UCLA Department of Statistics
Statistical Consulting Center

Regression in R
Part I:
Simple Linear Regression

Denise Ferrari & Tiffany Head
denise@stat.ucla.edu tiffany@stat.ucla.edu

Feb 10, 2010
Objective

The main objective of this mini-course is to show how to perform Regression Analysis in R. Prior knowledge of the basics of Linear Regression Models is assumed.
Installing R on a Mac

1. Go to http://cran.r-project.org/ and select MacOS X
2. Select to download the latest version: 2.8.1 (2008-12-22)
3. Install and Open. The R window should look like this:
For help with any function in R, put a question mark before the function name to determine what arguments to use, examples and background information.

```
1 ?plot
```
When downloading data from the internet, use `read.table()`.

In the arguments of the function:

- **header**: if TRUE, tells R to include variables names when importing
- **sep**: tells R how the entries in the data set are separated
  - `sep = ","`: when entries are separated by COMMAS
  - `sep = \t`: when entries are separated by TAB
  - `sep = " "`: when entries are separated by SPACE

```r
data <- read.table("http://www.stat.ucla.edu/data/moore/TAB1-2.DAT", header=FALSE, sep = " ")
```
Data from Your Computer

- Check the current R working folder:

  ```
  getwd()
  ```

- Move to the folder where the data set is stored (if different from (1)). Suppose your data set is on your desktop:

  ```
  setwd("~/Desktop")
  ```

- Now use `read.table()` command to read in the data:

  ```
  data <- read.table(<name>, header=TRUE, sep="" )
  ```
R Data Sets

- To use a data set available in one of the R packages, install that package (if needed).
- Load the package into R, using the `library()` command:
  ```r
  library(MASS)
  ```
- Extract the data set you want from that package, using the `data()` command. Let’s extract the data set called `Pima.tr`:
  ```r
  data(Boston)
  ```
When to Use Regression Analysis?

Regression analysis is used to describe the relationship between:

- A single response variable: \( Y \); and
- One or more predictor variables: \( X_1, X_2, \ldots, X_p \)
  - \( p = 1 \): Simple Regression
  - \( p > 1 \): Multivariate Regression
The Variables

**Response Variable**

The response variable $Y$ must be a *continuous* variable.

**Predictor Variables**

The predictors $X_1, \ldots, X_p$ can be *continuous*, *discrete* or *categorical* variables.
Initial Data Analysis I

Does the data look like as we expect?

Prior to any analysis, the data should always be inspected for:

- Data-entry errors
- Missing values
- Outliers
- Unusual (e.g. asymmetric) distributions
- Changes in variability
- Clustering
- Non-linear bivariate relationships
- Unexpected patterns
Initial Data Analysis II

Does the data look like as we expect?

We can resort to:

- **Numerical summaries:**
  - 5-number summaries
  - correlations
  - etc.

- **Graphical summaries:**
  - boxplots
  - histograms
  - scatterplots
  - etc.
Loading the Data

Example: Diabetes in Pima Indian Women \(^1\)

- Clean the workspace using the command: `rm(list=ls())`
- Download the data from the internet:
  ```r
  ```
- Name the variables:
  ```r
  colnames(pima) <- c("npreg","glucose","bp","triceps","insulin","bmi","diabetes","age","class")
  ```

\(^1\)Data from the UCI Machine Learning Repository

pima-indians-diabetes.names
Having a peek at the Data

Example: Diabetes in Pima Indian Women

- For small data sets, simply type the name of the data frame.
- For large data sets, do:

```r
head(pima)
```

```
npreg glucose bp triceps insulin bmi diabetes age class
1   6  148  72   35    0  33.6   0.627   50    1
2   1   85  66   29    0  26.6   0.351   31    0
3   8  183  64    0    0  23.3   0.672   32    1
4   1   89  66   23   94  28.1   0.167   21    0
5   0  137  40   35 168  43.1   2.288   33    1
6   5  116  74    0    0  25.6   0.201   30    0
```
Univariate summary information:

- Look for unusual features in the data (data-entry errors, outliers): check, for example, min, max of each variable

```r
summary(pima)
```

<table>
<thead>
<tr>
<th></th>
<th>npreg</th>
<th>glucose</th>
<th>bp</th>
<th>triceps</th>
<th>insulin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>0.000</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>1.000</td>
<td>1st Qu.: 99.0</td>
<td>1st Qu.: 62.0</td>
<td>1st Qu.: 0.0</td>
<td>1st Qu.: 0.0</td>
</tr>
<tr>
<td>Median</td>
<td>3.000</td>
<td>Median :117.0</td>
<td>Median : 72.0</td>
<td>Median :23.0</td>
<td>Median : 30.5</td>
</tr>
<tr>
<td>Mean</td>
<td>3.845</td>
<td>Mean :120.9</td>
<td>Mean : 69.1</td>
<td>Mean :20.54</td>
<td>Mean : 79.8</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>6.000</td>
<td>3rd Qu.:140.2</td>
<td>3rd Qu.: 80.0</td>
<td>3rd Qu.:32.0</td>
<td>3rd Qu.:127.2</td>
</tr>
<tr>
<td>Max.</td>
<td>17.000</td>
<td>Max. :199.0</td>
<td>Max. :122.0</td>
<td>Max. :99.0</td>
<td>Max. :846.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>bmi</th>
<th>diabetes</th>
<th>age</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>0.0</td>
<td>Min. :0.0780</td>
<td>Min. :21.00</td>
<td>Min. :0.0000</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>27.30</td>
<td>1st Qu.:0.2437</td>
<td>1st Qu.:24.00</td>
<td>1st Qu.:0.0000</td>
</tr>
<tr>
<td>Median</td>
<td>32.00</td>
<td>Median :0.3725</td>
<td>Median :29.00</td>
<td>Median :0.0000</td>
</tr>
<tr>
<td>Mean</td>
<td>31.99</td>
<td>Mean :0.4719</td>
<td>Mean :33.24</td>
<td>Mean :0.3490</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>36.60</td>
<td>3rd Qu.:0.6262</td>
<td>3rd Qu.:41.00</td>
<td>3rd Qu.:1.0000</td>
</tr>
<tr>
<td>Max.</td>
<td>67.10</td>
<td>Max. :2.4200</td>
<td>Max. :81.00</td>
<td>Max. :1.0000</td>
</tr>
</tbody>
</table>
Coding Missing Data I

Example: Diabetes in Pima Indian Women

- Variable “npreg” has maximum value equal to 17
  - unusually large but not impossible
- Variables “glucose”, “bp”, “triceps”, “insulin” and “bmi” have minimum value equal to zero
  - in this case, it seems that zero was used to code missing data
Coding Missing Data II

Example: Diabetes in Pima Indian Women

R code for missing data

- Zero should **not** be used to represent missing data
  - it’s a valid value for some of the variables
  - can yield misleading results

- Set the missing values coded as zero to NA:

```
1   pima$glucose[pima$glucose==0] <- NA
2   pima$bp[pima$bp==0] <- NA
3   pima$triceps[pima$triceps==0] <- NA
4   pima$insulin[pima$insulin==0] <- NA
5   pima$bmi[pima$bmi==0] <- NA
```
Coding Categorical Variables

Example: Diabetes in Pima Indian Women

- Variable “class” is categorical, not quantitative

R code for categorical variables

- Categorical should not be coded as numerical data
  - problem of “average zip code”

- Set categorical variables coded as numerical to factor:

```r
1  pima$class <- factor (pima$class)
2  levels(pima$class) <- c("neg", "pos")
```
### Initial Data Analysis

#### Final Coding

Example: Diabetes in Pima Indian Women

```r
summary(pima)
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
<th>NA’s</th>
</tr>
</thead>
<tbody>
<tr>
<td>npreg</td>
<td>0.00</td>
<td>1.00</td>
<td>3.00</td>
<td>3.85</td>
<td>6.00</td>
<td>17.0</td>
<td>5.0</td>
</tr>
<tr>
<td>glucose</td>
<td>44.0</td>
<td>99.0</td>
<td>117.0</td>
<td>121.7</td>
<td>141.0</td>
<td>199.0</td>
<td>1.0</td>
</tr>
<tr>
<td>bp</td>
<td>24.0</td>
<td>64.0</td>
<td>72.0</td>
<td>72.4</td>
<td>80.0</td>
<td>122.0</td>
<td>5.0</td>
</tr>
<tr>
<td>triceps</td>
<td>7.00</td>
<td>22.00</td>
<td>29.00</td>
<td>29.15</td>
<td>36.00</td>
<td>99.0</td>
<td>0.0</td>
</tr>
<tr>
<td>insulin</td>
<td>14.0</td>
<td>76.25</td>
<td>125.00</td>
<td>155.5</td>
<td>190.00</td>
<td>846.0</td>
<td>0.0</td>
</tr>
<tr>
<td>bmi</td>
<td>18.20</td>
<td>27.50</td>
<td>32.30</td>
<td>32.46</td>
<td>36.60</td>
<td>67.10</td>
<td>0.0</td>
</tr>
<tr>
<td>diabetes</td>
<td>0.0780</td>
<td>0.2437</td>
<td>0.3725</td>
<td>0.4719</td>
<td>0.6262</td>
<td>2.4200</td>
<td>0.0</td>
</tr>
<tr>
<td>age</td>
<td>21.00</td>
<td>24.00</td>
<td>29.00</td>
<td>33.24</td>
<td>41.00</td>
<td>81.00</td>
<td>0.0</td>
</tr>
<tr>
<td>class</td>
<td>neg:500</td>
<td>pos:268</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Graphical Summaries

Example: Diabetes in Pima Indian Women

- Univariate

```r
1 # simple data plot
2 plot(sort(pima$bp))
3 # histogram
4 hist(pima$bp)
5 # density plot
6 plot(density(pima$bp, na.rm=TRUE))
```
Graphical Summaries

Example: Diabetes in Pima Indian Women

- **Bivariate**

```r
# scatterplot
plot(triceps~bmi, pima)

# boxplot
boxplot(diabetes~class, pima)
```
Linear regression with a single predictor

**Objective**

Describe the relationship between two variables, say $X$ and $Y$ as a straight line, that is, $Y$ is modeled as a linear function of $X$.

**The variables**

$X$: explanatory variable (horizontal axis)

$Y$: response variable (vertical axis)

After data collection, we have pairs of observations:

$$(x_1, y_1), \ldots, (x_n, y_n)$$
Linear regression with a single predictor

Example: Production Runs (Taken from Sheather, 2009)

Loading the Data:

```r
```

<table>
<thead>
<tr>
<th>Case</th>
<th>RunTime</th>
<th>RunSize</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>195</td>
<td>175</td>
</tr>
<tr>
<td>2</td>
<td>215</td>
<td>189</td>
</tr>
<tr>
<td>3</td>
<td>243</td>
<td>344</td>
</tr>
<tr>
<td>4</td>
<td>162</td>
<td>88</td>
</tr>
<tr>
<td>5</td>
<td>185</td>
<td>114</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>230</td>
<td>337</td>
</tr>
<tr>
<td>19</td>
<td>208</td>
<td>146</td>
</tr>
<tr>
<td>20</td>
<td>172</td>
<td>68</td>
</tr>
</tbody>
</table>

Variables:

- **RunTime (Y)**: time taken (in minutes) for a production run
- **RunSize (X)**: number of items produced in each run

We want to be able to describe the production run time as a linear function of the number of items in the run.
Linear regression with a single predictor I

Example: Production Runs

The scatter plot allows one to check if the linear relationship is supported by the data.

1 attach(production)
2 plot(RunTime ~ RunSize)
Simple linear regression model

The regression of variable $Y$ on variable $X$ is given by:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \ldots, n$$

where:

- Random Error: $\epsilon_i \sim N(0, \sigma^2)$, independent
- Linear Function: $\beta_0 + \beta_1 x_i = E(Y|X = x_i)$

**Unknown parameters**

- $\beta_0$ (Intercept): point in which the line intercepts the $y$-axis;
- $\beta_1$ (Slope): increase in $Y$ per unit change in $X$. 
Estimation of unknown parameters

We want to find the equation of the line that “best” fits the data. It means finding $b_0$ and $b_1$ such that the fitted values of $y_i$, given by

$$\hat{y}_i = b_0 + b_1 x_i,$$

are as “close” as possible to the observed values $y_i$.

**Residuals**

The difference between the observed value $y_i$ and the fitted value $\hat{y}_i$ is called residual and is given by:

$$e_i = y_i - \hat{y}_i$$
Estimation of unknown parameters II

Least Squares Method

A usual way of calculating $b_0$ and $b_1$ is based on the minimization of the sum of the squared residuals, or residual sum of squares (RSS):

$$RSS = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - b_0 - b_1 x_i)^2$$
Fitting a simple linear regression in R I

Example: Production Runs

The parameters $b_0$ and $b_1$ are estimated by using the function `lm()`:

```r
1 # Fit the regression model using the function `lm()`: 
production.lm <- lm(RunTime ~ RunSize, data = production)

2 # Use the function `summary()` to get some results:

3 summary(production.lm)
```

Denise Ferrari & Tiffany Head  denise@stat.ucla.edu tiffany@stat.ucla.edu

Regression in R I  UCLA SCC
Fitting a simple linear regression in R II

Example: Production Runs

The output looks like this:

```
Call:
  lm(formula = RunTime ~ RunSize, data = production)

Residuals:
     Min       1Q   Median       3Q      Max

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 149.74770   8.32815  17.980 6.00e-13 ***
RunSize      0.25924    0.03714   6.980 1.61e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 16.25 on 18 degrees of freedom
Multiple R-squared:  0.7302,   Adjusted R-squared:  0.7152
F-statistic: 48.72 on 1 and 18 DF,  p-value: 1.615e-06
```
Fitted values and residuals

- Fitted values obtained using the function `fitted()`
- Residuals obtained using the function `resid()`

```r
# Create a table with fitted values and residuals
data.frame(production, fitted.value=fitted(production.lm), residual=resid(production.lm))
```

<table>
<thead>
<tr>
<th>Case</th>
<th>RunTime</th>
<th>RunSize</th>
<th>fitted.value</th>
<th>residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>195</td>
<td>175</td>
<td>195.1152</td>
<td>-0.1152469</td>
</tr>
<tr>
<td>2</td>
<td>215</td>
<td>189</td>
<td>198.7447</td>
<td>16.2553496</td>
</tr>
<tr>
<td>3</td>
<td>243</td>
<td>344</td>
<td>238.9273</td>
<td>4.0726679</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>172</td>
<td>68</td>
<td>167.3762</td>
<td>4.6237657</td>
</tr>
</tbody>
</table>

\[
\hat{y}_1 = 149.75 + 0.26 \times 175 = 195.115 \\
e_1 = 195 - 195.115 = -0.115
\]
Fitted values and residuals I

When there are missing data

Missing data need to be handled carefully. Using the `na.exclude` method:

```r
# Load the package that contains the data
library(ISwR)
data(thuesen); attach(thuesen)
# Option for dealing with missing data
options(na.action=na.exclude)
# Now fit the regression model as before
velocity.lm <- lm(short.velocity~blood.glucose)
# Create a table with fitted values and residuals
data.frame(thuesen, fitted.value=fitted(velocity.lm), residual=resid(velocity.lm))
```
## Fitted values and residuals II

When there are missing data

<table>
<thead>
<tr>
<th>blood.glucose</th>
<th>short.velocity</th>
<th>fitted.value</th>
<th>residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15.3</td>
<td>1.76</td>
<td>1.433841 0.326158532</td>
</tr>
<tr>
<td>2</td>
<td>10.8</td>
<td>1.34</td>
<td>1.335010 0.004989882</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>16</td>
<td>8.6</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>8.8</td>
<td>1.12</td>
<td>1.291085 -0.171085074</td>
</tr>
<tr>
<td>24</td>
<td>9.5</td>
<td>1.70</td>
<td>1.306459 0.393541161</td>
</tr>
</tbody>
</table>
Analysis of Variance (ANOVA) I

The ANOVA breaks the total variability observed in the sample into two parts:

\[
\text{Total variability} \ (\text{TSS}) = \text{Variability explained by the model} \ (\text{SSreg}) + \text{Unexplained variability} \ (\text{RSS})
\]
Analysis of Variance (ANOVA) II

In R, we do:

```r
1 anova(production.lm)
```

Analysis of Variance Table

Response: RunTime

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RunSize</td>
<td>1</td>
<td>12868.4</td>
<td>12868.4</td>
<td>48.717</td>
<td>1.615e-06</td>
</tr>
<tr>
<td>Residuals</td>
<td>18</td>
<td>4754.6</td>
<td>264.1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Measuring Goodness of Fit I

Coefficient of Determination, $R^2$

- represents the proportion of the total sample variability explained by the regression model.
- for simple linear regression, the $R^2$ statistic corresponds to the square of the correlation between $Y$ and $X$.
- indicates of how well the model fits the data.

From the ANOVA table:

$$R^2 = \frac{12868.4}{12868.4 + 4754.6} = 0.7302$$

which we can also find in the regression summary.
Measuring Goodness of Fit II

**Adjusted \( R^2 \)**

The adjusted \( R^2 \) takes into account the number of degrees of freedom and is preferable to \( R^2 \).

From the ANOVA table:

\[
R_{adj}^2 = 1 - \frac{4754.6/18}{(12868.4 + 4754.6)/(18 + 1)} = 0.7152
\]

also found in the regression summary.

**Attention**

Neither \( R^2 \) nor \( R_{adj}^2 \) give direct indication on how well the model will perform in the prediction of a new observation.
Confidence and prediction bands I

Confidence Bands
Reflect the uncertainty about the regression line (how well the line is determined).

Prediction Bands
Include also the uncertainty about future observations.

Attention
These limits rely strongly on the assumption of normally distributed errors with constant variance and should not be used if this assumption is violated for the data being analyzed.
Confidence and prediction bands II

Predicted values are obtained using the function `predict()`.

1. `# Obtaining the confidence bands:
2. predict(production.lm, interval="confidence")`

<table>
<thead>
<tr>
<th>fit</th>
<th>lwr</th>
<th>upr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>195.1152</td>
<td>187.2000</td>
</tr>
<tr>
<td>2</td>
<td>198.7447</td>
<td>191.0450</td>
</tr>
<tr>
<td>3</td>
<td>238.9273</td>
<td>225.4549</td>
</tr>
</tbody>
</table>

... 

| 20   | 167.3762 | 154.4448  | 180.3077 |
Confidence and prediction bands III

```r
# Obtaining the prediction bands:
predict(production.lm, interval="prediction")
```

<table>
<thead>
<tr>
<th>fit</th>
<th>lwr</th>
<th>upr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>195.1152</td>
<td>160.0646</td>
</tr>
<tr>
<td>2</td>
<td>198.7447</td>
<td>163.7421</td>
</tr>
<tr>
<td>3</td>
<td>238.9273</td>
<td>202.2204</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>167.3762</td>
<td>130.8644</td>
</tr>
</tbody>
</table>
Confidence and prediction bands IV

For plotting:

1. `# Create a new data frame containing the values of X at which we want the predictions to be made`
   ```r
   pred.frame <- data.frame(RunSize= seq(55, 345, by=10))
   ```
2. `# Confidence bands`
   ```r
   pc <- predict(production.lm, int="c", newdata=pred.frame)
   ```
3. `# Prediction bands`
   ```r
   pp <- predict(production.lm, int="p", newdata=pred.frame)
   ```
Confidence and prediction bands V

```r
# Plot
require(graphics)

# Standard scatterplot with extended limits
plot(RunSize, RunTime, ylim=range(RunSize, pp, na.rm=T))
pred.Size <- pred.frame$RunSize

# Add curves
matlines(pred.Size, pc, lty=c(1,2,2), lwd=1.5, col=1)
matlines(pred.Size, pp, lty=c(1,3,3), lwd=1.5, col=1)
```
Confidence and prediction bands VI

![Graph showing confidence and prediction bands for RunSize vs RunTime]
The simple dummy variable regression is used when the predictor variable is not quantitative but categorical and assumes only two values.
Dummy Variable Regression I

Example: Change over time (Taken from Sheather, 2009)

Loading the Data:

```r
```

Variables:

<table>
<thead>
<tr>
<th>Method</th>
<th>Changeover</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Existing</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td>Existing</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>Existing</td>
<td>39</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>New</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>New</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>New</td>
<td>35</td>
<td>1</td>
</tr>
</tbody>
</table>

Change-over ($Y$): time (in minutes) required to change the line of food production

New ($X$): 1 for the new method, 0 for the existing method

We want to be able to test whether the change-over time is different for the two methods.
### Dummy Variable Regression II

#### Example: Change over time (Taken from Sheather, 2009)

```r
attach(changeover)
#
Summary:
summary(changeover)
```

<table>
<thead>
<tr>
<th>Method</th>
<th>Changeover</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Existing</td>
<td>Min. : 5.00</td>
<td>Min. : 0.0</td>
</tr>
<tr>
<td>New</td>
<td>1st Qu.: 11.00</td>
<td>1st Qu.: 0.0</td>
</tr>
<tr>
<td></td>
<td>Median : 15.00</td>
<td>Median : 0.0</td>
</tr>
<tr>
<td></td>
<td>Mean : 16.59</td>
<td>Mean : 0.4</td>
</tr>
<tr>
<td></td>
<td>3rd Qu.: 21.00</td>
<td>3rd Qu.: 1.0</td>
</tr>
<tr>
<td></td>
<td>Max. : 40.00</td>
<td>Max. : 1.0</td>
</tr>
</tbody>
</table>

We need to recode the X variable (New) to factor:
Dummy Variable Regression III

Example: Change over time (Taken from Sheather, 2009)

1. `changeover$New <- factor(changeover$New)`
2. `summary(changeover)`

<table>
<thead>
<tr>
<th>Method</th>
<th>Changeover</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Existing:72</td>
<td>Min. : 5.00 0:72</td>
<td></td>
</tr>
<tr>
<td>New :48</td>
<td>1st Qu.:11.00 1:48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Median :15.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean :16.59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3rd Qu.:21.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Max. :40.00</td>
<td></td>
</tr>
</tbody>
</table>
Dummy Variable Regression IV

Example: Change over time (Taken from Sheather, 2009)

Plotting the data:

```r
1 plot(Changeover ~ New)
```
Dummy Variable Regression V

Example: Change over time (Taken from Sheather, 2009)

Fitting the linear regression:

```r
1 # Fit the linear regression model
2 changeover.lm <- lm(Changeover~New, data=changeover)
3 # Extract the regression results
4 summary(changeover.lm)
```
Dummy Variable Regression VI

Example: Change over time (Taken from Sheather, 2009)

The output looks like this:

Call:
`lm(formula = Changeover ~ New, data = changeover)`

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>-10.861</td>
<td>-5.861</td>
<td>-1.861</td>
<td>4.312</td>
<td>25.312</td>
</tr>
</tbody>
</table>

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|---------|
| (Intercept)    | 17.8611  | 0.8905     | 20.058  | <2e-16  *** |
| New1           | -3.1736  | 1.4080     | -2.254  | 0.0260  * |

---

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

Residual standard error: 7.556 on 118 degrees of freedom
Multiple R-squared:  0.04128, Adjusted R-squared:  0.03315
F-statistic: 5.081 on 1 and 118 DF,  p-value: 0.02604
Dummy Variable Regression VII

Example: Change over time (Taken from Sheather, 2009)

Analysis of the results:

- There’s significant evidence of a reduction in the mean change-over time for the new method.
- The estimated mean change-over time for the new method ($X = 1$) is:

$$\hat{y}_1 = 17.8611 + (-3.1736) \times 1 = 14.7 \text{ minutes}$$

- The estimated mean change-over time for the existing method ($X = 0$) is:

$$\hat{y}_0 = 17.8611 + (-3.1736) \times 0 = 17.9 \text{ minutes}$$
Diagnostics

Assumptions

The assumptions for simple linear regression are:

- $Y$ relates to $X$ by a linear regression model:
  \[ Y = \beta_0 + \beta_1 X + \epsilon \]

- the errors are independent and identically normally distributed with mean zero and common variance.
Diagnostics

What can go wrong?

Violations:

- In the linear regression model:
  - linearity (e.g. quadratic relationship or higher order terms)
- In the residual assumptions:
  - non-normal distribution
  - non-constant variances
  - dependence
  - outliers

Checks:

⇒ look at plot of residuals vs. X
⇒ look at plot of residuals vs. fitted values
⇒ look at residuals Q-Q norm plot
Validity of the regression model I

Example: The Anscombe’s data sets (Taken from Sheather, 2009)

```
# Loading the data:

attach(anscombe)

# Looking at the data:
anscombe
```

<table>
<thead>
<tr>
<th>case</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
<th>y1</th>
<th>y2</th>
<th>y3</th>
<th>y4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>8.04</td>
<td>9.14</td>
<td>7.46</td>
<td>6.58</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>6.95</td>
<td>8.14</td>
<td>6.77</td>
<td>5.76</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>4.82</td>
<td>7.26</td>
<td>6.42</td>
<td>7.91</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>8</td>
<td>5.68</td>
<td>4.74</td>
<td>5.73</td>
<td>6.89</td>
</tr>
</tbody>
</table>
Validity of the regression model II

Example: The Anscombe’s data sets (Taken from Sheather, 2009)

```r
# Fitting the regressions
a1.lm <- lm(y1~x1, data=anscombe)
a2.lm <- lm(y2~x2, data=anscombe)
a3.lm <- lm(y3~x3, data=anscombe)
a4.lm <- lm(y4~x4, data=anscombe)

# Plotting
# For the first data set
plot(y1~x1, data=anscombe)
abline(a1.lm, col=2)
```
Validity of the regression model III

Example: The Anscombe’s data sets (Taken from Sheather, 2009)

For all data sets, the fitted regression is the same:

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

All models have \( R^2 = 0.67, \hat{\sigma} = 1.24 \) and the slope coefficients are significant at < 1% level. To check that, use the `summary()` function on the regression models.
Residual Plots I

Checking assumptions graphically

- Residuals vs. X

1. For the first data set
2. `plot(resid(a1.lm) ~ x1)`

### Data set 1
![Residuals vs. x1](image1.png)

### Data set 2
![Residuals vs. x2](image2.png)

### Data set 3
![Residuals vs. x3](image3.png)

### Data set 4
![Residuals vs. x4](image4.png)
Residual Plots II
Checking assumptions graphically

- Residuals vs. fitted values

1. `# For the first data set`
2. `plot(resid(a1.lm)~fitted(a1.lm))`

Data set 1
Residuals vs. Fitted values

Data set 2
Residuals vs. Fitted values

Data set 3
Residuals vs. Fitted values

Data set 4
Residuals vs. Fitted values
Leverage (or influential) points and outliers

Leverage points are those which have great influence on the fitted model, that is, those whose $x$-value is distant from the other $x$-values.

- **Bad leverage point**: if it is also an outlier, that is, the $y$-value does not follow the pattern set by the other data points.
- **Good leverage point**: if it is not an outlier.

Denise Ferrari & Tiffany Head  
denise@stat.ucla.edu  
tiffany@stat.ucla.edu
Leverage (or influential) points and outliers II

**Bad Leverage Point**

\[ \hat{y} = 0.07 - 0.08x \]

**Good Leverage Point**

\[ \hat{y} = -1.83 - 0.96x \]
Standardized residuals are obtained by dividing each residual by an estimate of its standard deviation:

\[ r_i = \frac{e_i}{\hat{\sigma}(e_i)} \]

To obtain the standardized residuals in R, use the command `rstandard()` on the regression model.

**Leverage Points**

- Good leverage points have their standardized residuals within the interval \([-2, 2]\)
- **Outliers** are leverage points whose standardized residuals fall outside the interval \([-2, 2]\)
Diagnostics

Leverage (or influential) points and outliers I

How to deal with them

- Remove invalid data points
  ⇒ if they look unusual or are different from the rest of the data

- Fit a different regression model
  ⇒ if the model is not valid for the data
  – higher-order terms
  – transformation
Leverage (or influential) points and outliers II
How to deal with them

Data set containing outliers:
Leverage (or influential) points and outliers III
How to deal with them

After their removal:
Normality and constant variance of errors

Normality and Constant Variance Assumptions

These assumptions are necessary for inference:

- hypothesis testing
- confidence intervals
- prediction intervals

⇒ Check the Normal Q-Q plot of the standardized residuals.
⇒ Check the Standardized Residuals vs. X plot.

Note
When these assumptions do not hold, we can try to correct the problem using data transformations.
**Diagnostics**

**Normality and constant variance checks**

**Example: Production Runs**

```
1 # Regression model
2 production.lm <- lm(RunTime~RunSize, data=production)
3 # Residual plots
4 plot(production.lm)
```
When to use transformations?

Transformations can be used to correct for:

- non-constant variance
- non-linearity
- non-normality

The most common transformations are:

- Square root
- Log
- Power transformation
Example of correction: non-constant variance

Example: Cleaning Data (Taken from Sheather, 2009)

Variables:

Rooms (Y): number of rooms cleaned
Crews (X): number of crews

We want to be able to model the relationship between the number of rooms cleaned and the number of crews.

1. # Load the data
3. attach(cleaning)
Example of correction: non-constant variance II

Example: Cleaning Data (Taken from Sheather, 2009)

```r
# Regression model
cleaning.lm <- lm(Rooms ~ Crews, data=cleaning)

# Plotting data and regression line
plot(Rooms~Crews)
abline(cleaning.lm, col=2)
```
Example of correction: non-constant variance III

Example: Cleaning Data (Taken from Sheather, 2009)

```r
1 # Diagnostic plots
2 plot(cleaning.lm)
```

1. Residuals vs Fitted
2. Normal Q-Q
3. Scale-Location
4. Residuals vs Leverage
Example of correction: non-constant variance IV

Example: Cleaning Data (Taken from Sheather, 2009)

1. # Applying square root transformation (counts)
2. sqrtRooms <- sqrt(Rooms)
3. sqrtCrews <- sqrt(Crews)
4. # Regression model on the transformed data
5. sqrt.lm <- lm(sqrtRooms ~ sqrtCrews)
6. # Plotting data and regression line
7. plot(sqrtRooms ~ sqrtCrews)
8. abline(sqrt.lm, col=2)
Example of correction: non-constant variance V

Example: Cleaning Data (Taken from Sheather, 2009)

```r
# Diagnostic plots
plot(sqrt.lm)
```

![Diagnostic plots](image)
Online Resources for R

Download R: http://cran.stat.ucla.edu

Search Engine for R: rseek.org

R Reference Card: http://cran.r-project.org/doc/contrib/Short-refcard.pdf

UCLA Statistics Information Portal: http://info.stat.ucla.edu/grad/

UCLA Statistical Consulting Center http://scc.stat.ucla.edu
References

- P. Daalgard
  Introductory Statistics with R,

- B.S. Everitt and T. Hothorn
  A Handbook of Statistical Analysis using R,

- J.J. Faraway
  Practical Regression and Anova using R,
  www.stat.lsa.umic.edu/~faraway/book
References II

J. Maindonald and J. Braun
Data Analysis and Graphics using R – An Example-Based Approach,

[Sheather, 2009] S.J. Sheather
A Modern Approach to Regression with R,
DOI: 10.1007/978-0-387-09608-7-3,

Denise Ferrari & Tiffany Head  denise@stat.ucla.edu tiffany@stat.ucla.edu
Regression in R I  UCLA SCC
Online Resources for R

- **Next Week:**
  - Nonlinear Regression (Feb 15, Monday)

- **Week After Next:**
  - Survival Analysis in R (February 22, Monday)
  - Spatial Statistics in R (February 24, Wednesday)

- For a schedule of all mini-courses offered please visit: http://scc.stat.ucla.edu/mini-courses
Thank you.

Any Questions?
Exercise in R I

Airfares Data (Taken from Sheather, 2009)

The data set for this exercise can be found at:

http://www.stat.tamu.edu/~sheather/book/docs/datasets/airfares.txt

It contains information on one-way airfare (in US$) and distance (in miles) from city A to 17 other cities in the US.
Exercise in R II

Airfares Data (Taken from Sheather, 2009)

1. Fit the regression model given by:

   \[
   \text{Fare} = \beta_0 + \beta_1 \text{Distance} + \epsilon
   \]

2. Critique the following statement:

   *The regression coefficient of the predictor variable (Distance) is highly statistically significant and the model explains 99.4% of the variability in the \(Y\)-variable (Fare).

   Thus this model is highly effective for both understanding the effects of Distance on Fare and for predicting future values of Fare given the value of the predictor variable.*

3. Does the regression model above seem to fit the data well? If not, describe how the model can be improved.