EE2211 : Introduction to Machine Learning (Fall 2020)
Least Squares Estimation and the Projection Theorem

Vincent Y. F. Tan

In this brief document, we will present one interesting example of least squares estimation. We will also unify the concepts of least squares estimation and the least norm solution.

1 Predicting the Outcome of Presidential Elections: An Application of Least Squares Estimation

While most of news this year has been dominated by Covid-19, there is still the little matter of another event that has monumental importance to the world—the quadrennial Presidential Elections in the United States of America, pitting the incumbent President Donald J. Trump (R) against former Vice President Joe Biden (D). We will put what we have learned to do linear regression on four potentially important factors that influence the incumbent's winning (or losing) margin. This analysis is based on Nate Silver's study in November 2011, one year before President Obama beat Mitt Romney in the 2012 elections. The original article, which contains much more analysis, can be found here—`https://fivethirtyeight.blogs.nytimes.com/2011/11/18/which-economic-indicators-best-predict-presidential-elections/`.

We consider four economic indicators:

(a) Real GDP growth rate x_1;

(b) Change in non-farm payrolls x_2;

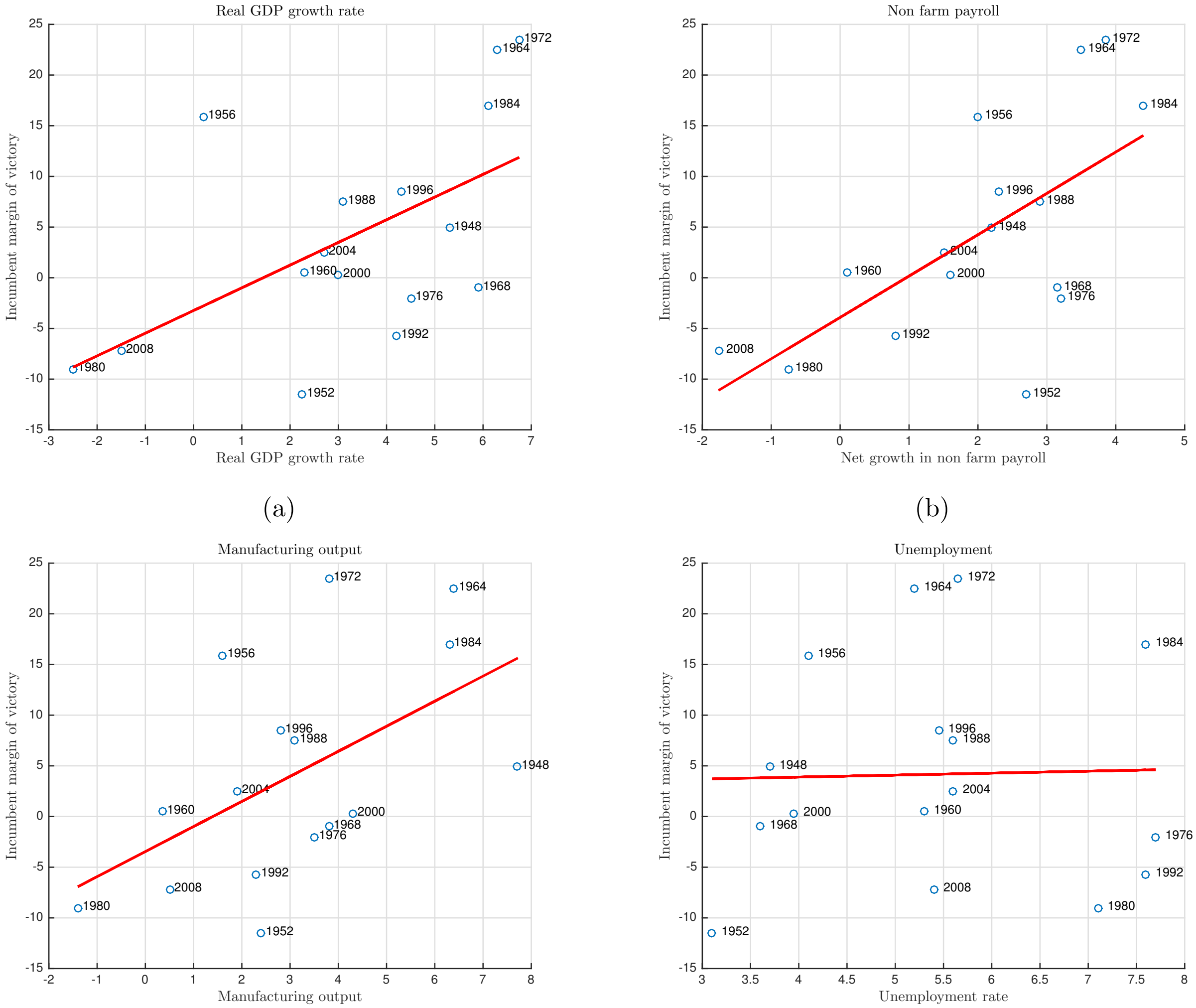
(c) ISM (Institute of Supply Management) manufacturing index x_3;

(d) Unemployment rate x_4.

We consider predicting the target variable—the *incumbent party's margin of victory* y. This can be negative if the incumbent party wins fewer votes than the challenger. For example, in 1996 the incumbent Bill Clinton (D) beat Bob Dole (R) by 8.5% so $y = 8.5$. In 2008, Barack Obama (D) beat John McCain (R) by 7.2% and the incumbent was a Republican (George W. Bush) so $y = -7.2$. Scatter plots of each of the unnormalized or raw indicators against the incumbent party's margin of victory are shown in Fig. 1. We have data of all the above economic indicators and margins of victory from 1948 (Truman against Dewey) to 2008 (Obama against McCain), constituting 16 presidential elections. Do note that even if a candidate wins the popular vote, s/he may not be elected president as one needs to win the so-called electoral college instead of the popular vote. For example, George W. Bush became president-elect in 2000 even though he lost the popular vote to Al Gore and more recently, Donald J. Trump became president-elect in 2016 even though he lost the popular vote (by 3 million votes) to Hillary R. Clinton.

We do a min-max normalization of the features to ensure that all of them are in $[0,1]$. We include the offset or bias term to obtain our design matrix $\mathbf{X}$ and target vector $\mathbf{y}$, i.e.,

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} \\ 1 & x_{2,1} & x_{2,2} & x_{2,3} & x_{2,4} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{16,1} & x_{16,2} & x_{16,3} & x_{16,4} \end{bmatrix} \in [0,1]^{16\times 5} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{16} \end{bmatrix} \in \mathbb{R}^{16}. \tag{1}$$



(a)

(b)

(c)

(d)

Figure 1: Scatter plots of incumbent party's victory margin against various economic factors

Hence, $x_{2,2}$ corresponds to the change in non-farm payrolls of the 1952 election (Eisenhower vs Stevenson).

In Fig. 2, we show the regression lines for each feature against y. For example, for Fig. 2(a), we only consider the matrix $\mathbf{X}(:,1:2) \in [0,1]^{16,2}$ (first two columns of $\mathbf{X}$) and learn the vector

$$\hat{\mathbf{w}}_1 = \left(\mathbf{X}(:,1:2)^\top \mathbf{X}(:,1:2)\right)^{-1} \mathbf{X}(:,1:2)^\top \mathbf{y}. \tag{2}$$

From the plots, we visually see that the first three economic indicators seem to have correlation with y. Unemployment rate, somewhat surprisingly, has minimal impact on y. In statistics, a common measure of linear dependence is the coefficient of determination or R^2 value. This is the proportion of the variance in the dependent variable y that is predictable from the independent variable x_i. You do not need to know what the formula is or how to use it but all that is needed to appreciate is that the closer the value is to one, the larger the linear dependence. The R^2 values for each of the variables is tabulated in Table 1. Again the values confirm that the first three independent variables are rather correlated to y.

Now, let us use the variables x_1, x_2 and x_4 as our variables to do linear regression, omitting x_3 (Manufacturing output) for simplicity. The design matrix is thus $\mathbf{X}' = [\mathbf{X}(:,1:3)\ \mathbf{X}(:,5)] \in [0,1]^{16\times 4}$ (we omit the

Variable	GDP Growth Rate	Non-Farm Payroll	Manufacturing Output	Unemployment Rate
R^2	0.3134	0.3984	0.2892	0.0007

Table 1: R^2 for each of the economic indicators

4th column corresponding to x_3). Then we estimate

$$\hat{\mathbf{w}}_{1,2,4} = \big((\mathbf{X}')^\top \mathbf{X}'\big)^{-1}(\mathbf{X}')^\top \mathbf{y} = \begin{bmatrix} -12.5102 \\ 4.1531 \\ 21.5454 \\ 1.9871 \end{bmatrix}. \tag{3}$$

So the offset term if -12.51 and the coefficients associated to GDP growth rate, change in non-farm payroll and unemployment rate are $4.15, 21.55$ and 1.99 respectively. Again, we see that the last feature seems to have a small effect due to the small coefficient.

Suppose we want to predict the winning margin in 2004 (the 15th data sample). So as not to bias our model, we remove this data point from $\mathbf{X}' \in [0,1]^{16\times 4}$ and $\mathbf{y} \in \mathbb{R}^{16}$. Call the new design matrix and dependent vector $\mathbf{X}'' \in [0,1]^{15\times 4}$ and $\mathbf{y}'' \in \mathbb{R}^{15}$ respectively. The coefficient vector learned in the absence of the 2004 data point is

$$\hat{\mathbf{w}}''_{1,2,4} = \big((\mathbf{X}'')^\top \mathbf{X}''\big)^{-1}(\mathbf{X}'')^\top \mathbf{y}'' = \begin{bmatrix} -12.5310 \\ 4.1519 \\ 21.5617 \\ 1.9822 \end{bmatrix}. \tag{4}$$

Great! The removal of one data example did not affect the regression coefficients too much. Now let's try to predict the winning margin in 2004. We have

$$\hat{y}_{15} = \underbrace{\begin{bmatrix} 1 \\ 0.5622 \\ 0.5285 \\ 0.5435 \end{bmatrix}^\top}_{\text{2004's normalized feature vector}} \hat{\mathbf{w}}''_{1,2,4} = 2.2748\% \tag{5}$$

From the economic indicators, this means that George W. Bush (the incumbent in 2004) is expected to win by 2.27%. In actual fact, he won the popular vote by 2.4%, which means that our model did pretty well! Some natural questions.

1. It seems from Fig. 1(d) that the unemployment rate is not a very good predictor of the incumbent's margin of victory. What if we removed it from the model? Do you expect predictions to improve? Try it for yourself using the csv file, which is provided with this tutorial.

2. What if we included manufacturing output as a variable in the model? Do you expect predictions to improve?

3. Why was our prediction of 2004's result in (5) so good? From Fig. 1, we see that the data for 2004 are very close to the individual regression lines. What if we tried to predict a result in a very "atypical" year, e.g., 1956?

4. Redo everything for the electoral college to predict the president-elects (which is arguably more important). Before collecting results and doing any machine learning, do you expect your results to be more or less accurate than predicting the winning margins?

5. We predicted a result that is known (2004's result), which is not so interesting. Try predicting Donald Trump's winning or losing margin against Joe Biden this Fall. For this, you need to know how and where to get reliable data. Let me know if you find something interesting!

2 Projection Theorem (Advanced and Optional)

In this section, we unify the solutions of the least squares estimator and the least norm solution via the Projection Theorem.

Theorem 1. *Let $\mathcal{M} \subset \mathcal{V}$ be a subspace of a vector space $\mathcal{V}$. Then the solution $\hat{\mathbf{m}}$ to the optimization problem*

$$\textit{minimize} \quad \|\mathbf{y} - \mathbf{m}\| \quad \textit{subject to} \quad \mathbf{m} \in \mathcal{M} \tag{6}$$

is unique and satisfies

$$(\mathbf{y} - \hat{\mathbf{m}}) \perp \mathcal{M}. \tag{7}$$

Conversely, if $\hat{\mathbf{m}}$ satisfies (7)*, it is the optimal solution to* (6)*.*

This means that the optimal solution $\hat{\mathbf{m}}$ is such that the induced "error vector" $\mathbf{y} - \hat{\mathbf{m}}$ is orthogonal to the any vector that lies in the subspace $\mathcal{M}$. The proof is given in the appendix.

2.1 Application to Least Squares Estimation

Let us see how to apply this to the least squares estimation problem, a problem considered and solved by Legendre and Gauss. In the previous set of notes, we saw how to do this by means of differentiating functions of vectors. It turns out that you neither need to remember nor derive these formulate—i.e., differentiation is not necessary. Recall that in the least squares problem, we are given a matrix $\mathbf{X} \in \mathbb{R}^{m\times d}$ with full column rank in which $m > d$ and a vector $\mathbf{y} \in \mathbb{R}^d$ and considered the problem

$$\text{minimize} \quad \|\mathbf{y} - \mathbf{X}\mathbf{w}\|. \tag{8}$$

In other words, we want to find $\mathbf{w}$ in the range of $\mathbf{X}$ such that $\mathbf{X}\mathbf{w}$ is closest to $\mathbf{y}$. By the Projection Theorem, $\hat{\mathbf{w}}$ is optimal if and only if

$$(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) \perp \mathcal{R}(\mathbf{X}). \tag{9}$$

Note that $\mathbf{X}\hat{\mathbf{w}}$ plays the role of $\hat{\mathbf{m}}$ in the Projection Theorem. However, (9) means that

$$\mathbf{x}_i^\top(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) = 0, \qquad \forall\, i = 1, 2, \ldots, d, \tag{10}$$

where $\{\mathbf{x}_i\}_{i=1}^d \subset \mathbb{R}^m$ is the set of d columns of $\mathbf{X}$. Stacking the conditions in (10) together in matrix form, we obtain

$$\begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_d^\top \end{bmatrix} (\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) = \mathbf{0} \quad \Longleftrightarrow \quad \mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) = \mathbf{0} \quad \Longleftrightarrow \quad \hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}. \tag{11}$$

Note that the inverse of $\mathbf{X}^\top\mathbf{X}$ exists because $\mathbf{X}$ has full column rank. This recovers the least squares solution without knowledge of differentiation.

2.2 Application to the Least Norm Problem

Recall that in the least norm problem, we are given a $\mathbf{X} \in \mathbb{R}^{m\times d}$ with full row rank matrix in which $m < d$ and a vector $\mathbf{y} \in \mathbb{R}^d$ and considered the problem

$$\text{minimize} \quad \|\mathbf{w}\| \quad \text{subject to} \quad \mathbf{X}\mathbf{w} = \mathbf{y}. \tag{12}$$

By the Rouché-Capelli Theorem, we see that since $\text{rank}(\mathbf{X}) = \text{rank}([\mathbf{X}\ \mathbf{y}]) = m < d$, there are infinitely many solutions to the system $\mathbf{X}\mathbf{w} = \mathbf{y}$. Let $\mathbf{w}_{\text{p}} := \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y}$ be one of the solutions (check that this is indeed a solution!). By writing $\mathbf{w} = \mathbf{w}_{\text{p}} - \mathbf{w}_0$ for some $\mathbf{w}_0$, above optimization problem can be rewritten as

$$\text{minimize} \quad \|\mathbf{w}_{\text{p}} - \mathbf{w}_0\| \quad \text{subject to} \quad \mathbf{X}(\mathbf{w}_{\text{p}} - \mathbf{w}_0) = \mathbf{y} \tag{13}$$

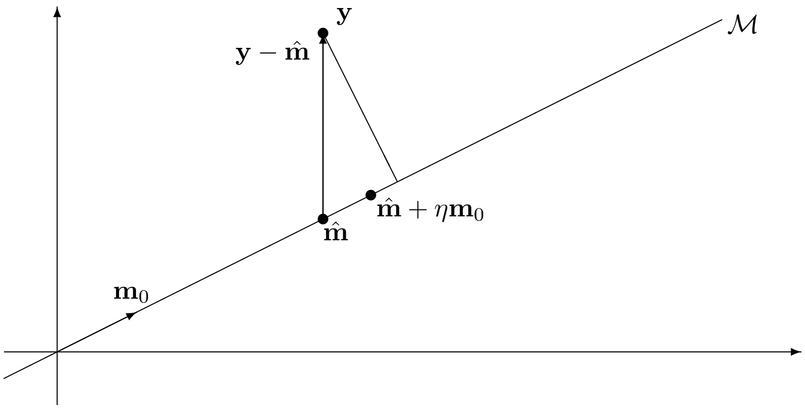

$\mathbf{y}$
$\mathcal{M}$
$\mathbf{y}-\hat{\mathbf{m}}$
$\hat{\mathbf{m}}+\eta\mathbf{m}_0$
$\hat{\mathbf{m}}$
$\mathbf{m}_0$

Figure 2: Illustration of the proof of the Projection Theorem. If $\mathbf{y}-\hat{\mathbf{m}}$ is not orthogonal to the subspace $\mathcal{M}$, shift $\hat{\mathbf{m}}$ in the direction $\mathbf{m}_0$ by an "amount" η. The distance to $\mathbf{y}$ will be reduced. In this figure, $\eta = (\mathbf{y}-\hat{\mathbf{m}})^\top \mathbf{m}_0 > 0$ as the angle between $\mathbf{y}-\hat{\mathbf{m}}$ and $\mathbf{m}_0$ is acute.

where the optimization variable is now $\mathbf{w}_0$. Our objective is to show that the optimal $\mathbf{w}_0 = \mathbf{0}$, which would validate that $\mathbf{w}_\mathrm{p}$ is the least norm solution. Since $\mathbf{X}\mathbf{w}_\mathrm{p} = \mathbf{y}$, the optimization problem in (13) is the same as

$$\text{minimize} \quad \|\mathbf{w}_\mathrm{p} - \mathbf{w}_0\| \quad \text{subject to} \quad \mathbf{X}\mathbf{w}_0 = \mathbf{0}. \tag{14}$$

This means that

$$\text{minimize} \quad \|\mathbf{w}_\mathrm{p} - \mathbf{w}_0\| \quad \text{subject to} \quad \mathbf{w}_0 \in \mathcal{N}(\mathbf{X}). \tag{15}$$

By the Projection Theorem,

$$(\mathbf{w}_\mathrm{p} - \mathbf{w}_0) \perp \mathcal{N}(\mathbf{X}). \tag{16}$$

Since

$$\mathbf{w}_0 \in \mathcal{N}(\mathbf{X}) \quad \text{and} \quad \mathbf{w}_\mathrm{p} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y} \tag{17}$$

we have

$$0 \overset{(16)}{=} (\mathbf{w}_\mathrm{p} - \mathbf{w}_0)^\top \mathbf{w}_0 \overset{(17)}{=} (\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y} - \mathbf{w}_0)^\top \mathbf{w}_0 \overset{(17)}{=} \mathbf{y}^\top \left((\mathbf{X}\mathbf{X}^\top)^{-1}\right)^\top \underbrace{\mathbf{X}\mathbf{w}_0}_{=\mathbf{0}} - \mathbf{w}_0^\top \mathbf{w}_0. \tag{18}$$

Thus, $\|\mathbf{w}_0\|^2 = 0$ which means that $\mathbf{w}_0 = \mathbf{0}$ as desired. This proof uses the Projection Theorem in (16) to show that the minimum norm solution $\mathbf{w} = \mathbf{w}_\mathrm{p} - \mathbf{w}_0$ of $\mathbf{X}\mathbf{w} = \mathbf{y}$ is such that $\mathbf{w}_0 = \mathbf{0}$.

Appendix

Proof of Projection Theorem. Assume that $\mathcal{M} \neq \{\mathbf{0}\}$ otherwise the claim is immediate. Suppose, to the contrary, there exists some $\mathbf{m}_0 \in \mathcal{M} \setminus \{\mathbf{0}\}$ of unit norm such that $(\mathbf{y}-\hat{\mathbf{m}})^\top \mathbf{m}_0 =: \eta \neq 0$. Then, we claim that the vector $\hat{\mathbf{m}} + \eta\mathbf{m}_0 \in \mathcal{M}$ yields a better solution in the sense that its distance from $\mathbf{y}$ is strictly smaller. (Why can we assume $\|\mathbf{m}_0\| = 1$ and why is $\hat{\mathbf{m}} + \eta\mathbf{m}_0$ in the subspace $\mathcal{M}$?)

Consider,

$$\begin{aligned}
\|\mathbf{y} - (\hat{\mathbf{m}} + \eta\mathbf{m}_0)\|^2 &= \|(\mathbf{y}-\hat{\mathbf{m}}) - \eta\mathbf{m}_0\|^2 && (19)\\
&= \|\mathbf{y}-\hat{\mathbf{m}}\|^2 - 2\eta(\mathbf{y}-\hat{\mathbf{m}})^\top\mathbf{m}_0 + \eta^2\|\mathbf{m}_0\|^2 && (20)\\
&\overset{(a)}{=} \|\mathbf{y}-\hat{\mathbf{m}}\|^2 - 2\eta^2 + \eta^2\|\mathbf{m}_0\|^2 && (21)\\
&\overset{(b)}{=} \|\mathbf{y}-\hat{\mathbf{m}}\|^2 - 2\eta^2 + \eta^2 < \|\mathbf{y}-\hat{\mathbf{m}}\|^2, && (22)
\end{aligned}$$

where (a) is due the definition of η and (b) is because $\mathbf{m}_0$ has unit norm. Thus, we have found a vector in $\mathcal{M}$, namely $\hat{\mathbf{m}} + \eta\mathbf{m}_0$, that has strictly smaller distance to $\mathbf{y}$, contradicting the optimality of $\hat{\mathbf{m}}$. See Fig. 2 for an illustration of this proof. $\square$