Chapter 4

Errors in Predictors

The regression model $Y = X\beta + \varepsilon$ allows for Y being measured with error by having the ε term, but what if the X is measured with error? In other words, what if the X we see is not the X used to generate Y?

Consider the simple regression (x_i, y_i) for $i = 1, \ldots n$.

$$\begin{aligned} y_i &= \eta_i + \varepsilon_i \\ x_i &= \xi_i + \delta_i \end{aligned}$$

where the errors ε and δ are independent. Suppose the true underlying relationship is

$$\eta_i = \beta_0 + \beta_1 \xi_i$$

but we only see (x_i, y_i). Putting it together, we get

$$y_i = \beta_0 + \beta_1 x_i + (\varepsilon_i - \beta_1 \delta_i)$$

Suppose we use least squares to estimate β_0 and β_1. Let's assume $E\varepsilon_i = E\delta_i = 0$ and that $\text{var}\ \varepsilon_i = \sigma^2$, $\text{var}\ \delta_i = \sigma^2_\delta$. Let

$$\sigma^2_\xi = \sum(\xi_i - \bar{\xi})^2/n \qquad \sigma_{\xi\delta} = cov(\xi, \delta)$$

where ξ are the true values of X and not random variables but we could (theoretically since they are not observed) compute statistics using their values. Now $\hat{\beta}_1 = \sum(x_i - \bar{x})y_i / \sum(x_i - \bar{x})^2$ and after some calculation we find that

$$E\hat{\beta}_1 \approx \beta_1 \frac{(\sigma^2_\xi + \sigma_{\xi\delta})}{(\sigma^2_\xi + \sigma^2_\delta + 2\sigma_{\xi\delta})}$$

If there is no relation between ξ and δ, this simplifies to

$$E\hat{\beta}_1 \approx \beta_1 \frac{\sigma^2_\xi}{(\sigma^2_\xi + \sigma^2_\delta)} = \beta_1 \frac{1}{1 + \sigma^2_\delta/\sigma^2_\xi}$$

So in general $\hat{\beta}_1$ will be biased (regardless of the sample size and typically towards zero). If σ^2_δ is small relative to σ^2_ξ then the problem can be ignored. In other words, if the variability in the errors of observation of X are small relative to the range of X then we need not be concerned. If not, it's a serious problem and other methods such as fitting using orthogonal rather than vertical distance in the least squares fit should be considered.

For prediction, measurement error in the x's is not such a problem since the same error will apply to the new x_0 and the model used will be the right one.

For multiple predictors, the usual effect of measurement errors is to bias the $\hat{\beta}$ in the direction of zero.

One should not confuse the errors in predictors with treating X as a random variable. For observational data, X could be regarded as a random variable, but the regression inference proceeds conditional on a fixed value for X. We make the assumption that the Y is generated conditional on the fixed value of X. Contrast this with the errors in predictors case where the X we see is not the X that was used to generate the Y.

For real data, the true values of the parameters are usually never known, so it's hard to know how well the estimation is working. Here we generate some artificial data from a known model so we know the true values of the parameters and we can tell how well we do: `runif()` generates uniform random numbers and `rnorm()` generates standard normal random numbers.

Because you will get different random numbers, your results will not exactly match mine if you try to duplicate this.

```
> x <- 10*runif(50)
> y <- x+rnorm(50)
> gx <- lm(y ~ x)
> summary(gx)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1236     0.2765   -0.45     0.66
x             0.9748     0.0485   20.09   <2e-16

Residual standard error: 1.02 on 48 degrees of freedom
Multiple R-Squared: 0.894,      Adjusted R-squared: 0.891
F-statistic:  403 on 1 and 48 degrees of freedom,       p-value:    0
```

True values of the regression coeffs are 0 and 1 respectively. What happens when we add some noise to the predictor?

```
> z <- x + rnorm(50)
> gz <- lm(y ~ z)
> summary(gz)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3884     0.3248     1.2     0.24
z             0.8777     0.0562    15.6   <2e-16

Residual standard error: 1.27 on 48 degrees of freedom
Multiple R-Squared: 0.835,      Adjusted R-squared: 0.832
F-statistic:  244 on 1 and 48 degrees of freedom,       p-value:    0
```

Compare the results - notice how the slope has decreased. Now add even more noise:

```
> z2 <- x+5*rnorm(50)
> gz2 <- lm(y ~ z2)
> summary(gz2)
Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.736      0.574    4.77  1.8e-05
z2             0.435      0.101    4.32  7.7e-05

Residual standard error: 2.66 on 48 degrees of freedom
Multiple R-Squared: 0.28,      Adjusted R-squared: 0.265
F-statistic: 18.7 on 1 and 48 degrees of freedom,      p-value: 7.72e-05
```

Compare again — the slope is now very much smaller. We can plot all this information in Figure 4.1.

```
> matplot(cbind(x,z,z2),y,xlab="x",ylab="y")
> abline(gx,lty=1)
> abline(gz,lty=2)
> abline(gz2,lty=5)
```

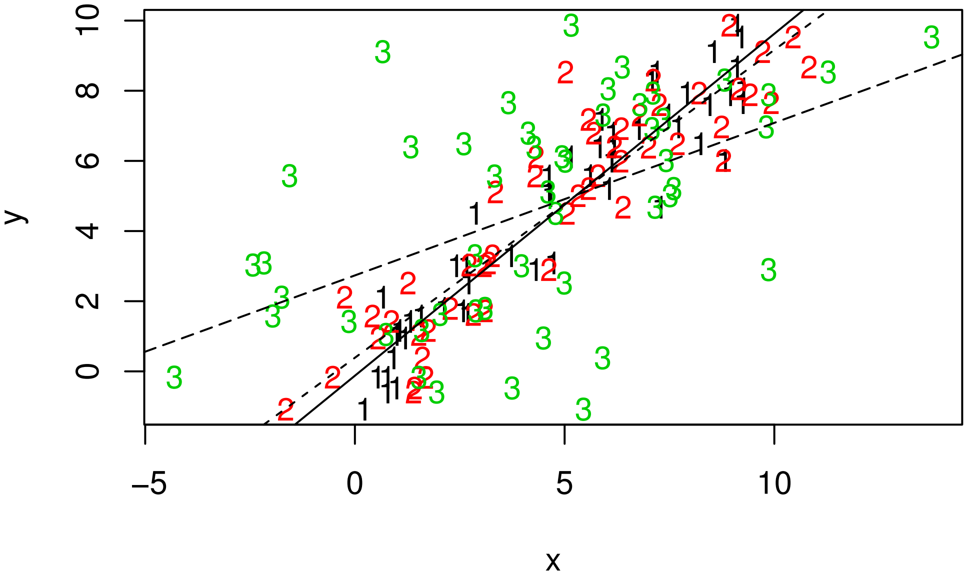

Figure 4.1: Original x shown with “1”, with small error as “2” and with large error as “3”. The regression lines for the no measurement error, small error and large error are shown as solid, dotted and dashed lines respectively.

This was just one realization - to get an idea of average behavior we need to repeat the experiment (I’ll do it 1000 times here). The slopes from the 1000 experiments are saved in the vector bc:

```
> bc <- numeric(1000)
> for(i in 1:1000){
+ y <- x + rnorm(50)
+ z <- x + 5*rnorm(50)
+ g <- lm(y ˜ z)
+ bc[i] <- g$coef[2]
+ }
```

Now look at the distribution of `bc`.

```
> summary(bc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.0106  0.2220  0.2580  0.2580  0.2950  0.4900
```

Given that the variance of a standard uniform random variable is 1/12, $\sigma^2_\delta = 25$ and $\sigma^2_\xi = 100/12$, we'd expect the mean to be 0.25. Remember that there is some simulation variation and the expression for the bias is only approximation, so we don't expect them to match exactly.