M8S3 - Applied Regression

Professor Jarad Niemi

STAT 226 - Iowa State University

December 6, 2018

Regression analysis procedure

1. Determine scientific question, i.e. why are you collecting data
2. Collect data (at least two variables per individual)
3. Identify explanatory and response variables
4. Plot the data
5. Run regression
6. Assess regression assumptions
7. Interpret regression output

Two examples:

- Inflation vs Unemployment
- Frozen Foods: Sales vs Visibility

Inflation vs Unemployment

Definition

Inflation is a systained increase in the price level of goods and services in an economy over a period of time. Unemployment percentage is calculated by dividing the number of unemployed individuals by all individuals currently in the labor force.

Scientific question:

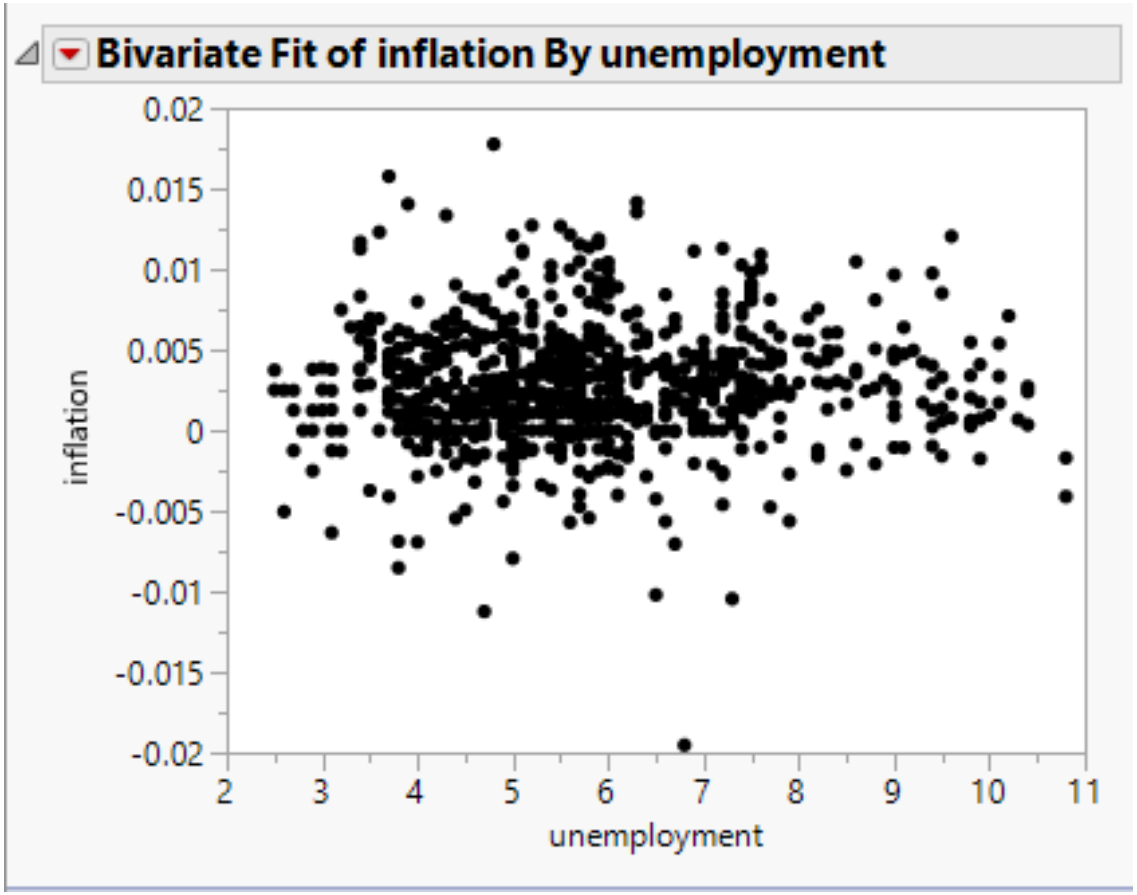
- What is the relationship between inflation and unemployment?
- Economic theory suggests lower unemployment leads to higher inflation. Is there evidence in the U.S. to support this theory?

Data

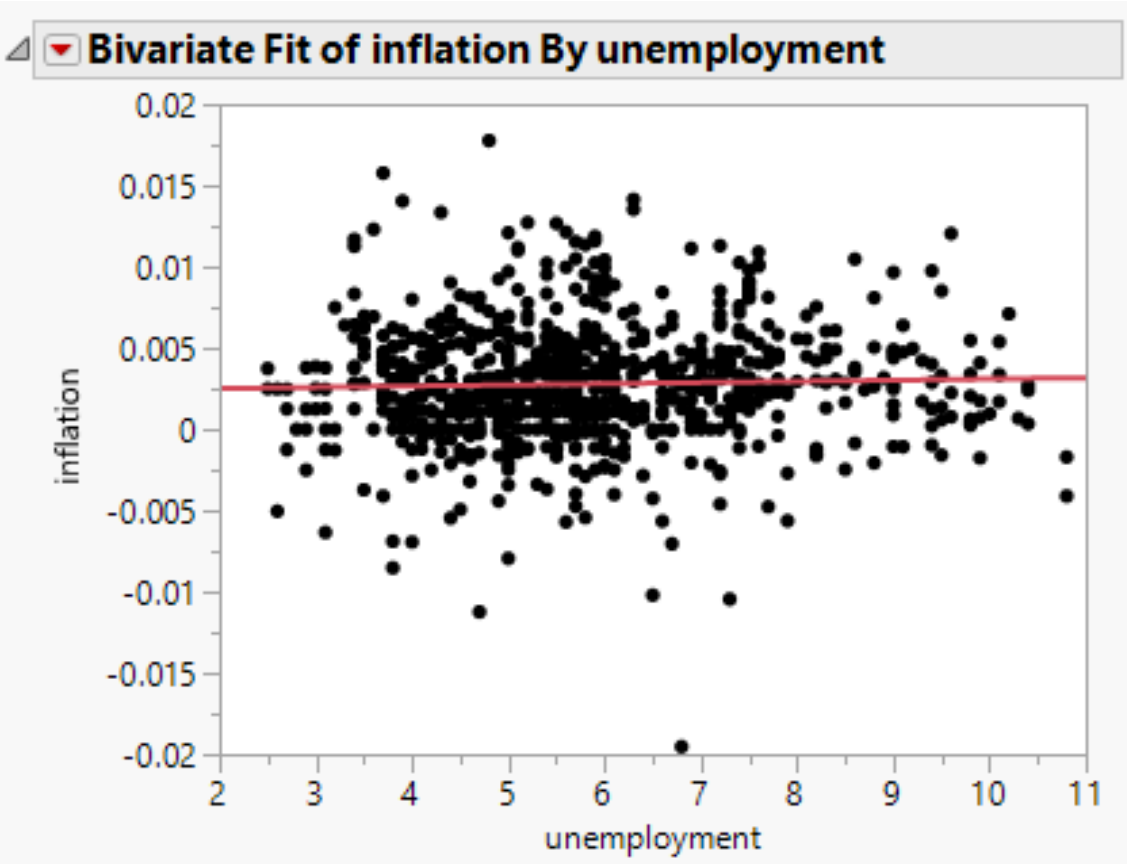
Obtained from `https://www.bls.gov/`:

	Year	month	inflation	unemployment
1	1948	Jan	0.0112676056	3.4
2	1948	Feb	-0.008522727	3.8
3	1948	Mar	-0.002849003	4
4	1948	Apr	0.0140449438	3.9
5	1948	May	0.0069735007	3.5
6	1948	Jun	0.0069252078	3.6
7	1948	Jul	0.0123119015	3.6
8	1948	Aug	0.0040871935	3.9
9	1948	Sep	0	3.8
10	1948	Oct	-0.004103967	3.7
11	1948	Nov	-0.006887052	3.8
12	1948	Dec	-0.006934813	4
13	1949	Jan	-0.001388889	4.3
14	1949	Feb	-0.011235955	4.7

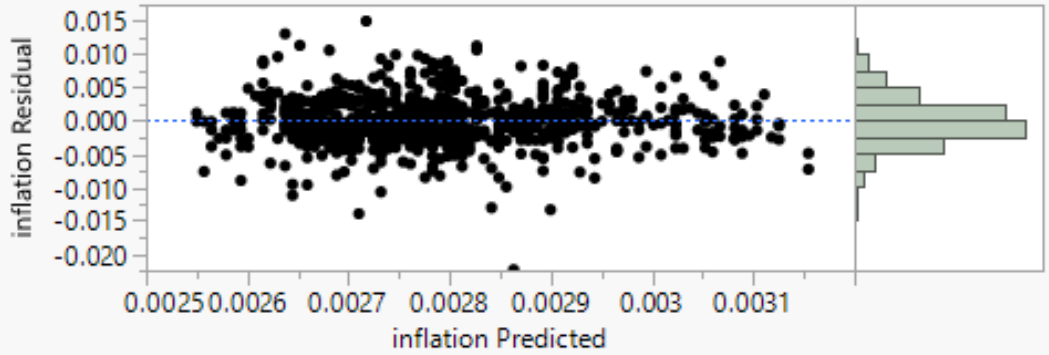
Plot


Bivariate Fit of inflation By unemployment
0.02
0.015
0.01
0.005
0
-0.005
-0.01
-0.015
-0.02
inflation
2
3
4
5
6
7
8
9
10
11
unemployment

Regression

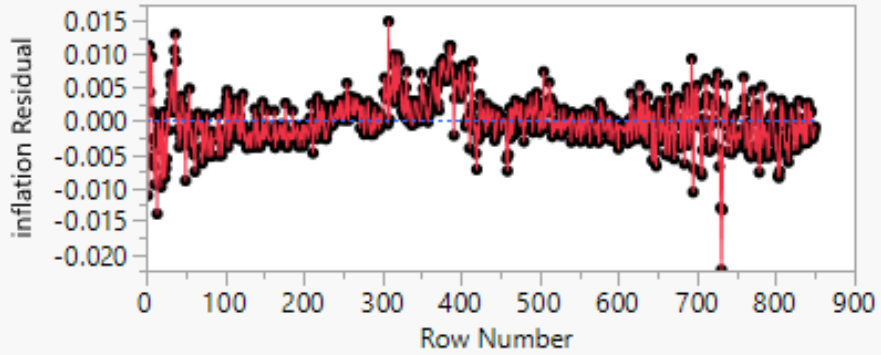

Bivariate Fit of inflation By unemployment
0.02
0.015
0.01
0.005
0
-0.005
-0.01
-0.015
-0.02
inflation
2
3
4
5
6
7
8
9
10
11
unemployment

Residual by Predicted Plot

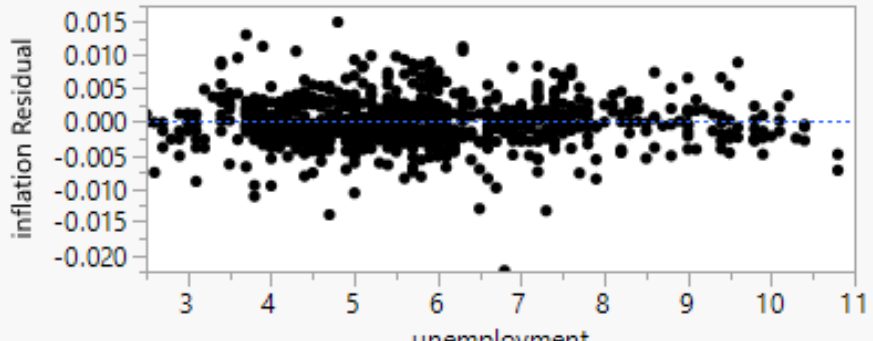


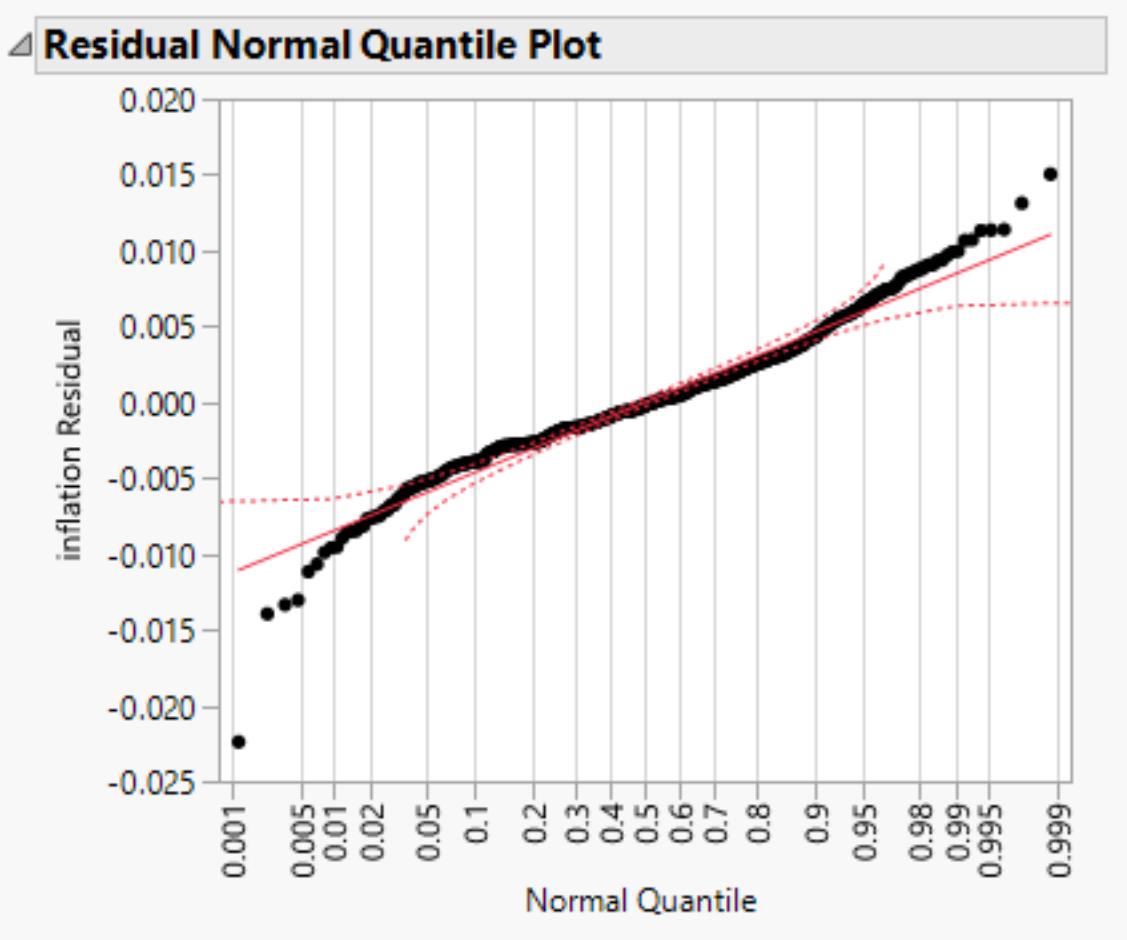
Actual by Predicted Plot

Residual by Row Plot



Residual by X Plot




Residual Normal Quantile Plot
inflation Residual
0.020
0.015
0.010
0.005
0.000
-0.005
-0.010
-0.015
-0.020
-0.025
0.001
0.005
0.01
0.02
0.05
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.95
0.98
0.99
0.995
0.999
Normal Quantile

Regression

Linear Fit

inflation = 0.0023679 + 7.2832e-5*unemployment

Summary of Fit

RSquare	0.001076
RSquare Adj	-0.0001
Root Mean Square Error	0.003636
Mean of Response	0.002788
Observations (or Sum Wgts)	850

Lack Of Fit

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0023679	0.000457	5.18	<.0001*
unemployment	7.2832e-5	7.621e-5	0.96	0.3395

Confidence intervals

Critical value for 80% confidence interval

$$t_{848,0.1} < t_{100,0.1} = 1.29$$

Intercept

$$0.0023679 \pm 1.29 \times 0.000457 = (0.0018, 0.0030)$$

Slope

$$0.000072832 \pm 1.29 \times 0.00007621 = (-0.000025, 0.000171)$$

Hypothesis tests

Scientific question: Economic theory suggests lower unemployment leads to higher inflation. Is there evidence in the U.S. to support this theory?

Hypothesis test:

$$H_0 : \beta_1 = 0 \qquad \text{vs} \qquad H_a : \beta_1 < 0$$

The point estimate for the slope (7.3e-5) is not consistent with this alternative hypothesis. Thus the p-value for this hypothesis test is $1 - (0.3395/2) \approx 0.83$.

Sales vs Visibility

Definition

Item_Outlet_Sales is the sales revenue for the particular product at a particular outlet for a given period of time. Item_Visibility is the % of total display area of all products in a store allocated to the particular product.

Scientific question:

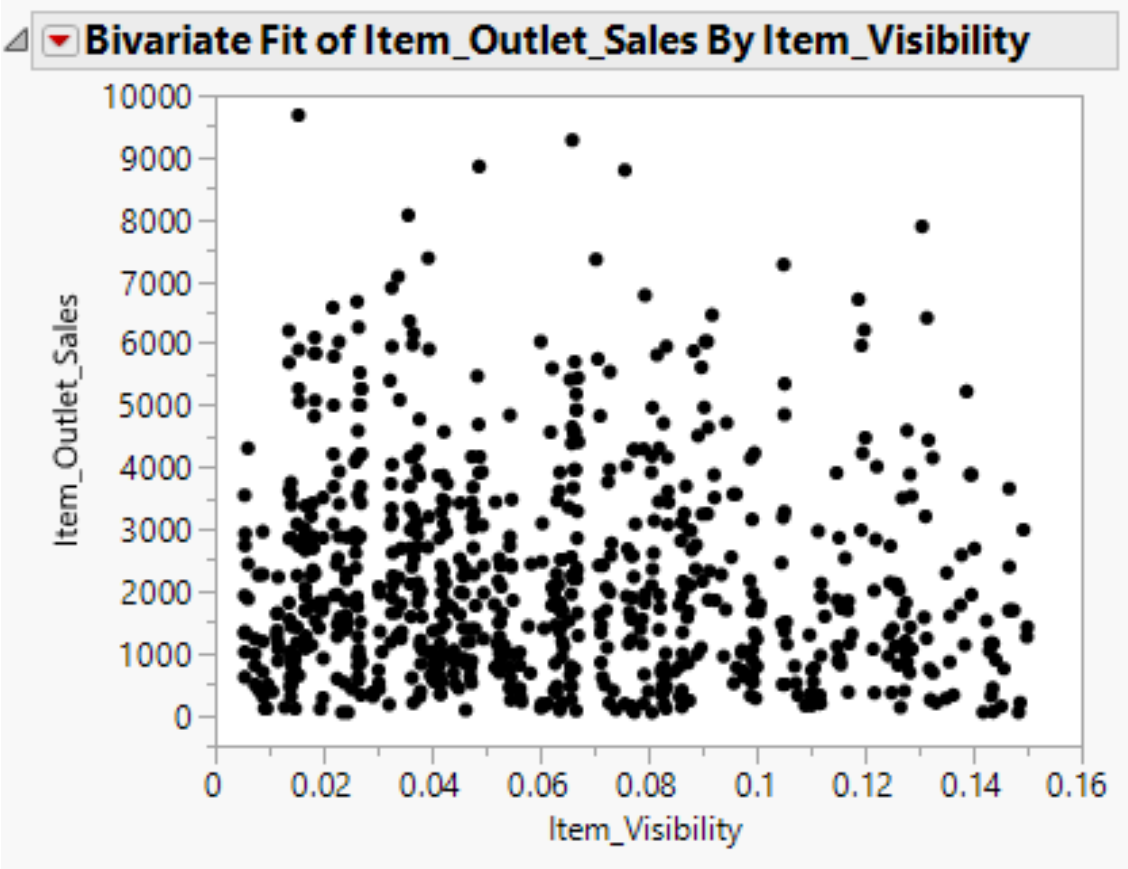
- What is the relationship between visibility and sales for frozen foods?
- Marketing theory suggests that increased visibility should increase sales.

Data

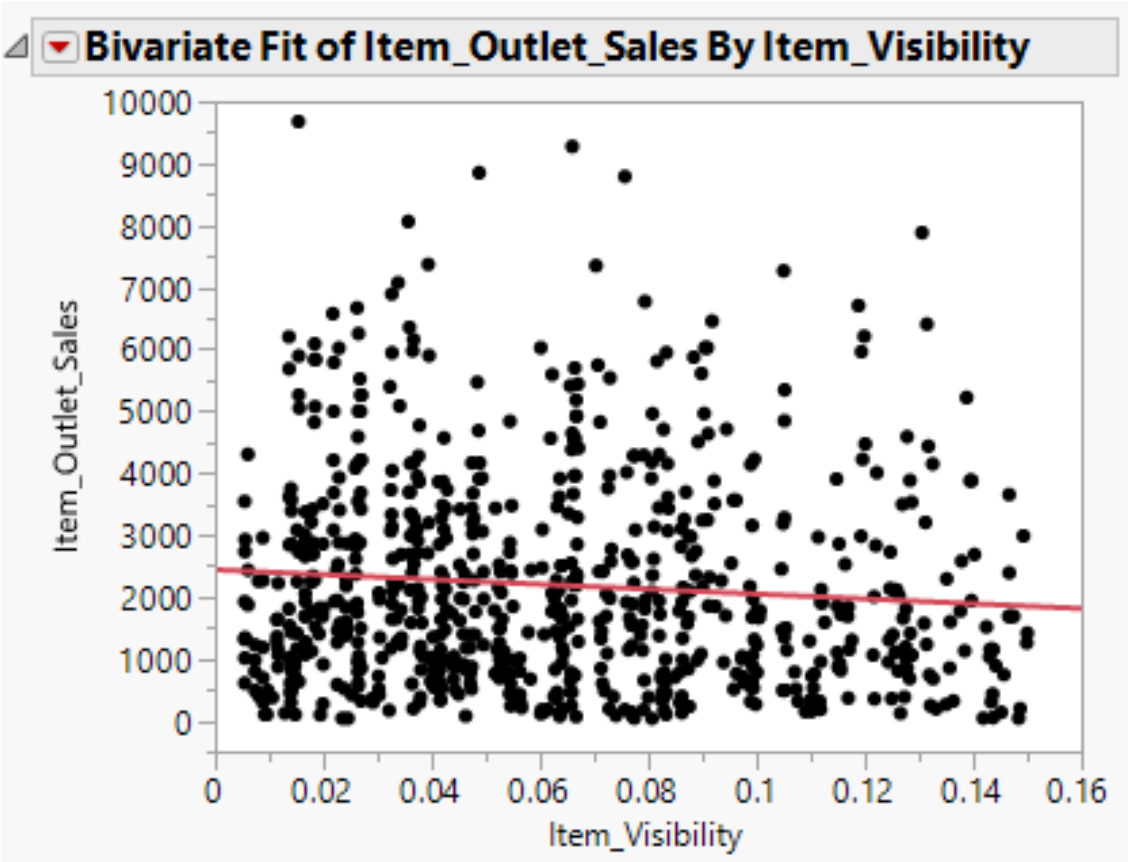
Obtained from `https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/`:

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
1	FDH17	16.2	Regular	0.016687114	Frozen Foods	96.9726	OUT045	2002	NA	Tier 2	Supermarket ...	1076.5986
2	FDU28	19.2	Regular	0.09444959	Frozen Foods	187.8214	OUT017	2007	NA	Tier 2	Supermarket ...	4710.535
3	FDR28	13.85	Regular	0.025896485	Frozen Foods	165.021	OUT046	1997	Small	Tier 1	Supermarket ...	4078.025
4	FDC29	8.39	Regular	0.024205661	Frozen Foods	114.0176	OUT046	1997	Small	Tier 1	Supermarket ...	2290.352
5	FDL04	19	Low Fat	0.112556507	Frozen Foods	104.9622	OUT017	2007	NA	Tier 2	Supermarket ...	1587.933
6	FDN04	11.8	reg	0.014087057	Frozen Foods	180.3344	OUT046	1997	Small	Tier 1	Supermarket ...	1427.4752
7	FDU04	•	Low Fat	0.009714595	Frozen Foods	120.0414	OUT019	1985	Small	Tier 1	Grocery Store	487.3656
8	FDF41	12.15	Low Fat	0.131383762	Frozen Foods	246.046	OUT049	1999	Medium	Tier 1	Supermarket ...	1231.73
9	FDT28	13.3	Low Fat	0.063695084	Frozen Foods	151.0708	OUT045	2002	NA	Tier 2	Supermarket ...	1805.6496
10	FDS52	8.89	low fat	0.005505481	Frozen Foods	102.4016	OUT017	2007	NA	Tier 2	Supermarket ...	2732.4432
11	FDD17	7.5	Low Fat	0.032677678	Frozen Foods	239.0906	OUT049	1999	Medium	Tier 1	Supermarket ...	5942.265
12	FDL40	17.7	Low Fat	0.01161096	Frozen Foods	95.041	OUT035	2004	Small	Tier 2	Supermarket ...	868.869

Plot

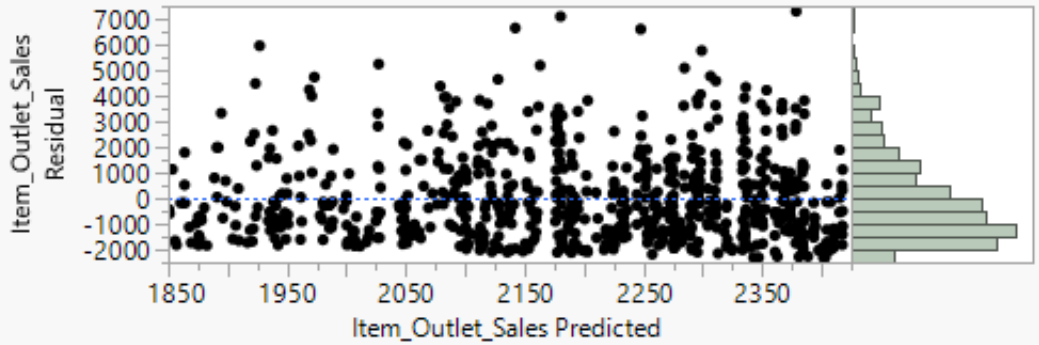

Bivariate Fit of Item_Outlet_Sales By Item_Visibility
Item_Outlet_Sales
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
Item_Visibility

Regression


Bivariate Fit of Item_Outlet_Sales By Item_Visibility
Item_Outlet_Sales
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
Item_Visibility

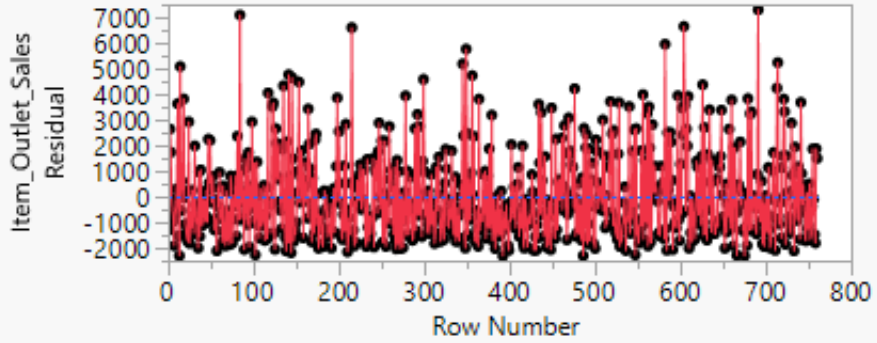
Diagnostics Plots

Residual by Predicted Plot

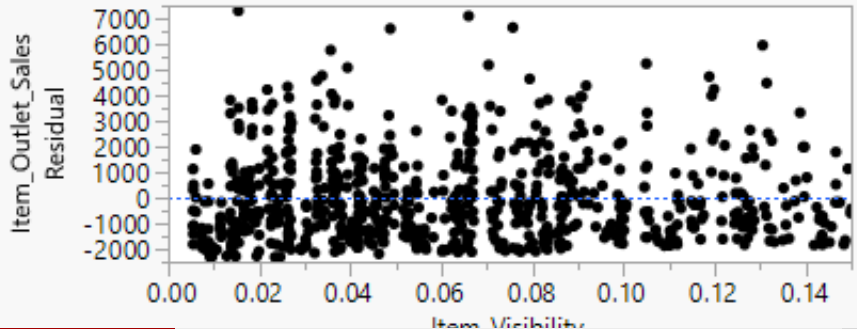


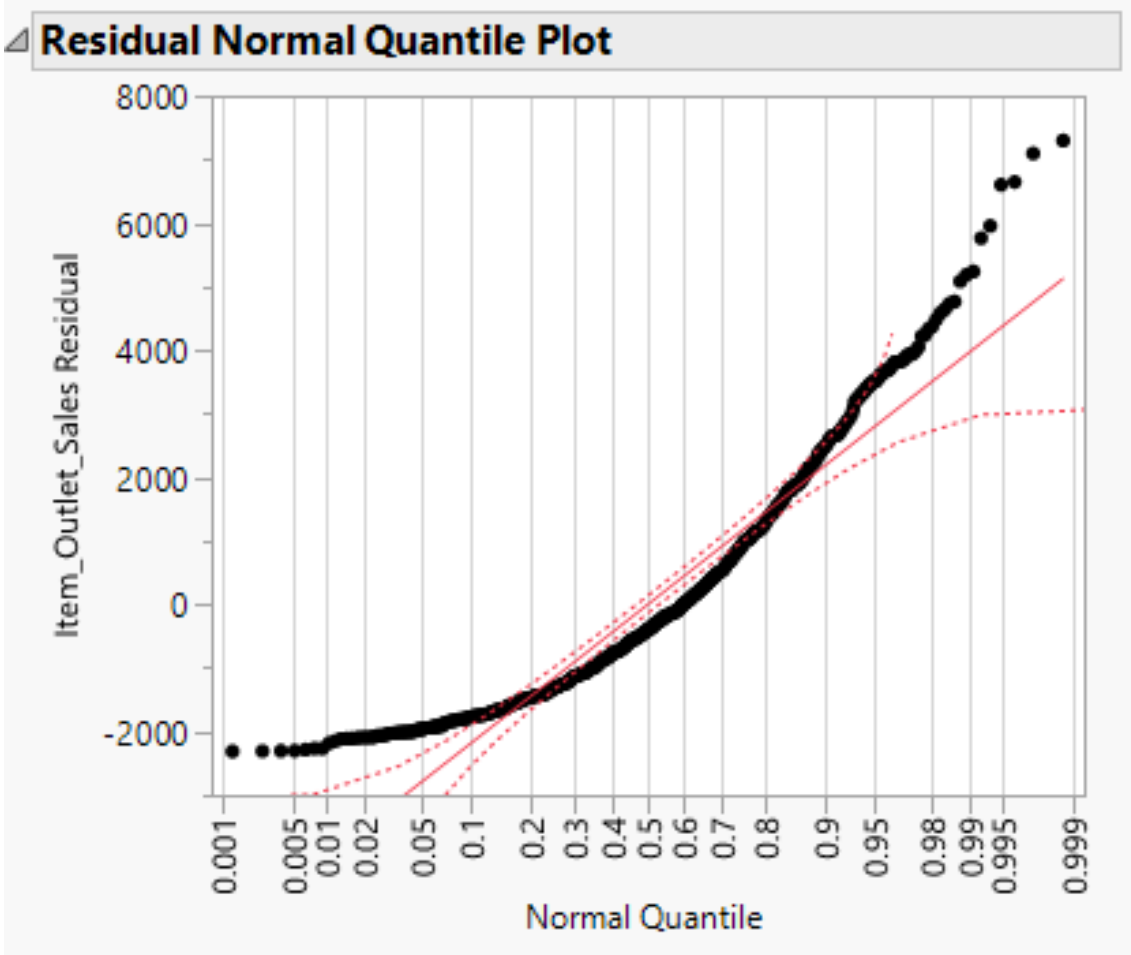
Actual by Predicted Plot

Residual by Row Plot



Residual by X Plot




Residual Normal Quantile Plot
Item_Outlet_Sales Residual
8000
6000
4000
2000
0
-2000
0.001
0.005
0.01
0.02
0.05
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.95
0.98
0.99
0.995
0.999
Normal Quantile

Regression

Linear Fit

Item_Outlet_Sales = 2439.0525 - 3923.0176*Item_Visibility

Summary of Fit

RSquare	0.007636
RSquare Adj	0.006327
Root Mean Square Error	1703.866
Mean of Response	2191.78
Observations (or Sum Wgts)	760

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2439.0525	119.5942	20.39	<.0001*
Item_Visibility	-3923.018	1624.367	-2.42	0.0160*

Confidence intervals

Critical value for 95% confidence interval

$$t_{758,0.1} < t_{100,0.1} = 1.984$$

Intercept

$$2439.0525 \pm 1.984 \times 119.5942 \approx (2200, 2680)$$

Slope

$$-3923.018 \pm 1.984 \times 1624.367 = (-7150, -700)$$

Hypothesis tests

Scientific question: Marketing theory suggests that increased visibility should increase sales.

Hypothesis test:

$$H_0 : \beta_1 = 0 \qquad \text{vs} \qquad H_a : \beta_1 > 0$$

The point estimate for the slope (-3923) is not consistent with this alternative hypothesis. Thus the p-value for this hypothesis test is $1 - (0.016/2) \approx 0.99$.