Statistics 203: Introduction to Regression and Analysis of Variance

Multiple Linear Regression: Inference & Polynomial

Jonathan Taylor
Today

- Inference: trying to “reduce” model.
- Polynomial regression.
- Splines + other bases.
Summary of last class

- \[ Y_{n \times 1} = X_{n \times p} \beta_{p \times 1} + \varepsilon_{n \times 1} \]

- \[ \hat{Y} = HY, \quad H = X(X^t X)^{-1}X^t \]

- \[ e = (I - H)Y \]

- \[ \|e\|^2 \sim \sigma^2 \chi_{n-p}^2 \]

- Generally, if \( P \) is a projection onto a subspace \( \tilde{L} \) such that \( P(X\beta) = 0 \), then

\[ \|PY\|^2 = \|P(X\beta + \varepsilon)\|^2 = \|P\varepsilon\|^2 \sim \sigma^2 \chi_{\text{dim}\tilde{L}}^2. \]
$R^2$ for multiple regression

\[ SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \|Y - \hat{Y}\|^2 \]

\[ SSR = \sum_{i=1}^{n} (\bar{Y} - \hat{Y}_i)^2 = \|\hat{Y} - \bar{Y}1\|^2 \]

\[ SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \|Y - \bar{Y}1\|^2 \]

\[ R^2 = \frac{SSR}{SST} \]
As we add more and more variables to the model – even random ones, $R^2$ will go to 1.

Adjusted $R^2$ tries to take this into account by replacing sums of squares by “mean” squares

$$R^2_a = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{MSE}{MST}.$$ 

Here is an example.
Inference in multiple regression

- $F$-statistics.
- Dropping a subset of variables.
- General linear hypothesis.
Testing $H_0 : \beta_2 = 0$

- Can be tested with a $t$-test:

$$T = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)}.$$

- Alternatively, using an $F$-test with a “full” and “reduced” model
  - (F) $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$
  - (R) $Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$

$F$-statistic: under $H_0 : \beta_2 = 0$

$$SSE_F = \|Y - \hat{Y}_F\|^2 \sim \sigma^2 \chi^2_{n-3}$$

$$SSE_R = \|Y - \hat{Y}_R\|^2 \sim \sigma^2 \chi^2_{n-2}$$

$$SSE_F - SSE_R = \|\hat{Y}_F - \hat{Y}_R\|^2 \sim \sigma^2 \chi^2_1$$

and $SSE_F - SSE_R$ is independent of $SSE_F$ (see details).
Testing $H_0 : \beta_2 = 0$

- Under $H_0$

$$F = \frac{(SSE_F - SSE_R)/1}{SSE_F/(n-3)} \sim F_{1,n-3}.$$ 

- Reject $H_0$ at level $\alpha$ if $F > F_{1,n-3,1-\alpha}$.
Some details

- \( SSE_F \sim \sigma^2 \chi^2_{n-3} \) if the full model is correct, and
- \( SSE_R \sim \sigma^2 \chi^2_{n-2} \) if \( H_0 \) is correct because

\[
H_FY = H_F(X\beta + \epsilon) = X\beta + H_F\epsilon
\]

\[
H_RY = H_R(X\beta + \epsilon) = X\beta + H_R\epsilon \quad \text{(under } H_0)\]

If \( H_0 \) is false \( SSE_R \) is \( \sigma^2 \) times a non-central \( \chi^2_{n-2} \).

- Why is \( SSE_R - SSE_F \) independent of \( SSE_F \)?

\[
SSE_R - SSE_F = \| Y - H_RY \|^2 - \| Y - H_FY \|^2
\]

\[
= \| H_RY - H_FY \|^2 \quad \text{(Pythagoras)}
\]

\[
= \| H_R\epsilon - H_F\epsilon \|^2 \quad \text{(under } H_0)\]

\((H_R - H_F)\epsilon\) is in \( L_F \), the subspace of the full model while \( e_F = (I - H_F)\epsilon\) is in \( L_F^\perp \), the orthogonal complement of the full model – therefore \( e_F \) is independent of \((H_R - H_F)\epsilon\).
Overall goodness of fit

- Testing

\[ H_0 : \beta_1 = \beta_2 = 0. \]

- Two models:
  - (F) \( Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \)
  - (R) \( Y_i = \beta_0 + \varepsilon_i \)

- \( F \)-statistic, under \( H_0 \):

\[
F = \frac{ (SSE_R - SSE_F)/2 }{ SSE_F/(n - 3) } = \frac{ \| (H_R - H_F)Y \|^2 / 2 }{ \| (I - H_F)Y \|^2 / (n - 3) } \sim F_{2,n-3}.
\]

- Reject \( H_0 \) if \( F > F_{1-\alpha,2,n-3} \).

- Details: same as before.
Dropping subsets

- Suppose we have the model

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \]

and we want to test whether we can simplify the model by dropping variables, i.e. testing

\[ H_0 : \beta_{j1} = \ldots = \beta_{jk} = 0. \]

- Two models:
  - (F) – above
  - (R) – model with columns \( X_{j1}, \ldots, X_{jk} \) omitted from the design matrix.

- Under \( H_0 \)

\[ F = \frac{(SSE_R - SSE_F)/(df_R - df_F)}{SSE_F/df_F} \sim F_{df_R - df_F, df_F} \]

where \( df_F \) and \( df_R \) are the “residual” degrees of freedom of the two models.
In previous slide: we had to fit two models, and we might want to test more than just whether some coefficients are zero.

Suppose we want to test

\[ H_0 : C_{k \times p} \beta_{p \times 1} = h_{k \times 1} \]

Specifying the reduced model can be difficult.

Under \( H_0 \)

\[ C\hat{\beta} - h \sim N \left( 0, \sigma^2 C(X^t X)^{-1} C^t \right). \]

As long as \( C(X^t X)^{-1} C^t \) is invertible

\[ (C\hat{\beta} - h)^t (C(X^t X)^{-1} C^t)^{-1} (C\hat{\beta} - h) = SSE_R - SSE_F \sim \sigma^2 \chi_k^2. \]

\( F \)-statistic

\[ F = \frac{(SSE_F - SSE_R)/(df_R - df_F)}{SSE_F / df_F} \sim F_{df_R - df_F, df_F}. \]
Another fact about multivariate normal

- Suppose that \( Z_{k \times 1} \sim N(0, \Sigma_{k \times k}) \) where \( \Sigma \) is invertible. Then
  \[
  Z^t \Sigma^{-1} Z \sim \chi_k^2.
  \]

- Why? Let \( \Sigma^{-1/2} \) be a square root of \( \Sigma^{-1} \), i.e. \( \Sigma^{-1/2} \) is a symmetric matrix such that
  \[
  \Sigma^{-1/2} \Sigma^{-1/2} = I_{k \times k}
  \]
  \[
  \Sigma^{-1/2} \Sigma^{-1/2} = \Sigma^{-1}.
  \]

- Then,
  \[
  \Sigma^{-1/2} Z \sim N(0, I_{k \times k})
  \]
  and
  \[
  Z^t \Sigma^{-1} Z = \|\Sigma^{-1/2} Z\|^2 \sim \chi_k^2.
  \]
Polynomial models

- So far, we have considered models that are linear in the $x$'s.
- We could have regression model be linear in known functions of $x$: example polynomials.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_k X_i^k + \varepsilon_i.$$ 

- Here is an example.
Polynomial models

- Caution should be used in degree of polynomial used: it is easy to overfit the model.
- Useful when there is reason to believe relation is nonlinear.
- Easy to add polynomials in more than two variables to the regression: *interactions*.
- Although polynomials can approximate any continuous function (Bernstein’s polynomials) there are sometimes better bases. For instance, regression model may not be polynomial, but only “piecewise” polynomial.
- Design matrix $X$ can become ill-conditioned which can cause numerical problems.
Spline models

- Spline models are piecewise polynomials functions, i.e. on an interval between “knots” \((t_i, t_{i+1})\) the spline \(f(x)\) is polynomial but the coefficients change within each interval.

- Example: cubic spline with knows at \(t_1 < t_2 < \cdots < t_h\)

\[
f(x) = \sum_{j=0}^{3} \beta_{0j} x^j + \sum_{i=1}^{h} \beta_i (x - t_i)^3_+
\]

where

\[
(x - t_i)_+ = \begin{cases} 
    x - t_i & \text{if } x - t_i \geq 0 \\
    0 & \text{otherwise.}
\end{cases}
\]

- Here is an example.

- Conditioning problem again: \(B\)-splines are used to keep the model subspace the same but have the design less ill-conditioned.

- Other bases one might use:
  - Fourier: \(\sin\) and \(\cos\) waves.
  - Wavelet: space/time localized basis for functions.