Stat 102

Introduction to Business Statistics
Class 15

Richard P. Waterman

Wharton

Table of contents I

Today's module

Topics to be covered in this module:

- Last time
- Correlated predictors – collinearity
- The market model
- Diagnostics for collinearity
- Fix-ups for collinearity
- Summary
- Next time

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

Collinearity

- Definition: correlation between the X-variables.
- Consequence: it is difficult to establish which of the X-variables are most important in the regression (they all look the same).
- Visually the regression plane becomes very unstable (sausage in space, legs on the table).
- Key formula:

$$\text{Multiple regression: } SE(b_k) \approx \frac{\sigma}{\sqrt{n}} \times \frac{1}{s_{x_k(adjusted)}}$$

- Contrast this to simple regression where

$$\text{Simple regression: } SE(b_1) \approx \frac{\sigma}{\sqrt{n}} \times \frac{1}{s_{x_1}}$$

- The difference is in whether the standard deviation of x is *adjusted*.

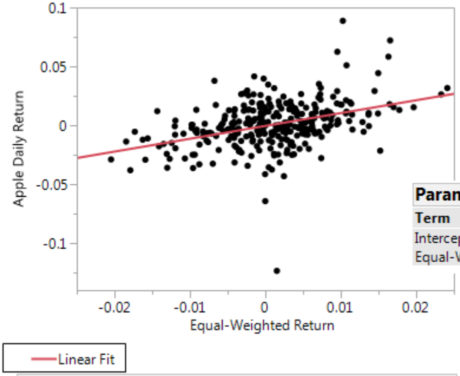
What is $s_{x_k(adjusted)}$?

- The unique variation in x_k that is not explained by the other explanatory variables.
- The standard deviation of the residuals from the regression of x_k against all of the other x's.

The market model

Consider the regression of Apple return against the `Equal weighted market return`.

Bivariate Fit of Apple Daily Return By Equal-Weighted Return



Summary of Fit

RSquare	0.158841
RSquare Adj	0.15611
Root Mean Square Error	0.018157
Mean of Response	0.00053
Observations (or Sum Wgts)	310

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.000506	0.00104	-0.49	0.6269
Equal-Weighted Return	1.0902031	0.142952	7.63	<.0001*

Linear Fit

Apple Daily Return = -0.000506 + 1.0902031*Equal-Weighted Return

Interpretations

- The slope: when the market goes up by an additional one percent, then Apple can be expected to increase by 1.09%.
- The intercept: on days when the market doesn't move, Apple can be expected to fall by 0.05% (but it is not significant).
- R^2: 16% of the risk in Apple is explained by the **market**.
- 1 - R^2: 84% of the risk in Apple is not explained by the market, that is, it is **specific** to Apple.

Getting ready for collinearity

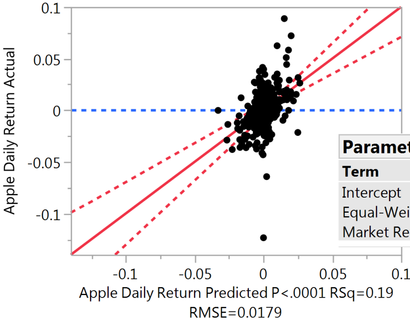
- The standard deviation of `Equal weighted market return` is 0.0072256.
- The standard error of the slope is approximately:

$$se(b_1) \approx \frac{RMSE}{\sqrt{n}} \times \frac{1}{s_{x_1}} = \frac{0.018157}{\sqrt{310}} \times \frac{1}{0.0072256} = 0.1427215$$

- Of course, you can see the exact standard error on the output: 0.142952, which is really close to 0.1427215, proving our approximation is very good.

Introducing a second correlated variable

Actual by Predicted Plot



Summary of Fit

RSquare	0.189643
RSquare Adj	0.184364
Root Mean Square Error	0.01785
Mean of Response	0.00053
Observations (or Sum Wgts)	310

Parameter Estimates

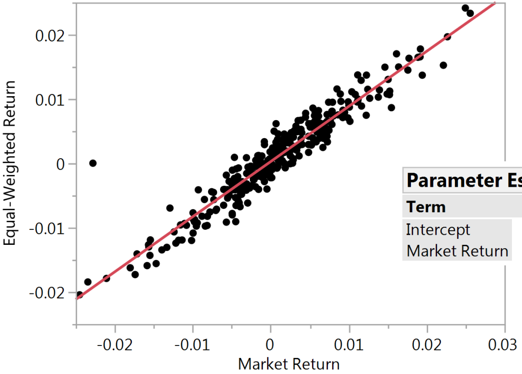
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.000226	0.001026	-0.22	0.8261
Equal-Weighted Return	-0.433086	0.467548	-0.93	0.3550
Market Return	1.4349853	0.420076	3.42	0.0007*

- The regression slope for `Equal weighted return` has **switched sign**, its standard error has **exploded** and it is **no longer** statistically significant.
- This is what gross collinearity can do to a regression analysis.

Why did the standard error explode?

Find $s_{x_k(adjusted)}$.
Regress `Equal weighted return` against `Market return` and find the standard deviation of the residuals. Of course, that's just the RMSE in this regression of x_1 against x_2.

Bivariate Fit of Equal-Weighted Return By Market Return



Summary of Fit

RSquare	0.909648
RSquare Adj	0.909355
Root Mean Square Error	0.002175
Mean of Response	0.00095
Observations (or Sum Wgts)	310

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0002534	0.000124	2.04	0.0421*
Market Return	0.8569163	0.015388	55.69	<.0001*

Putting it together

All terms refer to the `Equal weighted return` variable.

Model	sd	std.err slope
SRM	0.007226	0.1429
MRM	0.002175	0.4675

- The standard error of the slope has increased because the standard deviation of x has become much smaller.
- Recall that the standard deviation of x is in the denominator of the standard error formula.
- The standard deviation fell from 0.007226 to 0.002175 and the standard error increased from 0.1429 to 0.4675.

The Variance Inflation Factor

- This is a numeric summary of the extent of collinearity in a multple regression.
- Each variable gets its own VIF.
- The VIF is the price you pay for collinearity: the increase of the variance in the estimated regression coefficient, due to the presence of collinearity.
- An approximation with a nice interpretation:

$$VIF(X_k) \approx \left(\frac{s_{x_k}}{s_{x_k(adjusted)}}\right)^2.$$

- The exact formula:

$$VIF(X_k) = \frac{1}{1 - R^2(X_k v.X_1, \cdots, X_{k-1})}.$$

- When all the x's are all uncorrelated the VIFs are all 1 (perfection). As the collinearity increases so do the VIFs. VIFs above 10 are a warning signal to take some action.

VIFS in the market model

Using the approximation:

$$VIF(X_k) \approx \left(\frac{s_{x_k}}{s_{x_k(adjusted)}}\right)^2 = \left(\frac{0.007226}{0.002175}\right)^2 = 11.038.$$

From the JMP [1] output:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-0.000226	0.001026	-0.22	0.8261	.
Equal-Weighted Return	-0.433086	0.467548	-0.93	0.3550	11.067828
Market Return	1.4349853	0.420076	3.42	0.0007*	11.067828

Note that the approximation (11.038) is close to the true value (11.0678), so the VIF's interpretation as the ratio of the standard deviations is justified.

[1] You get the VIFs by right clicking in the parameter estimates table, choosing columns, then VIF

Diagnostics for collinearity

- Thin ellipses in the scatterplot matrix. (High correlation.)
- Counter-intuitive signs on the slopes.
- Large standard errors on the slopes (there's little information on them).
- [2]Collapsed leverage plots.
- High Variance Inflation Factors. The increase in the variance of the slope estimate due to collinearity.

$$VIF(X_k) = \frac{1}{1 - R^2(X_k v. X_1, \cdots, X_{k-1})}.$$

- Insignificant t-statistics even though over all regression is significant (ANOVA F-test).

[2]To be discussed in the next class

Fix ups for collinearity

- **Ignore** it. OK if sole objective is prediction in the range of the data.
- **Combine** collinear variables in a meaningful way.
- **Delete** variables. OK if extremely correlated.

Summary

- Collinearity:
  1. Definition
  2. Consequences
  3. Diagnostics
  4. Fix-ups
- The market model

Next time

- Leverage plots
- Categorical predictor variables