Stat 102

Introduction to Business Statistics
Class 12

Richard P. Waterman

Wharton

Table of contents I

Today's module

Topics to be covered in this module:

- Last time
- Broken regression assumptions
  1. Lack of independence of adjacent residuals
  2. Lack of constant variance of residuals
  3. Lack of normality of residuals
- Outliers and what to do with them
- Summary
- Next time

Last time

- Transformations of confidence intervals
- Prediction in regression
  1. Confidence bands for the true regression line
  2. Prediction intervals for a new observation

Key approximation formula:
An approximate 95% prediction interval for a new observation, y_{new} is given by

$$\hat{y}_{new} \pm 2RMSE.$$

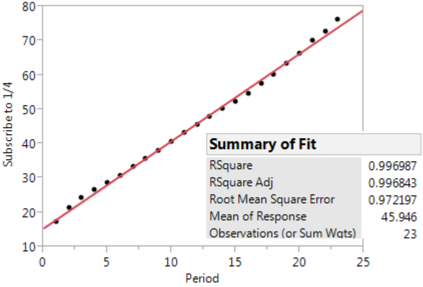
This approximation only works well within the observed range of the x-variable (that is, the region where the statistical extrapolation penalty has not yet kicked-in).

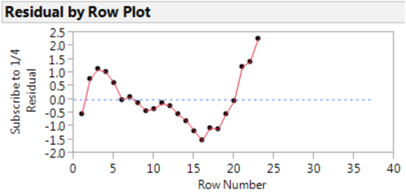
The regression assumptions reviewed

- ϵ_i are independent.
- ϵ_i are mean zero and have constant variance, $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma_\epsilon^2$ for all i (constant variance).
- ϵ_i are approximately normally distributed.

Independence of residuals and time-series

When you are working with a time-series it is essential to plot the residuals against time.
Using the cell-phone data we have a good fit and high R^2, but the residuals are problematic.





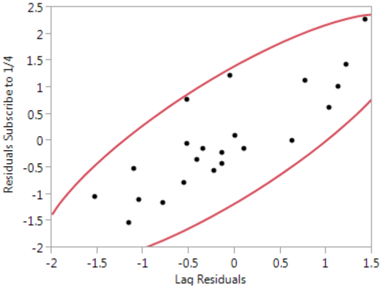
The problem

- Adjacent residuals are correlated. Knowing one residual is positive suggests its neighbor is likely to be positive. Knowing one is negative suggests its neighbor is likely to be negative.
- If the residuals were independent, then knowing the value of one residual should give no information about its neighbor.
- We describe the residual plot as *tracking* or *meandering*.
- Tracking residuals are common with time-series so always be on the look out for it.

A plot of adjacent residuals

- The term *lagged residual* means to drop the residual back in time.
- The lag-one residual is written as e_{t-1}.
- In JMP you create the lagged residual with the formula: Lag found in the Row group of functions in the formula dialog. You can set different lags, but we will just use lag one: Lag[resids, 1].

Figure 1: The scatterplot of adjacent residuals: e_t $v.$ e_{t-1}



A plot of adjacent residuals

- Notice that the residuals are correlated with their lags.
- If the residuals were independent then this plot would have no structure.
- Because the correlation in this plot is positive, we call this problem *positive auto-correlation*.

Consequences of a lack of independence

A lack of independence:

- If we have positive autocorrelation and we ignore it, then we are over-optimistic about the information content in the data. We think that there is less noise than there really is. for example, confidence intervals will be too narrow.
- We summarize this idea as a *false sense of precision*.

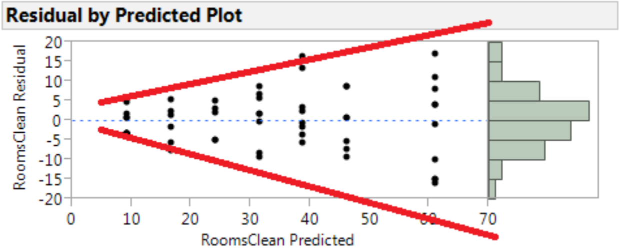
Fix-ups:

- Consider modeling *differences*: Regress $y_t - y_{t-1}$ against $x_t - x_{t-1}$.
- Consider adding *lagged residuals* to the model and running a multiple regression. (More on this when we do time-series, Ch. 27 in Stine).

Non-constant variance

This is common when an x-variable is a proxy for size.

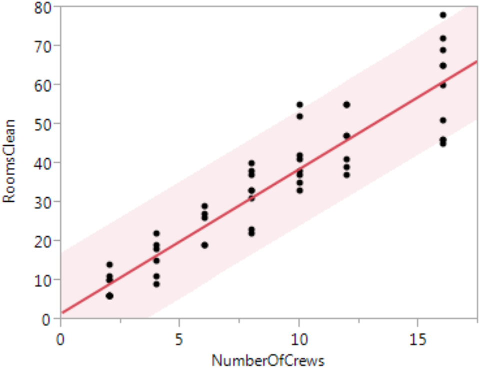
Figure 2: Cleaning data residuals



Notice how the residuals *fan out* breaking the constant variance assumption.

Non-constant variance

Figure 3: Cleaning data with 95% predictions intervals



Notice how the prediction intervals (calculated assuming constant variance) are too wide for small x, and too narrow for large x.

Consequences of violation of the assumptions

Non-constant variance:

- Incorrectly quantify the true uncertainty.
- Prediction intervals are inaccurate.
- Least squares is **unbiased**, but ...
- Least squares is **inefficient**: if you understood the structure of $\{\sigma_i^2\}$ better you could get better estimates of β_0 and β_1.

There is an advanced technique called **weighted least squared (WLS)** that is efficient.

It weights the observations by the inverse of their variances – the more precise the data, the more weight that point gets.

Fix-ups for non-constant variance

The fancy name for non-constant variance: *heteroscedasticity*.
Fix-ups include:

- Consider transforming Y with the log or square root. These are called variance stabilizing transforms in this context.
- Consider *size normalizing* the Y variable. That is, run the regression using Y/X where X is a "size variable".
- Consider ignoring the problem as Ordinary Least Squares is at least unbiased.

Consequences of violation of the assumptions

The Normality assumption (on the error terms):

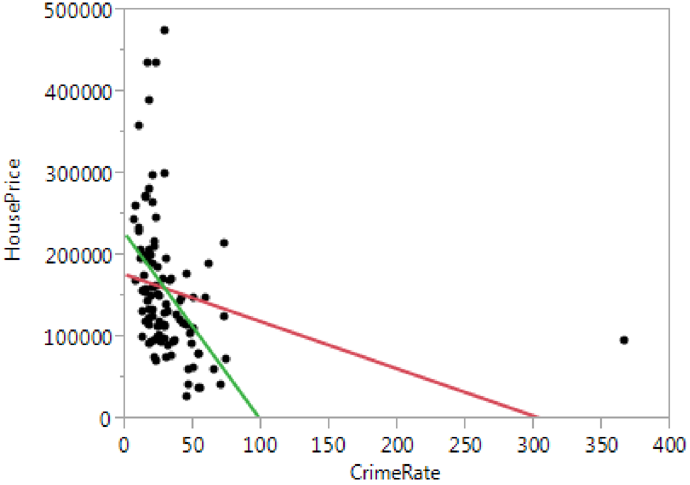
- If the ϵ_i are symmetric then the Normality assumption is not a big deal for estimation of the slope and intercept and related inference, because a Central Limit Theorem will save the day.
- Prediction intervals that are based on the Empirical Rule will be sensitive to this Normality assumption though, because they assume Normality.
- If the ϵ_i are really skewed and you only have a small amount of data then it is all up the creek.

Outliers in regression

- These are atypical observations.
- They may be atypical in either the x-direction, or the y-direction (or both).
- Points atypical in the x-direction are known as *high leverage* points.
- Points that are atypical in the y-direction are those with extreme residuals.
- If you think that a point in a regression is contaminating the regression, then refit the regression with the point removed, and review the magnitude of the residual.
- The issue is that a highly leveraged point will drag the regression line toward it, and so, diminish the magnitude of the residual.

The Philadelphia data set

Figure 4: The impact of the leveraged point on the regression fit



The Philadelphia data set

Comments:
Why do we get outliers? Some reasons:

- Corrupted data.
- Inadvertent mixing of another population, eg cars, tanks and scooters.
- Variable definition problems.

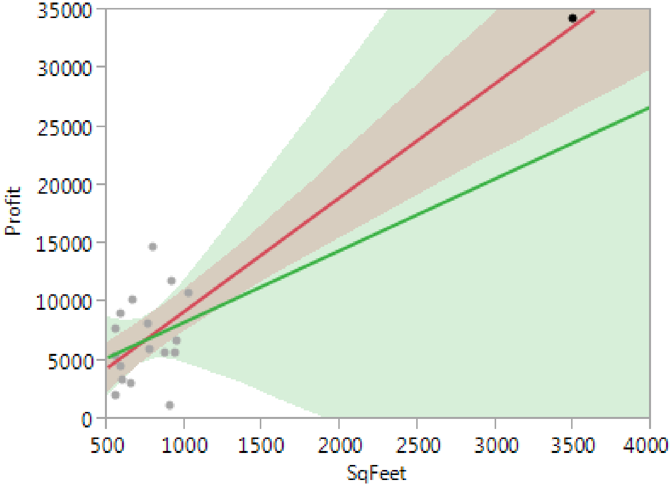
Do not automatically remove outliers. They are often the most important and most informative observations in the data set.

> *If you remove everything from your analysis that doesn't meet your initial expectations, what will you be left with?*

Try to explain the outliers. If you can, you have almost certainly learned something new.
Beware leveraged data-points, they have the potential (but not the guarantee) to drive the whole regression.

The Cottages data set

Figure 5: The impact of the leveraged point in the cottages dataset



The impact of a leveraged point on the regression summaries

Measure	With outlier	Without outlier
R-squared	0.78	0.075
RMSE	3570	3634
Slope	9.75	6.14
SE(slope)	1.30	5.56

Without the leveraged data point:

- R^2 "evaporates".
- The standard error of the slope "explodes".

Conclusion: leveraged data points convey a large amount of information.

Module summary

- Problems with regression assumptions
  1. Lack of independence of adjacent residuals
  2. Lack of constant variance of residuals
  3. Lack of normality of residuals
- Outliers and what to do with them

Next time

- More on auto-correlation + review.

Unbiasedness and efficiency

- For an estimator $\hat{\theta}$ of a parameter θ, we call it *unbiased* if $E(\hat{\theta}) = \theta$. That is, we are right on average.
- For an unbiased estimator $\hat{\theta}$, we call it *inefficient* if there exists another unbiased estimator $\tilde{\theta}$, such that $Var(\tilde{\theta}) < Var(\hat{\theta})$.
  That is, the other estimator is more concentrated about the true value, θ.