# Regularization:

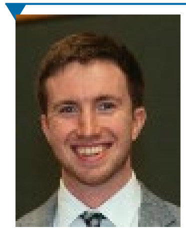
## Shrinkage Methods

June 28, 2021
Michael Regier
Tim Hagan

SERVE | ADD VALUE | INNOVATE

# Outline

- Presenters
- Motivation
  - Why Regularization
  - Example: Brief introduction
- Foundational Ideas
- Methodological Framework
  - Background
  - Ridge Regression
  - Lasso Regression
  - Elastic Net Regression
- Example: 1974 Motor Trend US Data



# Tim Hagan

## Senior Data Scientist, Insurance Analytics

### BACKGROUND

Tim Hagan is a Senior Data Scientist with Verisk Analytics, ISO Insurance Analytics, based in Buffalo Grove.  He has 5+ years of experience working in analytics, leading projects in the conversation and insurance spaces. He brings a fresh perspective to insurance analytics and has worked on mainly personal lines projects.

Prior to Verisk, Tim worked as a Sr. Behavioral Science Analyst. Most of his work involved NLP and putting structure around unstructured conversations.

### EDUCATION

BA in Psychology; Business/Economics, Wheaton College, IL, USA
MS in Statistics; Texas A&M, College Station, TX, USA (Currently Pursuing)

### SELECT EXPERIENCES

**Property**

- Build models to improve PPC effectiveness and assess performance
- Organizes internal and external data to provide complete picture of property loss experience
- Define requirements for and works with companies providing analysis files
- Define requirements for data onboarding specialists to create property analytic object

**Auto**

- Built control models used for:
- Building and improving new and current products

**Conversation Analytic Suites**

- Led and contributed to deployments to deliver customized text analytics to large companies in the following spaces:
  - Healthcare
  - Education
  - Insurance
  - Hospitality
  - Internet
  - Banking

### EXPERTISE

**Functional**

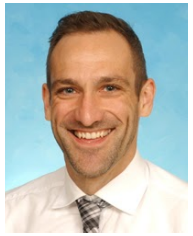
- Analytics | Data Science | Machine Learning
- Underwriting  | Rating Models
- Fire Severity | PPC
- Risk Segmentation | Risk Classification
- Econometrics

**Industry**

- Insurance
- Conversation

**Computation/Data Tools:**

- R
- Python
- SQL
- SAS
- AWS



# Michael Regier, PhD

## Director of Insurance Analytics, Personal Lines

## BACKGROUND

Michael Regier is the Director of Insurance Analytics, Personal Lines, with Verisk Analytics, ISO Insurance Analytics. His home office is Buffalo Grove, IL and is based in Anchorage, AK. He is a Ph.D. statistician with 16+ years of experience consulting, leading projects in the medical, academic, government, research, and insurance spaces. He brings a cross-functional and trans-disciplinary perspective to insurance analytics and has worked on both personal and commercial lines projects.

Prior to Verisk, Dr. Regier was a tenured Associate Professor, Dept. Biostatistics, working on the effect and correction of corrupted data structures for machine learning algorithms, effects of analytic architectures on statistical and Machine Learning methods, simulation study design, graphical interpretation of machine learning models, omics and clinical research. Other areas of research have included causal inference, graph theory for statistical inference, likelihood theory, EM Algorithm, epidemiological modeling, propensity scoring, missing data and measurement error.

## PROFESSIONAL DESIGNATIONS AND ACTIVITIES

Michael is an active member of the American Statistical Association and the IEEE professional societies. He was on the ASA Conference on Statistical Practice steering committee (2018-2021).

He has 37 peer reviewed publications, 12 technical reports and white papers, and over 70 presentations, seminars, and workshops, and has taught over 35 courses and lecture series ranging from mathematical statistics to programming.

## EDUCATION

- Postdoctoral Fellow, McGill University, Department of Epidemiology, Biostatistics and Occupational Health.
- PhD in Statistics, University of British Columbia
- MSc in Statistics, University of British Columbia

## SELECTED PROJECTS

**ISO Cyber Risk Solution**
Analytic support for both refreshment and redevelopment of various analytic components to the ISO Cyber Risk Solution rating models. Integrated novel machine learning techniques to address stability, monitoring and maintenance. Participated in customer and regulatory conversations.

**General Liability**
Developed a solution, based on functional clustering, to identify macro-level risk groups that will provide a more pragmatic yet refined characterization of general liability class group risk curve families.

**Professional Liability**
Developed a suite of analytics supporting an ISO Underwriting R&D project, integrating best in class from software and data engineering, data science, and statistics to support product maintenance. Participated in conversations with insurance partners.

**Personal Lines**
Leading teams for updates to 360 Value RCE, expansion of SmartSource capabilities, enhancements to the Risk Analyzer suite of products. He is working with his team in the areas of novel methodologies, addressing to social dimension of machine learning, and exploring way to better interpret complex methodologies.

## INTERESTS AND EXPERTISE

- Data Science | Analytics | Machine Learning
- Underwriting | Rating Models
- Statistical Practice |Theory | Coarsened Data
- Experimental Design | Causal Inference
- Simulation Studies
- Analytic Process Assessment and Architecture
- Stochastic Processes| AI | Rational Agents
- Risk Consulting and Analysis
- Insurance: Property/Casualty
- Clinical and Population Health
- Life Sciences | Clinical Statistical Methodology
- Academia | Consulting
- Government Health Organizations
- Grant and proposal writing

# Motivation

SERVE | ADD VALUE | INNOVATE

# Settings for Regularization Methods

- Regularization methods are
  - Commonly introduced within the context of regression methods, and
  - Commonly presented as a tool for model selection.
  - Can mitigate problems associated with collinearity

- Regularization methods have utility for
  - *Prediction*: Can reduce variability while maintaining low model bias in terms of the bias-variance trade-off.
  - *Interpretability*: These methods can assist in removing irrelevant or obfuscating variables – parsimonious models.
    - Parsimony: The state of being stingy.
    - Parsimonious models: Stingy with the number of variables retained in the final model.

- Regularization as model selection
  - An option for predictor (regressor) subset selection
  - Subset selection examples: Purposeful, stepwise, AIC, BIC, F- and t-test, best subset model space search.

# Example: 1974 Motor Trend US Data

- This dataset was extracted from the 1974 Motor Trend US magazine and comprises fuel consumption and 10 aspects of automobile design and performance for 32 cars (1973-74 models).

- Formula: mpg ~ cyl + disp + hp + drat + … + carb

- 4 Methods
  - OLS
  - Ridge
  - Lasso
  - Elastic Net

- Comparison of the methods

| **Variables** | |
|---|---|
| mpg | Miles/(US) gallon |
| cyl | Number of cylinders |
| disp | Displacement (cu.in.) |
| hp | Gross horsepower |
| drat | Rear axle ratio |
| wt | Weight (1000 lbs) |
| qsec | ¼ mile time |
| vs | Engine (0 = V-shaped, 1 = straight) |
| am | Transmission (0 = automatic, 1 = manual) |
| gear | Number of forward gears |
| carb | Number of carburetors |

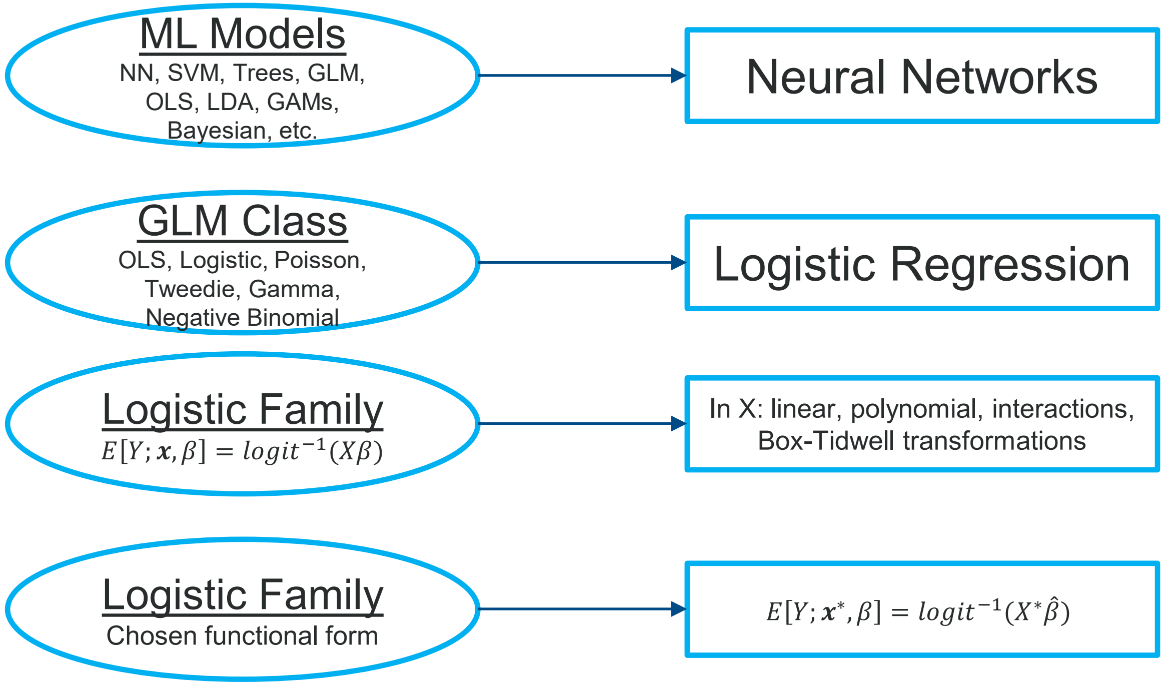
# Foundational Ideas

Verisk

SERVE | ADD VALUE | INNOVATE

# Terminology: Model Selection

- Model selection: The process of identifying a specific realization of a model from a set, class, or family of models.
  - Choosing neural networks from the set of all possible machine learning methodologies.
  - Choosing logistic regression from the class of all GLM models.
  - Choosing a specific functional form from the family of logistic regression models.
  - Choosing a specific set of estimated parameters for a specific functional form of a logistic regression model.
  - e.g. $\widehat{\mathcal{M}}: \{\mathcal{Y}, \mathcal{X}\} \rightarrow span(\mathcal{X}^*) s.t. \widehat{\Theta} = \operatorname*{argmin}_{\Theta} h\big(\widehat{\mathcal{M}}(\Theta), \mathcal{M};\ \mathcal{F}^*_{\mathcal{M}}\big)$



ML Models
NN, SVM, Trees, GLM, OLS, LDA, GAMs, Bayesian, etc.

Neural Networks

GLM Class
OLS, Logistic, Poisson, Tweedie, Gamma, Negative Binomial

Logistic Regression

Logistic Family
$E[Y; \boldsymbol{x}, \beta] = logit^{-1}(X\beta)$

In X: linear, polynomial, interactions, Box-Tidwell transformations

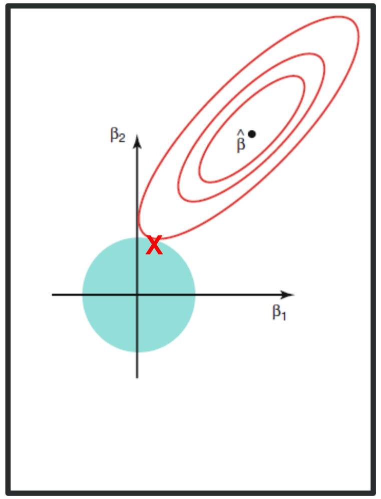
Logistic Family
Chosen functional form

$E[Y; \boldsymbol{x}^*, \beta] = logit^{-1}(X^*\hat{\beta})$

# Terminology: Regularization, Shrinkage

- *Regularization*: The process of adding information, such as a constraint, to solve ill-defined problems.
  - A common formulation for regularization is using a Lagrange Multiplier or constraint function.
  - Lagrange Multiplier

$$h(\boldsymbol{x},\lambda) = \underbrace{f(\boldsymbol{x})}_{\text{Objective Function}} + \underbrace{\lambda g(\boldsymbol{x})}_{\text{Constraint Function}}$$



  - Constraint function (example)

$$h(\boldsymbol{x},\lambda) = \underbrace{f(\boldsymbol{x})}_{\text{Objective Function}} \text{ subject to } \underbrace{g(\boldsymbol{x}) \leq s}_{\text{Constraint Function}}$$

  - The constraint functions bring clarity and focus to a problem.
    - e.g. model parsimony and/or minimize impact of latent data structure.

- *Shrinkage*: Another name for regularization when the constraint function forces (i.e. shrinks) coefficient estimates towards 0

# Modeling : The Human Component

- Check yourself at the IDE
  - We come with preconceived notions:
    - Favorite models (e.g. Model choice biases)
    - Personal thoughts/perceptions (e.g. Personal experience, media)
    - Educational predisposition (e.g. Actuarial vs. Statistical/Mathematical vs. Information Systems)

- Objective, rational model building requires:
  - Understanding, acknowledging, accepting, and challenging inherent predispositions, and
  - Embracing humility.

- We bring ourselves to every modeling exercise.
- The belief of analytic objectivity is naïve.
- Know yourself, know your data, know your question, and know the weaknesses.

# The “Linear” Model

**Origins:**

- The concept of a linear model is based on the linear algebra concepts of linear spaces, spans, and linear combinations of vectors: $\vec{w} = \sum_{i=1}^{n} a_i \vec{x}_i$.

**Examples:**

- OLS: $E[Y;\boldsymbol{x},\beta] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p + \epsilon = X\beta + \epsilon$

- GLM: $E[Y;\boldsymbol{x},\beta] = g^{-1}\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p\right) = g^{-1}(X\beta)$

- Polynomial OLS: $E[Y;\boldsymbol{x},\beta] = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \cdots + \beta_d x_1^d + \epsilon$

- Step Function OLS: $E[Y;\boldsymbol{x},\beta] = E[Y;\boldsymbol{x},\beta] = \beta_0 + \beta_1 C(x_1) + \beta_2 C(x_1) + \cdots + \beta_k C(x_1) + \epsilon$, where
  - $C_0(x_1) = I(x_1 < c_1)$
  - $C_1(x_1) = I(c_1 \leq x_1 < c_2)$
  - …
  - $C_k(x_1) = I(c_k \leq x_1)$

- Fractional Polynomial GLM: $E[Y;\boldsymbol{x},\beta] = g^{-1}[\phi(x;\beta,q)]$, where
  - $\phi(x;\beta,q) = \sum_{j=0}^{m} \beta_j h_j(x)$, where $h_j(x) = \begin{cases} x^{q_j}, & if\ q_j \neq q_{j-1} \\ x^{q_{j-1}} log x, & if\ q_j = q_{j-1} \end{cases}$

- Basis Expansion OLS: $E[Y;\boldsymbol{x},\beta] = \beta_0 + \beta_1 b_2(x_1) + \beta_2 b_2(x_1) + \cdots + \beta_k b_k(x_1) + \epsilon$

- Spline (GLM form): $E[Y;\boldsymbol{x},\beta] = \beta_0 + \beta_1 h_2(x_1) + \beta_2 h_2(x_1) + \cdots + \beta_k h_k(x_1) + \epsilon$

- CART: $f(x) = \sum_{m=1}^{M} c_m I(x \in \mathbb{R}_m)$

*Components of a linear model*

$$\underbrace{E[Y;\boldsymbol{x},\beta]}_{\text{Random Component}} = \underset{\substack{\uparrow \\ \text{Inverse Link Function}}}{g^{-1}} \underbrace{(X\beta)}_{\text{Systematic Component}}$$

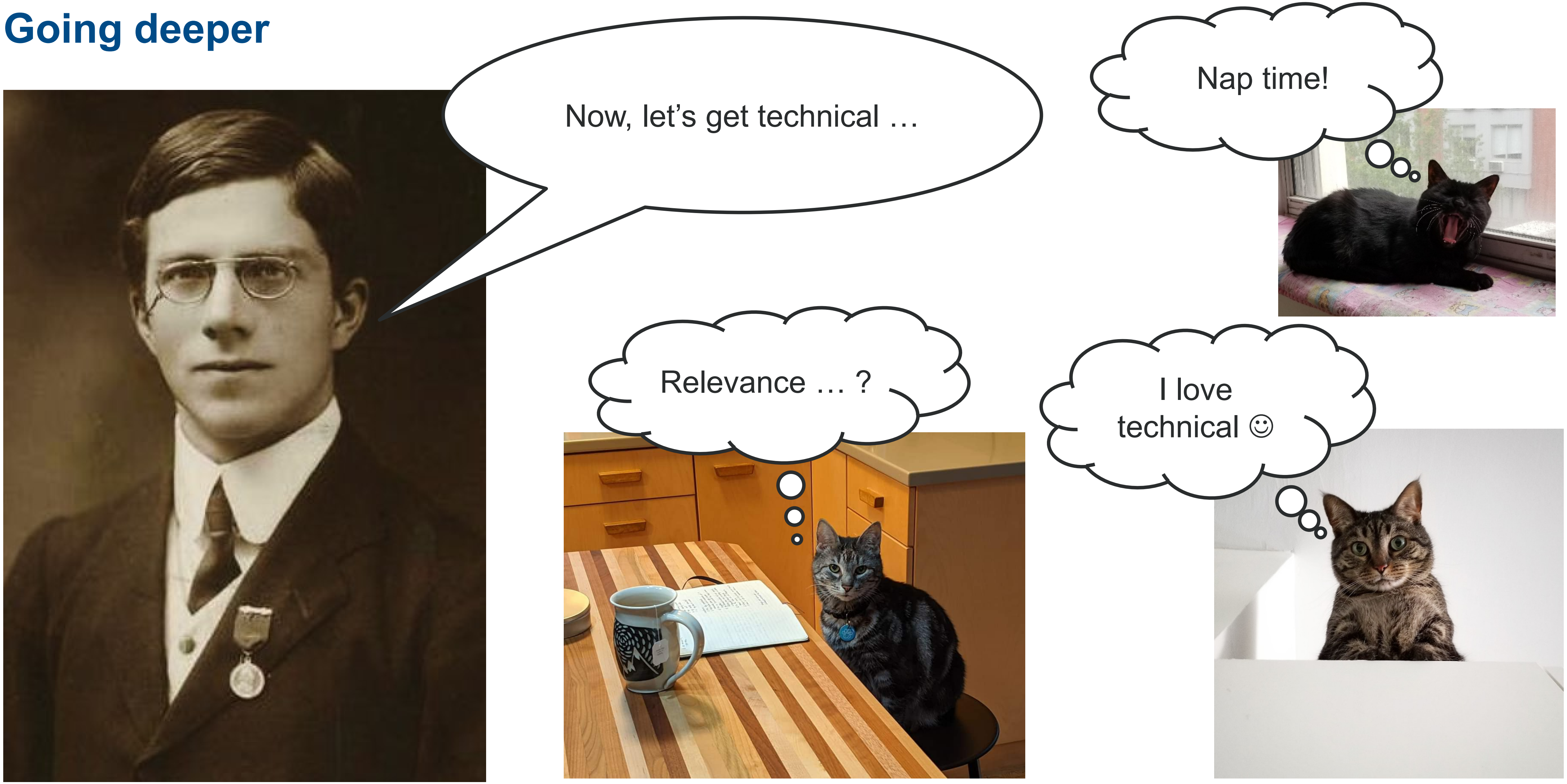
# MSE: A Familiar Model Assessment Friend

- Recall: The mean squared error (MSE) for an *estimator W of the parameter θ* is defined as

$$\underbrace{E_\theta(W-\theta)^2}_{\substack{\text{Expected squared} \\ \text{“distance” between} \\ \text{the estimator and} \\ \text{the parameter}}} = \underbrace{Var_\theta(W)}_{\substack{\text{Variance of the} \\ \text{estimator,} \\ Var_\theta(W) \\ = E_\theta\left((W-\theta)^2\right)}} + \underbrace{(Bias_\theta W)}_{\substack{\text{Squared Bias of} \\ \text{the estimator,} \\ Bias_\theta(W) \\ = E_\theta(W-\theta)}}^2$$

- Reducing the MSE of an estimator requires an overall reduction in the balance between its variance and bias.
- For an *unbiased estimator* where $Bias_\theta(W)$=*0*,
- For unbiased estimators, reducing the MSE requires only the minimization of the estimator’s variance.

# Going deeper



# Bias-Variance Trade-off

- When assessing model performance for predictive models, the MSE reveals more complexity.
- Assume:
  - $Y = f(X) + \epsilon$.
  - $E(\epsilon) = 0$.
  - $Var(\epsilon) = \sigma_\epsilon^2$, where the variance is induced by $Y \sim G(\boldsymbol{\theta})$.
  - $\hat{f}(X)$ is the estimated model, often written as $\hat{y} = \hat{f}(X)$.
  - $X = x_0$ is an input point that is fixed for which a prediction is desired.

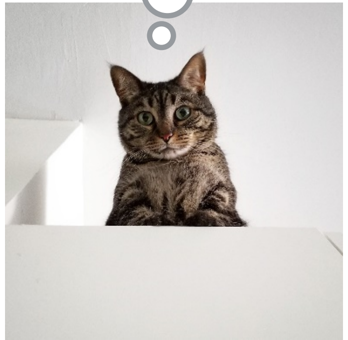
- Given an input, the prediction error is

$$Err(x_0) = E_{X,Y\sim G}\left(\left(Y - \hat{f}(X)\right)^2 \middle| X = x_0\right)$$

$$= \sigma_\epsilon^2 + E_X((E_Y(Y|x) - f(X))^2|X = x_0)$$

$$= \underbrace{\sigma_\epsilon^2}_{\text{Irreducible Error determined by } Y} + \underbrace{Bias_X(f(X))^2}_{\text{Squared Bias of the estimator } f(X)} + \underbrace{Var_X(f(X))}_{\text{Variance of the estimator } f(X)}$$

So … there are 2 variances, and a bias associated with the model!



- We can choose a model, $f(X)$ such that we reduce bias and/or variance associated with our model estimator $f(X)$.
- We cannot choose a model $f(X)$ that reduces variance in $Y$, $\sigma_\epsilon^2$. This variance is determined by the underlying distribution of the data.

# Bias-Variance Trade-off: Why should I care?

- Assume you have the same data and two models: $M_1, M_2$.
- Let both models be polynomials of degree $p_1$ and $p_2$ respectively, where $p_1 < p_2$, e.g. a quadratic and a cubic.
- Due to the lower complexity of $M_1$, this model will be a poorer fit to the data than $M_2$. This indicates that

$$Bias_X\left(f_X(x|M_1)\right) > Bias_X\left(f_X(x|M_2)\right).$$

- Although $M_1$ has larger bias than $M_2$, it is more stable with respect to changes in the underlying data. This indicates that

$$Var_X\left(f_X(x|M_1)\right) < Var_X\left(f_X(x|M_2)\right).$$

- Balancing the bias-variance trade-off is a central exercise in predictive modeling.
- It forces us to re-think how bias and variance is manifest for machine learning models
- It is reasonable that a “biased estimator” (model) may have meaningful variance reduction to produce an overall desirable predictive error, $Err(x_0)$
- Conversely, a bias free estimator – “perfect predictor” – may have so much variability as to render it useless for any sensible prediction due to high instability to the underlying data.

Ah, a slightly imperfect model may be better because it provides stability when used on “new” data – it is generalizable!

# Two Sides of the Same Model

There are two fundamental uses of a model.

1. Association

   - Identify a set of variables (explanatory variables, regressors) that *explain* the variation observed in Y, and
   - Estimate the effect of each variable in the explanation.

   - The explained variation is characterized by the systematic component and the remaining noise is “unexplained” error.

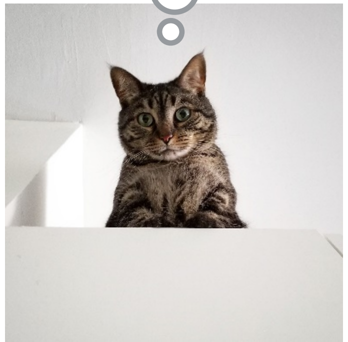
   - Associations may be purely relational to even causal, depending on the study design (architecture)

2. Prediction

   - Predict (estimate) the response value we would see given a set of known variables (e.g. predictors, regressors)

   - If I have  a new observation (e.g. house, car, person) and some attributes, what can I expect to see.

The intended use and questions affect how I use and potentially build the model!



# Regularization/Shrinkage: Methodological Framework

SERVE | ADD VALUE | INNOVATE

# OLS: Recalling Key Assumptions

- The OLS model is defined as

$$Y = \left(\sum_{j=0}^{p} \beta_j x_j\right) + \epsilon, \text{ where } x_0 \equiv 1.$$

- It is a **linear model**, because it is **linear in $\boldsymbol{\beta}$**.

- We $\underbrace{\text{characterize}}_{\text{Association}}$/$\underbrace{\text{predict}}_{\text{Prediction}}$ $\hat{y} = E(Y|X = x) = \left(\sum_{j=0}^{p} \beta_j x_j\right)$ using least squares to estimate $\beta$

$$\underbrace{\underset{\beta}{\operatorname{argmin}}}_{\text{Find the } \beta \text{ from all possible } \beta \text{ that gives the minimum value.}}\underbrace{(y - X\beta)'(y - X\beta)}_{\text{Residual sum of squares – matrix form}}$$

- Assumptions:
  1. $Y$ and $\epsilon$ are random variables, distributed from the same family.
  2. $X$ is a set of known constants. These are observed and have no random variation associated with them.
  3. $E(\epsilon) = 0$ for all observational units $\Leftrightarrow E(Y|X = x) = \left(\sum_{j=0}^{p} \beta_j x_j\right)$.
  4. $\text{Var}(\epsilon) = \sigma^2$ for all observational units $\Leftrightarrow Var(Y|X = x) = \sigma^2$.
  5. $Cov(\epsilon) = Diag(\sigma^2)$ and is 0 for all off diagonal $\Leftrightarrow Cov(Y|X = x) = Diag(\sigma^2)$ with 0 of diagonals.
  6. $X = \{x | x_j \perp x_{k,} \ \forall\, j \neq k\}$ and $\forall x_j \in X,\ x_j \notin X^* \subseteq span(X/x_j)$; regressors are independent and not a linear combination of other regressors.

# GLM: Key Assumptions

- GLMs are an extension or generalization of the OLS model.

Link Function ⟶ $$g\underbrace{(E(Y|X=x))}_{\text{Random Component}} = \underbrace{\textstyle\sum_{j=0}^{p}\beta_j x_j,}_{\text{Systematic Component}} \text{ where } x_0 \equiv 1.$$

- The GLM assumptions are
  1. $Y$ is random variables, distributed from the *exponential* family: $f(Y|\theta,\phi) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$.
  2. $X$ are observed and have no random variation associated with them.
  3. $E(Y) = b'(\theta)$ for all observational units.
  4. $\mathrm{Var}(Y) = b''(\theta)a(\phi)$ for all observational units.
  5. $Cov(Y)$has 0 for all off diagonal.
  6. $X = \{x | x_j \perp x_{k,} \ \forall\, j \neq k\}$ and $\forall x_j \in X,\ x_j \notin X^* \subseteq span(X/x_j)$; regressors are independent and not a linear combination of other regressors.
  7. *The link function is any monotonic differentiable function.*

- *Notes*
  - GLM estimation theory uses the likelihood method, hence the requirement of exponential family for estimator derivation.
  - Fitting a GLM (estimation) uses an equivalence between the maximum likelihood theory and weighted least squares: $\underset{\beta}{\operatorname{argmin}}(y - X\beta)'W(y - X\beta)$

# Ridge Regression: Definition

- From least squares estimation, we recall that $\beta$ is estimated using the argument that minimizes

$$RSS = \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p}\beta_j x_{ij}\right)^2, \text{ where } x_0 \equiv 1.$$

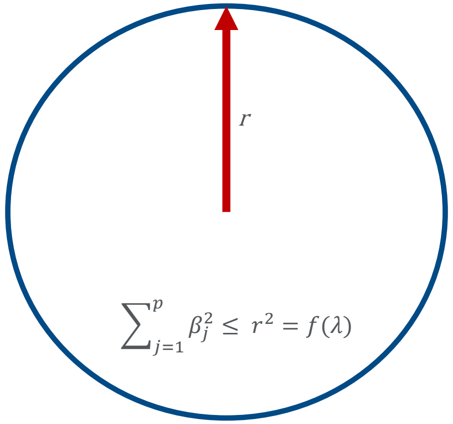
- Ridge regression constrains the minimization to the circle defined by $\sum_{j=1}^{p}\beta_j^2$.
- $\beta$ is estimated using

$$\hat{\beta}^R = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p}\beta_j^2$$

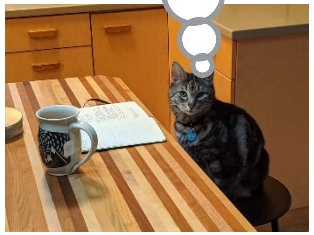
$$= \underset{\beta}{\operatorname{argmin}}\, RSS + \lambda \underbrace{\sum_{j=1}^{p}\beta_j^2}_{\text{Shrinkage penalty } (\iota_2),\ \text{where } \lambda\ \geq 0 \text{ is a tuning parameter}}$$

Shrinkage penalty only applied to $\beta_1, \ldots, \beta_p$

- With $\lambda = 0$, the ridge regression is the least squares estimate.
- As $\lambda \longrightarrow \infty$, ridge regression puts more emphasis on the penalty. Coefficients will approach 0.
- We expect that ridge regression will produce a different set of coefficients with different $\lambda$.
- $\beta_0$ is not included in the constraint because it is a measure of the mean response.



So … shrinking the size of the circle means I am increasing the penalty for the squared sum of coefficients. This is the same as $\lambda$ being large.



# Ridge Regression: Two Formulations

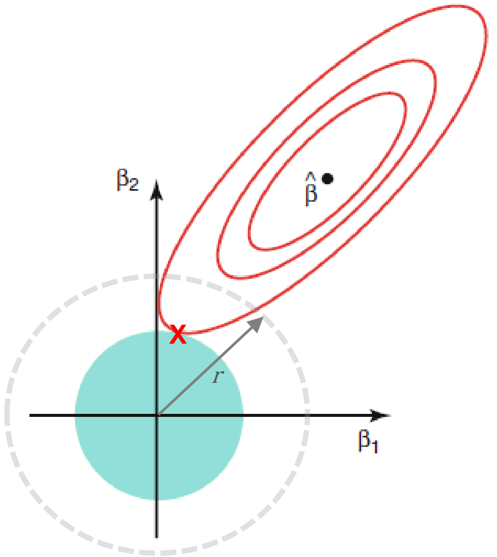
- Constrained optimization for ridge regression can be viewed through two lenses.

1. Lagrange Multiplier:

$$\hat{\beta}^R = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left( y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

2. Constraint optimization (non-Lagrange formulation)

$$\hat{\beta}^R = \underset{\beta}{\operatorname{argmin}} \textstyle\sum_{i=1}^{n} \left( y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \le s$$



2-dimensional example of how the constraint function shrinks parameters. Estimated $\beta = \{\beta_1, \beta_2\}$ is located at the X.

Note:

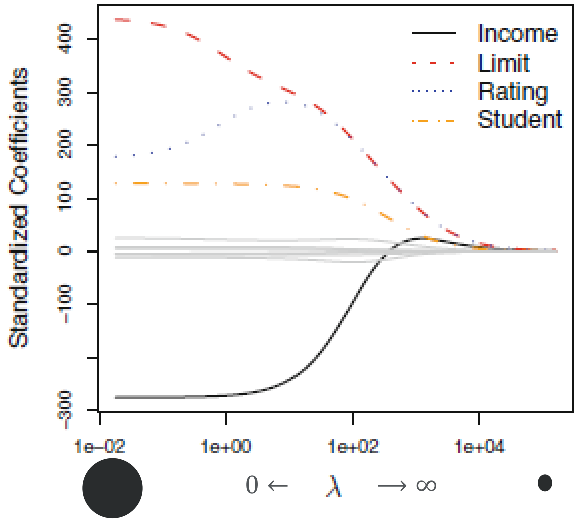
- $s = h(r)$, e.g. $s = r^2$
- As $r$ increases, it will eventually cover the unconstrained $\hat{\beta}$, which is equivalent to $\lambda \rightarrow 0$.
- The dual manner by which the constrained optimization is posed reveals that there exists a circle of size $s$ that touches a constant level of the RSS curve, and the tangent lines for each are parallel. This is equivalent of finding the stationary point (first derivative) of the Lagrange equation.

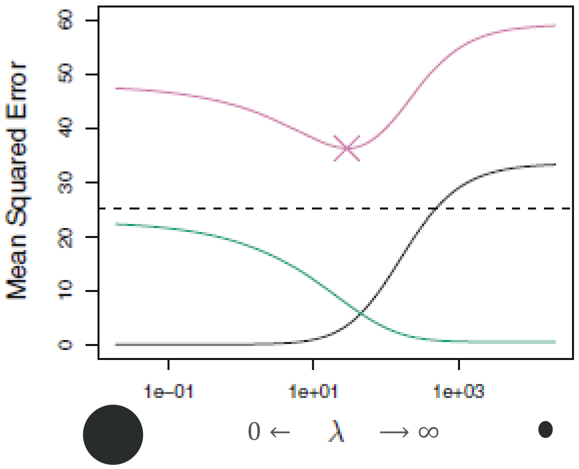
# The Effect of the Tuning Parameter

- Common plots for examining the effect of the tuning parameter are:
  - Standardized Coefficients against $\lambda$
  - MSE against $\lambda$

- Standardized Coefficients against $\lambda$
  - $\lambda = 0$ is the equivalent to the OLS solution.
  - Coefficients shrink as $\lambda \longrightarrow \infty$
  - The “function” of shrinking need not be monotonic.
  - Due to the non-monotonic nature of shrinkage, we should not expect the same selection or “mix” of coefficients for each lambda.

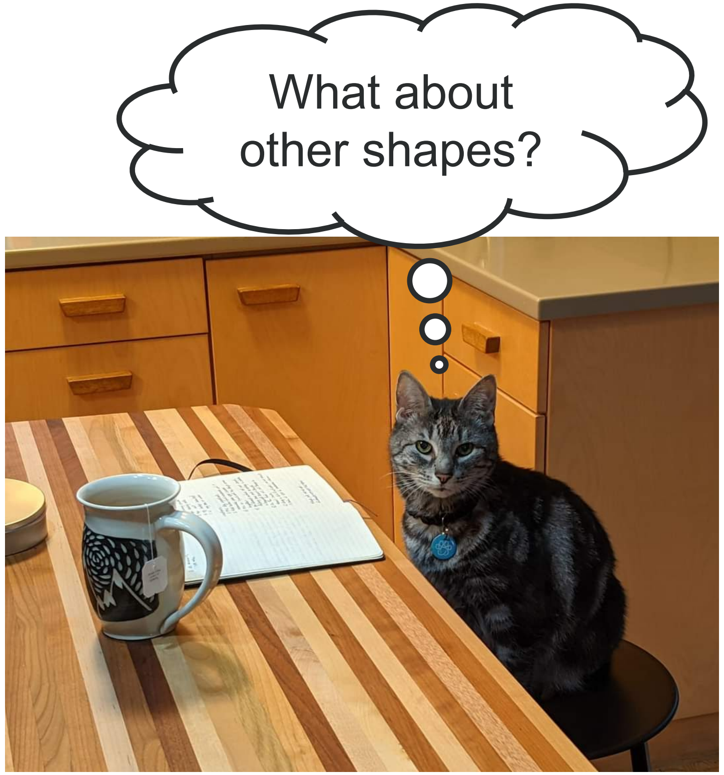
- MSE against $\lambda$
  - Black: Squared Bias
  - Green: Variance
  - Test MSE: Purple
  - In general, as $\lambda \longrightarrow \infty$ the ridge regression “fit” to the data decreases.
    - Decreased variance (green), but increased bias (black)
    - More stable, but less flexibly for characterizing the specifics of the data.
  - The test MSE reveals the optimal $\lambda$ that balances the Bias-Variance trade-off.





Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* 2nd edition. Springer.

# Methodology



# Lasso Regression: Definition

- From least squares estimation, we recall that $\beta$ is estimated using the argument that minimizes

$$RSS = \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} \beta_j x_{ij}\right)^2, \text{ where } x_0 \equiv 1.$$

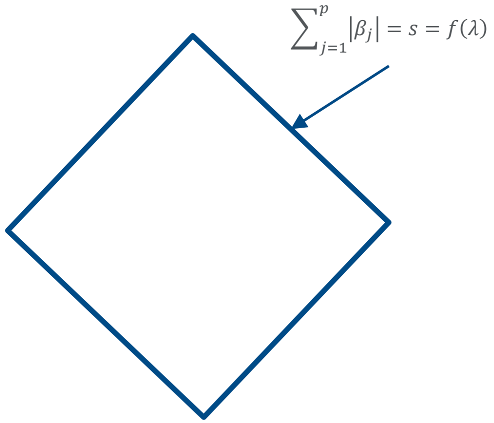
- Lasso regression constrains the minimization to the region (polytope) defined by $\sum_{j=1}^{p}|\beta_j|$.
- $\beta$ is estimated using

$$\hat{\beta}^L = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j|$$

$$= \underset{\beta}{\operatorname{argmin}}\, RSS + \underbrace{\lambda \sum_{j=1}^{p}|\beta_j|}_{\text{Shrinkage penalty } (\iota_1),\ \text{where } \lambda\ \geq 0 \text{ is a tuning parameter}}$$

Shrinkage penalty only applied to $\beta_1, \ldots, \beta_p$

- With $\lambda = 0$, the lasso produces the least squares estimate.
- As $\lambda \longrightarrow \infty$, then the lasso puts more emphasis on the penalty; coefficients will *reach* 0.
  - Lasso performs variable selection, because of this property.
- $\beta_0$ is not included in the constraint because it is a measure of the mean response.



So … shrinking the “size” of the polytope (making *s* smaller), means I am increasing the penalty for the sum of absolute coefficients. This is the same as $\lambda$ being large.

# Lasso Regression: Two Formulations

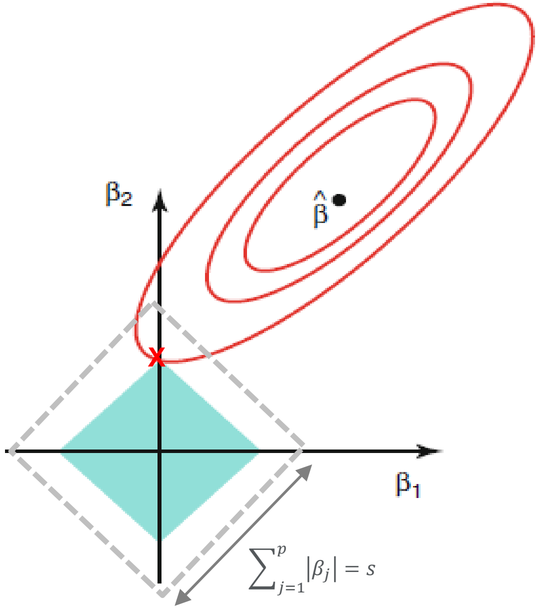
- Constrained optimization for lasso regression can be viewed through two lenses.

1. Lagrange Multiplier:

$$\hat{\beta}^L = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left( y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

2. Constraint optimization (non-Lagrange formulation)

$$\hat{\beta}^L = \underset{\beta}{\operatorname{argmin}} \textstyle\sum_{i=1}^{n} \left( y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \le s$$



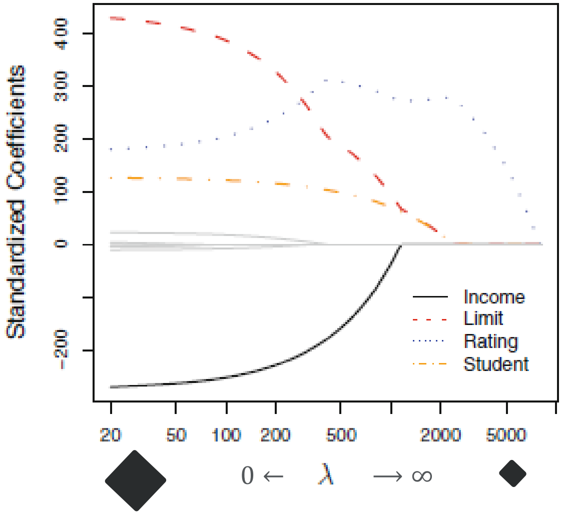
2-dimensional example of how the constraint function shrinks parameters. Estimated $\beta = \{\beta_1, \beta_2\}$ is located at the X.

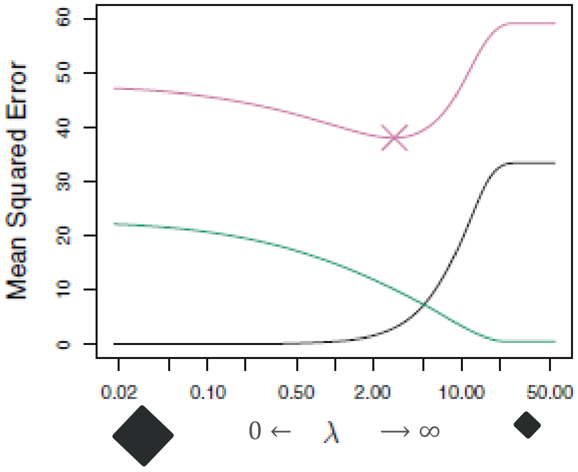
Note: The point of “intersection” between the parameter space and the constraint space is a vertex that is aligned with an axis of the parameter space. It is this property that facilitates variable selection for the Lasso.

# The Effect of the Tuning Parameter: Variable Selection

- Standardized Coefficients against $\lambda$
  - $\lambda = 0$ is the equivalent to the OLS solution
  - Coefficients: as $\lambda \longrightarrow \infty$, then $\hat{\beta} \rightarrow 0$
  - As $\lambda \longrightarrow \infty$ then the fully specified model tends to the null model (e.g. intercept only)
  - The “function” of shrinking need not be monotonic
  - We anticipate when a coefficient is removed (i.e. $\hat{\beta} = 0$), it will not re-enter into the model.

- MSE against $\lambda$
  - Black: Squared Bias
  - Green: Variance
  - Test MSE: Purple
  - In general, as $\lambda \longrightarrow \infty$ the lasso regression fit decreases
    - Decreased variance (green), but increased bias (black)
    - More stable, but less flexibly for characterizing the data
    - Here we observe that as $\lambda \longrightarrow \infty$ the model tends to the grand average (intercept only model)
  - The test MSE reveals the optimal $\lambda$ that balances the Bias-Variance trade-off.





Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. 2nd edition. Springer.

# Comparing Ridge and Lasso

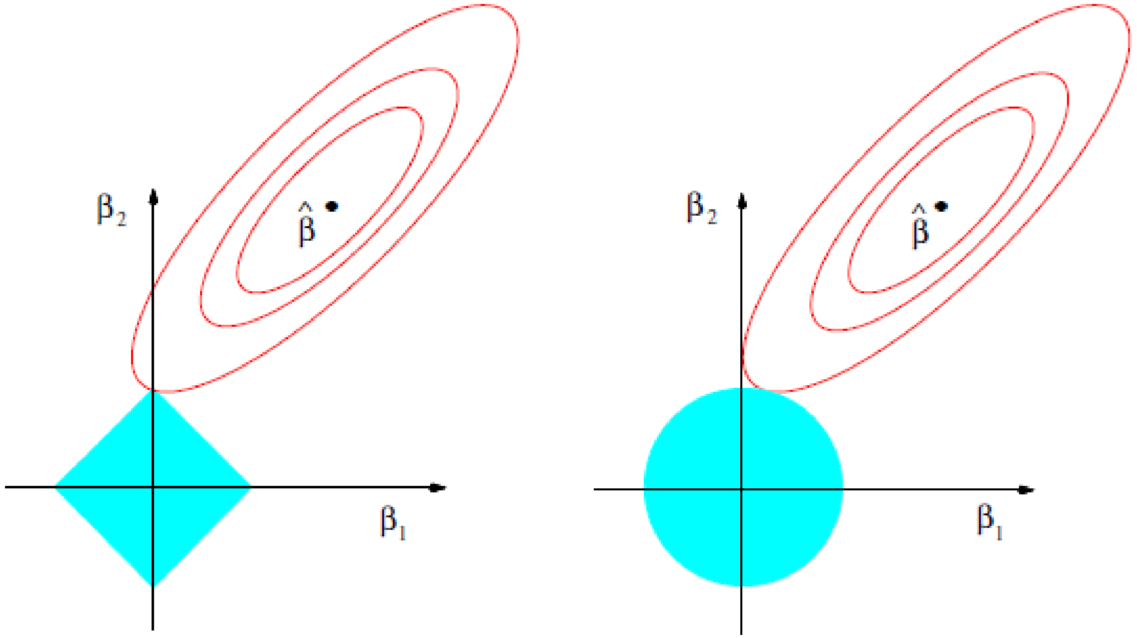


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* 2nd edition. Springer.

- The Lasso performs variable selection producing simpler models. This may ease interpretation.

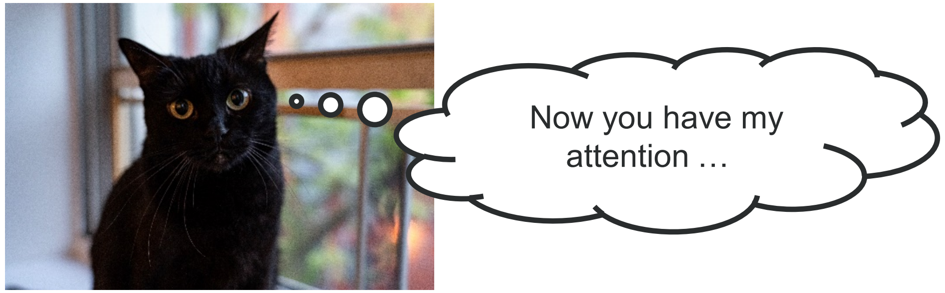
- The Lasso demands that some of the coefficients will eventually reach zero and the as $\lambda \longrightarrow \infty$ then the fully specified model tends to the null model. *This is an implicit assumption that the Ridge does not contain.*

- Choosing between the two is an exercise in comparing which approach optimizes the minimization of the MSE. It is contextual that should not involve “preferred” approaches.
  - Lasso may perform better where there are a few dominating effects
  - Ridge may perform better when the response is a function of many effects.

- Both approaches help to reduce variance, in terms of the Bias-Variance trade-off.

- Both have comparable computational costs.

# Methodology



# Elastic Net Regression: Definition

- From least squares estimation, we recall that $\beta$ is estimated using the argument that minimizes

$$RSS = \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p}\beta_j x_{ij}\right)^2, \text{ where } x_0 \equiv 1.$$

- The Elastic Net regression constrains the minimization to the region defined by $\sum_{j=1}^{p}\left(\alpha\beta_j^2 + (1-\alpha)\left|\beta_j\right|\right)$

$$\hat{\beta}^L = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\left(\alpha\beta_j^2 + (1-\alpha)\left|\beta_j\right|\right)$$

$$= \underset{\beta}{\operatorname{argmin}}\, RSS + \lambda\sum_{j=1}^{p}\underbrace{\left(\alpha\beta_j^2 + (1-\alpha)\left|\beta_j\right|\right)}_{\text{Shrinkage penalty is a mixture of } \iota_1 \text{ and } \iota_2 \text{ norms, where } \alpha, \lambda \geq 0 \text{ are tuning parameters}}$$

Shrinkage penalty only applied to $\beta_1, \dots, \beta_p$

- OLS, Ridge, Lasso are members of the Elastic Net family of models. Originally viewed as a generalization of the Lasso.
- With $\lambda = 0$, the elastic net produces the "familiar" least squares estimate.
- With $\lambda > 0$ and $\alpha = 1$, the elastic net reduces to ridge regression
- With $\lambda > 0$ and $\alpha = 0$, the elastic net reduces to lasso regression
- $\beta_0$ is not included in the constraint because it is a measure of the mean response.

# Elastic Net Constraint Observations

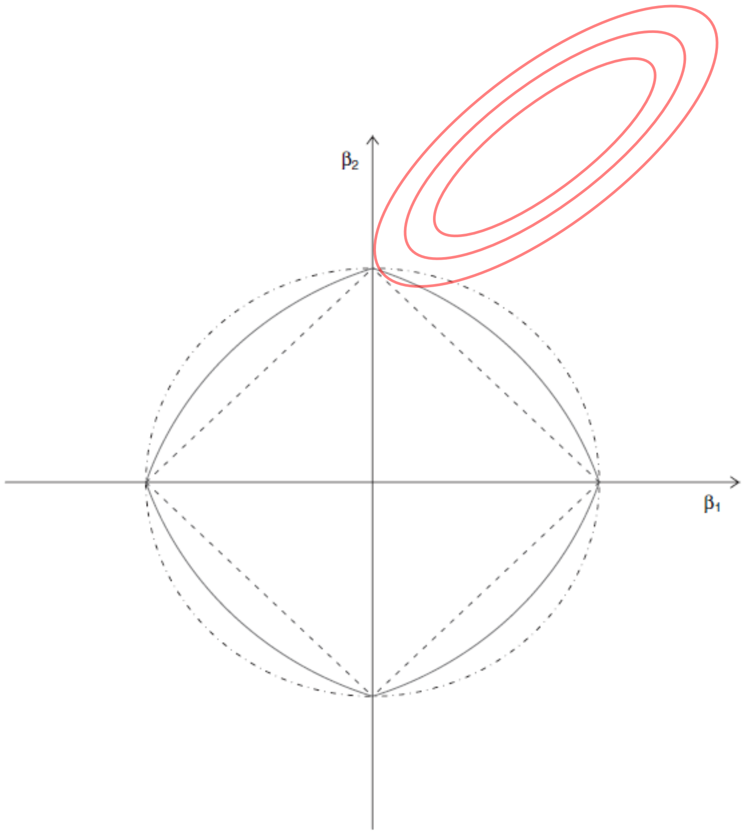
- Edge singularities are observed at the vertices.
  - This picks up a property of the lasso.

- The edges are strictly convex, with the strength of convexity varying with $\alpha$.
  - The constraint is strictly convex when $\alpha > 0$.

- Observations about the function as $\lambda \to 0$ and as $\lambda \longrightarrow \infty$ remain but are modified as a function of $\alpha$.
  - e.g. As $\lambda \longrightarrow \infty$, and as $\alpha \to 1$, then the effect of selection decreases, shifting from model selection when $\alpha = 0$, to coefficient shrinkage when $\alpha = 1$.

- Model sparsity (parsimony) is a function of $\alpha$. As $\alpha \to 0$, sparsity increases. As as $\alpha \to 0$, and $\lambda \longrightarrow \infty$, the elastic net tends to the grand mean model (intercept only).

- Model stability (e.g. lower variance) with parsimony may be achievable by balancing the effects of the lasso $(\iota_1)$ with the ridge $(\iota_2)$. The consequence is that sparse solutions are not fully realized in the presence of collinearity.
  - The presence of the ridge loss is critical to providing stability in the presence of collinearity.



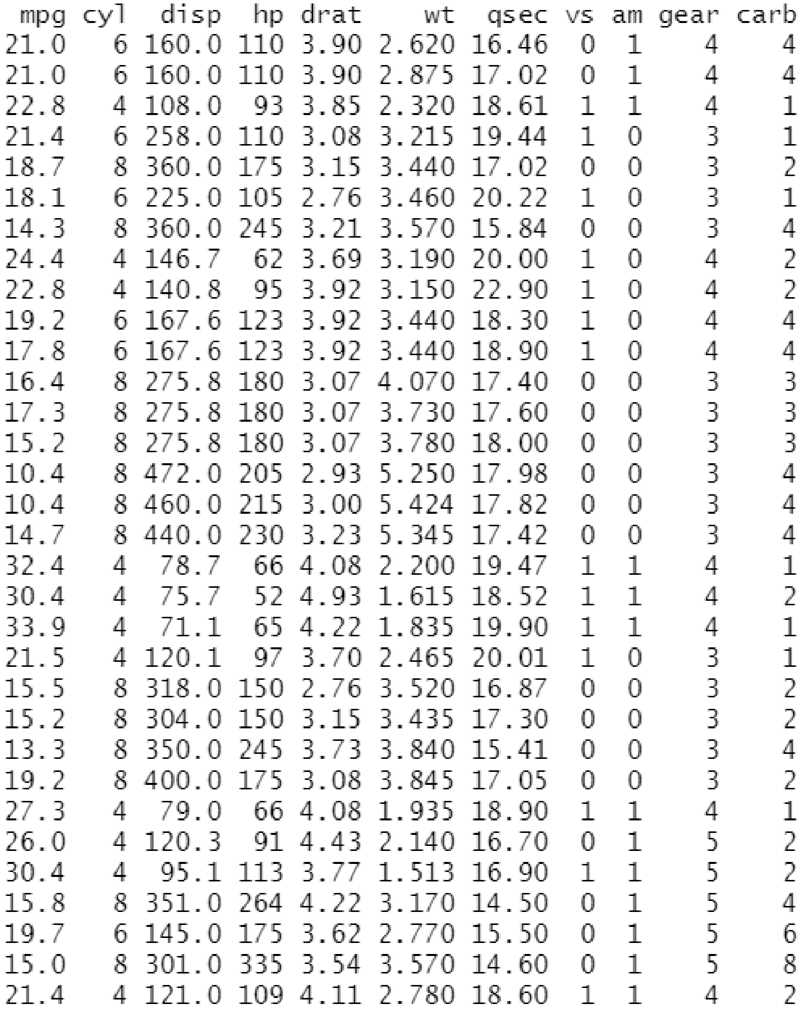
Elastic Net constraint with $\alpha = 0.5$.

# Examples

SERVE | ADD VALUE | INNOVATE

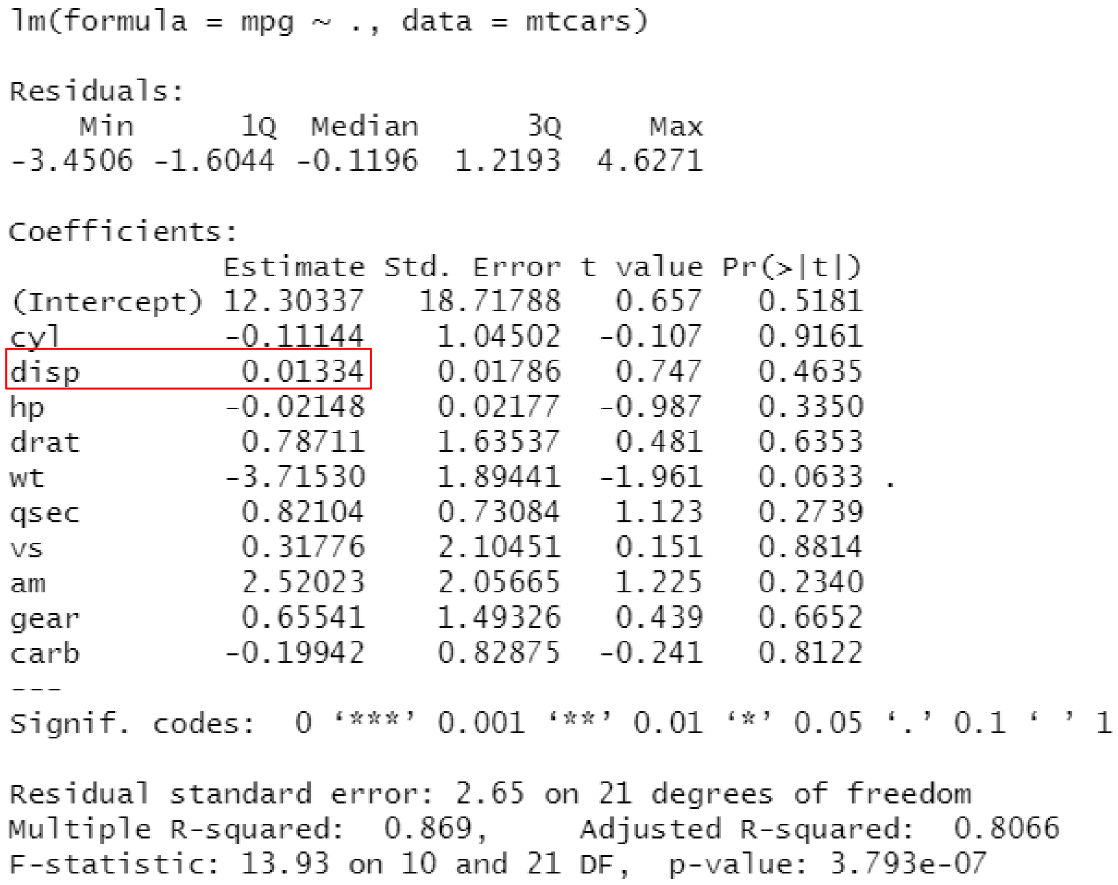
# Motor Trend Cars Dataset

| Variables | |
|---|---|
| mpg | Miles/(US) gallon |
| cyl | Number of cylinders |
| disp | Displacement (cu.in.) |
| hp | Gross horsepower |
| drat | Rear axle ratio |
| wt | Weight (1000 lbs) |
| qsec | 1/4 mile time |
| vs | Engine (0 = V-shaped, 1 = straight) |
| am | Transmission (0 = automatic, 1 = manual) |
| gear | Number of forward gears |
| carb | Number of carburetors |

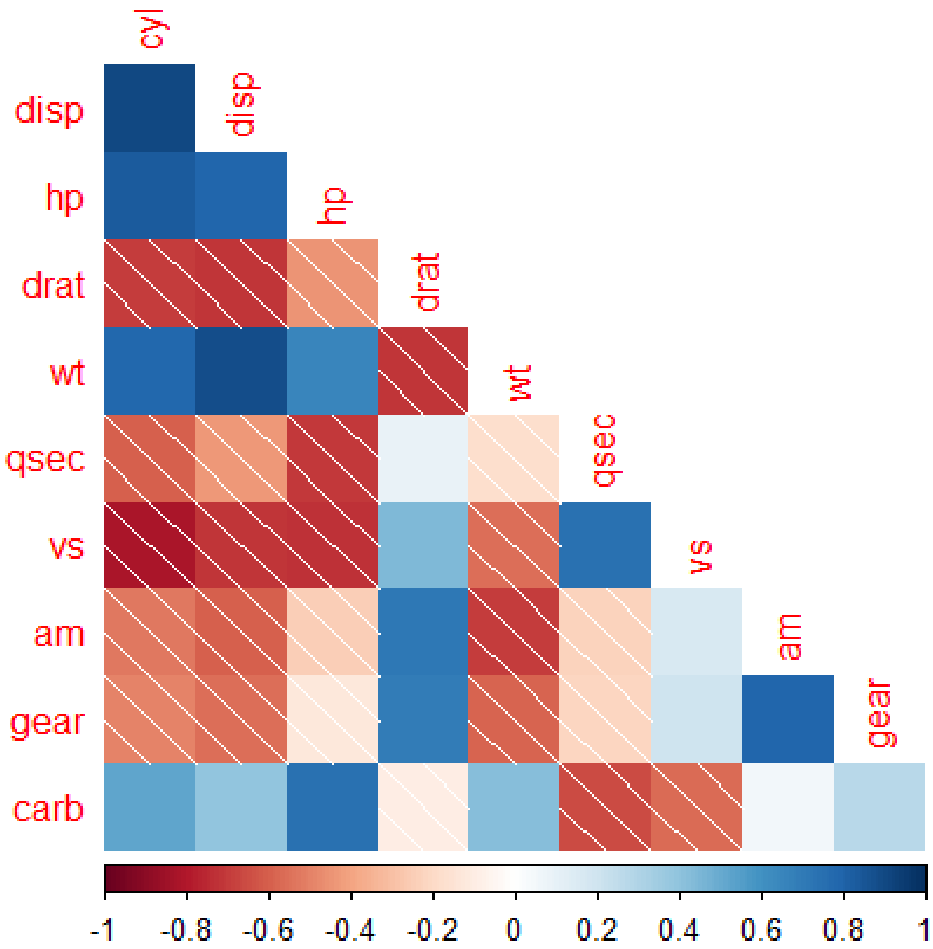


```
 mpg cyl  disp  hp drat    wt  qsec vs am gear carb
21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

# OLS Solution

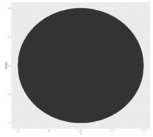


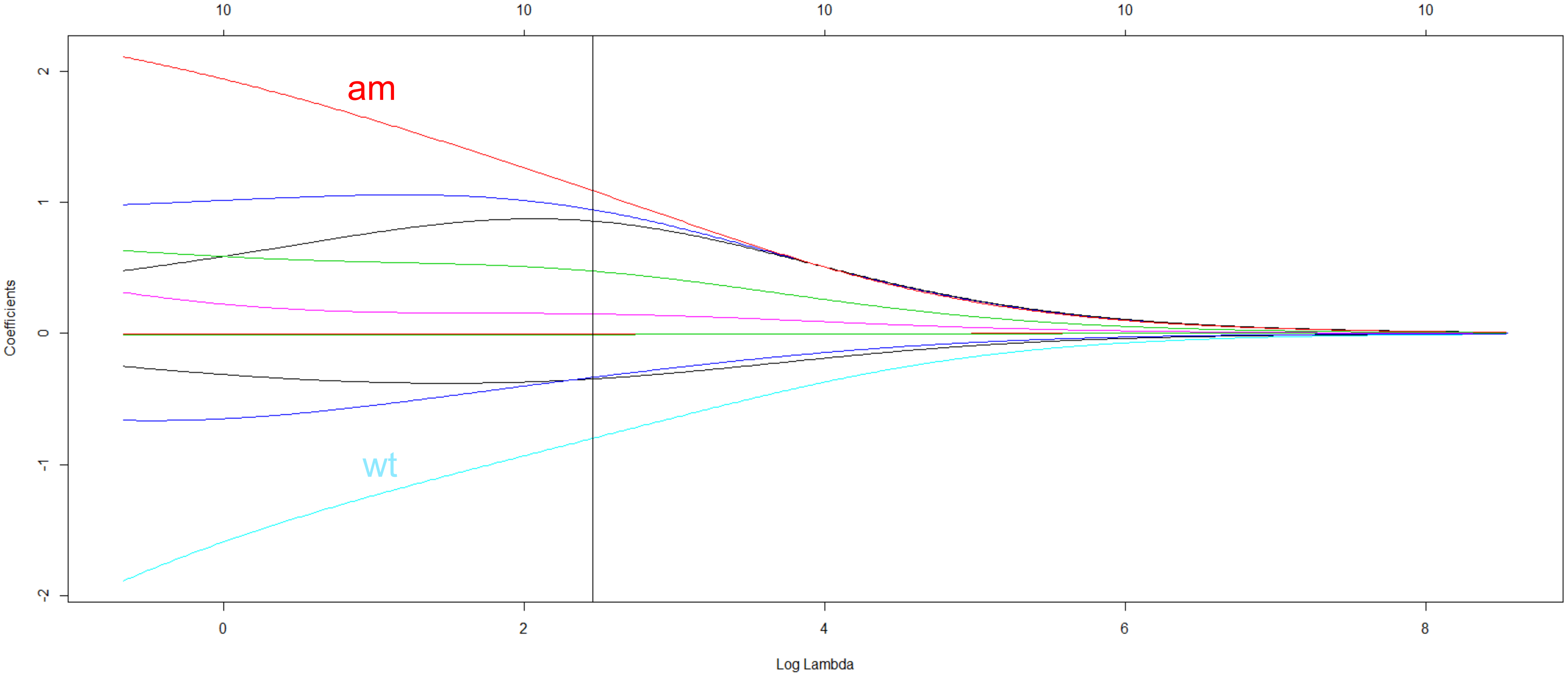
```
lm(formula = mpg ~ ., data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
cyl         -0.11144    1.04502  -0.107   0.9161
disp         0.01334    0.01786   0.747   0.4635
hp          -0.02148    0.02177  -0.987   0.3350
drat         0.78711    1.63537   0.481   0.6353
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739
vs           0.31776    2.10451   0.151   0.8814
am           2.52023    2.05665   1.225   0.2340
gear         0.65541    1.49326   0.439   0.6652
carb        -0.19942    0.82875  -0.241   0.8122
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```



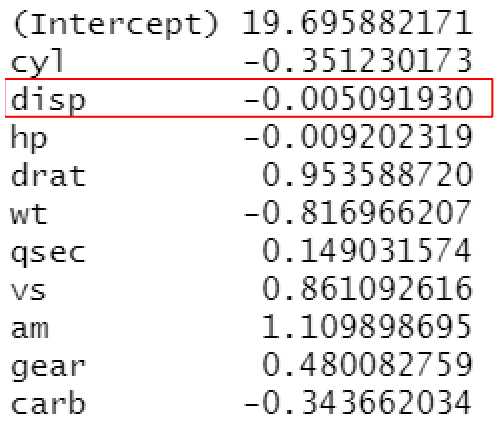
# Ridge Regression

## lambda.1se = 11.65





$$\lambda\sum_{j=1}^{p}\beta_j^2$$

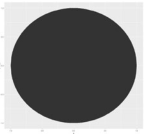


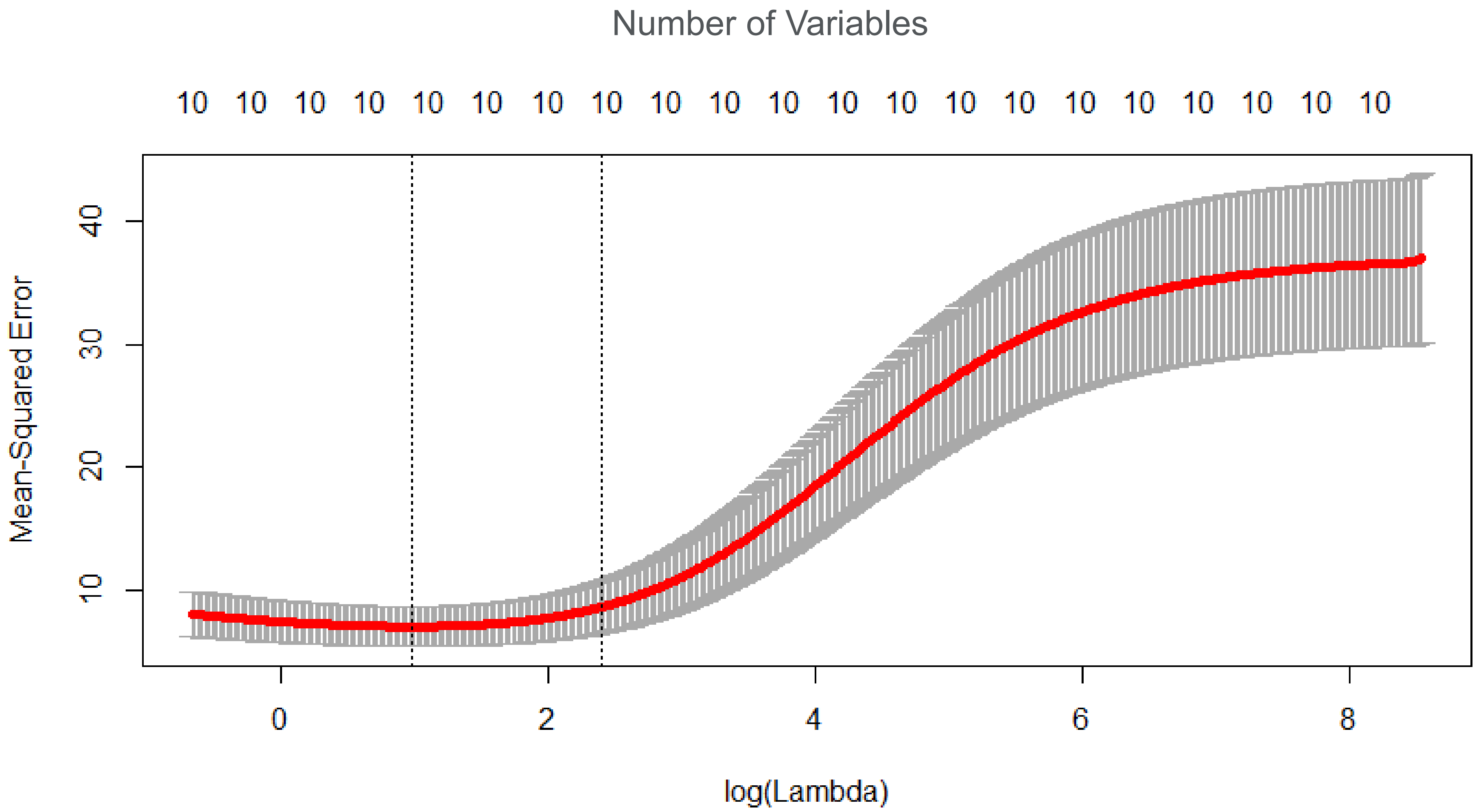
```
(Intercept) 19.695882171
cyl         -0.351230173
disp        -0.005091930
hp          -0.009202319
drat         0.953588720
wt          -0.816966207
qsec         0.149031574
vs           0.861092616
am           1.109898695
gear         0.480082759
carb        -0.343662034
```

SERVE | ADD VALUE | INNOVATE

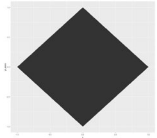
# Ridge Regression

## lambda.1se = 11.65

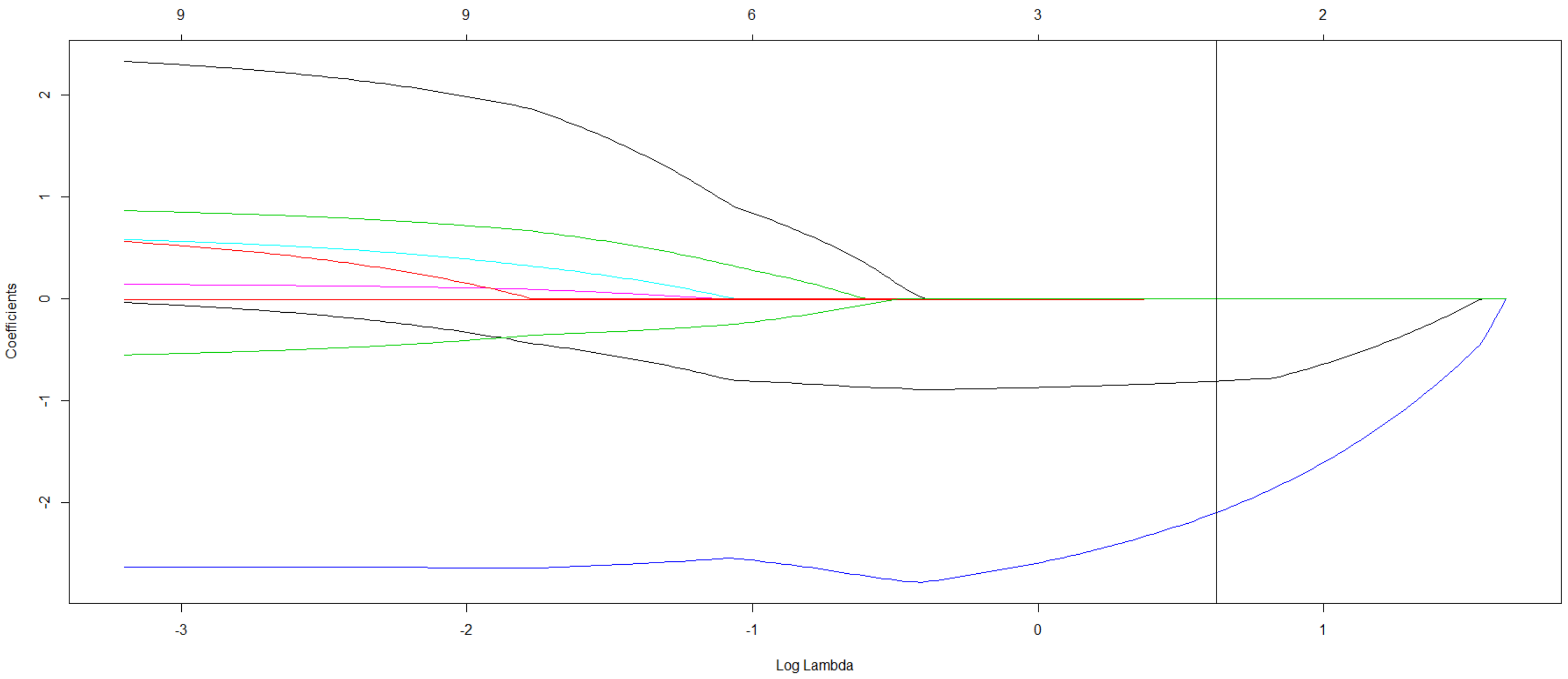




# Lasso Regression



## lambda.1se = 1.4

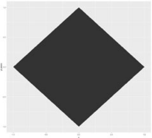


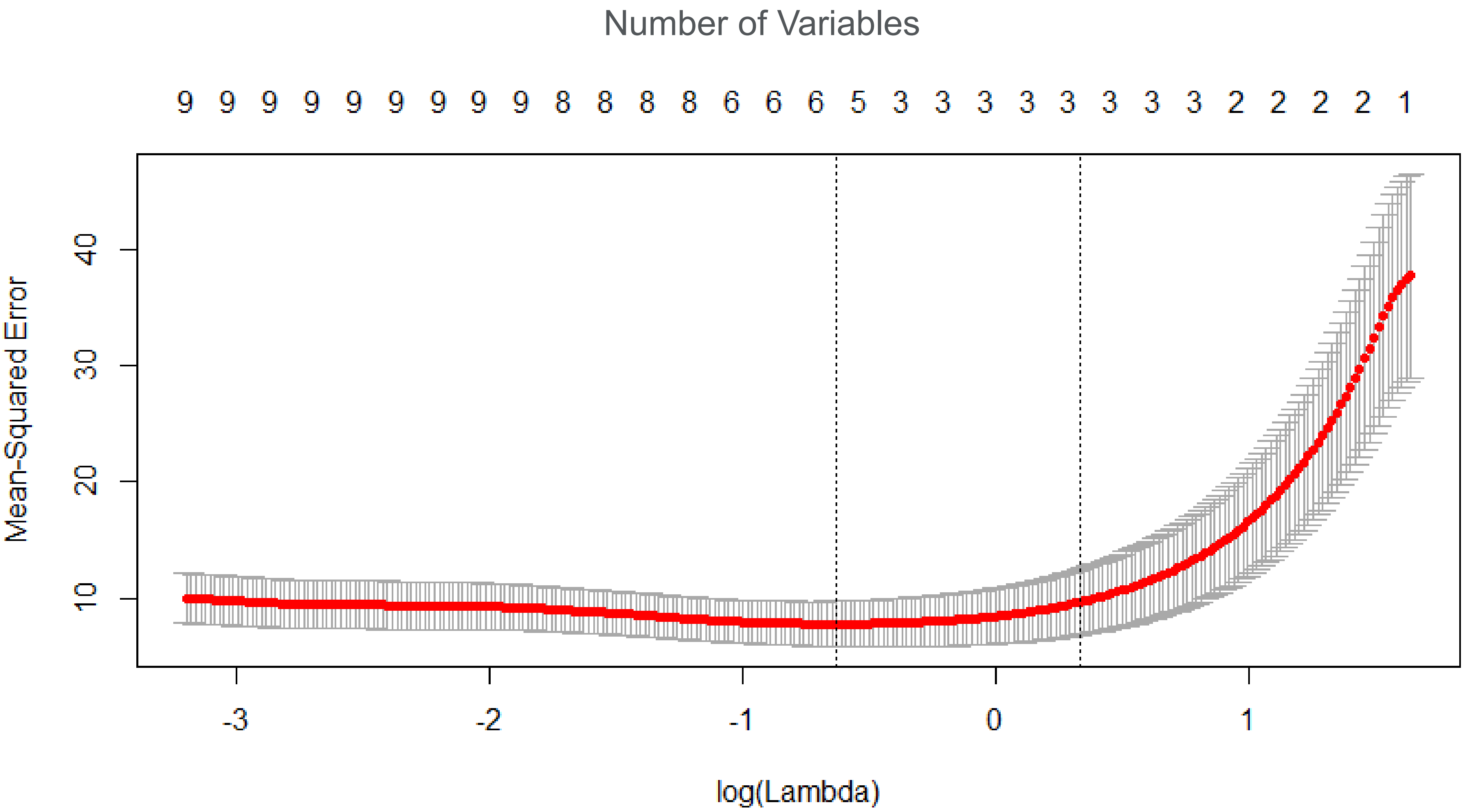
$$\lambda \sum_{j=1}^{p} |\beta_j|$$

```
(Intercept) 33.979385170
cyl         -0.844122926
disp         .
hp          -0.007051019
drat         .
wt          -2.372042343
qsec         .
vs           .
am           .
gear         .
carb         .
```

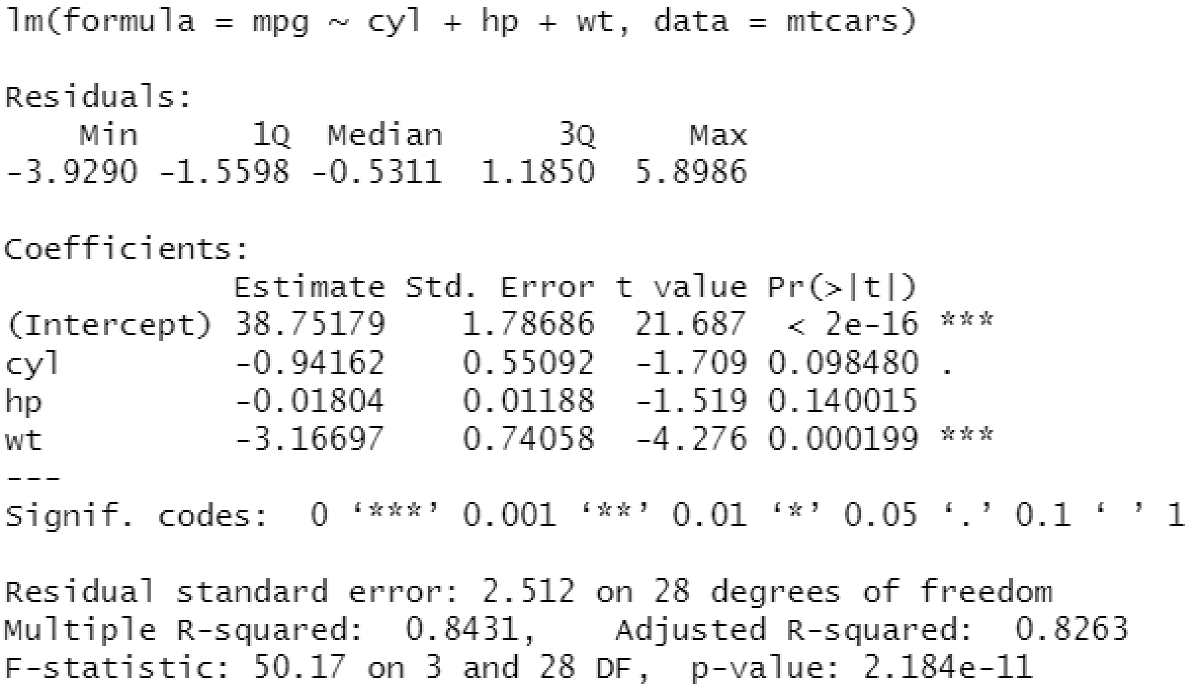
# Lasso Regression

## lambda.1se = 1.4





# Refitted Lasso

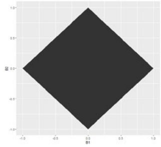


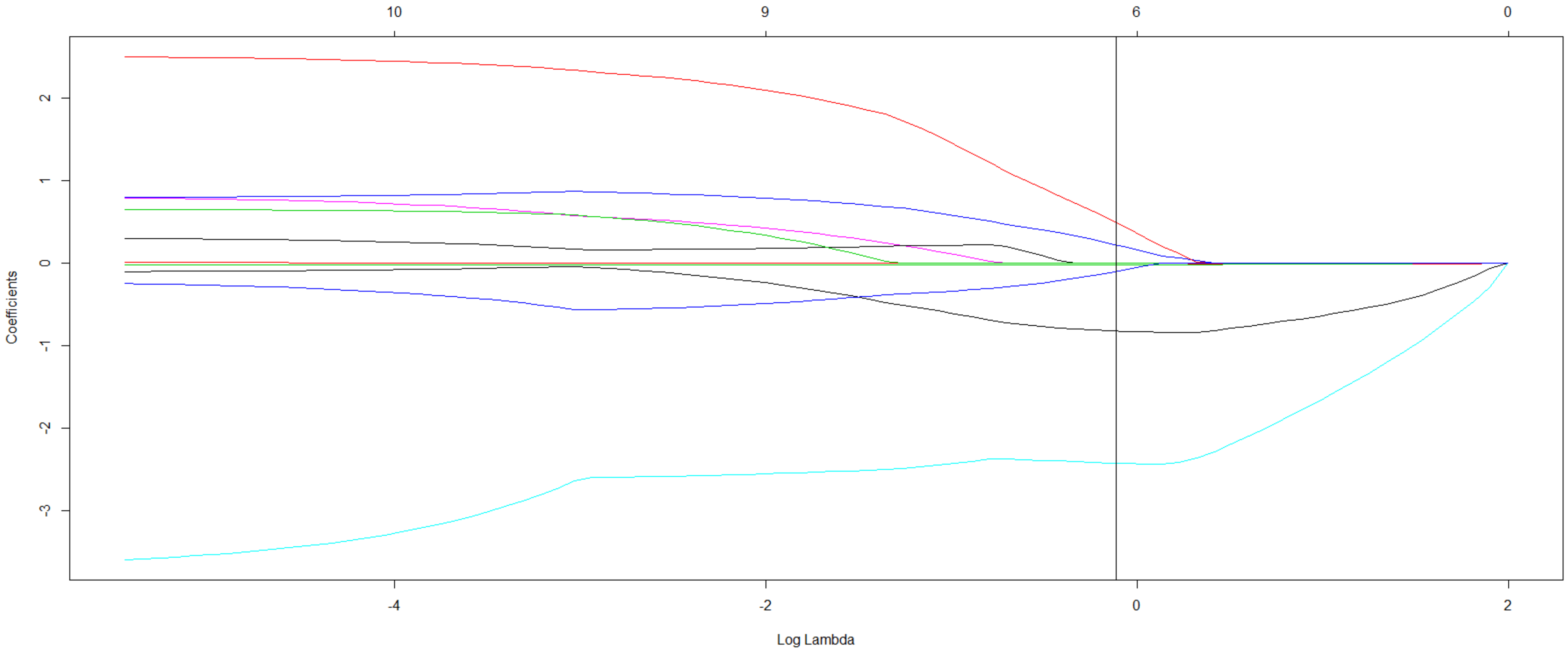
```
lm(formula = mpg ~ cyl + hp + wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9290 -1.5598 -0.5311  1.1850  5.8986 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
cyl         -0.94162    0.55092  -1.709 0.098480 .  
hp          -0.01804    0.01188  -1.519 0.140015    
wt          -3.16697    0.74058  -4.276 0.000199 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.512 on 28 degrees of freedom
Multiple R-squared:  0.8431,	Adjusted R-squared:  0.8263 
F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11
```

| Variable | Lasso | Refit Lasso (OLS) |
|---|---|---|
| cyl | -0.84412 | -0.94162 |
| hp | -0.00705 | -0.01804 |
| wt | -2.37204 | -3.16697 |

# Elastic Net

## Lambda = 0.9      alpha = .7

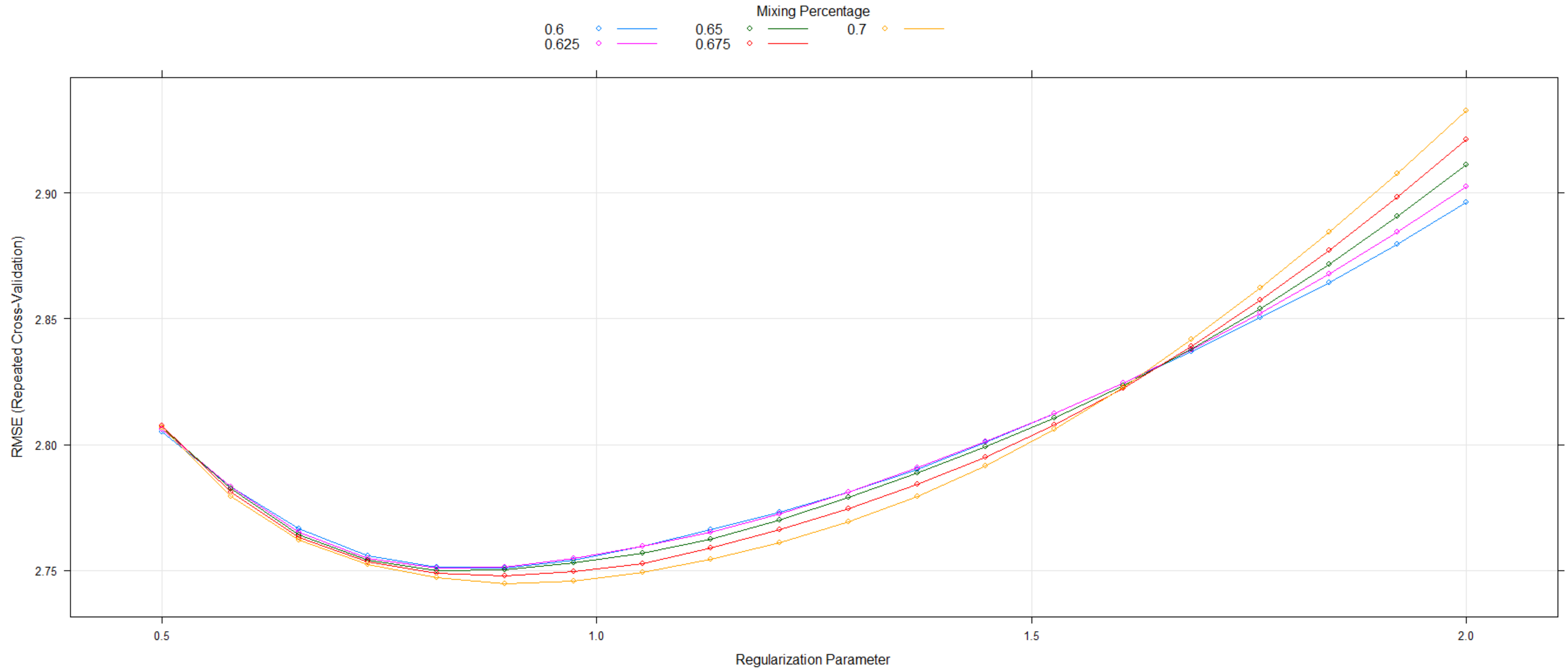




$$\lambda \sum_{j=1}^{p} \left( \alpha \beta_j^2 + (1-\alpha)\left|\beta_j\right| \right)$$

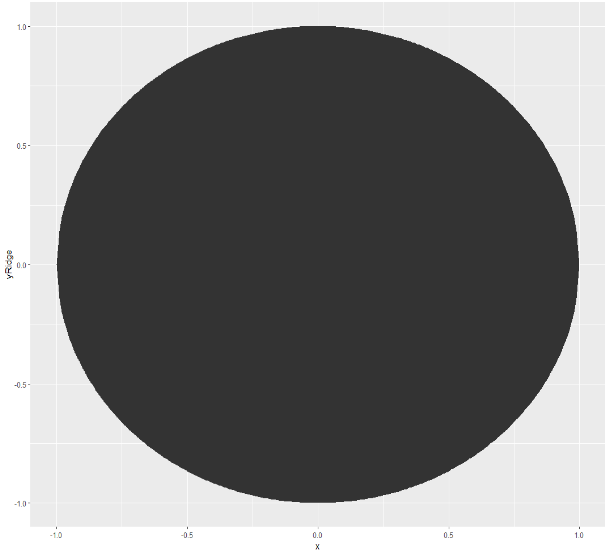
```
(Intercept) 34.29992189
cyl         -0.81531458
disp         .
hp          -0.01450304
drat         0.22667633
wt          -2.41950282
qsec         .
vs           .
am           0.49693991
gear         .
carb        -0.09604077
```

# Elastic Net


Mixing Percentage
0.6
0.625
0.65
0.675
0.7
RMSE (Repeated Cross-Validation)
2.90
2.85
2.80
2.75
0.5
1.0
1.5
2.0
Regularization Parameter

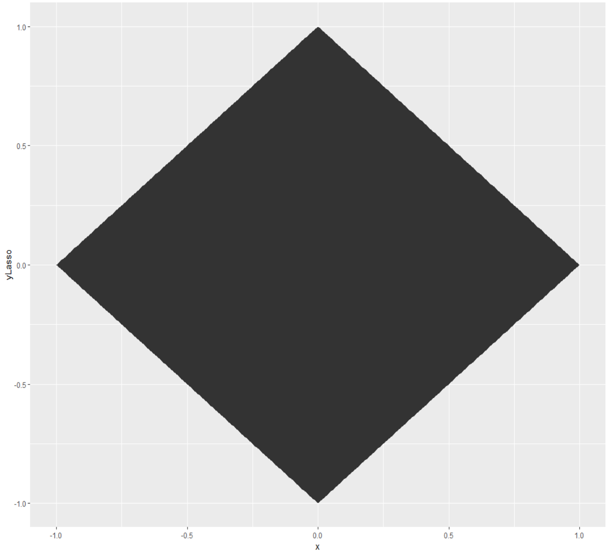
# Final Comparison

## Ridge



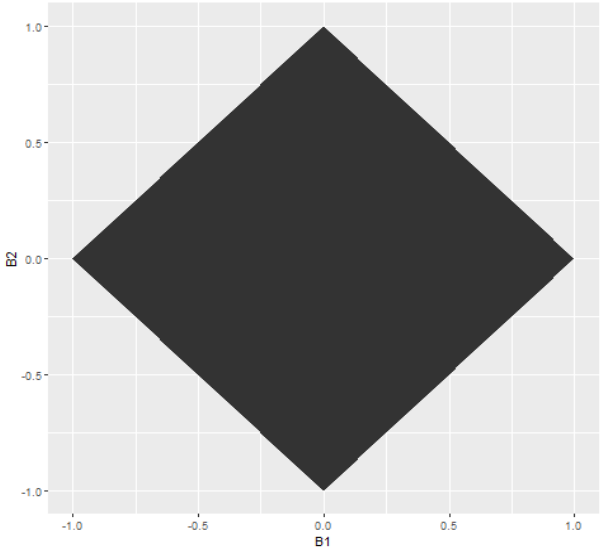
```
(Intercept) 19.695882171
cyl         -0.351230173
disp        -0.005091930
hp          -0.009202319
drat         0.953588720
wt          -0.816966207
qsec         0.149031574
vs           0.861092616
am           1.109898695
gear         0.480082759
carb        -0.343662034
```

## Lasso



```
(Intercept) 33.979385170
cyl         -0.844122926
disp         .
hp          -0.007051019
drat         .
wt          -2.372042343
qsec         .
vs           .
am           .
gear         .
carb         .
```

## Elastic Net



```
(Intercept) 34.29992189
cyl         -0.81531458
disp         .
hp          -0.01450304
drat         0.22667633
wt          -2.41950282
qsec         .
vs           .
am           0.49693991
gear         .
carb        -0.09604077
```

# Final Summary: A Reminder of Why We Should Care

- Regularization methods are
  - Commonly introduced within the context of regression methods
  - Commonly presented as a tool for model selection
  - Can mitigate problems associated with collinearity

- Regularization methods have utility for
  - *Prediction*: Can reduce variability while maintaining low model bias in terms of the bias-variance trade-off
  - *Interpretability*: These methods can assist in removing irrelevant or obfuscating variables – parsimonious models
    - Parsimony: The state of being stingy
    - Parsimonious models: Stingy with the number of variables retained in the final model

- Regularization as model selection
  - An option for predictor (regressor) subset selection
  - Subset selection examples: Purposeful, stepwise, AIC, BIC, F- and t-test, best subset model space search

# References

SERVE | ADD VALUE | INNOVATE

# References

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*.  2nd edition. Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning.* New York: Springer.
- Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics*. John Wiley & Sons.
- Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. CRC press.
- McCullagh, P., & Nelder, J. A. (1999). *Generalized linear models*. Chapman & Hall/CRC.
- Casella, G., & Berger, R. L. (2002). *Statistical inference, 2nd edition*. Duxbury Thomson Learning.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: series B (statistical methodology)*, *67*(2), 301-320.
- Frank, I. and Friedman, J. (1993) A statistical view of some chemometrics regression tools. *Technometrics*, 35,109–148.
- De Mol, C., De Vito, E., & Rosasco, L. (2009). Elastic-net regularization in learning theory. *Journal of Complexity*, *25*(2), 201-230.
- Zhou, Q., Chen, W., Song, S., Gardner, J., Weinberger, K., & Chen, Y. (2015, February). A reduction of the elastic net to support vector machines with an application to GPU computing. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 29, No. 1)