S&DS 242/542: Homework 10

Due Wednesday, April 16, at 1PM

1. **Heteroscedastic errors I.** Consider a linear model for (X, Y) where $X \sim f_X$ for some distribution f_X, and conditional on $X = x$,

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad \text{where} \qquad \varepsilon \sim \mathcal{N}(0, v(x))$$

for a known variance function $v : \mathbb{R} \to (0, \infty)$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be IID observations distributed according to this model.

(a) Show that the maximum likelihood estimators $(\hat{\beta}_0, \hat{\beta}_1)$ are the minimizers of a weighted least-squares objective.

(b) Show that these estimators are unbiased for (β_0, β_1).

2. **Prediction interval.** Consider the standard linear model

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

where $X \sim f_X$ is independent of $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Assume for simplicity that $\sigma^2 > 0$ is known. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be IID observations distributed according to this model, and let $(\hat{\beta}_0, \hat{\beta}_1)$ be the least squares estimators of (β_0, β_1).

(a) Consider any fixed test point $x \in \mathbb{R}$. Show that the prediction $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ has a normal distribution conditional on $X_1, \ldots, X_n$, and derive its mean and variance. Use this to construct a confidence interval $I = I(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ such that

$$\mathbb{P}[\beta_0 + \beta_1 x \in I] = 1 - \alpha.$$

(b) Consider $Y_{\text{test}} = \beta_0 + \beta_1 X_{\text{test}} + \varepsilon$ where $(X_{\text{test}}, Y_{\text{test}})$ is an independent observation with the same distribution as (X_i, Y_i). Show that $\hat{\beta}_0 + \hat{\beta}_1 X_{\text{test}} - Y_{\text{test}}$ has a normal distribution conditional on $X_1, \ldots, X_n, X_{\text{test}}$, and derive its mean and variance. Use this to construct an interval $I = I(X_1, \ldots, X_n, Y_1, \ldots, Y_n, X_{\text{test}})$ such that

$$\mathbb{P}[Y_{\text{test}} \in I] = 1 - \alpha.$$

(Such an interval is called a *prediction interval* for Y.)

3. **Two-sample t-test revisited.** Let $Y_1, \ldots, Y_n \overset{IID}{\sim} \mathcal{N}(\mu_0, \sigma^2)$ and $Y_{n+1}, \ldots, Y_{n+m} \overset{IID}{\sim} \mathcal{N}(\mu_1, \sigma^2)$ be two independent samples, with unknown means $\mu_0, \mu_1 \in \mathbb{R}$ and an unknown common variance $\sigma^2 > 0$. Consider testing

$$H_0 : \mu_0 = \mu_1 \qquad \text{vs.} \qquad H_1 : \mu_1 > \mu_0$$

(a) Explain why this is equivalent to testing

$$H_0' : \beta_1 = 0 \qquad \text{vs.} \qquad H_1' : \beta_1 > 0$$

in a simple linear regression model with $n + m$ observations, where $x_1 = \ldots = x_n = 0$ and $x_{n+1} = \ldots = x_{n+m} = 1$ are fixed (non-random), and

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

with normal errors $\varepsilon_1, \ldots, \varepsilon_{n+m} \overset{IID}{\sim} \mathcal{N}(0, \sigma^2)$.

(b) Let

$$\bar{Y}^{(0)} = \frac{1}{n} \sum_{i=1}^{n} Y_i, \qquad \bar{Y}^{(1)} = \frac{1}{m} \sum_{i=n+1}^{n+m} Y_i.$$

Show that the least-squares estimator for β_1 and unbiased estimator of the residual variance σ^2 in the linear model of part (a) are given by

$$\hat{\beta}_1 = \bar{Y}^{(1)} - \bar{Y}^{(0)}, \qquad \hat{\sigma}^2 = \frac{1}{n+m-2} \left(\sum_{i=1}^{n} (Y_i - \bar{Y}^{(0)})^2 + \sum_{i=n+1}^{n+m} (Y_i - \bar{Y}^{(1)})^2 \right).$$

Explain why the t-test of H_0' vs. H_1' in this linear model is exactly the two-sample t-test of H_0 vs. H_1.

4. **Heteroscedastic errors II.** Consider the linear model of Problem 1, where $X \sim f_X$ and, conditional on $X = x$,

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad \text{where} \qquad \varepsilon \sim \mathcal{N}(0, v(x)).$$

Suppose now that $v(x)$ is unknown, and let $(\hat{\beta}_0, \hat{\beta}_1)$ be the ordinary (unweighted) least-squares estimators of (β_0, β_1).

(a) Derive a formula for the standard deviation of $\hat{\beta}_1$ conditional on $X_1, \ldots, X_n$, and show that this depends on $v(X_1), \ldots, v(X_n)$. Suggest a plug-in estimate for this standard deviation that is obtained by estimating $v(X_i)$ using the squared residual $(\hat{\beta}_0 + \hat{\beta}_1 X_i - Y_i)^2$.

(b) Perform a simulation that compares your standard error estimate in part (a) to the usual standard error estimate in linear regression software, as follows: Suppose the true parameters are $(\beta_0, \beta_1) = (0, 1)$, and $v(x) = x^2$. Generate $X_1, \ldots, X_{100} \overset{IID}{\sim} \text{Uniform}(0, 1)$ and $Y_1, \ldots, Y_{100}$ from this model, and fit

$$Y = \beta_0 + \beta_1 x + \text{error}$$

using any standard linear regression software to obtain the least-squares estimators $(\hat{\beta}_0, \hat{\beta}_1)$ and their standard errors. Compute also your standard error estimate in part (a).

Report the empirical standard deviation of $\hat{\beta}_1$ across 10000 simulations, and plot two histograms of the estimated standard error of $\hat{\beta}_1$ using these two different methods. Summarize briefly your findings.

[In R, given data vectors `X` and `Y`, you may fit the linear regression using

```
model = lm(Y ~ X)
```

The least-squares estimate $\hat{\beta}_1$ is

```
summary(model)[["coefficients"]][["X","Estimate"]]
```

and the estimated standard error of $\hat{\beta}_1$ is

```
summary(model)[["coefficients"]][["X","Std. Error"]]
```

]