LTCC Advanced Course: Introduction to Semiparametric Modelling Lecture 1

Clifford Lam

Department of Statistics
London School of Economics and Political Science

Introduction

The purpose of regression : To study the relationship among variables.

- Economists or Econometricians want to know what kind of factors are driving the economy, how effective.
- Biologists want to investigate certain gene activities in reaction to different drugs.
- Meteorologists want to predict if a tsunami or hurricane coming in the next week/month.

Results : Can find out the effect of various factors, and predict new outcomes.

Example : Management of a retirement fund

A company called BRI (fake name) sells a particular type of retirement plan to small firms. A prediction of the year-end dollar amount contributed to each plan can help make internal revenue and cost projections.

- Data on each firm are available on several attributes from last year.
- BRI also wants to know if a specifically trained sales representative is effective in increasing contributions to the retirement plans.

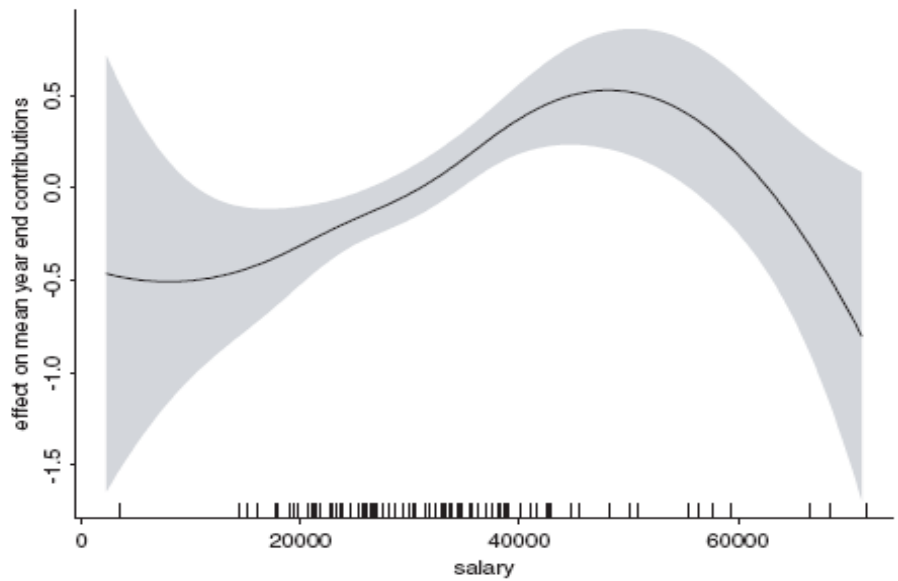

Figure: An estimated log-year-end contribution against mean salary of a firm.

Example : Term structure of interest rate

- Corporations raise money by issuing bonds, which is a contract requiring that entity to pay to the bond holder principal plus interest when the bond expires, called **maturity**.
- Interest rates on bonds depend upon their maturities. For zero-coupon bonds, simple arguments (you do not need to know the details) shows that

$$P(0) = P(T)\exp\left(-\int_0^T r(x)\,dx\right),$$

where $P(0)$ is the current price of the bond, $P(T)$ is the par value (T is the time to maturity) and $r(t)$ is the interest rate with time t to maturity for the bond.

- Letting $y_i = 100P(0)/P(T_i)$, $i = 1, \ldots, n$, where n is the number of bonds in the data (US STRIPS, time to maturity calculated at Dec 31, 1995), we have

$$\log(y_i) = \log(100) - \int_0^{T_i} r(x)\,dx.$$

In particular,

$$r(T_i) = -\frac{d\log(y_i)}{dT_i}.$$

- Hence the term structure of interest rate can be studied by looking at the slope of the $\log(y)$ vs T graph.

Example : Term structure of interest rate

- How can we estimate the interest rate from the following graphs? Is the interest rate constant?
- Local linear regression can give estimates of slope at different points directly.

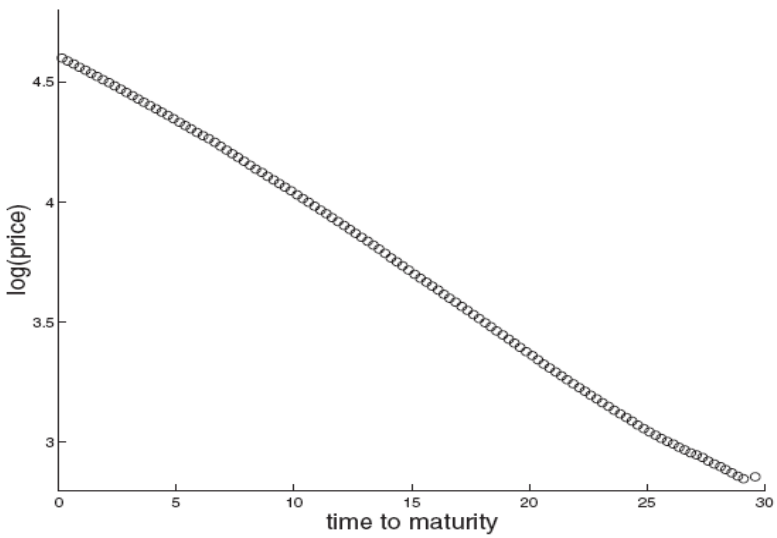

Figure: Log price vs time to maturity.

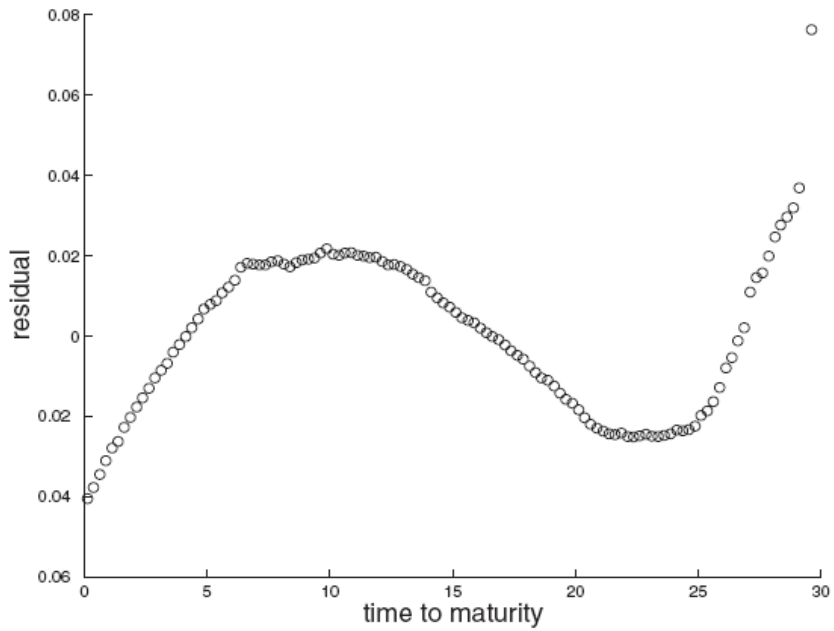

Figure: Residuals after a linear fit.

Example : Women weights and heights

Open R and do the following:

```
attach(women)
plot(height, weight)
reg = lm(weight ~ height)
abline(reg)
```

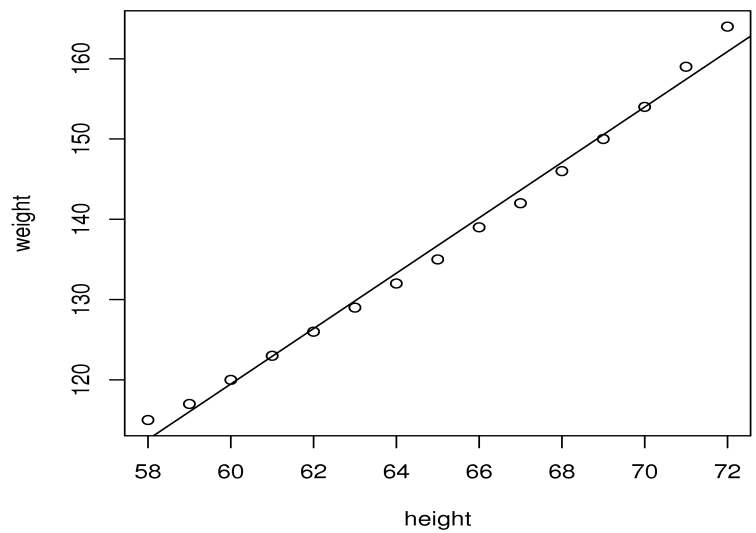

Figure: Weight vs height for the American women heights and weights data, and the fitted linear regression line.

Example : Women weights and heights

Clearly a linear fit is not enough, although the strength of the linear relationship is strong (check using `summary(reg)`). How about adding a quadratic term for height? The model is then

$$E(\text{weight}) = \beta_0 + \beta_1 * \text{height} + \beta_2 * \text{height}^2.$$

```
reg2 = lm(weight~height + I(height^2))
reg2$coeff
fun1 = function(x) 261.87818 - 7.34832*x + 0.08306*x^2
x = seq(52,73,0.01)
points(x, fun1(x), type="l", col="blue")
```

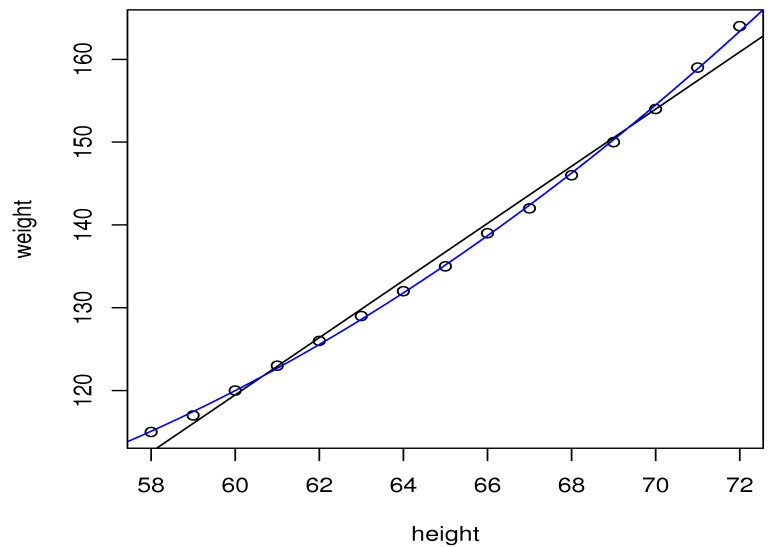

Figure: Same as previous graph, but with fitted quadratic curve.

Polynomial regression

- The quadratic regression considered for the women weights and heights data is an example of **polynomial regression**. Since the model is linear in the parameters (it is only quadratic in height, but not the β_j's), it is still a linear model.
- The general p-th degree regression model for a scatterplot (x_i, y_i), $1 \leq i \leq n$, is

$$y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_p x_i^p + \epsilon_i, \quad E(\epsilon_i) = 0.$$

- For inference purpose, we usually assume $\epsilon_i \sim N(0, \sigma_\epsilon^2)$, so that the model is **homoscedastic**, i.e. the error variance is constant throughout, irrespective of x. This is true for the women weights and heights data when a cubic curve is fitted (and the cubic term is significant as well (sensible?); use `summary(reg3)` to see).

```
reg3 = update(reg2, .~.+I(height^3)); plot(height, reg3$resid)
```

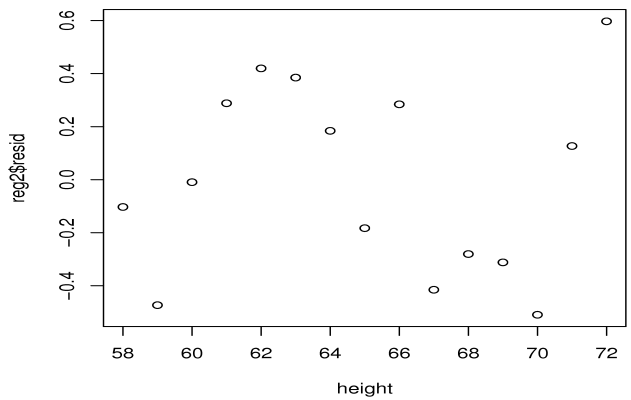

Figure: Residuals against height after a quadratic curve is fitted.

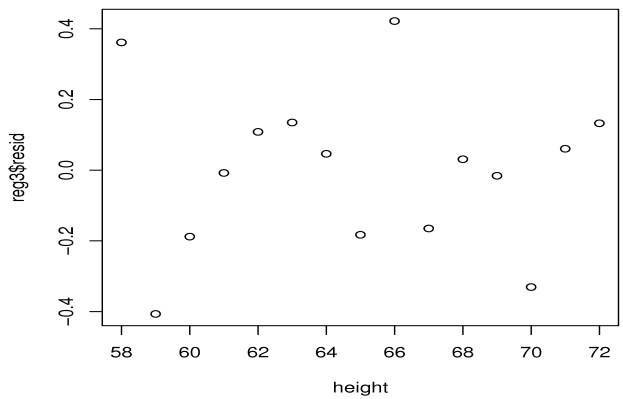

Figure: Residuals against height after a cubic curve is fitted.

LIDAR data - where polynomial regression fails

- The technique LIDAR (light detection and ranging) uses the reflection of laser-emitted light to detect chemical compounds in the atmosphere. The LIDAR data here is for detection of mercury.
- The response variable is `logratio`, the logarithm of the ratio of received light from two laser sources. One source had the resonance frequency of mercury, and the other one had not. The predictor variable is `range`, the distance traveled before the light is reflected back to its source.
- Scientists are interested in the estimated curve and its derivative.

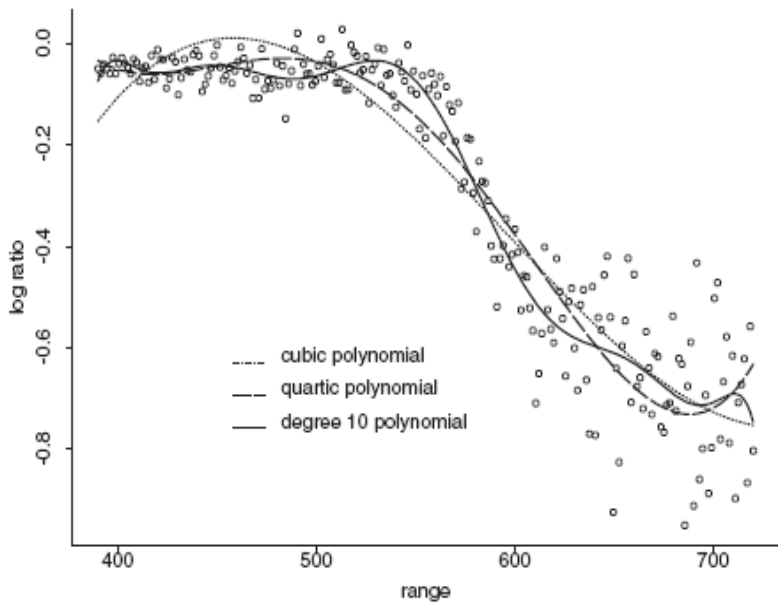

Figure: Higher degree polynomial fits to the LIDAR data.

- Lower degree polynomial do not go through the data cloud well.
- Higher degree polynomial goes through well, but with lots of wiggles which are characteristics of high degree polynomial. Very problematic when the derivative of the curve is also of interest.
- Another feature is evidence of **heteroscedasticity** - non-constant variance across `range`.

Scatterplot smoothing for the LIDAR data

- Consider two extremes - **interpolation** of data, and linear regression. Interpolation treated every data point as a **knot**. For any two adjacent knots, the best straight line through the data points in between is to be fitted - straight line exactly passing through the two data points as adjacent knots.
- Linear regression has two extremities of the data cloud as knots, and the best straight line is fitted for the data points in between the two knots.
- There are usually better fit to the data between these two extremes - **linear spline regression**, assuming, for $\kappa_1, \ldots, \kappa_K$ being K knots,

$$E(y|x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} b_k (x - \kappa_k)_+.$$

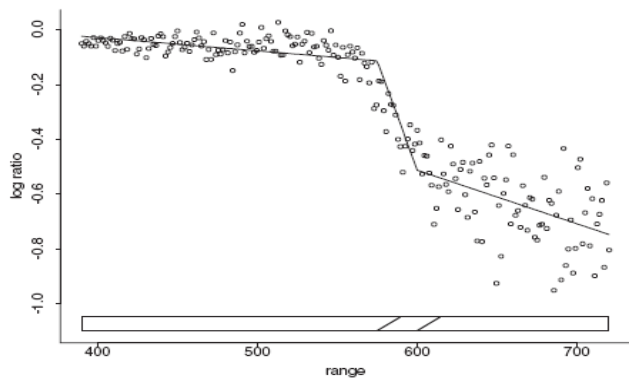

Figure: Linear spline fit to LIDAR data with two knots.

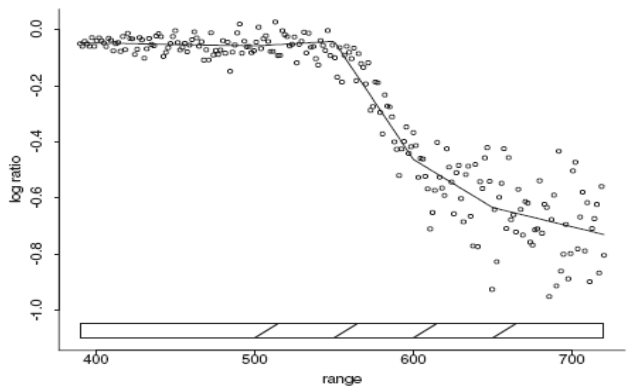

Figure: Linear spline fit to LIDAR data with four knots.

Linear model theories - OLS estimator

The multiple regression model we assume for the data is

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

where $\epsilon_i \sim WN(0, \sigma^2)$ (white noise with mean 0 and variance σ^2). In matrix form, the model is written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^{\mathrm{T}}$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}}$,

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}.$$

The ordinary least square (OLS) estimator minimizes

$$\sum_{i=1}^{n} (y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta})^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2.$$

The derivative of the above with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial}{\partial \boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = -2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}.$$

Hence $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$, provided $\mathbf{X}$ is of full rank (so that $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ exists).

Equivalence of MLE and OLS estimator for β

For maximum likelihood estimator (MLE), we usually further assume that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$. Then the log-likelihood function is

$$\ell(\boldsymbol{\beta}, \sigma^2) = \text{const.} - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2.$$

Differentiating with respect to $\boldsymbol{\beta}$, it is easy to see that it is equivalent to doing a least square estimation, and hence MLE is equivalent to OLS estimator under the assumption $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$.
To estimate σ^2, differentiating the log-likelihood with respect to σ^2 and set the derivative to zero, we get

$$\hat{\sigma}^2 = \frac{1}{n} \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2.$$

This is a **biased** estimator for σ^2. Denote $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} := \mathbf{H}\mathbf{y}$, which is the fitted value of $\mathbf{y}$. Here $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ is called the hat or projection matrix, projecting $\mathbf{y}$ onto the column space of $\mathbf{X}$. Then

$$\hat{\sigma}^2 = \frac{1}{n} \|\mathbf{y} - \mathbf{H}\mathbf{y}\|_2^2 = \frac{1}{n} \mathbf{y}^{\mathrm{T}} (\mathbf{I} - \mathbf{H})^{\mathrm{T}} (\mathbf{I} - \mathbf{H}) \mathbf{y} = \frac{1}{n} \mathbf{y}^{\mathrm{T}} (\mathbf{I} - \mathbf{H}) \mathbf{y},$$

since $\mathbf{H}$ and $\mathbf{I} - \mathbf{H}$ are idempotent and symmetric.

Sum of squares of a linear model

- When a model contains the intercept term β_0, then the design matrix $\mathbf{X}$ has its first column being all ones. Most of the time we are only interested in the effects of the other variables rather than the constant term. One measure is to calculate the "total variation" explained by those variables. To do this, the total variation of the data, after **the effect of the constant term is taken out**, is defined by

$$\text{Total corrected SS} = \sum (y_i - \bar{y})^2.$$

- The variation explained by all the variables (other than the constant term) is the **sum of squares due to regression**, defined as

$$\text{SS(reg)} = \sum (\hat{y}_i - \bar{y})^2,$$

  where $\hat{y}_i$ is the i-th element in $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

- The residual sum of square is defined as

$$\text{RSS} = \sum (y_i - \hat{y}_i)^2.$$

  In exercise 1 you will show that

$$\text{Total corrected SS} = \text{SS(reg)} + \text{RSS}.$$

Hypothesis testing and confidence interval for a regression parameter

Suppose we want to test

$$H_0 : \beta_2 = 0 \longleftrightarrow H_1 : \beta_2 \neq 0.$$

We need to assume $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ for this purpose. Then $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Since we have the OLS (or MLE, they are equivalent under normality assumption) estimator being $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ which is **linear** in $\mathbf{y}$, $\hat{\boldsymbol{\beta}}$ must be normally distributed. Can show easily then

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}).$$

In exercise 1 you will show that $\hat{\boldsymbol{\beta}}$ is independent of S^2, the unbiased estimator of σ^2, and

$$S^2 = \frac{\text{RSS}}{n-p} \sim \sigma^2\chi^2_{n-p},$$

where $p = k+1$, the total number of regression parameters. Hence we can use

$$T = \frac{\hat{\beta}_2 - \beta_2}{\sigma\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{22}}} \Big/ \sqrt{\frac{(n-p)S^2}{\sigma^2(n-p)}} = \frac{\hat{\beta}_2 - \beta_2}{S\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{22}}} \sim t_{n-p}$$

for construction of confidence interval or hypothesis testing of β_2.

Hypothesis testing for more regression parameters

Suppose now we want to test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \longleftrightarrow H_1 : \text{Not all } \beta_i \text{ are 0 for } i \geq 1.$$

This is the usual hypothesis to see if the regression as a whole is significant. The t-test cannot be used. In exercise 1 you will show that under the null hypothesis above,

$$\frac{\mathsf{SS(reg)}}{\sigma^2} \sim \chi^2_{p-1}.$$

You will also show that SS(reg) is independent of RSS. Hence under the null hypothesis,

$$F = \frac{\mathsf{SS(reg)}/(p-1)}{\mathsf{RSS}/(n-p)} \sim F_{p-1,n-p}.$$

When only one parameter is being tested, e.g. $H_0 : \beta_2 = 0$, we can construct a t-test like before. But we can also construct an F-test, with numerator under H_0 to be

$$\mathsf{RSS}_{\mathsf{smaller}} - \mathsf{RSS}_{\mathsf{larger}} = \frac{\hat{\beta}_2^2}{[(\mathbf{X}^T\mathbf{X})^{-1}]_{22}} \sim \sigma^2 \chi^2_1.$$

Fortunately, both tests will give exactly the same result, as $t^2_{n-p} \equiv F_{1,n-p}$.
In general, when testing a smaller model (with p_1 parameters) under the null hypothesis against a larger model (with $p_2 > p_1$ parameters),

$$F = \frac{(\mathsf{RSS}_{\mathsf{smaller}} - \mathsf{RSS}_{\mathsf{larger}})/(p_2 - p_1)}{\mathsf{RSS}/(n-p)} \sim F_{p_2-p_1,n-p}.$$

Prediction intervals for new observation

A new observation (at $\mathbf{x}$) comes in as

$$y = \mathbf{x}^T \boldsymbol{\beta} + \epsilon,$$

and we predict with $\hat{y} = \mathbf{x}^T \hat{\boldsymbol{\beta}}$. Variability comes from ϵ and $\hat{y}$.

- $\text{var}(\hat{y}) = \mathbf{x}^T \text{var}(\hat{\boldsymbol{\beta}})\mathbf{x} = \sigma^2 \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}$.
- Hence $\hat{y} - y = \mathbf{x}^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \epsilon \sim N(0, \sigma^2 + \sigma^2 \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x})$.
- A $(1-\delta)$ prediction interval for a new observation at $\mathbf{x}$ is

$$\hat{y} \pm t_{n-p,\delta/2} S \sqrt{1 + \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}}.$$

Pointwise variability bands

At each $\mathbf{x}$, the fitted value for the response variable is

$$\hat{y} = \mathbf{x}^T \hat{\beta}.$$

Under normality assumption, since $\hat{\boldsymbol{\beta}} \sim N(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$,

$$\hat{y} \sim N(\mathbf{x}^T\boldsymbol{\beta}, \sigma^2\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}).$$

This can be used to construct a 95% confidence interval for $\mathbf{x}^T\boldsymbol{\beta}$ if σ is known:

$$[\hat{y} \pm z_{0.025}\sigma\sqrt{\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}}].$$

If σ is unknown, we construct the t-statistic with s substituting σ, and so a 95% confidence interval for $\mathbf{x}^T\boldsymbol{\beta}$ is

$$[\hat{y} \pm t_{n-p;0.025}s\sqrt{\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}}].$$

When at each $\mathbf{x}$ we concerned a confidence interval is constructed for $\mathbf{x}^T\boldsymbol{\beta}$, a **pointwise variability band** is resulted. In practice, we just use the factor 2 in place of $z_{0.025}$ or $t_{n-p;0.025}$ for an approximate confidence band.

Family-wise variability bands

- The variability band introduced before is **pointwise**. Consider $\mathbf{x}_1^T\boldsymbol{\beta}$ having a 95% confidence interval I_1, and $\mathbf{x}_2^T\boldsymbol{\beta}$ having a 95% confidence interval I_2. We can say **separately** that $\mathbf{x}_1^T\boldsymbol{\beta}$ lies in I_1 with 95% confidence, and $\mathbf{x}_2^T\boldsymbol{\beta}$ lies in I_2 with 95% confidence.

  However, we **cannot** say that

  $$\mathbf{x}_1^T\boldsymbol{\beta} \text{ lies in } I_1 \text{ and } \mathbf{x}_2^T\boldsymbol{\beta} \text{ lies in } I_2 \text{ with 95\% confidence.}$$

- Mathematically, if $[L(\mathbf{x}), U(\mathbf{x})]$ is a pointwise 95% variability band for $\mathbf{x} \in \mathbb{X}$, then

  $$P(L(\mathbf{x}) \leq \mathbf{x}^T\boldsymbol{\beta} \leq U(\mathbf{x})) \geq 95\% \text{ for all } \mathbf{x} \in \mathbb{X}.$$

- In later lectures, we will see how to construct a **family-wise variability band**, where we can say that

  $$P(L(\mathbf{x}) \leq \mathbf{x}^T\boldsymbol{\beta} \leq U(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathbb{X}) \geq 95\%.$$