STATISTICAL FOUNDATIONS

Heike Hofmann

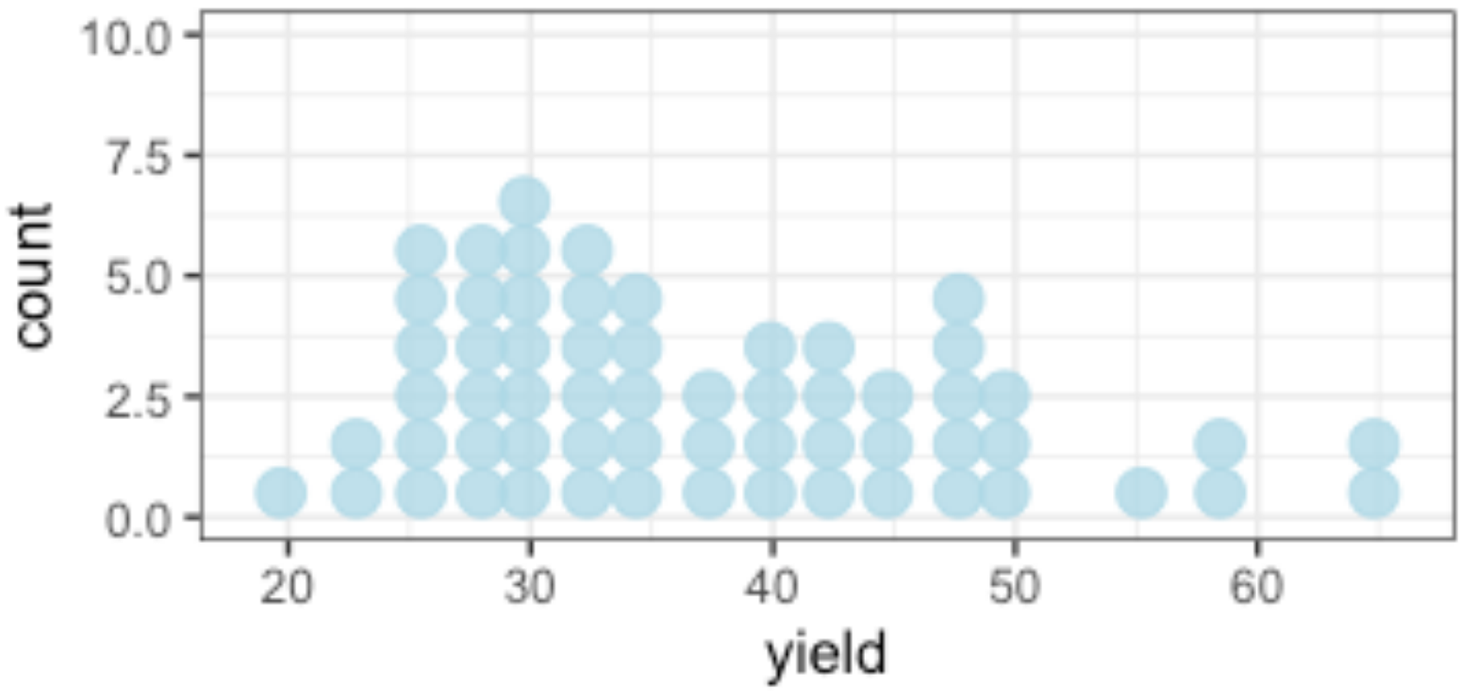
DAY 2

- Estimation of means and proportions (MLE)
- Confidence Intervals
- (Normal) Linear Models
- Predictions

ESTIMATION OF PARAMETERS

THE SITUATION

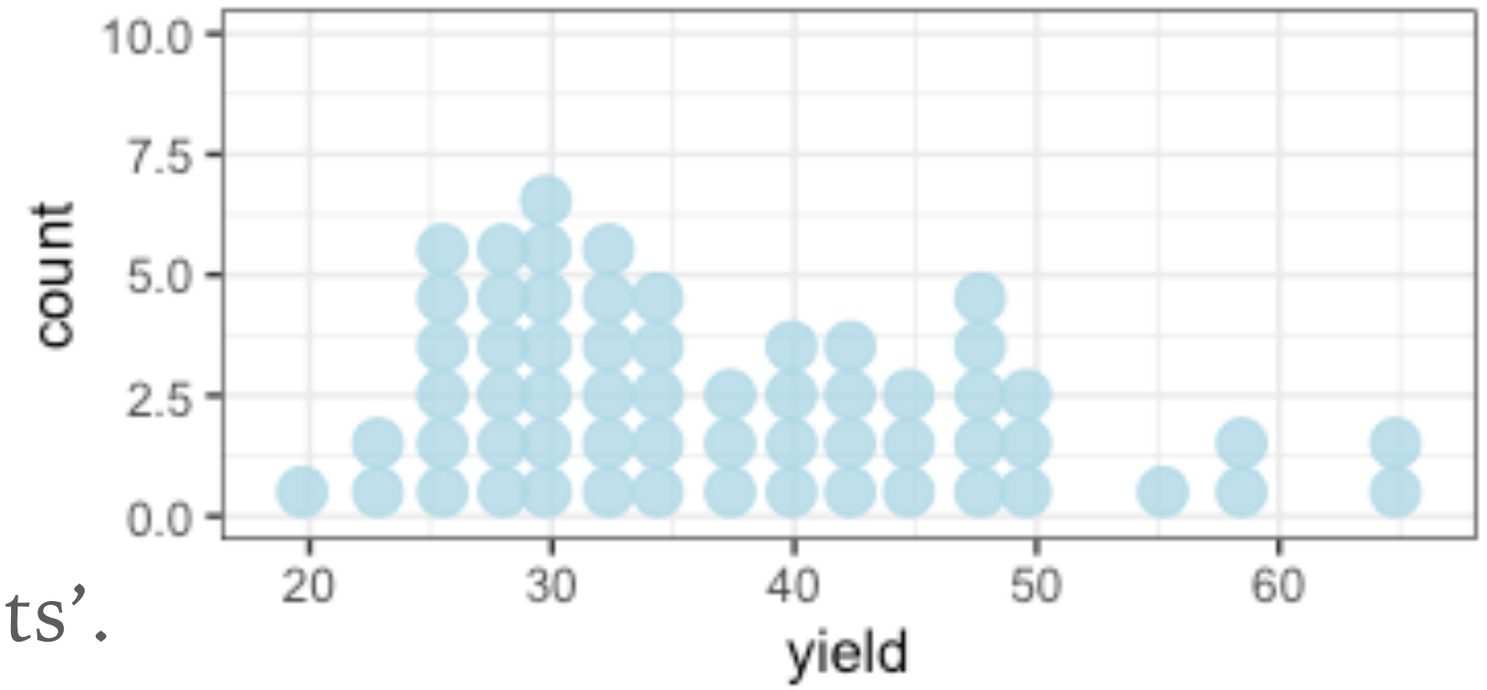
- We have measurements (barley yields) from an ag experiment
- Let us assume barley yield follows a normal distribution
- The shape of yield looks mostly uni-modal and mostly symmetric (cough)



- Which values should we use for μ and σ^2?

MAXIMUM LIKELIHOOD ESTIMATION

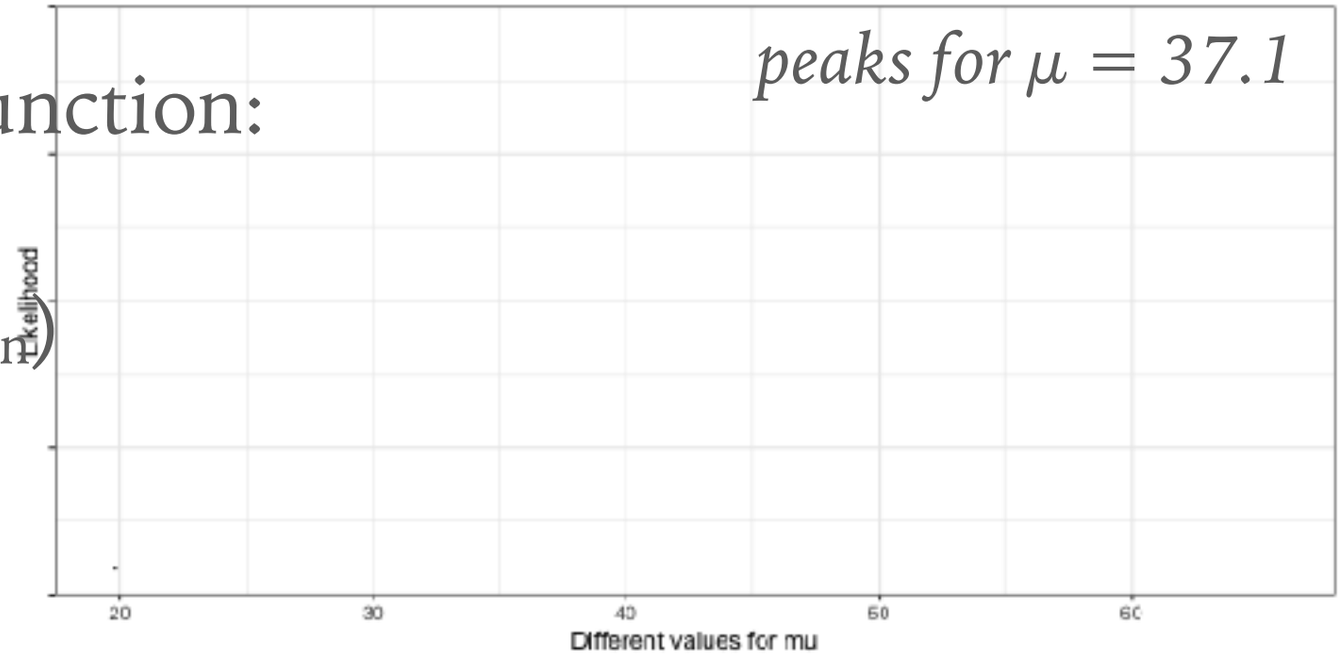
- Basic idea:
  move the density function around (by using different parameters) and see where it best 'fits'.



varying μ from 20 to 65

- Setup the **Likelihood** function:
  $L(\mu, \sigma^2 \mid x_1, \ldots, x_n) =$
  $f_{\mu, \sigma}(x_1) \cdot f_{\mu, \sigma}(x_2) \cdot \ldots \cdot f_{\mu, \sigma}(x_n)$

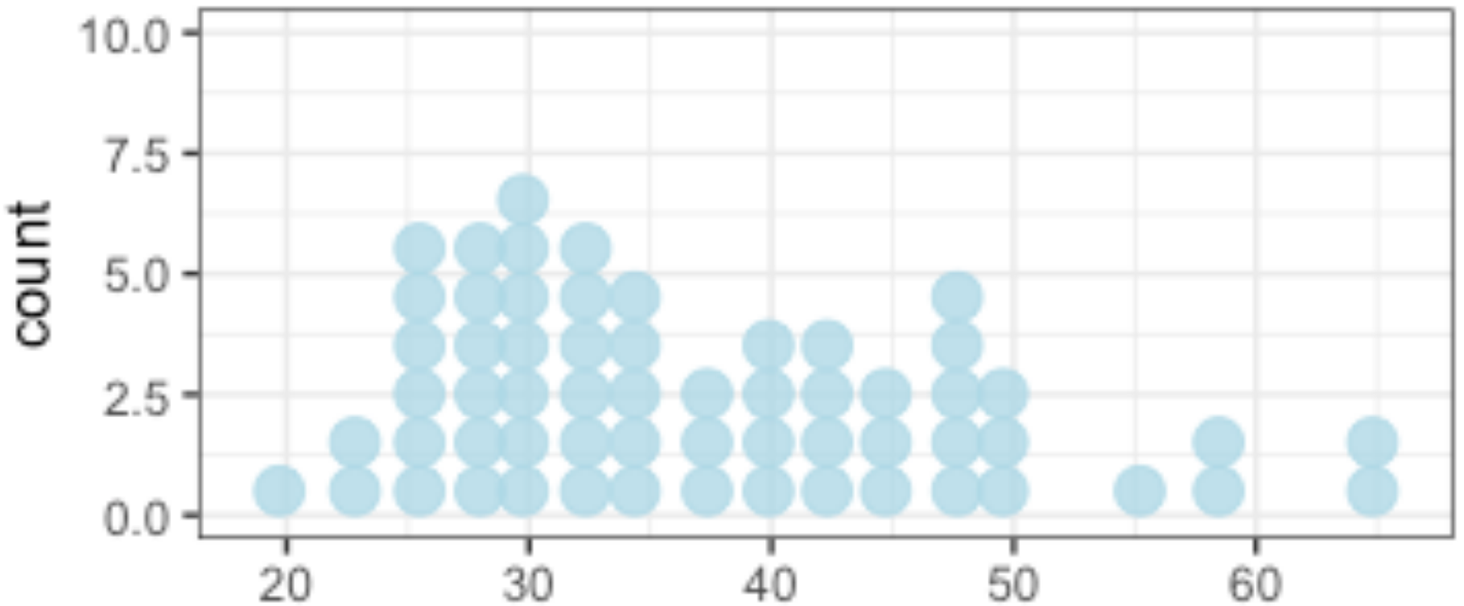
  very similar to probability, but with focus on parameters



peaks for $\mu = 37.1$

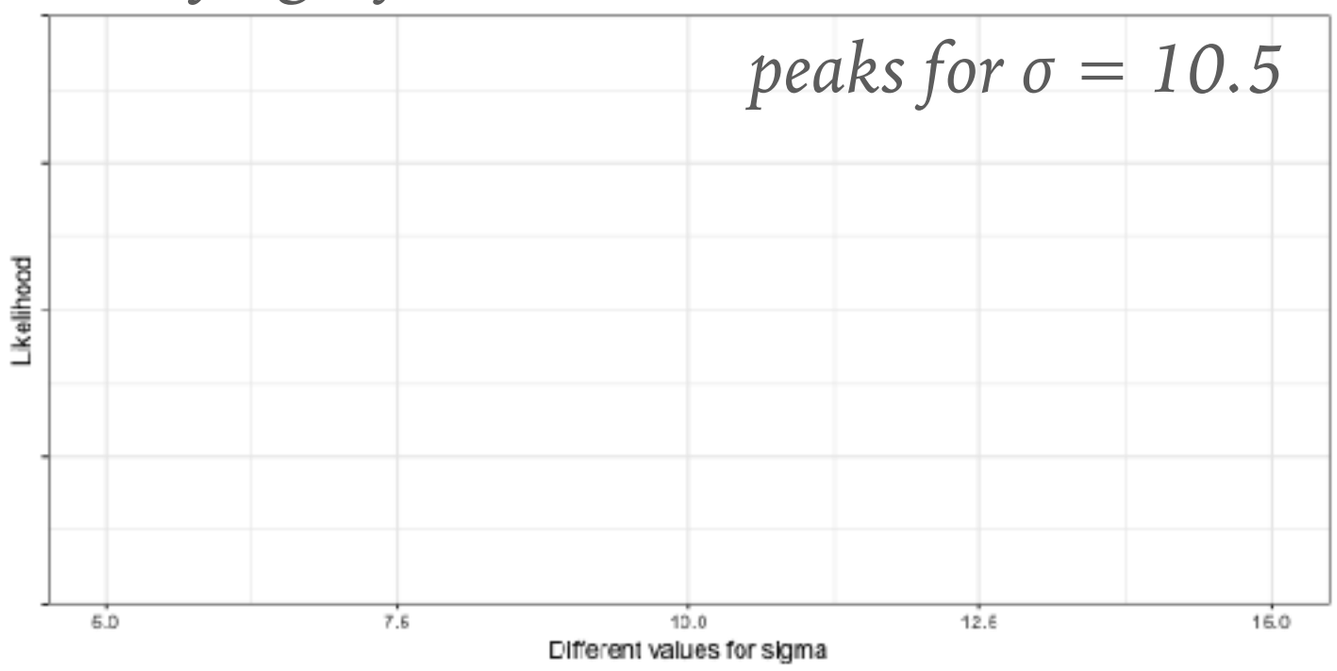
MAXIMUM LIKELIHOOD ESTIMATION

- Likelihood peaks for $\mu = 37.1$



- Now vary σ:

varying σ from 5 to 15



MAXIMUM LIKELIHOOD ESTIMATION

- Assume data $x_1, \ldots, x_n$ are observations of RVs $X_i \sim F_\theta$
- Setup the **Likelihood** function:
  $L(\theta \mid x_1, \ldots, x_n) = f_\theta(x_1) \cdot f_\theta(x_2) \cdot \ldots \cdot f_\theta(x_n)$
- Find a maximum of L in θ by getting derivative w.r.t θ, set to zero, and solve for θ.
- The Maximum Likelihood Estimator $\hat{\theta}$ of θ is defined as:

$$\hat{\theta} = \arg\max_{\theta} L(\theta \mid x_1, \ldots, x_n)$$

MAXIMUM LIKELIHOOD ESTIMATOR OF NORMAL

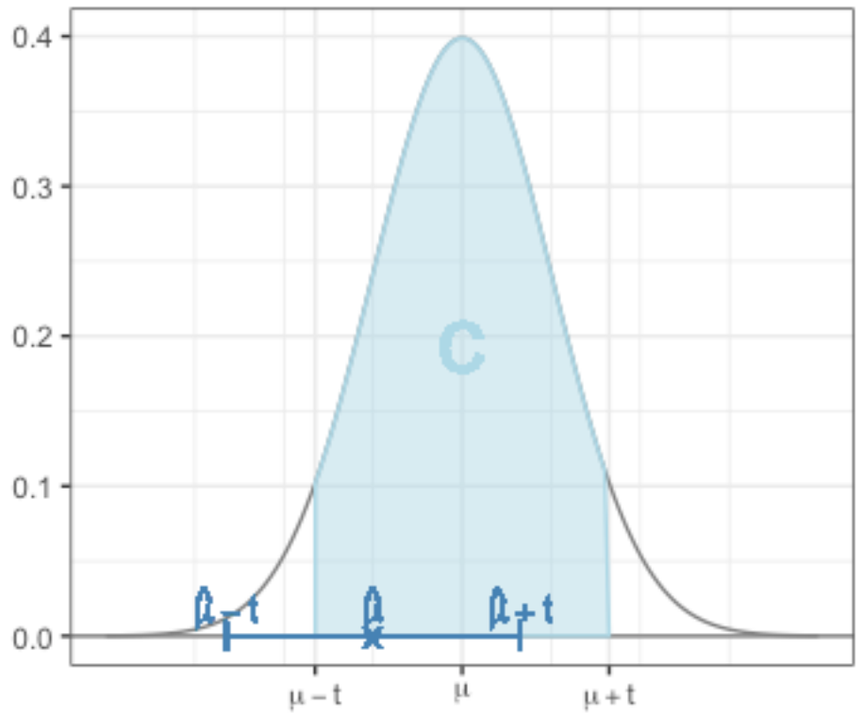
- The MLE for μ and σ^2 of a normal distribution are

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \text{and} \qquad \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

- What can we say about these estimators?
- They are both averages of i.i.d. RVs!
- Central Limit Theorem tells us about their approximate distribution!
- $\hat{\mu}$ is average of i.i.d RVs and therefore $\hat{\mu} \sim N(\mu, \sigma^2/n)$

MAXIMUM LIKELIHOOD ESTIMATOR OF NORMAL

- $\hat{\mu}$ is average of i.i.d RVs and therefore $\hat{\mu} \sim N(\mu, \sigma^2/n)$



- For all $\hat{\mu} \in (\mu - t, \mu + t)$ we know that $\mu \in (\hat{\mu} - t, \hat{\mu} + t)$
- This is the basis of a confidence interval for μ

CONFIDENCE INTERVALS

CONFIDENCE INTERVALS

- The c·100% Confidence interval of θ is defined as

$$(\hat{\theta}-t,\hat{\theta}+t)$$

where $P(|\hat{\theta}-\theta| < t) > c$

- For μ we get a c·100% confidence interval as

$$\left(\bar{X} - z \cdot \frac{\sigma}{\sqrt{n}}, \bar{X} + z \cdot \frac{\sigma}{\sqrt{n}}\right)$$ if σ is known

that's a pretty big assumption

z are the critical values from N(0,1):

	0.5(c+1)	z
c = 0.9	0.95	1.645
c = 0.95	0.975	1.960
c = 0.99	0.995	2.576

CONFIDENCE INTERVALS – UNKNOWN SIGMA

- The c·100% Confidence interval of θ is defined as

$$(\hat{\theta}-t,\hat{\theta}+t)$$

where $P(|\hat{\theta}-\theta| < t) > c$

- For μ we get a c·100% confidence interval as

$$\left(\bar{X} - t\frac{s}{\sqrt{n}}, \bar{X} + t\frac{s}{\sqrt{n}}\right)$$

for unknown σ we substitute with s

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

t are the critical values from a t distribution with (n-1) degrees of freedom: `qt((c+1)/2, df = n-1)`

EXAMPLE: MEAN YIELD OF BARLEY

- What is a 95% confidence interval for the mean yield μ of barley?

- Assume $\sigma = 10.5$, $\hat{\mu} = 37.1$, and $n = 60$

1.96 is 95% critical value of N(0,1)

95% C.I. for μ is $(37.1 - 1.96 \cdot 10.5/\sqrt{60},\ 37.1 + 2.65) =$
$= (34.4, 39.8)$

- For unknown σ we estimate (from the data) $s = 10.6$

$t = 2.00$ is critical value of t_{59}

95% C.I. for μ is $(37.1 - 2.00 \cdot 10.6/\sqrt{60},\ 37.1 + 2.74) =$
$= (34.4, 39.8)$

difference in C.I.s only shows in 2nd significant digit

YOUR TURN

Transformations are often used to stabilize variances (or make the distribution 'more' normal). being able to back-transform is important for interpretability.

Work on the questions by yourself

LOG TRANSFORM

- Barley yields look a bit skewed to the right (too many extreme cases of high yields compared to very low yields).
- The following code gives access to the data:

```
data(barley, package="lattice")
b31 <- subset(barley, year == 1931)
```

- Log-transform `yield` to `lyield` and draw a histogram or stacked dotplot. Is the result less skew?
- Calculate a 95% C.I. for the mean log yield.
- Back-transform the result and compare to the previous intervals.

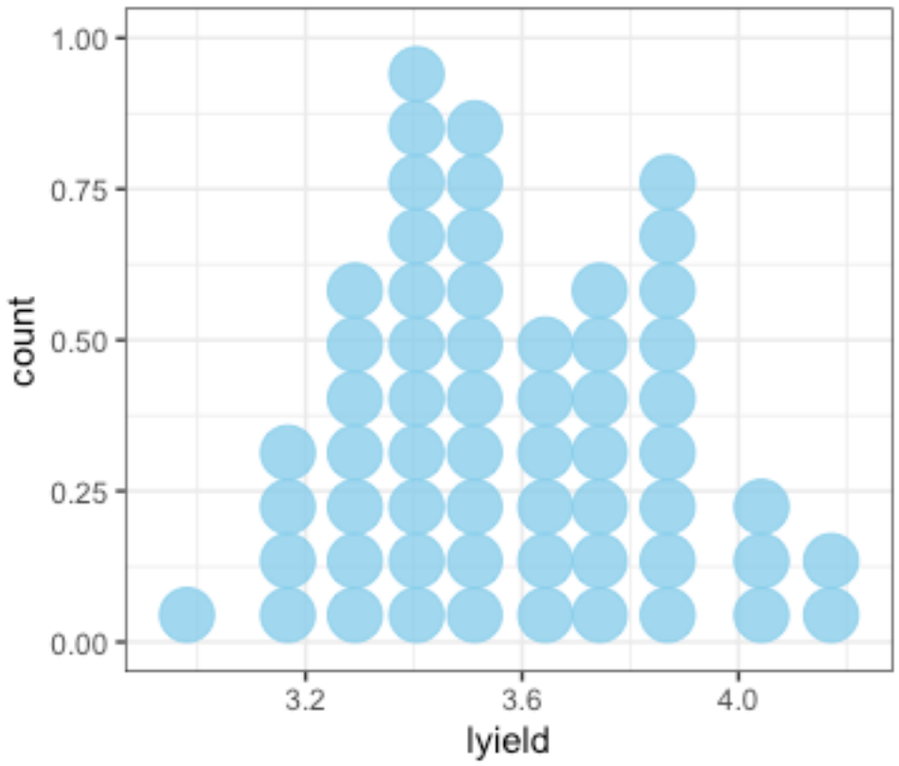
YOUR TURN SOLUTION – LOG TRANSFORM

- Get the data:

```
library(tidyverse)
data(barley, package="lattice")
b31 <- barley %>%
  filter(year==1931)
```

- Modify & plot:

```
b31$lyield <- log(b31$yield)
b31 %>%
  ggplot(aes(x = lyield)) +
  geom_dotplot(binwidth=.1, fill="skyblue",
                            colour = "skyblue",
               alpha = 0.8) +
  theme_bw()
```



- 95% C.I.

```
mu <- mean(b31$lyield); s <- sd(b31$lyield); t <- qt(.975, 59)

mu + c(-1,1)*t*s/sqrt(60)
[1] 3.503186 3.646208
exp(mu + c(-1,1)*t*s/sqrt(60))
[1] 33.22113 38.32904
```

Compare to previous C.I

(34.4, 39.8)

QUESTIONS?

NORMAL LINEAR MODELS

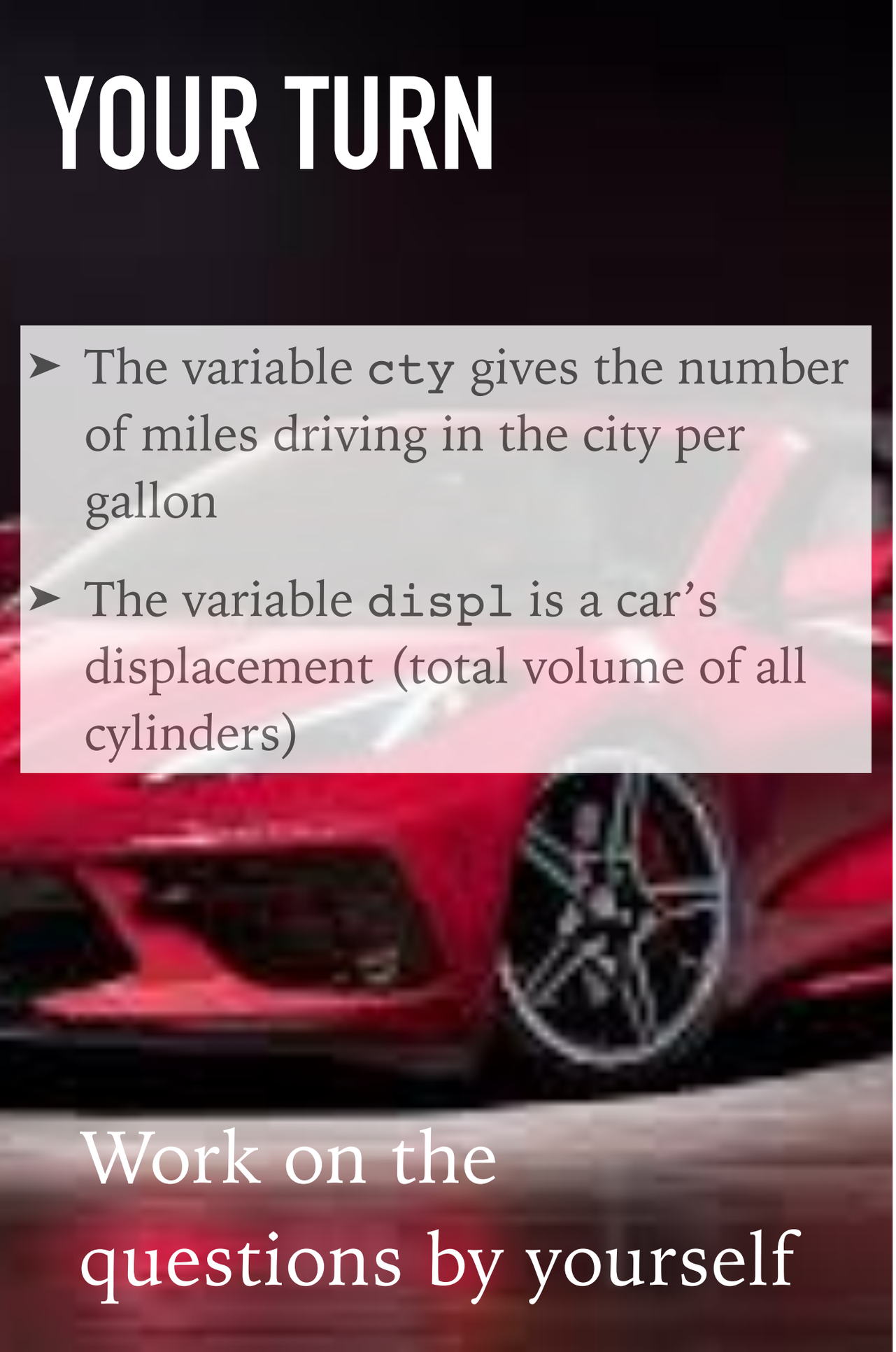
LINEAR REGRESSION

- Assume we have observations from RVs X_1, … X_k, and we are trying to approximate the behavior of RV Y by finding a function g such that $Y \approx g(X_1, \ldots X_k)$
- Simplest case: $k = 1$ and g is linear $g(x) = ax + b$

- Always: draw scatterplot of X and Y
- Sometimes: transform X and/or Y to get to linear relationship

YOUR TURN

- The variable `cty` gives the number of miles driving in the city per gallon
- The variable `displ` is a car's displacement (total volume of all cylinders)



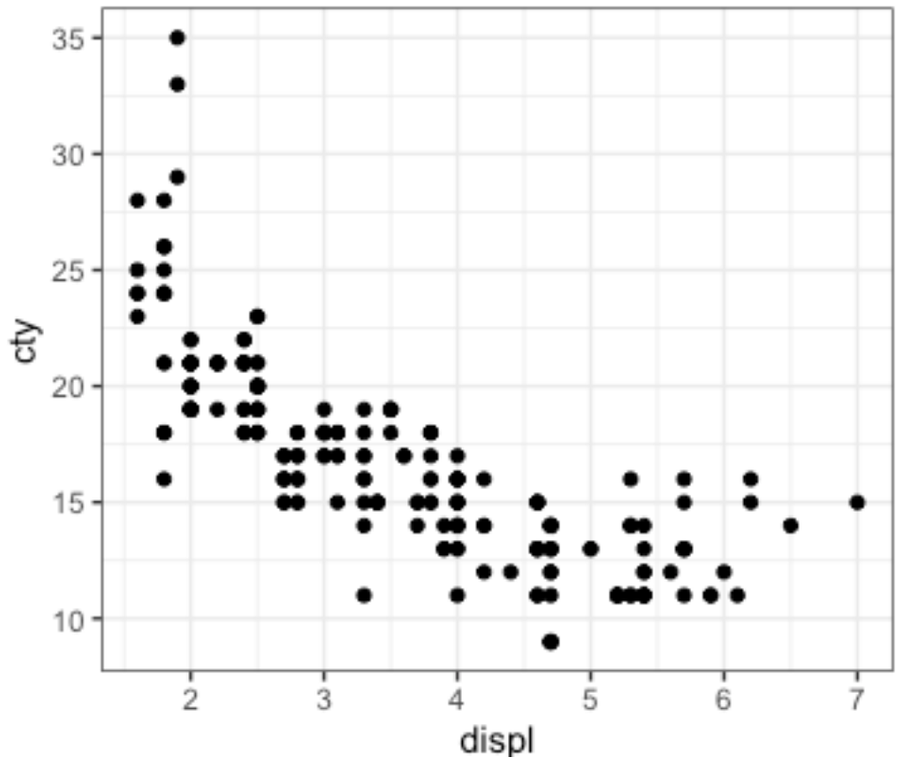
Work on the questions by yourself

DISPLACED?

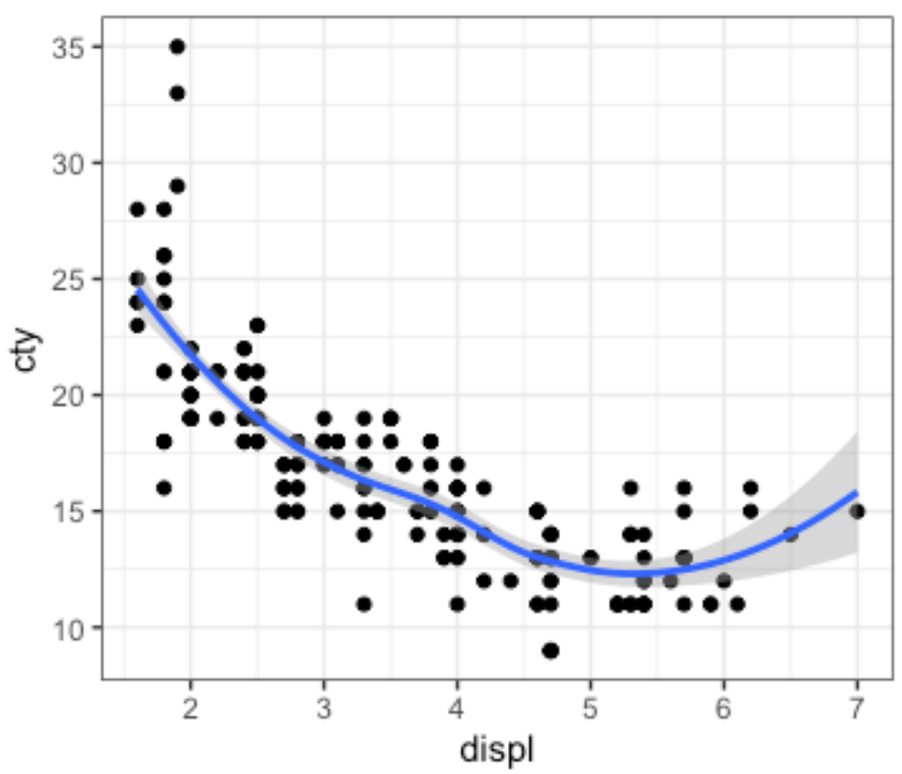
- The following code gives access to data on mileage of cars:
  `data(mpg, package="ggplot2")`
- Draw a scatterplot of mileage in the city by displacement
- What relationship do you see
- Suggest a transformation to make the relationship 'more linear'.

YOUR TURN SOLUTION – DISPLACED

```
library(ggplot2)


mpg %>%
  ggplot(aes(x = displ, y = cty)) +
  geom_point() +
  theme_bw()
```

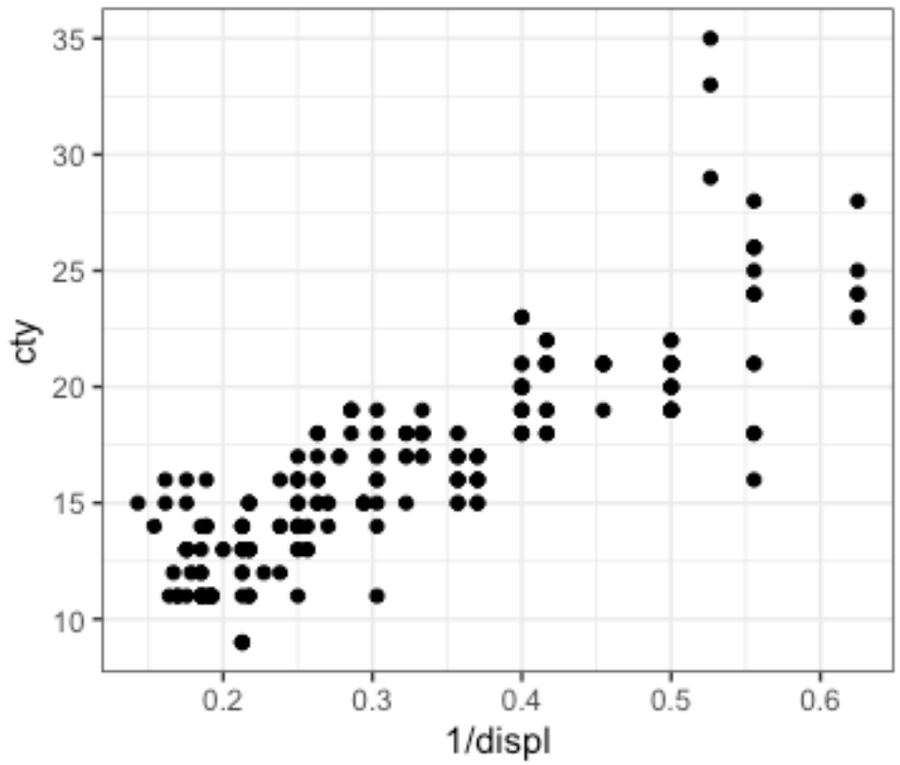



```
mpg %>%
  ggplot(aes(x = displ, y = cty)) +
  geom_point() +
  theme_bw() +
  geom_smooth()
```

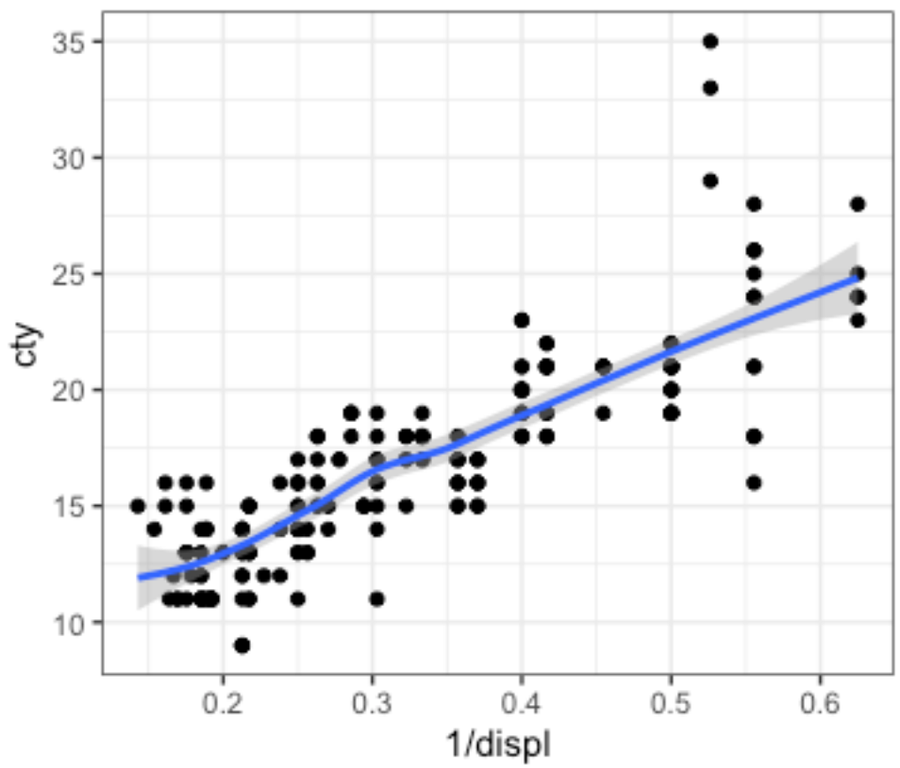


YOUR TURN SOLUTION – DISPLACED (2)

```
mpg %>%
  ggplot(aes(x = 1/displ, y = cty)) +
  geom_point() +
  theme_bw()
```

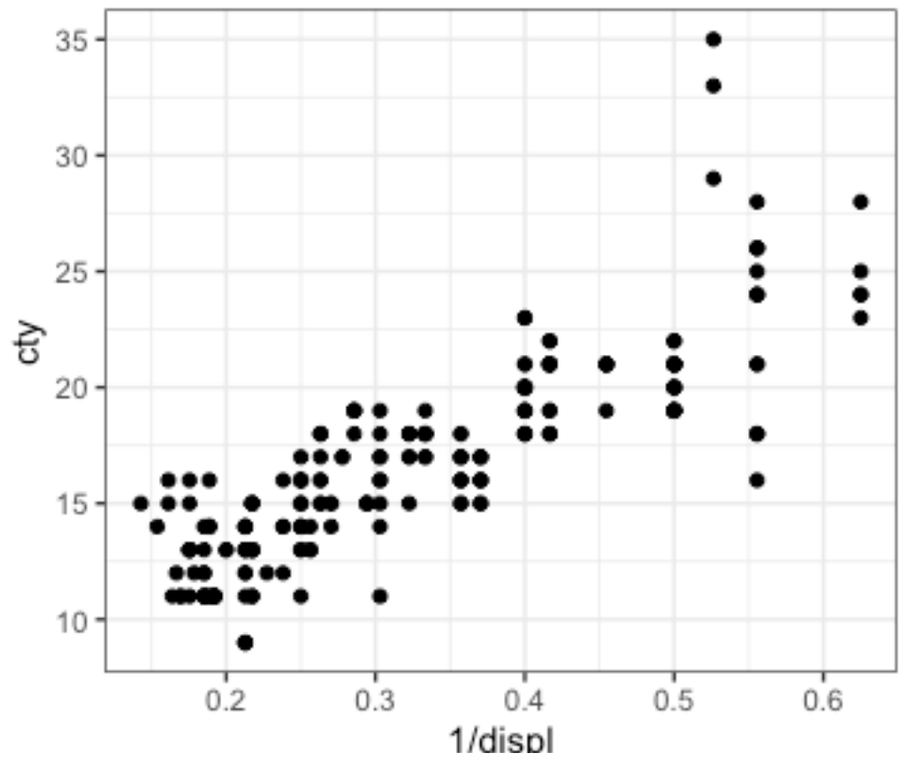



```
mpg %>%
  ggplot(aes(x = 1/displ, y = cty)) +
  geom_point() +
  theme_bw() +
  geom_smooth()
```



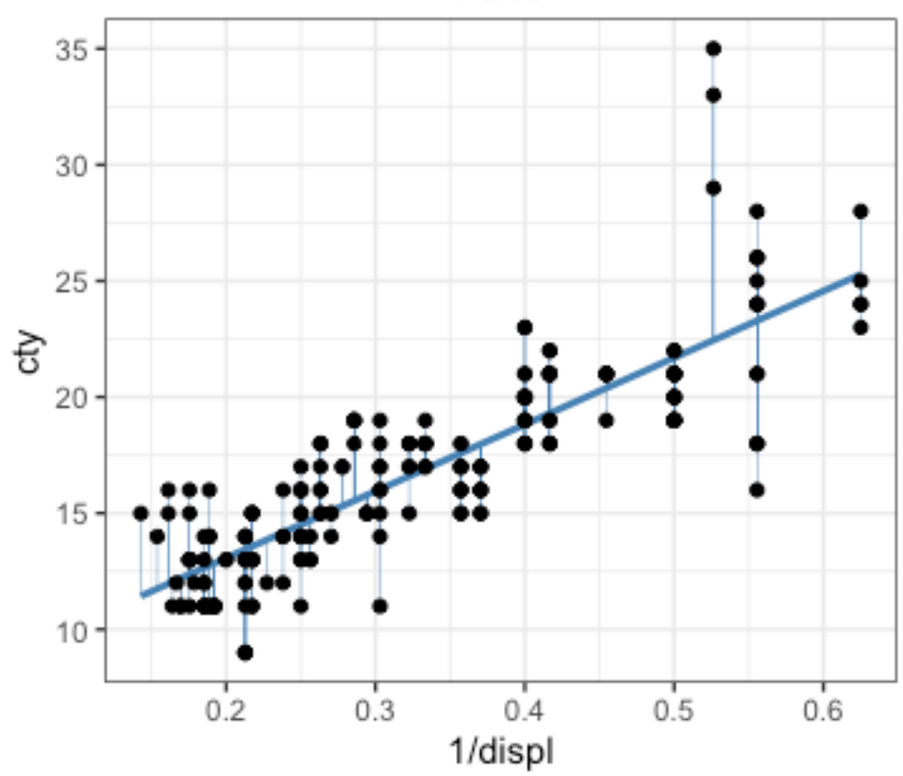
LINEAR REGRESSION

- Assume $Y \approx aX + b$
- How do we determine a and b?



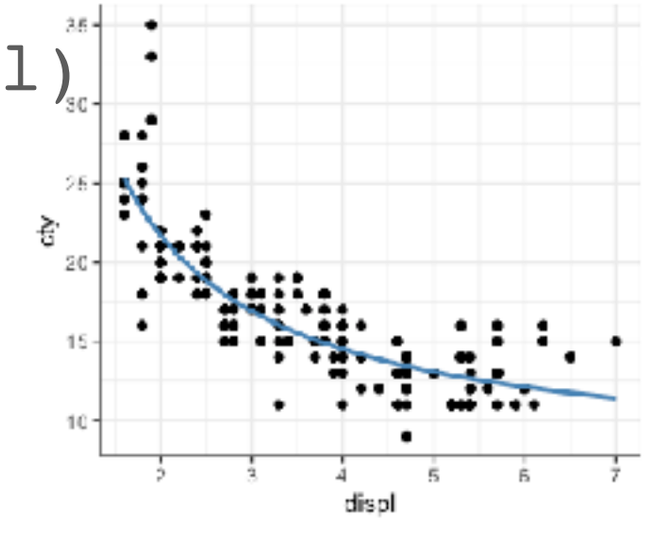
- Least squares: $Q(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$

  a and b are estimated such that Q(a,b) is minimized



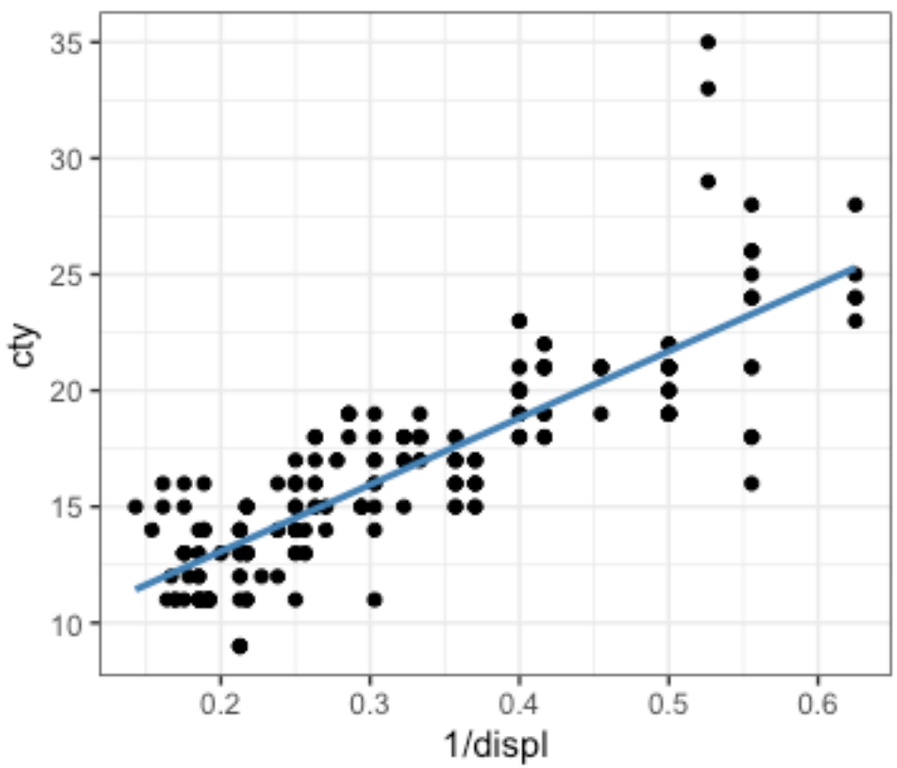
LINEAR MODELS IN R

- `mpg <- mpg %>% mutate(displ_inv = 1/displl)`




```
mod <- lm(cty~displ_inv, data = mpg)
summary(mod)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.3309     0.4253   17.24   <2e-16 ***
displ_inv    28.7243     1.2000   23.94   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- fitted values:
  `mpg$fitted <- fitted(mod)`



Interpretation: intercept of 7.33 is reached when 1/displ is zero

irrelevant/questionable?

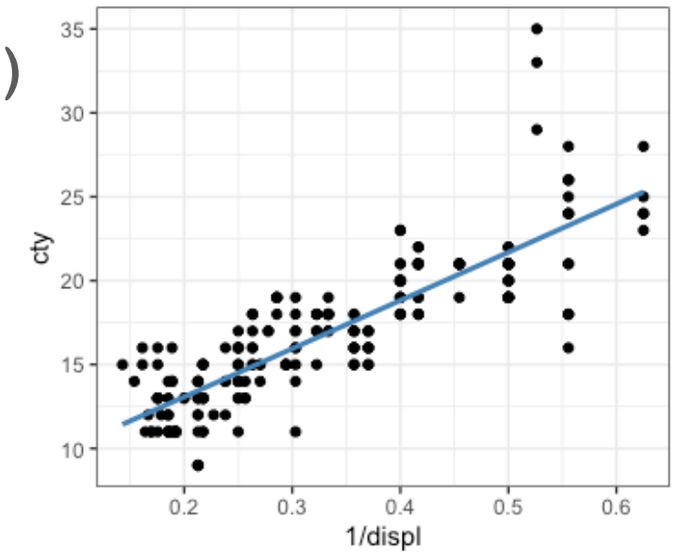
on average we see an increase of 28.7 miles per gallon in cty when 1/displ is increased by 1 unit

R SQUARED

- `mpg <- mpg %>% mutate(displ_inv = 1/displ)`

  `mod <- lm(cty~displ_inv, data = mpg)`

  `Multiple R-squared:  0.7118,`
  `Adjusted R-squared:  0.7106`



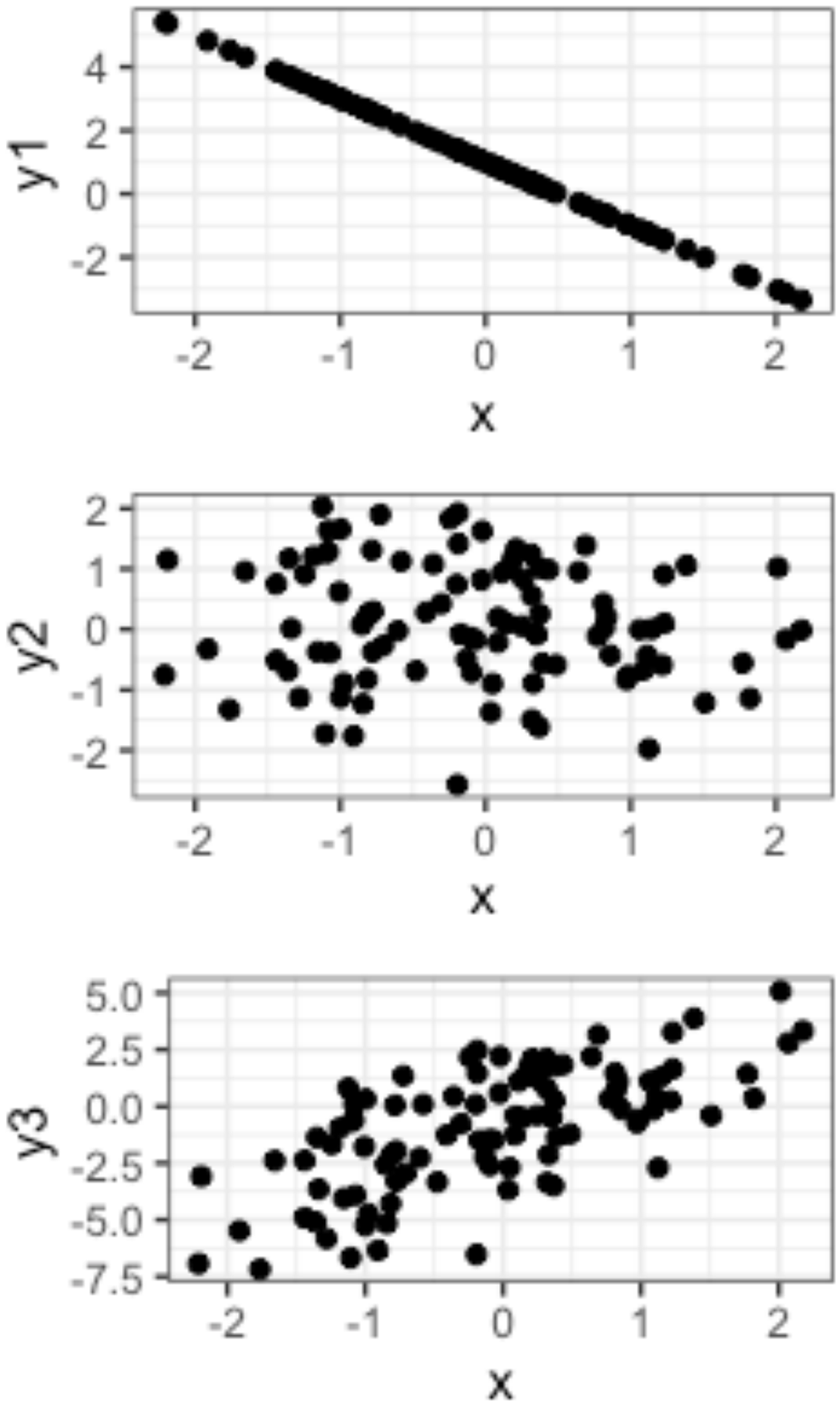
- R^2 is the "Coefficient of determination"
- R^2 is a measure of how much of the total variability in Y is explained by the covariate X:

$$R^2 = \frac{TSS - SSE}{TSS} \qquad SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \;\; where \;\; \hat{y}_i = ax_i + b$$

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- $0 \leq R^2 \leq 1$, with larger R^2 indicating a better linear fit

THE GOOD, THE BAD, AND THE . . . ACCEPTABLE?



- What is a good R^2 value for a model?
- $R^2 = 1$ indicates a perfect linear fit between X and Y, i.e. Y is a line in X
- $R^2 = 0$ indicates that X contributes nothing to model that the mean of Y doesn't explain.
- Depending on the application, a high or low R^2 might be misleading as a measure in itself:
  low R^2 indicates a lot of extra variability, but predictor X might be valuable

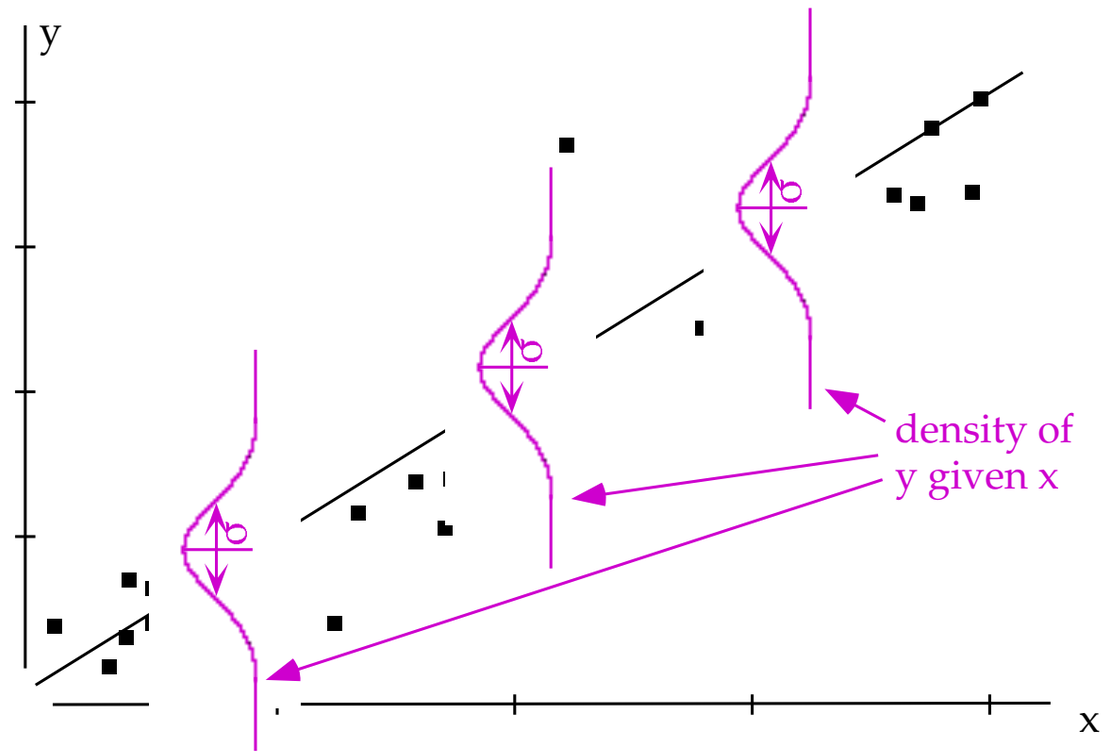
NORMAL MODEL

- Let us be more precise about the $Y \approx aX + b$:

$$Y = aX + b + \varepsilon$$

We will assume that ε denotes the error, and $\varepsilon \sim N(0, \sigma^2)$

- This also means that $Y|X$ is normal



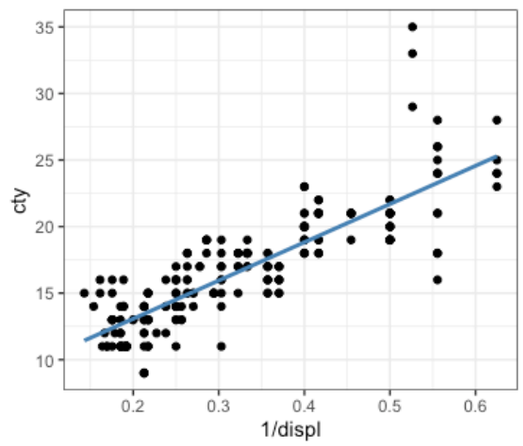
NORMAL MODEL

- A normal assumption for errors allows us to assess goodness of fit of the overall model as well as significance of predictors:

```
lm(formula = cty ~ displ_inv, data = mpg)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.3309     0.4253   17.24   <2e-16 ***
displ_inv     28.7243     1.2000   23.94   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.29 on 232 degrees of freedom
Multiple R-squared:  0.7118,Adjusted R-squared:  0.7106
F-statistic:   573 on 1 and 232 DF,  p-value: < 2.2e-16
```

Test, if parameter is equal to 0

Test that predictors are not needed

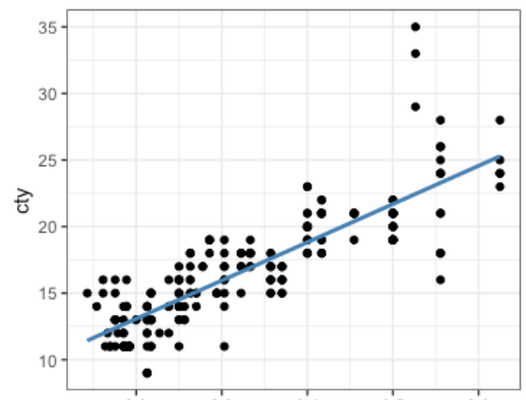


CONFIDENCE INTERVALS OF MODEL PARAMETERS

- A normal assumption for errors allows us to assess goodness of fit of the overall model as well as significance of predictors:

```
lm(formula = cty ~ displ_inv, data = mpg)
```

Test, if parameter is equal to 0



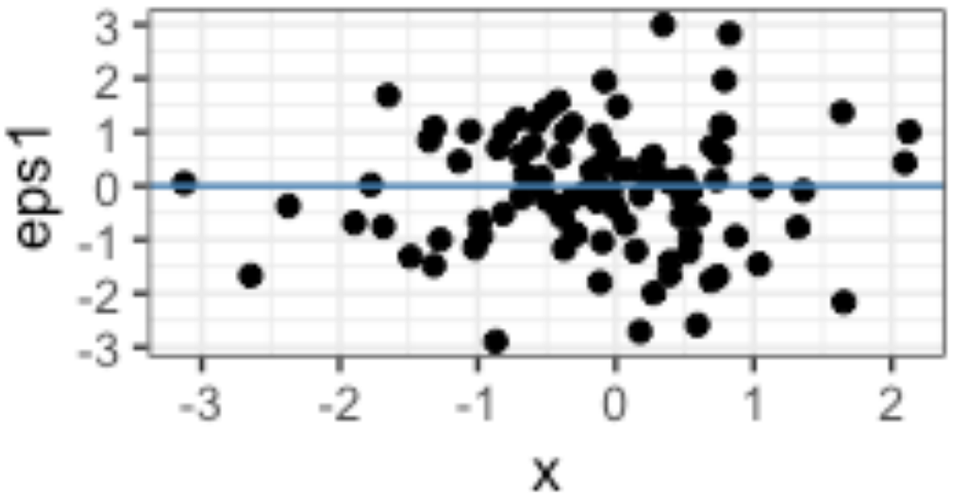

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.3309     0.4253   17.24   <2e-16 ***
displ_inv     28.7243     1.2000   23.94   <2e-16 ***
---
```

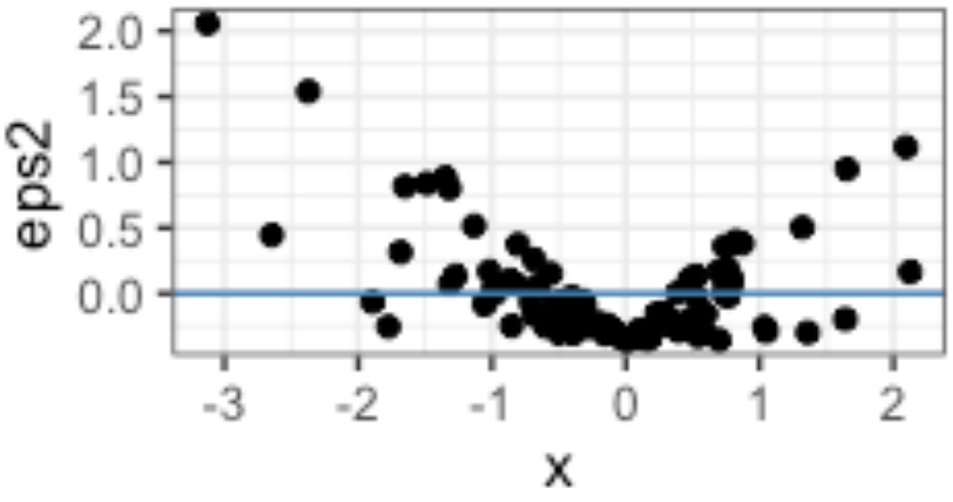
Confidence intervals for parameters:

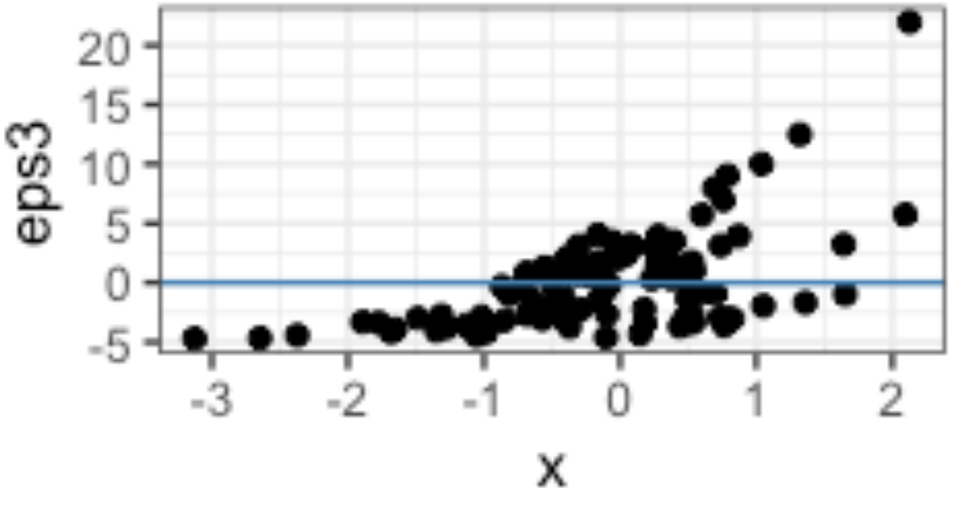
```
confint(mod, level = 0.95)
                2.5 %    97.5 %
(Intercept)  6.49306  8.16877
displ_inv   26.36009 31.08859
```

$$\hat{a} \pm t^*_{n-2} se(\hat{a})$$

NORMAL MODEL – RESIDUAL PLOTS







- We need to check for homogeneity in the variance of residuals
- Residuals should be spread symmetrically along each covariate
- Other structures indicate missing non-linear terms in a covariate (middle)
- Or the need for variance-stabilizing transformations (bottom)

YOUR TURN

- Run the line to read the data:
  `crabs <- read.csv("http://ggobi.org/book/data/australian-crabs.csv")`
- FL is frontal lip
  RW is rear width
  CL is carapace length
  CW is carapace width

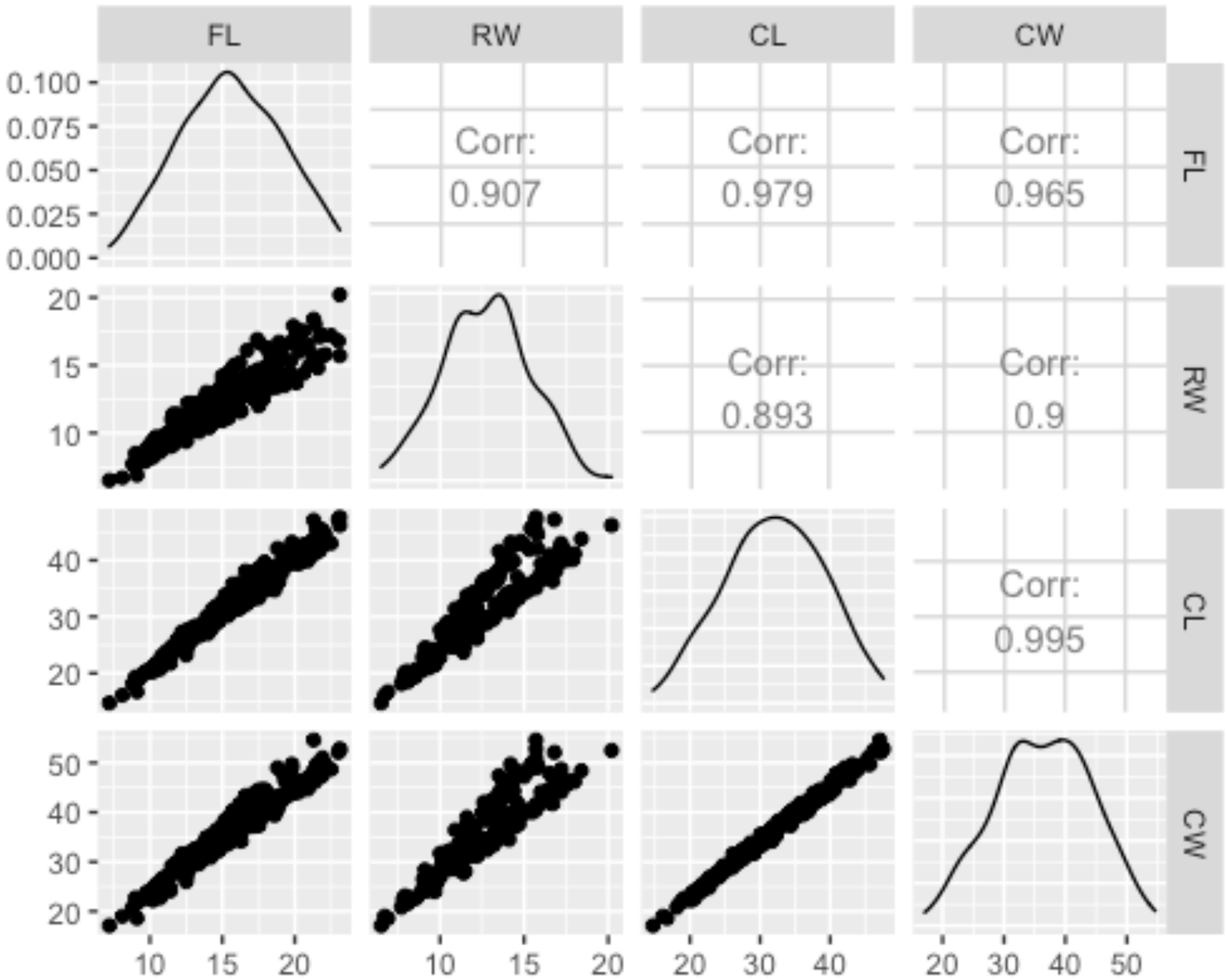
Work on the questions by yourself

ROCK CRABS

- Investigate the relationships between the physical measurements of the crabs (in scatterplots)
- Run a linear model of rear width in carapace length
- Assess residuals for homogeneity of variance
- Asses residuals by sex and species

YOUR TURN SOLUTION - ROCK CRABS

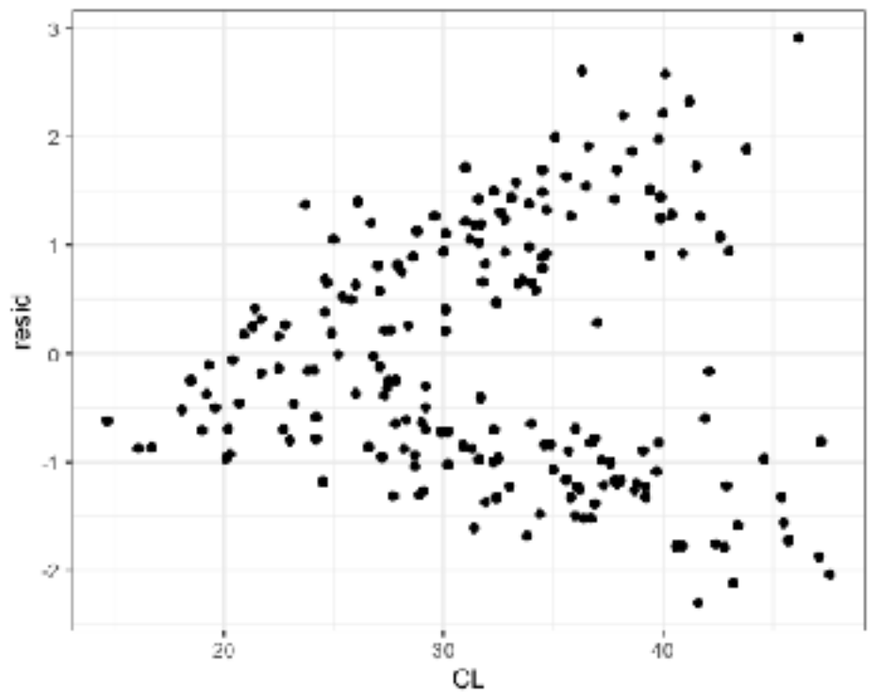
```
crabs <- read.csv("http://ggobi.org/book/data/australian-crabs.csv")
library(GGally)
ggpairs(crabs %>% select(FL, RW, CL, CW))
```



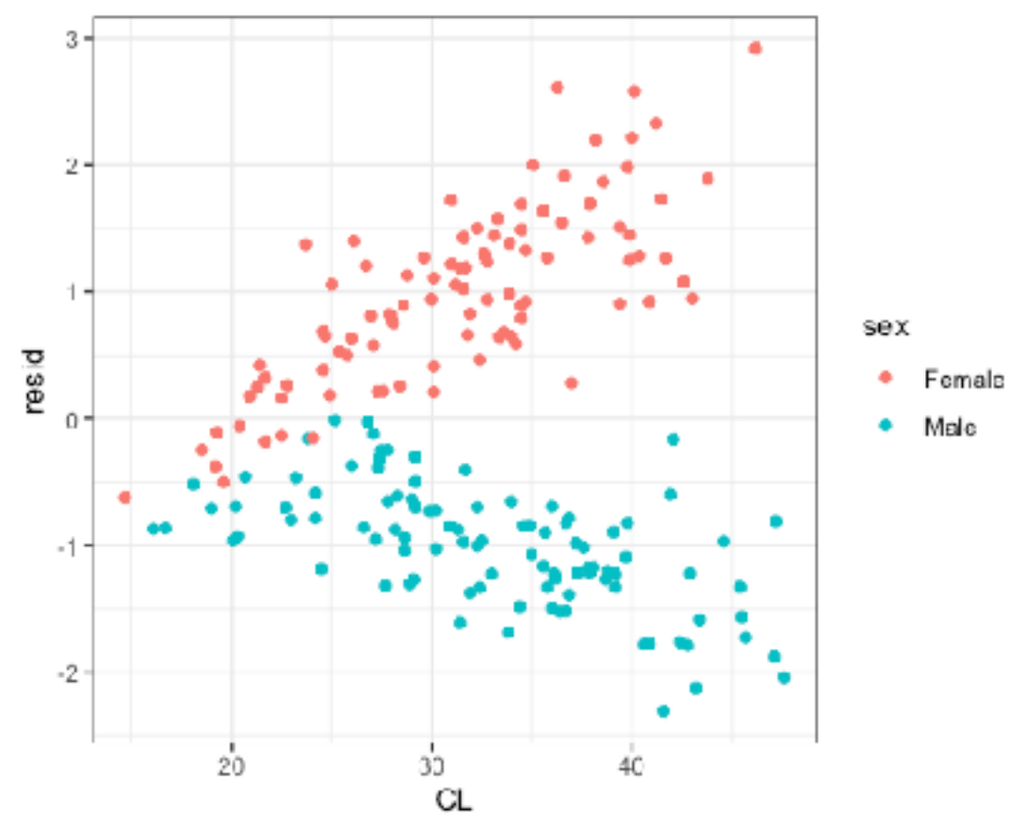
YOUR TURN SOLUTION - ROCK CRABS

```
m1 <- lm(RW~CL, data = crabs)

crabs$fitted <- fitted(m1)

crabs$resid <- resid(m1)
```

```
crabs %>%

  ggplot(aes(x = CL, y = resid)) +
  geom_point() +
  theme_bw()
```



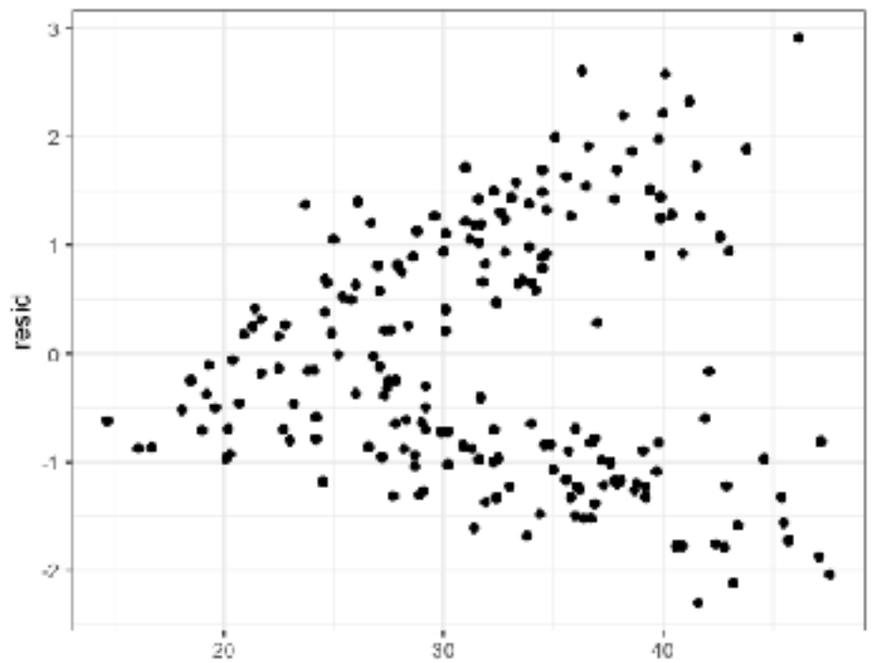

```
crabs %>%

  ggplot(aes(x = CL, y = resid)) +
  geom_point(aes(colour = sex)) +

  theme_bw()
```



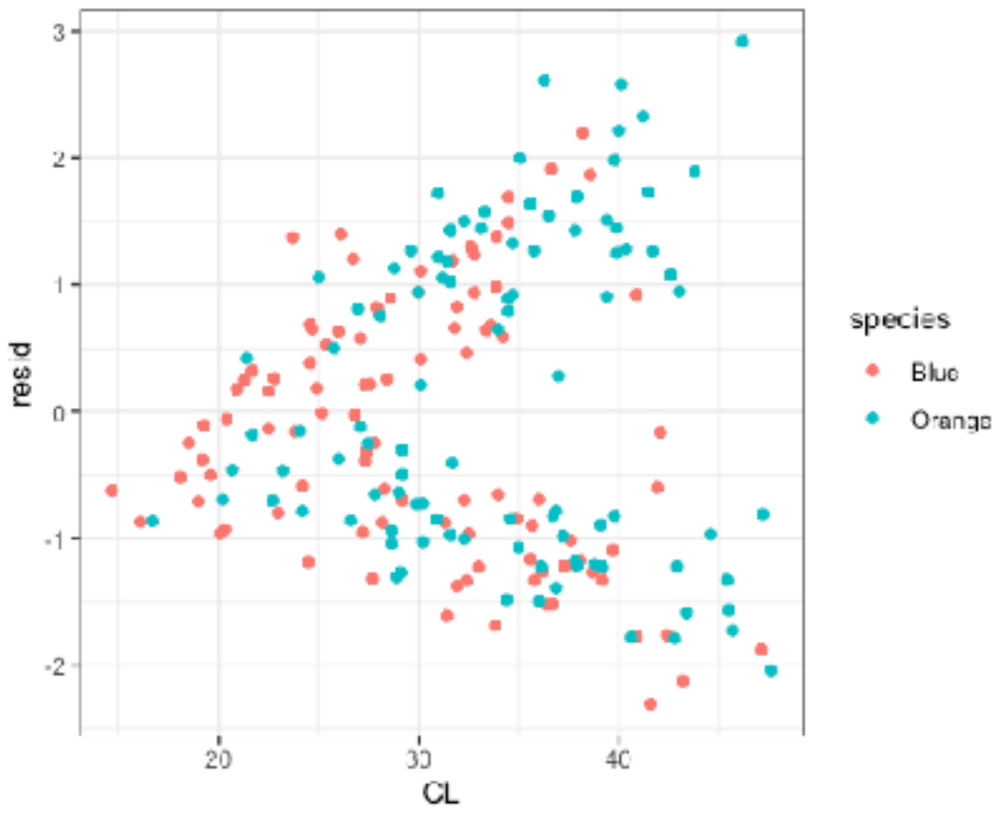
YOUR TURN SOLUTION - ROCK CRABS

```
m1 <- lm(RW~CL, data = crabs)

crabs$fitted <- fitted(m1)

crabs$resid <- resid(m1)
```

```
crabs %>%

  ggplot(aes(x = CL, y = resid)) +
  geom_point() +
  theme_bw()
```




```
crabs %>%

  ggplot(aes(x = CL, y = resid)) +
  geom_point(aes(colour = species)) +

  theme_bw()
```



MULTIPLE COVARIATES IN NORMAL MODELS

- For continuous RVs X_1, …, X_k the extension of the linear model is straightforward:

$$Y = a_1X_1 + a_2X_2 + \ldots + a_kX_k + \varepsilon$$

- no real problem - model is just bigger, estimates are solution from minimizing $Q(a_1, \ldots, a_k, b)$

- Ordinal discrete RVs, where we suspect a linear relationship with Y, can be treated the same as continuous RVs.

- But what about discrete variables such as gender or species?

FACTOR VARIABLES IN NORMAL MODELS

- We can include the factor variable in the R code as an additive variable:

```
m2 <- lm(RW ~ CL + sex, data = crabs)

summary(m2)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.903304   0.196005   14.81   <2e-16 ***
CL           0.337490   0.005955   56.67   <2e-16 ***
sexMale     -2.000198   0.084581  -23.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5948 on 197 degrees of freedom
Multiple R-squared:  0.9471,  Adjusted R-squared:  0.9466
```

Just a single effect for a binary factor

Female is the reference category

CATEGORICAL VARIABLES IN LINEAR MODELS

- Let X be a discrete RV with J levels.
- We create J dummy variables d_1 to d_J as $d_i = I(X = j)$ (indicator variable is 1 if X is level j and 0 otherwise)
- Model:

$$Y = a_2d_2 + \ldots + a_Jd_J + \varepsilon$$

a_1d_1 are left out! d1 is used as reference level and a1 is set to 0

- R is creating these dummy variables on the fly - we don't have to do anything!
- For binary variables such as gender or species only one parameter estimate will show up (the other one is set to 0 as a reference)

FACTOR VARIABLES IN NORMAL MODELS

- We can include the factor variable in the R code as an additive variable:

```
m2 <- lm(RW ~ CL + sex, data = crabs)

summary(m2)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.903304   0.196005   14.81   <2e-16 ***
CL           0.337490   0.005955   56.67   <2e-16 ***
sexMale     -2.000198   0.084581  -23.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5948 on 197 degrees of freedom
Multiple R-squared:  0.9471,  Adjusted R-squared:  0.9466
```

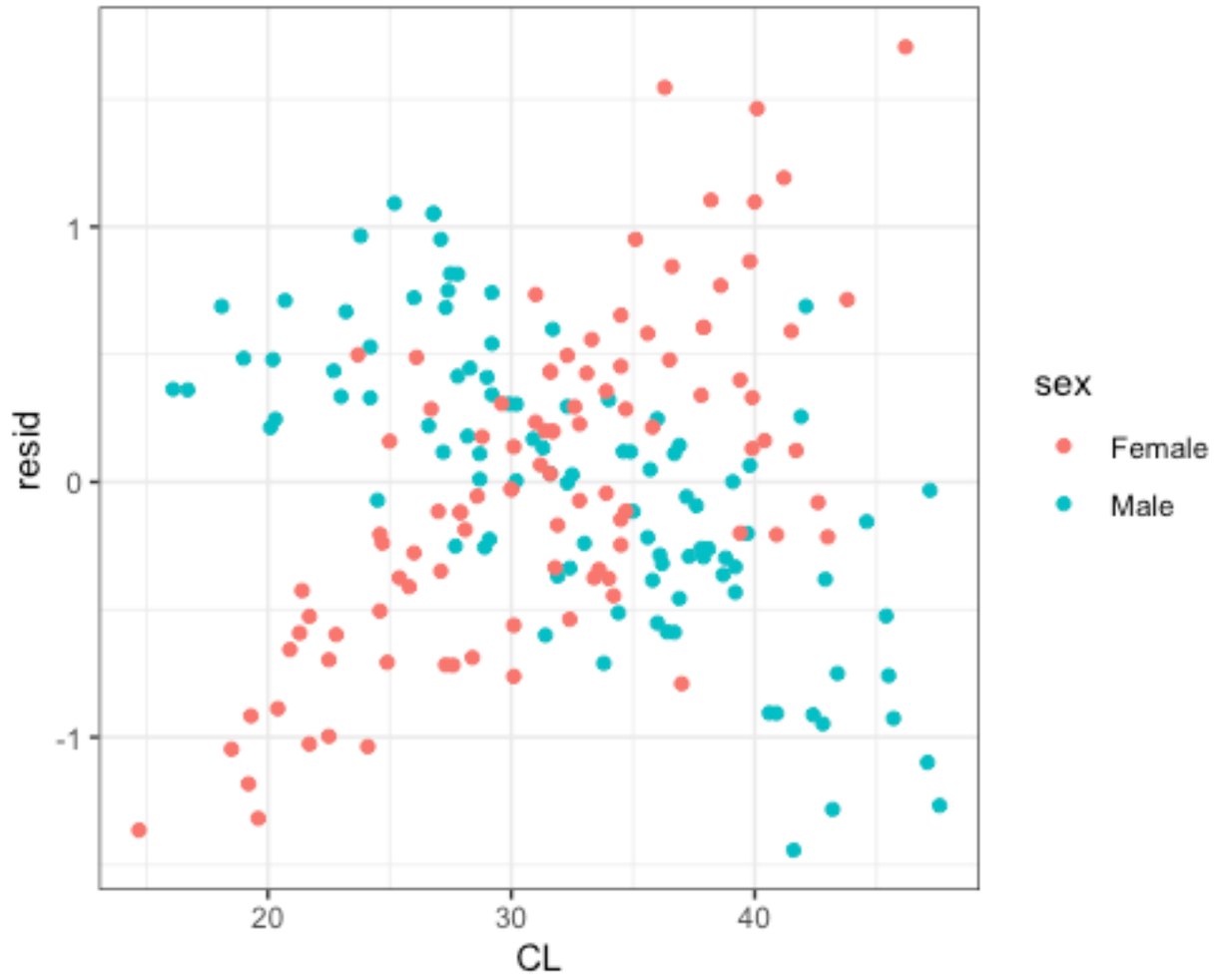
Just a single effect for a binary factor

Female is the reference category

THE RESIDUAL PLOT STILL HAS UNWANTED STRUCTURE!

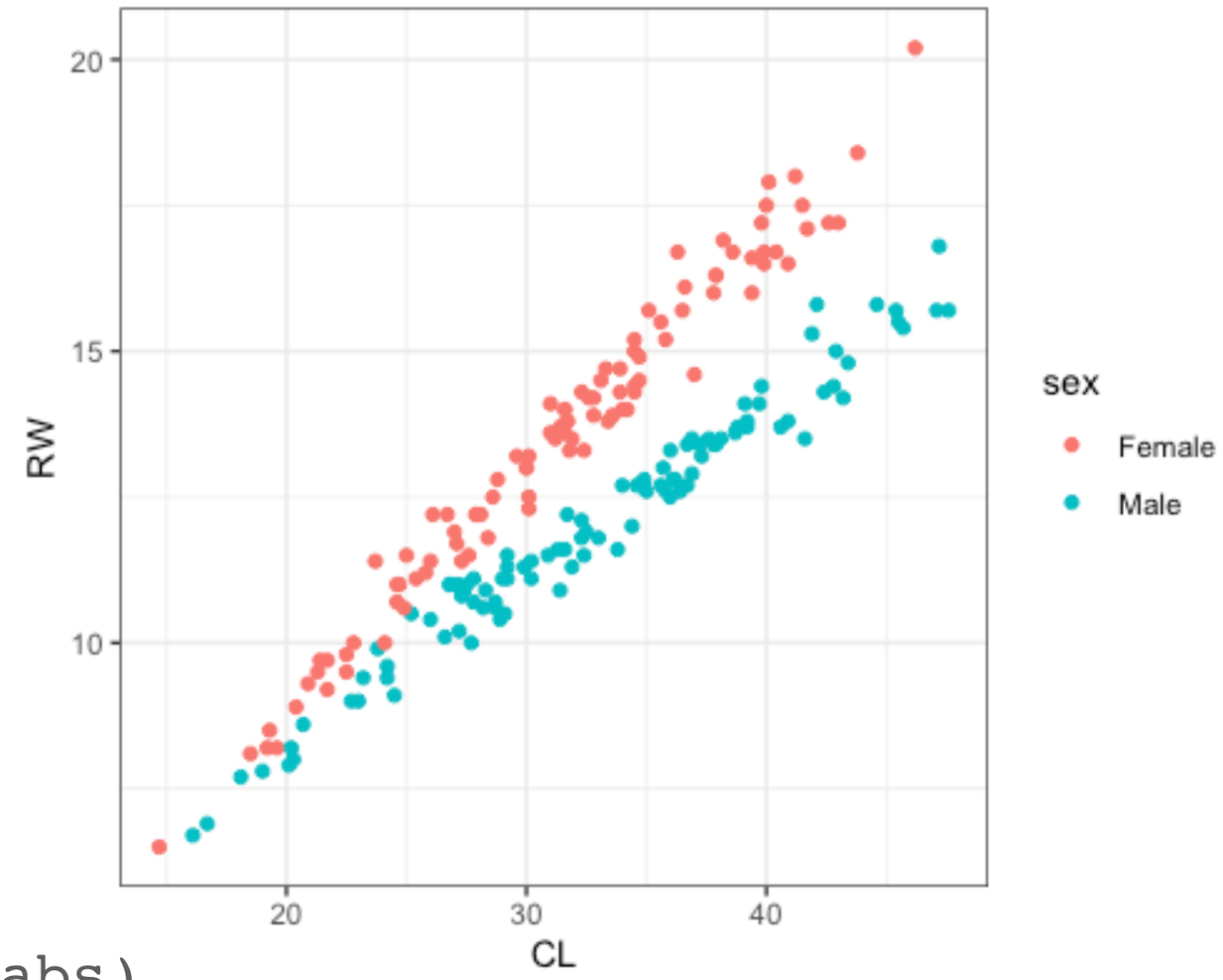
```
m2 <- lm(RW ~ CL + sex, data = crabs)
crabs$fitted <- fitted(m2)
crabs$resid <- resid(m2)

crabs %>%
ggplot(aes(x = CL, y = resid)) +
geom_point(aes(colour = species)
theme_bw()
```

There are different slopes in CL for the different genders



FITTING AN INTERACTION TERM

- X1 * X2 denotes interaction effect and all main effects
- X1:X2 denotes interaction effect only




```
m3 <- lm(RW ~ CL * sex, data = crabs)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.836288   0.202594   4.128 5.41e-05 ***
CL           0.403403   0.006319  63.840  < 2e-16 ***
sexMale      1.809718   0.278398   6.500 6.49e-10 ***
CL:sexMale  -0.118967   0.008489 -14.014  < 2e-16 ***
```

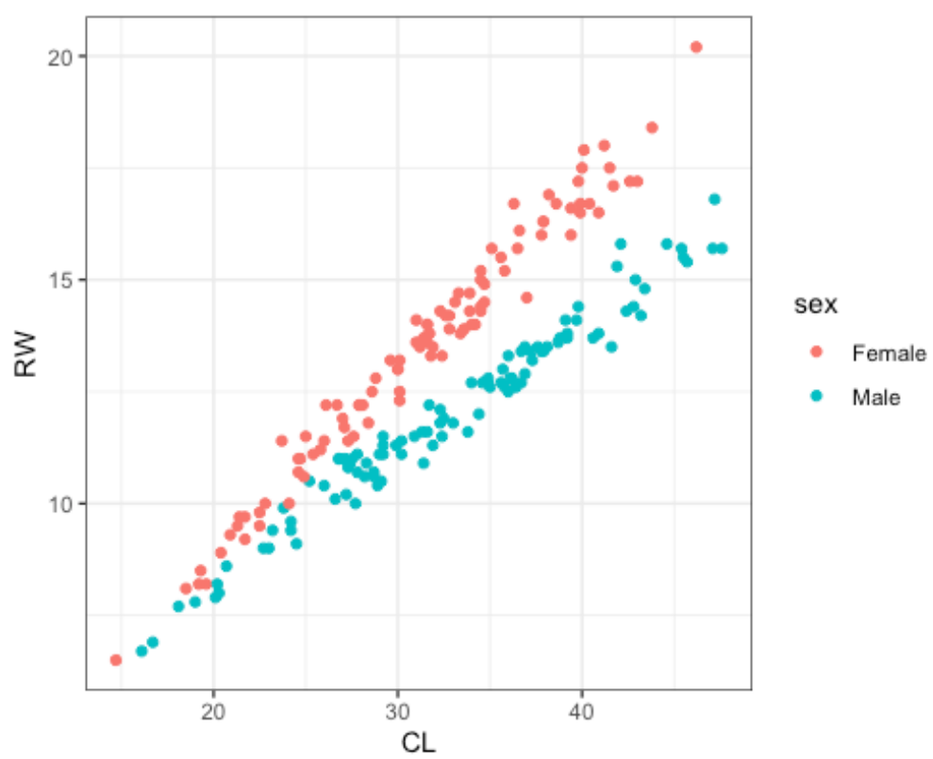
YOUR TURN

- Run the line to read the data:
  `crabs <- read.csv("http://ggobi.org/book/data/australian-crabs.csv")`
- FL is frontal lip
  RW is rear width
  CL is carapace length
  CW is carapace width

Work on the questions by yourself

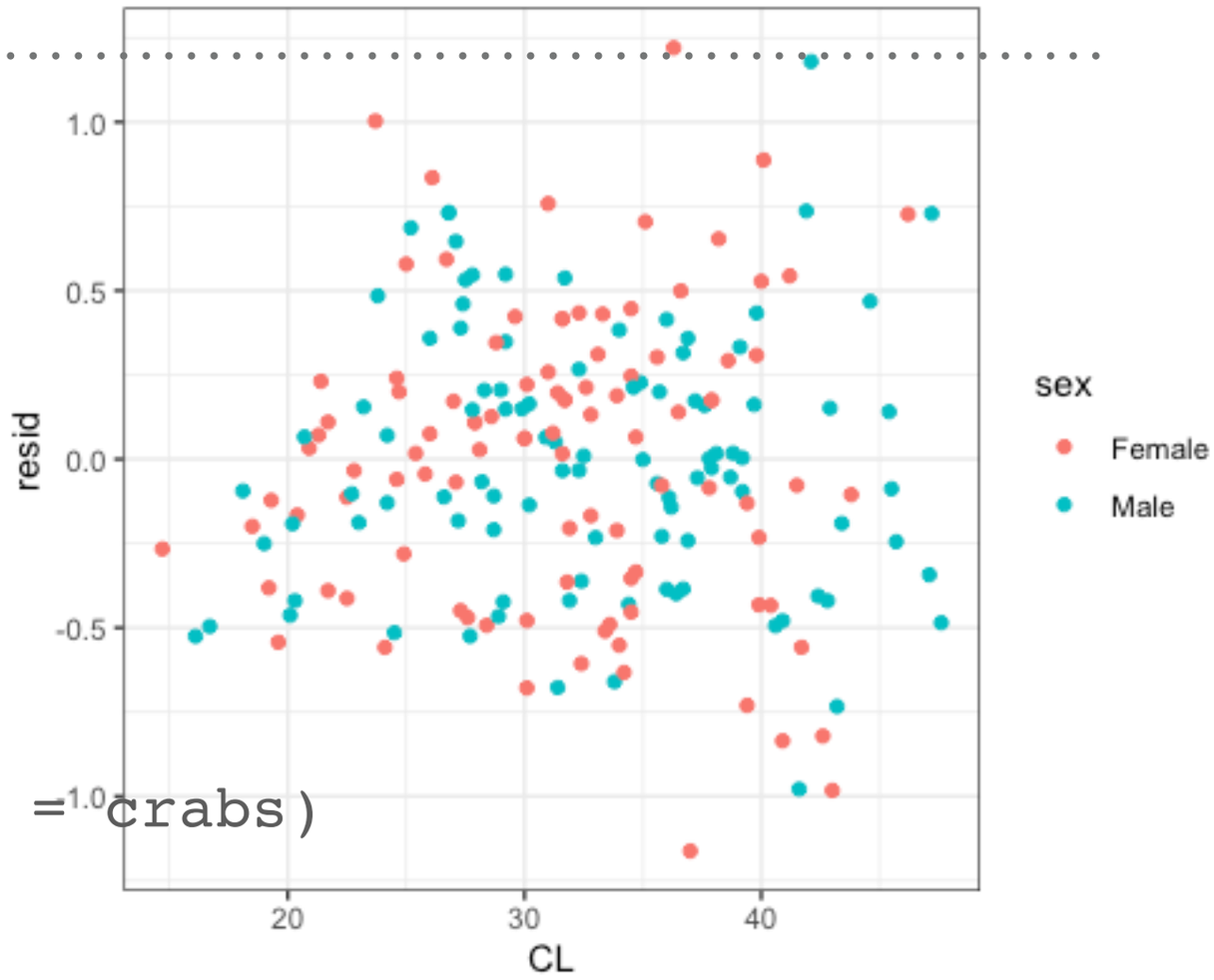
ROCK CRABS AGAIN

- Fit a linear model of rear width in carapace length and sex of the crab
- Find a model specification that allows you to directly compare the slopes of the two lines in the scatterplot below
- Make sure that the fitted values are the same for this model and model m3



YOUR TURN SOLUTION – ROCK CRABS

- use CL:sex to describe the interaction between CL and sex




```
m3b <- lm(RW ~ sex + CL:sex, data = crabs)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.836288   0.202594   4.128 5.41e-05 ***
sexMale      1.809718   0.278398   6.500 6.49e-10 ***
sexFemale:CL 0.403403   0.006319  63.840  < 2e-16 ***
sexMale:CL   0.284436   0.005669  50.172  < 2e-16 ***
```

PREDICTIONS

using models

PREDICTION OF MEANS AND SINGLE VALUES

- We use linear models to make two kinds of predictions:
  - mean responses
  - prediction of a future observation

- Mean response:
  what would be the **expected rear width** of a rock crab with a carapace length of x?
- Prediction of a future observation:
  what would be the **rear width** of a rock crab with a carapace length of x?

ROCK CRABS – MODEL MEANS

- Rear width of a crab is now determined as a function of the crab's sex and carapace length as
  $RW = 0.84 + 1.81\ I(male) + 0.4 \cdot CL \cdot I(female) + 0.28 \cdot CL \cdot I(male)$
- Based on this formula, what is the expected rear width of a male crab with carapace length of 35mm?
- $RW = 0.8363 + 1.8097 + 0.2844 \cdot 35 = 12.6$ (mm)
- Prediction intervals?
  `predict(model, newdata, interval)`

ROCK CRABS - MODEL MEANS

➤ predict(model, newdata, interval)

```
new.data = data.frame(sex="Male", CL = 35)

predict(m3b, newdata = new.data, interval = "confidence")
       fit      lwr      upr
1 12.60125 12.51474 12.68777
```

interval for the mean response

```
predict(m3b, newdata = new.data, interval = “prediction")
       fit      lwr      upr
1 12.60125 11.76563 13.43687
```

wider, because of the added uncertainty of a single prediction

QUESTIONS?

THANK YOU FOR YOUR ATTENTION!