Linear Models

Week 8: Linear models

- Review of the simple linear model
- Multiple linear regression: linear models with many variables
- Case study: urban tree cover

This week: multiple linear regression

Objective: extend the simple linear model to multiple explanatory variables.

- **Review of the simple linear model**
  - The simple linear model in matrix form
  - Estimation and uncertainty quantification
  - Parameter interpretation
- **Multiple linear regression**
  - The linear model in matrix form
  - Estimation and uncertainty quantification
- **Case study: urban tree cover**
  - Background
  - Model 1: summer temperatures, tree cover, and income
    - model fitting calculations, step by step
    - model visualization
    - interpretation of results
  - Model 2: adding population density
    - categorical variable encodings
    - results and interpretation

Multiple linear regression

- The linear model in matrix form
- Estimation and uncertainty quantification

Extending the simple linear model

The simple linear model was:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \begin{cases} i = 1, \ldots, n \\ \epsilon_i \sim N\left(0, \sigma^2\right) \end{cases} \quad \text{(simple linear m}$$

fitted value

residual

It's called 'simple' because it only has a single explanatory variable x_i.

The **linear model** is a direct extension of the simple linear model to $p-1$ variables $x_{i1}, \ldots, x_{i,p-1}$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

$$= \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \mathcal{E}$$

Other names: this is sometimes also called the *multiple regression model* or *multiple linear model*.

The linear model in matrix form

The linear model is often written observation-wise in *indexed* form as above: i indexes the observations. However, it's much more concise in matrix form:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

This is shorthand for:

$$\mathbf{y}: \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \mathbf{X}: \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{bmatrix} \times \beta: \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \epsilon: \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$Var(\varepsilon) = \sigma^2 I$$

The linear model in matrix form

Carrying out the arithmetic on the right-hand side:

$$\underset{n\times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}} = \underset{n\times 1}{\begin{bmatrix} \beta_0 + \beta_1 x_{11} + \cdots + \beta_{p-1} x_{1,p-1} + \epsilon_1 \\ \beta_0 + \beta_1 x_{21} + \cdots + \beta_{p-1} x_{2,p-1} + \epsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \cdots + \beta_{p-1} x_{n,p-1} + \epsilon_n \end{bmatrix}}$$

This is exactly the model relationship, written for each i.

Estimation in the MLR

Estimation and uncertainty quantification are **exactly the same as in the simple linear model**.

The OLS estimates are:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_{p-1} \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

An estimate of the error variance is:

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_{p-1}x_{i,p-1}\right)^2 = \frac{1}{n-p}\underbrace{\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)}_{\text{measure of residual variation}}$$

The simple linear model was the special case with $p = 2$.

$$\hat{B} = \underbrace{(X^T X)^{-1} X^T}_{\text{Fixed and known.}} \underbrace{y}_{\text{Random Variable}}$$

$\hat{B}$: Random Variable (estimator)

Unbiased: $E[\hat{B}] = B$

Variance: $Var(\hat{B}) =$

$$Var((X^T X)^{-1} X^T y)$$

$$(X^T X)^{-1} X^T Var(y|X) X (X^T X)^{-1}$$

$$= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1}$$

Uncertainty quantification

In the case of the multiple linear model, the variances and covariances are a $p \times p$ (instead of 2×2) matrix:

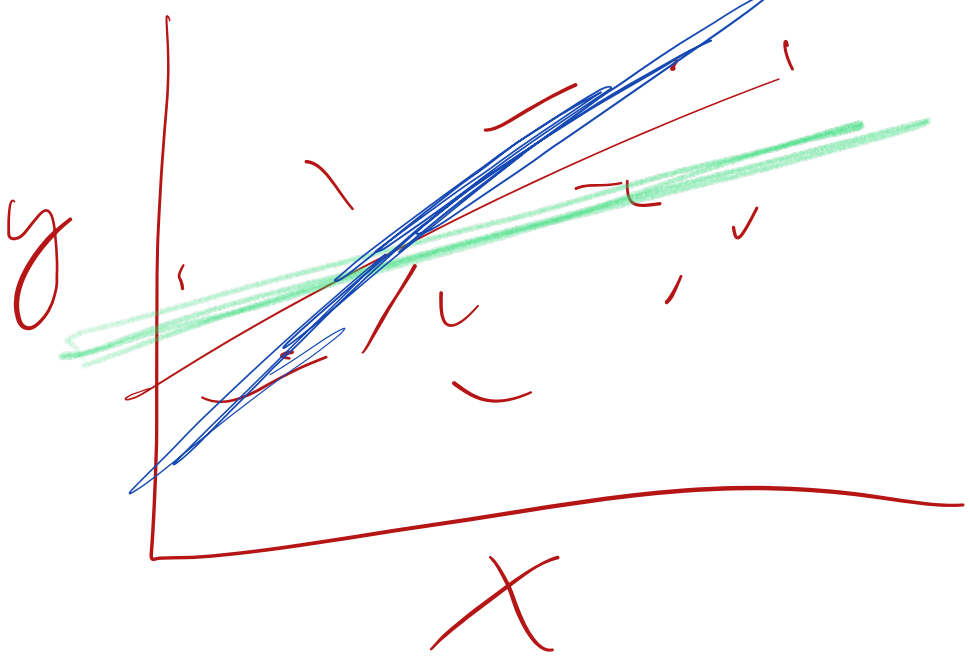
$$\mathbf{V} = \begin{bmatrix} \operatorname{var}\hat{\beta}_0 & \operatorname{cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) & \cdots & \operatorname{cov}\left(\hat{\beta}_0, \hat{\beta}_{p-1}\right) \\ \operatorname{cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) & \operatorname{var}\hat{\beta}_1 & \cdots & \operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_{p-1}\right) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}\left(\hat{\beta}_0, \hat{\beta}_{p-1}\right) & \operatorname{cov}\left(\hat{\beta}_1, \hat{\beta}_{p-1}\right) & \cdots & \operatorname{var}\hat{\beta}_{p-1} \end{bmatrix}$$

This matrix is again estimated by plugging in $\hat{\sigma}^2$ (the estimate) for σ^2:

$$\hat{\mathbf{V}} = \begin{bmatrix} \hat{v}_{11} & \hat{v}_{12} & \cdots & \hat{v}_{1p} \\ \hat{v}_{21} & \hat{v}_{22} & \cdots & \hat{v}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{v}_{p1} & \hat{v}_{p2} & \cdots & \hat{v}_{pp} \end{bmatrix} = \hat{\sigma}^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

The square roots of the diagonal elements give *standard errors*:

$$\mathrm{SE}(\hat{\beta}_0) = \sqrt{\hat{v}_{11}}\,, \quad \mathrm{SE}(\hat{\beta}_1) = \sqrt{\hat{v}_{22}}\,, \quad \cdots \quad \mathrm{SE}(\hat{\beta}_{p-1}) = \sqrt{}$$



$\hat{\beta}_0 \uparrow \quad \hat{\beta}_1 \downarrow$

$$V(\hat{\beta}) = \begin{pmatrix} \hat{v}_{11} & \hat{v}_{12} \\ \hat{v}_{21} & \hat{v}_{22} \end{pmatrix}$$

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \hat{v}_{12}$$

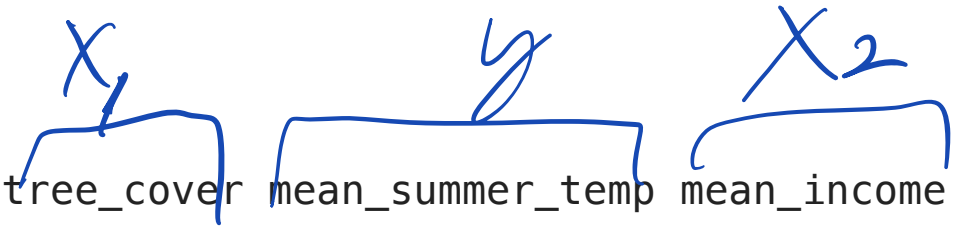
negative

Case study: urban tree cover

- Background
- Model 1: summer temperatures, tree cover, and income
  - model fitting calculations, step by step
  - model visualization
  - interpretation of results
- Model 2: adding population density
  - categorical variable encodings
  - results and interpretation

Urban tree cover data

The following are data on urban tree cover in the San Diego area:



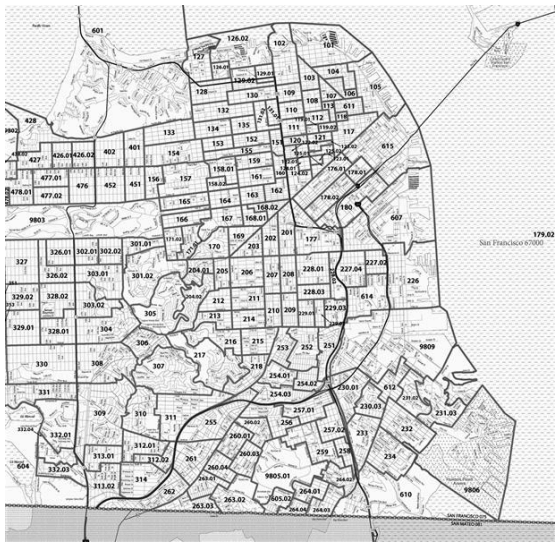

```
# A tibble: 5 × 7
  Name   census_block_GEOID tree_cover mean_summer_temp mean_income
income_level
  <chr>               <dbl>      <dbl>            <dbl>       <dbl> <fct>
1 San D…            6.07e13     0.0107             31.8       95779 high
2 San D…            6.07e13     0.224              31.2       73690 high
3 San D…            6.07e13     0.134              33.7       28860 low
4 San D…            6.07e13     0.0467             35.9       38496 medium
5 San D…            6.07e13     0.0829             33.2       16382 very low
# ℹ 1 more variable: pop_density <fct>
```

Source: McDonald RI, Biswas T, Sachar C, Housman I, Boucher TM, Balk D, et al. (2021) The tree cover and temperature disparity in US urbanized areas: Quantifying the association

Observational units

The **observational units** are census blocks.

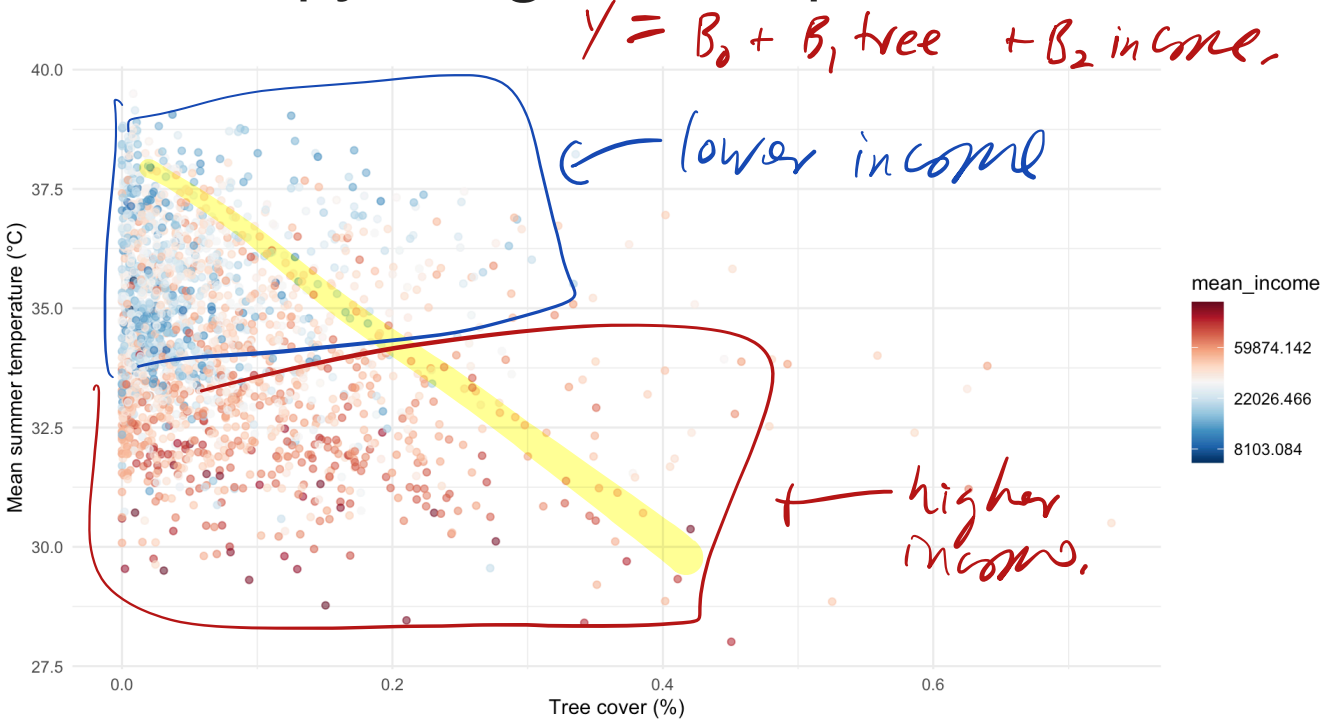
Census blocks are pretty small, about the size of a city block. To get an idea of the geographic scale, here's a map of census tracts in San Francisco. Each tract contains several census blocks:



The data comprise observations on a random sample of 1,998 census blocks in San Diego.

Tree cover and summer temperatures

Tree canopy mitigates temperature in urban areas.



For this reason, urban forestry projects have been advocated as a strategy for mitigating the effects of climate change.

Modeling objectives

Here we'll try to quantify the association between temperature, tree cover, income, and population density using multiple regression.

Let's start with a MLR model with just two explanatory variables, tree cover and income:

$$\underbrace{\text{temp}_i}_{y_i} = \beta_0 + \beta_1 \underbrace{\text{cover}_i}_{x_{i1}} + \beta_2 \underbrace{\text{income}_i}_{x_{i2}} + \epsilon_i \qquad i = 1, \ldots, 199$$

Fitting the model

```
1 summer_temp_mod <- lm(mean_summer_temp ~ tree_cover + mean_income, trees)
2 summary(summer_temp_mod)
```

```
Call:
lm(formula = mean_summer_temp ~ tree_cover + mean_income, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1218 -1.2376 -0.1302  1.2357  4.5536

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.647e+01  7.967e-02 457.745  < 2e-16 ***
tree_cover  -2.816e+00  4.406e-01  -6.391 2.05e-10 ***
mean_income -4.899e-05  1.890e-06 -25.920  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.653 on 1993 degrees of freedom
Multiple R-squared:  0.3029,    Adjusted R-squared:  0.3022
F-statistic:   433 on 2 and 1993 DF,  p-value: < 2.2e-16
```

Error variance estimate

Next we'll calclulate $\hat{\sigma}^2$. For this we need the fitted values:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

And residuals:

$$\mathbf{e} = \left(\mathbf{y} - \mathbf{X}\hat{\beta}\right) = \mathbf{y} - \hat{\mathbf{y}}$$

Error variance estimates

Then the error variance estimate is:

$$\hat{\sigma}^2 = \frac{1}{n-p}\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right) = \frac{1}{n-p}\mathbf{e}'\mathbf{e} = \frac{n-1}{n-p}S_e^2$$

```
# error variance estimate
n <- nrow(trees)
p <- 3

sigmasqhat <- var(resid) * (n - 1) / (n - p)
sigmasqhat
```

```
[1] 2.731703
```

Uncertainty quantification

Lastly, we'll compute the coefficient estimate standard errors:

$$\sqrt{\hat{v}_{jj}} \quad \text{from} \quad \hat{\mathbf{V}} = \hat{\sigma}^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

```
              [,1]          [,2]          [,3]
[1,]  6.347872e-03 -6.467020e-03 -1.180553e-07
[2,] -6.467020e-03  1.941401e-01 -2.255469e-07
[3,] -1.180553e-07 -2.255469e-07  3.571842e-12
```

```
[1] 7.967354e-02 4.406133e-01 1.889932e-06
```

Quality of fit: R^2

There are many metrics that measure fit quality. The most common is the **proportion of variation explained by the model**, which is *denoted by* R^2:

$$R^2 = \frac{\text{reduction in variation}}{\text{total variation}} = \frac{\tilde{\mathbf{y}}'\tilde{\mathbf{y}} - \mathbf{e}'\mathbf{e}}{\tilde{\mathbf{y}}'\tilde{\mathbf{y}}} \qquad \text{where } \tilde{\mathbf{y}} = \mathbf{y} - \bar{\mathbf{y}}$$

$$R^2 = \frac{\text{Var(Fit)}}{\text{Var(total)}} = \frac{\text{Var}(X\beta)}{\text{Var}(Y)}$$

$$1 - \frac{\text{Var}(\varepsilon)}{\text{Var}(Y)}$$

Quality of fit: R^2

```r
# 'by hand'
y <- trees$mean_summer_temp
y_ctr <- y - mean(y)
(t(y_ctr) %*% y_ctr - t(resid) %*% resid) / (t(y_ctr) %*% y_ctr)
```

```
          [,1]
[1,] 0.3028967
```

```r
# Or alculate R^2
var(fitted) / var(y)
```

```
[1] 0.3028967
```

```r
## Or...
summary(summer_temp_mod)$r.squared
```

```
[1] 0.3028967
```

Interpretation: 30% of variation in mean summer temperature is explained by tree cover and income.

Model Summaries

Now we've computed relevant quantities – estimates and standard errors – but we have yet to report these in an organized fashion.

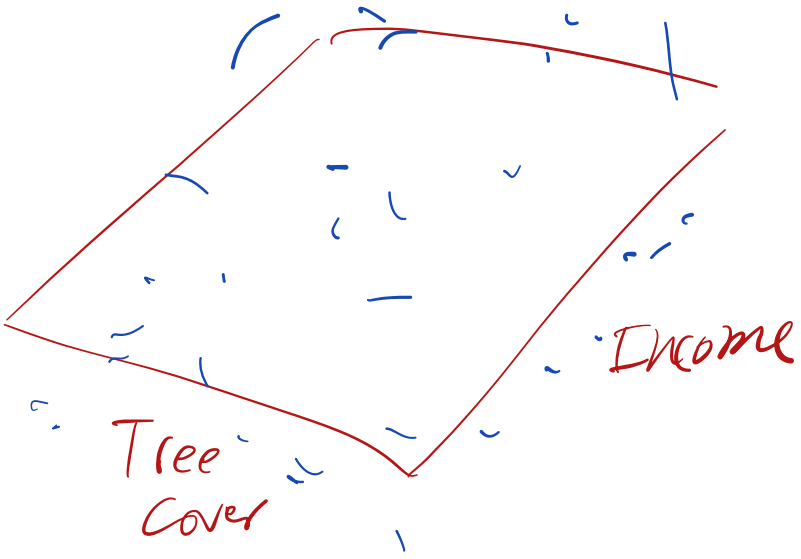
We can get this from `summary` in on the fit model but it's not very neat

```
Call:
lm(formula = mean_summer_temp ~ tree_cover + mean_income, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1218 -1.2376 -0.1302  1.2357  4.5536

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.647e+01  7.967e-02 457.745  < 2e-16 ***
tree_cover  -2.816e+00  4.406e-01  -6.391 2.05e-10 ***
mean_income -4.899e-05  1.890e-06 -25.920  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.653 on 1993 degrees of freedom
Multiple R-squared:  0.3029,    Adjusted R-squared:  0.3022
F-statistic:   433 on 2 and 1993 DF,  p-value: < 2.2e-16
```

Model Summaries

The `gt` package can organize the data nicely

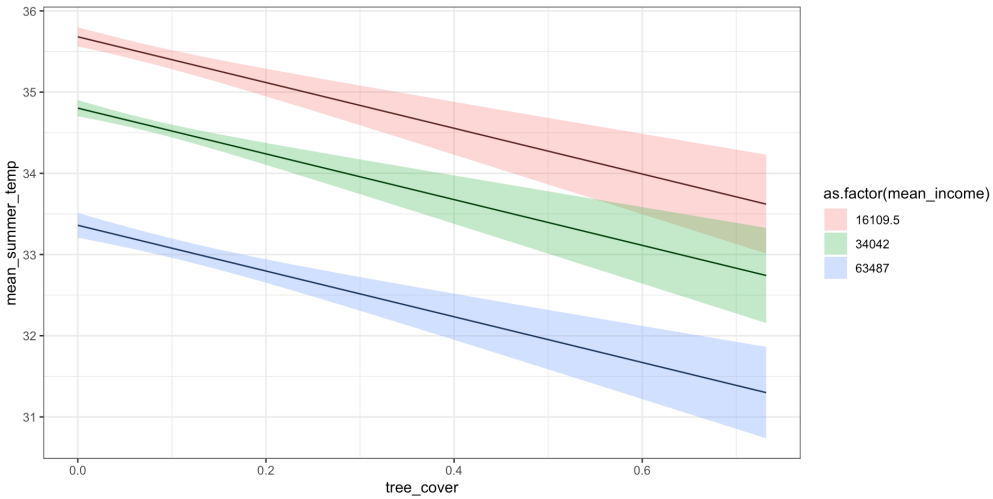
Coefficient Estimates with 95% Confidence Intervals		
Variable	Lower 95% CI	Upper 95% CI
(Intercept)	36.300	36.600
tree_cover	−3.680	−1.950
mean_income	-5.27×10^{-5}	-4.53×10^{-5}


Income
Tree
Cover

Visualization

Here are trend lines and error bands for three values of income (10th, 50th, and 90th percentiles).

The error bands represent the variation of the lines – how much they might change if we sampled different census blocks. (**Not** the variation of temperatures.)



Interpretation

So what have we learned from this exercise?

```
Call:
lm(formula = mean_summer_temp ~ tree_cover + mean_income, data = trees)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1218 -1.2376 -0.1302  1.2357  4.5536

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.647e+01  7.967e-02 457.745  < 2e-16 ***
tree_cover  -2.816e+00  4.406e-01  -6.391 2.05e-10 ***
mean_income -4.899e-05  1.890e-06 -25.920  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.653 on 1993 degrees of freedom
Multiple R-squared:  0.3029,    Adjusted R-squared:  0.3022
F-statistic:   433 on 2 and 1993 DF,  p-value: < 2.2e-16
```

Tree Cover
.01 = 1%
1 = 100%

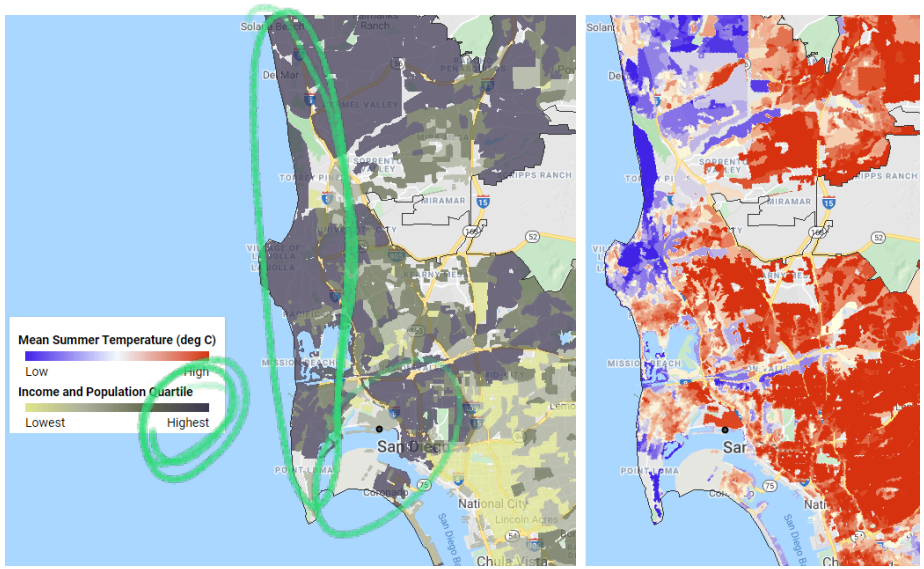
1% increase in tree cover is associated w/ a .028 degree drop in Temp.

- Among census blocks in San Diego, a 1% increase in tree canopy cover is associated with a decrease in average summer temperatures of 3.15 degrees Celsius, after accounting for mean income of the census block.
- Among census blocks in San Diego, a $10K increase in mean income is associated with a decrease in average summer temperatures of 0.5 degrees Celsius, after accounting for tree canopy cover.
- There's still lots of unexplained local variation in average summer temperatures; low R^2.

Explaining associations

There are various mechanisms by which tree canopy reduces temperatures: providing shade; improving air quality. But what explains the weaker association between income and temperature?

Well, one possibility is that property values are higher near the ocean.

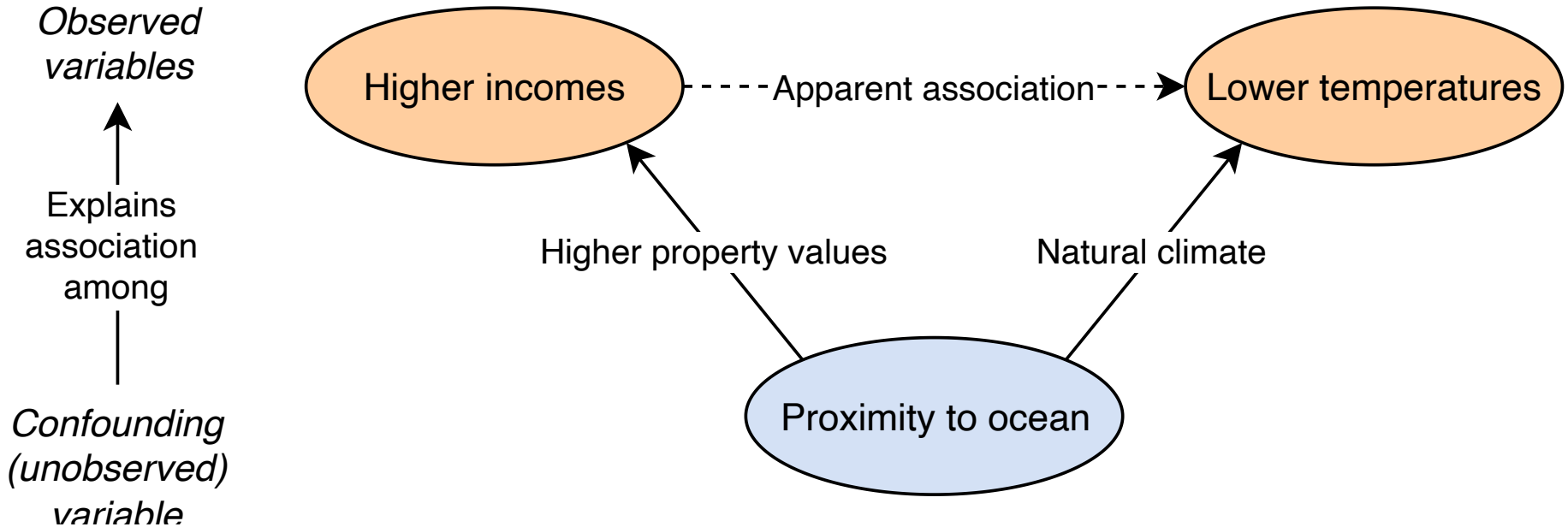


So income may simply be a proxy for distance to the ocean; and temperatures are cooler near the ocean. This is an excellent example of **confounding**!

Confounding

In statistical jargon, we'd say that proximity to the ocean is a *confounding factor*:

- it is correlated with higher incomes (the explanatory variable);
- it is also correlated with lower temperatures (the response);
- and so it 'explains away' the apparent association.


Observed variables
Explains association among
Confounding (unobserved) variable
Higher incomes
Apparent association
Lower temperatures
Higher property values
Natural climate
Proximity to ocean

MLR with categorical variables

What if we also incorporated the population density factor (a categorical variable)?

```
# A tibble: 5 × 7
  Name   census_block_GEOID tree_cover mean_summer_temp mean_income income_level
  <chr>               <dbl>      <dbl>            <dbl>       <dbl> <fct>
1 San D…            6.07e13     0.0107             31.8       95779 high
2 San D…            6.07e13     0.224              31.2       73690 high
3 San D…            6.07e13     0.134              33.7       28860 low
4 San D…            6.07e13     0.0467             35.9       38496 medium
5 San D…            6.07e13     0.0829             33.2       16382 very low
# ℹ 1 more variable: pop_density <fct>
```

Well, it might be natural to think we could just append this column as another variable x_{i3} :

$$\underbrace{\text{temp}_i}_{y_i} = \beta_0 + \beta_1 \underbrace{\text{cover}_i}_{x_{i1}} + \beta_2 \underbrace{\text{income}_i}_{x_{i2}} + \beta_3 \underbrace{\text{density}_i}_{x_{i3}} + \epsilon_i \qquad i = 1, \ldots, 1998$$

But this doesn't quite make sense, because the *values* of density_i would be *words*! So…

$$\beta_3 \times \text{low} = ?$$

So we'll need to represent the categorical variable differently to include it in the model.

Indicator variable encoding

"One Hot"

The solution to this issue is to **encode each level** of the categorical variable using an ***indicator***: a function whose value is zero or one to indicate a condition.

If we want to indicate whether a census block is of low population density, we can use the indicator:

$$I(\text{density} = \text{low}) = \begin{cases} 1 \text{ if population density is low} \\ 0 \text{ otherwise} \end{cases}$$

We can encode the levels of `pop_density` using a collection of indicators:

	pop_density	pop_densitylow	pop_densitymedium	pop_densityhigh
1	low	1	0	0
2	very low	0	0	0
3	medium	0	1	0
4	very low	0	0	0
5	medium	0	1	0
6	very low	0	0	0
7	low	1	0	0
8	low	1	0	0
9	low	1	0	0
10	low	1	0	0

This captures all the information about the categorical variable in quantitative terms.

The MLR model with indicators

The model with the encoded population density variable is:

$$\underbrace{\text{temp}_i}_{y_i} = \beta_0 + \beta_1 \underbrace{\text{cover}_i}_{x_{i1}} + \beta_2 \underbrace{\text{income}_i}_{x_{i2}} + \beta_3 \underbrace{\text{low}_i}_{x_{i3}} + \beta_4 \underbrace{\text{med}_i}_{x_{i4}} + \beta_5 \underbrace{\text{high}_i}_{x_{i5}} + \epsilon_i \qquad i = 1, \ldots$$

only one of these is "turned on" for any observation

The *effect* of doing this is to allow the model to have different intercepts for each population density group.

$$\text{density} = \text{very low} \implies \mathbb{E}\text{temp}_i = \underbrace{\beta_0}_{\text{intercept}} + \beta_1 \text{cover}_i + \beta_2 \text{income}_i \qquad \text{density} = \text{low}$$

↳ $\beta_0 + \beta_1 \text{Cover} + \beta_2 \text{income}$

"high" $\beta_0 + \beta_1 \text{Cover} + \beta_2 \text{income} + \beta_5 \times \text{high}$

In matrix form

The explanatory variable matrix $\mathbf{X}$ for this full model ('full' because it includes all variables) will be of the form:

$$\mathbf{X} = \left[\begin{array}{c:cccc} 1 & \text{cover}_1 & \text{income}_1 & \text{low}_1 & \text{med}_1 & \text{high}_1 \\ 1 & \text{cover}_2 & \text{income}_2 & \text{low}_2 & \text{med}_2 & \text{high}_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \text{cover}_{1998} & \text{income}_{1998} & \text{low}_{1998} & \text{med}_{1998} & \text{high}_{1998} \end{array}\right]$$

Here's everything to the right of the dashed partition:

	tree_cover	mean_income	pop_densitylow	pop_densitymedium	pop_densityhigh
1	0.01068702	95779	1	0	0
2	0.22361858	73690	0	0	0
3	0.13386349	28860	0	1	0
4	0.04669126	38496	0	0	0

Fit summary

The remaining calculations are all the same as before. (You can see the source code for these slides if you're interested in reviewing the details.)

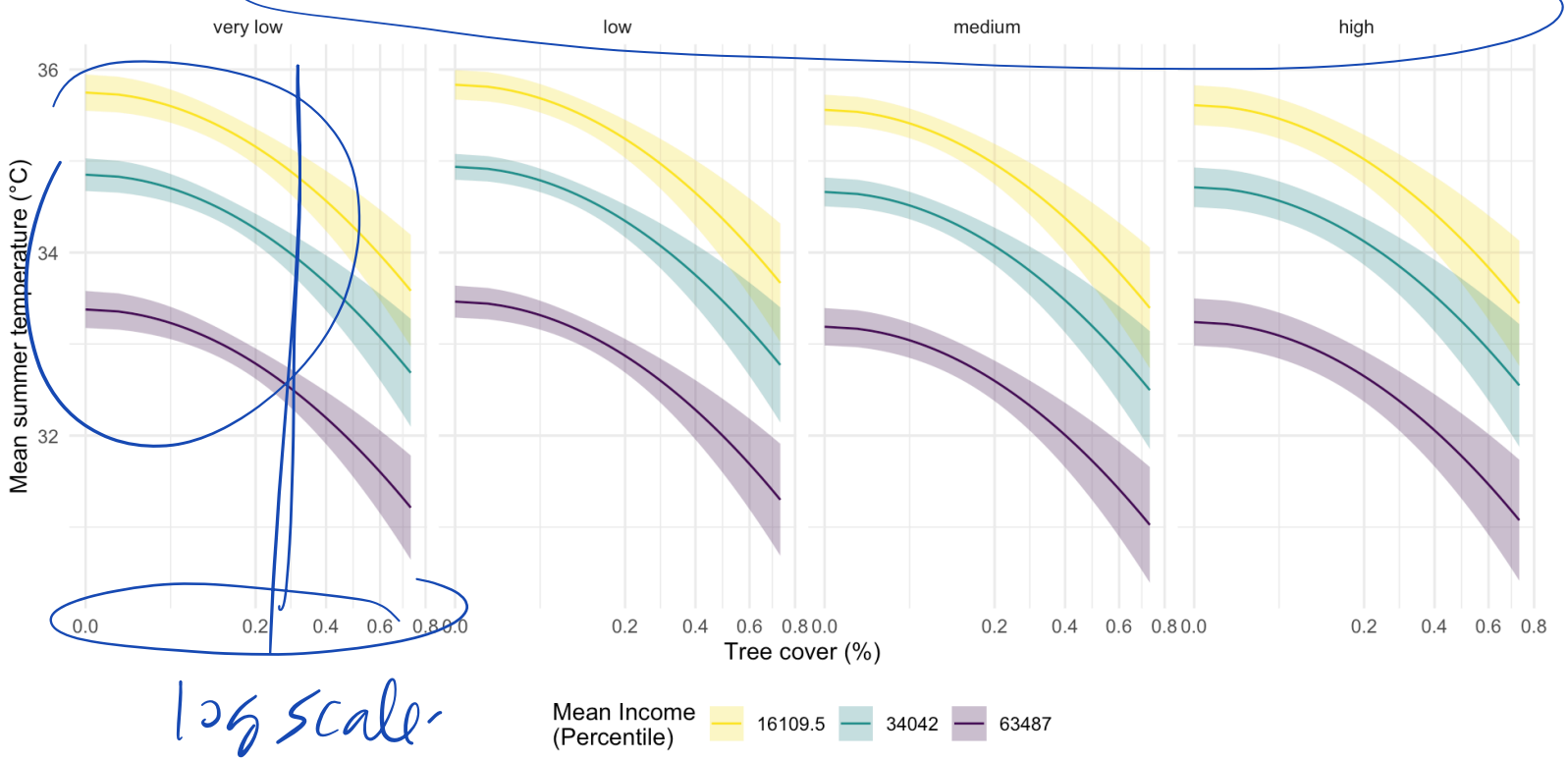
Here is the model fit summary:

term	estimate	std.error	statistic	p.value
(Intercept)	3.655395e+01	1.159275e-01	315.3173786	0.000000e+00
tree_cover	-2.959149e+00	4.616625e-01	-6.4097664	1.814791e-10
mean_income	-5.002453e-05	1.935154e-06	-25.8504109	2.745531e-127
pop_densitylow	8.574775e-02	9.637409e-02	0.8897386	3.737138e-01
pop_densitymedium	-1.887223e-01	1.080009e-01	-1.7474148	8.071965e-02
pop_densityhigh	-1.377282e-01	1.338675e-01	-1.0288399	3.036799e-01

- Estimates for the cover and income coefficients are about the same.
- The population density variable changes the intercept by about ± 0.15 degrees Celsius, depending on the density level.
- So the association between temperature and population density appears negligible.

Model visualization

The estimated trends and error bands for the same three income levels and each level of population density are shown below, without the data scatter:



Comments on scope of inference

The data in this case study are from a *random sample* of census blocks in the San Diego urban area.

They are therefore representative of *all* census blocks in the San Diego urban area (population).

So the results are generliazable, meaning:

- The model approximates the *actual* associations between summer temperatures, tree cover, income, and population density in the region.

Summary

This week focused on extending the simple linear model to multiple explanatory variables.

- The **linear model** represents a quantitative response variable y_i as a linear function of several explanatory variables and a random error:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

- When represented in matrix form $\mathbf{y} = \mathbf{X}\beta + \epsilon$, all calculations are the same as in the simple linear model.
  - OLS estimates: $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
  - Error variance: $\hat{\sigma}^2 = \frac{1}{n-p}\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)$
  - Uncertainty quantification: $\hat{\mathbf{V}} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$
- Model fitting and interpretation was illustrated in a case study of urban tree cover in San Diego.
  - Estimation and uncertainty quantification calculations
  - Model visualization
  - Encoding categorical variables
  - Parameter interpretation