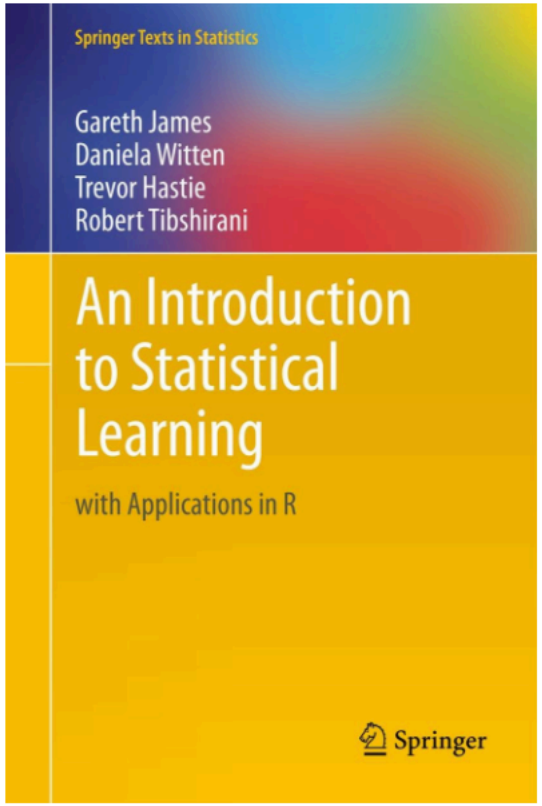


Textbook:

An Introduction to Statistical Learning

Lecture Slides:

Stanford Stats 202: Data Mining and Analysis

Spring 17’ ESC Statistical Data Analysis

REFERENCE

Reading:

An Introduction to Statistical Learning

chapter 3.1 Simple Linear Regression
chapter 3.2 Multiple Linear Regression
chapter 3.3.1 Qualitative Predictors
chapter 3.3.2 Extensions of the Linear Model

Table of Contents

Simple Linear Regression

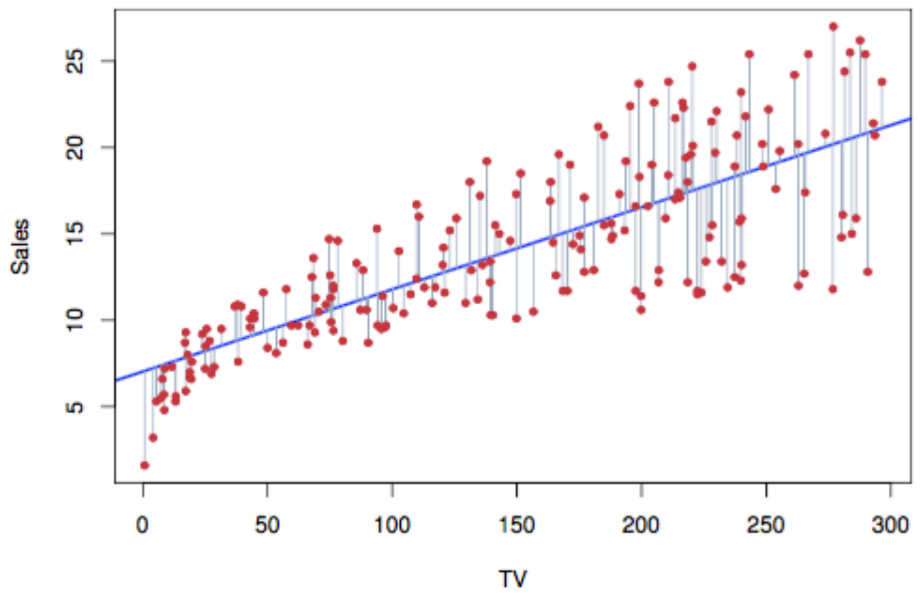


Figure 3.1

- We assume a model using two parameters or coefficients β_0 and β_1 where β_0 represents the intercept and β_1 represents the slope

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- We assume error ϵ as

$$\epsilon_i \sim iid\ N(0, \sigma^2)$$

- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we predict Y using

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

Estimates $\hat{\beta}_0$ and $\hat{\beta}_1$: Least Square Method

- The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to minimize the residual sum of squares (RSS):

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- A little calculus shows that the minimizers of the RSS are:

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

Accuracy of $\hat{\beta}_0$ and $\hat{\beta}_1$

- Standard error of an estimator can be obtained as

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

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

- Then 95% confidence intervals are

$$\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0)$$

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$$

Hypothesis Test

- An important question: Does a relationship between Y and X exist?
- We can perform hypothesis tests on the coefficients

$$\begin{cases} H_0 : \text{There is no relationship between } X \text{ and } Y \\ H_1 : \text{There is some relationship between } X \text{ and } Y \end{cases}$$

- This can be expressed as

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

- Under H_0, the *t-statistic* can be defined as

$$t = \frac{\hat{\beta}_1}{\sqrt{\sigma^2 / \sum (x_i - \bar{x})^2}} \sim t(n-2)$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

TABLE 3.1. *For the* Advertising *data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of \$1,000 in the TV advertising budget is associated with an increase in sales by around 50 units (Recall that the* sales *variable is in thousands of units, and the* TV *variable is in thousands of dollars).*

Interpreting the Hypothesis Test

- If we reject the null hypothesis, can we conclude that there is significant evidence of a linear relationship?
  - No. A quadratic relationship may be a better fit

- If we don't reject the null hypothesis, can we assume there is no relationship between X and Y?
  - No. This test is only powerful against certain monotone alternatives. There could be more complex non-linear relationships.

Interpretation of Simple Linear Regression

- $\hat{\beta}_0$: Intercept of the regression function

  that is, $E(Y|X = 0)$

- $\hat{\beta}_1$: Slope of the regression function

$$\hat{\beta}_1 = \frac{\partial E(Y \mid X = x)}{\partial x}$$

  that is, the increment of $E(Y|X = x)$ when X increases 1 unit from x

Multiple Linear Regression

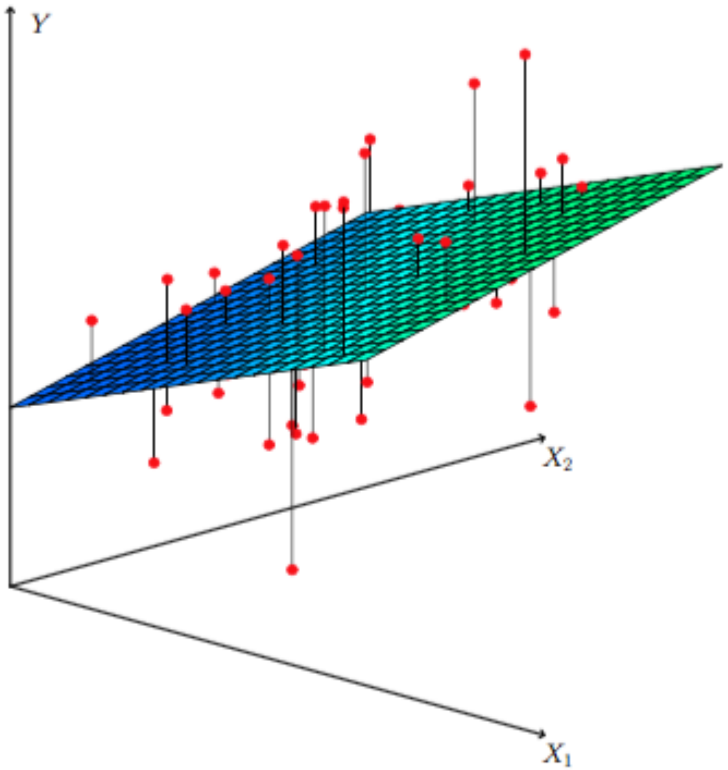


Figure 3.4

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, \sigma) \quad \text{i.i.d.}$$

or, in matrix notation:

$$E\mathbf{y} = \mathbf{X}\beta,$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\beta = (\beta_0, \ldots, \beta_p)^T$ and $\mathbf{X}$ is our usual data matrix with an extra column of ones on the left to account for the intercept.

The Estimates $\widehat{\beta}$

- Our goal is to minimize the RSS (training error):

$$\begin{aligned}\text{RSS} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i,1} - \cdots - \beta_p x_{i,p})^2.\end{aligned}$$

- This is minimized by the vector $\hat{\beta}$ (next page):

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

- This only exists when $X^T X$ is invertible. This requires $n \geq p$.

- Same as simple linear regression; we estimate coefficients by minimizing the *RSS*

$$\text{Arg Min RSS} = \sum_{i=1}^{n}(y_i - \hat{y})^2 = (Y - X\beta)^T(Y - X\beta)$$

- Differentiate by β and make it 0 to minimize,

$$\frac{\partial\, RSS}{\partial \beta} = -2X^TY + 2X^TX\beta = 0$$

thus,

$$\hat{\beta} = (X^TX)^{-1}X^TY$$

Testing a Group of Variables

- F-test:

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0.$$

RSS_0 is the residual sum of squares for the model in H_0.

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

- Special case: $q = p$. Test whether any of the predictors are related to Y.
- Special case: $q = 1$, exclude a single variable. Test whether this variable is related to Y after linearly correcting for all other variables. Equivalent to t-test in R output.

Interpretation of Multiple Linear Regression

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

$\hat{\beta}_j$ represents;

1. $\hat{\beta}_j = \frac{\partial E(Y|X)}{\partial X_j}$: the change of $E(Y)$ per unit increase in X_j while X_k $(k = 0, 1, 2, \ldots p\ , k \neq j)$'s held constant.
2. the additional contribution of X_j on Y, after X_j and Y has been adjusted for X_k $(k = 0, 1, 2, \ldots p\ , k \neq j)$

How many variables are important?

- When choosing a subset of the predictors, we have 2^p choices. We cannot test every possible subset!

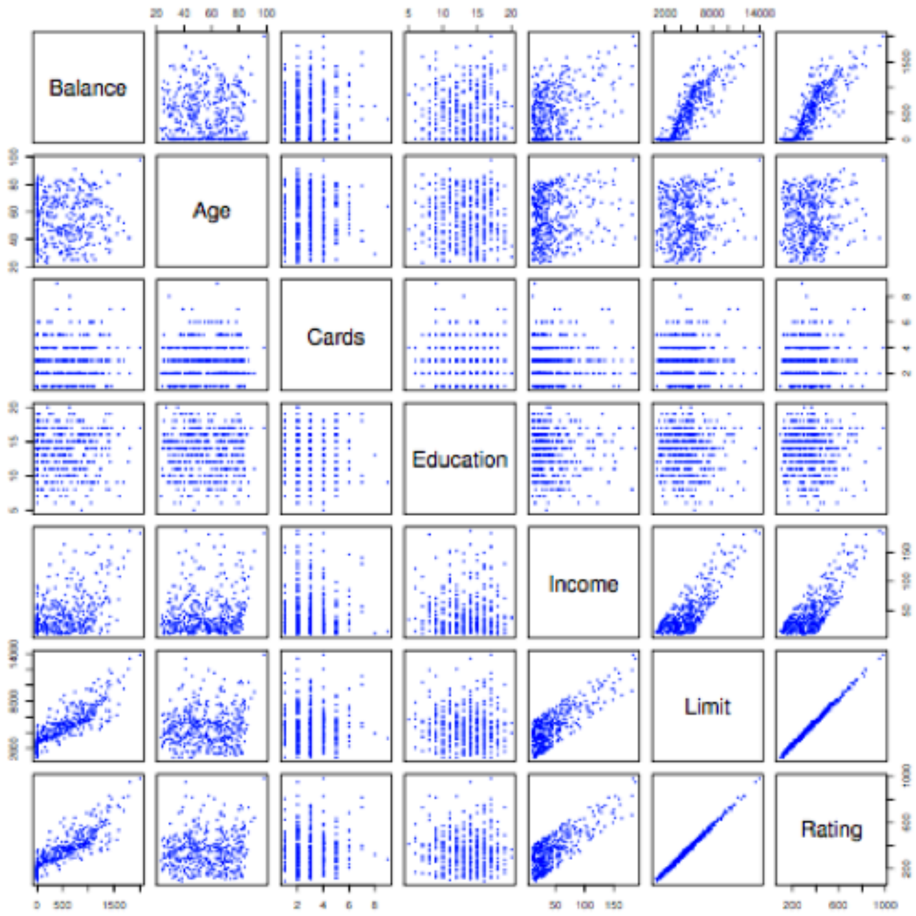
- Instead we will use a **stepwise approach:**
  - Construct a sequence of p models with increasing number of variables.
  - Select the best model among them.

Three Variants of Stepwise Selection

- **Forward Selection:** Starting from a null model (the intercept), include variables one at a time, minimizing the RSS at each step.
- **Backward Selection:** Starting from the full model, eliminate variables one at a time, choosing the one with the largest t-test p-value at each step.
- **Mixed Selection:** Starting from a null model, include variables one at a time, minimizing the RSS at each step. If the p-value for some variable goes beyond a threshold, eliminate the variable.

Categorical or Qualitative Predictors

Example: Credit Dataset



- There are 4 qualitative variables:
  - `gender`: male, female
  - `student`: student or not
  - `status`: married, single, divorced
  - `ethnicity`: African, American, Asian, Caucasian

- For each qualitative predictor, e.g. `status`:
  - Choose a baseline category, e.g. single
  - For every other category, define a new predictor:
    - $X_{married}$ is 1 if the person is married and 0 otherwise.
    - $X_{dicorced}$ is 1 if the person is divorced and 0 otherwise.
- The model will be:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_7 X_7 + \beta_{\text{married}} X_{\text{married}} + \beta_{\text{divorced}} X_{\text{divorced}} + \varepsilon.$$

- $\beta_{married}$ is the relative effect on balance for being married compared to the baseline category.

- The model fit $\hat{f}$ and predictions $\hat{f}(x_0)$ are independent of the choice of the baseline category.

- However, the interpretation of parameters and associated hypothesis tests depend on the baseline category.
  - **Solution**: To check whether `status` is important, use an F-test for the hypothesis $\beta_{married} = \beta_{divorced} = 0$. This does not depend on the coding of the baseline category.

The function `predict` in R output predictions from a linear model; eg. $x_0 = (5, 10, 15)$:

```
> predict(lm.fit,data.frame(lstat=(c(5,10,15))),
        interval="confidence")
    fit   lwr   upr
1 29.80 29.01 30.60
2 25.05 24.47 25.63
3 20.30 19.73 20.87
```

"Confidence intervals" reflect the uncertainty on $\hat{\beta}$; ie. confidence interval for $f(x_0)$.

```
> predict(lm.fit,data.frame(lstat=(c(5,10,15))),
        interval="prediction")
    fit    lwr   upr
1 29.80 17.566 42.04
2 25.05 12.828 37.28
3 20.30  8.078 32.53
```

"Prediction intervals" reflect uncertainty on $\hat{\beta}$ and the irreducible error ε as well; i.e. confidence interval for y_0.

Goodness of the Fit

- To assess the fir, we focus on the residuals.
  - $R^2 = corr(Y, \hat{Y})$, always increases as we add more variables.
  - The residual standard error (RSE) does not always improve with more predictors:

$$\text{RSE} = \sqrt{\frac{1}{n-p-1}\text{RSS}}.$$

  - Visualizing the residuals can reveal phenomena that are not accounted for by the model.

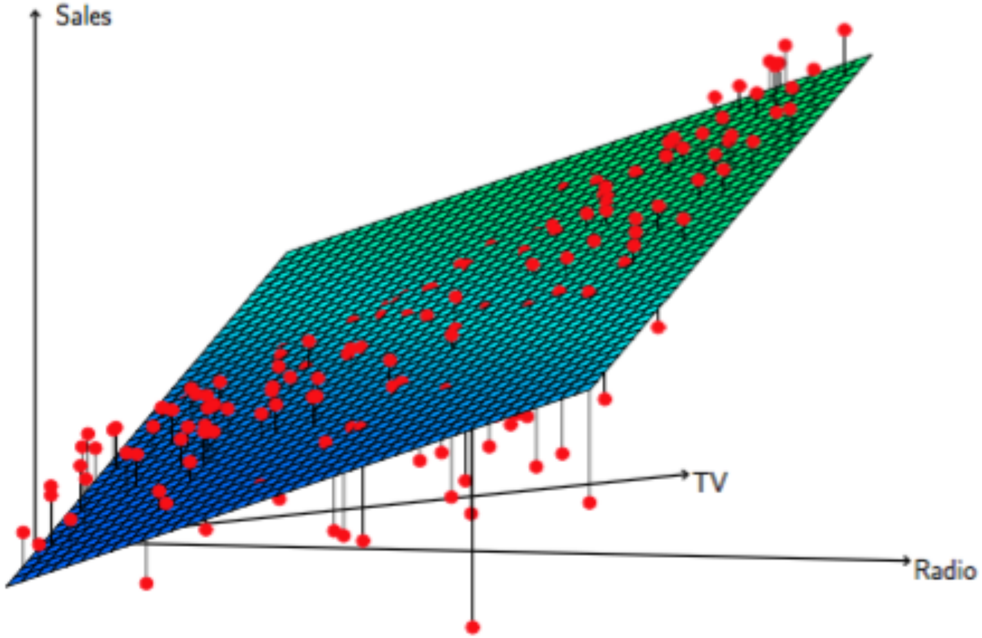
Interactions between Predictors

- Linear regression has an additive assumption:

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{tv} + \beta_2 \times \texttt{radio} + \varepsilon$$

- i.e. An increase of $100 dollars in TV ads causes a fixed increase sales, regardless of how much you spend on radio ads.

- When we visualize the residuals, we see a pronounced non-linear relationship:



- One way to deal with this is to include multiplicative variables in the model:

$$\mathtt{sales} = \beta_0 + \beta_1 \times \mathtt{tv} + \beta_2 \times \mathtt{radio} + \beta_3 \times (\mathtt{tv} \cdot \mathtt{radio}) + \varepsilon$$

- The interaction variable is high when both `tv` and `radio` are high.

R makes it easy to include interaction variables in the model:

```
> lm.fit=lm(Sales~.+Income:Advertising+Price:Age,data=Carseats)
> summary(lm.fit)

Call:
lm(formula = Sales ~ . + Income:Advertising + Price:Age, data =
    Carseats)

Residuals:
   Min      1Q  Median      3Q     Max
-2.921  -0.750   0.018   0.675   3.341

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.575565   1.008747    6.52  2.2e-10 ***
CompPrice           0.092937   0.004118   22.57  < 2e-16 ***
Income              0.010894   0.002604    4.18  3.6e-05 ***
Advertising         0.070246   0.022609    3.11  0.00203 **
Population          0.000159   0.000368    0.43  0.66533
Price              -0.100806   0.007440  -13.55  < 2e-16 ***
ShelveLocGood       4.848676   0.152838   31.72  < 2e-16 ***
ShelveLocMedium     1.953262   0.125768   15.53  < 2e-16 ***
Age                -0.057947   0.015951   -3.63  0.00032 ***
Education          -0.020852   0.019613   -1.06  0.28836
UrbanYes            0.140160   0.112402    1.25  0.21317
USYes              -0.157557   0.148923   -1.06  0.29073
Income:Advertising  0.000751   0.000278    2.70  0.00729 **
Price:Age           0.000107   0.000133    0.80  0.42381
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hierarchy

- Sometimes it is the case that an interaction term has a very small p-value, but the associated main effects do not.
- ***The Hierarchy Principle***:
  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.
- With significant interaction term, it does not matter whether main effect coefficient is 0 or not. Interaction terms are hard to interpret in a model without main effect - their meaning is changed.
- The interaction terms also contain main effects, if the model has no main effect terms.