

Applied Regression Modeling: A Business Approach

Chapter 3: Multiple Linear Regression Sections 3.1–3.3.2

by Iain Pardoe

3.1 Probability model for (X_1, X_2, \dots) and Y	2
Multiple linear regression	2
Multiple linear regression model	3
3D-scatterplot of (Y, X_1, X_2)	4
Multiple linear regression equation	5
3.2 Least squares criterion	6
Estimating the model	6
HOMES3 data	7
Scatterplot matrix	8
Multiple linear regression model	9
Computer output	10
3.3 Model evaluation	11
Evaluating fit numerically	11
Regression standard error, s	12
Calculating R^2	13
Interpreting R^2	14
Disadvantage of R^2 for model building	15
Adjusted R^2	16
Using adjusted R^2	17
SHIPDEPT data	18
Adjusted R^2 for SHIPDEPT data	19
Multiple correlation	20
Low correlation between Y and X_1	21
High correlation between Y and X_1	22

Multiple linear regression

- Y is a quantitative *response* variable (a.k.a. dependent, outcome, or output variable).
- (X_1, X_2, \dots) are quantitative *predictor* variables (a.k.a. independent/input variables, or covariates).
- Important to identify variables and define them carefully, e.g.:
 - Y is final exam score, out of 100;
 - X_1 is time spent partying during last week of term, in hours;
 - X_2 is average time spent studying during term, in hours per week.
- How much do we expect Y to change by when we change the values of X_1 and/or X_2 ?
- What do we expect the value of Y to be when $X_1 = 7.5$ and $X_2 = 1.3$?

© Iain Pardoe, 2006

2 / 22

Multiple linear regression model

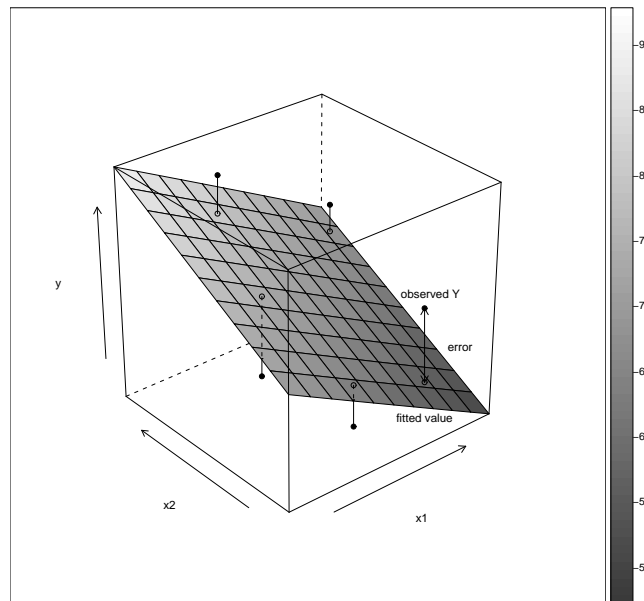
- Suppose exam score is (on average) 70 minus 1.6 times party hours plus 2.0 times study hours.
 - $E(Y | (X_{1i}, X_{2i})) = 70 - 1.6X_{1i} + 2.0X_{2i}$,
where $E(Y | (X_{1i}, X_{2i}))$ means “the expected value of Y given that $X_1 = X_{1i}$ and $X_2 = X_{2i}$ ”.
- Individual scores can deviate from this expected value by an amount e_i (called a “random error”).
 - $Y_i | (X_{1i}, X_{2i})$
 $= 70 - 1.6X_{1i} + 2.0X_{2i} + e_i \quad (i = 1, \dots, n)$
 $= \text{deterministic part} + \text{random error}.$
- Error, e_i , represents variation in Y due to factors other than X_1 and X_2 which we haven't measured, e.g., quantitative skills, exam-taking ability.
- Example: $E(Y) = 70 - 1.6(7.5) + 2.0(1.3) = 60.6$. If $Y = 65$, then $e = 65 - 60.6 = 4.4$.

© Iain Pardoe, 2006

3 / 22

3D-scatterplot of (Y, X_1, X_2)

Y = exam score, X_1 = party time, X_2 = study time



© Iain Pardoe, 2006

4 / 22

Multiple linear regression equation

- Population: $E(Y | (X_1, X_2, \dots)) = b_0 + b_1X_1 + b_2X_2 + \dots$
- Interpretation:
 - b_0 : expected Y -value when $X_1 = X_2 = \dots = 0$;
 - b_1 : "slope in the X_1 -direction" (i.e., when X_2, X_3, \dots are held constant);
 - b_2 : "slope in the X_2 -direction" (i.e., when X_1, X_3, \dots are held constant).
- Sample: $\hat{Y} = \hat{b}_0 + \hat{b}_1X_1 + \hat{b}_2X_2 + \dots$
 - How can we estimate $\hat{b}_0, \hat{b}_1, \hat{b}_2, \dots$?

© Iain Pardoe, 2006

5 / 22

Estimating the model

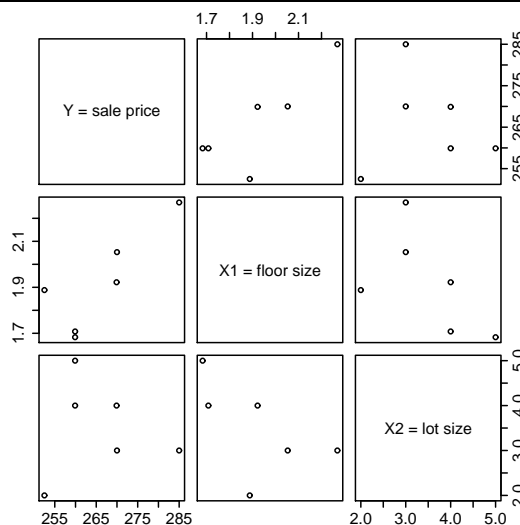
- Model: $E(Y) = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k$.
- Estimate: $\hat{Y} = \hat{b}_0 + \hat{b}_1X_1 + \hat{b}_2X_2 + \dots + \hat{b}_kX_k$.
- Obtain $\hat{b}_0, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_k$ by finding best fit "hyperplane" (using least squares).
- Mathematically, minimize sum of squared errors:

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^n \hat{e}_i^2 \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^n (Y_i - \hat{b}_0 - \hat{b}_1X_{1i} - \hat{b}_2X_{2i} - \dots - \hat{b}_kX_{ki})^2. \end{aligned}$$

- Can use calculus (partial derivatives), but we'll use computer software to find $\hat{b}_0, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_k$.

HOMES3 data

Y	252.5	259.9	259.9	269.9	270.0	285.0
X ₁	1.888	1.683	1.708	1.922	2.053	2.269
X ₂	2	5	4	4	3	3



Scatterplot matrix

- A matrix of scatterplots showing all bivariate relationships in a multivariate dataset (e.g., previous slide).
- However, patterns *cannot* tell us whether a multiple linear regression model can provide a useful mathematical approximation to these bivariate relationships.
- Primarily useful for identifying any strange patterns or odd-looking values that might warrant further investigation *before* we start modeling.
- Home price–floor size example:
no odd values to worry about.

© Iain Pardoe, 2006

8 / 22

Multiple linear regression model

- Propose this multiple linear regression model:

$$\begin{aligned} Y &= E(Y) + e \\ &= b_0 + b_1X_1 + b_2X_2 + e. \end{aligned}$$

- Random errors, e , represent variation in Y due to factors other than X_1 and X_2 that we haven't measured, e.g., numbers of bedrooms/bathrooms, property age, garage size, or nearby schools.
- Use least squares to estimate the deterministic part of the model, $E(Y)$, as $\hat{Y} = \hat{b}_0 + \hat{b}_1X_1 + \hat{b}_2X_2$.
 - i.e., use statistical software to find the values of b_0 , b_1 , and b_2 that minimize $SSE = \sum_{i=1}^n (Y_i - \hat{b}_0 - \hat{b}_1X_{1i} - \hat{b}_2X_{2i})^2$.

© Iain Pardoe, 2006

9 / 22

Computer output

Parameters^a

Model	Estimate	Std. Error	t-stat	Pr(> t)
1 (Intercept)	122.357	14.786	8.275	0.004
X1	61.976	6.113	10.139	0.002
X2	7.091	1.281	5.535	0.012

^a Response variable: Y.

- Fitted model: $\hat{Y} = 122.36 + 61.98X_1 + 7.09X_2$.
- Expect $Y = \hat{b}_0$ when $X_1 = X_2 = 0$, but *only* if this makes sense and we have data close to $X_1 = X_2 = 0$
- Expect Y to change by \hat{b}_1 when X increases by one and other predictor X -variables stay constant, i.e., expect sale price to increase \$6200 when floor size increases 100 sq. feet and lot size stays constant.
- Expect Y to change by \hat{b}_2 when X increases by one and other predictor X -variables stay constant, i.e., expect sale price to increase \$7090 when lot size increases one category and floor size stays constant.

© Iain Pardoe, 2006

10 / 22

3.3 Model evaluation

11 / 22

Evaluating fit numerically

Three methods:

- How close are the actual observed Y -values to the model-based fitted values, \hat{Y} ?
 - Calculate *regression standard error*, s (3.3.1).
- How much of the variability in Y have we been able to explain with our model?
 - Calculate *coefficient of determination*, R^2 (3.3.2).
- How strong is the evidence of our modeled relationship between Y and (X_1, X_2, \dots) ?
 - Estimate/test *regression parameters*, b_1, b_2, \dots
 - Globally (3.3.3), in subsets (3.3.4), individually (3.3.5).

© Iain Pardoe, 2006

11 / 22

Regression standard error, s Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.986 ^a	0.972	0.953	2.4753

^a Predictors: (Intercept), X1, X2.

- Regression standard error, s , estimates the std. dev. of the multiple linear regression random errors:

$$s = \sqrt{\frac{\text{SSE}}{n-k-1}}$$

- Unit of measurement for s is the same as unit of measurement for Y .
- Approximately 95% of the observed Y -values lie within plus or minus $2s$ of their fitted \hat{Y} -values.
- $2s = 4.95$, so expect to predict an unobserved sale price from particular floor and lot size values to within approx. $\pm \$4950$ (at a 95% confidence level).

© Iain Pardoe, 2006

12 / 22

Calculating R^2

- Without model, estimate Y with sample mean m_Y .
- With model, estimate Y using fitted \hat{Y} -value.
- How much do we reduce our error when we do this?
- Total error without model:
 $\text{TSS} = \sum_{i=1}^n (Y_i - m_Y)^2$, variation in Y about m_Y .
- Remaining error with model:
 $\text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$, unexplained variation.
- Proportional reduction in error: $R^2 = \frac{\text{TSS} - \text{SSE}}{\text{TSS}}$.
- Home price–floor size example: $R^2 = 0.972$.
- 97.2% of the variation in sale price (about its mean) can be explained by a multiple linear regression relationship between sale price and (floor size, lot size).

© Iain Pardoe, 2006

13 / 22

Interpreting R^2

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.986 ^a	0.972	0.953	2.4753

^a Predictors: (Intercept), X1, X2.

- R^2 measures the proportion of variation in Y (about its mean) that can be explained by a multiple linear regression relationship between Y and (X_1, X_2, \dots) .
- If $TSS = SSE$ then $R^2 = 0$: using (X_1, X_2, \dots) to predict Y hasn't helped and we may as well use m_Y to predict Y regardless of the (X_1, X_2, \dots) values.
- If $SSE = 0$ then $R^2 = 1$: using (X_1, X_2, \dots) allows us to predict Y perfectly (with no random errors).
- Such extremes rarely occur and usually R^2 lies between zero and one, with higher values of R^2 corresponding to better fitting models.

© Iain Pardoe, 2006

14 / 22

Disadvantage of R^2 for model building

- Model building: what is the best way to model the relationship between Y and (X_1, X_2, \dots, X_k) ?
 - e.g., should we use all k predictors, or just a subset?
- Consider a sequence of *nested* models, with each model in the sequence adding predictors to the previous model.
- Which model would R^2 say is the "best" model? The final model with k predictors.
- Geometrical argument: start with a regression line on a 2D-scatterplot, then add a second predictor to make the line a plane in a 3D-scatterplot.
- In other words, R^2 always increases (or stays the same) as you add predictors to a model.

© Iain Pardoe, 2006

15 / 22

Adjusted R^2

- R^2 has a clear interpretation since it represents the proportion of variation in Y (about its mean) explained by a multiple linear regression relationship between Y and (X_1, X_2, \dots) .
- *But*, R^2 is not appropriate for finding a model that captures the major, important population relationships without overfitting every slight twist and turn in the sample relationships.
- We need an alternate criterion, which penalizes models that contain too many unimportant predictor variables:
$$\text{adjusted } R^2 = 1 - \left(\frac{n-1}{n-k-1} \right) (1 - R^2).$$
- In practice, we can obtain the value for adjusted R^2 directly from statistical software.

© Iain Pardoe, 2006

16 / 22

Using adjusted R^2

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.826 ^a	0.682	0.603	7.1775
2	0.986 ^a	0.972	0.953	2.4753

^a Predictors: (Intercept), X1.

^a Predictors: (Intercept), X1, X2.

- Since adjusted R^2 is 0.603 for the single-predictor model, but 0.953 for the two-predictor model, the two-predictor model is better than the single-predictor model (according to this criterion).
- In other words, there is no indication that adding $X_2 = \text{lot size}$ to the model causes overfitting.
- What happens to R^2 and s ?

© Iain Pardoe, 2006

17 / 22

SHIPDEPT data

Y (labor hours)	X_1 (weight shipped)	X_2 (truck proportion)	X_3 (average weight)	X_4 (week)
100	5.1	90	20	1
85	3.8	99	22	2
...
85	4.8	58	25	20

- $Y = \text{weekly labor hours}$
- $X_1 = \text{total weight shipped in thousands of pounds}$
- $X_2 = \text{proportion shipped by truck}$
- $X_3 = \text{average shipment weight in pounds}$
- $X_4 = \text{week}$
- Compare two models:
 - $E(Y) = b_0 + b_1X_1 + b_3X_3;$
 - $E(Y) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4.$

© Iain Pardoe, 2006

18 / 22

Adjusted R² for SHIPDEPT data

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.899 ^a	0.808	0.786	8.815
2	0.905 ^a	0.820	0.771	9.103

^a Predictors: (Intercept), X1, X3.

^a Predictors: (Intercept), X1, X2, X3, X4.

- Since adjusted R² is 0.786 for the two-predictor model, but 0.771 for the four-predictor model, the two-predictor model is better than the four-predictor model (according to this criterion).
- In other words, there is a suggestion that adding X_2 = truck proportion and X_4 = week to the model causes overfitting.
- What happens to R² and s ?

© Iain Pardoe, 2006

19 / 22

Multiple correlation

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.986 ^a	0.972	0.953	2.4753

^a Predictors: (Intercept), X1, X2.

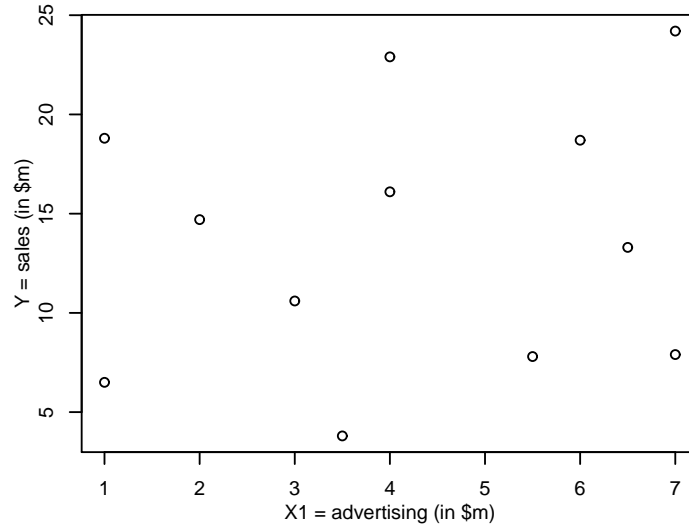
- The multiple correlation coefficient, multiple R , measures the strength and direction of linear association between the observed Y -values and the fitted \hat{Y} -values from the model.
- Multiple linear regression: multiple $R = +\sqrt{R^2}$.
 - e.g., $0.986 = \sqrt{0.972}$ for the home price–floor size example above.
- Beware: intuition about correlation can be seriously misleading when it comes to multiple linear regression (see next two slides).

© Iain Pardoe, 2006

20 / 22

Low correlation between Y and X_1

X_1 can still be a useful predictor of Y in a MLR model.

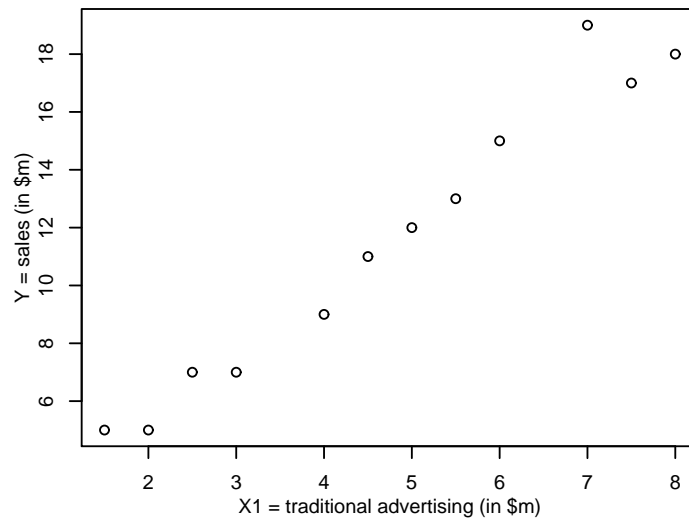


© Iain Pardoe, 2006

21 / 22

High correlation between Y and X_1

X_1 may be a poor predictor of Y in a MLR model.



© Iain Pardoe, 2006

22 / 22