UCLA Department of Statistics
Statistical Consulting Center

Introduction to Regression in R

Masanao Yajima
yajima@stat.ucla.edu

November 16th, 2010
1 Introduction
   - Overview and goal of this class
   - What is Linear Regression Analysis?
   - When To Use Linear Regression Analysis?
   - What To Be Careful Using Regression?

2 Estimation

3 Example: Child’s IQ and Mother’s IQ

4 Linear Transformation

5 Post-Fit Inference

6 Assumptions and Validation

7 Appendix

8 Resources

Masanao Yajima yajima@stat.ucla.edu
My goal for you today

- I have put more information in these slides than I can cover.
- But there are few things I will definitely talk about that I really want you to take home with you.
  - Interpretation of regression model
  - Importance of Centering and scaling of variables
  - Efficiency of using display() and coefplot()
What is Linear Regression Analysis?

• Linear regression is a method that summarizes how the average values of numerical outcome variable vary over subpopulations defined by linear functions of predictors. (Gelman and Hill 2007)

• (Linear regression model) offers a concise summary of the mean of the response variable as a function of the explanatory variable through two parameters: the slope and the intercept of the line. (Ramsey and Schafer 2002)
When To Use Linear Regression Analysis?

- We may use it for
  - Description: to see on average how much groups differ on certain outcome variable.
  - Prediction: to predict, on average, expected change in the outcome with change in the predictors.

- Some examples are:
  - Predicting average height of a child from their parents’ height
  - Looking at the average amount of yield of crop for given rainfall and temperature
  - Describing average Income of a person, stratified by education, income of the parents, and other social economical informations
  - etc...
What To Be Careful When Using Regression?

- **Making causal claim**
  - Causal claim is a claim such as "By changing predictor 'A', we can change outcome B."
  - This is legitimate only when we have no other predictor(s) that also affect the outcome B.
    - which is possible under special settings (e.g. experiment), or
    - under very strong assumptions
  - hence nothing in this class will be about causality.

- **Extrapolating beyond given data**
  - Regression model is good as the support of the data, be careful
    - when predicting value in the future
    - making claim for combination of predictor not in data
    - looking at the model where data is scarce
Convention Used in (Linear) Regression Analysis

- Regression is called,
  - Simple (Linear) Regression: When number of explanatory variable is 1
  - Multivariate (Linear) Regression: When number of explanatory variable is more than 1

In this class we will call it just regression.

- Following letters are used to make the notation simple
  - \( Y \): response variable; and
  - \( X_1, X_2, \ldots, X_p \): predictor variables

But we will try not to use these notations as much as possible for the following reasons
  - keeps the context of the model which makes interpretation easy
  - we can code information in the name of the variable

Masanao Yajima yajima@stat.ucla.edu
1. Introduction

2. Estimation
   - Introduction to Estimation

3. Example: Child’s IQ and Mother’s IQ

4. Linear Transformation

5. Post-Fit Inference

6. Assumptions and Validation

7. Appendix

8. Resources

Masanao Yajima yajima@stat.ucla.edu

Regression in R I

UCLA SCC
Estimation of unknown parameters I

- Given a model,

\[ \text{outcome}[i] = \alpha + \beta_1 \text{predictor}_1[i] + \cdots + \beta_p \text{predictor}_p[i] + \text{error} \]

- we want to find \( \alpha \) and \( \beta(s) \) such that the fitted values of \( \text{outcome}[i] \), given by

\[ \hat{\text{outcome}}[i] = \hat{\alpha} + \hat{\beta}_1 \text{predictor}_1[i] + \cdots + \hat{\beta}_p \text{predictor}_p[i], (i = 1, \ldots, n) \]

are as “close” as possible to the observed values \( \text{outcome}[i] \).

- The \( \hat{\cdot} \) notation is there to indicate that the it is the fitted value.
Estimation of unknown parameters II

Residuals

The difference between the observed value $\text{outcome}[i]$ and the fitted value $\hat{\text{outcome}}[i]$ is called residual and is given by:

$\text{residual}[i] = \text{outcome}[i] - \hat{\text{outcome}}[i]$
Estimation of unknown parameters III

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):

\[
RSS = \sum_i \text{residual}[i]^2 \\
= \sum_i (\text{outcome}[i] - \hat{\text{outcome}}[i])^2 \\
= \sum_i (\text{outcome}[i] - \alpha - \beta_1 \text{predictor}_1[i] - \cdots - \beta_p \text{predictor}_p[i])^2
\]
Fitting a regression model in R I

To fit a linear regression model in R you use the `lm()` function

```r
<fit object> <- lm( <outcome> ~ <predictor 1> + ... + <predictor p> )
```

To look at the fitted linear regression model we will use `display()` function in arm package

```r
display( <fit object > )
```

Another option is to use the `summary()` function.

```r
summary( <fit object > )
```

However we will stick with `display()` since it is easier to interpret.
You can also directly obtain the fitted value, residual, and estimated coefficient(s) by

\[
\text{fitted( <fit object > )}
\]
\[
\text{resid( <fit object > )}
\]
\[
\text{coef( <fit object > )}
\]

Adding a regression line in a plot is simple also, you first plot the outcome vs predictor then call

\[
\text{abline( <fit object > )}
\]
1 Introduction

2 Estimation

3 Example: Child’s IQ and Mother’s IQ
   - One Binary Predictor
   - One Continuous Predictor
   - Continuous and Binary
   - Interaction

4 Linear Transformation

5 Post-Fit Inference

6 Assumptions and Validation

7 Appendix

8 Resources

Masanao Yajima yajima@stat.ucla.edu
We will work with a cognitive test scores of three- and four-year-old children and characteristics of their mothers from a survey of adult American women and their children (subsample of National Longitudinal Survey of Youth). ¹

```r
> kidiq <- read.dta("kidiq.dta")
> attach(kidiq)
> kidiq

           kid_score mom_hs mom_iq mom_work mom_age
     1 65 1 121.11753 4 27
     2 98 1  89.36188 4 25
     3 85 1 115.44316 4 27
     4 83 1  99.44964 3 25
     5 115 1  92.74571 4 27

...```

¹Data and code are from Data Analysis Using Regression and Multilevel/Hierarchical Models by Gelman and Hill 2007

Masanao Yajima yajima@stat.ucla.edu
Regression in R UCLA SCC
Let’s fit our first regression model.

We will start with a simple model, then gradually build on it.

As an illustrative purpose, we will start with a binary variable indicating whether mother graduated from high school or not \( \text{mom\_hs} \) as a predictor.

\[
\text{kid\_score} = \alpha + \beta_{\text{hs}} \text{mom\_hs} + \text{error}
\]
Linear Regression With One Binary Predictor II

- Here is how to fit the model in R.

  ```r
  > fit.0 <- lm ( kid_score ~ mom_hs )
  > display( fit.0 )
  
  lm(formula = kid_score ~ mom_hs)
  
  coef.est  coef.se
  (Intercept)  77.55     2.06
  mom_hs       11.77     2.32
  
  ---
  
  n = 434, k = 2
  residual sd = 19.85, R-Squared = 0.06
  
  - For now just look at the estimated coefficients \( \hat{\alpha} = 77.55 \) and \( \hat{\beta}_{hs} = 11.77 \)
Linear Regression With One Binary Predictor III

- Now that we have a model lets try to interpret it.

$$\hat{\text{kid\_score}} = 77.55 + 11.77 \text{mom\_hs}$$

- for a mother with high school education
  - No ($\text{mom\_hs} = 0$): expected IQ of a child is about 78.
  - Yes ($\text{mom\_hs} = 1$): expected IQ of a child is about 89.

- Hence you see that regression coefficient $\beta_{\text{hs}}$ for a binary predictor is just a difference in the mean of the 2 groups.

```r
> mean(kid_score[mom_hs==0])
[1] 77.54839

> mean(kid_score[mom_hs==1])
[1] 89.31965

> coef(fit.0)[1]+coef(fit.0)[2]*c(0,1)
1    2
77.54839 89.31965
```
Linear Regression With One Binary Predictor IV

```r
kidscore.jitter <- jitter(kid_score)

jitter.binary <- function(a, jitt=.05){
  ifelse (a==0, runif (length(a), 0, jitt),
          runif (length(a), 1-jitt, 1))
}

jitter.mom_hs <- jitter.binary(mom_hs)

plot(jitter.mom_hs, kidscore.jitter,
     xlab="Mother completed high school",
     ylab="Child test score", pch=20,
     xaxt="n", yaxt="n")

axis (1, seq(0,1))
axis (2, c(20,60,100,140))
abline (fit.0)
```
Linear Regression With One Continuous Predictor

Next we will model the child’s IQ from mother’s IQ score.

\[ \text{kid} \_\text{score} = \alpha + \beta_{\text{iq}} \times \text{mom} \_\text{iq} + \text{error} \]

We fit the model the same way we did for a binary variable

```r
> fit.1 <- lm (kid_score ~ mom_iq)
> display(fit.1)

lm(formula = kid_score ~ mom_iq)

<table>
<thead>
<tr>
<th>coef.est</th>
<th>coef.se</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>25.80</td>
</tr>
<tr>
<td>mom_iq</td>
<td>0.61</td>
</tr>
</tbody>
</table>

---

n = 434, k = 2
residual sd = 18.27, R-Squared = 0.20

Masanao Yajima yajima@stat.ucla.edu
Linear Regression With One Continuous Predictor II

- Have the interpretation changed for the model with a continuous variable?

\[ \hat{\text{kid\_score}} = 25.80 + 0.61 \text{mom\_hs} \]

- Yes they have!
  - Intercept: expected IQ of a child for mother with IQ of 0 (is this even possible?).
  - Coefficient \( \beta_{iq} \): expected increase in child’s IQ with every unit increase in mother’s IQ.
    (that seems really small, but how different are mothers with difference of just 1 IQ?)

- Although we were able to fit a model the model seems bit hard to interpret...
Linear Regression With One Continuous Predictor III

- Can we do better?
- Yes! we center and scale the variables
  - centering: shifting the variable to a meaningful center so it is easier to interpret the coefficient(s)
  - scaling: re-scaling the variable to a meaningful unit
- Centering becomes more important when we add the interaction term.
- For the current case with mother’s IQ
  - centering: we can center the mother’s IQ by subtracting the mean IQ of the mothers
  - scaling: we can divide the mother’s IQ by unit that might be meaningful full say 10

Masanao Yajima yajima@stat.ucla.edu

Regression in R I UCLA SCC
Linear Regression With One Continuous Predictor IV

- We will refit the model using the centered and scaled variable.

  ```r
  > mom_iq_mod = (mom_iq - mean(mom_iq))/10
  > fit.2 <- lm (kid_score ~ mom_iq_mod)
  > display(fit.2)
  
  lm(formula = kid_score ~ mom_iq_mod)
  coef.est coef.se
  (Intercept) 86.80 0.88
  mom_iq_mod 6.10 0.59
  
  ---
  n = 434, k = 2
  residual sd = 18.27, R-Squared = 0.20
  ```

- How has the interpretation changed?
  - Intercept: expected IQ of a child for mother with mean IQ
  - Coefficient $\beta_{iq}$: expected increase in child’s IQ with every 10 points increase in mother’s IQ.

- much better? scale and center matters!

Masanao Yajima yajima@stat.ucla.edu
We will go a step further and combine both a continuous and a binary predictors:

\[ \text{kid\_score} = \alpha + \beta_{hs}\text{mom\_hs} + \beta_{iq}\text{mom\_iq\_cs} + \text{error} \]

We fit the model in the same way:

```r
> fit.3 <- lm (kid_score ~ mom_hs + mom_iq_mod)
> display(fit.3)

lm(formula = kid_score ~ mom_hs + mom_iq_mod)

coef.est coef.se
(Intercept) 82.12 1.94
mom_hs 5.95 2.21
mom_iq_mod 5.64 0.61
---
n = 434, k = 3
residual sd = 18.14, R-Squared = 0.21
```
Regression With a Continuous and a Binary Predictors

- Here is the fitted model

\[
\hat{\text{kid\_score}} = 82.12 + 5.95 \text{mom\_hs} + 5.64 \text{mom\_iq}\_cs
\]

- How has the interpretation changed?
  - Intercept \( \alpha \): expected IQ of a child for mother with mean IQ that did NOT graduate high school
  - Coefficient \( \beta_{hs} \): expected increase in child’s IQ for a mother graduating high school keeping the other variables fixed.
  - Coefficient \( \beta_{iq} \): expected increase in child’s IQ with every 10 points increase in mother’s IQ keeping the other variables fixed.

- You may have noticed that although we fit one regression model, because of the binary predictor we are actually fitting 2 regression models.
Regression With a Continuous and a Binary Predictors

Here is the plot of the regression line of child’s IQ score for the mothers who graduated from high school (blue) and who did not graduate high school (red).

```r
fit.3.5 <- lm(kid_score ~ mom_hs + mom_iq)
plot(mom_iq, kid_score, xlab="Mother IQ score", ylab="Child test score", pch=20, xaxt="n", yaxt="n", type="n")
curve(coef(fit.3.5)[1] + coef(fit.3.5)[2] + coef(fit.3.5)[3]*x, add=TRUE, col="blue")
curve(coef(fit.3.5)[1] + coef(fit.3.5)[3]*x, add=TRUE, col="red")
points(mom_iq[mom_hs==1], kid_score[mom_hs==1], col=rgb(0,0,1,alpha=0.7))
points(mom_iq[mom_hs==0], kid_score[mom_hs==0], col=rgb(1,0,0,alpha=0.7))
axis(1, c(80,100,120,140))
axis(2, c(20,60,100,140))
```
We will go one last step and add an interaction term

\[
\text{kid\_score} = \alpha + \beta_{hs}\text{mom\_hs} + \beta_{iq}\text{mom\_iq}_{cs} + \beta_{hsiq}\text{mom\_hs} : \text{mom\_iq}_{cs} + \text{error}
\]

Interaction term can be coded using ":"

\[
> \text{fit.4} \leftarrow \text{lm} (\text{kid\_score} \sim \text{mom\_hs} + \text{mom\_iq\_mod} + \text{mom\_hs:mom\_iq\_mod})
\]

\[
> \text{display(fit.4)}
\]

\[
\text{lm(formula = kid\_score} \sim \text{mom\_hs} + \text{mom\_iq\_mod} + \text{mom\_hs:mom\_iq\_mod}) \\
\quad \text{coef.est coef.se}
\]

\[
(\text{Intercept}) & 85.41 & 2.22 \\
\text{mom\_hs} & 2.84 & 2.43 \\
\text{mom\_iq\_mod} & 9.69 & 1.48 \\
\text{mom\_hs:mom\_iq\_mod} & -4.84 & 1.62 \\
\]

---

n = 434, k = 4
residual sd = 17.97, R-Squared = 0.23
a Continuous and a Binary Predictors + Interaction

- Here is the fitted model (It’s not as bad as it looks)

\[
kid\_score = 85.41 + 2.84 \text{mom\_hs} + 9.69 \text{mom\_iq}_{cs} - 4.84 \text{mom\_hs} : \text{mom\_iq}_{cs}
\]

- Intercept \( \alpha \): expected IQ of a child for mother with mean IQ that did NOT graduate high school
- Coefficient \( \beta_{hs} \): expected increase in child’s IQ for a mother graduating high school keeping the other variables fixed.
- Coefficient \( \beta_{iq} \): expected increase in child’s IQ with every 10 points increase in mother’s IQ keeping the other variables fixed.
- Coefficient \( \beta_{hsiq} \): difference of the \( \beta_{iq} \) for mothers who graduated high school and did not.

- Let’s look closely at what that means
a Continuous and a Binary Predictors + Interaction

- Since \(\text{mom}_{hs}\) only takes value 0 or 1
  - For mother who did not graduate high school
    \[
    \hat{\text{kid}}_{hs=0} = 85.41 + 9.69\text{mom}_{iqcs}
    \]
  - For mother who did graduate high school
    \[
    \hat{\text{kid}}_{hs=1} = 85.41 + 2.84 + 9.69\text{mom}_{iqcs} - 4.84\text{mom}_{iqcs}
    = 88.25 + 4.85\text{mom}_{iqcs}
    \]

- Looking at these model separately, we see that mothers who did graduate high school does have, on average, child with higher IQ than mothers who did not graduate from high school, evaluated at the mean IQ for the mothers.
- However, effect of mother’s IQ is larger on expected child IQ for mothers who did not graduate from high school.
Interaction

a Continuous and a Binary Predictors + Interaction

- Again, here is the plot of the regression line of child’s IQ score for the mothers who graduated from high school (blue) and who did not graduate high school (red).

```r
fit.4.5 <- lm (kid_score ~ mom_hs + mom_iq + mom_hs:mom_iq)
plot(mom_iq, kid_score, xlab="Mother IQ score", ylab="Child test score", pch=20, xaxt="n", yaxt="n", type="n")
curve (coef(fit.4.5)[1] + coef(fit.4.5)[2] + (coef(fit.4.5)[3] + coef(fit.4.5)[4])*x, add=TRUE, col="blue")
curve (coef(fit.4.5)[1] + coef(fit.4.5)[3]*x, add=TRUE,col="red")
points (mom_iq[mom_hs==1], kid_score[mom_hs==1], col=rgb(0,0,1,alpha=0.7))
points (mom_iq[mom_hs==0], kid_score[mom_hs==0], col=rgb(1,0,0,alpha=0.7))
axis (1, c(80,100,120,140))
axis (2, c(20,60,100,140))
```
What have we learned so far?

- So far we have fit models with
  - Binary predictor: Different intercepts with no slope
  - Continuous predictor: One slope
  - Binary + Continuous: Different intercepts with same slope
  - Binary, Continuous, and Interaction: different slope + intercepts

- By adding a predictor we are able to model more feature of the data we are trying to understand.
1 Introduction

2 Estimation

3 Example: Child’s IQ and Mother’s IQ

4 Linear Transformation
   - Log transformation
   - Other transformation
   - Categorical variable

5 Post-Fit Inference

6 Assumptions and Validation

7 Appendix

8 Resources

Masanao Yajima yajima@stat.ucla.edu
Regression in R I UCLA SCC
Transformations when and how to use them I

- Linear transformation does NOT affect the fit of the regression model.
- However, if used correctly, it improves the interpretability of the coefficients.
- We saw an example earlier of centering and scaling of the predictor variable.
- We will also talk about how to transform the outcome variables.
Centering and scaling I

- As we saw earlier that centering and scaling of the predictor variable increases the interpretability of the estimated coefficients greatly. There are few ways to center and scale.

- Centering decides where the baseline will be for your model
  - Subtract the mean
  - Subtract meaningful value

- Scaling changes the unit of the predictor variable
  - Divide by meaningful unit
  - Divide by standard deviation (combined with mean centering gives you z-score)
  - Divide by 2 standard deviation (coherent estimate when you have binary variable)
Log transformation

Logarithmic transformation I

- Log transformation is the most popular transformation.
- When we take log of the outcome variable

\[
\log(outcome[i]) = \alpha + \beta_1 predictor_1[i] + \cdots + \beta_p predictor_p[i]
\]

- It is same as taking the exponent of the right hand side

\[
outcome[i] = e^{\alpha + \beta_1 predictor_1[i] + \cdots + \beta_p predictor_p[i]}
\]

\[
= e^\alpha e^{\beta_1 predictor_1[i]} \cdots e^{\beta_p predictor_p[i]}
\]

- so now, \( e^{\beta_j} \) becomes multiplied with every unit increase in \( predictor_j \).
Let's look at an example of predicting income from height of an individual.

This is the model we have:

\[
\log(\text{earning}[i]) = \alpha + \beta_{\text{height}} \text{height}[i]
\]

we fit the model

```r
> log.earn <- log(earn)
> earn.logmodel.1 <- lm(log.earn ~ height)
> display(earn.logmodel.1)
```

```
lm(formula = log.earn ~ height)
  coef.est coef.se
(Intercept)  5.78     0.45
 height       0.06     0.01
---
 n = 1192, k = 2
 residual sd = 0.89, R-Squared = 0.06
```
Log transformation

Logarithmic transformation III

- Thus our fitted model looks like

\[
\begin{align*}
\log(\hat{\text{earning}}[i]) & = 5.78 + 0.06 \text{height}[i] \\
\text{earning}[i] & = e^{5.78} (e^{0.06})^{\text{height}[i]} \\
& = 323.76 \times 1.062^{\text{height}[i]}
\end{align*}
\]

- Interpretation for this model is
  - For person with height of 0, he/she is expected to make $324 on average.
  - With increase of every inch in height, earning goes up by about 6%.
  - notice that coefficient estimate 0.06 and increase of 6% is very close, this is what you get for using natural log.
Categorical variable to indicator variable. I

- square root transformation: used when logarithm transformation is too extreme

- problem with square root is it’s not easy to interpret the result and it doesn’t work well with negative numbers.
It is often the case that you have categorical variable that does not have clear ordering.

In such a case we can transform categorical variable into multiple binary variable.

For example if we have

- $mom\_work = 1$: mother did not work in first three years of child’s life
- $mom\_work = 2$: mother worked in second or third year of child’s life
- $mom\_work = 3$: mother worked part-time in first year of child’s life
- $mom\_work = 4$: mother worked full-time in first year of child’s life.

Masanao Yajima yajima@stat.ucla.edu
This is what the data looks like originally:

```r
> mom_work
[1] 4 4 4 3 4 1 4 3 1 1 1 4 4 4 2 1 3 3 4 3 4 ........
```

We will use the `as.factor()` function to tell R that this a categorical variable

```r
> as.factor(mom_work)
[1] 4 4 4 3 4 1 4 3 1 1 1 4 4 4 2 1 3 3 4 3 4 ........
Levels: 1 2 3 4
```

Although they look the same, internally we have converted the variable into indicator variable.

```
> mom_work
[1,] 1 0 0 0
[2,] 0 0 0 1
[3,] 0 1 0 0
[4,] 0 0 1 0
```
Categorical variable to indicator variable. III

- So we are going to fit a model as

\[
\text{kid\_score} = \alpha + \beta_{w2}\text{mom\_work2} + \beta_{w3}\text{mom\_work3} + \beta_{w4}\text{mom\_work4} + \text{error}
\]

- In R we do

```r
> fit.5 <- lm (kid_score ~ as.factor(mom_work))
> display(fit.5)

lm(formula = kid_score ~ as.factor(mom_work))

coef.est coef.se
(Intercept) 82.00 2.31
as.factor(mom_work)2 3.85 3.09
as.factor(mom_work)3 11.50 3.55
as.factor(mom_work)4 5.21 2.70
---

n = 434, k = 4
residual sd = 20.23, R-Squared = 0.02
```
Categorical variable to indicator variable. IV

- This is the fitted model

\[
\hat{\text{kid\_score}} = 82.00 + 3.85 \text{mom\_work2} + 11.50 \text{mom\_work3} \\
+ 5.21 \text{mom\_work4}
\]

- You have to be careful about the interpretation.
- Notice R did not estimate the coefficient for \( \text{mom\_work} = 1 \)
- By default R takes a category and makes it a default.
- So intercept in this estimation is the coefficient for \( \text{mom\_work} = 1 \).
- All other coefficients are relative difference in the mean from kids with \( \text{mom\_work} = 1 \) for each group.

Masanao Yajima yajima@stat.ucla.edu
Regression in R I UCLA SCC
Categorical variable to indicator variable... V

- For a child with \( \text{mom}_\text{work} = 1 \): mother did not work in first three years of child’s life

\[
\hat{\text{kid}}_\text{score} = 82.00 + 3.85(0) + 11.50(0) + 5.21(0) = 82.00
\]

- For a child with \( \text{mom}_\text{work} = 2 \): mother worked in second or third year of child’s life

\[
\hat{\text{kid}}_\text{score} = 82.00 + 3.85(1) + 11.50(0) + 5.21(0) = 85.85
\]

- For a child with \( \text{mom}_\text{work} = 3 \): mother worked part-time in first year of child’s life

\[
\hat{\text{kid}}_\text{score} = 82.00 + 3.85(0) + 11.50(1) + 5.21(0) = 93.50
\]

- For a child with \( \text{mom}_\text{work} = 4 \): mother worked full-time in first year of child’s life.

\[
\hat{\text{kid}}_\text{score} = 82.00 + 3.85(0) + 11.50(0) + 5.21(1) = 87.21
\]
Categorical variable to indicator variable. VI

```r
kidscore.jitter <- jitter(kid_score)
momwork.jitter <- jitter(mom_work)
plot(momwork.jitter, kidscore.jitter,
xlab="Mother work types", ylab="Child test score",
pch=20, xaxt="n", yaxt="n")
axis(1, seq(1,4))
axis(2, c(20,60,100,140))
alpha = coef(fit.5)[1]
beta= c(0, coef(fit.5)[2:4])
for(i in 1:4){
  lines(c(i-0.3,i+0.3),
       c(alpha+beta[i],alpha+beta[i]),col="red")
}
```

Masanao Yajima yajima@stat.ucla.edu

Regression in R I UCLA SCC
Introduction

Estimation

Example: Child’s IQ and Mother’s IQ

Linear Transformation

Post-Fit Inference
- Statistical Inference
- Prediction
- Goodness of Fit

Assumptions and Validation

Appendix

Resources

Masanao Yajima yajima@stat.ucla.edu

Regression in R I UCLA SCC
If you have ever taken other regression class you may be wondering, where is the “*”?

It is much too common in many discipline of science to be concerned only with the statistical “significant” predictors.

If the goal of the model fitting is in prediction this might be important, but for interpretation this is not critical.

I will not go into criticism of the statistical significance. You can find that on the Wikipedia.

- Statistical significance is different from practical significance.
- Statistical significance is a artifact of sample size.
Statistical Inference

- If you still insist on the classical way, `summary()`, `oneway.test()`, and `aov()` might be the function you want.
- But, better way to look at the coefficient is using the `coefplot()` of `arm` library.
- Each dot is a point estimate and bar around it is 95 percent confidence interval.

```
coefplot(savings.lm)
```
Statistical Inference

- Why is it better?
  - You can immediately see which coefficients are significantly different from 0.
  - When 95 percent confidence interval crosses 0 it is not statistically significant at 5 percent significance level.
  - You can see the relative size of each of the coefficients.
  - You can see relative size of the uncertainty for each of the coefficient estimate.
  - It gives you more information than table in much shorter time.
  - Caution: when you are making comparison of more than one pair of parameters, you do need to consider the multiple comparison issue.
  - Under the Bayesian framework, these estimates are the posterior intervals and so you are allowed to make comparisons directly.

Masanao Yajima yajima@stat.ucla.edu
Regression in R I UCLA SCC
Confidence and prediction bands I

Example: Savings Data (Taken from Faraway, 2002)

```r
library(faraway)
data(savings)
attach(savings)
head(savings)
```

<table>
<thead>
<tr>
<th>sr</th>
<th>pop15</th>
<th>pop75</th>
<th>dpi</th>
<th>ddpi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>11.43</td>
<td>29.35</td>
<td>2.87</td>
<td>2329.68</td>
</tr>
<tr>
<td>Austria</td>
<td>12.07</td>
<td>23.32</td>
<td>4.41</td>
<td>1507.99</td>
</tr>
<tr>
<td>Belgium</td>
<td>13.17</td>
<td>23.80</td>
<td>4.43</td>
<td>2108.47</td>
</tr>
<tr>
<td>Bolivia</td>
<td>5.75</td>
<td>41.89</td>
<td>1.67</td>
<td>189.13</td>
</tr>
<tr>
<td>Brazil</td>
<td>12.88</td>
<td>42.19</td>
<td>0.83</td>
<td>728.47</td>
</tr>
<tr>
<td>Canada</td>
<td>8.79</td>
<td>31.72</td>
<td>2.85</td>
<td>2982.88</td>
</tr>
</tbody>
</table>

Masanao Yajima yajima@stat.ucla.edu

Regression in R I UCLA SCC
Confidence and prediction bands II

Example: Savings Data (Taken from Faraway, 2002)

```r
# Fitting the model with all predictors
savings.lm <- lm(sr~pop15+pop75+dpi+ddpi, data=savings)
display(savings.lm)

lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)

coef.est  coef.se
(Intercept)  28.57     7.35
pop15       -0.46     0.14
pop75       -1.69     1.08
dpi          0.00     0.00
ddpi         0.41     0.20

---
n = 50, k = 5
residual sd = 3.80, R-Squared = 0.34
```
**Confidence and prediction bands III**

**Example: Savings Data (Taken from Faraway, 2002)**

Confidence Bands

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

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

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

<table>
<thead>
<tr>
<th></th>
<th>fit</th>
<th>lwr</th>
<th>upr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>10.566420</td>
<td>8.573419</td>
<td>12.559422</td>
</tr>
<tr>
<td>Austria</td>
<td>11.453614</td>
<td>8.796229</td>
<td>14.110999</td>
</tr>
<tr>
<td>Belgium</td>
<td>10.951042</td>
<td>8.685716</td>
<td>13.216369</td>
</tr>
<tr>
<td>Malaysia</td>
<td>7.680869</td>
<td>5.724711</td>
<td>9.637027</td>
</tr>
</tbody>
</table>
```

Masanao Yajima yajima@stat.ucla.edu

Regression in R I UCLA SCC
Confidence and prediction bands IV

Example: Savings Data (Taken from Faraway, 2002)

Prediction Bands

Include also the uncertainty about future observations.

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

<table>
<thead>
<tr>
<th>Country</th>
<th>fit</th>
<th>lwr</th>
<th>upr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>10.56642</td>
<td>2.65239</td>
<td>18.48045</td>
</tr>
<tr>
<td>Austria</td>
<td>11.45361</td>
<td>3.34673</td>
<td>19.56049</td>
</tr>
<tr>
<td>Belgium</td>
<td>10.95104</td>
<td>2.96408</td>
<td>18.93800</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malaysia</td>
<td>7.68087</td>
<td>-0.22396</td>
<td>15.58570</td>
</tr>
</tbody>
</table>

Masanao Yajima yajima@stat.ucla.edu

Regression in R I UCLA SCC
Confidence and prediction bands V
Example: Savings Data (Taken from Faraway, 2002)

Attention
- these limits rely strongly on the assumption of independence and normally distributed errors with constant variance and should not be used if these assumptions are violated for the data being analyzed.
- the confidence and prediction bands only apply to the population from which the data were sampled.
Another way to which we will not go into the detail is to graphically display the uncertainty in the model is by simulation.

```r
fit.2 <- lm (kid_score ~ mom_iq)
display(fit.2)

fit.2.sim <- sim(fit.2)
plot (mom_iq, kid_score, xlab="Mother IQ score", ylab="Child test score", pch=20)
for (i in 1:10){
   curve (fit.2.sim$coef[i,1] + fit.2.sim$coef[i,2]*x, add=TRUE, col="gray")
}
curve (coef(fit.2)[1] + coef(fit.2)[2]*x, add=TRUE, col="red")
```
Measuring Goodness of Fit I

Coefficient of Determination, $R^2$

- $R^2$ represents the proportion of the total sample variability (sum of squares) explained by the regression model.
- Indicates of how well the model fits the data.

Adjusted $R^2$

- $R^2_{adj}$ represents the proportion of the mean sum of squares (variance) explained by the regression model.
- It takes into account the number of degrees of freedom and is preferable to $R^2$. 

Masanao Yajima yajima@stat.ucla.edu

Regression in R I UCLA SCC
Both $R^2$ and $R^2_{adj}$ are given in the regression summary.

Neither $R^2$ nor $R^2_{adj}$ give direct indication on how well the model will perform in the prediction of a new observation.

$R^2$ increases as the number of predictor increases

$R^2_{adj}$ is affected by the sample size, small value may just be indication of small sample.

The use of these statistics is more legitimate in the case of comparing different models for the same data set.
<table>
<thead>
<tr>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
</tbody>
</table>

**Assumptions and Validation**
- Assumptions
- Validation

Masanao Yajima  
yajima@stat.ucla.edu

Regression in R I  
UCLA SCC
Assumptions I

- Here is a list of assumptions that you may want to remember when using regression models.

- Validity
  - Problem
    - Inadequate model, lacking important outcome or predictor
    - overgeneralization due to extrapolation from the sample
  - Make sure that
    - Outcome measure reflects the phenomenon of interest
    - Model has all the relevant predictors
    - Model generalizes to cases to which it will be applied

- Additivity and Linearity
  - Problem
    - The relation between the variables may not be linear nor additive
Assumptions II

- There is some observations in the data that makes linearity infeasible
- Possible solution
  - Transformation
  - Adding interaction, categorical predictor, and exponentiated terms to make the relation linear
  - Data cleaning, although you need to be very careful when you choose to do this.

- Independence of the Errors
  - Problem
    - Does not cause bias in the estimates but standard errors may be affected.
    - Hence your prediction and testing becomes unreliable
  - There is a way out
    - but requires sophisticated modeling beyond the scope of this class.
Assumptions III

- **Constant Variance**
  - **Problem**
    - Estimate is unbiased, but prediction and standard error may not be accurate
  - **Possible solution may be**
    - Transformation of the outcome variable
    - using Weighted Least Squares

- **Normality of the Errors**
  - **Problem**
    - Not too critical for estimation
    - May be a problem for long tails and small to moderate sample size
  - **Possible solution**
    - Graphical Diagnostics: Histogram, QQplot
    - Normality tests
Residual plot

Good way to diagnose a regression model is by plotting the residuals.

```r
## Fit the model
fit.2 <- lm (kid_score ~ mom_iq)
resid <- fit.2$residuals
sd.resid <- sd(resid)

plot (mom_iq, resid, xlab="Mother IQ score", ylab="Residuals", pch=20)
abline (sd.resid,0,lty=2)
abline(0,0)
abline (-sd.resid,0,lty=2)
```
Validity of the regression model I
Example: The Anscombe's data sets (Taken from Sheather, 2009)

```
1 # Loading the data:
2 anscombe <- read.table("http://www.stat.tamu.edu/~sheather/book/docs/datasets/anscombe.txt", h=T, sep=" ")
3 attach(anscombe)
4 # Looking at the data:
5 anscombe

case x1 x2 x3 x4 y1 y2 y3 y4
1 10 10 10 8 8.04 9.14 7.46 6.58
2 8 8 8 8 6.95 8.14 6.77 5.76
...  
10 7 7 7 8 4.82 7.26 6.42 7.91
11 5 5 5 8 5.68 4.74 5.73 6.89
```

Masanao Yajima yajima@stat.ucla.edu
Regression in R I
UCLA SCC
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)`

![Residual Plots I](image)
Residual Plots II
Checking assumptions graphically

- Residuals vs. fitted values

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

Data set 1
Residuals
```
4 6 8 10 12 14
```
Fitted values
```
4 6 8 10 12 14
```
Data set 2
Residuals
```
4 6 8 10 12 14
```
Fitted values
```
4 6 8 10 12 14
```
Data set 3
Residuals
```
4 6 8 10 12 14
```
Fitted values
```
4 6 8 10 12 14
```
Data set 4
Residuals
```
4 6 8 10 12 14
```
Fitted values
```
4 6 8 10 12 14
```
Non-constant variance I
Example: Galapagos Data (Taken from Faraway, 2002)

```r
# Loading the data
library(faraway)
data(gala)
attach(gala)

# Fitting the model
gala.lm <- lm(Species~Area+Elevation+Scruz+Nearest+Adjacent, data=gala)

# Residuals vs. fitted values
plot(resid(gala.lm)~fitted(gala.lm), xlab="Fitted values", ylab="Residuals", main="Original Data")
```

Masanao Yajima yajima@stat.ucla.edu
Non-constant variance II
Example: Galapagos Data (Taken from Faraway, 2002)
Non-constant variance III
Example: Galapagos Data (Taken from Faraway, 2002)

```r
# Applying square root transformation:
sqrtSpecies <- sqrt(Species)
gala.sqrt <- lm(sqrtSpecies ~ Area + Elevation + Scruz + Nearest + Adjacent)

# Residuals vs. fitted values
plot(resid(gala.sqrt) ~ fitted(gala.sqrt), xlab="Fitted values", ylab="Residuals", main="Transformed Data")
```
Non-constant variance IV

Example: Galapagos Data (Taken from Faraway, 2002)
1 Introduction

2 Estimation

3 Example: Child’s IQ and Mother’s IQ

4 Linear Transformation

5 Post-Fit Inference

6 Assumptions and Validation

7 Appendix
   - Common Errors To Check
     - Coding
     - Graphical Summaries

8 Resources

Masanao Yajima yajima@stat.ucla.edu

Regression in R I UCLA SCC
Common Errors To Check I

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
Common Errors To Check II

We can resort to:

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

- **Graphical summaries:**
  - boxplots
  - histograms
  - scatterplots
  - etc.

Masanao Yajima yajima@stat.ucla.edu
Loading the Data

Example: Diabetes in Pima Indian Women

- 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")
  ```

\(^2\)Data from the UCI Machine Learning Repository


Masanao Yajima yajima@stat.ucla.edu
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:

```
head(pima)
```

<table>
<thead>
<tr>
<th>npreg</th>
<th>glucose</th>
<th>bp</th>
<th>triceps</th>
<th>insulin</th>
<th>bmi</th>
<th>diabetes</th>
<th>age</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>148</td>
<td>72</td>
<td>35</td>
<td>0</td>
<td>33.6</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>85</td>
<td>66</td>
<td>29</td>
<td>0</td>
<td>26.6</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>183</td>
<td>64</td>
<td>0</td>
<td>0</td>
<td>23.3</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>89</td>
<td>66</td>
<td>23</td>
<td>94</td>
<td>28.1</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>137</td>
<td>40</td>
<td>35</td>
<td>168</td>
<td>43.1</td>
<td>33</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>116</td>
<td>74</td>
<td>0</td>
<td>0</td>
<td>25.6</td>
<td>30</td>
<td>0</td>
</tr>
</tbody>
</table>
Common Errors To Check

Numerical Summaries
Example: Diabetes in Pima Indian Women

- 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>99.0</td>
<td>62.0</td>
<td>0.00</td>
<td>0.0</td>
</tr>
<tr>
<td>Median</td>
<td>3.000</td>
<td>117.0</td>
<td>72.0</td>
<td>23.00</td>
<td>30.5</td>
</tr>
<tr>
<td>Mean</td>
<td>3.845</td>
<td>120.9</td>
<td>69.1</td>
<td>20.54</td>
<td>79.8</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>6.000</td>
<td>140.2</td>
<td>80.0</td>
<td>32.00</td>
<td>127.2</td>
</tr>
<tr>
<td>Max.</td>
<td>17.000</td>
<td>199.0</td>
<td>122.0</td>
<td>99.00</td>
<td>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>0.00</td>
<td>0.0780</td>
<td>0.21</td>
<td>0.0000</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>27.30</td>
<td>0.2437</td>
<td>24.0</td>
<td>0.0000</td>
</tr>
<tr>
<td>Median</td>
<td>32.00</td>
<td>0.3725</td>
<td>29.0</td>
<td>0.0000</td>
</tr>
<tr>
<td>Mean</td>
<td>31.99</td>
<td>0.4719</td>
<td>33.24</td>
<td>0.3490</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>36.60</td>
<td>0.6262</td>
<td>41.00</td>
<td>1.0000</td>
</tr>
<tr>
<td>Max.</td>
<td>67.10</td>
<td>2.4200</td>
<td>81.00</td>
<td>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:

```r
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:

```
1  pima$class <- factor (pima$class)  
2  levels(pima$class) <- c("neg", "pos") 
```
## Final Coding

### Example: Diabetes in Pima Indian Women

```r
summary(pima)
```

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>npreg</td>
<td>0.000</td>
<td>1.000</td>
<td>3.000</td>
<td>3.845</td>
<td>6.000</td>
<td>17.00</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>
</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>
</tr>
<tr>
<td>triceps</td>
<td>7.00</td>
<td>22.0</td>
<td>29.0</td>
<td>29.15</td>
<td>36.0</td>
<td>99.00</td>
</tr>
<tr>
<td>insulin</td>
<td>14.00</td>
<td>76.25</td>
<td>125.0</td>
<td>155.55</td>
<td>190.0</td>
<td>846.00</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>
</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>
</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>
</tr>
<tr>
<td>class</td>
<td>neg:500</td>
<td>pos:268</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NA's : 5.0 NA's : 35.0 NA's : 227.00 NA's : 374.00

Masanao Yajima yajima@stat.ucla.edu

Regression in R I UCLA SCC
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
1 # scatterplot
2 plot(triceps ~ bmi, pima)
3 # boxplot
4 boxplot(diabetes ~ class, pima)
```
1 Introduction

2 Estimation

3 Example: Child’s IQ and Mother’s IQ

4 Linear Transformation

5 Post-Fit Inference

6 Assumptions and Validation

7 Appendix

8 Resources
   - Online Resources for R
   - References

Masanao Yajima yajima@stat.ucla.edu

Regression in R I UCLA SCC
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

Masanao Yajima  yajima@stat.ucla.edu
References I

- A. Gelman and J. Hill
  Data Analysis Using Regression and Multilevel/Hierarchical Models
  Cambridge University Press; 2006

- F. Ramsey and D. Schafer
  The Statistical Sleuth: A Course in Methods of Data Analysis
  Duxbury Press; 2002

- P. Daalgard
  Introductory Statistics with R,
References II

B.S. Everitt and T. Hothorn

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

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

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