Lecture 4

- Diagnostics for residuals

Diagnostics for residuals

- Model: $Y_i = \beta_0 + \beta_1 X_i + \xi_i$
- Predicted values: $\hat{Y}_i = b_0 + b_1 X_i$
- Residuals: $e_i = Y_i - \hat{Y}_i$
- So, $Y_i = \hat{Y}_i + e_i$
- The e_i should be similar to the ξ_i
- The model assumes ξ_i iid $N(0, \sigma^2)$

Plot Plot

PLOT
PLOT
PLOT

Plot Plot

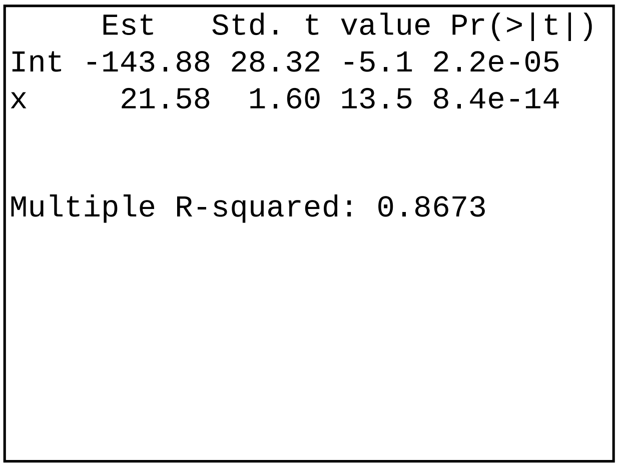
Questions addressed by diagnostics for residuals

- Is the relationship linear?
- Does the variance depend on X?
- Are there outliers?
- Do the errors depend on order (_n_)
- Are the errors normal?
- Are the errors dependent?

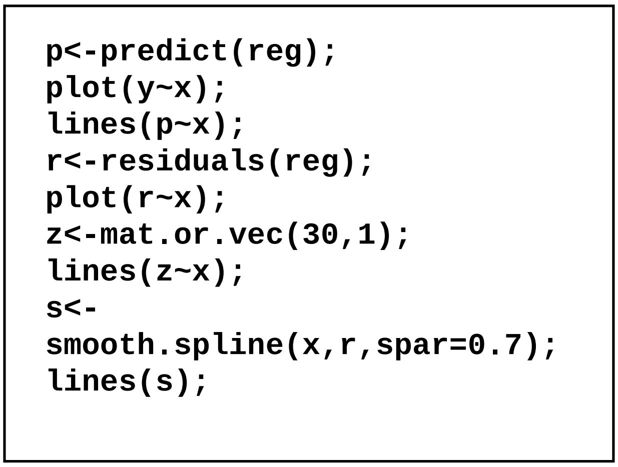
Is the Relationship Linear?

- Plot Y vs X
- Plot e vs X
- Plot of e vs X will emphasize deviations from linear pattern

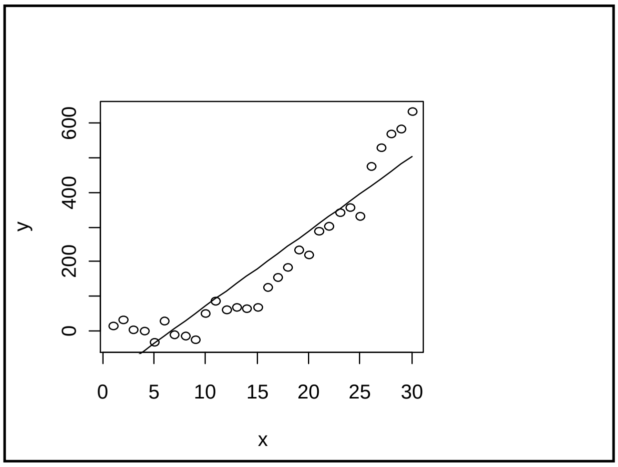
```
x<-seq(1:30);
e<-rnorm(30);
y<-x^2-10*x+30+25*e;
reg<-lm(y~x);
summary(reg);
```

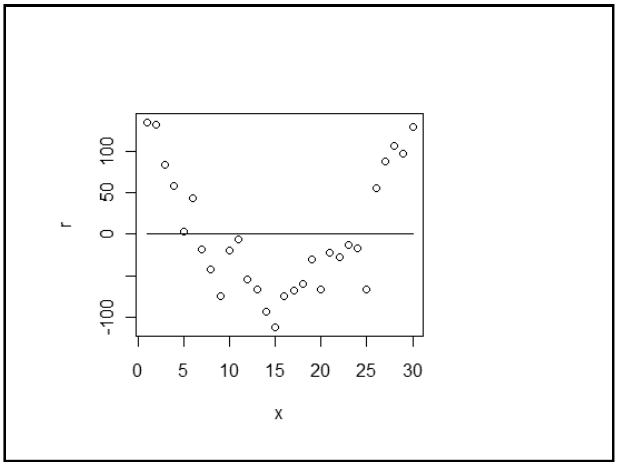


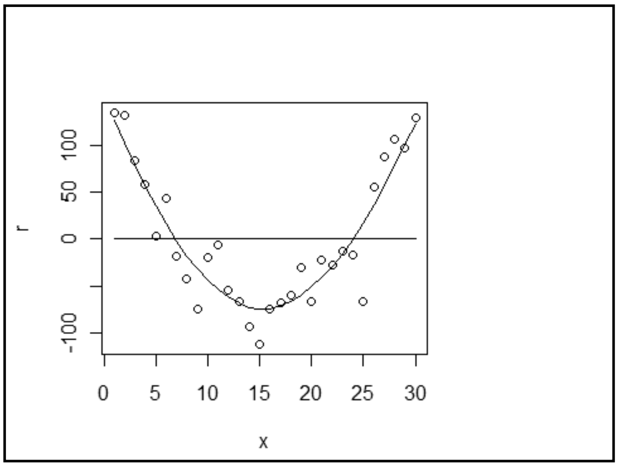
```
     Est   Std. t value Pr(>|t|)
Int -143.88 28.32 -5.1 2.2e-05
x      21.58  1.60 13.5 8.4e-14


Multiple R-squared: 0.8673
```

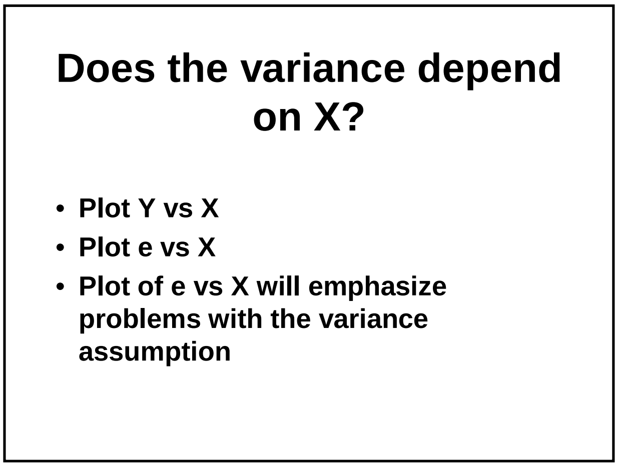


```
p<-predict(reg);
plot(y~x);
lines(p~x);
r<-residuals(reg);
plot(r~x);
z<-mat.or.vec(30,1);
lines(z~x);
s<-
smooth.spline(x,r,spar=0.7);
lines(s);
```







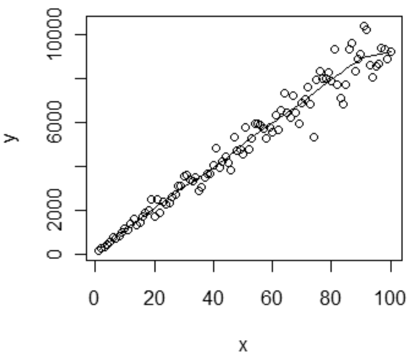


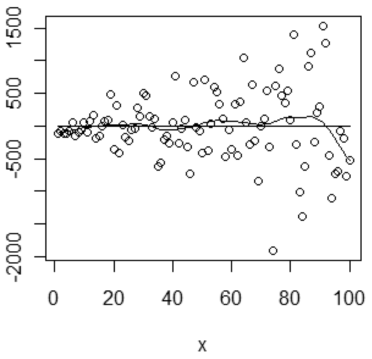
Does the variance depend on X?

- **Plot Y vs X**
- **Plot e vs X**
- **Plot of e vs X will emphasize problems with the variance assumption**

```
x<-seq(1:100);
y<-100*x+30+10*x*rnorm(100);
reg<-lm(y~x);
r<-residuals(reg);

plot(y~x);
s<-smooth.spline(x,y, spar=0.7);
lines(s);
```

```
plot(y~x);
s<-smooth.spline(x,y, spar=0.7);
lines(s);
plot(r~x);
z<-mat.or.vec(100,1);
lines(z~x);
s<-smooth.spline(x,r, spar=0.7);
lines(s);
```

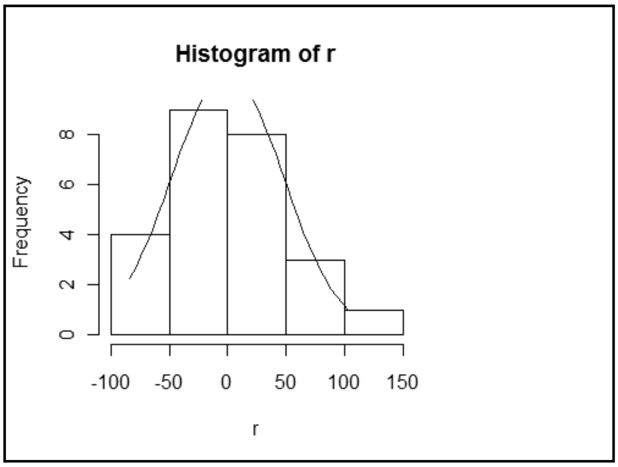


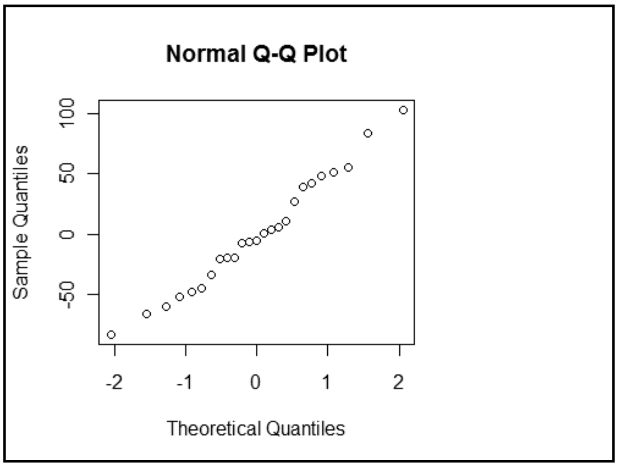


Are the errors normal?

- The *real* question is whether the distribution of the errors is far enough away from normal to invalidate our confidence intervals and significance tests
- Look at the distribution of the residuals
- Use a normal quantile plot

```
time<-read.table('CH01TA01.txt',
col.names=c("size", "hours"));
reg1<-lm(hours~size, time);
r<-residuals(reg1);
h<-hist(r);m<-mean(r);s<-sd(r);
xfit<-
seq(min(r),max(r),length=40);
d<-dnorm(xfit,m,s);
d <-
d*diff(h$mids[1:2])*length(r)
lines(d~xfit, col='blue');
qqnorm(r)
```





Dependent Errors

- Usually we see this in a plot of residuals vs time order
- We can have trends and/or cyclical effects

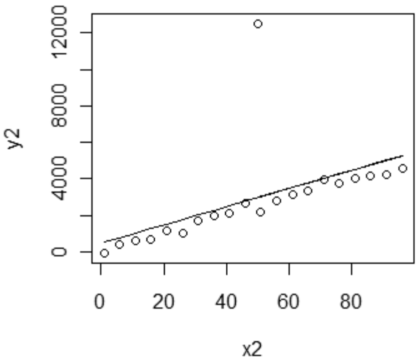
Are there outliers?

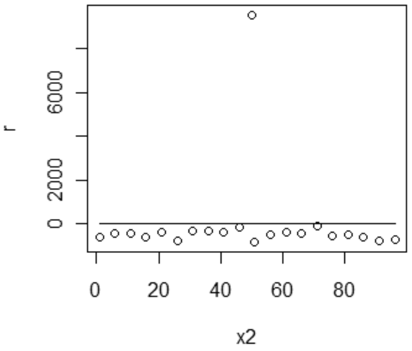
- Plot Y vs X
- Plot e vs X
- Plot of e vs X should emphasize an outlier

```
x<-seq(from=1, to=100, by=5);
y<-30+50*x+200*rnorm(20);
x1<-50;
y1<-30+50*50+10000;
x2<-c(x,x1);
y2<-c(y,y1);
reg1<-lm(y~x);
reg2<-lm(y2~x2);
summary(reg1);
summary(reg2);
```

```
   Est    Std.    t Pr(>|t|)
Int 44.92 88.71  0.51 0.619
x   49.08  1.57 31.22 <2e-16 **

Int 480.73 981.37 0.49 0.6298
x2   49.94  17.48 2.86 0.0101 *

Residual standard error: 202.7
Residual standard error: 2254
```

```
p<-predict(reg2);
plot(y2~x2);
lines(p~x2);
r<-residuals(reg2);
plot(r~x2);
z<-mat.or.vec(21,1);
lines(z~x2);
```



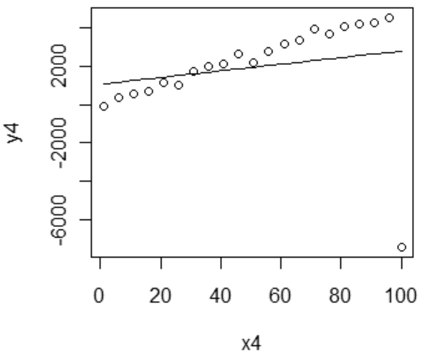


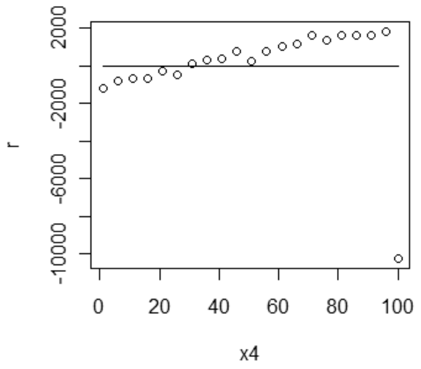
Different kinds of outliers

- **The outlier in the last example *influenced* the intercept**
- **but not the slope**
- **It inflated all of our standard errors**
- **Here is an example of an outlier that *influences* the slope**

```
x3<-100;
y3<-30+50*50-10000;
x4<-c(x,x3);
y4<-c(y,y3);
reg3<-lm(y4~x4);
summary(reg3);
```

```
    Est    Std.    t Pr(>|t|)
Int 44.92 88.71  0.51 0.619
x   49.08  1.57 31.22 <2e-16 ***


     Est     Std.    t  Pr(>|t|)
Int 1074.47 1112.32 0.966 0.346
x4    17.26   18.78 0.919 0.370
```





Outliers

- An outlier can be *influential* for the estimation of some model parameters
- And not influential for others
- Outliers usually inflate the variance and decrease our chances of finding statistically significant results

Other plots

- You can also look at
  - the stem plot and
  - the boxplot

More plots

- Plot residuals vs
  - Time (order)
  - Other candidate explanatory variables
- Look for
  - Overall patterns
  - outliers

Significance tests for normality

- H_0: data are an iid sample from a single normal population
- H_1: data are *not* an iid sample from a single normal population

Significance tests for normality? (2)

We have many choices for a significance testing procedure

- **Shapiro-Francia is a good choice**

```
library(nortest)
sf.test


Shapiro-Francia normality
test

data:  r
W = 0.9831, p-value = 0.8807
```

Other significance tests for model assumptions

- **Durbin-Watson test for serially correlated errors (dwtest {lmtest})**
- **Breusch-Pagan test for homogeneity of variance (bptest{lmtest})**

Comments on plots vs significance tests for model assumptions

- **Plots are more likely to lead to a remedy**
- **Significance tests results are very dependent on the sample size; with sufficiently large samples we can reject most null hypotheses**

Lack of fit

- **When we have repeated observations at different values of X, we can do a significance test for nonlinearity**
- **We will do details when we get to ANOVA**
- **Basic idea is to compare two models**
- **Plot with a smoothing function is usually a better approach**

Nonlinear relationships

- **We can model many nonlinear relationships with linear models, some have several explanatory variables (multiple linear regression)**
  - **Quadratic $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \xi$**
  - **$Y = \beta_0 + \beta_1 \log(X) + \xi$**

Nonlinear (2)

- Sometimes we transform a nonlinear problem into a linear form
- eg if Y = $\beta_0 \exp(\beta_1 X) + \xi$
- we could consider the linear model
- $\log(Y) = \log(\beta_0) + \beta_1 X + \xi$
- Note that we have changed our assumption about the error

Nonlinear (3)

- We can perform a nonlinear regression analysis
- R PROC NLS

Non constant error variance

- Sometimes we model the way in which the error variance changes (eg it may be linearly related to X)
- We can use a weighted analysis
- Use a weight option in PROC LM

Non normal errors

- Transformations often help
- Use a procedure that allows different distributions for the error term
- SAS PROC GLM

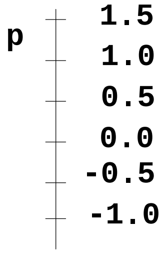
GLM (1)

- Possible distributions of Y:
- Binomial (binary data)
- Poisson
- Gamma (exponential)
- Inverse gaussian

GLM (2)

- Allows you to specify the link function g(μ)=EY in the model
- $g(\mu)=b_0+ b_1x_1+ \ldots+b_kx_k$

Ladder of Reexpression (transformations)

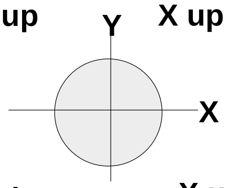


Transformation is x^p

Circle of Transformations

X down, Y up

X up, Y up



X down, Y down

X up, Y down

Box-Cox Transformations

- Also called power transformations
- $Y' = Y^{\lambda}$
- or $Y' = (Y^{\lambda} - 1)/\lambda$
- In the second form, the limit as λ approaches zero is the (natural) log

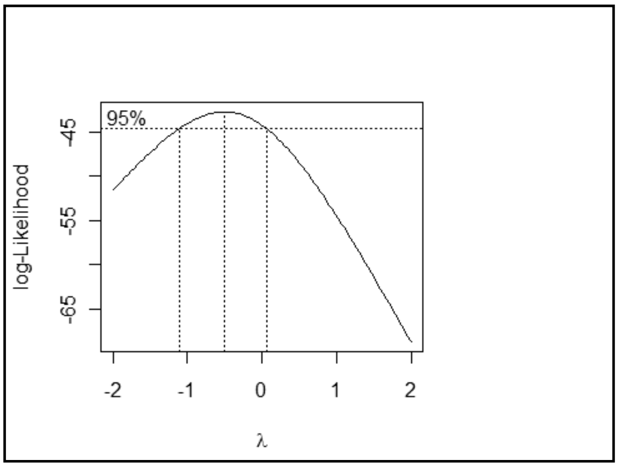
Important Special Cases

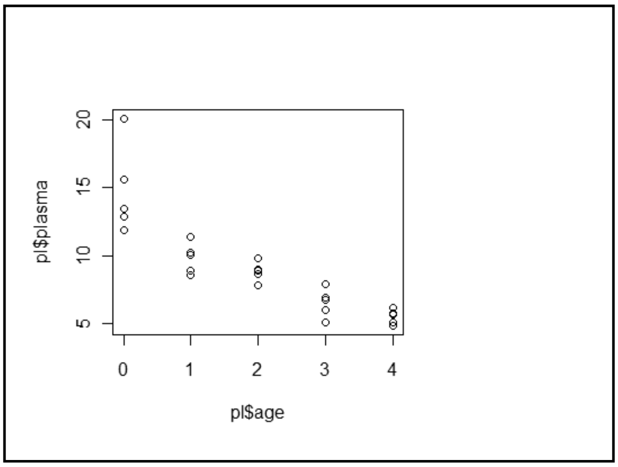
- $\lambda = 1$, $Y' = Y^1$, no transformation
- $\lambda = .5$, $Y' = Y^{1/2}$, square root
- $\lambda = -.5$, $Y' = Y^{-1/2}$, one over square root
- $\lambda = -1$, $Y' = Y^{-1} = 1/Y$, inverse
- $\lambda = 0$, $(Y' = (Y^{\lambda} - 1)/\lambda)$, log is the limit

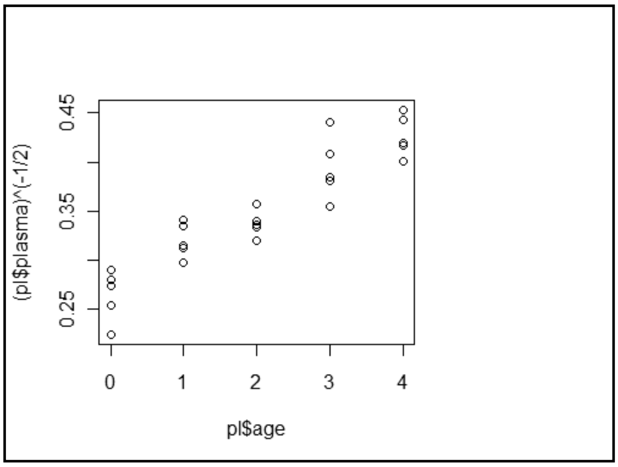
Box-Cox Details

- We can estimate λ by including it as a parameter in a non linear model
- $Y^{\lambda} = \beta_0 + \beta_1 X + \xi$
- and using the method of maximum likelihood
- Boxcox{MASS}

```
pl<-read.table('plasma.txt',
col.names=c("age", "plasma"));
boxcox(pl$plasma~pl$age)
```


95%
log-Likelihood
-45
-55
-65
-2
-1
0
1
2
λ


pl$plasma
20
15
10
5
0
1
2
3
4
pl$age


(pl$plasma)^(-1/2)
0.45
0.35
0.25
0
1
2
3
4
pl$age