**Modeling: Thinking about Assumptions and Checking Conditions**

data=read.table("C:/data R/percentfat.txt",header=TRUE)

attach(data)

plot(Pctfat~waist)



**Linearity Assumption (Check with Straight Enough Condition)**

Note: if the “true” relationship is far from linear our entire analysis will be useless.

**Independence Assumption**

The errors in the true underlying regression model must be mutually independent.

Usually when we care about inference (tests, CIs and PIs), it’s because our regression model might apply to a larger population. In this case we want to check the **Randomization Condition** (that the individuals are a representative sample from that population).

We can also check displays of the regression residuals for evidence of patterns, trends, clumping, any of which will suggest a failure of independence.

If the explanatory is related to time, a common violation of the Independence Assumption is for the errors to be correlated.

data=read.table("C:/data R/percentfat.txt",header=TRUE)

attach(data)

plot(Pctfat~waist)

model=lm(Pctfat~waist)

par(mfrow=c(2,1))

plot(Pctfat~waist)

plot(resid(model)~waist)



I can also use a time series plot in order to look patterns that “override” the Independence Assumption.

data=read.table("C:/data R/percentfat.txt",header=TRUE)

attach(data)

plot(Pctfat~waist)

model=lm(Pctfat~waist)

par(mfrow=c(2,1))

plot(resid(model)~waist)

abline(lm(resid(model)~waist))

plot(ts(resid(model)))



**Equal Variance Assumption (Check with “Does the Plot Thicken? Condition).**

The variability of y should be the same for all values of x. We will take a look at RMSE (the standard deviation of the residuals). We need the RMSE to build CIs, PIs and Tests. The standard deviation of the residuals is the building block for the standard errors of all regression estimators.

Now only if the scatter of the residuals is the same everywhere. The RMSE “pools” info across all the individual distributions at each x-value, and pooled estimates are appropriate only if they combine information for groups with the same variance.

Practically we can check: **Does the Plot Thicken? Condition**

**Make a residual plot against the fitted values:**

data=read.table("C:/data R/percentfat.txt",header=TRUE)

attach(data)

plot(Pctfat~waist)

model=lm(Pctfat~waist)

par(mfrow=c(3,1))

plot(Pctfat~waist)

plot(resid(model)~waist)

plot(resid(model)~fitted(model))



**Normal Population Assumption**

We assume the errors around the idealized regression line (the beta notation) at each value of x follow a Normal model. We need this assumption so that we can use a t-distribution (procedures) for inference.

As we have other times when we’ve used t procedures, we will settle for the residuals satisfying the **Nearly Normal Condition and the Outlier Condition.**

**Look at a Histogram and Normal Probability Plot of the residuals.**

data=read.table("C:/data R/percentfat.txt",header=TRUE)

attach(data)

plot(Pctfat~waist)

model=lm(Pctfat~waist)

par(mfrow=c(2,2))

plot(Pctfat~waist)

plot(resid(model)~fitted(model))

hist(resid(model))

qqnorm(resid(model))



data=read.table("C:/data R/percentfat.txt",header=TRUE)

attach(data)

plot(Pctfat~waist)

model=lm(Pctfat~waist)

par(mfrow=c(2,2))

plot(Pctfat~waist)

plot(resid(model)~fitted(model))

abline(lm(residuals(model)~fitted(model)))

hist(resid(model))

qqnorm(resid(model))





***We can make it better than Minitab.***

data=read.table("C:/data R/percentfat.txt",header=TRUE)

attach(data)

model=lm(Pctfat~waist)

par(mfrow=c(2,3))

plot(Pctfat~waist)

plot(resid(model)~waist)

abline(lm(resid(model)~waist))

plot(ts(resid(model)))

plot(resid(model)~fitted(model))

abline(lm(residuals(model)~fitted(model)))

hist(resid(model))

qqnorm(resid(model))

