Case study: Alcohol metabolism

Packages for this section

Metabolizing alcohol

- It is believed that women have a lower tolerance for alcohol compared to men, and they develop alcohol-related liver disease more easily than men (on average).
- Italian researchers developed a theory about why this is. To test their theory, they collected some data (variables on next page).

The variables

- `Subject`: subject number (ignored by us)
- `Metabol`: first-pass metabolism of alcohol in stomach (millimoles per litre-hour), response
- `Gastric`: gastric alcohol dehydrogenase activity in stomach (micromoles per minute per gram of tissue)
- `Sex`: whether subject was `Female` or `Male` (categorical)
- `Alcohol`: whether subject was an alcoholic or not (assessed by how much alcohol they drank and how frequently). Values `Alcoholic`, `Non-alcoholic` (categorical).

Comments

- one response variable, `Metabol`
- *several* explanatory variables
  - hence, *multiple* regression
  - some of them are categorical.
- according to theory, more gastric activity should mean higher metabolism
- metabolism is expected to be higher for males than females even at same level of gastric activity
- may be an effect of alcoholism, but direction not predicted by theory.

The data

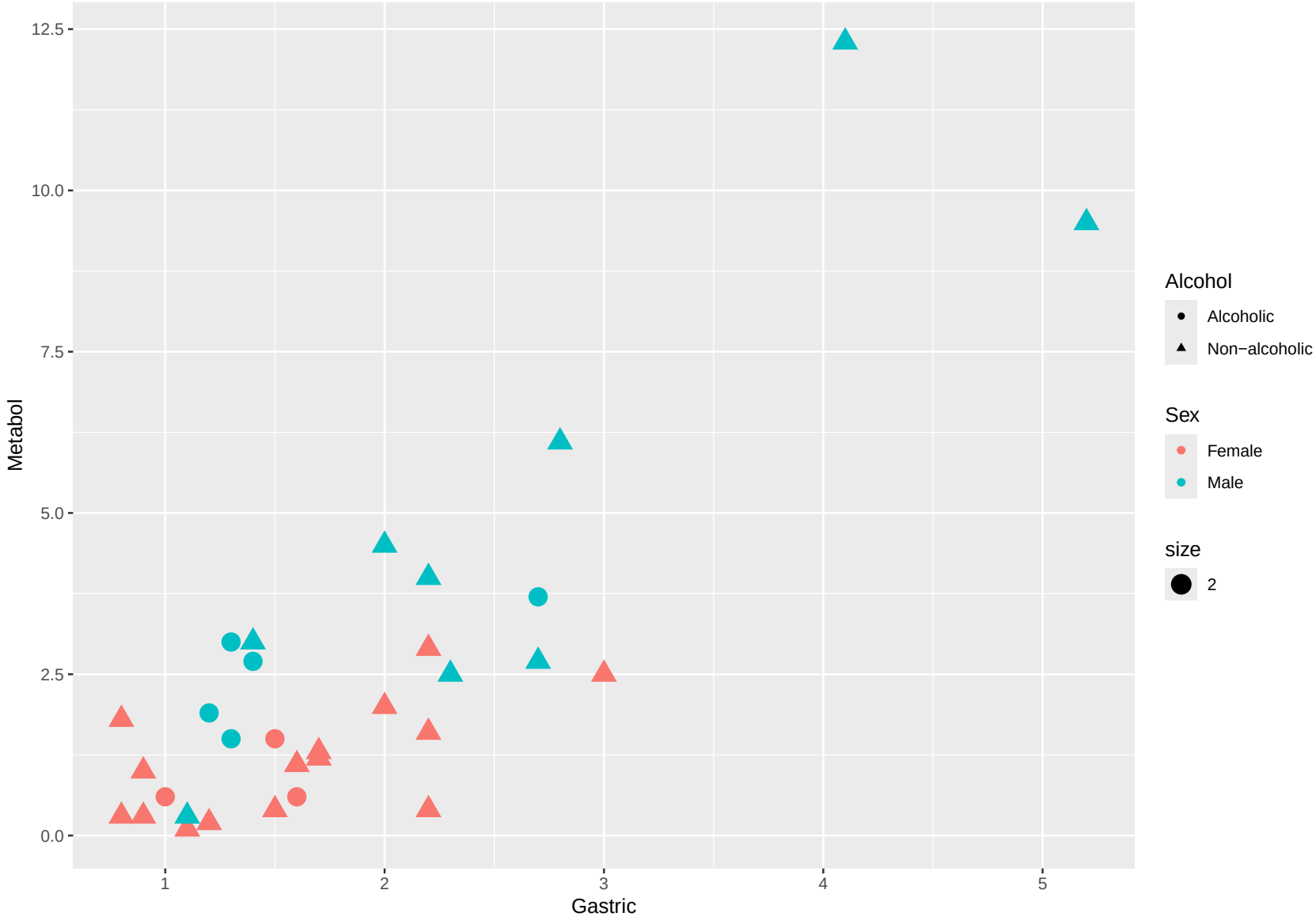
```
# A tibble: 32 x 5
   Subject Metabol Gastric Sex    Alcohol
     <dbl>   <dbl>   <dbl> <chr>  <chr>
 1       1     0.6     1   Female Alcoholic
 2       2     0.6     1.6 Female Alcoholic
 3       3     1.5     1.5 Female Alcoholic
 4       4     0.4     2.2 Female Non-alcoholic
 5       5     0.1     1.1 Female Non-alcoholic
 6       6     0.2     1.2 Female Non-alcoholic
 7       7     0.3     0.9 Female Non-alcoholic
 8       8     0.3     0.8 Female Non-alcoholic
 9       9     0.4     1.5 Female Non-alcoholic
10      10     1       0.9 Female Non-alcoholic
# i 22 more rows
```

A graph

- Have two quantitative variables (suggests scatterplot), but also two categorical ones (suggests colours, and ???)
- Can also use `shape` as well as colour to distinguish observations.
- Hence:

```
ggplot(alc, aes(x = Gastric, y = Metabol,
                colour = Sex, shape = Alcohol)) +
  geom_point()
```

The graph


12.5
10.0
7.5
5.0
2.5
0.0
Metabol
1
2
3
4
5
Gastric
Alcohol
Alcoholic
Non−alcoholic
Sex
Female
Male
size
2

Comments

- As gastric activity increases, metabolism also increases.
- Trend is apparently linear and of moderate strength.
- Two points unusually high on both variables, but seem to be on the trend.
- Male points (blue) mostly above female ones (red) for similar `Gastric`
- No apparent effect of `Alcohol`.

Fit regression

```
Call:
lm(formula = Metabol ~ Gastric + Sex + Alcohol, data = alc)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3119 -0.6339 -0.0927  0.6070  4.5629

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           -2.0159     0.6348  -3.175  0.00362 **
Gastric                1.9466     0.2884   6.749  2.5e-07 ***
SexMale                1.6535     0.5514   2.999  0.00564 **
AlcoholNon-alcoholic   0.1183     0.6008   0.197  0.84527
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.354 on 28 degrees of freedom
Multiple R-squared:  0.7657,    Adjusted R-squared:  0.7406
F-statistic: 30.51 on 3 and 28 DF,  p-value: 5.748e-09
```

Comments

- no effect of `Alcohol` (suggests removing)
- significant effects of both `Gastric` and `Sex`
- both slopes positive (as expected)
- but, before we go further, check residuals:
  - vs. fitted values
  - normal quantile plot
  - *also*, vs. each explanatory variable (if any trends, have form of relationship with that explanatory variable wrong).

Interpretation of Estimates

- There is one intercept, and *each* explanatory variable has a slope that expresses effect of that variable, *all else equal*.
- Slope for `Gastric` (quantitative) says that if `Gastric` increases by 1, and everything else same, `Metabol` predicted to increase by 1.95.
- Each categorical variable has a "baseline" category, the first one alphabetically: `Female`, `Alcoholic`.
- For categorical explanatory variables, the slope says how `Metabol` compares for the named category vs. the baseline:
  - for males, `Metabol` is 1.65 higher than for females, all else equal
  - for non-alcoholics, `Metabol` is 0.12 higher than for alcoholics, all else equal
- Intercept is the predicted value of `Metabol` when all the quantitative variables are 0 and all the categorical variables are at baseline (not usually very interesting).

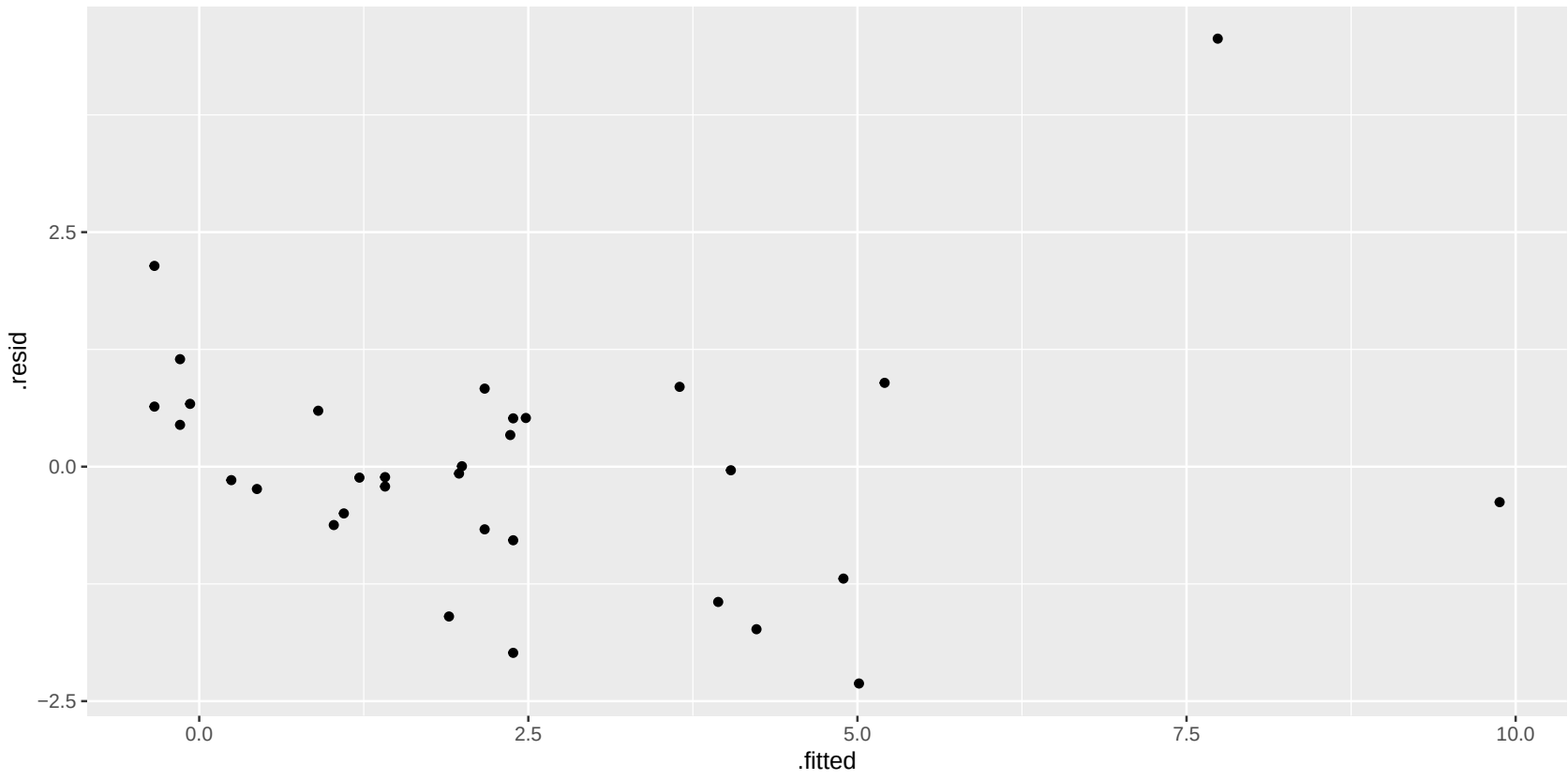
Create dataframe for plots

```
alc.1 %>% augment(alc) -> alc.1a
glimpse(alc.1a)
```

```
Rows: 32
Columns: 11
$ Subject    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,~
$ Metabol    <dbl> 0.6, 0.6, 1.5, 0.4, 0.1, 0.2, 0.3,~
$ Gastric    <dbl> 1.0, 1.6, 1.5, 2.2, 1.1, 1.2, 0.9,~
$ Sex        <chr> "Female", "Female", "Female", "Fem~
$ Alcohol    <chr> "Alcoholic", "Alcoholic", "Alcohol~
$ .fitted    <dbl> -0.06926514, 1.09870918, 0.9040467~
$ .resid     <dbl> 0.66926514, -0.49870918, 0.5959532~
$ .hat       <dbl> 0.17712194, 0.19504634, 0.18979002~
$ .sigma     <dbl> 1.371461, 1.374635, 1.372891, 1.32~
$ .cooksd    <dbl> 1.597846e-02, 1.021005e-02, 1.4003~
$ .std.resid <dbl> 0.544915310, -0.410544719, 0.48900~
```

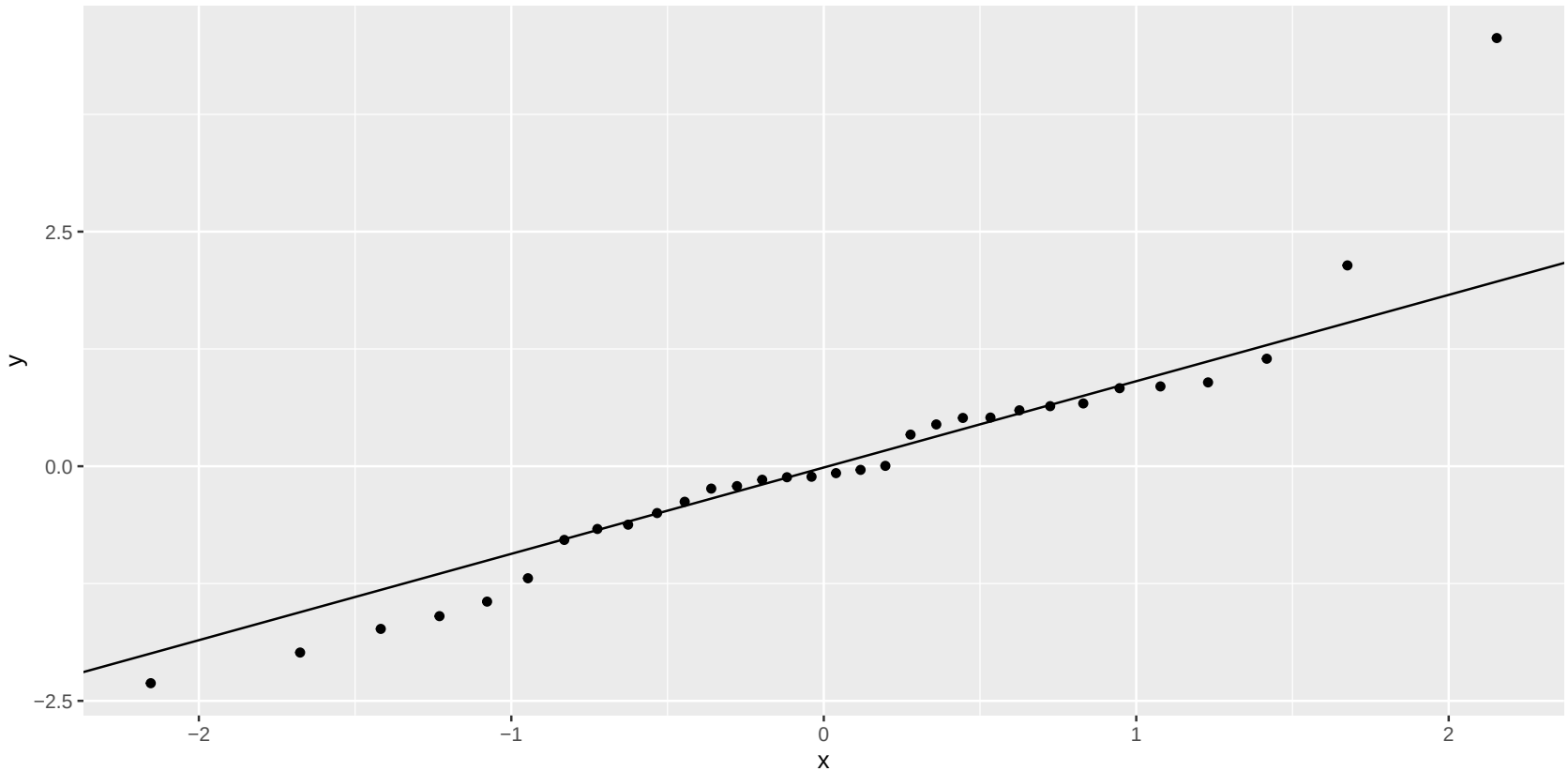
Residuals vs fitted

```
ggplot(alc.1a, aes(x = .fitted, y = .resid)) + geom_point()
```



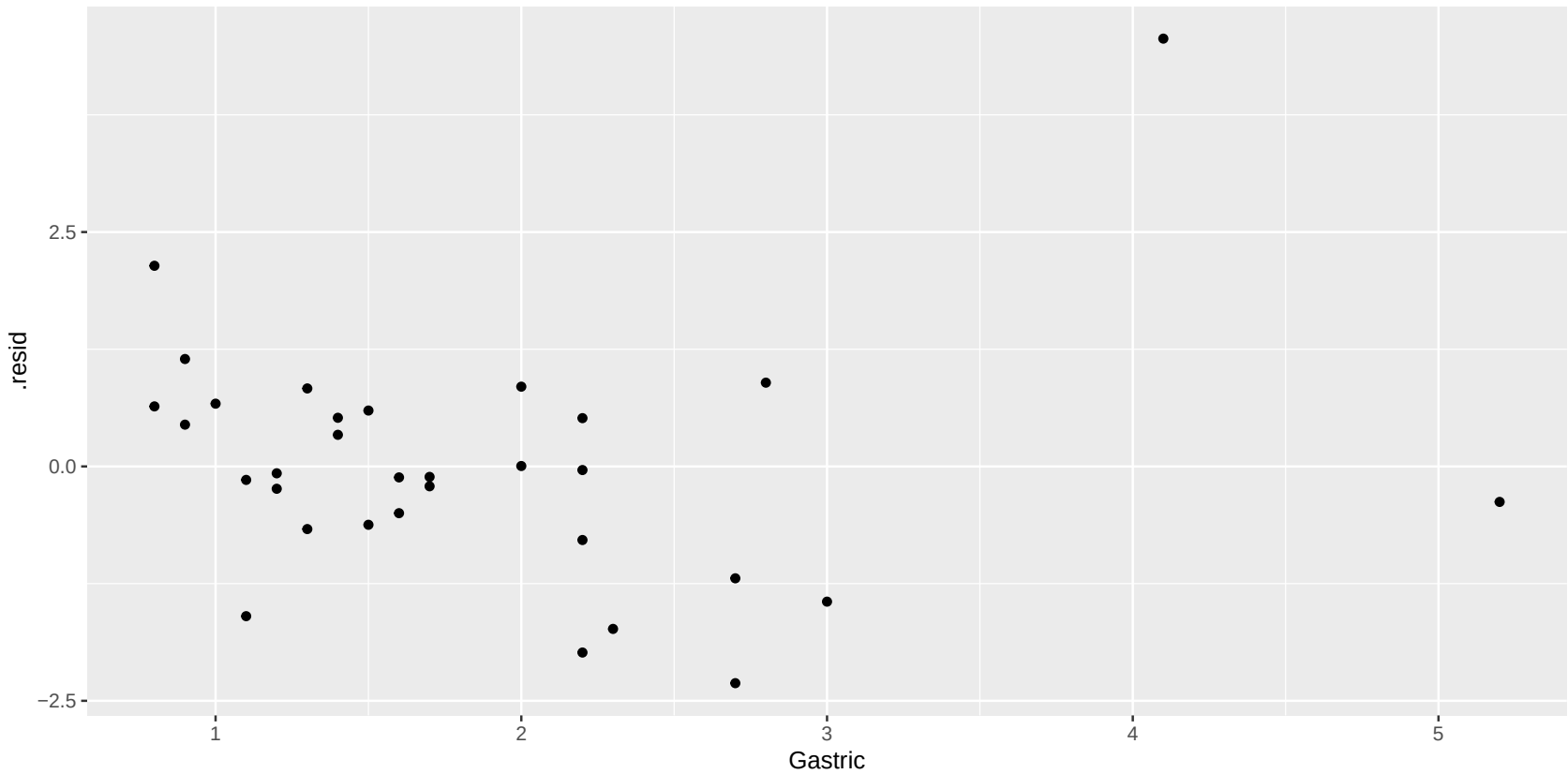
Normal quantile plot

```
ggplot(alc.1a, aes(sample = .resid)) +
  stat_qq() + stat_qq_line()
```



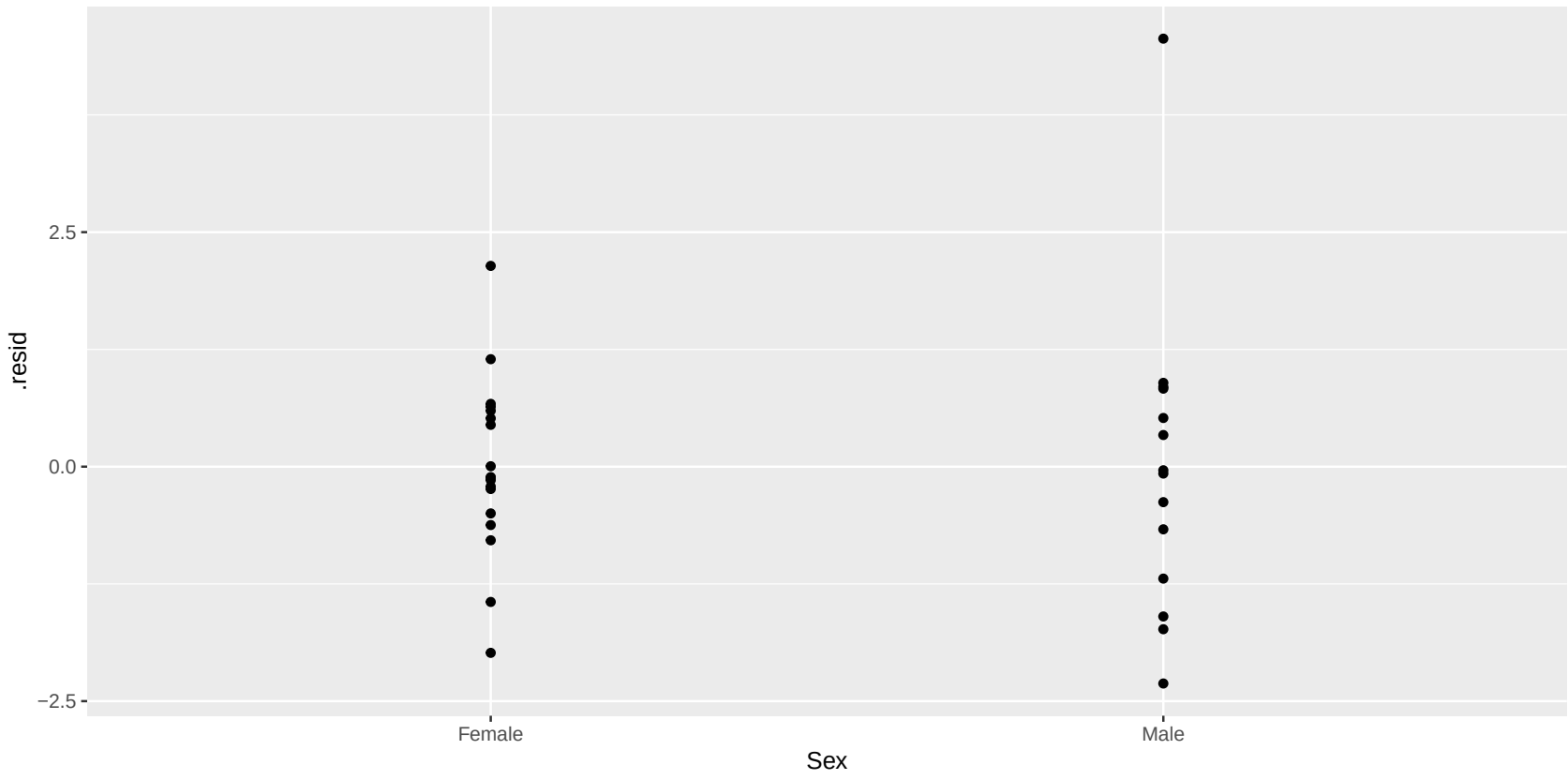
Residuals vs Gastric

```
ggplot(alc.1a, aes(x = Gastric, y = .resid)) + geom_point()
```



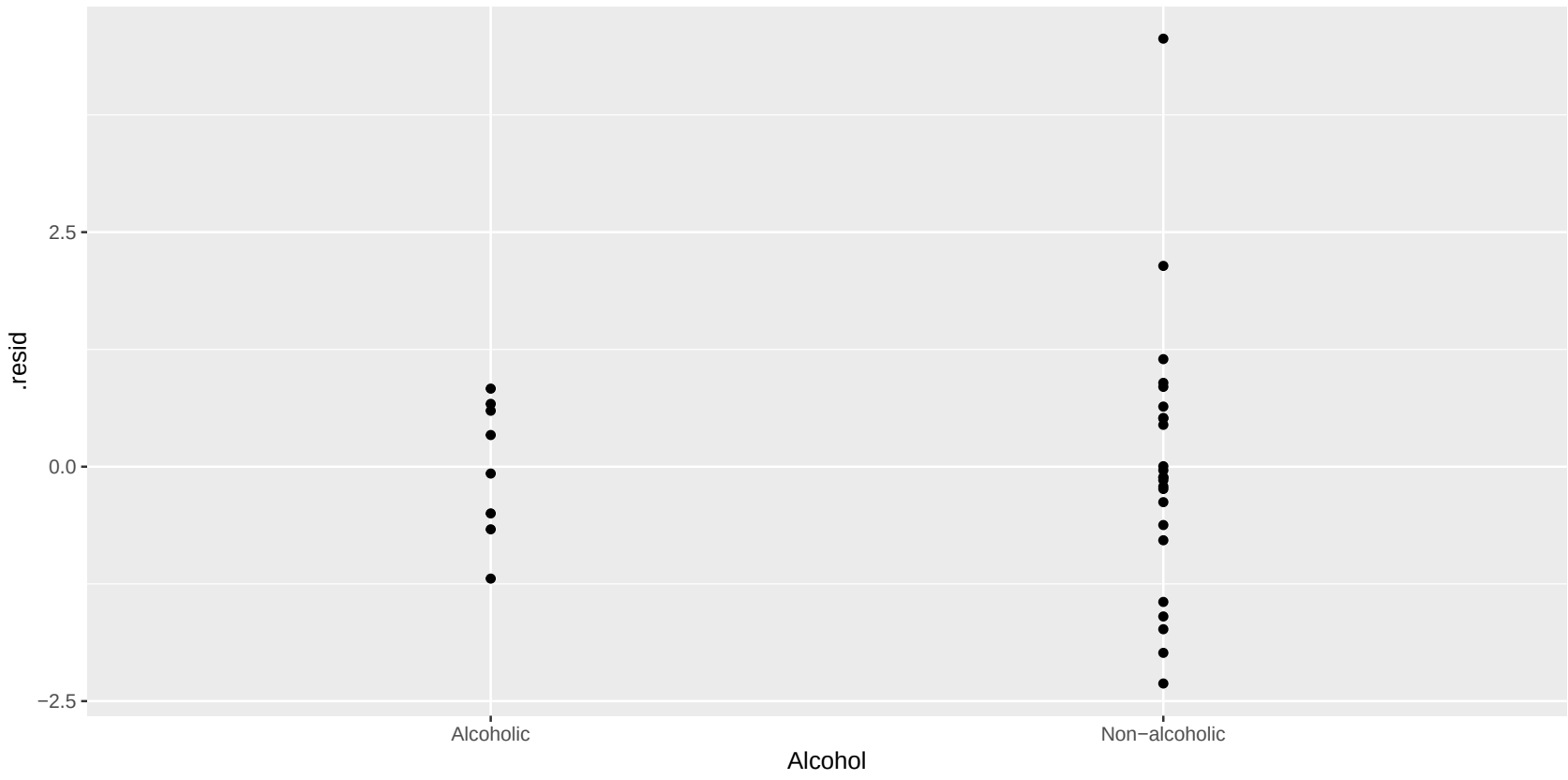
Residuals vs Sex

```
ggplot(alc.1a, aes(x = Sex, y = .resid)) + geom_point()
```



Residuals vs `Alcohol`

```
ggplot(alc.1a, aes(x = Alcohol, y = .resid)) + geom_point()
```



Comments

- residuals vs fitted and vs `Gastric`: suggestion of curve?
- residuals vs `Sex`: males more spread out
- residuals vs `Alcohol`: non-alcoholics more spread out

What to do next?

- curves plus unequal spreads suggest transformation of response
- problems with only one or two explanatory variables suggest transforming just those (our problems here are bigger than this)
- if we have some theory about relationship, can use that to guide transformation (eg. take logs, but don't have that here)
- can use *Box-Cox* to suggest good transformation.
  - aims to find transformation of response that promotes straightness plus equal spread (will also often reduce influence of large values).

Ladder of powers

from here:

Power	Name	Comment
2	Square of data values	Try with unimodal distributions that are skewed to the left.
1	Raw data	Data with positive and negative values and no bounds are less likely to benefit from re-expression.
½	Square root of data values	Counts often benefit from a square root re-expression. For counted data, start here.
"0"	We'll use logarithms here	Measurements that cannot be negative often benefit from a log re-expression.
-1/2	Reciprocal square root	An uncommon re-expression, but sometimes useful.
-1	The reciprocal of the data	Ratios of two quantities (e.g., mph) often benefit from a reciprocal.

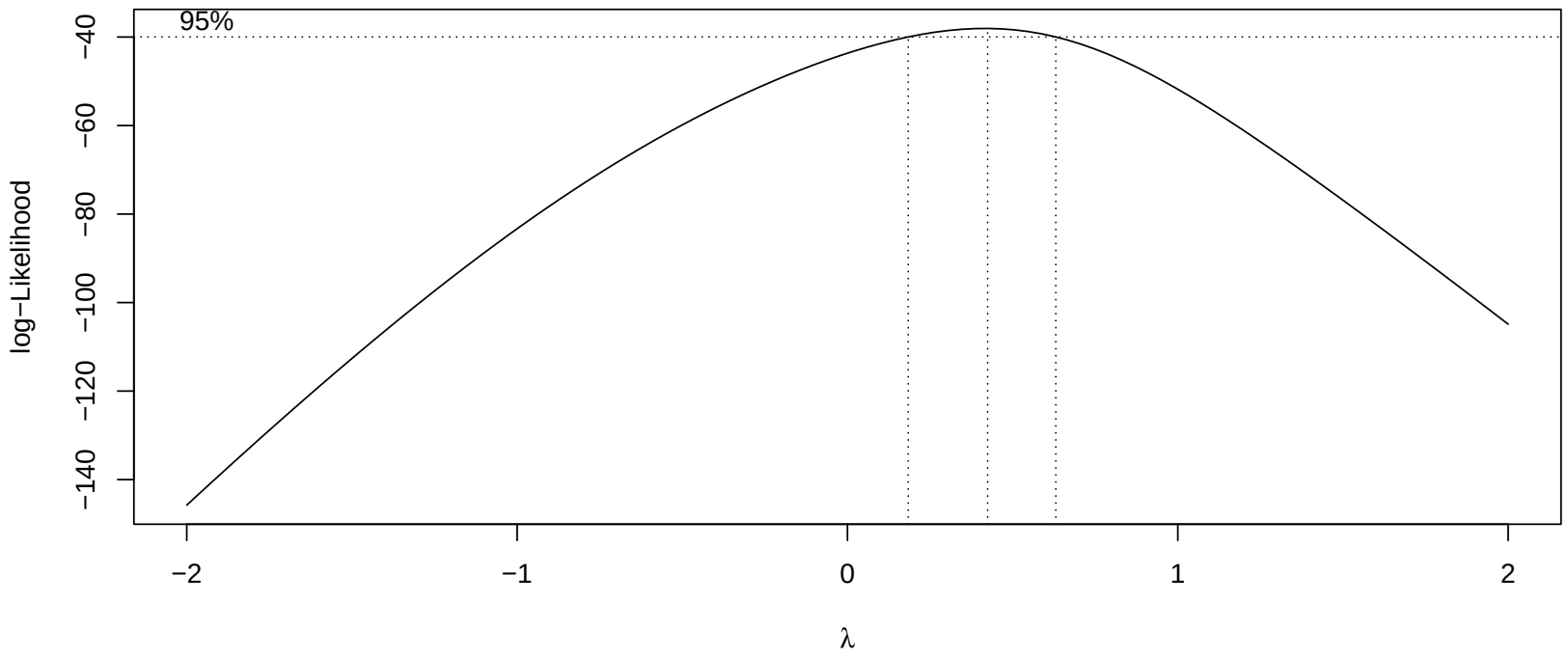
Box-Cox

- Box and Cox developed a method for choosing a transformation from the ladder of powers.
  - They worked out how to estimate the transformation as a power, but making zero be log (since power zero makes no sense otherwise)
- In package `MASS` as `boxcox`
- Technical detail: `MASS` also has a `select` that we don't want to have interfere with the `tidyverse` `select`, so load `MASS` as shown:

Running Box-Cox

- use `boxcox` with a model formula (such as you would use in `lm`):

```
boxcox(Metabol ~ Gastric + Sex + Alcohol, data = alc)
```



Comments

- Output is a graph.
- Peak of curve is single best power transformation
- Outer vertical lines mark 95% CI for transformation power
- Goal: find value from ladder of powers that is within CI
- Here that is 0.5, square root
- Note that 1, “do nothing”, and 0, “take logs”, are not supported by data.

Re-do regression

- with square root of `Metabol` as response:

```
alc.2 <- lm(sqrt(Metabol) ~ Gastric + Sex + Alcohol,
            data = alc)
```

Output

```
Call:
lm(formula = sqrt(Metabol) ~ Gastric + Sex + Alcohol, data = alc)

Residuals:
     Min       1Q   Median       3Q      Max
-0.74769 -0.26586  0.01346  0.25088  0.75600

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.22929    0.17494   1.311 0.200607
Gastric               0.50406    0.07948   6.342 7.33e-07 ***
SexMale               0.55856    0.15195   3.676 0.000995 ***
AlcoholNon-alcoholic -0.04690    0.16556  -0.283 0.779042
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3731 on 28 degrees of freedom
Multiple R-squared:  0.768, Adjusted R-squared:  0.7431
F-statistic: 30.89 on 3 and 28 DF,  p-value: 5.03e-09
```

Remove Alcohol:

```
alc.3 <- lm(sqrt(Metabol) ~ Gastric + Sex, data = alc)
summary(alc.3)
```

```
Call:
lm(formula = sqrt(Metabol) ~ Gastric + Sex, data = alc)

Residuals:
     Min       1Q   Median       3Q      Max
-0.77319 -0.23494  0.02865  0.26615  0.74255

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.20185    0.14334   1.408 0.169709
Gastric      0.49655    0.07373   6.735 2.17e-07 ***
SexMale      0.57286    0.14103   4.062 0.000338 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3671 on 29 degrees of freedom
Multiple R-squared:  0.7673,    Adjusted R-squared:  0.7513
F-statistic: 47.81 on 2 and 29 DF,  p-value: 6.584e-10
```

Comments; set up to check residuals again

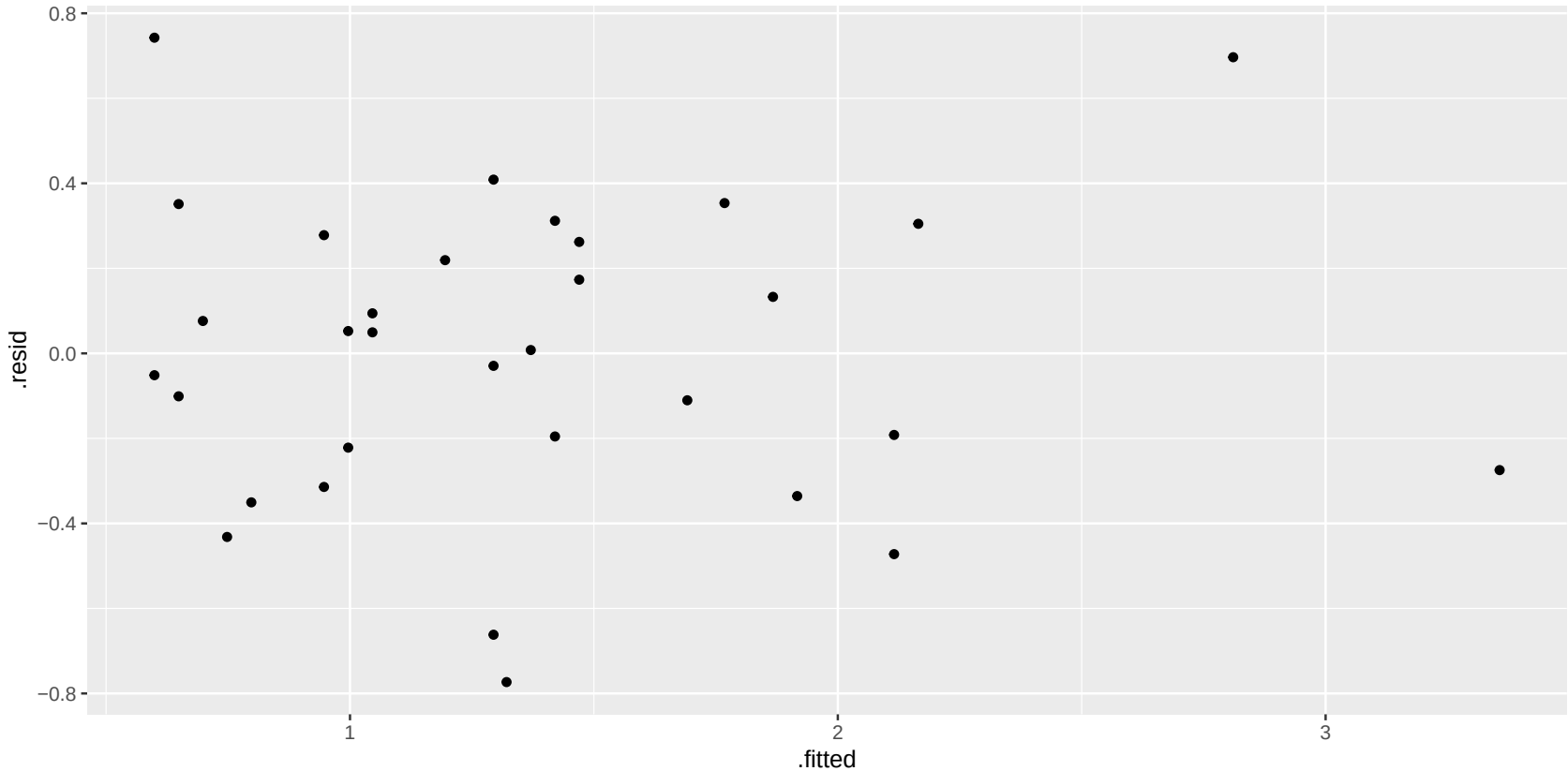
- `Alcohol` was not significant in the regression, so removed it. If you have more than one non-significant explanatory variable, remove only the *one* least significant one (highest P-value), refit, re-evaluate.
- Everything else significantly adds to the regression, so cannot remove anything else.

```
alc.3 %>% augment(alc) -> alc.3a
```

- Expect the residual plots to be a lot better.
- In principle, check again:
  - residuals vs. fitted
  - normal quantile plot of residuals
  - residuals against each explanatory

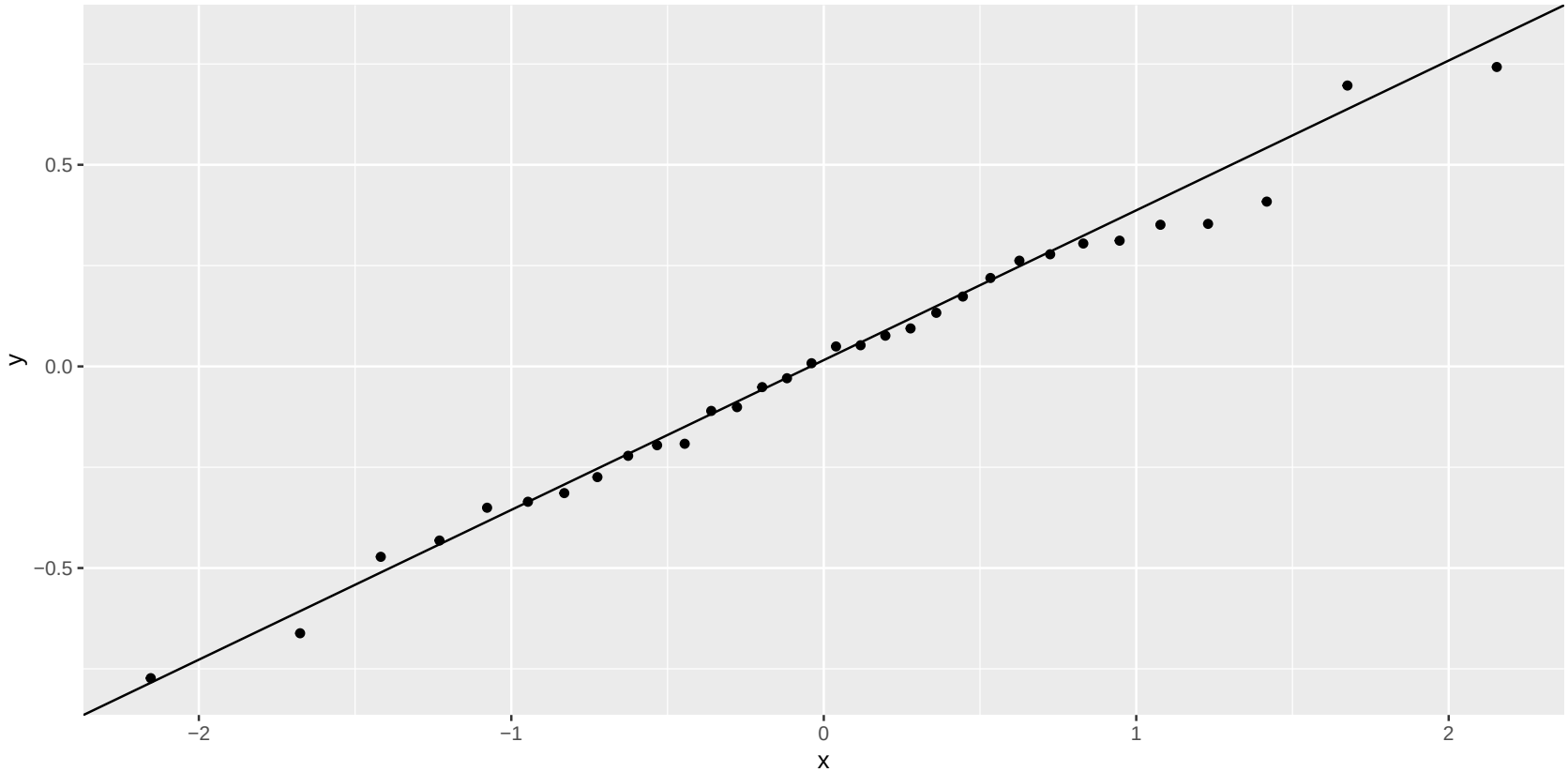
Residuals against fitted

```
ggplot(alc.3a, aes(x = .fitted, y = .resid)) +
  geom_point()
```



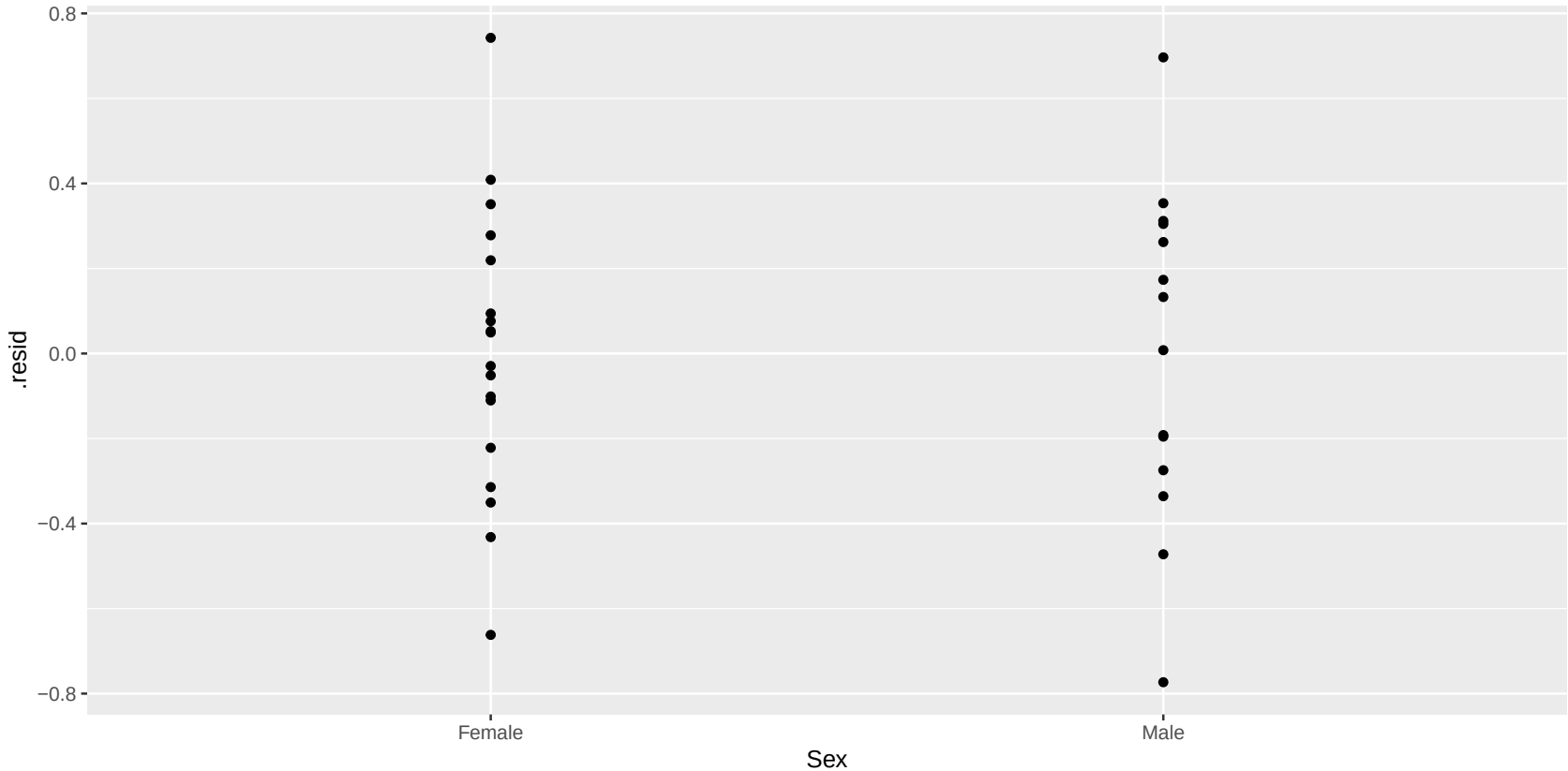
Normal quantile plot of residuals

```
ggplot(alc.3a, aes(sample = .resid)) +
  stat_qq() + stat_qq_line()
```



Residuals against Sex

```
ggplot(alc.3a, aes(x = Sex, y = .resid)) + geom_point()
```



Comments

- Fitted vs residual looks less curved
- Residuals look much more normal (high values brought down)
- Residuals vs `Sex` have no outliers and look equally spread

Back to regression output

```
glance(alc.3)
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <
1     0.767         0.751 0.367      47.8 6.58e-10     2  -
11.8  31.5  37.4
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
tidy(alc.3)
```

```
# A tibble: 3 x 5
  term        estimate std.error statistic     p.value
  <chr>          <dbl>     <dbl>     <dbl>       <dbl>
1 (Intercept)    0.202    0.143       1.41 0.170
2 Gastric        0.497    0.0737      6.74 0.000000217
3 SexMale        0.573    0.141       4.06 0.000338
```

Comments

- R-squared is reasonably high
- Slope for `Gastric` is significantly positive: as gastric activity increases, metabolism increases (for everybody)
- Slope for `SexMale` positive: compared to baseline `Sex` (Female), metabolism is higher for Males *even allowing for the effect of* `Gastric`.
- We removed `Alcohol`: being alcoholic does not affect metabolism over and above the other explanatory variables.
- This all seems to support the researchers' theory.

More details of researchers' theory xxx

(see Sleuth 3, case 11.1.1)