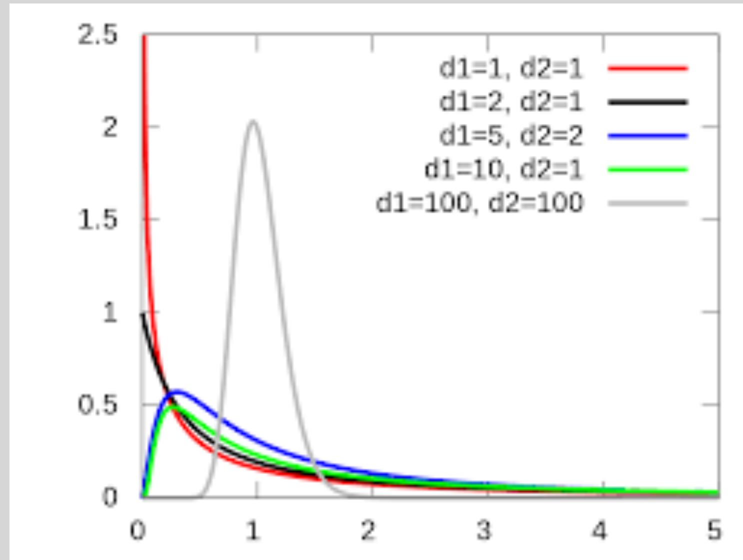
For our case, we will plot the ratio of the mean of $SSreg = SYY - RSS$ and the $\hat{\sigma}^2 = \frac{RSS}{N-5}$, i.e., $F_p = \frac{SSreg/4}{\hat{\sigma}^2} = 11.9924$

We will use the *F-distribution* to verify or disprove the **null hypothesis** that the mean of *Fuel* does not depend on any of the terms in $\hat{y}$.

F-test

Statistics: “F” Distribution

Given two χ^2 random variables χ_N^2 and χ_M^2 then the random ratio variable $(\chi_N^2$ /N)/(χ_M^2/D) has an F distribution with N & D degrees of freedom. These are described, obviously, as the numerator and denominator degrees of freedom, respectively. Like the χ^2, the F distribution is positive and skewed to the right.

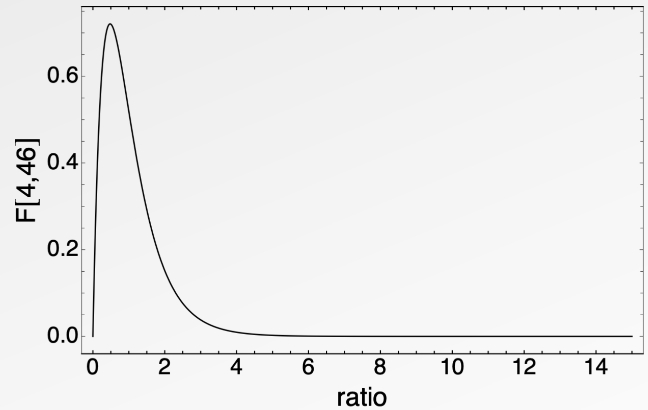


Again, inferences for multiple regression models, analysis of variance models and multivariate means make use of this distribution.

More can be found on p. 20 of Jobson, Vol 1.

The *F-distribution* $F(\{p, N-(1+p)\}, F_p) = F(\{4{,}46\}, F_p)$.

Using the distribution, we find that the probability for $F_p = 11.9924$ that the null hypothesis is true is $PR(>F) = 8.8\times10^{-7}$, which indicates that the null hypothesis is not true
and *Fuel* **does** depend on the terms in $\hat{y}$.



Source	dF	SS	MSS	F	p-value
Regression	p	SS_{reg}	SSreg/1	$MSS_{reg}/\hat{\sigma}^2$	9×10^{-7}
Residual	N-(p+1)	RSS	$\hat{\sigma}^2 = \dfrac{RSS}{N-5}$		
Total	N-1	SYY			

F distribution

Suppose we want to see how important taxes are in determining fuel use.

We already have RSS and $\hat{\sigma}^2$ including tax.

Remove column of tax data from X and solve for $\hat{\beta}_{notax}$ and then calculate RSS_{notax} and $\hat{\sigma}^2_{notax}$ in the same way as before. Note that there are $df = 51 - (1 + 3) = 47$ degrees of freedom for the $notax$ case and 46 for the full solution.

We find:

Source	df	SS	MS	F	p-value
no tax	47	211964			
with tax	46	193700			
difference	1	18264	18264	4.34	0.043

There is modest evidence that the coefficient for *Tax is different from 0.*

Can do with other variables as well.

Effects of specific variables

More generally, the whole point of the statistical tests is all about *hypothesis testing*. The most common approach is the NULL Hypothesis, called H_0, which means that the proposed model being tested can/should be rejected. Before obtaining an answer to such a hypothesis, however, we must choose the probability or confidence level for rejection. A very common choice is *p=0.05*, which says that we are confident to the 95 % level that the NULL hypothesis can be rejected i.e., it’s wrong! You can say that there is only a 5 % that we are making the wrong choice by rejecting the NULL hypothesis and accepting that the model is meaningful. The latter is known as accepting the Alternative Hypothesis, H_1. Type I Error (incorrectly rejecting the NULL) is also known as False Positive and Type II Error (incorrectly rejecting the Alternative) is also known as False Negative.

As remarked on this website, https://data-flair.training/blogs/hypothesis-testing-in-r/, “A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you can reject it. A large p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail by rejecting it.” A standard test is, e.g., the “t test” that can be used to decide if two different samples have the same mean values; in R, this is the `t.test` procedure.

Hypothesis Testing

The lecture on (part of Monday and) Wednesday will be a tutorial on the use of R, which will be useful for the first homework assignment.

It will be given using the usual Zoom link.

The instructor will be Dr. Amit Verma.

Next Lecture