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**. Make relevant statistical tests, or calculate 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 $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- $\mathbf{b}$ is a noisy estimate of $\boldsymbol{\beta}$, meaning that we cannot find $\boldsymbol{\beta}$ exactly due to measurement error. Similarly, $\mathbf{e}$ 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.

- Plot the residuals against time

- Plot the difference of residuals – look for runs of +ve & -ve values.

- Plot e_i against e_{i-1} for $i = 2, \ldots, n$. Look for positive correlation to show that +ve & -ve residuals cluster together

- Try fitted versus residuals, but points with similar fitted values may have very different times, so this may not show a temporal pattern.

Residuals for unemployment vs life expectancy

- Recall the linear model relating life expectancy to unemployment:

```
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 <- read.table(file="life_expectancy.txt",header=TRUE)
L <- subset(L,L$Year %in% U$Year)
L_fit <- lm(Total~Year,data=L)
L_detrended <- L_fit$residuals
lm1 <- lm(L_detrended~U_detrended)
```

this removes years with L but not U measured.

note: the $e_1,..,e_n$ are already residuals from regressing residual life expectancy & unemployment after detrending,

- 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

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

so $h_2, \ldots h_n$ are residuals of residuals of residuals.

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 $e_1, .., e_n$ are constructed to have sample mean zero.

If e_{i-1} is zero, we expect e_i to be around zero.

this is y_i this is x_i

Question 8.3. How would you fit the linear model

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

for the residuals from `lm1<-lm(L_detrended~U_detrended)` in R?

```
e <- resid(lm1)   or   e<- lm1$residuals
```
note names(lm1) would tell you to look for lm1$residuals.
```
n <- length(e)
y <- e[2:n]
x <- e[1:(n-1)]
lm2 <- lm(y~x-1)
```
↑ this tells R not to include an intercept.

if the code breaks, it's wrong!

or plot the fitted model.

or check model.matrix(lm2) to see the design matrix.

```
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 -0.28669804
## 18 -0.36158257
## 19 -0.01849112

summary(lm2)$coef

##       Estimate Std. Error  t value     Pr(>|t|)
## lag_e  0.86763 0.06248349 13.88575 2.941982e-21
```

lag is a technical term for one time point previously. There is a lag() function in R. Our model here could be called a lag regression model.

this is an abbreviated version of the full details of a hypothesis test.

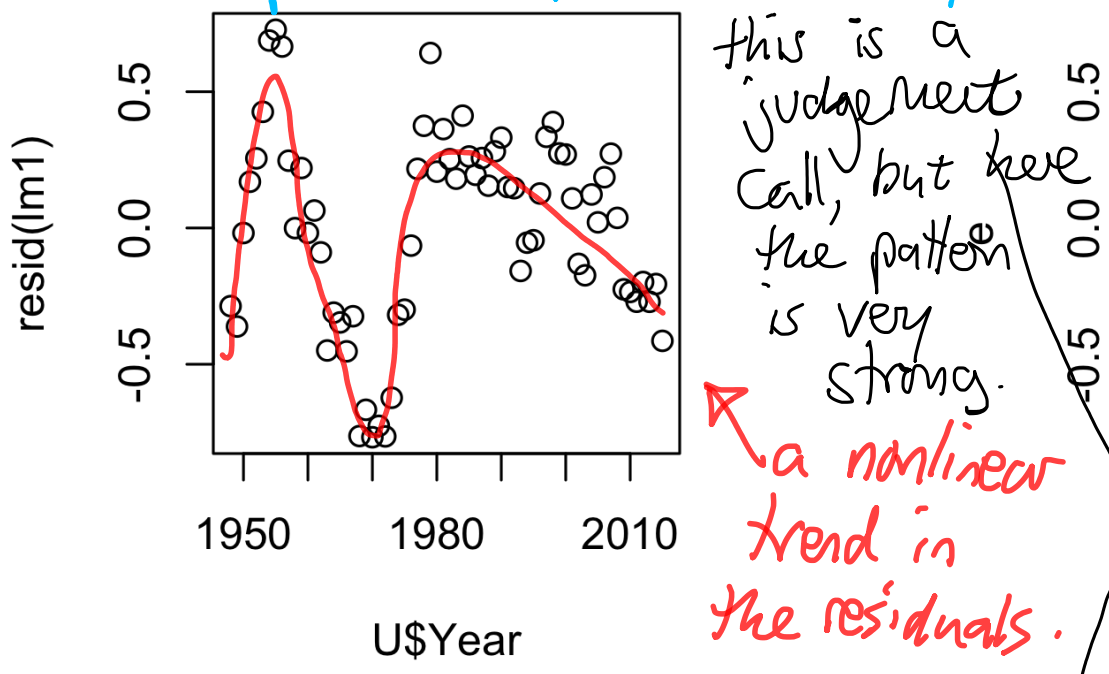
Question 8.4. What do you conclude from this analysis?

There is a clear association between residuals at neighboring lags. The p-value lets us reject the null hypothesis corresponding to a probability model where the true coefficient is zero.

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

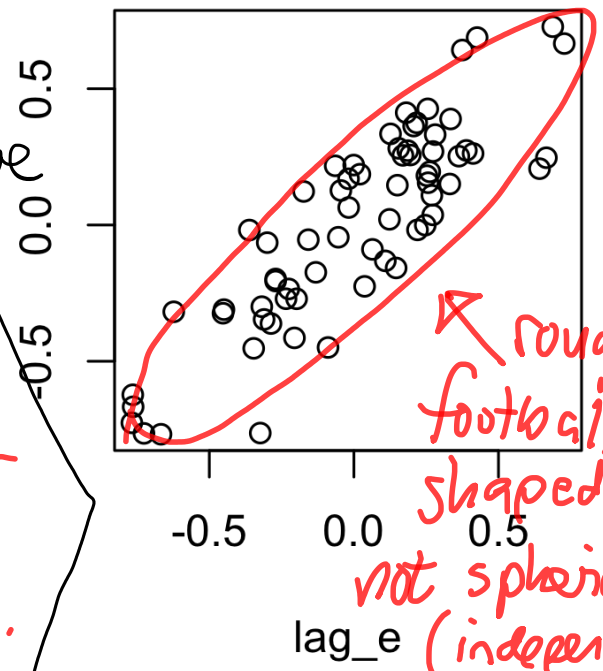
```
plot(lag_e,e)
```

these plots complement the previous hypothesis test.



this is a judgement call, but here the pattern is very strong.

a nonlinear trend in the residuals.



roughly football shaped, but not spherical (independent)

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?

Our probability model has a serious problem. The residuals show a strong pattern. We should be suspicious of probability calculations, such as confidence intervals, computed using this model.

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 8.6. Can we do a better data analysis? How?

Coming soon!

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 8.7. 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?

Outliers can be the most informative datapoints. They tell you something new that you didn't expect (if they are not errors!)

If you have good reason to believe they are errors, and you explain this clearly, it is okay to remove them.

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

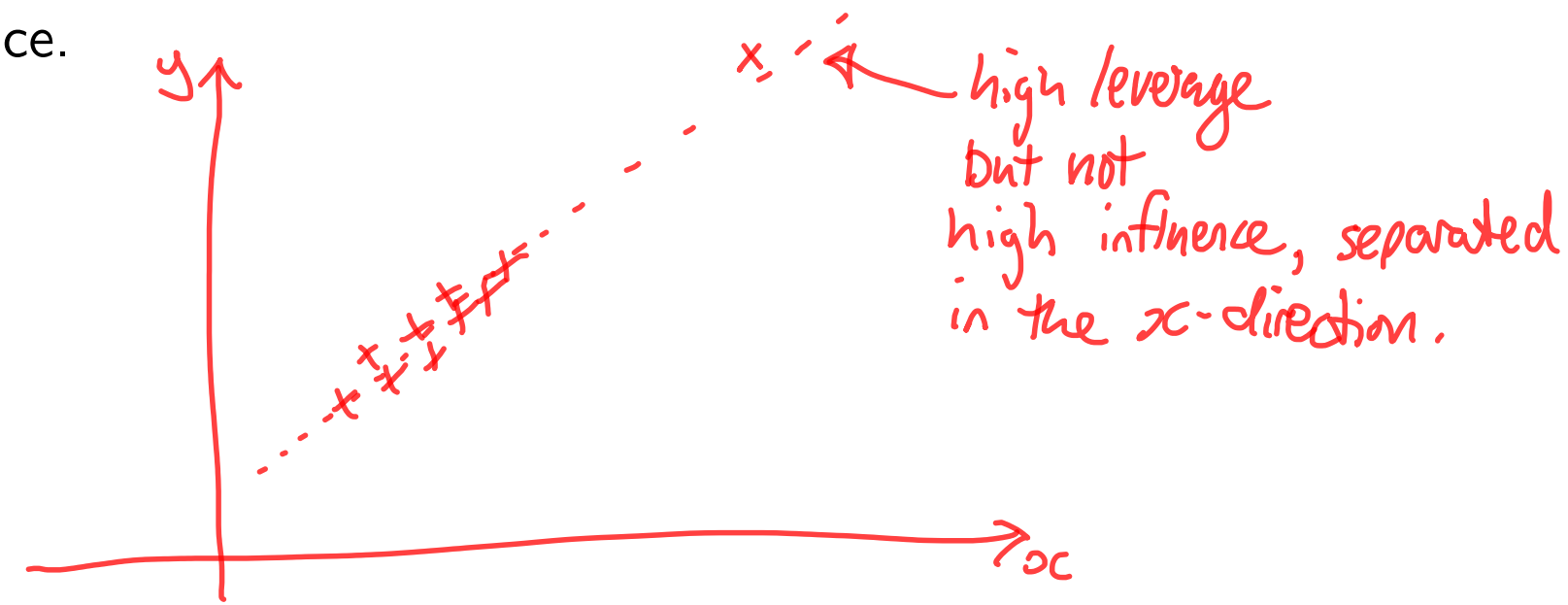
(i) Removing a point to make a plot look better for publication is falsification. This has intent. Carelessly failing to report removal of an outlier is still falsification.

(ii) Provide sound justification for removing data. At a minimum, report that data were removed — reproducibility of the results from the raw data.

Leverage and influence

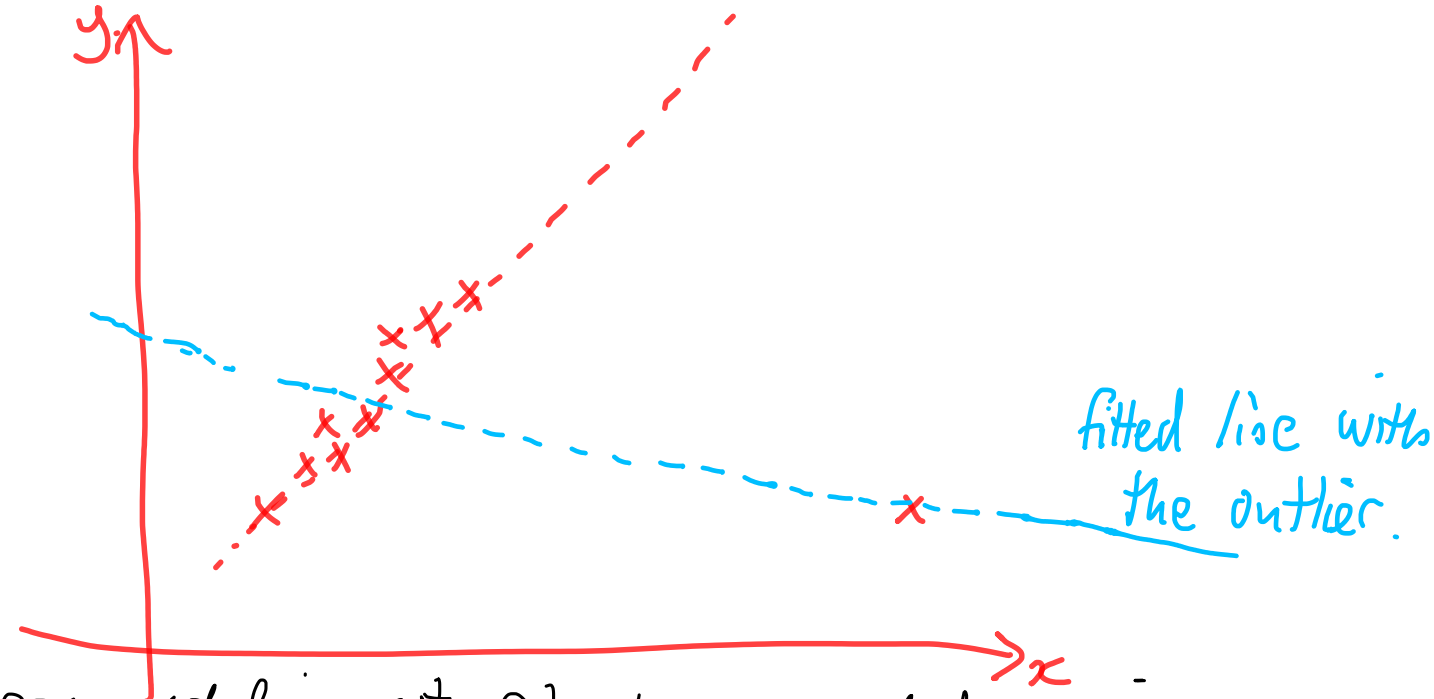
- A data point has high **leverage** if its explanatory variables are distant from much of the rest of the data, so the point plays a relatively large role in determining the fitted values.
- Leverage of a point i depends only on the design matrix $\mathbb{X} = [x_{ij}]_{n \times p}$, and primarily on $x_{i1}, \ldots, x_{ip}$.
- 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 8.9. 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.



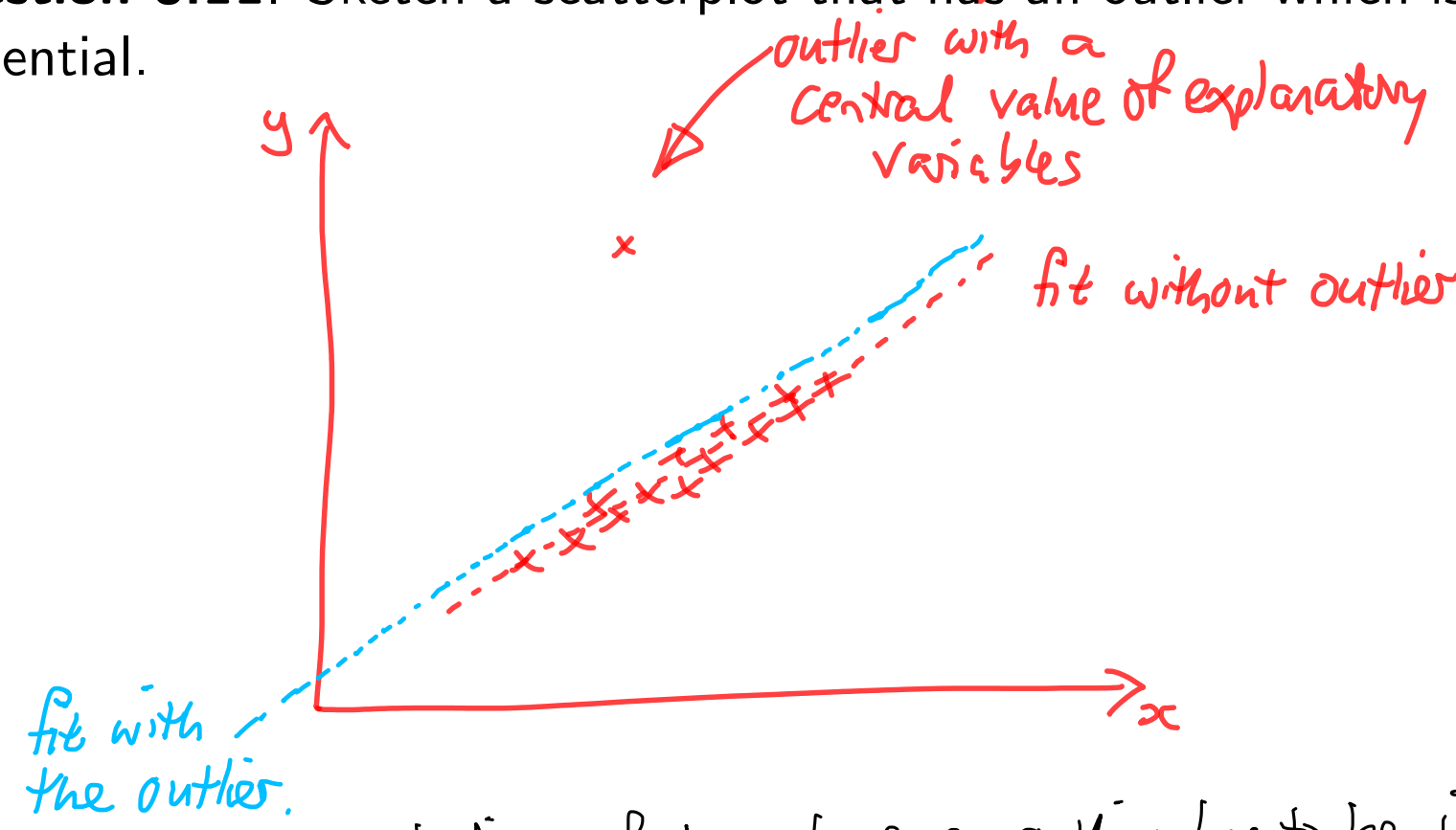
note: the definition of outlier is a large residual, inconsistent with the measurement model. Formally, we shouldn't talk of "outliers in the x-direction."

Question 8.10. Sketch a scatterplot that has a point of high leverage which is also a point of high influence.



The linear model is not robust: one data point can completely change the fitted model. In practice, don't trust a result depending on only one data point! Plot the data!

Question 8.11. Sketch a scatterplot that has an outlier which is not influential.



Note. Interpretation of how large an outlier has to be in order to be problematic, or how much it has to affect the model fit, is context dependent.

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

- A histogram of the residuals assesses the normal measurement error model.
- 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

Very commonly, larger measurements have larger errors. Measurement error is typically proportional to the size of the quantity.

- 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 help surprisingly often in this case.
- Other approaches to deal with this problem are beyond this course, though you now have the necessary background to learn and use these methods. These involve models with a different measurement model from the constant variance model $\epsilon_1, \ldots, \epsilon_n \sim \text{iid normal}(0, \sigma)$.

Logarithms turn proportional (i.e., multiplicative) error into additive error. It is usually worth considering taking logs of non-negative variables.