2/2/23
Day 4

## 3.2: Multiple Linear Regression

If predictors are correlated, then using separate regression curves can mislead you.

(3.19) $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$

### 3.2.1: Estimating the Regression Coefficients

We again choose $\vec{\beta}$ to minimize

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \vec{\beta} \cdot (1 \oplus x_i))$$

Say
- $X_1$ and $X_2$ are correlated,
- $Y$ regressed on $X_1$ has a significant $\beta_1$
- $Y$ regressed on $X_2$ has a significant $\beta_2$
- $Y$ regressed on $(X_1, X_2)$ has significant $\beta_1$, insignificant $\beta_2$

This example shows how individual modeling can fool you.

### 3.2.2: Some Important Questions

1) Is at least one of the predictors $X_1, \ldots, X_p$ useful in predicting the response?
2) Do all predictors help explain $Y$, or only a subset thereof?
3) How well does the model fit the data?
4) Given a set of predictor values, what response value should we predict, and how accurate is the prediction?

**One: Is there a relationship between the response and the predictors?**

- In the multiple regression setting with $p$ predictors, we answer this by doing the hypothesis test

$H_0: \beta_1 = \dots = \beta_p = 0$

$H_a: \beta_1 \neq 0$ or $\dots$ or $\beta_p \neq 0$

To perform this hypothesis test, we compute the ~~hypothesis~~ F-statistic

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

If the linear model assumptions are correct, then
$\mathbb{E}(RSS/(n-p-1)) = \sigma^2 = Var(\varepsilon)$.
and if $H_0$ is true then
$\mathbb{E}((TSS-RSS)/p) = \sigma^2$

If $H_0$ is true, then $F \approx 1$. If $H_a$ is true, then $\mathbb{E}((TSS-RSS)/p) > \sigma$, so $F > 1$.

- The smaller $n$ is, the larger $F$ needs to be to reject $H_0$.
- When $H_0$ is true and $\varepsilon \sim N(0, \sigma^2)$, $F$ follows an $F$ distribution.

Sometimes we want to test if a subset of parameters are 0; i.e.,

$H_0: \beta_{p-q+1} = \beta_{p-q+2} = \dots = \beta_p = 0$

First where we're testing to drop the last $q$ variables. Fit a model with the $n-q$ variables remaining. Then, let $RSS_0$ be the residual sum of squares for this smaller model. Now, calculate

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

It is improper to look at individual t-statistics when ~~doing~~ testing for the significance of multiple predictors.

## Two: Deciding on important variables

Once we have a significant F, we ask which predictors are significant.

Variable selection is studied in Ch. 6. This is an overview.

Mallow's $C_p$, AIC, BIC, adjusted $R^2$

- Forward selection: Start w/ null, iteratively add variables until some stopping criterion is met.
- Backward selection: Start w/ full model, iteratively remove according to largest p value. until halt.
- Mixed selection: Start w/ null model, add variables. If a p value gets too high, drop it. Keep going.

## Three: Model fit

- RSE and $R^2$
- $R^2$ always increases w/ a new variable, but if the improvement is small, that's evidence for excluding that variable.
- We're interested in big ~~fatness~~ improvements in $R^2$ and RSE.

In general, RSE is

$$RSE = \sqrt{\frac{1}{n-p-1} RSS},$$

so if the decrease in RSS is small compared to the increase in $\frac{1}{n-p-1}$ as $p \uparrow$, then RSE can increase.