Data Set: annualtemp.csv

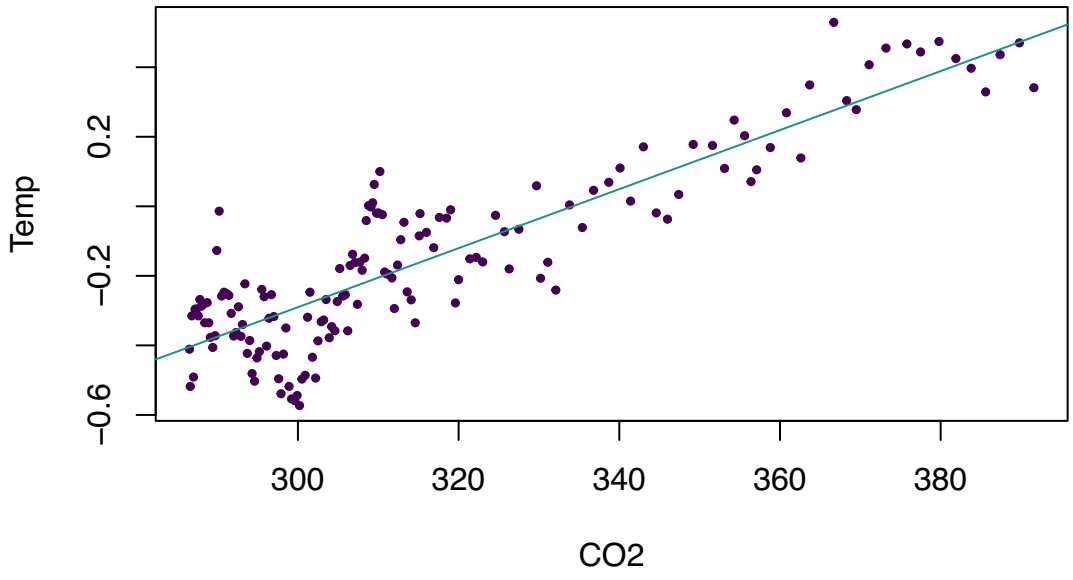
```
> y = read.csv("annual_temp.csv", header=TRUE)
> head(y,3)
```

Sea Surface Temperature

	Year	Temp	CO2	CH4	NO2	Irradiance	Nino_SST	Volcano
1	1861	-0.411	286.5	838.2	288.9	1361.097	26.74233	0.00281
2	1862	-0.518	286.6	839.6	288.9	1360.987	26.39426	0.00859
3	1863	-0.315	286.8	840.9	289.0	1360.837	26.16013	0.01318

Check: Relationship between CO_2 & Temperature

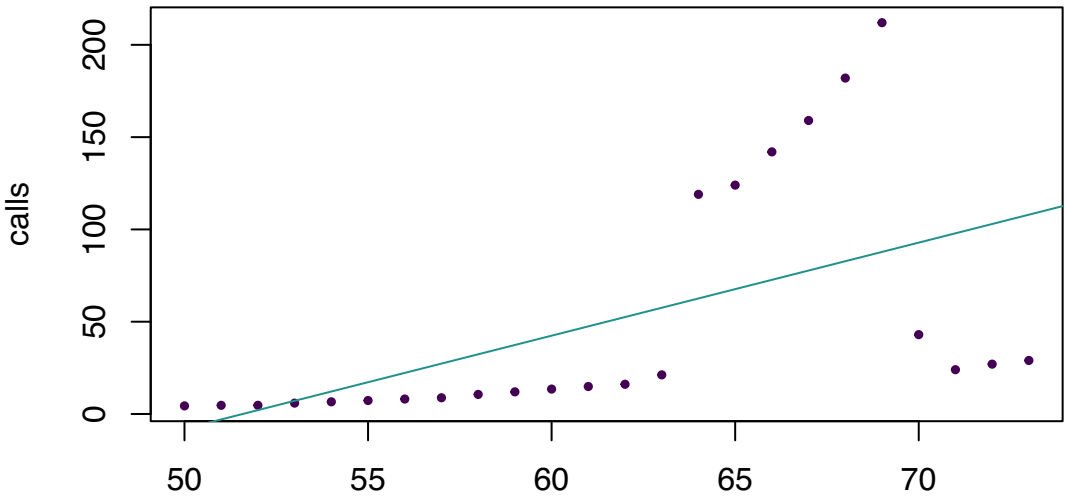
```
> plot(Temp ~ CO2,  data=y, pch=19,cex=.5, col="#440154")
> abline(lm(Temp ~ CO2, data=y), col="#21918c")
```



Ex: Do summary(lm(Temp ~ CO2, data=y)) in R.

Simple Linear Regression: Belgium Phone Calls

```
> require(MASS)
> plot(calls ~ year,  data=phones, pch=19,cex=.5, col="#440154")
> abline(lm(calls ~ year, data=phones), col="#21918c")
```



- Clearly data between 64-69, reflect a different relationship.
- One could try other lines (i.e. mean-deviation line) ...

Simple Linear Regression

Solve:

$\underset{m,c}{\operatorname{argmin}} \sum_{i=1}^{n} |y_i - mx_i - c|$

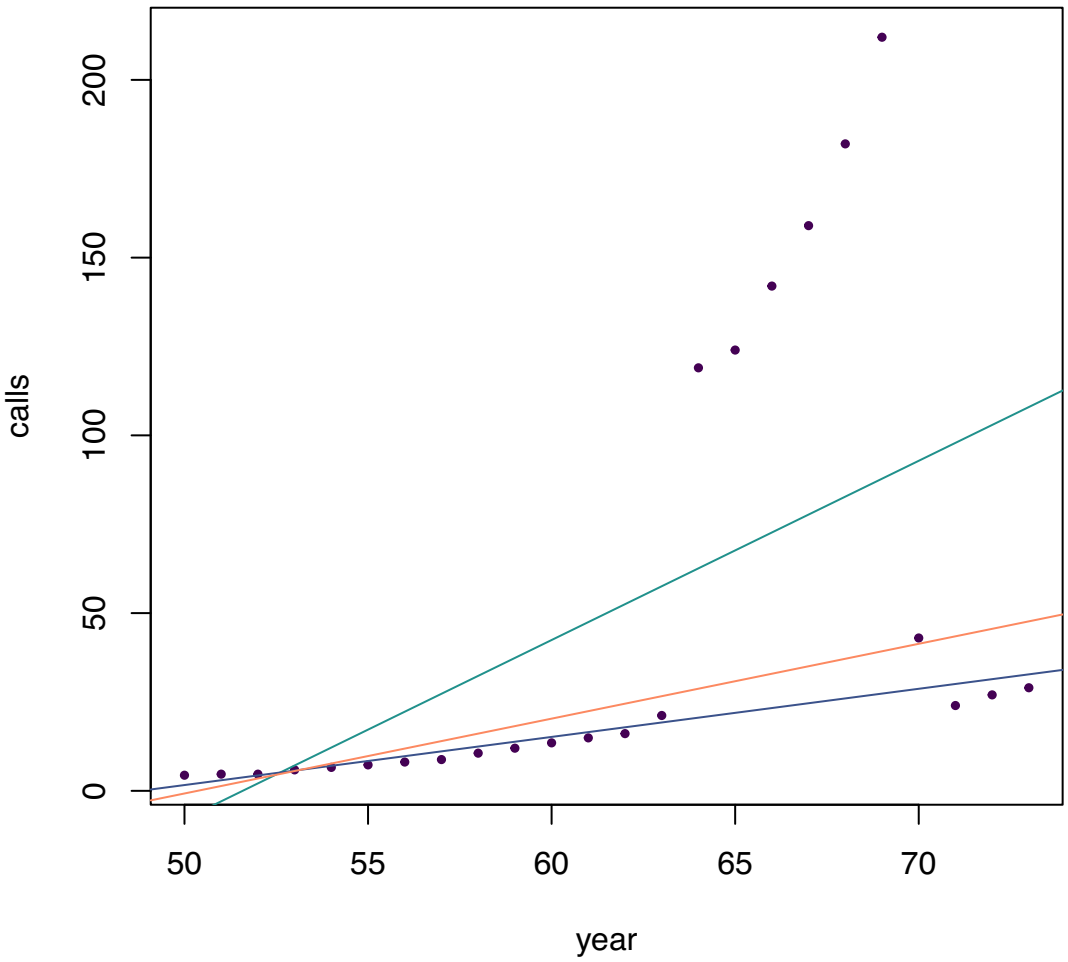
- Our own absolute deviation line:

```
> ABSMINLINE = function(x)
+ { with (phones, sum(abs(calls- x[1] -x[2]*year)))
+ }
> OPTIMAL = optim(c(0,0), fn = ABSMINLINE)
```

- Plotted `lm`, `rlm`, `absline`

```
> abline(lm(calls ~ year, data=phones), col="#21918c")
> abline(OPTIMAL$par, col="#3b528b")
> abline(rlm(calls ~ year, data=phones), col="#fc8961")
```

Ex: check in R about this line

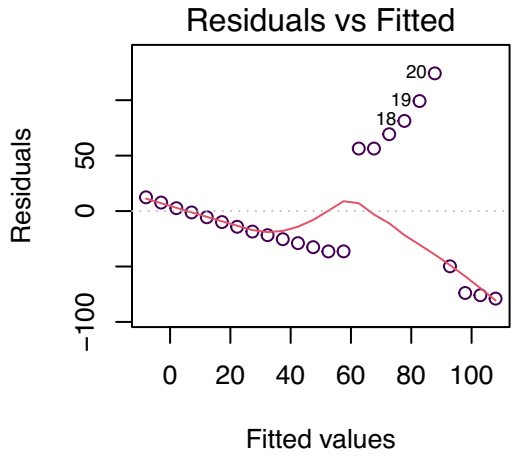


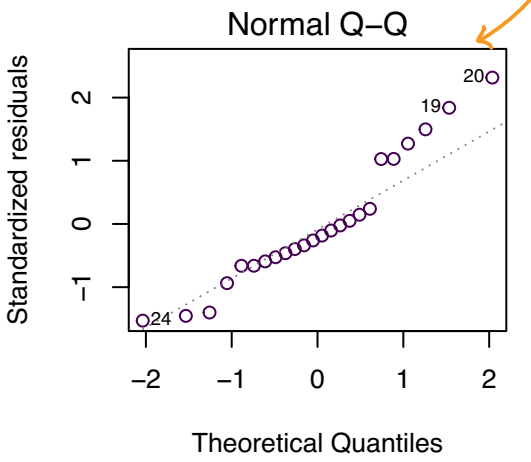
are there better fits ?: Shown you these four graphs.

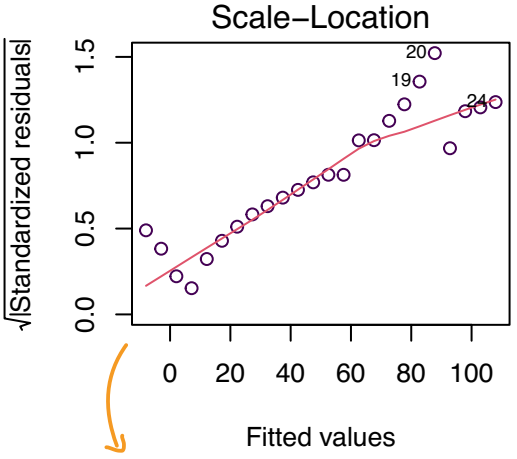
checks if Residuals have linear pattern

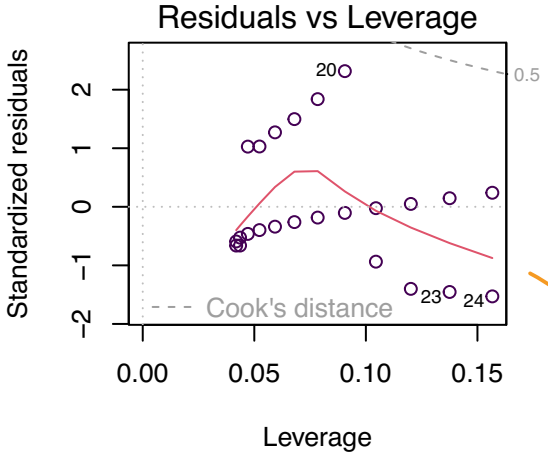
```
> par(mfrow=c(2,2))
> plot(lm(calls ~ year, data=phones), col="#440154")
```

checks if residuals are normal (ish) distributed









Homoscedasticity

checks if variance among residuals is same.

- line needs to be horizontal

Points that lie outside dashed lines (Cook distance) are influential

Simple Linear Regression

- Model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

  with ε_i being i.i.d Normal$(0, \sigma^2)$.

- Estimators:
  - $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
  - $\hat{\beta}_1 = \dfrac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \text{correlation}(x, y) \dfrac{S_y}{S_x}$

$$= \sum_{i=1}^{n} \frac{x_i}{S_x^2} (y_i - \bar{y}) \quad \Rightarrow \hat{\beta}_1 \text{ is Normally distributed}$$

- $E[\hat{\beta}_1] = \sum_{i=1}^{n} \frac{x_i}{S_x^2} (\beta_0 + \beta_1 x_i - \beta_0 - \beta_1 \bar{x})$

$$= \beta_1$$

- $\text{Var}[\hat{\beta}_1] = \frac{\sigma^2}{S_x^2 n}$ Ex

- if

$$\text{RSS} \equiv \text{Residual Sum of Squares} := \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Then one can show:

- $E[RSS] = (n-2)\sigma^2$

and consequently that

- $$\frac{(n-2)\text{RSS}}{\sigma^2} \sim \chi^2_{n-2}$$

It turns out that

- RSS and $\hat{\beta}_1$ are independent

$\Rightarrow$

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\text{RSS}}{S_{xx}^2}}} \sim t_{n-2}.$$

We can now test for β_1!

Simple Linear Regression: Testing

- Test Statistic:

$$T := \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\text{RSS}}{S_{xx}^2}}}$$

- Null Hypothesis: $\beta_1 = 0$ versus Alternate Hypothesis: $\beta_1 \neq 0$

- Decide on level of significance: α

- Compute p-value:

$$\mathbb{P}(\mid t_{n-2} \mid \geq \mid T \mid)$$

- Reject Null Hypotheis:

if p-value is less than α

and Confidence Interval for β_1:

The interval

$$\left(\hat{\beta}_1 - t_{n-2}(0.25)\sqrt{\frac{\text{RSS}}{S_{xx}^2}}, \hat{\beta}_1 + t_{n-2}(0.25)\sqrt{\frac{\text{RSS}}{S_{xx}^2}}\right)$$

–the 95% confidence interval for β_1 with

$$\mathbb{P}(t_{n-2} \geq t_{n-2}(0.25)) = 0.025.$$

Paired Samples

Test of Means:

- Let $n \geq 1$, $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ be two samples, that are paired. Then the test Statistic:

$$T := \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_Y)}{\frac{S}{\sqrt{n}}}$$

where

$$S^2 = \frac{1}{n-1}(\sum_{i=1}^{n}(z_i - \bar{z})^2) \qquad \text{with } z_i = x_i - y_i$$

- $T \sim t_{n-1}$
- We can think of it as single sample test of equality of means with data from $x_i - y_i$.

Two Sample Difference in Proportion-test

- Want to test if proportion of success $p_1 = p_2$ between two populations.
- Let $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$
- Under Null Hypothesis the statistic is

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}},$$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Large n_1, n_2 assume normality for Z

Two sample test of means, unequal sample size

Let $n, m \geq 1$, $X_1, X_2, \ldots, X_n$ be i.i.d. Normal(μ_X,σ_1^2) and $Y_1, Y_2, \ldots, Y_m$ be i.i.d. Normal(μ_Y,σ_2^2).

Variance Unknown but assumed equal, $\sigma_1 = \sigma_2$:

- Test Statistic:

$$T := \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_Y)}{S_{pooled}\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{\text{Observed} - \text{Expected}}{\text{Standard Error}}$$

where

$$S_{pooled}^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

- $T \sim t_{n+m-2}$

Variance Unknown but assumed unequal, $\sigma_1 \neq \sigma_2$, (Welch approximation)

- Test Statistic:

$$T := \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_Y)}{S_{pooled}}$$

where

$$S_{pooled}^2 = \frac{S_X^2}{n} + \frac{S_Y^2}{m}$$

- $T \sim t_d$ with $d = \left(\frac{S_X^2}{n} + \frac{S_Y^2}{m}\right)^2 \cdot \left(\frac{1}{n-1}\left(\frac{S_X^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{S_Y^2}{m}\right)^2\right)^{-1}$.