BIOST 527 - HOMEWORK 2

KATIE WOOD

Problem 1

1. From (3), we have

$$g^*(x) = \arg\min_g \mathbb{E}_{Y|X}[(Y - g(X))^2|X = x], \quad \forall x \in \mathcal{X}$$

Then noting that $g : \mathcal{X} \to \mathbb{R}$ (since $\mathcal{Y} = \mathbb{R}$), let $g(x) = [g(X)|X = x] = c \in \mathbb{R}$ be arbitrary. Following the hint, we then prove:

$$\begin{aligned}
\mathbb{E}[(Y - c)^2] &= \mathbb{E}[(Y - \mathbb{E}(Y) + \mathbb{E}(Y) - c)^2] \\
&= \mathbb{E}[(Y - \mathbb{E}(Y))^2 + (\mathbb{E}(Y) - c)^2 + 2(Y - \mathbb{E}(Y))(\mathbb{E}(Y) - c)] \\
&= \mathbb{E}[(Y - \mathbb{E}(Y))^2] + \mathbb{E}[(\mathbb{E}(Y) - c)^2] + 2(\mathbb{E}(Y) - c)\cancel{\mathbb{E}[(Y - \mathbb{E}(Y))]} \\
&= \mathrm{Var}(Y) + (\mathbb{E}(Y) - c)^2
\end{aligned}$$

Conditioning on $X = x$, this becomes:

$$\mathbb{E}_{Y|X}[(Y - g(X))^2|X = x] = \mathrm{Var}(Y|X = x) + (\mathbb{E}(Y|X = x) - g(x))^2$$

Now, plugging this result into our equation for $g^*(x)$ above:

$$\begin{aligned}
g^*(x) &= \arg\min_g [\mathrm{Var}(Y|X = x) + (\mathbb{E}(Y|X = x) - g(x))^2] \\
&= \arg\min_g [(\mathbb{E}(Y|X = x) - g(x))^2]
\end{aligned}$$

Since $\mathrm{Var}(Y|X = x)$ does not depend on g. Then since $(\mathbb{E}(Y|X = x) - g(x))^2 \geq 0$, we minimize by setting

$$\begin{aligned}
(\mathbb{E}(Y|X = x) - g^*(x))^2 &= 0 \\
\Rightarrow g^*(x) &= \mathbb{E}(Y|X = x)
\end{aligned}$$

2. From (3), we have

$$\begin{aligned}
g^*(x) &= \arg\min_g \mathbb{E}_{Y|X}(|Y - g(X)||X = x) \\
&= \arg\min_g \mathbb{E}_{Y|X}(|(Y|X = x) - g(x)|)
\end{aligned}$$

Now let $y_1, ..., y_n \in \mathbb{R}$ denote the values Y takes on when $X = x$. Then

$$g^*(x) = \arg\min_g \sum_{i=1}^{n} |y_i - g(x)| P(Y = y_i|X = x)$$

Empirically, we simply observe Y taking on the values $y_1, ..., y_n$ when $X = x$. Thus, from an empirical standpoint, $P(Y = y_i|X = x) = \frac{1}{n}$ for all $i = 1, ..., n$. Therefore,

$$\begin{aligned} g^*(x) &= \arg\min_g \frac{1}{n} \sum_{i=1}^{n} |y_i - g(x)| \\ &= \arg\min_g \sum_{i=1}^{n} |y_i - g(x)| \end{aligned}$$

By way of contradiction, assume that $g^*(x) \neq \text{median}[Y|X = x]$, and, without loss of generality, assume $g^*(x) < \text{median}[Y|X = x]$. Denote $\hat{n}$ the number of data points y_i for which $y_i < g^*(x)$, and (thus) by $n - \hat{n}$ denote the number of data points y_i for which $y_i > g^*(x)$. Then

$$\begin{aligned} \sum_{i=1}^{n} |y_i - g^*(x)| &= \sum_{y_i < g^*(x)} |y_i - g^*(x)| + \sum_{y_i > g^*(x)} |y_i - g^*(x)| \\ &= \sum_{i=1}^{\hat{n}} |y_i - g^*(x)| + \sum_{i=\hat{n}+1}^{n} |y_i - g^*(x)| \end{aligned}$$

But for some $g^{**}(x) = g^*(x) + \epsilon$, where $\epsilon \leq y_{\hat{n}+1} - g^*(x)$, we have

$$\begin{aligned} \sum_{i=1}^{n} |y_i - g^{**}(x)| &= \sum_{i=1}^{\hat{n}} |y_i - (g^*(x) + \epsilon)| + \sum_{i=\hat{n}+1}^{n} |y_i - (g^*(x) + \epsilon)| \\ &= \sum_{i=1}^{\hat{n}} |y_i - g^*(x)| + \hat{n}\epsilon + \sum_{i=\hat{n}+1}^{n} |y_i - g^*(x)| - (n - \hat{n})\epsilon \\ &= \sum_{i=1}^{n} |y_i - g^*(x)| + (2\hat{n} - n)\epsilon \end{aligned}$$

and since $\hat{n} < \frac{n}{2}$ by our assumption, we find

$$\sum_{i=1}^{n} |y_i - g^{**}(x)| < \sum_{i=1}^{n} |y_i - g^*(x)|$$

which is a contradiction. Thus, $g^*(x) = \text{median}[Y|X = x]$.

3. Again from (3), we have

$$g^*(x) = \arg\min_g \mathbb{E}_{Y|X}[\mathbb{I}(Y \neq g(X))|X = x]$$

Then since $\mathcal{Y} = \{0, 1\}$,

$$g^*(x) = \arg\min_g \sum_{y' \in \{0,1\}} \mathbb{I}(y' \neq g(x)) P(Y = y'|X = x)$$

If $g(x) \notin \{0, 1\}$,

$$\begin{aligned} \sum_{y' \in \{0,1\}} \mathbb{I}(y' \neq g(x)) P(Y = y'|X = x) &= \sum_{y' \in \{0,1\}} P(Y = y'|X = x) \\ &= 1 \end{aligned}$$

On the other hand, if $g(x) \in \{0, 1\}$,

$$\begin{aligned}\sum_{y' \in \{0,1\}} \mathbb{I}(y' \neq g(x)) P(Y = y'|X = x) &= P(Y \neq g(x)|X = x) \\ &= 1 - P(Y = g(x)|X = x) \\ &\leq 1\end{aligned}$$

We minimize the left-hand side of this equation by making $P(Y = g(x)|X = x)$ as large as possible, which we can do by setting

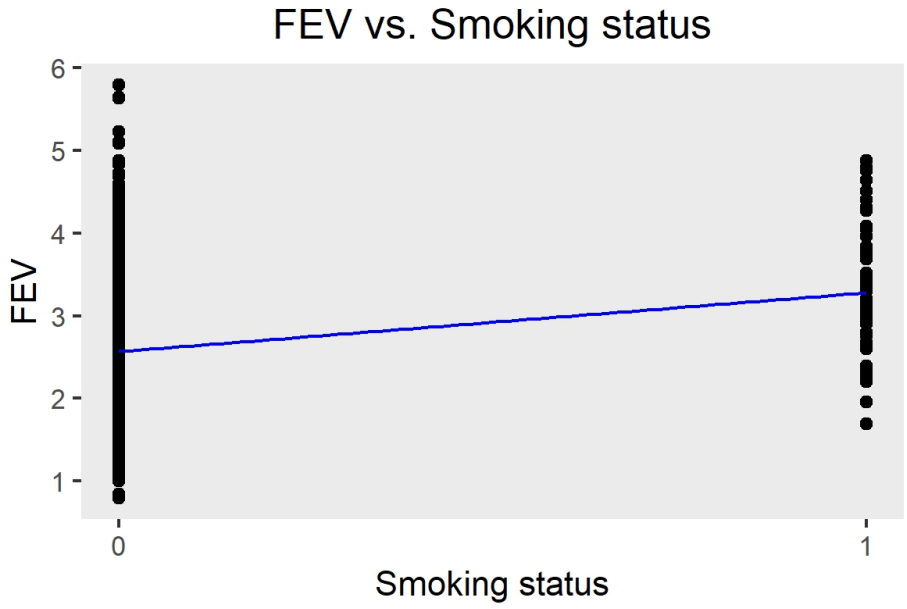
$$g^*(x) = \arg \max_{y' \in \{0,1\}} P(Y = y'|X = x)$$

Problem 2

1. Assuming the linear model

   $$Y = \beta_1 X + \beta_0 + \epsilon$$

   for the data, we find:



Coefficients	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	2.56614	0.03466	74.037	$< 2 \times 10^{-16}$
X	0.71072	0.10994	6.464	1.99×10^{-10}

   Our estimate for β_1 is thus $\hat{\beta}_1 = 0.71072$. With the p-value 1.99×10^{-10}, we would reject the null hypothesis that there is no association ($\beta_1 = 0$) between Y and X at a significance level of $\alpha = 0.05$ (or even lower). This fit seems to provide strong evidence of a positive ($\beta_1 > 0$) association between Y (FEV) and X (smoking status).
2. The results in part 1 do not allow us to conclude that smoking impairs lung function in children because the association we found shows the reverse: children who smoke tend to have *higher* FEV. Also, we cannot conclude that smoking causes lower or higher FEV with this evidence because an association does not imply

causation. In particular, there could be confounding variables (height and age) that influence both smoking status and FEV in ways we have not yet accounted for.

3. Starting from

$$Y = \beta_1 X + g^*(Z) + \epsilon$$

we condition both sides on Z, noting that ϵ is independent of Z:

$$Y|Z = \beta_1 X|Z + g^*(Z)|Z + \epsilon$$

Then we take the expectation of both sides, and use linearity of expectation on the right-hand side:

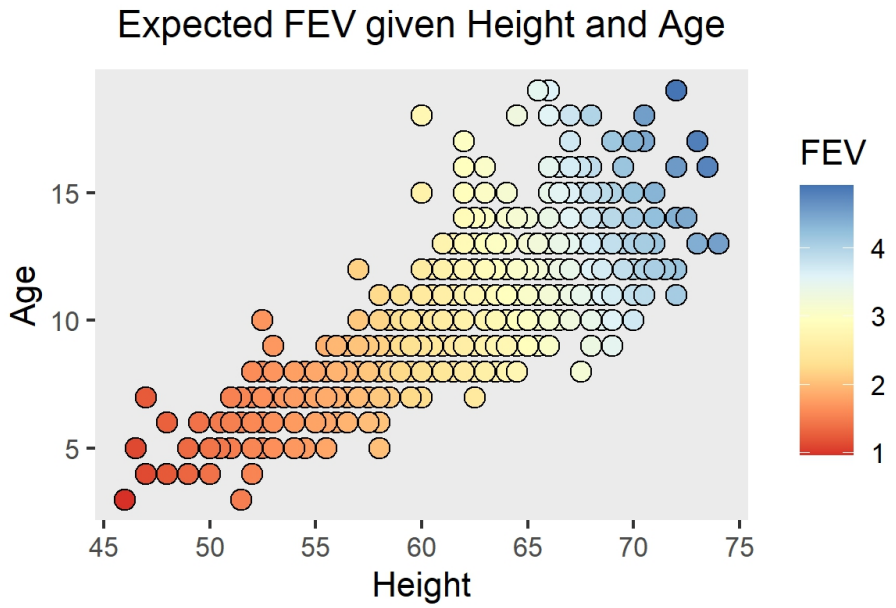
$$\begin{aligned}\mathbb{E}(Y|Z) &= \mathbb{E}(\beta_1 X|Z + g^*(Z) + \epsilon) \\ &= \beta_1 \mathbb{E}(X|Z) + \mathbb{E}(g^*(Z)|Z) + \cancel{\mathbb{E}(\epsilon)} \\ &= \beta_1 \mathbb{E}(X|Z) + g^*(Z)\end{aligned}$$

where the last line follows because g^* is a deterministic function and thus given Z, we know $g^*(Z)$. Next, we subtract the above equation from the equation for the model:

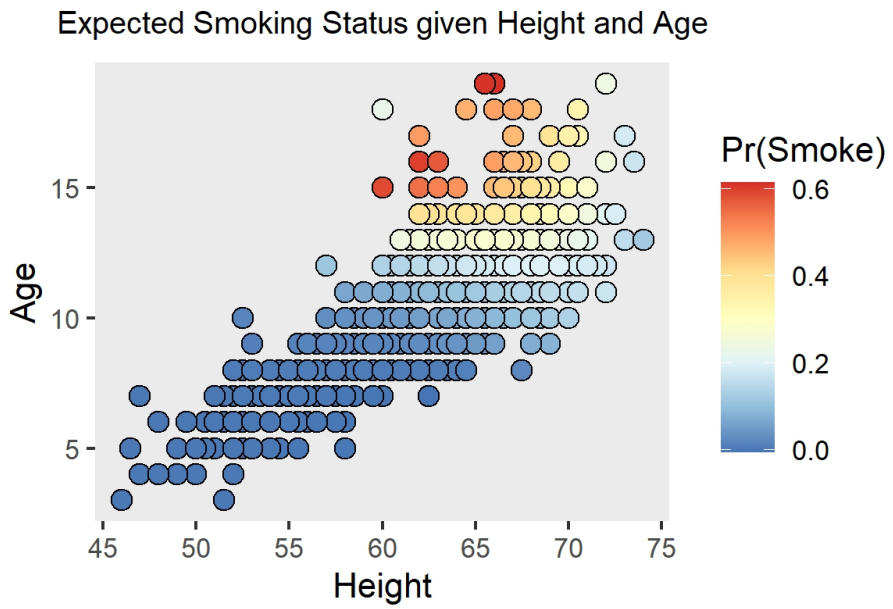
$$\begin{aligned}Y - \mathbb{E}(Y|Z) &= \beta_1 X + \cancel{g^*(Z)} + \epsilon - \beta_1 \mathbb{E}(X|Z) - \cancel{g^*(Z)} \\ &= \beta_1 (X - \mathbb{E}(X|Z)) + \epsilon\end{aligned}$$

which is the Robinson transformation.

4. See attached code.
5. We first plot $\mathbb{E}(Y|Z)$:



Next we plot $\mathbb{E}(X|Z)$:



We notice several important trends in the above plots. First, FEV appears to increase with height. Next, the probability of being a smoker appears to increase with age (we note that the estimate for $\mathbb{E}(X|Z)$ lies in $[0, 1]$, and since $X \in \{0, 1\}$, we can interpret $\mathbb{E}(X|Z)$ as the probability that a child's smoking status $= 1$). Lastly, we notice that height tends to increase with age. These three trends explain why we initially saw a positive association between FEV and smoking status: older children, who tend to be taller, have higher FEVs, and it is these same older children who are also more likely to smoke.

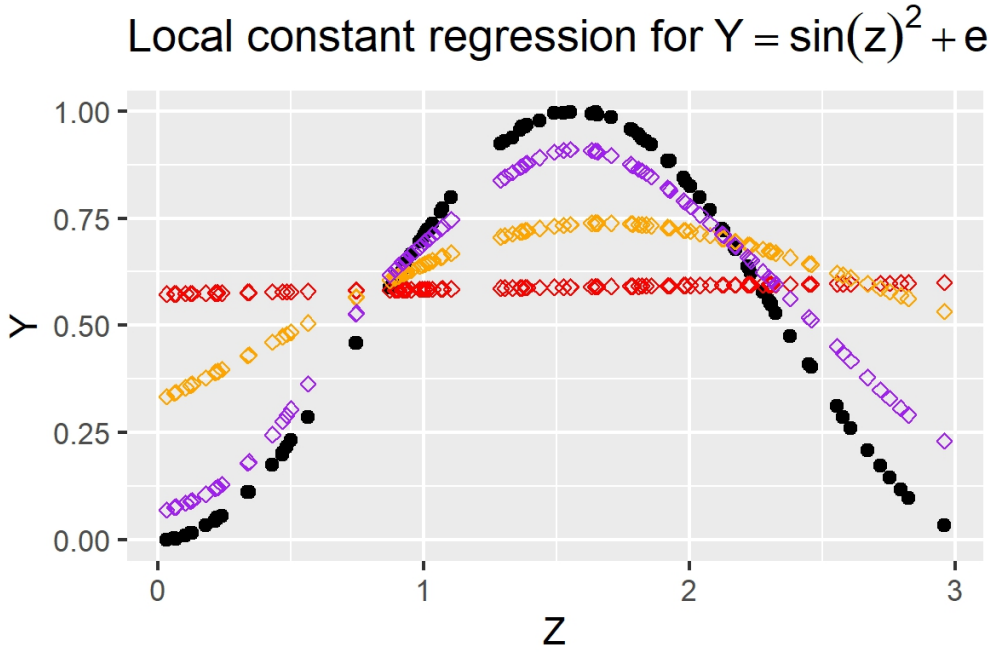
6. Below is the regression output for the Robinson procedure:

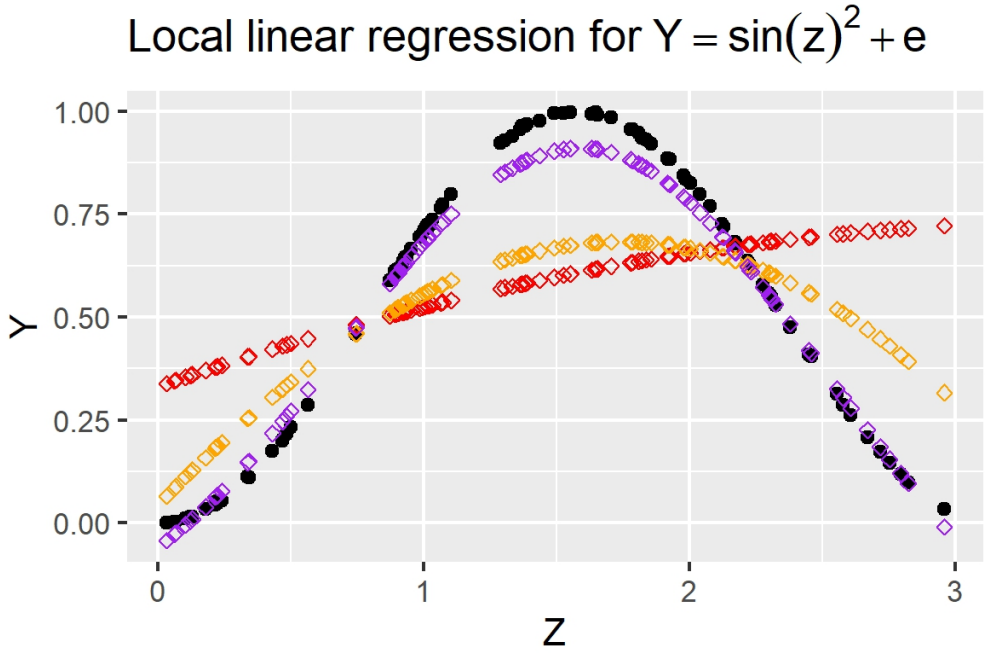
Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr($>|t|$)**
(Intercept)	-0.01272	0.01527	-0.833	0.4051
res2	-0.14882	0.05856	-2.542	0.0113

Thus, our estimate of β_1 is $\hat{\beta}_1 = -0.14882$, and the p-value associated with the hypothesis test $\beta_1 = 0$ is 0.0113. We would reject the hypothesis that there is no association ($\beta_1 = 0$) at a significance level of $\alpha = 0.05$. Thus, we find evidence of a negative association between Y (FEV) and X (smoking status) through the Robinson procedure. This is the opposite of the association we found using the model in equation (4) because using the Robinson procedure we have now controlled for the confounding variables. The true association between FEV and smoking status, when controlling for height and age, is likely a negative one.

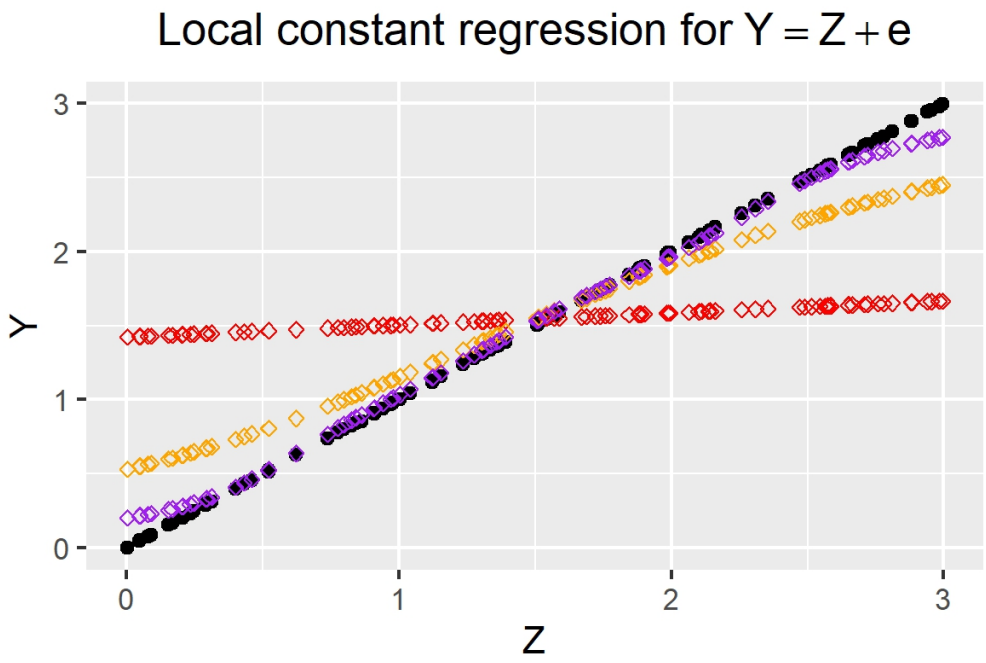
Problem 3

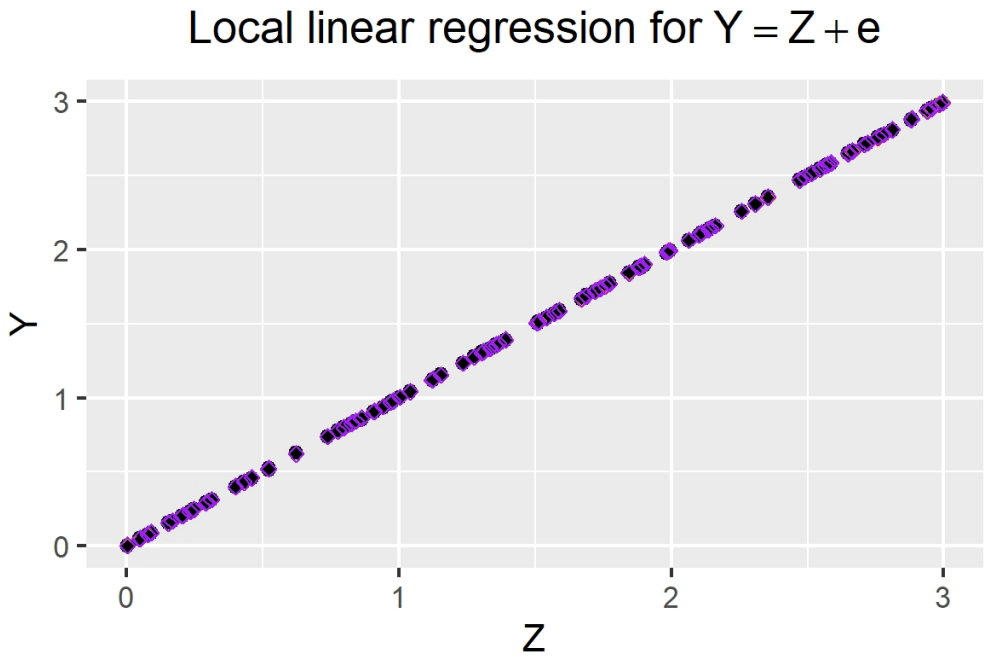
Nonparametric model Below, we plot the Nadaraya-Watson (local constant) and local linear regression for various choices of the bandwidth (red = wide, orange = medium, purple = narrow) for the data generated from the sine function:





Next, we plot the same procedure applied to the linear function:





We make the following observations about the success of our local constant and local linear regressions. First, for the sine-generated data, the local constant regression outperforms the local linear regression near the boundaries. This makes sense because the derivative of the underlying function is approaching zero near these boundaries. Next, for the linearly generated data, the local linear regression performs near-perfectly for all values of the bandwidth, while the local constant regression under-performs especially near the boundaries, since the derivative of the underlying function is not going to zero in this case.

Partially linear model First we apply Robinson's procedure to estimate β_1 from the data, having set $\beta_1 = 0$ in the model. We find:

Coefficients	Estimate	Std. Error	t value	Pr($>|t|$)
(Intercept)	0.0001254	0.0001772	0.708	0.481
res2	-0.0001373	0.0002193	-0.626	0.533

Our p-value here is 0.533, and thus we would not reject the null hypothesis that $\beta_1 = 0$ at a significance level of $\alpha = 0.05$. This makes sense because we set $\beta_1 = 0$, so we do not expect an association between Y and X.

Next, we regenerate the data 200 times and each time apply Robinson's procedure. We expect that if the procedure is successful, we will reject the null hypothesis at a significance level of $\alpha = 0.05$ about 5% of the time, that is, only by chance. Indeed, the results showed that we only rejected the null hypothesis 6 times ($6/200 = 3\%$).

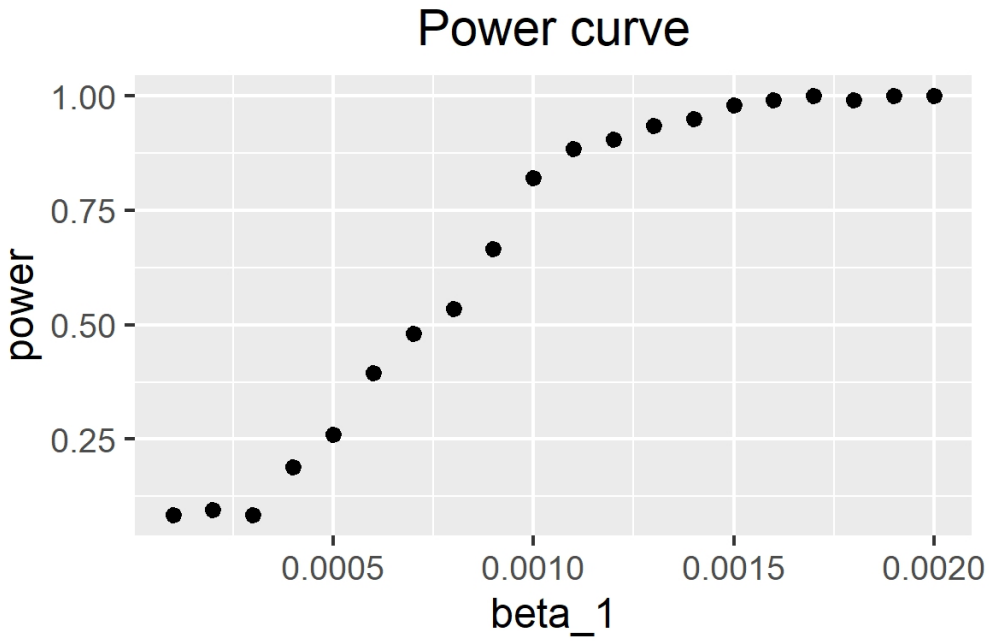
Next, we set $\beta_1 = 1$, and regenerate the data 200 more times, applying Robinson's procedure each time. We expect to reject the null hypothesis very often, since we have now enforced an association between Y and X. Indeed, the results showed that we rejected the null hypothesis at a significance level of $\alpha = 0.05$ all 200 times.

[Extra credit] Partially linear model with dependence In my version of the model, I enforced the following association between X and Z:

$$Z \sim \text{Unif}(0, 3)$$
$$X \sim \mathcal{N}(\mathbb{E}(Z), \text{Var}(Z))$$

Setting $\beta_1 = 0$, I regenerated the data 200 times, each time applying Robinson's procedure. I rejected the null hypothesis at a significance level of $\alpha = 0.05$ exactly 8 times. Since the null hypothesis is true, we expect to reject it by chance about 5% of the time, and indeed, $8/200 = 4\%$ is close to that expectation.

Next, for each value of $\beta_1 \in \{0.0001, 0.0002, ..., 0.00020\}$, I again regenerated the data 200 times and applied Robinson's procedure. I found the following power curve:



As we might expect, the power increases quickly as the true value of β_1 diverges from the null hypothesis value $\beta_1 = 0$. For the smallest values of β_1, we find powers of about $7-9\%$, which is close to what we would expect since in theory the power should approach α as $\beta_1 \to 0$. That the Robinson procedure yields powers $\geq 80\%$ for β_1 as small as 0.001

shows that it is able to accurately detect even slight associations between the variables Y and X.