Multiple linear regression: Inference, Part II

Patrick Breheny

February 1

Introduction

- Today in lab we're going to apply the formulas we derived last time to our ozone data and go through several examples of quantifying the variability of estimates and predictions
- We'll also take a closer look at what exactly is meant by "linear" regression and linear-versus-nonlinear dependence among the explanatory variables

Residuals in R

- Let's begin by re-fitting our model from last time, storing the fit, and inspecting various components of the fit:

```
fit <- lm(Ozone~Solar+Wind+Temp+Day)
fit$coefficients
fit$fitted.values
fit$residuals
fit$rank
fit$df.residual
```

- Note that

```
n <- nrow(ozone)
p <- fit$rank
n-p
```

is equal to `fit$df.residual`

Residuals in SAS

- In SAS, one can see the residuals and fitted values by passing along a `P` option to the `MODEL` statement:

```
PROC REG DATA=ozone;
  MODEL Ozone = Solar Wind Temp Day / P;
RUN;
```

- Note that the residual degrees of freedom and residual sum of squares are also reported

Estimating σ^2

- We showed last time that dividing the residual sum of squares by $n-p$ produces an unbiased estimator of σ^2:
    - In R,

      ```
      sig2 <- sum(fit$residuals^2)/fit$df.residual
      sig <- sqrt(sig2)
      ```
    - In SAS, $\hat{\sigma}$ is reported as "Root MSE" (the residual sum of squares is also referred to as the "squared error", and dividing by $n-p$ is akin to taking the "mean squared error")
- Note that the standard deviation of ozone concentrations is 33.3, whereas $\hat{\sigma} = 21.0$

Estimating the variance of $\hat{\beta}$

- Now we can estimate the variance of $\hat{\beta}$:

```
X <- as.matrix(cbind(1,ozone[,-1]))
VarB <- sig2*solve(crossprod(X))
```

- Alternatively, the function `summary` computes additional information about the least squares fit:

```
summ <- summary(fit)
summ$sigma
summ$cov.unscaled
summ$sigma^2*summ$cov.unscaled
```

- In SAS, the you can pass the `COVB` option to the `MODEL` statement to obtain the estimated variance-covariance matrix of $\hat{\beta}$

Estimating the variance of $\hat{\beta}$

- Now that we have $\widehat{\mathrm{Var}}(\hat{\beta})$, we are in a position to quantify the variability of our estimates, as well as combinations of estimates
- An obvious place to start is with the standard errors of our regression coefficients:

  `sqrt(diag(VarB))`
- Note that this agrees with the reported standard errors from `summary(fit)` and `PROC REG`

Variance of linear combinations

- However, we can also estimate the variance/standard error of combinations of parameters
- Suppose we are interested in some linear combination of parameters $\boldsymbol{\lambda}^T\boldsymbol{\beta}$:

$$\mathrm{Var}(\boldsymbol{\lambda}^T\hat{\boldsymbol{\beta}}) = \boldsymbol{\lambda}^T\mathrm{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{\lambda}$$

- So, for instance, suppose we wanted to know about the effect on ozone concentrations of simultaneously lowering the wind speed by 5 mph and raising the temperature by 10 degrees

Variance of linear combinations in R/SAS

- In R,

```
lambda <- c(0,0,-5,10,0)
crossprod(lambda,fit$coefficients)
sqrt(t(lambda) %*% VarB %*% lambda)
```

- So the effect of this change in the weather will be to raise ozone concentrations on average 34.9 ppb $\pm$ 3.15 ppb
- The ESTIMATE statement in SAS accomplishes the same thing, although for some inexplicable reason, it is not available in PROC REG; you have to use PROC GLM:

```
PROC GLM Data=ozone;
    MODEL Ozone = Solar Wind Temp Day;
    ESTIMATE '-5*Wind+10*Temp' Wind -5 Temp 10;
RUN;
```

The point of the off-diagonal elements

- Note that we would not get the right answer if we ignored the covariance between $\hat{\beta}_3$ and $\hat{\beta}_4$:

```
25*VarB[3,3] + 100*VarB[4,4]
```

- Furthermore, the uncertainty in estimating the effect of dropping wind speed and raising temperature is not the same as the uncertainty involved in raising wind speed and raising temperature:

```
lambda <- c(0,0,5,10,0)
sqrt(t(lambda) %*% VarB %*% lambda)
```

- The intuitive explanation for this is that wind speed and temperature were negatively correlated, so there is a lot more information in the data set about what would happen if one was raised and the other lowered than if they were both raised

Prediction

- Let's revisit our two sample days from last week:
  - A: Solar=180, Wind=15, Temp=70, Day=274
  - B: Solar=180, Wind=5, Temp=90, Day=274
- We could predict the average ozone concentration of these two days using

  ```
  a <- c(1,180,15,70,274)
  b <- c(1,180,5,90,274)
  ```

  in place of `lambda`
- This would indicate that Day A can expect to have an ozone concentration of 5.2 ± 5.4, while Day B can expect to have an ozone concentration of 74.9 ± 4.3

Prediction (cont'd)

- This estimate of variability does not, however, accurate represent the uncertainty concerning the actual concentration of day 274
- The $\pm$ number only takes into account our uncertainty about the mean ozone concentration, not the inherent daily variability in ozone levels
- The actual variability of the ozone concentration of day 274 is the larger number

$$\mathrm{Var}(\mathbf{x}^T\hat{\boldsymbol{\beta}} + \epsilon) = \mathbf{x}^T\mathrm{Var}(\hat{\boldsymbol{\beta}})\mathbf{x} + \sigma^2$$

Prediction in R/SAS

- So in R,

  ```
  sqrt(t(a) %*% VarB %*% a + sig2)
  ```

- In SAS, you can add observations to the data set, and then request intervals for the mean with CLM and intervals for individual days with CLI:

  ```
  PROC REG DATA=ozone;
    MODEL Ozone = Solar Wind Temp Day / P CLM CLI;
  RUN;
  ```

R^2 in R/SAS

- Finally, let's calculate R^2:

```
var(Ozone)
var(fit$residuals) + var(fit$fitted.values)
TSS <- crossprod(Ozone-mean(Ozone))
RSS <- crossprod(fit$residuals)
MSS <- crossprod(fit$fitted.values-mean(fit$fitted.valu
MSS/TSS
cor(fit$fitted.values,Ozone)^2
```

- R^2 is also reported by default with `summary(fit)` and by `PROC REG`

Interpretation of R^2

- The fact that our model is able to explain 62% of the variability in ozone concentrations is reassuring that our model fits the data reasonably well
- If, on the other hand, $R^2 = .08$ (not at all uncommon), we might have doubts
- A low R^2 could be caused simply by large random effects and inherent unpredictability, but it could also be a signal of a bad model which leaves out many important factors
- Furthermore, if there are important factors left out of the model, perhaps they are confounders that would alter the model's conclusion if they were incorporated

Interpretation of R^2 (cont'd)

- However, it bears reminding that a high R^2 does not rule out the possibility of confounding or prove that the model is correct
- For example, over the period 1950-1999, the correlation in the U.S. between deaths from lung cancer and the purchasing power of the dollar was 0.95 (*i.e.*, $R^2 = .9$)
- Inflation, however, does not cause lung cancer!

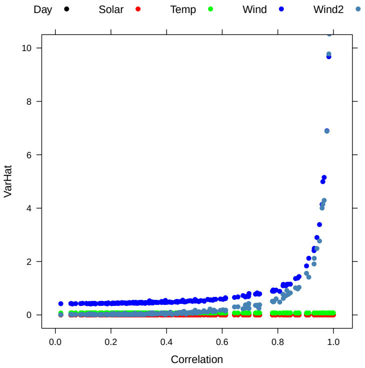
Close to linear dependence

- We have said that linearly dependent variables cause problems in linear regression, and seen the kinds of error messages they provoke in SAS and R
- Do highly correlated, but not strictly dependent variables cause problems?
- Indeed they do; try

```
Wind2 <- Wind + rnorm(n,mean=0,sd=20)
cor(Wind,Wind2)
summ <- summary(lm(Ozone~Solar+Wind+Temp+Day))
summ2 <- summary(lm(Ozone~Solar+Wind+Temp+Day+Wind2))
diag(summ$sigma^2*summ$cov.unscaled)
diag(summ2$sigma^2*summ2$cov.unscaled)
```

Close to linear dependence (cont'd)

- Not much increase in the variance of $\hat{\beta}_{Wind}$ ...
- However, as we decrease the SD of the random noise (and thereby increase the correlation between `Wind` and `Wind2`), the variance increases without bound



Nonlinear functions do not cause problems

- However, it is important to note that it is only *linear* dependence that causes problems
- For example, suppose we introduce

```
WindSq <- Wind^2
summary(lm(Ozone~Solar+Wind+WindSq+Temp+Day))
```

- Even though `Wind` and `WindSq` are completely dependent, this does not cause any problems (quite the contrary: R^2 goes up from 62% to 70%)

"Linear" regression?

- But wait, if we've got terms like $\texttt{Wind}^2$ in the model, is our model still "linear"?
- Yes, the model is still considered to be linear, because it's still linear with respect to the parameters $\{\beta_j\}$, and therefore estimation and inference work in exactly the same way, regardless of whether or not the variables happen to be nonlinear transformations of each other
- The same goes for transformations of the outcome variable as well

Transformation

- So, for example, you may have been troubled by our earlier result that the mean ozone concentration for Day A was 5.2 $\pm$ 5.4, as this would seem to suggest that negative ozone concentrations are likely
- One way to enforce positive values is to model the log of the ozone concentrations:

```
fit <- lm(log(Ozone)~Solar+Wind+Temp+Day)
summary(fit)
```

- Any resulting predictions or estimates would then be on the log scale, and once the inverse transformation was applied, would necessarily be positive

Factors

- One final issue while we're on the topic of transformations is the issue of categorical explanatory variables (sometimes called *factors*)
- Suppose we're studying the relationship between x and y, but we wish to adjust for gender (which can take on one of two values, "Male" or "Female")
- We of course need to quantify this for our model; one way of doing this is to introduce *indicator variables* (also called *dummy variables*): `Male` $=1$ if `Gender=''Male''`, 0 if `Gender=''Female''`

Linear dependence among factors

- An indicator variable `Female` could also be created, but caution is in order:

  `Female = 1 - Male`

  and thus, assuming that we have an intercept in our model, the two variables will be linearly dependent
- One option, of course, is to eliminate the intercept; this would mean that the coefficient β_{Male} would be the intercept for the males, while β_{Female} would be the intercept for the females

Linear dependence among factors (cont'd)

- The other option would be to only include the coefficient for males
- This model is functionally equivalent to the other model (all the fitted values, residuals, R^2, etc. will be identically the same), but the meaning of the regression coefficients will be different
- Now, β_0 will be the intercept for the females, and $\beta_0 + \beta_{Male}$ will be the intercept for the males
- We will go into more detail, with real examples, next Tuesday