Tutorial 12: Various types of model checks

Johannes Karreth

RPOS 517, Day 12

This tutorial shows you:

- how to specify quadratic terms in regression models
- how to use residuals to interpret model quality

Note on copying & pasting code from the PDF version of this tutorial: Please note that you may run into trouble if you copy & paste code from the PDF version of this tutorial into your R script. When the PDF is created, some characters (for instance, quotation marks or indentations) are converted into non-text characters that R won't recognize. To use code from this tutorial, please type it yourself into your R script or you may copy & paste code from the *source file* for this tutorial which is posted on my website.

Note on R functions discussed in this tutorial: I don't discuss many functions in detail here and therefore I encourage you to look up the help files for these functions or search the web for them before you use them. This will help you understand the functions better. Each of these functions is well-documented either in its help file (which you can access in R by typing `?ifelse`, for instance) or on the web. The *Companion to Applied Regression* (see our syllabus) also provides many detailed explanations.

As always, please note that this tutorial only accompanies the other materials for Day 12 and that you are expected to have worked through the reading for that day before tackling this tutorial.

Nonlinear relationships: quadratic terms

So far, we have not encountered serious violations of the assumption of linearity - a linear relationship between predictors and outcome. But this assumption simply means that we impose a linear structure on the relationship between x and y. Coefficient estimates from a regression model will not reveal this.

Theoretical example

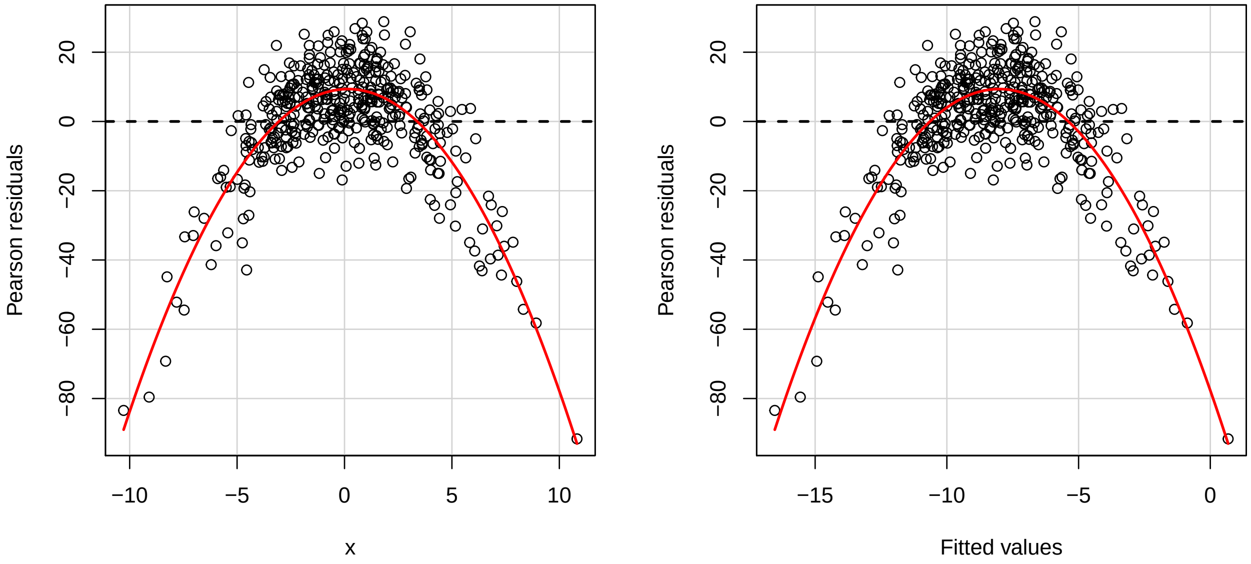
Theories might often make predictions of the form, "as x increases, y first increases, and then drops again". An example for this is the Kuznets curve in economics, suggesting that as countries developed, income inequality first increased, peaked, and then decreased (summarized, for instance, in Acemoglu and Robinson 2002). This implies a so-called curvilinear relationship between economic development and inequality: both poor and rich countries have low inequality, but middle-income countries should exhibit high levels of inequality.

Example with simulated data

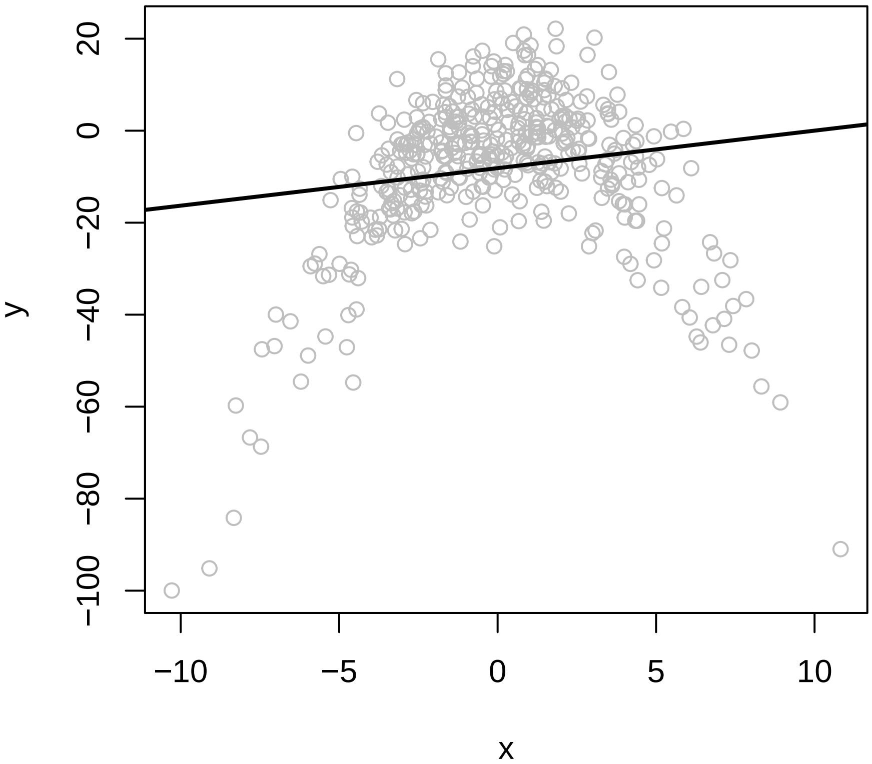
Take the following example:

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.650  -5.757   3.239   9.980  28.822
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -8.1480     0.8235  -9.895  < 2e-16 ***
## x             0.8155     0.2559   3.186  0.00155 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.02 on 425 degrees of freedom
## Multiple R-squared:  0.02333,    Adjusted R-squared:  0.02103
## F-statistic: 10.15 on 1 and 425 DF,  p-value: 0.001547
```

Perhaps you might notice the low R^2 value, but that itself is not indicative of problems. Examining the residual plots, however, reveals that the the model produces residuals that are grouped below 0 at low and high values of x:

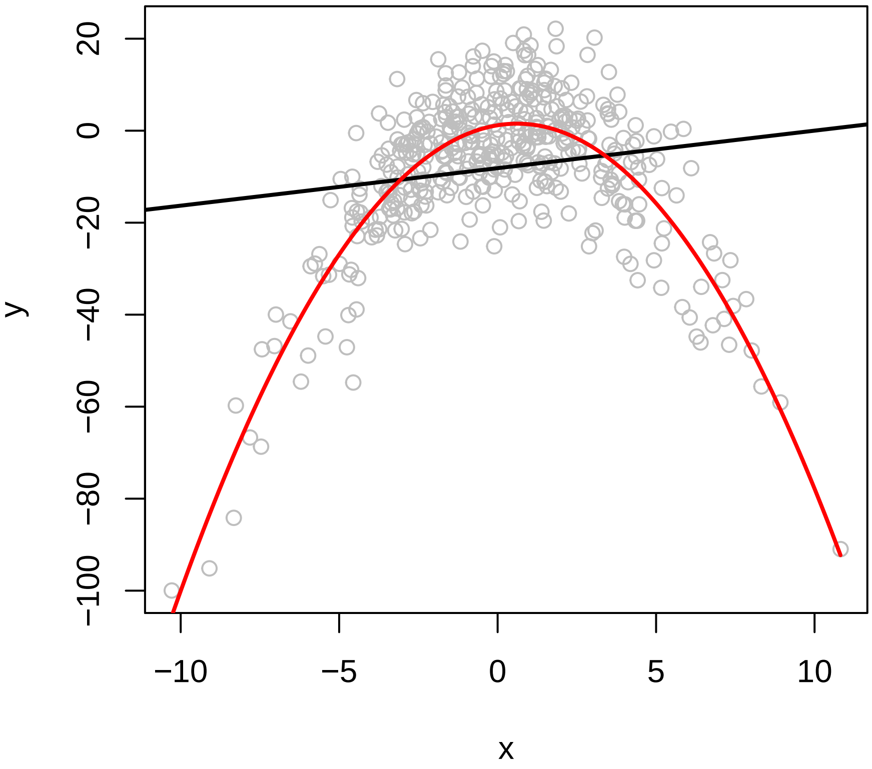


In this case, the residuals strongly indicate a (negative) quadratic relationship between x and y. Plotting x against y and adding the regression line from the model we just fit shows that imposing a linear relationship misses the true structure of the data quite a bit:



To account for a quadratic relationship in the regression model, we enter a squared term of the predictor (x^2), using the code `I(x^2)`.

```
##
## Call:
## lm(formula = y ~ x + I(x^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.166  -6.533   0.121   6.870  24.074
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.18336    0.55212   2.143   0.0327 *
## x            1.11205    0.14338   7.756 6.59e-14 ***
## I(x^2)      -0.90137    0.02946 -30.601  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.511 on 424 degrees of freedom
## Multiple R-squared:  0.6956, Adjusted R-squared:  0.6942
## F-statistic: 484.4 on 2 and 424 DF,  p-value: < 2.2e-16
```



Example with survey data

A second, less dramatic but still consequential example with real-world data illustrates the point as well. Here I'm using a dataset on the wages of 3680 Norwegian survey respondents. The data are taken from the European Social Survey's training module at http://essedunet.nsd.uib.no/cms/topics/multilevel/ch1/5.html.

```
wages.dat <- read.csv("http://www.jkarreth.net/files/ess_wages.csv")
summary(wages.dat)
```

```
##       wage           edyears           age           female
##  Min.   : 25.00   Min.   : 0.000   Min.   :16.0   Min.   :0.0000
##  1st Qu.: 71.00   1st Qu.: 1.000   1st Qu.:30.0   1st Qu.:0.0000
##  Median : 83.33   Median : 3.000   Median :39.0   Median :0.0000
##  Mean   : 90.11   Mean   : 2.651   Mean   :39.5   Mean   :0.4696
##  3rd Qu.:102.45   3rd Qu.: 3.000   3rd Qu.:48.0   3rd Qu.:1.0000
##  Max.   :343.75   Max.   :11.000   Max.   :74.0   Max.   :1.0000
##                   egp
##  Lower service class:1044
##  Routine non-manual :1139
##  Skilled workers    : 607
##  Unskilled workers  : 598
##  Upper service class: 292
##
```

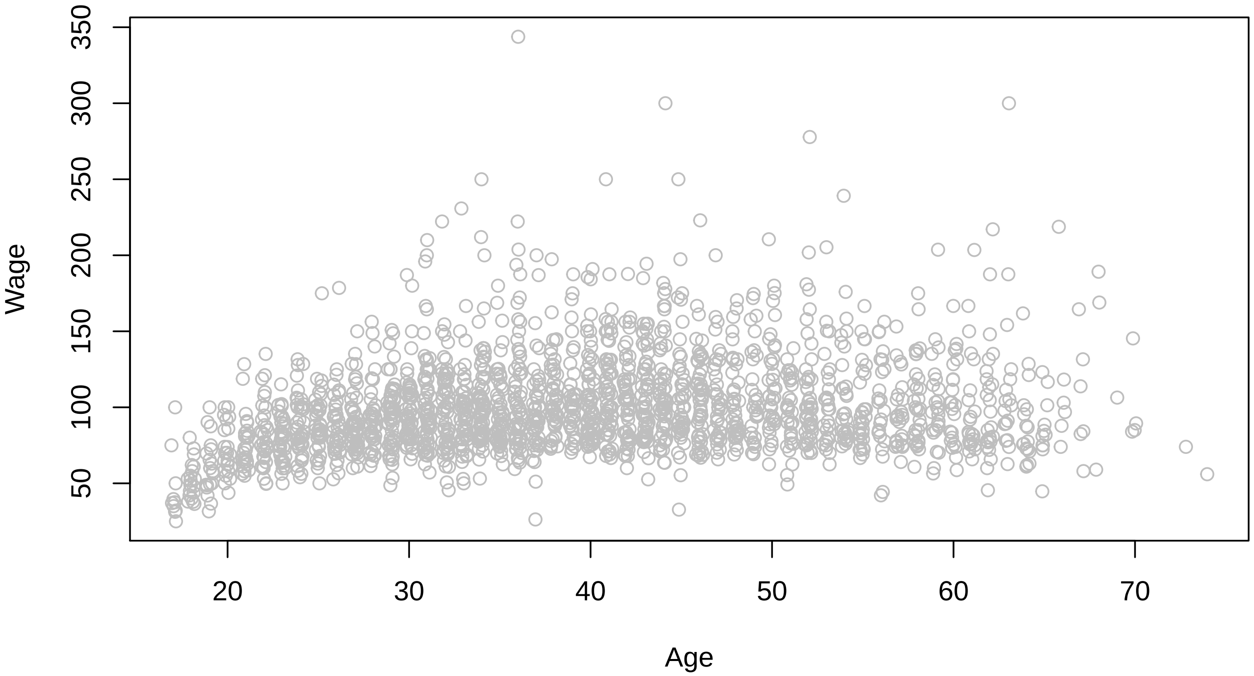
These data contain the following variables:

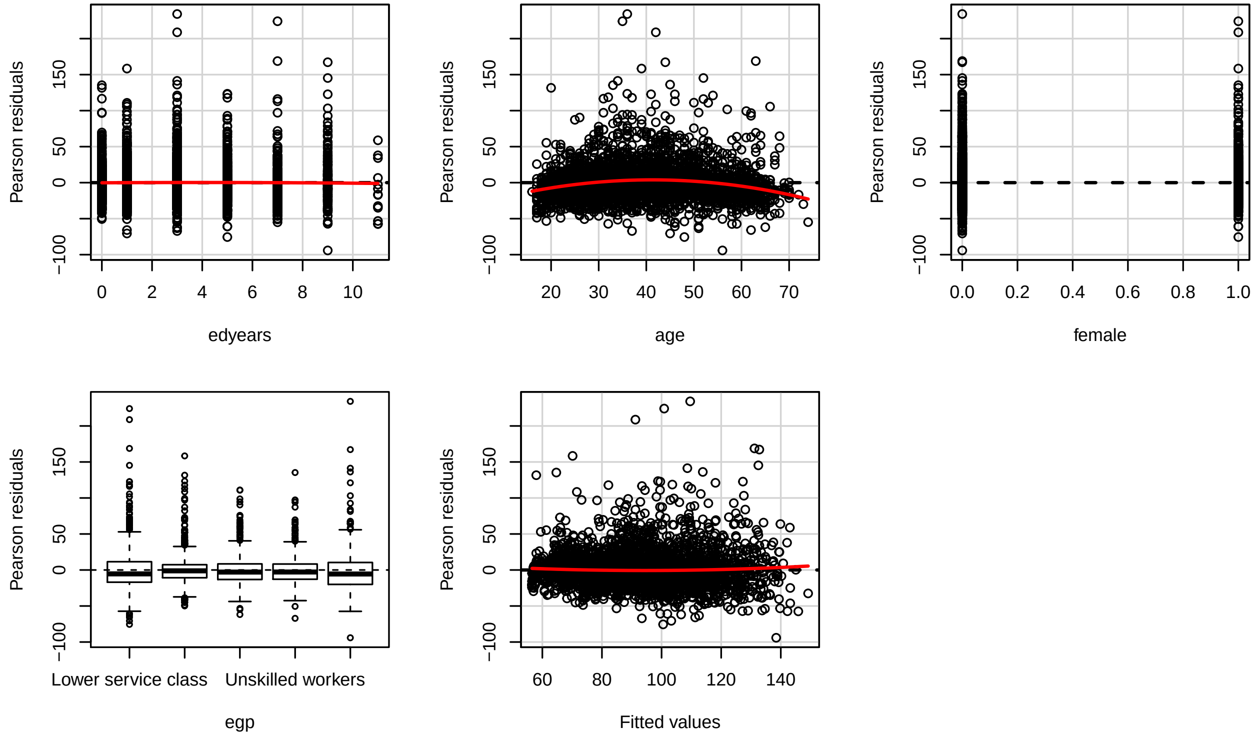
Variable	Description
wage	Hourly wage in Norwegian kronor
edyears	Years of non-compulsory education
age	Age in years
female	1 if female, 0 if male
egp	Type of occupation

Fitting a model assuming linear relationships between each predictor and respondents' wages returns the following results:

```r
mod <- lm(wage ~ edyears + age + female + egp, data = wages.dat)
summary(mod)
```

```
##
## Call:
## lm(formula = wage ~ edyears + age + female + egp, data = wages.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -94.100 -13.761  -2.858   8.719 234.155
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             78.81362    1.94612  40.498   <2e-16 ***
## edyears                  3.23882    0.21025  15.404   <2e-16 ***
## age                      0.47133    0.03313  14.228   <2e-16 ***
## female                 -17.10797    0.94132 -18.174   <2e-16 ***
## egpRoutine non-manual  -13.17149    1.21909 -10.804   <2e-16 ***
## egpSkilled workers     -12.60142    1.42218  -8.861   <2e-16 ***
## egpUnskilled workers   -12.57560    1.44120  -8.726   <2e-16 ***
## egpUpper service class   4.09760    1.68588   2.431   0.0151 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.5 on 3672 degrees of freedom
## Multiple R-squared:  0.3484, Adjusted R-squared:  0.3471
## F-statistic: 280.4 on 7 and 3672 DF,  p-value: < 2.2e-16
```





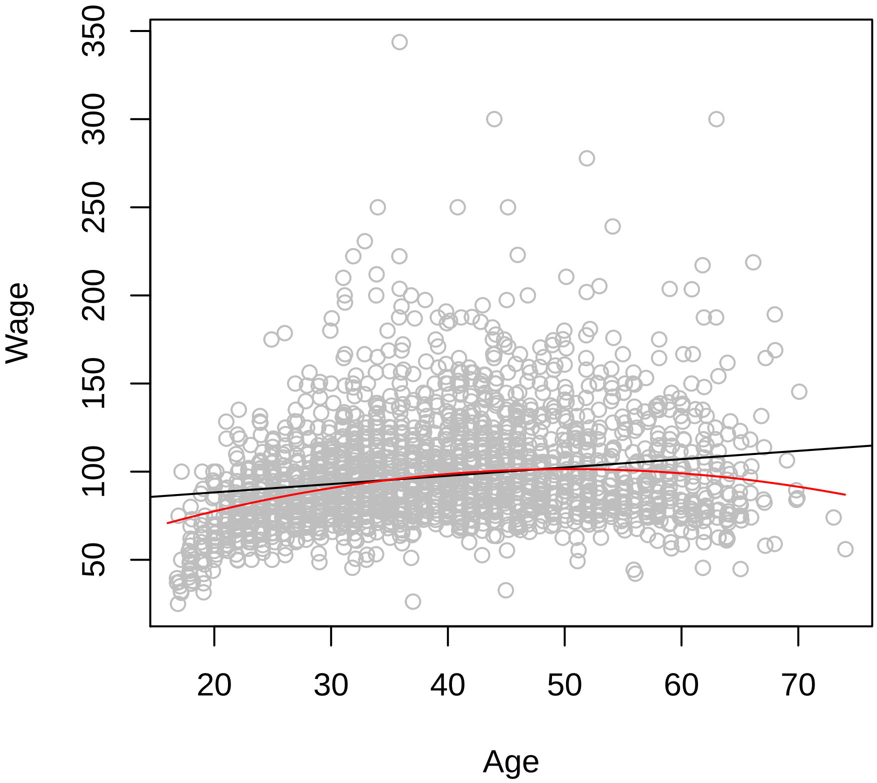
```
##              Test stat Pr(>|t|)
## edyears         -0.488    0.626
## age            -10.724    0.000
## female           0.635    0.525
## egp                 NA       NA
## Tukey test       2.283    0.022
```

Here, the residual plot for the age variable does suggest a nonlinear relationship. Adding it to the model returns the following results:

```r
mod2 <- lm(wage ~ edyears + age + I(age^2) + female + egp, data = wages.dat)
summary(mod2)
```

```
##
## Call:
## lm(formula = wage ~ edyears + age + I(age^2) + female + egp,
##     data = wages.dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -90.899 -13.342  -2.936   8.242 231.857
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             37.785293   4.279082   8.830  < 2e-16 ***
## edyears                  3.139944   0.207269  15.149  < 2e-16 ***
## age                      2.615905   0.202624  12.910  < 2e-16 ***
## I(age^2)                -0.025977   0.002422 -10.724  < 2e-16 ***
## female                 -16.956516   0.927151 -18.289  < 2e-16 ***
## egpRoutine non-manual  -11.588935   1.209635  -9.581  < 2e-16 ***
## egpSkilled workers     -11.104731   1.407543  -7.889 3.97e-15 ***
## egpUnskilled workers   -10.905792   1.427856  -7.638 2.80e-14 ***
## egpUpper service class   4.181041   1.660322   2.518   0.0118 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.13 on 3671 degrees of freedom
## Multiple R-squared:  0.3682, Adjusted R-squared:  0.3668
## F-statistic: 267.4 on 8 and 3671 DF,  p-value: < 2.2e-16
```

And showing the regression line from the first model (black) and second (red) illustrates the meaning of the squared term.



Using residuals to check and improve models

You've already worked with residuals a lot, but here is one more way they can be used to detect potential omitted variables in a regression model. As an example, we'll work with the "Angell" data from John Fox's "car" package. See http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-3E/datasets/Angell.pdf for a description of these data, which come from R. C. Angell (1951), "The moral integration of American Cities," American Journal of Sociology, 57 (part 2): 1-140. The dataset contains 43 rows; each is one American city. The variables are:

Variable	Description
`moral`	Composite of crime rate and welfare expenditures.
`hetero`	Ethnic heterogenity, from percentages of nonwhite and foreign-born white residents.
`mobility`	Geographic mobility: from percentages of residents moving into and out of the city.
`region`	E = Northeast, MW = Midwest, S = Southeast, W = West

The outcome variable is Angell's morality index.

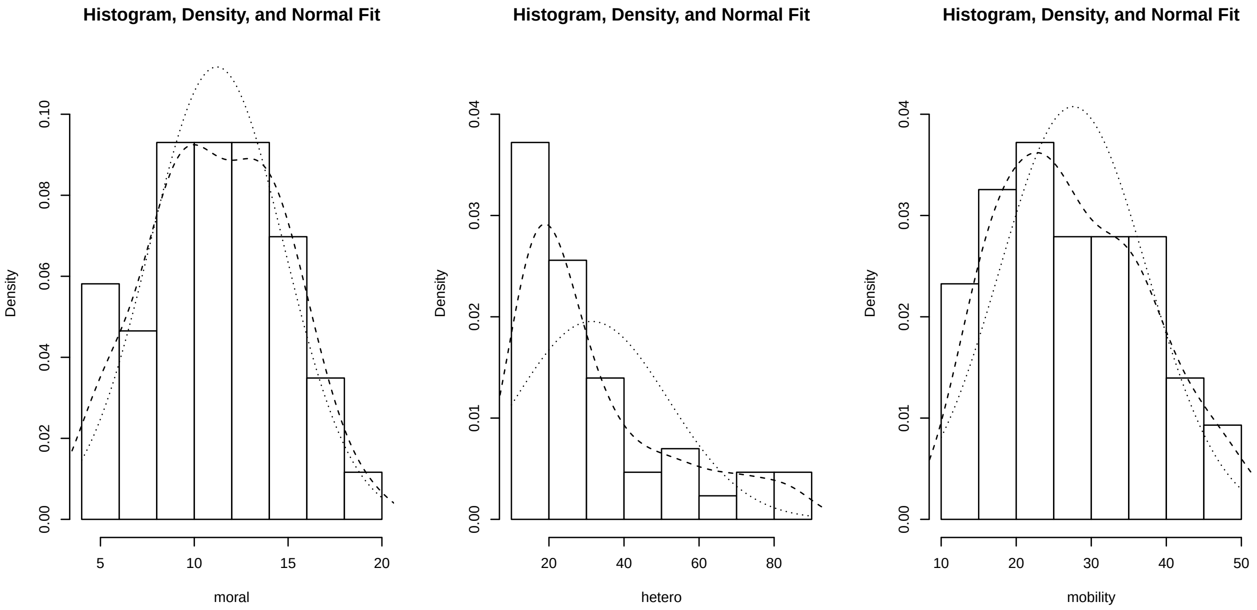
```
angell.dat <- read.csv("http://www.jkarreth.net/files/angell.csv")
summary(angell.dat)
```

```
##      moral           hetero         mobility      region        city
##  Min.   : 4.20   Min.   :10.60   Min.   :12.10   E : 9   Akron     : 1
##  1st Qu.: 8.70   1st Qu.:16.90   1st Qu.:19.45   MW:14   Atlanta   : 1
##  Median :11.10   Median :23.70   Median :25.90   S :14   Baltimore : 1
##  Mean   :11.20   Mean   :31.37   Mean   :27.60   W : 6   Birmingham: 1
##  3rd Qu.:13.95   3rd Qu.:39.00   3rd Qu.:34.80           Bridgeport: 1
##  Max.   :19.00   Max.   :84.50   Max.   :49.80           Buffalo   : 1
##                                                          (Other)   :37
```

```
library(psych)
```

```
##
## Attaching package: 'psych'
##
## The following object is masked from 'package:car':
##
##     logit
```

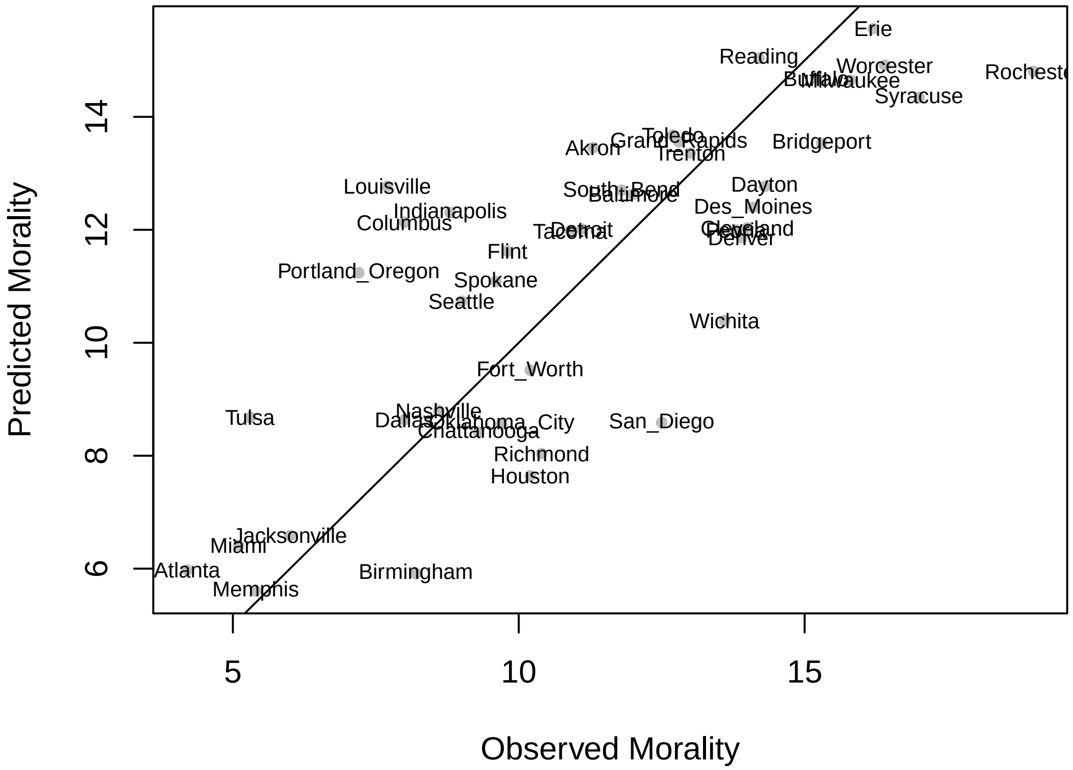
```
multi.hist(angell.dat[, c(1:3)], ncol = 3)
```



```
mod <- lm(moral ~ hetero + mobility, data = angell.dat)
```

Plotting observed against predicted values can reveal which observations the model over- and under-predicts:

```
angell.dat$predicted <- fitted(mod)
plot(x = angell.dat$moral, y = angell.dat$predicted,
     xlab = "Observed Morality", ylab = "Predicted Morality", pch = 20,
     col = "gray")
abline(a = 0, b = 1)
text(x = angell.dat$moral, y = angell.dat$predicted,
     labels = angell.dat$city, cex = 0.7)
```



In the plot above, the line shows the perfect concordance between observed and predicted morality. For observations to the left of the line, the model predicts higher than actual morality; for observations to the right of the line, the model predicts lower than actual morality. Patterns in this plot can suggest potential omitted variables that should be in the model. Can you think of any omitted variables that this particular plot suggests?

Other formal checks for violations of the regression assumptions

Most of the diagnostics we've discussed so far relied on visual interpretation of residual plots. Statisticians have developed a larger number of more formal tests for violations of the regression assumptions; most of them are discussed in *AR*. Often, you can "eyeball" violations of the assumptions in residual plots, but more formal tests can be useful. Overall, though, the important question you should usually ask is whether potential violations of the OLS assumptions **bias** estimation results and whether they affect the standard errors for coefficient estimates.