Agenda

Linear regression as a probability model

1. Administrative
2. Linear regression with one covariate
3. Joint posteriors
4. Interpretation of log-scale coefficients
5. ***Hands on:*** *Working with posterior samples in R*

ADMINISTRATIVE

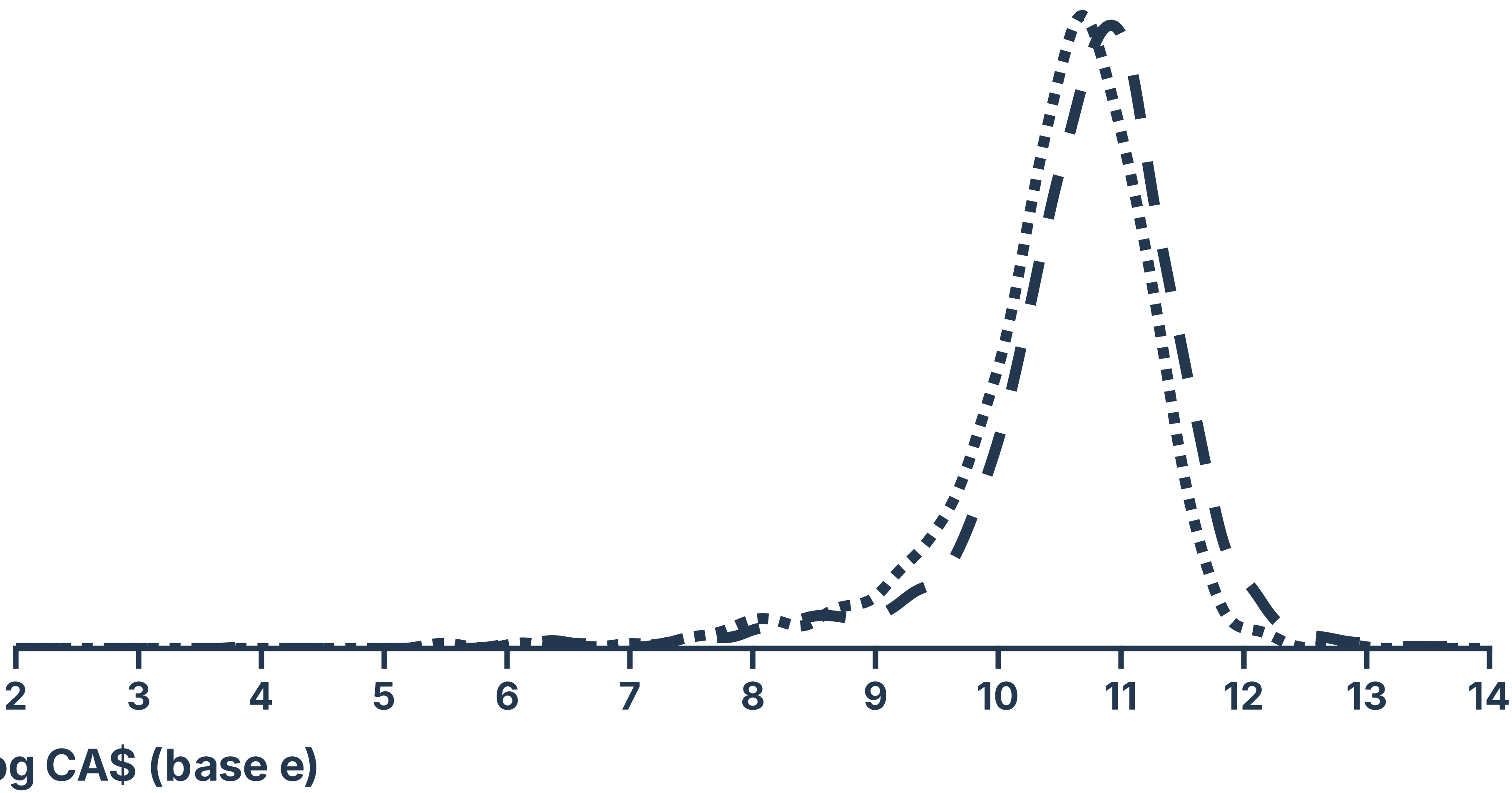
Labs with TA

- Leacock 808
  (for the rest of the term)
- Mondays, 10am-11am

Worksheet

- Check in
- Due *this Wednesday*, Jan 22 by midnight
- Peer assessments due by Monday, Jan 27

MODELING INCOME BY SEX



Note:
Canadian Income Survey (CIS) uses the Labour Force Survey (LFS) *sex* variable, which asks respondents for their sex “assigned at birth” and requires respondents to answer either “male” or “female.” While the LFS includes a *gender* item, this is not available in the CIS.

Model from last week:

Entire population has one mean and one standard deviation

$$y_i \sim \mathrm{Norm}(\mu, \sigma)$$

Regression:

Standard linear regression allows mean to vary depending on respondent

$$y_i \sim \mathrm{Norm}(\mu_i, \sigma)$$

Each observation (i) can have a different value for μ_i

Regression:

Standard linear regression allows mean to vary depending on respondent

$$y_i \sim \mathrm{Norm}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta m_i$$

$m_i = 0$ for female respondents
$m_i = 1$ for male respondents

$\mu_i = \alpha$ for female respondents
$\mu_i = \alpha + \beta$ for male respondents

Regression:

Standard linear regression allows mean to vary depending on respondent

$$
\begin{aligned}
y_i &\sim \mathrm{Norm}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta m_i \\
\alpha &\sim \mathrm{Norm}(0, 30) \\
\beta &\sim \mathrm{Norm}(0, 30) \\
\sigma &\sim \mathrm{Unif}(0, 50)
\end{aligned}
$$

Prior for each of the parameters

Regression:

Standard linear regression allows mean to vary depending on respondent

Stochastic relationship

Deterministic relationship

$$
\begin{aligned}
y_i &\sim \mathrm{Norm}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta m_i \\
\alpha &\sim \mathrm{Norm}(0, 30) \\
\beta &\sim \mathrm{Norm}(0, 30) \\
\sigma &\sim \mathrm{Unif}(0, 50)
\end{aligned}
$$

MODELING INCOME BY SEX

No predictors

$$
\boxed{\begin{aligned}
y_i &\sim \mathrm{Norm}(\mu, \sigma) \\
\mu &\sim \mathrm{Norm}(0, 30) \\
\sigma &\sim \mathrm{Unif}(0, 50)
\end{aligned}}
$$

One predictor

$$
\boxed{\begin{aligned}
y_i &\sim \mathrm{Norm}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta m_i \\
\alpha &\sim \mathrm{Norm}(0, 30) \\
\beta &\sim \mathrm{Norm}(0, 30) \\
\sigma &\sim \mathrm{Unif}(0, 50)
\end{aligned}}
$$

Same model, three* representations:

$y_i \sim \mathrm{Norm}(\mu_i, \sigma)$		$y_i = \alpha + \beta m_i + \varepsilon_i$
$\mu_i = \alpha + \beta m_i$	$y_i \sim \mathrm{Norm}(\alpha + \beta m_i, \sigma)$	$\varepsilon_i \sim \mathrm{Norm}(0, \sigma)$
$\alpha \sim \mathrm{Norm}(0, 30)$	$\alpha \sim \mathrm{Norm}(0, 30)$	$\alpha \sim \mathrm{Norm}(0, 30)$
$\beta \sim \mathrm{Norm}(0, 30)$	$\beta \sim \mathrm{Norm}(0, 30)$	$\beta \sim \mathrm{Norm}(0, 30)$
$\sigma \sim \mathrm{Unif}(0, 50)$	$\sigma \sim \mathrm{Unif}(0, 50)$	$\sigma \sim \mathrm{Unif}(0, 50)$

* *at least* three

JOINT POSTERIOR

When we estimate this model, we get a single ***joint*** posterior distribution for ***all three*** parameters:

$$\Pr(\alpha, \beta, \sigma | D)$$

What can we do with a joint posterior?

JOINT POSTERIOR

$$\Pr(\alpha, \beta, \sigma | D)$$

Data:
Sample of 3,181 working adults in Canada

1. Describe the marginal posterior distributions
   $\text{Prob}(\alpha|D); \text{Prob}(\beta|D); \text{Prob}(\sigma|D)$

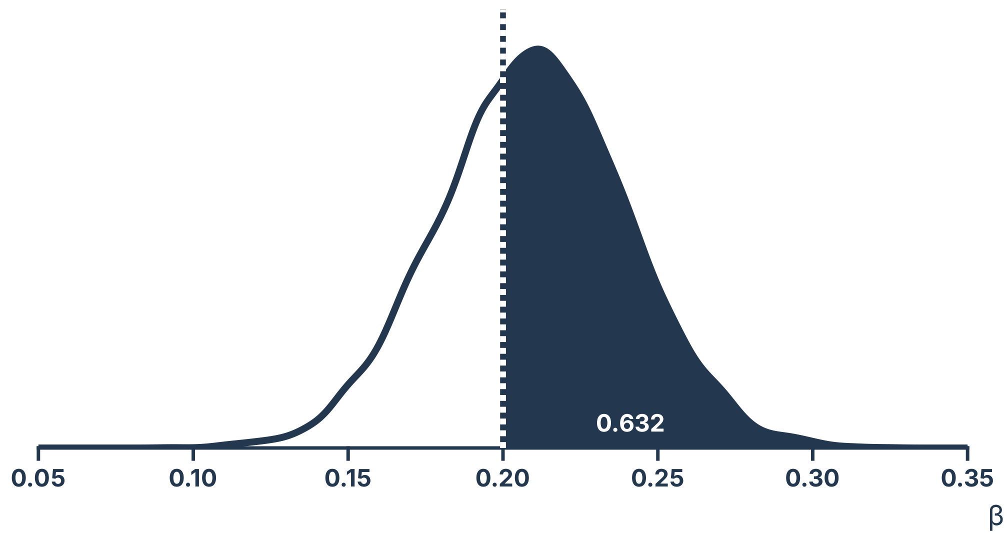
	Mean	Std. dev	2.5%	97.5%
α	10.46	0.02	10.42	10.51
β	0.21	0.03	0.15	0.27
σ	0.85	0.01	0.83	0.87

JOINT POSTERIOR

$$\Pr(\alpha, \beta, \sigma | D)$$

Data:
Sample of 3,181 working adults in Canada

1. Describe the marginal posterior distributions
   $\text{Prob}(\alpha|D); \text{Prob}(\beta|D); \Pr(\sigma|D)$
2. Describe posterior probability of theoretically relevant scenarios
   $\Pr(\beta \geq 0.2|D) = 0.628$

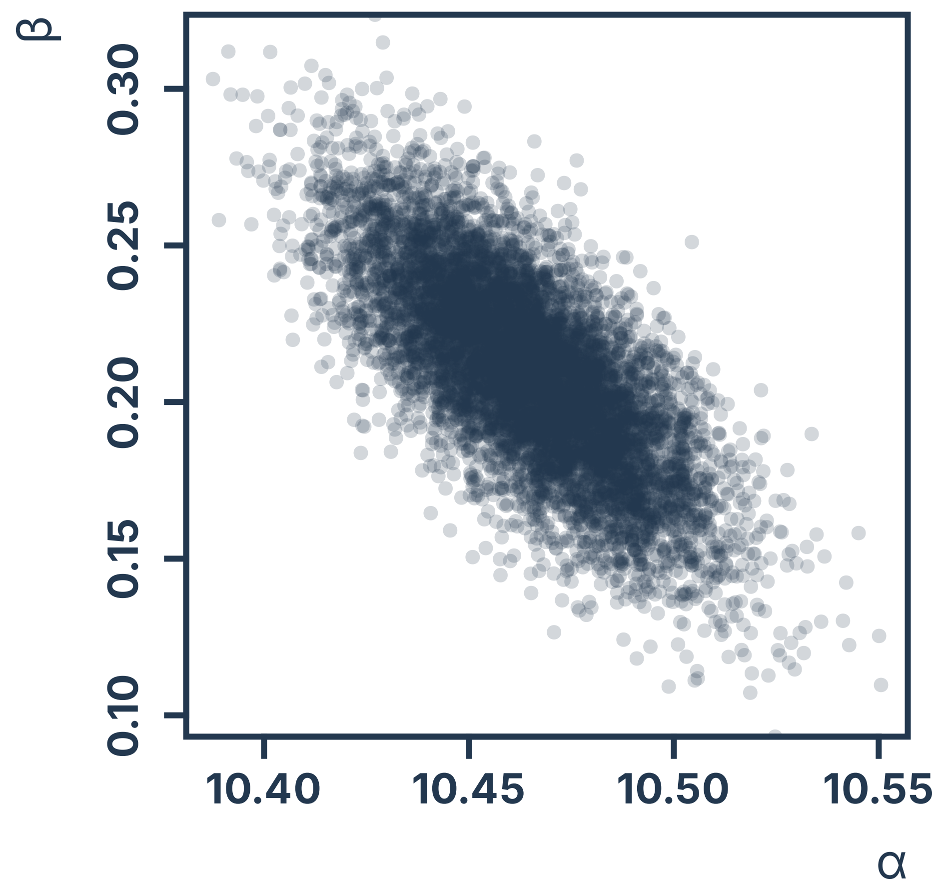


JOINT POSTERIOR

$$\Pr(\alpha, \beta, \sigma | D)$$

Data:
Sample of 3,181 working adults in Canada

1. Describe the marginal posterior distributions
   $\text{Prob}(\alpha|D); \text{Prob}(\beta|D); \Pr(\sigma|D)$
2. Describe posterior probability of theoretically relevant scenarios
   $\Pr(\beta \geq 0.2|D) = 0.628$
3. Describe the 'partial' joint posterior distribution



LOG-SCALE COEFFICIENTS

$$y_i \sim \mathrm{Norm}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta m_i$$

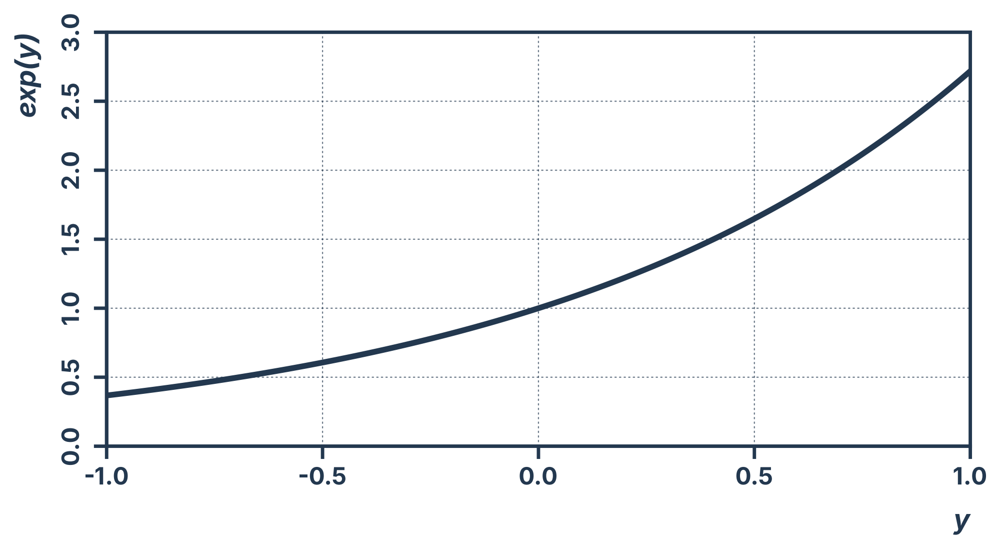
	Mean	Std. dev	2.5%	97.5%
α	10.46	0.02	10.42	10.51
β	0.21	0.03	0.15	0.27
σ	0.85	0.01	0.83	0.87

$$\exp(E(\alpha)) = e^{E(\alpha)} = e^{\hat{\alpha}} \approx 35,054$$
$$e^{\hat{\beta}} \approx 1.233$$
$$e^{\hat{\alpha}+\hat{\beta}} = e^{\hat{\alpha}} \times e^{\hat{\beta}} \approx 43,220$$

In general: if the outcome variable is on a log-scale, then exponentiating coefficient estimates ($e^{\hat{\alpha}}$) gives *multiplicative* factors

$$e^{\hat{\beta}} \approx 1.233$$

These results suggest that men make about 22.3% more than women on average



ADDING COVARIATES

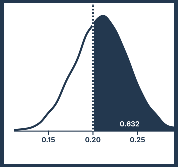
From here, we can add covariates to model income however we like

$$
\begin{aligned}
y_i &\sim \text{Norm}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_1 m_i + \beta_2 age_i + \beta_3 college_i \\
\alpha &\sim \text{Norm}(0, 30) \\
\beta_1 &\sim \text{Norm}(0, 30) \\
\beta_2 &\sim \text{Norm}(0, 30) \\
\beta_3 &\sim \text{Norm}(0, 30) \\
\sigma &\sim \text{Unif}(0, 50)
\end{aligned}
$$

Compact notation:

$$
\begin{aligned}
y_i &\sim \text{Norm}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_1 m_i + \beta_2 age_i + \beta_3 college_i \\
\alpha, \beta_1, \beta_2, \beta_3 &\sim \text{Norm}(0, 30) \\
\sigma &\sim \text{Unif}(0, 50)
\end{aligned}
$$

Image credit



Figures by Peter McMahan (source code)