Regression: $y_i = \alpha + \beta x_i + \varepsilon_i$

minimize $\text{SSE} = \sum (y_i - \hat{y}_i)^2 \Rightarrow \hat{y} = \hat{\alpha} + \hat{\beta} x \quad \hat{\alpha}, \hat{\beta} = \text{OLS}$

Analysis of Variance (ANOVA):

$\text{SST} = \text{SSexp.} + \text{SSunexp.}$

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

Variability in $y$ is reduced to, after $x$ is accounted for.

Regression is a method of error reduction.

E.g. The variability in Table length ($\pm 10$ cm) can be reduced to, say $\pm 3$, after we take into account that Table length, $y$, is related to temperature $x$, as $y = \alpha + \beta x + \varepsilon$.

SSexp, SSunexp are found from the above equations, i.e.

$$\text{lm.1} = \text{lm} (y \sim x)$$

$$\text{SSmod} = \sum \left(\text{predict (lm.1)} - \text{mean (y)}\right)^2 \quad \text{by hand}$$

$$\text{SSexp} = \text{SSmod}$$

$$\text{SSunexp} = \sum \left(\text{predict (lm.1)} - y\right)^2 \quad \text{by hand}$$

or by anova(lm.1) (by R)

$$\text{SSexp}/\text{SST} = R^2 = \text{percent of var. in } y \text{ explained by } x. \quad \text{(goodness-of-fit)}$$

$$\sqrt{\frac{\text{SSunexp}}{n-2}} = s_e = \text{typical error made by regression.} \quad \text{by R.}$$

$\hat{y}(x) \pm s_e$
Picture for the ANOVA decomposition:

So, when there is a (linear) relationship between $x$ & $y$, then some portion of the variation in $y$ can be attributed to (or explained by) $x$. That portion is $SS_{\text{exp.}}$, and the (unexplained) rest is $SS_{\text{unexp.}} = SSE$.

So the variability in $y$, $SST$, is reduced to $SSE$.

When there is no relationship between $x$ and $y$, then the fig looks like below. Note that this situation is equivalent to the situation where we have data only on $y$, but not on $x$ at all.

In that case, the best prediction for every case is $\bar{y}$ (see below):
Example (same as in last few lectures):

\[ SST = \sum (y_i - \bar{y})^2 = \ldots = 6251.2 \]

\[ SSE = \sum (y_i - \hat{y}_i)^2 = \quad \text{last column in table in prev. lecture} \]
\[ = (-1.5)^2 + (5.1)^2 + (11.5)^2 + (-30.3)^2 + (5.1)^2 = 1307 \]

\[ R^2 = \text{Coef. of det.} = \frac{SST - SSE}{SST} = \frac{6251.2 - 1307}{6251.2} = 0.79. \]

Conclusion: 79% of the variability in \( y \) (weight, table's length) is due to the linear relation with \( x \) (height or temperature).

The other piece of the decomposition:

\[ S_e = \sqrt{\frac{1307}{5 - 2}} = 20.9 \text{ pounds} \]

Conclusion: The typical deviation of the \( y \) values (weight/height or Table's length) (i.e. error or residual) about the fit is about 20.9 pounds.

Report weight (or Table length): \( \hat{y} \pm 20.9 \) with \( R^2 = 0.79 \)

or ICP

or \( \ldots -755 + 13.3X \) (or height or EV or ...)
If there is no $x$ data, then the OLS prediction $\hat{y}$ is just $\bar{y}$.

I.e.,

$$\hat{y} = \bar{y} \quad \text{⇒ \ no \ } x,$$

**What are the $R^2$ and $S_e$?**

$$S_{\text{e}}^2 = \frac{\text{SSE}}{n-2} = \frac{\sum (y_i - \hat{y})^2}{n-2} = \frac{\sum (y_i - \bar{y})^2}{n-2} = \frac{(n-1) S_y^2}{n-2} \Rightarrow S_e \sim S_y$$

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum (y_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2} = 0 \Rightarrow R^2 = 0.$$

So, if we use $\bar{y}$ as our prediction, then $R^2 = 0$ (Bad), and $S_e \sim S_y$, i.e. the typical error ~ typical dev. in $y$, i.e. nothing gained.

Another situation when nothing is gained is if we make random predictions, e.g. $\hat{y}_i$ = random. Suppose the mean and the var. of these random predictions are the same as those of observations, i.e. $\hat{y}_i = \text{random with} \ \bar{y} = \bar{y}, \ S_{\hat{y}} = S_y$. The picture is

But now, something strange happens:

Although one can use the formula for $R^2$ to arrive at a number, that number does not have the usual interpretation (i.e. percentage of var. in $y$, explained by $x$), because $\hat{y}_i$ = random are not OLS predictions. So, we don't have the ANOVA decomposition at all. Same objection applies to $S_e$.

Again, the ANOVA decomposition is correct only for OLS $\hat{y}$; $\hat{y}$ = random are not OLS predictions.

These both have equal/comparable $S_{\hat{y}}$ (i.e. $R^2$).

But the blue one has lower $S_e$.

This doesn't contradict ANOVA, because the red $\hat{y}$ is not OLS.

In short, both have equal precisions, but blue is more accurate.
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 $S_e$:

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

Root mean squared error

See examples in Lab.

---

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

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

Important notation: $\hat{y}$ symbol, as in our / many books

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

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}(\sum (y_i - \bar{y})^2)}$$

$$\text{Note } (0.88916)^2 = 0.79 \text{ (See } R^2 \text{ in prev lecture)}$$

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

But only in simple linear regression, i.e. $y = ax + bx$.

In everything else we will do next, $R^2 \neq (r)^2$. 
So far, we've considered situations where $x$ and $y$ 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, $x \rightarrow \log(x)$ often straightens scatter plots that look like this.

\[ y \quad \xrightarrow{\text{Log}} \quad \log x \]

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

I.e. $y = \alpha + \beta (\log x) \quad \text{not} \quad y = \alpha + \beta x$

$\rightarrow$ and decompose (i.e. ANOVA) as before.

\[ \begin{align*} 
  \text{SST} &= \text{SSExp} + \text{SSUnexp} \\
  &\quad \text{where} \quad x \quad \text{is replaced by} \quad \log(x). \\
  &\quad \downarrow \quad \text{by log}(x) 
\end{align*} \]

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

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

The best rule is to try different ones, and check the scatterplot.
2) If the relationship is not monotonic?

\[ \hat{y} = \alpha + \beta_1 x + \beta_2 x^2 \quad \hat{y} = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \]

These are examples of polynomial regression.

\[ \text{lm}(y \sim x + I(x^2) + I(x^3) + \ldots) \]

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, \( \ldots \).

The only difference is that \( R^2 \neq (\hat{R})^2 \)

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

Q1: Suppose \( n = 3 \), and we fit \( y = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \). Then \( R^2 \) will be

\[ R^2 = \frac{SS_{\text{exp}}}{SS_{\text{T}}} = \frac{SS_{\text{T}} - SSE}{SS_{\text{T}}} = 1 - \frac{SSE}{SS_{\text{T}}} = 1 - 0 = 1 \]

a) 0  b) \frac{1}{2}  c) 1  d) depends on the data.
Q: If we want a really "good fit" to the scatter plot, why not just fit a really high-order polynomial?

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

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

A: Hard Question!

Try testing your model on independent/new data. Google "cross-validation" or "bootstrap".
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 ($R^2$, Se, 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:

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + ...$$
Assess performance with ANOVA.

→ Extrapolate cautiously! Remember the -755 pound person!

→ And Do not overfit! It's bad for predictions.

Also, recall how to manipulate eqns like these:
E.g. $y = \alpha + \beta \ln x$

\[
\begin{align*}
(y - \alpha)/\beta &= \ln x \\
\Rightarrow \quad y &= \ln e^x + \ln \beta = \ln(e^x \beta) = e \cdot e^x \beta
\end{align*}
\]

A additive/multiplicative errors:

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

Mult. $y = \alpha \varepsilon^x \beta$ $\Rightarrow$ $\ln y = \alpha + \beta \ln x + \varepsilon$

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

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) vs. y**  
Interpret! Does it look "random" about x-axis?  
````
d) Compute R^2, and interpret.
e) Compute se, and interpret.
f) Do you need to consider polynomial regression? Or transforming variables? If so, do it!

**Slide This for now.**
a) Read the data file sin_data.txt from the course website, and make a scatter plot 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 scatter plot 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 $y = \sin(2\pi \text{ period})$. 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 scatter plot of the difference vs. x.

d) Now you are ready to plot a line through the previous scatter plot, 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 scatter plot in part c.

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