BUS 41100 Applied Regression Analysis

Week 1: Introduction, Simple Linear Regression

Data visualization, conditional distributions, correlation, and least squares regression

Max H. Farrell
The University of Chicago Booth School of Business
The basic problem

Available data on two or more variables

Formulate a model to predict or estimate a value of interest

Use estimate to make a (business) decision
Regression: What is it?

- Simply: The most widely used statistical tool for understanding relationships among variables

- A conceptually simple method for investigating relationships between one or more factors and an outcome of interest

- The relationship is expressed in the form of an equation or a model connecting the outcome to the factors
Regression in business

- Optimal portfolio choice:
  - Predict the future joint distribution of asset returns
  - Construct an optimal portfolio (choose weights)

- Determining price and marketing strategy:
  - Estimate the effect of price and advertisement on sales
  - Decide what is optimal price and ad campaign

- Credit scoring model:
  - Predict the future probability of default using known characteristics of borrower
  - Decide whether or not to lend (and if so, how much)
Regression in everything

Straight prediction questions:
- What price should I charge for my car?
- What will the interest rates be next month?
- Will this person like that movie?

Explanation and understanding:
- Does your income increase if you get an MBA?
- Will tax incentives change purchasing behavior?
- Is my advertising campaign working?
Data Visualization

Example: pickup truck prices on Craigslist

We have 4 dimensions to consider.

```r
> data <- read.csv("pickup.csv")
> names(data)
[1] "year"  "miles"  "price"  "make"
```

A simple summary is

```r
> summary(data)
```

<table>
<thead>
<tr>
<th></th>
<th>year</th>
<th>miles</th>
<th>price</th>
<th>make</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>1978</td>
<td>1500</td>
<td>1200</td>
<td>Dodge:10</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>1996</td>
<td>70958</td>
<td>4099</td>
<td>Ford :12</td>
</tr>
<tr>
<td>Median</td>
<td>2000</td>
<td>96800</td>
<td>5625</td>
<td>GMC  :24</td>
</tr>
<tr>
<td>Mean</td>
<td>1999</td>
<td>101233</td>
<td>7910</td>
<td></td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>2003</td>
<td>130375</td>
<td>9725</td>
<td></td>
</tr>
<tr>
<td>Max.</td>
<td>2008</td>
<td>215000</td>
<td>23950</td>
<td></td>
</tr>
</tbody>
</table>
First, the simple histogram (for each continuous variable).

```r
> par(mfrow=c(1,3))
> hist(data$year)
> hist(data$miles)
> hist(data$price)
```

Data is “binned” and plotted bar height is the count in each bin.
We can use scatterplots to compare two dimensions.

```r
> par(mfrow=c(1,2))
> plot(data$year, data$price, pch=20)
> plot(data$miles, data$price, pch=20)
```
Add color to see another dimension.

```r
> par(mfrow=c(1,2))
> plot(data$year, data$price, pch=20, col=data$make)
> legend("topleft", levels(data$make), fill=1:3)
> plot(data$miles, data$price, pch=20, col=data$make)
```
Boxplots are also super useful.

```r
> boxplot(price ~ make, ylab="Price ($)", main="Make")
> boxplot(price ~ year_boxplot, ylab="Price ($)", main="Year")
```

The box is the **Interquartile Range** (IQR; i.e., $25^{th}$ to $75^{th}$ %), with the median in bold. The **whiskers** extend to the most extreme point which is no more than 1.5 times the IQR width from the box.
Regression is what we’re really here for.

```r
> plot(data$year, data$price, pch=20, col=data$make)
> abline(lm(price ~ year),lwd=1.5)
```

- Fit a line through the points, but how?
- `lm` stands for linear model
- Rest of the course: formalize and explore this idea
Conditional distributions

Regression models are really all about modeling the conditional distribution of $Y$ given $X$.

Why are conditional distributions important?

We want to develop models for forecasting. What we are doing is exploiting the information in the conditional distribution of $Y$ given $X$.

The conditional distribution is obtained by “slicing” the point cloud in the scatterplot to obtain the distribution of $Y$ conditional on various ranges of $X$ values.
Conditional v. marginal distribution

Consider a regression of house price on size:

“slice” of data

conditional distribution of price given $3 < \text{size} < 3.5$

regression line

marginal distribution of price
Key observations from these plots:

- **Conditional distributions** answer the forecasting problem: if I know that a house is between 1 and 1.5 1000 sq.ft., then the conditional distribution (second boxplot) gives me a point forecast (the mean) and prediction interval.

- The **conditional means** (medians) seem to line up along the regression line.

- The conditional distributions have much smaller dispersion than the marginal distribution.
This suggests two general points:

- If $X$ has no forecasting power, then the marginal and conditionals will be the same.
- If $X$ has some forecasting information, then conditional means will be different than the marginal or overall mean and the conditional standard deviation of $Y$ given $X$ will be less than the marginal standard deviation of $Y$. 
Intuition from an example where $X$ has no predictive power.

House price v. number of stop signs ($Y$) within a two-block radius of a house ($X$)

See that in this case the marginal and conditionals are not all that different
Predicting house prices

Problem:
- Predict market price based on observed characteristics

Solution:
- Look at property sales data where we know the price and some observed characteristics.
- Build a decision rule that predicts price as a function of the observed characteristics.

⇒ We have to define the variables of interest and develop a specific quantitative measure of these variables
What characteristics do we use?

- Many factors or variables affect the price of a house
  - size of house
  - number of baths
  - garage, air conditioning, etc.
  - size of land
  - location

- Easy to quantify price and size but what about other variables such as location, aesthetics, workmanship, etc?
To keep things super simple, let’s focus only on size of the house.

The value that we seek to predict is called the dependent (or output) variable, and we denote this as

\[ Y = \text{price of house (e.g. thousands of dollars)} \]

The variable that we use to guide prediction is the explanatory (or input) variable, and this is labelled

\[ X = \text{size of house (e.g. thousands of square feet)} \]
What do the data look like?

```r
> size <- c(.8,.9,1,1.1,1.4,1.4,1.5,1.6, 1.8,2,2.4,2.5,2.7,3.2,3.5)
> price <- c(70,83,74,93,89,58,85,114, 95,100,138,111,124,161,172)
> plot(size, price, pch=20)
```
Appears to be a linear relationship between price and size:

- as size goes up, price goes up.

Fitting a line by the “eyeball” method:

```
> abline(35, 40, col="red")
```
Recall that the equation of a line is:

\[ \hat{Y} = b_0 + b_1 X \]

where \( b_0 \) is the intercept and \( b_1 \) is the slope.

In the house price example

- our “eyeball” line has \( b_0 = 35, b_1 = 40 \).
- predict the price of a house when we know only size
  - just read the value off the line that we’ve drawn.
- The intercept value is in units of \( Y \) ($1,000).
- The slope is in units of \( Y \) per units of \( X \)
  ($1,000/1,000 sq ft).
Recall how the slope ($b_1$) and intercept ($b_0$) work together graphically.

\[ Y = b_0 + b_1X \]
What is a good line?

Can we do better than the eyeball method?

We desire a strategy for estimating the slope and intercept parameters in the model \( \hat{Y} = b_0 + b_1 X \).

That involves

- choosing a criteria, i.e., quantifying how good a line is
- and matching that with a solution, i.e., finding the best line subject to that criteria.
Although there are lots of ways to choose a criteria

- only a small handful lead to solutions that are “easy” to compute,

- and which have nice statistical properties (more later).

Most reasonable criteria involve measuring the amount by which the fitted value obtained from the line differs from the observed value of the response value(s) in the data. This amount is called the residual.

- Good lines produce small residuals.

- Good lines produce accurate predictions.
The line is our predictions or fitted values: $\hat{Y}_i = b_0 + b_1 X_i$. The residual $e_i$ is the discrepancy between the fitted $\hat{Y}_i$ and observed $Y_i$ values.

Note that we can write $Y_i = \hat{Y}_i + (Y_i - \hat{Y}_i) = \hat{Y}_i + e_i$. 
Least squares

A reasonable goal is to minimize the size of all residuals:

- If they were all zero we would have a perfect line.
- Trade-off between moving closer to some points and at the same time moving away from other points.

Since some residuals are positive and some are negative, we need one more ingredient.

- $|e_i|$ treats positives and negatives equally.
- So does $e_i^2$, which is easier to work with mathematically.

Least squares chooses $b_0$ and $b_1$ to minimize $\sum_{i=1}^{n} e_i^2$. 
Choose the line to minimize the sum of the squares of the residuals,

\[ \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - [b_0 + b_1 X_i])^2. \]
R’s `lm` command provides a least squares fit.

```r
> reg <- lm(price ~ size)
> reg

Call:
  lm(formula = price ~ size)

Coefficients:
  (Intercept)     size
       38.88       35.39

▶ `lm` stands for “linear model”; it’ll be our workhorse
The least squares line is different than our eyeballed line

... but why do we like it better?
Properties of the least squares fit

Developing techniques for model validation and criticism requires a deeper understanding of the least squares line.

The fitted values ($\hat{Y}_i$) and “residuals” ($e_i$) obtained from the least squares line have some special properties.

- From now on “obtained from the least squares line” will be implied (and therefore not repeated) whenever we talk about $\hat{Y}_i$ and $e_i$.

Let’s look at the housing data analysis to figure out what some of these properties are . . .

...but first, review covariance and correlation.
Correlation and covariance

\[ \text{Cov}(X, Y) = \mathbb{E} [(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \]

\(X\) and \(Y\) vary with each other around their means.
\[ \text{corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \]
Correlation only measures **linear** relationships:

- \( \text{corr}(X, Y) = 0 \) does not mean the variables are unrelated!

Also be careful with influential observations.
Least squares properties

1. The fitted values are **perfectly** correlated with the inputs.

```r
> plot(size, reg$fitted, pch=20, xlab="X", +       ylab="Fitted Values")
> text(x=3, y=80, col=2, cex=1.5, +       paste("corr(y.hat, x) =", cor(size, reg$fitted)))
```

![Graph showing fitted values perfectly correlated with inputs](chart.png)

corr(y.hat, x) = 1
2. The residuals have **zero** correlation with inputs, i.e. “stripped of all linearity”.

```r
> plot(size, reg$fitted-price, pch=20, xlab="X", ylab="Residuals")
> text(x=3.1, y=26, col=2, cex=1.5,
+       paste("corr(e, x) = ", round(cor(size, reg$fitted-price),2)))
> text(x=3.1, y=19, col=4, cex=1.5,
+       paste("mean(e) = ", round(mean(reg$fitted-price),0)))
> abline(h=0, col=8, lty=2)
```
Intuition for the relationship between $\hat{Y}$, $e$, and $X$?

- Let's consider some “crazy” alternative line:
This is a bad fit! We are underestimating the value of small houses and overestimating the value of big houses.

\[ \text{corr}(e, x) = -0.7 \]
\[ \text{mean}(e) = 1.8 \]

Clearly, we have left some predictive ability on the table!
As long as the correlation between $e$ and $X$ is non-zero, we could always adjust our prediction rule to do better:

$$\min \sum_{i=1}^{n} e_i^2 \quad \text{equivalent to} \quad \text{corr}(e, X) = 0 \quad \text{and} \quad 1/n \sum_{i=1}^{n} e_i = 0$$

We need to exploit all of the (linear!) predictive power in the $X$ values and put this into $\hat{Y}$,

- leaving no “$X$ness” in the residuals.

**In Summary:** $Y = \hat{Y} + e$ where:

- $\hat{Y}$ is “made from $X$”; $\text{corr}(X, \hat{Y}) = 1$;
- $e$ is unrelated to $X$; $\text{corr}(X, e) = 0$. 
What is a good line?  

Statistics version!

In a happy coincidence, the least squares line makes good statistical sense too. To see why, we need a model.

Simple Linear Regression (SLR) Model:

\[ Y = \beta_0 + \beta_1 X + \varepsilon \]

What’s important right now?  

- It is a model, so we are assuming this relationship holds for some fixed, but unknown values of \( \beta_0, \beta_1 \).
- The error \( \varepsilon \) is independent of \( X \), mean zero.

(more next week)
The model $Y = \beta_0 + \beta_1 X + \varepsilon$ is something we assume but don’t know. Can we estimate it?

$$\text{cov}(X, Y) = \text{cov}(X, \beta_0 + \beta_1 X + \varepsilon)$$
$$= \text{cov}(X, \beta_1 X)$$
$$= \beta_1 \text{var}(X)$$

Re-write this to get: $\beta_1 = \frac{\text{cov}(X, Y)}{\text{var}(X)} = \text{corr}(X, Y) \frac{\sigma_y}{\sigma_x}$. 

▶ Interpretation: $\beta_1$ is correlation in units $Y$ per units $X$.

But still unknown!

▶ Replace with analogues from the data:

$$b_1 = \frac{\text{sample } \text{cov}(X, Y)}{\text{sample } \text{var}(X)} = r_{xy} \frac{s_y}{s_x}$$
To summarize:

R’s `lm(Y ~ X)` function

- finds the coefficients $b_0$ and $b_1$ characterizing the “least squares” line $\hat{Y} = b_0 + b_1X$.
- That is it minimizes $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$.
- Equivalent to: $\text{corr}(e, X) = 0$ & $\frac{1}{n} \sum_{i=1}^{n} e_i = 0$

The least squares formulas are (cf slide 40)

$$b_1 = r_{xy} \frac{s_y}{s_x} \quad \text{and} \quad b_0 = \bar{Y} - b_1 \bar{X}.$$
Example: wage data

Data:

- 39 demographic groupings of 6000 households with the male head earning less than $15,000 annually in 1966.

Problem:

- Estimate the relationship between pay and labor supply.
- Use this information to influence social policy decisions and the debate on a guaranteed national wage.

Possible solution:

- Fit a linear model for the effect of pay on labor supply for the working poor.
Read in the data, and change to $X$ and $Y$ for convenience.

```r
> D <- read.csv("wages.csv")
> Y <- D$HRS
> X <- D$RATE
```

Use correlation to slope the line

```r
> b1 <- cor(X,Y)*sd(Y)/sd(X)
> b1
[1] 80.93679
```

Make sure that $\bar{X}$ and $\bar{Y}$ are on the line.

```r
> b0 <- mean(Y) - mean(X)*b1
> b0
[1] 1913.009
```
> plot(X,Y, xlab="rate", ylab="hours")
> abline(b0, b1, col="red")
The residual errors are defined as $e_i = Y_i - (b_0 + b_1 X_i)$.

```
> e <- Y - (b0 + b1*X)
> plot(X, e, xlab="rate", ylab="residuals", pch=20)
> abline(h=0, col=8, lwd=2)
```

What does this plot of $e_i$ versus $X_i$ say about the fit?
How does labor supply change with pay?

We’ve estimated the linear model:

$$\text{hours} = 1913 + 81 \times \text{rate} + e$$

**It increases:** Every dollar extra per hour leads to an increase of 81 expected annual hours worked.

What would this mean for the debate (at the time of the study) on national wage/negative income tax?
Questions remain ...

1. Does this apply today?
2. Does the same hold for those who make more money?
3. How good is this estimate?
4. Is pay rate the only variable of importance?
5. Do higher wages cause more work?
Steps in a regression analysis

1. State the problem
2. Data collection
3. Model fitting & estimation (*this class*)
   3.1 Model specification (linear? logistic?)
   3.2 Select potentially relevant variables
   3.3 Model fitting (least squares)
   3.4 Model validation and criticism
   3.5 Back to 3.1? Back to 2?
4. Answering the posed questions

But that oversimplifies a bit;

- it is more iterative, and can be more art than science.
Course Overview

W1: Introduction, The simple linear regression (SLR) model
W2: Inference for SLR
W3: Multiple linear regression (MLR)
W4: Logistic Regression
W5: Time Series Data
W6: Midterm Exam & Lecture: Panel and Clustered Data
W7: Regression Issues and Diagnostics
W8: Model Building 1
W9: Model Building 2 & Causal Inference
W10: Discrete Choice and Count Data (if time)
W11: Final exam and Project due
Administrative details

Syllabus
▶ Lots of info! Please check before you ask. Pretty please.

Slides
▶ Not a stand-alone resource for learning

Course Website
▶ All class material is already posted
▶ HW solutions posted after due date

Piazza Q&A
▶ Post all questions here!
▶ Student & Instructor Answers

More help
▶ Office hours
▶ See syllabus for: books, R help, etc
Your work

Turned-in work: clear, concise, and on message
  ▶ Fewer plots usually better
  ▶ Results and analysis, not output/code

Homework
  ▶ Not exam practice! Not similar at all
  ▶ Reinforce & extend ideas, challenge you
  ▶ Open-ended analysis

Exams
  ▶ Narrower scope
  ▶ Test core concepts/abilities
  ▶ Look at sample exams to get a sense of style

Project: Your glimpse at real life!
Glossary of symbols

- \( X \) = input, explanatory variable, or covariate.
- \( Y \) = output, dependent variable, or response.
- \( s_{xy} \) is covariance and \( r_{xy} \) is the correlation, \( s_x \) and \( s_y \) are standard deviation of \( X \) and \( Y \) respectively.
- \( r_{xy} = \frac{s_{xy}}{s_x s_y} \).
- \( b_0 \) = least squares estimate of the intercept
- \( b_1 \) = least squares estimate of the slope
- \( \hat{Y} \) is the fitted value \( b_0 + b_1 X \)
- \( e_i \) is the residual \( Y_i - \hat{Y}_i \).