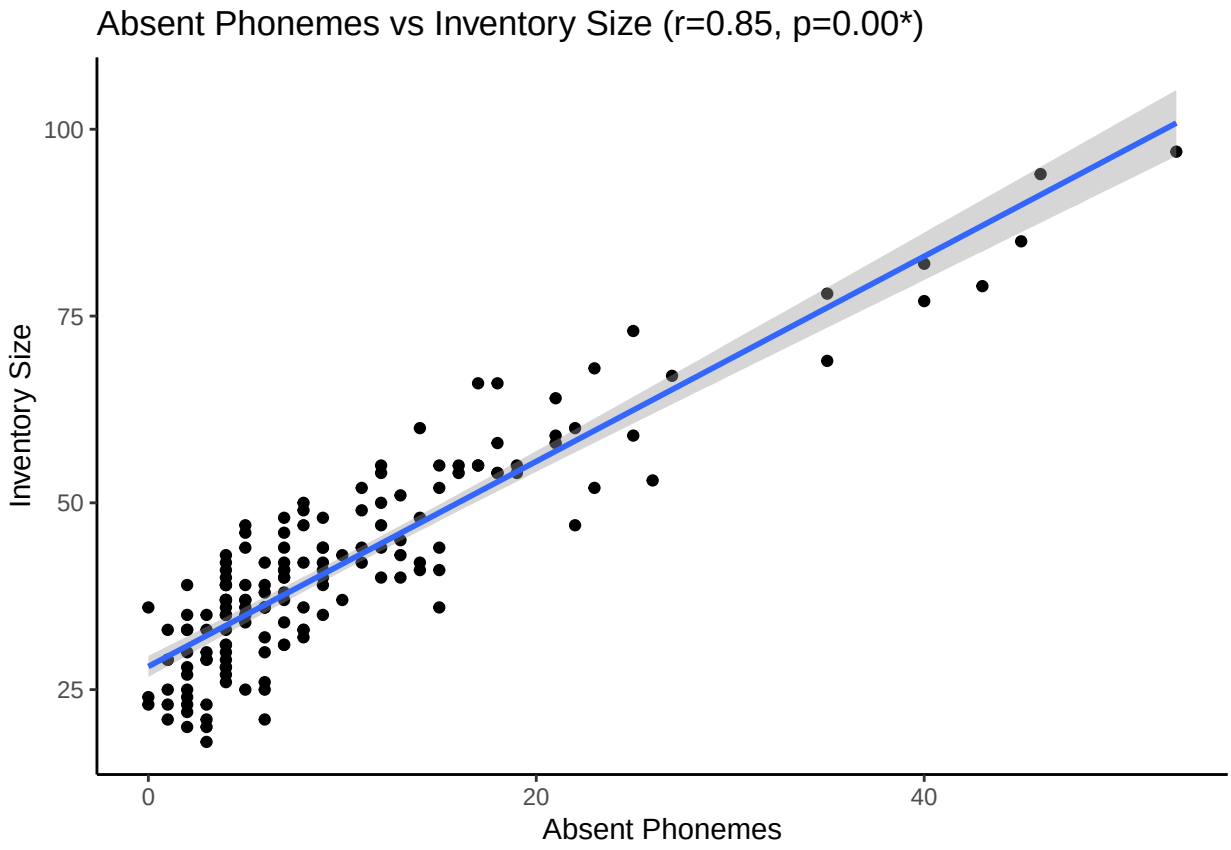
Correlation Results

Do languages with a bigger inventory have more absent?

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

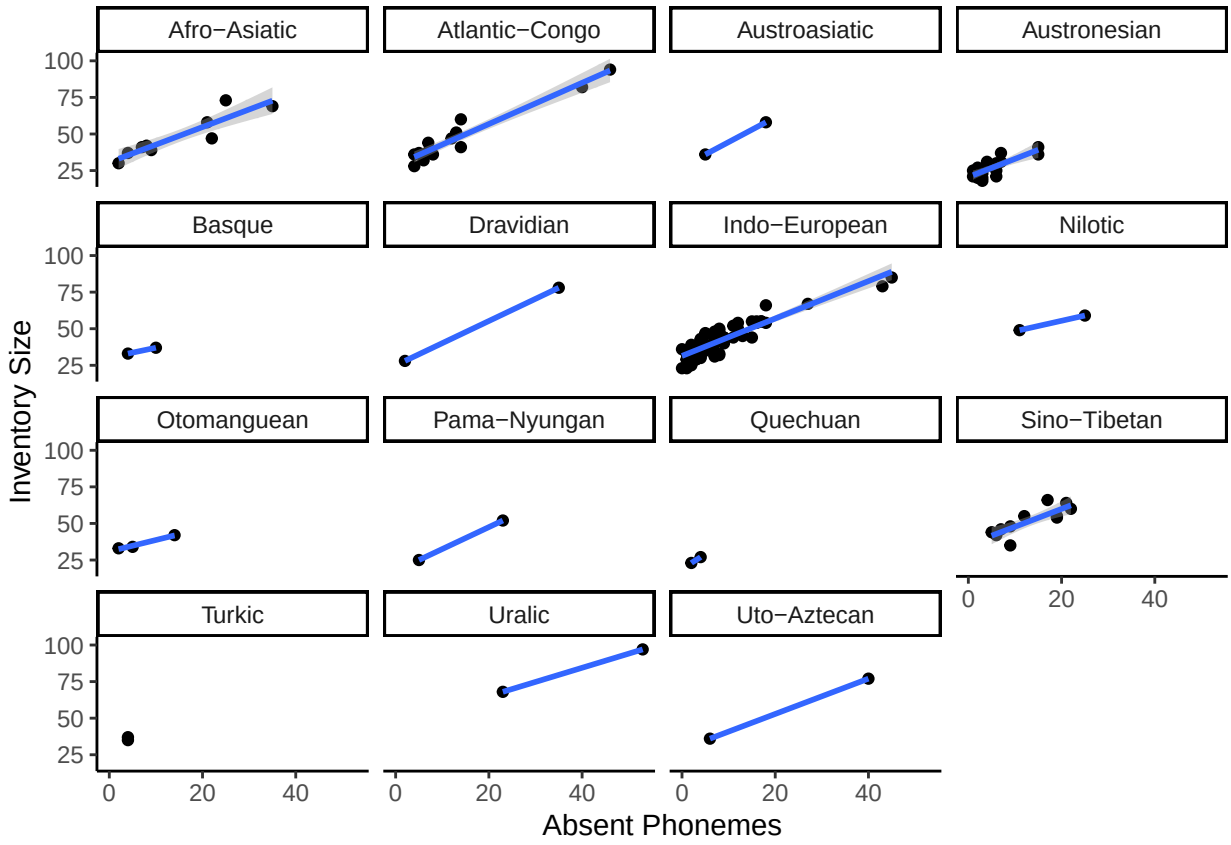
```
## Warning in cor.test.default(x, y, method = "spearman"): Cannot compute
## exact p-value with ties
```




```
## 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
```

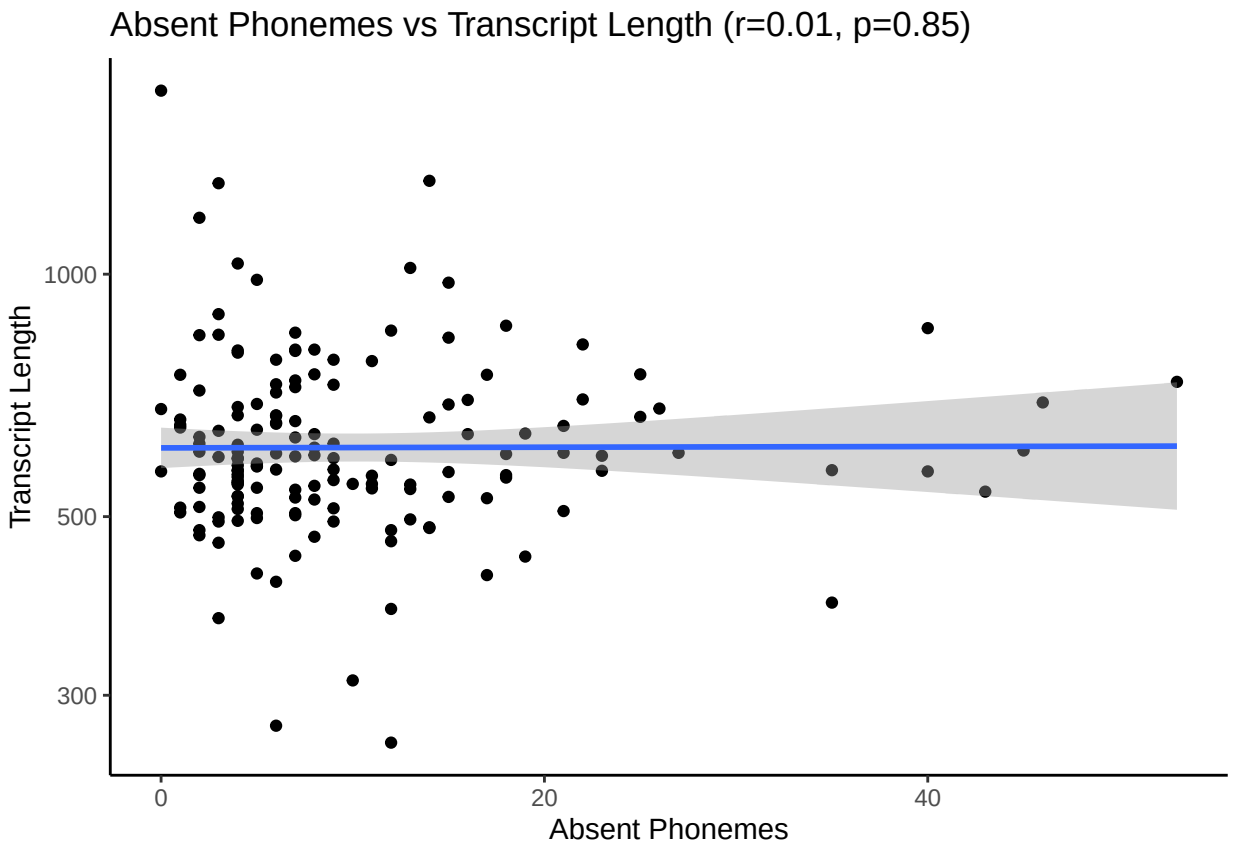
Faceted by family



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.

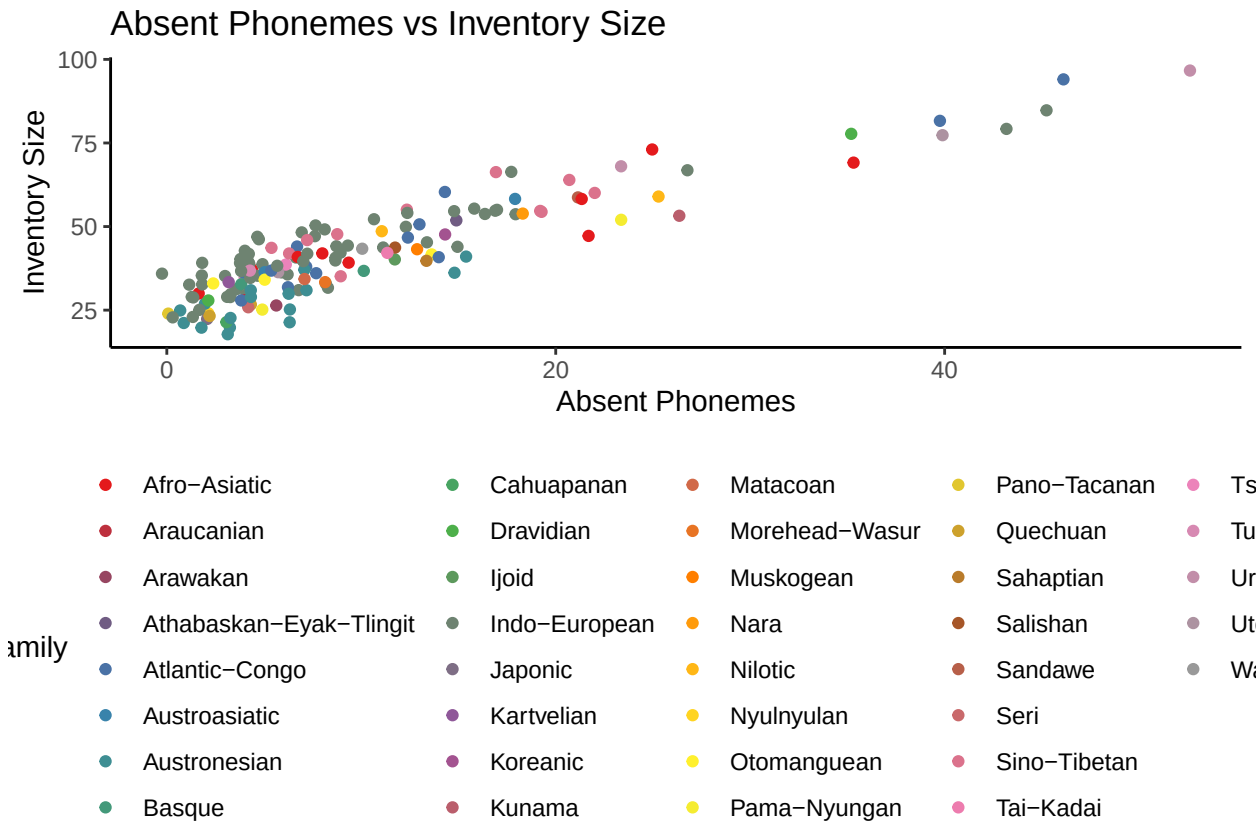
```
## Warning in cor.test.default(x, y, method = "spearman"): Cannot compute
## exact p-value with ties
```




```
## 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
```

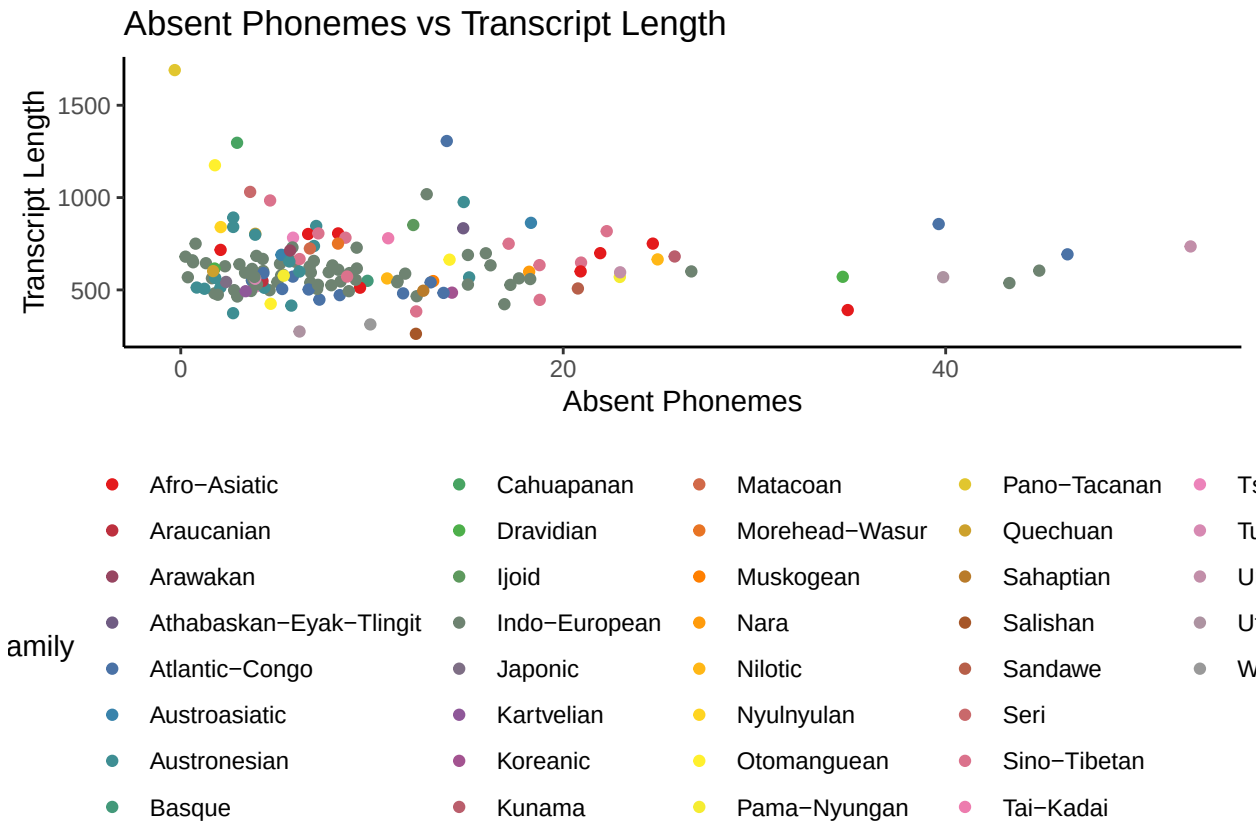
```
library(RColorBrewer)
colorCount = length(unique(cov$Family))
getPalette = colorRampPalette(brewer.pal(9, "Set1"))

p <- ggplot(cov, aes(x=Unobserved, y=InventoryLength, color=Family))
p <- p + geom_jitter()
p <- p + scale_color_manual(values = getPalette(colorCount))
p <- p + ggtitle("Absent Phonemes vs Inventory Size")
p <- p + xlab('Absent Phonemes') + ylab("Inventory Size")
p <- p + theme_classic()
p <- p + theme(legend.position="bottom")
#p <- p + guides(fill=guide_legend(nrow=2))

plot(p)
```




```
pdf('scatter-absences_vs_inventory_length-colored.pdf')
print(p)
x <- dev.off()
```

```
p <- ggplot(cov, aes(x=Unobserved, y=TranscriptLength, color=Family))
p <- p + geom_jitter()
p <- p + scale_color_manual(values = getPalette(colorCount))
p <- p + ggtitle("Absent Phonemes vs Transcript Length")
p <- p + xlab('Absent Phonemes') + ylab("Transcript Length")
p <- p + theme_classic()
p <- p + theme(legend.position="bottom")
#p <- p + guides(fill=guide_legend(nrow=2))
```

```
plot(p)
```




```
pdf('scatter-absences_vs_transcript_length-colored.pdf')
print(p)
x <- dev.off()
```