# STAT 527: Assignment #1

Philip Pham

April 16, 2020

## Problem 1

(a) *Proof.* This follows from the definition of variance after some algebra.

$$
\begin{aligned}
&\mathrm{E}\left[\left(\hat{f}(x) - f(x)\right)^2\right] = \mathrm{E}\left[\left[\left(\hat{f}(x) - \mathrm{E}\left[\hat{f}(x)\right]\right) + \left(\mathrm{E}\left[\hat{f}(x)\right] - f(x)\right)\right]^2\right] \\
&= \mathrm{E}\left[\left(\hat{f}(x) - \mathrm{E}\left[\hat{f}(x)\right]\right)^2\right] + 2\,\mathrm{E}\left[\left(\hat{f}(x) - \mathrm{E}\left[\hat{f}(x)\right]\right)\left(\mathrm{E}\left[\hat{f}(x)\right] - f(x)\right)\right] + \mathrm{E}\left[\left(\mathrm{E}\left[\hat{f}(x)\right] - f(x)\right)^2\right] \\
&= \operatorname{var}\left(\hat{f}(x)\right) + \left(\mathrm{E}\left[\hat{f}(x)\right] - f(x)\right)^2 + 2\left(\mathrm{E}\left[\hat{f}(x)\right] - f(x)\right)\mathrm{E}\left[\left(\hat{f}(x) - \mathrm{E}\left[\hat{f}(x)\right]\right)\right] \\
&= \left(\mathrm{E}\left[\hat{f}(x)\right] - f(x)\right)^2 + \operatorname{var}\left(\hat{f}(x)\right).
\end{aligned}
$$

□

(b) *Proof.* The prediction error is

$$
\begin{aligned}
&\mathrm{E}\left[\left(y_{\text{new}} - \hat{f}\left(x_{\text{new}}\right)\right)^2\right] = \mathrm{E}\left[\left(f\left(x_{\text{new}}\right) + \epsilon_{\text{new}} - \hat{f}\left(x_{\text{new}}\right)\right)^2\right] \\
&= \mathrm{E}\left[\left(\epsilon_{\text{new}} + \left(f\left(x_{\text{new}}\right) - \hat{f}\left(x_{\text{new}}\right)\right)\right)^2\right] \\
&= \mathrm{E}\left[\epsilon_{\text{new}}^2\right] + 2\,\mathrm{E}\left[\epsilon_{\text{new}}\left(f\left(x_{\text{new}}\right) - \hat{f}\left(x_{\text{new}}\right)\right)\right] + \left(\mathrm{E}\left[\hat{f}(x_{\text{new}})\right] - f(x_{\text{new}})\right)^2 + \operatorname{var}\left(\hat{f}\left(x_{\text{new}}\right)\right) \\
&= \sigma^2 + \left(\mathrm{E}\left[\hat{f}(x_{\text{new}})\right] - f(x_{\text{new}})\right)^2 + \operatorname{var}\left(\hat{f}\left(x_{\text{new}}\right)\right).
\end{aligned}
$$

There is an additional $\sigma^2$ term to account for the variance of our new observation. □

(c) *Proof.*

$$\begin{aligned}
a \geq \mathrm{E}[L] &= \int_0^\infty t \; \mathrm{d}L(t) \\
&= \int_0^{a/\epsilon} t \; \mathrm{d}L(t) + \int_{a/\epsilon}^\infty t \; \mathrm{d}L(t) \\
&\geq \int_0^{a/\epsilon} t \; \mathrm{d}L(t) + \frac{a}{\epsilon}\int_{a/\epsilon}^\infty \mathrm{d}L(t) \\
&\geq \frac{a}{\epsilon}\int_{a/\epsilon}^\infty \mathrm{d}L(t) = \frac{a}{\epsilon}\,\mathrm{P}\left(L > \frac{a}{\epsilon}\right).
\end{aligned}$$

Thus, we have that

$$\frac{a}{\epsilon}\,\mathrm{P}\left(L > \frac{a}{\epsilon}\right) \leq a \Leftrightarrow \mathrm{P}\left(L > \frac{a}{\epsilon}\right) \leq \epsilon \Leftrightarrow \mathrm{P}\left(\frac{L}{a} > \frac{1}{\epsilon}\right) \leq \epsilon.$$

□

(d) *Proof.* Note that $y = X\beta + \epsilon$, where $\epsilon \sim \mathrm{N}\left(0, \sigma^2 I\right)$, so

$$\begin{aligned}
\hat{\beta} &= \left(X^\top X\right)^{-1} X^\top y_i \\
&= \left(X^\top X\right)^{-1} X^\top (X\beta + \epsilon) \\
&= \beta + \left(X^\top X\right)^{-1} X^\top \epsilon \\
\hat{\beta} - \beta &= \left(X^\top X\right)^{-1} X^\top \epsilon.
\end{aligned}$$

Thus, we have that

$$\begin{aligned}
\mathrm{E}\left[\hat{\beta} - \beta\right] &= \mathrm{E}\left[\left(X^\top X\right)^{-1} X^\top \epsilon\right] = \mathrm{E}\left[\left(X^\top X\right)^{-1} X^\top\right]\mathrm{E}[\epsilon] = 0 \\
\mathrm{var}\left(\hat{\beta} - \beta\right) &= \mathrm{E}\left[\left(X^\top X\right)^{-1} X^\top \epsilon\epsilon^\top X\left(X^\top X\right)^{-1}\right] = \sigma^2\,\mathrm{E}\left[\left(X^\top X\right)^{-1}\right].
\end{aligned}$$

If the design was fixed ($\{x_i\} \subset \mathbb{R}^p$ were deterministic), then we simply have $\mathrm{var}\left(\hat{\beta} - \beta\right) = \sigma^2\left(X^\top X\right)^{-1}$, so

$$\hat{\beta} - \beta \sim \mathrm{N}\left(0, \sigma^2\left(X^\top X\right)^{-1}\right)$$

With a random design, we appeal to Slutsky's theorem. $X^\top X = \sum_{i=1}^n x_i^\top x_i$, so by strong law of large numbers $\left(X^\top X\right)/n \xrightarrow{\text{a.s.}} \Sigma$, where $\Sigma$ is a constant.
$X^\top \epsilon = \sum_{i=1}^n x_i \epsilon_i$, where the $x_i\epsilon_i$ are i.i.d. and $\mathrm{var}(x_i\epsilon_i) = \sigma^2\Sigma$.
By CLT, we have that

$$\frac{X^\top \epsilon}{\sqrt{n}} \xrightarrow{d} \mathrm{N}\left(0, \sigma^2\Sigma\right)$$

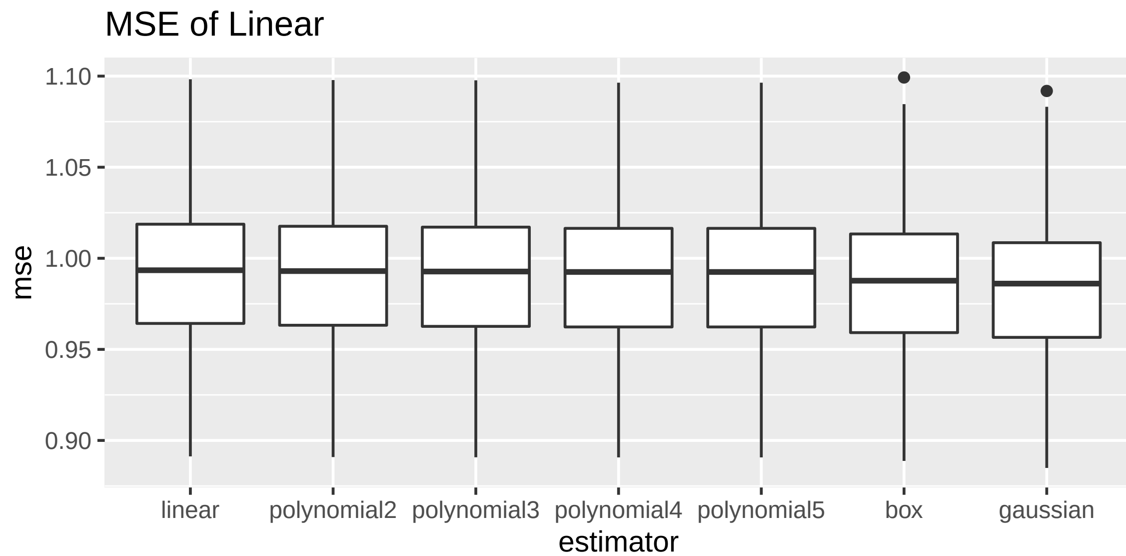
So, we can now apply Slutsky's theorem to show that

$$\begin{aligned}
\sqrt{n}\left(\hat{\beta} - \beta\right) = \left(\frac{X^\top X}{n}\right)^{-1}\frac{X^\top\epsilon}{\sqrt{n}} \xrightarrow{d} \Sigma^{-1}\,\mathrm{N}\left(0, \sigma^2\Sigma\right) &= \mathrm{N}\left(0, \sigma^2\Sigma^{-1}\Sigma\Sigma^{-1}\right) \\
&= \mathrm{N}\left(0, \sigma^2\Sigma^{-1}\right)
\end{aligned}$$

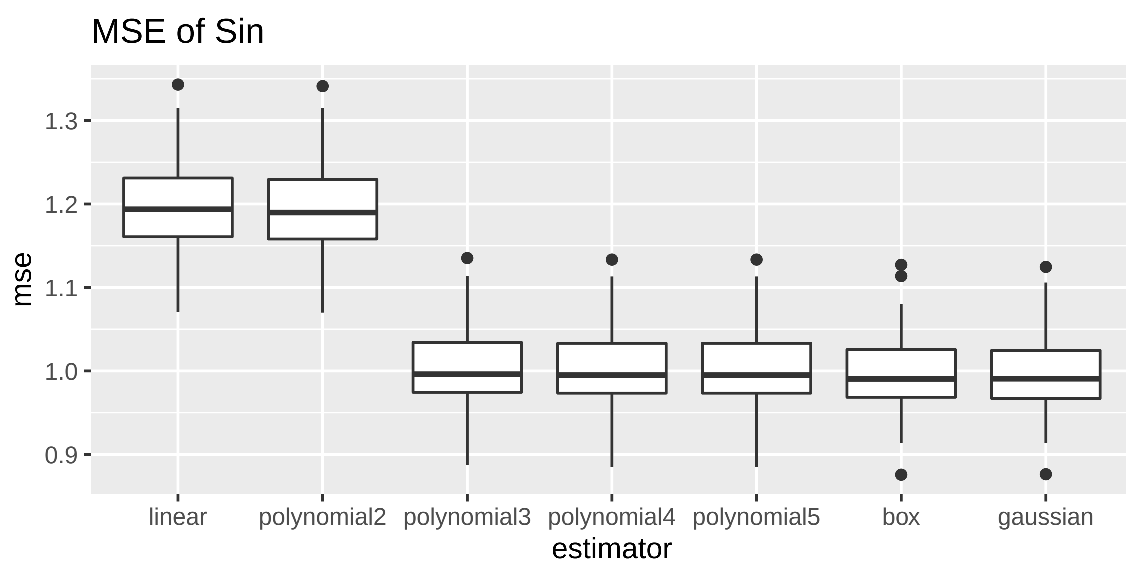
as desired. $\square$

## Problem 2

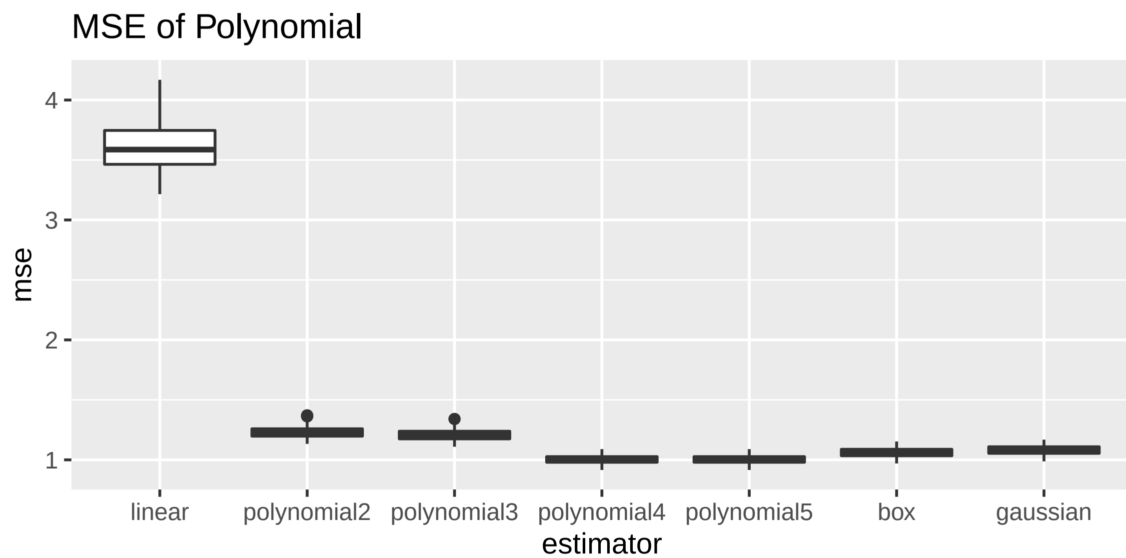
(a) $f(x) = 2x$



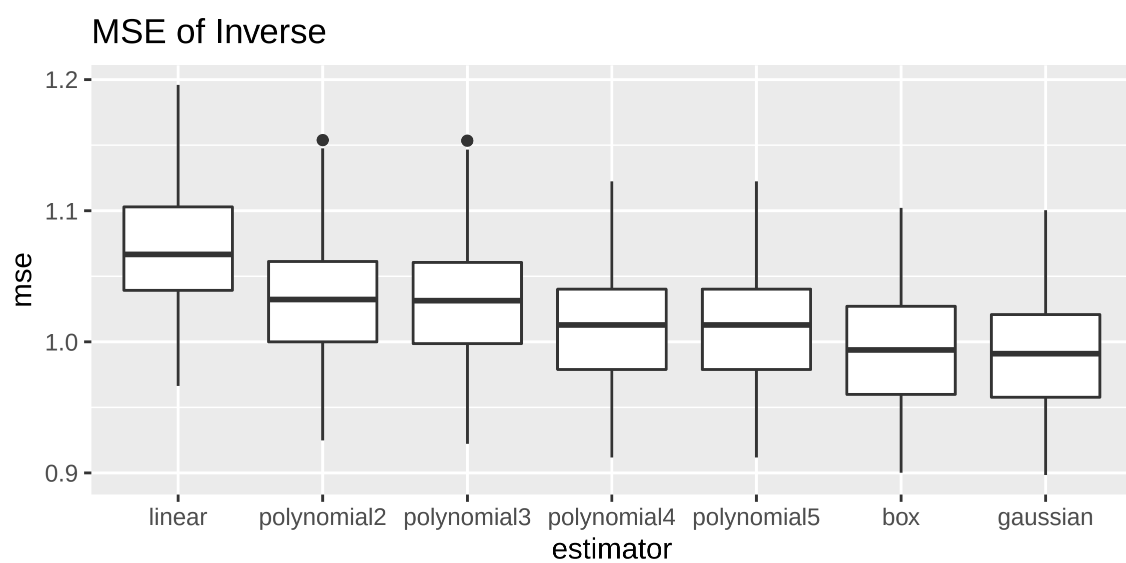

(b) $f(x) = \sin(\pi x)$



(c) $f(x) = 2x + x^3 - 6x^4$




(d) $f(x) = 1/\left(1 + (5x)^2\right)$




$n = 1000$ and 100 simulations were run for each case. Almost all the estimators achieve the optimal MSE of 1 when $f$ is linear.

When $f$ is sinusoidal, the linear and degree 2 polynomial can no longer model the data.

When $f$ is a higher-degree polynomial, the linear estimator does terribly. Unsuprisingly, the degree 4 and degree 5 polynomials do the best.

When $f$ is an inverse polynomial function, increasing the degree of the polynomial estimator helps, but only the non-parametric models are able to achieve the theoretical best error of 1.