05/09/2014 David ? (Manager, Quantitative Analyst) Phone Technical 45min

Intro to David
Youtube team, 2yrs, it's the first job after post-doc in biostat, math undergrad, stat phd
12 people now, look to add 2 more in the next few months
we do complex modelling than other QA's at google, a few phd level statistician in the team

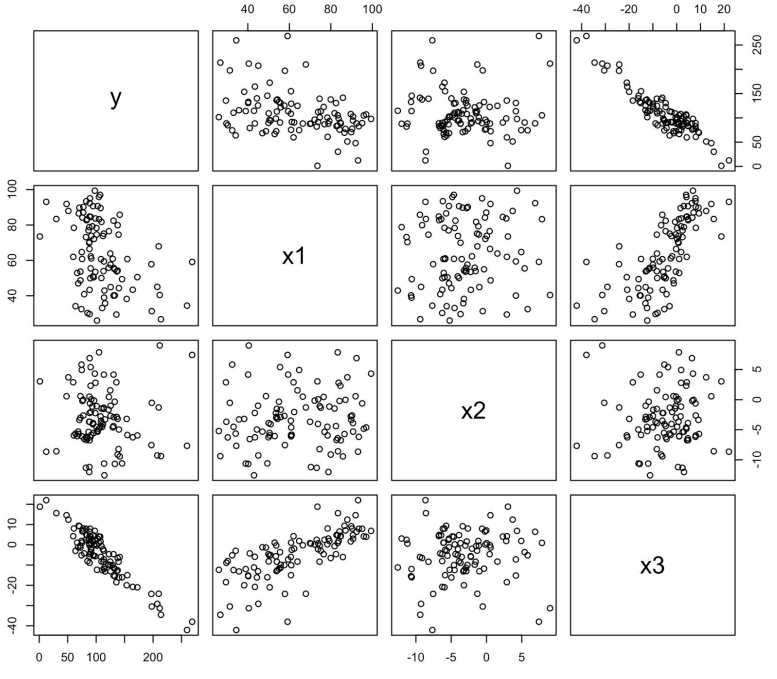
Tell me about a data analysis project you did
Bitcoin, blabla

What software system did you use for data analysis
R and Hadoop, Rhipe

Given the following values, explain what SE is to a person with no stat background
sample mean = 25
SE = 1.7
# mean measures the average, SE measures the variability, e.g., 5 and 45 vs 24 and 26
Does SE measure variability of the sample values or the statistic
both?
How is SE computed
compute sample variance, then sample standard deviation, then sd/sqrt(n)

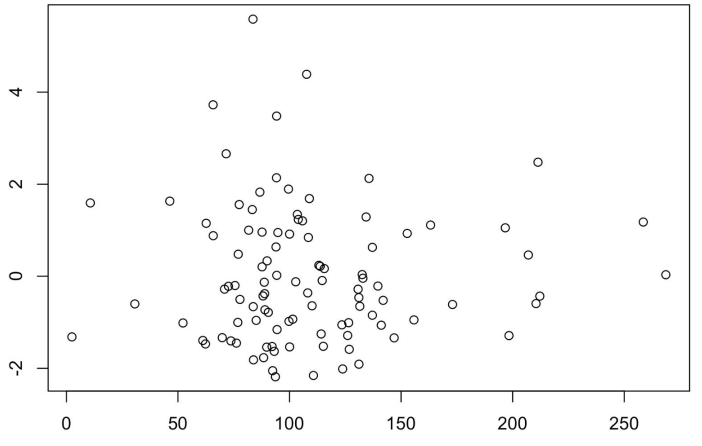
Explain paired t-test
blabla

Given the following data and scatterplot matrix, what do you think about fitting a multiple linear regression model
outcome y and predictors x1, x2, and x3

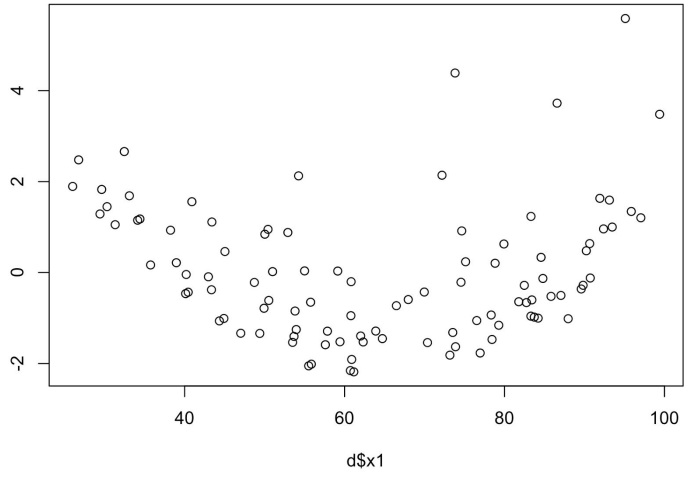


all var are numeric, there is a strong negative correlation between y and x3, there might be correlation for x1 and x3, look out for multicollinearity
How do you fit a multiple linear regression model
in R: fit = lm(y ~ x1+x2+x3, data=df)
How do you check the model fit

```
# the following code, then david asked me to interpret the results, notice the quadratic
pattern in x1
summary(fit)
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1878 -1.0552 -0.2827  0.9531  5.5888

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  5.588149   0.780419    7.160 1.62e-10 ***
x1           1.236599   0.010758  114.946  < 2e-16 ***
x2          -0.001377   0.032074   -0.043    0.966
x3          -4.999049   0.018285 -273.404  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.494 on 96 degrees of freedom
Multiple R-squared: 0.9989,       Adjusted R-squared: 0.9989
F-statistic: 2.94e+04 on 3 and 96 DF,  p-value: < 2.2e-16

plot(fitted(fit), residual(fit))
```

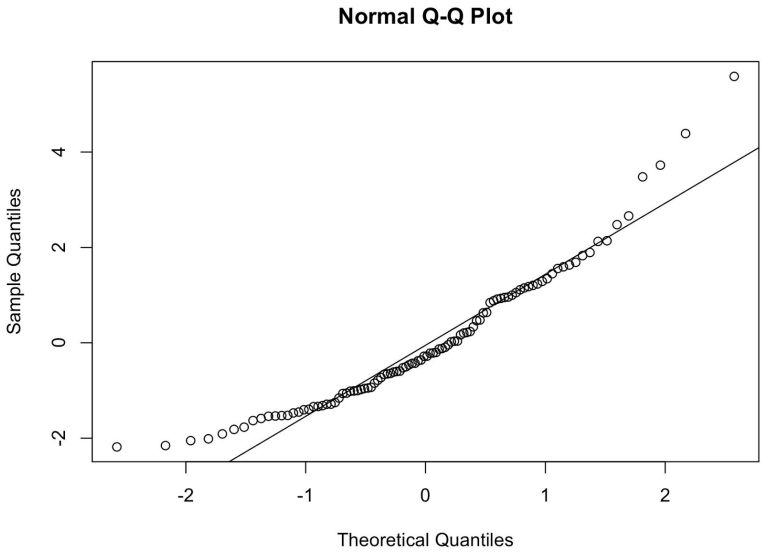



```
plot(df$x1, residual(fit))
```




```
qqnorm(residual(fit))
```

qqline()



fit2 = lm(y ~ x1 + x2 + x3 + I(x1^2), data=df)

Case study: in google consumer survey
How could I estimate the approximate fraction of users who click randomly?
give out surveys where the order of items are swapped
So this can show whether they always clicked on the first one
yes, also i'm thinking of converting this to a classification problem, do you have labels?
We do not, but we can add questions to the survey
we can have questions with unknown/obvious answers, to see if they get it wrong, but this is not too statistical
Yea, this is more of problem solving than statistics, but there are people doing that …
Now, given the following, suppose we find out the fraction of random clicks, how to recover the actual probability
25% of users click randomly
Do you own a car?
66% - Yes
34% - No
p_obs(yes) = p(rand) * 0.5 + p(sin) * p(yes), so 0.66 = .25 *.5 + .75*p(yes) => p(yes) =
Good, we do not need the algebra

Case study: anomaly detection in logs
given 200 time series of metrics, how to detect anomaly
not very familiar with time series, but in general, we decompose time series into trend, seasonal pattern, days of week pattern, within a day morning/afternoon/evening pattern; for anomaly detection, the idea is to find out what normal traffic looks like, we can study the normal traffic to get an idea of what normal behavior looks like, then compare new event to the normal

Questions for interviewer
how did you do with the anomaly detection project
similar simple model as you suggested
will candidate be assigned to positions or do I have a choice
assigned, if you join youtube, then you will be in my team
how are projects generated

varies, from engineering, you can also work on short term (one day) projects you find interesting
projects individual oriented
mostly, there are cooperations
big data opportunity
in my team, we almost exclusively work on big data, we use tools to pool data from SQL like database and analyze it in R, we also use other tools