Stat 102

Introduction to Business Statistics

Class 14

Richard P. Waterman

Wharton

Table of contents I

Today's module

Topics to be covered in this module:

- Last time
- The usual suspects: R^2, RMSE
- Adjusted-R^2
- Prediction in multiple regression
- Checking assumptions in multiple regression
- Inference and the model building process
- Summary
- Next time

Last time

- Motivation for multiple regression
- Marginal v. partial association
- Interpretation of regression coefficients
- The scatterplot matrix and Mahalanobis plot
- The path diagram

R^2, RMSE

- R^2 is now: the proportion of variability in Y explained by the regression model.
- RMSE still measures the standard deviation of the residuals.
- But now, the equation for RMSE is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-(k+1)}},$$

where k is the number of variables in the model.

- The same relationship between R^2 and RMSE still holds:

$$R^2 \approx 1 - \frac{RMSE^2}{s_y^2}.$$

A critique of R^2

- R^2 always [1]goes up when you add additional variables to the model.
- Even if you throw junk into the model, R^2 is going up.
- Therefore chasing R^2 is a fools errand: it will encourage you to put more and more variables into the model until R^2 equals 100%.
- Adjusted R^2 is a version of R^2 that includes a penalty term for model complexity – adjusted R^2 can go down as variables are added to a model.
- I will not make a big deal about adjusted R^2 because the model with the highest adjusted R^2 is the same as the one with the smallest RMSE.
- Though we prefer models with low RMSE, if you throw junk into the model then RMSE can in fact increase and penalize you for doing so, and that is good.

[1]Technically, it can never goes down, it could stay the same. You will never see this in practice though.

Predictions and residuals

- As in simple regression, it is the case that:

$$y_i = \hat{y}_i + e_i.$$

- This says that the regression decomposes the data into two pieces, the signal $\hat{y}_i$ and the noise e_i.
- You can [2]save predicted values and residuals through the "save" option in the multiple regression output.
- By hand: for a car that weights 300lbs and has 300 horsepower what do you predict the GPM1000M˙City to be?
-

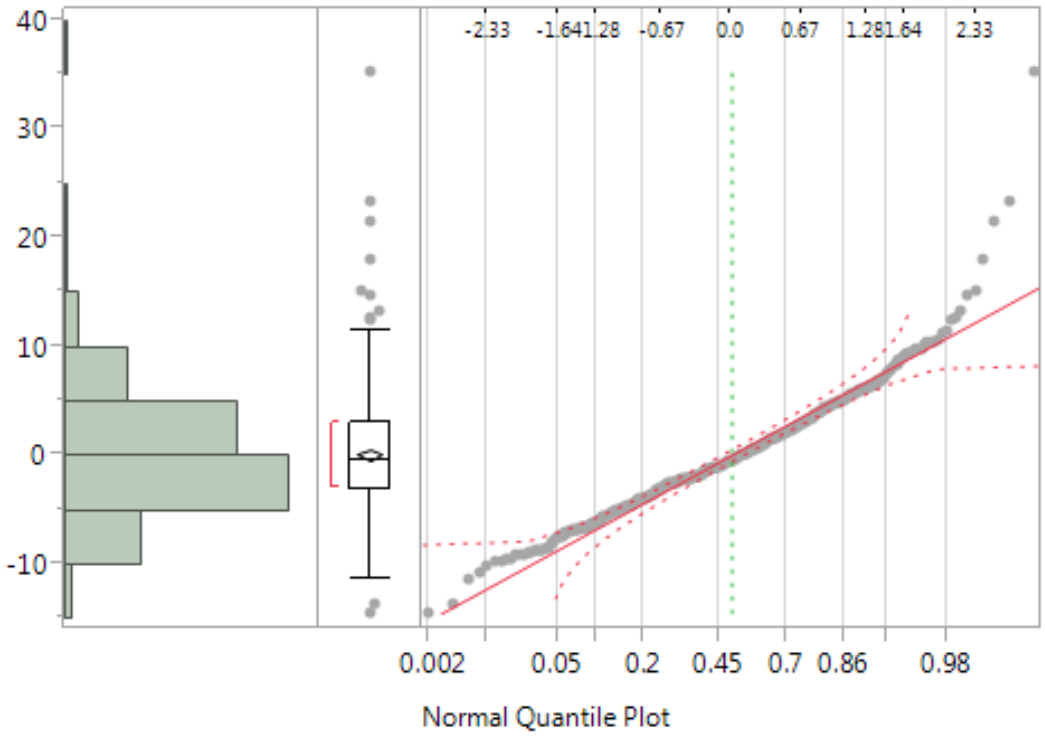
$$10.079054 + 0.0079 \times 3000 + 0.08009 \times 300 = 57.883.$$

- Or, in real life: create a new row in the spreadsheet, add in 3000 and 300 for weight and horsepower and view the calculated prediction in the prediction column.

[2]Click on the title bar in the MRM output, choose "Save columns", then Prediction Formula, then Residuals.

Residual plots

- JMP creates a default residual plot of e_i against $\hat{y}_i$.
- If you save the residuals into the spreadsheet you can create residuals plots against any you choose, and create a normal quantile plot of the residuals.



The residuals are still a little right-skewed here.

Market segmentation

- A marketing project identified a list of affluent customers for a new phone. Should the company target promotion towards the younger or older members of this list?
- To answer this question, the marketing firm obtained a sample of 75 consumers and asked them to rate their likelihood of purchase on a scale of 1 to 10. The data is in smartphone.JMP where Age and Income of consumers were also recorded.

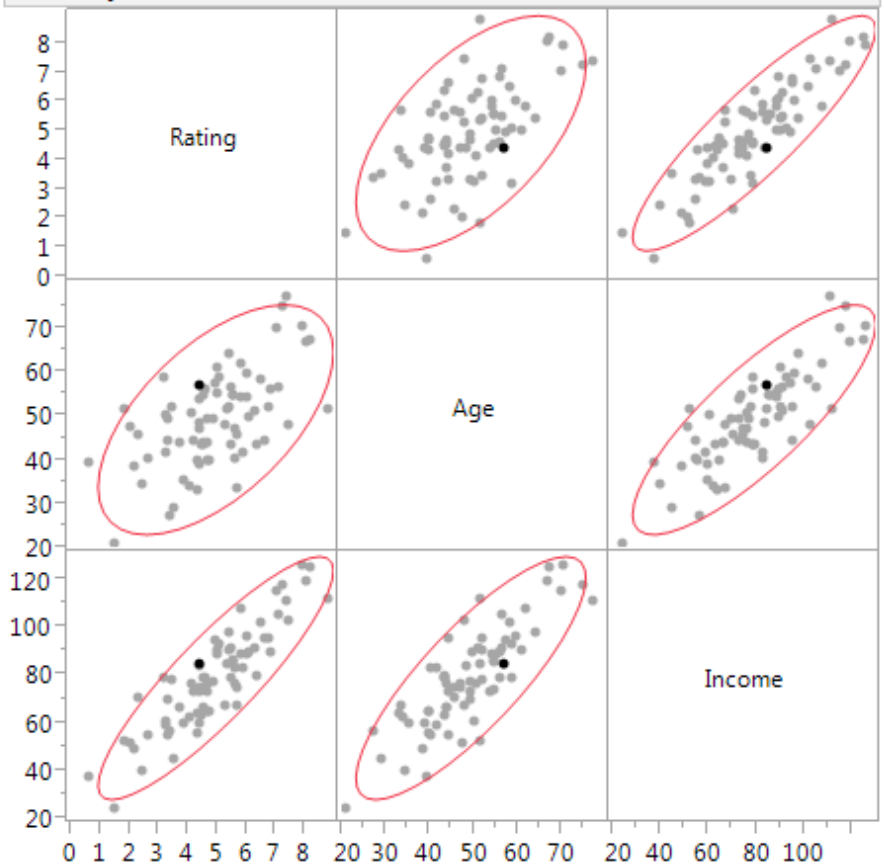
Market segmentation

Multivariate

Correlations

	Rating	Age	Income
Rating	1.0000	0.5867	0.8845
Age	0.5867	1.0000	0.8286
Income	0.8845	0.8286	1.0000

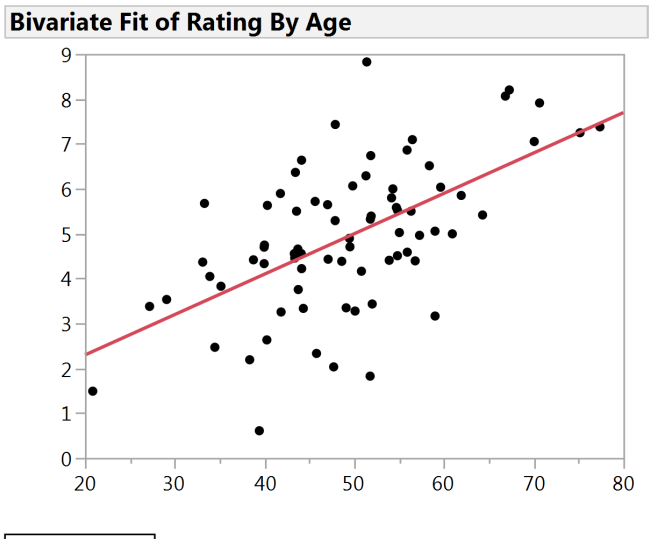
Scatterplot Matrix



Market segmentation

- Note that both age and income are positively correlated with rating.
- Older people tend to give higher ratings, so should we therefore market to the old?

The simple regression output of rating against age (it is positive and significant):



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.4923901	0.733017	0.67	0.5039
Age	0.0900056	0.014541	6.19	<.0001*

Market segmentation

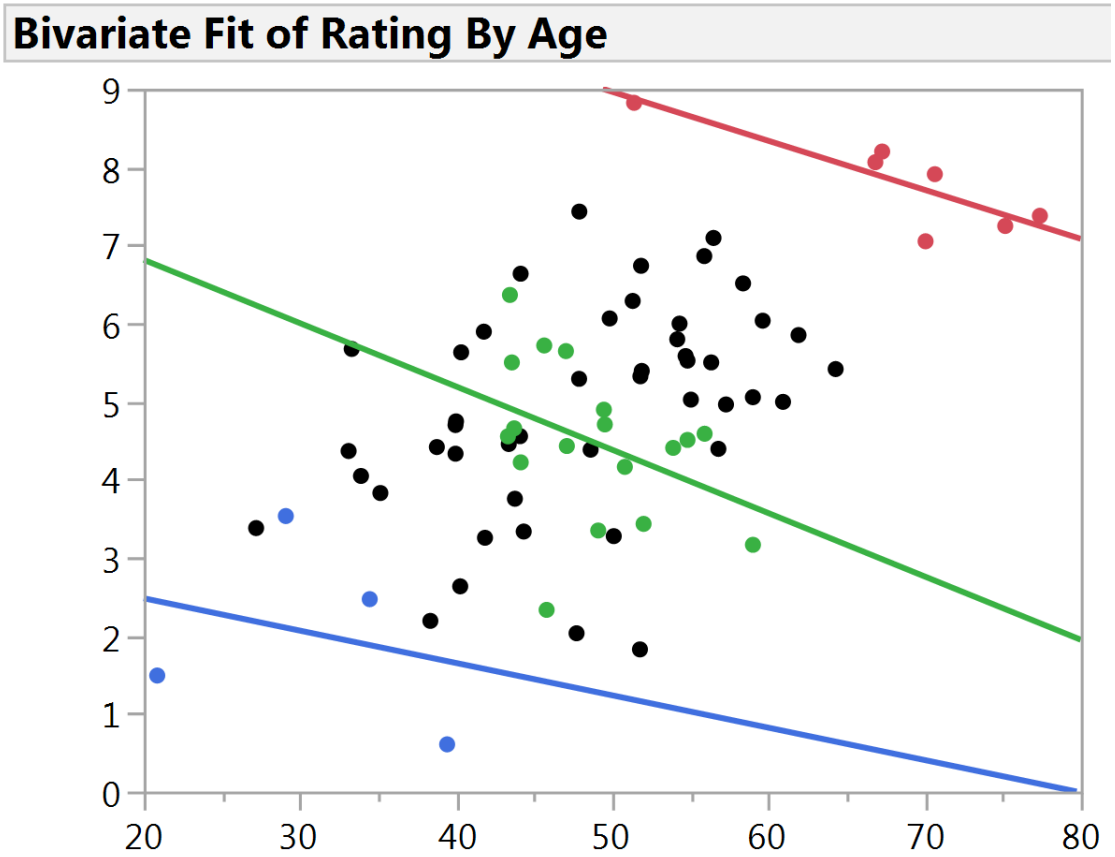
- The multiple regression of rating against age and income now has a significant partial regression coefficient for age, but it is negative!

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.517709	0.352139	1.47	0.1459
Age	-0.071628	0.012479	-5.74	<.0001*
Income	0.1006638	0.00644	15.63	<.0001*

The resolution

- "Controlling" for income using the "by" command:



- Within each income level there is a negative association with age.
- As there is a list of *affluent* customers, you should market to the younger on that list.

Hypothesis testing in multiple regression

Recall the regression assumptions. If they are seriously broken, then these tests could be misleading.
Three flavors. They all test whether slopes are equal to zero or not. They differ in the number of slopes we are looking at simultaneously.

1. Test **all** the regression coefficients at once.
2. Test a **single** regression coefficient (slope).
3. Test a **subset** of the regression coefficients (more than one, but not all of them – the Partial F-test).

Test **all** the regression coefficients at once

Stine: p. 611.

- It is sometimes called "the whole model test".
- Look for the F-statistic in the ANOVA table.
- The hypothesis test in English: do any of the X-variables in the model explain any of the variability in the Y-variable?
- It essentially tests whether the observed R^2 is significantly greater than 0.
- A small p-value says YES, big p-value says NO.
- Note that the test does not identify which variables are important.
- If you answer this question as NO then it's back to the drawing board – none of your variables are any good!

Test **all** the regression coefficients at once

Reading the p-value from the ANOVA table:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	79089.131	39544.6	1407.981
Error	477	13397.023	28.1	**Prob > F**
C. Total	479	92486.154		<.0001*

Test a **single** regression coefficient (slope)

Stine: p.613.

- Look for the t-statistic.
- The hypothesis test in English: does this variable add any explanatory power to the model that already includes all the other X-variables?
- Small p-value says YES, big p-value says NO.
- This is a "last in test". Does this variable, in the presence of the other add significant explanatory power to the model?
- Note: the order you enter variables in a regression has no impact on their t-statistics.

Test a **subset** of the regression coefficients

- A video on how to use the custom test dialog is on Canvas as is a PDF document.
- The test in English: do any of the X-variables in the **subset** under consideration explain any of the variability in Y?
- We read the p-value from the JMP output to answer this one.

You must be able to answer this question: "why not do a whole bunch of t-tests rather than one partial F-test?" Answer: the partial F-test is an **honest** simultaneous test. Honest means that it controls the family-wise Type I error rate.

The model fitting process

- There is a discussion in Stine in Ch. 23.5.
- I have added a document to Canvas, with a detailed discussion.

Model building philosophy in this course

- Keep it as simple as possible (parsimony).
- Make sure everything is interpretable (especially any transformations).
- After having met the above criteria go for biggest R-squared, smallest RMSE and the model that makes most sense (signs on regression slopes).

Next time

- The usual suspects: R^2, RMSE
- Adjusted-R^2
- Checking assumptions in multiple regression
- Inference and the model building process