Linear Regression - Introduction

Objectives

After this lecture you should be able to:

- Define supervised learning
- Define parametric models
- Define linear regression
    - Difference between simple and multiple linear regression
- Interpret linear regression coefficients
- Describe how coefficients in linear regression are determined
- Assess the accuracy of the coefficients
- Assess the accuracy of a linear regression model
- Interpret an Ordinary Least Squares linear regression summary from Statsmodels

Supervised vs. Unsupervised Learning

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.

X, y -> predicting y based on the values in X

X, y a.k.a
features, target
independent, dependent
exogenous, endogenous
predictors, response

Example capstone: [Avalanche Prediction]

Unsupervised learning is a type of self-organized … learning that helps find previously unknown patterns in data set without pre-existing labels.

X -> understanding structure in X
Example capstone: [Spice Blends]

-Wikipedia

Parametric vs Non-parametric models

A machine learning algorithm can be supervised or unsupervised, and parametric or non-parametric.

A **parametric** algorithm

- has a fixed number of parameters
- makes assumptions about the structure of the data
- will work well if the assumptions are correct!
- common examples: linear regression, neural networks, statistical distributions defined by a finite set of parameters

A **non-parametric** algorithm

- uses a flexible number of parameters, and the number of parameters often grows as it learns from more data.
- makes fewer assumptions about the data
- common examples: K-Nearest Neighbors, decision trees

source - Jason Brownlee

Linear Regression

Given a data set $\{y_i,\, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ of n statistical units, a linear regression model assumes that **the relationship between the dependent variable y and the p-vector of regressors x is linear**.

Assuming that there is only one predictor leads to modeling the relationship using **simple linear regression**:

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

More than one predictor leads to **multiple linear regression**:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Y response

X predictors

β parameters/coefficients (what we need to figure out from *fitting* data)

ϵ irreducible error

Linear Regression

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

$$Y \approx \beta_0 + \beta_1 X$$

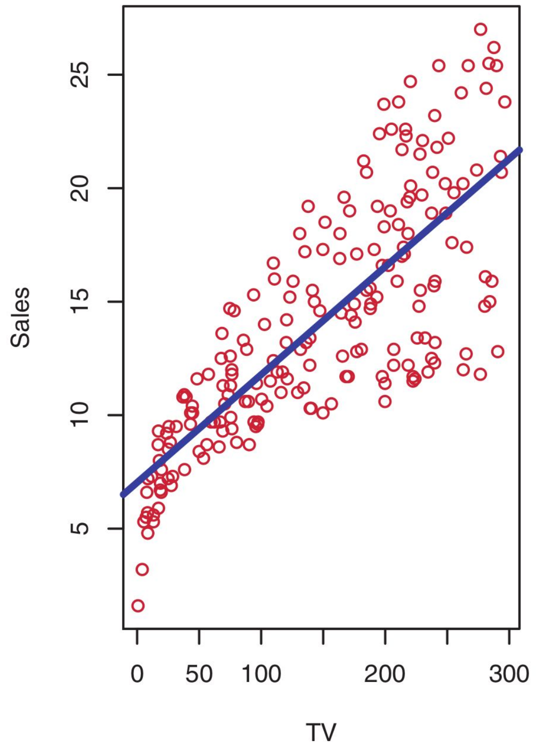
$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}$$

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

$\hat{y} =$ 7 + 0.05 x blue line ->

^ estimated value of unknown parameter, predicted value

Sales as a *f*(TV advertising)



source: [ISLR]

Interpreting linear regression coefficients

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

How to interpret $\hat{\beta}_1$?

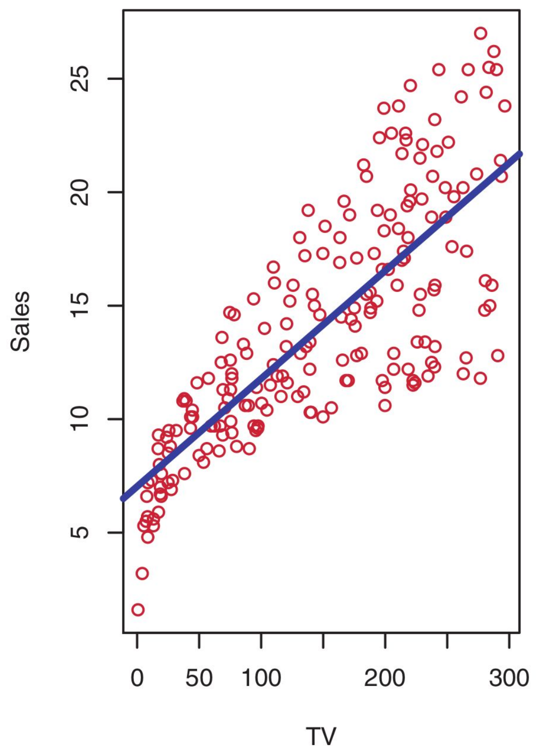
Holding all else constant, a 1 unit increase in x_1 increases the response $\hat{y}$ by $\hat{\beta}_1$.

$\hat{y} =$ 7 + 0.05 x

$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}$

Increasing TV by 1 increases sales by 0.05.

Sales as a *f*(TV advertising)



source: ISLR

Determining the coefficients (*fitting* the model)

The coefficients β are unknown. We'd like to find values for them so that the resulting prediction (line in simple regression, multidimensional plane in multiple regression) is as close to the training data as possible.

This is a mathematical optimization problem.

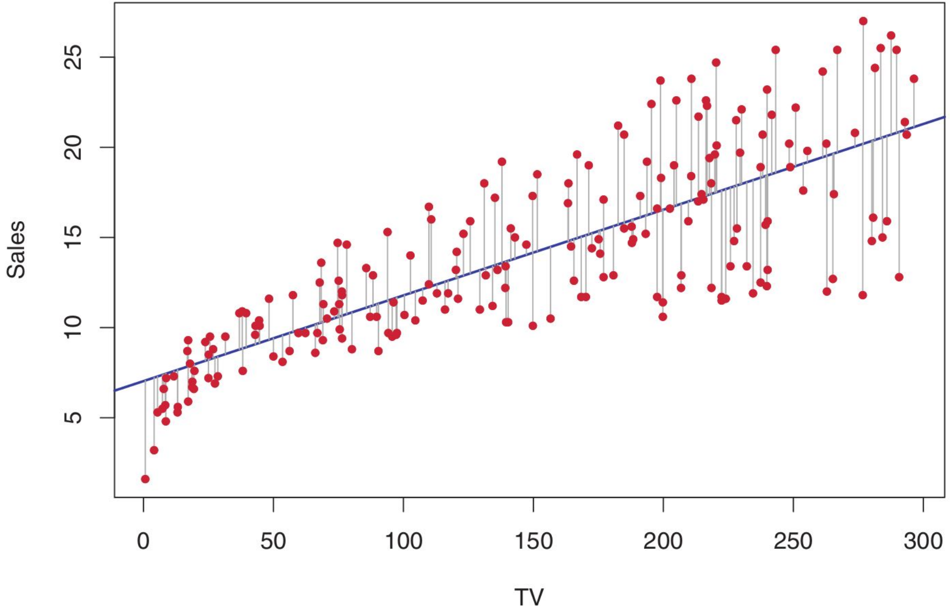
In mathematical optimization, we need an objective function to maximize or minimize.

In machine learning, we typically talk about making loss/cost functions. The goal of the optimization is to find the parameters/coefficients/weights that minimize the cost/loss function.

Cost function: Residual Sum of Squares (RSS)

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

$y_i - \hat{y}_i$ is a residual



Mean squared error is the average RSS:

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

source: ISLR

Least squares: solve β by minimizing squared residuals

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

Any given set of parameters/coefficients will give a value of RSS. We'd like to find the set that give the minimum RSS.

This will be the minimum of the cost function (a surface in multidimensional space).

If the cost function is simple enough, you can take its derivative, set it equal to zero, and solve for the coefficients -> yields an analytic solution for the coefficients. (Great, this is fast!)

If the cost function is more complex (very often the case), resort to numerical optimization. Commonly, gradient descent (more later).

Least squares analytic solution for linear regression

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

See derivation again.

Make matrices clear

Design Matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p-1} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p-1} \end{bmatrix}$$

Target:

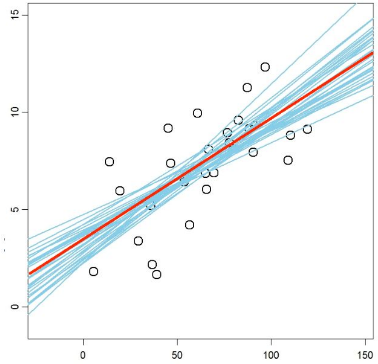
$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Coefficient matrix β:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$$

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

Assessing the usefulness/accuracy of your coefficients



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

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

$$\sigma = \mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}}$$

	Recall	**Here**
Setup Hypothesis	H_0: $\mu = 100$	$H_0 : \beta_1 = 0$
Sample Statistic	$\bar{x}$	$\hat{\beta}_1$
Test Statistic	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$
Confidence Interval	$(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}})$	$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$

Test if X has effect on Y

source: ISLR

Assessing the accuracy/usefulness of your model

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

Residual sum of squares - ok, but depends on *n*

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Mean squared error - better, but in squared units of response

$$\mathrm{RMSD}(\hat{\theta}) = \sqrt{\mathrm{MSE}(\hat{\theta})}$$

Root mean square error (or deviation) - in units of response (similar to RSE previous slide)

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

R squared, ranges from 0 to 1 (but doesn't penalize for model complexity), where $\mathrm{TSS} = \sum (y_i - \bar{y})^2$

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n-1}{n-p-1}$$

Adjusted R-squared (0 to 1), penalizes for model complexity according to number of predictors, *p*

source: Wikipedia & [ISLR]

Assessing the accuracy/usefulness of your model

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

Residual sum of squares - ok, but depends on *n*

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Mean squared error - better, but in squared units of response

$$\mathrm{RMSD}(\hat{\theta}) = \sqrt{\mathrm{MSE}(\hat{\theta})}$$

Root mean square error (or deviation) - in units of response

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

R squared, ranges from 0 to 1 (but doesn't penalize for model complexity), where $\mathrm{TSS} = \sum (y_i - \bar{y})^2$

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Adjusted R-squared (0 to 1), penalizes for model complexity according to number of predictors, *p*

source: Wikipedia & ISLR

Assessing the accuracy/usefulness of your model

There were hypothesis tests checking if coefficients in the model were significant (not zero).

The F-statistic is a hypothesis test where the Null Hypothesis is that **all the coefficients are = 0** (so model is not useful).

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

The alternative is that at least one of the coefficients is not 0.

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

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

For no relationship between predictors and response (H0), F statistic is near 1.

source: ISLR

Interpretation of OLS summary from Statsmodels.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.933
Model:                            OLS   Adj. R-squared:                  0.928
Method:                 Least Squares   F-statistic:                     211.8
Date:                Mon, 03 Nov 2014   Prob (F-statistic):           6.30e-27
Time:                        14:45:06   Log-Likelihood:                -34.438
No. Observations:                  50   AIC:                             76.88
Df Residuals:                      46   BIC:                             84.52
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
x1             0.4687      0.026     17.751      0.000         0.416     0.522
x2             0.4836      0.104      4.659      0.000         0.275     0.693
x3            -0.0174      0.002     -7.507      0.000        -0.022    -0.013
const          5.2058      0.171     30.405      0.000         4.861     5.550
==============================================================================
Omnibus:                        0.655   Durbin-Watson:                   2.896
Prob(Omnibus):                  0.721   Jarque-Bera (JB):                0.360
Skew:                           0.207   Prob(JB):                        0.83[illegible]
Kurtosis:                       3.026   Cond. No.                         221.
==============================================================================
```

Proportion of Variance Explained by model is 93.3%

Measure of the significance of the fit …my model isn't utterly useless ☺

There is an approximately 95% chance that [0.275, 0.693] will contain the true value of β2

Each coefficient is really significant. Can also think of this as a Partial F-test.

"The average effect on Y of a one unit increase in X_2, holding all other predictors (X_1 & X_3) fixed, is 0.4836"

- However, interpretations are generally pretty hazardous due to correlations among predictors.
- p-values for each coefficient ≈ 0, so might be okay here

Note: Magnitude of the Beta coefficients is NOT how to determine whether predictor contributes. Why?

Statsmodel’s OLS demonstration

statsmodels_ols_demonstration.ipynb

Objectives

After this lecture you should be able to:

- Define supervised learning
- Define parametric models
- Define linear regression
    - Difference between simple and multiple linear regression
- Interpret linear regression coefficients
- Describe how coefficients in linear regression are determined
- Assess the accuracy of the coefficients
- Assess the accuracy of a linear regression model
- Interpret an Ordinary Least Squares linear regression summary from Statsmodels