Reviewing Basic Statistics I - Simple Linear Regression

Objectives

- Perform a simple linear regression with R
  - plot time series data
  - fit a linear model to a set of ordered pairs

The Mauna Loa CO_2 Data

```
plot(co2, main="Atmospheric CO2 Concentration")
```

- The response (i.e. CO_2 concentration) of the i^{th} observation may be denoted by the random variable Y_i
- This response depends upon the explanatory variable x_i in a linear way, with some noise added, as

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- error term ε_i
  - measurement error
  - lack of knowledge of other important influences,
  - etc.

- (Often reasonable!) assumptions:
  - the errors are normally distributed and, on average, zero;
  - the errors all have the same variance (they are homoscedastic), and
  - the errors are unrelated to each other (they are independent across observations).

$$Q = \sum(\text{observed} - \text{predicted})^2$$

$Y_i = i^{th}$ observed response variable

$\hat{Y}_i = i^{th}$ predicted response variable $=$ slope $\cdot x_i +$ intercept

- Develop your linear model:

```
(co2.linear.model = lm(co2 ~ time(co2)))
```

```
coefficients:
(Intercept)    time(co2)
-2249.774      1.307
```

- Plot your line with your data: (not a great model)

```
plot(co2, main = "Atmospheric CO2 Concentration with Fitted line")
  abline(co2.linear.model)
```

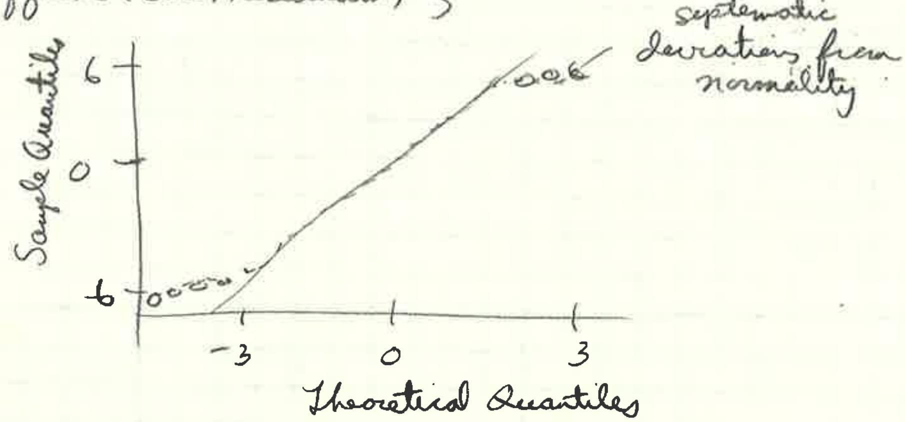
Reviewing Basic Statistics II: More Linear Regression

- Perform a simple linear regression with R

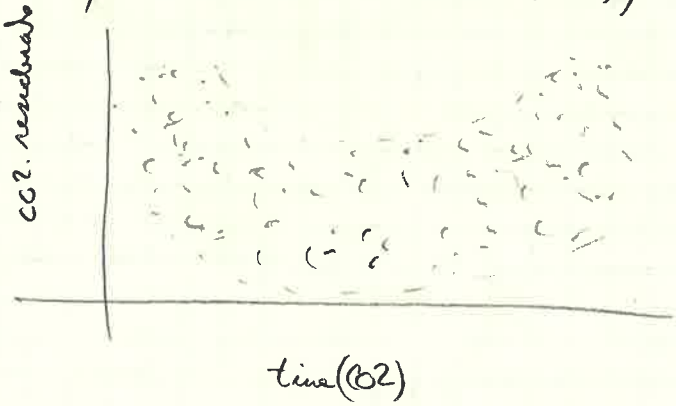
  - assess normality of residuals

```
co2.lm = lm(co2 ~ time(co2))

co2.residuals = resid(co2.lm)
hist(co2.residuals, main = "Histogram of Residuals")

qqnorm(co2.residuals)
qqline(co2.residuals)
```

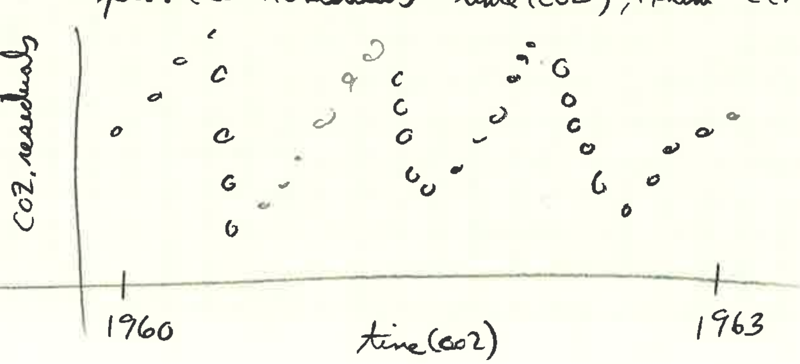
} Normal Q-Q Plot




```
plot(co2.residuals ~ time(co2))
```




```
plot(co2.residuals ~ time(co2), xlim = c(1960, 1963), main = "Zoomed resid")
```



Reviewing Basic Statistics III - Inference

Objectives:
- Develop a Graphical Intuition
- Perform a Hypothesis Test Concerning Means

- The Gossett Data

help(sleep)

- 20 observations on 3 variables
  - [,1] extra numeric increase in hours of sleep
  - [,2] group factor drug given
  - [,3] ID factor patient ID

- Plot your Data!

(boxplot)
- plot(extra ~ group, data = sleep, main = "Extra Sleep by Group")
- attach(sleep)
- extra.1 = extra[group == 1]
- extra.2 = extra[group == 2]

- Test your Hypothesis!

t.test(extra.1, extra.2, paired = TRUE, alternative = "two.sided")

Paired t-test:

- data: extra by group
- t = -4.0621, df = 9, p-value = 0.002833
- alternative hypothesis: true difference in means is not equal to 0
- 95% confidence interval (CI): [-2.4598858, -0.7001142]
- sample estimates: mean of the differences = -1.58

- Unpack this Output

H_0: Mean response is the same for both drugs $\Leftrightarrow \mu_{drug_1} - \mu_{drug_2} \equiv \mu_{diff} = 0$

H_1: Mean response is not the same for both drugs $\Leftrightarrow \mu_{drug_1} - \mu_{drug_2} = \mu_{diff} \neq 0$

$\alpha \equiv P(\text{Type I error}) = 0.05$

$$t = \frac{\bar{d} - 0}{sd/\sqrt{n}} = \frac{-1.58 - 0}{1.229995483/\sqrt{10}} = -4.062127683$$

sd of differences

$\bar{d} \equiv$ average of differences = difference of averages

$s_d \equiv$ standard deviation of differences

$n \equiv$ sample size

$p = 0.00283289$

$p = 2 * pt(-4.062127683, 9)$

$p < \alpha \Rightarrow$ reject H_0
$p > \alpha \Rightarrow$ do not reject H_0

- General Framework for Hypothesis Tests

- State clearly what your variables are (define your terms).
- State the null and alternative hypotheses.
- Decide upon a level of significance, α.
- Compute a test statistic (z, t, χ^2, F are popular).
- Find the p-value corresponding to your test statistic (for left/right/or two tailed test).
- Form a conclusion: if $p < \alpha$ (improbable data) reject H_0, otherwise do not reject. We typically do not accept, just like the courts never say that someone is innocent.

- Confidence Interval

A common form for a CI: Estimate ± Table Value • (Estimated) Standard Error

$$\bar{d} = \pm t_{\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}$$

- Our Data

$$-1.58 \pm \underset{qt(0.975,9)}{\underset{||}{2.262157}} \cdot \frac{1.229995483}{\sqrt{10}} = (-2.459686, -0.7001143)$$

- Recall: - standard error is the standard deviation of a sampling distribution.
  - statistic (something we compute from data)
  - parameter (a numerical descripted about a distribution or population),
  - Type I and Type II errors,
  - etc.

Reviewing Basic Statistics IV - Measuring Linear Association with the Correlation Function

Objectives:
- plot data pairwise to visually explore the associations between variables
- calculate and interpret covariance and correlation

- Girth, Height and Volume for Black Cherry Trees

```
> help(trees)
> pairs(trees, pch=21, bg=c("red"))
> cov(trees)
```

	Girth	Height	Volume
Girth	9.847914	10.38333	49.88812
Height	10.383333	40.60000	62.66000
Volume	49.888118	62.66000	270.20280

```
> cor(trees)
```

	Girth	Height	Volume
Girth	1.0000	0.5192801	0.9671194
Height	0.5192801	1.000	0.5982497
Volume	0.9671194	0.5982497	1.000

- Formulas

- For random variables, $\mathrm{cov}[X,Y] \equiv \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$
- For data sets, when we estimate covariance, $\mathrm{cov} \equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$
- For random variables, $\rho(X,Y) = \mathbb{E}\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]$
- For data sets, when we estimate correlation,

$$r \equiv \hat{\rho} \equiv \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

$$SSX \equiv \sum(x_i-\bar{x})^2 = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2$$

$$SSY \equiv \sum(y_i-\bar{y})^2 = \sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2$$

$$SSXY \equiv \sum(x_i-\bar{x})(y_i-\bar{y}) = \sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i$$

$$\Rightarrow \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{\sqrt{\frac{SSX}{n-1}}}\right)\left(\frac{y_i-\bar{y}}{\sqrt{\frac{SSY}{n-1}}}\right)$$