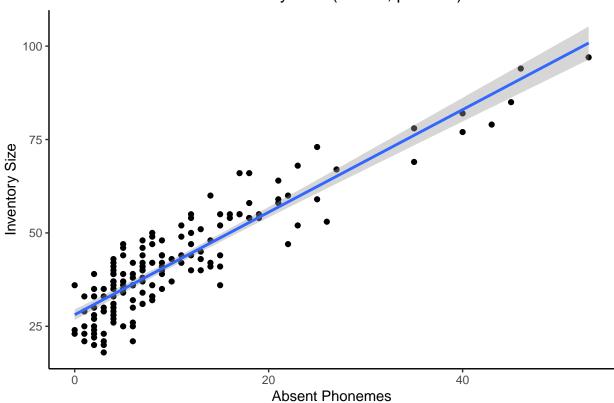
Correlation Results

Do languages with a bigger inventory have more absent?

Languages with more phonemes have a much much higher rate of missing phonemes.

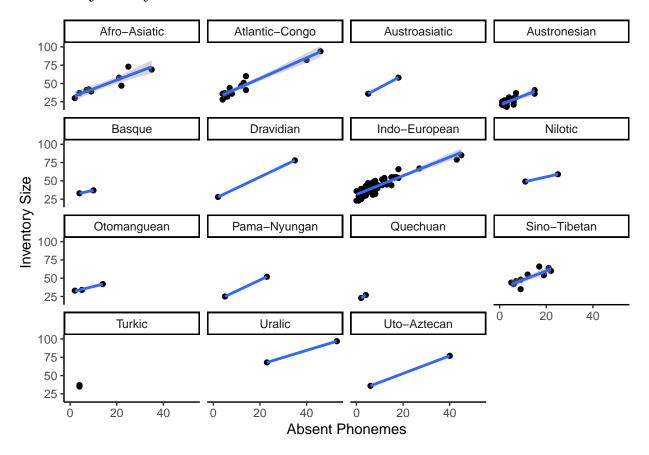
```
## Warning in cor.test.default(x, y, method = "spearman"): Cannot compute
## exact p-value with ties
```

Absent Phonemes vs Inventory Size (r=0.85, p=0.00*)



```
## Warning in cor.test.default(cov$Unobserved, cov$InventoryLength, method =
## "spearman"): Cannot compute exact p-value with ties
##
## Spearman's rank correlation rho
##
## data: cov$Unobserved and cov$InventoryLength
## S = 97467, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
## rho
## 0.8517301</pre>
```

Faceted by family

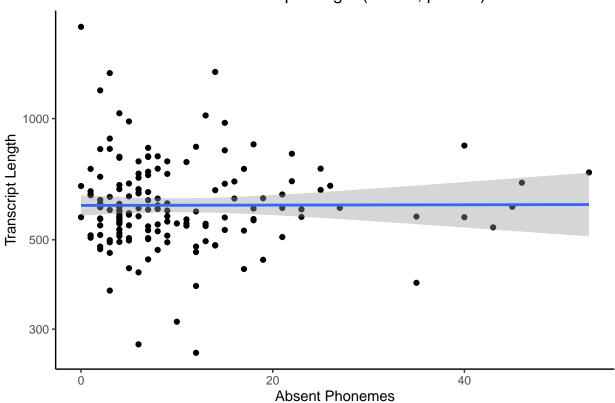


Does coverage get better with a longer transcript?

- Coverage is not really getting much better as transcripts get longer.
- Only a very weak and non-significant effect.
- Note the log scale on y axis as transcript length spans an order of magnitude.

Warning in cor.test.default(x, y, method = "spearman"): Cannot compute
exact p-value with ties

Absent Phonemes vs Transcript Length (r=0.01, p=0.85)



```
## Warning in cor.test.default(cov$Unobserved, cov$TranscriptLength, method =
## "spearman"): Cannot compute exact p-value with ties
##
## Spearman's rank correlation rho
##
## data: cov$Unobserved and cov$TranscriptLength
## S = 647501, p-value = 0.8516
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
## rho
## 0.01499702
```

```
library(RColorBrewer)
colorCount = length(unique(cov$Family))
getPalette = colorRampPalette(brewer.pal(9, "Set1"))
p <- ggplot(cov, aes(x=Unobserved, y=InventoryLength, color=Family))</pre>
p <- p + geom_jitter()</pre>
p <- p + scale_color_manual(values = getPalette(colorCount))</pre>
p <- p + ggtitle("Absent Phonemes vs Inventory Size")</pre>
p <- p + xlab('Absent Phonemes') + ylab("Inventory Size")</pre>
p <- p + theme_classic()</pre>
p <- p + theme(legend.position="bottom")</pre>
#p <- p + guides(fill=guide_legend(nrow=2))</pre>
plot(p)
        Absent Phonemes vs Inventory Size
    100
 Inventory Size
     75
     50
     25
                                           20
                                                                           40
                                            Absent Phonemes
           Afro-Asiatic
                                       Cahuapanan
                                                          Matacoan
                                                                                Pano-Tacanan
                                                                                                   Ts
           Araucanian
                                       Dravidian
                                                          Morehead-Wasur
                                                                                Quechuan
                                                                                                   Tu
           Arawakan
                                       Ijoid
                                                          Muskogean
                                                                                Sahaptian
                                                                                                   Ur
           Athabaskan-Eyak-Tlingit
                                       Indo-European
                                                          Nara
                                                                                Salishan
                                                                                                   Ut
amily
           Atlantic-Congo
                                       Japonic
                                                          Nilotic
                                                                                Sandawe
                                                                                                   Wa
           Austroasiatic
                                       Kartvelian
                                                          Nyulnyulan
                                                                                Seri
           Austronesian
                                       Koreanic
                                                          Otomanguean
                                                                                Sino-Tibetan
           Basque
                                                          Pama-Nyungan
                                                                                Tai-Kadai
                                       Kunama
pdf('scatter-absences_vs_inventory_length-colored.pdf')
print(p)
x <- dev.off()
p <- ggplot(cov, aes(x=Unobserved, y=TranscriptLength, color=Family))</pre>
p <- p + geom_jitter()</pre>
p <- p + scale_color_manual(values = getPalette(colorCount))</pre>
p <- p + ggtitle("Absent Phonemes vs Transcript Length")</pre>
```

p <- p + xlab('Absent Phonemes') + ylab("Transcript Length")</pre>

p <- p + theme_classic()</pre>

p <- p + theme(legend.position="bottom")
#p <- p + guides(fill=guide_legend(nrow=2))</pre>

