LECTURE 2, PART II: BREAKING ASSUMPTIONS

Text references: Shalizi Sections 2.2, 2.3, 2.4

Changing Slopes

Look back at the population regression coefficients. Note that **the best β in linear regression** (determining the slope) **depends on the distribution of the predictor** X, through both the terms $\mathrm{Var}(X)$ and $\mathrm{Cov}(X, Y)$. *If* the true model is indeed linear, i.e., $r(X) = \beta^T X$, then this dependence goes away, as

[*HW Exercise:* Show that $\mathrm{Cov}(X, \epsilon) = 0$ without the additional assumption that ϵ and X are independent.]

But if the true model is *not* linear, i.e., $r(X)$ is not really a linear function of X, then this is not true, and the the population slope coefficients depend on the distribution of X

Q: What does this mean in practice?

Well, if we are applying linear regression to a case in which the true relationship nonlinear (say by means of approximation, which is always the case to some extent), then our coefficient estimates will depend on which predictor values we observe.

For example, if $Y = \sqrt{X} + \varepsilon$ (with $\varepsilon \sim N(0, \sigma^2)$, independent of X), then $\beta = \mathrm{Var}(X)^{-1}\mathrm{Cov}(X, Y)$, and in practice $\widehat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$ is going to depend highly on the distribution of X, as we will see in R Demo 2.1.

R Demo 2.1 (Slope varies with location)

(a) Suppose

$$Y = \sqrt{X} + \epsilon$$

with $\epsilon \sim \mathcal{N}(0, 0.05^2)$ (i.e. the standard deviation of the noise was 0.05). Generate and plot data from three different distributions for X: $X \sim \mathrm{Unif}(0, 1)$ (*green squares*), $X \sim \mathcal{N}(0.5, 0.01)$ (*blue circles*), and $X \sim \mathrm{Unif}(2, 3)$ (*red triangles*). Show the location of the actual sample points with axes tick marks.

(b) Fit linear regression lines and add them, in matching colors.

(c) Combine the data and fit a regression line to all the data. Compare to the true regression function. What are your conclusions? (Note that R^2 is

not a good measure of the quality of the fit; see Shalizi Sec 2.2.1.1)

[Ref: Shalizi Sec 2.2.1]

Omitted Variables

What happens if we assume the linear model

$$Y = \beta_0 + \beta^T X + \varepsilon, \tag{1}$$

but in reality the relationship is

$$Y = \beta_0 + \beta^T X + \gamma^T Z + \widetilde{\varepsilon}. \tag{2}$$

where Z is a variable that depends on X (and e.g. $\widetilde{\varepsilon} \sim \mathcal{N}(0, \widetilde{\sigma}^2)$)?

Q: When do you think it is OK to omit a variable Z from the model (2)? Discuss.

Hint 1: How is the regression of Y on X defined?

Hint 2: Our formulas for the population regression coefficients tell us that in (2),

$$\begin{pmatrix} \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Z) \\ \mathrm{Cov}(Z,X) & \mathrm{Var}(Z) \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Z,Y) \end{pmatrix}.$$

Some tedious algebra gives the following:

$$\begin{aligned} \beta &= \frac{\mathrm{Cov}(X,Y) - \mathrm{Cov}(X,Z)\mathrm{Cov}(Z,Y)/\mathrm{Var}(Z)}{\mathrm{Var}(X) - \mathrm{Cov}(X,Z)^2/\mathrm{Var}(Z)}, \\ \gamma &= \frac{\mathrm{Cov}(Z,Y) - \mathrm{Cov}(Z,X)\mathrm{Cov}(X,Y)/\mathrm{Var}(X)}{\mathrm{Var}(Z) - \mathrm{Cov}(Z,X)^2/\mathrm{Var}(X)}. \end{aligned}$$

Note that even small correlations between X and Z can lead to large differences when fitting the regression model (1); in the R working examples we will see that changing $\mathrm{Cor}(X,Z)$ from 0.1 to -0.1 can cause a jump in the coefficients in (1).

R Demo 2.2 (Omitted Variables Bias)

(a) Suppose X and Z are both $\mathcal{N}(0, 1)$, but with a positive correlation of 0.1. Generate

$$Y \sim \mathcal{N}(X + Z, 0.01)$$

and show a scatterplot of all three variables together ($n = 100$).

(b) Now change the correlation between X and Z to -0.1. **The marginal distributions, and the distribution of Y given X and Z, are unchanged.** Show a scatterplot of all three variables together ($n = 100$).

(c) Now omit Z and plot Y versus X for both cases; use tick-marks on the axes to show the marginal distributions of X and Y. (Use black for the data with positive correlation between X and Z; use blue for the data with negative correlation between X and Z.) What do you see?

(d) Regress Y on X with Z omitted for the two cases. (Use black for the data with positive correlation between X and Z; use blue for the data with negative correlation between X and Z.) What do you see?

[Ref: Shalizi Sec 2.2.2]

Variable Transformations

Imagine our model is

$$Y = \log(X) + \varepsilon$$

[See Shalizi Figures 2.5 and 2.6]. To use linear regression, we could either transform Y or X:

Which is better? That depends on the application/context and the properties of the error/noise terms, but in many cases it makes sense to transform the predictors. A few notes:

1. Transforming the response changes the scale in which the model is fit, i.e., the model for $\widetilde{Y} = \exp(Y)$ is

$$\widetilde{Y} = X \cdot \exp(\varepsilon),$$

   which is a multiplicative error model, not additive. (There are some cases where transforming responses, using e.g. $\sqrt{Y}, \log(Y), \log(Y + c), 1/Y$, actually brings the data closer to satisfying assumptions like normality and/or equal variance.)

2. For a situation like $Y = \log(X) + Z^{1/3} + \varepsilon$, it's not at all obvious how to

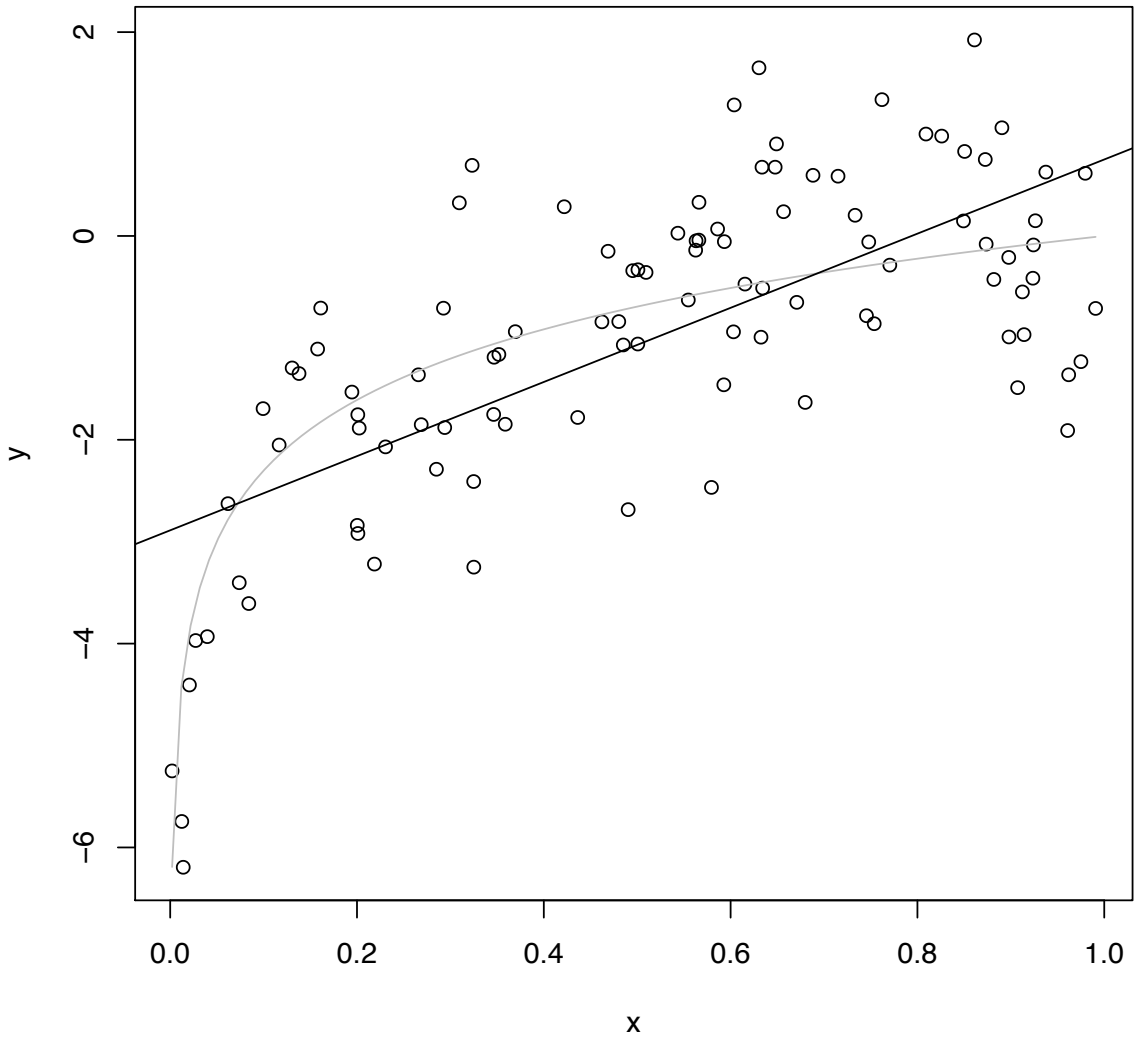
transform Y, but it is easy to transform the predictor variables (different transformations for X and Z)

3. There is a very versatile class of models based on transforming predictors.

Regarding the third point, consider a model of the form

If we take $f_i(X) = X_i$, then we get back linear regression. But this encapsulates a lot more than just linear regression; e.g., we could include *interaction terms* via $f_i(X) = X_j X_k$, we could include *polynomial terms* via $f_i(X) = X_i^k$, and so on.

Two basic strategies: first, fix some dictionary of functions $f_1, \ldots f_m$ ahead of time, or second, try to estimate appropriate ones from data. We will see in future lectures how to carry out both of these approaches.




```
x <- runif(100)
y <- rnorm(100,mean=log(x),sd=1)
plot(y~x)
curve(log(x),add=TRUE,col="grey")
abline(lm(y~x))
```

FIGURE 2.5: *Sample of data for* $Y|X \sim \mathcal{N}(\log X, 1)$. *(Here* $X \sim \text{Unif}(0,1)$*, and all logs are natural logs.) The true, logarithmic regression curve is shown in grey (because it's not really observable), and the linear regression fit is shown in black.*

Figure 1: Shalizi Figure 2.5

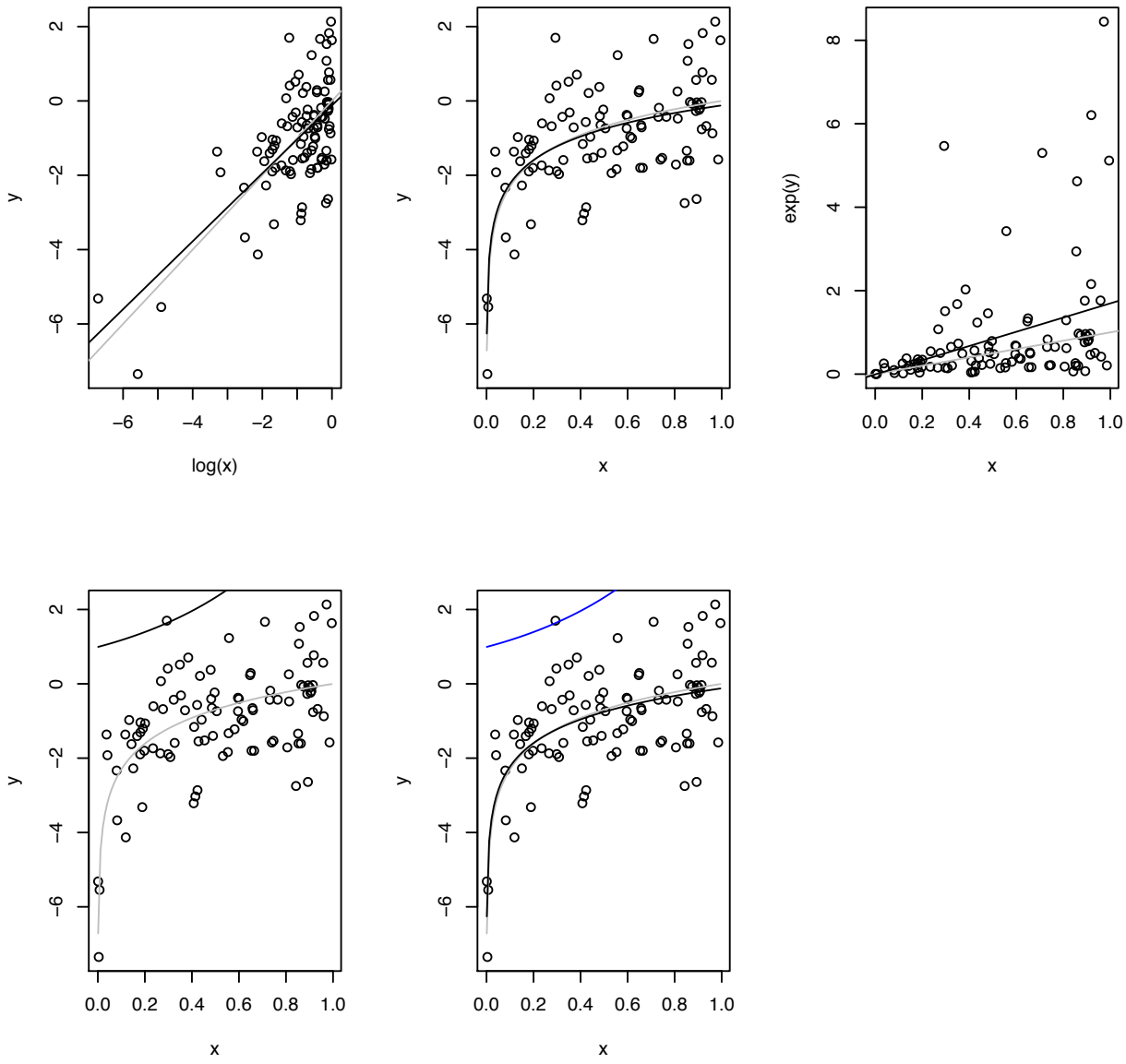


Figure 2.6 Transforming the predictor (left column) and the response (right) in the data from Figure 2.5, shown in both the transformed coordinates (top) and the original coordinates (middle). The bottom figure super-imposes the two estimated curves (transformed X in black, transformed Y in blue). The true regression curve is always in grey. (R code deliberately omitted; reproducing this is Exercise 2.4.)

Figure 2: Shalizi Figure 2.6