Analysis of variance approach to regression

If x is useless, i.e. $\beta_1 = 0$, then $E(Y_i) = \beta_0$. In this case β_0 is estimated by $\bar{Y}$. The i*th* deviation about this *grand* mean can be written:

$$\overbrace{Y_i - \bar{Y}}^{\text{deviation about } \textit{grand} \text{ mean}} = \overbrace{\hat{Y}_i - \bar{Y}}^{\text{explained by model}} + \overbrace{Y_i - \hat{Y}_i}^{\text{slop left over}}$$

Our regression uses line explains how Y varies with x. We are interested in *how much* variability in the $Y_1, \ldots, Y_n$ is soaked up by the regression model.

This can be represented mathematically by *partitioning* the *total sum of squares* (SSTO):

$$SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

SSTO is a measure of the total (sample) variation of Y ignoring x.

Note: SSTO $= (n-1)S_Y^2$.

The sum of squares *explained by the regression line* is given by

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2.$$

The sum of squared errors measures how much Y *varies around the regression line*

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

It happily turns out that

$$SSR + SSE = SSTO.$$

An aside...

$$\text{Let } \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \hat{\mathbf{Y}} = \begin{bmatrix} b_0 + b_1 x_1 \\ b_0 + b_1 x_2 \\ \vdots \\ b_0 + b_1 x_n \end{bmatrix}, \text{ and } \mathbf{1}_n = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

Certainly this is always true:

$$(\hat{\mathbf{Y}} - \mathbf{1}_n \bar{Y}) + (\mathbf{Y} - \hat{\mathbf{Y}}) = (\mathbf{Y} - \mathbf{1}_n \bar{Y}).$$

This is just a vector version of the first expression on the first slide.

The pythagorean theorem gives us the decomposition of the total sum of squares. Since $\hat{\mathbf{Y}} - \mathbf{1}_n\bar{Y}$ is *orthogonal* to $\mathbf{Y} - \hat{\mathbf{Y}}$, (shown later), we have

$$||\hat{\mathbf{Y}} - \mathbf{1}_n\bar{Y}||^2 + ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = ||\mathbf{Y} - \mathbf{1}_n\bar{Y}||^2,$$

that is SSR+SSE=SSTO.

It is worthwhile to verify SSR=$||\hat{\mathbf{Y}} - \mathbf{1}_n\bar{Y}||^2$, etc. Recall that for any vector $\mathbf{v} = (v_1, \ldots, v_n)'$ the length of the vector is $||\mathbf{v}|| = \sqrt{a_1^2 + \cdots + a_n^2}$ (thank's again, Pythagorus).

(end of aside)

Restated: The variation in the data (SSTO) can be divided into two parts: the part explained by the model (SSR), and the slop that's left over, i.e. unexplained variability (SSE).

Associated with each sum of squares are their degrees of freedom (df) and mean squares, forming a nice table:

Source	SS	df	MS	$E(MS)$
Regression	SSR$= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	1	$\frac{SSR}{1}$	$\sigma^2 + \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2$
Error	SSE$= \sum_{i=1}^n (Y_i - \hat{Y})^2$	$n-2$	$\frac{SSE}{n-2}$	σ^2
Total	SSTO$= \sum_{i=1}^n (Y_i - \bar{Y})^2$	$n-1$		

Note: $E(MSR) > E(MSE) \Leftrightarrow \beta_1 \neq 0$. Loosely, we expect MSR to be larger than MSE when $\beta_1 \neq 0$.

So testing whether the simple linear regression model explains a significant amount of the variation in Y is equivalent to testing $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$.

Consider the *ratio* MSR/MSE. If $H_0 : \beta_1 = 0$ is true, then this should be near one. In fact

$$F^* = \frac{MSR}{MSE} \sim F_{1,n-2} \text{ when } H_0 : \beta_1 = 0 \text{ is true.}$$

So $E(F^*) = (n-2)/(n-4)$ which goes to one as $n \to \infty$ (when $\beta_1 = 0$).

This leads to an F-test of $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ using $F^* = MSR/MSE$:

If $F^* > F_{1,n-2}(1 - \alpha/2)$ then reject $H_0 :\ beta_1 = 0$ at significance level α.

Note: $F^* = (t^*)^2$ and that the F-test is completely equivalent (is exactly the same) to the Wald t-test based on $t^* = b_1/se(b_1)$ for $H_0 : \beta_1$.

SAS Example...

General linear test

Note that if $H_0 : \beta_1 = 0$ holds our *reduced model* is

$$Y_i = \beta_0 + \epsilon_i.$$

It can be show that the least-squares estimate of β_0 in this reduced model is $\hat{\beta}_0 = \bar{Y}$.

Thus SSE for the reduced model is

$$SSE(R) = \sum_{i=1}^{n} (Y_i - \bar{Y})^2,$$

which is the SSTO from the *full model*.

Note that the SSE(R) can never be less than the SSE(F), the sum of squared errors from the full model. Including a predictor can *never explain less variation* in Y, only as much or more. So...

$$SSE(R) \geq SSE(F).$$

If SSE(R) is only a little more than SSE(F), the predictor is not helping much (and so the reduced model may be adequate).

We can generally test this with an F-test:

$$F^* = \frac{\left[\frac{SSE(R)-SSE(F)}{df_R - df_F}\right]}{\left[\frac{SSE(F)}{df_F}\right]},$$

and reject H_0 : *reduced model holds* if $F^* > F_{df_R - df_F, df_F}(1 - \alpha/2)$. This idea/test will be used often in complex regression models with multiple predictors. “Full model / reduced model ”

R^2 and r

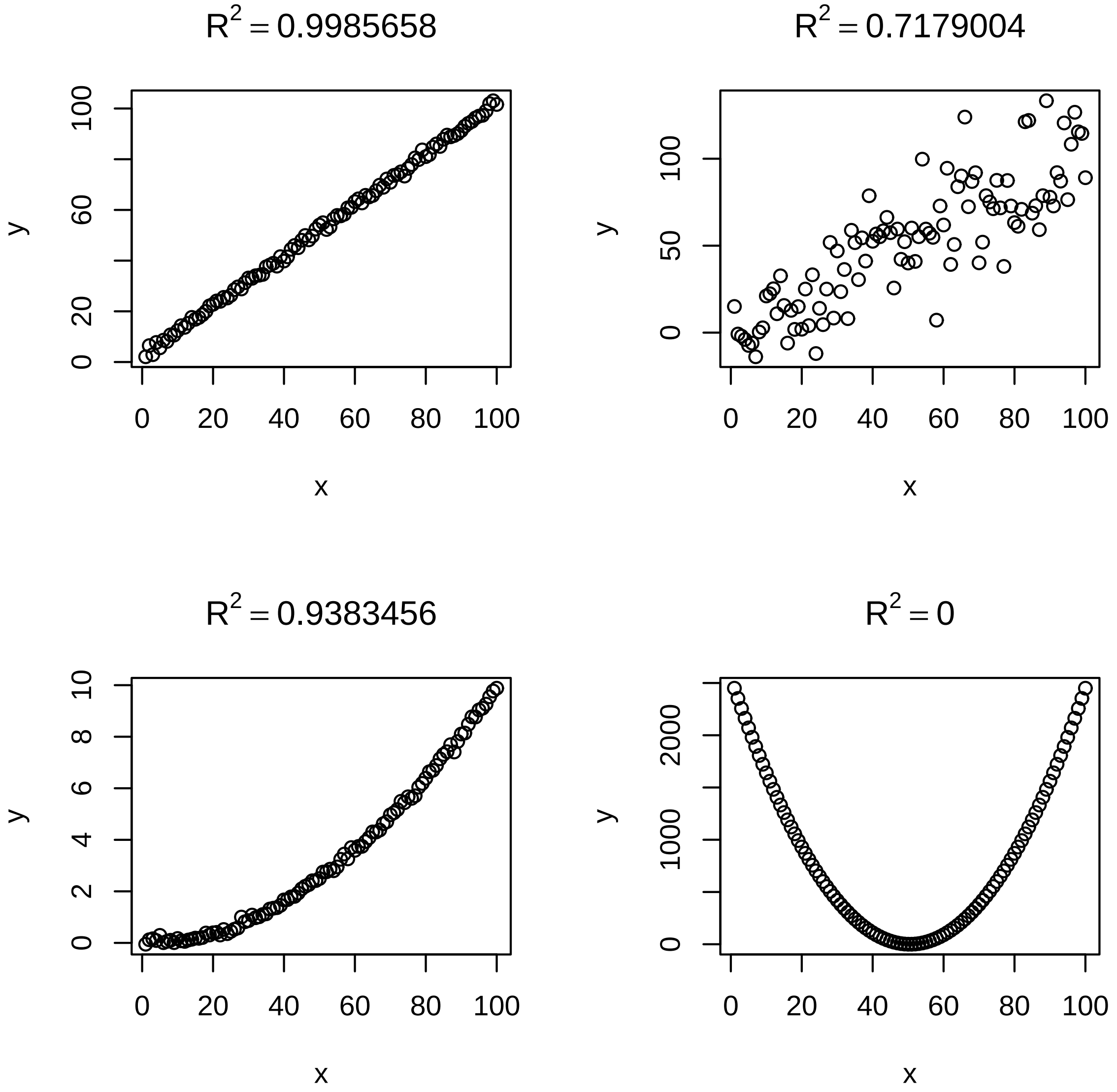
The *coefficient of determination* is

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO},$$

the proportion of total sample variation in Y that is explained by its linear relationship with x. Note:

- $0 \le R^2 \le 1$.
- $R^2 = 1 \Rightarrow$ data perfectly linear.
- $R^2 = 0 \Rightarrow$ regression line horizontal ($b_1 = 0$).

The closer R^2 is to one, the greater the linear relationship between x and Y.


$R^2 = 0.9985658$
$R^2 = 0.7179004$
$R^2 = 0.9383456$
$R^2 = 0
y
x
0
20
40
60
80
100

Note: Let

$$r = \text{corr}(\mathbf{x}, \mathbf{Y}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

be the sample correlation between x and Y. Then $R^2 = r^2$. So $\sqrt{R} = \sqrt{SSR/SSTO}$ is equal to $|r|$.

Note: $b_1 > 0 \Leftrightarrow r > 0$ and $b_1 < 0 \Leftrightarrow r < 0$. So $r = \sqrt{R^2}(b_1/|b_1|)$.

As usual:

- r near $0 \Rightarrow$ little linear association between x and Y
- r near $1 \Rightarrow$ strong positive, linear association between x and Y
- r near $-1 \Rightarrow$ strong negative, linear association between x and Y

Cautions about R^2 and r

- R^2 could be close to one, but the $E(Y_i)$ may not lay on a line. (Why? Which plot?)
- R^2 may not be close to one, but a line is best for $E(Y_i)$ (Why? Which plot?)
- R^2 could be essentially zero, but x and Y could be highly related. (Why? Which plot?)

  Example: Toluca data...

Correlation models

In the regression model

- The x values are assumed to be known constants, and
- We generally want to predict Y from x.

If we simply have two continuous variables X and Y without neither being a natural response/predictor, a correlation model can be used.

Example: For the Toluca data, say we are interested in simply determining whether lot size and work hours are linearly related.

If appropriate, we could assume that X and Y have a bivariate normal normal distribution with parameters μ_x, μ_y, σ_x, σ_y, and ρ. Then

$$\left[\begin{array}{c} X_i \\ Y_i \end{array}\right] \sim N_2\left(\left[\begin{array}{c} \mu_x \\ \mu_y \end{array}\right], \left[\begin{array}{cc} \sigma_x^2 & \sigma_x\sigma_y\rho \\ \sigma_x\sigma_y\rho & \sigma_y^2 \end{array}\right]\right).$$

Investigation of linear association between X and Y is done through inferences about $\rho = \mathrm{corr}(X_i, Y_i)$. A point estimator of ρ is r as defined a few slides back, the sample correlation.

r is the MLE under normality, but also a consistent moment-based estimator in general (not assuming normality).

Testing $H_0 = 0$ is equivalent to testing $H_0 : \beta_1 = 0$ in the regression of Y on x.

A large-sample CI for ρ uses Fishers z-transformation:

$$z' = 0.5 \log \left(\frac{1+r}{1-r} \right).$$

In large samples, a $(1-\alpha)100\%$ CI for $\zeta = E(z')$ is

$$z' \pm z(1-\alpha/2)\sqrt{1/(n-3)}.$$

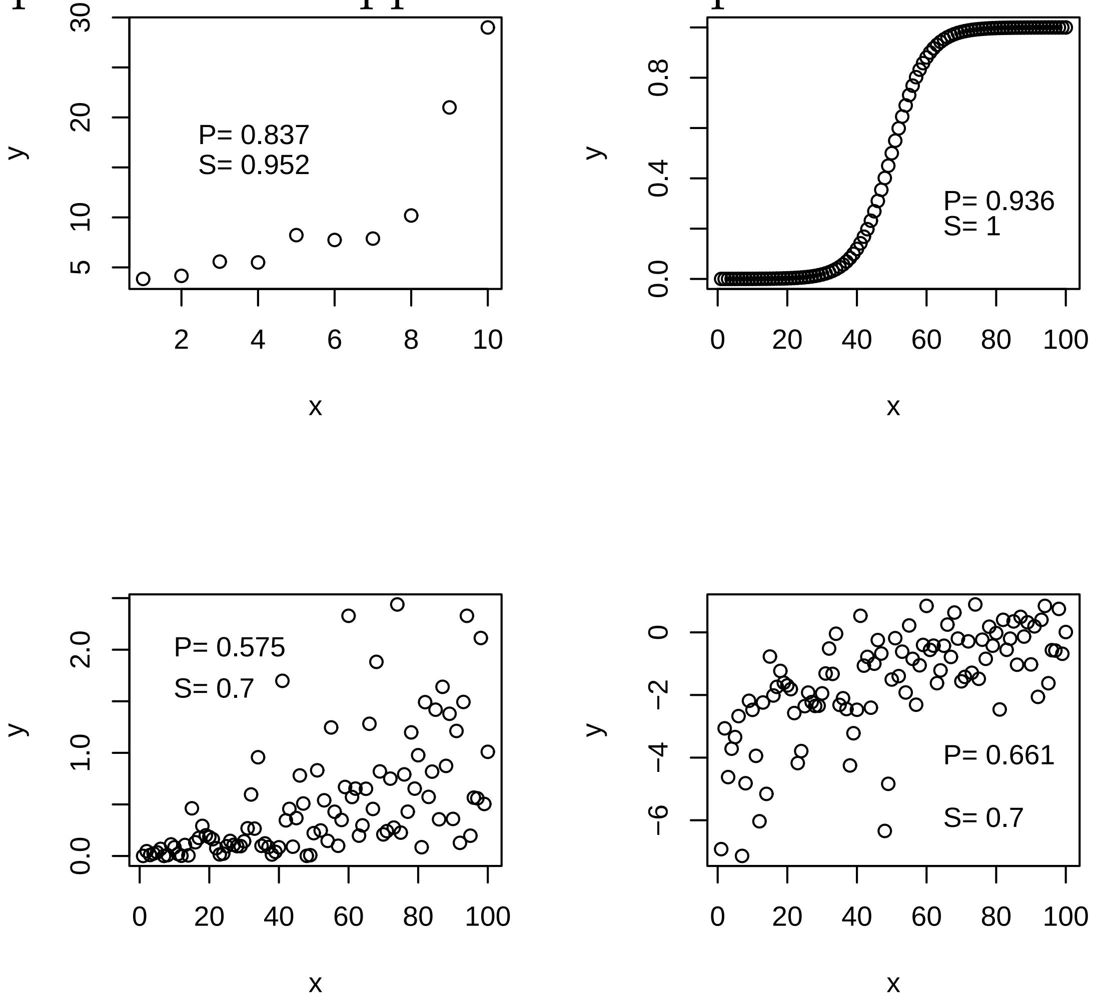
Then use Table B.8 in the back of your text to back-transform endpoints to get a CI for ρ.

Here, $z(1-\alpha/2) = \Phi^{-1}(1-\alpha/2)$.

The **Spearman** rank correlation coefficient replaces the X values with their ranks, replaces the Y values with their ranks, then carries out a (standard Pearson, described in last slide) correlation analysis on the ranks.

Example: Toluca data in R.

Pearson (P) versus Spearman (S). Last plot takes log of each Y in 2nd to last plot. What happens to the Spearman correlation?



Cautions about regression

- When predicting *future values*, the conditions affecting Y and x should remain similar for the prediction to be trustworthy.
- Beware of extrapolation: predicting Y_h for x_h far outside the range of x in the data. The relationship may not hold outside of the observed x-values.
- Concluding that x and Y are linearly related (that $beta_1 \neq 0$) does not imply a cause and effect relationship between x and Y.
- Beware of making multiple predictions or inferences simultaneously unless using an appropriate procedure (e.g. Scheffe's method). One needs to consider both the individual Type I error and the "family error rate."

- The least squares estimates are *not unbiased* if x is measured with error – in fact coefficients are biased towards zero. Slightly more advanced techniques are needed (see Section 4.2, p. 172).
- We have not discussed model checking and diagnostics. These will come next when we start adding more predictors to the model. For simple linear regression, in *most cases* a scatterplot tells us all we need to know about (i) linear mean and (ii) homoscedastic (constant variance) errors. (iii) A QQ plot to assess normality can be examined for the residuals $e_1, \ldots, e_n$.