Simple Regression Basics

A Quick Overview

Simple Regression Versus Correlation

- **Pearson Correlation analysis** quantifies the strength of the linear relationship between two continuous variables.
    - However, no distinction is made between dependent (target) versus independent (predictor) variable
- **Simple linear regression** defines (mathematically) the linear relationship between a *continuous* dependent (target) variable and a *continuous* predictor (independent or explanatory) variable.
    - Later you will see that we may use *categorical (nominal)* variables as predictors as well.

Simple Regression Objectives

The objectives of simple linear regression are to:

- Assess the significance of the predictor (independent) variable in **explaining** the variability or behavior of the response variable
- **Predict** the values of the response (dependent) variable given the values of the predictor (independent or explanatory) variable.

3

Simple Regression Example

- A new credit card marketer wants to predict the number of credit cards owned by a family
- Why would he like to do this?
  - The more number of credit cards a family owns, it may be likely that the family may be open to new credit card offers.
  - Assume the marketer has a prospect file of 1 million names. The file has names/addresses of each family as well as say the family size. What can the database marketer do if he is not willing to send a new credit card offer to everyone of the 1 million names.
    - If he knows the correlation between family size and number of credit cards is +0.25?
    - If he has an equation such as: Number of credit cards = 1.5 + 2 * Family Size

4

Understanding Simple Regression

- Suppose the marketer did not know about simple regression prediction
- But, he has data on a few customers about how many credit cards they own.
- How would he predict for the **prospect file of 1 million**, how many credit cards are owned by each prospect family?

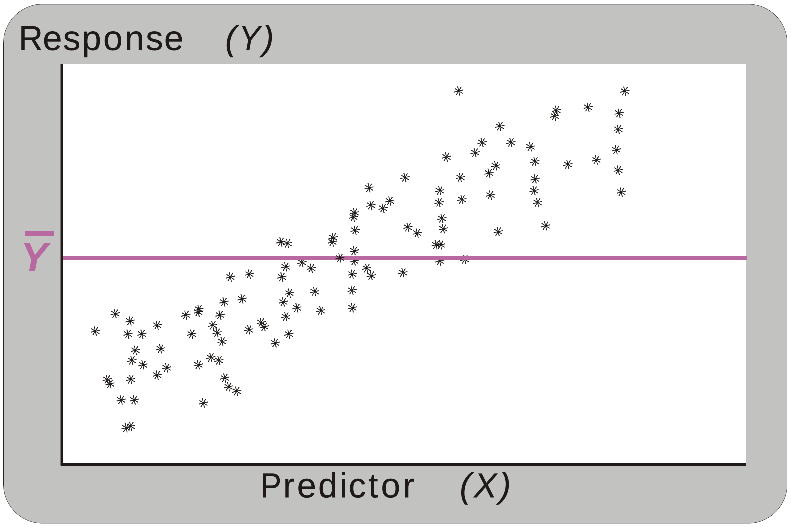
5

Credit Card Data (Dependent variable is Number of Credit Cards (CC))

Family ID	Num. CC	Baseline Prediction	Error	Square Error
1	4	7	-3	9
2	6	7	-1	1
3	6	7	-1	1
4	7	7	0	0
5	8	7	+1	1
6	7	7	0	0
7	8	7	+1	1
8	10	7	+3	9
Total	**56**			**22**

Average no. of CC used = 56/8 = 7
So, if we have no other information, our best prediction for number of CC owned by a family would be **7.**
How good is the prediction of 7 cards for each household? How do we quantify?
The key question is: can we do better than the baseline prediction, if we have independent variable(s) that can reduce the square error in prediction?

The Baseline Model



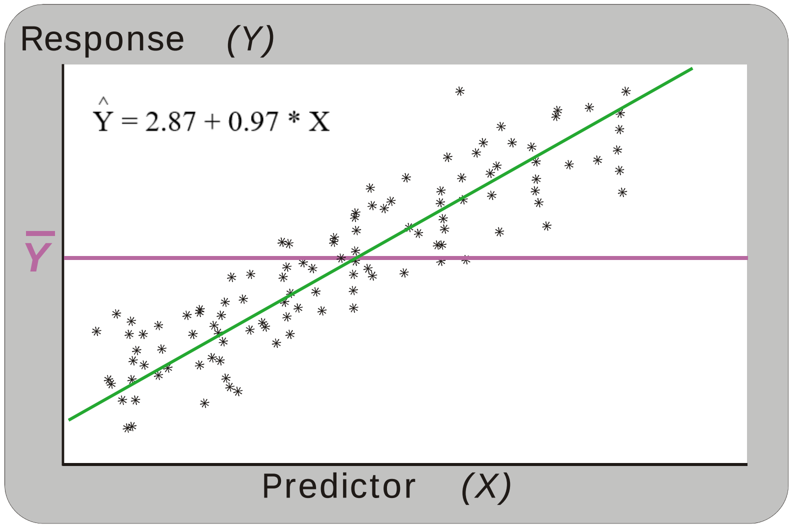
7

Does Family Size Help us with Prediction?

- Assume we have data on family size (X).
- Let's say we come up with a model (shown as prediction equation) using family size. Did it do better than our baseline prediction?
  - Prediction equation, $\hat{Y} = 2.87 + 0.97 * X$

Key Question: How much better? Can we quantify it?

The Baseline Model versus Regression Model



9

Simple Regression

Mechanics and Interpretation

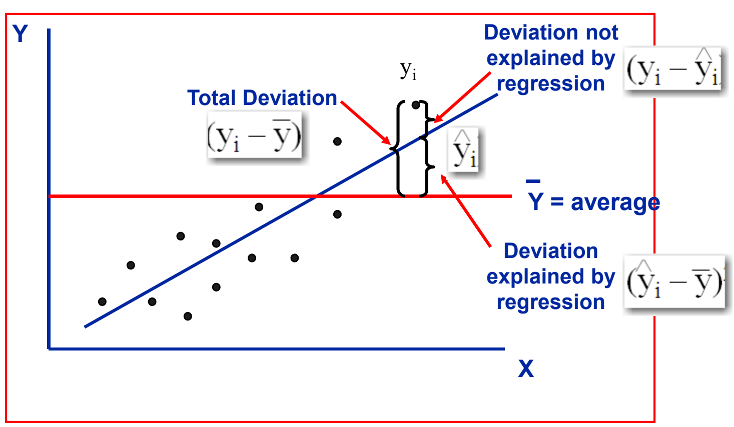
10

Linear Regression Model

- Regression Model (Equation): $Y_i = \beta_0 + \beta_1 X_i + e_i$
  - Y_i is the ‘ith’ value for the dependent variable
  - X_i is the ‘ith’ value for the independent variable
  - β_0 and β_1 are the **intercept** and **slope** of the regression line
  - e_i is the error associated with the equation representing relationship between Y and X

11

Least Squares Regression Line and Explanation



$$\hat{Y} = 2.87 + 0.97 * X$$

Family ID	Number of Credit Cards Used	Family Size	Simple Regression Prediction	Prediction Error	Prediction Error Squared
1	4	2	4.81	−.81	.66
2	6	2	4.81	1.19	1.42
3	6	4	6.75	−.75	.56
4	7	4	6.75	.25	.06
5	8	5	7.72	.28	.08
6	7	5	7.72	−.72	.52
7	8	6	8.69	−.69	.48
8	10	6	8.69	1.31	1.72
Total					5.50

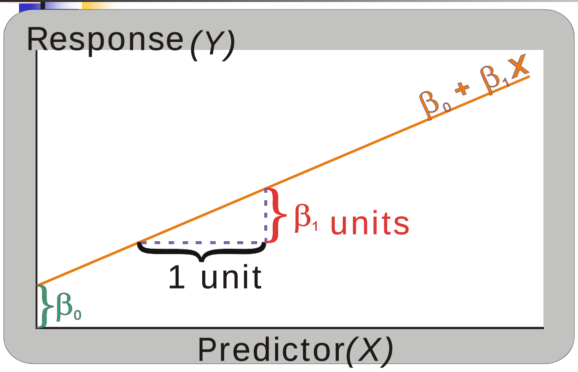
Assume we have data on family size. Let's say we come up with a model (shown as prediction equation) using family size. Did it do better than our baseline prediction?

Key Question: How much better? Can we quantify it?

SST = SSE + SSR

- SST (Total Deviation) = $\Sigma (y_i - \bar{y})^2$: what does it mean?
- SSE (Unexplained) = $\Sigma (\hat{y}_i - y_i)^2$: what does it mean?
- SSR (Explained by regression) = $\Sigma (\hat{y}_i - \bar{y})^2$: what does this mean?
- In this example, SST = 22, SSE = 5.5 and SSR = 16.5
- Coefficient of determination R^2 = (SSR/SST) = (SST-SSE)/SST = 16.5/22 = 75%
- What does it mean to have R^2 = 0% or 75% or 100%

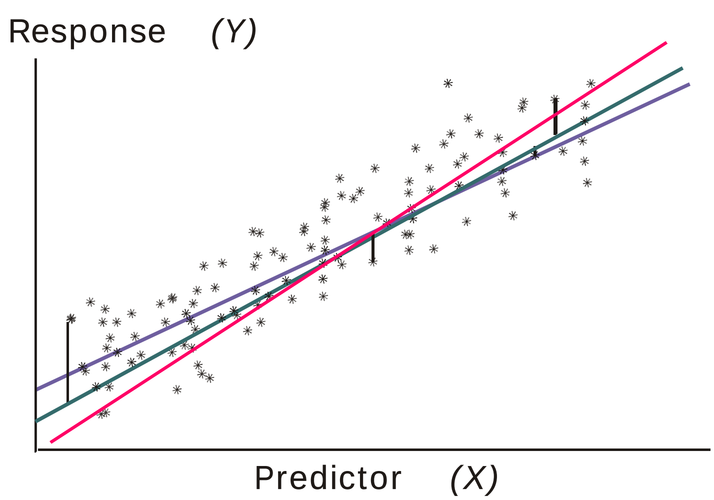
Linear Regression Model Interpretation



- Interpretation of regression coefficients:
  - β_0 represents the value of Y when X is 0 (!)
  - β_1 represents how much Y will change by when X is changed by one unit.
  - Positive β_1 means, as X increases by 1 unit, Y also increases by β_1 units.
  - Negative β_1 means, as X increases by 1 unit, Y decreases by β_1 units

15

Method of Least Squares – Which is the best Line?



16

Mechanics of Regression

- Regression coefficients are determined by the method of least squares.
- The idea is relatively simple. We want to find values for β_0 and β_1 such that the line ($Y = \beta_0 + \beta_1$ x) best fits the sample data.
- The 'best fit' is defined as the line for which the sum of squared vertical distances (SSE) of all sample points from that line is minimized
  - Mathematically that means we will differentiate the SSE with respect to X and set that equal to 0 to solve for the regression coefficients (β_0 and β_1)
- We will leave the actual mathematical calculation to be handled by computer programs and focus on understanding

The least squares estimates of β_0 and β_1 are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

17

Simple Regression Interpretation

- Regression output provides two sets of statistical tests:
  - Test for overall model (the Analysis of Variance, ANOVA table)
  - Test for each coefficient (the Parameter Estimates table)
- Hypothesis for overall model test:
  - H_0: The regression model does not explain the relationship between the dependent and the independent variable in the population any better than the baseline model.
  - H_1: H_0 is not true.
  - Decision about H_0 in hypothesis test for overall model is made using p-value based on F-statistic in the ANOVA Table
- Think of what it means if you can not reject **this null** hypothesis

18

Simple Regression Interpretation (Contd.)

- Hypothesis for test of coefficients:
  - H_0: The regression coefficient for X (β_1) equals 0 in the population.
  - H_1: H_0 is not true.
- Decision about H_0 in hypothesis test for coefficients is made based on the p-value based on t-statistic

19

Simple Regression Interpretation (contd.)

- R-Square (values between 0, 1) in simple regression provides a summary measure of variance (uncertainty) explained in the dependent variable
- Rules of thumb: 0-10% low, 10-50% moderate, 50% or more high.

20

Simple Regression

Basic Demonstration using JMP

Data Set: Ecommerce

- The data is a sample from customers of an ecommerce company. Variables and descriptions are given below:

Columns (9/0)

- ID
- Spend_thisyr
- Age
- Gender
- HomeOwner
- HomeValue
- Income
- Spend_lastyr
- Number_of_emails

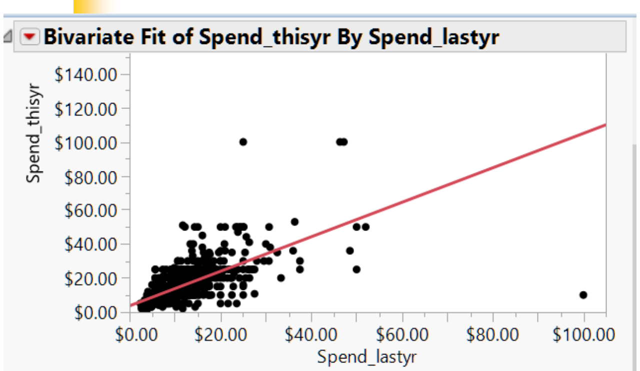
ID : Customer identification number
Spend_thisyr: Amount spent by customer this year ($)
Age: Age in years
Gender : M (Male), F (Female)
HomeOwner: Owner or Renter
HomeValue: Value of home ($)
Income : Annual Income ($)
Spend_lastyr: Amount spent by customer last year ($)
Number_of_emails: Number of emails sent to customer

Questions and JMP Procedure

- Questions:
  - How well does spend_lastyr explain spend_this yr?
  - How well does spend_lastyr predict spend_this yr?
- JMP> Analyze >Fit Y by X> Spend_thisyr as Y, Response > Spend_lastyr as X, Factor>Run
- Red triangle next to Bivariate Fit…> Fit line

23

Results



Linear Fit

Linear Fit

Spend_thisyr = 3.6895388 + 1.0134793*Spend_lastyr

Summary of Fit

RSquare	0.304305
RSquare Adj	0.303608
Root Mean Square Error	10.8761
Mean of Response	15.72191
Observations (or Sum Wgts)	1000

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	51637.67	51637.7	436.5364
Error	998	118052.89	118.3	Prob > F
C. Total	999	169690.56		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.6895388	0.670776	5.50	<.0001*
Spend_lastyr	1.0134793	0.048507	20.89	<.0001*

24

Results (contd.)

- Hypothesis for overall model test:
  - H_0: The regression model **does not explain** the relationship between the dependent and the independent variable in the population **any better than the baseline** model.
  - H_1: H_0 is not true.
  - Decision about H_0 in hypothesis test for overall model is made using p-value based on F-statistic in the ANOVA Table

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	51637.67	51637.7	436.5364
Error	998	118052.89	118.3	**Prob > F**
C. Total	999	169690.56		<.0001*

25

Results (contd.)

- Hypothesis for test of coefficients:
  - H_0: The regression coefficient for X (β_1) equals 0 in the population.
  - H_1: H_0 is not true.
- Decision about H_0 in hypothesis test for coefficients is made based on the p-value based on t-statistic

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.6895388	0.670776	5.50	<.0001*
Spend_lastyr	1.0134793	0.048507	20.89	<.0001*

How do we explain estimate for Intercept and Spend_lastyr?

26

Simple Regression

Prediction and Diagnostics

27

Prediction Using Regression equation

- Regression equation, $Y = \beta_0 + \beta_1$ x
- Once the values of intercept and slope (β_0 , β_1) are known, it is simple to calculate predicted value of Y (dependent or target variable) for any given value of X (independent variable).
- Some issues to keep in mind are:
  - Be careful about going beyond the range of X-values observed in the data that ere used to calculate the values of intercept and slope (β_0 , β_1).
    - Rule of thumb: 25% beyond observed range may be OK.
  - However, note that we are working with sample numbers and hence the values of the regression parameters (intercept and slope) will change from sample to sample!
    - Best to use confidence intervals for predictions
    - Software will do the calculations – you just ask for it!

28

Diagnosing Regression Model Performance

- Once a regression model is run, we get the regression equation ($Y = \beta_0 + \beta_1 X$).
- Then, this equation is used to predict Y for each observation by plugging-in the X-value for each observation.
- The difference between the actual Y-values and the predicted Y-values is called the residual (or, error)
  - The residual is calculated for each observation.
- These residuals are the primary tools for diagnosing regression model performance.

29

More on Residuals

- Large residual for an observation means that the model is not predicting well for that observation
  - That could be a cause for concern and perhaps need more exploration
- But, how do we know what is large?
  - Use standardized (studentized) residuals
  - If these are more than 3, then there may be cause for concern?
    - Why?

30

What to Do with Really Large Residuals

- The observation with a large residual may be **an outlier**
- We need to try to figure out why the model is not working for this observation
  - If I have access to data, I will go back and first check if there was any error in data entry
  - If there was no error, I need to think hard if this observation should be retained for the analysis (is there something peculiar about this observation)?
  - At the very least, I will rerun the regression by deleting the most severe outlier and compare results between the two regressions (with and without outliers)
    - If the results are similar then perhaps we have less to worry about.

31

Simple Regression Demo

Prediction and Diagnostics using JMP

32

Prediction

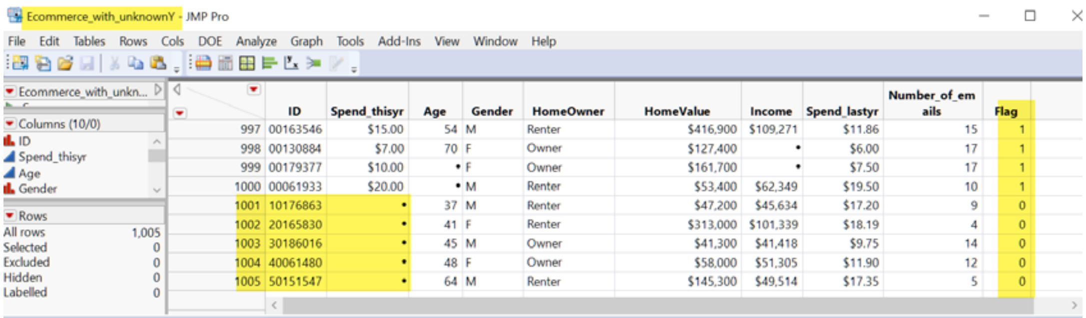
- Regression Model: Spend_thisyr = 3.6895388 +**1.01**34793*Spend_lastyr
- Suppose I want to predict how much customer A will spend this year, if I know A spent $10 last year
- Predicted Spend_thisyr for A = 3.6895388 +1.0134793*10 = 13.82
- What if I want to predict how much customer B will spend this year, if I know B spent $11 last year
- Predicted Spend_thisyr for B = 3.6895388 +1.0134793*11 = 14.83
- Difference between B and A is =14.83 - 13.82 = **1.01** or, coefficient of Spend_lastyr (within rounding error)!
- JMP: Red triangle next to Linear Fit > Save Predicted > Save Residuals

Confidence Intervals

- You will get two confidence intervals!
- JMP: Red triangle next to Linear Fit > Individual Confidence Limit Formula
  - For customer A, who spent $10 last year, this is the confidence interval for prediction of Spend_thisyr
- JMP: Red triangle next to Linear Fit > Mean Confidence Limit Formula
  - This is the confidence interval for prediction of **mean** of Spend_thisyr for *all those* who spent $10 last year
- Question for you to ponder: Which of the above two confidence intervals will be wider?

How About Making Predictions on Data Where You Don't Know the Target (Dependent) variable?

- We can do it two different ways:
  - By adding the new data to the current table and identifying those new data as not to be used in modeling but to be predicted
  - By running codes (scripts) on an external data set
- Data set: Ecommerce_with_unknownY (with a new Flag variable)

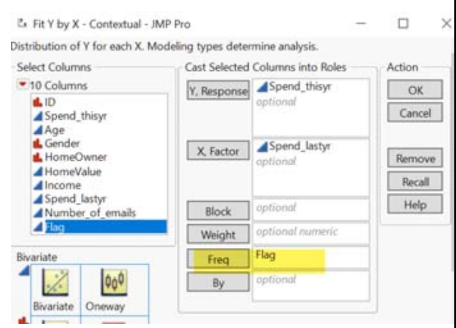


	ID	Spend_thisyr	Age	Gender	HomeOwner	HomeValue	Income	Spend_lastyr	Number_of_emails	Flag
997	00163546	$15.00	54	M	Renter	$416,900	$109,271	$11.86	15	1
998	00130884	$7.00	70	F	Owner	$127,400	•	$6.00	17	1
999	00179377	$10.00	•	F	Owner	$161,700	•	$7.50	17	1
1000	00061933	$20.00	•	M	Renter	$53,400	$62,349	$19.50	10	1
1001	10176863	•	37	M	Renter	$47,200	$45,634	$17.20	9	0
1002	20165830	•	41	F	Renter	$313,000	$101,339	$18.19	4	0
1003	30186016	•	45	M	Owner	$41,300	$41,418	$9.75	14	0
1004	40061480	•	48	F	Owner	$58,000	$51,305	$11.90	12	0
1005	50151547	•	64	M	Renter	$145,300	$49,514	$17.35	5	0

35

Making Predictions on New Data

- Open the new data in JMP
- JMP > JMP> Analyze >Fit Y by X> Spend_thisyr as Y, Response > Spend_lastyr as X, Factor> Select Flag and move it to Freq > OK
- Red triangle next to Bivariate Fit…> Fit line
- JMP: Red triangle next to Linear Fit > Save Predicted > Save Residuals
- JMP: Red triangle next to Linear Fit > Mean Confidence Limit Formula
- JMP: Red triangle next to Linear Fit > Individual Confidence Limit Formula



36

Identifying Large Residuals

- JMP: Red triangle next to Linear Fit > Save Studentized residuals
- JMP: Analyze > Distribution > Studentized Residuals as Y, Columns > OK
- What should we do at this point?

37

Pearson Correlation Coefficient and Simple Regression R-Square

- Run Pearson correlation between Spend_lastyr and Spend_thisyr
  - Correlation is 0.5516
- Regression R-square is 0.3043

38