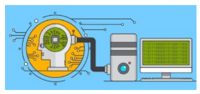
Applied Data Science

Session 15-28: Applied Machine Learning

Dr. Soharab Hossain Shaikh



1

Supervised Learning
::
Regression
:
Linear Regression

2

Regression Models

- Relationship between one dependent variable and explanatory variable(s)
- Use equation to set up relationship
  - Numerical Dependent (Response) Variable
  - one or More Numerical or Categorical Independent (Explanatory) Variables
- Used mainly for Prediction & Estimation of a **continuous numeric variable**

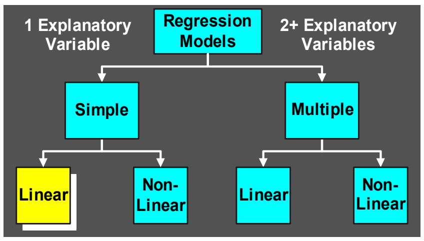
3

Regression Modeling Steps

- 1. Hypothesize Deterministic Component
  - Estimate Unknown Parameters
- 2. Specify Probability Distribution of Random Error Term
  - Estimate Standard Deviation of Error
- 3. Evaluate the fitted Model
- 4. Use Model for Prediction & Estimation

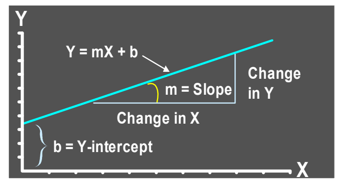
4

Types of Regression Models



5

Linear Equations



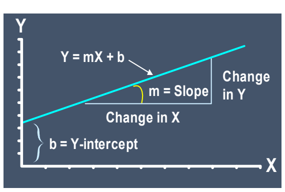
Linear relationship between the predictors/independent variable (X) and the response/dependent variable (Y) is assumed.

6

Simple Linear Regression

Linear Equation : y = β_0 + β_1*x

Living area (feet^2)	Price (1000\$s)
2104	400
1600	330
2400	369
1416	232
3000	540
⋮	⋮



Linear relationship between the predictors/independent variable (x) and the response/dependent variable (y) is assumed.

7

Multiple Linear Regression

More than one predictor…

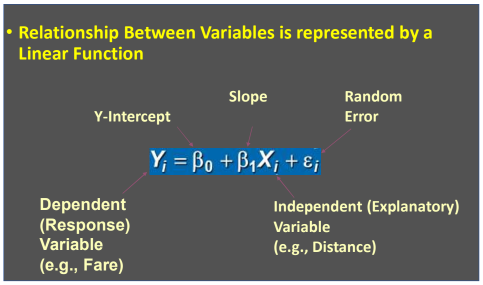
Living area (feet^2)	#bedrooms	Price (1000\$s)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540
⋮	⋮	⋮

y = β_0 + β_1*x$_1$ + β_2*x$_2$

Each regression coefficient is the amount of change in the outcome variable that would be expected per one-unit change of the predictor, if all other variables in the model were held constant.

8

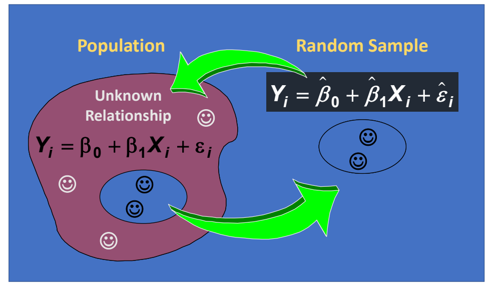
Linear Regression Model



38

9

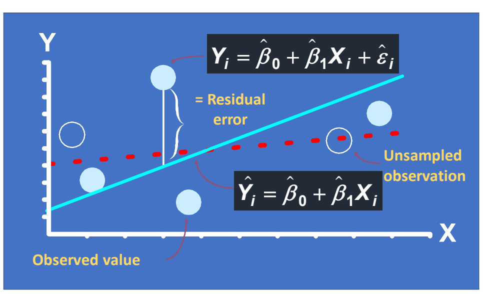
Population & Sample Regression Models



10

10

Linear Regression Model

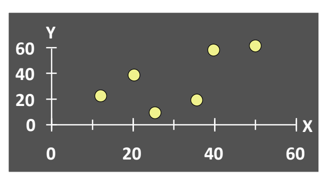


11

11

Scatter Plot

- 1. Plot of All (X_i, Y_i) Pairs
- 2. Suggests How Well Model Will Fit

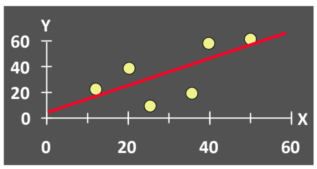


12

12

Estimate Parameters

How would you draw a line through the points? How do you determine which line 'fits best'?

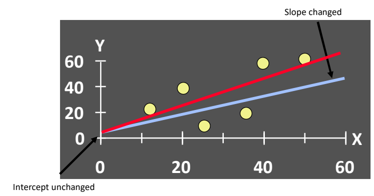


13

13

Estimate Parameters

How would you draw a line through the points? How do you determine which line 'fits best'?

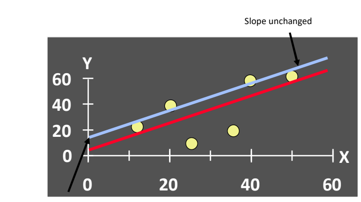


14

14

Estimate Parameters

How would you draw a line through the points? How do you determine which line 'fits best'?

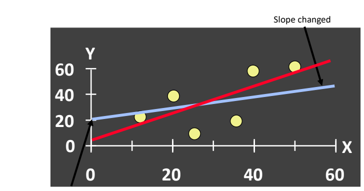


15

15

Estimate Parameters

How would you draw a line through the points? How do you determine which line 'fits best'?



16

16

Least Squares

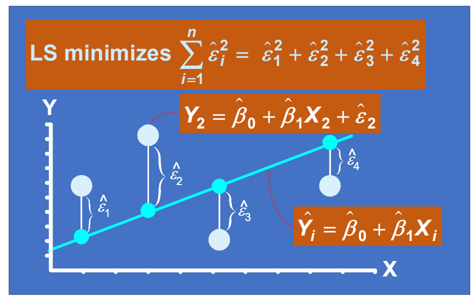
- 1. 'Best Fit' Means difference between actual y values & predicted $\hat{y}$ values are a minimum.
- *But* Positive Differences Off-Set Negative. So square errors!
- 2. LS Minimizes the Sum of the Squared Differences (errors) (SSE)

$$\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

17

17

Least Squares Graphically



18

18

Assumptions

- Linear regression assumes that…
  - 1. The relationship between X and Y is linear
  - 2. Y is distributed normally at each value of X
  - 3. The variance of Y at every value of X is the same (homogeneity of variances)
  - 4. The observations are independent

48

19

Residual Analysis: Check Assumptions

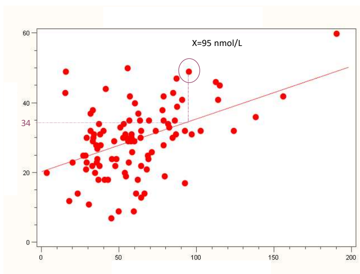
$$e_i=Y_i-\hat{Y}_i$$

- The residual for observation ***i***, $\mathbf{e_i}$, is the difference between its observed and predicted value.
- Check the assumptions of regression by examining the residuals
  - Examine for linearity assumption
  - Examine for constant variance for all levels of X (homoscedasticity)
  - Evaluate normal distribution assumption
  - Evaluate independence assumption
- Graphical Analysis of Residuals
  - Can plot residuals vs. X

49

20

Residual = Observed - Predicted

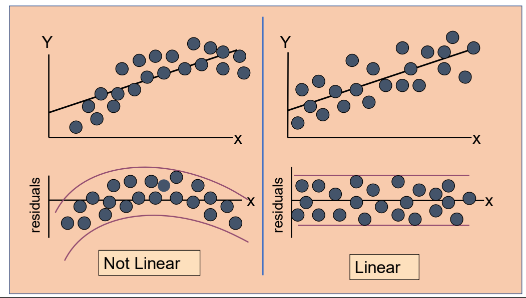


$$y_i = 48$$
$$\hat{y}_i = 34$$
$$y_i - \hat{y}_i = 14$$

50

21

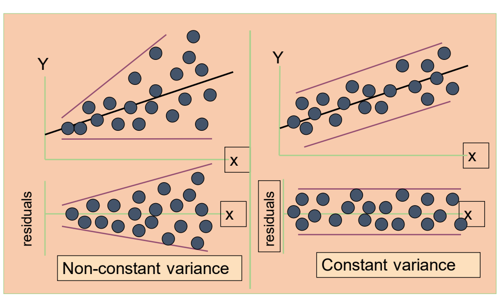
Residual Analysis for Linearity



60

22

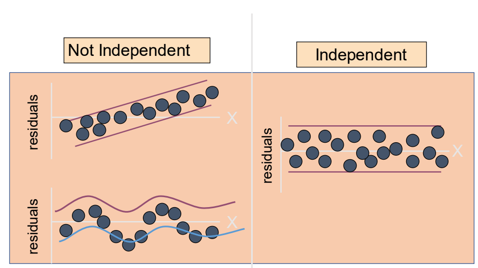
Residual Analysis for Homoscedasticity



61

23

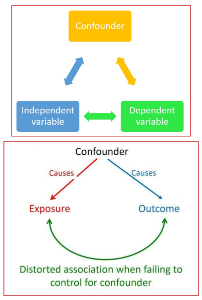
Residual Analysis for Independence

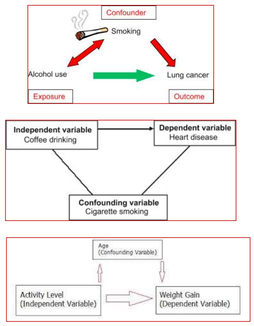


62

24

Confounding Variable





25

Multivariate Regression Pitfalls

- **Multicollinearity**: two variables that measure the same thing or similar things (e.g., weight and BMI) are both included in a multiple regression model; they will, in effect, cancel each other out and generally destroy your model.
- **Residual confounding**: we cannot completely wipe out confounding simply by adjusting for variables in multiple regression unless variables are measured with zero error (which is usually impossible).
- **Overfitting**: In multivariate modeling, you can get highly significant but meaningless results if you put too many predictors in the model.

64

26

Least Square Regression: An Example

27

Predicting Electricity Use

What will peak power consumption be in Pittsburgh tomorrow?

Difficult to build an “a priori” model from first principles to answer this question

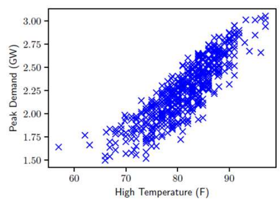
But, relatively easy to record past days of consumption, plus additional features that affect consumption (i.e., weather)

Date	High Temperature (F)	Peak Demand (GW)
2011-06-01	84.0	2.651
2011-06-02	73.0	2.081
2011-06-03	75.2	1.844
2011-06-04	84.9	1.959
...	...	...

28

Plot Consumption vs. Temperature

Plot of high temperature vs. peak demand for summer months (June – August) for past six years



29

Hypothesis: Linear Model

Let's suppose that the peak demand approximately fits a *linear model*

$$\text{Peak\_Demand} \approx \theta_1 \cdot \text{High\_Temperature} + \theta_2$$

Here θ_1 is the "slope" of the line, and θ_2 is the intercept

How do we find a "good" fit to the data?

Many possibilities, but natural objective is to minimize some difference between this line and the observed data, e.g. squared loss

$$E(\theta) = \sum_{i \in \text{days}} (\theta_1 \cdot \text{High\_Temperature}^{(i)} + \theta_2 \ - \text{Peak\_Demand}^{(i)})^2$$

30

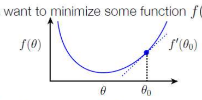
How do we find the parameters?

How do we find the parameters θ_1, θ_2 that minimize the function

$$E(\theta) = \sum_{i \in \text{days}} (\theta_1 \cdot \text{High\_Temperature}^{(i)} + \theta_2 \ - \text{Peak\_Demand}^{(i)})^2$$

$$\equiv \sum_{i \in \text{days}} (\theta_1 \cdot x^{(i)} + \theta_2 \ - y^{(i)})^2$$

General idea: suppose we want to minimize some function $f(\theta)$



Derivative is slope of the function, so negative derivative points "downhill"

31

Computing the Derivatives

What are the derivatives of the error function with respect to each parameter θ_1 and θ_2?

$$\begin{aligned}
\frac{\partial E(\theta)}{\partial \theta_1} &= \frac{\partial}{\partial \theta_1} \sum_{i \in \text{days}} (\theta_1 \cdot x^{(i)} + \theta_2 - y^{(i)})^2 \\
&= \sum_{i \in \text{days}} \frac{\partial}{\partial \theta_1} (\theta_1 \cdot x^{(i)} + \theta_2 - y^{(i)})^2 \\
&= \sum_{i \in \text{days}} 2(\theta_1 \cdot x^{(i)} + \theta_2 - y^{(i)}) \cdot \frac{\partial}{\partial \theta_1} \theta_1 \cdot x^{(i)} \\
&= \sum_{i \in \text{days}} 2(\theta_1 \cdot x^{(i)} + \theta_2 - y^{(i)}) \cdot x^{(i)} \\
\frac{\partial E(\theta)}{\partial \theta_2} &= \sum_{i \in \text{days}} 2(\theta_1 \cdot x^{(i)} + \theta_2 - y^{(i)})
\end{aligned}$$

32

Finding the best θ

To find a good value of θ, we can repeatedly take steps in the direction of the negative derivatives for each value

Repeat:

$$\theta_1 := \theta_1 - \alpha \sum_{i \in \text{days}} 2(\theta_1 \cdot x^{(i)} + \theta_2 - y^{(i)}) \cdot x^{(i)}$$

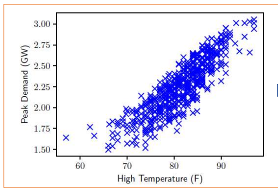
$$\theta_2 := \theta_2 - \alpha \sum_{i \in \text{days}} 2(\theta_1 \cdot x^{(i)} + \theta_2 - y^{(i)})$$

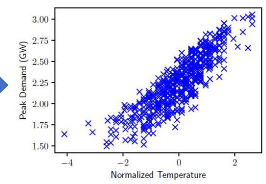
where α is some small positive number called the *step size*

This is the *gradient decent algorithm*, the workhorse of modern machine learning

33

Gradient Descent

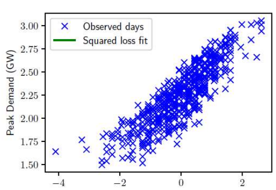




Normalize input by subtracting the mean and dividing by the standard deviation

34

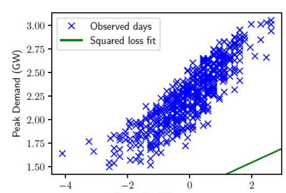
Gradient Descent - Iteration 1



$\theta = (0.00, 0.00)$
$E(\theta) = 1427.53$
$(\frac{\partial E(\theta)}{\partial \theta_1}, \frac{\partial E(\theta)}{\partial \theta_2}) = (-151.20, -1243.10)$

35

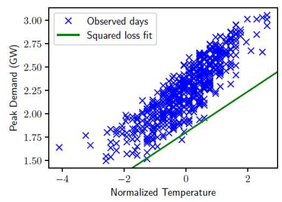
Gradient Descent - Iteration 2



$\theta = (0.15, 1.24)$
$E(\theta) = 292.18$
$(\frac{\partial E(\theta)}{\partial \theta_1}, \frac{\partial E(\theta)}{\partial \theta_2}) = (-67.74, -556.91)$

36

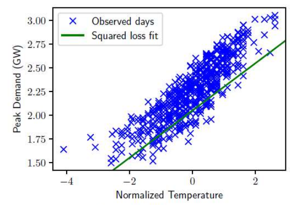
Gradient Descent - Iteration 3



$\theta = (0.22, 1.80)$
$E(\theta) = 64.31$
$(\frac{\partial E(\theta)}{\partial \theta_1}, \frac{\partial E(\theta)}{\partial \theta_2}) = (-30.35, -249.50)$

37

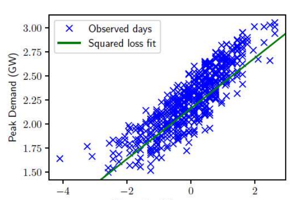
Gradient Descent - Iteration 4



$\theta = (0.25, 2.05)$
$E(\theta) = 18.58$
$(\frac{\partial E(\theta)}{\partial \theta_1}, \frac{\partial E(\theta)}{\partial \theta_2}) = (-13.60, -111.77)$

38

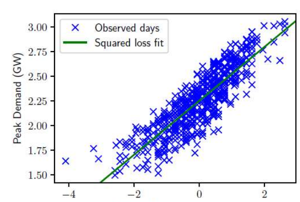
Gradient Descent - Iteration 5



$\theta = (0.26, 2.16)$
$E(\theta) = 9.40$
$(\frac{\partial E(\theta)}{\partial \theta_1}, \frac{\partial E(\theta)}{\partial \theta_2}) = (-6.09, -50.07)$

39

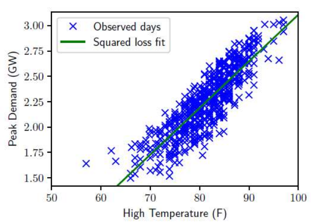
Gradient Descent - Iteration 10



$\theta = (0.27, 2.25)$
$E(\theta) = 7.09$
$(\frac{\partial E(\theta)}{\partial \theta_1}, \frac{\partial E(\theta)}{\partial \theta_2}) = (-0.11, -0.90)$

40

Fitted Line in Original Coordinates



41

Making Prediction

Importantly, our model also lets us make *predictions* about new days

What will the peak demand be tomorrow?

If we know the high temperature will be 72 degrees (ignoring for now that this is *also* a prediction), then we can predict peak demand to be:

$$\text{Predicted\_demand} = \theta_1 \cdot 72 + \theta_2 = 1.821 \text{ GW}$$

(requires that we rescale θ after solving to “normal” coordinates)

Equivalent to just “finding the point on the line”

42

Extensions

What if we want to add additional features, e.g. day of week, instead of just temperature?

What if we want to use a different loss function instead of squared error (i.e., absolute error)?

What if we want to use a non-linear prediction instead of a linear one?

We can easily reason about all these things by adopting some additional notation…

43

Least Squares

Using our new terminology, plus matrix notion, let's see how to solve linear regression with a squared error loss

Setup:

- Linear hypothesis function: $h_\theta(x) = \sum_{j=1}^{n} \theta_j \cdot x_j$
- Squared error loss: $\ell(\hat{y}, y) = (\hat{y} - y)^2$
- Resulting machine learning optimization problem:

$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \theta_j \cdot x_j^{(i)} - y^{(i)} \right)^2 \equiv \underset{\theta}{\text{minimize}}\ E(\theta)$$

44

Derivative of the Least Squares Objective

Compute the partial derivative with respect to an arbitrary model parameter θ_j

$$\begin{aligned}\frac{\partial E(\theta)}{\partial \theta_k} &= \frac{\partial}{\partial \theta_k}\sum_{i=1}^{m}\left(\sum_{j=1}^{n}\theta_j \cdot x_j^{(i)} - y^{(i)}\right)^2 \\ &= \sum_{i=1}^{m}\frac{\partial}{\partial \theta_k}\left(\sum_{j=1}^{n}\theta_j \cdot x_j^{(i)} - y^{(i)}\right)^2 \\ &= \sum_{i=1}^{m}2\left(\sum_{j=1}^{n}\theta_j \cdot x_j^{(i)} - y^{(i)}\right)\frac{\partial}{\partial \theta_k}\sum_{j=1}^{n}\theta_j \cdot x_j^{(i)} \\ &= \sum_{i=1}^{m}2\left(\sum_{j=1}^{n}\theta_j \cdot x_j^{(i)} - y^{(i)}\right)x_k^{(i)}\end{aligned}$$

45

Gradient Descent Algorithm

1. Initialize $\theta_k := 0,\ k = 1, \ldots, n$
2. Repeat:
   - For $k = 1, \ldots, n$:

$$\theta_k := \theta_k - \alpha\sum_{i=1}^{m}2\left(\sum_{j=1}^{n}\theta_j \cdot x_j^{(i)} - y^{(i)}\right)x_k^{(i)}$$

Note: do not actually implement it like this, you'll want to use the matrix/vector notation we will cover soon

46

The Gradient

It is typically more convenient to work with a vector of all partial derivatives, called the **gradient**

For a function $f: \mathbb{R}^n \to \mathbb{R}$, the gradient is a vector

$$\nabla_\theta f(\theta) = \begin{bmatrix}\frac{\partial f(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial f(\theta)}{\partial \theta_n}\end{bmatrix} \in \mathbb{R}^n$$

47

Gradient in Vector Notation

We can actually *simplify* the gradient computation (both notationally and computationally) substantially using matrix/vector notation

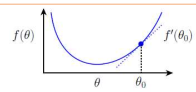
$$\begin{aligned}\frac{\partial E(\theta)}{\partial \theta_k} &= 2\sum_{i=1}^{m}\left(\sum_{j=1}^{n}\theta_j \cdot x_j^{(i)} - y^{(i)}\right)x_k^{(i)} \\ \Longleftrightarrow \nabla_\theta E(\theta) &= 2\sum_{i=1}^{m}x^{(i)}\left({x^{(i)}}^T\theta - y^{(i)}\right)\end{aligned}$$

Putting things in this form also make it more clear how to analytically find the optimal solution for last squares

48

Solving Least Squares

Gradient also gives a condition for optimality:
- Gradient must equal zero



Solving for $\nabla_\theta E(\theta) = 0$:

$$2\sum_{i=1}^{m} x^{(i)}\left({x^{(i)}}^T\theta - y^{(i)}\right) = 0$$

$$\Rightarrow \left(\sum_{i=1}^{m} x^{(i)}{x^{(i)}}^T\right)\theta - \sum_{i=1}^{m} x^{(i)}y^{(i)} = 0$$

$$\Rightarrow \theta^\star = \left(\sum_{i=1}^{m} x^{(i)}{x^{(i)}}^T\right)^{-1}\left(\sum_{i=1}^{m} x^{(i)}y^{(i)}\right)$$

49

Matrix Notation

Let's define the matrices

$$X = \begin{bmatrix} - & {x^{(1)}}^T & - \\ - & {x^{(2)}}^T & - \\ & \vdots & \\ - & {x^{(m)}}^T & - \end{bmatrix}, \; y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

Then

$$\nabla_\theta E(\theta) = 2\sum_{i=1}^{m} x^{(i)}\left({x^{(i)}}^T\theta - y^{(i)}\right) = 2X^T(X\theta - y)$$

$$\Longrightarrow \theta^\star = (X^TX)^{-1}X^Ty$$

50

How to *build* a Regression Model in *Scikit-Learn*

51

Linear Regression in Scikit-Learn

scikit-learn

sklearn.linear_model.LinearRegression

```
class sklearn.linear_model.LinearRegression(*, fit_intercept=True, normalize=False, copy_X=True, n_jobs=None)
```

- Build an estimator model like LinearRegression()
- Use fit() function to ***Train*** the model with **Training Dataset**
- Use predict() function to ***Test/Evaluate*** the model with **Test Dataset**
- Use predict() function to make ***prediction/inference*** on **New Unseen Data**

52

Solve a real-life problem with a *Regression* using *Scikit-Learn*

53

Progression of Diabetes Dataset

- *Feature/Attributes*: First 10 columns are numeric predictive values -- *age, sex, body mass index, average blood pressure,* and *six blood serum measurements* were obtained for each of n = 442 diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline.
- *Target*: Column 11 is a *quantitative measure of disease progression one year after baseline*
- Total number of Instances: 442

Attribute/Feature Information:
- *Age*
- *Sex*
- *Body mass index*
- *Average blood pressure*
- *S1*
- *S2*
- *S3*
- *S4*
- *S5*
- *S6*

Note: Each of these 10 feature variables have been mean centered and scaled by the standard deviation times `n_samples` (i.e. the sum of squares of each column totals 1).

Source URL: https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html

54

Remember - p-Value and Model Coefficients???

p-values and their importance in interpreting regression results

$$Sales = \beta_0 + \beta_1 Price + \beta_2 AdExp + \beta_3 PromExp$$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	197798832.8	65932944	40.56262	1.0848E-08
Residual	20	32509212.11	1625461		
Total	23	230308045			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-25096.83	24859.61131	-1.009542	0.324773	-76953.0734	26759.408
Price ($)	-5055.27	526.3995537	-9.603484	6.22E-09	-6153.32009	-3957.22
Adexp ('000$)	648.61214	209.0048787	3.103335	0.005602	212.635603	1084.5887
Promexp ('000$)	1802.611	392.8485427	4.588565	0.000178	983.143256	2622.0787

H_0: $\beta_0 = 0$
H_A: $\beta_0 \neq 0$

H_0: $\beta_1 = 0$
H_A: $\beta_1 \neq 0$

$\boxed{H_0: \beta_2 = 0 \quad H_A: \beta_2 \neq 0}$

H_0: $\beta_3 = 0$
H_A: $\beta_3 \neq 0$

- Reject the Null hypothesis that $\beta_2 = 0$.
- Advertising expenditure is an important variable in explaining Sales.

If *p-value* is less than the level of significance (α default 0.05) – 95% confidence interval; we Reject the *Null Hypothesis*.

55

Let's Go to the *Coding Demo*

56

57

To be continued in the next session…..

58