Announcements

- Make sure to sign in on the google form (linked here)
- Pset 7 due November 4 at 5 pm
- Midterm 2 November 11 through November 18

Weighted least squares regression

This question is based on an October 29th conversation with Skyler Wu.

Consider a least squares model where, rather than weighting all residuals equally, we are going to assign different weights to different residuals. That is, we want to minimize

$$\sum_{i=1}^{n}[w_i(Y_i - \hat{Y}_i)]^2$$

Equivalently, letting $\mathbf{W}$ be a diagonal matrix of weights, letting $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$ with $\vec{\epsilon} \sim \text{MVN}_n(0, \sigma^2 I_n)$, and using the fact that $\hat{\vec{Y}} = \mathbf{X}\hat{\vec{\beta}}$, we want to minimize

$$||\mathbf{W}(\vec{Y} - \mathbf{X}\hat{\vec{\beta}})||^2$$

Expanding and taking the derivative gives the following:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \hat{\vec{\beta}}}((\vec{Y} - \mathbf{X}\hat{\vec{\beta}})^T\mathbf{W}^T\mathbf{W}(\vec{Y} - \mathbf{X}\hat{\vec{\beta}})) \\ &= -2\mathbf{X}^T\mathbf{W}^T\mathbf{W}(\vec{Y} - \mathbf{X}\hat{\vec{\beta}}) \\ \implies \mathbf{X}^T\mathbf{W}^T\mathbf{W}\mathbf{X}\hat{\vec{\beta}} &= \mathbf{X}^T\mathbf{W}^T\mathbf{W}\vec{Y} \\ \implies \hat{\vec{\beta}} &= (\mathbf{X}^T\mathbf{W}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^T\mathbf{W}\vec{Y} \end{aligned}$$

This is our new weighted least-squares regression $\hat{\vec{\beta}}$, which we will be studying in this problem.

Here are a few facts that will be useful here and on the homework:

- For matrices $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$, of allowable dimensions, $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
- If $\mathbf{A}$ is of full column rank, $\mathbf{A}^T\mathbf{A}$ is symmetric and invertible
- If $\mathbf{A}$ is symmetric, and $\mathbf{B}$ is of allowable dimensions, $\mathbf{B}^T\mathbf{AB}$ is symmetric
- For an invertible and symmetric matrix $\mathbf{A}$, $\mathbf{A}^{-1} = (\mathbf{A}^{-1})^T$
- For $\vec{Y} = \vec{c} + \mathbf{B}\vec{X}$ with $\vec{c}$ and $\mathbf{B}$ constant and $\vec{X}$ random, $E(\vec{Y}) = \vec{c} + \mathbf{B}E(\vec{X})$
- $\text{Cov}(\vec{X})$ is an $n \times n$ matrix whose i, j entry is $\text{Cov}(X_i, X_j)$
- For $\vec{Y} = \vec{c} + \mathbf{B}\vec{X}$ with $\vec{c}$ and $\mathbf{B}$ constant and $\vec{X}$ random, $\text{Cov}(\vec{Y}) = \mathbf{B}\text{Cov}(\vec{X})\mathbf{B}^T$

1. Verify that using the usual weights for least squares regression, this formula reduces to the usual estimator for $\hat{\vec{\beta}}$.
2. Find the bias of $\hat{\vec{\beta}}$ for $\vec{\beta}$.
3. Find the variance-covariance matrix of $\hat{\vec{\beta}}$ in matrix form. When will this equal the variance-covariance matrix of $\hat{\vec{\beta}}$ in OLS regression?

Ridge, LASSO, optimizing λ, and β trajectories

This question will deal with a data set of country-level statistics from this source with an explanation of the data encoding found here.

A few useful columns:

- `spi_ospi`: Overall social progress index on 0-100 scale
- `mad_gdppc`: GDP per capita
- `wdi_internet`: Percent of population using the internet
- `wdi_birth`: Birth rate per 1000 people
- `wdi_chexppgdp`: Current health expenditures as percent of GDP
- `wdi_elerenew`: Percent of total electricity output that's renewable
- `wdi_lifexp`: Life expectancy at birth
- `wdi_wip`: Proportion of seats held by women in national parliaments
- `wdi_popurb`: Percentage of total population that is urban
- `wdi_imig`: Proportion of people born outside the country in which they live

1. Find a well-tuned Ridge regression model via `cv.glmnet` for predicting `spi_ospi`: consider all main predictors above and all 2-way interactions of these predictors.

```
library(glmnet)
set.seed(139)

# Variables to be used
columns <- c("mad_gdppc", "wdi_internet", "wdi_birth", "wdi_chexppgdp",
             "wdi_elerenew", "wdi_lifexp", "wdi_wip", "wdi_popurb", "wdi_imig")

# TODO: Model matrix for glmnet

# TODO: Scale the model matrix

# TODO: Run cross validation
```

2. Plot the average MSE on the validation sets against the λ's you considered in the previous part. Report the best λ and justify this choice using this plot.

```
# TODO: Plot MSE
```

3. Provide the $\hat{\beta}$ trajectory plot of the main effects from this model (plot each β_j as a function of λ as a line, and do this for all 11 main effects). Interpret what you see in 2-3 sentences.

```
# TODO: Plot beta trajectories
```

4. Fit a well-tuned LASSO regression model: examine main effects of predictors and all their 2-way interactions.

```
# TODO: Run cross validation
```

5. Plot the average MSE on the validation sets against the λ's you considered in the previous part. Report the best λ and justify this choice using this plot.

```
# TODO: Plot MSE
```

6. Provide the $\hat{\beta}$ trajectory plot of the main effects from this LASSO model (plot each β_j as a function of λ as a line, and do this for all 11 main effects). Compare this to the ridge trajectories.

```
# TODO: Plot trajectories
```

7. Choose a best regularized/penalized regression model and briefly justify your choice.

```
# TODO: Choose best model
```

Penalization functions

Recall that for both Ridge and LASSO, we are trying to minimize something of the form:

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + p(\hat{\vec{\beta}})$$

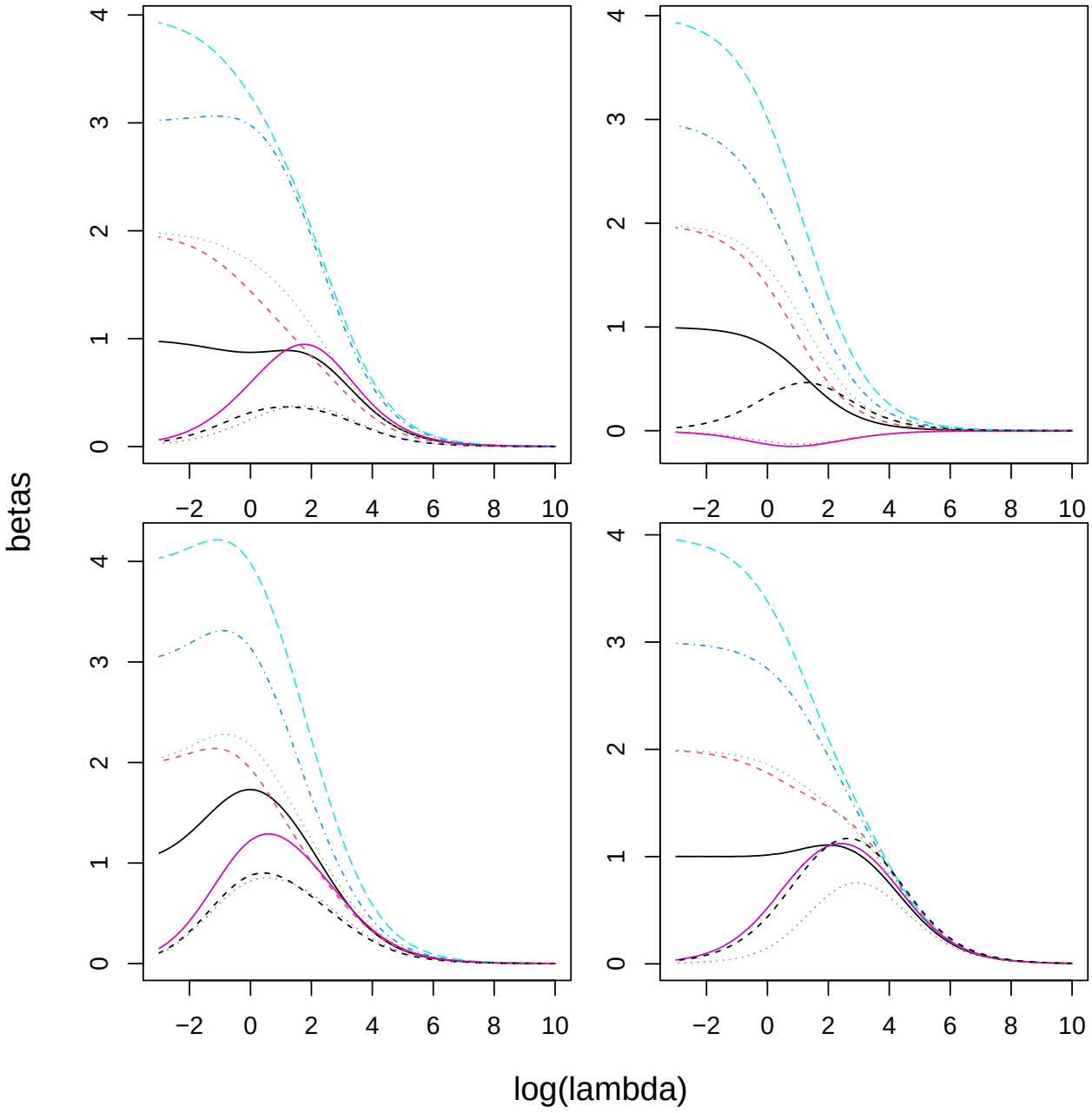
State whether the following functions could or couldn't be used as penalization functions for $\hat{\vec{\beta}}$. If they could, provide a context in which this might be a useful penalization function; if not, explain why it would give undesired behavior.

1. $p(\hat{\vec{\beta}}) = \sum_{i=1}^{k} \hat{\beta}_i$
2. $p(\hat{\vec{\beta}}) = \sum_{i=1}^{k} \hat{\beta}_i^{\,4}$
3. $p(\hat{\vec{\beta}}) = \sum_{i=1}^{k} \log(\hat{\beta}_i)$
4. $p(\hat{\vec{\beta}}) = \sum_{i=1}^{k} \log(|\hat{\beta}_i|)$
5. $p(\hat{\vec{\beta}}) = \sum_{i=1}^{k} 1/|\hat{\beta}_i|$
6. $p(\hat{\vec{\beta}}) = -\sum_{i=1}^{k} 1/|\hat{\beta}_i|$
7. What general requirements do we need for a penalization function?
8. Write a valid penalization function that we haven't studied before.

Miscellaneous

1. For what λs would LASSO and Ridge give the same model?
2. Below are four $\hat{\beta}$ trajectory plots. Each comes from a data set with 50 data points. One trajectory comes from data with no built-in correlation between the predictors; one comes from data with moderate and equal correlation among all the predictors; one comes from data with moderate random (but fixed) correlation among the predictors; and one is fake (and impossible). Determine which is which.

```
## Loading required package: ggplot2
```


betas
log(lambda)
0
1
2
3
4
−2
0
2
4
6
8
10