STAC67: Regression Analysis

Lecture 16

Thibault Randrianarisoa

November 6, 2026

University of Toronto Scarborough

Multicollinearity

When the predictors are uncorrelated

Eight work crews: crew size X_1, bonus pay X_2, productivity Y. The predictors are uncorrelated below: $r(X_1, X_2) = 0$.

X_1	4	4	4	4	6	6	6	6
X_2	2	2	3	3	2	2	3	3
Y	42	39	48	51	49	53	61	60

When the predictors are uncorrelated

Eight work crews: crew size X_1, bonus pay X_2, productivity Y. The predictors are uncorrelated below: $r(X_1, X_2) = 0$.

X_1	4	4	4	4	6	6	6	6
X_2	2	2	3	3	2	2	3	3
Y	42	39	48	51	49	53	61	60

Model	$\hat{\beta}_1$	$\hat{\beta}_2$
$Y \sim X_1$	5.375	
$Y \sim X_2$		9.250
$Y \sim X_1 + X_2$	5.375	9.250

When the predictors are uncorrelated

Eight work crews: crew size X_1, bonus pay X_2, productivity Y. The predictors are uncorrelated below: $r(X_1, X_2) = 0$.

X_1	4	4	4	4	6	6	6	6
X_2	2	2	3	3	2	2	3	3
Y	42	39	48	51	49	53	61	60

Model	$\hat{\beta}_1$	$\hat{\beta}_2$
$Y \sim X_1$	5.375	
$Y \sim X_2$		9.250
$Y \sim X_1 + X_2$	5.375	9.250

The slopes do not change, and neither do the extra sums of squares: $SSR(X_1) = SSR(X_1 \mid X_2) = 231.125$ and $SSR(X_2) = SSR(X_2 \mid X_1) = 171.125$. The order no longer matters.

When the predictors are correlated

In the acceleration data, the four predictors move together:

```
       wt   hp disp  cyl
wt   1.00 0.86 0.93 0.90
hp   0.86 1.00 0.90 0.84
disp 0.93 0.90 1.00 0.95
cyl  0.90 0.84 0.95 1.00
```

When the predictors are correlated

In the acceleration data, the four predictors move together:

```
       wt   hp disp  cyl
wt   1.00 0.86 0.93 0.90
hp   0.86 1.00 0.90 0.84
disp 0.93 0.90 1.00 0.95
cyl  0.90 0.84 0.95 1.00
```

The coefficient of weight depends on which other predictors are in the model:

Model	$\hat{\beta}_{\text{wt}}$	$s\{\hat{\beta}_{\text{wt}}\}$
`acc ~ wt`	-0.00136	0.00015
`acc ~ wt + hp`	0.00229	0.00021
`acc ~ wt + hp + disp`	0.00306	0.00029
`acc ~ wt + hp + disp + cyl`	0.00309	0.00029

When the predictors are correlated

In the acceleration data, the four predictors move together:

```
       wt   hp disp  cyl
wt   1.00 0.86 0.93 0.90
hp   0.86 1.00 0.90 0.84
disp 0.93 0.90 1.00 0.95
cyl  0.90 0.84 0.95 1.00
```

The coefficient of weight depends on which other predictors are in the model:

Model	$\hat{\beta}_{\text{wt}}$	$s\{\hat{\beta}_{\text{wt}}\}$
`acc ~ wt`	-0.00136	0.00015
`acc ~ wt + hp`	0.00229	0.00021
`acc ~ wt + hp + disp`	0.00306	0.00029
`acc ~ wt + hp + disp + cyl`	0.00309	0.00029

The sign changes once horsepower is in (Lecture 10), and the standard error doubles as more correlated predictors are added.

Why the variances inflate

With two predictors, the variance of a slope is (stated without proof)

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{11}\,(1 - r_{12}^2)}, \qquad S_{11} = \sum_i (X_{i1} - \bar{X}_1)^2,$$

where r_{12} is the correlation between the two predictors. With one predictor, it was σ^2 / S_{11}.

Why the variances inflate

With two predictors, the variance of a slope is (stated without proof)

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{11}\,(1 - r_{12}^2)}, \qquad S_{11} = \sum_i (X_{i1} - \bar{X}_1)^2,$$

where r_{12} is the correlation between the two predictors. With one predictor, it was σ^2 / S_{11}.

The factor $1/(1 - r_{12}^2)$ is 1 when $r_{12} = 0$, and explodes as $|r_{12}| \to 1$.

Displacement and cylinders: $r = 0.951$, a factor of 10.4 on the variance:

	`acc ~ disp`	`acc ~ disp + cyl`
$s\{\hat{\beta}_{\text{disp}}\}$	0.0011	0.0036

The standard error is multiplied by about $\sqrt{10.4} \approx 3.2$.

Why the variances inflate

With two predictors, the variance of a slope is (stated without proof)

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{11}\,(1 - r_{12}^2)}, \qquad S_{11} = \sum_i (X_{i1} - \bar{X}_1)^2,$$

where r_{12} is the correlation between the two predictors. With one predictor, it was σ^2 / S_{11}.

The factor $1/(1 - r_{12}^2)$ is 1 when $r_{12} = 0$, and explodes as $|r_{12}| \to 1$.

Displacement and cylinders: $r = 0.951$, a factor of 10.4 on the variance:

	`acc ~ disp`	`acc ~ disp + cyl`
$s\{\hat{\beta}_{\text{disp}}\}$	0.0011	0.0036

The standard error is multiplied by about $\sqrt{10.4} \approx 3.2$.

With more predictors, $1/(1 - R_k^2)$, where R_k^2 comes from regressing X_k on the other predictors, is the *variance inflation factor*, later in the course. For displacement among the four predictors, it is 19.3.

Significant alone, but not once the others are in

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20.0553	0.4061	49.3866	0
cyl	-0.8226	0.0708	-11.6248	0

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.8864	0.6018	29.7228	0.0000
disp	-0.0173	0.0036	-4.7785	0.0000
cyl	0.1884	0.2225	0.8467	0.3977

Significant alone, but not once the others are in

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20.0553	0.4061	49.3866	0
cyl	-0.8226	0.0708	-11.6248	0

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.8864	0.6018	29.7228	0.0000
disp	-0.0173	0.0036	-4.7785	0.0000
cyl	0.1884	0.2225	0.8467	0.3977

Alone, cylinders are strongly related to acceleration ($t = -11.6$). Once displacement is in the model, their coefficient is not significant ($t = 0.85$, $p = 0.40$), and its sign even flips.

Significant alone, but not once the others are in

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20.0553	0.4061	49.3866	0
cyl	-0.8226	0.0708	-11.6248	0

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.8864	0.6018	29.7228	0.0000
disp	-0.0173	0.0036	-4.7785	0.0000
cyl	0.1884	0.2225	0.8467	0.3977

Alone, cylinders are strongly related to acceleration ($t = -11.6$). Once displacement is in the model, their coefficient is not significant ($t = 0.85$, $p = 0.40$), and its sign even flips.

The t test asks whether cylinders add anything once displacement is known. With $r = 0.95$, they carry almost the same information, so the answer is: very little.

What multicollinearity does, and does not do

It does not violate any assumption: $\hat{\beta}$ is still unbiased, and, unless the dependence is exact, $\boldsymbol{X}'\boldsymbol{X}$ can still be inverted. Fitted values and predictions inside the range of the data stay valid.

What multicollinearity does, and does not do

It does not violate any assumption: $\hat{\beta}$ is still unbiased, and, unless the dependence is exact, $\boldsymbol{X}'\boldsymbol{X}$ can still be inverted. Fitted values and predictions inside the range of the data stay valid.

It does:

- inflate the standard errors of the coefficients;
- make the coefficients unstable: they change, and can change sign, depending on which predictors are in the model;
- blur the meaning of "the others held fixed", when the predictors rarely move separately;
- make the sequential sums of squares depend heavily on the order;
- allow every individual t test to be non-significant while the overall F test rejects.

What multicollinearity does, and does not do

It does not violate any assumption: $\hat{\beta}$ is still unbiased, and, unless the dependence is exact, $\boldsymbol{X}'\boldsymbol{X}$ can still be inverted. Fitted values and predictions inside the range of the data stay valid.

It does:

- inflate the standard errors of the coefficients;
- make the coefficients unstable: they change, and can change sign, depending on which predictors are in the model;
- blur the meaning of "the others held fixed", when the predictors rarely move separately;
- make the sequential sums of squares depend heavily on the order;
- allow every individual t test to be non-significant while the overall F test rejects.

Remedies, later in the course: drop or combine predictors, or use ridge regression.

Qualitative predictors

Coding a qualitative predictor

Many predictors are qualitative (categorical): the type of a company, a treatment group, a region. They enter the model through indicator variables (dummy variables):

$$D_i = \begin{cases} 1 & \text{if observation } i \text{ is in the group,} \\ 0 & \text{otherwise.} \end{cases}$$

Coding a qualitative predictor

Many predictors are qualitative (categorical): the type of a company, a treatment group, a region. They enter the model through indicator variables (dummy variables):

$$D_i = \begin{cases} 1 & \text{if observation } i \text{ is in the group,} \\ 0 & \text{otherwise.} \end{cases}$$

With a single binary predictor, the model is

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i : \qquad E(Y_i) = \begin{cases} \beta_0 & \text{if } D_i = 0, \\ \beta_0 + \beta_1 & \text{if } D_i = 1. \end{cases}$$

β_0 is the mean response of the group coded 0, and β_1 is the difference between the two group means.

Coding a qualitative predictor

Many predictors are qualitative (categorical): the type of a company, a treatment group, a region. They enter the model through indicator variables (dummy variables):

$$D_i = \begin{cases} 1 & \text{if observation } i \text{ is in the group,} \\ 0 & \text{otherwise.} \end{cases}$$

With a single binary predictor, the model is

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i : \qquad E(Y_i) = \begin{cases} \beta_0 & \text{if } D_i = 0, \\ \beta_0 + \beta_1 & \text{if } D_i = 1. \end{cases}$$

β_0 is the mean response of the group coded 0, and β_1 is the difference between the two group means.

Least squares gives $\hat{\beta}_0 = \bar{Y}_0$ and $\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0$, the sample means of the groups. The t test of $H_0 : \beta_1 = 0$ is the two-sample t test with a pooled variance.

Example: how fast do insurance firms innovate?

For 20 insurance firms: the number of months before the firm adopted an innovation (`months`), and whether it is a stock company ($D = 1$, 10 firms) or a mutual company ($D = 0$, 10 firms).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.7	2.92	5.719	0.000
stock	5.4	4.13	1.308	0.207

Example: how fast do insurance firms innovate?

For 20 insurance firms: the number of months before the firm adopted an innovation (`months`), and whether it is a stock company ($D = 1$, 10 firms) or a mutual company ($D = 0$, 10 firms).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.7	2.92	5.719	0.000
stock	5.4	4.13	1.308	0.207

$\hat{\beta}_0 = 16.7$ months is the mean of the mutual firms, and $\hat{\beta}_1 = 5.4$ months is the difference: stock firms took longer, on average.

Example: how fast do insurance firms innovate?

For 20 insurance firms: the number of months before the firm adopted an innovation (`months`), and whether it is a stock company ($D = 1$, 10 firms) or a mutual company ($D = 0$, 10 firms).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.7	2.92	5.719	0.000
stock	5.4	4.13	1.308	0.207

$\hat{\beta}_0 = 16.7$ months is the mean of the mutual firms, and $\hat{\beta}_1 = 5.4$ months is the difference: stock firms took longer, on average.

But $t = 1.31$ ($p = 0.21$): on its own, the type of firm shows no significant effect. We come back to this after adding the size of the firm.

A factor with k levels

A qualitative predictor with k levels needs $k-1$ indicators. The level without an indicator is the baseline: its mean is β_0, and each coefficient is the difference between a level and the baseline:

$$Y_i = \beta_0 + \beta_1 D_{i1} + \cdots + \beta_{k-1} D_{i,k-1} + \epsilon_i,$$

$$E(Y_i) = \begin{cases} \beta_0 & \text{at the baseline,} \\ \beta_0 + \beta_j & \text{at the level of } D_j. \end{cases}$$

A factor with k levels

A qualitative predictor with k levels needs $k-1$ indicators. The level without an indicator is the baseline: its mean is β_0, and each coefficient is the difference between a level and the baseline:

$$Y_i = \beta_0 + \beta_1 D_{i1} + \cdots + \beta_{k-1} D_{i,k-1} + \epsilon_i,$$

$$E(Y_i) = \begin{cases} \beta_0 & \text{at the baseline,} \\ \beta_0 + \beta_j & \text{at the level of } D_j. \end{cases}$$

Why not k indicators? With an intercept, the k indicators add up to the column of 1s: $D_{i1} + \cdots + D_{ik} = 1$ for every i. The columns of $\boldsymbol{X}$ are linearly dependent and $\boldsymbol{X}'\boldsymbol{X}$ cannot be inverted (Problem Set 5, Exercise 3).

A factor with k levels

A qualitative predictor with k levels needs $k-1$ indicators. The level without an indicator is the baseline: its mean is β_0, and each coefficient is the difference between a level and the baseline:

$$Y_i = \beta_0 + \beta_1 D_{i1} + \cdots + \beta_{k-1} D_{i,k-1} + \epsilon_i,$$

$$E(Y_i) = \begin{cases} \beta_0 & \text{at the baseline,} \\ \beta_0 + \beta_j & \text{at the level of } D_j. \end{cases}$$

Why not k indicators? With an intercept, the k indicators add up to the column of 1s: $D_{i1} + \cdots + D_{ik} = 1$ for every i. The columns of $\boldsymbol{X}$ are linearly dependent and $\boldsymbol{X}'\boldsymbol{X}$ cannot be inverted (Problem Set 5, Exercise 3).

The choice of baseline does not change the fitted values; it only changes which differences the coefficients report. R takes the first level in alphabetical order.

Example: plant weights in three groups

Dried weights of 30 plants: a control group and two treatments, 10 plants each (`PlantGrowth`, in `R`).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.032	0.197	25.527	0.000
grouptrt1	-0.371	0.279	-1.331	0.194
grouptrt2	0.494	0.279	1.772	0.088

Example: plant weights in three groups

Dried weights of 30 plants: a control group and two treatments, 10 plants each (`PlantGrowth`, in R).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.032	0.197	25.527	0.000
grouptrt1	-0.371	0.279	-1.331	0.194
grouptrt2	0.494	0.279	1.772	0.088

Baseline: `ctrl`, with mean 5.032. The coefficients are the differences `trt1` $-$ `ctrl` $= -0.371$ and `trt2` $-$ `ctrl` $= 0.494$.

Example: plant weights in three groups

Dried weights of 30 plants: a control group and two treatments, 10 plants each (`PlantGrowth`, in R).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.032	0.197	25.527	0.000
grouptrt1	-0.371	0.279	-1.331	0.194
grouptrt2	0.494	0.279	1.772	0.088

Baseline: `ctrl`, with mean 5.032. The coefficients are the differences `trt1` $-$ `ctrl` $= -0.371$ and `trt2` $-$ `ctrl` $= 0.494$.

Are the three group means equal? $H_0 : \beta_1 = \beta_2 = 0$ is the overall F test of this model: $F^* = 4.85$ on 2 and 27 df, $p = 0.016$. We reject: the group means are not all equal (here `trt1` and `trt2` lie on either side of the control).

This is the one-way analysis of variance.

Why not code the levels $1, 2, \ldots, k$?

Back to the plants: code the groups `ctrl`, `trt1` and `trt2` as $L = 1, 2, 3$ in one column:

$$Y_i = \beta_0 + \beta_1 L_i + \epsilon_i : \qquad E(Y_i) = \beta_0 + \beta_1, \quad \beta_0 + 2\beta_1, \quad \beta_0 + 3\beta_1.$$

Going from one group to the next always adds the same amount β_1: the levels are forced to be ordered and equally spaced.

Why not code the levels $1, 2, \ldots, k$?

Back to the plants: code the groups `ctrl`, `trt1` and `trt2` as $L = 1, 2, 3$ in one column:

$$Y_i = \beta_0 + \beta_1 L_i + \epsilon_i : \qquad E(Y_i) = \beta_0 + \beta_1, \quad \beta_0 + 2\beta_1, \quad \beta_0 + 3\beta_1.$$

Going from one group to the next always adds the same amount β_1: the levels are forced to be ordered and equally spaced.

For these data:

	`ctrl`	`trt1`	`trt2`
Sample means	5.03	4.66	5.53
Fitted means, codes $1, 2, 3$	4.83	5.07	5.32
Fitted means, codes $2, 1, 3$	5.07	4.64	5.51

Why not code the levels $1, 2, \ldots, k$?

Back to the plants: code the groups `ctrl`, `trt1` and `trt2` as $L = 1, 2, 3$ in one column:

$$Y_i = \beta_0 + \beta_1 L_i + \epsilon_i : \qquad E(Y_i) = \beta_0 + \beta_1, \quad \beta_0 + 2\beta_1, \quad \beta_0 + 3\beta_1.$$

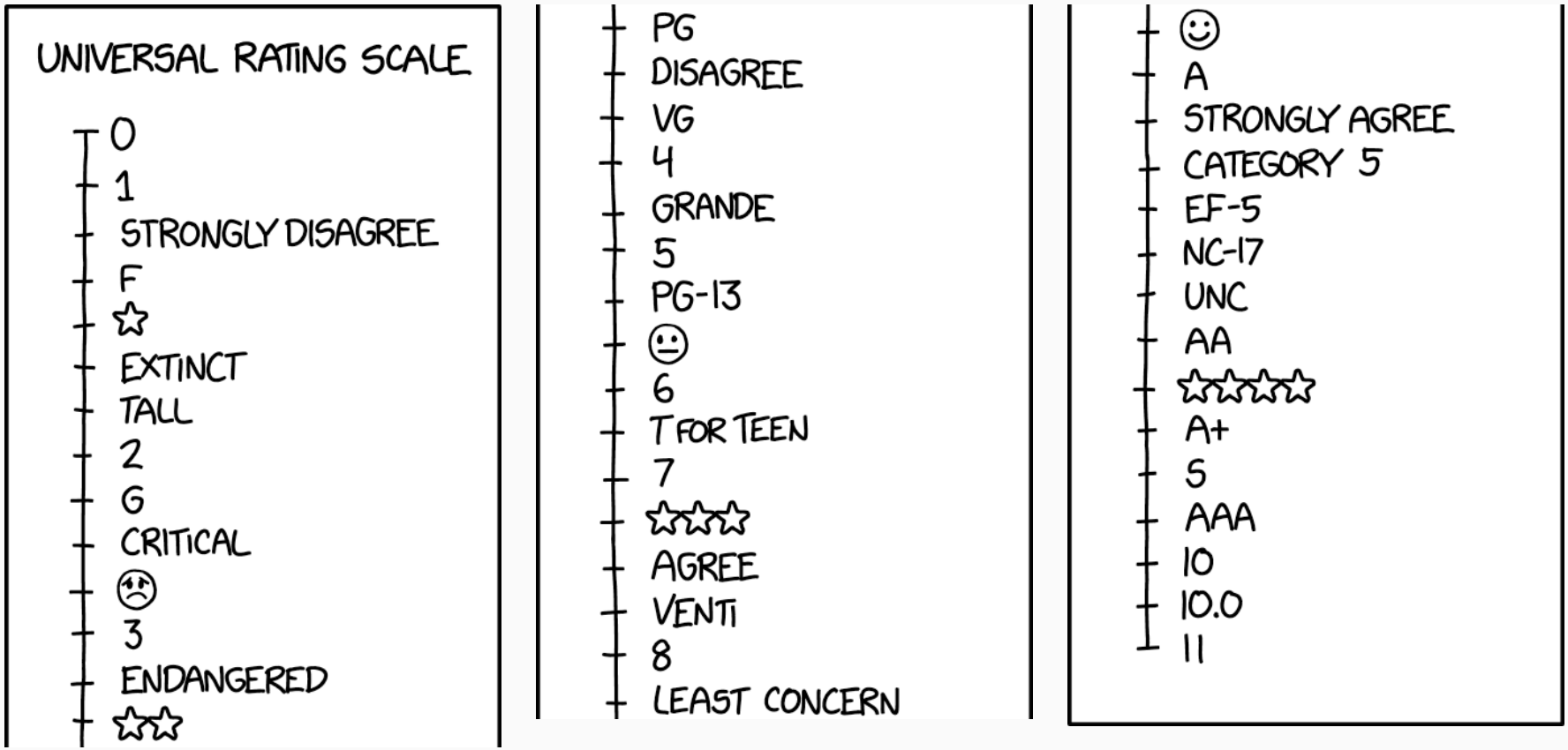
Going from one group to the next always adds the same amount β_1: the levels are forced to be ordered and equally spaced.

For these data:

	`ctrl`	`trt1`	`trt2`
Sample means	5.03	4.66	5.53
Fitted means, codes $1, 2, 3$	4.83	5.07	5.32
Fitted means, codes $2, 1, 3$	5.07	4.64	5.51

The fit misses the means, and it changes when the labels, which are arbitrary, are swapped. With indicators, $E(Y_i) = \beta_0,\ \beta_0 + \beta_1,\ \beta_0 + \beta_2$: each level has its own mean, and the fitted means are the sample means, whatever the labels.

One scale for everything?



Coding the levels $1, 2, \ldots, k$ imposes an order, and equal gaps, that the categories do not have.

xkcd.com/2329 (CC BY-NC 2.5)

One quantitative and one qualitative predictor

Parallel lines

With a quantitative predictor X and an indicator D:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \epsilon_i.$$

Parallel lines

With a quantitative predictor X and an indicator D:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \epsilon_i.$$

Group	Mean response	
$D = 0$	$E(Y) = \beta_0 + \beta_1 X$	intercept β_0, slope β_1
$D = 1$	$E(Y) = (\beta_0 + \beta_2) + \beta_1 X$	intercept $\beta_0 + \beta_2$, slope β_1

Two lines with the same slope: they are parallel with a vertical distance β_2 apart.

Parallel lines

With a quantitative predictor X and an indicator D:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \epsilon_i.$$

Group	Mean response	
$D = 0$	$E(Y) = \beta_0 + \beta_1 X$	intercept β_0, slope β_1
$D = 1$	$E(Y) = (\beta_0 + \beta_2) + \beta_1 X$	intercept $\beta_0 + \beta_2$, slope β_1

Two lines with the same slope: they are parallel with a vertical distance β_2 apart.

- β_1: change in the mean response per unit of X, in either group.
- β_2: difference between the mean responses of the two groups at the same value of X.
- $H_0 : \beta_2 = 0$ (a t test): one line for both groups.

The insurance firms, with their size

Add the size of the firm (total assets, in millions of dollars):

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.8741	1.8139	18.6751	0
size	-0.1017	0.0089	-11.4430	0
stock	8.0555	1.4591	5.5208	0

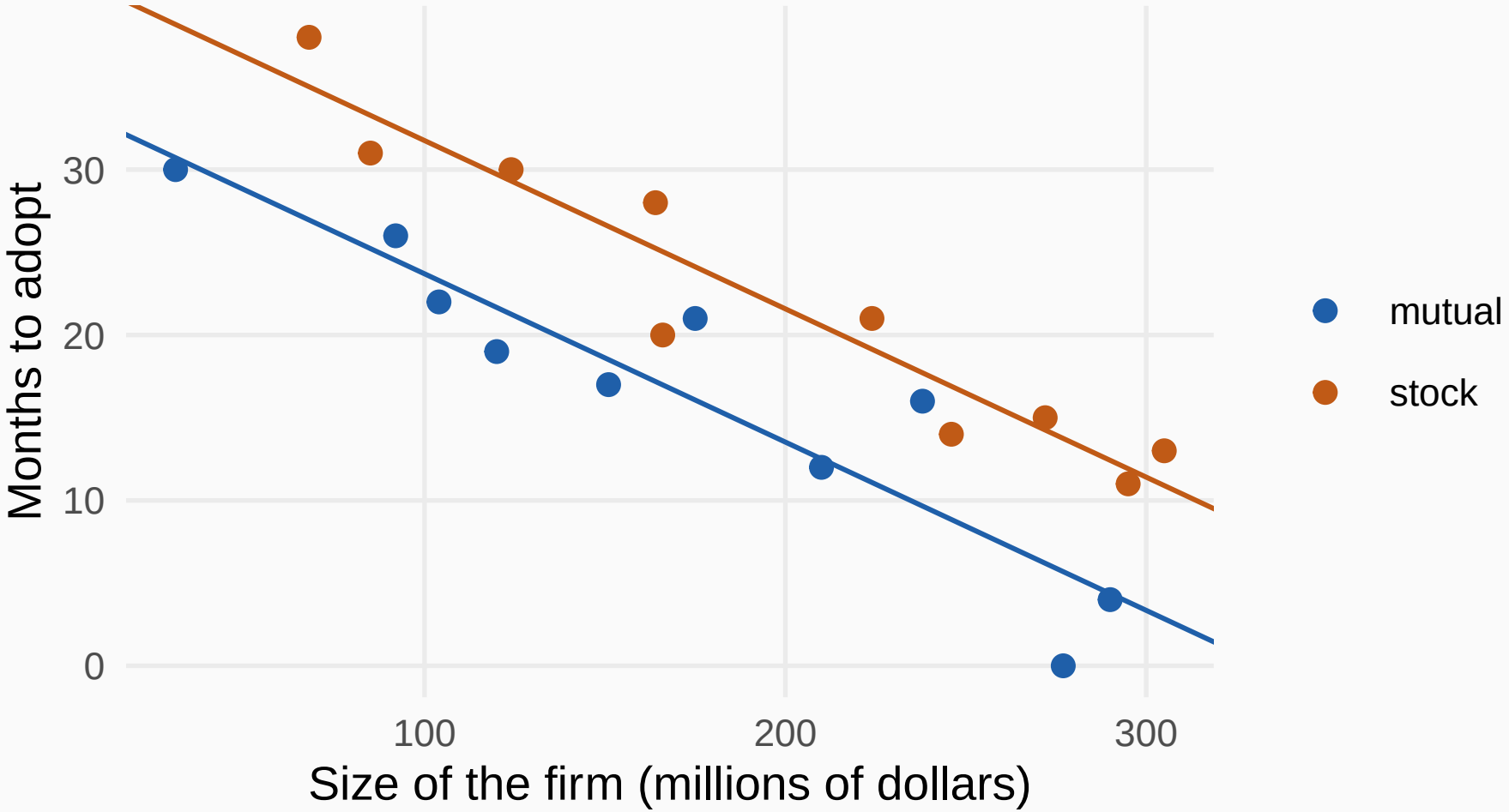
The insurance firms, with their size

Add the size of the firm (total assets, in millions of dollars):

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.8741	1.8139	18.6751	0
size	-0.1017	0.0089	-11.4430	0
stock	8.0555	1.4591	5.5208	0

At a given size, stock firms take about 8.1 months longer than mutual firms ($t = 5.52$, $p < 0.001$). In both groups, each extra million dollars of assets shortens the delay by 0.102 months.

The two fitted lines



Mutual: $\hat{Y} = 33.87 - 0.1017\,X$. Stock: $\hat{Y} = 41.93 - 0.1017\,X$.

Why the type of firm mattered only once size was in

Model	$\hat{\beta}$ for `stock`	$s\{\hat{\beta}\}$	$\sqrt{MSE}$
`months` $\sim$ `stock`	5.40	4.13	9.23
`months` $\sim$ `size + stock`	8.06	1.46	3.22

Why the type of firm mattered only once size was in

Model	$\hat{\beta}$ for `stock`	$s\{\hat{\beta}\}$	$\sqrt{MSE}$
`months` $\sim$ `stock`	5.40	4.13	9.23
`months` $\sim$ `size + stock`	8.06	1.46	3.22

- Size explains most of the variation in the delay, so $\sqrt{MSE}$ drops from 9.2 to 3.2 months, and the standard error with it.
- Stock firms in the sample are larger (195 against 169 million on average), and larger firms innovate sooner. Comparing the raw means mixes the two effects; the model compares firms of the same size.

Why the type of firm mattered only once size was in

Model	$\hat{\beta}$ for `stock`	$s\{\hat{\beta}\}$	$\sqrt{MSE}$
`months` $\sim$ `stock`	5.40	4.13	9.23
`months` $\sim$ `size + stock`	8.06	1.46	3.22

- Size explains most of the variation in the delay, so $\sqrt{MSE}$ drops from 9.2 to 3.2 months, and the standard error with it.
- Stock firms in the sample are larger (195 against 169 million on average), and larger firms innovate sooner. Comparing the raw means mixes the two effects; the model compares firms of the same size.

The same idea as "holding the other predictors fixed" in Lecture 10, with a qualitative predictor.

With k groups

With a factor of k levels and one quantitative predictor:

$$Y_i = \beta_0 + \beta_1 X_i + \gamma_2 D_{i2} + \cdots + \gamma_k D_{ik} + \epsilon_i :$$

k parallel lines, one per level, with intercepts β_0 (baseline) and $\beta_0 + \gamma_j$.

With k groups

With a factor of k levels and one quantitative predictor:

$$Y_i = \beta_0 + \beta_1 X_i + \gamma_2 D_{i2} + \cdots + \gamma_k D_{ik} + \epsilon_i :$$

k parallel lines, one per level, with intercepts β_0 (baseline) and $\beta_0 + \gamma_j$.

"Do the groups differ, at a given X?" is $H_0 : \gamma_2 = \cdots = \gamma_k = 0$: $k - 1$ restrictions, tested with full and reduced models, as for the extra sums of squares:

$$F^* = \frac{\big(SSE(X) - SSE(X, D_2, \ldots, D_k)\big)/(k-1)}{MSE(X, D_2, \ldots, D_k)},$$

which follows an $F(k-1,\ n-k-1)$ distribution under H_0.

With k groups

With a factor of k levels and one quantitative predictor:

$$Y_i = \beta_0 + \beta_1 X_i + \gamma_2 D_{i2} + \cdots + \gamma_k D_{ik} + \epsilon_i :$$

k parallel lines, one per level, with intercepts β_0 (baseline) and $\beta_0 + \gamma_j$.

"Do the groups differ, at a given X?" is $H_0 : \gamma_2 = \cdots = \gamma_k = 0$: $k - 1$ restrictions, tested with full and reduced models, as for the extra sums of squares:

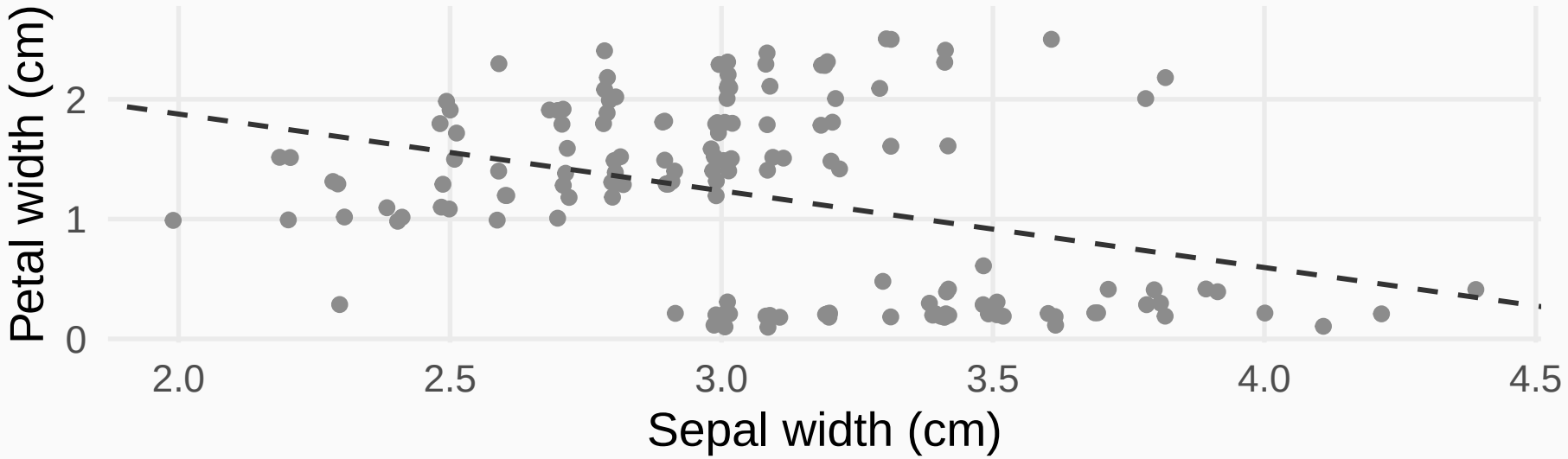
$$F^* = \frac{\big(SSE(X) - SSE(X, D_2, \ldots, D_k)\big)/(k-1)}{MSE(X, D_2, \ldots, D_k)},$$

which follows an $F(k-1,\ n-k-1)$ distribution under H_0.

In R: `anova(lm(y ~ x), lm(y ~ x + group))`, with `group` a factor.

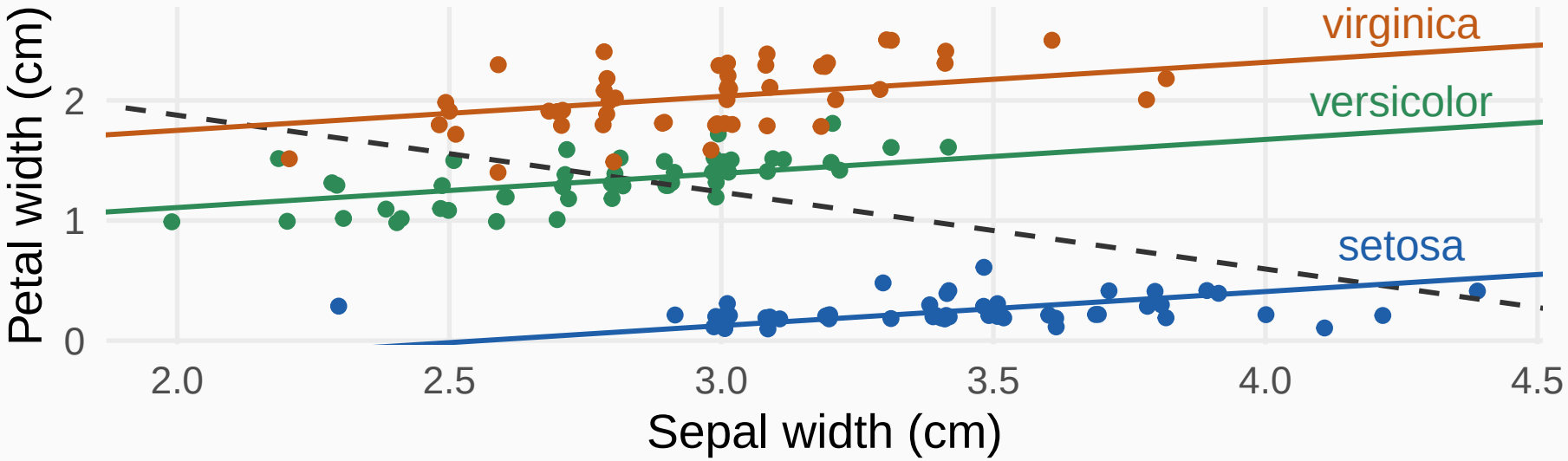
Simpson's paradox

150 iris flowers, 50 from each of three species (`iris`, in R):



Simpson's paradox

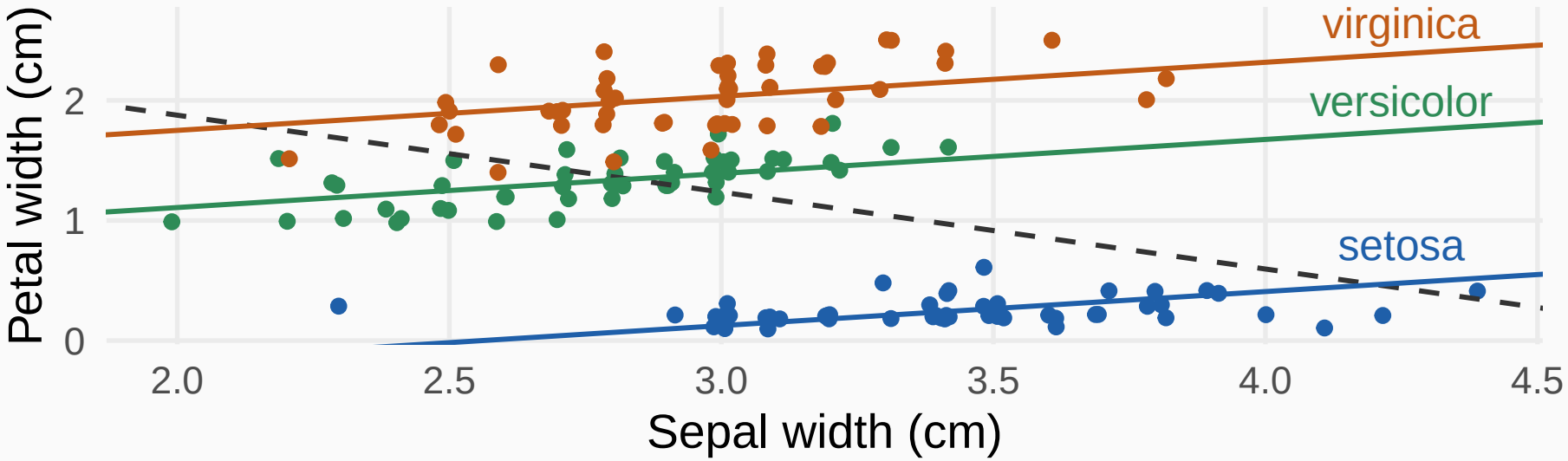
150 iris flowers, 50 from each of three species (`iris`, in R):



Ignoring the species (dashed): slope -0.64 ($t = -4.79$). With the species, three parallel lines: slope 0.28 ($t = 6.44$). Both are significant. Within a species, wider sepals go with wider petals.

Simpson's paradox

150 iris flowers, 50 from each of three species (`iris`, in R):



Ignoring the species (dashed): slope -0.64 ($t = -4.79$). With the species, three parallel lines: slope 0.28 ($t = 6.44$). Both are significant. Within a species, wider sepals go with wider petals.

This is Simpson's paradox: an association that reverses once a third variable is taken into account. Setosa flowers have the widest sepals but by far the narrowest petals, so the dashed line mixes the comparison within a species with the differences between species.

Interactions

When the slopes may differ

The parallel-lines model assumes that X has the same effect in both groups: the model is additive. An interaction term, the product $D \cdot X$, relaxes this:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 D_i X_i + \epsilon_i.$$

When the slopes may differ

The parallel-lines model assumes that X has the same effect in both groups: the model is additive. An interaction term, the product $D \cdot X$, relaxes this:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 D_i X_i + \epsilon_i.$$

Group	Mean response	
$D = 0$	$E(Y) = \beta_0 + \beta_1 X$	intercept β_0, slope β_1
$D = 1$	$E(Y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X$	intercept $\beta_0 + \beta_2$, slope $\beta_1 + \beta_3$

When the slopes may differ

The parallel-lines model assumes that X has the same effect in both groups: the model is additive. An interaction term, the product $D \cdot X$, relaxes this:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 D_i X_i + \epsilon_i.$$

Group	Mean response	
$D = 0$	$E(Y) = \beta_0 + \beta_1 X$	intercept β_0, slope β_1
$D = 1$	$E(Y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X$	intercept $\beta_0 + \beta_2$, slope $\beta_1 + \beta_3$

- β_3: how much the slope of X changes from group 0 to group 1.
- β_2: the difference between the groups at $X = 0$ only. With an interaction, the difference between the groups depends on X: it is $\beta_2 + \beta_3 X$.

Two tests

Should it be the same slope? $H_0 : \beta_3 = 0$, a t test on one coefficient. If it is not rejected, the parallel-lines model is adequate.

Two tests

Should it be the same slope? $H_0 : \beta_3 = 0$, a t test on one coefficient. If it is not rejected, the parallel-lines model is adequate.

Should it be the same line? $H_0 : \beta_2 = \beta_3 = 0$: two restrictions, tested with the full and reduced models of Lecture 15:

$$F^* = \frac{\big(SSE(X) - SSE(X, D, DX)\big)/2}{MSE(X, D, DX)} \sim F(2,\ n-4) \text{ under } H_0.$$

Two tests

Should it be the same slope? $H_0 : \beta_3 = 0$, a t test on one coefficient. If it is not rejected, the parallel-lines model is adequate.

Should it be the same line? $H_0 : \beta_2 = \beta_3 = 0$: two restrictions, tested with the full and reduced models of Lecture 15:

$$F^* = \frac{\big(SSE(X) - SSE(X, D, DX)\big)/2}{MSE(X, D, DX)} \sim F(2,\ n-4) \text{ under } H_0.$$

In R, `y ~ x * d` fits x, d and their product $x{:}d$; `anova(lm(y ~ x), lm(y ~ x * d))` gives the test of the same line.

Insurance firms: is the slope the same?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.8384	2.4406	13.8645	0.0000
size	-0.1015	0.0131	-7.7786	0.0000
stock	8.1313	3.6541	2.2253	0.0408
size:stock	-0.0004	0.0183	-0.0228	0.9821

Insurance firms: is the slope the same?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.8384	2.4406	13.8645	0.0000
size	-0.1015	0.0131	-7.7786	0.0000
stock	8.1313	3.6541	2.2253	0.0408
size:stock	-0.0004	0.0183	-0.0228	0.9821

Same slope? $\hat{\beta}_3 = -0.0004$, $t = -0.02$, $p = 0.98$: no evidence that size acts differently in the two types of firm. The parallel lines are adequate.

Insurance firms: is the slope the same?

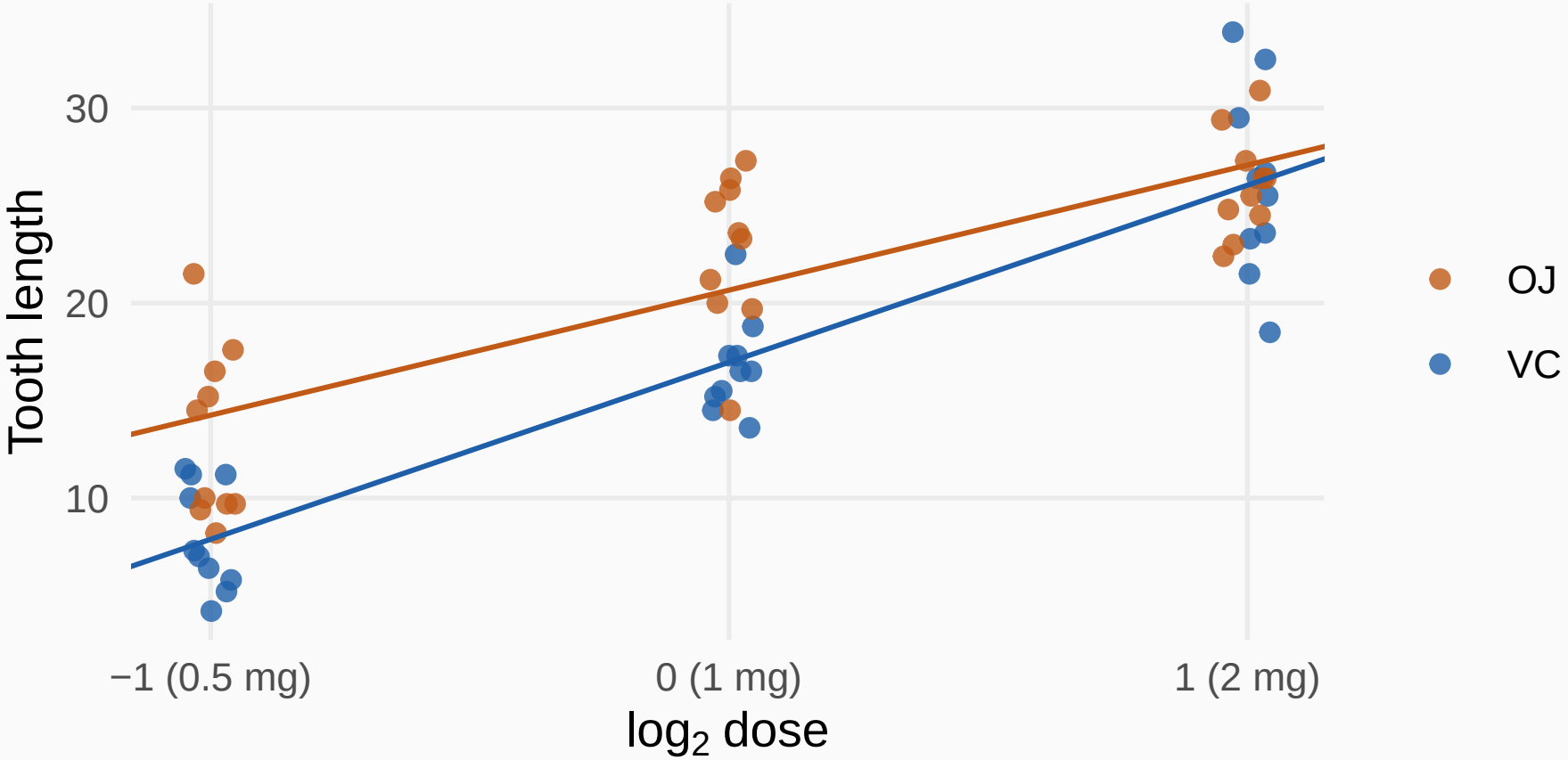
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.8384     2.4406 13.8645   0.0000
size         -0.1015     0.0131 -7.7786   0.0000
stock         8.1313     3.6541  2.2253   0.0408
size:stock   -0.0004     0.0183 -0.0228   0.9821
```

Same slope? $\hat{\beta}_3 = -0.0004$, $t = -0.02$, $p = 0.98$: no evidence that size acts differently in the two types of firm. The parallel lines are adequate.

Same line? $F^* = 14.34$ on 2 and 16 df ($p = 0.0003$): one line for both types is rejected; the intercepts differ.

Example: vitamin C and tooth growth

Tooth growth of 60 guinea pigs (`ToothGrowth`, in R): vitamin C at a dose of 0.5, 1 or 2 mg per day, given as orange juice (OJ) or as ascorbic acid (VC). We use $X = \log_2(\text{dose}) = -1, 0, 1$.



The lines of the interaction model: ascorbic acid starts lower but rises faster.

The interaction model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20.663	0.679	30.425	0.000
ldose	6.415	0.832	7.712	0.000
suppVC	-3.700	0.960	-3.852	0.000
ldose:suppVC	2.665	1.176	2.266	0.027

The interaction model

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     20.663      0.679  30.425    0.000
ldose            6.415      0.832   7.712    0.000
suppVC          -3.700      0.960  -3.852    0.000
ldose:suppVC     2.665      1.176   2.266    0.027
```

- Orange juice (baseline): slope 6.42 per doubling of the dose. Ascorbic acid: 6.42 + 2.66 = 9.08.
- The interaction is significant: $t = 2.27$, $p = 0.027$ (equivalently $F^* = 5.13 = t^2$ in `anova(additive, interaction)`).

The interaction model

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    20.663      0.679  30.425    0.000
ldose           6.415      0.832   7.712    0.000
suppVC         -3.700      0.960  -3.852    0.000
ldose:suppVC    2.665      1.176   2.266    0.027
```

- Orange juice (baseline): slope 6.42 per doubling of the dose. Ascorbic acid: 6.42 + 2.66 = 9.08.
- The interaction is significant: $t = 2.27$, $p = 0.027$ (equivalently $F^* = 5.13 = t^2$ in `anova(additive, interaction)`).

Difference VC $-$ OJ: $-3.70 + 2.665\,X$: -6.36 at 0.5 mg, -3.70 at 1 mg, -1.04 at 2 mg.
The coefficient of `suppVC` alone is the difference at 1 mg only.

Interaction between two qualitative predictors

Tooth growth with the dose as a factor (3 levels) and the supplement (2 levels): 6 cells. Mean tooth length in each cell:

	0.5 mg	1 mg	2 mg
OJ	13.23	22.70	26.06
VC	7.98	16.77	26.14
OJ − VC	5.25	5.93	-0.08

Interaction between two qualitative predictors

Tooth growth with the dose as a factor (3 levels) and the supplement (2 levels): 6 cells. Mean tooth length in each cell:

	0.5 mg	1 mg	2 mg
OJ	13.23	22.70	26.06
VC	7.98	16.77	26.14
OJ $-$ VC	5.25	5.93	-0.08

Without interaction (additive model), the OJ $-$ VC difference would be the same at every dose. With interaction, each cell has its own mean: $2 \times 3 = 6$ parameters instead of $1 + 1 + 2 = 4$.

Interaction between two qualitative predictors

Tooth growth with the dose as a factor (3 levels) and the supplement (2 levels): 6 cells. Mean tooth length in each cell:

	0.5 mg	1 mg	2 mg
OJ	13.23	22.70	26.06
VC	7.98	16.77	26.14
OJ $-$ VC	5.25	5.93	-0.08

Without interaction (additive model), the OJ $-$ VC difference would be the same at every dose. With interaction, each cell has its own mean: $2 \times 3 = 6$ parameters instead of $1 + 1 + 2 = 4$.

Test of the interaction: the 2 extra parameters, $F^* = 4.11$ on 2 and 54 df ($p = 0.022$): the supplement matters less at the highest dose.

Three rules when using interactions

- Keep the main effects. When $D \cdot X$ is in the model, keep D and X too, even if their own tests are not significant: without them, the two lines are forced through constraints that rarely make sense.

Three rules when using interactions

- Keep the main effects. When $D \cdot X$ is in the model, keep D and X too, even if their own tests are not significant: without them, the two lines are forced through constraints that rarely make sense.

- Interpret the coefficients at the right place. With an interaction, the coefficient of D is the group difference where $X = 0$. If $X = 0$ is far from the data, it means little. Centring X (using $X - \bar{X}$) makes it the difference at the average X.

Three rules when using interactions

- Keep the main effects. When $D \cdot X$ is in the model, keep D and X too, even if their own tests are not significant: without them, the two lines are forced through constraints that rarely make sense.

- Interpret the coefficients at the right place. With an interaction, the coefficient of D is the group difference where $X = 0$. If $X = 0$ is far from the data, it means little. Centring X (using $X - \bar{X}$) makes it the difference at the average X.

- Count the parameters. Each interaction with a factor of k levels adds $k - 1$ parameters; with several predictors, they add up quickly (and $(X'X)^{-1}$ may not exist anymore).

Exercise

For the insurance firms, `lm(months ~ size * stock)` gives, with X the size of the firm and $D = 1$ for stock firms:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.8384     2.4406 13.8645   0.0000
size         -0.1015     0.0131 -7.7786   0.0000
stock         8.1313     3.6541  2.2253   0.0408
size:stock   -0.0004     0.0183 -0.0228   0.9821
```

1. Write down the fitted line of each type of firm.
2. Estimate the difference in mean months between a stock and a mutual firm of size 200.
3. Which model would you report, this one or the parallel-lines model, and why?

Exercise: solution

1. Mutual ($D = 0$): $\hat{Y} = 33.84 - 0.1015\, X$. Stock ($D = 1$): $\hat{Y} = 41.97 - 0.1019\, X$.

Exercise: solution

1. Mutual ($D = 0$): $\hat{Y} = 33.84 - 0.1015\, X$. Stock ($D = 1$): $\hat{Y} = 41.97 - 0.1019\, X$.
2. $\hat{\beta}_2 + \hat{\beta}_3 \times 200 = 8.13 - 0.0004 \times 200 = 8.05$ months.

Exercise: solution

1. Mutual ($D = 0$): $\hat{Y} = 33.84 - 0.1015\,X$. Stock ($D = 1$): $\hat{Y} = 41.97 - 0.1019\,X$.
2. $\hat{\beta}_2 + \hat{\beta}_3 \times 200 = 8.13 - 0.0004 \times 200 = 8.05$ months.
3. The parallel-lines model. In the table, the test of $H_0 : \beta_3 = 0$ (same slope) gives $t = -0.02$, $p = 0.98$: no evidence that the slopes differ, and they are almost equal (-0.1015 and -0.1019). The simpler model is easier to interpret: one slope, and one difference between the types, 8.1 months ($t = 5.52$), that holds at every size.

Summary

- Correlated predictors inflate the variances of the coefficients and make them hard to interpret, without violating any assumption.
- A qualitative predictor with k levels enters through $k - 1$ indicators, and each coefficient compares a level with the baseline. Testing that all the group means are equal is the one-way analysis of variance (global F-test).
- With a quantitative predictor and an indicator: parallel lines, and the coefficient of the indicator compares the groups at the same X.
- An interaction lets the effect of one predictor depend on another: $D \cdot X$ (different slopes), or two factors (cell means).
- Same slope: t test on the interaction. Same line: F test of full against reduced model.

Next week: comparing regression functions across groups, interactions between two quantitative predictors, then polynomial regression and transformations.