IEOR E4650 Business Analytics

Session 4: Linear Regression

Spring 2018

Prof. Adam Elmachtoub

What Data Looks Like

	X_1	X_2	$\ldots$	X_p	Y
x_1	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_2					
$\cdots$					
x_n					

- (x_i, y_i) is the i^{th} row of data, where x_i vector length p and $y_i \in \mathbb{R}$
- $X_1, X_2, \ldots, X_p$ and Y are *each* scalar random variables, and we denote the random vector $X := (X_1, X_2, \ldots, X_p)$
- One should think of x_i as a sample from X
- One should think of y_i as a sample from Y that depends on x_i
- X_i are also called input variables/independent variables/features/attributes/covariates/predictors/regressors/factors
- Y is also called output variable/outcome/response/label/dependent variable

Main Types of Columns

	X_1	X_2	$\ldots$	X_p	Y
x_1	...	...	...	...	...
x_2					
...					
x_n					

- Continuous: *quantitative*, a number like weight or length, the values can be sorted
- Discrete: *qualitative*, categorical such as 'cat' or 'dog', '0' or '1', $\{0, 1, 2, 3\}$, {January, February,..., December}, typically no inherent ordering on values

Main Goal of Supervised Learning: Prediction

Explanatory/Input variables vs. Response/Output variables

1. Obtain some kind of model based on observations, or training data $\{(x_i, y_i), i = 1, 2, \ldots, n\}$, through a process called learning (or estimation).

2. Use that model to predict something about data you haven't seen before, but that comes from the same distribution as the training data, called test data.

Types of Supervised Learning

- Regression: predict a continuous output variable

- Classification: predict a discrete/categorical output variable

Linear Regression: a fundamental starting point for all regression methods.

Linear Regression

- Linear regression is an approach to model the relationship between a scalar, continuous, dependent variable Y and one or more explanatory variables denoted $X \in \mathbb{R}^p$
  - The case of one explanatory variable is called simple linear regression.
  - For more than one explanatory variable, it is called multiple linear regression.
- Linear regression is a tool for predicting a quantitative response.
- Linear regression is useful and widely used statistical learning method.

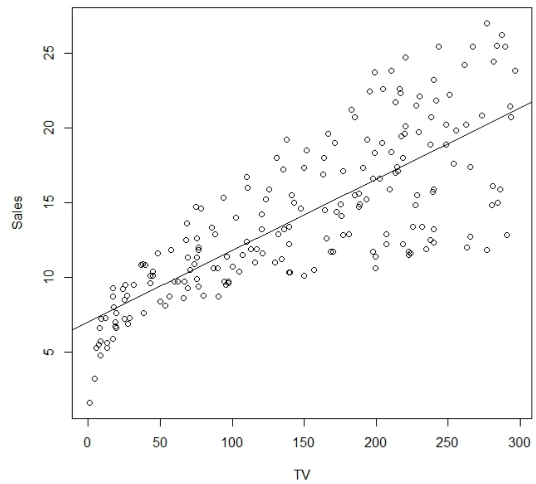
Example: Advertising

- `Sales` (thousand units), advertising budget on `TV ($), Radio ($), Newspaper ($)`
- 200 records

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
...				

Advertising Data: plots in-class exercise

```
> ad=read.csv("Advertising.csv")
> attach(ad)
> plot(TV,Sales)
```



Questions about Advertising Data: Discussion

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we estimate the effect of each media on sales?
- How accurately can we predict future sales?
- Is the relationship linear?

Multiple Linear Regression

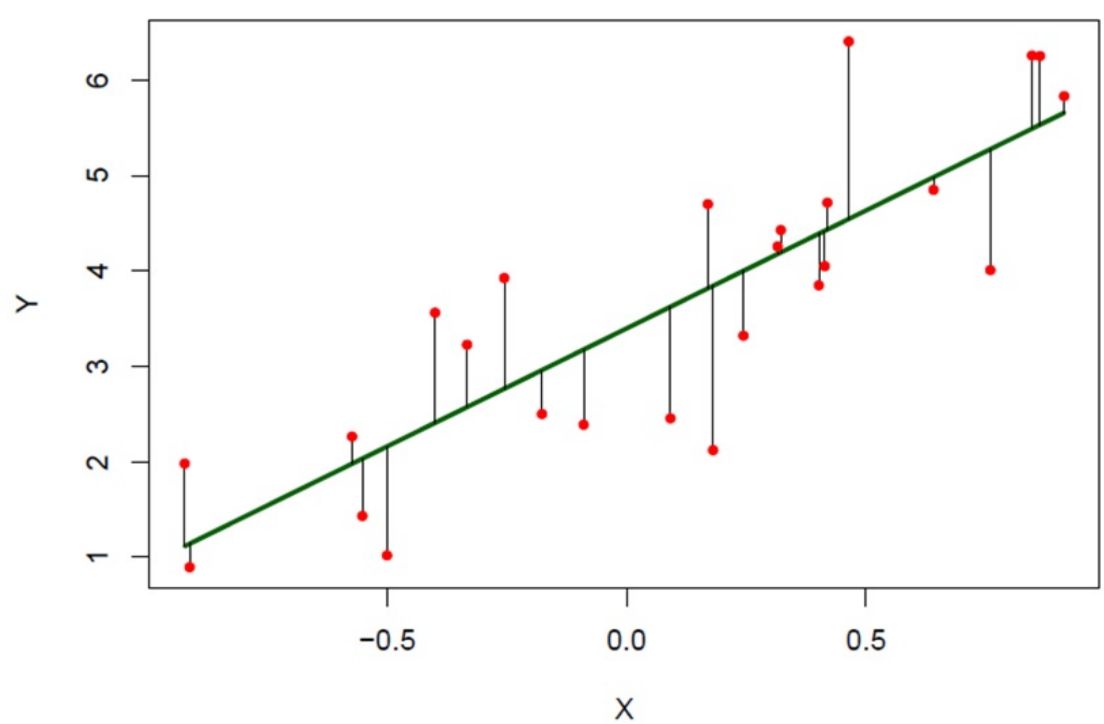
- Multiple linear regression model

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \\
Y &\approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p
\end{aligned}
$$

  - where ϵ denotes the irreducible error, i.e., a mean zero random variable with variance σ^2
  - X_j represents the j-th predictor and β_j quantifies the association between that variable and the response.
  - We interpret β_j as the average effect on Y of a one unit increase in X_j , holding all other predictors fixed.
  - $\beta_0, \beta_1, \ldots, \beta_p$ are unknown constants
  - Use training data to produce estimates $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$, then prediction $\widehat{y_i}$ for input x_i is

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_p x_{ip}.$$

Prediction vs. Observation



Least Squares Approach

- Denote n observation pairs as follows
$$(x_1, y_1),\ (x_2, y_2), \ldots,\ (x_n, y_n),$$
where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$.
- Given estimates, we make predictions using the formula
$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_p x_{ip}.$$
- Let $e_i = y_i - \widehat{y_i}$ be the residual for prediction i
- Residual sum of squares (RSS) and mean squared error (MSE) are
$$RSS = e_1^2 + \ldots e_n^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \text{ and } MSE = \frac{1}{n} RSS$$
- Least Squares idea: Find $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$ that minimize RSS (MSE)
$$RSS = \sum_{i=1}^{n} \left(y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_p x_{ip})\right)^2.$$

Algorithm to find the Regression Coefficients

- Efficient computer codes exist to estimate $\widehat{\beta} = [\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p]^T$
- Define
$$A = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \text{ and } b = \begin{bmatrix} y_1 \\ y_2 \\ .. \\ y_n \end{bmatrix}$$
- Estimation formula:
$$\hat{\beta} = (A^T A)^{-1} A^T b$$
- Important: $\hat{\beta}$ is an unbiased estimator of β! This means that if we repeated this procedure many times with different datasets, then $E[\hat{\beta}] = \beta$.

Advertising Data: Multiple Linear Regression

```
> my.lm.4 = lm(Sales~TV+Radio+Newspaper)
> summary(my.lm.4)
Residuals:
Min       1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate  Std.Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0  *** 0.001 ** 0.01 * 0.05  0.1  1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared: 0.8972,Adjusted R-squared: 0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```


Accuracy of the Model: RSE

- The quality of a linear regression is a function of the RSS and the number of predictors p used. Less RSS is better (less bias) and small p is better (less variance)
- Two most common measures of accuracy are: the residual standard error (RSE) and the adjusted R^2 statistic.
- RSE formula:

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

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

- Note: One can estimate the variance σ^2 by RSE^2

R^2: the Coefficient of Determination

- R^2 formula

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS},$$

where TSS is the total sum of squares, i.e., $TSS = \sum_{i=1}^{n}(y_i - \overline{y})^2$ where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

- R^2 measures the proportion of variability in Y that can be explained using X.
  - An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression.
  - A number near 0 indicates that the regression did not explain much of the variability in the response.

Adjusted R^2

- By adding variables to a model, the *residual sum of squares* (RSS) decreases, so the R^2 increases.
- The adjusted R^2 for a model with p predictors and $p+1$ estimated coefficients,

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}.$$

It introduces a penalty for the number of estimated coefficients.

- While the R^2 can never decrease as more variables are added to the model, the adjusted R^2 with too many unneeded variables can actually decrease.

Hypothesis: Multiple Linear Regression

- Coefficient estimates $\widehat{\beta_0}, \ldots, \widehat{\beta_p}$ depend on the data, can be different when we change the data
- Is there ANY relationship between the response and predictors?
- Null hypothesis is $H_0 : \ \beta_1 = \beta_2 = \ldots = \beta_p = 0.$
- Alternative hypothesis H_a : at least one β_j is non-zero.
- This hypothesis test is performed by computing the F-statistic,

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

- Large F-statistic provides evidence against the null hypothesis H_0.

Advertising Data: Multiple Linear Regression

```
> my.lm.4 = lm(Sales~TV+Radio+Newspaper)
> summary(my.lm.4)
Residuals:
Min       1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
            Estimate   Std.Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
TV           0.045765   0.001395  32.809   <2e-16 ***
Radio        0.188530   0.008611  21.893   <2e-16 ***
Newspaper   -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0  *** 0.001 ** 0.01 * 0.05  0.1  1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared: 0.8972,Adjusted R-squared: 0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Understanding the Regression Summary

- call: this shows how lm() was called when it created the model.
- Residuals statistics: Min and Max; first quartile (1Q) and third quartile (3Q); median
- Coefficients
  - The column labeled Estimate contains the estimated regression coefficients as calculated by ordinary least squares.
  - The column labeled Std. Error (SE) is the standard error (SE) of the estimated coefficient. This is the std. dev. of the estimate. The column labeled t value is the t statistic from which the p-value was calculated.
  - The p-value represents the probability of what we observed if the true coefficients were 0. It gauges the likelihood that the coefficient is not significant, so **smaller** p-value means that it's more likely that the coefficient is significant.

Understanding the Regression Summary, Cont.

- Residual standard error (RSE): this reports the standard error of the residuals – that is, the sample standard deviation.
- R^2 and adjusted R^2: R^2 is a measure of the model's quality; the adjusted R^2 accounts for the number of variables in your model and so is a more realistic assessment of its effectiveness.
- F statistic: the F statistic tells you whether the model is significant or insignificant. The model is significant if any of the coefficients are nonzero. Conventionally, a p-value of less than 0.05 indicates that the model is likely significant (one or more β_i are nonzero)
- Most people look at the R^2 statistic first. The statistician wisely starts with the F statistic, for if the model is not significant then nothing else matters.

Confidence Interval for Coefficients

- A 95% confidence interval for an estimate (of any kind) means that if the CI procedure is repeated many times, the true value is within that interval 95% of the time
- Thus, if our CI is in that 95%, then the true estimate is somewhere in the interval we computed... but we don't know if we are in that 95%
- The 95% confidence intervals for $\widehat{\beta}_i$ are approximately:

$$[\widehat{\beta}_i - 2 \cdot SE(\widehat{\beta}_i),\ \widehat{\beta}_i + 2 \cdot SE(\widehat{\beta}_i)]$$

- In order to obtain a confidence interval for the coefficient estimates, we can use the `confint()` command.

```
> confint(my.lm.4,level=.9)
                     5%          95%
(Intercept)  2.42340953 3.454369213
TV           0.04345935 0.048069943
Radio        0.17429853 0.202761502
Newspaper   -0.01074031 0.008665319
```


Confidence and Prediction Intervals for Y

- The `predict()` function can be used to produce *confidence* intervals for the prediction of `Sales` for a given value of `TV`. This gives a range for the *average* value of Y given X. ($E[Y|X]$, i.e., assumes 0 noise)

```
> predict(my.lm.4,data.frame(TV=c(75,140), Radio=c(40,50), Newspaper=c(30,35))
,interval="confidence",level=.98)
fit      lwr      upr
1 13.88131 13.37571 14.38691
2 18.73613 18.14301 19.32925
```

- The `predict()` function can be used to produce *prediction* intervals for the prediction of `Sales` for a given value of `TV`. This gives a range for the *observed* values of Y given X. ($Y|X$, includes noise ϵ)

```
> predict(my.lm.4,data.frame(TV=c(75,140), Radio=c(40,50), Newspaper=c(30,35))
,interval="prediction",level=.98)
fit      lwr      upr
1 13.88131  9.89571 17.86692
2 18.73613 14.73848 22.73378
```

- As expected, the confidence and prediction intervals are centered around the same point, but the latter are substantially wider.

Extensions of the Linear Model

- Adding interaction terms
  - Standard linear regression model with two variable

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

  - Interaction term

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

- Non-linear Relationships: example

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

- These models are still linear in β!
- The `lm()` function can also accommodate non-linear transformations of the predictors.
- We can create a predictor X^2 by using $I(X^2)$
- Example: > my.lm5=lm(Sales~ TV +I(TV^2))

Interaction Terms

- The syntax $x_1 : x_2$ tells R to include an interaction term
- The syntax $x_1 * x_2$ simultaneously includes x_1, x_2 and the interaction term $x_1 \times x_2$
- Example:

```
> my.lm2=lm(Sales~TV:Radio,data=ad)
> my.lm2=lm(Sales~TV*Radio,data=ad)
> my.lm3=lm(Sales~TV:Radio+TV,data=ad)
> my.lm4=lm(Sales~TV:Radio+Radio,data=ad)
```

- Look at RSE and adjusted R^2 on test data to pick best model

Example: Toyota Used-Car Prices

```
> toyota=read.csv("ToyotaCorolla.csv")
> toyota[1:5,]
```

```
  Price Age    KM FuelType HP MetColor Automatic   CC Doors Weight
1 13500  23 46986   Diesel 90        1         0 2000     3   1165
2 13750  23 72937   Diesel 90        1         0 2000     3   1165
3 13950  24 41711   Diesel 90        1         0 2000     3   1165
4 14950  26 48000   Diesel 90        0         0 2000     3   1165
5 13750  30 38500   Diesel 90        0         0 2000     3   1170
```

- FuelType is *NOT* quantitative!

Example: Toyota Used-Car Prices, Cont.

- We create indicator variables for the categorical variable
- Use function `levels()` to check categorical variables: FuelType with its three nominal outcomes: CNG, Diesel, and Petrol

```
> v1=rep(1,length(toyota$FuelType))
> v2=rep(0,length(toyota$FuelType))
> toyota$FuelType1=ifelse(toyota$FuelType=="CNG",v1,v2)
> toyota$FuelType2=ifelse(toyota$FuelType=="Diesel",v1,v2)
> auto=toyota[-4]
```

```
> auto[1:3,]
  Price Age    KM HP MetColor Automatic   CC Doors Weight FuelType1 FuelType2
1 13500  23 46986 90        1         0 2000     3   1165         0         1
2 13750  23 72937 90        1         0 2000     3   1165         0         1
3 13950  24 41711 90        1         0 2000     3   1165         0         1
```

Example: Toyota Used-Car Prices, Cont.

- Play with Data

```
> plot(Price~Weight, data=auto)
> plot(Price~KM)
> plot(Price~Automatic)
```

- Plot residuals

```
> m11=lm(Price~Age+KM,data=auto)
> summary(m11)
> plot(m11$res~m11$fitted)
> hist(m11$res)
```


Example: Toyota Used-Car Prices, Cont.

- Regression w.r.t all variables

```
> m2=lm(Price~.,data=auto)
> summary(m2)
```

- Regression w.r.t all variables but some

```
> m3=lm(Price~.-MetColor,data=auto)

> m4=lm(Price~.-MetColor-Automatic,data=auto)
> summary(m4)
```

Additional Comments: Toyota Used-Car Prices

- Introduce quadratic terms

```
> auto$Age2=auto$Age^2
> auto$KM2=auto$KM^2
```

- Run regression, compare them

```
> m11=lm(Price~Age+KM,data=auto)
> summary(m11)

> m12=lm(Price~Age+Age2+KM+KM2,data=auto)
> summary(m12)

> m13=lm(Price~Age+Age2+KM,data=auto)
> summary(m13)
```

- Residual plots

```
> m11=lm(Price~Age+KM,data=auto)
> summary(m11)
> plot(m11$res~m11$fitted)
> hist(m11$res)
```


Appendix

Regression Plots

- The `abline( )` function can be used to draw any line, not just the least squares regression line.
- To draw a line with `intercept` a and `slope` b, we type `abline(a,b)`.
- The `lwd=3` command causes the width of the regression line to be increased by a factor of 3
- We can also use the `pch` option to create different plotting symbols.

```
> plot(Sales~TV,data=ad)
> abline(my.lm,lwd=3)
> plot(Sales~TV,col="red")
> plot(Sales,TV,col="red")
> plot(Sales,TV,pch=20)
> plot(Sales,TV,pch="$")
> plot(1:20,1:20,pch=1:20)
```


How to deal with a large number of dummy variables?

- For the Toyota data set, we have shown how to turn "CNG", "Diesel", and "Petrol" into two dummy variables.
- An alternative is to use the psych library.
- The commands to do this are:

```
library(psych)
newcolumns = dummy.code(Dummycolumn)
newdataset = cbind(olddataset, newcolumns)
```