My Hobby: Extrapolating

As you can see, by late next month you'll have over four dozen husbands. Better get a bulk rate on wedding cake.

Number of Husbands

Yesterday

Today

Linear Regression

Anna Leontjeva

anna.leontjeva@ut.ee
Which of the following is most related to linear regression?

1) Information Gain  
2) Linear Atavism  
3) Regression to Mean  
4) Method of Least Squares
Introduction to Linear Regression

Linear regression is an approach to modeling the relationship between a response variable $Y$ and one or more explanatory variables denoted $X$ (predictors), e.g., regression is the study of dependence.

A response variable $Y$ must be continuous.

The case of one explanatory variable is called simple regression.
More than one explanatory variable is multiple regression.
```
> head(heights[,c(3,4)])
Mheight_cm  Dheight_cm
1   151.638  139.954
2   147.828  143.510
3   153.924  142.240
4   154.178  144.272
5   156.972  142.240
6   140.970  147.066
```
Sketch on each plot what you think is the best-fitting line for predicting $y$ from $x$. 
# Short quiz

<table>
<thead>
<tr>
<th>pic</th>
<th>y prediction</th>
<th>Residual sum of squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• Cross at the average $y$-value for each $x$ and draw the best-fitting line to the crosses
• Re-compute the $y$ prediction and sum of squared errors.
Linear Regression Function

\[ E(Y \mid X = x) = \beta_0 + \beta_1 x \]
Linear Regression Function

\[ E(Y | X = x) = \beta_0 + \beta_1 x \]

Mean function
Intercept
Slope
Linear Regression Function

\[ E(Y | X = x) = \beta_0 + \beta_1 x \]

- Intercept
- Slope
- Mean function

\[ E(D\text{height} | M\text{height} = x) = \beta_0 + \beta_1 x \]

Intercept and slope are unknown, want to estimate
Linear regression function

\( \beta_0 = \text{intercept} \)
Residuals (Errors)

\[ \hat{e}_i = y_i - \hat{y}_i \]

Figure 2.2  A scatter plot of data with a line of best fit and the residuals identified
Objective function

residual sum of squares (RSS, SSE):

\[ RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \]

Ordinary Least Squares (OLS)
Minimization

\[
\frac{\partial \text{RSS}(\beta_0, \beta_1)}{\beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0
\]

\[
\frac{\partial \text{RSS}(\beta_0, \beta_1)}{\beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0
\]

\[
\beta_0 n + \beta_1 \sum x_i = \sum y_i
\]

\[
\beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i
\]

\[
\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}
\]

\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
\]
Example

\[ b_1 = \frac{\text{sum}((x - \text{mean}(x))(y - \text{mean}(y)))}{\text{sum}((x - \text{mean}(x))^2)} \]

\[ [1] \ 0.541747 \]

\[ b_0 = \text{mean}(y) - b_1 \times \text{mean}(x) \]

\[ [1] \ 75.99029 \]
Example

\[ b1 = \frac{\text{sum}\left((x - \text{mean}(x)) \ast (y - \text{mean}(y))\right)}{\text{sum}\left((x - \text{mean}(x))^2\right)} \]

[1] 0.541747

\[ b0 = \text{mean}(y) - b1 \ast \text{mean}(x) \]

[1] 75.99029

\[ b1 = \frac{\text{cov}(x, y)}{\text{var}(x)} \]
Example

$$\text{lm}(y \sim x)$$
Example

Call:
`lm(formula = Dheight_cm ~ Mheight_cm, data = heights)`

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-18.7894</td>
<td>-3.8828</td>
<td>0.0914</td>
<td>3.7900</td>
<td>22.9934</td>
</tr>
</tbody>
</table>

Coefficients:

|                  | Estimate | Std. Error | t value | Pr(>|t|) |
|------------------|----------|------------|---------|---------|
| (Intercept)      | 75.99029 | 4.12107    | 18.44   | <2e-16 *** |
| Mheight_cm       | 0.54175  | 0.02596    | 20.87   | <2e-16 *** |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.756 on 1373 degrees of freedom
Multiple R-squared: 0.2408, Adjusted R-squared: 0.2402
F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16
Example

Call:
`lm(formula = Dheight_cm ~ Mheight_cm, data = heights)`

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-18.7894</td>
<td>-3.8828</td>
<td>0.0914</td>
<td>3.7900</td>
<td>22.9934</td>
</tr>
</tbody>
</table>

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 75.99029 | 4.12107    | 18.44   | <2e-16 *** |
| Mheight_cm     | 0.54175  | 0.02596    | 20.87   | <2e-16 *** |

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.756 on 1373 degrees of freedom
Multiple R-squared: 0.2408, Adjusted R-squared: 0.2402
F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16

\[ y = 75.99 + 0.54 M\_height\_cm \]
Multiple regression

Usually we have more than one variable:

\[ Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + e_i \]

or in matrix notation:

\[ Y = X\beta + e \]
Matrix notation

\[ Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \]

\[ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \text{and} \quad e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} \]

n observations, p explanatory variables,
\( \dim(Y) = n \times 1, \dim(X) = n \times (p+1), \dim(\beta) = (p+1) \times 1, \)
\( \dim(e) = n \times 1 \)
OLS for multiple regression

\[ RSS = (y - X\beta)^T (y - X\beta) = (y^T - \beta^T X^T) (y - X\beta) = \]
\[ = y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta = \]
\[ = y^T y - 2y^T X\beta + \beta^T X^T X\beta \]

\[ \frac{\partial RSS}{\partial \beta} = -2X^T y + 2X^T X\beta = 0 \]

\[ X^T X\beta = X^T y \]

\[ \beta = (X^T X)^{-1} X^T y \]
Example

\[ b = \text{solve}(t(X) \cdot X) \cdot t(X) \cdot y \]
Example

\[ b = \text{solve}(t(X) \times X) \times t(X) \times y \]

\[ b = \text{ginv}(X) \times y \]

\[ \text{lm}(y \sim X) \]

\[ \text{lm}(y \sim x1 + x2) \]
Types of predictors

• The **intercept** (model can be with or without);
  \[ \text{lm}(y \sim x_1 + x_2 - 1) \]

• **Transformations** of predictors
  \[ \text{lm}(y \sim x_1 + \log(x_2)) \]

• **Polynomials**
  \[ \text{lm}(y \sim x_1 + I(x_2^2)) \]

• **Interactions** and other combinations of predictors
  \[ \text{lm}(y \sim x_1/x_2) \]

• **Dummy variables** and factors
  \[ \text{lm}(y \sim \text{is\_male}) \]
Polynomials

the predictors are a single predictor, $x$, and its polynomial powers ($x^2$, $x^3$, etc.)
Polynomials

$$m2 \leftarrow \text{lm}(\text{Salary} \sim \text{Experience} + \text{I}((\text{Experience})^2), \text{data} = \text{prof})$$
Quiz: What does it mean: linear?

In which case we cannot use linear regression?

1. \[ y = \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_3 \]

2. \[ y = \beta_1 x_1 + \beta_1^2 x_2 + \beta_3 x_3 \]

3. \[ y = \beta_1 x_1 x_2 x_3 \]
Quiz: What does it mean: linear?
**Dummy variables**

- Are binary variables (i.e., 0 or 1) created from a variable with the higher level of measurement (categorical variable):

<table>
<thead>
<tr>
<th>Eye color</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>1</td>
</tr>
<tr>
<td>Blue</td>
<td>2</td>
</tr>
<tr>
<td>Grey</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Eye color</th>
<th>Is_Brown</th>
<th>Is_Blue</th>
<th>Is_Grey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Blue</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Grey</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Dummy variables

- Are binary variables (i.e., 0 or 1) created from a variable with the higher level of measurement (categorical variable):

<table>
<thead>
<tr>
<th>Eye color</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>1</td>
</tr>
<tr>
<td>Blue</td>
<td>2</td>
</tr>
<tr>
<td>Grey</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Eye color</th>
<th>Is_Brown</th>
<th>Is_Blue</th>
<th>Is_Grey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Blue</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Grey</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Call:  
`lm(formula = salary ~ yrs.since.phd + sex, data = Salaries)`

Residuals:  
```
  Min     1Q   Median     3Q    Max
-84167  -19735   -2551  15427  102033
```

Coefficients:  
```
                 Estimate Std. Error t value  Pr(>|t|)
(Intercept)     85181.8     4748.3  17.939  <2e-16 ***
yrs.since.phd   958.1       108.3   8.845  <2e-16 ***
sexMale         7923.6     4684.1  1.692    0.0915 
```

Salary for males: $85181.8 + 958.1 \text{ yrs.since.phd} + 7923.6 \times 1 = 93105.4 + 958.1 \text{ yrs.since.phd}$

Salary for females: $85181.8 + 958.1 \text{ yrs.since.phd} + 7923.6 \times 0 = 85181.8 + 958.1 \text{ yrs.since.phd}$
Diagnostics

\[ \hat{y} = 3.0 + 0.5x. \]
Leverage points

Demo: [http://www.stat.sc.edu/~west/javahtml/Regression.html](http://www.stat.sc.edu/~west/javahtml/Regression.html)

![Graphs showing good and bad leverage points](image-url)

- **Good leverage point**:
  \[y = 2.946 + 0.97958x, \quad r = 0.99840766\]
  \[y = 1.9750478 + 1.0067168x, \quad r = 0.9969587\]

- **Bad leverage point**:
  \[y = 2.946 + 0.97958x, \quad r = 0.99840766\]
  \[y = 72.57289 + 0.53454214x, \quad r = 0.52276415\]
• a **leverage point** is an observation that has an extreme value on one or more explanatory variables.

• a point is a **bad leverage point** if its $Y$-value does not follow the pattern set by the other data points.

• a **bad leverage point** is a leverage point which is also an outlier.
Standardized residuals

\[ \text{Standardized residual}_i = \frac{\text{Residual}_i}{\text{St. deviation of residuals}} \]
Goodness-of-fit-measures

• R-squared
  (square of the sample correlation coefficient between the outcomes and their predicted values)

• Coefficient Significance:
  (used to test the hypothesis that the true value of the coefficient is non-zero, in order to confirm that the independent variable really belongs in the model)

• Measures on the test set (RSS, R-squared)
Over- and underfitting
Regularization

• Simple objective function:
  \[ \min(\text{Error}) \]

• … with regularization:
  \[ \min(\text{Error} + \lambda \text{ Complexity}) \]
Regularization

• Simple objective function:
  \[ \text{min}(\text{Error}) \]

• … with regularization:
  \[ \text{min}(\text{Error} + \lambda \text{ Complexity}) \]

Penalty for more complex models: with larger values of lambda, greater penalty – more compact model
Regularization

• OLS objective function:
  \[ \min(\sum e^2) \]

• OLS with regularization (Ridge regression):
  \[ \min(\sum e^2 + \lambda \sum \beta_i^2) \]
Regularization

• OLS objective function:
  \[ \min(\sum e^2) \]
  
  \[ \beta = (X^T X)^{-1}X^T y \]

• OLS with regularization (Ridge regression):
  \[ \min(\sum e^2 + \lambda \sum \beta_i^2) \]
  
  \[ \beta = (X^T X + \lambda I)^{-1}X^T y \]
Literature

- A modern approach to Regression with R, Simon Sheather;
- Applied linear regression, Weisberg