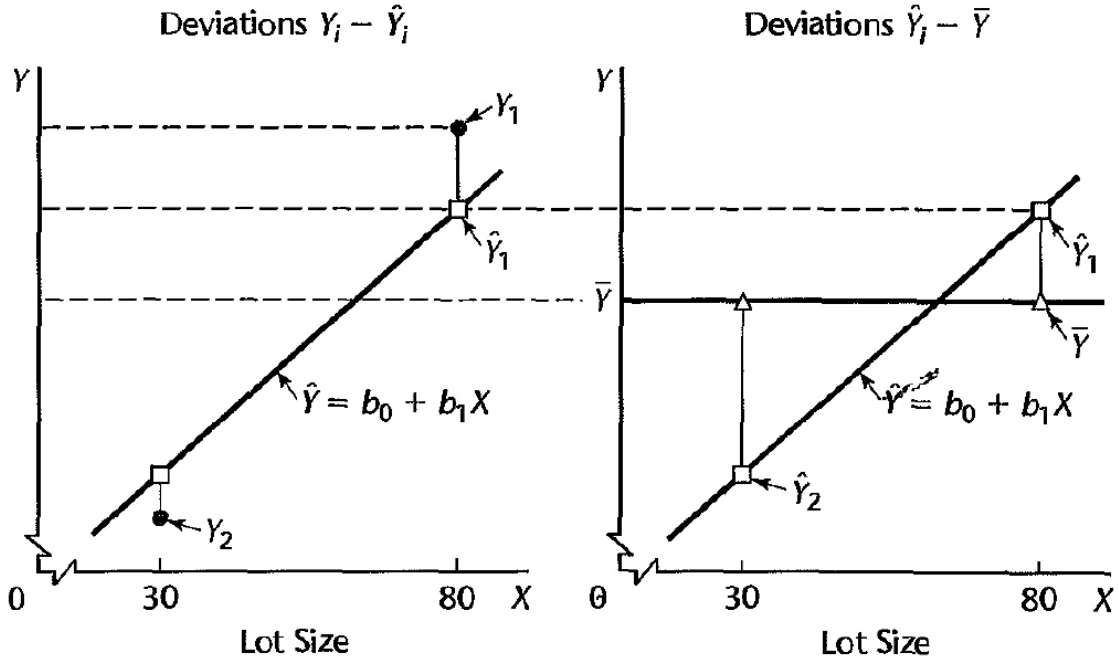
# Chapter 2

Yen-Yi Ho

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

# 2.7 Analysis of variance approach to regression (pp. 63–72)




The total deviation can be partitioned into two parts:

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

Our regression 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.

# Partitioning the SSTO

Two sources of variability (model & model error) go into the *total sum of squares* (SSTO):

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

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

$$SSR + SSE = SSTO.$$

# Analysis of variance (ANOVA) table

$$\hat{Y} = \beta_0 + \beta_1 X_i$$

**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$= \sum_{i=1}^{n}(Y_i - \hat{Y})^2$ | $n-2$ | $\frac{SSE}{n-2}$ | $\sigma^2$ |
| Total | SSTO$= \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ | $n-1$ | | |

# Another test of $H_0 : \beta_1 = 0$

- **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 $H_a : \beta_1 \neq 0$ (or more properly: $H_0 : \beta_1^2 = 0$ versus $H_a : \beta_1^2 > 0$)
- Consider the *ratio* $MSR/MSE$. If $H_0 : \beta_1 = 0$ is true, then this should be near one. In fact

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

Refer to the text's discussion of Cochran's Theorem (pp. 69-70) for intuition on why $SSR/\sigma^2$ and $SSE/\sigma^2$ have $\chi^2$ distributions.

# F-test in ANOVA table

This leads to an F-test of $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ using $F^* = MSR/MSE$:
If $F^* > F_{1,n-2}(1-\alpha)$ then reject $H_0 : \beta_1 = 0$ at significance level $\alpha$.
**Note**: $F^* = (t^*)^2$ and so the F-test is completely equivalent to the Wald t-test based on $t^* = b_1/se(b_1)$ for $H_0 : \beta_1 = 0$.

Toluca data:

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 252378 | 252378 | 105.88 | <.0001 |
| Error | 23 | 54825 | 2383.71562 | | |
| Corrected Total | 24 | 307203 | | | |

# $F^* = (t^*)^2$ in simple linear regression (page 69)

$H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$

$$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n}[b_1(X_i - \overline{X})]^2 = b_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$F^* = \frac{MSR}{MSE} = \frac{b_1^2}{\frac{MSE}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$$

$$t^* = \frac{b_1}{\sqrt{\frac{MSE}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}}$$

Notice that

$$E(SSR) = E(b_1^2) \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2$$

# 2.8 General linear test (pp. 72–73)

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 $\beta_0$ 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*.

# Form of test statistic and $F$-distribution

**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 $F^* > F_{df_R-df_F,df_F}(1-\alpha)$.
This idea/test will be used often in complex regression models with multiple predictors.

# 2.9 $R^2$ and $r$ (pp. 74–77)

The *coefficient of determination* is

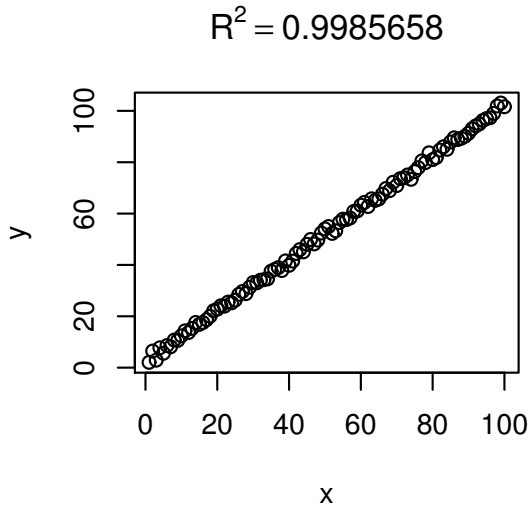
$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO},$$

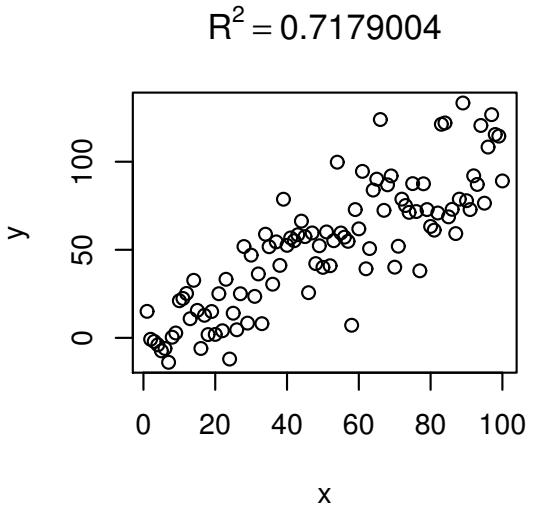
the proportion of total sample variation in $Y$ that is explained by its linear relationship with $X$. Note:

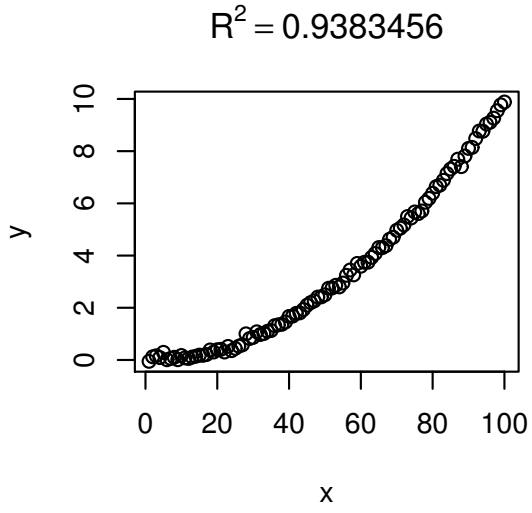
- $0 \leq R^2 \leq 1$.
- $R^2 = 1 \Rightarrow$ data perfectly linear.
- $R^2 = 0 \Rightarrow$ regression line horizontal ($b_1 = 0$).

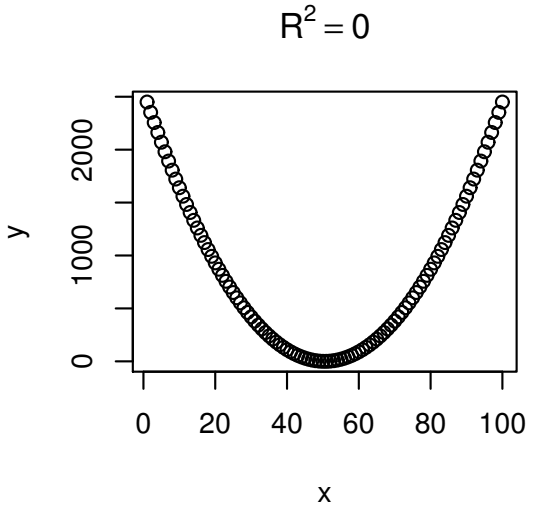
The closer $R^2$ is to one, the greater the linear relationship between $X$ and $Y$.

# $R^2$ for different data sets



# Sample correlation $r$ in a simple linear regression

**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 $r = \sqrt{R^2} \times \text{sign}(b_1)$.
**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?)

# Cautions about $R^2$ and $r$

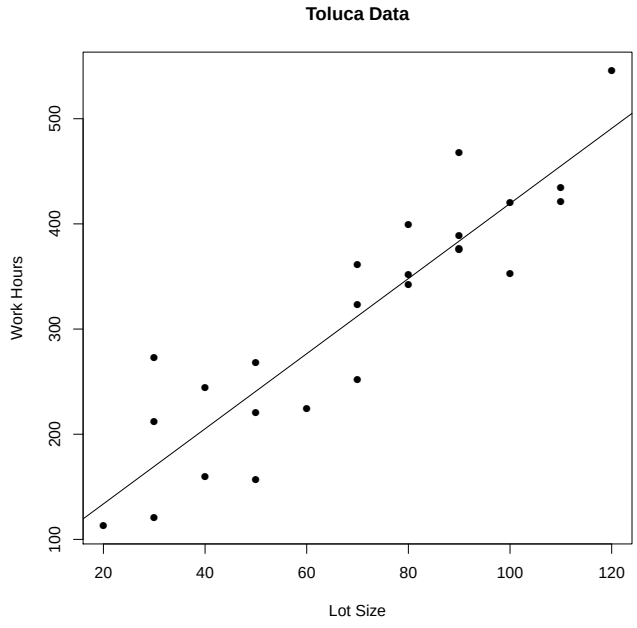
- $R^2$ could be close to one, but the $E(Y_i)$ may not lay on a line. (Why? Which plot?)
- $R^2$ may not be close to one, but a line is best for $E(Y_i)$ (Why? Which plot?)
- $R^2$ could be essentially zero, but $X$ and $Y$ could be highly related. (Why? Which plot?)

- R-squared cannot determine whether the coefficient estimates and predictions are biased, which is why we must assess the residual plots.

# 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?)

- R-squared cannot determine whether the coefficient estimates and predictions are biased, which is why we must assess the residual plots.
- R-squared does not indicate whether a regression model is adequate. You can have a low R-squared value for a good model, or a high R-squared value for a model that does not fit the data!




Toluca data:

| | | | |
|---|---|---|---|
| Root MSE | 48.82331 | R-Square | 0.8215 |
| Dependent Mean | 312.28000 | Adj R-Sq | 0.8138 |
| Coeff Var | 15.63447 | | |

# Correlation models

In the regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

- 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.

# 2.11 Bivariate normal correlation (pp. 78–87)

- If appropriate, we could assume that $X$ and $Y$ have a bivariate normal normal distribution with parameters $\mu_x$, $\mu_y$, $\sigma_x$, $\sigma_y$, and $\rho$.
- Then

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

- Investigation of linear association between $X$ and $Y$ is done through inferences about $\rho = \text{corr}(X_i, Y_i)$.
- A point estimator of $\rho$ is $r$ as defined a few slides back, the sample correlation.
- $r$ is the MLE under normality, but also an estimator in general (not assuming normality).

# Test $H_0 : \rho = 0$ and CI for $\rho$

Testing $H_0 : \rho = 0$ is equivalent to testing $H_0 : \beta_1 = 0$ in the regression of $Y$ on $X$.
A large-sample CI for $\rho$ uses Fishers z-transformation:

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

A large sample $(1-\alpha)100\%$ CI for $\log\left(\frac{1+\rho}{1-\rho}\right)$ is

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

Then back-transform endpoints to get a CI for $\rho$.
Here, $z(1-\alpha/2) = \Phi^{-1}(1-\alpha/2)$.

# Spearman rank estimate (pp 87–89)

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.

The Spearman coefficient is robust to outlying observations. It is also invariant to monotonic transformations in either $X$ or $Y$.

# R code and output, Toluca

```
> p<-cor.test(WorkHour, LotSize, method="pearson")
> p

Pearson's product-moment correlation

data:  WorkHour and LotSize
t = 10.29, df = 23, p-value = 4.449e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7965202 0.9583070
sample estimates:
      cor
0.9063848

> s<-cor.test(WorkHour, LotSize, method="spearman")
> s

Spearman's rank correlation rho

data:  WorkHour and LotSize
S = 253.88, p-value = 7.079e-10
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9023542
```

# Cautions about regression (pp. 77–78)

- 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 $Y$. Important! **association does not imply causation**.
- 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."

# More cautions...

- The least squares estimates are *not unbiased* if $X$ is measured with error – in fact coefficients are biased towards zero (see Section 4.5, p. 165-168).
- 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 $e_1, \ldots, e_n$.