8. Model diagnostics

- We know how to estimate parameters and make hypothesis tests for linear models.
- We know how to make predictions, with uncertainty estimates, using linear models.
    1. What if our conclusions depend on our choice of model?
    2. What if our model is a poor description of the data?
    3. What if a much better model exists?
    4. What if the model describes some parts of the data okay, but not other parts?
- How can we answer these questions?
    1. **Graphical investigations**. Make informative plots.
    2. **Quantitative investigations**. Find informative tests, or other interpretable statistics.

Looking for patterns in the residuals

- Recall that the **residuals** for a linear model are $e_1, \ldots, e_n$ in the linear model $\mathbf{y} = \mathbb{X}\mathbf{b} + \mathbf{e}$.
- Residuals estimate the measurement errors $\epsilon_1, \ldots, \epsilon_n$ in the probability model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- Since $\mathbf{b}$ is a noisy estimate of β, $\mathbf{e}$ is a noisy estimate of $\boldsymbol{\epsilon}$.
- The specification that $\epsilon_1, \ldots, \epsilon_n \sim \text{iid normal}(0, \sigma)$ implies the measurement errors have no pattern.
- Any pattern, or association with some other variable, that we can find in the residuals contradicts the model and could lead to improvements.
- The search for patterns in the residuals can take creativity and persistence.

Residuals for time series data

- A fairly common type of data has points collected through time. This type of data is called a **time series**.
- For example, the annual data we investigated on unemployment and life expectancy are both time series.
- Time series might be expected to have measurements at points close in time that are more similar than those distant in time. If this is true of residuals, the pattern is inconsistent with the iid measurement error model.

Question 8.1. How can we look for temporal patterns in the residuals? Think of (at least) two plots to make.

(i) A timeplot plots residuals against time; (ii) plotting e_i against e_{i-1} for $i = 2, \ldots, n$ looks for correlations between neighboring residuals.

Residuals for unemployment vs life expectancy

- Recall the linear model relating life expectancy to unemployment:

```
lm1 <- lm(L_detrended~U_detrended)
```

- Some graphical investigations of the residuals follow on the next slide.
- One way to see if the residuals have statistically noticeable dependence is to fit a linear model to the residuals $e_{1:n}$ of the form

$$e_i = be_{i-1} + h_i, \quad \text{for } i = 2, 3, \ldots, n,$$

where h_i is the residual error when e_{i-1} is used to predict e_i.

Question 8.2. Why do we not need an intercept here?

The residuals are centered on zero by their construction, as long as there is an intercept in the model.

Question 8.3. How would you fit this linear model for the residuals in R?

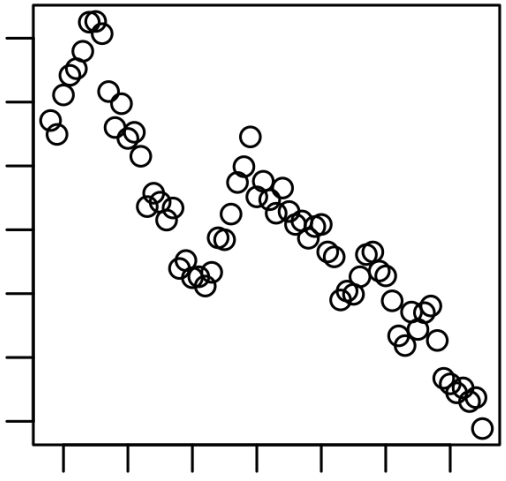
```
n<-length(L_detrended)\e<-resid(lm1)[2:
n]\lag_e<-resid(lm1)[1:(n-1)]lm(e~lag_e-1)
```

```r
n <- length(resid(lm1))
e <- resid(lm1)[2:n]
lag_e <- resid(lm1)[1:(n-1)] # NOTE WE NEED 1:(n-1) NOT 1:n-1
lm2 <- lm(e~lag_e-1)
head(model.matrix(lm2),3)

##        lag_e
## 17 0.8556642
## 18 0.7466793
## 19 1.0556704

summary(lm2)$coef

##        Estimate Std. Error  t value     Pr(>|t|)
## lag_e 0.9898371 0.03167559 31.24921 2.834997e-41
```

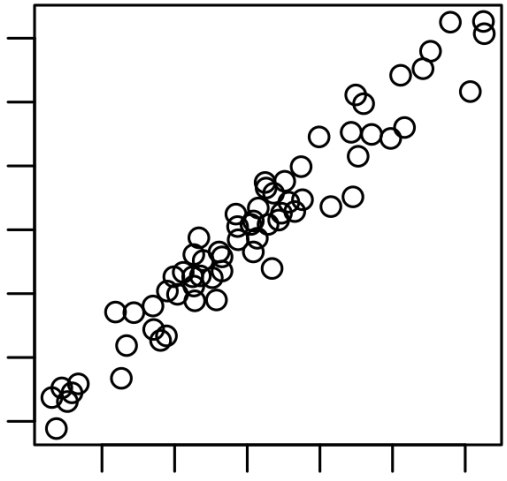
Question 8.4. What do you conclude from this analysis?

The residuals are much more highly correlated than can be explained by chance variation.

```
plot(U$Year,resid(lm1))
```



```
plot(lag_e,e)
```



Question 8.5. What do you these plots tell you about (i) the least squares estimate of the association between changes of life expectancy and unemployment; (ii) its standard error and confidence interval?

The model assumption of independent measurement errors is far from the reality. All standard errors and confidence intervals built on this assumption are therefore brought into question.

Why do the detrended residuals have a trend?

- Recall the code we used to construct `L_detrended` and `U_detrended`

```
L <- read.table(file="life_expectancy.txt",header=TRUE)
L_fit <- lm(Total~Year,data=L)
L_detrended <- L_fit$residuals
U <- read.table(file="unemployment.csv",sep=",",header=TRUE)
U_annual <- apply(U[,2:13],1,mean)
U_detrended <- lm(U_annual~U$Year)$residuals
L_detrended <- subset(L_detrended,L$Year %in% U$Year)
```

Question 8.6. We removed a linear trend from both life expectancy and unemployment. What does that mean? What is the equation for the model that we have fitted?

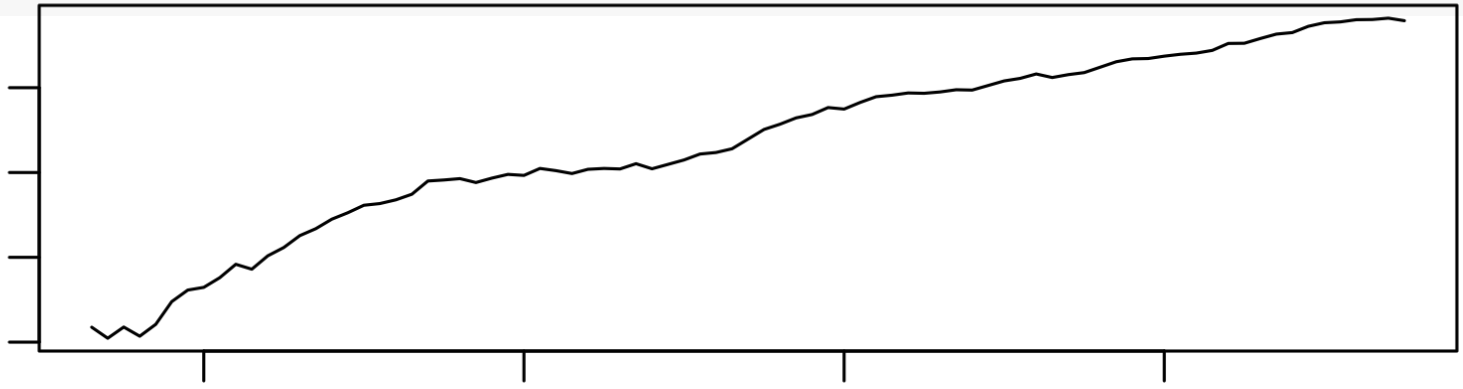
It means we fitted a sample model

$$y_i = a + bi + e_i, \quad i = 1, 2, \ldots, n$$

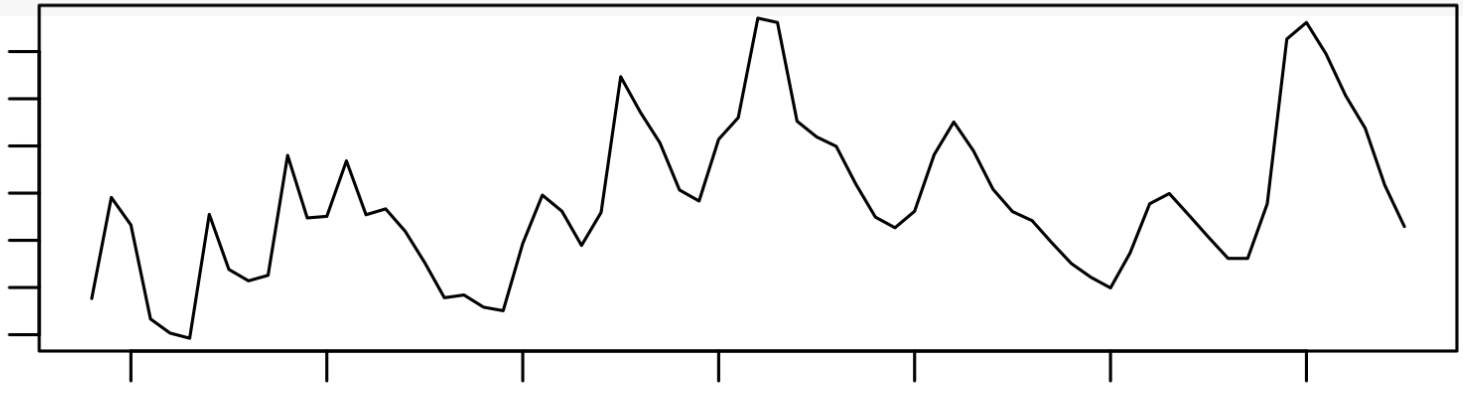
with y_i being either the life expectancy or unemployment for the ith year in the dataset. The detrended data are the residuals e_i, $i = 1, \ldots, n$ from this regression.

Question 8.7. It is a statistical detective puzzle to figure out how the residuals from `lm1` can have a linear trend when all the variables are detrended. Any ideas? Plotting the variables may give a clue.
Look at the axes. Careful reading of the code reveals (after plenty of head-scratching) that the detrending was done before subsetting `L_detrended` to make the times match `U_detrended`. The removed part was trending up and the remaining part trends down.

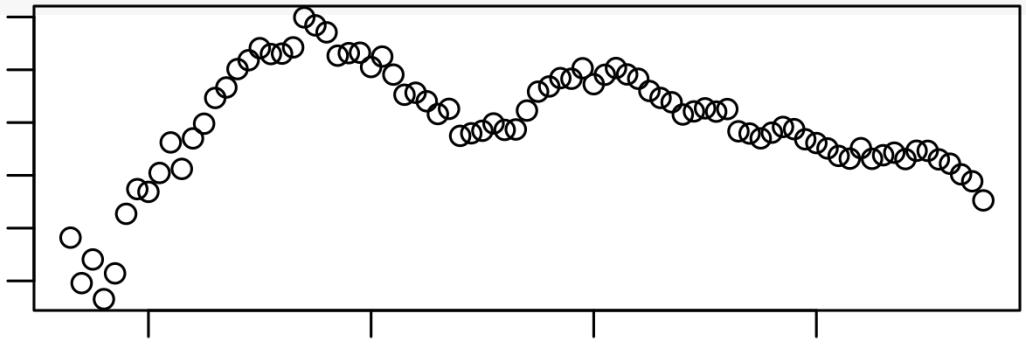
```r
plot(Total~Year,data=L,type="l") # L is life expectancy
```



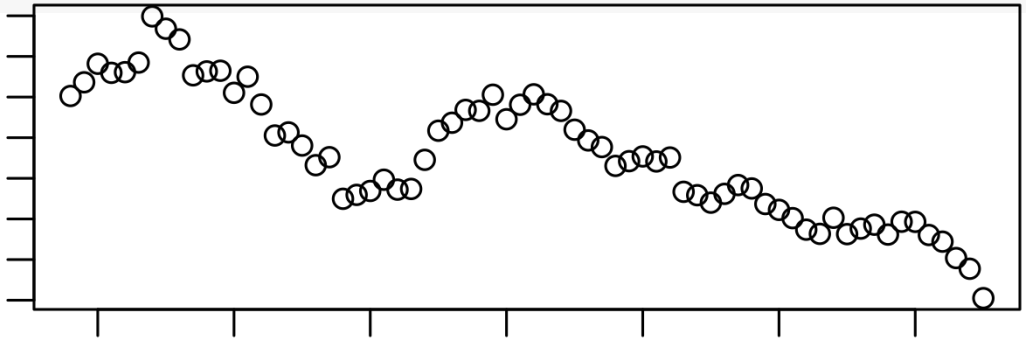
```r
plot(U_annual~Year,data=U,type="l") # U is unemployment
```



```
L_fit <- lm(Total~Year,data=L)
L_detrended <- L_fit$residuals
plot(L_detrended~Year,data=L)
```



```
L_detrended <- subset(L_detrended,L$Year %in% U$Year)
plot(L_detrended~Year,data=U)
```



Rescuing the life expectancy/unemployment analysis

- We have found a serious problem with our linear model analysis.
- From a statistically significant coefficient, we inferred counter-intuitively that higher unemployment is associated with above-trend life expectancy.
- **A p-value is only as good as the probability model producing it.**
- We have found that the probability model we used is seriously defective. It is based on assumptions that are substantially violated.
- This doesn't necessarily mean that the result is right or wrong.
- It means we haven't yet found a good argument either way.
- This topic is of current interest:
  `https://www.nytimes.com/2017/10/16/upshot/how-a-healthy-economy-can-shorten-life-spans.html`

Question 8.8. Can we do a better data analysis? How?

We will find out one way later.

Outliers

- Sometimes one, or a few, points are inconsistent with a model that explains the rest of the data nicely.
- These points are called **outliers**.
- Our first responsibility is to notice them.
- Our second responsibility is to work out whether they affect the conclusions of the analysis. If they don't, the issue becomes unimportant.

Question 8.9. It is tempting to remove clear outliers from the data analysis on the assumption that they are errors. When is that reasonable? When is it a bad decision?

If you have good reason to believe they are errors, and you explain clearly in your analysis what you did, that is okay. However, outliers can be the most informative points: they can tell you about some unexpected special situation. Unexpected data points can be scientific discoveries: Alexander Fleming noticed an accidental mold contamination killed his cultured acteria.

Outliers and responsible scientific conduct

- **Falsification** is the manipulation of research materials, equipment, or processes or changing or omitting data or results such that the research is not accurately represented in the research record (`https://en.wikipedia.org/wiki/Scientific_misconduct`).

Question 8.10. How could inappropriate treatment of outliers lead to charges of falsification? What can a careful data analyst do to avoid that?

Leverage and influence

- A data point has high **leverage** if its explanatory variables are distant from much of the rest of the data, so the point plays a relatively large role in determining the fitted values.
- Leverage of a point i depends only on the design matrix $\mathbb{X} = [x_{ij}]_{n \times p}$, and primarily on $x_{i1}, \ldots, x_{ip}$.
- A point has high **influence** if removing that point leads to large changes in the parameter estimates and fitted values.
- Influence depends on both $\mathbb{X}$ and $\mathbf{y}$.
- An outlier with high leverage is a point of very high influence.

Question 8.11. Sketch a scatterplot (i.e., a plot of $\mathbf{y}$ against a single explanatory vector $\mathbf{x}$) that has a point of high leverage, but not high influence.

A cluster of x-values and an outlier in the x direction which is close to the linear regression line in the y direction. Archimedes: "Give me a lever and a place to stand and I will move the earth."

Question 8.12. Sketch a scatterplot that has a point of high leverage which is also a point of high influence.

A cluster of x-values and an outlier in the x direction which is also an outlier from the linear regression line in the y direction.

Question 8.13. Sketch a scatterplot that has an outlier which is not influential.

An outlier in the middle of a simple linear regression scatterplot.

Practical strategies for influence and leverage

- A small collection of points with unusual and extreme values of the explanatory variables will likely have high leverage and may also have high influence.
- Try removing these points to see if that changes the conclusions of your data analysis. If it does, then hard thought is required.
- A measure of influence is **Cook's distance**, which is computed for a model `lm1` by `cooks.distance(lm1)`.
- We are not going to study Cook's distance carefully. You can investigate the points which have the highest Cook's distance. For example, you can see the effect of removing these points on your conclusions.

Normality

- If the number of points is fairly large (say, more than 30) the estimates of the coefficients in the linear model have a **central limit theorem**.
- Recall that a basic central limit theorem says that the average of many independent identically distributed (iid) random variables approximately follows a normal distribution.
- The least squares estimates of coefficients can be thought of as a kind of averaging of the data. This argument suggests (correctly!) that the distribution of these estimates should also follows a central limit theorem.
- Measurement error with very long tails may lead to observations that look like outliers. They may also behave like outliers, and potentially have high influence.
- Usually, because of the central limit theorem, normality of errors is not especially important. It is more important for prediction intervals.

Non-constant variance

- Our usual probability model assumes (in addition to normality and independence) that the measurement errors have constant variance.
- Plotting the residuals (say, against fitted values or against time or against some other variable) may show that the spread of the residuals is larger in some places than others.
- Taking the logarithm of non-negative data may help in this case.