15.063: Communicating with Data

Summer 2003

MIT SLOAN SCHOOL OF MANAGEMENT

Recitation 6

Linear Regression

Today’s Content

- Linear Regression
- Multiple Regression
- Some Problems

Linear Regression

- Why?
- What is it?
- Pros?
- Cons?

Linear Regression

Data: (x_i, y_i) $i = 1, 2, \ldots, n$

Model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

- $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are i.i.d. random variables, $N(0, \sigma)$
- β_0 = "baseline" value of Y (i.e., value of Y if X is 0)
- β_1 = "slope" of line (change in Y per unit change in X)

Comments:

- $E(Y_i \mid x_i) = \beta_0 + \beta_1 x_i$
- Standard Deviation $(Y_i \mid x_i) = \sigma$
- Relationship is linear (described by a "line")

Linear Regression

The "best" regression line is the one that chooses b_0 and b_1 to minimize the residual sum of squares

Regression coefficients: b_0 and b_1 are <u>sample estimates</u> for β_0 and β_1

Estimate for Y_i : $\hat{y}_i = b_0 + b_1 x_i$ (predicted value)

Residual: $e_i = y_i - \hat{y}_i$ (observed - predicted)

Residual sum of squares: $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Problem 1: Linear Regression

Juan Martinez is an analyst at a leading mutual fund company. Juan has been asked to develop a regression model for **predicting the market value of a firm** using **total sales as input**.

A) Construct a simple linear regression model to predict the market value of a firm based on the data that Juan has gathered.

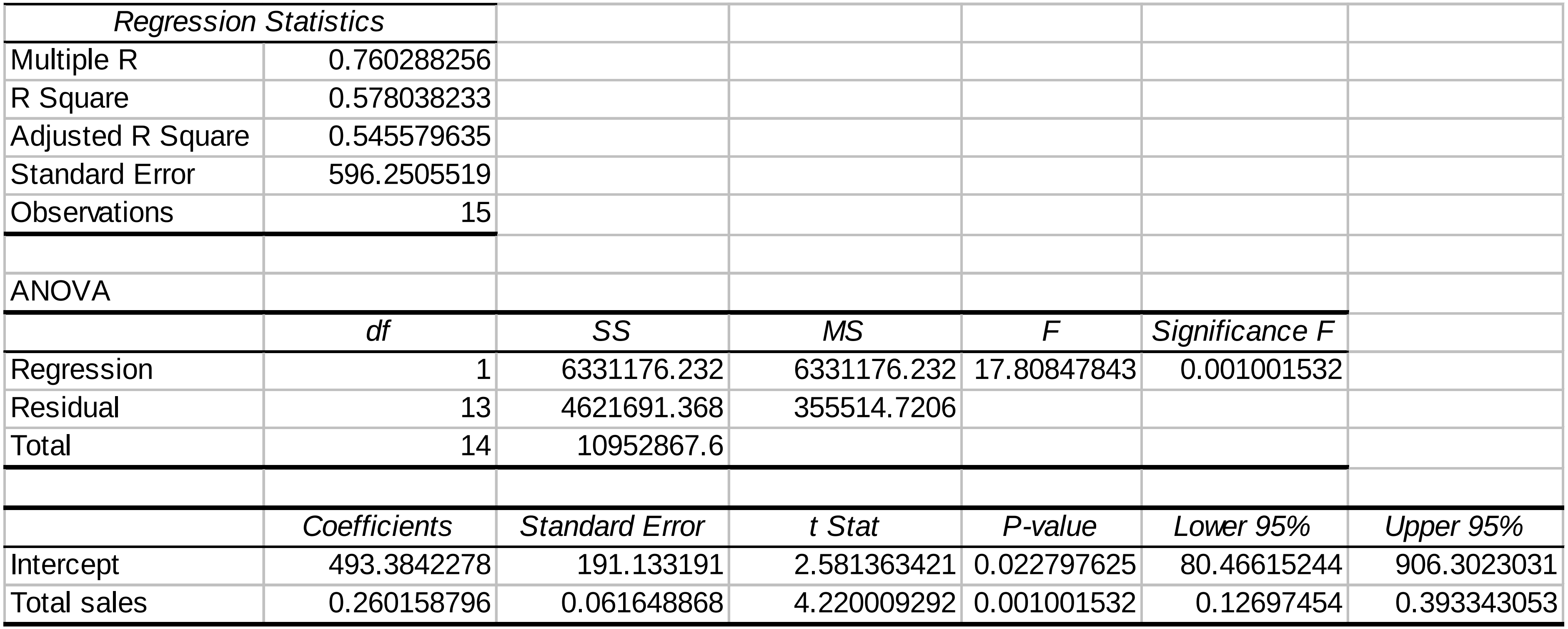
B) Examine the regression output of your regression model. Compute the 95% confidence interval for the regression coefficient.

C) Plot the relation as well as the residuals of the regression.

D) Suppose a firm has sales of $3,500 million. What is your prediction of the firm's market value?

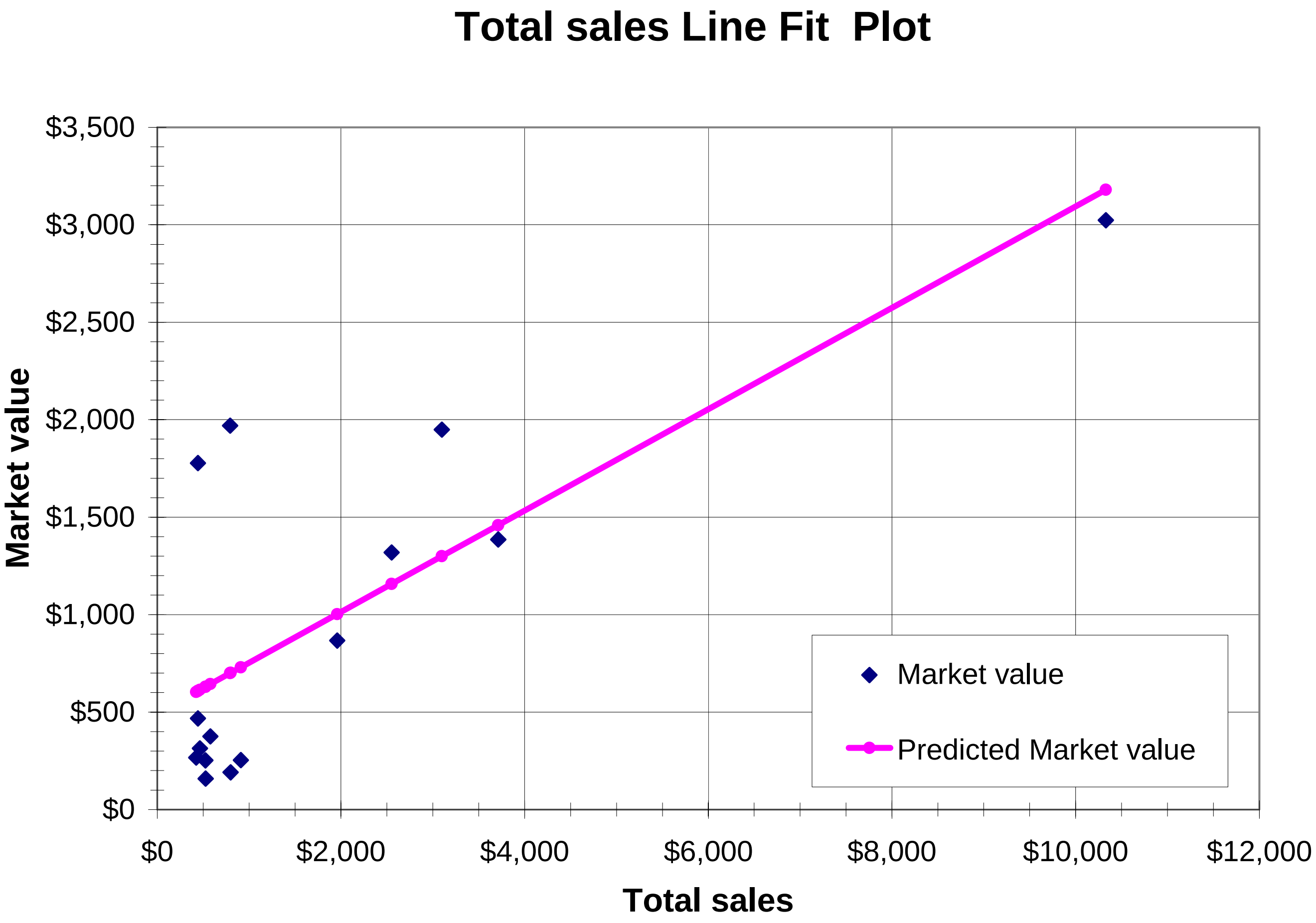
Problem 1: Linear Regression

- Output of regression model
- What are the 95% confidence interval for the regression coefficients?

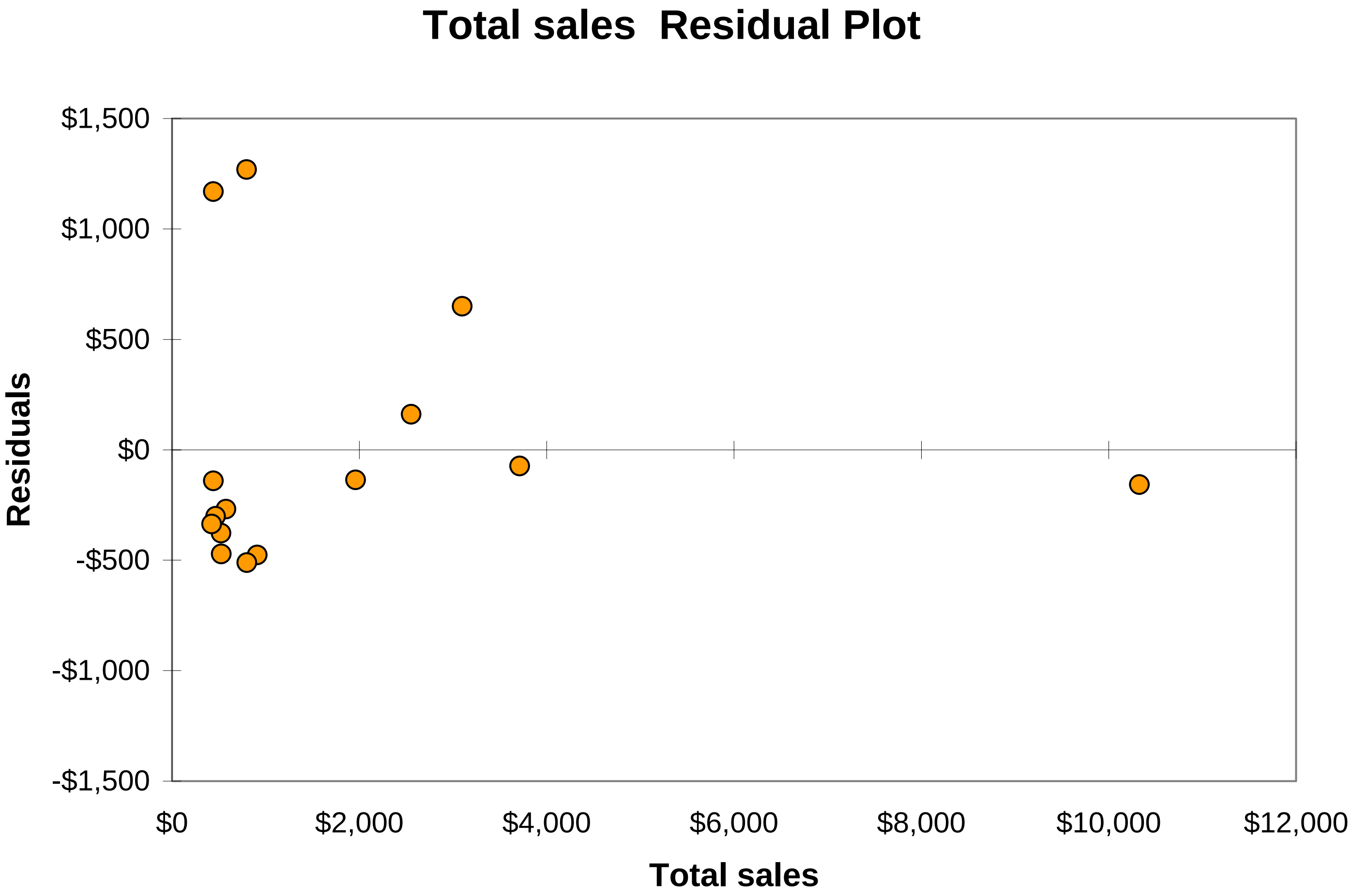


Regression Statistics						
Multiple R	0.760288256					
R Square	0.578038233					
Adjusted R Square	0.545579635					
Standard Error	596.2505519					
Observations	15					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	6331176.232	6331176.232	17.80847843	0.001001532	
Residual	13	4621691.368	355514.7206			
Total	14	10952867.6				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	493.3842278	191.133191	2.581363421	0.022797625	80.46615244	906.3023031
Total sales	0.260158796	0.061648868	4.220009292	0.001001532	0.12697454	0.393343053

Graphs for the Problem



Graphs for the Problem



Multiple Regression

In general, there are many factors (= k) that affect the dependent variable:

Independent variables: $x_{1i}, x_{2i}, \ . \ . \ . \ , \ x_{ki}$ $i = 1,2,...,n$

Dependent variables: Y_i $i = 1,2,...,n$

Model: $Y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + \varepsilon_i$

$\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are iid random variables, $N(0, \sigma)$

Goal: Choose $b_0, b_1, \ . \ . \ . \ , b_k$ to minimize the residual sum of squares

$\hat{y}_i = b_0 + b_1 x_{1i} + \ . \ . \ . \ + b_k x_{ki}$ $e_i = y_i - \hat{y}_i$

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

Regression Output

1) Regression coefficients: $b_0, b_1, \dots, b_k$

- sample estimates of $\beta_0, \beta_1, \dots, \beta_k$

2) Standard error, s: estimate of σ

- a measure of the amount of "noise" in the model

3) Degrees of freedom: $n - (k + 1) = n - k - 1$

- n pieces of data; used up (k + 1) degrees of freedom to form sample estimates $b_0, b_1, \dots, b_k$

Regression Output

4) Standard errors of the coefficients: s_{b0} , s_{b1} , ... , s_{bk}

- Used to get confidence intervals for β_m.
- Same role as the estimate “s” of the standard deviation of the sample mean.

5) t-Statistic: $t_m = \frac{b_m}{s_{b_m}}$

- A measure of the statistical significance of each individual x_m in accounting for the variability of Y.

Regression Output

6) Confidence Intervals:

- The α% confidence interval for β_m is

$$(b_m \; - \; c \times s_{bm}, \; b_m \; + \; c \times s_{bm}),$$

where c is such that P(- c < T < c) = α/100

- The test: *Is β_m significantly different from zero?*

can be answered by checking if 0 is inside the confidence interval.

Regression Output

7) Coefficient of determination: R^2

- Measures the amount of variation that is accounted for by the x variables:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$= 1 - \frac{\text{Variation not accounted for by x variables}}{\text{Total variation}}$$

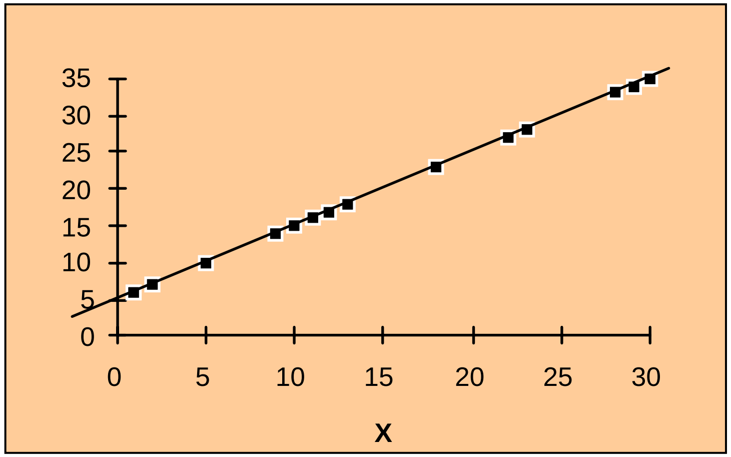
$$= \frac{\text{Variation that is accounted for by x variables}}{\text{Total variation}}$$

= Percentage of variation that is *accounted for* by the x variables

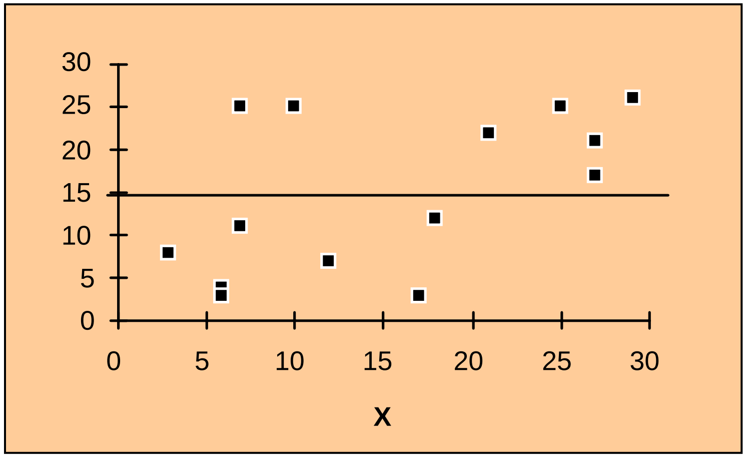
($\bar{y}$ is the sample mean of y_i's.)

Regression Output

- R^2 takes values between 0 and 1:



$R^2 = 1$



$R^2 = 0$

- $R^2 = \text{corr}(X,Y)^2$
- Measures the how good the fit was.

Validating the regression model

- $Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$ $(i = 1, \ldots, n)$
- $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are i.i.d. random variables with distrib $N(0,\sigma)$

1) Linearity

- If k = 1 (simple regression), can check visually from scatter plot

2) Solution makes sense

- The signs of the regression estimates follow intuition.
- The confidence intervals of the estimates don't include zero.

Validating the regression model

3) The R^2 of the regression is high enough

- A “good” value of R^2 depends on the situation (for example, the intended use of the regression, and complexity of the problem)

- Users of regression tend to be fixated on R^2, but it’s not the whole story. It is important that the regression model is “valid.”

4) Normality of ε_i

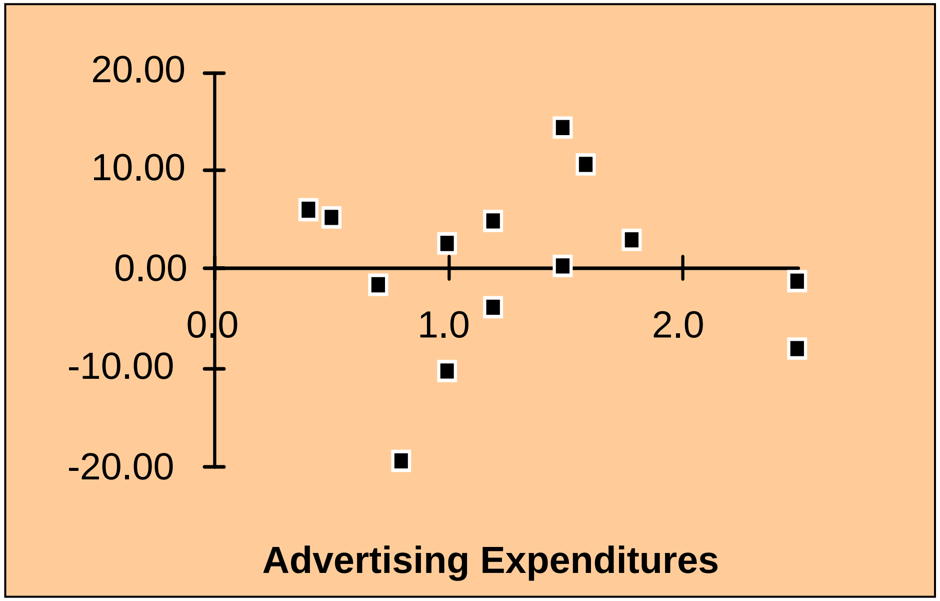
- Plot a histogram of residuals ($e_i = y_i - \hat{y}_i$)

- Usually, results are fairly robust with respect to this assumption.

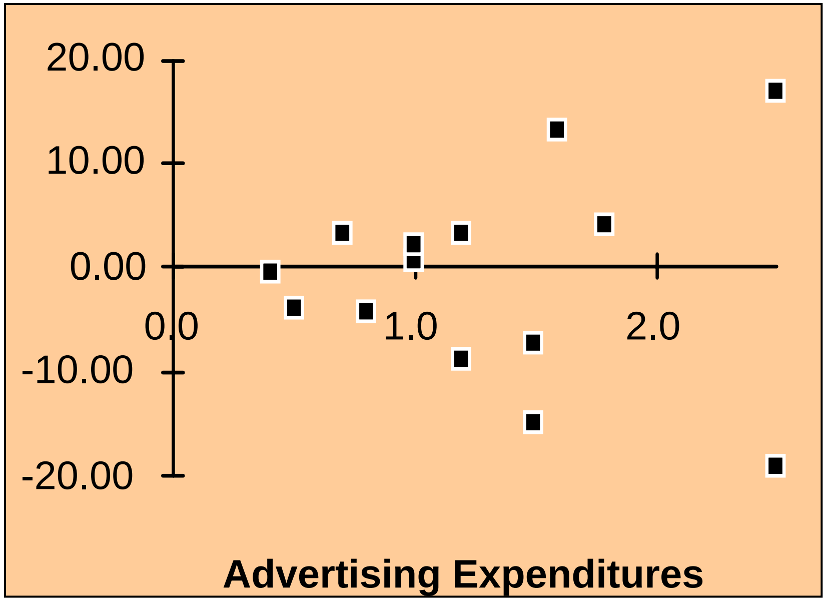
Validating the regression model

5) Heteroscedasticity

- Error terms have constant standard deviation? (i.e., Std Dev(ε_i) = σ for all i?)
- Check scatter plot of residuals vs. estimates and independent variables.



No evidence of heteroscedasticity

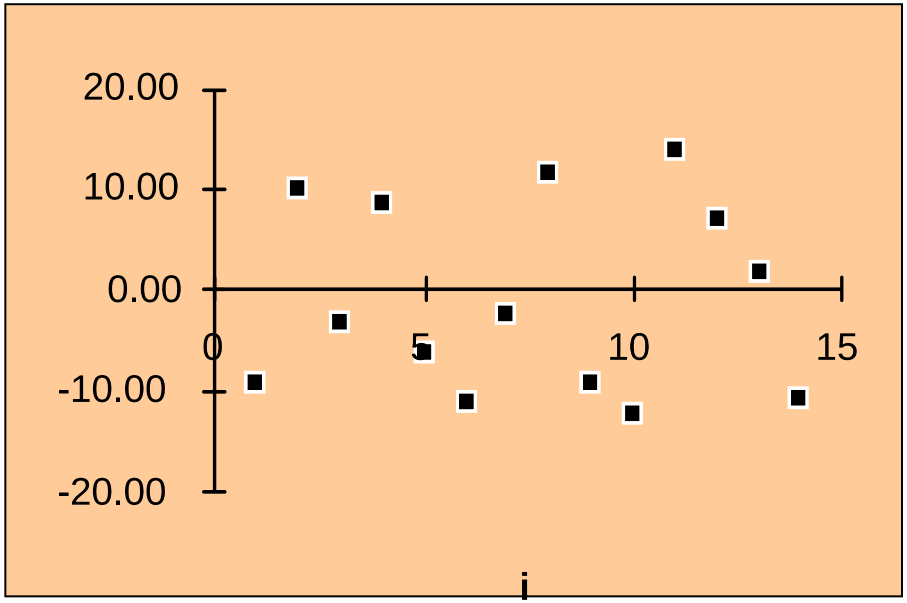


Evidence of heteroscedasticity

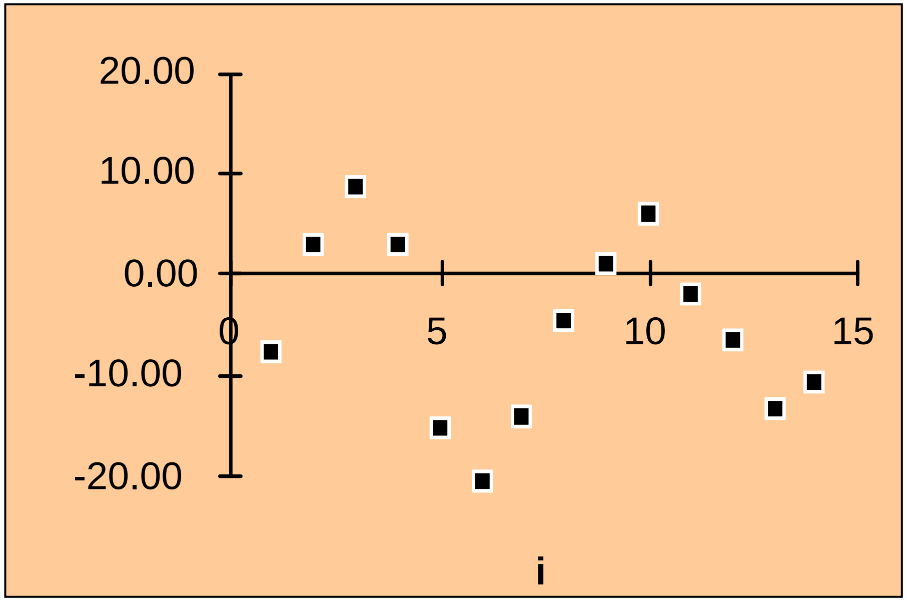
Validating the regression model

6) Autocorrelation

- Are error terms independent?
- Plot residuals in order, e.g., e_1, e_2, . . . , e_n, (a "time series plot") and check for patterns.



No evidence of autocorrelation



Evidence of autocorrelation

Validating the regression model

7) Multicolinearity

- If R^2 is high but a confidence interval of the estimate of an independent variable contains zero, it could mean that there is multicolinearity.
- Check the correlation matrix to see if one of the independent variables should be discarded.

Problem 2: Multiple Regression

Jack, the master brewer at Hubbard Brewing Co. (HBC), believes that beer **sales depend on ingredients, taste, advertising and investment**. He performed a multiple regression analysis from data of last 50 types of beer produced and sold by HBC. The variables are:

- H=Hops (keeps fresh, adds bitterness) (oz)
- M=Malt (adds sweetness, color and alcohol) (lbs.)
- A=Annual Advertising ($)
- B=Bitterness (rating 1=not bitter to 10=very bitter)
- I=Initial investment ($)

Problem 2: Multiple Regression

(a) From Jack’s model, what is the formula for annual sales?

Sales = 37 H + 1,319 M + 0.049 A - 63 B + 53 I - 13,707

(\$000) (oz) (lb) (\$) (1-10) (M\$)

Problem 2: Multiple Regression

(b) Calculate 95% CI for each of the coefficients and interpret what each interval implies.

- 95% CI for H: $0 \in [-48 , 122]$
- 95% CI for M: $0 \notin [994 , 1643]$
- 95% CI for A: $0 \notin [.04, .059]$
- 95% CI for B: $0 \in [-230, 104]$
- 95% CI for I: $0 \in [-215 , 321]$
- The two that remain have positive estimates, which is reasonable.

Problem 2: Multiple Regression

(c) Based on the computer output provided, and your results from part (b), evaluate the model. In particular, which variables are significant and how might you change the model, if at all.

Malt and Annual Advertising are statistically significant, since the 95% CI does not contain the value 0.

→ Consider the following modifications:

• drop Initial investment (non-significant),

• drop Bitterness (non-significant, and correlated with Malt, which is significant, and has higher correlation with Annual sales. Correlation with malt may create multicollinearity),

• drop Hops (non-significant).

Then re-fit the two-factor regression model ...

Problem 2: Multiple Regression

(d) Jack has come up with a new beer that he believes will be the most bitter ale ever sold: Final Examber Ale. The recipe calls for 13oz of hops and 7lbs of malt and will register 10 on the bitterness scale. Jack would like to spend $150,000 on advertising and $700,000 on initial investments. Why might you not want to use the model to estimate sales for this new beer?

1) Bitterness of 10 exceeds the highest value in the data set used for the model. This can be a problem. Therefore, extrapolating beyond the range of data could yield to misleading results.

2) As discussed in (c) this is not the best model as it contains non-significant and correlated variables...

Problem 2: Multiple Regression

(e) Jack wants to use this model to evaluate four new beers. Jack currently has only $2.5M for initial investments and he wants to produce the beers that maximize annual sales. Which combination should he produce? You will need to compute the expected sales for the new beers. Indicate Jack's expected annual sales.

1) Expected sales for Wang's Great Ale =

37 x 12 + 1,319 x 8 + .049 x 150,000 - 63 x 6 + 53 x 1.2 - 13,707 = $4,377

2) The combination that maximizes expected annual sales such that total initial investment is less than $2.5 M is to produce:

Wang's Great Ale	I=1.2	Exp. Sales = $4,377
Sloan Stout	I=1.0	Exp. Sales = $2,935
Total	I=2.2	Exp. Sales = $7,312

Checklist for evaluating a regression model

- **Linearity.** If there is only one independent variable, construct a scatter-plot of the data to check for linearity. Otherwise, use common sense to decide if a linear relationship is reasonable.

- **Signs of Regression Coefficients.** Check to see that the signs make intuitive sense.

- **Confidence intervals.** Check that the confidence intervals don’t contain zero.

- **R^2.** Check if the value of R^2 is reasonably high.

- **Normality.** Check that the residuals are approximately Normally distributed by constructing a histogram of residuals.

Checklist for evaluating a regression model

- **Heteroscedasticity.** Check for heteroscedasticity of the residuals by plotting the residuals with the observed values of each of the independent variables.

- **Autocorrelation.** If the data are time-dependent, plot the residuals over time to check for any apparent patterns.

- **Multicolinearity.** If you suspect that two independent variables are correlated, run a regression on the observations of these variables to see if they are strongly correlated. Examine a scatter-plot of these two variables.

The End.