# STAT2401: Analysis of Experiments
# Simple Linear Regression

John W. Lau

Dept. of Math. & Stat.

April 13, 2019

Original-Authors: Ed Cripps, Kevin Murray, and Russel John
Special Thanks: Kenyon Ng

## Simple Linear Regession

Relationship Between Two Variables

- Measure a pair of variables on a sample of experimental units or subjects:
  - height and weight of a random sample of people
  - income of wife and husband in a sample of couples
  - GDP and Debt of a sample of countries
  - response and dose of drug in a sample of patients
- The Variables:

$$\begin{aligned} Y &\quad - \quad \text{response variable or dependent variable} \\ X &\quad - \quad \text{explanatory variable or independent variable} \end{aligned}$$

- Two related but distinct questions of interest:
  - Are $Y$ and $X$ related?
  - Can $Y$ be predicted from $X$?

Which Variable is Which and Does It Matter?

- Context will usually determine which variable is $X$ and which $Y$
- $Y$ is the variable of interest and $X$ the variable which is being used in an attempt to explain what might be contributing to the variability in Y
- The variability is assumed to be in the $Y$'s and the $X$'s are taken as fixed
- The explanation of $Y$ is conditional on the values of $X$ observed
- Reversing the roles of $X$ and $Y$ does change the nature of the problem and will lead to different results and conclusions

Simple Linear Regression Model

- The model assumes the expected value of $Y$ for a given $X$ is a straight line i.e.

$$E(Y|X) = \mu(Y|X) = \beta_0 + \beta_1 X$$

- Parameters $\beta_0$ and $\beta_1$ are constants which do not depend on X
- It also assumes the distribution of Y for any X is Normal and that

$$Var(Y|X) = \sigma^2$$

  does not depend on $X$
- If $\beta_1 = 0$ then $Y$ and $X$ are unrelated, for then the distribution of $Y$ given $X$ does not depend on $X$
- The problem of establishing a relationship devolves to determining whether $\beta_1 = 0$ is plausible, based on the magnitude of some sample estimate of $\beta_1$

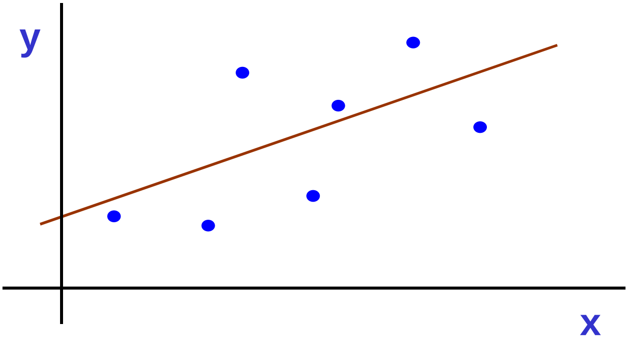
Simple Linear Regression Model of a Sample of $(X, Y)$ Pairs

- For a given set of observations $(X_i, Y_i)$ , $i = 1 \ldots n$

  $$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

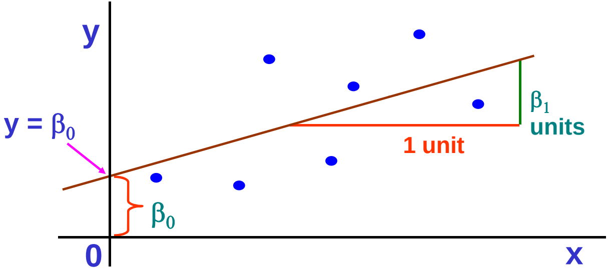
  where $\beta_0$ and $\beta_1$ are fixed parameters to be estimated and $\epsilon_i$ are independent, identically distributed Normal random variables with mean 0 and constant variance $\sigma^2$
- All variability is in the response variable
- The model describes the conditional distribution of Y at the X points observed in the sample
- At each point, we sample a value from a Normal distribution whose mean lies on the straight line determined by $\beta_0$ and $\beta_1$ and whose variance remains constant.

Given a scatter plot of the data




Find the straight line which "best" fits the data

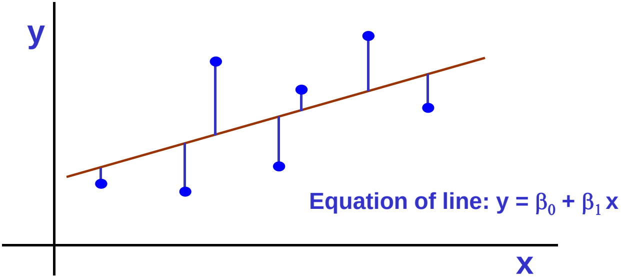
Equation of the Line: $y = \beta_0 + \beta_1 x$




- $\beta_0$ is the *intercept*; $y = \beta_0$ when $x = 0$
- $\beta_1$ is the *slope*; if $x$ increases 1 unit, $y$ changes by 1 units

# Simple Linear Regession

Least Squares Regression Line of $Y$ on $X$




- Vertical displacement of points from line are *residuals*
- The sum of the squared residuals is a measure of how close the line goes to the points
- Choose $\beta_0$ and $\beta_1$ to minimize the sum of the squared residuals

Normal Equations (Optional)

- The sum of the squared residuals (the objective function) to be minimized is

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

- Differentiating the objective function with respect to $\beta_0$ and $\beta_1$ and setting both derivatives to 0 gives the Normal equations:

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) &= 0\\ \sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) &= 0\end{aligned}$$

  This implies

$$\begin{aligned}\hat{\beta}_0 + \hat{\beta}_1 \bar{X} &= \bar{Y}\\ \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} X_i^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i Y_i\end{aligned}$$

  where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

Normal Equations

- The first Normal equation gives

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

  Substituting this into the second gives

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i (Y_i - \bar{Y})}{\sum_{i=1}^{n} X_i (X_i - \bar{X})}$$

- Noting that $\sum_{i=1}^{n} (Y_i - \bar{Y}) = 0 = \sum_{i=1}^{n} (X_i - \bar{X})$ leads to two more expressions

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Fitted Values and Unbiasedness

- The fitted value for an observation $Y_i$ is the estimate of $E(Y_i)$, namely the point on the least squares regression line at $X = X_i$ :

$$\hat{\mu}(Y|X_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i = \bar{Y} - \hat{\beta}_1(X_i - \bar{X})$$

- The latter formulation shows the least squares regression line passes through the point $(\bar{X}, \bar{Y})$
- From the model, $E(Y_i - \bar{Y}) = \beta_1(X_i - \bar{X})$ and so

$$E(\hat{\beta}_1) = \frac{\sum_{i=1}^n X_i E(Y_i - \bar{Y})}{\sum_{i=1}^n X_i(X_i - \bar{X})} = \beta_1$$

  and

$$E(\hat{\beta}_0) = E(\bar{Y}) - E(\hat{\beta}_1)\bar{X} = \beta_0 + \beta_1\bar{X} - \beta_1\bar{X} = \beta_0$$

- The fitted value is thus an unbiased estimate of $E(Y_i)$, since

$$E(\hat{\mu}(Y|X_i)) = \beta_0 + \beta_1 X_i = \mu(Y|X_i)$$

Absolute Magnitude of $\beta_1$

- From any of the given formulas for $\beta_1$ it is clear that it is not a dimensionless constant
- If $Y$ is measured in kg and $X$ in m, the units of $\beta_1$ would be kg/m
- Changing the units of $Y$ to g would increase the absolute magnitude of $\beta_1$ a thousandfold
- So the absolute magnitude of $\beta_1$ is no measure of the strength of the relationship between $Y$ and $X$, nor can it be used on its own to determine if $Y$ is related to $X$
- One way of proceeding is to standardize the $X$ and $Y$ sample values before fitting the least square regression line
- To standardize a sample, subtract the mean of the sample from all observations and then divide them all by the standard deviation of the sample

Example: Ponds Institute

- Price (per ounce) of 36 facial cleansers and a preference rating for each based on scores from a panel of 90 women oblivious to brand

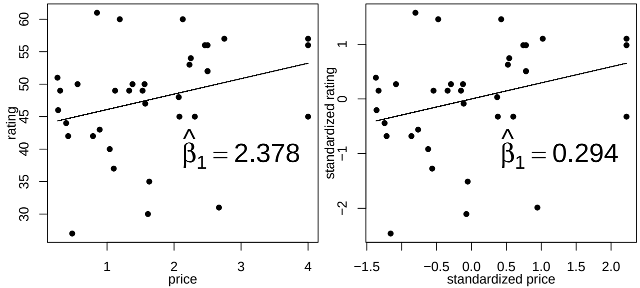
```
> load("facecleanser.RData")
> str(facecleanser)
'data.frame': 36 obs. of  4 variables:
 $ price    : num  0.85 2.13 1.19 2.75 4 4 2.5 2.46 2.25 2.23 ...
 $ rating   : num  61 60 60 57 57 56 56 56 54 53 ...
 $ Z.price  : num  -0.803 0.426 -0.477 1.021 2.221 ...
 $ Z.rating: num  1.58 1.46 1.46 1.1 1.1 ...
 - attr(*, "variable.labels")= Named chr  "" "" "Zscore(PRICE)" "Zscore(RATING
  ..- attr(*, "names")= chr  "price" "rating" "zprice" "zrating"
 - attr(*, "codepage")= int 1252
```

The Standardized Price `Z.price` and Standardized rating `Z.rating` are given by

```
> facecleanser$Z.price = with(facecleanser,(price-mean(price))/sd(price))
> facecleanser$Z.rating = with(facecleanser,(rating-mean(rating))/sd(rating))
```

## Example: Ponds Institute

```
> with(facecleanser,plot(rating~price,pch=16))
> with(facecleanser,lines(fitted(lm(rating~price))~price))
> with(facecleanser,plot(Z.rating~Z.price,pch=16))
> with(facecleanser,lines(fitted(lm(Z.rating~Z.price))~Z.price))
```



Sample Correlation Coefficient

- Let

$$Z_{X_i} = \frac{X_i - \bar{X}}{S_X} \text{ and } Z_{Y_i} = \frac{Y_i - \bar{Y}}{S_Y}$$

be the standardized samples where

$$S_X = \sqrt{\tfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}\,, S_Y = \sqrt{\tfrac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\,,$$

$\bar{X} = \sum_{i=1}^{n} X_i$, and $\bar{Y} = \sum_{i=1}^{n} Y_i$.

- Note $\sum_{i=1}^{n}(Z_{X_i} - \bar{Z}_X)^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{S_X^2} = n-1$ and $\bar{Z}_X = \frac{1}{n}\sum_{i=1}^{n} Z_{X_i} = 0$ and similarly for $Z_{Y_i}$
- The slope of the least squares regression line of $Z_Y$ on $Z_X$ is called the sample correlation coefficient, written $r_{XY}$ (or $R$)

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Sample Correlation Coefficient

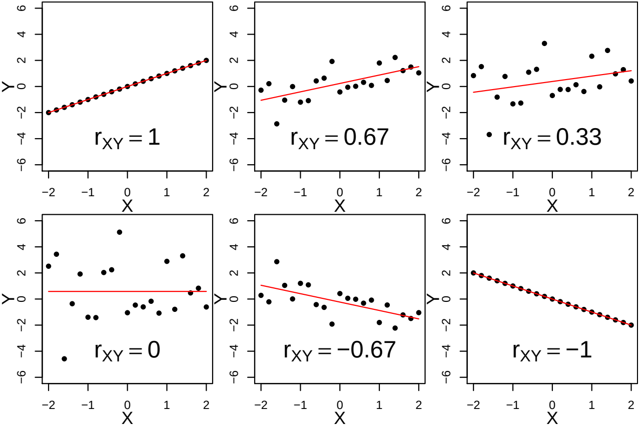
- Replacing $X_i$ and $Y_i$ by $Z_{X_i}$ and $Z_{Y_i}$ in the formula, that is to say if we consider $Z_{Y_i} = \beta_0 + \beta_1 Z_{X_i} + \epsilon_i$, $\hat{\beta}_1$ gives

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(Z_{X_i} - \bar{Z}_X)(Z_{Y_i} - \bar{Z}_Y)}{\sum_{i=1}^{n}(Z_{X_i} - \bar{Z}_X)^2} \\
&= \frac{\sum_{i=1}^{n} \frac{(X_i - \bar{X})}{s_X} \frac{(Y_i - \bar{Y})}{s_Y}}{n-1} \\
&= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = r_{XY}
\end{aligned}
$$

  - $|r_{XY}| \leq 1$ (Cauchy-Schwarz inequality)
  - $r_{XY}$ is $+1$ or -1 if $Y_i$ and $X_i$ lie in a straight line
  - $r_{XY}$ is 0 if $Y_i$ and $X_i$ are not linearly related

Scatterplots of Standardized $X$ and $Y$ with Various Values of $r_{XY}$

- Some Scatterplots



$r$ As Sample Estimate of $\rho$

- The correlation coefficient of two random variables $Y$ and $X$ is

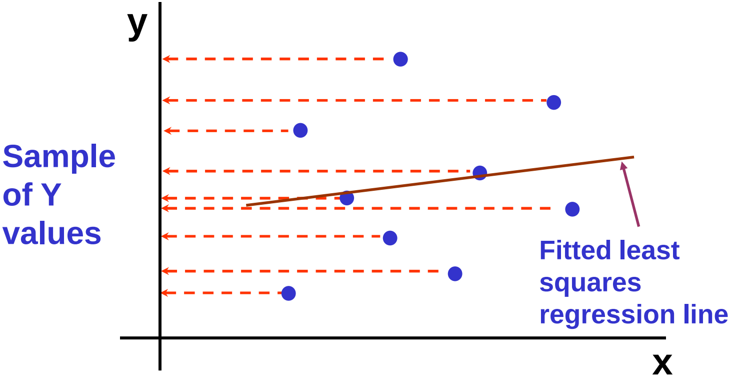
$$\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

  where $Cov(X, Y) = E(X - E(X))(Y - E(Y))$

- $\rho$ is a measure of association between the two variables
- If $X$ and $Y$ are independent then $\rho_{XY} = 0$, but more conditions (like joint Normality) must be imposed on the $Y$ and $X$ for $\rho_{XY} = 0$ to imply independence
  In this, both $Y$ and $X$ are variables in their own right with a joint probability distribution
- It is the particular properties of joint Normality which allows the analysis to proceed via the conditional distribution of $Y$ given $X$

Components of Total Variability

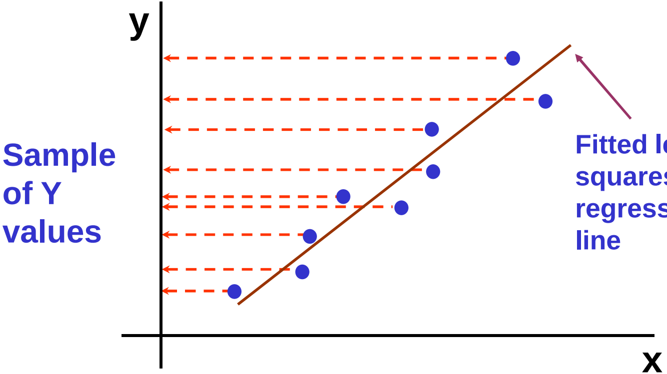
- What is contributing to the variability in the $Y$'s?




Large variability about the line and almost no contribution from the $X$'s

Components of Total Variability

- Exactly the same variability in the $Y$'s




Little variability about the line; $Y$ inherits its variability from the $X$'s

The Arithmetic of the Decomposition (Optional)

- The fitted value for $Y_i$ given $X_i$

$$\hat{\beta}_0 + \hat{\beta}_1 X_i$$

- The residual is given by

$$Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

- The difference between $Y_i$ and $\bar{Y}$

$$\begin{array}{rcll}
Y_i - \bar{Y} & = & (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) - (\bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 X_i) & \Leftarrow \quad (\text{Total}) \\
 & = & (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) & \Leftarrow \quad (\text{Residual}) \\
 & & + \ (\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y}) & \Leftarrow \quad (\text{Regression}) \\
 & = & (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) & \Leftarrow \quad (\text{Residual}) \\
 & & + \ (\hat{\beta}_0 + \hat{\beta}_1 X_i - (\hat{\beta}_0 + \hat{\beta}_1 \bar{X})) & \Leftarrow \quad (\text{Regression}) \\
 & = & (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) & \Leftarrow \quad (\text{Residual}) \\
 & & + \ \hat{\beta}_1 (X_i - \bar{X}) & \Leftarrow \quad (\text{Regression})
\end{array}$$

The Arithmetic of the Decomposition (Optional)

- Putting "sum of squares" in both sides

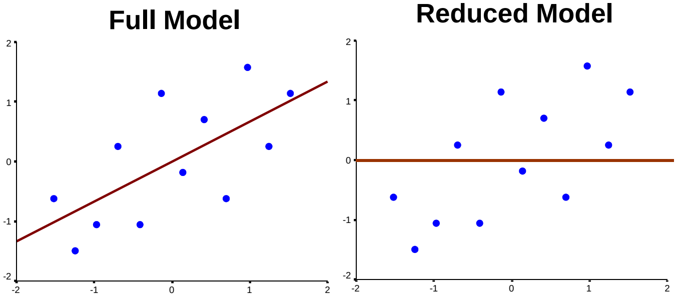
$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{Total Sum of Squares}} = \underbrace{\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}_{\text{Residual Sum of Squares}} + \underbrace{\hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2}_{\text{Regression Sum of Squares}}$$

because the cross term is zero, i.e.

$$\begin{aligned}
& 2\hat{\beta}_1 \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(X_i - \bar{X}) \\
= \;& 2\hat{\beta}_1 \sum_{i=1}^{n} Y_i(X_i - \bar{X}) - 2\hat{\beta}_1^2 \sum_{i=1}^{n} X_i(X_i - \bar{X}) \\
= \;& 2\hat{\beta}_1 \left(\hat{\beta}_1 \times \sum_{i=1}^{n} X_i(X_i - \bar{X})\right) - 2\hat{\beta}_1^2 \sum_{i=1}^{n} X_i(X_i - \bar{X}) = 0
\end{aligned}$$

Competing Models

- Full Model and Reduced Model




| | Full Model | Reduced Model |
|---|---|---|
| Fitted value: | $\hat{\beta}_0 + \hat{\beta}_1 X_i$: 2 parameters | $\bar{Y}$: 1 parameter |
| Residual: | $Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ | $Y_i - \bar{Y}$ |
| Residual SS: | Residual SS on $n-2$ df | Total SS on $n-1$ df |

Extra-Sum-of-Squares Principle:

- Difference in Residual SS (Reduced $-$ Full) $=$ Regression SS
- Difference in Residual DF (Reduced $-$ Full) $= 1$
- Regression MS/Residual MS (Full) $=$ $F$-ratio

## Example: Ponds Institute

- ```
  > F1 = lm(rating~price,data=facecleanser)
  > anova(F1)
  Analysis of Variance Table

  Response: rating
            Df  Sum Sq Mean Sq F value  Pr(>F)
  price      1  214.72 214.723  3.2239 0.08146 .
  Residuals 34 2264.50  66.603
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
  ```
  ```

### Example: Ponds Institute

- ```
  > summary(F1)

  Call:
  lm(formula = rating ~ price, data = facecleanser)

  Residuals:
      Min      1Q  Median      3Q     Max
  -19.061  -3.607   2.235   4.666  15.268

  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
  (Intercept)   43.711      2.616  16.712   <2e-16 ***
  price          2.378      1.325   1.796   0.0815 .
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 8.161 on 34 degrees of freedom
  Multiple R-squared:  0.08661,Adjusted R-squared:  0.05974
  F-statistic: 3.224 on 1 and 34 DF,  p-value: 0.08146
  ```
  ```

Estimating $\sigma^2$ and the SE's

- The Residual MS is an unbiased estimate of $\sigma^2$. Later we will see
$$E\left[\frac{1}{n-2}\sum_{i=1}^n(Y_i-\hat{\beta}_0-\hat{\beta}_1X_i)^2\right]=\sigma^2$$
- Using $Y_i=\beta_0+\beta_1X_i+\epsilon_i$ in the last formula for $\hat{\beta}_1$ gives
$$\hat{\beta}_1=\frac{\sum_{i=1}^n(X_i-\bar{X})Y_i}{\sum_{i=1}^n(X_i-\bar{X})^2}=\frac{\sum_{i=1}^n(X_i-\bar{X})(\beta_0+\beta_1X_i+\epsilon_i)}{\sum_{i=1}^n(X_i-\bar{X})^2}=\beta_1+\frac{\sum_{i=1}^n(X_i-\bar{X})\epsilon_i}{\sum_{i=1}^n(X_i-\bar{X})^2}$$
and so
$$Var(\hat{\beta}_1)=Var(\hat{\beta}_1-\beta_1)=\frac{\sum_{i=1}^n(X_i-\bar{X})^2\sigma^2}{(\sum_{i=1}^n(X_i-\bar{X})^2)^2}=\frac{\sigma^2}{\sum_{i=1}^n(X_i-\bar{X})^2}$$
- (Optional) This immediately leads to an SE for $\hat{\beta}_1$ and shows that
$$\begin{aligned}E\left[\hat{\beta}_1^2\sum_{i=1}^n(X_i-\bar{X})^2\right] &= E[\hat{\beta}_1^2]\sum_{i=1}^n(X_i-\bar{X})^2\\ &= (Var(\hat{\beta}_1)+(E[\hat{\beta}_1])^2)\sum_{i=1}^n(X_i-\bar{X})^2\\ &= \sigma^2+\beta_1^2\sum_{i=1}^n(X_i-\bar{X})^2\end{aligned}$$
So, the Regression MS is also an estimate of $\sigma^2$ when $\beta_1=0$

Coefficient of Determination (or `Multiple R-squared`)

- Definition

$$R^2 = \frac{\text{Regression SS}}{\text{Total SS}} = 1 - \frac{\text{Residual SS}}{\text{Total SS}}$$

  - Often expressed as a percentage
  - A.k.a. Percentage of variance accounted for by regression
  - Close to 1 (or 100%) if relationship is strong; precise prediction of individuals, as opposed to average, is possible
  - For simple linear regression, $R =$ coefficient of correlation
  - For a given response variable, the smaller the Residual SS (i.e. the better the model fits) the larger $R^2$

Adjusted $R^2$ (or `Adjusted R-squared`)

- Definition

$$\text{Adjusted } R^2 = 1 - \frac{\text{Residual MS}}{\text{Total MS}} = 1 - \frac{\hat{\sigma}^2}{S_Y^2}$$

  - For a given response variable, the smaller the estimate of $\sigma^2$, the larger Adj $R^2$
  - "Adjusted" for differing model df when there are many $X$ variables and thus many models to chose from
  - Penalizes $R^2$ for always increasing when more explanatory variables are put into the model
  - It has no role in simple linear regression
  - Adj $R^2$ can be negative

## Simple Linear Regession

Towards a SE of a Fitted Value (Optional)

- We begin with our previous incarnation of $\hat{\beta}_1$, namely

$$\hat{\beta}_1 - \beta_1 = \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \sum_{i=1}^n (X_i - \bar{X})\epsilon_i$$

and the variable $\bar{Y}$ minus its expected value

$$\bar{Y} - \beta_0 - \beta_1 \bar{X} = \frac{1}{n} \sum_{i=1}^n \epsilon_i$$

Then

$$\begin{aligned} Cov(\bar{Y}, \hat{\beta}_1) &= E\Big[(\bar{Y} - \beta_0 - \beta_1\bar{X})(\hat{\beta}_1 - \beta_1)\Big] \\ &= \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \frac{1}{n} \sum_{j=1}^n \sum_{i=1}^n (X_i - \bar{X}) E[\epsilon_i \epsilon_j] \\ &= \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}) E[\epsilon_i^2] \\ &= \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}) \sigma^2 = 0 \end{aligned}$$

## Simple Linear Regession

SE of the Mean of $Y$ at any $X$, of a Fitted Value and of $\beta_0$

- Knowing $Var(\bar{Y})$, $Var(\hat{\beta}_1)$ and $Cov(\bar{Y}, \hat{\beta}_1) = 0$ enables the calculation of the $Var(\hat{\mu}(Y|X_0))$ for any point $X = X_0$ (which includes $X =$ an observed $X_i$) as a simple sum of the variances:

$$\begin{aligned} Var\left[\hat{\mu}(Y|X_0)\right] &= Var\left[\hat{\beta}_0 + \hat{\beta}_1 X_0\right] = Var\left[\bar{Y} - \beta_1 \bar{X} + \hat{\beta}_1 X_0\right] \\ &= Var\left[\bar{Y} - \hat{\beta}_1(\bar{X} - X_0)\right] \\ &= \sigma^2\left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] \end{aligned}$$

- This leads to a SE for $\hat{\mu}(Y|X_0)$ which can be used to construct confidence intervals using the $t_{n-2}$ distribution
- Putting $X = 0$ in the above expression gives

$$Var\left[\hat{\mu}(Y|X = 0)\right] = Var\left[\hat{\beta}_0\right] = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

Calculating Confidence Intervals for $\beta_0$ and $\beta_1$

- $100(1-\alpha)\%$ Confidence intervals $\beta_0$ and $\beta_1$ are, respectively

$$
\begin{aligned}
\hat{\beta}_0 &\pm t_{n-2,\alpha/2} \times \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right]} \\
\hat{\beta}_1 &\pm t_{n-2,\alpha/2} \sqrt{\hat{\sigma}^2 \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2}}
\end{aligned}
$$

- In particular, we use the notation

$$
\begin{aligned}
\mathrm{SE}(\hat{\beta}_0) &= \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right]} \\
\mathrm{SE}(\hat{\beta}_1) &= \sqrt{\hat{\sigma}^2 \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2}}
\end{aligned}
$$

and $\hat{\sigma}^2$ is the Residual MS

# Simple Linear Regession

## Example: Ponds Institute

- ```
  > confint(F1,level=0.95)
                   2.5 %   97.5 %
  (Intercept) 38.3955349 49.02635
  price       -0.3135386  5.07000
  ```

  or

  ```
  > ## need library("Rcmdr")
  > Confint(F1,level=0.95)
              Estimate      2.5 %   97.5 %
  (Intercept) 43.71094 38.3955349 49.02635
  price        2.37823 -0.3135386  5.07000
  ```
  ```

Calculating Confidence Intervals for the Mean of $Y$ at Any Point $X_0$

- A $100(1-\alpha)\%$ confidence interval for $\mu(Y|X_0)$ is

$$\hat{\beta}_0 + \hat{\beta}_1 X_0 \ \pm \ t_{n-2,\alpha/2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(X_0-\bar{X})^2}{\sum_{i=1}^n (X_i-\bar{X})^2}\right]}$$

- Here we use the notation

$$\mathrm{SE}(\hat{\beta}_0 + \hat{\beta}_1 X_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(X_0-\bar{X})^2}{\sum_{i=1}^n (X_i-\bar{X})^2}\right]}$$

  and $\hat{\sigma}^2$ is the Residual MS.

- In particular, when $X_0 = 0$,

$$\mathrm{SE}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i-\bar{X})^2}\right]}$$

Example: Ponds Institute

- At $X_0 = 1$ and $X_0 = 3$:

```
> predict(F1,data.frame(price=c(1,3)),interval="confidence")
       fit      lwr      upr
1 46.08917 42.76392 49.41442
2 50.84563 46.35804 55.33322
```

Confidence Bands

- As $X$ changes, the upper and lower bounds of the $100(1-\alpha)\%$ C.I. for $\mu(Y|X)$ trace out curves in $X$

$$\hat{\beta}_0 + \hat{\beta}_1 X \ \pm \ t_{n-2,\alpha/2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(X-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right]}$$

  These curves can be superimposed on the scatterplot of $Y_i$ vs $X_i$ to give some idea of the level of precision in the estimate of the $\mu(Y|X)$ at any point

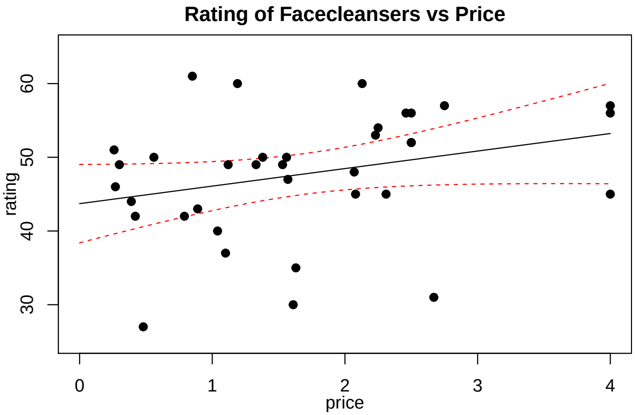
Example: Ponds Institute

- ```
  > new = data.frame(price=seq(0,4,0.1))
  > CIs = predict(F1,new,interval="confidence")
  > matplot(new$price,CIs,lty=c(1,2,2),col=c("black","red","red"),
  +         type="l",ylab="rating",main="Rating of Facecleansers vs Price",
  +         ylim=c(25,65))
  > points(rating~price,data=facecleanser,pch=21,bg="black")
  ```
  ```

## Simple Linear Regession

Example: Ponds Institute

- The plot




Interval width increases with the distance from the mean of the $X$'s; precision diminishes as the data becomes scarcer

Predicting The Impossible

- Up to this point we have been attempting to predict the average of $Y$ at for any given $X$
- What if we were to predict a new observation at $X = X_0$?
- The new observation would be

$$Y_{\text{new}} = \beta_0 + \beta_1 X_0 + \epsilon$$

  where $\epsilon$ is Normally distributed with mean 0 and variance $\sigma^2$ and independent of the sample values $Y_i$
- The estimate of $Y_{\text{new}}$, say

$$\hat{Y}_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

Prediction Intervals

- the expected value and variance of the difference $Y_{\text{new}} - \hat{Y}_{\text{new}}$

$$\begin{aligned} E(Y_{\text{new}} - \hat{Y}_{\text{new}}) &= 0 \\ Var(Y_{\text{new}} - \hat{Y}_{\text{new}}) &= \sigma^2 + Var(\hat{Y}_{\text{new}}) \end{aligned}$$

where

$$Var(\hat{Y}_{new}) = \sigma^2 \left[ \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]$$

- Estimating $\sigma^2$ by the Residual MS leads to a $t_{n-2}$ distribution for $Y_{new} - \hat{Y}_{new}$, which gives the $100(1-\alpha)\%$ CI for $Y_{new}$ as

$$\hat{\beta}_0 + \hat{\beta}_1 X_0 \ \pm \ t_{n-2,\alpha/2} \sqrt{\hat{\sigma}^2 \left[ 1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]}$$

Such an interval is usually termed a $100(1-\alpha)\%$ prediction interval, to distinguish it from a confidence interval of a fixed parameter

Prediction Intervals

- As $\hat{\sigma}^2\left[\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right]\to 0$ when $n\to\infty$, The first term dominates the SE and the prediction interval is effectively

$$\hat{\beta}_0+\hat{\beta}_1X_0 \ \pm\ t_{n-2,\alpha/2}\sqrt{\hat{\sigma}^2}$$

as might be expected from the assumption that the variance of an observation at any given $X$ is constant

## $R^2$ Or Not $R^2$

- For some regressionistas, $R^2$ dominates the discussion of results and no model will be countenanced unless it exceeds 90%
- Such dizzy heights are seldom seen in any data collected from living things
- When considering summary output, you need to bear in mind the aim of the regression
- Certainly $R^2$ needs to be very high if you're contemplating estimating a new observation
- For those who merely wish to establish a relationship or are happy estimating the mean response, $R^2$ can take a back seat
- Also be mindful that $R^2$ is on a squared scale, so that an $R^2$ of 50% corresponds to an $r$ of 0.7, which is a relatively high correlation; almost certainly significant

## Simple Linear Regession

Example: Cholesterol Data Set

- Data from 1109 West Australians with measurements pertaining to body mass, cholesterol, blood pressure and other data of medical obsession
- Is BMI related to cholesterol?

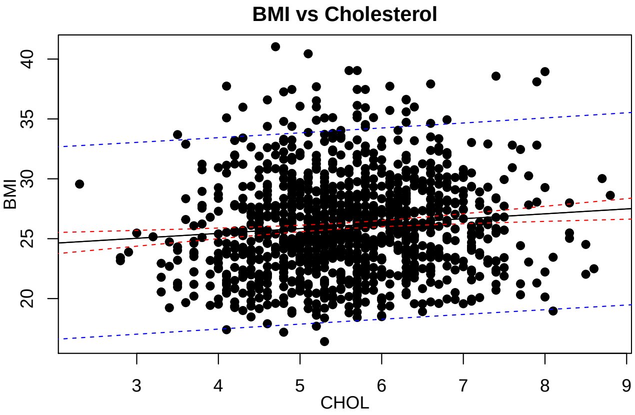
```
> load("cholesterol.RData")
> str(cholesterol)
'data.frame': 1109 obs. of  8 variables:
 $ AGE   : num  32 40 39 37 46 44 51 50 49 49 ...
 $ BMI   : num  24.2 26.3 25.1 28.7 26.3 ...
 $ CHOL  : num  4.7 5.8 5.5 5.6 5.9 5.8 5.4 6 4.9 7.2 ...
 $ DBP   : num  70 70 70 80 80 84 90 80 84 90 ...
 $ HEIGHT: num  175 183 182 183 185 180 170 173 187 178 ...
 $ SEX   : Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 2 ...
 $ WAIST : num  82 93 91 98 95 84 82 92 95 89 ...
 $ WEIGHT: num  74 88 83 96 90 76 72 75 98 89 ...
 - attr(*, "variable.labels")= Named chr  "age" "BMI" "cholesterol" "DBP" ...
  ..- attr(*, "names")= chr  "AGE" "BMI" "CHOL" "DBP" ...
```

## Example: Cholesterol Data Set

- ```
  > cholesterol[1:18,]
     AGE   BMI CHOL DBP HEIGHT  SEX WAIST WEIGHT
  1   32 24.16  4.7  70    175 male    82     74
  2   40 26.28  5.8  70    183 male    93     88
  3   39 25.06  5.5  70    182 male    91     83
  4   37 28.67  5.6  80    183 male    98     96
  5   46 26.30  5.9  80    185 male    95     90
  6   44 23.46  5.8  84    180 male    84     76
  7   51 24.91  5.4  90    170 male    82     72
  8   50 25.06  6.0  80    173 male    92     75
  9   49 28.02  4.9  84    187 male    95     98
  10  49 28.09  7.2  90    178 male    89     89
  11  55 26.51  5.1 100    178 male   105     84
  12  52 29.98  4.8  80    178 male   102     95
  13  54 29.32  7.1  80    180 male   107     95
  14  61 31.26  4.6  80    185 male   109    107
  15  59 24.22  6.4  70    170 male    92     70
  16  61 23.77  4.7  90    180 male    89     77
  17  67 28.73  6.3  80    175 male    98     88
  18  67 25.03  8.3  90    181 male    97     82
  ```
  ```

Example: Cholesterol Data Set

- The plot




This is a large data set of high variability. The sample size works in favour of the CI (red), but the PI (blue) is left out in the cold.

Example: Cholesterol Data Set

- The code

```
> C1 = lm(BMI~CHOL,data=cholesterol)
> plot(BMI~CHOL,data=cholesterol,pch=21,bg="black",main="BMI vs Cholesterol")
> new = data.frame(CHOL=seq(2,10,0.1))
> CIs = predict(C1,new,interval="confidence")
> PIs = predict(C1,new,interval="predict")
> matpoints(new$CHOL,CIs,lty=c(1,2,2),col=c("black","red","red"),type="l")
> matpoints(new$CHOL,PIs,lty=c(1,2,2),col=c("black","blue","blue"),type="l")
```

Example: Cholesterol Data Set

- The Fitted Model

```
> C1 = lm(BMI~CHOL,data=cholesterol)
> anova(C1) ## Analysis of Variance Table
Analysis of Variance Table

Response: BMI
            Df  Sum Sq Mean Sq F value    Pr(>F)
CHOL         1   184.3 184.333  11.141 0.0008726 ***
Residuals 1107 18316.1  16.546
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Cholesterol Data Set

- The Fitted Model

```
> summary(C1)


Call:
lm(formula = BMI ~ CHOL, data = cholesterol)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5632 -2.9024 -0.4118  2.4582 15.3019

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.8078     0.6919  34.409  < 2e-16 ***
CHOL          0.4086     0.1224   3.338 0.000873 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.068 on 1107 degrees of freedom
Multiple R-squared:  0.009964,Adjusted R-squared:  0.009069
F-statistic: 11.14 on 1 and 1107 DF,  p-value: 0.0008726
```

Example: Cholesterol Data Set

- Conclusion

  - ```
    > confint(C1,level=0.95)
                    2.5 %     97.5 %
    (Intercept) 22.4501727 25.1653519
    CHOL         0.1683972  0.6487605
    ```

    or

    ```
    > ## need library("Rcmdr")
    > Confint(C1,level=0.95)
                  Estimate      2.5 %     97.5 %
    (Intercept) 23.8077623 22.4501727 25.1653519
    CHOL         0.4085789  0.1683972  0.6487605
    ```

    As cholesterol increases by 1, mean BMI is estimated to increase by between 0.168 and 0.649
  - Note there is no causal relationship implied or intended to be implied by this statement
    ```

Example: Cholesterol Data Set

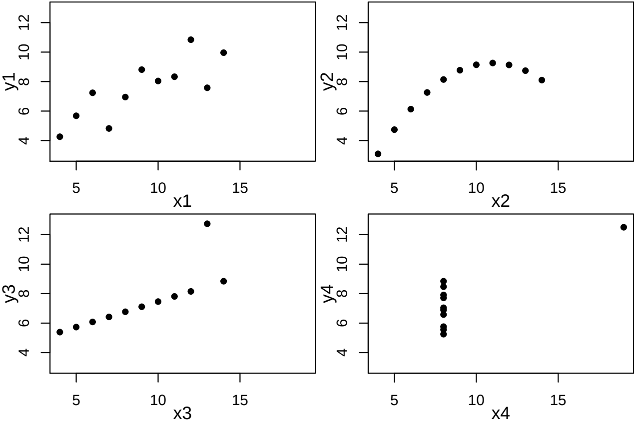
- Conclusion
  - Whether such an increase is of practical significance is not something statistics has an opinion on
  - Clearly in this example, there is a great deal of "unexplained" variability; 99% in fact
  - There are many other variables involved in determining BMI. Incorporating those variables into the equation to obtain a better explanation is the province of multiple regression
- Disclaimer
  - The interpretations of the examples presented so far are only valid provided the model assumptions are valid
  - Nothing in the numbers presented so far will tell you if this is the case or not

## Anscombe Quartet

- ```
  > load("anscombe.RData")
  > anscombe
     x1    y1 x2   y2 x3    y3 x4    y4
  1   4  4.26  4 3.10  4  5.39  8  5.56
  2   5  5.68  5 4.74  5  5.73 19 12.50
  3   6  7.24  6 6.13  6  6.08  8  5.25
  4   7  4.82  7 7.26  7  6.42  8  6.89
  5   8  6.95  8 8.14  8  6.77  8  5.76
  6   9  8.81  9 8.77  9  7.11  8  8.84
  7  10  8.04 10 9.14 10  7.46  8  6.58
  8  11  8.33 11 9.26 11  7.81  8  8.47
  9  12 10.84 12 9.13 12  8.15  8  7.91
  10 13  7.58 13 8.74 13 12.74  8  7.71
  11 14  9.96 14 8.10 14  8.84  8  7.04
  ```
  ```

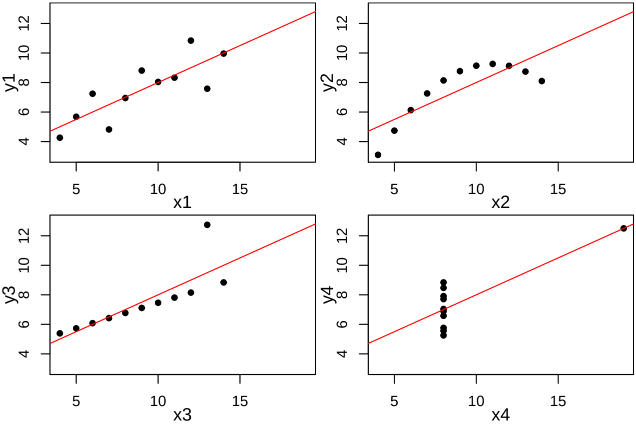
Anscombe Quartet

- Scatterplots



Anscombe Quartet

- The fits




Equation of line is $y = 3 + 0.5x$ with $R^2 = 67\%$ in each case

### Virtually Identical Output

- R-output

```
> A1 = lm(y1~x1,data=anscombe)
> summary(A1)


Call:
lm(formula = y1 ~ x1, data = anscombe)

Residuals:
     Min       1Q   Median       3Q      Max
-1.92127 -0.45577 -0.04136  0.70941  1.83882

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0001     1.1247   2.667  0.02573 *
x1            0.5001     0.1179   4.241  0.00217 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6665,Adjusted R-squared:  0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

### Virtually Identical Output

- R-output

```
> A2 = lm(y2~x2,data=anscombe)
> summary(A2)


Call:
lm(formula = y2 ~ x2, data = anscombe)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9009 -0.7609  0.1291  0.9491  1.2691

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.001      1.125   2.667  0.02576 *
x2             0.500      0.118   4.239  0.00218 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared:  0.6662,Adjusted R-squared:  0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179
```

## Virtually Identical Output

- R-output

```
> A3 = lm(y3~x3,data=anscombe)
> summary(A3)


Call:
lm(formula = y3 ~ x3, data = anscombe)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1586 -0.6146 -0.2303  0.1540  3.2411

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0025     1.1245   2.670  0.02562 *
x3            0.4997     0.1179   4.239  0.00218 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6663,Adjusted R-squared:  0.6292
F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
```

## Virtually Identical Output

- R-output

```
> A4 = lm(y4~x4,data=anscombe)
> summary(A4)


Call:
lm(formula = y4 ~ x4, data = anscombe)

Residuals:
   Min     1Q Median     3Q    Max
-1.751 -0.831  0.000  0.809  1.839

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0017     1.1239   2.671  0.02559 *
x4            0.4999     0.1178   4.243  0.00216 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.236 on 9 degrees of freedom
Multiple R-squared:  0.6667,Adjusted R-squared:  0.6297
F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

Important Lesson

- Nothing in the regression output, be it $R^2$, adjusted $R^2$, Residual MS, $F$ ratio, $p$-value, regression coefficients or any other number, will alert you to possibly serious inadequacies in the analysis
- The adequacy of the fitted model and potential problems must be assessed independently of standard output
- Diagnostics:
  - Plot the data
  - Plot Standardized Residuals vs Fitted Values
  - Plot Standardized Residuals vs $X$
  - Leverages and Cook's Distances
  - Q-Q plot of Standardized Residuals

Plotting Standardized Residuals vs Fitted Values or $X$

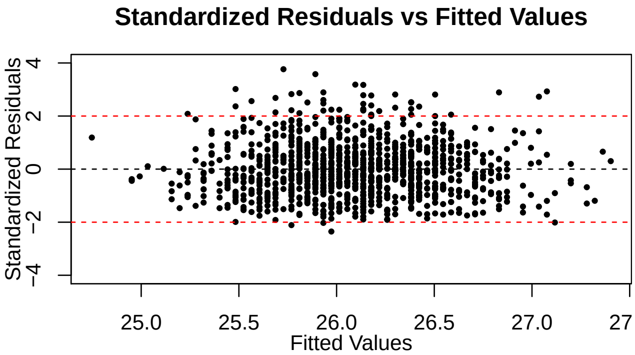
- Check Assumptions and Model Fit:
  - Check constant variance; do the points fan out as $X$ increases
  - Check for large residuals; 95% should lie within the tramlines of $y = \pm 2$
  - Check for skewness and model fit; equal numbers of positive and negative residuals at each $X$
  - Check the adequacy of the straight line to model the data; there should be no curvature apparent in the plot
  - Check for unusual patterns or clusters; the plot should look as "random" as a shotgun blast
  - Check for points of high leverage; points whose $X$ value is extreme

## Simple Linear Regession

Example: Cholesterol Data Set

- BMI on CHOL Standardized Residuals vs Fits Plot

```
> plot(rstandard(C1)~fitted(C1),pch=21,bg="black",main="Standardized Residuals
+      xlab="Fitted Values",ylab="Standardized Residuals")
> abline(h=c(-2,0,2),col=c(2,1,2),lty=2)
```

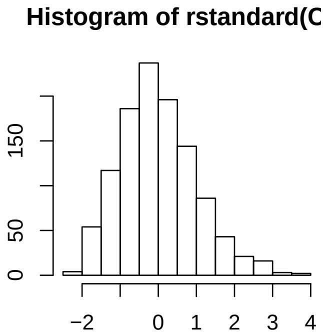


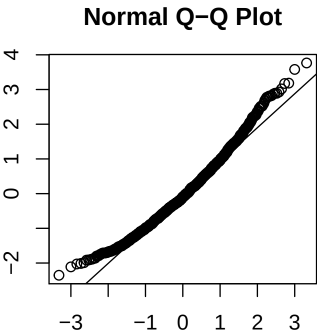

Many overly large positive residuals. Skew distribution of standardized residuals. Some evidence of increasing variance.

Example: Cholesterol Data Set

- BMI on CHOL QQ Plot of Standardized Residual

```
> hist(rstandard(C1))
> qqnorm(rstandard(C1))
> qqline(rstandard(C1))
```



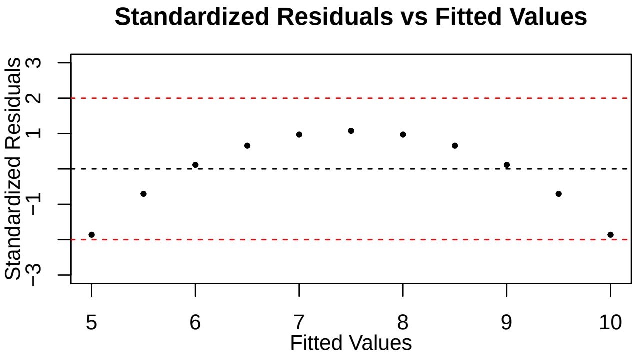




Residuals are skewed to right; not Normal. A log transform of BMI and possibly also CHOL is suggested.

Example

- Anscombe's `Y2` on `X2` Standardized Residuals vs Fits Plot

```
> plot(rstandard(A2)~fitted(A2),pch=21,bg="black",main="Standardized Residuals
+      xlab="Fitted Values",ylab="Standardized Residuals")
> abline(h=c(-2,0,2),col=c(2,1,2),lty=2)
```

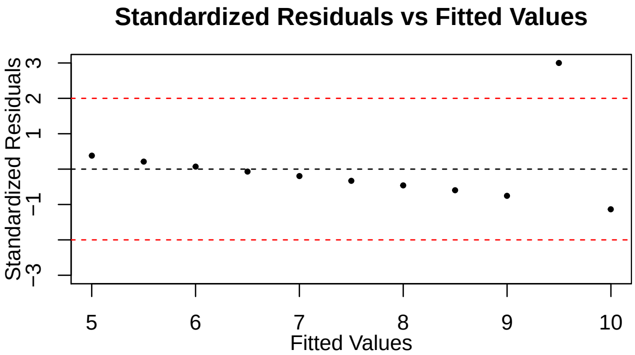



Quadratic trend in the residuals. Straight line fit is inadequate. Add quadratic term to the model.

Example

- Anscombe's `Y3` on `X3` Standardized Residuals vs Fits Plot

```
> plot(rstandard(A3)~fitted(A3),pch=21,bg="black",main="Standardized Residuals
+       xlab="Fitted Values",ylab="Standardized Residuals")
> abline(h=c(-2,0,2),col=c(2,1,2),lty=2)
```




Outlier has pulled line towards it. Linear trend still exists in the residuals. Fitted slope does not match the majority of the data.

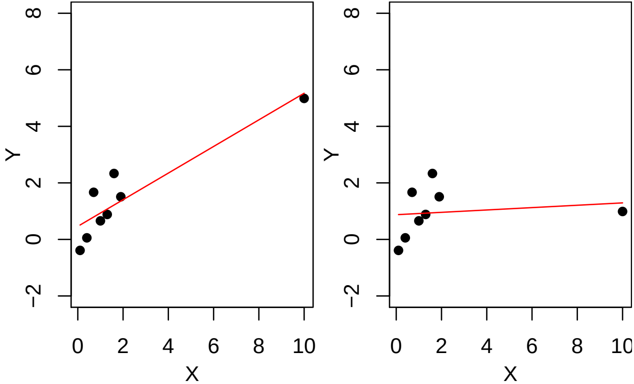
## Simple Linear Regession

Problematic Points of Two Types (outliers or/and points of high leverage)

- Individual points in a regression can be problematic because their $Y$ value is a long way from where it should be (such points are outliers)
- Equally problematic are points whose $X$ value is a long way from the bulk of the other $X$'s (these are points of high leverage)
- Presence of either type of points can have a substantial effect on the analysis and its conclusions
- Often their presence indicates that an inadequate model has been fitted and should not immediately set in train a massive cull of offending points

Scatterplots of Standardized $X$ and $Y$ with Various Values of $r_{XY}$

- High Leverage Points




Points at some remove from the other $X$'s. (Whither thou goest I will go)

Towards the Leverages

- The leverage of a sample point $(X_i, Y_i)$ is derived from the variance of its residual

$$Var(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = \sigma^2 + Var(\hat{\beta}_0 + \hat{\beta}_1 X_i) - 2Cov(Y_i, \hat{\beta}_0 + \hat{\beta}_1 X_i)$$

Towards the Leverages (Optional)

- The variance term we know as the variance of a fitted value

$$Var(\hat{\beta}_0 + \hat{\beta}_1 X_i) = \sigma^2 \left[ \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \right]$$

- The covariance term involves the product of

$$\begin{aligned} Y_i - E[Y_i] &= \epsilon_i \\ \hat{\beta}_0 + \hat{\beta}_1 X_i - E[\hat{\beta}_0 + \hat{\beta}_1 X_i] &= (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1) X_i \\ &= \frac{1}{n} \sum_{j=1}^{n} \epsilon_j + (\hat{\beta}_1 - \beta_1)(X_i - \bar{X}) \end{aligned}$$

because

$$\begin{aligned} (\hat{\beta}_0 + \hat{\beta}_1 \bar{X}) - \bar{Y} &= (\hat{\beta}_0 + \hat{\beta}_1 \bar{X}) - (\beta_0 + \beta_1 \bar{X} + \frac{1}{n} \sum_{j=1}^{n} \epsilon_j) \\ \Rightarrow \qquad \hat{\beta}_0 - \beta_0 &= \frac{1}{n} \sum_{j=1}^{n} \epsilon_j - (\hat{\beta}_1 - \beta_1) \bar{X} \end{aligned}$$

Towards the Leverages (Optional)

- The calculation uses the "$\epsilon$" formulation of

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{j=1}^{n}(X_j - \bar{X})\epsilon_j}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

- The Covariance Term

$$\begin{aligned}
Cov(Y_i, \hat{\beta}_0 + \hat{\beta}_1 X_i) &= E\left[(Y_i - E[Y_i])(\hat{\beta}_0 + \hat{\beta}_1 X_i - E[\hat{\beta}_0 + \hat{\beta}_1 X_i])\right] \\
&= E\left[\epsilon_i\left(\frac{1}{n}\sum_{j=1}^{n}\epsilon_j + (\hat{\beta}_1 - \beta_1)(X_i - \bar{X})\right)\right] \\
&= E\left[\epsilon_i\left(\frac{1}{n}\sum_{j=1}^{n}\epsilon_j + \frac{\sum_{j=1}^{n}(X_j - \bar{X})\epsilon_j}{\sum_{j=1}^{n}(X_j - \bar{X})^2}(X_i - \bar{X})\right)\right] \\
&= \frac{1}{n}\sum_{j=1}^{n}E[\epsilon_i\epsilon_j] + \frac{\sum_{j=1}^{n}(X_j - \bar{X})(X_i - \bar{X})E[\epsilon_i\epsilon_j]}{\sum_{j=1}^{n}(X_j - \bar{X})^2} \\
&= \frac{1}{n}E[\epsilon_i^2] + \frac{(X_i - \bar{X})^2 E[\epsilon_i^2]}{\sum_{j=1}^{n}(X_j - \bar{X})^2} \\
&= \sigma^2\left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}\right]
\end{aligned}$$

Variance of a Residual and Expected Residual MS

- The variance of the residual at $(X_i, Y_i)$ is thus

$$Var(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = \left[1 - \left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2}\right)\right]\sigma^2$$

- (Optional) This not only gives us the leverages but also clears an item in the pending tray:

$$\begin{aligned} E\left[\texttt{Residual SS}\right] &= E\left[\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2\right] \\ &= \sum_{i=1}^n E\left[(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i - E[Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i])^2\right] \\ &= \sum_{i=1}^n Var(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \\ &= \left[n - \frac{n}{n} - \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2}\right]\sigma^2 = (n-2)\sigma^2 \end{aligned}$$

showing the Residual MS is an unbiased estimate of $\sigma^2$

Leverages

- The $i$th leverage is denoted $h_i$. By definition

$$Var(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = (1 - h_i)\sigma^2$$

  and so

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

- Some maximization under constraint shows $0 < h_i < 1$
- The further $X_i$ is from $\bar{X}$ the closer its leverage to 1
- If the leverage is very close to 1, the residual is 0 with probability close to 1 and the point's predicted value is close to its observed value, irrespective of any input from the rest of the data
- A point is defined to have high leverage if

$$h_i > 2 \ \times \ \text{average leverage} = 2\frac{p}{n}$$

  where $p$ = Regression df + 1 = 2 for simple linear regression

Example: Huber's Data

- ```
  > load("huber.RData")
  > huber
     x  BadY  GoodY
  1 -4  2.48   2.48
  2 -3  0.73   0.73
  3 -2 -0.04  -0.04
  4 -1 -1.44  -1.44
  5  0 -1.32  -1.32
  6 10  0.00 -11.40
  ```
- The point of contention is at $X = 10$; otherwise `GoodY = BadY`
  ```

Example: Huber's Data

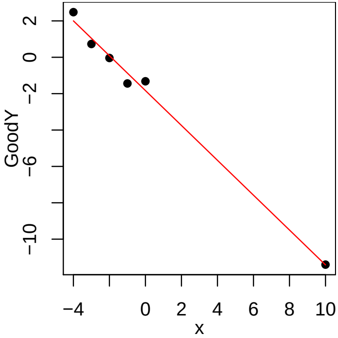
- $n = 6$ and so the leverage cut-off point is $4/6 = 0.667$

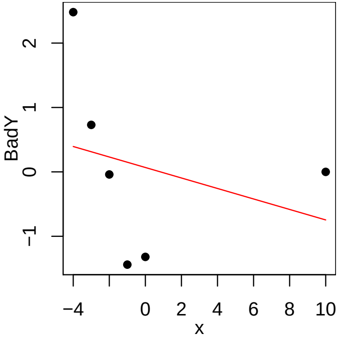
```
> H1 = lm(GoodY~x,data=huber)
> H2 = lm(BadY~x,data=huber)
> cbind('H1 stdres'=rstandard(H1),'H1 hats'=hatvalues(H1),
+       'H2 stdres'=rstandard(H2),'H2 hats'=hatvalues(H2))
   H1 stdres   H1 hats  H2 stdres   H2 hats
1  1.1792919 0.2897436  1.5967356 0.2897436
2 -0.7454671 0.2358974  0.3079765 0.2358974
3 -0.2902758 0.1974359 -0.1953457 0.1974359
4 -1.2964502 0.1743590 -1.1287795 0.1743590
5  1.1650975 0.1666667 -0.9811785 0.1666667
6  0.1273613 0.9358974  1.9016428 0.9358974
```

leverages (or hat values) are the same because the $X$'s are the same

Example: Huber's Data

- Comparsion (for the 6th point)



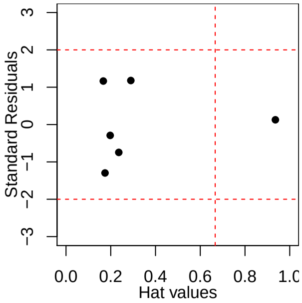

`stdres = 0.127` & `hats = 0.936`
Point follows the trend of the rest of the data
It's a point of high leverage but not an outlier
No action necessary

`stdres = 1.902` & `hats = 0.936`
Point does not follow the trend of the rest of the data
It's a point of high leverage and an outlier ($|$`stdres`$| > 2$, almost)
Fit a different model

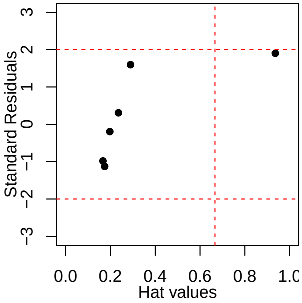
Example: Huber's Data

- Comparsion (Better Pictures)

Model: `H1`




Model: `H2`



Example: Huber's Data

- Code

```
> plot(rstandard(H1)~hatvalues(H1),pch=16,xlim=c(0,1),ylim=c(-3,3),
+       xlab="Hat values",ylab="Standard Residuals")
> abline(h=c(-2,2),col=c(2,2),lty=2)
> abline(v=4/6,col=2,lty=2)
> plot(rstandard(H2)~hatvalues(H2),pch=16,xlim=c(0,1),ylim=c(-3,3),
+       xlab="Hat values",ylab="Standard Residuals")
> abline(h=c(-2,2),col=c(2,2),lty=2)
> abline(v=4/6,col=2,lty=2)
```

Huber's Process and Standardization

- A point can have high leverage or not, or have a standardized residual exceeding 2 (roughly) in magnitude or not
- From a leverage perspective, only high leverage points which are also outliers need actioning
- This is not to say outliers per se don't need actioning
- It is important that the standardized residuals used in determining whether a point of high leverage is good or bad are the individually standardized residuals and not globally standardized residuals. E.g. `Excel` and `SPSS` both use $\sqrt{\hat{\sigma}^2}$ and not $\sqrt{(1 - h_i)\hat{\sigma}^2}$ when "standardizing". `SPSS` and The Statistical Sleuth reserve the terminology "studentized" for their properly standardized residuals

Cook's Distance – Combining Leverage & Standardized Residual Information

- The good leverage point, bad leverage point process works well except when there is no relationship among the rest of data. E.g. In Anscombe's 4th data set the lone point has residual $= 0$ and leverage $= 1$

```
> cbind('y4'=anscombe$y4,'x4'=anscombe$x4,
+       'A4 res'=residuals(A4),'A4 hats'=hatvalues(A4))
      y4 x4 A4 res A4 hats
1   5.56  8 -1.441     0.1
2  12.50 19  0.000     1.0
3   5.25  8 -1.751     0.1
4   6.89  8 -0.111     0.1
5   5.76  8 -1.241     0.1
6   8.84  8  1.839     0.1
7   6.58  8 -0.421     0.1
8   8.47  8  1.469     0.1
9   7.91  8  0.909     0.1
10  7.71  8  0.709     0.1
11  7.04  8  0.039     0.1
```

Cook's Distance – Combining Leverage & Standardized Residual Information

- For a data point $(X_i, Y_i)$, Cook's Distance

$$D_i = p^{-1}(\text{standardized residual}_i)^2 \times \frac{h_i}{1 - h_i}$$

  where $p = \text{Regression df} + 1 = 2$ in simple linear regression
- Cook's Distance is large if a point has a large standardized residual, a large leverage or both
- If $D_i > 1$ the point is considered to have undue influence
- Calculate the predicted values from the regression including case $i$ and the predicted values from the regression excluding case $i$
- Calculate the sum of the squared differences between these predictions and divide by $p\hat{\sigma}^2$
- This is $D_i$, Cook's distance for case $i$

## Example: Huber's Data (Cook's Distances in R)

- ```
  > cooks.distance(H1)
           1          2          3          4          5          6
  0.28366861 0.08578246 0.01036426 0.17747401 0.13574522 0.11841263
  > (1/2)*rstandard(H1)^2*hatvalues(H1)/(1-hatvalues(H1))
           1          2          3          4          5          6
  0.28366861 0.08578246 0.01036426 0.17747401 0.13574522 0.11841263
  > cooks.distance(H2)
             1            2            3            4            5            6
   0.520037540  0.014641199  0.004693794  0.134536856  0.096271128 26.398591892
  > (1/2)*rstandard(H2)^2*hatvalues(H2)/(1-hatvalues(H2))
             1            2            3            4            5            6
   0.520037540  0.014641199  0.004693794  0.134536856  0.096271128 26.398591892
  ```

  The not quite bad but definitely naughty leverage point of model H2 is clearly overstepping Cook's line in the sand
  ```

Example: Huber's Data (Alternative Derivation of Cook's Distance)

- Alternatively you can get a raft of in influence measures, together with a flag if any overstep some line

```
> influence.measures(H2)

Influence measures of
 lm(formula = BadY ~ x, data = huber) :

   dfb.1_     dfb.x   dffit cov.r   cook.d   hat inf
1  1.1124 -9.56e-01  1.4667 0.329  0.52004 0.290   *
2  0.1261 -8.13e-02  0.1500 2.218  0.01464 0.236
3 -0.0775  3.33e-02 -0.0843 2.173  0.00469 0.197
4 -0.5320  1.14e-01 -0.5442 1.000  0.13454 0.174
5 -0.4361  1.78e-17 -0.4361 1.230  0.09627 0.167
6  8.5733  1.84e+01 20.3160 0.255 26.39859 0.936   *
```

## Example: Anscombe's Y4 on X4

```
> influence.measures(A4)
Influence measures of
 lm(formula = y4 ~ x4, data = anscombe) :

     dfb.1_   dfb.x4   dffit cov.r   cook.d hat inf
1  -0.25432  0.12769 -0.4235 0.974 8.39e-02 0.1
2   0.00000  0.00000     NaN   NaN      NaN 1.0   *
3  -0.32505  0.16320 -0.5413 0.795 1.24e-01 0.1
4  -0.01788  0.00898 -0.0298 1.403 4.98e-04 0.1
5  -0.21353  0.10721 -0.3556 1.078 6.23e-02 0.1
6   0.34734 -0.17439  0.5784 0.742 1.37e-01 0.1
7  -0.06827  0.03428 -0.1137 1.366 7.17e-03 0.1
8   0.26029 -0.13069  0.4334 0.958 8.72e-02 0.1
9   0.15149 -0.07606  0.2523 1.225 3.34e-02 0.1
10  0.11654 -0.05851  0.1941 1.294 2.03e-02 0.1
11  0.00628 -0.00315  0.0105 1.406 6.15e-05 0.1
```

## Simple Linear Regession

Example: Cargo Data

- Relationship between the volume of a ship's cargo loaded and unloaded ($X$) and the time in hours spent in port ($Y$)

```
> load("cargoes.RData")
> str(cargoes)
'data.frame': 31 obs. of  2 variables:
 $ Tonnage: num  268 294 329 353 363 507 529 536 547 663 ...
 $ Time   : num  11 13 13 15 20 11 11 22 20 13 ...
```

## Example: Cargo Data

- ```
  > cargoes[1:18,]
     Tonnage Time
  1      268   11
  2      294   13
  3      329   13
  4      353   15
  5      363   20
  6      507   11
  7      529   11
  8      536   22
  9      547   20
  10     663   13
  11     851    9
  12    1328   15
  13    1486   28
  14    1732   24
  15    1849   17
  16    2213   17
  17    2790   43
  18    2829   30
  ```
  ```

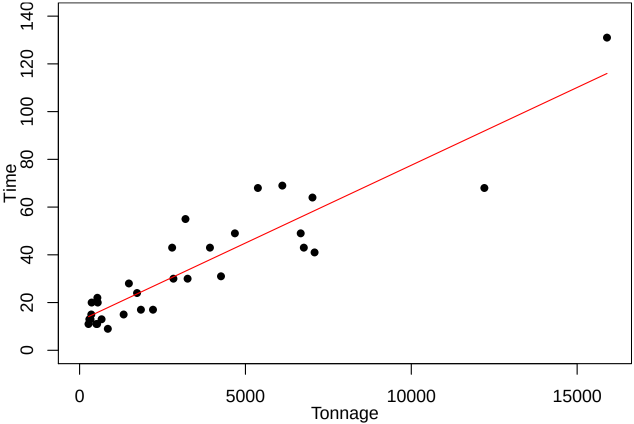
### Example: Cargo Data

- 
```
> cargoes[19:31,]
   Tonnage Time
19    3192   55
20    3256   30
21    3930   43
22    4263   31
23    4682   49
24    5375   68
25    6112   69
26    6666   49
27    6760   43
28    7021   64
29    7084   41
30   12203   68
31   15900  131
```

## Simple Linear Regession

Example: Cargo Data

- ```
  > with(cargoes,plot(Time~Tonnage,xlim=c(0,16000),ylim=c(0,140),pch=16))
  > with(cargoes,lines(fitted(lm(Time~Tonnage))~Tonnage,col=2))
  ```




Variance appears to be increasing and there are two points of obvious high leverage - but will they be good or bad?
  ```

## Example: Cargo Data

- ```
  > D1 = lm(Time~Tonnage,data=cargoes)
  > summary(D1)

  Call:
  lm(formula = Time ~ Tonnage, data = cargoes)

  Residuals:
      Min       1Q   Median       3Q      Max
  -23.882   -6.397   -1.261    5.931   21.850

  Coefficients:
               Estimate Std. Error t value Pr(>|t|)
  (Intercept) 12.344707   2.642633   4.671 6.32e-05 ***
  Tonnage      0.006518   0.000531  12.275 5.22e-13 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

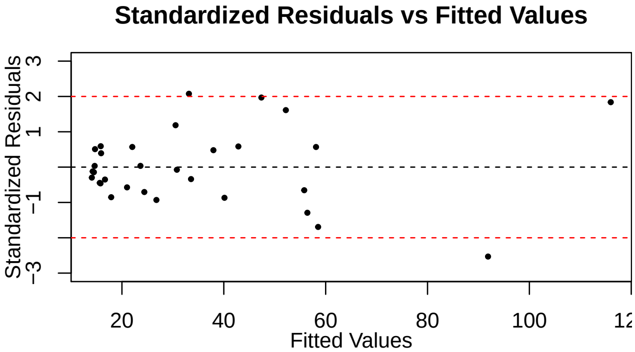
  Residual standard error: 10.7 on 29 degrees of freedom
  Multiple R-squared:  0.8386,Adjusted R-squared:  0.833
  F-statistic: 150.7 on 1 and 29 DF,  p-value: 5.218e-13
  ```

What part of this output can be trusted, apart from the arithmetic?
  ```

# Simple Linear Regession

## Example: Cargo Data

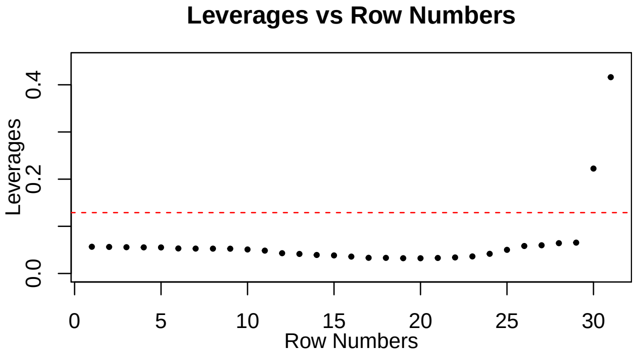
```
> plot(rstandard(D1)~fitted(D1),pch=21,bg="black",main="Standardized Residuals vs Fitted Values",
+      xlab="Fitted Values",ylab="Standardized Residuals")
> abline(h=c(-2,0,2),col=c(2,1,2),lty=2)
```




A clear case of increasing variance, with both high leverage points living dangerously. No obvious evidence of curvature.

## Example: Cargo Data

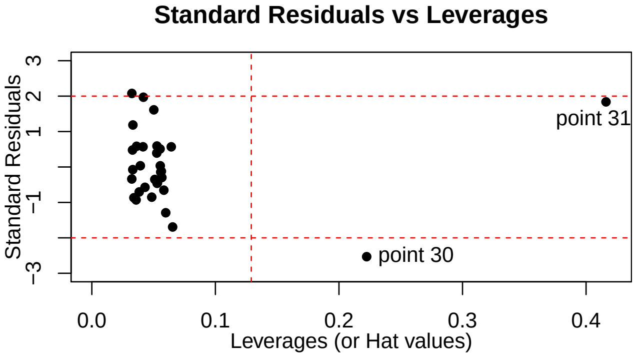
- ```
  > plot(hatvalues(D1)~c(1:31),pch=21,bg="black",main="Leverages vs Row Numbers",
  +      xlab="Row Numbers",ylab="Leverages")
  > abline(h=2*2/31,col=2,lty=2)
  ```




The line in the sand is $2 \times 2/31 = 0.129$.
  ```

## Example: Cargo Data

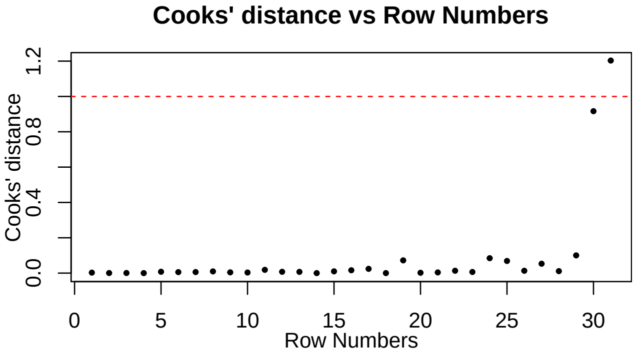
- ```
  > plot(rstandard(D1)~hatvalues(D1),pch=16,main="Standard Residuals vs Leverages",
  +       xlab="Leverages",ylab="Standard Residuals")
  > abline(h=c(-2,2),col=c(2,2),lty=2)
  > abline(v=2*2/31,col=2,lty=2)
  ```




Point 30 is definitely bad leverage point and 31 somewhat naughty to say the least.
  ```

Example: Cargo Data

- ```
  > plot(cooks.distance(D1)~c(1:31),pch=21,bg="black",main="Cooks' distance vs Row Numbers",
  +      xlab="Row Numbers",ylab="Cooks' distance")
  > abline(h=1,col=2,lty=2)
  ```



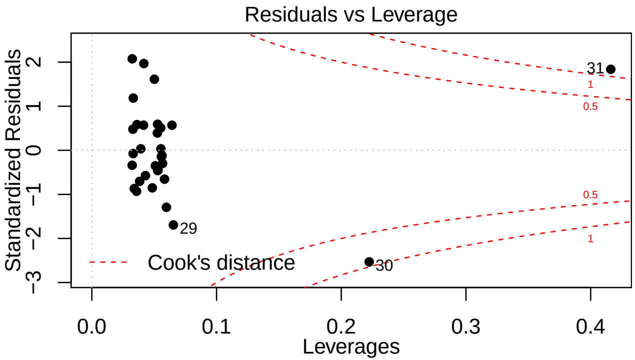

Point 31 is now in influential and 30 just under the radar. Both probably need a stern talking to.
  ```

# Simple Linear Regession

Example: Cargo Data

- One from the `plot(D1)` Stable

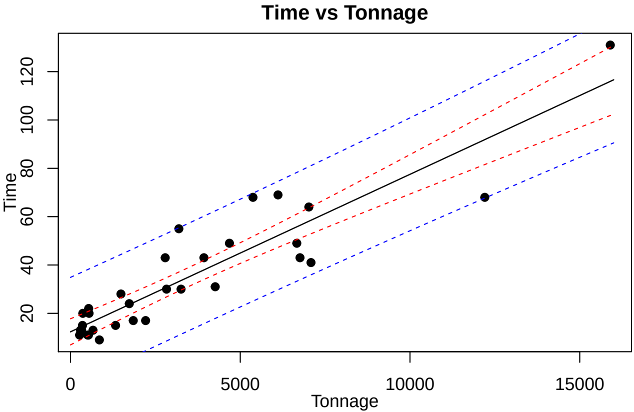
```
> plot(D1,5,add.smooth=FALSE,pch=16)
```




A combination standardized residual, leverage and Cook's distance plot. Points beyond the red lines are influential. (better to use `plot(D1)` to get all the plots instead)

Example: Cargo Data

- The plot




Variance is overestimated at low tonnages and underestimated at high tonnages. The latter is worse in the context. (What Can Be Done? Transformation)

## Simple Linear Regession

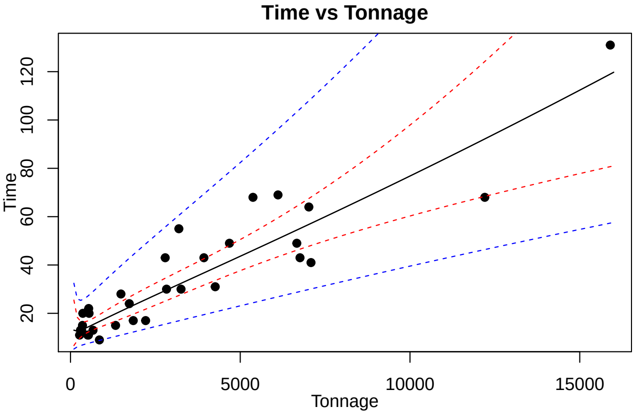
Example: Cargo Data

- The code

```
> plot(Time~Tonnage,data=cargoes,pch=21,bg="black",main="Time vs Tonnage")
> new = data.frame(Tonnage=seq(0,16000,100))
> CIs = predict(D1,new,interval="confidence")
> PIs = predict(D1,new,interval="predict")
> matpoints(new$Tonnage,CIs,lty=c(1,2,2),col=c("black","red","red"),type="l")
> matpoints(new$Tonnage,PIs,lty=c(1,2,2),col=c("black","blue","blue"),type="l"
```

Example: Cargo Data

- log transformation in both sides




Quadratic fit of `log(Time)` on `log(Tonnage)` transformed back to the original scale. The fitted line is almost identical but the estimates of precision very different. (need to check assumptions for sure)

Example:

- The code

```
> plot(Time~Tonnage,data=cargoes,pch=21,bg="black",main="Time vs Tonnage")
> new = data.frame(Tonnage=seq(0,16000,100))
> D1log = lm(log(Time)~log(Tonnage)+I(log(Tonnage)^2),data=cargoes)
> CIs = exp(predict(D1log,new,interval="confidence"))
> PIs = exp(predict(D1log,new,interval="predict"))
> matpoints(new$Tonnage,CIs,lty=c(1,2,2),col=c("black","red","red"),type="l")
> matpoints(new$Tonnage,PIs,lty=c(1,2,2),col=c("black","blue","blue"),type="l"
```

## Simple Linear Regession

Transforming $Y$ and/or $X$ variables ("O, that way madness lies" King Lear)

- That's what we do
  1. Increasing variance: Transform $Y$
     The Hierarchy of transforms:
     - square root – balanced but nasty
     - log – assertive but nice
     - inverse – aggressive and obnoxious
  2. Bad leverage points: Transform $X$ (also see 5.)
  3. Skewness in Standardized Residuals: Transform $Y$
     Often accompanies 1. together with presence of outliers
  4. Skewness in $X$ values: Transform $X$
     Often accompanies 2.
  5. Incorrect functional form: Transform $Y$ and/or $X$ or add polynomial terms to the model