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)

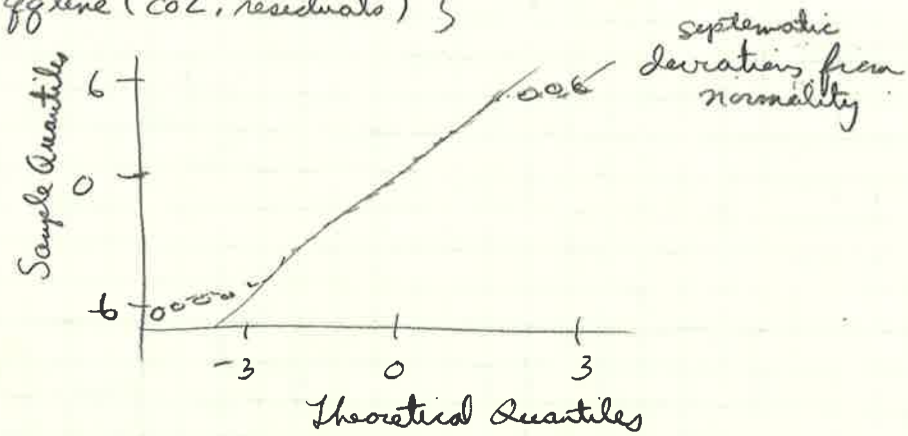
```
plot(co2, main = "Atmospheric CO2 Concentration with Fitted line")
abline(co2.linear.model)
```

Reviewing Basic Statistics II: More Linear Regression

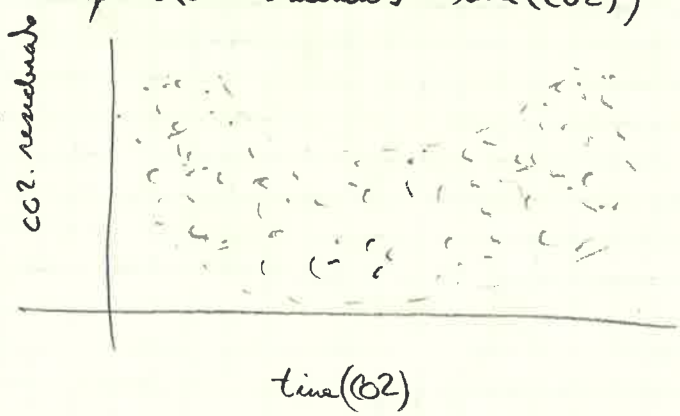
- Perform a simple linear regression with R
  - assess normality of residuals

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

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

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

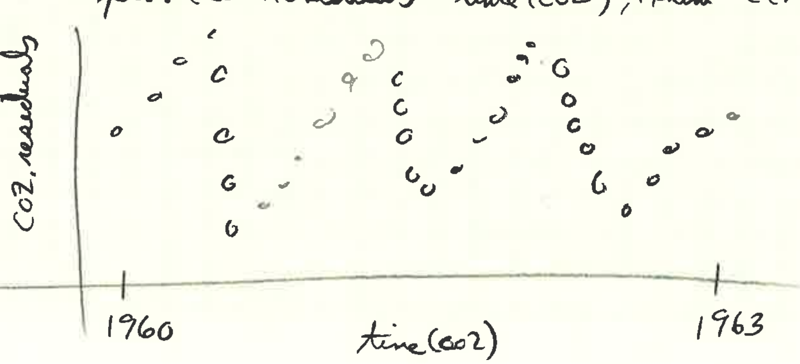
$\Big\}$ Normal Q-Q Plot




```
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

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

- Recall: - standard error is the standard deviation of a sampling distribution.
  - statistic (something we compute from data)
  - parameter (a numerical descriptor 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, $\operatorname{cov}[X,Y] \equiv \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$
- For data sets, when we estimate covariance, $\operatorname{cov} \equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$
- 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,

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

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

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

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

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

Week 2: Visualizing Time Series, and Beginning to Model T.S.

Notes for week 2 are in slide handouts

Week 3: Stationarity, MA(q) and AR(p) processes

Part 1: Stationarity: generalizing from an individual to a group.

Stationarity – Intuition and Definition

Objectives:
- Be able to explain why stationarity is crucial in formulating a model from data
- Find the mean, variance, and covariance function in a few simple stochastic processes

- Ensembles and Realizations

· A stochastic process is a complicated thing! To fully specify its structure we would need to know the joint distribution of the full set of r.v.'s.

· We usually just have one sequentially observed data set and must infer the properties of the generating process from this single trajectory.

- Mean, Variance, and Autocovariance Functions

Mean function: $\mu(t) \equiv \mu_t \equiv E[X(t)]$
Variance function: $\sigma^2(t) \equiv \sigma_t^2 \equiv V[X(t)]$

X_1	X_2	X_3	$\cdots$	X_N
$E[X_1]=\mu_1$	$E[X_2]=\mu_2$	$E[X_3]=\mu_3$		$E[X_N]=\mu_N$
$V[X_1]=\sigma_1^2$	$V[X_2]=\sigma_2^2$	$V[X_3]=\sigma_3^2$		$V[X_N]=\sigma_N^2$

· White Noise IID r.v.'s

Mean function: $\mu(t) = \text{const}$
Variance function: $\sigma^2(t) = \sigma^2 \text{(const)}$
Autocovariance Function: $\gamma(t_1, t_2) = \begin{cases} 0, & t_1 \neq t_2 \\ \sigma^2, & t_1 = t_2 \end{cases}$

- Estimation

How can we infer the properties of a stochastic process from a single realization?

- Strict Stationarity : Definition

We say a process is Strictly Stationary if the Joint Distribution of $X(t_1), X(t_2), \ldots, X(t_k)$ is the same as the Joint Distribution of $X(t_1+\tau), X(t_2+\tau), \ldots, X(t_k+\tau)$

- Strict Stationarity : Implications

Implication: Distribution of $X(t_1)$ same as Distribution of $X(t_1+\tau)$

Implication: The r.v.'s are identically distributed, though not necessarily independent

Implication: Mean function: $\mu(t) = \mu$
Variance Function: $\sigma^2(t) = \sigma^2$

Implication: Joint Distribution of $X(t_1), X(t_2)$ same as J.D. of $X(t_1+\tau), X(t_2+\tau)$, that is, the joint distribution depends only on the lag spacing, so

Autocovariance Function: $\gamma(t_1, t_2) = \gamma(t_2 - t_1) = \gamma(\tau)$
(ACF)

- Weak Stationarity : Definition

We say a process is weakly stationary if
Mean Function: $\mu(t) = \mu$
ACF: $\gamma(t_1, t_2) = \gamma(t_2 - t_1) = \gamma(\tau)$

Implication: Constant Variance
So much easier, but still useful!

Stationarity - First Examples... White Noise and Random Walks

Objectives: Develop some examples of Stationary Processes: white noise, random walks, introduction to moving averages.

- White Noise is Stationary!

Consider a discrete family of iid normal r.v.'s (often Gaussian)

$X_t \sim iid\,(0, \sigma^2)$

$X_t \sim iid\,N(0, \sigma^2)$

3.3

Mean function $\mu(t) = 0$ is obviously constant, so consider $\gamma(t_1, t_2) = \begin{cases} 0, & t_1 \neq t_2 \\ \sigma^2, & t_1 = t_2 \end{cases}$

- Random Walks are not stationary!
  - Start with IID r.v.'s $Z_t \sim iid(\mu, \sigma^2)$. Build a walk with t steps:

$$X_1 = Z_1$$
$$X_2 = X_1 + Z_2 = Z_1 + Z_2$$
$$X_3 = X_2 + Z_3 = Z_1 + Z_2 + Z_3$$
$$\vdots$$
$$X_t = X_{t-1} + Z_t = \sum_{i=1}^{t} Z_i$$

$$E[X_t] = E\left[\sum_{i=1}^{t} Z_i\right] = \sum_{i=1}^{t} E[Z_i] = t\mu$$
$$V[X_t] = V\left[\sum_{i=1}^{t} Z_i\right] = \sum_{i=1}^{t} V[Z_i] = t \cdot \sigma^2$$

Notes: Independent r.v.'s have variances which add. All r.v.'s have means which add

- Moving Average Processes are Stationary!
  Start with iid r.v's $Z_t \sim iid(0, \sigma^2)$.
  MA(q) process: $X_t = \beta_0 Z_t + \beta_1 Z_{t-1} + \cdots + \beta_q Z_{t-q}$

  q tells us how far back to look along the white noise sequence for our weighted average.

Stationarity - First examples ... ACF of a Moving Average

- Objectives: Develop the ACF of a Moving Average Process

  - Moving Average Processes are Stationary (cont'd)!
  Look at the covariance at two locations along a MA process:

$$cov[X_t, X_{t+k}] = E[X_t X_{t+k}] - E[X_t]E[X_{t+k}]$$
$$E[X_t] = E[X_{t+k}] = 0 \Rightarrow cov[X_t X_{t+k}] = E[X_t X_{t+k}]$$
$$cov[X_t X_{t+k}] = E[(\beta_0 Z_t + \cdots + \beta_q Z_{t-q}) \cdot (\beta_0 Z_{t+k} + \cdots + \beta_q Z_{t+k-q})]$$

Intuition: Since the underlying Z_t are independent, we shouldn't get contributions to the covariance except where X_t and X_{t+k} share building blocks.

3.4

More formally, consider: $\operatorname{cov}[X_t, X_{t+k}] = E\left[(\beta_0 Z_t + \dots + \beta_q Z_{t-q})(\beta_0 Z_{t+k} + \dots + \beta_q Z_{t+k-q})\right]$

Now, expand the product:

$$E\left[(\beta_0 Z_t + \dots + \beta_q Z_{t-q})(\beta_0 Z_{t+k} + \dots + \beta_q Z_{t+k-q})\right]$$

$$= E\left[\begin{array}{c} \beta_0\beta_0 Z_t Z_{t+k} + \beta_0\beta_1 Z_t Z_{t+k-1} + \dots + \beta_0\beta_q Z_t Z_{t+k-q} \\ + \\ \beta_1\beta_0 Z_{t-1} Z_{t+k} + \beta_1\beta_1 Z_{t-1} Z_{t+k-1} + \dots + \beta_1\beta_q Z_{t-1} Z_{t+k-q} \\ + \dots + \\ \beta_q\beta_0 Z_{t-q} Z_{t+k} + \beta_q\beta_1 Z_{t-q} Z_{t+k-1} + \dots + \beta_q\beta_q Z_{t-q} Z_{t+k-q} \end{array}\right]$$

When the subscripts in the products agree, we get a contribution. When the subscripts disagree we get 0. If $k > q$, the r.v.'s are too far away to get a contribution.

Intuition: $k=0$

$$E\left[\begin{array}{c} \beta_0\beta_0 Z_t Z_t + \beta_0\beta_1 Z_t Z_{t-1} + \dots + \beta_0\beta_q Z_t Z_{t-q} + \beta_1\beta_0 Z_{t-1} Z_t + \beta_1\beta_1 Z_{t-1} Z_{t-1} + \dots + \beta_1\beta_q Z_{t-1} Z_{t-q} \\ + \dots + \\ \beta_q\beta_0 Z_{t-q} Z_t + \beta_q\beta_1 Z_{t-q} Z_{t-1} + \dots + \beta_q\beta_q Z_{t-q} Z_{t-q} \end{array}\right]$$

$$= \beta_0\beta_0\sigma^2 + 0 + \dots + 0 + 0 + \beta_1\beta_1\sigma^2 + 0 + \dots + 0 + \dots + 0 + 0 + \dots + \beta_q\beta_q\sigma^2$$

$$= \sigma^2 \sum_{i=0}^{q} \beta_i^2$$

Intuition: $k=1$

$$E[\cdots] = \sigma^2 \sum_{i=0}^{q-1} \beta_i \beta_{i+1}$$

Intuition: $k=q$

$$E[\cdots] = \sigma^2 \beta_0 \beta_q$$

Generic $k \le q$

$$\operatorname{Cov}[X_t, X_{t+k}] = \sigma^2 \sum_{i=0}^{q-k} \beta_i \beta_{i+k} \quad (\text{no } t \text{ dependence})$$