Chapter 3 Linear Regression

Very old and dull technique, but powerful.
Questions to ask from business and feasibility of predictive modeling perspective

1) Is there a relation between resources (to put into predictors) and target? - business
2) How strong? - business
3) Which predictors contribute to target? - modeling
4) How accurately can we estimate the effect based on historical data? - modeling
5) How accurately can we predict the future based on the model? - modeling
6) Is the relationship linear? - modeling
7) Is there a synergy (i.e. if resources are divided between two predictors, do they have more impact that allocating to any one of them)? - business and statistics question

3.1 Simple Linear Regression: One predictor variable

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

Least square is the most used method to estimate betas based on some n observations. (It is a overdefined system of equations as per Linear Algebra)

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X.

$e_i = y_i - \hat{y}_i$ represents the ith *residual*— Residual sum of square: RSS

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

Least square tries to minimize RSS. The minimizers (using calculus on the equation for y hat)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad (3.4)$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

$\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

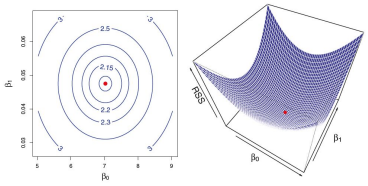

FIGURE 3.2. *Contour and three-dimensional plots of the RSS on the* This kind of plots we saw in Andrew Ngs's course. It is called gradient descent.

3.1.2 Assessing the Accuracy of the Coefficient Estimates

$$Y = \beta_0 + \beta_1 X + \epsilon. \qquad (3.5)$$

The error term epsilon is catch all for what we might miss with simple model.
Equation 3.5 defines *population regression* line which is best approximation to the true linear relationship between X and Y

large population. For example, suppose that we are interested in knowing the population mean μ of some random variable Y. Unfortunately, μ is unknown, but we do have access to n observations from Y, which we can write as $y_1, \ldots, y_n$, and which we can use to estimate μ. A reasonable estimate is $\hat{\mu} = \bar{y}$, where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean. The sample mean and the population mean are different, but in general the sample mean will provide a good estimate of the population mean. In the same way, the unknown coefficients β_0 and β_1 in linear regression define the population regression line. We seek to estimate these unknown coefficients using $\hat{\beta}_0$ and $\hat{\beta}_1$ given in (3.4). These coefficient estimates define the least squares line.

If we estimate betas from huge number of datasets, then average of such least squares estimations will be almost equal to true betas. Can be better explained with picture below

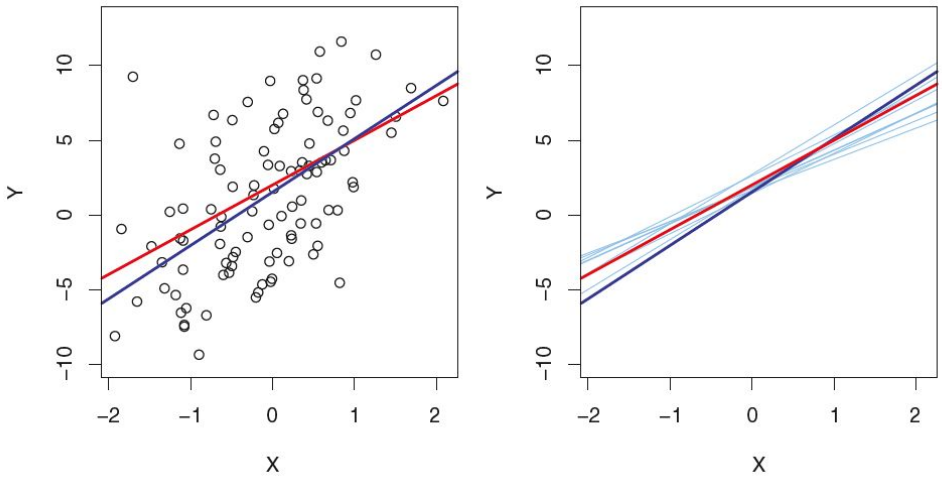

FIGURE 3.3. *A simulated data set.* Left: *The red line represents the true relationship, $f(X) = 2 + 3X$, which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for $f(X)$ based on the observed data, shown in black.* Right: *The population regression line is again shown in red, and the least squares line in dark blue. In light blue, ten least squares lines are shown, each computed on the basis of a separate random set of observations. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.*

Recall for the population mean estimation, we have standard error SE from standard dev sigma

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}, \qquad (3.7)$$

Standard error tells us the average amount by which estimate differs from the true mean.

estimate $\hat{\mu}$ differs from the actual value of μ. Equation 3.7 also tells us how this deviation shrinks with n—the more observations we have, the smaller the standard error of $\hat{\mu}$. In a similar vein, we can wonder how close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to the true values β_0 and β_1. To compute the standard errors associated with $\hat{\beta}_0$ and $\hat{\beta}_1$, we use the following formulas:

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad (3.8)$$

where $\sigma^2 = \mathrm{Var}(\epsilon)$. For these formulas to be strictly valid, we need to as- We need to assume that all epsilons are uncorrelated with common variance sigma^2.

If xi's are more spread out from mean, denominator is large and SE(beta1) is much smaller. Intuitively this means we have more leverage to estimate slope.

If mean(x) is zero, then beta0 is mean(y).

The estimate of sigma is called RSE - residual standard error , and equal to sqrt(RSS/(n-2))

Sigma for true population is not known, so it is estimated from data. Hence SE are also estimated values.

That is, there is approximately a 95 % chance that the interval

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1) \right] \quad (3.10)$$

will contain true Beta1

Similarly for Beta0

Hypothesis testing:

$$H_0 : \beta_1 = 0$$

versus

$$H_a : \beta_1 \neq 0,$$

with alpha = 5% for 95% confidence.

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}, \quad (3.14)$$

this measures the number of standard deviations that beta1 hat may be away from zero.

response. We *reject the null hypothesis*—that is, we declare a relationship to exist between X and Y—if the p-value is small enough. Typical p-value cutoffs for rejecting the null hypothesis are 5 or 1 %. When $n = 30$, these correspond to t-statistics (3.14) of around 2 and 2.75, respectively.

p-value is understood while studying inferential statistics

3.1.3: Assessing the accuracy of the model:

Once we reject null hypothesis, we measure quality of error, using Residual Standard Error (RSE) and R^2 statistics and even F-statistics

RSE is an estimate of standard deviation of epsilon. Practically it means the following. In data science problems, how these translate to reality is most important.

In the case of the advertising data, we see from the linear regression output in Table 3.2 that the RSE is 3.26. In other words, actual sales in each market deviate from the true regression line by approximately 3,260 units, on average. Another way to think about this is that even if the model were correct and the true values of the unknown coefficients β_0 and β_1 were known exactly, any prediction of sales on the basis of TV advertising would still be off by about 3,260 units on average. Of course, whether or not 3,260 units is an acceptable prediction error depends on the problem context. In the advertising data set, the mean value of `sales` over all markets is approximately 14,000 units, and so the percentage error is $3{,}260/14{,}000 = 23\,\%$.

The RSE is considered a measure of the *lack of fit* of the model (3.5) to the data. If the predictions obtained using the model are very close to the true outcome values—that is, if $\hat{y}_i \approx y_i$ for $i = 1, \ldots, n$—then (3.15) will be small, and we can conclude that the model fits the data very well. On the other hand, if $\hat{y}_i$ is very far from y_i for one or more observations, then the RSE may be quite large, indicating that the model doesn't fit the data well.

RSE is measured based on Y. It is not clear what is good RSE. R^2 is a number between 0 to 1, it is the proportion of variance explained.

To calculate R^2, we use the formula

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \qquad (3.17)$$

$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$ $\text{TSS} = \sum(y_i - \bar{y})^2$ is the *total sum of squares*,

$\text{TSS} - \text{RSS}$ measures the amount of variability in the response that is explained (or removed) by performing the regression, and R^2 measures the *proportion of variability in Y that can be explained using X*. An R^2 statistic that is

An R^2 number close to 0 could indicate that linear regression model was wrong or inherent variance sigma^2 was very high.
In certain problems like Physics, observations can come with very small residual error. In this case R^2 can be very high. In problems related to biology, psychology, marketing R^2 can be as low as 0.1, or below

3.2 Multiple Linear Regression

We should be taking all variables together and try fitting a model. Why not fit simple regression for each one of them? Answer below:

However, the approach of fitting a separate simple linear regression model for each predictor is not entirely satisfactory. First of all, it is unclear how to make a single prediction of sales given levels of the three advertising media budgets, since each of the budgets is associated with a separate regression equation. Second, each of the three regression equations ignores the other two media in forming estimates for the regression coefficients. We will see shortly that if the media budgets are correlated with each other in the 200 markets that constitute our data set, then this can lead to very misleading estimates of the individual media effects on sales.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \tag{3.19}$$

3.2.1 Estimating the Regression Coefficients

As was the case in the simple linear regression setting, the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ in (3.19) are unknown, and must be estimated. Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, we can make predictions using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p. \tag{3.21}$$

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Same least square approach to minimize

Multicollinearity is very well explained. Bottomline is if there are predictor variables X1 and X2 (along with few others), with target Y:

a) Suppose a simple regression gives non zero coefficients with small p-value, for both of these X1, X2
b) Multiple regression makes X1's coefficient almost zero, and with high p-value:

This means, X2 is the one influencing Y directly. X1 and X2 are correlated and X1 is just a surrogate for X2. So, when a simple regression is run, it is not aware of the other variables, hence just gives relationship between X1 and Y.

Note: It is ok/important to include target variable too while finding correlation matrix, drawing heatmap?

3.2.2 Some Important Questions

When we perform multiple linear regression, we usually are interested in answering a few important questions.

1. *Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?*
2. *Do all the predictors help to explain Y, or is only a subset of the predictors useful?*
3. *How well does the model fit the data?*
4. *Given a set of predictor values, what response value should we predict, and how accurate is our prediction?*

Answering first question: Is at least one of the variables useful?
Hypothesis testing to determine if indeed there's a relationship between X's and Y

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

versus the alternative

$$H_a : \text{ at least one } \beta_j \text{ is non-zero.}$$

This hypothesis test is performed by computing the *F-statistic*,

F-statistic

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)}, \tag{3.23}$$

where, as with simple linear regression, $\text{TSS} = \sum(y_i - \bar{y})^2$ and $\text{RSS} = \sum(y_i - \hat{y}_i)^2$. If the linear model assumptions are correct, one can show that

$$E\{\text{RSS}/(n-p-1)\} = \sigma^2$$

and that, provided H_0 is true,

$$E\{(\text{TSS} - \text{RSS})/p\} = \sigma^2.$$

Hence, when there is no relationship between the response and predictors, one would expect the F-statistic to take on a value close to 1. On the other hand, if H_a is true, then $E\{(\text{TSS} - \text{RSS})/p\} > \sigma^2$, so we expect F to be greater than 1.

Interesting point: How large F-statistic should be, if one were to reject Ha? It depends on n and p. If n was large, an F-stat even a little greater than 1 can provide evidence against H0, and if n is small we need a very large (several 100s to Million) to provide evidence against H0. When H0 is true and epsilons are normal, then F-statistic of coefficients follows a F-distribution.

Sometimes we might want to test hypothesis H0 that a subset of coefficients are zero, say q of them. Page number 77.
RSS for a second model dropping those q variables is RSS0, then F-stat is give by
(Note that when you drop some variables, residuals is higher than not dropping)

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n-p-1)}. \tag{3.24}$$

Pg no 77 contains answer to why we need F-stat, when we have t-stat for each predictor.
Brief answer is in presence of many variables like 100, there's 100% chance that 5 of them will have p-value below 0.05, just by chance. F-stat adjusts for the number of variables, and hence regardless of the number of variables, it has only 5% chance that p-value is below 0.05 by chance.

The approach of using an F-statistic to test for any association between the predictors and the response works when p is relatively small, and certainly small compared to n. However, sometimes we have a very large number of variables. If $p > n$ then there are more coefficients β_j to estimate than observations from which to estimate them. In this case we cannot even fit the multiple linear regression model using least squares, so the

[7] The square of each t-statistic is the corresponding F-statistic.

F-statistic cannot be used, and neither can most of the other concepts that we have seen so far in this chapter. When p is large, some of the approaches discussed in the next section, such as *forward selection*, can be used. This

Second Question: Deciding on important variables:

Forward Selection: Begin with null model. An intercept only. Fit p simple linear regression models, and add (to the null model) the variable that results in least RSS. Then add the next variable that has least RSS among the remaining and add it. Continue until stopping criteria are satisfied

Backward Selection: We start with all variables, remove the one with highest p-value. Then fit p-1 variable model and remove one with largest p-value. Continued until a stop condition is met

- *Mixed selection*. This is a combination of forward and backward selection. We start with no variables in the model, and as with forward selection, we add the variable that provides the best fit. We continue to add variables one-by-one. Of course, as we noted with the `Advertising` example, the p-values for variables can become larger as new predictors are added to the model. Hence, if at any point the p-value for one of the variables in the model rises above a certain threshold, then we remove that variable from the model. We continue to perform these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

mixed selection

Backward selection cannot be used if $p > n$, while forward selection can always be used. Forward selection is a greedy approach, and might include variables early that later become redundant. Mixed selection can remedy this.

Third question: Model fit:

Recall that in simple regression, R^2 is the square of the correlation of the response and the variable. In multiple linear regression, it turns out that it equals $\text{Cor}(Y, \hat{Y})^2$, the square of the correlation between the response and the fitted linear model; in fact one property of the fitted linear model is that it maximizes this correlation among all possible linear models.

Hence R^2 close to 1 indicates it(fitted model) is closely correlated to response.

R^2 will never decrease with additional variables, even if they are correlated with already existing and hence contribute very weakly. (Here R^2 as per the training set we are talking about, it is not guaranteed not to decrease with new/unseen before test data)

RSE shows similar trend to R^2.

Remember that RSS decreases with additional variables, but RSE might increase, as it is sqrt(RSS/n-p-1)

Important: Always plot residuals...This is not “just to observe whether they are random enough to fit through a linear regression” and hence validate our assumptions. It is to see if there’s synergy/interaction between variables.

can be useful to plot the data. Graphical summaries can reveal problems with a model that are not visible from numerical statistics. For example,

We see that some observations lie above and some observations lie below the least squares regression plane. In particular, the linear model seems to overestimate `sales` for instances in which most of the advertising money was spent exclusively on either `TV` or `radio`. It underestimates `sales` for instances where the budget was split between the two media. This pro-

This clearly means that combinations of variables actually result in bigger effect than just summing their impact, and this interaction should be captured in a separate variable. The resulting hyperplane may not look linear anymore, but it is still linear in the coefficients. Hence it is still Linear Regression

Four: Predictions:
There are three sorts of uncertainty associated with predictions.

1. Coefficients are only estimates of true population coefficients. We can have confidence interval to determine how close Yhat (Estimated) is to f(X) (True)
2. Model bias: Assumed that the samples can be fit into f(X)
3. Even if we knew true f(X), predicting new values from this is not a perfect process.

because of the random error ϵ in the model (3.21). In Chapter 2, we referred to this as the *irreducible error*. How much will Y vary from $\hat{Y}$? We use *prediction intervals* to answer this question. Prediction intervals are always wider than confidence intervals, because they incorporate both the error in the estimate for $f(X)$ (the reducible error) and the uncertainty as to how much an individual point will differ from the population regression plane (the irreducible error).

3.3 Other Considerations in Linear Regression Model

3.3.1: Qualitative predictors: One hot encoding: Convert a variables that can attain a finite set of values, say K types of values, into K-1 different predictor variables, that can take either 1 and 0.
3.3.2 Extensions of Linear Model: Additive and Linear are the two assumptions of Linear Regression Model. (Additive means, effect of Xj on Y is independent of other predictors, Linear means effect is same betai*Xi irrespective of value of Xi)
Removing additive effect: Add interaction term. Xi*Xj, if there's an interaction between those two.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

According to this model, if we increase X_1 by one unit, then Y will increase by an average of β_1 units. Notice that the presence of X_2 does not alter this statement—that is, regardless of the value of X_2, a one-unit increase in X_1 will lead to a β_1-unit increase in Y. One way of extending this model to allow for interaction effects is to include a third predictor, called an *interaction term*, which is constructed by computing the product of X_1 and X_2. This results in the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon. \quad (3.31)$$

For example, suppose that we are interested in studying the productivity of a factory. We wish to predict the number of units produced on the basis of the number of production lines and the total number of workers. It seems likely that the effect of increasing the number of production lines will depend on the number of workers, since if no workers are available to operate the lines, then increasing the number of lines will not increase production. This suggests that it would be appropriate to include an interaction term between lines and workers in a linear model to predict units. Suppose that when we fit the model, we obtain

$$\begin{aligned} \texttt{units} &\approx 1.2 + 3.4 \times \texttt{lines} + 0.22 \times \texttt{workers} + 1.4 \times (\texttt{lines} \times \texttt{workers}) \\ &= 1.2 + (3.4 + 1.4 \times \texttt{workers}) \times \texttt{lines} + 0.22 \times \texttt{workers}. \end{aligned}$$

In other words, adding an additional line will increase the number of units produced by $3.4 + 1.4 \times$ workers. Hence the more workers we have, the stronger will be the effect of lines.

Check 88 and 89 pages for sales example of how the effect of all three is not additive.

the associated main effects (in this case, TV and radio) do not. The *hierarchical principle* states that *if we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.* In other words, if the interaction be-

hierarchical principle

Concept of interaction applies as well to qualitative variables as well.
Example on how it looks

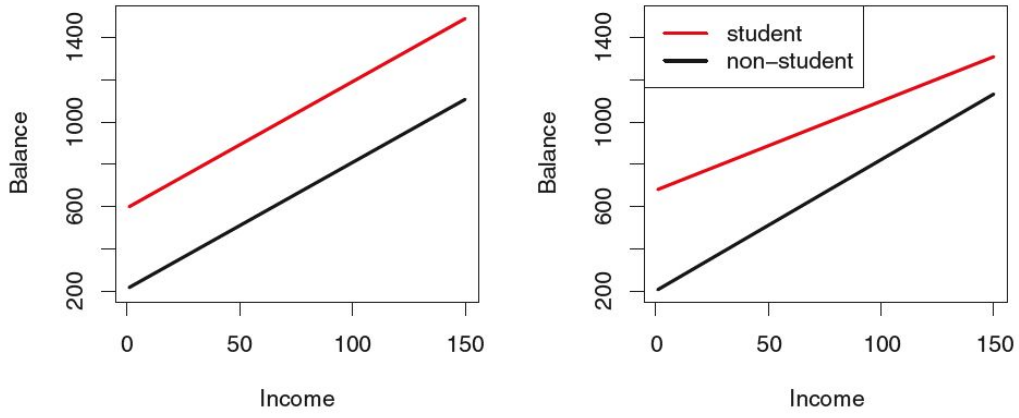

FIGURE 3.7. *For the* Credit *data, the least squares lines are shown for prediction of* balance *from* income *for students and non-students.* Left: *The model (3.34) was fit. There is no interaction between* income *and* student. Right: *The model (3.35) was fit. There is an interaction term between* income *and* student.

Notice the different slopes in second picture.

Without interaction term:

$$\begin{aligned}\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases}\beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student}\end{cases} \\ &= \beta_1 \times \texttt{income}_i + \begin{cases}\beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student.}\end{cases}\end{aligned} \tag{3.34}$$

With interaction term:

$$\begin{aligned}\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases}\beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student}\end{cases} \\ &= \begin{cases}(\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student}\end{cases}\end{aligned} \tag{3.35}$$

Removing Linear assumption:

3.3.3 Potential Problems

When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:

1. *Non-linearity of the response-predictor relationships.*
2. *Correlation of error terms.*
3. *Non-constant variance of error terms.*
4. *Outliers.*
5. *High-leverage points.*
6. *Collinearity.*

Parametric Vs Non parametric:
https://machinelearningmastery.com/parametric-and-nonparametric-machine-learning-algorithms/
kNN, Decision-Tree,SVM - non parametric.
Regressions or anything with closed form is parametric