Simple and Multiple Linear Regression

Reeves Johnson

University of Virginia
McIntire School of Commerce

Why Regression? Accuracy, Flexibility, and Insight

- Testing a new Like button design against the current one.
- You could analyze this A/B test using a t-test on click rates.
- Regression analysis provides more *accurate* estimates of the effect and more *flexibility* by allowing you to include (control) variables also related to click behavior.

Why Regression? Accuracy, Flexibility, and Insight

- (Multiple) regression estimates the effect of the new button design *above and beyond* or *controlling for* all other predictors.
- What does this mean as it concerns this A/B test?
- Once you've estimated the model, you can predict click rates for different user segments, create ranges of plausible outcomes with prediction intervals, and gain deeper insights into user behavior.

University of Virginia
McIntire School of Commerce

The Simple Linear Regression Model

A simple linear regression model of the *population* can be expressed as:

$$y = \beta_0 + \beta_1 x_1 + \epsilon,$$

where:

- y is the **response** ('dependent variable')
- x_1 is the **predictor variable** ('indpendent variable')
- ϵ is the **error term**
- Parameters β_0 and β_1 are the population intercept and slope

The *Estimated* Model

After obtaining a sample and estimating the model the simple linear model is written as:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1$$

where:

- $\hat{b}_0$ and $\hat{b}_1$ are **estimates** for the intercept and slope
- $\hat{y}$ is the **predicted value** of the response variable.

The difference between an observed value, y_i, and its predicted value, $\hat{y}_i$, is called a **residual** and is denoted e_i:

$$e_i = y_i - \hat{y}_i.$$

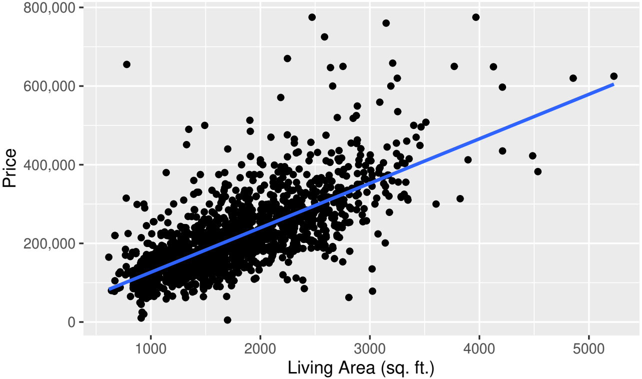
The 'Best Fitting' Line

The desired curve *fits the data the best* by the 'least sqaures principles':

- The residuals, e_i, are collectively as 'small' as possible relative to the curve.
- Equivalently, the vertical distances of observed and predicted values are as 'small' as possible.

The Least-Squares Line Depicted

The least-squares regression line fit to housing data:



Interpreting Estimates: All Important!!

How do we interpret our intercept $\hat{b}_0$ and slope $\hat{b}_1$ estimates?

```
Call:
lm(formula = price ~ living_area, data = houses)

Residuals:
    Min      1Q  Median      3Q     Max
-277022  -39371   -7726   28350  553325

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13439.394   4992.353   2.692  0.00717 **
living_area   113.123      2.682  42.173  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 69100 on 1726 degrees of freedom
Multiple R-squared:  0.5075,    Adjusted R-squared:  0.5072
F-statistic:  1779 on 1 and 1726 DF,  p-value: < 2.2e-16
```

Practice Interpreting Regression Parameters

As the nature of investing shifted in the 1990s, the relationship between mutual fund monthly performance (*Return*) as a percentage and money flowing (*Flow*) into mutual funds ($ million) shifted. Using 1990s data, *Flow* was regressed on *Return*, yielding an estimated equation of:

$$\widehat{Flow} = 56 + 209\,Return.$$

- Interpret the intercept in the linear model.
- Interpret the slope in the linear model.

R^2 as Quantifying Improved Model Fit

All models fall between 2 extremes: Perfect fits and no fits. The **coefficient of determination**, or R^2, is a measure of how well the model fits the data and ranges from 0 to 1.

Think of R^2 as the improvement in model fit compared to a *null model* consisting of just an intercept: $\hat{y} = \hat{b}_0$.

Null Model: Simplest possible model in regression. The predicted value for each data point is $\bar{y}$. (The regression line is a horizontal line at $\bar{y}$.)

The null model assumes $\bar{y}$ is the best estimate for *all* x-values so that $\hat{b}_0 = \bar{y}$.

R^2 as Quantifying Improved Model Fit

Variation Accounted For: R^2 measures the percentage of the total variation in y that's accounted for by your regression model. In other words, it quantifies how much better your model is at predicting the response variable compared to the null model.

Improvement in Model Fit: Adding predictors to the model serves to capture patterns that improve predictions compared to the null model. R^2 tells you how much better your model's predictions are compared to the predictions of the null model.

The Coefficient of Determination, R^2

Coefficient of Determination: Percentage of the variation in y that's accounted for by its regression on x.[1]

Think of R^2 graphically as the difference between the variation of the data around $\bar{y}$ and the variation around the regression line (i.e., the variation of the residuals). The graphical 'formula' is:

$$R^2 = \frac{\mathrm{var}(\text{mean}) - \mathrm{var}(\text{line})}{\mathrm{var}(\text{mean})},$$

where the difference is divided by $\mathrm{var}(\text{mean})$ to keep R^2 bounded in the range $[0, 1]$.

[1] In simple linear regression, R^2 equals the correlation coefficient squared, r^2.

Optional: A Mathematical Definition of R^2

Translating the graphical 'formula' to an actual formula you have:

$$R^2 = \frac{SST - SSE}{SST},$$

where SST is the total sum of squares (of y) and SSE is the sum of squared errors, and are defined as:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \text{ and } SSE = \sum_{i=1}^{n} e_i^2$$

What's a 'Good' R^2?

There's no value of R^2 that automatically determines that a model is 'good'.

Data from scientific experiments in controlled settings often have R^2 in the 80% to 90% range.

Data from observational studies may have an acceptable R^2 in the 30% to 50% range.

There are few contexts in which you'd choose one model over another based on their respective R^2's: Don't use R^2 to choose models; use your brain to build the right model.

Multiple Regression

A standard multiple regression model of the *population* can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon,$$

where:

- y is the **response**
- x_i is a **predictor** $(i = 1, \ldots, k)$
- ϵ is the **error term**
- Parameters β_0 and β_i $(i = 1, \ldots, k)$ are the *population* intercept and coefficients

The corresponding *estimated* equation is expressed as:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_k x_k.$$

Conceptually and Practically Important Slide!

Compare the simple and multiple linear regression models:

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \epsilon_{\text{slr}} \\
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon_{\text{mlr}}.
\end{aligned}
$$

Note Well: The simple linear regression model assumed that all influences on y that are *not* x_1 are contained in its error term ϵ_{slr}.

Conceptually and Practically Important Slide!

Compare the simple and multiple linear regression models:

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \epsilon_{\text{slr}} \\
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon_{\text{mlr}}.
\end{aligned}
$$

Note Well: The simple linear regression model assumed that all influences on y that are *not* x_1 are contained in its error term ϵ_{slr}. That is:

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \epsilon_{\text{slr}} \\
y &= \beta_0 + \beta_1 x_1 + \underbrace{\beta_2 x_2 + \cdots + \beta_k x_k + \epsilon_{\text{mlr}}}_{\epsilon_{\text{slr}}}
\end{aligned}
$$

Important: I didn't type this out to show that one formulation is a special case of another; this has far-reaching consequences on our modeling practices. Let's talk about this!

Example: Predicting Credit Scores

- Loan officer's chief concern is default risk and credit scores are indicators of default risk.
- Let's predict potential borrower's credit scores by their gross incomes.[2]

$$\text{CreditScore} = \beta_0 + \beta_1 \text{GrossIncome} + \epsilon.$$

[2]Gross income is measured in $1,000$'s of USD.

SLR Regression Results

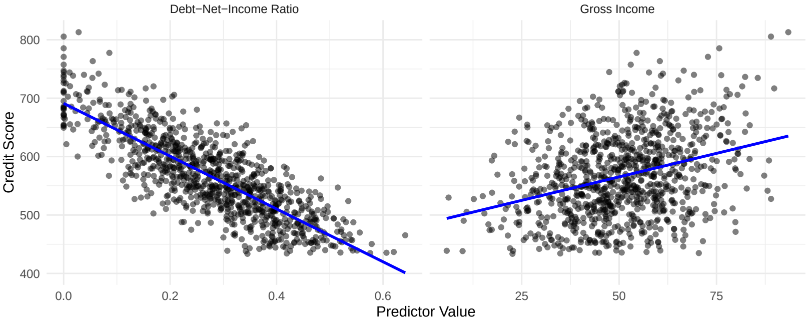
How do you interpret the intercept $\widehat{b}_0$ and coefficient $\widehat{b}_1$ estimates?

```
Call:
lm(formula = credit_score ~ income, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-151.391  -29.188    0.256   34.002  145.613 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 458.55160    4.09322  112.03   <2e-16 ***
income        2.41046    0.07585   31.78   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 46.98 on 992 degrees of freedom
Multiple R-squared:  0.5045,    Adjusted R-squared:  0.504 
F-statistic:  1010 on 1 and 992 DF,  p-value: < 2.2e-16
```

Adding Another Predictor: Multiple Linear Regression

- Add prospective borrower's debt to net income ratio as a predictor.
- Why? High gross income with high debt-income ratio might indicate higher risk than moderate income with low debt-income ratio.

$$\text{CreditScore} = \beta_0 + \beta_1 \text{GrossIncome} + \beta_2 \text{DebtNetIncomeRatio} + \epsilon.$$



MLR Regression Results

- How do you interpret the regression estimates now?
- What changes do you notice from the simple linear regression?
- Also note the coefficient estimate of gross income changed.

```
Call:
lm(formula = credit_score ~ income + debt_to_income, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-91.568 -20.077   0.339  19.935  90.412 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    591.7436     4.4212  133.84   <2e-16 ***
income           2.0670     0.0482   42.88   <2e-16 ***
debt_to_income  -4.7749     0.1266  -37.70   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.22 on 977 degrees of freedom
Multiple R-squared:  0.7547,	Adjusted R-squared:  0.7542 
F-statistic:  1503 on 2 and 977 DF,  p-value: < 2.2e-16
```

Predicting Credit Scores of the Model

- We have model to predict credit scores.
- Now use the estimates to make predictions credit.
- Let's use the Shiny app with R to find predicted credit scores for different values of income and debt-income ratios.