

A Price Prediction Model for Building Blocks

Elizabeth Ann Tiedeman

Department of Mathematics and Computer Science

Dr. John C. Kern, Advisor

Duquesne University

December 1, 2005

Objective and Outline

Objective: Apply a least-squares regression model to collected LEGO data in order to predict the price of a given set.

1. Data Collection and Exploratory Analysis
2. Apply the Linear Regression Strategy
 - ii. Line of best fit
 - iii. Formula derivation
3. Linear Regression in Matrix Terminology
4. Multiple Linear Regression
 - i. Indicator Variables
 - ii. Linear Independence
 - iii. Interaction Terms
5. Conclusions
6. References

The Building Blocks

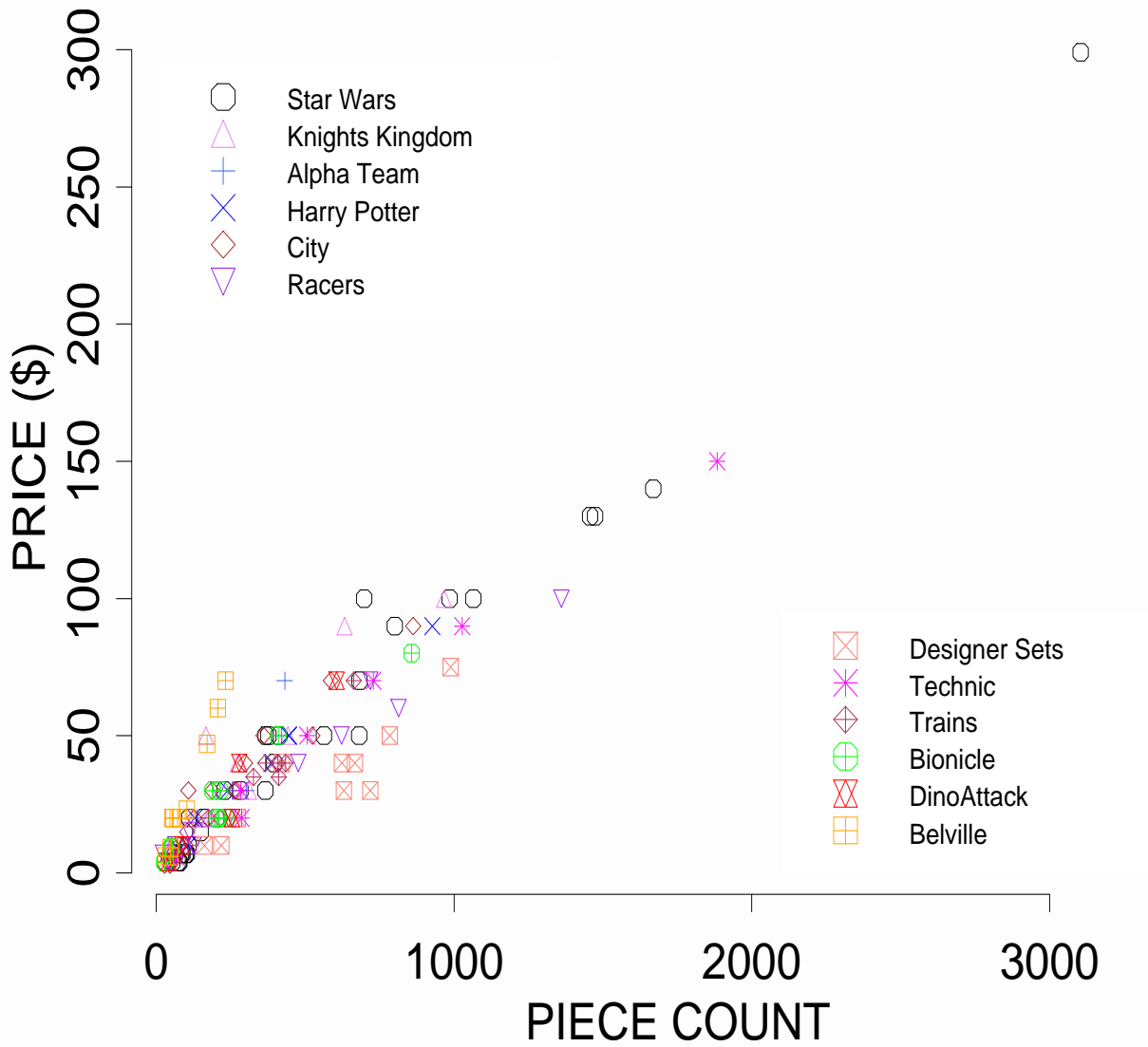
Data acquired per LEGO set (176 sets):

1. price (price)
2. piece count (pieces)
3. figurine count (minis)
4. genre

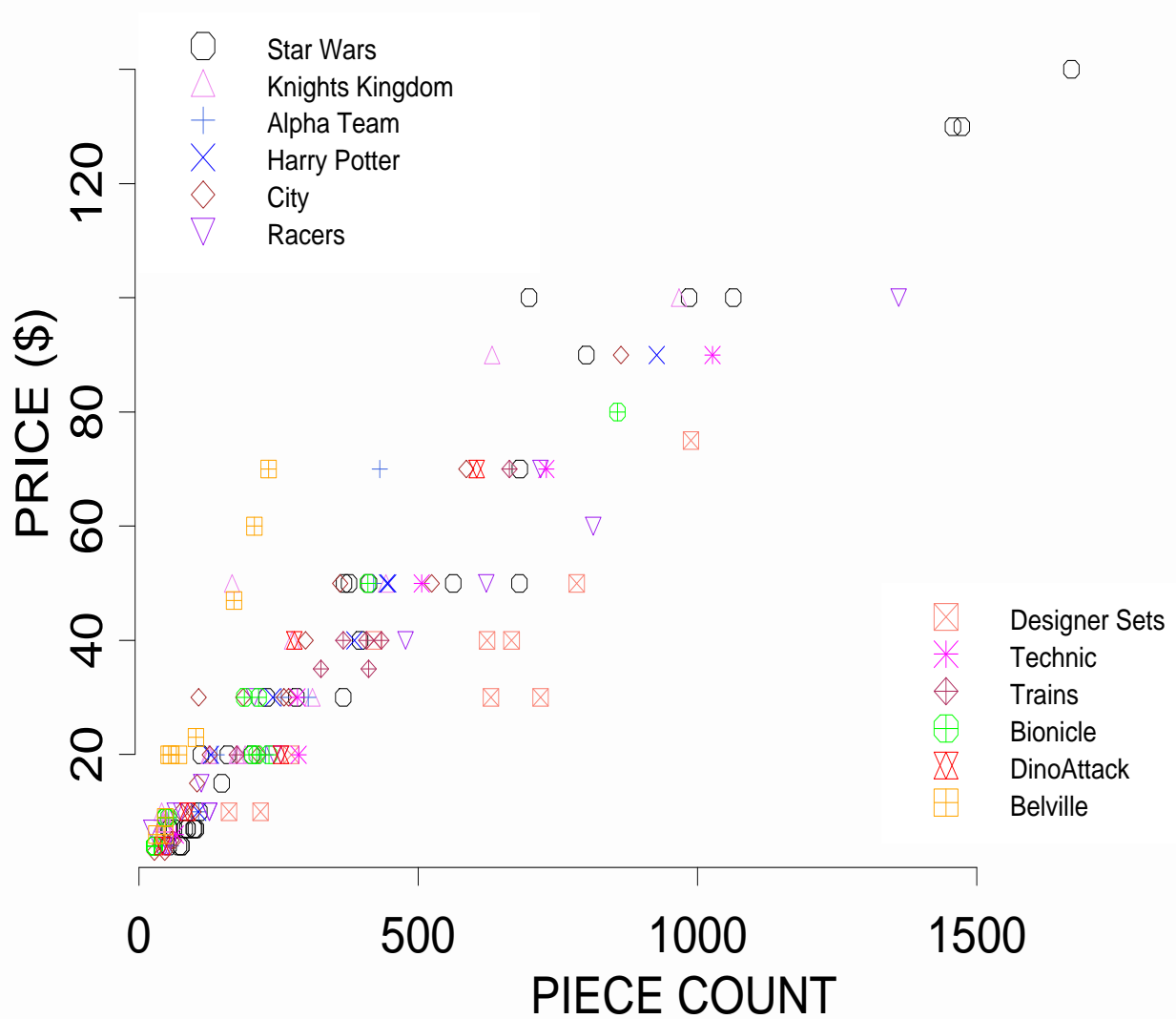
Star Wars (G1)	Designer Sets (G7)
Knights Kingdom (G2)	Technic (G8)
Alpha Team (G3)	Trains (G9)
Harry Potter (G4)	Bionicle (G10)
City (G5)	DinoAttack (G11)
Racers (G6)	Belville (G12)

The coordinate representing the piece count and price of the i^{th} set is denoted (x_i, y_i) .

The Data: Price vs. Piece Count



Price vs. Piece Count Refocused



Least-Squares Regression Model

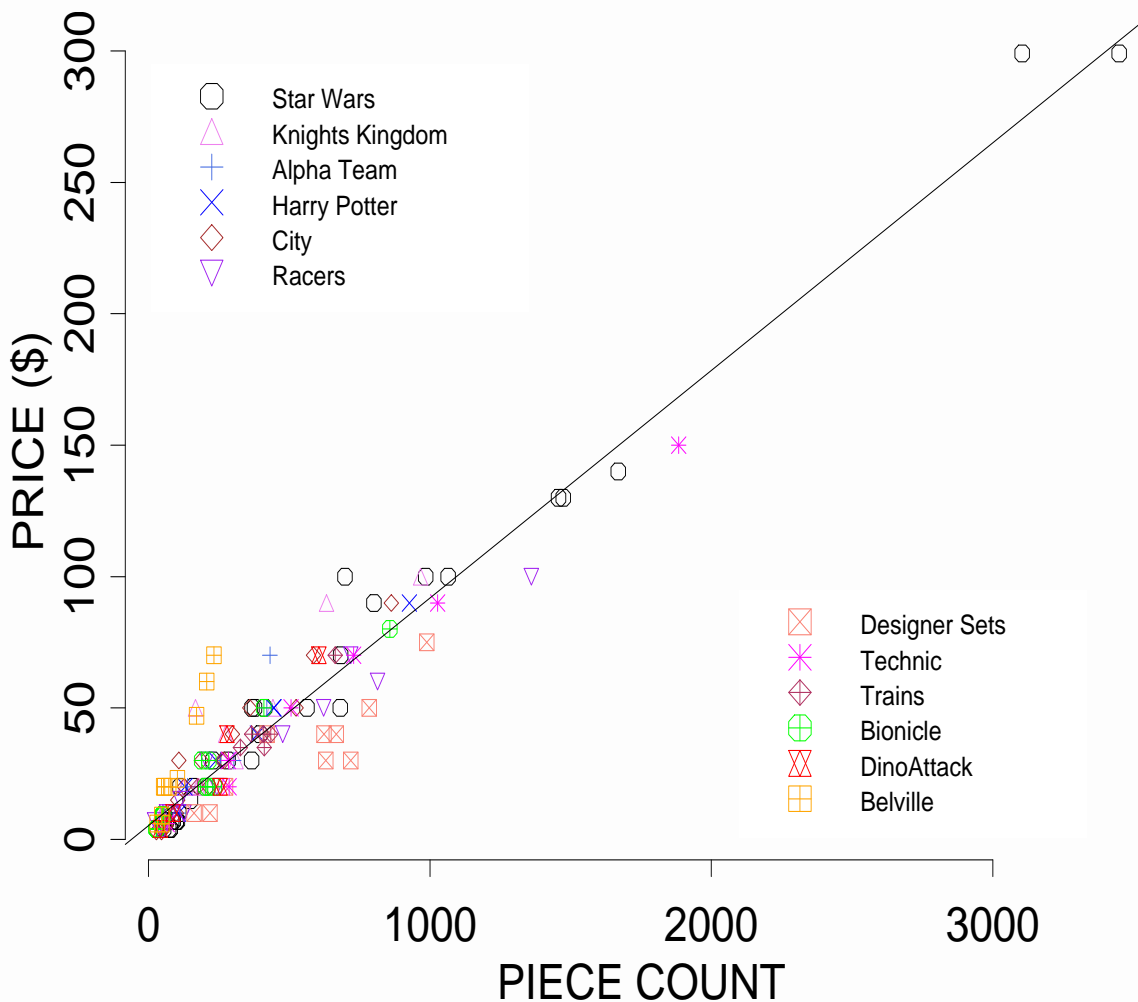
- The price prediction \hat{y} for a given piece count x is linear in x :

$$\hat{y} = \beta_0 + \beta_1 x$$

- Unknown coefficients β_0, β_1 are estimated by values b_0 and b_1 that minimize the sum of the squared vertical distances (residuals)

$$y_i - (b_0 + b_1 x_i)$$

- $\hat{y} = 5.2837 + 0.0866x$ minimizes the sum of the squared residuals for this example.



Determining b_0 and b_1

Find b_0 and b_1 that minimize

$$f(b_1, b_0) = \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2$$

by solving simultaneous equations

$$\frac{\partial f}{\partial b_0} = 0, \quad \frac{\partial f}{\partial b_1} = 0.$$

This gives:

$$b_0 = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$
$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

Matrix Representation for b_0 and b_1

Let

$$b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Then $(Y - Xb)$ is equivalent to

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} y_1 - (b_0 + b_1 x_1) \\ y_2 - (b_0 + b_1 x_2) \\ \vdots \\ y_n - (b_0 + b_1 x_n) \end{bmatrix}$$

Thus

$$\begin{aligned} f(b_1, b_0) &= \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2 = (Y - Xb)^T (Y - Xb) \\ &\Rightarrow X^T X b = X^T Y \\ &\Rightarrow b = (X^T X)^{-1} (X^T Y) \end{aligned}$$

Inference in Linear Regression

- Test whether true slope coefficient β_1 is different from 0. Formally:

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

- If H_0 is true, b_1 is not “too far” from zero. Furthermore:

$$b_1 \sim t_{n-2}$$

Provided

1. The data are independent
 2. Values of y at a particular x^* are $N(\beta_0 + \beta_1 x^*, \sigma^2)$
 3. σ^2 does not depend on x (homoskedasticity).
- For our price vs. piece example:

	Coefficient	Std. Error	t value	$Pr(> t)$
intercept	5.2837	0.9693	5.4510	0.0000
pieces	0.0866	0.0017	50.4725	0.0000

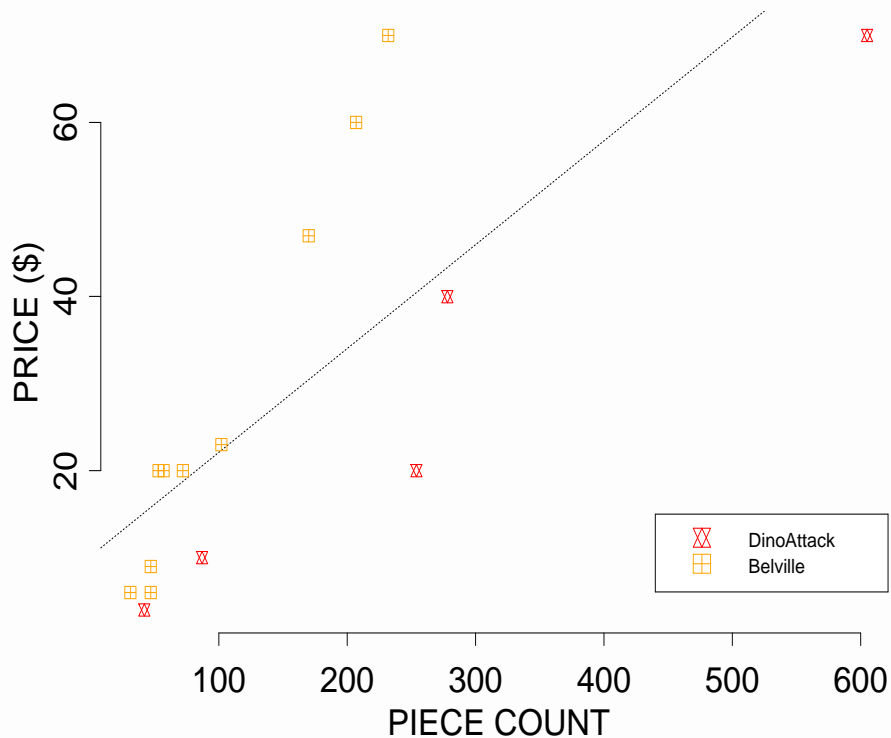
Adding More Predictors

- Can price prediction be improved by incorporating genre?
- Indicator example DinoAttack(G11) vs. Belville(G12):

$$\hat{y} = b_0 + b_1x$$

	Coefficient	Std. Error	t value	$Pr(> t)$
intercept	10.1628	5.6431	1.8009	0.0949
pieces	0.1193	0.0267	4.4660	0.0006

- Residual Std. Error: 15.15
- Multiple R-Squared: 0.6054



$$\text{price} = 10.1628 + 0.1193(\text{pieces})$$

Indicator Variables

Now let

$$\text{price} = b_0 + b_1(\text{pieces}) + b_2(\text{genre})$$

where

$$\text{genre} = \begin{cases} 0 & \text{if the set is not Belville} \\ 1 & \text{if the set is Belville} \end{cases}$$

Note $b_2(\text{genre})$ allows intercepts to differ.

The X matrix is changed accordingly

$$X = \begin{bmatrix} 1 & x_{(1,1)} & x_{(1,2)} \\ 1 & x_{(2,1)} & x_{(2,2)} \\ \vdots & \vdots & \vdots \\ 1 & x_{(n,1)} & x_{(n,2)} \end{bmatrix}$$

$x_{(ij)}$ = set i 's value of variable j .

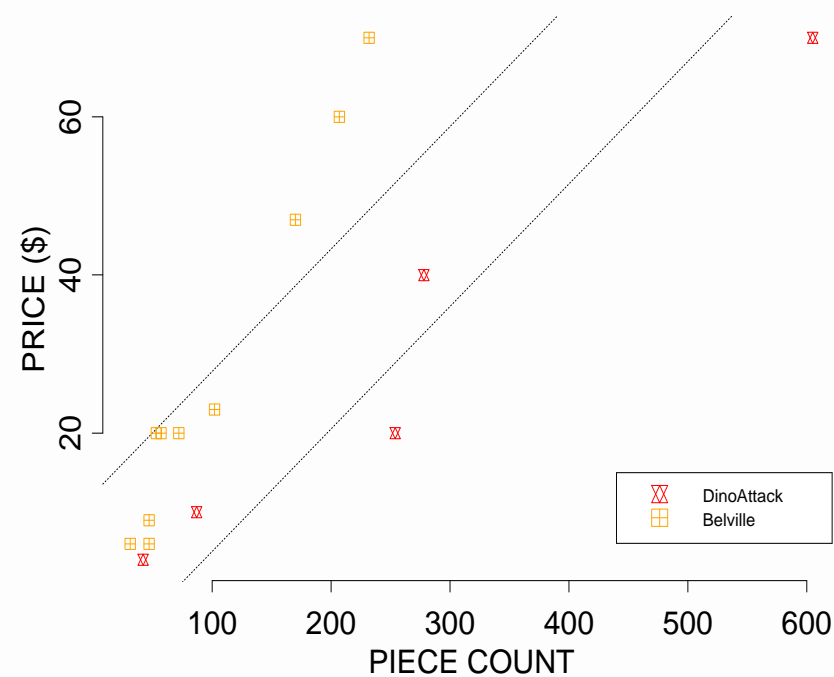
The vector b of (now three) least-squares coefficients is still given by

$$b = (X^T X)^{-1} X^T Y.$$

Belville vs. DinoAttack: Indicator Variable Regression

	Coefficient	Std. Error	t value	$Pr(> t)$
intercept	-10.4402	8.0069	-1.3039	0.2167
pieces	0.1549	0.0238	6.5113	0.0000
G12	22.7576	7.3955	3.0772	0.0096

- Residual Std. Error: 11.79
- Multiple R-Squared: 0.7794



$$\text{price} = 10.4402 + 0.1549(\text{pieces}) + 22.7576(\text{G12})$$

Therefore:

$$\text{Belville price} = -10.4402 + 0.1549(\text{pieces}) + 22.7576$$

$$\text{DinoAttack price} = -10.4402 + 0.1549(\text{pieces})$$

Interaction Terms

- Improve our formula by allowing completely different regression lines for the two genres:

$$\text{price} = b_0 + b_1(\text{pieces}) + b_2(\text{genre}) + b_3(\text{genre} * \text{pieces})$$

- $b_3(\text{genre} * \text{pieces})$ allows for the slopes to differ

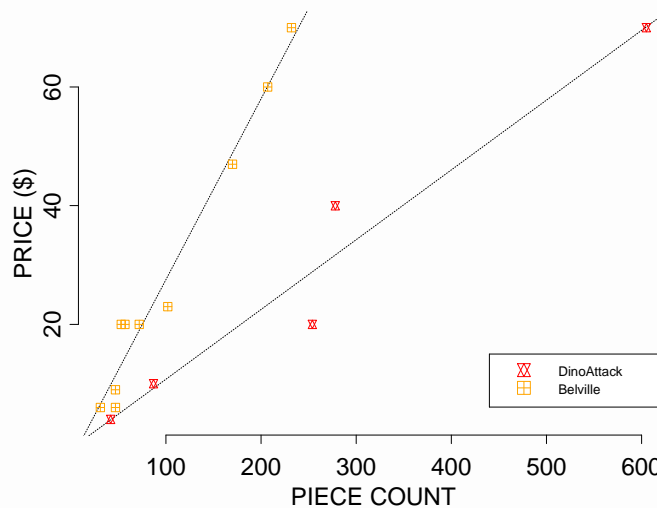
$$X = \begin{bmatrix} 1 & x_{(1,1)} & x_{(1,2)} & x_{(1,3)} \\ 1 & x_{(2,1)} & x_{(2,2)} & x_{(2,3)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{(n,1)} & x_{(n,2)} & x_{(n,3)} \end{bmatrix}$$

Belville vs. DinoAttack: Interaction Term Added

The new output that results is:

	Coefficient	Std. Error	t value	$Pr(> t)$
intercept	-0.9944	3.7644	-0.2641	0.7965
pieces	0.1176	0.0117	10.0498	0.0000
G12	-1.9370	4.7503	-0.4078	0.6913
piece*G12	0.1871	0.1871	7.1376	0.0000

- Residual Std. Error: 5.19
- Multiple R-Squared: 0.9608



$$\text{price} = -0.9944 + 0.1176(\text{pieces}) - 1.9370(\text{G12}) + 0.1871(\text{pieces} * \text{G12})$$

Therefore:

$$\begin{aligned} \text{Belville price} &= -0.9944 + 0.1176(\text{pieces}) - 1.9370(\text{G12}) \\ &\quad + 0.1871(\text{pieces} * \text{G12}) \end{aligned}$$

$$\text{DinoAttack price} = -0.9944 + 0.1176(\text{pieces})$$

Eliminating Non-Significant Indicators

Data from the previous model:

	Coefficient	Std. Error	t value	$Pr(> t)$
intercept	-0.9944	3.7644	-0.2641	0.7965
pieces	0.1176	0.0117	10.0498	0.0000
G12	-1.9370	4.7503	-0.4078	0.6913
piece*G12	0.1871	0.1871	7.1376	0.0000

- Residual Std. Error: **5.19**
- Multiple R-Squared: **0.9608**

- $0.6913 > 0.05$ thus we can eliminate the indicator variable *genre*

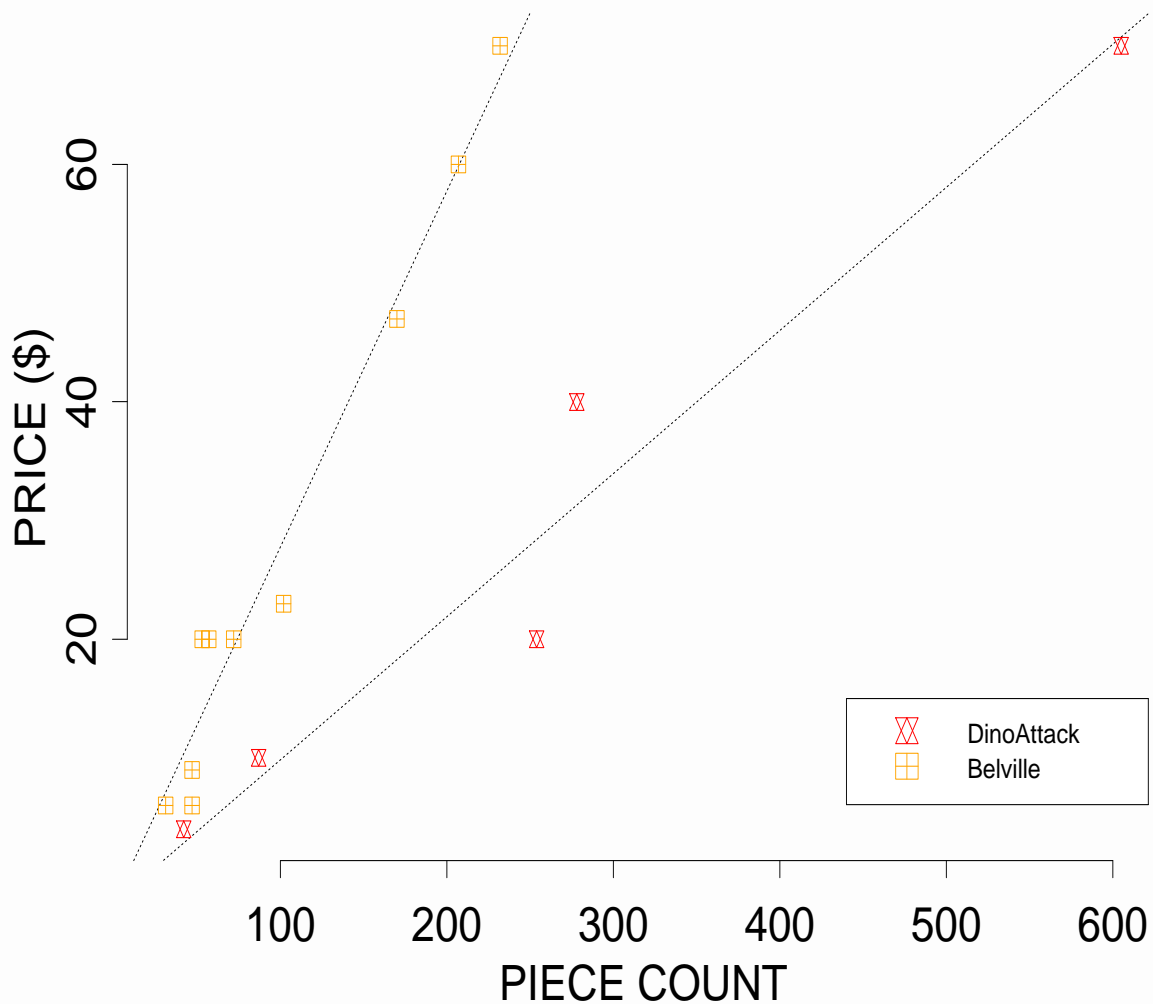
Removing G12 from the model gives:

	Coefficient	Std. Error	t value	$Pr(> t)$
intercept	-2.2108	2.2148	-0.9982	0.3379
pieces	0.1206	0.0088	13.6684	0.0000
piece*G12	0.1793	0.0173	10.3490	0.0000

- Residual Std. Error: **5.006**
- Multiple R-Squared: **0.9602**

Result: We are able to eliminate the indicator of “genre” because there is little evidence that its coefficient β_2 is different from zero.

Belville vs. DinoAttack: A “Final” Model



$$\text{price} = -2.2108 + 0.1206(\text{pieces}) + 0.1793(\text{pieces} * G12)$$

Therefore:

$$\text{Belville price} = -2.2108 + 0.1206(\text{pieces}) + 0.1793$$

$$\text{DinoAttack price} = -2.2108 + 0.1206(\text{pieces})$$

Concept Applications: The Complete Model

The full model:

$$\begin{aligned} \text{price} = & b_0 + b_1(\text{piece}) + b_2(\text{minis}) \\ & + b_3(\text{G2}) + \cdots + b_{13}(\text{G12}) \\ & + b_{14}(\text{pieces} * \text{G2}) + \cdots + b_{24}(\text{pieces} * \text{G12}) \\ & + b_{25}(\text{minis} * \text{G2}) + \cdots + b_{33}(\text{minis} * \text{G12}) \end{aligned}$$

Through backward elimination of non-significant predictors, we arrive at the model:

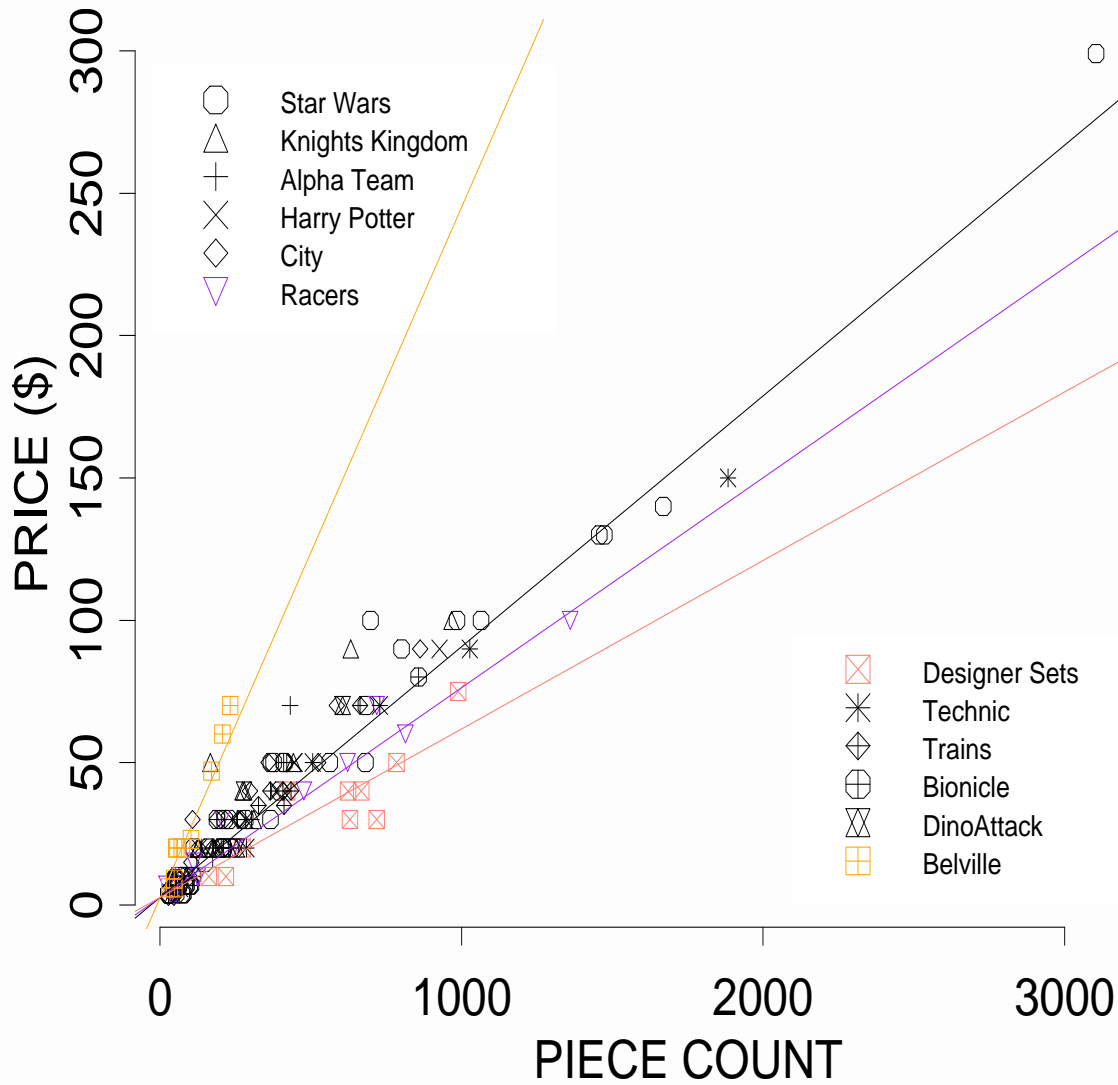
$$\begin{aligned} \text{price} = & 2.6642 + 0.0881(\text{pieces}) + 1.2975(\text{minis}) - 0.0144(\text{G6} * \text{pieces}) \\ & - 0.0289(\text{G7} * \text{pieces}) + 0.1544(\text{G12} * \text{pieces}) \end{aligned}$$

Model output from S-plus:

	Coefficient	Std. Error	t value	$Pr(> t)$
intercept	2.6642	0.6779	3.9303	0.0001
pieces	0.0881	0.0011	77.3763	0.0000
minis	1.2975	0.1600	8.1077	0.0000
G6*pieces	-0.0144	0.0035	-4.1056	0.0001
G7*pieces	-0.0289	0.0036	-7.9684	0.0000
G12*pieces	0.1544	0.0172	8.9575	0.0000

- Residual Std. Error: 6.562
- Multiple R-Squared: 0.9755

Price Prediction Model (5 predictors)



- Price prediction for Belville:
 $\text{Price} = 2.6642 + 0.2425(\text{pieces}) + 1.2975(\text{minis})$
- Price prediction for Designer Sets:
 $\text{Price} = 2.6642 + 0.0592(\text{pieces}) + 1.2975(\text{minis})$
- Price prediction for Racers:
 $\text{Price} = 2.6642 + 0.0737(\text{pieces}) + 1.2975(\text{minis})$
- Price prediction for any other genre:
 $\text{Price} = 2.6642 + 0.0881(\text{pieces}) + 0.1.2975(\text{minis})$

Conclusions

- A LEGO set containing zero legos (an empty box) would sell for \$2.66 (95% confidence interval 2.66 ± 1.35).
- In addition to the initial \$2.66 most genres cost 8.81¢ per piece (95% confidence interval 8.81 ± 0.22).
- Belville is the most expensive of the genres (An additional 15.44¢ per piece).
- Racers are slightly cheaper than the average set with each piece costing 1.44¢ less than the majority.
- Similarly, Designer Sets are cheaper with each piece costing 2.89¢ less than the majority.
- Each mini adds \$1.30 (95% confidence interval $1.30 \pm .32$) to the expected price of a LEGO set, regardless of genre.
- S-plus automated variable selection suggested a similar (but larger) model.
- Future Research: Define a model that recognizes the “\$9.99” price structure.
- How will the close of Legoland theme park impact future pricing?

References

1. Bluman, Allan G. *Elementary Statistics: A Step by Step Approach*. 3rd ed. New York: The McGraw-Hill Companies, Inc., 1998.
2. Goossens, Michel., Mittelbach, Frank., and Samarin, Alexander. *The Latex Companion*. Reading, Addison Wesley Longman, Inc., 1994.
3. Kleinbaum, David G., et al. *Applied Regression Analysis and Other Multivariable Methods*. 3rd ed. New York: Brooks/Cole Publishing Company, 1998.
4. Kutner, Michael H., Neter, John., Wasserman, William. *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs*. 3rd ed. Boston: Richard D. Irwin, Inc., 1990.