7. 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 are also known as **residual errors** or **errors**.
- Residuals estimate the measurement errors $\epsilon_1, \ldots, \epsilon_n$ in the model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- Statistical properties of $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ are usually carried out using a probability model that $\epsilon_1, \ldots, \epsilon_n$ are independent identically distributed (iid) $N[0, \sigma]$.
- iid random variables 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. How can we look for temporal patterns in the residuals? Think of (at least) two plots to make.

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

$$e_i = \beta e_{i-1} + h_i, \quad \text{i=2,3,...,n},$$

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

Question 7.1. Why do we not need an intercept here?

Question 7.2. How would you fit this linear model for the residuals in R?

```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)
```

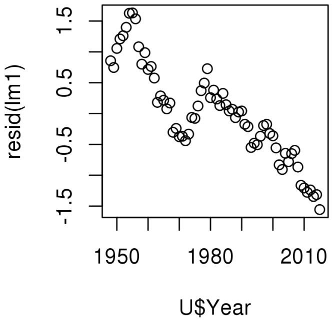
```
##        lag_e
## 17 0.8556642
## 18 0.7466793
## 19 1.0556704
```

```r
summary(lm2)$coef
```

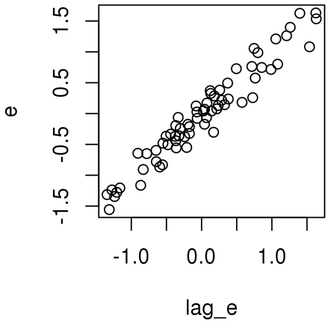
```
##         Estimate Std. Error  t value     Pr(>|t|)
## lag_e 0.9898371 0.03167559 31.24921 2.834997e-41
```

Question 7.3. What do you conclude from this analysis?

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




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



Question 7.4. 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?

Why do the detrended residuals have a trend?

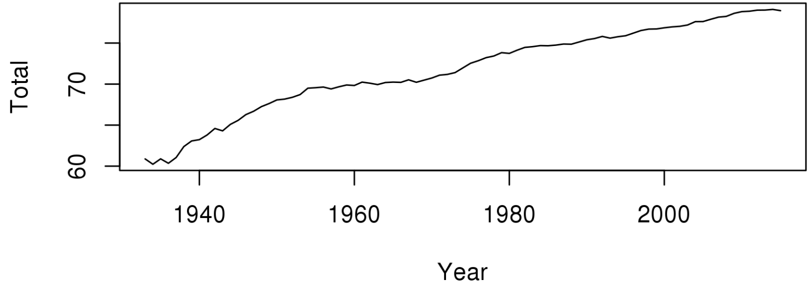
- Recall the code we used to construct `L_detrended` and `U_detrended`

```
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)
```

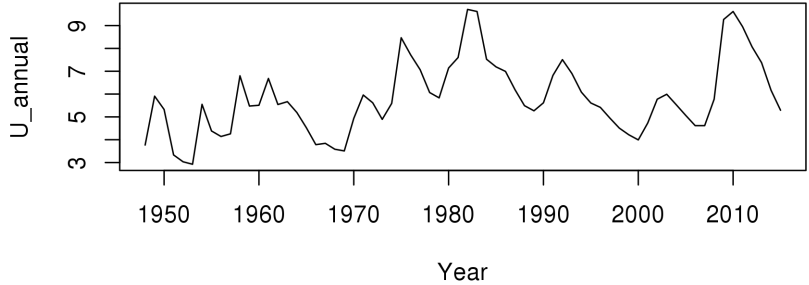
Question 7.5. We removed a linear trend from both life expectancy and unemployment. What does that mean? What is the equation for the model that we have fitted?

Question 7.6. 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.

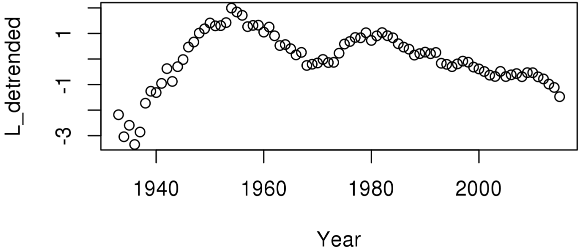
```
plot(Total~Year,data=L,type="l") # L is life expectancy
```



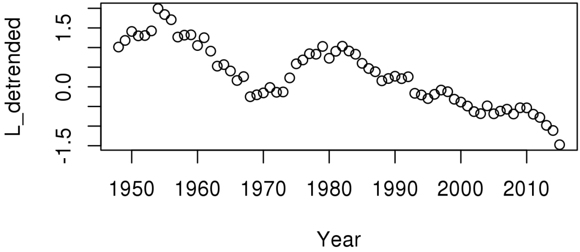

```
plot(U_annual~Year,data=U,type="l") # U is unemployment
```



```
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 does mean 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 7.7. Can we do a better data analysis? How?

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 less urgent).

Question 7.8. 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 dangerous?

Outliers and responsible scientific conduct

- According to the National Science Foundation, **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 7.9. 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 7.10. 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.

Question 7.11. Sketch a scatterplot that has a point of high leverage which is also a point of high influence.

Question 7.12. Sketch a scatterplot that has an outlier which is not influential.