Initial Modeling

Nick Browen, Eric Ortiz

brief reminder of the context/levels/response variable

Notes on this document

We made 11 separate models in this document to see how the data could be modeled. Depending on a few questions we have, our final model will be either Model 9, Model 10, or Model 11.

Residual plots for these models are included at the bottom of the document.

Model 9 how is sex coded?

Here, we use centered variables on original scales. As predictors, we include distance (quantitative, 100m Dash = 0), centered BMI, sex, year (0 = 1896), centered GDP (in $Billions), and interactions ~~between~~ distance*BMI, distance*sex, and distance*GDP. We are treating level 1 as the individual athletes and level 2 as the countries they represent.

We ran into a few issues with the scaling of factors, namely distance and time in seconds (response). Each event is on a drastically different scale, since athletes compete in events of distances 100m, 200m, 400m, 800m, 5k. In later models we used a log transformation on these variables.

convert to speed? ok adjust for event?

Model 10

Here, we used the same model as Model 9, except that distance has now undergone a log transformation. However, when we plug in values for a male running the 100m dash, the expected completion time is -615 seconds, which doesn't make sense. is distance quantitative or categorical?

Model 11

Here, we used the same model as model 10, but now our response, time in seconds, has undergone a log transformation as well. This resulted in centered BMI, logdist100:c_BMI, and logdist100:sexW no longer being significant. However, this did a better job at predicting an average male's 100m dash time at 2.46 log(seconds) = e^2.46 = 11.72 seconds. This makes a lot more sense now. across the board?

Next Steps:

(because transformed the response)

Since AIC/BIC are not effective in comparing these models, we will be looking at residual plots to decide which of these to use.

Model 1 = Random Intercepts

```
mod1 <- lme(data = track, fixed = timeSecs ~ 1, random = ~1|country2)
mod1
```

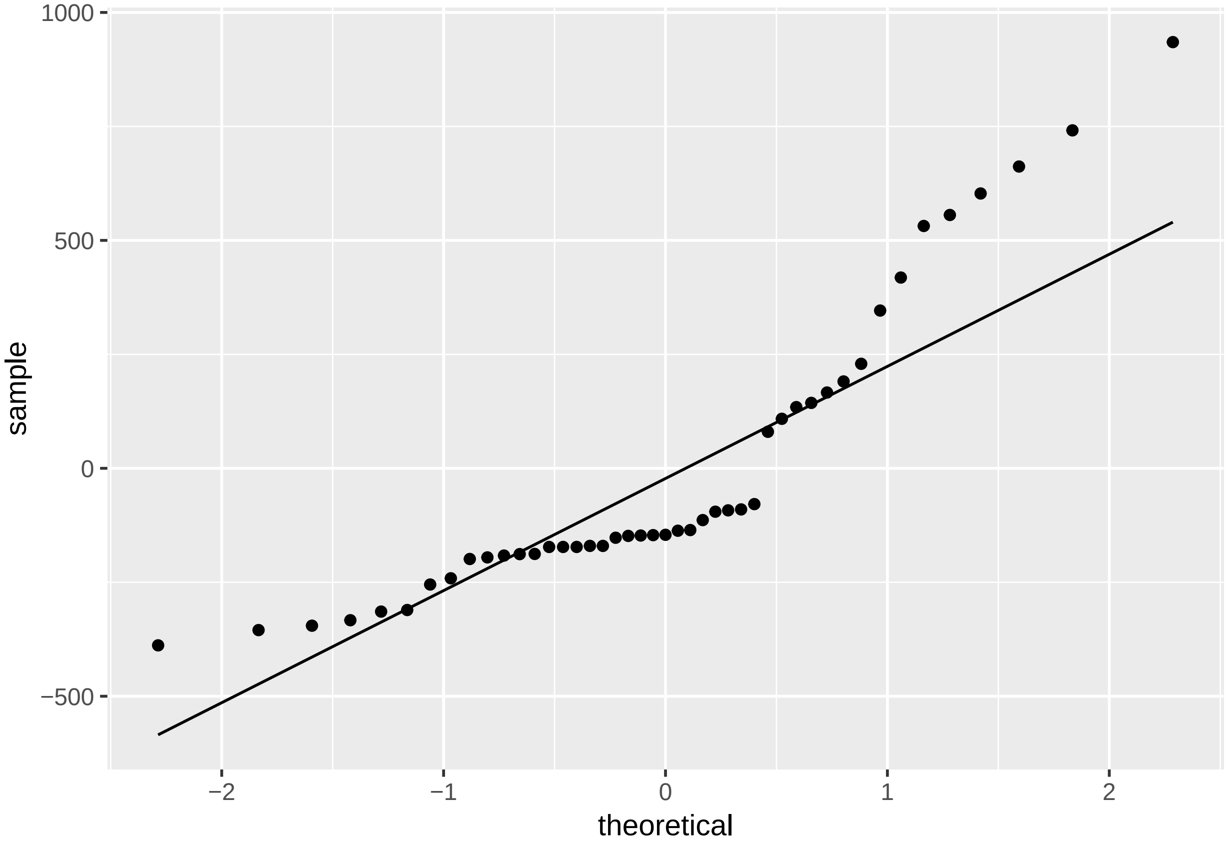
```
## Linear mixed-effects model fit by REML
##   Data: track
##   Log-restricted-likelihood: -4394.977
##   Fixed: timeSecs ~ 1
## (Intercept)
##    423.1011
##
## Random effects:
##  Formula: ~1 | country2
##         (Intercept) Residual
## StdDev:    383.6371 419.2617
##
## Number of Observations: 585
## Number of Groups: 45
```

```
print(VarCorr(mod1), comp=c("Variance", "Std.Dev."), digits = 2)
```

```
## country2 = pdLogChol(1)
##             Variance StdDev
## (Intercept) 147177.4 383.6371
## Residual    175780.4 419.2617
```

ICC = 147177.4/(147177.4+175780.4) = 0.456 so 45.6% of the variation in finishing times is due to country-to-country variation. If we randomly selected two athletes from the same country, their finishing times would be 45.6% correlated. the estimated correlation coefficient is 0.456?

Intercept: The average finishing time of an athlete from the average country is 7.052 minutes. (predicted)



We might have some caution with the model since the random effects are not approximately normal, but we will procede with this in mind.

Mode1 2 = Random Intercepts + fixed effect distance

```
mod2 <- lme(data = track, fixed = timeSecs ~ dist, random = ~1|country2)
mod2
```

```
## Linear mixed-effects model fit by REML
##   Data: track
##   Log-restricted-likelihood: -2957.623
##   Fixed: timeSecs ~ dist
## (Intercept)        dist
## -24.2852599   0.1771274
```

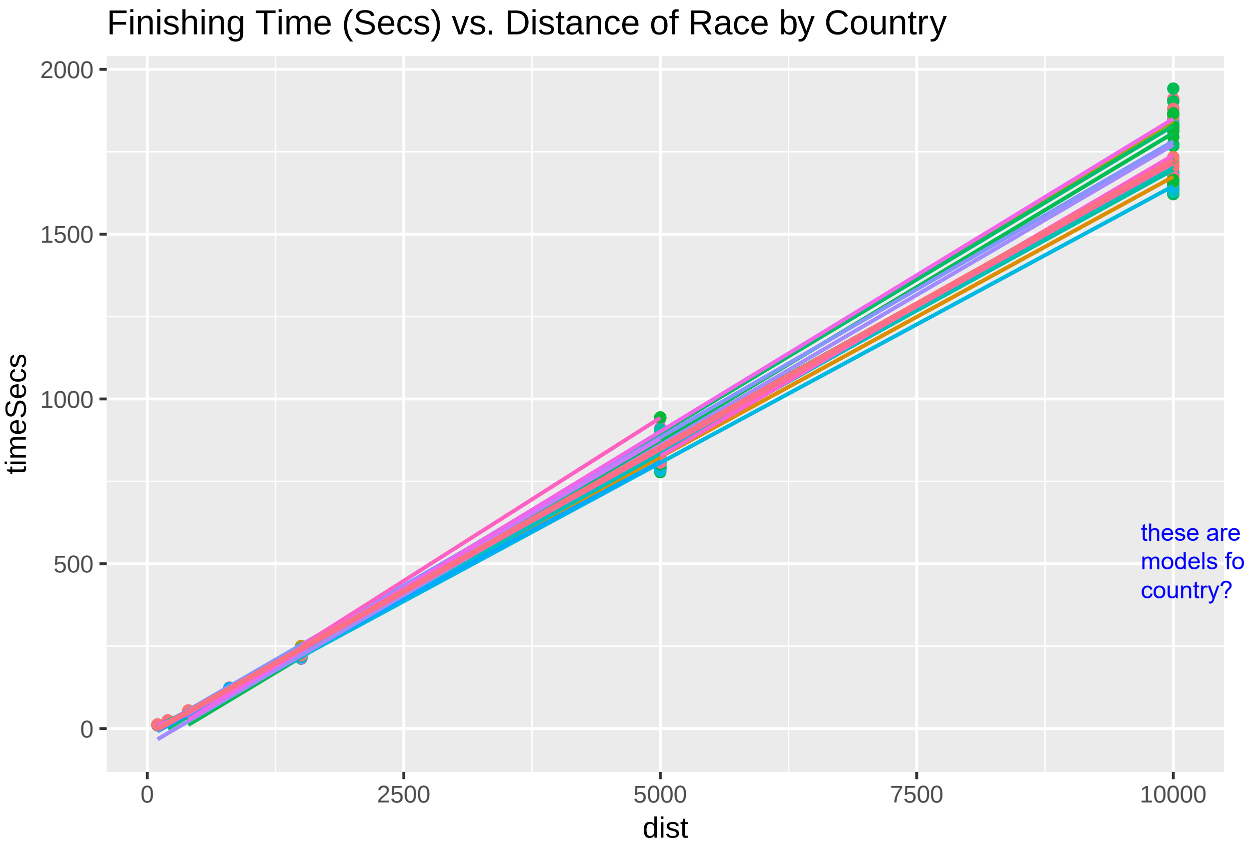
is finish time linearly related to distance?

```
##
## Random effects:
##  Formula: ~1 | country2
##         (Intercept) Residual
## StdDev:    16.42762 36.54942
##
## Number of Observations: 585
## Number of Groups: 45
```

```
print(VarCorr(mod2), comp=c("Variance", "Std.Dev."), digits = 2)
```

```
## country2 = pdLogChol(1)
##             Variance  StdDev
## (Intercept)  269.8667 16.42762
## Residual    1335.8604 36.54942
```

Both variances went down extremely. The within country variation went from $\sigma^2 = 175780.4$ to $\sigma^2 = 1335.9$, decreasing it by 99.2%. After adjusting for the distance of the race, our ICC went down to 16.8%. If we randomly selected two athletes from the same country, their finishing times would be 16.8% correlated, after adjusting for the distance of their race. good adjustment, but would still interpret as correlation coefficient :)



these are the fitted models for each country?

Model 3 = Random Intercepts, Random Slopes for distance

```
mod3 <- lme(data = track, fixed = timeSecs ~ dist, random = ~ dist|country2)
mod3
```

```
## Linear mixed-effects model fit by REML
##   Data: track
##   Log-restricted-likelihood: -2910.962
##   Fixed: timeSecs ~ dist
## (Intercept)        dist
## -23.2630898   0.1754018
##
## Random effects:
##  Formula: ~dist | country2
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev       Corr
## (Intercept)  8.232628737 (Intr)
## dist         0.007090814 -0.776
## Residual    33.256101511
##
## Number of Observations: 585
## Number of Groups: 45
```

```
print(VarCorr(mod3), comp=c("Variance", "Std.Dev."), digits = 2)
```

```
## country2 = pdLogChol(dist)
##             Variance     StdDev       Corr
## (Intercept) 6.777618e+01  8.232628737 (Intr)
## dist        5.027964e-05  0.007090814 -0.776
## Residual    1.105968e+03 33.256101511
```

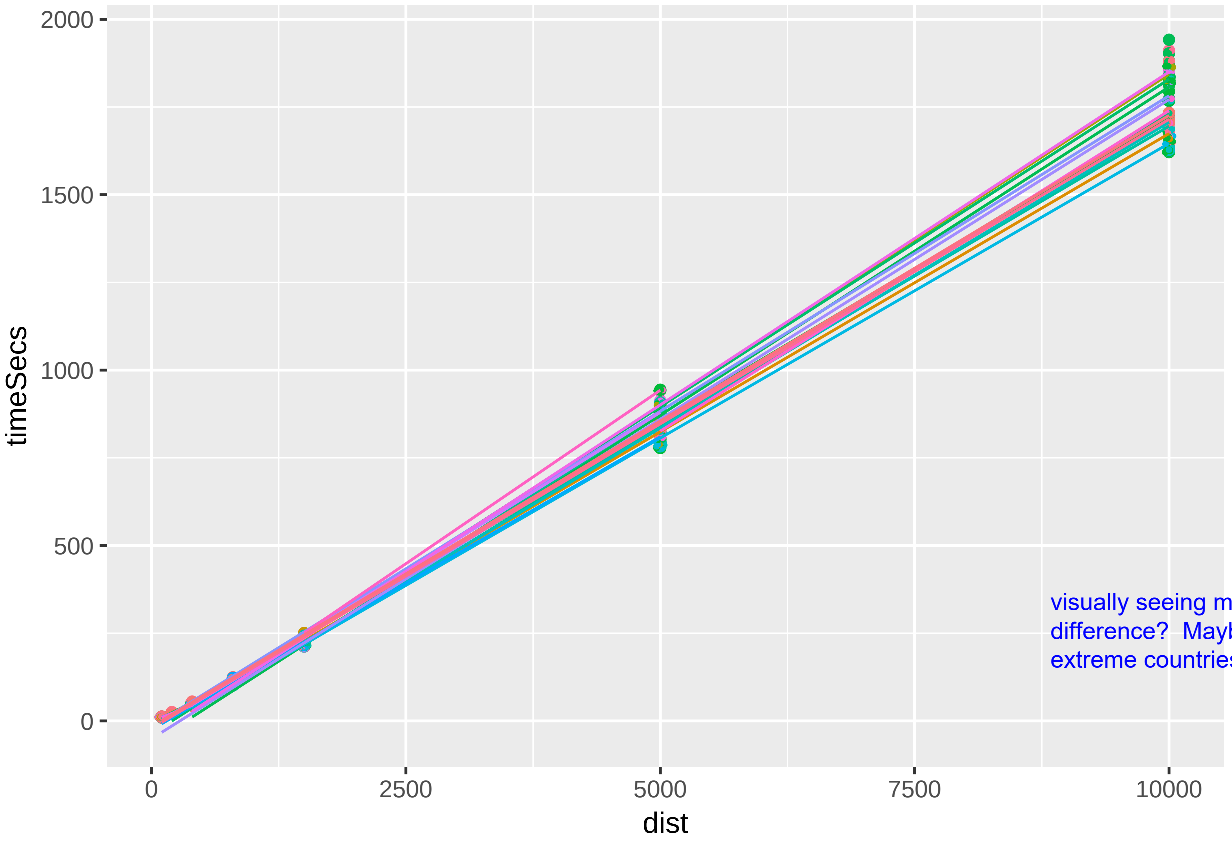
There was a $\frac{1335.8604-1105.968}{1335.8604} * 100 = 17.2\%$ decrease in σ^2, within country variation, by allowing random slopes for distance by country.

```
anova(mod2,mod3)
```

```
##      Model df      AIC      BIC    logLik   Test L.Ratio p-value
## mod2     1  4 5923.246 5940.718 -2957.623
## mod3     2  6 5833.925 5860.134 -2910.962 1 vs 2 93.3209  <.0001
```

Random slopes for distance makes the model significantly better. The relationship between time and distance is different country to country (every 1 meter increase corresponds to a different increase in finishing time depending on the country).

kinda surprising? Maybe sure to compare these analyses with your EDA results too!



visually seeing much difference? Maybe find the extreme countries and discuss?

Model building steps: First, we fit a random intercepts model treating country as random level 2 units. Next, we added distance as a fixed effect because there will obviously be a lot of variation in finishing times due to distance. Then we reassessed our more accurate ICC. Then, we allowed the slopes for distance to be random across country. This was helpful in explaining within country variation. Now we will fit a model with all the fixed effects our EDA indicated would be useful in explaining finishing times. After looking at that, we will assess and systematically remove the largest non-signficant terms one by one.

within or between?

Fixed effects to add from EDA:

- distance,height, weight, height x distance, weight x distance, year, age, age x year, age x dist, sex, sex x dist, GDP, logPOP
- keep in mind height and weight are very correlated (could try BMI = weight/(height^2)) and that gdp and logPOP are very correlated (could try GDP/capita = GDP/POP) and gdp and year are correlated

Model 4

```
mod4 <- lme(data = track,
            fixed = timeSecs ~ dist100 + c_height + c_weight + c_height*dist100 + c_weight*dist100 +
              year1896 + c_age + c_age*year1896 + c_age*dist100 + sex + sex*dist100 + c_logpop + c_gdp,
            random = ~ dist100|country2)
summary(mod4)
```

```
## Linear mixed-effects model fit by REML
##  Data: track
##        AIC      BIC    logLik
##   5323.445 5401.698 -2643.722
##
```

```
## Random effects:
##  Formula: ~dist100 | country2
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev       Corr
## (Intercept) 11.300150100 (Intr)
## dist100      0.006714674 -0.942
## Residual    20.009734766
##
## Fixed effects: timeSecs ~ dist100 + c_height + c_weight + c_height * dist100 +      c_weight * dist10
##                      Value Std.Error  DF   t-value p-value
## (Intercept)      17.889705  4.271045 527   4.18860  0.0000
## dist100           0.171779  0.001433 527 119.91080  0.0000
## c_height         -0.022202  1.825382 527  -0.01216  0.9903
## c_weight          6.681171  2.034559 527   3.28384  0.0011
## year1896         -0.395314  0.042379 527  -9.32796  0.0000
## c_age             1.198796  2.740630 527   0.43742  0.6620
## sexW             20.109938  3.436543 527   5.85179  0.0000
## c_logpop         -1.245292  1.770208 527  -0.70347  0.4821
## c_gdp             6.158151  1.415637 527   4.35009  0.0000
## dist100:c_height -0.004209  0.000558 527  -7.54313  0.0000
## dist100:c_weight  0.002241  0.000807 527   2.77560  0.0057
## year1896:c_age   -0.030169  0.030695 527  -0.98289  0.3261
## dist100:c_age     0.000365  0.000267 527   1.36391  0.1732
## dist100:sexW      0.014610  0.001037 527  14.08687  0.0000
##  Correlation:
##                  (Intr) dst100 c_hght c_wght yr1896 c_age  sexW   c_lgpp
## dist100          -0.557
## c_height         -0.018 -0.010
## c_weight         -0.054  0.083 -0.649
## year1896         -0.736  0.053 -0.057 -0.086
## c_age             0.201 -0.035 -0.073  0.073 -0.242
## sexW             -0.045  0.115  0.181  0.394 -0.334  0.125
## c_logpop          0.168 -0.118 -0.084  0.083 -0.037  0.106 -0.059
## c_gdp             0.412 -0.019  0.120 -0.174 -0.483 -0.088 -0.014 -0.360
## dist100:c_height  0.086 -0.012 -0.461  0.352 -0.139  0.080  0.035 -0.006
## dist100:c_weight -0.114  0.117  0.224 -0.430  0.187 -0.095 -0.261  0.004
## year1896:c_age   -0.174  0.032  0.078 -0.055  0.199 -0.915 -0.101 -0.071
## dist100:c_age    -0.063 -0.058  0.033 -0.076  0.115 -0.269 -0.086 -0.034
## dist100:sexW      0.021  0.011 -0.075 -0.204  0.078 -0.030 -0.450 -0.029
##                  c_gdp  dst100:c_h dst100:c_w y1896: dst100:c_g
## dist100
## c_height
## c_weight
## year1896
## c_age
## sexW
## c_logpop
## c_gdp
## dist100:c_height -0.007
## dist100:c_weight  0.021 -0.612
## year1896:c_age    0.053 -0.062      0.075
## dist100:c_age     0.034 -0.088      0.080      0.048
## dist100:sexW      0.060  0.063      0.531      0.011  0.022
##
```

```
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max 
## -4.959872437 -0.434931695 -0.007872736  0.374403693  8.056420510 
## 
## Number of Observations: 585
## Number of Groups: 45
```

Suprisingly height is not statistically significant with a huge p-value of 0.9903. GDP is signficant with a t-value of 4.35 but a very low coefficient and a std error of 0, which is weird (later discovered this is because GDP is in the billions so the coefficient was very very small). Less surprising, logpop is not significant. We are going to refit almost the same model, but with BMI in place of height and weight (multicollinear, height becomes significant when weight is dropped) and GDP/pop in place of GDP and logPOP.

standardize?

Model 5

```
mod5 <- lme(data = track,
            fixed = timeSecs ~ dist100 + c_BMI + c_BMI*dist100 + year1896 + c_age +
              c_age*year1896 + c_age*dist100 + sex + sex*dist100 + c_gdp_pop,
            random = ~ dist100|country2)
summary(mod5)
```

```
## Linear mixed-effects model fit by REML
##  Data: track 
##        AIC      BIC    logLik
##   5370.394 5435.684 -2670.197
## 
## Random effects:
##  Formula: ~dist100 | country2
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev       Corr  
## (Intercept) 11.470258780 (Intr)
## dist100      0.007089808 -0.933
## Residual    21.011535363       
## 
## Fixed effects: timeSecs ~ dist100 + c_BMI + c_BMI * dist100 + year1896 + c_age +      c_age * year18
##                    Value Std.Error  DF   t-value p-value
## (Intercept)    13.478959  4.527568 530   2.97709  0.0030
## dist100         0.172492  0.001486 530 116.08043  0.0000
## c_BMI           4.333177  1.330373 530   3.25711  0.0012
## year1896       -0.348624  0.048044 530  -7.25629  0.0000
## c_age           2.225986  2.856117 530   0.77937  0.4361
## sexW           14.693626  2.903846 530   5.06006  0.0000
## c_gdp_pop       3.002130  1.488306 530   2.01715  0.0442
## dist100:c_BMI   0.002016  0.000483 530   4.17592  0.0000
## year1896:c_age -0.040038  0.032112 530  -1.24683  0.2130
## dist100:c_age   0.000289  0.000280 530   1.03089  0.3031
## dist100:sexW    0.019175  0.000905 530  21.19561  0.0000
##  Correlation: 
##                (Intr) dst100 c_BMI  yr1896 c_age  sexW   c_gdp_ d100:_B
## dist100        -0.511                                                  
## c_BMI          -0.117  0.102                                           
## year1896       -0.792  0.057 -0.001                                    
## c_age           0.211 -0.020  0.052 -0.258                             
## sexW           -0.015  0.092  0.463 -0.255  0.181
## c_gdp_pop       0.564 -0.056 -0.191 -0.645  0.009 -0.056
## dist100:c_BMI  -0.076  0.058 -0.457  0.141 -0.090 -0.293  0.051
## year1896:c_age -0.194  0.018 -0.037  0.232 -0.915 -0.166 -0.044  0.065
## dist100:c_age  -0.068 -0.055 -0.067  0.107 -0.268 -0.087  0.002  0.092
## dist100:sexW    0.042 -0.068 -0.236  0.061 -0.034 -0.477  0.067  0.556
##                y1896: ds100:_
## dist100
## c_BMI
## year1896
## c_age
## sexW
## c_gdp_pop
## dist100:c_BMI
## year1896:c_age
## dist100:c_age   0.047
## dist100:sexW    0.008  0.050
##
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max
## -5.189647914 -0.444364689 -0.005206271  0.374535646  8.375570994
##
## Number of Observations: 585
## Number of Groups: 45
```

Looks like all the terms with age are statistically insignificant (age, year x age, distance x age) So, next we will remove them and refit the model.

Model 6

```
mod6 <- lme(data = track,
            fixed = timeSecs ~ dist100 + c_BMI + c_BMI*dist100 + year1896 + sex + sex*dist100 + c_gdp_po
            random = ~ dist100|country2)
summary(mod6)
```

```
## Linear mixed-effects model fit by REML
##  Data: track
##        AIC      BIC    logLik
##   5349.475 5401.769 -2662.738
##
## Random effects:
##  Formula: ~dist100 | country2
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev      Corr
## (Intercept) 11.48201198 (Intr)
## dist100      0.00717439 -0.933
## Residual    20.99959247
##
## Fixed effects: timeSecs ~ dist100 + c_BMI + c_BMI * dist100 + year1896 + sex +      sex * dist100 +
##                    Value Std.Error  DF   t-value p-value
## (Intercept)   12.748961  4.428559 533   2.87881  0.0042
## dist100        0.172547  0.001497 533 115.23479  0.0000
## c_BMI          4.365171  1.325862 533   3.29233  0.0011
## year1896      -0.340366  0.046405 533  -7.33475  0.0000
## sexW          14.362158  2.851198 533   5.03724  0.0000
## c_gdp_pop      2.859676  1.481910 533   1.92972  0.0542
## dist100:c_BMI  0.002002  0.000480 533   4.17515  0.0000
## dist100:sexW   0.019127  0.000902 533  21.19741  0.0000
##  Correlation:
##               (Intr) dst100 c_BMI  yr1896 sexW   c_gdp_ d100:_
## dist100       -0.522
## c_BMI         -0.133  0.100
## year1896      -0.781  0.057  0.015
## sexW          -0.056  0.093  0.460 -0.217
## c_gdp_pop      0.578 -0.060 -0.193 -0.669 -0.061
## dist100:c_BMI -0.057  0.060 -0.452  0.119 -0.281  0.052
## dist100:sexW   0.050 -0.068 -0.233  0.053 -0.480  0.063  0.555
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -5.2576587 -0.4374543 -0.0124798  0.3836700  8.4847422
##
## Number of Observations: 585
## Number of Groups: 45
```

GDP/pop is slightly insignificant (p-value = 0.0542). From the EDA and model 4, we saw that GDP had a relationship with finishing time. And model 4 showed that population (logpop) doesn't have a significant relationship with finishing time. So we are going to replace GDP/pop with GDP.

Model 7

```r
mod7 <- lme(data = track,
            fixed = timeSecs ~ dist100 + c_BMI + c_BMI*dist100 + year1896 + sex + sex*dist100 + c_gdp,
            random = ~ dist100|country2)
summary(mod7)
```

```
## Linear mixed-effects model fit by REML
##  Data: track
##        AIC      BIC   logLik
##   5335.141 5387.435 -2655.57
##
## Random effects:
##  Formula: ~dist100 | country2
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev      Corr
## (Intercept) 12.19196725 (Intr)
## dist100      0.00725988 -0.933
## Residual    20.69122960
##
## Fixed effects: timeSecs ~ dist100 + c_BMI + c_BMI * dist100 + year1896 + sex +      sex * dist100 +
##                   Value Std.Error  DF   t-value p-value
## (Intercept)   18.262589  4.393242 533   4.15697  0.0000
## dist100        0.172310  0.001502 533 114.70484  0.0000
## c_BMI          3.953808  1.302899 533   3.03462  0.0025
## year1896      -0.390955  0.042444 533  -9.21111  0.0000
## sexW          14.335837  2.820151 533   5.08336  0.0000
## c_gdp          6.042805  1.390912 533   4.34449  0.0000
## dist100:c_BMI  0.002002  0.000472 533   4.23809  0.0000
## dist100:sexW   0.019200  0.000889 533  21.59564  0.0000
##  Correlation:
##               (Intr) dst100 c_BMI  yr1896 sexW   c_gdp  d100:_
## dist100       -0.550
## c_BMI         -0.100  0.098
## year1896      -0.756  0.062 -0.037
## sexW          -0.034  0.091  0.457 -0.264
## c_gdp          0.549 -0.070 -0.152 -0.593 -0.030
## dist100:c_BMI -0.082  0.062 -0.450  0.160 -0.280  0.016
## dist100:sexW   0.036 -0.066 -0.230  0.078 -0.480  0.044  0.553
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -5.12372663 -0.42585719 -0.01250783  0.37647640  8.42318257
##
## Number of Observations: 585
## Number of Groups: 45
```

Yes, gdp is significant, but it has a coefficient oof 0 and std error of 0. Looking closer and using lmer output, the coefficient is actually 1.882×10^{-12}. We will to use GDP in billions of dollars, to make this a more usefull variable.

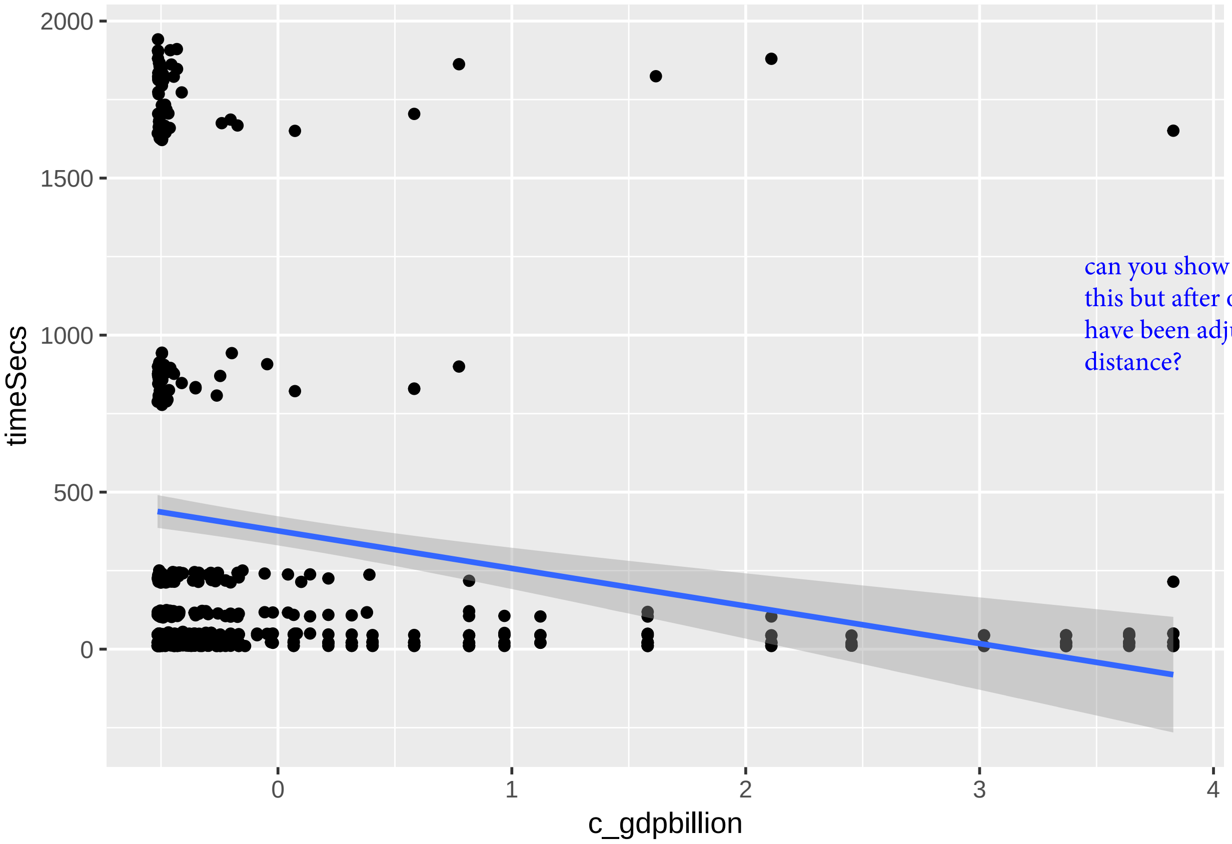
Model 8

```
mod8 <- lme(data = track,
            fixed = timeSecs ~ dist100 + c_BMI + c_BMI*dist100 + year1896 + sex + sex*dist100 + c_gdpbil
            random = ~ dist100|country2)
summary(mod8)
```

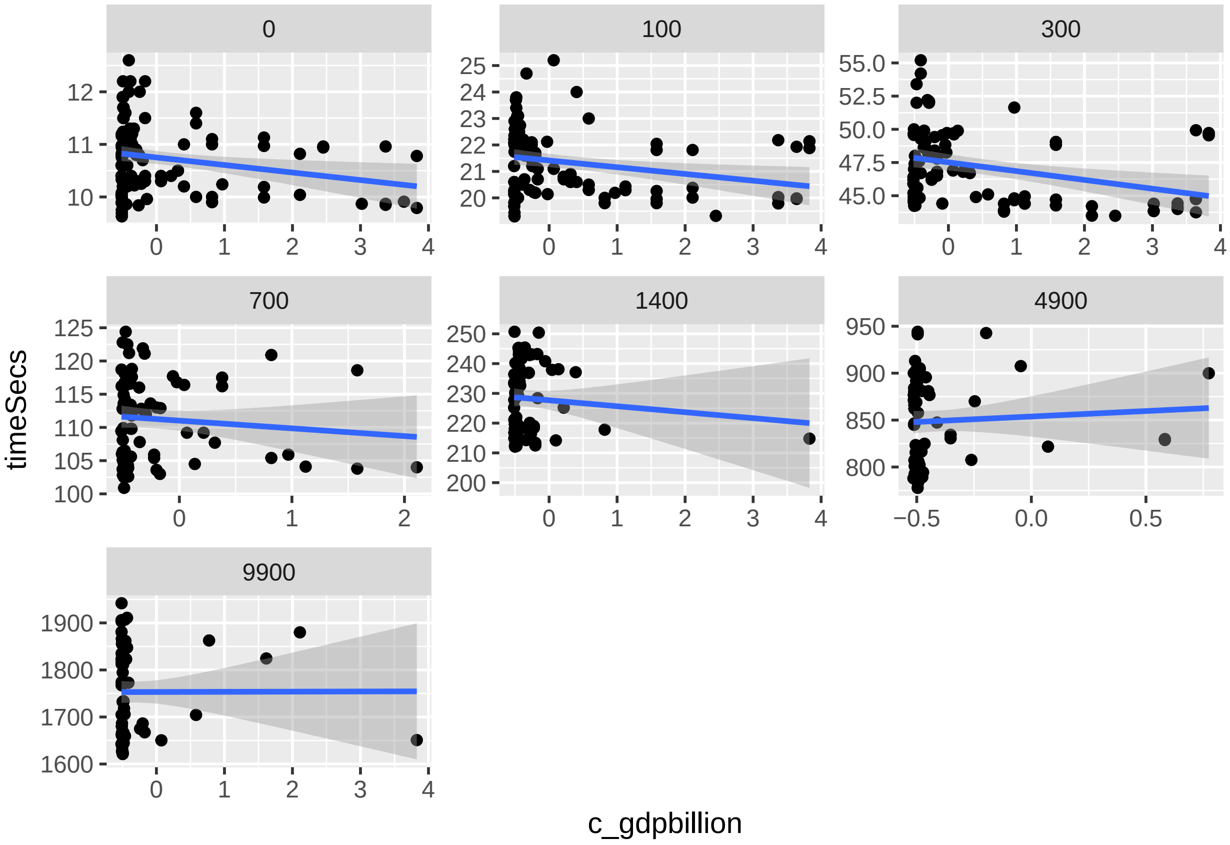
```
## Linear mixed-effects model fit by REML
##  Data: track
##        AIC      BIC   logLik
##   5335.141 5387.435 -2655.57
##
## Random effects:
##  Formula: ~dist100 | country2
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev      Corr
## (Intercept) 12.19196724 (Intr)
## dist100      0.00725988 -0.933
## Residual    20.69122960
##
## Fixed effects: timeSecs ~ dist100 + c_BMI + c_BMI * dist100 + year1896 + sex +      sex * dist100 + 
##                    Value Std.Error  DF   t-value p-value
## (Intercept)    18.262589  4.393242 533   4.15697  0.0000
## dist100         0.172310  0.001502 533 114.70484  0.0000
## c_BMI           3.953808  1.302899 533   3.03462  0.0025
## year1896       -0.390955  0.042444 533  -9.21111  0.0000
## sexW           14.335837  2.820151 533   5.08336  0.0000
## c_gdpbillion    6.042805  1.390912 533   4.34449  0.0000
## dist100:c_BMI   0.002002  0.000472 533   4.23809  0.0000
## dist100:sexW   0.019200  0.000889 533  21.59564  0.0000
##  Correlation:
##               (Intr) dst100 c_BMI  yr1896 sexW   c_gdpb d100:_
## dist100       -0.550
## c_BMI         -0.100  0.098
## year1896      -0.756  0.062 -0.037
## sexW          -0.034  0.091  0.457 -0.264
## c_gdpbillion   0.549 -0.070 -0.152 -0.593 -0.030
## dist100:c_BMI -0.082  0.062 -0.450  0.160 -0.280  0.016
## dist100:sexW   0.036 -0.066 -0.230  0.078 -0.480  0.044  0.553
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -5.12372663 -0.42585719 -0.01250783  0.37647640  8.42318257
##
## Number of Observations: 585
## Number of Groups: 45
```

We have a model here with all fixed effects being statistically significant.

Although its odd that gdp's coefficient is positive. The graph of time vs. gdp shows a negative relationship (as expected).



can you show a graph like this but after observations have been adjusted for distance?



Events from the 100m to the 1500m have a negative relationship with time and gdp, BUT the distances with higher times have slightly positive or almost 0 slopes, which could be making the whole gdp coefficient positive. We will try including the interaction of dist x gdp.

Model 9

```
mod9 <- lme(data = track,
            fixed = timeSecs ~ dist100 + c_BMI + c_BMI*dist100 + year1896 + sex + sex*dist100 + c_gdpbil
            random = ~ dist100|country2)
summary(mod9)
```

```
## Linear mixed-effects model fit by REML
##  Data: track
##        AIC      BIC    logLik
##   5345.162 5401.791 -2659.581
##
## Random effects:
##  Formula: ~dist100 | country2
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev       Corr
## (Intercept) 12.368407467 (Intr)
## dist100      0.007146399 -0.94
## Residual    20.639645552
##
## Fixed effects: timeSecs ~ dist100 + c_BMI + c_BMI * dist100 + year1896 + sex +      sex * dist100 + c
##                            Value Std.Error  DF   t-value p-value
## (Intercept)          20.002170  4.451730 532   4.49312  0.0000
## dist100               0.171875  0.001495 532 114.99009  0.0000
## c_BMI                 4.109750  1.302003 532   3.15648  0.0017
## year1896             -0.406194  0.042633 532  -9.52773  0.0000
## sexW                 14.496656  2.812158 532   5.15499  0.0000
## c_gdpbillion          7.341639  1.505467 532   4.87665  0.0000
## dist100:c_BMI         0.001794  0.000481 532   3.72810  0.0002
## dist100:sexW          0.019091  0.000887 532  21.51088  0.0000
## dist100:c_gdpbillion -0.001505  0.000701 532  -2.14656  0.0323
##  Correlation:
##                      (Intr) dst100 c_BMI  yr1896 sexW   c_gdpb d100:_B
## dist100              -0.574
## c_BMI                -0.090  0.087
## year1896             -0.758  0.088 -0.043
## sexW                 -0.032  0.088  0.457 -0.263
## c_gdpbillion          0.561 -0.130 -0.119 -0.595 -0.021
## dist100:c_BMI        -0.115  0.095 -0.452  0.186 -0.278 -0.068
## dist100:sexW          0.027 -0.057 -0.232  0.085 -0.481  0.021  0.552
## dist100:c_gdpbillion -0.175  0.160 -0.058  0.151 -0.018 -0.400  0.209
##                      d100:W
## dist100
## c_BMI
## year1896
## sexW
## c_gdpbillion
## dist100:c_BMI
## dist100:sexW
## dist100:c_gdpbillion  0.051
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -5.20198607 -0.43597692 -0.01434764  0.38412058  8.25926406
##
## Number of Observations: 585
## Number of Groups: 45
```

The interaction is significant. But the gdp coefficient would still have a positive slope for all races except the 10000m, which is not what we want.

Running mod9 with lme4 instead of nlme

```
lmer(data = track, timeSecs ~ dist100 + c_BMI + c_BMI*dist100 + year1896 + sex + sex*dist100 + c_gdpbil
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : unable to evaluate scaled gradient
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge: degenerate Hessian with 1
## negative eigenvalues
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: timeSecs ~ dist100 + c_BMI + c_BMI * dist100 + year1896 + sex +
##     sex * dist100 + c_gdpbillion + dist100 * c_gdpbillion + (dist100 |
##     country2)
##    Data: track
## REML criterion at convergence: 5345.818
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  country2 (Intercept) 32.66876
##           dist100      0.01385 -0.98
##  Residual             20.17797
## Number of obs: 585, groups:  country2, 45
## Fixed Effects:
##          (Intercept)              dist100                c_BMI
##            24.300807             0.171782             3.936587
##             year1896                 sexW         c_gdpbillion
##            -0.451518            15.044073             8.565514
##        dist100:c_BMI         dist100:sexW  dist100:c_gdpbillion
##             0.001834             0.019133            -0.001596
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
## convergence code 0; 2 optimizer warnings; 0 lme4 warnings
```

We can see lmer is throwing up a lot of errors.

I think the variable distance is causing some of the errors mentioned above. Particulary lmers: “Rescale variables?;Model is nearly unidentifiable: large eigenvalue ratio” Because dist’s values are 100,200,400,800,1500,5000,10000. Which is a huge range.Taking the log of distance would bring these all much much closer together. log(dist) = 4.61, 5.30, 5.99, 6.68, 7.31, 8.52, 9.21

Running mod9 with logdist (with lmer)

```
lmer(data = track, timeSecs ~ logdist100 + c_BMI + c_BMI*logdist100 + year1896 + sex + sex*logdist100 +
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: timeSecs ~ logdist100 + c_BMI + c_BMI * logdist100 + year1896 +
##     sex + sex * logdist100 + c_gdpbillion + c_gdpbillion * logdist100 +
##     (logdist100 | country2)
##    Data: track
## REML criterion at convergence: 7891.89
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  country2 (Intercept) 624.2
##           logdist100  206.5    -0.96
##  Residual             190.5
## Number of obs: 585, groups:  country2, 45
## Fixed Effects:
##              (Intercept)               logdist100                    c_BMI
##                 -615.452                  404.482                  106.686
##                 year1896                     sexW             c_gdpbillion
##                   -1.278                  191.809                  -32.565
##         logdist100:c_BMI          logdist100:sexW  logdist100:c_gdpbillion
##                  -48.690                  -54.672                   48.281
```

Model 10

Same as previous lmer model but using lme and saving it as mod10

```
mod10 <- lme(data = track,
            fixed = timeSecs ~ logdist100 + c_BMI + c_BMI*logdist100 + year1896 + sex + sex*logdist100
```

```r
              c_gdpbillion + c_gdpbillion*logdist100,
            random = ~ logdist100|country2)
summary(mod10)
```

```
## Linear mixed-effects model fit by REML
##  Data: track
##       AIC      BIC    logLik
##   7917.89 7974.519 -3945.945
##
## Random effects:
##  Formula: ~logdist100 | country2
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev   Corr
## (Intercept) 624.1502 (Intr)
## logdist100  206.4469 -0.96
## Residual    190.5049
##
## Fixed effects: timeSecs ~ logdist100 + c_BMI + c_BMI * logdist100 + year1896 +      sex + sex * logdi
##                             Value Std.Error  DF   t-value p-value
## (Intercept)             -615.4490 119.94842 532 -5.130947  0.0000
## logdist100               404.4821  38.28289 532 10.565611  0.0000
## c_BMI                    106.6864  17.86605 532  5.971459  0.0000
## year1896                  -1.2779   0.48534 532 -2.633003  0.0087
## sexW                     191.8102  38.67328 532  4.959761  0.0000
## c_gdpbillion             -32.5652  18.85753 532 -1.726909  0.0848
## logdist100:c_BMI         -48.6901   7.90372 532 -6.160400  0.0000
## logdist100:sexW          -54.6727  16.01728 532 -3.413355  0.0007
## logdist100:c_gdpbillion   48.2810  10.36712 532  4.657123  0.0000
##  Correlation:
##                         (Intr) lgd100 c_BMI  yr1896 sexW   c_gdpb l100:_B
## logdist100              -0.891
## c_BMI                    0.001  0.054
## year1896                -0.366  0.036 -0.186
## sexW                    -0.033  0.106  0.564 -0.232
## c_gdpbillion             0.266 -0.103 -0.103 -0.548 -0.098
## logdist100:c_BMI        -0.081  0.023 -0.756  0.237 -0.442 -0.024
## logdist100:sexW          0.036 -0.095 -0.416  0.061 -0.750  0.119  0.498
## logdist100:c_gdpbillion -0.036  0.111  0.108 -0.024  0.121 -0.589 -0.005
##                         l100:W
## logdist100
## c_BMI
## year1896
## sexW
## c_gdpbillion
## logdist100:c_BMI
## logdist100:sexW
## logdist100:c_gdpbillion -0.107
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -2.3925558 -0.5634912 -0.1256192  0.4028348  5.1532978
##
## Number of Observations: 585
## Number of Groups: 45
```

lmer now gives no errors! The sign for the GDP coefficient is now negative, which makes more sense. The unusual issue now is that intercept is negative.

For the 100m (logdist100 = 0), the predicted time for male athlete of average of everything is -615.45s. That's not good.

Comparing models with dist and logdist

```
AIC(mod9); AIC(mod10)
```

```
## [1] 5345.162

## [1] 7917.89
```

The AIC is much worse for the log(dist) model.

Model 11

can you support some of these decisions with graphs too?

Trying log(time) as the response

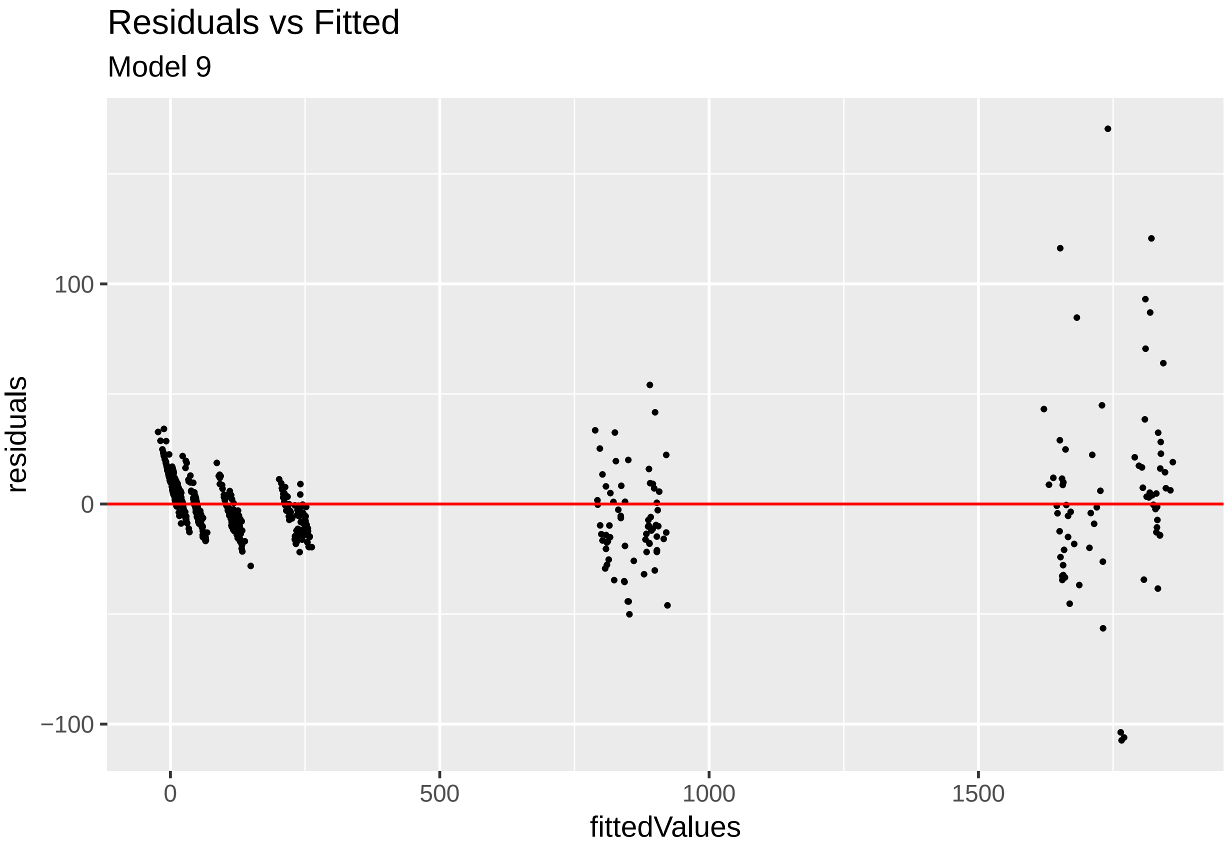
```
mod11 <- lme(data = track,
            fixed = logtimeSecs ~ logdist100 + c_BMI + c_BMI*logdist100 + year1896 +
              sex + sex*logdist100 + c_gdpbillion + c_gdpbillion*logdist100,
            random = ~ logdist100|country2)
summary(mod11)
```

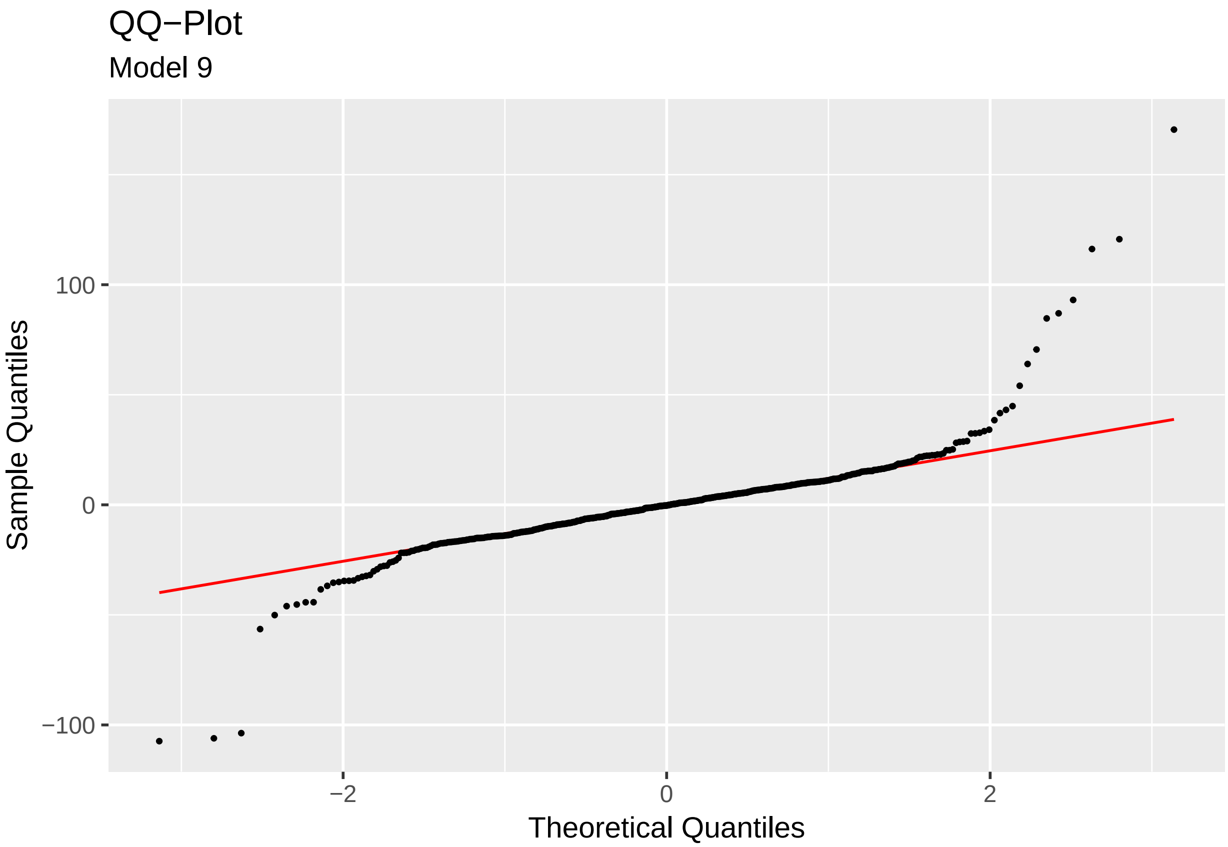
```
## Linear mixed-effects model fit by REML
##  Data: track
##         AIC       BIC   logLik
##   -1933.247 -1876.618 979.6236
##
## Random effects:
##  Formula: ~logdist100 | country2
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev     Corr
## (Intercept) 0.06341842 (Intr)
## logdist100  0.01547263 -0.97
## Residual    0.03921032
##
## Fixed effects: logtimeSecs ~ logdist100 + c_BMI + c_BMI * logdist100 + year1896 +      sex + sex * l
##                              Value   Std.Error  DF   t-value p-value
## (Intercept)              2.4616189 0.015352790 532 160.33691  0.0000
## logdist100               1.1120519 0.003570038 532 311.49581  0.0000
## c_BMI                   -0.0032275 0.003595938 532  -0.89755  0.3698
## year1896                -0.0015264 0.000092422 532 -16.51593  0.0000
## sexW                     0.1176975 0.007770546 532  15.14662  0.0000
## c_gdpbillion             0.0081033 0.003771490 532   2.14855  0.0321
## logdist100:c_BMI        -0.0018225 0.001543737 532  -1.18055  0.2383
## logdist100:sexW         -0.0003421 0.003137697 532  -0.10903  0.9132
## logdist100:c_gdpbillion -0.0037983 0.001876714 532  -2.02392  0.0435
##  Correlation:
##                         (Intr) lgd100 c_BMI  yr1896 sexW   c_gdpb l100:_B
## logdist100              -0.826
## c_BMI                   -0.013  0.113
## year1896                -0.547  0.103 -0.191
## sexW                    -0.064  0.211  0.556 -0.229
## c_gdpbillion             0.403 -0.212 -0.106 -0.551 -0.110
## logdist100:c_BMI        -0.106  0.028 -0.759  0.268 -0.442 -0.028
## logdist100:sexW          0.060 -0.167 -0.415  0.054 -0.755  0.142  0.508
## logdist100:c_gdpbillion -0.098  0.204  0.119  0.064  0.139 -0.628 -0.024
##                         l100:W
## logdist100
## c_BMI
## year1896
## sexW
## c_gdpbillion
## logdist100:c_BMI
## logdist100:sexW
## logdist100:c_gdpbillion -0.157
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.53932746 -0.77973368  0.06844736  0.75072189  4.06989613
##
## Number of Observations: 585
## Number of Groups: 45
```

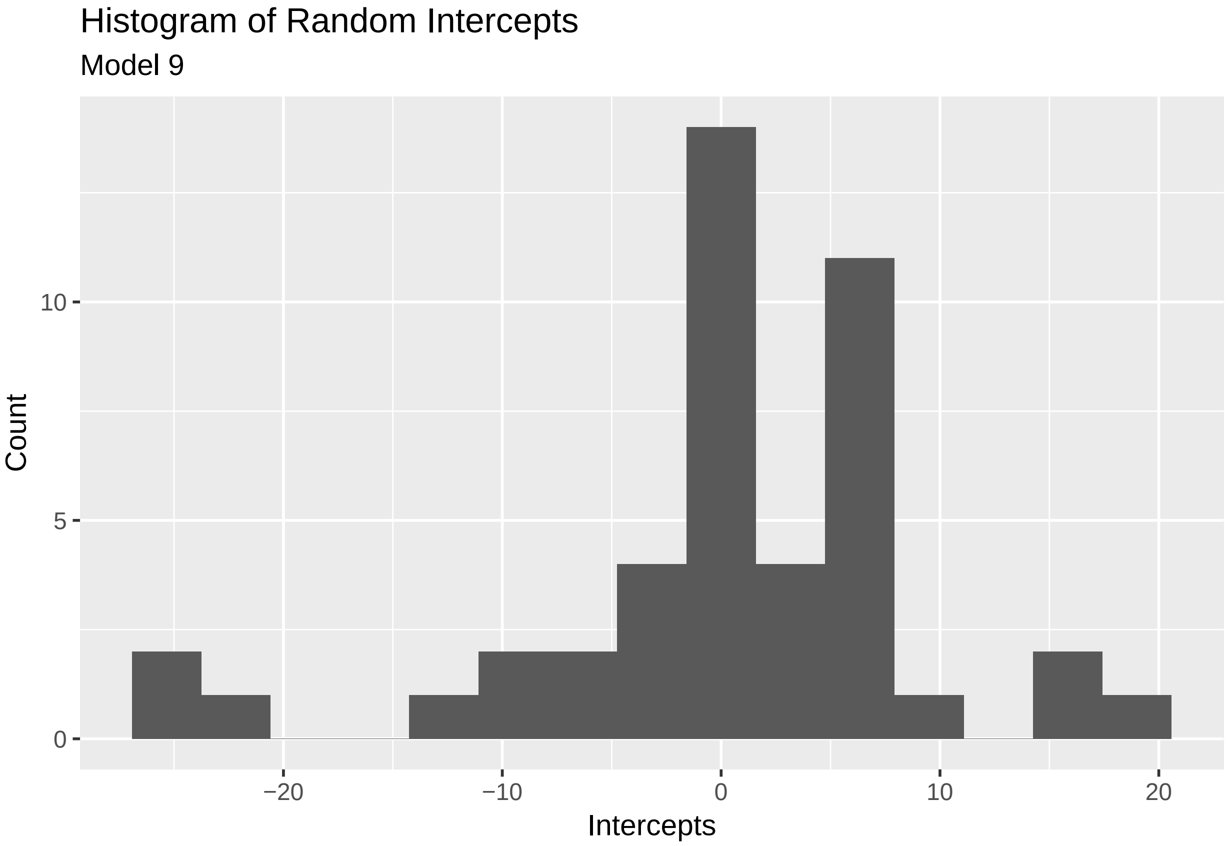
The predicted time for a male athlete's 100m (and average of everything else) is eˆ(2.462) = 11.72s which is good. But now, BMI, dist x BMI, dist x sex are not statistically significant.

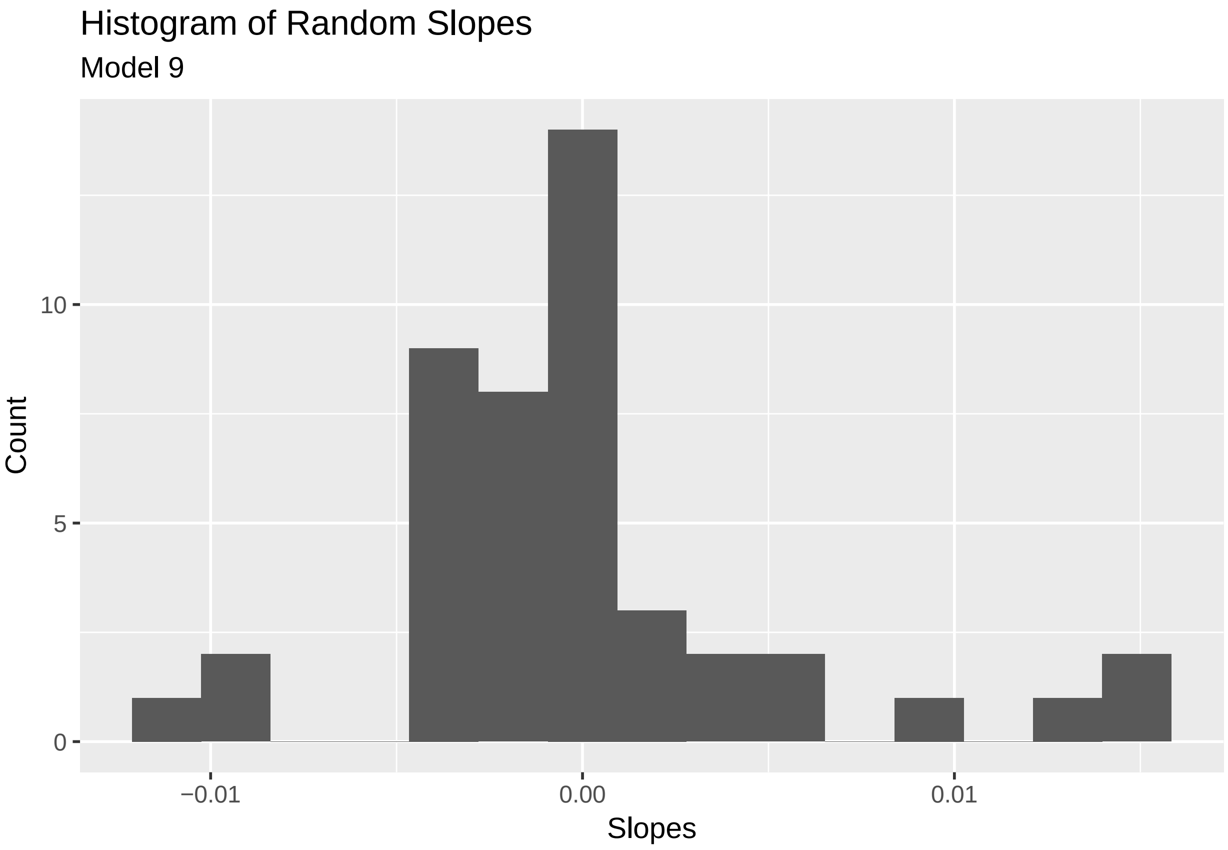
Comparing Residual plots from mod9 to mod11 mod9 = time ~ dist + … mod11 = logtime ~ logdist + …

Model 9 Residual Plots

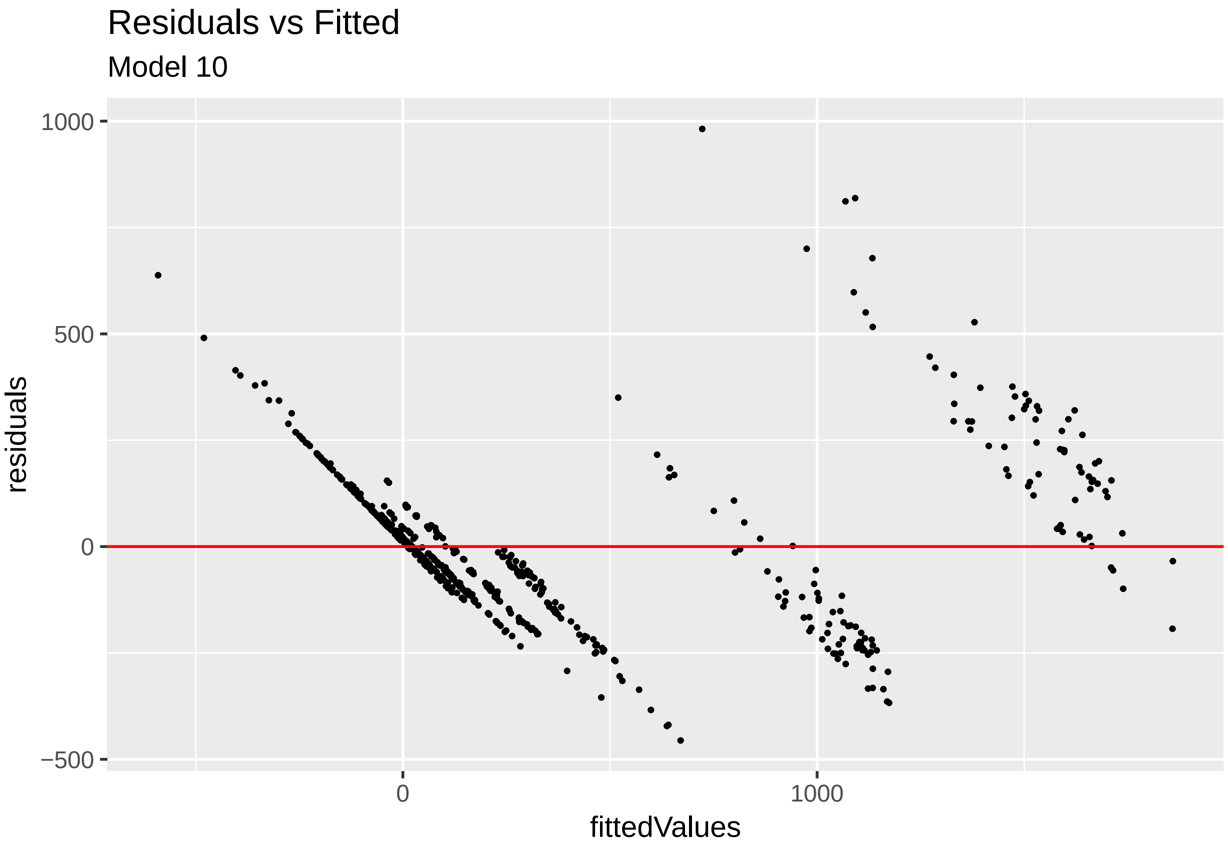


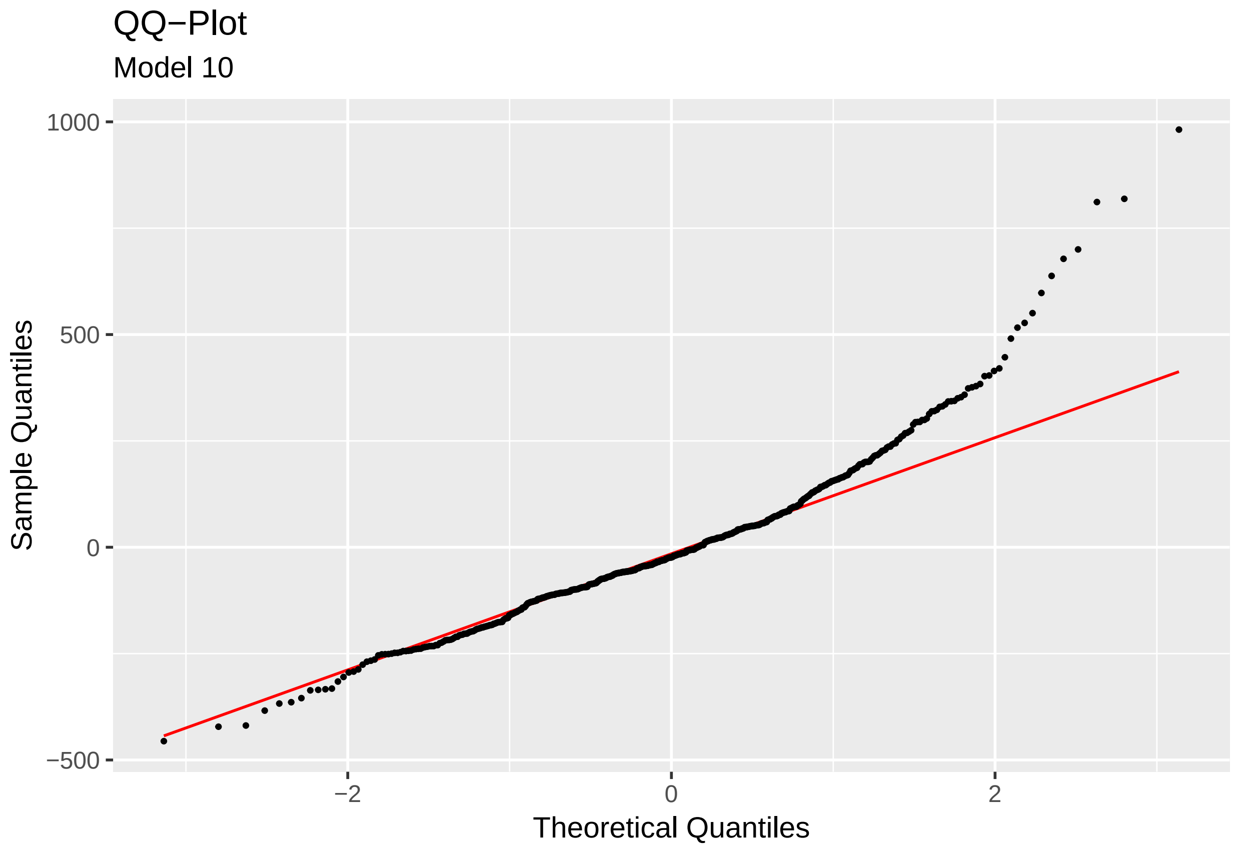

QQ−Plot
Model 9
Sample Quantiles
100
0
−100
−2
0
2
Theoretical Quantiles

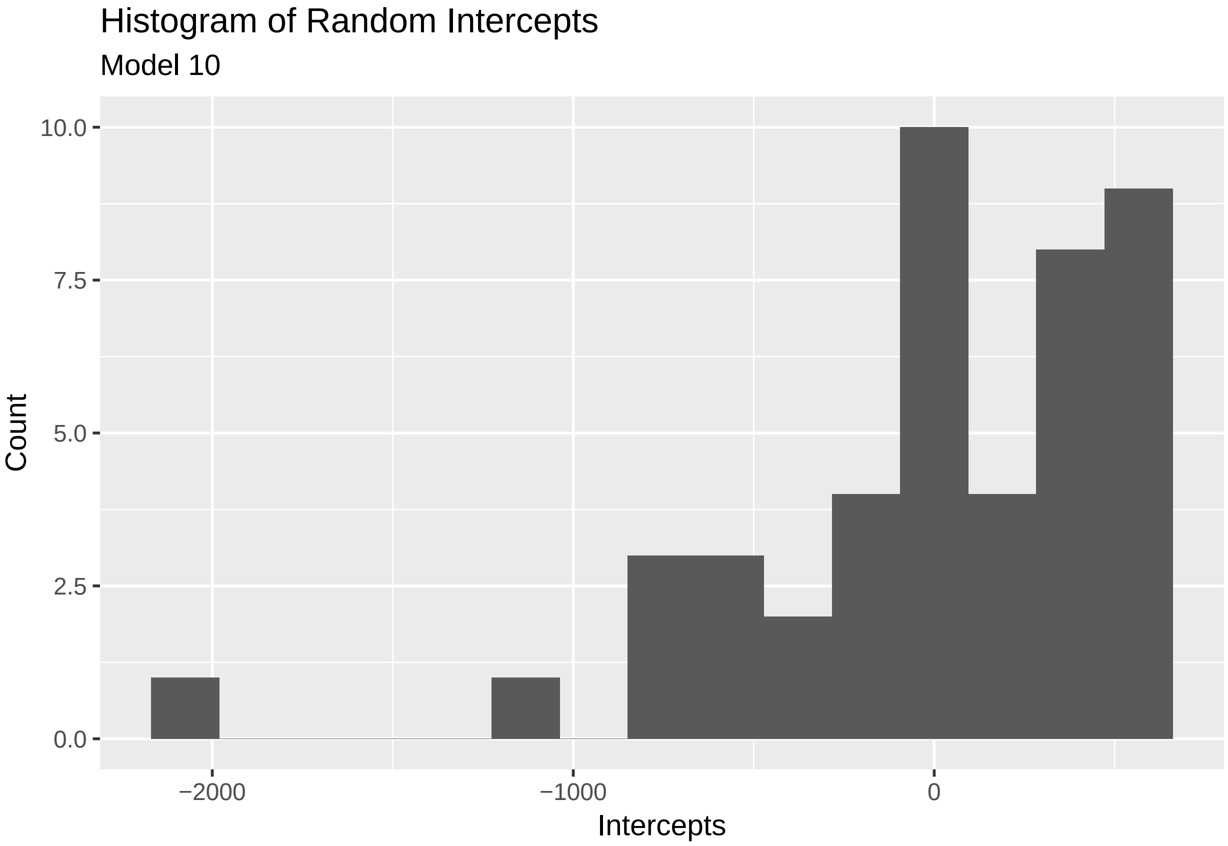

Histogram of Random Intercepts
Model 9
Count
10
5
0
−20
−10
0
10
20
Intercepts

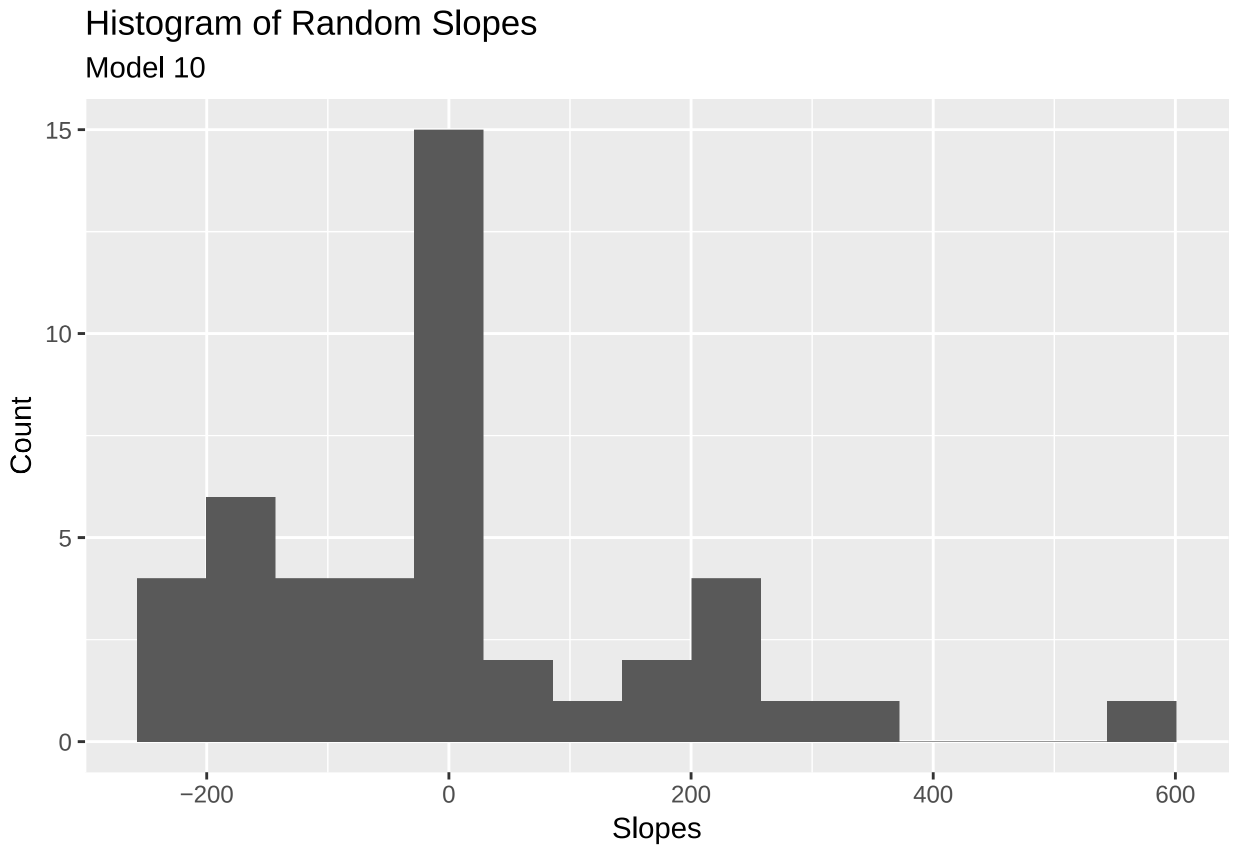

Histogram of Random Slopes
Model 9
Count
10
5
0
−0.01
0.00
0.01
Slopes

Model 10 Residual Plots

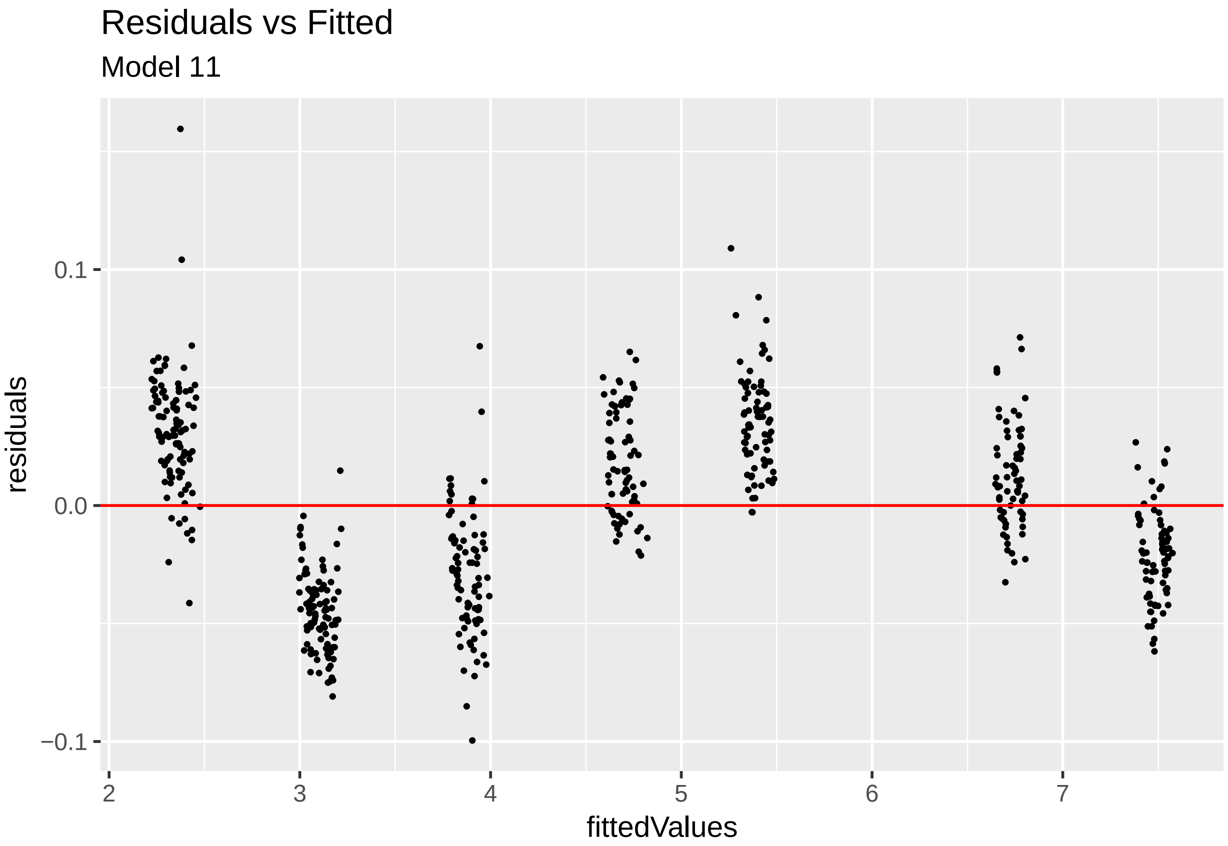


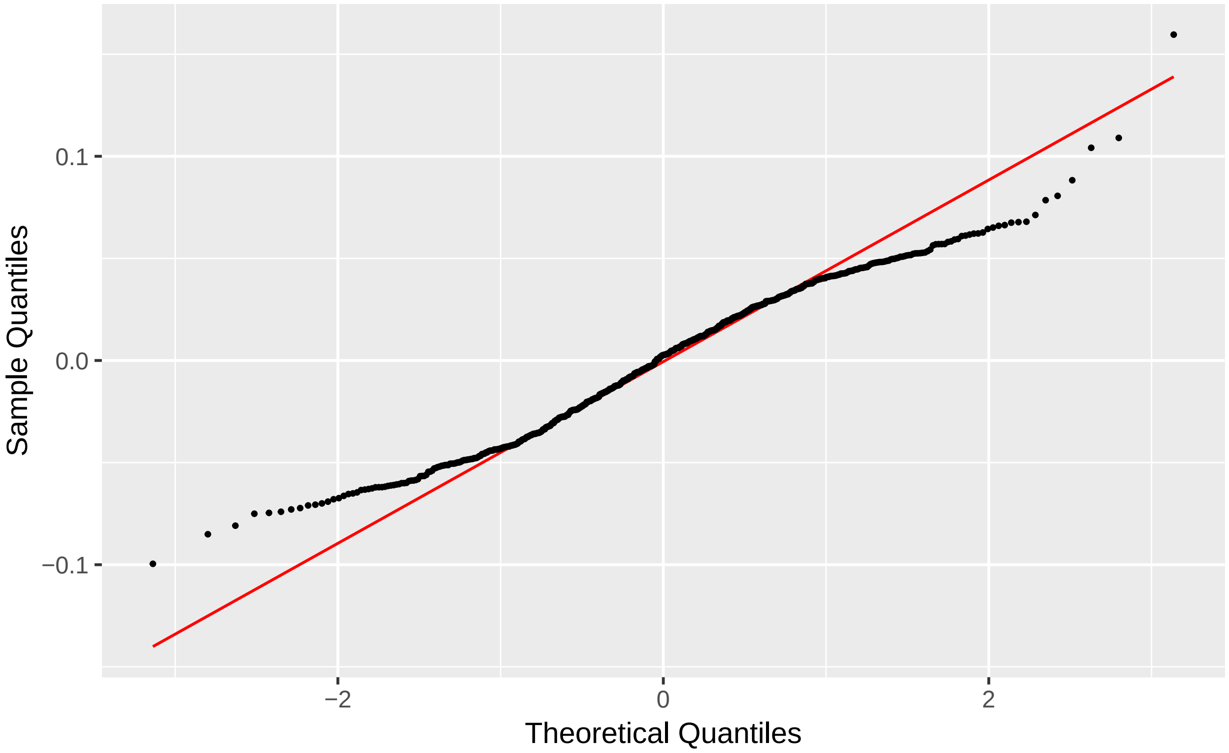

QQ−Plot
Model 10
Sample Quantiles
1000
500
0
−500
−2
0
2
Theoretical Quantiles


Histogram of Random Intercepts
Model 10
Count
10.0
7.5
5.0
2.5
0.0
−2000
−1000
0
Intercepts


Histogram of Random Slopes
Model 10
Count
15
10
5
0
−200
0
200
400
600
Slopes

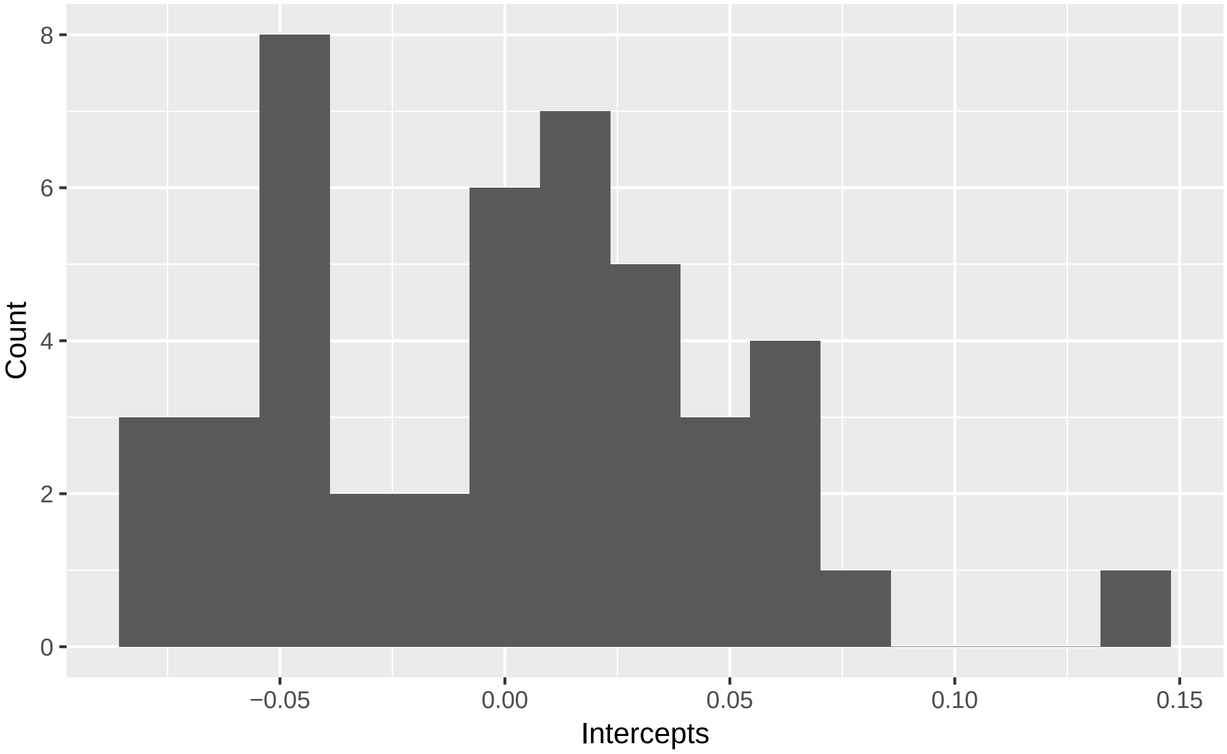
Model 11 Residual Plots





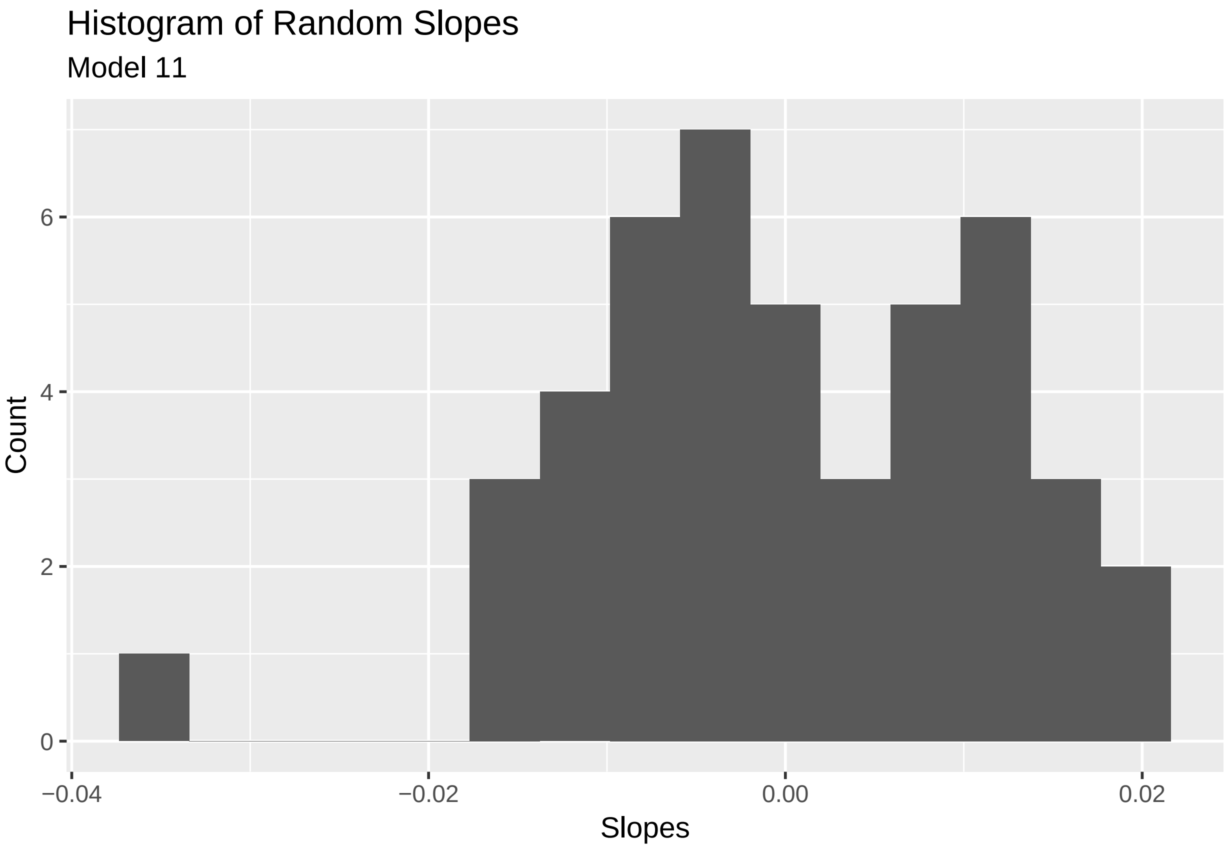
QQ−Plot

Model 11



Histogram of Random Intercepts

Model 11



Questions:

Asked in class 11/29

- Explain why we did logdist (lmer errors, very different scales)
- How can we test/compare models once we do log transformations for both explanatory and responses variables? cant compare AIC with different responses, resid plots
- Once we do logTime, AIC becomes negative. What do we do with that? more negative is better
- Once we do logTime, 3 variables become insignificant (1 main, 2 interactions) and we know at least the BMI variable should be significant. Does everything have to be log transformed?
- Just general tips for dealing with transformations
- Is it okay to have year in as a level 1 explanatory

does seem like you might have a couple of influential observations too? Might also consider using factors for some of the variables (small vs. large)?

Is every country represented every year?
Do you expect the "within group" correlation to be larger by year or by country?
Are you looking at multiple athletes per event per country each year or just the fastest? Or are year and athelete id kinda the same thing?