Lab 7 - Your Turn Solutions

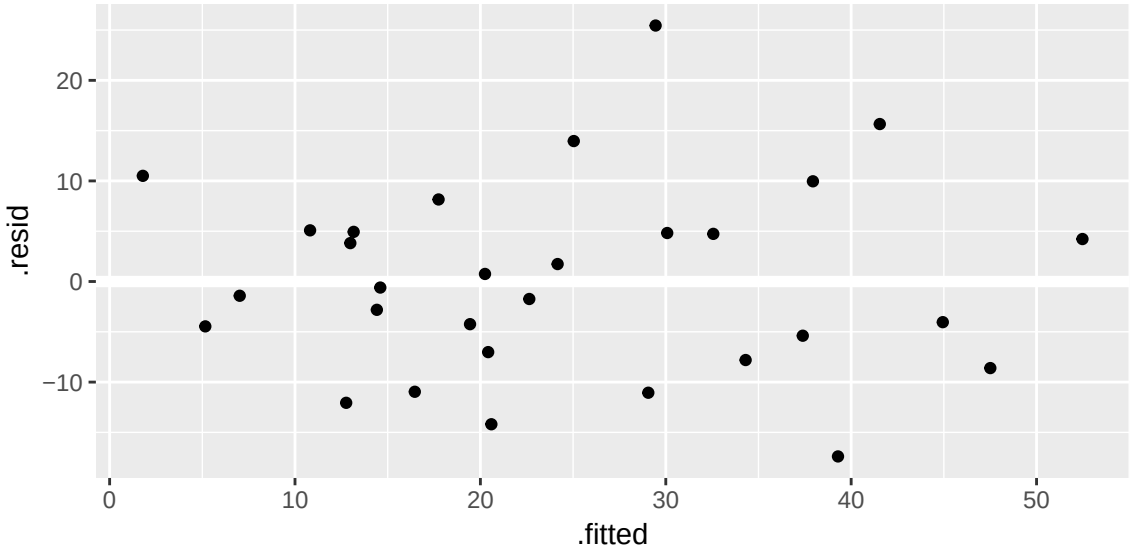
Matt Higham

February 12, 2016

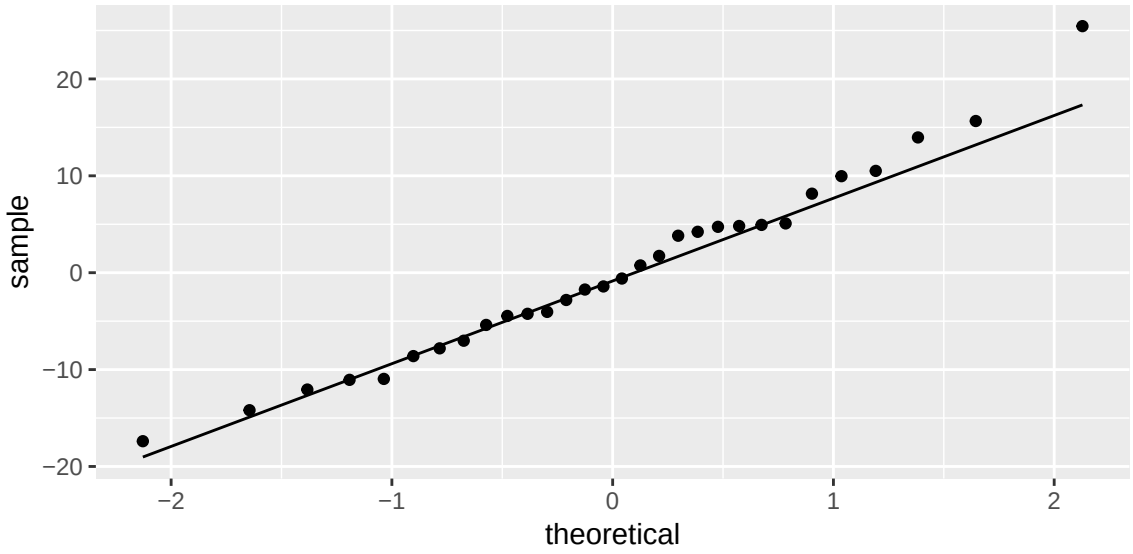
Some edits from Charlotte 2019-02-21

Cheddar Data Set

```
library(tidyverse)
library(broom)

data(cheddar, package = "faraway")
fit_cheddar <- lm(taste ~ Acetic + H2S + Lactic, data = cheddar)
fit_cheddar_df <- augment(fit_cheddar, data = cheddar)
ggplot(fit_cheddar_df, aes(.fitted, .resid)) +
  geom_hline(yintercept = 0, color = "white", size = 2) +
  geom_point()
```



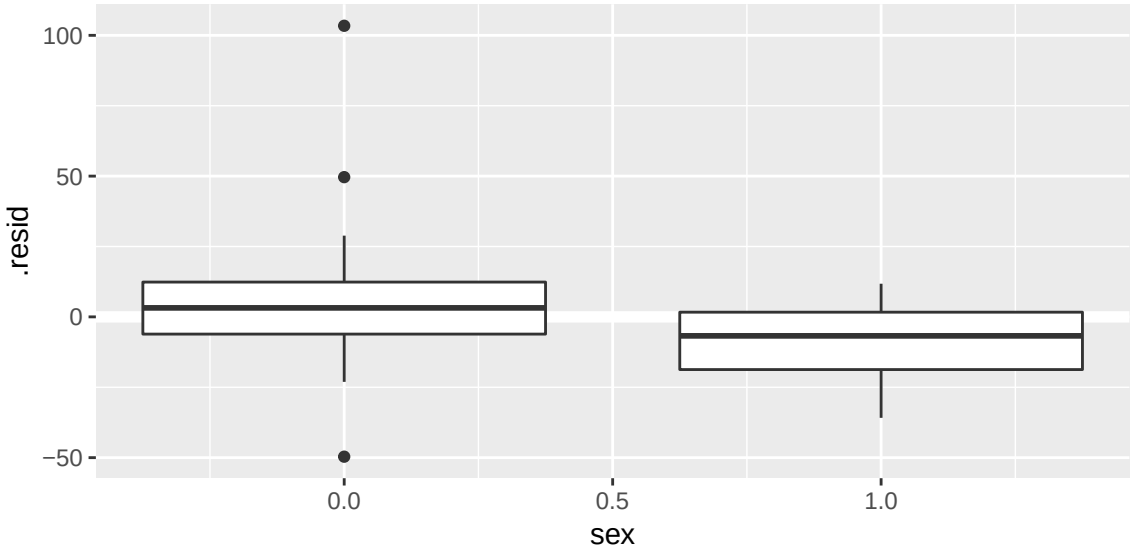

```
ggplot(fit_cheddar_df, aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line()
```



I do not see any strong violations of assumptions in any of these residual plots. There is perhaps a slight violation in the constant variance assumption. Taking a square root transformation may help the non-constant variance, but I do not think the violation is strong enough to warrant a transformation.

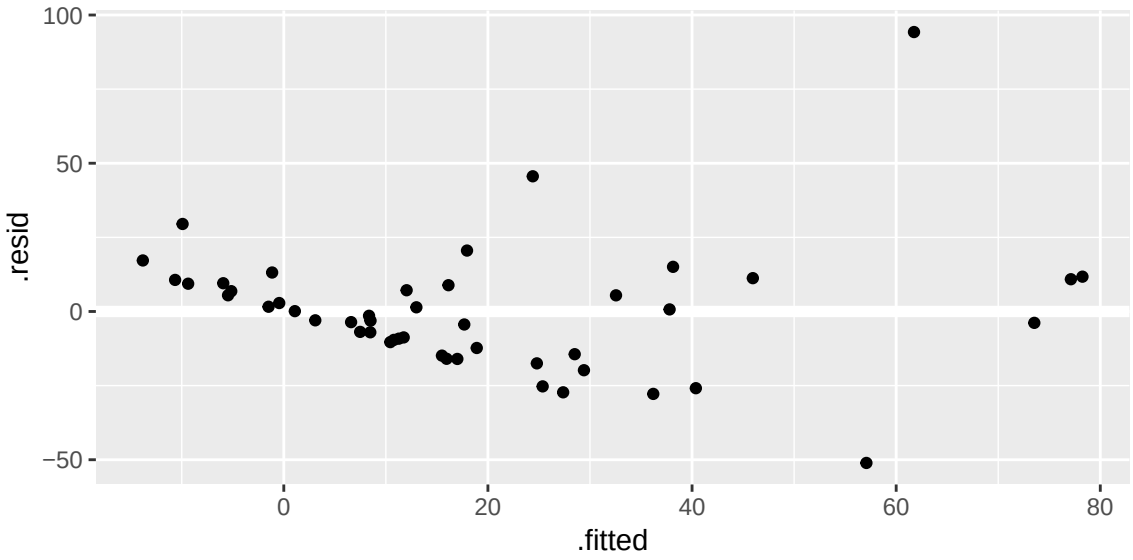
Gambling Data Set

```
data(teengamb, package = "faraway")
fit_teen <- lm(gamble ~ income + status + verbal, data = teengamb)
fit_teen_df <- augment(fit_teen, data = teengamb)

ggplot(fit_teen_df, aes(sex, .resid)) +
  geom_hline(yintercept = 0, color = "white", size = 2) +
  geom_boxplot(aes(group = sex))
```

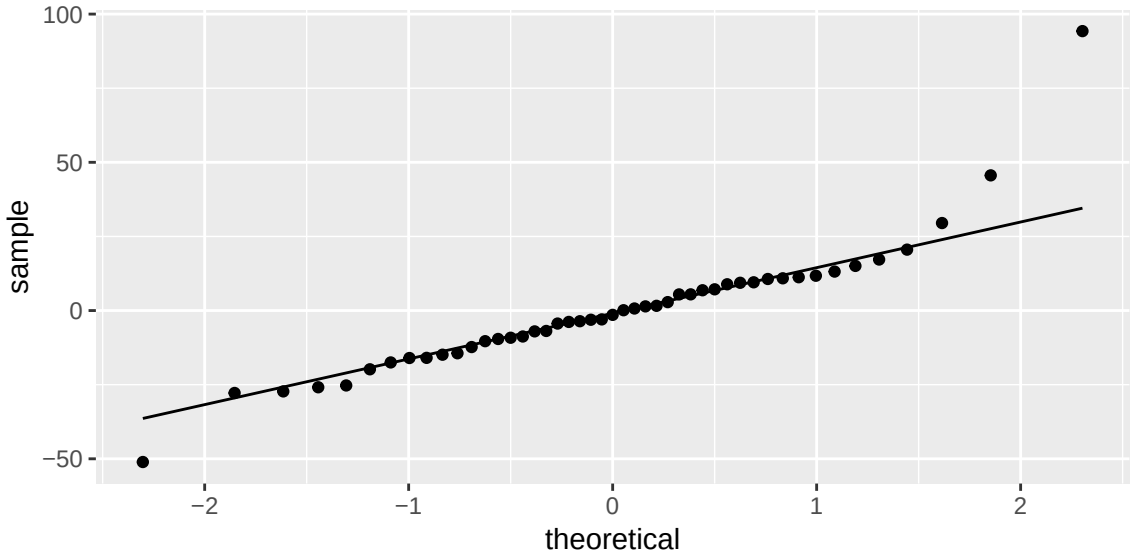


From the above graph of residuals versus sex, which was not included in the original model, we see that sex might be an important predictor of gambling since the residuals around the sex = 0 (the male sex) are generally greater than 0 and the residuals around the female sex are generally less than 0. Therefore, there is evidence that we should add sex back to the model.

```
fit_teen2 <- lm(gamble ~ income + status + verbal + sex, data = teengamb)
fit_teen_df2 <- augment(fit_teen2, data = teengamb)

ggplot(fit_teen_df2, aes(.fitted, .resid)) +
  geom_hline(yintercept = 0, color = "white", size = 2) +
  geom_point()
```

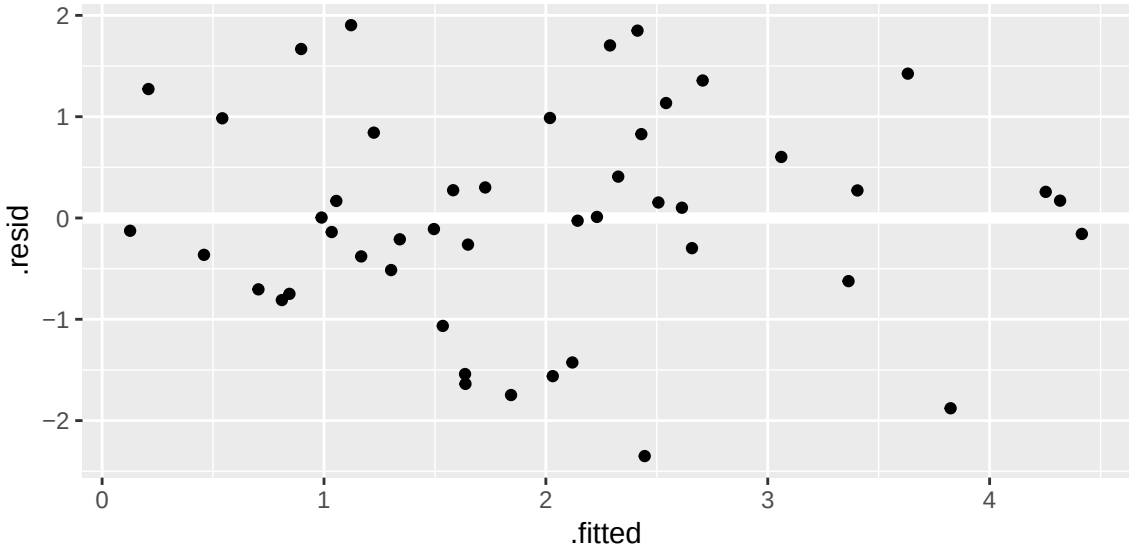



```
ggplot(fit_teen_df2, aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line()
```

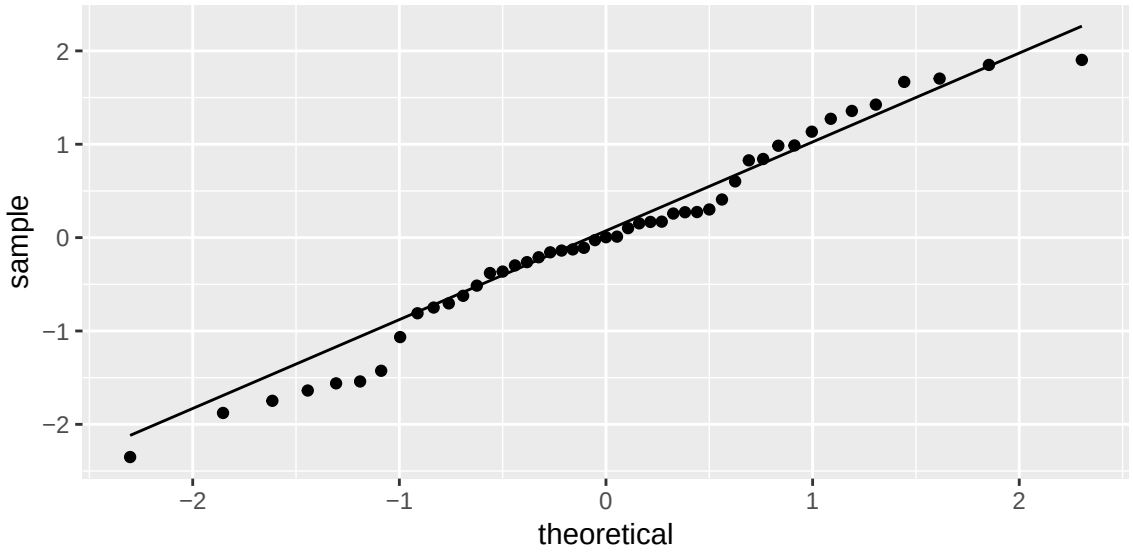


Even with sex in the model, we see evidence of non-constant variance. We can transform the response to fix some of these violations of the assumptions of our regression model.

```
fit_teen3 <- lm(log(gamble + 1) ~ income + status + verbal + sex, data = teengamb)
fit_teen_df3 <- augment(fit_teen3, data = teengamb)

ggplot(fit_teen_df3, aes(.fitted, .resid)) +
  geom_hline(yintercept = 0, color = "white", size = 2) +
  geom_point()
```




```
ggplot(fit_teen_df3, aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line()
```

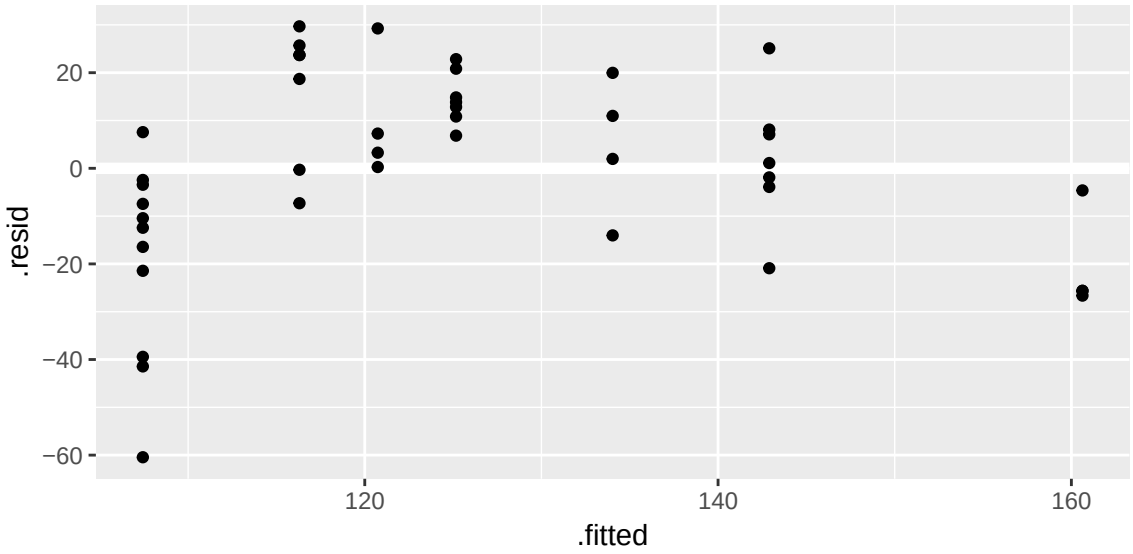


```
# ggplot(fit_teen_df3, aes(status, .resid)) +
#   geom_hline(yintercept = 0, color = "white", size = 2) +
#   geom_point()
#
# ggplot(fit_teen_df3, aes(income, .resid)) +
#   geom_hline(yintercept = 0, color = "white", size = 2) +
#   geom_point()
#
# ggplot(fit_teen_df3, aes(verbal, .resid)) +
#   geom_hline(yintercept = 0, color = "white", size = 2) +
#   geom_point()
```

Wowsers! These diagnostics look a lot better!

Corn Data Set

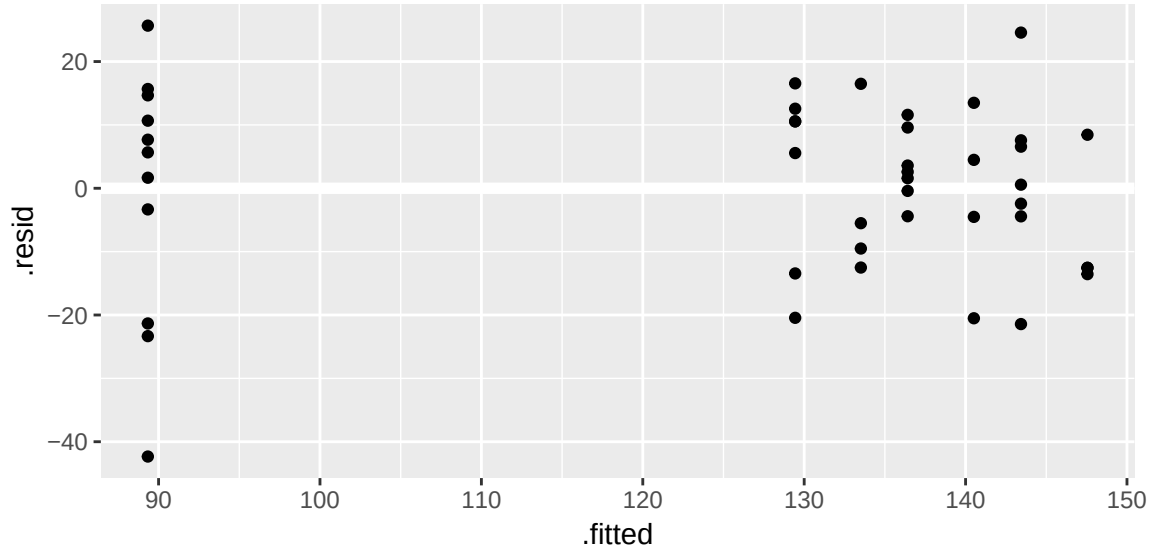
```
data(cornnit, package = "faraway")
fit_corn <- lm(yield ~ nitrogen, data = cornnit)
fit_corn_df <- augment(fit_corn, cornnit)

ggplot(fit_corn_df, aes(.fitted, .resid)) +
  geom_hline(yintercept = 0, color = "white", size = 2) +
  geom_point()
```



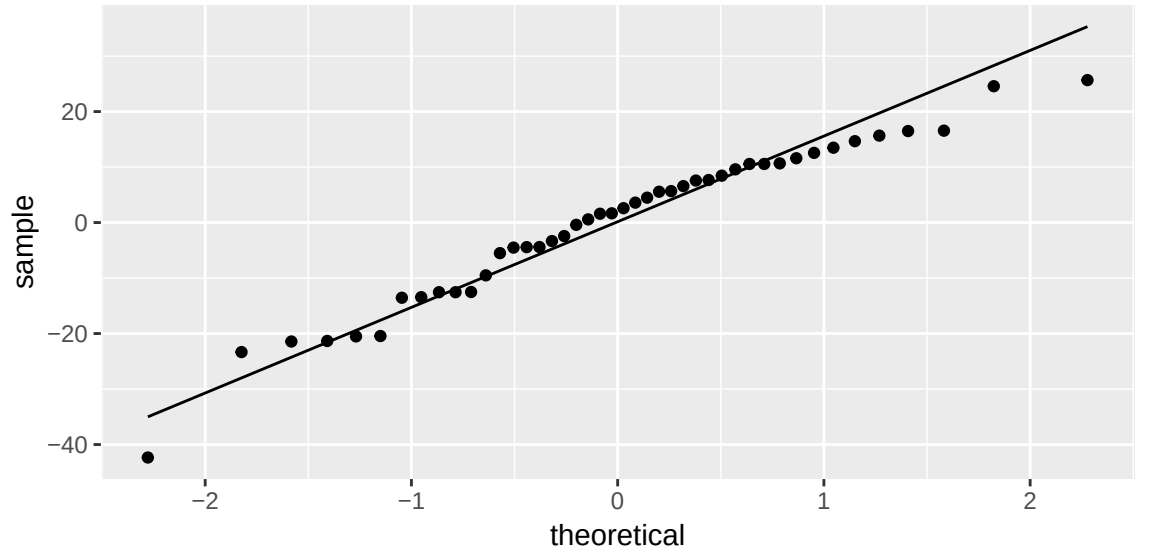
There is evidence of non-linearity in this residuals versus fitted plot! We can try many different types of transformations to fix this problem, or we can add a quadratic term to the model. Below is the residual versus fitted plot after log transforming the predictor, nitrogen.

```
fit_corn2 <- lm(yield ~ log(nitrogen + 1), data = cornnit)
fit_corn_df2 <- augment(fit_corn2, cornnit)

ggplot(fit_corn_df2, aes(.fitted, .resid)) +
```

```
  geom_hline(yintercept = 0, color = "white", size = 2) +
  geom_point()
```




```
ggplot(fit_corn_df2, aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line()
```



That seems to fix the linearity problem. Also, there is now little evidence of non-normality or non-constant variance.