Lecture 12 (Ch. 3)

From the 2 clicker questions:

Suppose we have data on variable y, only.
What's the "best" number for predicting y?
Sample mean of y, i.e. $\bar{y}$ report $\bar{y} \pm s_y$

Suppose we also have data on x, which is related to y.
What's the "best" number for predicting y?
The fitted value $\hat{y} = \hat{\alpha} + \hat{\beta}x$. report $\hat{y} \pm s_e$
Here, the number depends on x.

$\Rightarrow$ $\hat{y}(x) = ?$

We will improve on these, later.

Given data (x_i, y_i) $i = 1, 2, \ldots, n$
assume $y = \alpha + \beta x$,
which means $y_i = \alpha + \beta x_i + \epsilon_i$ ← errors.
minimize $SSE = \sum_{i=1}^{n} \epsilon_i^2$
to get $\hat{\alpha}, \hat{\beta}$ OLS estimates of α, β
predict: $\hat{y}(x) = \hat{\alpha} + \hat{\beta}x$ OLS fit to data.

Decompose Var. $S_{yy} = SST = SS_{exp} + SS_{unexp}$

$R^2 \equiv \frac{SS_{exp}}{SST}$ ~ goodness of fit.

$s_e = \sqrt{\frac{SS_{unexp}}{n-2}}$ = std. dev. of errors. ~ typical error.

Note: The idea of minimizing SSE (in fitting) translates to maximizing $SS_{explained}$ (in ANOVA of regression)

I say $\hat{\alpha}, \hat{\beta}$, SSE, book says a, b, SSResid (and SSE)

⇒ These quantities are generally formatted in an ANOVA Table. Look at p.121 and learn how to read the outputs to identify what you need. For example, some computer outputs may call R^2, Coeff. of determ., or r^2, R-sqd, ... Also, they may give RMSE, instead of s_e:

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1}{n-2}\sum_{i}^{n}(y_i - \hat{y}_i)^2}$$

Root ← funny mean ← error ← squared

see examples in Lab.

Why is R^2 written as R^2 ?!

IMPORTANT!

R^2 ← not a square of anything; at least not generally.

= symbol.

To see why it is written as R^2 (or even r^2), consider our example: (as in our/many books; height/weight)

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}} \text{ or } \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0.8891\text{6}$$

Note $(.8891\text{6})^2 = \underline{0.79}$ (see R^2 in prev lecture)

I.e. coeff. of deter. $(R^2) = (r)^2$

But only in simple linear regression. i.e. $y = \alpha + \beta x$.

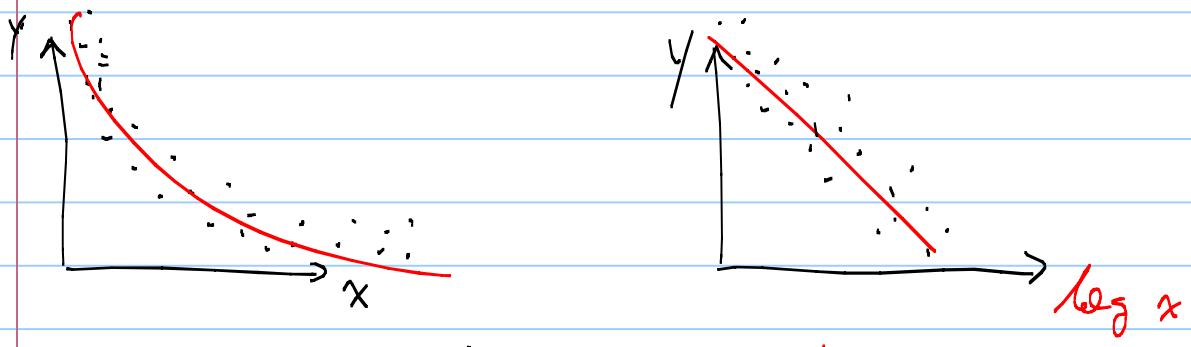
In everything else we will do next, $R^2 \neq (r)^2$.

Non linear relations

So far, we've considered situations where x & y are linearly related. If the relationship (in the scatterplot) is non linear, then there are 2 options:

1) If monotonic, then transform data:

For example, $x \longrightarrow \log(x)$ often straightens scatter plots that look like this



→ Then, we do regression on y vs. $\log(x)$.

I.e. $y = \alpha + \beta(\log x)$ not $y = \alpha + \beta x$

→ and decompose (ie. Anova) as before.

$SST = SS_{exp} + SS_{unexp}$ where x is replaced by $\log(x)$.

↳ by $\log(x)$

Usually, one (or some) of the following transformations straightens a scatterplot:

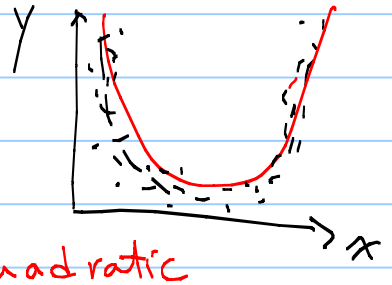
$\log x$, e^x, $\sqrt{x}$, $(x)^{1/3}$, same for y.

The best rule is to try different ones, and check the scatterplot.

Non linear relations

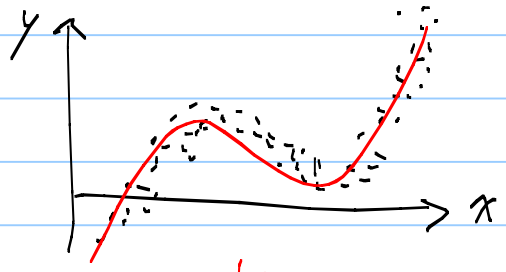
2) If the relationship is not monotonic?

E.g.



quadratic

or



cubic

Answer

$\hat{y} = \alpha + \beta_1 x + \beta_2 x^2$

$\hat{y} = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

These are examples of polynomial regression.

> in R `lm(y ~ x + I(x^2) + I(x^3) + ...)`

for R reasons.

As in simple linear regression, we can still decompose the total variability in y into explained and unexplained, and so, compute R^2, se, ...

The only difference is that $R^2 \neq (r)^2$

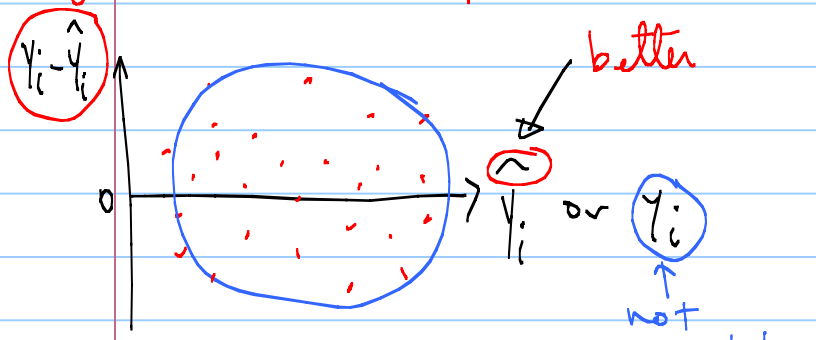
Note that with the same basic ideas we have learned so far, we can now fit (almost) any data.

Visual assessment of fit quality

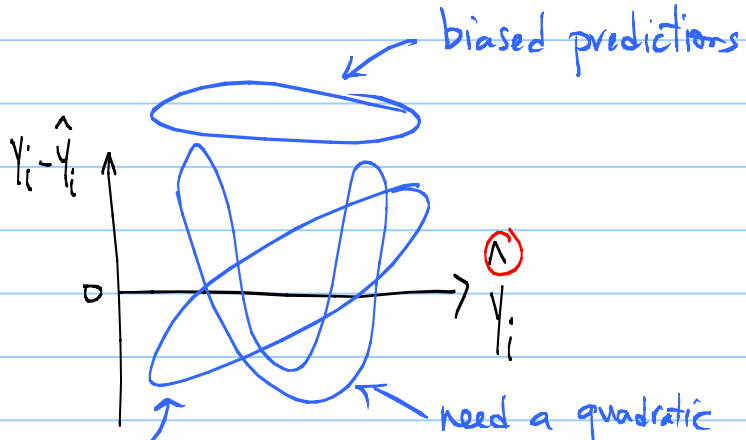
Residual Plot: (Mona Lisa)

count

errors (last column of "Table" before)



Random, symmetric about error = 0 line



under/over forecasting small/large y.

need a quadratic model.

GOOD Model/fit

∴ Nothing more to do

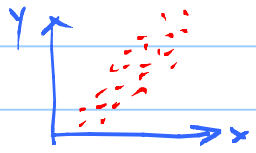
BAD Model/fit

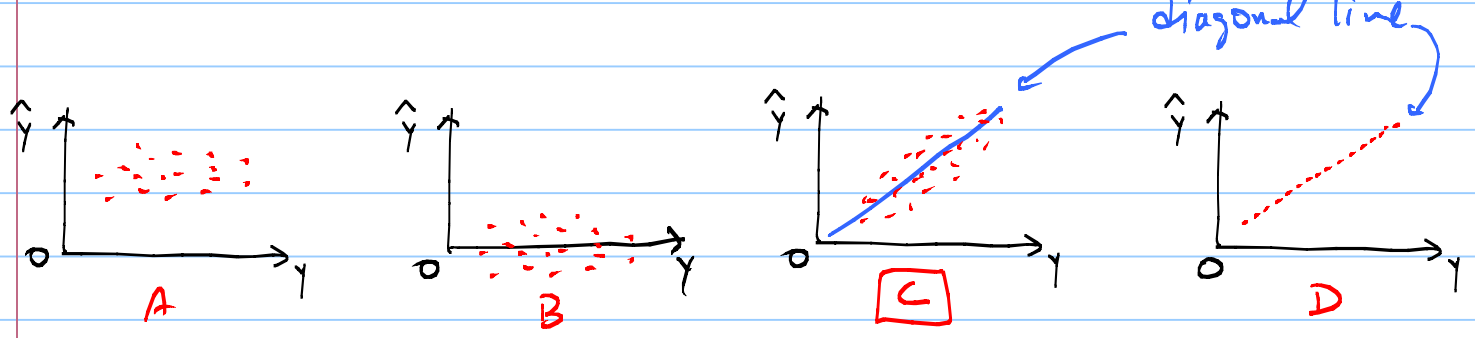
Things to do:

1) Transform data, or
2) fit diff. polynomial,

Q1: Instead of residual plots, some people look at the plot of the predictions ($\hat{y}$) vs. observations (y). Which of the following displays a good model/fit?

Suppose the data look like:





Summary: When you see data on (x, y),

→ Look at their scatterplot (and histograms, and ...)

→ If linear, do regression $y = \alpha + \beta x$

Assess performance with ANOVA (R^2, s_e, residual plots, ...)

→ If non linear, but monotonic,

Then transform x and/or y: E.g. $y = \alpha + \beta \log x$

Assess performance with ANOVA.

→ If non monotonic, Then polynomial regression:

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots$$

Assess performance with ANOVA.

→ Extrapolate Cautiously! Remember The -755 pound person!

Also, Recall how to manipulate eqns like These:

E.g. $y = \alpha + \beta \ln x$

↳ $(y - \alpha)/\beta = \ln x \Rightarrow x = e^{(y-\alpha)/\beta}$

↳ $y = \ln e^{\alpha} + \ln x^{\beta} = \ln(e^{\alpha} x^{\beta}) \Rightarrow e^{y} = e^{\alpha} x^{\beta}$

FYI

Also know about additive/multiplicative errors:

Additive $y = \alpha + \beta x + \epsilon$

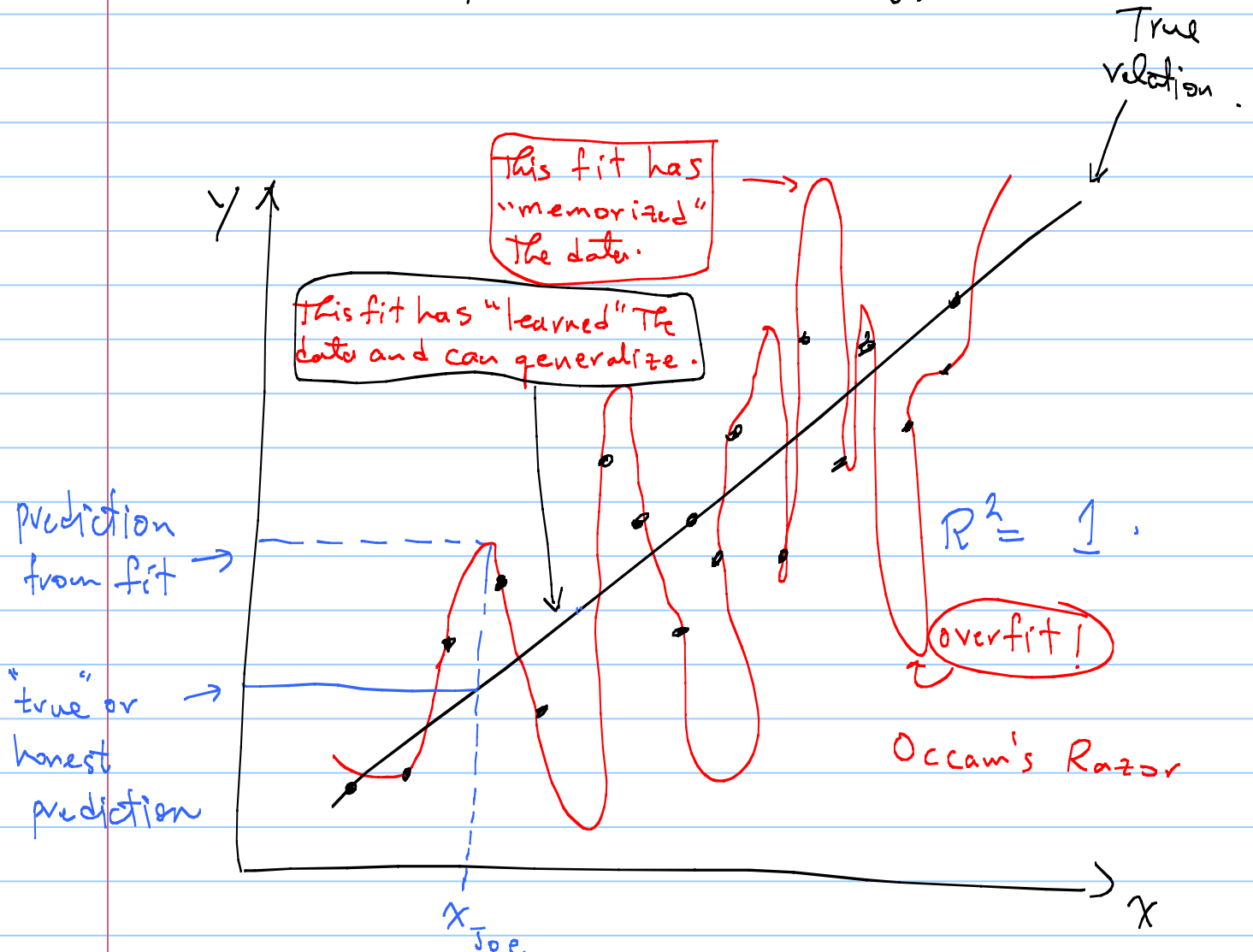
Mult. $y = \alpha \epsilon x^{\beta} \rightarrow \ln y = \alpha + \beta \ln x + \epsilon$

So a problem with multiplicative errors can be handled by doing linear regression on the log of all data.

Overfitting

Q = If we want a really "good fit" to the scatterplot why not just fit a really high-order polynomial?

A = Overfitting can lead to poor predictions on cases not present in the (training) data.



Moral: Don't overfit!

Q: How will you know if/when you have overfit? Hard Question!

A: Try testing your model on indep./new data.

Google "cross-validation" or "bootstrap"! Check Lab.

hw_lect12_1 By R

Read the data file transform_data.txt from the course website into R, and
a) make a scatterplot of y versus x.
b) transform x and/or y to linearize the relationship
c) Perform regression on the transformed data, and overlay the regression line on the scatterplot of the transformed data in part.

hw_lect12_2

The procedure for estimating the regression coefficients in polynomial regression is the same as before, i.e. by minimizing MSE w.r.t. $\alpha, \beta_1, \beta_2, \ldots$. Each derivative leads to a linear equation, and the system of equations can be uniquely solved to give $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \ldots$.

For this hw, consider a quadratic regression, and derive the linear equations that must be satisfied by $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2$. Write these equations in terms of the following means: $\bar{x}, \overline{x^2}, \overline{x^3}, \overline{x^4}, \overline{xy}, \overline{x^2y}, \bar{y}$

Do not solve the system of equations.

hw_lect12_3 By R

a) Read the data file bias_0_data.txt into R (it's on the course website), perform regression to predict y from x, make the scatterplot of the predictions versus the observed y values, and overlay a diagonal line (y-intercept=0, slope=1) on it. BUT, because we want to diagonal line to actually appear as diagonal, make sure the range of x and y values shown in the scatterplot is the same; in fact, set that range to (-6,6) for both x and y values. If you don't know how, check the prelabs, looking for xlim and ylim.

b) Now, read in the data file bias_1_data.txt, perform regression, and *overlay* on the previous plot (in part a) the scatterplot of predictions versus the observed y values. Make these points red. If done correctly, you will see that the predictions are now all positively biased (i.e., consistently shifted up).

c) The scatterplot in part a looks good in that it does not suggests any problems with the model. However, as discussed in class, the scatterplot in part b suggests a positive bias (the predictions are consistently higher than the observed values). Why? Hint: There is something about the data that is causing this bias. What is it?

hw_lect 12-4 | By R

In hw-A, you collected data which included data on 2 continuous variables. Call them x and y, depending on which variable you want to predict from the other. Now

a) Perform simple linear regression to estimate the regression coefficients, and interpret them.

b) Draw the regression line on the scatterplot of y vs. x

c) Make the residual plot of $(y-\hat{y})$ vs. $\hat{y}$
Interpret! Does it look "random" about x-axis?

d) Compute R^2, and interpret.

e) Compute s_e, and interpret.

f) Do you need to consider polynomial regression? or Transforming variables? If so, do it!

DO NOT DO This!

a) Read the data file sin_data.txt from the course website, and make a scatterplot of the y versus x.

b) The y values could be hourly temperature data at 100 different hours. In periodic situations like this the source of the periodic behavior is often known; for example, the 24-hour daily cycle. In fact, if you look carefuly, you will see a 24-hr period (i.e., the x distance from one peak to a neighboring peak). To confirm this, superimpose on the scatterplot in part a) a sine function with a period of 24, and an amplitude of 1, plotted at all integer x values from 1 to 100. Hint: the equation of the sine function is y = sin(2 pi / period). Don't worry if the sine function does not go "through" the data.

c) Take the difference between the y values of the data and the y values of the sine function; it doesn't matter which minus which. Then, make a scatterplot of the difference vsersus x.

d) Now you are ready to plot a line through the previous scatterplot, because if you've done things correctly, the periodic behavior will have disappeared by now. Find the equation of the OLS line, and overlay it on the previous scatterplot in part c.

e) report the R^2 and the s_e, and interpret both