

Comprendre une relation

Régression linéaire multiple avec les propriétés d'Ames

Aurélien Nicosia

Le fil de la matinée

Heure	Travail
8 h 45	Question de départ et brise-glace
9 h	Bloc 1: choisir avant d'ajuster
9 h 30	Mission 1 et débrief
10 h 10	Pause
10 h 20	Bloc 2: interpréter et prédire
10 h 50	Mission 2 et débrief
11 h 30	Bloc 3: mettre le modèle à l'épreuve
11 h 50	Mission finale et synthèse

Ce que vous produirez

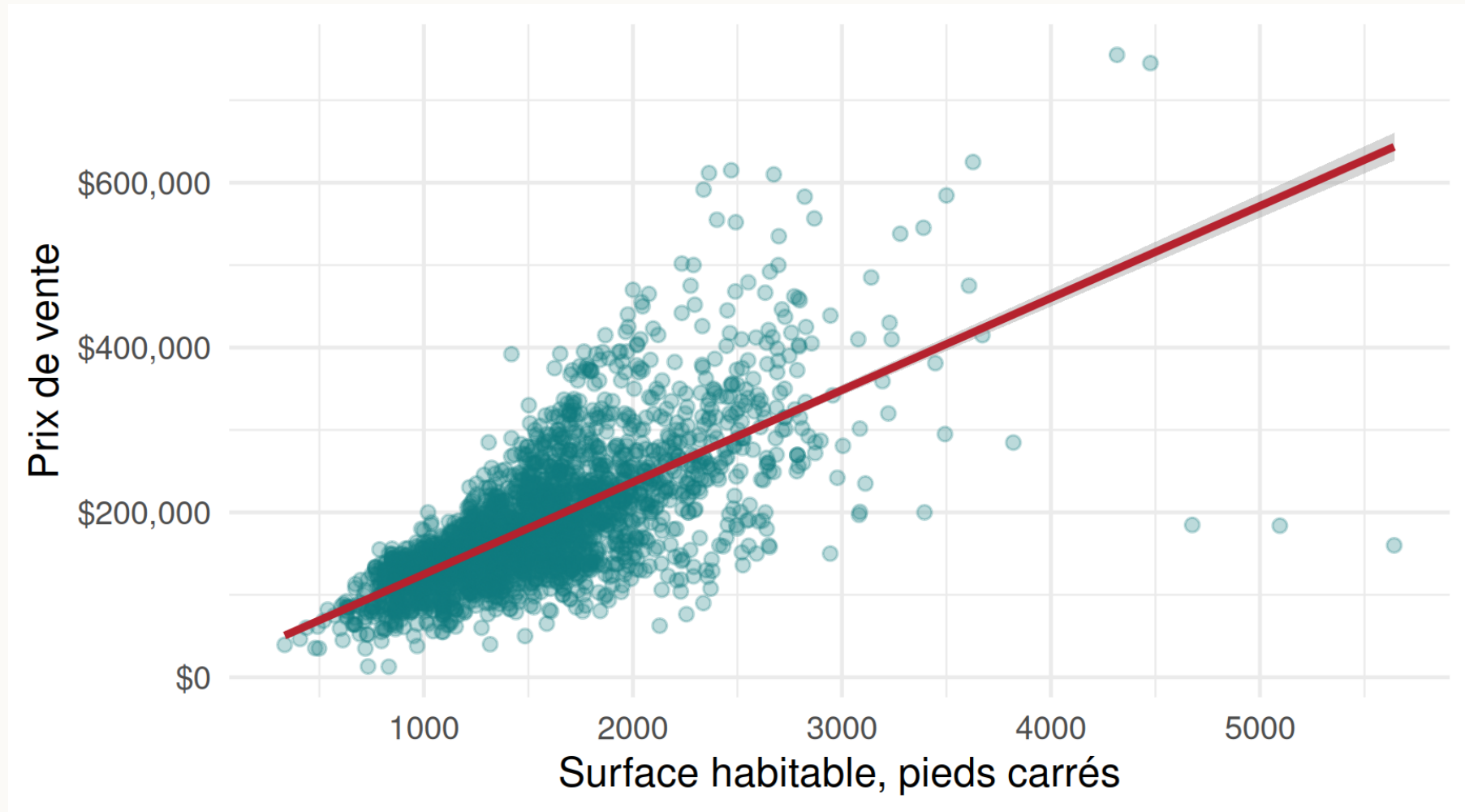
À 12 h 15, chaque équipe aura produit une fiche modèle de cinq phrases:

1. la question étudiée;
2. le modèle retenu;
3. le principal résultat conditionnel;
4. le diagnostic qui nuance ce résultat;
5. la limite qui encadre son utilisation.

Le but n'est pas seulement d'obtenir une équation, mais de décider quand lui faire confiance.

Qu'est-ce qui fait varier le prix d'une maison?

Une relation visible, mais pas encore comprise



Que faudrait-il vérifier avant de raconter cette relation?

Bloc 1

Choisir avant d'ajuster

De la question au petit ensemble de prédicteurs

Trois usages, trois critères de réussite

Usage	Question prioritaire
Décrire	Le résumé représente-t-il les données?
Expliquer une association	Le coefficient conditionnel est-il interprétable?
Prédire	L'erreur sur de nouvelles ventes est-elle acceptable?

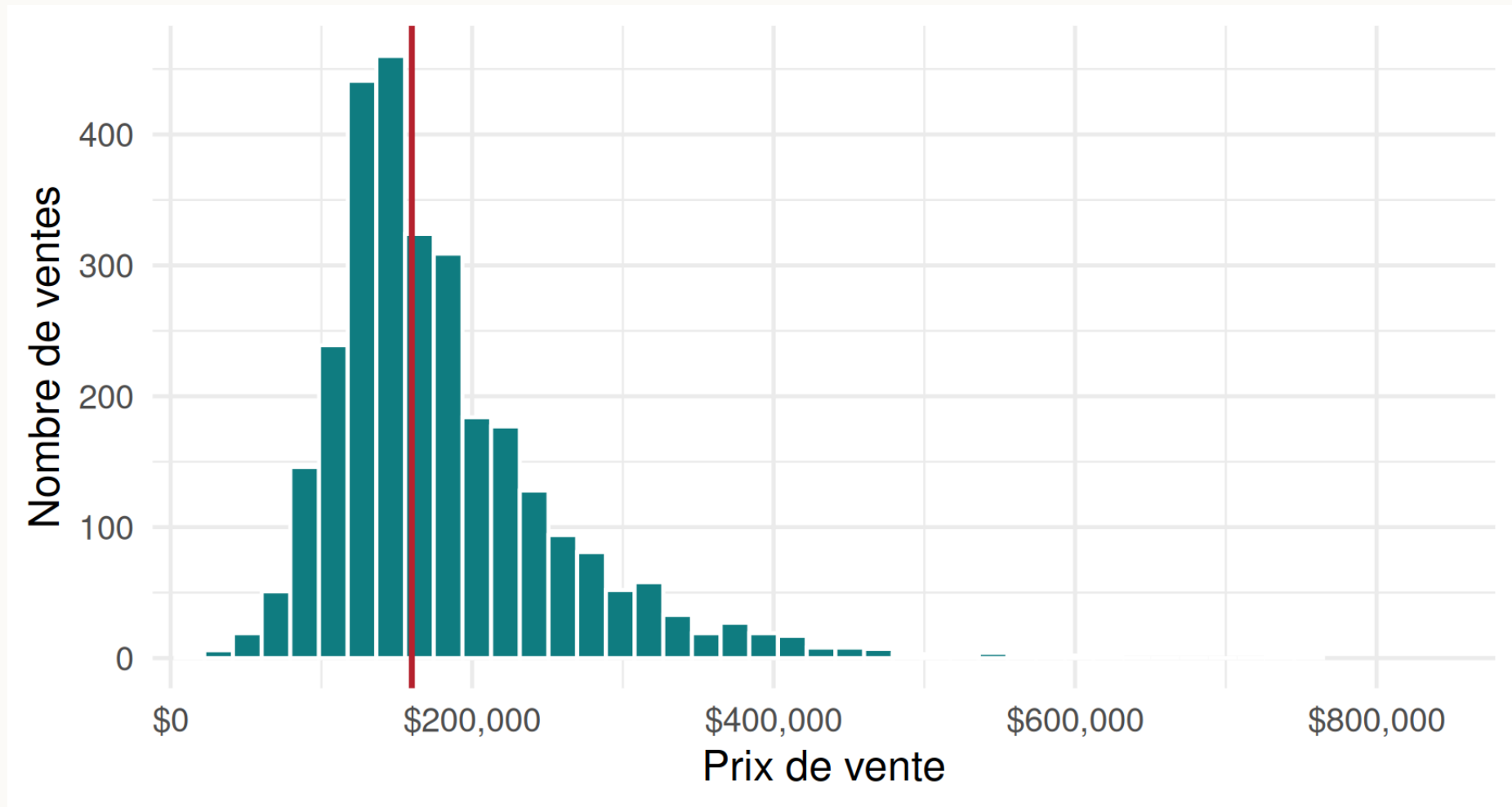
Un bon ajustement ne transforme pas des données observationnelles en expérience causale.

L'unité statistique fixe la portée

- 2930 ventes résidentielles
- 28 quartiers codés
- une période de 2006 à 2010 dans le jeu original

Une ligne n'est ni une personne, ni un quartier, ni la valeur actuelle d'une propriété.

Le prix brut est asymétrique



Médiane: \$160,000

Le type d'une variable détermine son rôle dans le modèle

Variable	Type	Traitement retenu
<code>gr_liv_area</code>	quantitative	pente par pied carré
<code>overall_qual</code>	ordinaire	score de 1 à 10
<code>year_built</code>	quantitative	pente par année
<code>garage_cars</code>	dénombrement	pente par place
<code>neighborhood</code>	nominale	indicateurs et référence

Raisons défendables: connaissance du problème, preuve graphique, contrôle d'un facteur et interprétabilité.

Simple et multiple ne répondent pas à la même question

Régression simple:

$$\text{prix} = \beta_0 + \beta_1 \text{surface} + \varepsilon$$

Régression multiple:

$$\text{prix} = \beta_0 + \beta_1 \text{surface} + \beta_2 \text{qualité} + \beta_3 \text{année} + \beta_4 \text{garage} + \varepsilon$$

Le coefficient de surface devient une comparaison entre propriétés semblables sur les autres variables du modèle.

Les moindres carrés choisissent la droite qui réduit les erreurs

Pour chaque vente:

$$\text{résidu}_i = y_i - \hat{y}_i$$

`lm()` choisit les coefficients qui minimisent:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

Les grandes erreurs comptent davantage parce qu'elles sont élevées au carré.

**Votre modèle commence avec
4, 10 ou 80 prédicteurs?**

Mission 1

Ouvrir la mission 1

25 minutes, en équipes de trois ou quatre.

À remettre:

1. une formule de modèle;
2. un graphique probant;
3. un verdict de trois phrases.

Avant d'exécuter `lm()`, prédisez le signe de chaque coefficient.

Débrief de la mission 1

Modèle commun retenu:

```
1 modele_prix <- lm(  
2   sale_price ~ gr_liv_area + overall_qual +  
3   year_built + garage_cars,  
4   data = ames  
5 )
```

Verdict attendu:

1. une ligne est une vente;
2. l'objectif initial est l'association conditionnelle;
3. les quatre variables sont lisibles, plausibles et imparfaites.

Pause

10 h 10 à 10 h 20

Avant de partir, écrivez une question qui reste floue.

Bloc 2

Interpréter et prédire

Lire un coefficient, puis communiquer l'incertitude

Le modèle est une moyenne conditionnelle

$$E(Y_i | X_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_4 X_{i4}$$

Chaque coefficient représente une différence moyenne attendue lorsque les autres variables du modèle sont maintenues constantes.

Cette phrase ne signifie pas:

1. que les propriétés sont identiques sur tout;
2. que toutes les variables importantes sont incluses;
3. que la relation est causale.

Ajuster avec `lm()`

```
1 modele_prix <- lm(  
2   sale_price ~ gr_liv_area + overall_qual +  
3   year_built + garage_cars,  
4   data = ames  
5 )  
6  
7 tidy(modele_prix, conf.int = TRUE)
```

A tibble: 5 × 7

	term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>	conf.low <dbl>	conf.high <dbl>
1	(Intercept)	-820149.	60058.	-13.7	3.32e- 41	-937908.	-702389.
2	gr_liv_area	57.3	1.82	31.5	4.59e-188	53.7	60.8
3	overall_qual	23960.	777.	30.8	3.85e-181	22437.	25483.
4	year_built	377.	31.6	11.9	4.17e- 32	315.	439.
5	garage_cars	14723.	1275.	11.5	3.43e- 30	12223.	17223.

```
1 glance(modele_prix)
```

A tibble: 1 × 12

	r.squared <dbl>	adj.r.squared <dbl>	sigma <dbl>	statistic <dbl>	p.value <dbl>	df <dbl>	logLik <dbl>	AIC <dbl>	BIC <dbl>
1	0.765	0.765	38761.	2379.	0	4	-35111.	70234.	70270.

i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

`tidy()` lit les coefficients. `glance()` résume le modèle entier.

Quatre associations positives

Variable	Coefficient	Borne basse	Borne haute
Surface	\$57	\$54	\$61
Qualité	\$23,960	\$22,437	\$25,483
Année	\$377	\$315	\$439
Garage	\$14,723	\$12,223	\$17,223

R carré ajusté: 0.765

Lire le coefficient de surface

Coefficient: \$57.3 par pied carré.

Interprétation utile:

À qualité, année de construction et capacité du garage identiques, 100 pieds carrés supplémentaires sont associés en moyenne à environ \$5,727 de plus sur le prix.

Le coefficient ne dit pas que l'ajout de 100 pieds carrés causerait ce gain.

Lire le coefficient de qualité

Coefficient: \$23,960 par point.

À surface, année de construction et capacité du garage identiques, un point supplémentaire de qualité globale est associé en moyenne à environ \$23,960 de plus.

Réserve: le codage 1 à 10 suppose le même incrément moyen entre tous les niveaux successifs.

Significatif, important et utile sont trois jugements distincts

Élément	Question posée
Coefficient	Quelle est la direction et la taille estimée?
Intervalle de confiance	Quelles valeurs restent compatibles avec le modèle?
Valeur p	Les données sont-elles compatibles avec un coefficient nul?
Utilité pratique	L'effet est-il substantiel dans le contexte?

Une petite valeur p ne garantit ni un effet important, ni une bonne prédiction.

Colinéarité: les prédicteurs se répètent-ils?

Variable	VIF
gr_liv_area	1.64
overall_qual	2.34
year_built	1.78
garage_cars	1.84

Ici, les VIF restent modérés. Cela ne valide pas les autres hypothèses du modèle.

Deux intervalles pour deux questions

Intervalle	Cible
Confiance	prix moyen de propriétés semblables
Prédiction	prix d'une nouvelle propriété individuelle

Le second est plus large car il combine:

1. l'incertitude sur la moyenne;
2. la variabilité individuelle autour de cette moyenne.

Une propriété type

1 500 pieds carrés, qualité 6, construite en 2000, garage de deux places.

	fit	lwr	upr
1	192778.7	190525.4	195031.9

	fit	lwr	upr
1	192778.7	116744	268813.3

Mission 2: auditer une estimation

Ouvrir la mission 2

25 minutes.

À remettre:

1. deux prédictions individuelles;
2. une décomposition de leur écart;
3. un message professionnel corrigé.

Débrief de la mission 2

Profil	Valeur ajustée	Intervalle individuel
Maison A	\$193,000	\$116,000 à \$269,000
Maison B	\$214,000	\$138,000 à \$290,000

L'écart ajusté est d'environ \$21,200, mais les caractéristiques poussent dans des directions opposées et l'incertitude individuelle reste large.

Appliquer un coefficient n'est pas répéter sa définition: il faut recomposer le profil, l'incertitude et la portée.

Bloc 3

Mettre le modèle à l'épreuve

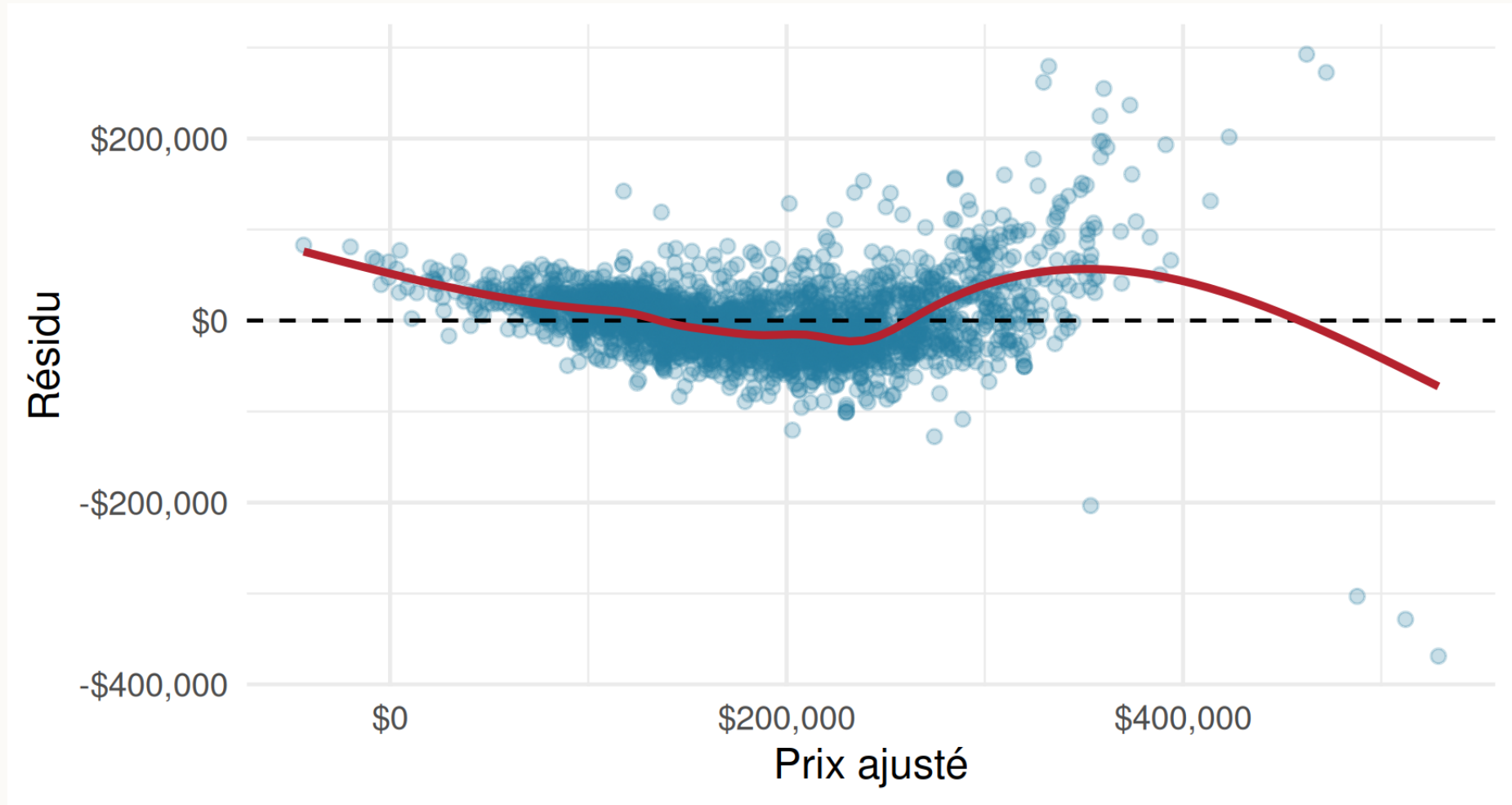
Résidus, influence, transformation et décision

Les hypothèses ne protègent pas toutes la même conclusion

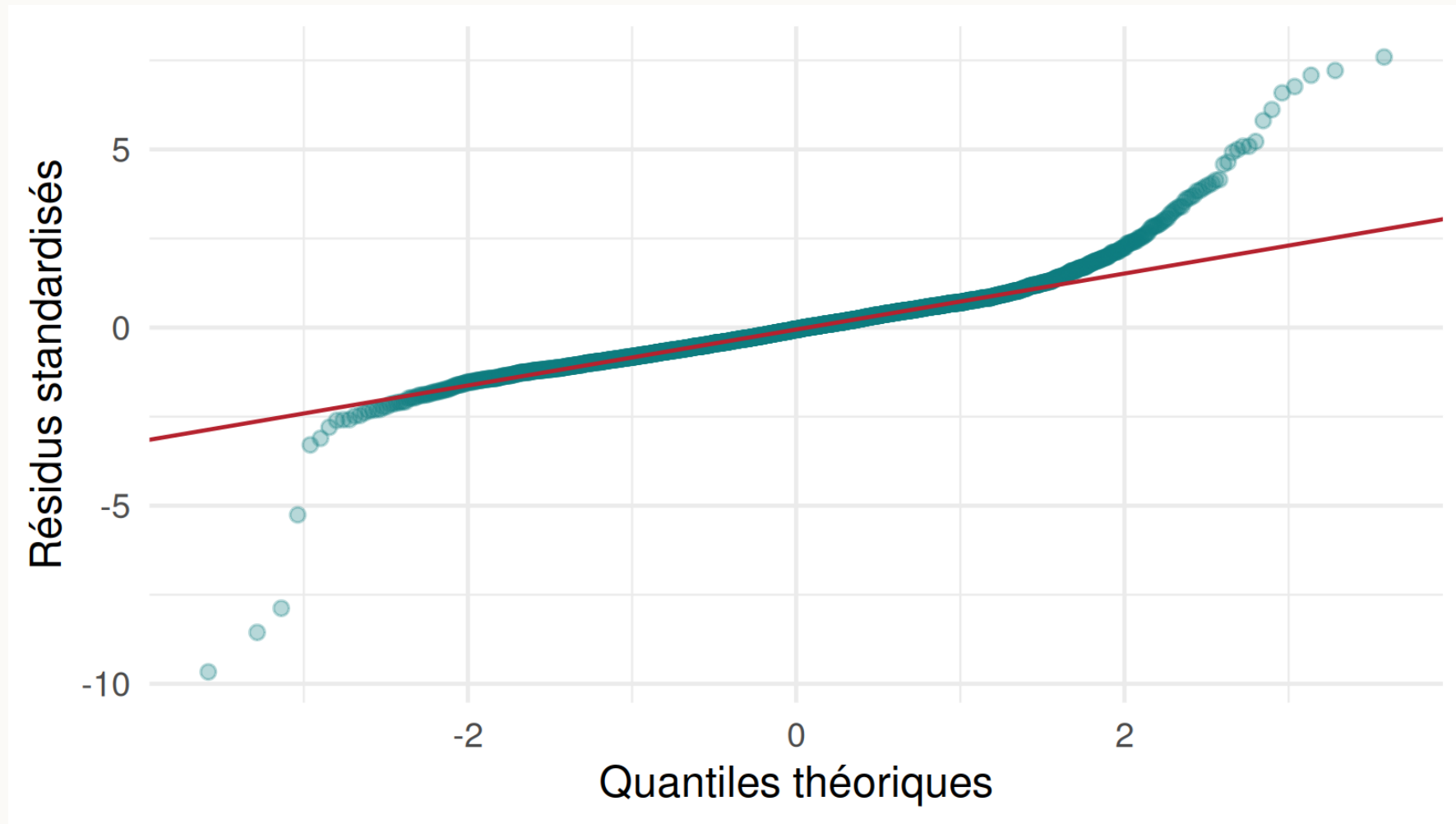
Hypothèse	Vérification principale
moyenne correctement structurée	résidus contre valeurs ajustées
variance à peu près constante	largeur verticale des résidus
erreurs indépendantes	plan de collecte, temps et espace
queues compatibles avec l'inférence	diagramme quantile-quantile
absence de domination individuelle	influence et sensibilité

La normalité concerne les erreurs, pas la distribution brute du prix.

Les résidus révèlent une structure



Le QQ-plot examine les queues



L'écart dans les queues concerne surtout l'inférence et les intervalles, pas l'existence même d'une association.

Influence: un signal à enquêter

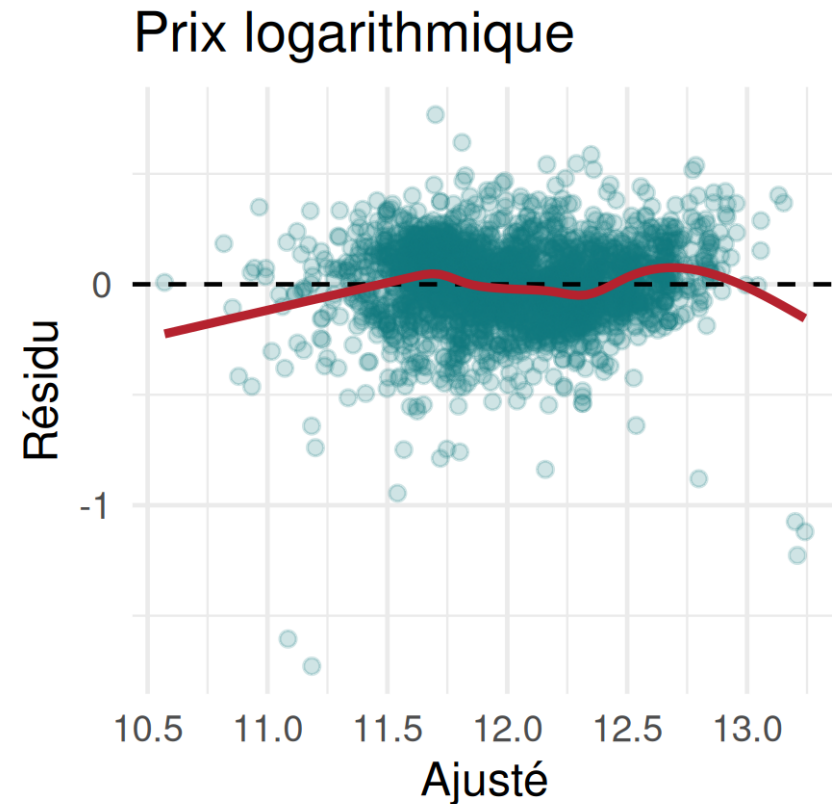
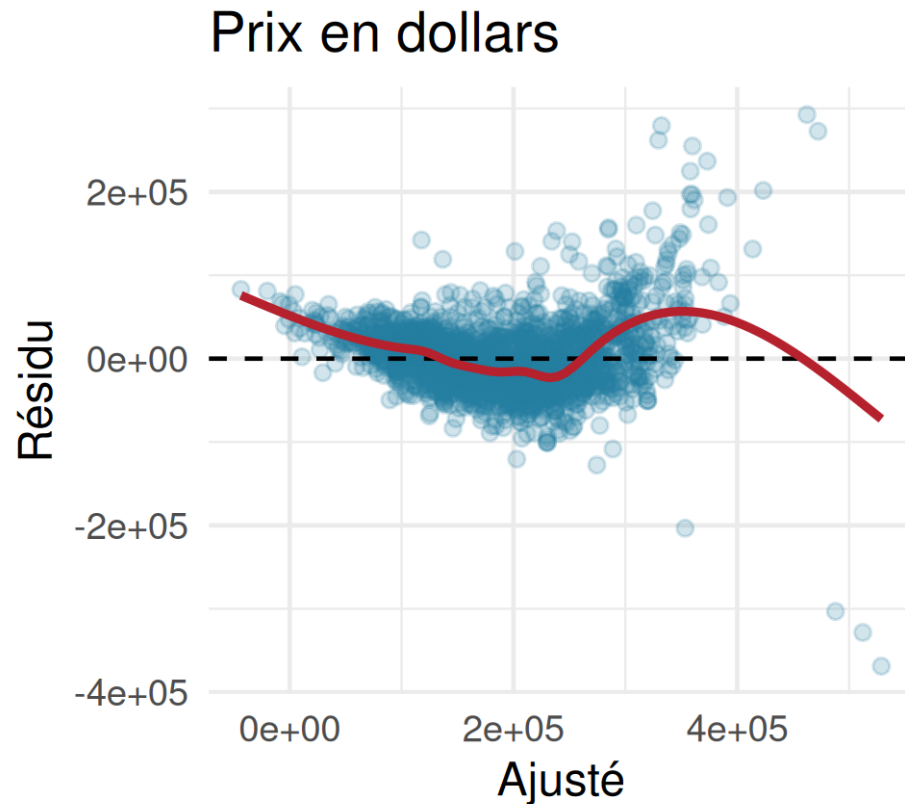
numero	sale_price	gr_liv_area	overall_qual	.coooksd
1499	\$160,000	5642	10	0.592
2181	\$183,850	5095	10	0.294
2182	\$184,750	4676	10	0.189
1768	\$755,000	4316	10	0.134
1761	\$745,000	4476	10	0.131

La vente 1499 est inhabituelle. Elle n'est pas automatiquement erronée.

Sensibilité: avec et sans la vente la plus influente

Variable	Toutes	Sans la vente 1499
gr_liv_area	57.3	60.2
overall_qual	23959.6	23775.3
year_built	376.9	388.0
garage_cars	14723.3	13798.2

Le logarithme change le diagnostic et la question



Mission finale

Ouvrir la mission finale

20 minutes.

À remettre:

1. une décision adaptée au mandat reçu;
2. une preuve diagnostique décisive;
3. une fiche modèle de cinq phrases.

Restitution éclair

Chaque équipe dispose de 45 secondes:

1. notre résultat principal;
2. le diagnostic qui a changé notre lecture;
3. ce que nous ne dirions pas.

Les autres équipes peuvent seulement ajouter une nuance ou poser une question.

Sept questions à emporter

1. Quelle est l'unité statistique?
2. Quel est l'objectif: expliquer ou prédire?
3. Pourquoi chaque variable est-elle dans le modèle?
4. Les coefficients gardent-ils leurs unités et leurs conditions?
5. Que montrent les résidus et l'influence?
6. Pour prédire, l'erreur a-t-elle été évaluée sur des données non utilisées pour ajuster?
7. La conclusion dépasse-t-elle les données?

À 12 h 15, votre modèle doit être défendable

Un modèle utile relie:

1. une question précise;
2. une comparaison conditionnelle;
3. une incertitude visible;
4. un diagnostic;
5. une limite explicite.

Demain: passer d'un prix continu à une probabilité.