Lecture 13 (ch. 3)

Summary

We will add Things here, below

Given data (x_i, y_i), minimize $SSE = \sum_i (y_i - \hat{y}_i)^2 \Rightarrow \hat{y} = \hat{\alpha} + \hat{\beta} x + \ldots$

OLS estimates of α, β.

The ANOVA decomposition in regression:

$$SST = SS_{exp.} + SS_{unexp.}$$

$$\sum (y_i - \bar{y})^2 \qquad \sum (\hat{y}_i - \bar{y})^2 \qquad \sum (y_i - \hat{y}_i)^2 = SSE.$$

Interpretation: Regression "reduces" The variability in y from SST to SSE, by accounting for (filtering out) The variability due to x.

$R^2 = \dfrac{SS_{expl}}{SST}$ ~ percent variability in y due to (or explained by) x.

$s_e = \sqrt{\dfrac{SSE}{n-2}}$ ~ typical variability in y NOT due to x. (ie. typical error).

It may help to see ANOVA "By hand:" "By R": lm.1 = lm(y~x), anova(lm.1)

lm.1 = lm(y ~ x)

sum((predict(lm.1) - mean(y))^2) $= SS_{expl.} = SS$ model

sum((y - predict(lm.1))^2) $= SS_{unexp} = SS$ Residual

Q: What's the "best" number for predicting y?

- if we have data on variable y, only.

A: Sample mean of y, ie. $\bar{y}$ report $\bar{y} \pm s_y$

We will improve on These, later. in Ch. 11

- if we also have data on x, which is related to y.

A: The fitted value $\hat{y}(x) = \hat{\alpha} + \hat{\beta} x$ report $\hat{y}(x) \pm s_e$ $< s_y$

Here, The prediction depends on x

A Couple of more comments about ANOVA:

⇒ when we write $SST = SS_{expl} + SS_{unexpl}$,

these SS quantities are generally formatted in an ANOVA Table. Look at p.124 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:

$$S_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\underbrace{\frac{1}{n-2}\sum_i^n}_{\text{funny mean}} \underbrace{(y_i - \hat{y}_i)}_{\text{error}}^2}$$

Root ← squared

see examples in Lab.

⇒ Why is R^2 written as R^2 (or even r^2, as in our book)?

IMPORTANT

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

as in our/many books

To see why it is written as R^2 (or even r^2), consider our example from previous lecture:

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

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

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

But only in simple linear regression.

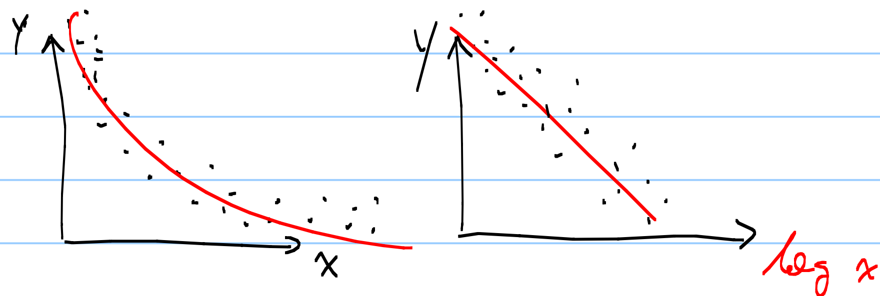
i.e. $y = \alpha + \beta x$.

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

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

1) If monotonic, then transform data:

For example, $x \to \log(x)$ often straightens scatterplots 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.

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

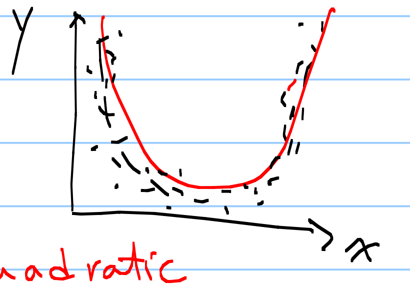
Usually, one of the following transformations straightens the scatterplot:

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

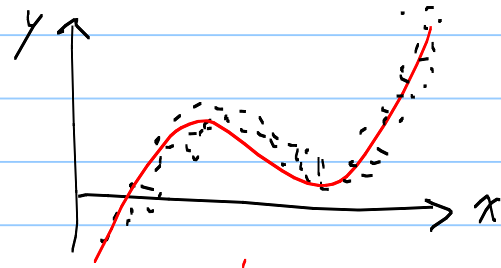
The best rule is to try different ones, and check the scatterplot.

2) If the relationship is not monotonic?

E.g.



or



A

quadratic

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

cubic

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

These are examples of polynomial regression.

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

for R reasons.

$\hat{\alpha}, \hat{\beta}_1$, etc. are obtained as before, by $\frac{\partial}{\partial \alpha}, \frac{\partial}{\partial \beta_1}, \frac{\partial}{\partial \beta_2}$ ...

See hw

As in simple linear regression, we can still decompose the total variability in y into explained and un-explained, .., R^2, s_e, ...

The only difference is that

$$SST = SS_{expl} + \overset{SSE}{\overset{\shortparallel}{SS_{unexp}}}$$

$R^2 \neq r^2$

$$s_e^2 = \frac{1}{n-(k+1)} \sum (y_i - \hat{y}_i)^2$$

of β's   The α.

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

But, these tools are dangerous (in the wrong hands).

In particular, look below, for "overfitting."

Summary

When you see data on (x, y)

→ Look at their scatter plot (and histograms, and ...)

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

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

→ 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!

→ And Do not overfit! It's bad for predictions (See below)

Learn how to manipulate eqns like These:

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

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

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

FYI

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.

$\boxed{\text{Some details of the above summary}}$

Given data (x_i, y_i) $i = 1, 2, \ldots, n$

assume $y = \alpha + \beta_1 x + \beta_2 x^2 + \cdots$ (or transform, $\sqrt{x}$, $\log y$, ...)

which means $y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \epsilon_i$

minimize $SSE = \sum_{i=1}^{n} \epsilon_i^2$

to get $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \ldots$ OLS estimates of $\alpha, \beta_1, \beta_2, \ldots$

predict: $\hat{y}(x) = \hat{\alpha} + \hat{\beta}_1 x + \hat{\beta}_2 x^2$ OLS fit to data.

ANOVA: $S_{yy} = SST = SS_{exp} + SS_{unexp}$

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

$S_e = \sqrt{\frac{SS_{unexp}}{n-(k+1)}}$ = std. dev. of errors.

~ goodness of fit.

~ typical error.

Notation:

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

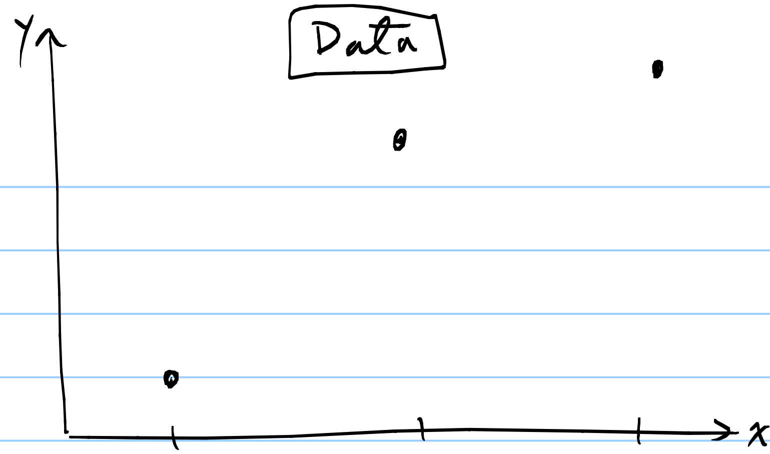
FYI

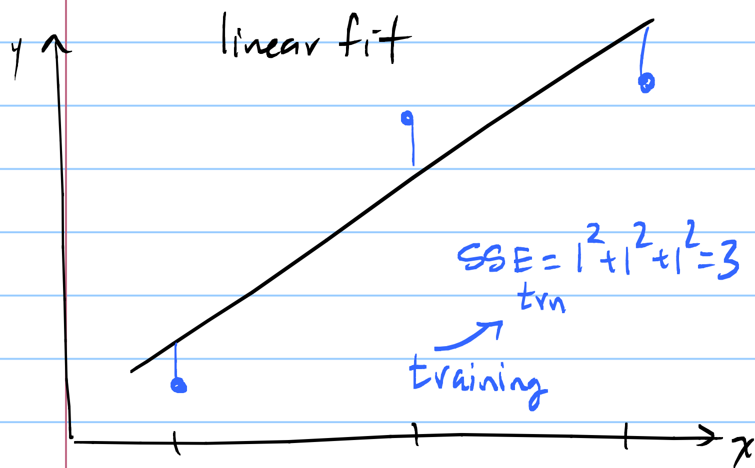
It may seem like the idea of minimizing SSE should be equivalent to maximizing SS_{exp}, because $SS_{exp} + SSE$ = Constant (SST). However, that argument is flawed because the decomposition $SS_{exp} + SSE = SST$ is itself true only when SSE is minimized. If SSE is not minimized, then there is a non-zero <u>cross-term</u>.

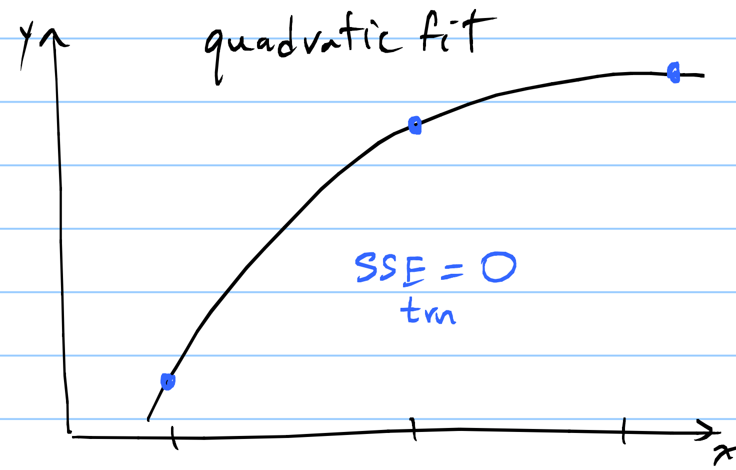
⇒ With the tools you have learned now, you can fit <u>any</u> data (x, y) regardless of how complicated (<u>nonlinear</u>) the relationship between x & y.

However, powerful tools can be dangerous

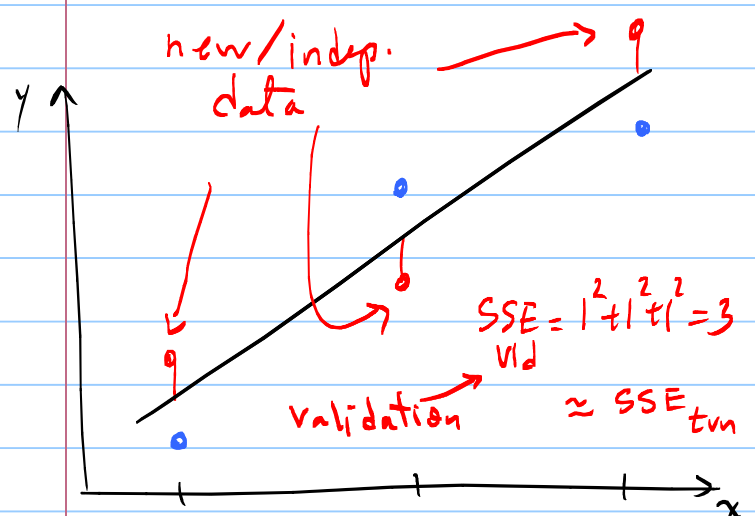
overfitting

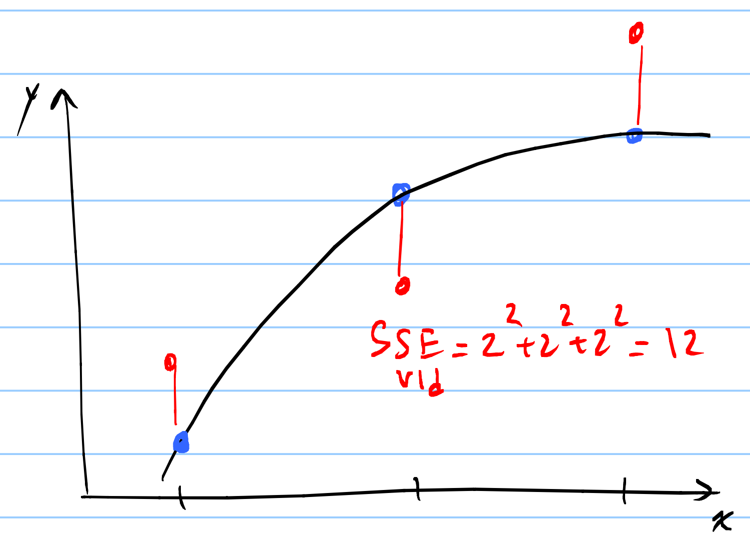






If the true relationship between x, y is linear, then new/independent data will be approximately random about the linear fit, e.g. red dots.





$\Downarrow$

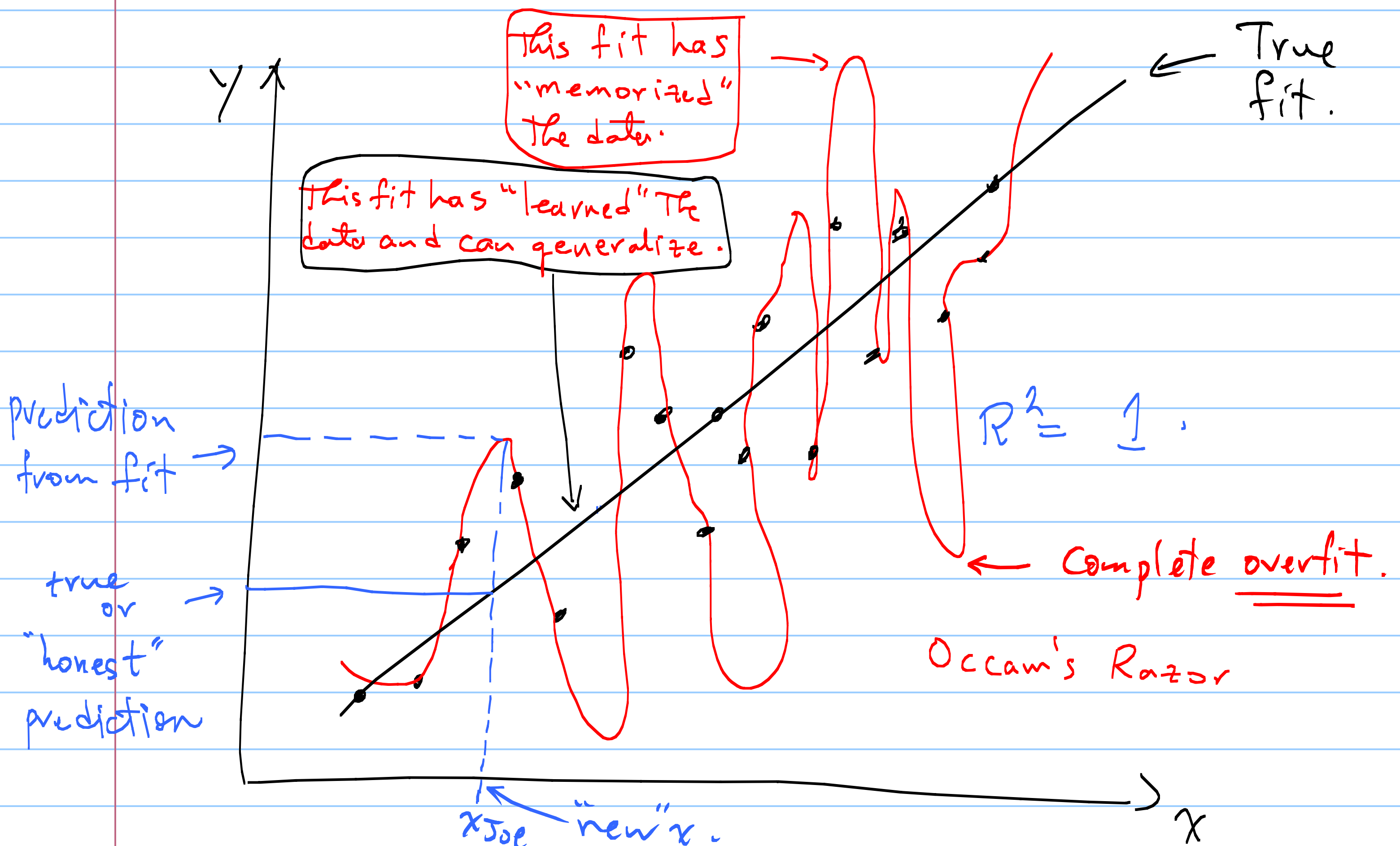
predictions are good

$\Downarrow$

predictions are poor.

i.e. it $\boxed{\text{overfit}}$ the trn set. $\Leftarrow$ The quadratic model "memorized" the training data ($SSE_{trn} = 0$), but did not learn/generalize.

The picture we usually see for overfitting is something like this:
But this is a bit wrong because it assumes a "new" case has a new x value. As you can see above, overfitting can occur even when the set of x values is fixed (across old and new data).



A: Overfitting will yield a model that does really well on the data used for developing the model (that data is called the training set) but it will do poorly on "new" data (called the test/validation set)
slightly different but nevermind

Overfitting is a "gray" thing. The above red curve shows complete overfitting. But lower levels of overfitting are still bad (in predictions). $\boxed{\text{Moral: Don't overfit!}}$

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_lect13_1

a) Read the data file transform_dat.txt from the course website into R, and
b) Make a scatterplot of y vs. x.
c) Transform x and/or y to linearize the relationship.
d) Perform regression on the transformed data, i.e., do (lm),
e) Overlay the corresponding line on the scatterplot
f) What percentage of the variability in the transformed y is explained by the transformed x, and what's the typical error in the prediction of the transformed y.

hw_lect13_2

The procedure for estimating the regression coefficients in polynomial regression is the same as before, i.e. by minimizing MSE with respect to alpha, beta_1, beta_2, ... Each derivative leads to a linear equation, and the system of equations can be uniquely solved to get alpha_hat, beta_1_hat, etc. For this hw, consider a quadratic regression,and derive the linear equations that must be satisfied by alpha_hat, beta_1_hat, etc. 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_lect13_3 (By R)

Return to the data you collected at the start of the quarter. Call the 2 continuous variables x and y, depending on which variable you want to predict from the other, and
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) Compute R^2 and interpret it
d) Compute s_e and interpret it
e) Do you need to consider polynomial regression? Or transforming variables? If so, do it!