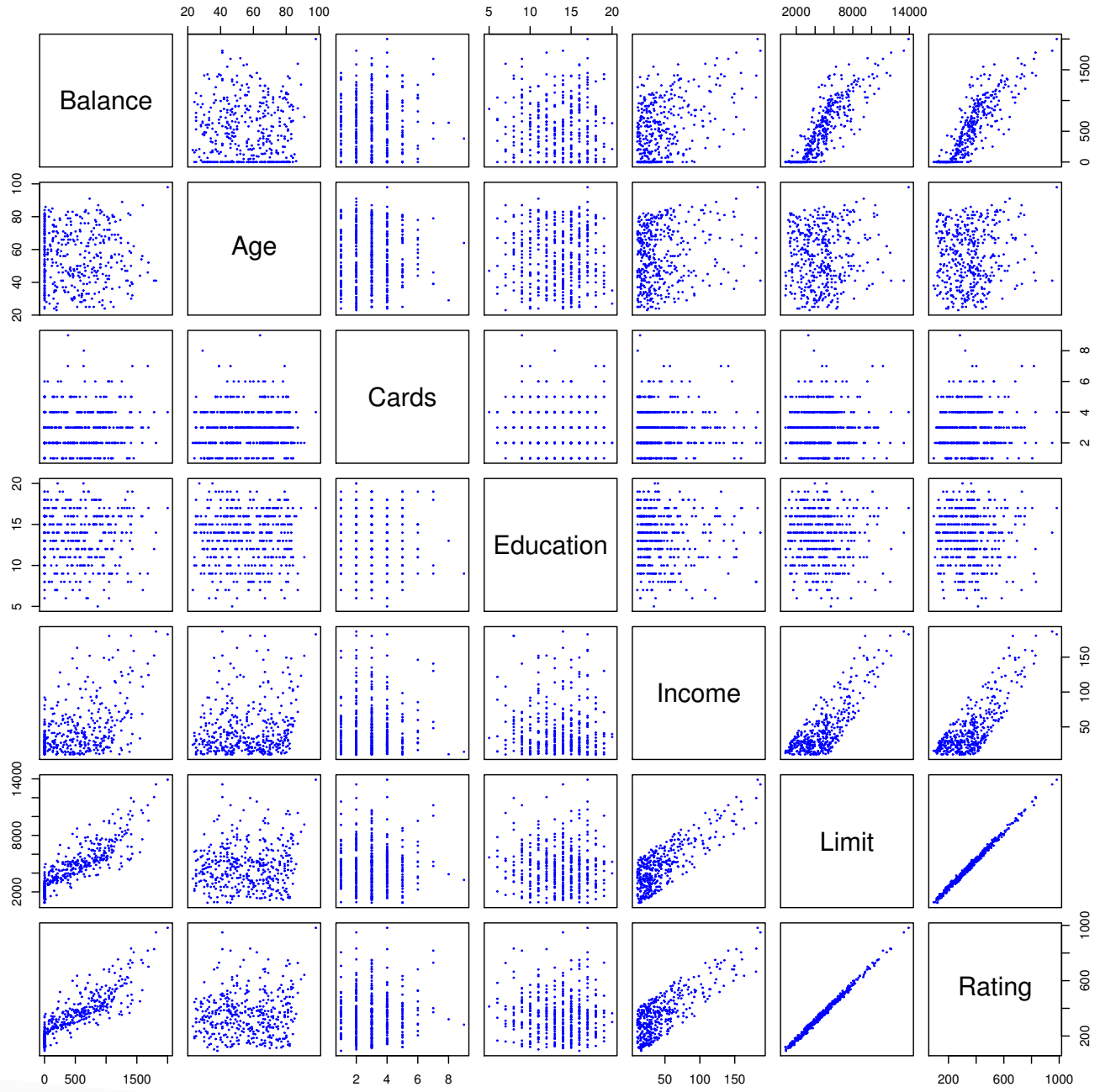
Using Multiple Linear Regression to Analyze the Credit Dataset


Balance
Age
Cards
Education
Income
Limit
Rating

Specifying the Model

Let y_i denote the balance of the i^{th} person.

$$y_i = \beta_0 + \sum_{p=1}^{P} x_{ip}\beta_p + \epsilon_i$$

$$= \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$$

$\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

Alternatively, we can write in matrix form as

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

$\mathbf{y} = (y_1, \ldots, y_n)', \boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'.$

Specifying the Model

How do we include the categorical predictors into a MLR for Balance?

For two categories (e.g. gender):

$$x_i = \begin{cases} 1 & \text{if person } i \text{ is Female} \\ 0 & \text{otherwise} \end{cases}$$

For more than two categories need one less indicator variable (e.g. ethnicity):

$$x_{i1} = \begin{cases} 1 & \text{if person } i \text{ is White} \\ 0 & \text{otherwise} \end{cases} \qquad x_{i2} = \begin{cases} 1 & \text{if person } i \text{ is Asian} \\ 0 & \text{otherwise} \end{cases}$$

Specifying the Model

How does a MLR Model help us with the credit dataset?

1. We can do predictions:

$$\hat{y}_0 = \mathbf{x}_0' \hat{\boldsymbol{\beta}}$$

2. We can easily interpret $\hat{\beta}_j$ to understand how balance interacts with other variables.

Fitting the Model

How do we obtain $\hat{\boldsymbol{\beta}}$?

1. Least Squares Approach: Let $\hat{\boldsymbol{\beta}}$ be the solution to:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 = \min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$\Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Fitting the Model

How do we obtain $\hat{\boldsymbol{\beta}}$?

2. Maximum Likelihood: Let $\hat{\boldsymbol{\beta}}$ maximize:

$$L(\boldsymbol{\beta} \mid y_1, \ldots, y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 \right\}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\}$$

$$\Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$
$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Why does this give the same answer as least squares?

Fitting the Model

How do we obtain $\hat{\boldsymbol{\beta}}$?

3. Bayesian Approach:

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{m}, \mathbf{S})$$

$$\Rightarrow \boldsymbol{\beta} \mid \mathbf{y}, \sigma^2 \sim \mathcal{N}((\mathbf{X}'\mathbf{X}/\sigma^2 + \mathbf{S}^{-1})^{-1}(\mathbf{X}'\mathbf{y}/\sigma^2 + \mathbf{S}^{-1}\mathbf{m}), (\mathbf{X}'\mathbf{X}/\sigma^2 + \mathbf{S}^{-1})^{-1})$$

Fitting the Model

How do we obtain $\hat{\sigma}^2$?

$$\hat{\sigma}^2 = \frac{1}{n-P-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

Note that this is an unbiased estimate (its not the MLE) and is often called the residual standard error.

Bayesians assume:

$$\sigma^2 \sim \mathcal{IG}(a, b)$$

$$\Rightarrow \sigma^2 \mid \boldsymbol{\beta}, \mathbf{y} \sim \mathcal{IG}\left(\frac{n}{2} + a, \frac{\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2} + b\right)$$

What Covariates to Include?

Should we just throw all the covariates in the model?

- Overfitting – Include covariates we shouldn't.
  - $\hat{\beta}$ is unbiased (use Eq. 2.50 in Rencher & Schaalje (2008)).
  - $\mathrm{Var}(\hat{\beta})$ is inflated.

What Covariates to Include?

Should we just throw all the covariates in the model?

- Underfitting – Not include covariates that we should.
  - Let $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$.

$$\begin{aligned}
\mathbb{E}\left((\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}\right) &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbb{E}(\mathbf{y}) \\
&= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2) \\
&= \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 \\
&= \text{Biased}
\end{aligned}$$

What Covariates to Include?

Best Subset Selection

1. Let $\mathcal{M}_0$ denote the model with no covariates.
2. For $p = 1, \ldots, P$
   a. Consider all models that have p covariates.
   b. Pick the "best" model and call it $\mathcal{M}_p$.
3. Pick "best" model from $\mathcal{M}_0, \ldots, \mathcal{M}_p$.

This is great but not realistic. Why?

What Covariates to Include?

Forward Selection Algorithm

1. Let $\mathcal{M}_0$ denote the model with no covariates.
2. For $p = 0, \ldots, P-1$
   a. Consider all models that add one to $\mathcal{M}_p$.
   b. Pick the "best" model and call it $\mathcal{M}_{p+1}$.
3. Pick "best" model from $\mathcal{M}_0, \ldots, \mathcal{M}_P$

What Covariates to Include?

Backward Elimination Algorithm

1. Let $\mathcal{M}_P$ denote the full model with all covariates.
2. For $p = P, p-1, \ldots, 1$
   a. Consider all models that delete one from $\mathcal{M}_p$.
   b. Pick the "best" model and call it $\mathcal{M}_{p-1}$.
3. Pick "best" model from $\mathcal{M}_0, \ldots, \mathcal{M}_P$

What Covariates to Include?

Hybrid Algorithm

Start with a null model. Add a variable then check to see if removing any of the included variables results in a “better” model.

What Covariates to Include?

What do we mean by a "best" model?

$$\text{AIC} = -2\log(\text{Like}) + 2P$$

$$C_p = \frac{1}{n}\left(\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2 + 2P\hat{\sigma}^2\right)$$

$$\text{BIC} = -2\log(\text{Like}) + P\log(n)$$

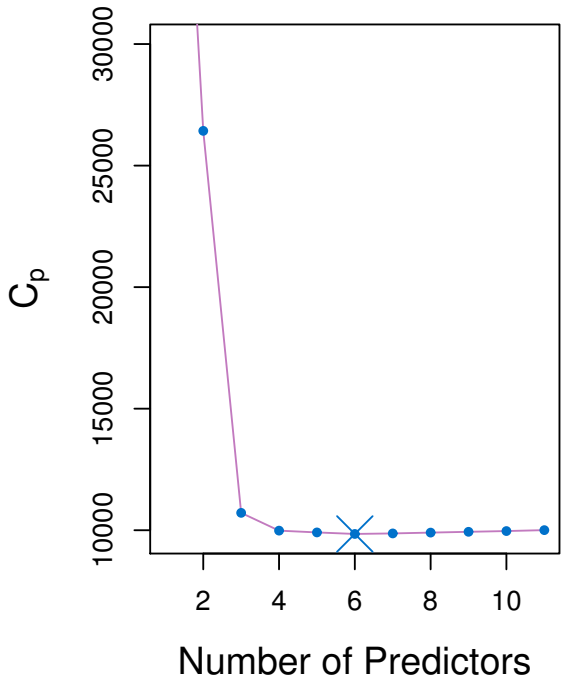
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

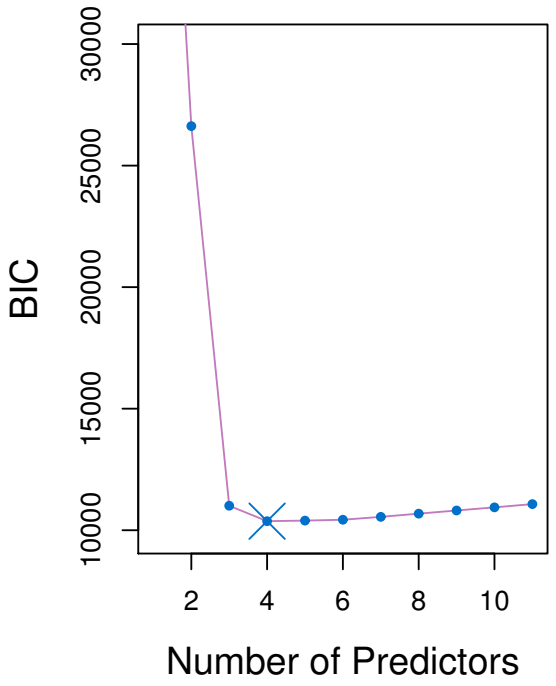
$$\text{Adjusted } R^2 = 1 - \frac{\left(\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2\right)/(n-P-1)}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)/(n-1)}$$

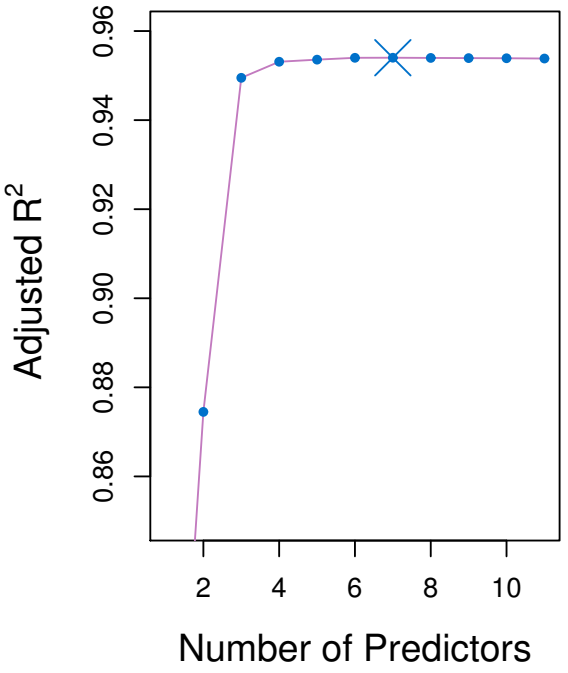
Training Set Error

Cross Validation Error

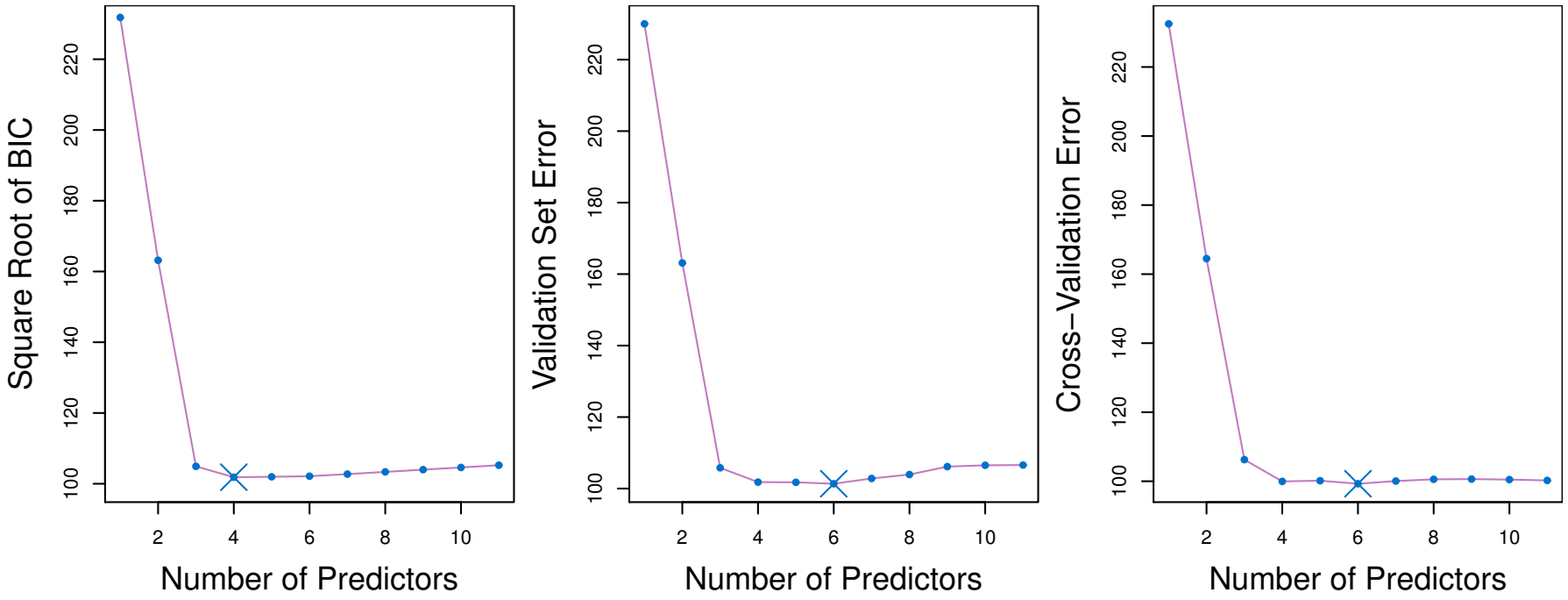
What Covariates to Include?







What Covariates to Include?



What Covariates to Include?

Bayesian Variable Selection:

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, \sigma^2 \sim \mathcal{N}\left(0, g\sigma^2 \left(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}\right)^{-1}\right)$$

$$\gamma_j = \begin{cases} 0 & \text{if } \beta_j = 0 \\ 1 & \text{if } \beta_j \neq 0 \end{cases} \sim \text{Bin}(1, \omega)$$

$$\gamma_j \mid \boldsymbol{\beta}, - \sim \text{Bin}\left(1, \frac{[\boldsymbol{\beta}_{\boldsymbol{\gamma}|\gamma_j=1} \mid \gamma_j = 1]\omega}{[\boldsymbol{\beta}_{\boldsymbol{\gamma}|\gamma_j=1} \mid \gamma_j = 1]\omega + [\boldsymbol{\beta}_{\boldsymbol{\gamma}|\gamma_j=0} \mid \gamma_j = 1](1-\omega)}\right)$$

What Covariates to Include?

Bayesian Model Selection:

$$\mathcal{M}_i \equiv \text{ Model } i \equiv \left\{ \overbrace{[\mathbf{y} \mid \boldsymbol{\theta}_i]}^{i^{th} \text{ Likelihood}} \quad , \quad \overbrace{[\boldsymbol{\theta}_i]}^{i^{th} \text{ Prior}} \right\}$$

$$[\mathcal{M}_i] \; i = 1, \ldots, M \equiv \text{ Prior Probability of Model } i$$

$$\Rightarrow [\mathcal{M}_i \mid \mathbf{y}] \propto [\mathbf{y} \mid \mathcal{M}_i][\mathcal{M}_i]$$

NASTY! → $$[\mathbf{y} \mid \mathcal{M}_i] = \int_{\boldsymbol{\Theta}} [\mathbf{y} \mid \boldsymbol{\theta}_i, \mathcal{M}_i][\boldsymbol{\theta}_i \mid \mathcal{M}_i] d\boldsymbol{\theta}_i$$

$$(\text{Bayes Factor})_{ij} = \frac{[\mathbf{y} \mid \mathcal{M}_i]}{[\mathbf{y} \mid \mathcal{M}_j]}$$

→ Closed form for linear models!

Do Any Covariates Matter?

Do any of the covariates in the credit data set explain balance?
We need to test:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_P$$
$$H_A : \text{At least one } \beta_j \neq 0$$

$$F = \frac{(\text{TSS} - \text{RSS})/P}{\text{RSS}/(n - P - 1)} \sim \mathcal{F}_{P,n-P-1}$$

Do Any Covariates Matter?

Which of the covariates in the credit data set explain balance?

We need to test:

$$H_0 : \beta_j = 0$$
$$H_A : \beta_j \neq 0$$

$$t = \frac{\hat{\beta}_j}{\mathrm{SE}(\hat{\beta}_j)} \sim t_{n-P-1}$$

$$\hat{\beta} \pm t^{\star}_{n-P-1}\mathrm{SE}(\hat{\beta}_j)$$

Remember: $\mathrm{SE}(\hat{\beta}_j) = \sqrt{\mathrm{Var}(\hat{\boldsymbol{\beta}})_{jj}}$

Do Any Covariates Matter?

How would a Bayesian tell which of the covariates in the credit data set explain balance?

Because $\Pr(\beta_j = 0 \mid \mathbf{y}) = 0$, need to use an interval (a,b) such that

$$\Pr(a \leq \beta_j \leq b \mid \mathbf{y}) = 0.95$$

How do we do predictions?

Point Estimate: $\hat{y}_0 = \mathbf{x}_0' \hat{\boldsymbol{\beta}} + \epsilon_0$

Confidence Interval:

$$\mathbf{x}_0' \hat{\boldsymbol{\beta}} \pm t^{\star}_{n-P-1} \mathrm{SE}(\hat{y}_0)$$

$$\mathrm{SE}(\hat{y}_0) = \sqrt{\mathrm{Var}(\hat{y}_0)} = \sqrt{\mathbf{x}_0' \mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) \mathbf{x}_0 + \mathrm{Var}(\epsilon_0)}$$

How do we do predictions?

How would a Bayesian do prediction?

Bayes Posterior Predictive Distribution:

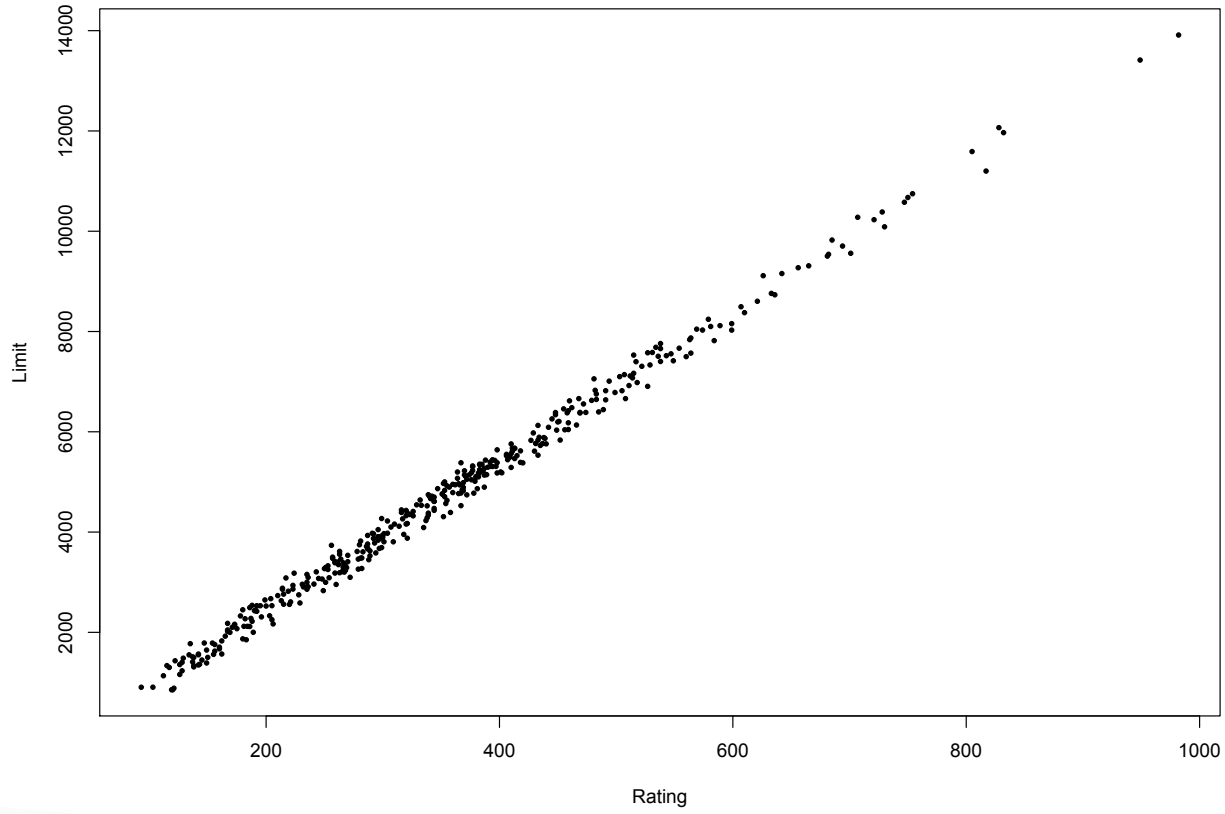
$$[y_0 \mid \mathbf{y}] = \int [y_0, \boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}] d\sigma^2 d\boldsymbol{\beta}$$

$$= \int [y_0 \mid \boldsymbol{\beta}, \sigma^2, \mathbf{y}][\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}] d\sigma^2 d\boldsymbol{\beta}$$

Intuitively, this is just a weighted sum where the weights are the posterior probability of that prediction.

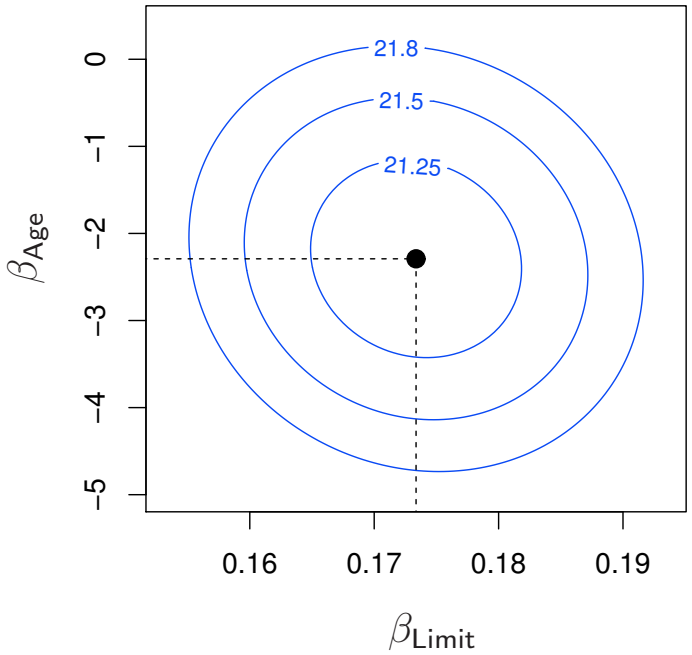
Issues Using a MLR Model for the Credit Dataset

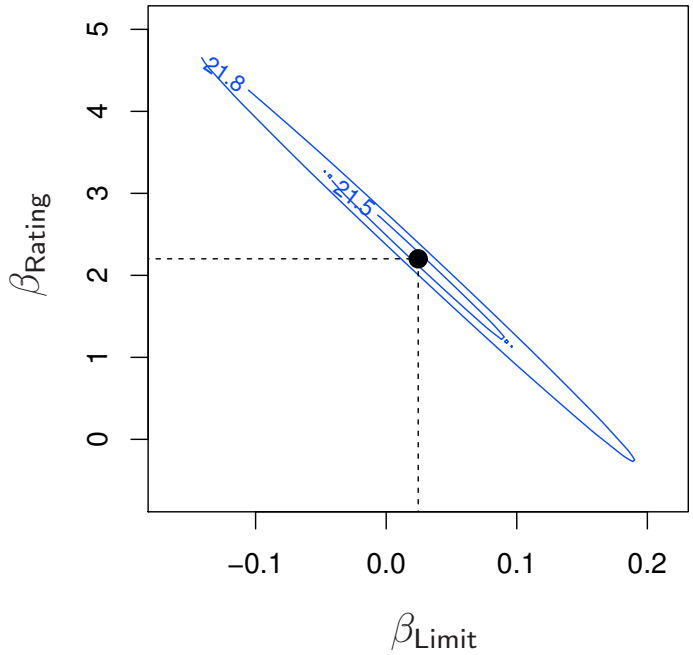
1. Collinearity



Issues Using a MLR Model for the Credit Dataset

Why is collinearity a problem?





$(\mathbf{X}'\mathbf{X})^{-1}$ is much less stable so the variability of $\hat{\boldsymbol{\beta}}$ increases.

Issues Using a MLR Model for the Credit Dataset

How do we diagnose collinearity?

1. Look at $\mathrm{Var}(\hat{\boldsymbol{\beta}})$
2. Use variance inflation factors:

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{j|-j}}$$

Issues Using a MLR Model for the Credit Dataset

How do we fix collinearity?

1. Orthogonalize:

$$\mathbf{x}_2^\star = \mathbf{x}_2 - \mathbf{x}_1 \left(\mathbf{x}_1'\mathbf{x}_1\right)^{-1} \mathbf{x}_1'\mathbf{x}_2$$

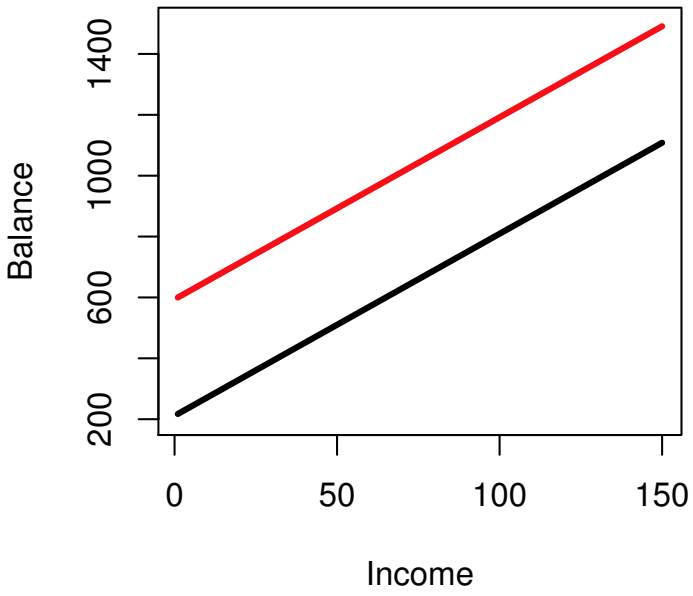
$$\mathbf{x}_1'\mathbf{x}_2^\star = 0$$

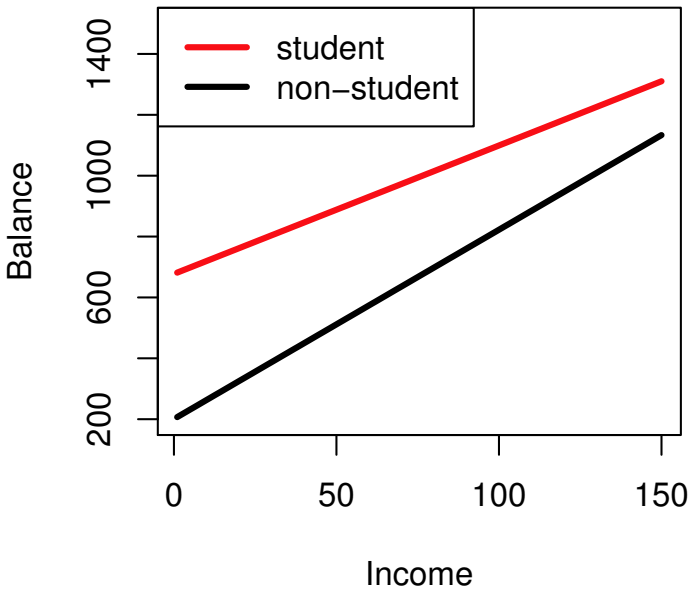
2. Bayesians would enforce prior correlation.

3. Combine them via principle components or partial least squares (more on this later).

Issues Using a MLR Model for the Credit Dataset

2. Interactions





Whether or not the customer is a student affects the slope of income.

Issues Using a MLR Model for the Credit Dataset

How do we include interactions?
Define:

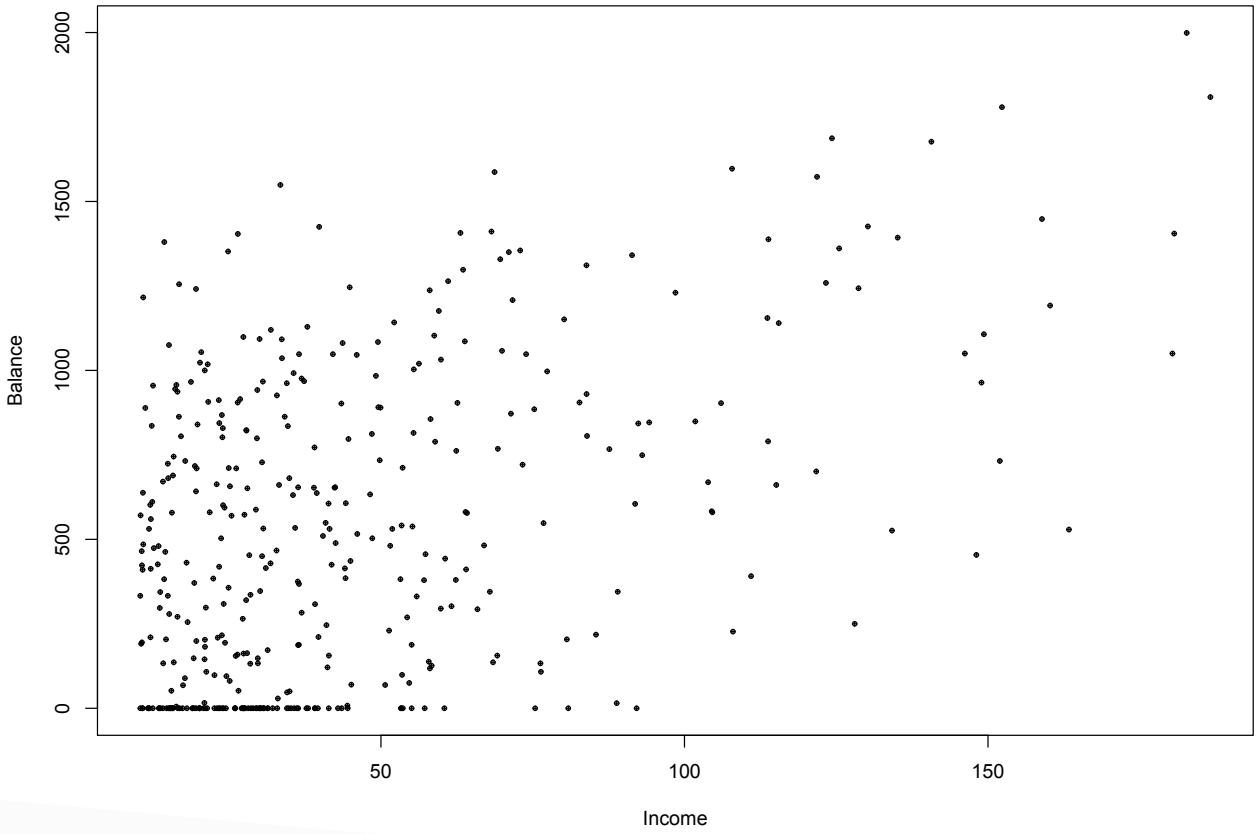
$$x_3 = x_1 \times x_2$$

For example,

$$\begin{aligned}\text{Balance}_i &= \beta_0 + \beta_1 \times \text{Inc}_i + \beta_2 \times \text{Student}_i + \beta_3 \times (\text{Inc}_i \times \text{Student}_i) \\ &= \begin{cases} \beta_0 + \beta_1 \times \text{Inc}_i & \text{if not student} \\ \beta_0 + \beta_2 + (\beta_1 + \beta_3) \times \text{Inc}_i & \text{if student} \end{cases}\end{aligned}$$

Issues Using a MLR Model for the Credit Dataset

3. Positive Support: $y_i \in [0, \infty)$



Issues Using a MLR Model for the Credit Dataset

We can't use a log-transform because $\log(0) = \infty$.

Latent (hidden) variable approach:

$$y_i = \max(0, z_i)$$

$$= \begin{cases} 0 & \text{if } z_i < 0 \\ z_i & \text{otherwise} \end{cases}$$

$$\mathbf{z} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\right)$$

Issues Using a MLR Model for the Credit Dataset

How do we estimate $\boldsymbol{\beta}$ now?

Remember: Likelihood is just the joint density function. So now we have:

$$[y_1, \ldots, y_n \mid \boldsymbol{\beta}] = \prod_{i=1}^{n} [y_i \mid \boldsymbol{\beta}]$$

$$= \left\{ \prod_{i:y_i>0} [y_i \mid \boldsymbol{\beta}] \right\} \left\{ \prod_{i:y_i=0} [y_i \mid \boldsymbol{\beta}] \right\}$$

Issues Using a MLR Model for the Credit Dataset

What is $[y_i \mid \boldsymbol{\beta}]$ when $y_i = 0$?

$$[y_i \mid \boldsymbol{\beta}] = \Pr(y_i = 0 \mid \boldsymbol{\beta}) = \Pr(z_i < 0 \mid \boldsymbol{\beta}) = \Phi\left(\frac{z_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)$$

$$[y_i \mid \boldsymbol{\beta}, \sigma^2] = \begin{cases} 0 & \text{if } y_i < 0 \\ \Phi\left(\frac{0 - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right) & \text{if } y_i = 0 \\ \phi\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right) & \text{if } y_i > 0 \end{cases}$$

Issues Using a MLR Model for the Credit Dataset

Let $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimate where our likelihood is now:

$$[y_1, \ldots, y_n \mid \boldsymbol{\beta}] = \prod_{i:y_i>0} \phi\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right) \prod_{i:y_i=0} \Phi\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)$$

Does $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$?

NO!!

So, what is it?

$$\hat{\boldsymbol{\beta}} \xrightarrow{d} \mathcal{N}(\boldsymbol{\beta}, \mathbf{I}^{-1}(\boldsymbol{\beta}))$$

Hard! Instead we should pull ourselves up by our bootstraps!

Issues Using a MLR Model for the Credit Dataset

Bootstrap Algorithm (Repeat MANY times):

For $b = 1, \ldots, B$ where B is large

1. Take a "bootstrap sample" of size K from the original n data points with replacement.
2. Calculate and retain $\hat{\boldsymbol{\beta}}^b$.

Theory tells us

$$\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}_j^b - \bar{\hat{\beta}}_j\right)^2} \approx \mathrm{SE}(\hat{\beta}_j)$$

Issues Using a MLR Model for the Credit Dataset

Bootstrapped Confidence Intervals:

Quantile Interval:

$$(\hat{\beta}_{\text{boot}}^{(0.025)}, \hat{\beta}_{\text{boot}}^{(0.975)})$$

Centered Interval: If $\hat{\beta} - \beta \approx \hat{\beta}_{\text{boot}} - \hat{\beta}$ then

$$\begin{aligned}
\Pr(\hat{\beta}_{\text{boot}}^{(0.025)} < \hat{\beta}_{\text{boot}} < \hat{\beta}_{\text{boot}}^{(0.975)}) &= \Pr(\hat{\beta}_{\text{boot}}^{(0.025)} - \hat{\beta} < \hat{\beta}_{\text{boot}} - \hat{\beta} < \hat{\beta}_{\text{boot}}^{(0.975)} - \hat{\beta}) \\
&\approx \Pr(\hat{\beta}_{\text{boot}}^{(0.025)} - \hat{\beta} < \hat{\beta} - \beta < \hat{\beta}_{\text{boot}}^{(0.975)} - \hat{\beta}) \\
&= \Pr(2\hat{\beta} - \hat{\beta}_{\text{boot}}^{(0.975)} < \beta < 2\hat{\beta} - \hat{\beta}_{\text{boot}}^{(0.025)}) \\
&= 0.95
\end{aligned}$$

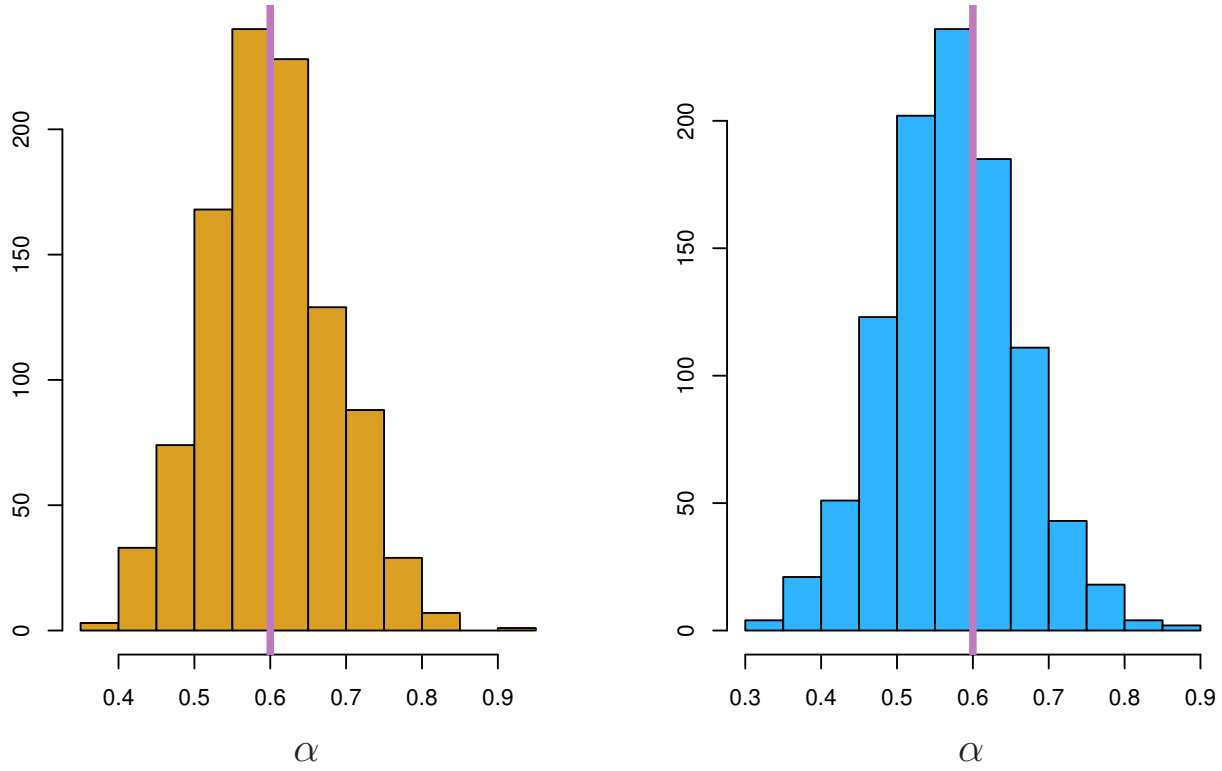
Issues Using a MLR Model for the Credit Dataset

Bootstrapped Standard Errors:

$$\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}_j^b - \bar{\hat{\beta}}_j\right)^2} \approx \mathrm{SE}(\hat{\beta}_j)$$

Bootstrapped Confidence Intervals:

Issues Using a MLR Model for the Credit Dataset



Issues Using a MLR Model for the Credit Dataset

Because we have to bootstrap to get our standard errors, shouldn't we just be Bayesian about this? Yes – in fact you should always be Bayesian about things.

How would a Bayesian do this?

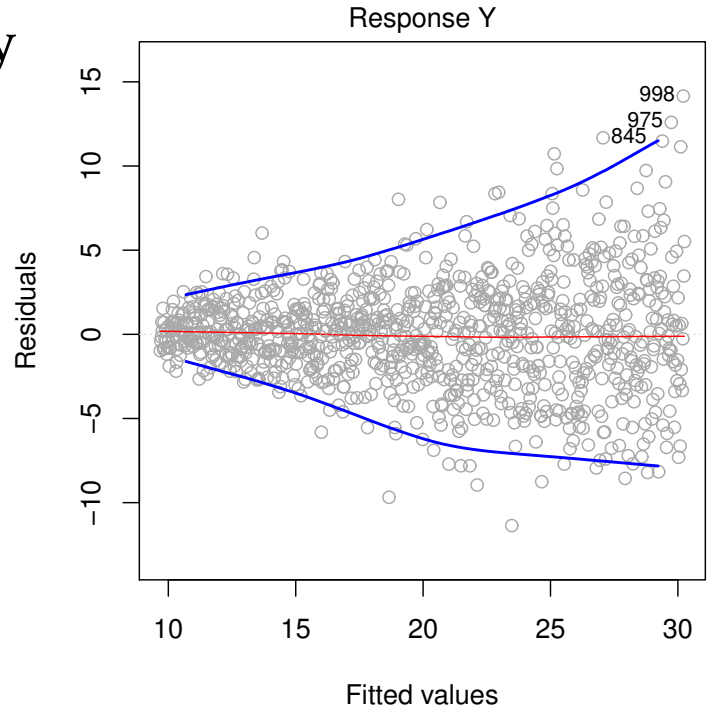
$$z_i \mid \boldsymbol{\beta}, \sigma^2 \sim \begin{cases} N^{-}\left(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2\right) & \text{if } y_i = 0 \\ y_i & \text{if } y_i > 0 \end{cases}$$

$$\boldsymbol{\beta} \mid \mathbf{z}, \sigma^2 \sim \text{ Normal (see previous)}$$

$$\sigma^2 \mid \boldsymbol{\beta}, \mathbf{z} \sim \text{ Inverse-gamma (see previous)}$$

(Potential) Issue Using a MLR Model for the Credit Dataset

4. Correlated Error Terms or Non-Constant Variance
   i. Two observations are related (e.g. time series, spatial problems)
   ii. Heteroskedasticity



(Potential) Issue Using a MLR Model for the Credit Dataset

Mathematically, the "true" model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{V})$$

But we fit,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$$

What's the problem?

1. $\hat{\boldsymbol{\beta}}$ is still unbiased
2. $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ is wrong!

(Potential) Issue Using a MLR Model for the Credit Dataset

	Range	0.2	0.4	0.6	0.8	1
Independent	Errors	0.94	0.81	0.68	0.60	0.52
Correlated	Errors	0.96	0.95	0.94	0.93	0.93

(Potential) Issue Using a MLR Model for the Credit Dataset

What do we do?

Recall, from your linear algebra class, the Cholesky decomposition (square root of a matrix) is:

$$\mathbf{V} = \mathbf{L}\mathbf{L}'$$

$$\mathbf{L} = \text{ lower triangular}$$

Transform (decorrelate) the data,

$$\begin{aligned} \mathbf{L}^{-1}\mathbf{y} &= \mathbf{L}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{L}^{-1}\boldsymbol{\epsilon} \\ &= \mathbf{X}^{\star}\boldsymbol{\beta} + \boldsymbol{\epsilon}^{\star} \\ \text{Var}(\boldsymbol{\epsilon}^{\star}) &= \mathbf{L}^{-1}\mathbf{V}(\mathbf{L}')^{-1} = \mathbf{I} \end{aligned}$$

(Potential) Issue Using a MLR Model for the Credit Dataset

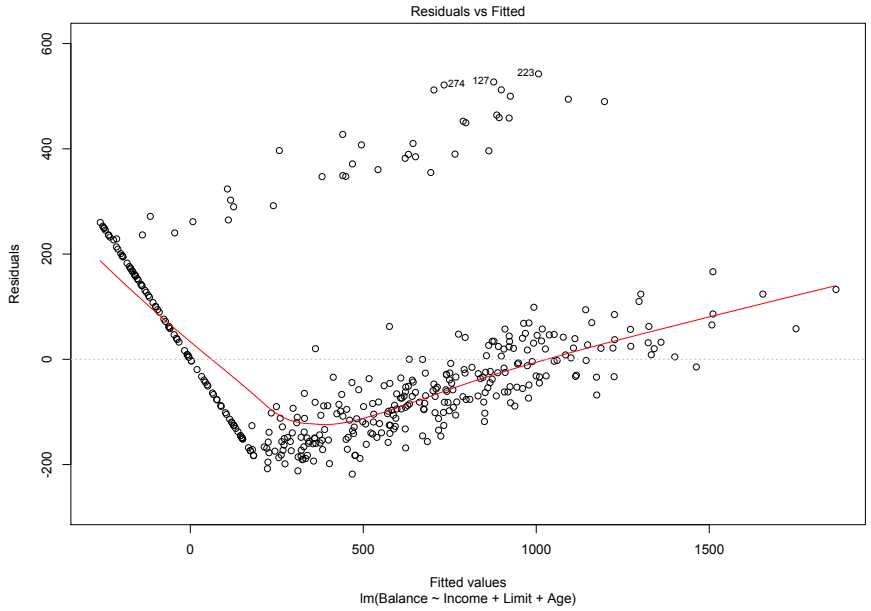
$$
\begin{aligned}
\Rightarrow \hat{\boldsymbol{\beta}} &= \left((\mathbf{X}^{\star})'\mathbf{X}^{\star}\right)^{-1}(\mathbf{X}^{\star})'\mathbf{y}^{\star} \\
&= \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\
\operatorname{Var}\left(\hat{\boldsymbol{\beta}}\right) &= \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}
\end{aligned}
$$

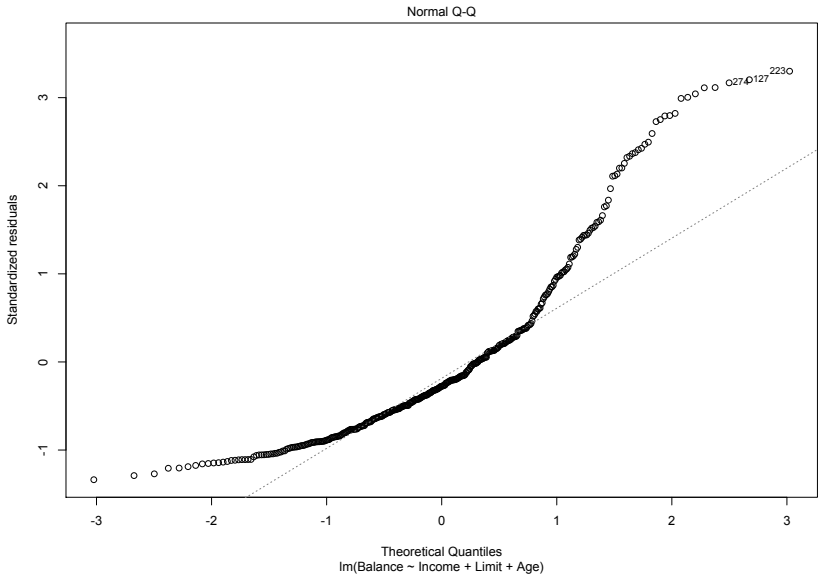
Inference can now proceed as usual (sort of).

What is $\mathbf{V}$? (More on this later in the class)

(Potential) Issue Using a MLR Model for the Credit Dataset

5. Heavy Tailed Errors (Outliers and Leverage Points)





(Potential) Issue Using a MLR Model for the Credit Dataset

5. Heavy Tailed Errors (Outliers and Leverage Points)

- Outliers - y_i is far from $\hat{y}_i$
  - Doesn't (necessarily) change $\hat{\boldsymbol{\beta}}$
  - Certainly inflates $\hat{\sigma}^2$
- Leverage Points - $\mathbf{x}_i$ is far from $\bar{\mathbf{x}}_i$
  - Has a major influence ("leverage") on $\hat{\boldsymbol{\beta}}$
  - Doesn't necessarily inflate $\hat{\sigma}^2$

(Potential) Issue Using a MLR Model for the Credit Dataset

How do we diagnose heavy-tailed errors?

1. Graphically
   - Fitted (predicted) vs. Residual Plots
   - QQ plots

2. Numerically

$$\text{Studentized (Standardized) Residual} = \frac{\hat{y}_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}}{\text{SE}(\hat{y}_i)}$$

$$\text{Leverage}(i) = \mathbf{H}_{ii} = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')_{ii}$$

$$\bar{\mathbf{H}}_{ii} = \frac{(P+1)}{n}$$

(Potential) Issue Using a MLR Model for the Credit Dataset

How do we diagnose heavy-tailed errors?

1. Graphically
   - Fitted (predicted) vs. Residual Plots
   - QQ plots

2. Numerically

$$\text{Cooks Distance}(i) = \frac{1}{P}\left(\frac{y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}}{\text{SE}(\hat{y}_i)}\right)^2\left(\frac{\mathbf{H}_{ii}}{(1-\mathbf{H}_{ii})}\right)$$

(Potential) Issue Using a MLR Model for the Credit Dataset

How do account for heavy-tailed errors?

M-Estimation

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} d(y_i, \mathbf{x}_i'\boldsymbol{\beta}),\ d \equiv \text{distance}$$

Hubers Method

$$d = \begin{cases} \left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)^2 & \text{if } \left|\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right| \leq c \\ c\left|\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right| - \frac{c^2}{2} & \text{otherwise.} \end{cases}$$

To minimize this, we can cast this as a heteroskedastic regression problem where $\mathbf{V}$ depends on the residual. Then, solve it iteratively (iteratively reweighted least squares).

(Potential) Issue Using a MLR Model for the Credit Dataset

t-Distributed Errors:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i \qquad \epsilon_i \overset{iid}{\sim} t_\nu(0, \sigma^2)$$

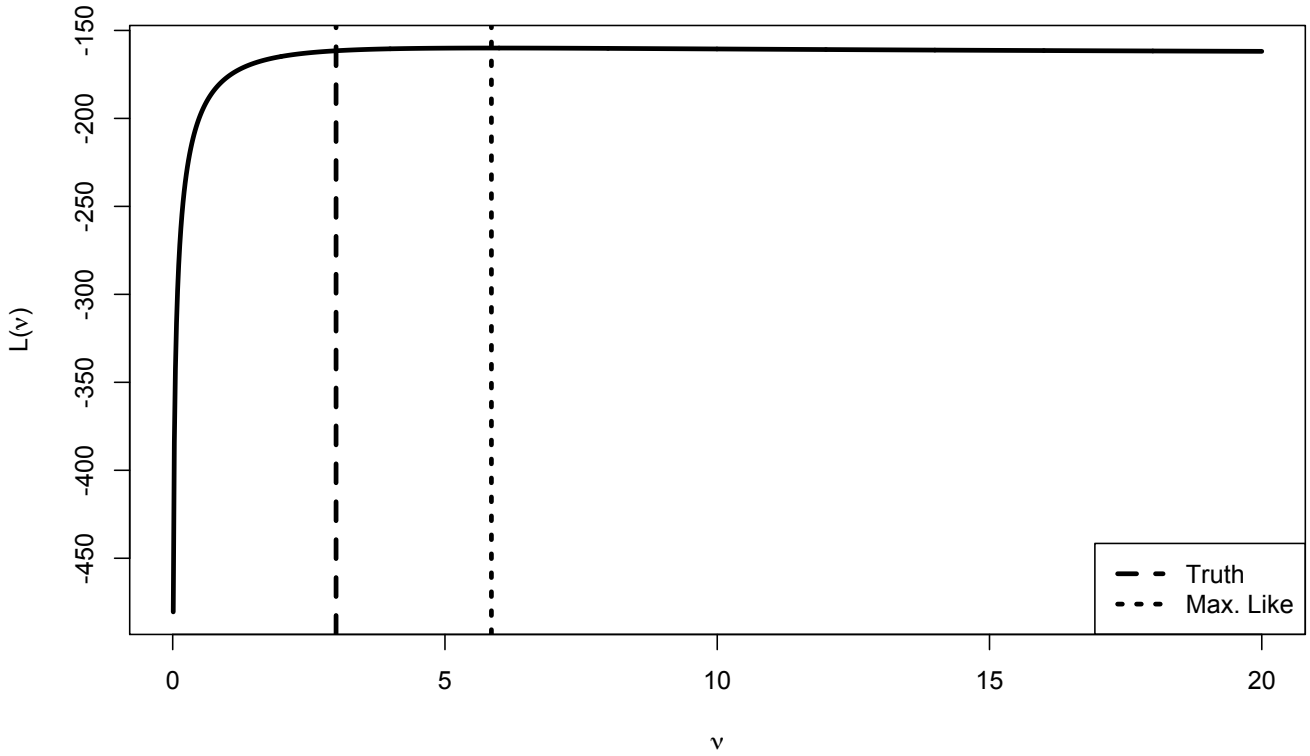
$$L(\boldsymbol{\beta}, \sigma) = \prod_{i=1}^{n} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu\sigma^2}} \left(1 + \frac{1}{\nu\sigma^2}\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2\right)^{-\frac{\nu+1}{2}}$$

Note:

1. σ^2 is not the "variance" but it's a "scale" parameter.
2. $\mathbf{x}_i'\boldsymbol{\beta}$ is the center (same as with the normal case)
3. ν is hard to estimate (identify), its typically fixed or given a strong informative prior (like a discrete prior)

(Potential) Issue Using a MLR Model for the Credit Dataset

t-Distributed Errors:



(Potential) Issue Using a MLR Model for the Credit Dataset

The Bayesian Way: Recall the identities,

If $X \mid Y \sim \mathcal{N}(0, Y^{-1})$ and $Y \sim \mathcal{G}\text{amma}(\nu/2, \nu/2)$ then the marginal distribution for $X \sim t_\nu(0, 1)$.

If $X \mid Y, \mu, \sigma^2 \sim \mathcal{N}(\mu, Y^{-1}\sigma^2)$ and $Y \sim \mathcal{G}\text{amma}(\nu/2, \nu/2)$ then the marginal distribution for $X \mid \mu, \sigma^2 \sim t_\nu(\mu, \sigma^2)$.

(Potential) Issue Using a MLR Model for the Credit Dataset

The Bayesian Way:

$$
\begin{aligned}
y_i &\overset{ind}{\sim} \mathcal{N}\left(\mathbf{x}_i'\boldsymbol{\beta}, \lambda_i^{-1}\sigma^2\right) \\
[\boldsymbol{\beta}, \sigma^2] &\propto \sigma^{-2} \\
\lambda_i &\overset{iid}{\sim} \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right) \\
\Rightarrow y_i \mid \boldsymbol{\beta}, \sigma^2 &\overset{ind}{\sim} t_\nu(\mathbf{x}_i'\boldsymbol{\beta}, \sigma)
\end{aligned}
$$

(Potential) Issue Using a MLR Model for the Credit Dataset

Gibbs Sampler:

$$\boldsymbol{\beta} \mid - \sim \text{Normal with GLS Estimate}$$

$$\sigma^2 \mid - \sim \mathcal{IG}\left(\frac{n}{2}, \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Lambda}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2}\right)$$

$$\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_n)$$

$$\lambda_i \mid - \overset{ind}{\sim} \mathcal{G}\left(\frac{\nu + 1}{2}, \frac{\nu + \sigma^{-1}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2}{2}\right)$$

Expectations for Credit Analysis

1. Variable Selection
2. Interactions
3. Validation of Predictions
4. Validation of Uncertainty Estimates

How would I analyze the credit dataset?

1. Forward variable selection – I like forward selection because it, typically, results in small models which are easy to interpret. I'd choose my model by comparing confidence interval width on my hold-out data because prediction is important.

2. Interactions – I think interactions b/n continuous variables are hard to interpret so I'd only consider interactions between categorical variables and continuous variables.

3. Validation of Predictions – I pick 10% of my data to leave out as a "test data set" and use it to choose my model.

4. Validation of Uncertainty Estimates – Make sure my final model has 95% coverage and small prediction interval width.

5. NOT EXPECTED – I'd also do the latent variable trick to account for all the zeros.