INTRODUCTION TO CLINICAL RESEARCH

Introduction to Linear Regression

Karen Bandeen-Roche, Ph.D.

July 23, 2013

Acknowledgements

• Marie Diener-West

• Rick Thompson

• ICTR Leadership / Team
Outline

1. Regression: Studying association between (health) outcomes and (health) determinants
2. Correlation
3. Linear regression: Characterizing relationships
4. Linear regression: Prediction
5. Future topics: multiple linear regression, assumptions, complex relationships

Introduction

• 30,000-foot purpose: Study association of continuously measured health outcomes and health determinants

• Continuously measured outcomes (“linear”)
  – No gaps
  – Total lung capacity (l) and height (m)
  – Birthweight (g) and gestational age (mos)
  – Systolic BP (mm Hg) and salt intake (g)
  – Systolic BP (mm Hg) and drug (trt, placebo)
Introduction

- 30,000-foot purpose: Study association of continuously measured health outcomes and health determinants

- Association
  - Connection
  - Determinant predicts the outcome
  - A query: Do people of taller height tend to have a larger total lung capacity (l)?

Example: Association of total lung capacity with height

Study: 32 heart lung transplant recipients aged 11-59 years

. list tlc height age in 1/10

+----------+-----+-----+-----+----------+-----+-----+
| tlc     | height | age |
+----------+-----+-----+-----+----------+-----+-----+
| 3.41     | 138  | 11  |
| 3.4      | 149  | 35  |
| 8.05     | 162  | 20  |
| 5.73     | 160  | 23  |
| 4.1      | 157  | 16  |
| 5.44     | 166  | 40  |
| 7.2      | 177  | 39  |
| 6.       | 173  | 29  |
| 4.55     | 152  | 16  |
| 4.83     | 177  | 35  |
+----------+-----+-----+-----+----------+-----+-----+
Introduction

- Two analyses to study association of continuously measured health outcomes and health determinants
  - Correlation analysis: Concerned with measuring the strength and direction of the association between variables. The correlation of X and Y (Y and X).
  - Linear regression: Concerned with predicting the value of one variable based on (given) the value of the other variable. The regression of Y on X.

Correlation Analysis

Some specific names for “correlation” in one’s data:
- r
- Sample correlation coefficient
- Pearson correlation coefficient
- Product moment correlation coefficient
Correlation Analysis

• Characterizes the extent of linear relationship between two variables, and the direction

  – How closely does a straight-line trend (non-flat) characterize the relationship of the two variables?
    • Exactly: $r = 1$ or -1
    • Not at all (e.g. flat relationship): $r=0$
    • $-1 \leq r \leq 1$

  – Does one variable tend to increase as the other increases ($r>0$), or decrease as the other increases ($r<0$)

Types of Correlation

(continued)
Examples of Relationships and Correlations

Correlation: Lung Capacity Example

$r = 0.865$
FYI: Sample Correlation Formula

\[ r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}} \]

Heuristic: If I draw a straight line through the vertical middle of scatter of points created by plotting y versus x, \( r \) divides the SD of the heights of points on the line by the SD of the heights of the original points.

Correlation – Closing Remarks

- The value of \( r \) is independent of the units used to measure the variables.
- The value of \( r \) can be substantially influenced by a small fraction of outliers.
- The value of \( r \) considered “large” varies over science disciplines:
  - Physics: \( r=0.9 \)
  - Biology: \( r=0.5 \)
  - Sociology: \( r=0.2 \)
- \( r \) is a “guess” at a population analog.
Linear regression

• Aims to predict the value of a health outcome, Y, based on the value of an *explanatory* variable, X.

  – What is the relationship between average Y and X?
    • The analysis “models” this as a line
    • We care about “slope”—size, direction
    • Slope=0 corresponds to “no association”

  – How precisely can we predict a given person’s Y with his/her X?

Linear regression – Terminology

• Health outcome, Y
  – Dependent variable
  – Response variable

• Explanatory variable (predictor), X
  – Independent variable
  – Covariate
Linear regression - Relationship

- Model: \( Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \)

Linear regression - Relationship

- In words
  - Intercept \( \beta_0 \) is mean \( Y \) at \( X=0 \)
  - … mean lung capacity among persons with 0 height
  - Recommendation: “Center”
    - Create new \( X^* = (X-165) \), regress \( Y \) on \( X^* \)
    - Then: \( \beta_0 \) is mean lung capacity among persons 165 cm
  - Slope \( \beta_1 \) is change in mean \( Y \) per 1 unit difference in \( X \)
  - … difference in mean lung capacity comparing persons who differ by 1 cm in height
  - … irrespective of centering
  - Measures association (=0 if slope=0)
Linear regression – Sample inference

• We develop best guesses at $\beta_0, \beta_1$ using our data
  – Step 1: Find the “least squares” line
    • Tracks through the middle of the data “as best possible”
    • Has intercept $b_0$ and slope $b_1$ that make sum of $[Y_i - (b_0 + b_1 X_i)]^2$ smallest
  – Step 2: Use the slope and intercept of the least squares line as best guesses
    • Can develop hypothesis tests involving $\beta_1, \beta_0$ using $b_1, b_0$
    • Can develop confidence intervals for $\beta_1, \beta_0$ using $b_1, b_0$

Linear regression – Lung capacity data
Linear regression – Lung capacity data

• In STATA - “regress” command:
  Syntax  “regress yvar xvar”

```
. regress tlc height

Source |       SS       df       MS              Number of obs =      32
-----------+------------------------------           F(  1,    30) =   89.12
Model |  93.7825029     1  93.7825029           Prob > F      =  0.0000
Residual |  31.5694921    30   1.0523164           R-squared     =  0.7482
-----------+------------------------------           Adj R-squared =  0.7398
Total |  125.351995    31  4.04361274           Root MSE      =  1.0258

------------------------------------------------------------------------------
  tlc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+----------------------------------------------------------------
  height |   .1417377    .015014     9.44   0.000     .1110749    .1724004
      _cons |  -17.10484   2.516234    -6.80   0.000    -22.24367     -11.966

```

TLC of -17.1 liters among persons of height = 0
If centered at 165 cm: TLC of 6.3 liters
among persons of height = 165

On average, TLC increases by 0.142 liters per cm
increase in height, or equivalently, by 1.42 liters
per 10 cm increase in height.
Linear regression – Lung capacity data

- Inference: p-value tests the null hypothesis that the coefficient = 0.

```
regress tlc height
Source |       SS       df       MS              Number of obs =      32
-------------+------------------------------           F(  1,    30) =   89.12
Model |  93.7825029     1  93.7825029           Prob > F      =  0.0000
Residual |  31.5694921    30   1.0523164           R-squared     =  0.7482
-------------+------------------------------           Adj R-squared =  0.7398
Total |  125.351995    31  4.04361274           Root MSE      =  1.0258
------------------------------------------------------------------------------
tlc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
height |   .1417377    .015014     9.44   0.000     .1110749    .1724004
_cons |  -17.10484   2.516234    -6.80   0.000    -22.24367     -11.966
------------------------------------------------------------------------------
```

We reject the null hypothesis of 0 slope (no linear relationship).
The data support a tendency for TLC to increase with height.

---

t=coef/std.err: bigger than 2 roughly corresponds to p<0.05

We reject the null hypothesis of 0 slope (no linear relationship).
The data support a tendency for TLC to increase with height.
Linear regression – Lung capacity data

- Inference: Confidence interval for coefficients; these both exclude 0.

```
. regress tlc height
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>93.7825029</td>
<td>1</td>
<td>93.7825029</td>
<td>F( 1, 30) = 89.12</td>
</tr>
<tr>
<td>Residual</td>
<td>31.5694921</td>
<td>30</td>
<td>1.0523164</td>
<td>R-squared = 0.7482</td>
</tr>
<tr>
<td>Total</td>
<td>125.351995</td>
<td>31</td>
<td>4.04361274</td>
<td>Root MSE = 1.0258</td>
</tr>
</tbody>
</table>

| tlc | Coef.   | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|-----|---------|-----------|------|-----|---------------------|
| height | .1417377 | .015014 | 9.44 | 0.000 | .1110749 - .1724004 |
| _cons  | -17.10484 | 2.516234 | -6.80 | 0.000 | -22.24367 - -11.966 |

We are 95% confident that the interval (0.111, 0.172) includes the true slope. Data are consistent with an average per-cm of height increase in TLC ranging between 0.111 and 0.172. The data support a tendency for TLC to increase with height.

Linear regression

- Aims to predict the value of a health outcome, Y, based on the value of an explanatory variable, X.

- What is the relationship between average Y and X?
  - The analysis “models” this as a line
  - We care about “slope”—size, direction
  - Slope=0 corresponds to “no association”

- How precisely can we predict a given person’s Y with his/her X?
Linear regression - Prediction

• What is the linear regression prediction of a given person’s Y with his/her X?
  – Plug X into the regression equation
  – The prediction “\( \hat{Y} \)" = \( b_0 + b_1X \)

• Data Model: \( Y_i = b_0 + b_1X_i + \varepsilon_i \)
Linear regression - Prediction

• What is the linear regression prediction of a given person’s Y with his/her X?
  – Plug X into the regression equation
    \[ \hat{Y} = b_0 + b_1 X \]
  – The prediction “\( \hat{Y} \)” = \( b_0 + b_1 X \)
  – The “residual” \( \epsilon = \text{data-prediction} = Y - \hat{Y} \)

  – Least squares minimizes the sum of squared residuals, e.g. makes predicted Y’s as close to observed Y’s as possible (in the aggregate)

Linear regression - Prediction

• How precisely does \( \hat{Y} \) predict Y?
  – Conventional measure: R-squared

  • Variance of \( \hat{Y} \) / Variance of Y
  • = Proportion of Y variance “explained” by regression
  • = squared sample correlation between \( \hat{Y} \) and Y

  • In examples so far (because only one X):
    = squared sample correlation between Y, X
Linear prediction – Lung capacity data

- Inference: Confidence interval for coefficients; these both exclude 0.

```
. regress tlc height
```

```
Source |       SS       df       MS              Number of obs =  32
-------------+------------------------------           F(  1,    30) =   89.12
Model |  93.7825029     1  93.7825029           Prob > F      =  0.0000
Residual |  31.5694921    30   1.0523164           R-squared     =  0.7482
-------------+------------------------------           Adj R-squared =  0.7398
Total |  125.351995    31  4.04361274           Root MSE      =  1.0258

------------------------------------------------------------------------------
  tlc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  height |   .1417377    .015014     9.44   0.000     .1110749    .1724004
     _cons |  -17.10484   2.516234    -6.80   0.000    -22.24367     -11.966
------------------------------------------------------------------------------
```

R-squared = 0.748: 74.8 % of variation in TLC is characterized by the regression of TLC on height. This corresponds to correlation of sqrt(0.748) = .865 between predictions and actual TLCs. This is a precise prediction.

A correlation of 0.8-0.9

(continued)
Correlation: Lung Capacity Example

$r=.865$

Linear regression - Prediction

- Cautionary comment: In ‘real life’ you’d want to evaluate the precision of your predictions in a sample different than the one with which you built your prediction model

- “Cross-validation”
• To study how mean TLC varies with height...
  – Could dichotomize height at median and compare TLC between two height groups using a two-sample t-test
Lung capacity example – two height groups

```
ttest tlc, by(height_above_med) unequal
Two-sample t test with unequal variances
```

<table>
<thead>
<tr>
<th>Group</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Err.</th>
<th>Std. Dev.</th>
<th>[5% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;= Median</td>
<td>16</td>
<td>3.024375</td>
<td>.3286601</td>
<td>1.314641</td>
<td>4.323858</td>
</tr>
<tr>
<td>&gt; Median</td>
<td>16</td>
<td>8.190625</td>
<td>.2974915</td>
<td>1.189966</td>
<td>7.516537</td>
</tr>
<tr>
<td>combined</td>
<td>32</td>
<td>6.097070</td>
<td>.5524750</td>
<td>2.010874</td>
<td>5.865203</td>
</tr>
<tr>
<td>diff</td>
<td>-3.12525</td>
<td>.4433043</td>
<td>-4.031973</td>
<td>-2.220527</td>
<td></td>
</tr>
</tbody>
</table>

Satterthwaite's degrees of freedom: 29.797

Ho: mean(Median) - mean(Above Med) = diff = 0
Hα: diff < 0
   t = -7.9522
   P < t = 0.0000

Hα: diff = 0
   t = -7.0922
   P > |t| = 0.0000

Hα: diff > 0
   P > t = 1.0000

Could replicate this analysis with SLR of TLC on X=1 if height > median and X=0 otherwise

More advanced topics
Regression with more than one predictor

• “Multiple” linear regression
  – More than one X variable (ex.: height, age)
  – With only 1 X we have “simple” linear regression

  $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

• Intercept $\beta_0$ is mean Y for persons with all Xs=0

• Slope $\beta_k$ is change in mean Y per 1 unit difference in $X_k$ among persons identical on all other Xs
More advanced topics
Regression with more than one predictor

• Slope $\beta_k$ is change in mean Y per 1 unit difference in $X_k$ among persons identical on all other Xs
  – i.e. holding all other Xs constant
  – i.e. “controlling for” all other Xs

• Fitted slopes for a given predictor in a simple linear regression and a multiple linear regression controlling for other predictors do NOT have to be the same
  – We’ll learn why in the lecture on confounding

More advanced topics
Assumptions

• Most published regression analyses make statistical assumptions

• Why this matters: p-values and confidence intervals may be wrong, and coefficient interpretation may be obscure, if assumptions aren’t approximately true

• Good research reports on analyses to check whether assumptions are met (“diagnostics”, “residual analysis”, “model checking/fit”, etc.)
More advanced topics
Linear Regression Assumptions

- Units are sampled independently (no connections such as familial relationship, residential clustering, etc.)
- Posited model for average Y-X relationship is correct
- Normally (Gaussian; bell-shaped) distributed responses for each X
- Variability of responses (Ys) the same for all X

Assumptions well met:
More advanced topics
Linear Regression Assumptions

Non-normal responses per X

Non-constant variability of responses per X
More advanced topics
Linear Regression Assumptions

Lung capacity example

![Graph showing linear relationship between height and residual values.]

More advanced topics
Types of relationships that can be studied

• ANOVA (multiple group differences)

• ANCOVA (different slopes per groups)
  – Effect modification: lecture to come

• Curves (polynomials, broken arrows, more)

• Etc.
Main topics once again

1. Regression: Studying association between (health) outcomes and (health) determinants
2. Correlation
3. Linear regression: Characterizing relationships
4. Linear regression: Prediction
5. Future topics: assumptions, model checking, complex relationships