Stat 102

Introduction to Business Statistics
Class 7

Richard P. Waterman

Wharton

Table of contents I

Today's module

Topics to be covered in this module:

- Last time
- The residual definition
- RMSE (s_e)
- Various residual plots
- R^2 and the quality of fit
- The relationship between R^2 and RMSE
- Summary
- Next time

Last time

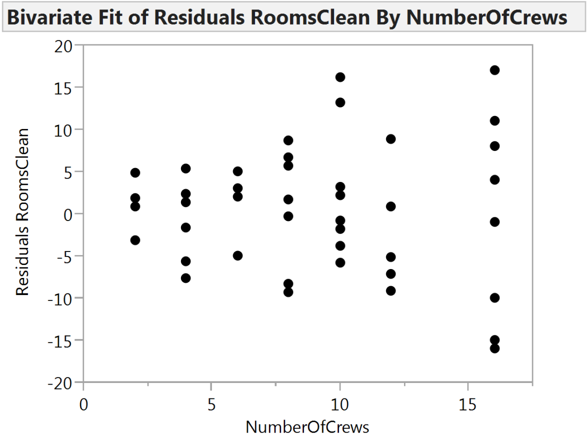
Main points:

- Global and local approaches to prediction
- Line definition
- The least squares fit to data
- The residuals
- Interpreting the slope and intercept of a regression

The residuals

- Every point has its own residual.
- It is the vertical distance from the point to the least squares line.
- Always look at a plot of the residuals, e, against x: the *residual plot*.
- The residuals should have no structure at all.
- They should look like a random swarm of points .

The residual plot



Not a disaster, but increasing variance is not ideal.

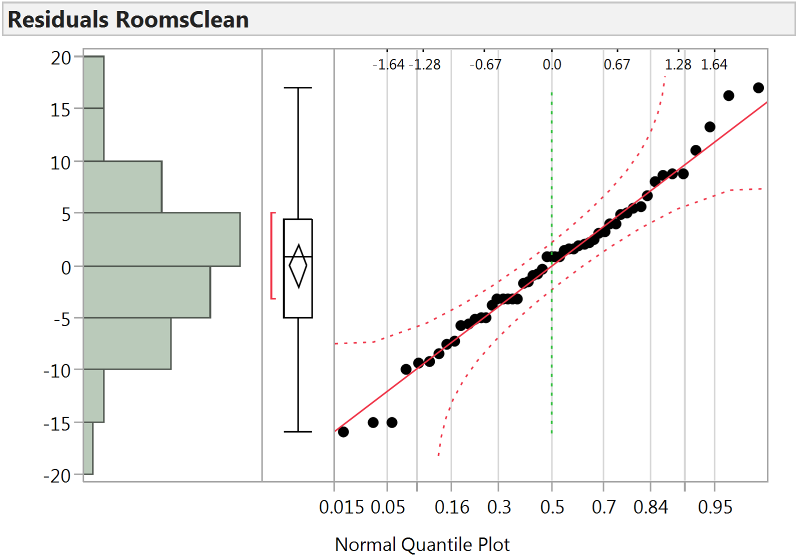
Facts about the residuals

- There are two key numerical summaries of data. The mean and standard deviation.
  1. Fact: the sample mean of the residuals is always exactly zero.
  2. The sample standard deviation of the residuals:

$$s_e = \sqrt{\frac{e_1^2 + e_2^2 + \cdots + e_n^2}{n-2}}.$$

- The (n - 2) in the denominator is there because we have estimated 2 parameters in the regression, the slope and intercept.
- s_e is a measure of the variation in y that is not explained by knowing x. That is, s_e measures the unexplained variation in y.
- Low values of s_e are good, and if you are choosing between models with the same outcome variable, then prefer models with the lower s_e.
- s_e is also known as **R**oot **M**ean **S**quared **E**rror (RMSE).

The residuals and the NQP



The numerical summaries of the residuals

Summary Statistics	
Mean	3.486e-15
Std Dev	7.2652969
Std Err Mean	0.9979653
Upper 95% Mea	2.0025638
Lower 95% Mean	-2.002564
N	53

Note that the sample mean is 0. It has to be as a by-product of the least squares methodology.

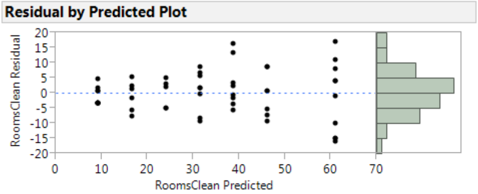
JMPs residual plots

JMP produces 5 residual plots in simple regression:

1. Residuals against predicted
2. Actual by predicted
3. Residuals against row
4. Residuals against x
5. Normal quantile plot of the residuals

Plot #1

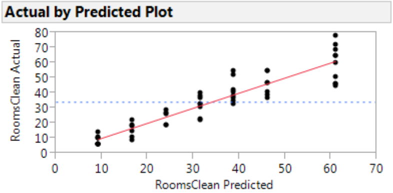
Plot: $e_i \quad v. \quad \hat{y}_i.$



- Hoping to see no structure.
- Useful in multiple regression, where $\hat{y}$ can be thought of as a *blend* of all the x's in the model.

Plot #2

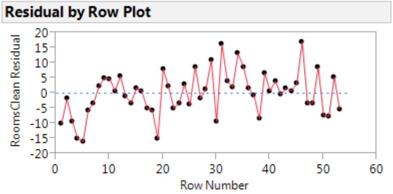
Plot: $y_i \quad v. \quad \hat{y}_i$.



- Hoping to see points tightly clustered about the 45^o line.
- A way of visualizing R^2 (works in multiple regression too). R^2 is the square of the correlation in this plot.
- More sophisticated: a *calibration* plot. If you see systematic departures from the 45^o line, then for some predicted values you are getting a biased estimate of $E(y)$, and you might want to fix this up.

Plot #3

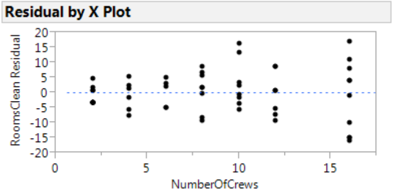
Plot: $e_i \quad v. \quad i$ (the row number).



- Only useful is there is a concept of *adjacency* between rows. This will be true for time-series.
- Look for systematic structure, particularly a tracking in the residuals which may indicate a lack of independence between neighboring residuals.

Plot #4

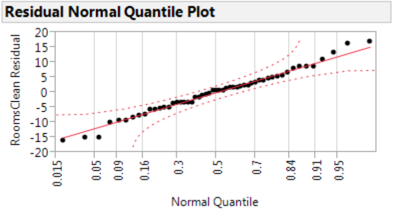
Plot: $e_i \quad v. \quad x_i$.



- The most *natural* residual plot.
- Equivalent in simple regression to plot #1.
- This is because $\hat{y}$ is a linear transform of x: $\hat{y}_i = b_0 + b_1 x_i$.
- If b_1 is negative then the orientation in the horizontal direction switches around between plots #1 and #4, but from the point of view of looking for *structure*, it is identical to #1.

Plot #5

Plot: Normal Quantile Plot of the residuals.



- When we start checking assumptions, this is the one to assess normality of the error terms.

R^2 and the quality of fit

- As the residuals are defined as $e = y - \hat{y}$, then so $y = \hat{y} + e$.
- The representation:

$$y = \hat{y} + e,$$

  shows that the model splits the observed data y into two parts: a systematic part $\hat{y}$, and a random component e.
- This is the

$$\text{Data} = \text{Signal} + \text{Noise},$$

  paradigm.

Summarizing the fit

- Define R^2 as $(r)^2$, that is the sample correlation squared.
- It is sometimes called the Coefficient of Determination, but R^2 is more common.
- Interpretation: the proportion of variability in y explained by the regression model.
- Facts about R^2:
  1. $0 \leq R^2 \leq 1$.
  2. An R^2 of 1 means perfect linear association.
  3. An R^2 of zero means no linear association.
  4. R^2 has no measurements units.
- All other things being equal, we prefer models with a higher R^2.

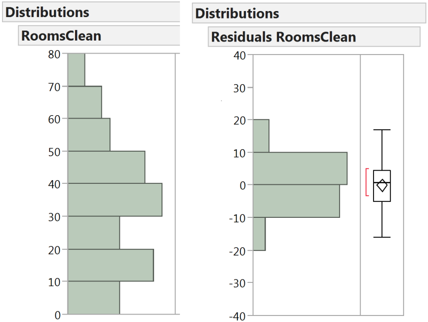
Visualizing R^2

- Plot a histogram of y and measure its variance, s_y^2.
- Plot a histogram of the residuals (e) and measure their variance s_e^2.
- Compare the two variances: if the variance of the residuals is small compared to the variance of the raw data y, then that is good, we have explained a lot of variation in y by using the model.
- In fact:

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

Visualizing R^2

Look at the variation in the raw data and compare it to the variation in the residuals.



Summary Statistics	
Mean	33.91
Std Dev	$\boxed{19.20}$
Std Err Mean	2.64
Upper 95% Mea	39.20
Lower 95% Mean	28.61
N	53.00

Summary Statistics	
Mean	0.00
Std Dev	$\boxed{7.27}$
Std Err Mean	1.00
Upper 95% Mea	2.00
Lower 95% Mean	-2.00
N	53.00

Notice that there is less variation in the residuals, 19.20 v. 7.27.

Verifying the approximation

- The approximation to R^2:

$$1 - \frac{7.27^2}{19.20^2} = 0.86.$$

- The exact answer from JMP:

Summary of Fit

RSquare	0.857
RSquare Adj	0.854
Root Mean Square Error	7.336
Mean of Response	33.906
Observations (or Sum Wgts)	53.000

- You don't ever have to use the approximation in practice, but it is helpful in understanding exactly what R^2 is measuring.

Things to think about when running regressions

- Think about lurking variables and be careful with your conclusions. Regression only identifies association and not causation.
- The association might be *spurious* because it may be driven by an omitted variable.
- Check that the association is approximately linear, otherwise the line doesn't make much sense.
- Inspect the residuals and hope to find no structure.

Module summary

1. The residual definition
2. RMSE (s_e)
3. Various residual plots
4. R^2 and the quality of fit
5. The relationship between R^2 and RMSE

Next time

- Dealing with curvature