431 Class 22

Thomas E. Love

2018-11-15

Today's Agenda

- Regression Comparison of Means, with Covariate Adjustment
- Project Study 2 Demonstration

Today's R Setup

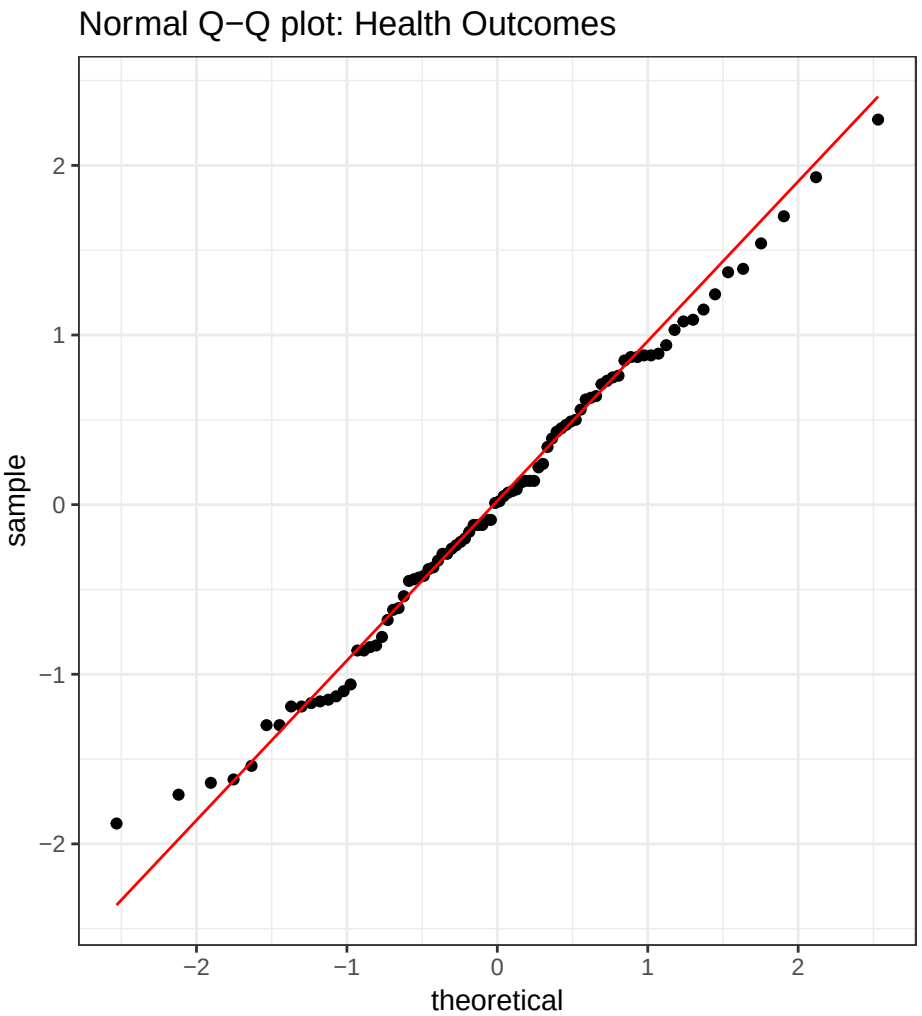
```
library(Hmisc); library(magrittr); library(broom)
library(readxl) # to read in .xlsx file
library(tidyverse) # always load tidyverse last
```

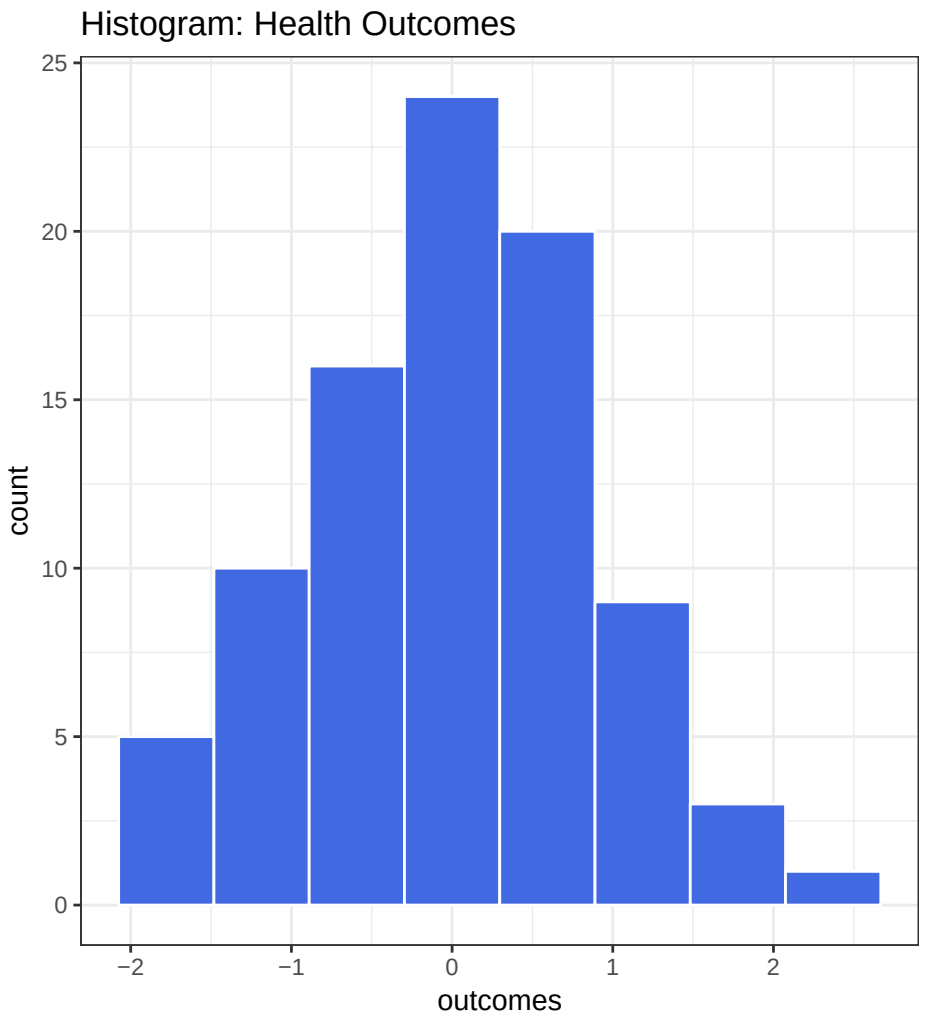
County Health Rankings Data for Ohio, 2018

Data Source:
http://www.countyhealthrankings.org/app/ohio/2018/downloads

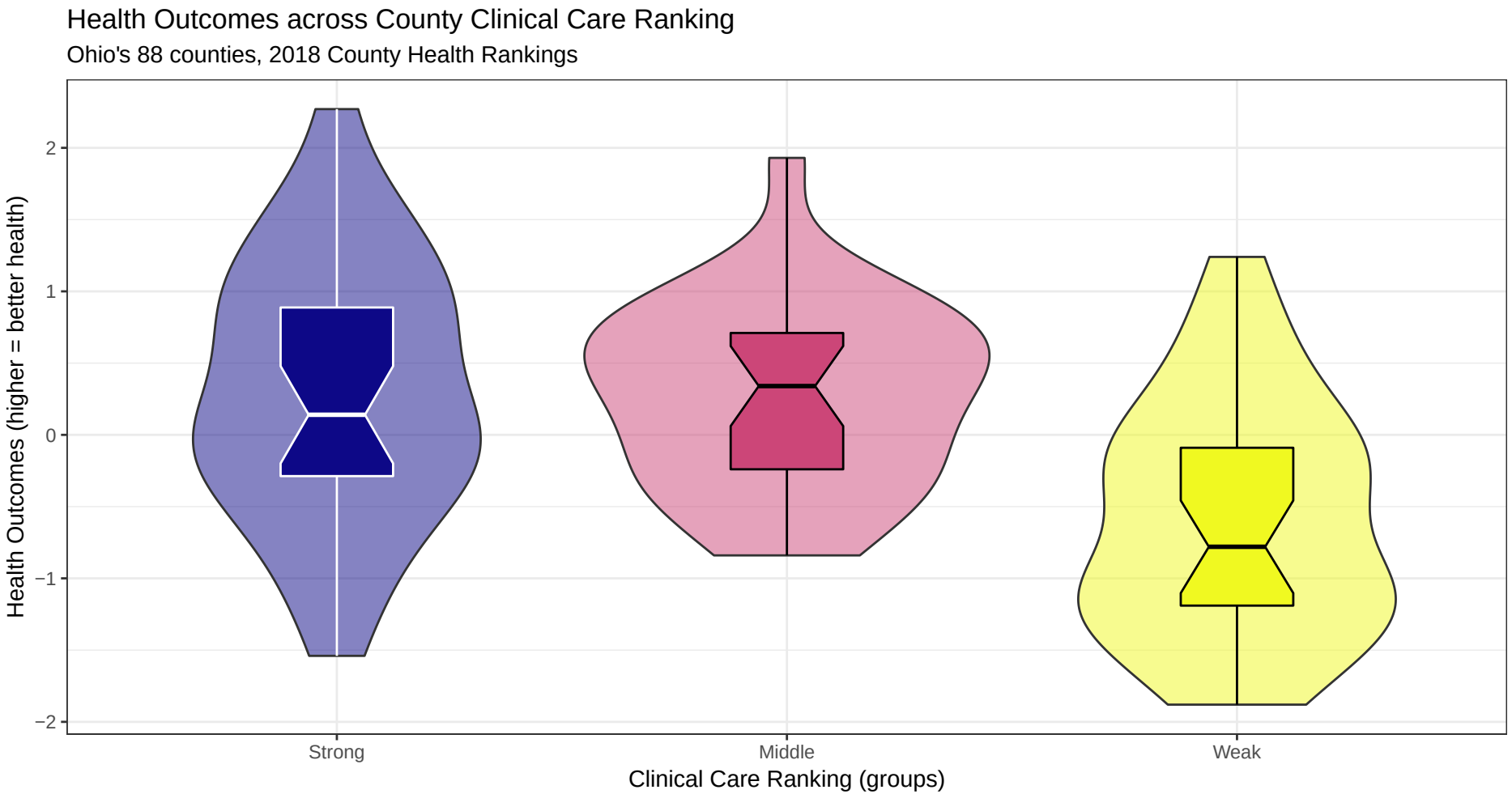
```
ohio18 <- read_xlsx("data/ohio_2018_rankings.xlsx") %>%
  mutate(behavior = cut2(rk_behavior, g = 4),
         clin_care = cut2(rk_clin_care, g = 3)) %>%
  mutate(behavior = fct_recode(behavior,
            "Best" = "[ 1,23)", "High" = "[23,45)",
            "Low" = "[45,67)", "Worst" = "[67,88]")) %>%
  mutate(clin_care = fct_recode(clin_care,
            "Strong" = "[ 1,31)", "Middle" = "[31,60)",
            "Weak" = "[60,88]")) %>%
  mutate(density = factor(density)) %>%
  select(FIPS, state, county, outcomes,
         behavior, clin_care, density, income)
```

Health Outcomes (Normally Distributed?)





Health Outcomes by Clinical Care Groups



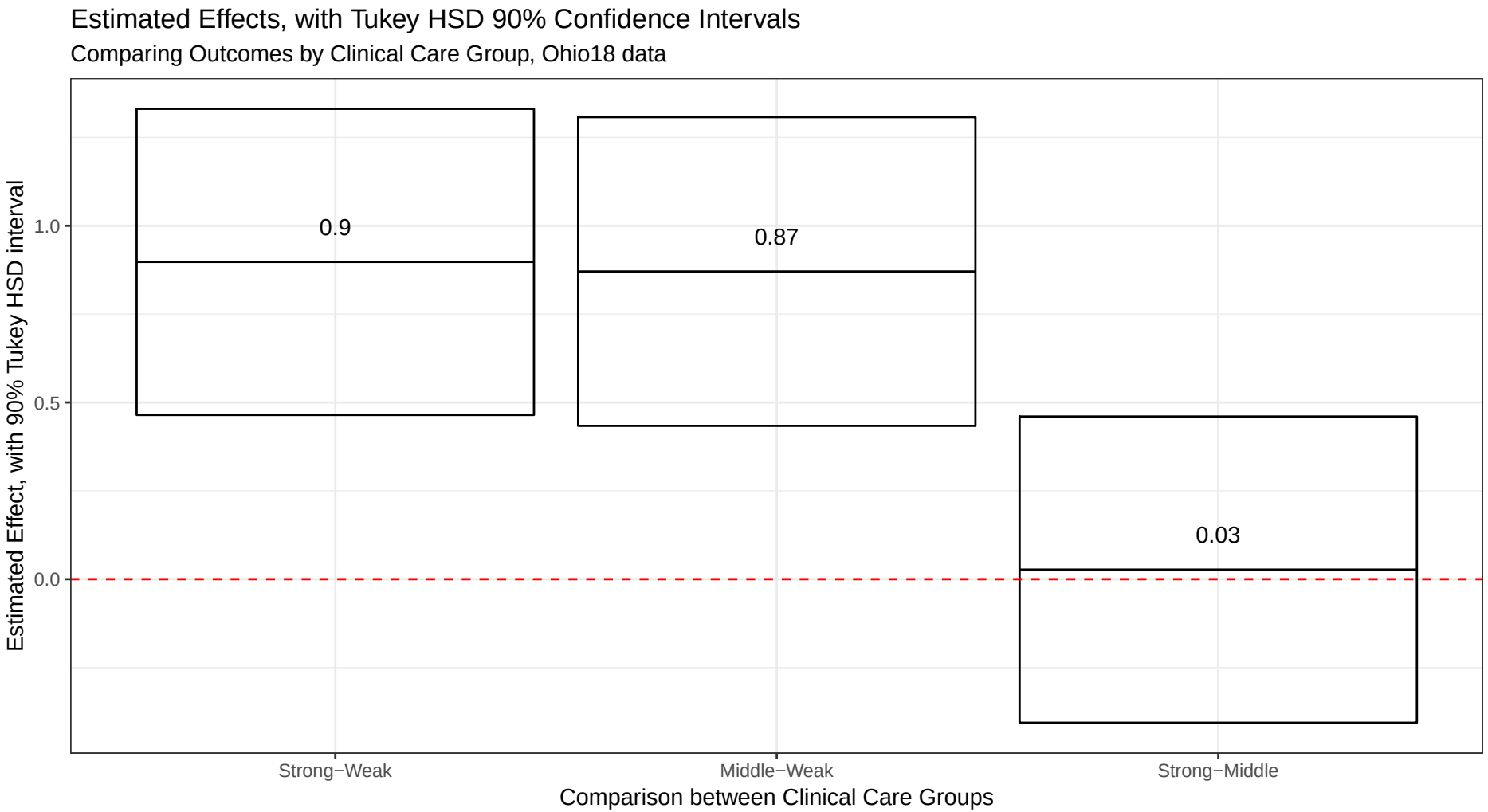
Unadjusted - ANOVA and 90% Tukey HSD intervals

```
model_unadj <- lm(outcomes ~ clin_care, data = ohio18)

tidy(anova(model_unadj))
```

```
# A tibble: 2 x 6
  term          df sumsq meansq statistic    p.value
  <chr>      <int> <dbl>  <dbl>     <dbl>      <dbl>
1 clin_care      2  15.2  7.61       11.9  0.0000276
2 Residuals     85  54.3  0.639      NA   NA
```

```
tukey_unadj <- tidy(TukeyHSD(aov(model_unadj),
                             ordered = TRUE,
                             conf.level = 0.90))
```

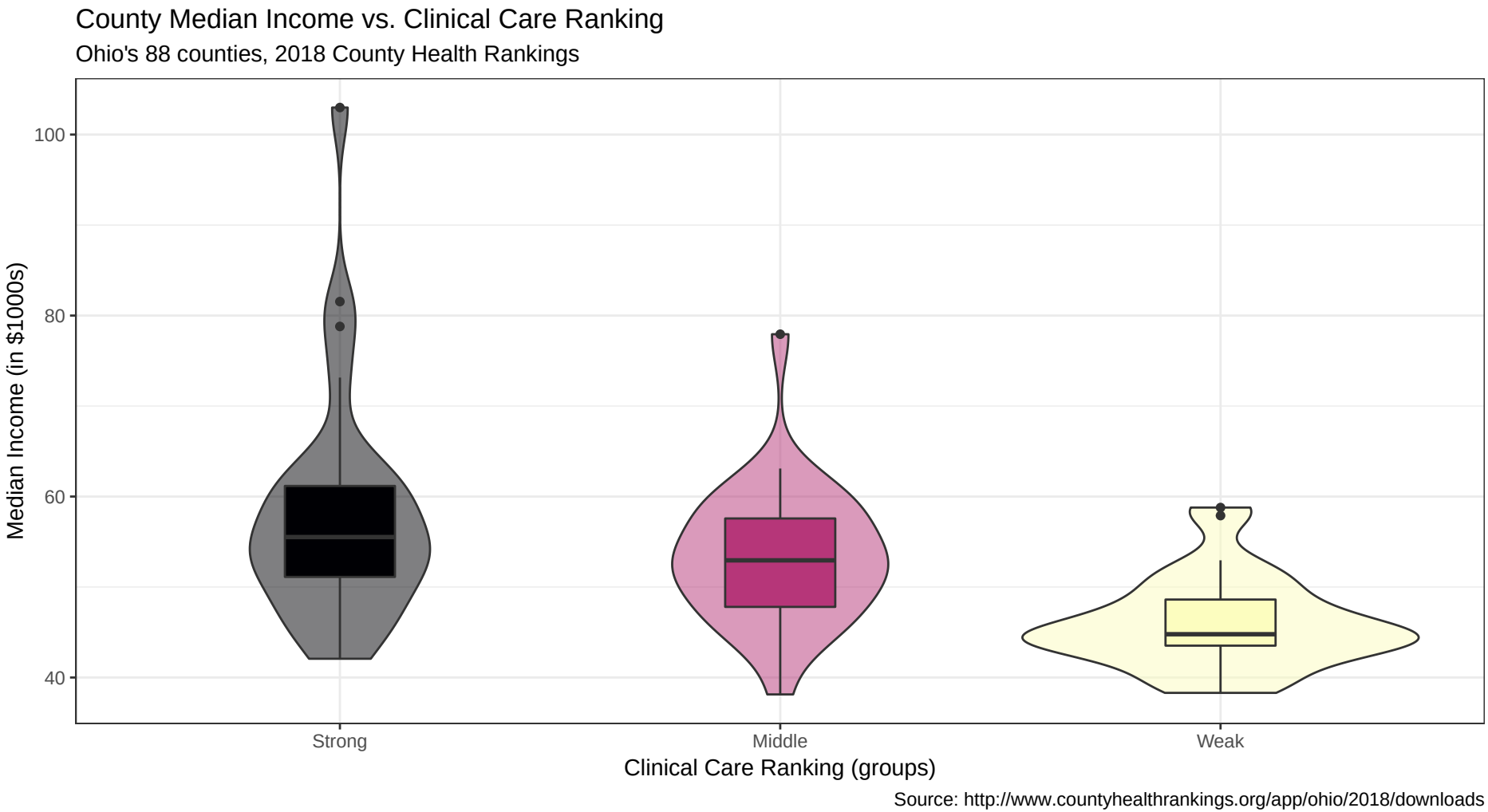
Tukey HSD results, unadjusted ANOVA



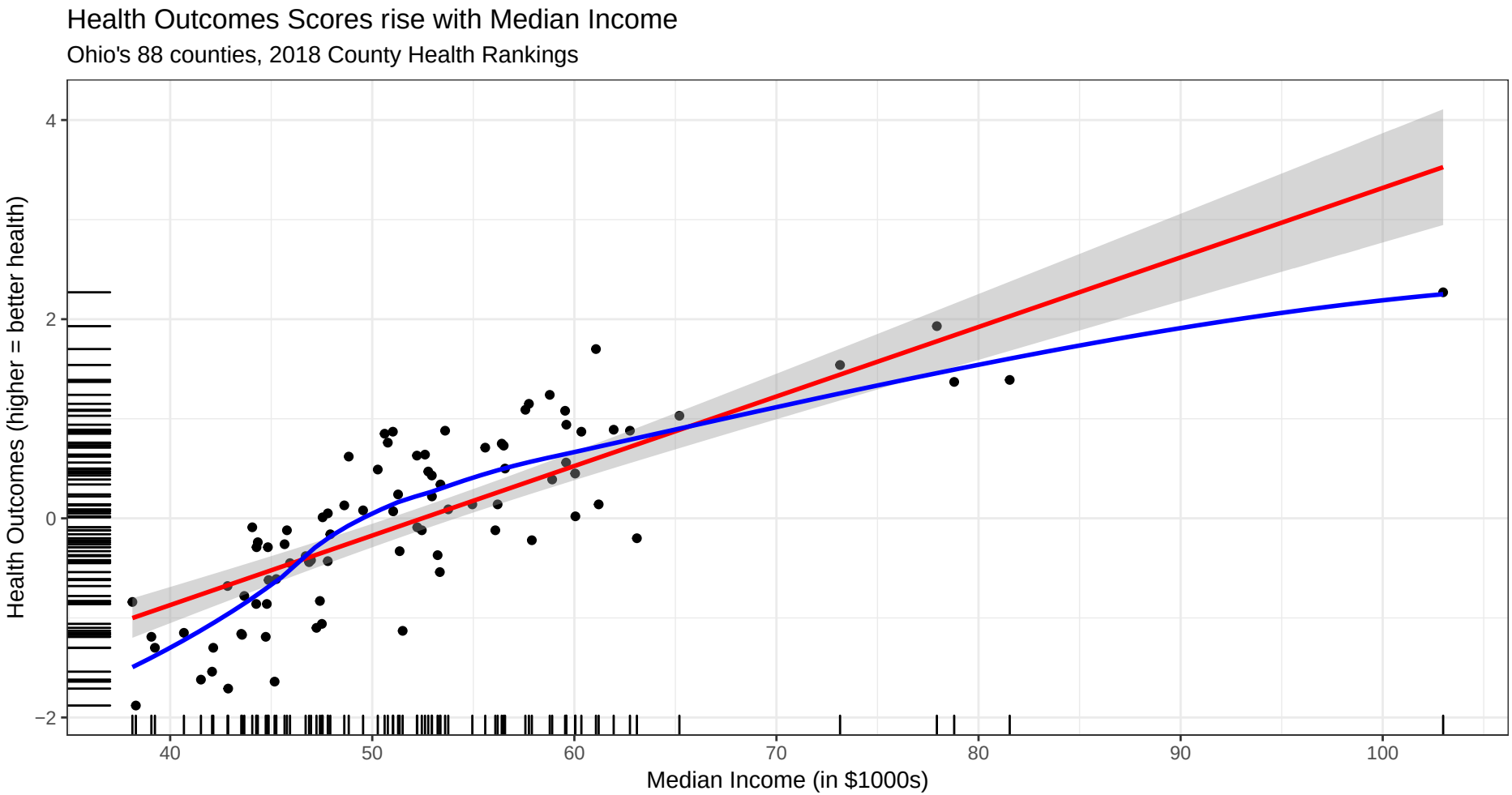
Our New Question

1. Do groups of counties defined by clinical care still show meaningful differences in average health outcomes, **after** adjustment for differences in their median income levels?

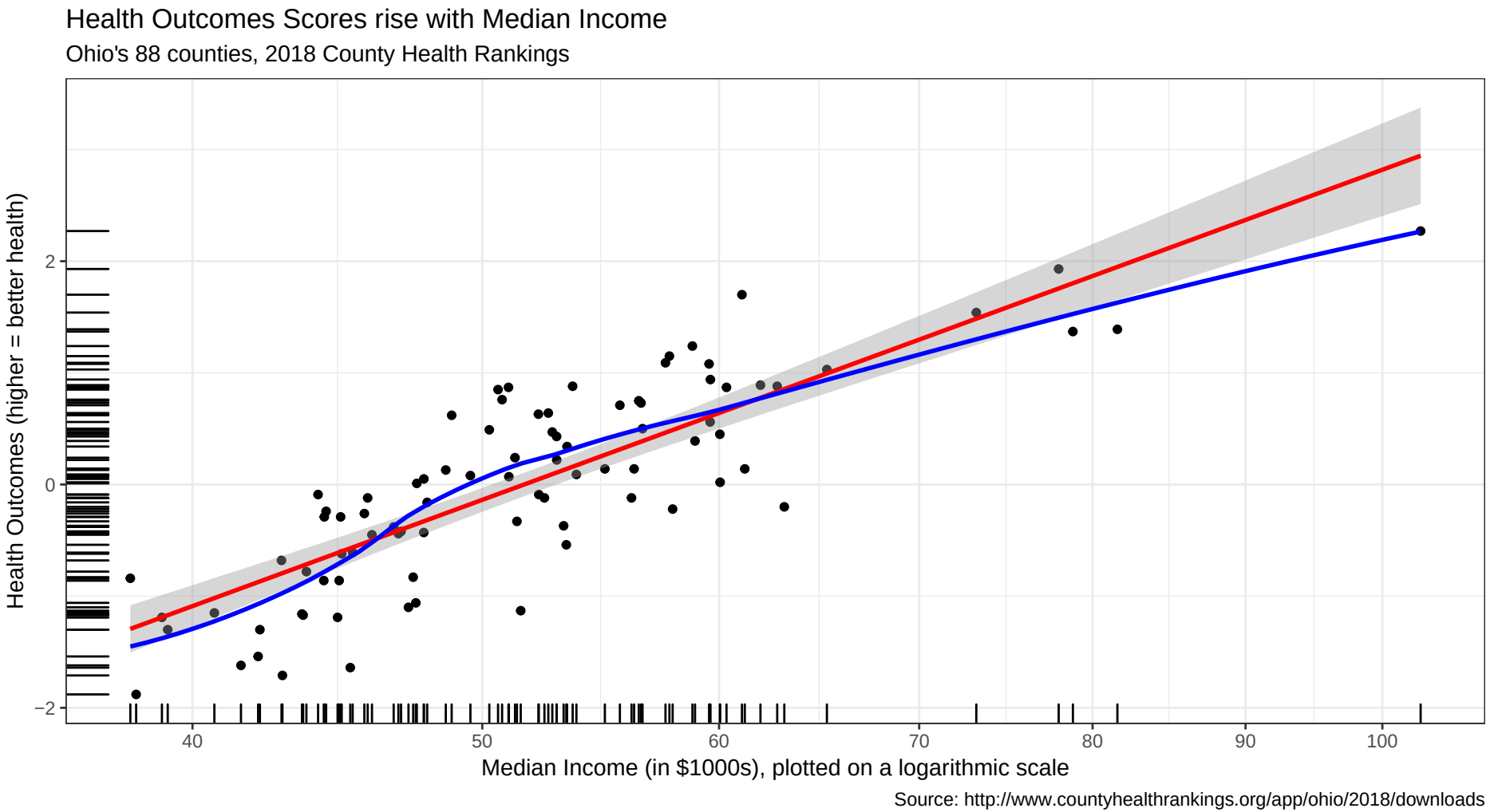
Income by Clinical Care Groups



Income vs. Outcome



Income (on the Log scale) vs. Outcome



Our New Model

```
ohio18 <- ohio18 %>%
  mutate(incK = income/1000)
```

```
model_adj1 <- lm(outcomes ~ incK + clin_care, data = ohio18)

anova(model_adj1)
```

```
Analysis of Variance Table

Response: outcomes
           Df Sum Sq Mean Sq  F value  Pr(>F)
incK        1 44.349  44.349 161.7268 < 2e-16 ***
clin_care   2  2.163   1.082   3.9446 0.02305 *
Residuals  84 23.035   0.274
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpreting the ANOVA table with a covariate

```
anova(lm(outcomes ~ incK + clin_care, data = ohio18)) %>%
  tidy() %>% knitr::kable(digits = 3)
```

term	df	sumsq	meansq	statistic	p.value
incK	1	44.349	44.349	161.727	0.000
clin_care	2	2.163	1.082	3.945	0.023
Residuals	84	23.035	0.274	NA	NA

- This ANOVA table tests the predictors, in order.
- The `incK` p value tests H_0: `incK` adds no predictive value to the model, as compared to a model with an intercept alone.
  - Compares the [intercept only] model to the `incK` model.
- The `clin_care` F and p value tests H_0: `clin_care` adds no incremental predictive value to the model that already includes `incK`.
  - Compares the `incK` to the `incK` and `clin_care` model.

What if we reverse the order in which we create the model?

This ANOVA table is sequential

```
anova(lm(outcomes ~ clin_care + incK, data = ohio18)) %>%
  tidy() %>% knitr::kable(digits = 3)
```

term	df	sumsq	meansq	statistic	p.value
clin_care	2	15.221	7.610	27.752	0
incK	1	31.292	31.292	114.111	0
Residuals	84	23.035	0.274	NA	NA

- Notice the change in *p* value for `clin_care`. That *p* value now compares [intercept only] to `clin_care`, ignoring the covariate.
  - We saw that result last time, in our ANOVA modeling.
- The `incK` test now assesses the incremental value of one predictor (`incK`) after you already have the other (`lin_care`) in the model.
- Either way, though, it looks like both `incK` and `clin_care` are useful. How much of the variation do they explain, together?

(Edited) Summary of the Adjusted Model

```
summary(model_adj1)
```

```
lm(formula = outcomes ~ incK + clin_care, data = ohio18)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.566447   0.374786  -9.516 5.43e-15 ***
incK             0.066894   0.006262  10.682  < 2e-16 ***
clin_careMiddle  0.275345   0.139278   1.977   0.0513 .
clin_careWeak   -0.105040   0.155254  -0.677   0.5005

Residual standard error: 0.5237 on 84 degrees of freedom
Multiple R-squared:  0.6688,    Adjusted R-squared:  0.657
F-statistic: 56.54 on 3 and 84 DF,  p-value: < 2.2e-16
```

What % of the variation in outcomes do incK and clin_care explain?

Unadjusted vs. Adjusted Model

```
glance(lm(outcomes ~ clin_care, data = ohio18)) %>%
  knitr::kable(digits = 3)
```

	r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik
value	0.219	0.2	0.799	11.907	0	3	-103.644

```
glance(lm(outcomes ~ clin_care + incK, data = ohio18)) %>%
  knitr::kable(digits = 3)
```

	r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik
value	0.669	0.657	0.524	56.539	0	4	-65.892

Does `incK` add a substantial amount of predictive value?

Predict the outcome at the average level of the covariate for each group

At the mean level of `incK`, 52.474, predict the values of `outcomes` for counties in each `clin_care` category.

```
new_dat <- data_frame(
  clin_care = c("Strong", "Middle", "Weak"),
  incK = rep(mean(ohio18$incK), 3))
new_dat
```

```
# A tibble: 3 x 2
  clin_care  incK
  <chr>     <dbl>
1 Strong     52.5
2 Middle     52.5
3 Weak       52.5
```

Predict the outcome at the average level of the covariate for each group, using a 90% prediction interval

```
preds_adj <- predict(model_adj1, newdata = new_dat,
        interval = "prediction", level = 0.90)

bind_cols(new_dat, data.frame(preds_adj)) %>%
  knitr::kable(digits = 3)
```

clin_care	incK	fit	lwr	upr
Strong	52.474	-0.056	-0.943	0.831
Middle	52.474	0.219	-0.667	1.105
Weak	52.474	-0.161	-1.050	0.727

Tukey HSD after covariate adjustment?

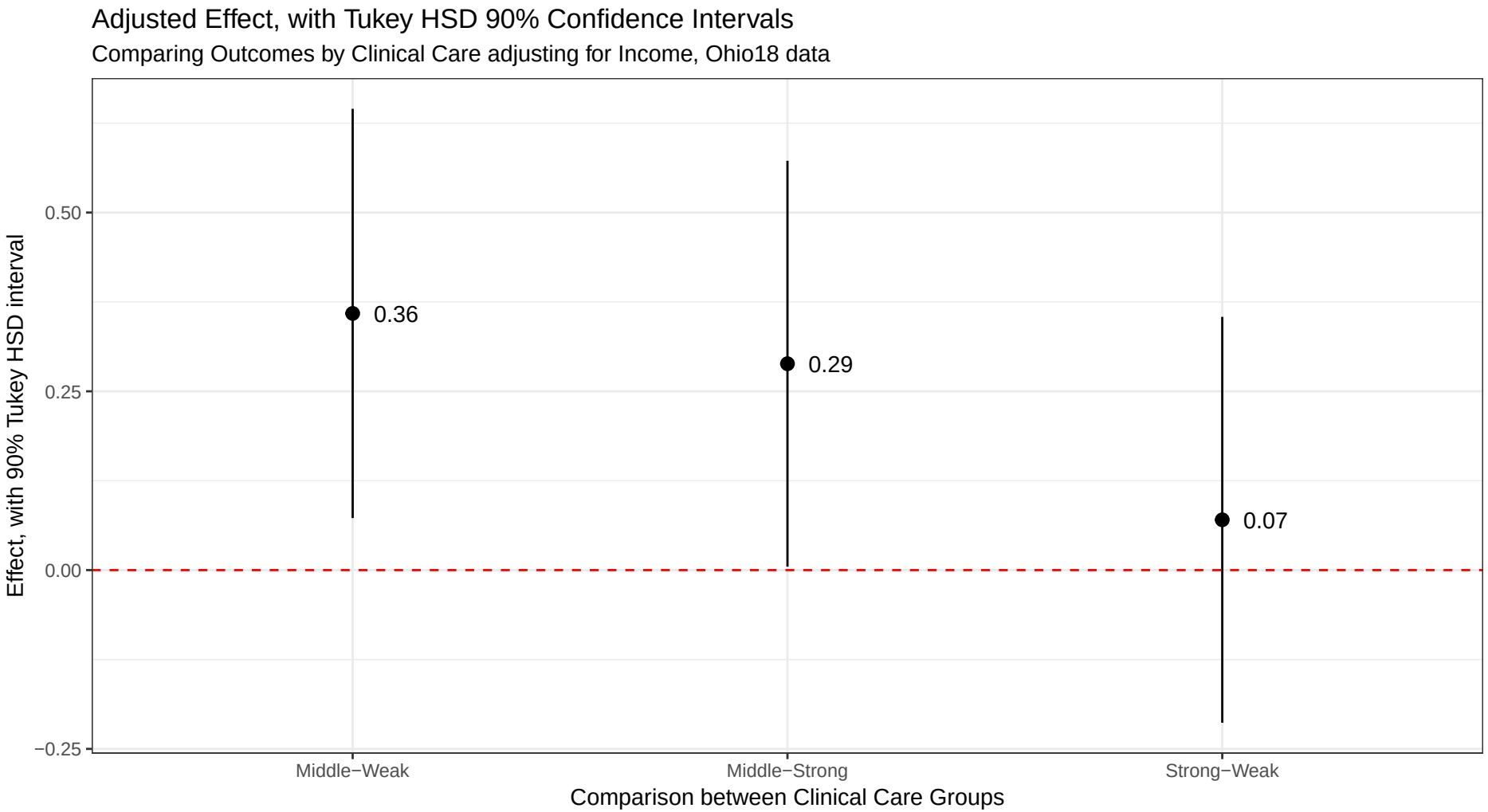
```
tukey_adj <- TukeyHSD(
  aov(outcomes ~ incK + clin_care, data = ohio18),
  which = "clin_care", ordered = TRUE, conf.level = 0.9) %>%
  tidy()
```

```
Warning in replications(paste("~", xx), data = mf):
non-factors ignored: incK
```

```
tukey_adj %>% knitr::kable(digits = 3)
```

term	comparison	estimate	conf.low	conf.high	adj.p.value
clin_care	Strong-Weak	0.070	-0.213	0.354	0.864
clin_care	Middle-Weak	0.359	0.073	0.645	0.029
clin_care	Middle-Strong	0.289	0.005	0.572	0.093

Tukey HSD results, after adjustment for `income`



Checking Regression Assumptions

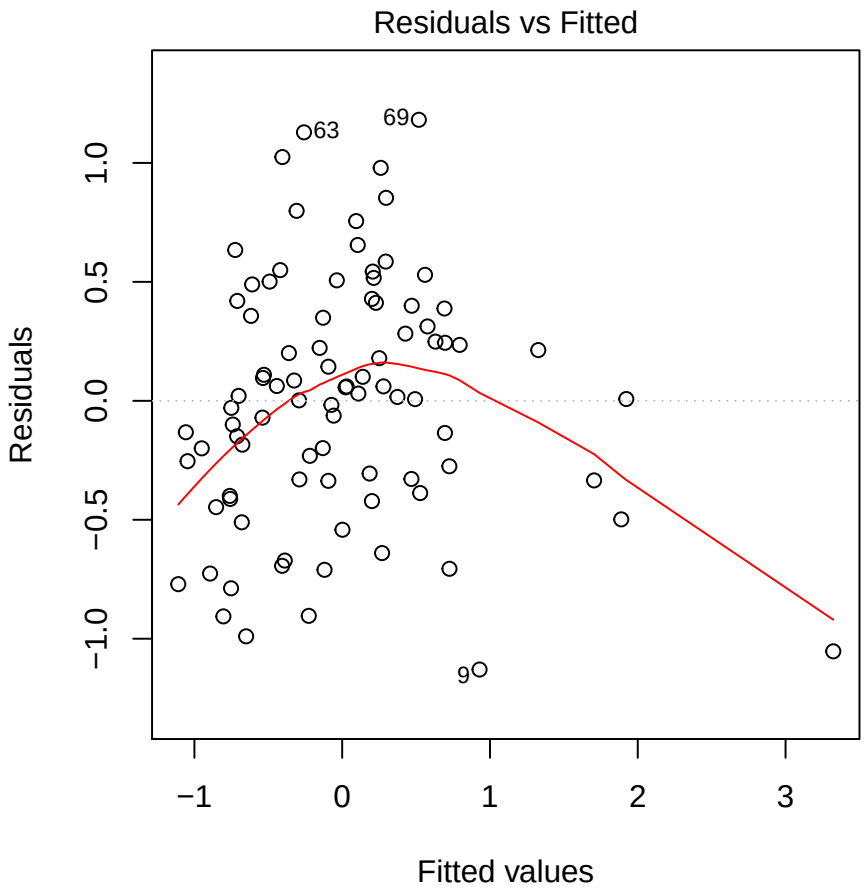
On the next slide, we'll build two quick plots...

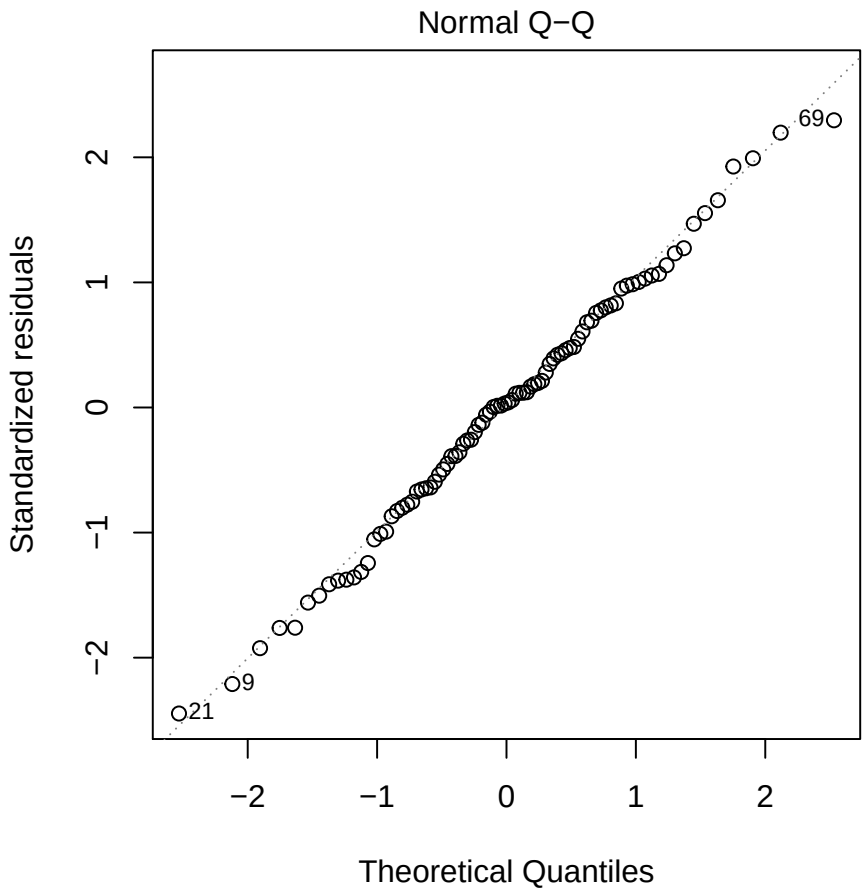
- If the residuals vs. fitted values plot looks like a fuzzy football, with no particular pattern or trend, then we're in good shape for the moment with the assumption of **linearity**.
- If the Normal Q-Q plot of standardized residuals looks like a straight line (so we'd assume a Normal model held for the residuals), then we're in good shape with the assumption of **Normality**.

How do these plots look?

Residual Plots

```
par(mfrow = c(1,2))
plot(model_adj1, which = 1:2)
```





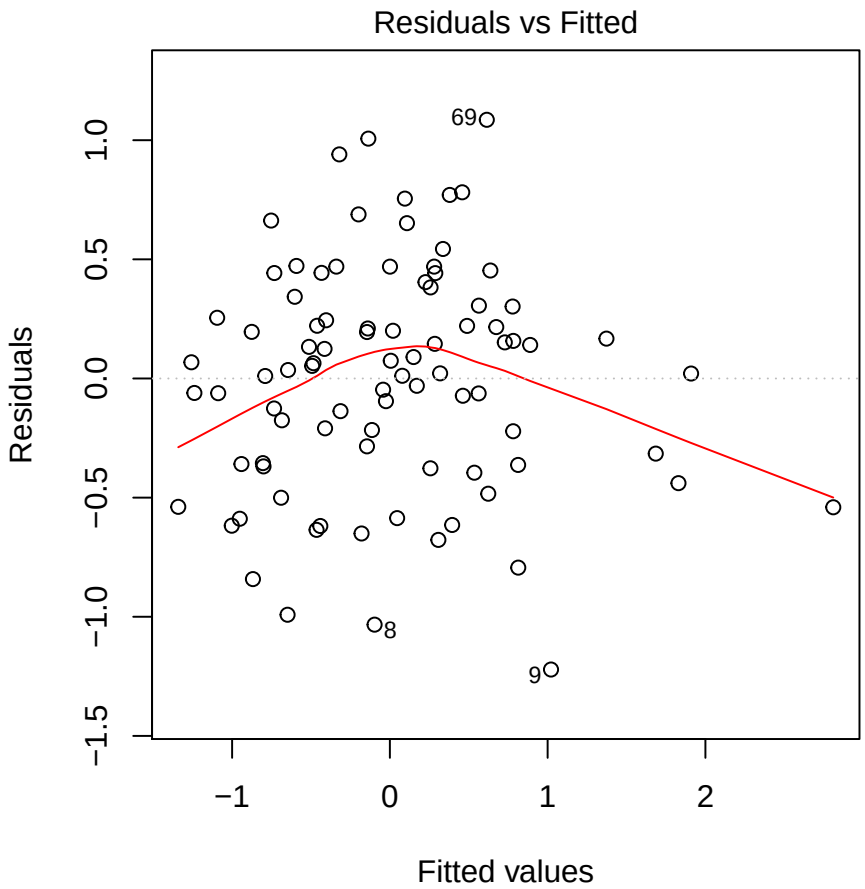
Would transforming the income data change things?

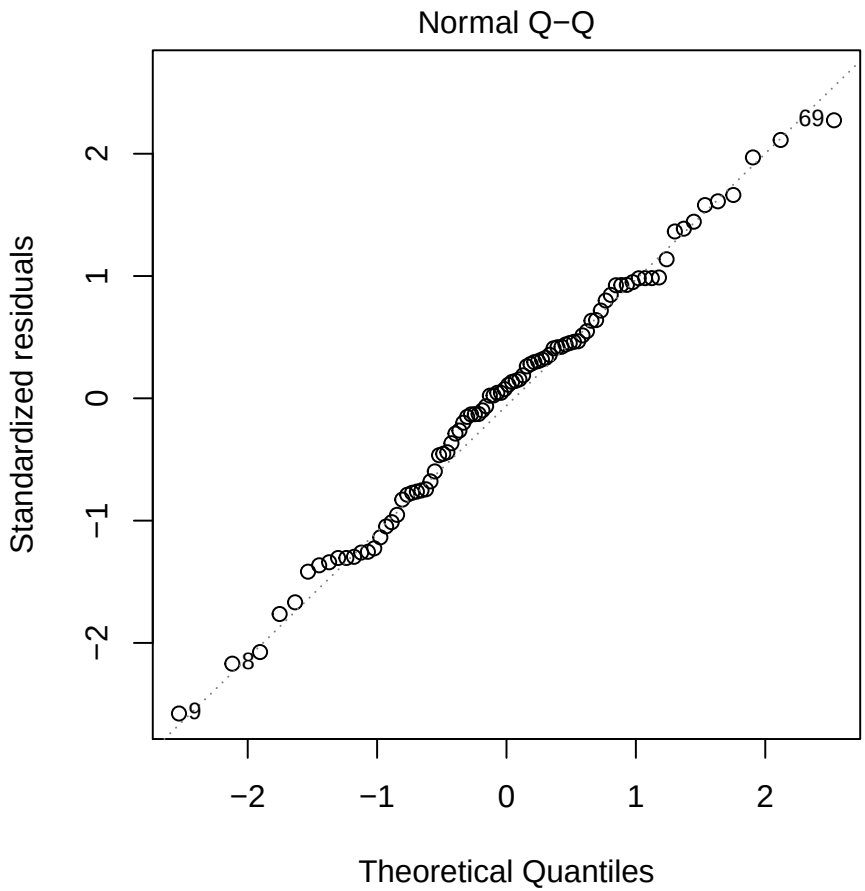
```
model_adj2 <- lm(outcomes ~ log(income) + clin_care,
                 data = ohio18)
anova(model_adj2)
```

```
Analysis of Variance Table

Response: outcomes
             Df Sum Sq Mean Sq  F value  Pr(>F)
log(income)   1 48.290  48.290 204.1115 < 2e-16 ***
clin_care     2  1.384   0.692   2.9245 0.05918 .
Residuals    84 19.873   0.237
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Impact of Transforming `income` data

```
par(mfrow = c(1,2))
plot(model_adj2, which = 1:2)
```





A New Model

Can we predict these health `outcomes` with a combination of `income` data and the county's `density` (defined as either Urban or Rural)?

```
model3 <- lm(outcomes ~ incK + density, data = ohio18)

tidy(model3) %>% knitr::kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-3.796	0.300	-12.637	0.000
incK	0.073	0.006	12.820	0.000
densityUrban	-0.382	0.159	-2.398	0.019

Our model3 summary, edited a little

```
summary(model3)
```

```
Call: lm(formula = outcomes ~ incK + density, data = ohio18)

Residuals:  Min      1Q  Median      3Q     Max
         -1.164 -0.364   0.074   0.358   1.008

Coefficients:    Estimate Std. Error  t value Pr(>|t|)
  (Intercept)      -3.796      0.300    -12.6   <2e-16 ***
  incK              0.073      0.006     12.8   <2e-16 ***
  densityUrban     -0.382      0.159     -2.4   0.0187 *

Residual standard error: 0.5269 on 85 degrees of freedom
Multiple R-squared:  0.6606,    Adjusted R-squared: 0.6527
F-statistic: 82.74 on 2 and 85 DF,  p-value: < 2.2e-16
```

ANOVA of our model3

```
anova(model3) %>% tidy() %>% knitr::kable(digits = 3)
```

term	df	sumsq	meansq	statistic	p.value
incK	1	44.349	44.349	159.725	0.000
density	1	1.597	1.597	5.752	0.019
Residuals	85	23.601	0.278	NA	NA

- The total Sum of Squares is 44.349 + 1.597 + 23.601 = 69.547.
  - Together, `incK` and `density` account for 44.349 + 1.597 = 45.946.
  - That is 66.06%, the same as the Multiple R^2 for the model.
- The residual mean square here (0.278) is the square of the residual standard error (0.5269) from the previous slide, and the degrees of freedom attributed to residuals there was also 85.
- The ANOVA F test on the previous screen (F = 82.74 on 2 and 85 df) combines the impact of both predictors.

Checking Regression Assumptions

The assumptions behind a linear regression model are, in order of importance:

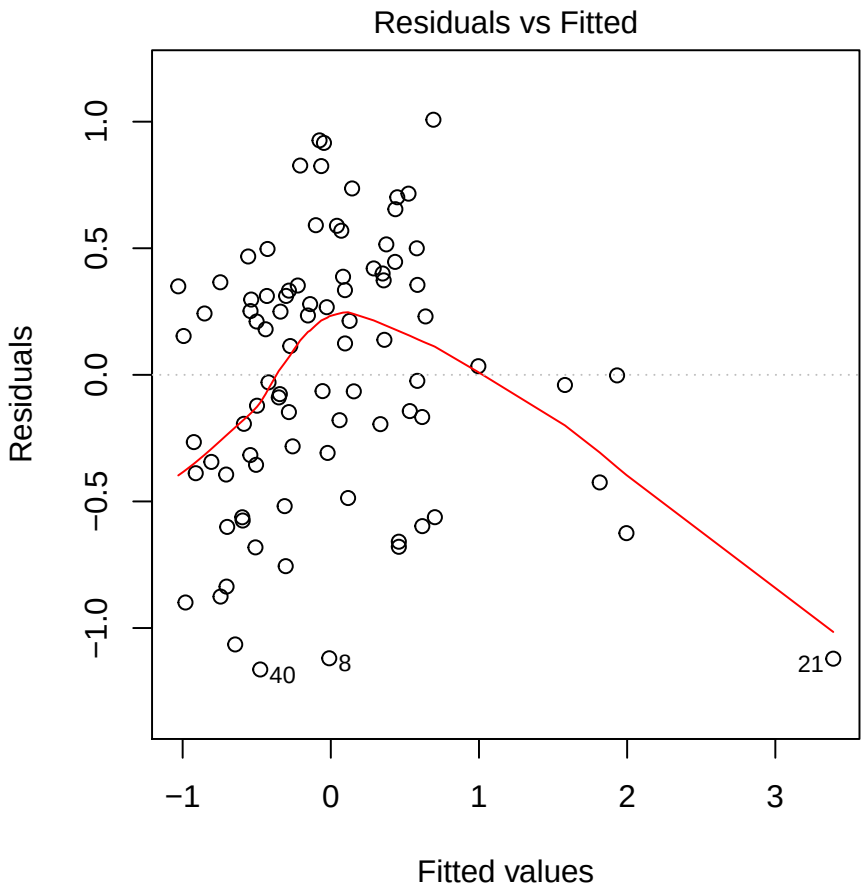
1. Linearity
2. Homoscedasticity (constant variance)
3. Independence
4. Normality

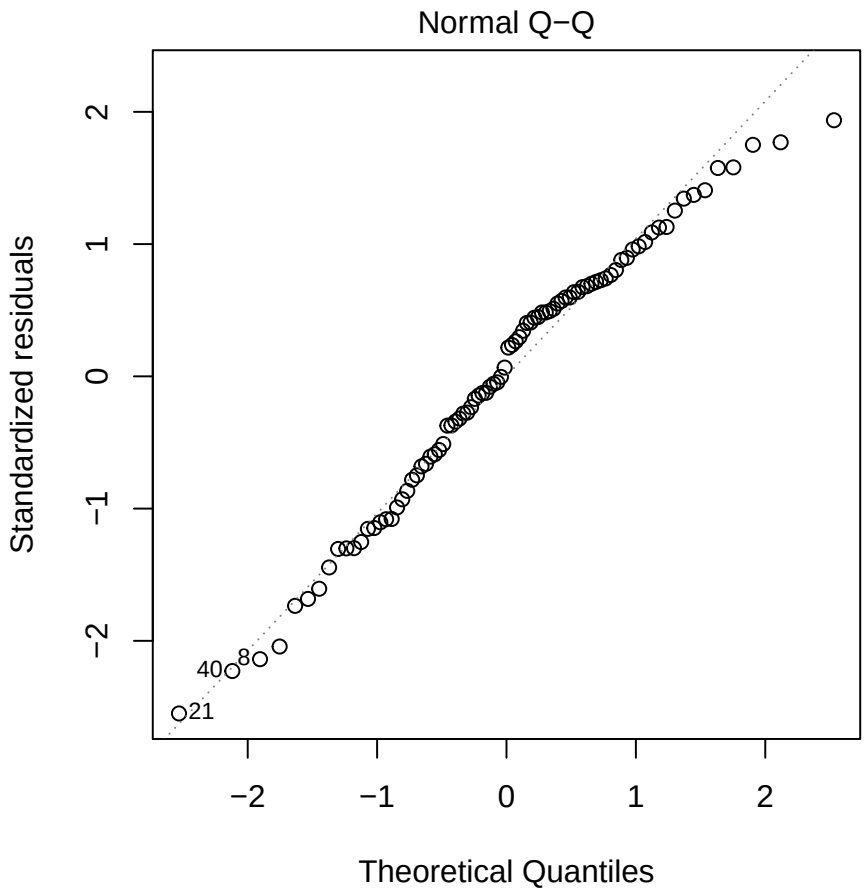
We build residual plots to check assumptions 1, 2, and 4. If the data are ordered in time or space, we will also think a bit about the independence assumption.

There are many ways to build residual plots using `ggplot2` but for now, we'll stick to base R and show you a very simple way to generate five plots of potential interest.

Residuals vs. Fitted, Normal Q-Q Plot

```
par(mfrow = c(1,2)); plot(model3, which = 1:2)
```





What county is County 21?

```
ohio18 %>% slice(21) %>% select(-FIPS, -state, -income) %>%
  knitr::kable()
```

county	outcomes	behavior	clin_care	density	incK
Delaware	2.27	Best	Strong	Urban	102.99

```
ohio18 %>% select(outcomes, income, density) %>% summary
```

```
    outcomes                income          density
 Min.   :-1.8800000   Min.   : 38131   Rural:74
 1st Qu.:-0.6125000   1st Qu.: 45230   Urban:14
 Median : 0.0150000   Median : 51157
 Mean   :-0.0001136   Mean   : 52474
 3rd Qu.: 0.6575000   3rd Qu.: 56821
 Max.   : 2.2700000   Max.   :102990
```

Where is Delaware County?

A map of Ohio, with Delaware County highlighted

Build it.

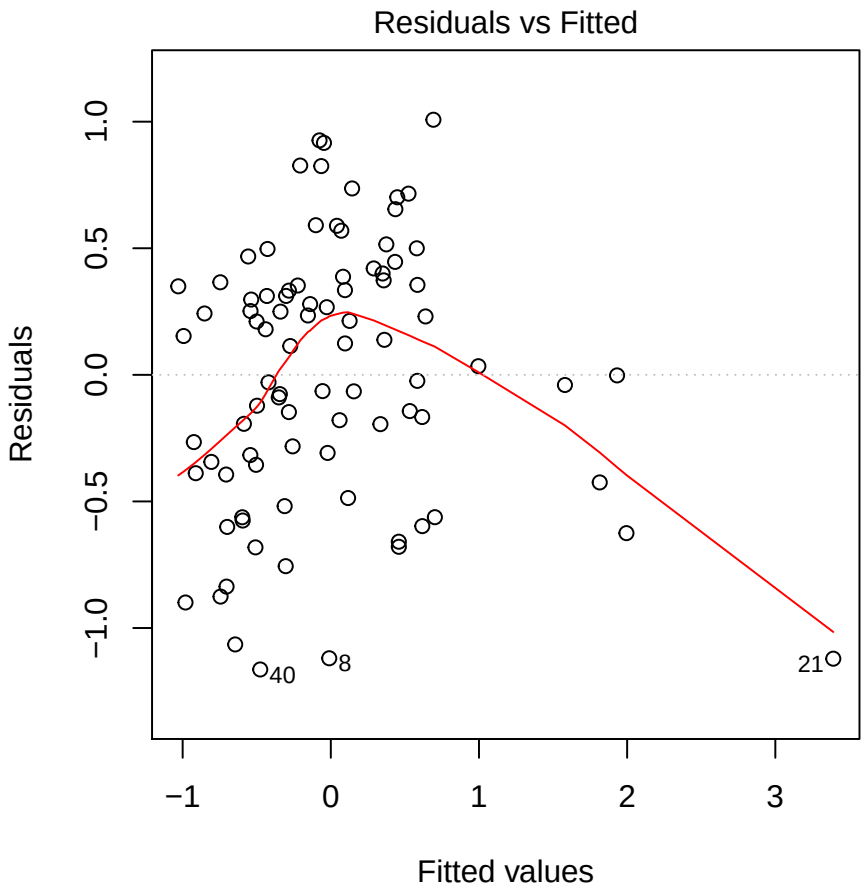
Augmenting the data with our model's results

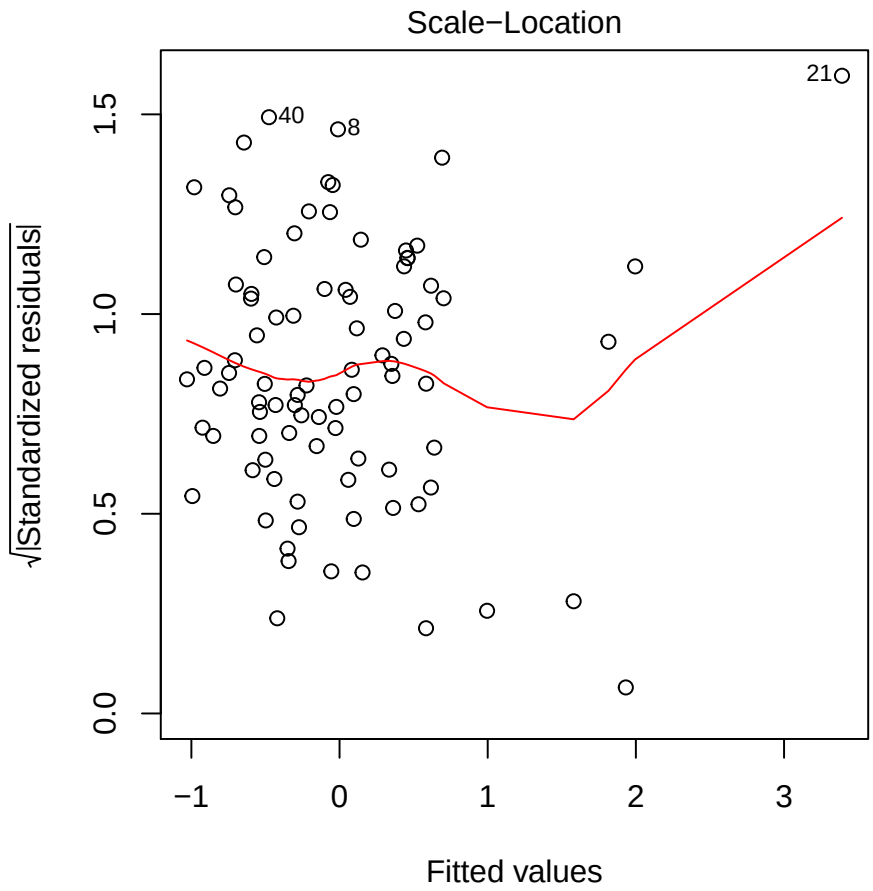
```
mod3_aug <- augment(model3) %>%
  bind_cols(ohio18 %>% select(county)) %>%
  select(county, everything())

slice(mod3_aug, c(18, 21)) %>% print.data.frame(digits=3)
```

```
    county outcomes  incK density .fitted .se.fit
1 Cuyahoga    -0.38  46.7   Urban  -0.746   0.157
2 Delaware     2.27 103.0   Urban   3.391   0.290
  .resid   .hat .sigma .cooksd .std.resid
1  0.366 0.0885  0.528  0.0171       0.727
2 -1.121 0.3035  0.509  0.9438      -2.549
```

Residuals vs. Fitted and Scale-Location Plot

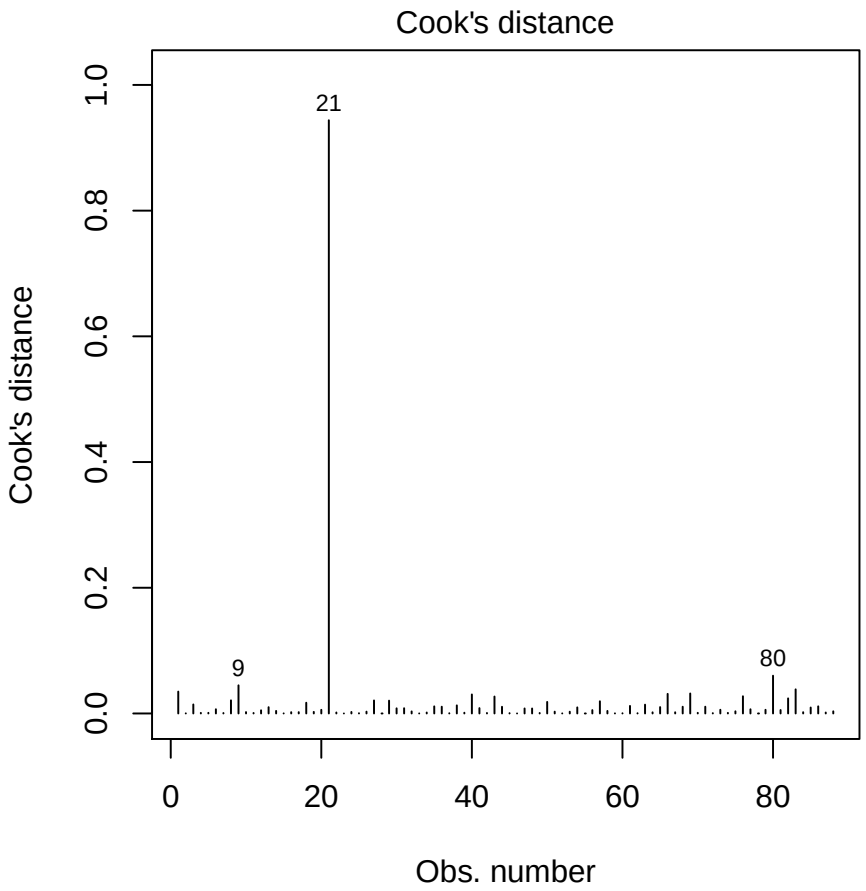
```
par(mfrow = c(1,2)); plot(model3, which = c(1,3))
```

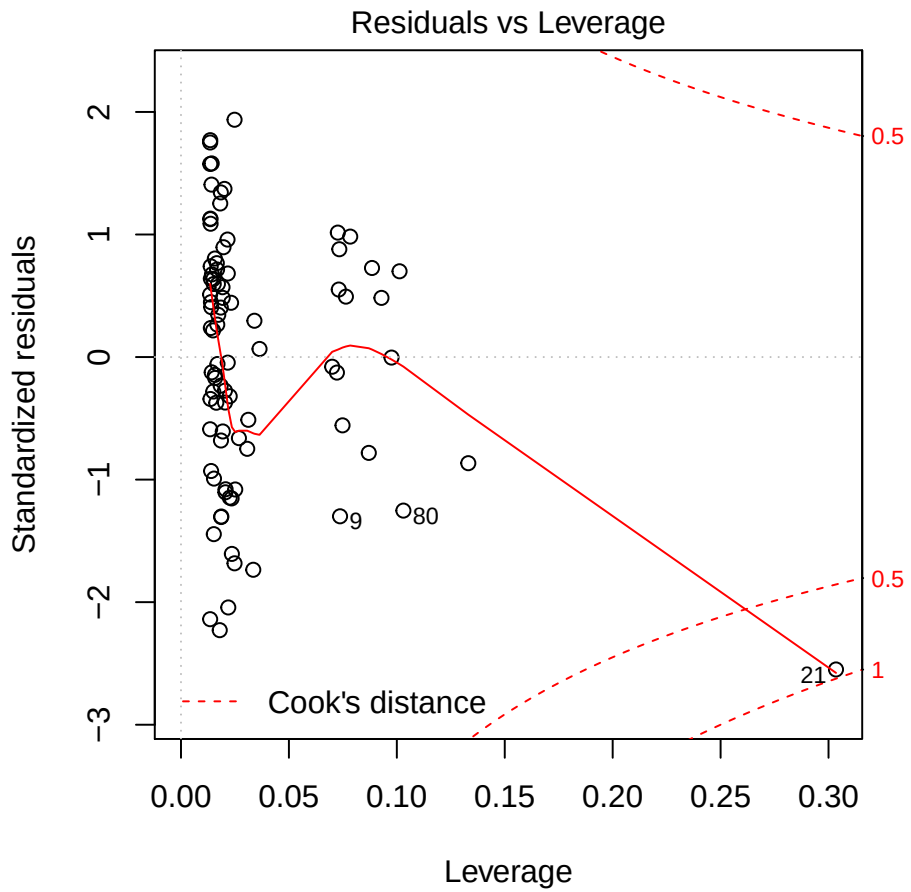




Cook's Distance and Influence Plot

```
par(mfrow = c(1,2)); plot(model3, which = 4:5)
```





Running the Model without Delaware County

```
model4 <- ohio18 %>% filter(county != "Delaware") %$%
    lm(outcomes ~ incK + density)

tidy(model4) %>% knitr::kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-4.228	0.334	-12.678	0.000
incK	0.082	0.006	12.804	0.000
densityUrban	-0.330	0.155	-2.123	0.037

Model 4 summary (no Delaware County)

```
summary(model4)
```

```
Call: lm(formula = outcomes ~ incK + density)

Residuals:    Min     1Q  Median      3Q     Max
           -1.121 -0.305  0.094   0.358   0.932

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)     -4.228      0.334 -12.678   <2e-16 ***
incK             0.082      0.006  12.804   <2e-16 ***
densityUrban    -0.330      0.155  -2.123   0.0367 *

Residual standard error: 0.5094 on 84 degrees of freedom
Multiple R-squared:  0.6612,    Adjusted R-squared:  0.6531
F-statistic: 81.97 on 2 and 84 DF,  p-value: < 2.2e-16
```

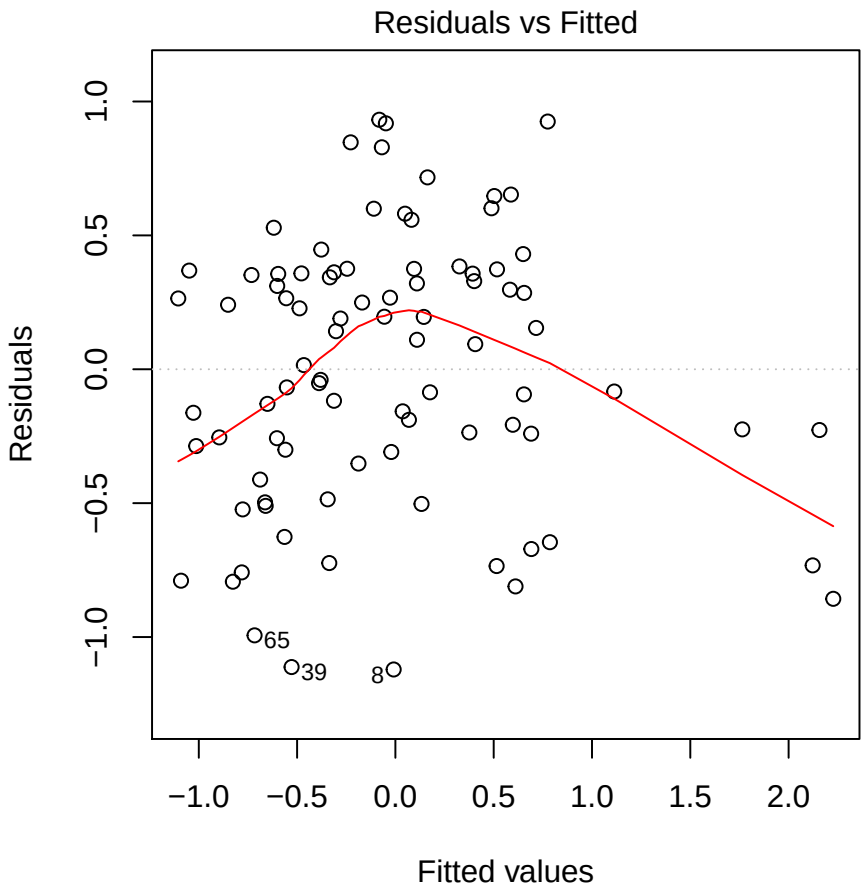
Model 4 ANOVA (no Delaware County)

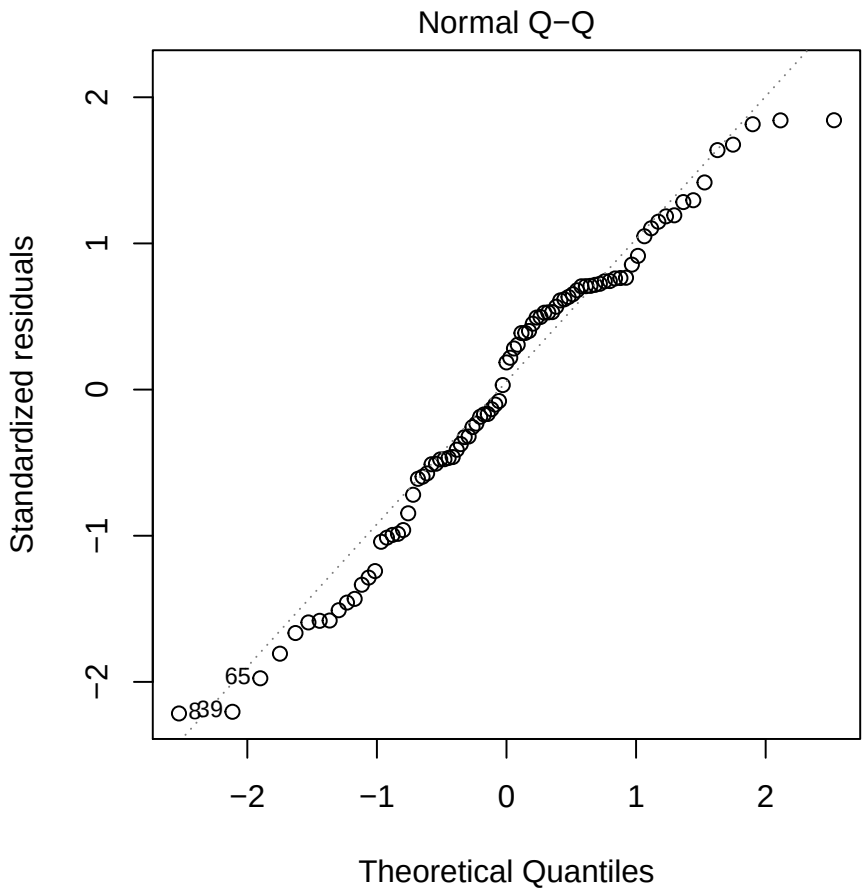
```
anova(model4) %>% tidy() %>% knitr::kable(digits = 3)
```

term	df	sumsq	meansq	statistic	p.value
incK	1	41.368	41.368	159.425	0.000
density	1	1.170	1.170	4.508	0.037
Residuals	84	21.797	0.259	NA	NA

Residuals vs. Fitted, Normal Q-Q Plot

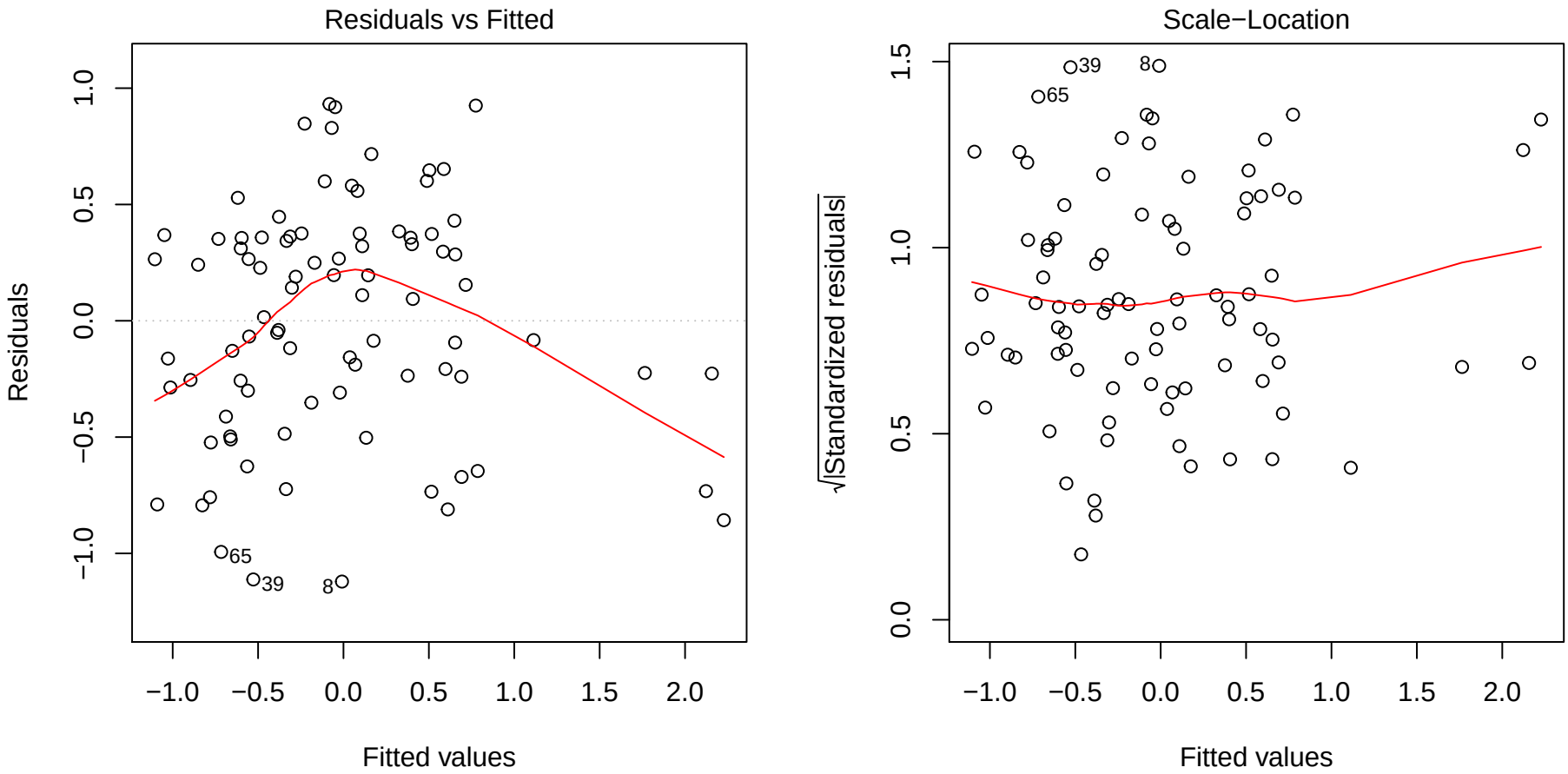
```
par(mfrow = c(1,2)); plot(model4, which = 1:2)
```





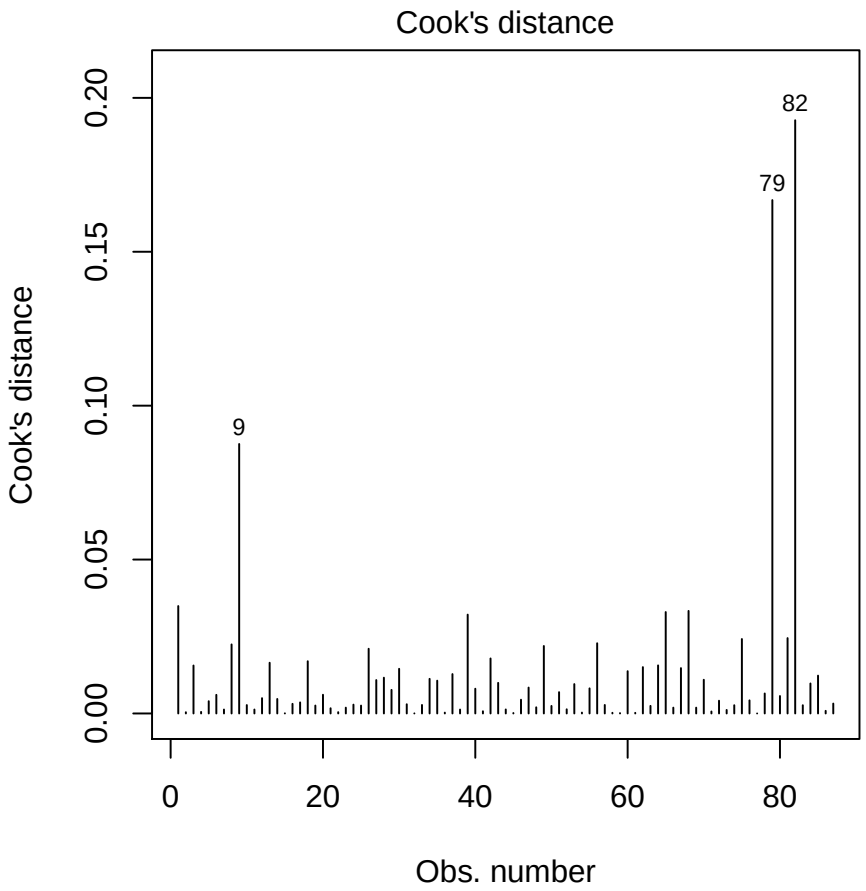
Residuals vs. Fitted and Scale-Location Plot

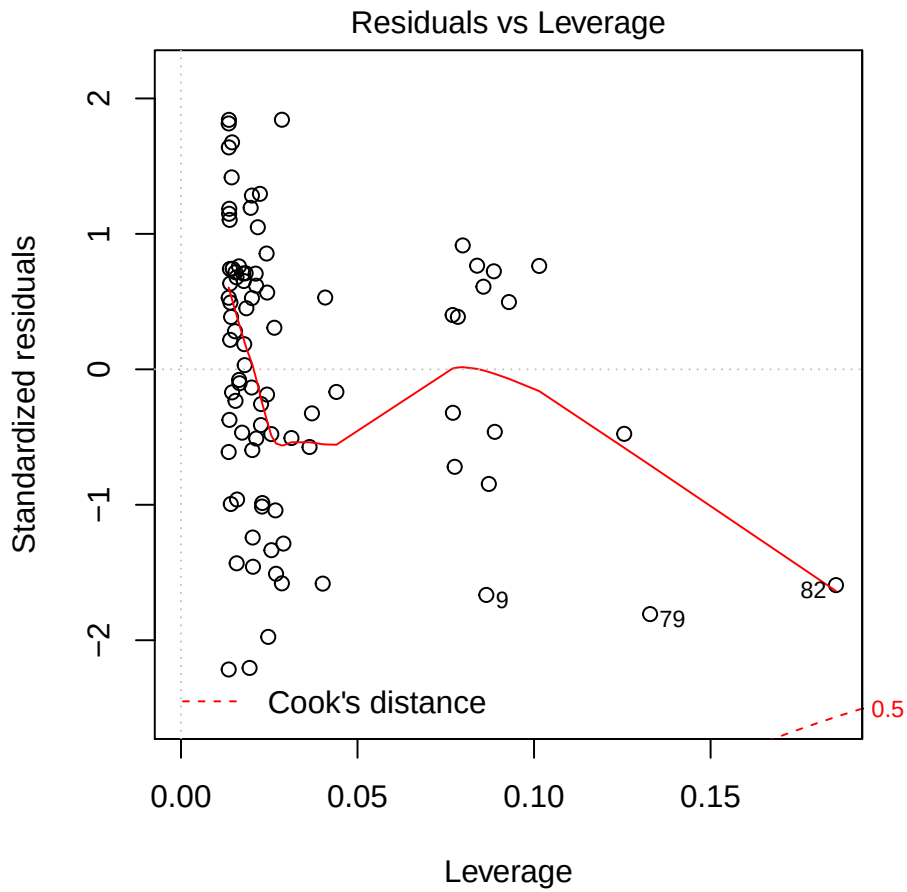
```
par(mfrow = c(1,2)); plot(model4, which = c(1,3))
```



Cook's Distance and Influence Plot

```
par(mfrow = c(1,2)); plot(model4, which = 4:5)
```





Study 2 Demonstration Project

- See github site for this.

Our Next Steps

- Building and Using a Scatterplot Matrix (review)
- Comparing Models, by splitting our data into training (model development) and test samples
  - Assessing training sample performance with adjusted R^2, AIC and BIC
  - Assessing test sample prediction errors with MAPE and MSPE
- Making good decisions when building regression models

Have a nice break.

Get your project moving along. Thanks.