Multiple linear regression

Chrysafis Vogiatzis

Lecture 32

Learning objectives

After lecture 32, we will be able to:

- Recall the ANOVA identity.
- Recall and use the R^2 and R^2_{adj} parameters to evaluate how good a regression is.
- Understand when to use and how to apply regression with multiple independent variables.
- Derive, use, and interpret the results of the last squares line for multiple independent variables.
- Perform hypothesis testing on multiple parameters of the least squares line.

Motivation: Maintenance fees

What happens when we are trying to derive a (linear) relationship between one dependent variable y and multiple $k > 1$ different independent variables x_j? Well, in that case, we need multiple different parameters (slopes), one for each independent variable!

For example, what is the linear relationship between the maintenance fees (costs y) of a bank as a function of the number of the new applications (x_1) and the number of outstanding loans (x_2)?

Motivation: realtor.com

Taken from , here are 8 recently (August 2019) sold homes in Urbana:

	Sq. ft.	Year built	Garages	#bedrooms	#bathrooms	Price
1	1547	1950	1	3	3	158500
2	1834	1957	0	4	2	183000
3	2520	1980	3	5	2.5	233000
4	985	1911	1	2	1	69000
5	1275	1968	0	3	1.5	118000
6	2337	1977	2	5	2	249900
7	1880	1967	2	3	2	175000
8	1943	1965	1	4	2.5	169900

Which one of the five predictor variables (sq. ft., year built, garages, #bedrooms, #bathrooms) is the least important for predicting price?

The ANOVA identity

Let us begin with an example of the calculations we will see in this section. During the previous lecture, we saw an example that led us to a regression line of $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 7.381 + 0.273 \cdot x$

For the given data (again check the previous lecture for all the details of this example), we finally got:

x	y	$\hat{y}$
1	7.6	7.654
9	10.24	9.838
2	7.3	7.927
7	8.97	9.292
6	8.74	9.019
7	8.99	9.292
8	9.93	9.565
1	8.47	7.654

We used that table to calculate SS_E (**the sum of squares of the error**) as:

$$SS_E = \sum (y_i - \hat{y}_i)^2 = (7.6 - 7.654)^2 + (10.24 - 9.838)^2 + \ldots = 1.629.$$

This would eventually be divided by $8 - 2 = 6$ degrees of freedom to estimate the **mean square error** (MS_E).

In a similar manner, we may define **total sum of squares** as the sum of squares of the differences between each *observed* value y_i versus the expectation:

$$SS_T = \sum (y_i - \overline{y})^2 .$$

We may also define the **regression sum of squares** as the sum of squares of the differences between each *fitted* value $\hat{y}_i$ versus the expectation:

$$SS_R = \sum (\hat{y}_i - \overline{y})^2 .$$

We then claim that:

$$\boxed{SS_T = SS_E + SS_R}$$

This is called the **Analysis of Variance** (ANOVA) identity and it is immensely useful when analyzing how good our regression is.

Using the ANOVA identity

In this example, we have already calculated the sum of squares of errors SS_E to be equal to 1.629. How about the total sum of squares SS_T and the regression sum of squares SS_R?

First, begin by calculate the average y value as

$$\overline{y} = \frac{\sum y_i}{n} = \frac{7.6 + 10.24 + \ldots + 8.47}{8} = \frac{70.24}{8} = 8.78.$$

Then:

- $SS_T = \sum (y_i - \overline{y})^2 = (7.6 - 8.78)^2 + (10.24 - 8.78)^2 + \ldots + (8.47 - 8.78)^2 = 7.2148$.
- Using the ANOVA identity: $SS_T = SS_R + SS_E \implies SS_R = SS_T - SS_E = 7.2148 - 1.629 = 5.5858$.

Note how we could have derived SS_R by applying the formula and getting that $SS_R = \sum (\hat{y}_i - \overline{y})^2 = (7.654 - 8.78)^2 + (9.838 - 8.78)^2 + \ldots + (7.654 - 8.78)^2 = 5.5858$, which is the same result.

We now proceed to define an easy to compute parameter that helps us estimate the quality of our regression line.

The R^2 parameter

We want to somehow quantify how "good" a regression is. We would like to establish some coefficient that tells us how closely our predictions $\hat{y}$ follow the real data (y). We call that parameter R^2 and allow it to be in $[0, 1]$m where a value of 1 implies that all data points fall on the regression line. Of course, we would like high values of R^2 and we hope that they imply a good fit of the regression line. Formally:

Definition 1 *R^2 is a measure of how much of the variability is accounted for by the regression model and is calculated as:*

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

Recall that:

***total**: $SS_T = \sum (y_i - \overline{y})^2$.*

***error**: $SS_E = \sum (y_i - \hat{y}_i)^2$.*

regression: $SS_R = \sum (\hat{y}_i - \overline{y})^2$.

R^2 calculations

What is the R^2 coefficient for the previous regression?

We have two ways to calculate it!

- $R^2 = \frac{SS_R}{SS_T} = 5.58587.2148 = 0.774$.
- $R^2 = 1 - \frac{SS_E}{SS_T} = 1 - \frac{1.629}{7.2148} = 0.774$.

So, *how high is good enough* for R^2? The answer is that (as so many other things that we have seen) "it depends!" We'll take another look at it (and an adjusted version) shortly.

Multiple linear regression

We now move to more than just one independent variable x. This should make sense, as in most practical cases our "future" depends on more than just one piece of information:

- Success in an exam is not only how much you've studied, but also a function of your physical and mental health, how well rested you are, luck, etc.
- The box office success of a movie is not only how good the movie is, but how much budget they've had for advertising, the recognition of the names starring and directing, etc.
- Any more examples?

Let us begin easy with just two predictor variables x_1, x_2. We need to extend our definitions from the simple case:

- We now have a triple [1] (x_{i1}, x_{i2}, y_i), $i = 1, \ldots, n$, that is a series of n data points with provided values for x_1, x_2, y.
- The main idea is still the same!

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

where:

  - β_0 is the intercept intercept;
  - β_1, β_2 are the slopes for x_1, x_2, respectively;
  - ϵ_i is the "noise" associated with point i.

[1] Contrast with the pair (x_i, y_i) earlier.

- Hence our goal is to find the "best" $\beta_0, \beta_1, \beta_2$ by optimizing the least squares function:

$$L = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2}\right)^2.$$

How to derive a solution here? Like earlier, we can take the proper derivatives and set them to zero! How many derivatives, though? Well, in this case, we need to take three derivatives:

$$\begin{aligned}
\frac{\partial L}{\partial \beta_0} = 0 &\implies -2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}\right) = 0 \\
\frac{\partial L}{\partial \beta_1} = 0 &\implies -2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}\right) x_{i1} = 0 \\
\frac{\partial L}{\partial \beta_2} = 0 &\implies -2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}\right) x_{i2} = 0
\end{aligned}$$

Or, simplifying:

$$\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} &= \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} &= \sum_{i=1}^{n} y_i x_{i1} \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i2} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} x_{i2} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}^2 &= \sum_{i=1}^{n} y_i x_{i2}
\end{aligned}$$

This is a system of equations with three unknowns and three equations; solvable under certain conditions. However, it is much more easily expressed in matrix form, no? Let us go back to the original regression line equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i.$$

Written in **matrix form**, we have:

$$y = X\beta + \epsilon$$

- $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$, $\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$

Once more, we wish to find $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ such that

$$L = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i 1 - \beta_2 x_{i2}\right)^2 = (y - X\beta)^T (y - X\beta)$$

is *minimized*. We may rewrite L as:

$$\begin{aligned} L &= (y - X\beta)^T (y - X\beta) = \\ &= y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta = \\ &= y^T y - 2\beta^T X^T y + \beta^T X^T X\beta \end{aligned}$$

We need to take the derivative as far as vector β is concerned:

$$\frac{\partial L}{\partial \beta} = 0 \implies -2X^T y + 2X^T X\beta = 0 \implies X^T X\beta = X^T y.$$

This last equality can be solved by taking the inverse $\left(X^T X\right)^{-1}$ and multiplying on the left [2] to obtain:

$$\boxed{\hat{\beta} = \left(X^T X\right)^{-1} X^T y}$$

[2] Why is that? Well, recall that $Ax = b$ can be solved as $x = A^{-1}b$, when matrix A is invertible!

Overall, we have shown that *in general* (not only for two predictor variables, but for as *many* as we would like to), we have $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$, which can be used

- in matrix form:
$$\hat{y} = X\hat{\beta},$$
- or in scalar form:
$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}, \quad \text{for all } i = 1, \ldots, n.$$

Like in simple linear regression $e_i = \hat{y}_i - y_i$ is the residual/error for each observation i.

Bank maintenance fee prediction

A small bank is hypothesizing that a lot of the fees they pay have to do with the number of loan applications they process every month as well as the number of outstanding loans they have going on. More specifically, they have collected data over the last 16 months that are presented in the following table.

What is the regression line they should use? How much money should they budget for their maintenance costs if they expect 100 applications and 13 outstanding loans this coming January?

Bank maintenance fee prediction

# Applications	# Outstanding	Cost
80	8	2256
93	9	2340
100	10	2426
82	12	2293
90	11	2330
99	8	2368
81	8	2250
96	10	2409
94	12	2364
93	11	2379
97	13	2440
95	11	2364
100	8	2404
85	12	2317
86	9	2309
87	12	2328

First, build matrix X and calculate $\left(X^T X\right)^{-1}$:

$$X = \begin{bmatrix} 1 & 93 & 9 \\ 1 & 100 & 10 \\ 1 & 82 & 12 \\ 1 & 90 & 11 \\ 1 & 99 & 8 \\ 1 & 81 & 8 \\ 1 & 96 & 10 \\ 1 & 94 & 12 \\ 1 & 93 & 11 \\ 1 & 97 & 13 \\ 1 & 95 & 11 \\ 1 & 100 & 8 \\ 1 & 85 & 12 \\ 1 & 86 & 9 \\ 1 & 87 & 12 \end{bmatrix}, \quad \left(X^T X\right)^{-1} = \begin{bmatrix} 14.176 & -0.130 & -0.223 \\ -0.130 & 1.429 \cdot 10^{-3} & -4.764 \cdot 10^{-5} \\ -0.223 & -4.764 \cdot 10^{-5} & 2.222 \cdot 10^{-2} \end{bmatrix}.$$

Finally, we calculate $X^T y$:

$$X^T y = \begin{bmatrix} 37577 \\ 3429550 \\ 385562 \end{bmatrix}.$$

Bank maintenance fee prediction

Combining all we get:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y = \begin{bmatrix} 1566.077 \\ 7.62 \\ 8.58 \end{bmatrix}.$$

This in turn gives us the regression line as:

$$\hat{y} = 1566.077 + 7.62 \cdot \#\text{new loans} + 8.58 \cdot \#\text{loans outstanding}.$$

For January then, we should expect to pay:

$$\hat{y}_{Jan} = 1566.077 + 7.62 \cdot 100 + 8.58 \cdot 13 = 2439.62.$$

The question we should be thinking about at this point: *does the ANOVA identity still hold?* And how can we use that to do hypothesis testing for the regression significance? While we are at it, what does regression significance mean for more than one predictor variables? Let us go ahead and answer all of these questions in the remainder of the lecture.

The ANOVA identity still holds:

$$SS_T = SS_R + SS_E.$$

Each of the three sum of squares is calculated the same way as before. The difference lies with the degrees of freedom:

- SS_T: $n-1$ degrees of freedom [3].
- SS_R: k degrees of freedom.
- SS_E: $n-k-1$ degrees of freedom [4].

Due to that, the mean squares are changed and are now equal to:

- MS_T: $\frac{SS_T}{n-1}$.
- MS_R: $\frac{SS_R}{k}$.
- MS_E: $\frac{SS_E}{n-k-1}$.

Now, back to the derivations from the previous class. We wanted to come up with an estimate for the (unknown!) noise standard deviation σ. We came up with:

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-2}.$$

[3] The same as before.

[4] Different, as we are now estimating $k+1$ parameters. What are those? They are the regression line intercept and slopes: $\beta_0, \beta_1, \ldots, \beta_k$.

Hopefully, you see where we are going with this: our MS_E is different, but other than that the derivation holds. Hence we estimate this standard deviation as:

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-k-1},$$

where the sum of squares of error is calculated as $SS_E = \sum (y_i - \hat{y}_i)^2$ or, in matrix form, as $SS_E = y^T y - \hat{\beta}^T X^T y$.

On to the significance of the regression. Recall that for a single predictor variable our significance testing was easy: either $\beta_1 = 0$ (the slope was zero, and hence insignificant) or not (the slope was nonzero and hence it is significant). When dealing with more than just one predictor variable, though, then all of them need to have zero slopes for the regression to be insignificant! This leads us to:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0, \text{ for at least one } j.$$

We now make the observation that **if the null hypothesis is true**, then the mean squares of the regression and the error are distributed following a χ^2 distribution, each with their own degrees of freedom:

- $SS_R/\sigma^2 \sim \chi^2_k$, where $SS_R = \sum (\hat{y}_i - \overline{y})^2$
- $SS_E/\sigma^2 \sim \chi^2_{n-k-1}$, where $SS_E = \sum (y_i - \hat{y}_i)^2$.

We are then comparing two population "variances" (for MS_R and MS_E) and the test statistic for that is:

$$F_0 = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{MS_R}{MS_E}$$

The rejection area is if $F_0 > f_{\alpha,k,n-k-1}$. Some software will also return a P-value, and the rejection criterion is simply whether P-value $< \alpha$.

Multiple linear regression significance

Null hypothesis:	Test statistic:	Distribution:
$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0.$	$F_0 = \frac{MS_R}{MS_E}.$	$F_0 \sim F_{k,n-k}.$

H_1	Rejection region
At least one $\beta_j \neq 0$	$F_0 > f_{\alpha,k,n-k-1}$

Finally, recall R^2: we have some unfinished business. We already defined $R^2 = 1 - \frac{SS_E}{SS_T}$. We make two observations about it:

- Observation #1: R^2 will always increase or stay the same with the addition of any predictor variable.
- Observation #2: This happens even when that predictor variable is associated with a β_j that is insignificant (i.e., the slope is zero).

We hence define an *adjusted R^2 model*, called R^2_{adj}, that will **penalize more complex regressions** (that is, the use of more predictor variables). Its definition?

$$R^2_{adj} = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)}.$$

Note how adding more predictor variables will lead to a bigger numerator in the fraction which in turn will cause R^2_{adj} to go down.

We claim that this adjusted version is more appropriate than the simple version of R^2. Why? Well, primarily because it does not necessarily increase with the addition of new predictor variables, and thus will not favor more complex models. Indeed, it will many times decrease when an insignificant variable is entered. When R^2 and R^2_{adj} differ by a lot, this is an indication that insignificant terms have been added.

Let us put these things to the test in an example on the regression line we got earlier in the bank example.

Testing significance

In the previous bank example, we already found the line as

$$\hat{y} = 1566.077 + 7.62 \cdot \#\text{new loans} + 8.58 \cdot \#\text{loans outstanding}.$$

Is the regression significant using $\alpha = 0.05$? What is R^2 and how does it compare with R^2_{adj}?
We begin with the calculations of the sum of squares:

- $SS_E = \sum_{i=1}^{16} (y_i - \hat{y}_i)^2 = 3479$
- $SS_R = \sum_{i=1}^{16} (\hat{y}_i - \overline{y})^2 = 44157$
- Using ANOVA, $SS_T = SS_R + SS_E = 47636$.

Now, on to calculate the ratio of the two mean squares:

$$F_0 = \frac{MS_R}{MS_E} = \frac{SS_R/2}{SS_E/13} = 82.5$$

Compared to $f_{\alpha,k,n-k-1} = f_{0.05,2,13} = 3.81$, we overwhelmingly reject. The regression is significant! Let us look at the two R^2 parameter calculations:

- $R^2 = 1 - \frac{SS_E}{SS_T} = 1 - 3479/44157 = 0.921$.
- $R^2_{adj} = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)} = 0.916$.

Note how close the two values are, an indication that no insignificant terms have been added.

What if we were interested in each individual coefficient one-by-one? That is, what if we wanted to check whether the number of new loans is significant; or whether the number of outstanding loans is significant? First of all, let us address why this is not the same question as the one we saw how to address earlier.

Consider a regression with k predictor variables: $k-1$ of them could be insignificant, and one of them could be very significant! Then, the regression as a whole is also significant. Because of that, it is a different question whether the whole regression is significant compared to whether each individual independent variable is significant.

So, if we are interested in whether a single variable is significant or not, this reverts back to checking whether the corresponding slope is zero or not.

$$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0.$$

The test statistic is the same as for simple linear regression:

$$T_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \cdot C_{jj}}}$$

- where C_{jj} is the j-th [5] diagonal element of $\left(X^T X\right)^{-1}$,
- and $\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-k-1}$.

[5] We assume here that the first row and first column element is C_{00}, i.e., we start counting from zero.

Finally, reject if $|T_0| > t_{\alpha/2,n-k-1}$. Note how the main difference from the simple linear regression to the multiple linear regression comes in the form of C_{jj} which replaces S_{xx}. [6] Let us put this to the test right away.

[6] See Lecture 30-31 for details on S_{xx}.

Multiple linear regression term single significance

Null hypothesis:	Test statistic:	Distribution:
$H_0 : \beta_j = 0.$	$T_0 = \frac{\hat{\beta}_j}{\sqrt{MS_E \cdot C_{jj}}}.$	$T_0 \sim T_{n-k-1}.$

H_1	Rejection region	CI region
$\beta_j \neq 0$	$\lvert T_0 \rvert > t_{\alpha/2,n-k-1}$	$\left[\hat{\beta}_j - t_{\alpha/2,n-k-1}\sqrt{MS_E \cdot C_{jj}}, \hat{\beta}_j + t_{\alpha/2,n-k-1}\sqrt{MS_E \cdot C_{jj}}\right]$

Testing significance one-by-one

Let us go back to the banking example from earlier. We already have that the regression line can be written as

$$\hat{y} = 1566.077 + 7.62 \cdot \#\text{new loans} + 8.58 \cdot \#\text{loans outstanding}.$$

- Is the number of new loans significant?
- Is the number of loans outstanding significant?

Use $\alpha = 0.05$. Recall that we already know that the regression is significant; again, though, this does not necessarily imply that both of them are significant!

Testing significance one-by-one

First of all, recall that:

- $SS_E = \sum_{i=1}^{16} (y_i - \hat{y}_i)^2 = 3479$
- $\hat{\sigma}^2 = MS_E = \frac{SS_E}{13} = 267.62.$

For $\hat{\beta}_1$ (number of new loans):

- We have $\left(X^T X\right)^{-1} = \begin{bmatrix} 14.176 & -0.130 & -0.223 \\ -0.130 & 1.429 \cdot 10^{-3} & -4.764 \cdot 10^{-5} \\ -0.223 & -4.764 \cdot 10^{-5} & 2.222 \cdot 10^{-2} \end{bmatrix}$.
- So..

$$C_{11} = 1.429 \cdot 10^{-3}.$$

Combining, we get

$$T_0 = \frac{7.62}{\sqrt{267.62 \cdot 1.429 \cdot 10^{-3}}} = 12.32.$$

Contrasting to $t_{0.025,13} = 2.16$, we reject. The number of new loans is **significant**.

On the other hand, for $\hat{\beta}_2$ (number of loans outstanding):

- Again, looking at $\left(X^T X\right)^{-1}$:

$$C_{22} = 2.222 \cdot 10^{-2}.$$

And we get that

$$T_0 = \frac{8.58}{\sqrt{267.62 \cdot 2.222 \cdot 10^{-2}}} = 3.52.$$

This leads to rejecting the null hypothesis and hence the number of loans outstanding is **also significant**. That said, there is something to be said about which one of the two predictor variables is more important to the regression, no?

We finish this lecture with one big, comprehensive example, solved over the last few pages.

One big comprehensive example

A real estate problem

Taken from realtor.com, here are 8 of the most recently sold homes in Urbana:

	Sq. ft.	Year built	Garages	#bedrooms	#bathrooms	Price
1	1547	1950	1	3	3	158500
2	1834	1957	0	4	2	183000
3	2520	1980	3	5	2.5	233000
4	985	1911	1	2	1	69000
5	1275	1968	0	3	1.5	118000
6	2337	1977	2	5	2	249900
7	1880	1967	2	3	2	175000
8	1943	1965	1	4	2.5	169900

Which one of the five predictor variables (sq. ft., year built, garages, #bedrooms, #bathrooms) is the least important for predicting price? Use $\alpha = 0.05$.

To solve this problem, we enumerate our steps in a way that makes it easier to memorize, understand, and interpret. Here we go:

1. Build $X = \begin{bmatrix} 1 & 1547 & 1950 & 1 & 3 & 3 \\ 1 & 1834 & 1957 & 0 & 4 & 2 \\ 1 & 2520 & 1980 & 3 & 5 & 2.5 \\ 1 & 985 & 1911 & 1 & 2 & 1 \\ 1 & 1275 & 1968 & 0 & 3 & 1.5 \\ 1 & 2337 & 1977 & 2 & 5 & 2 \\ 1 & 1880 & 1967 & 2 & 3 & 2 \\ 1 & 1943 & 1965 & 1 & 4 & 2.5 \end{bmatrix}$.

2. Calculate $X^T X = \begin{bmatrix} 8 & 14321 & 15675 & 10 & 29 & 16.5 \\ 14321 & 27474233 & 28123127 & 20469 & 55469 & 30798 \\ 15675 & 28123127 & 30716537 & 19654 & 56950 & 32377.5 \\ 10 & 20469 & 19654 & 20 & 40 & 22 \\ 29 & 55469 & 56950 & 40 & 113 & 62 \\ 16.5 & 30798 & 32377.5 & 22 & 62 & 36.75 \end{bmatrix}$

3. Compute $X^T y = \begin{bmatrix} 1356300 \\ 2629528500 \\ 2664759800 \\ 1946200 \\ 5318600 \\ 2944550 \end{bmatrix}$

4. Compute $\left(X^T X\right)^{-1} = \begin{bmatrix} 3654.00381 & 0.09848 & -1.93379 & -15.26988 & -6.28608 & 0.35632 \\ 0.09848 & 0.00002 & -0.00005 & -0.00238 & -0.00486 & -0.00139 \\ -1.93379 & -0.00005 & 0.00102 & 0.00826 & 0.00341 & -0.00038 \\ -15.26988 & -0.00238 & 0.00826 & 0.53030 & 0.63902 & 0.17674 \\ -6.28608 & -0.00486 & 0.00341 & 0.63902 & 1.85652 & 0.37698 \\ 0.35632 & -0.00139 & -0.00038 & 0.17674 & 0.37698 & 0.62852 \end{bmatrix}$

5. Find $\hat{\beta} = \left(X^T X\right)^{-1} X^T y = \begin{bmatrix} -322042.7 \\ 93.2 \\ 150.2 \\ -4111.9 \\ 7242.7 \\ 4538.8 \end{bmatrix}$

We finally get that the regression line is:

$$\boxed{\hat{y} = -322042.7 + 93.2 \cdot x_1 + 150.2 \cdot x_2 - 4111.9 \cdot x_3 + 7242.7 \cdot x_4 + 4538.8 \cdot x_5}$$

Let us get some of the estimator calculations out of the way now:

- $SS_E = \sum_{i=1}^{8} (y_i - \hat{y}_i)^2 = 1164261866.8$.
- $\hat{\sigma}^2 = MS_E = \frac{SS_E}{8-6} = 582130933.4$.
- $SS_T = \sum (y_i - \overline{y})^2 = 23582558750$.
- Using ANOVA, $SS_R = SS_T - SS_E = 23582558750 - 1164261866.8 = 22418296883.2$.

We now have *everything* we need to do five distinct hypothesis tests for each of the five predictor variables. Specifically, we have:

1. For the square footage:
   - $H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0$.
   - $T_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \cdot C_{jj}}} = \frac{93.2}{\sqrt{582130933.4 \cdot 0.00002}} = 0.964$.
2. For the year built:
   - $H_0 : \beta_2 = 0, \quad H_1 : \beta_2 \neq 0$.
   - $T_0 = \frac{150.2}{\sqrt{\hat{\sigma}^2 \cdot 0.00102}} = 0.195$.
3. For the garage spots:
   - $H_0 : \beta_3 = 0, \quad H_1 : \beta_3 \neq 0$.
   - $T_0 = \frac{-4111.9}{\sqrt{\hat{\sigma}^2 \cdot 0.5303}} = -0.234$.

4. For the # bedrooms:
   - $H_0 : \beta_4 = 0, \;\; H_1 : \beta_4 \neq 0.$
   - $T_0 = \frac{7242.7}{\sqrt{\hat{\sigma}^2 \cdot 1.85652}} = 0.22.$
5. For the # bathrooms:
   - $H_0 : \beta_5 = 0, \;\; H_1 : \beta_5 \neq 0.$
   - $T_0 = \frac{4538.8}{\sqrt{\hat{\sigma}^2 \cdot 0.62852}} = 0.237.$

Hm... Apparently all factors are in the "fail to reject" region; in essence, this means that all of them one-by-one can be viewed as insignificant.. Some more (e.g., the year build with a $T_0 = 0.195$) than others (e.g., the square footage with a $T_0 = 0.964$), but still all of them can be declared insignificant when compared to $t_{\alpha/2,n-k-1} = t_{0.025,2} = 4.303$ as for all of them we have that $|T_0| < t_{\alpha/2,n-k-1}$. So, is the regression *significant at all*?

We can answer that through an F test:

$$F_0 = MS_R/MS_E = \frac{\frac{SS_R}{k}}{\frac{SS_E}{n-k-1}} = \frac{4483659376.64}{582130933.4} = 7.7.$$

Checking the critical value we get that $F_0 \leq f_{\alpha,k,n-k-1} = f_{0.05,5,2} = 19.3$, which means that we indeed do not have a good regression in our hands.

Finally, we may calculate the R^2 and adjusted R^2 coefficients:

- $R^2 = 1 - SS_E/SS_T = 0.951.$
- $R^2_{adj} = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)} = 0.827.$

Note the difference between R^2 and the adjusted R^2_{adj} showcasing that some insignificant predictor variables have been added.