Lecture 16 (Ch. 3)

Summary:

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: $\hat{y} = \hat{\alpha} + \hat{\beta} x$ OLS fit to data.

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

$\frac{SS_{exp}}{SST} = R^2$

$S_e = \sqrt{\frac{SS_{unexp}}{n-2}}$ = std. dev. of errors.

$\boxed{\text{Typo p. 121}}$

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

IMPORTANT

R^2 = Coefficient of determination

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

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}} \quad \text{or} \quad \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = 0.88916$$

Note $(.88916)^2 = \underline{0.79}$ (see R^2 in prev lecture)

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

<u>But only</u> in <u>simple linear regression</u>.

i.e. $y = \alpha + \beta x$.

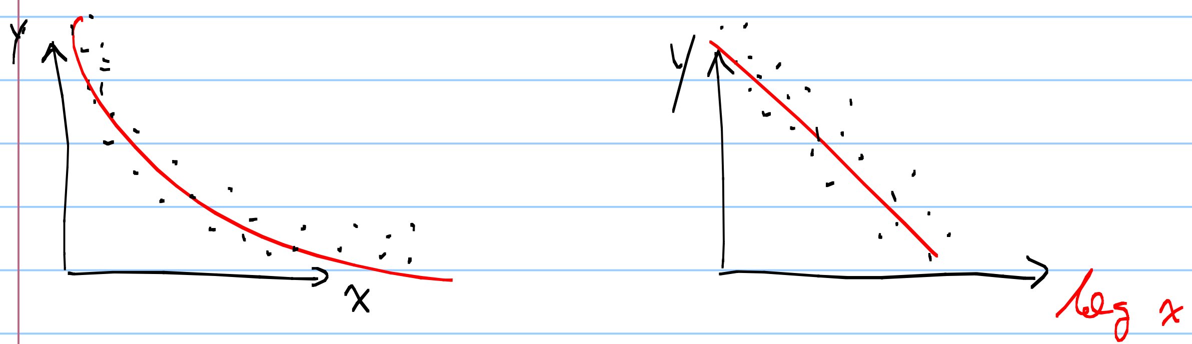
In everything else we will do next, $R^2 \neq (r)^2$.

Nonlinear relations

So far, we've considered situations where x & y are linearly related. If the relationship (in the scatterplot) is nonlinear, 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.

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.

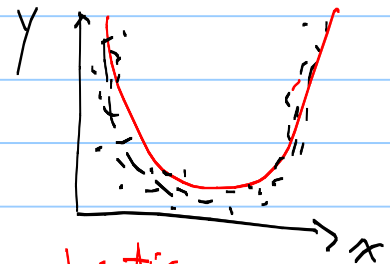
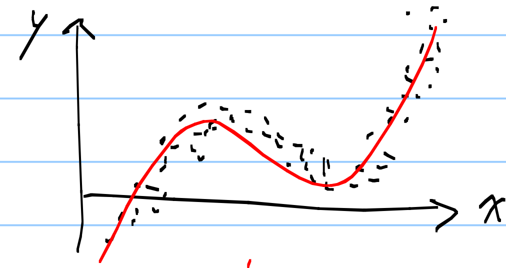
Q1: Suppose we have found that a scatterplot of $\sqrt{y}$ vs. $\sqrt{x}$ gives a linear pattern. So, we proceed to fit the model $\sqrt{y} = \alpha + \beta\sqrt{x}$. That model is equivalent to which of the following models?

A) $y \sim x$ $\boxed{\text{B) } y \sim x + \sqrt{x}}$ C) None of the above.

in R lingo

$\sqrt{y} = \alpha + \beta\sqrt{x} \Rightarrow y = (\alpha + \beta\sqrt{x})^2 = \alpha^2 + \beta^2 x + 2\alpha\beta\sqrt{x}$

2) If the relationship is not monotonic?

E.g.  or 

quadratic — cubic

A

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

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 un-explained, 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.

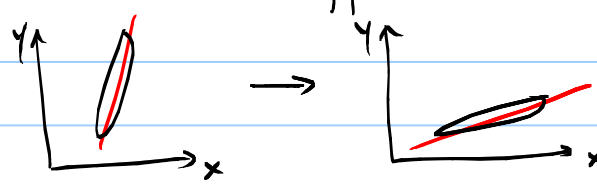
Important: Everything we've done is called linear regression even when we consider polynomial fits to the data.

e.g. $y = \alpha + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$

The reason is that "linear" refers to "linear in the parameters" (i.e. regression coefficients). This linearity is important because then the minimization of SSE leads to a system of linear equations that can be solved uniquely. But, it is also important to note that the linearity (of linear regression) does not prevent us from modeling complex/nonlinear relationships between x, y.

And, once again, r is a measure of association, the slope of the regression line is not. Think of a situation/scatterplot where r is large but slope is small.

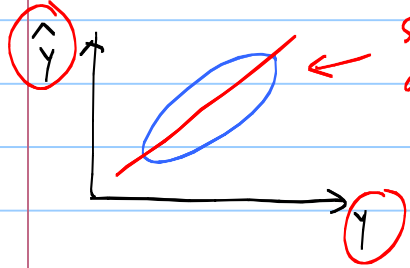
Or think about what happens if we switch $x \leftrightarrow y$:



slope changes but r does not.

Visual assessment

Plot $\hat{y}$ vs. y:

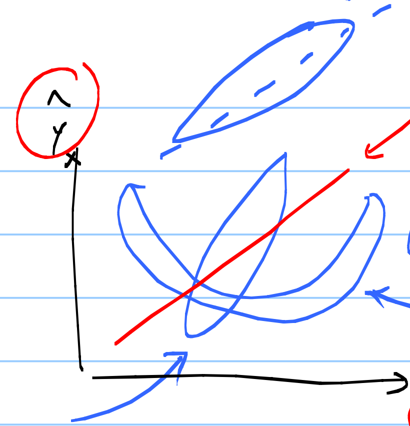


Random, symmetric about diagonal

$\boxed{\text{GOOD}}$

biased predictions (easy to fix)

diagonal



$\boxed{\text{BAD}}$

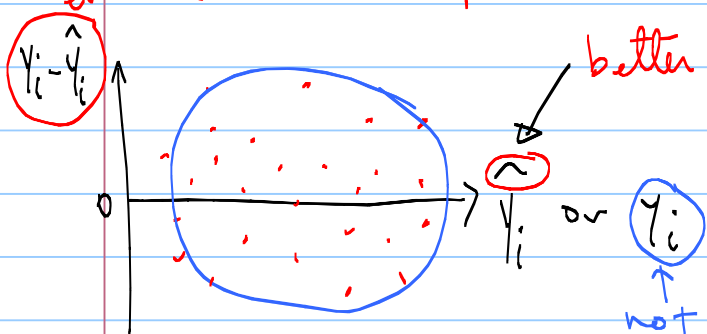
need a quadratic model.

under/over forecasting small/large y.

Residual Plot: (Mona Lisa)

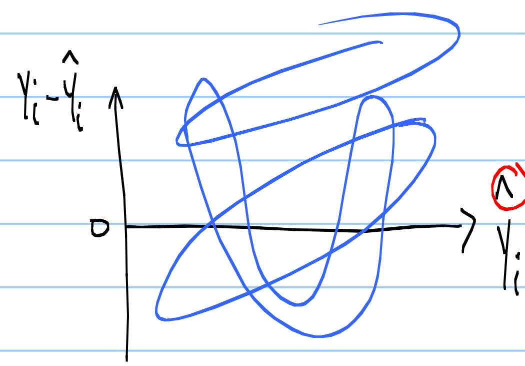
same

errors (last column of "Table" before)



Random, symmetric about error = 0 line. $\boxed{\text{GOOD}}$

$\therefore$ Nothing more to do



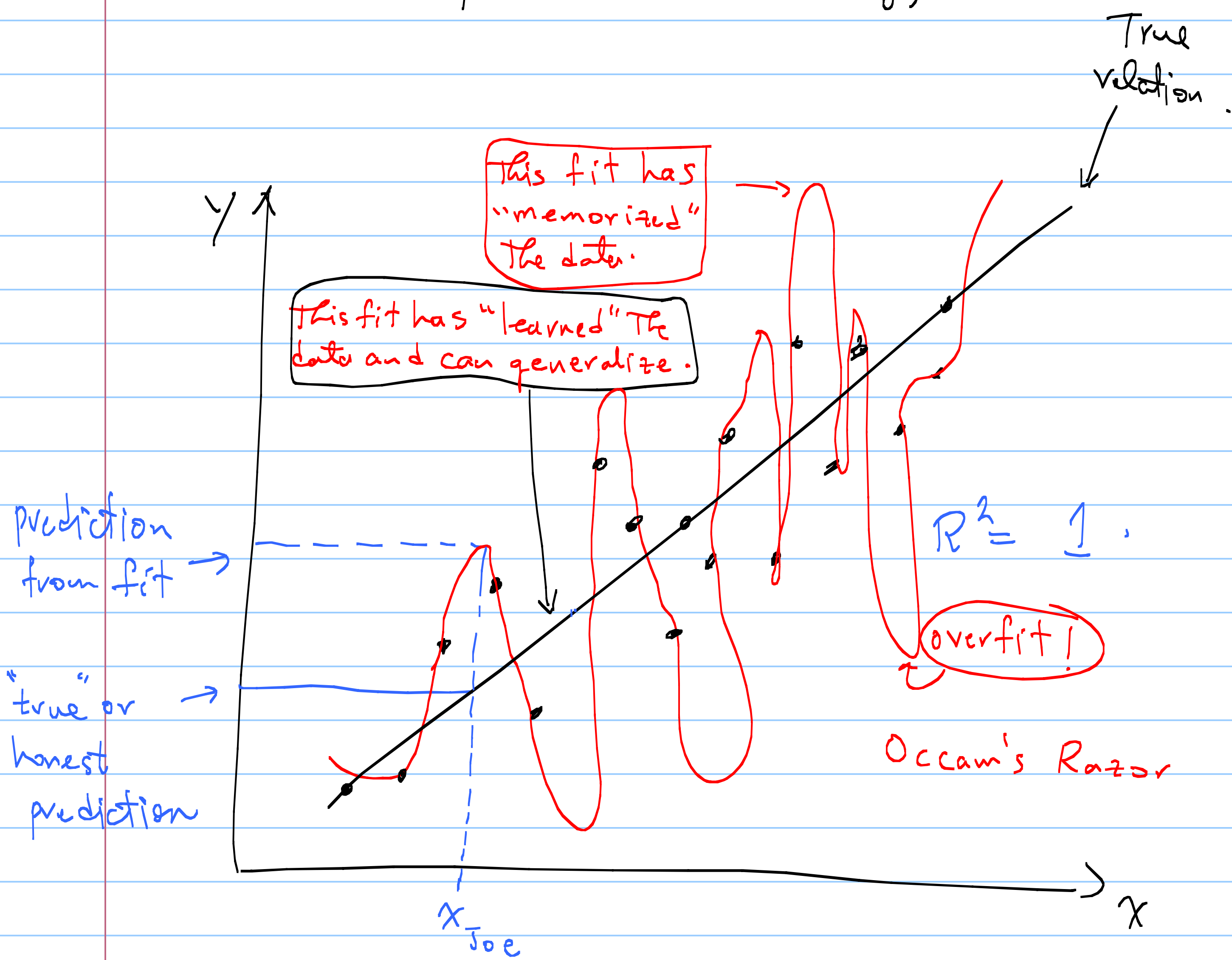
$\boxed{\text{BAD}}$

Things to do:

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

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!

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

→ Look at their scatter plot (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.

→ Do NOT Overfit!

→ Extrapolate Cautiously! Remember the -755 pound person! or "Life Before Earth"!

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.

hw_lect16-1 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_lect16-2 by R.

a) Read the data file transform_data.txt from the course website into R, and make a scatterplot of y versus x. Clearly, the relationship is nonlinear and monotonic. I can tell you that a good transformation that linearizes the relationship is to take the sqrt of both and x and y. Make a scatterplot of the transformed data.

b) Perform regression on the transformed data, and overlay the regression line on the scatterplot of the transformed data in part a).

c) Fit a regression model of the form y = alpha + beta_1 sqrt(x) + beta_2 x to the original (untransformed data).

d) In a clicker question I claimed that these two models are essentially equivalent. To check that, let's see if they make similar predictions. Make a scatterplot to compare their predictions. Just keep in mind that the second model predicts y, but the first model predicts sqrt(y).

hw_lect16-3 by R.

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 carefully, 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 hw_lect16-4

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: $\overline{x}, \overline{x^2}, \overline{x^3}, \overline{x^4}, \overline{xy}, \overline{x^2y}, \overline{y}$

Do not solve the system of equations.

---

hw_lect16-5 $\boxed{\text{By R}}$ $\boxed{\text{Optional}}$

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!