Simple Linear Regression

What Is Regression?

- Fitting a statistical model to data, in particular linear models.
- Understand the errors in the data and how they affect the estimated model.

Why Do Regression?

- Describe the relationship between two variables.
- Predict the value of one variable given the value of another.
- Learn how to control future values of a variable by controlling values of variables it's related to.

Regression Model

$$Y=\beta_0+\beta_1 X+\varepsilon$$

For particular points, $Y_i=\beta_0+\beta_1 X_i+\varepsilon_i, i=1,\ldots,n$.

We assume *X* is not random; if it is, condition on it.

If we have an infinite number of observations *Y*'s for each value *X*, the mean of the *Y*'s is $\bar{Y}=E[\bar{X}]$; if *X* is random, then $E[Y|X=x]$.

Fitted Value

$E[Y]=\beta_0+\beta_1 X$.

For each X_i , the fitted value $\hat{Y}_i=b_0+b_1 X_i$. Deviation from the line $Y_i-\hat{Y}_i=e_i$ is the residual.

Gauss-Markov Conditions

1. $E(\varepsilon_i)=0\Rightarrow E(Y_i)=\beta_0+\beta_1 X_i$.
2. $var(\varepsilon_i)=\sigma^2$ ($E(\varepsilon_i^2)=\sigma^2$).
3. ε_i 's are uncorrelated ($E(\varepsilon_i\varepsilon_j)=0$).

Or equivalently,

1. $E(Y_i)=\beta_0+\beta_1 X_i$.
2. $var(Y_i)=\sigma^2$.
3. Y_i 's are uncorrelated.

Method of Least-Squares

Define $S=\sum_{i=1}^{n}(y_i-b_0-b_1 x_i)^2$ to be the sum of square residuals (vertical deviations from the line).

Since β_0 , β_1 , σ^2 are parameters (constants we usually don't know), they are estimated by:

- $b_0=\bar{y}-b_1\bar{x}$,
- $b_1=\dfrac{\sum_{i=1}^{n}x_i y_i-n\bar{x}\bar{y}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}=\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$,
- $s^2=\sum_{i=1}^{n}\dfrac{e_i^2}{n-2}$.

Properties of Least Square Estimators

1. Unbiased: $E(b_0)=\beta_0$ and $E(b_1)=\beta_1$.
2. Linear in Y_i 's: can find constants c_i , d_i such that $b_0=\sum c_i Y_i$ and $b_1=\sum d_i Y_i$.

Gauss-Markov Theorem

Least squares estimate are the best (minimum variance) linear unbiased estimators (BLUE).

Properties of Fitted Regression Line

1. $\sum_{i=1}^{n} e_i=0$, where $e_i=Y_i-\hat{Y}_i$ are the residuals.
2. $\sum_{i=1}^{n} e_i^2$ is a minimum.
3. $\sum_{i=1}^{n} e_i X_i=0$.
4. $\sum_{i=1}^{n} \hat{Y}_i=n\bar{Y}$, that is the average of the predicted values ($\hat{Y}_i$) is the same as the average of the data values of *Y*.
5. $\sum_{i=1}^{n} e_i \hat{Y}_i=0$.

Note: $\sum_{i=1}^{n} e_i Y_i\neq 0$.

Correlation

Correlation gives rv's *X* and *Y* a symmetric measure of the strength of their linear relationship.

Pearson Product-Moment Correlation

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2-\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

This is which is the MLE of $\rho=\frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}$ when *X* and *Y* have a bivariate normal distribution.

Note: $b_1=r\frac{s_y}{s_x}$, where s_x and s_y are the standard deviations for *X* and *Y*.

Regression Analysis of Variance

Decomposition of Sum of Squares

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2=b_1^2\sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{n}e_i^2$$

- $\sum_{i=1}^{n}(Y_i-\bar{Y})^2$ = Total Sum of Squares (SSTO).
- $b_1^2\sum_{i=1}^{n}(X_i-\bar{X})^2$ = Regression Sum of Squares (SSR) – amount of variation in the *Y*'s that is explained by the least-square line.
- $\sum_{i=1}^{n}e_i^2$ = Error Sum of Squares (SSE).

ANOVA Table

Source	SS	df	MS
Regression Line	SSR = $b_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$	1	$\frac{\text{SSR}}{1} = \text{MSR}$
Residuals	SSE = $\sum_{i=1}^n e_i^2$	$n-2$	$\frac{\text{SSR}}{n-2} = \text{MSE} = s^2$
Total	SSTO = $\sum_{i=1}^n (Y_i - \bar{Y})^2$	$n-1$	

Coefficient of Variation

The coefficient of variation is $R^2 = \frac{\text{SSR}}{\text{SSTO}} = \frac{b^2 \sum (X_i - \bar{X})^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}, 0 \le R^2 \le 1$. It gives the percentage of variation in the *Y*'s explained by the regression line.

Note: $R^2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$ which is the square of sample correlation coefficient between *X* and *Y*.

Statistical Properties of b_0 and b_1

b_0 and b_1 are rv's because they are a function of the *Y*'s. So they have expected values and variances.

1. $E(b_0) = \beta_0$, $\text{var}(b_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right)$.
2. $E(b_1) = \beta_1$, $\text{var}(b_1) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}$.

Hypothesis Testing

Steps

1. Specify the null (H_0) and the alternative (H_A) hypothesis for parameter of interest.
2. Calculate a test statistic whose distribution is known.
3. Estimate a p-value. Assuming H_0 is correct, calculate the probability of the value of the test statistic for values more extreme (belongs to H_A). P-value gives evidence against H_0 .
4. If p-value is small, then either H_0 is correct and data are just rare, or H_0 is incorrect. The smaller the p-value, the stronger the evidence against H_0 .
   $p > 0.1$: no evidence against H_0 .
   $0.05 < p < 0.1$: some weak evidence against H_0 .
   $0.01 < p < 0.05$: moderate evidence against H_0 .
   $p < 0.01$: strong evidence against H_0 .

t-Test

If $X_{1,} \ldots, X_n$ are iid $N(\mu, \sigma^2)$ rv's, then:

1. $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

2. $\bar{X}$ is an estimator of μ , $s^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$ is an estimator of σ^2 .

3. If we standardize $\bar{X}$, we get $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$. If we estimate σ by *s*, then $\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t_n - 1$.

Hypothesis Test for β_1

$b_1 \sim N\left(\beta_{1,} \frac{\sigma^2}{\sum(X_i - \bar{X})^2}\right)$ because it is a linear combination of normal rv's.

Estimate $var(b_1)$ by $\frac{s^2}{\sum(X_i - \bar{X})^2}$ which has $n-2$ df. Thus, $\frac{b_1 - \beta_1}{\sqrt{\frac{s^2}{\sum(X_i - \bar{X})^2}}} \sim t_{n-2}$.

To test $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$, use test statistic $t_{obs} = \frac{b_1}{\sqrt{s^2 / \sum(X_i - \bar{X})^2}}$ which is from t_{n-2} if H_0 is true.

Confidence Intervals

Let $X_{1,} \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$. We want a confidence interval for μ .

Since $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$, we find $t_{n-1,\alpha/2}$ such that $P\left(\left|\frac{\bar{X}-\mu}{S/\sqrt{n}}\right| \leq t_{n-1,\alpha/2}\right) = 1-\alpha$. Then the interval $\bar{X} \pm t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}$ is a $100(1-\alpha)$ % C.I. for μ .

Not true: The probability that μ is in this interval is $(1-\alpha)$.

True: C.I. calculated from repeated sample from a $N(\mu, \sigma^2)$ of size n by this method will include μ $100(1-\alpha)$ % of the time.

Note: Quote C.I. to give an idea of the precision of an estimate.

Prediction

Least square line: $\hat{Y} = b_0 + b_1 X$.

For X_h (a given value of X), a $100(1-\alpha)$ % C.I. for the corresponding point on the line (estimate of $E(Y_h)$) is $\hat{Y}_h \pm t_{n-2,\alpha/2} s \sqrt{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}}}$.

Note: If the line is a good fit, s is small and narrow C.I.

Prediction Interval

We use the line to predict a new Y_h given a X_h .

A $100(1-\alpha)$ % prediction interval for Y_h corresponding to X_h is $\hat{Y}_h \pm t_{n-2,\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{S_{XX}}}$

Note: This is not a C.I. since C.I. are for parameters.

Confidence Band: Workman-Hotelling Procedure

Want to put a C.I. on the entire line, i.e. a confidence band around the line.

Replace t-critical values $t_{n-2,\alpha/2}$ in C.I. by $\sqrt{2F_{2,n-2,1-\alpha}}$, where $F_{2,n-2,1-\alpha}$ is the $100(1-\alpha)$ percentile from the *F* distribution with 2 and $n-2$ df.

Thus, the bounds at X_h is $\hat{Y}_h \pm W\,\text{var}(\hat{Y}_h) = \hat{Y}_h \pm W\sqrt{s^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)}$. To obtain the confidence band, join up all the C.I.'s.

The ANOVA F-Test

If $\beta_1 = 0$, then $E(\text{MSE}) = E(\text{MSR}) = \sigma^2$.

Test $H_0\colon \beta_1 = 0$ vs $H_A\colon \beta_1 \neq 0$, use test statistic $F_{obs} = \frac{\text{MSR}}{\text{MSE}}$. Since $\beta_1 \neq 0$ gives larger values of F_{obs} , deviations from $\beta_1 = 0$ are only in the right tail of *F* distribution.

Diagnostic and Remedial Measures

So Far

- Model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, n = 1, \ldots, n$.
- Assumption: Linear relationship is appropriate.
- LS line: $\hat{Y}_i = b_0 + b_1 X_i$.
- Gauss-Markov assumptions: $E(\epsilon_i) = 0$, $\text{var}(\epsilon_i) = \sigma^2$, ϵ_i and ϵ_j uncorrelated; can show b_0 and b_1 unbiased and BLUE.
- For testing, confidence intervals, assume ϵ_i normally distributed.

Residual Plots

$e_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1 X_i$ estimate of ϵ_i . What do we know about the e_i 's?

- $\sum e_i = 0$, $\sum e_i X_i = 0$, $\sum e_i \hat{Y}_i = 0$.
- e_i 's are random variables: $E(e_i) = 0$, $\text{var}(e_i) = \text{ugly}$, e_i 's not uncorrelated (since $\sum e_i = 0$).
- Distribution of the e_i 's is normal because e_i 's linear function of Y_i 's.

Plots of e_i 's are very useful for checking the assumptions of the model. Look for:

- Evidence that straight line is an appropriate model – no curvature, outlier, influential point.
- Constant variance.
- Normally distributed errors – otherwise p-values are wrong because test statistic doesn't have *t* or *F* distribution under $H_0\colon \beta_1 = 0$, and CI's have coverage other than $100(1-\alpha)$ %; especially important for prediction intervals.
- Lack of independence (if possible to check).
- Missing predictor variables (if possible to check).

Recommended plots: e_i vs $\hat{Y}_i$, e_i vs X_i , normal quantile-quantile plot.

Plot 1: e_i vs $\hat{Y}_i$

Look for:

1. Evidence of not constant variance.
2. Outliers – more exaggerated on residual plots than scatter plots.
3. Evidence of curvature – more exaggerated on residual plots than scatter plots.
4. A “good” plot has just random scatter about 0.

Plot 2: e_i vs X_i

- When model has only one predictor variable (simple linear regression), plots 1 and 2 are essentially the same because $\hat{Y}_i = b_0 + b_1 X_i$.
- As a general rule, curvature is assessed from this plot, where plot 1 is used for outliers and constant variance.

Plot 3: Normal Quantile-Quantile Plot of Residuals

- Used for assessing normality.
- Check after the other two.

Other Plots

- Univariate analysis of residuals (ex: box plot, stem-i-leaf plot): look for departures from normality.
- $|e_i|$ vs $\hat{Y}_i$: exaggerated non-constant variance.
- Residuals vs time or other sequence of observations.: a pattern indicates residuals may be correlated.
- Residual vs other predictor variables: want random scatter, otherwise this predictor variable should be considered in model.

Normal Quantile Plots

- To assess assumption of normality of errors.
- More sensitive to departures from normality than a histogram.

Idea of Construction

1. Order the e_i 's from smallest to largest: $e_{(1)} \le e_{(2)} \le \cdots \le e_{(n)}$.
2. Get values of corresponding quantile from a stander normal distribution: $E\left(Z_{(1)}\right), \ldots, E\left(Z_{(n)}\right)$ with $E\left(Z_{(i)}\right) = \Phi^{-1}\left(\frac{i}{n}\right)$.
3. Plot e_i vs $E\left(Z_{(i)}\right)$. If the normal assumption for ε_i is valid, the plot should be approximately a straight line.

Problem

$E\left(Z_{(n)}\right) = \Phi^{-1}\left(\frac{n}{n}\right) = \Phi^{-1}(1) = \infty \cdot$

One possible adjustment is $E\left(Z_{(i)}\right) = \Phi^{-1}\left(\frac{i}{n+1}\right)$.

Another possible adjustment (Blom) is $E\left(Z_{(i)}\right) = \Phi^{-1}\left(\frac{i-3/8}{n+3/4}\right)$.

Is Lack of Normality a Problem?

- Yes for prediction intervals.
- Less so for other tests and CI's.
- If tails are a little heavy but the distribution is of residuals is still symmetric, the *F* and *t* tests are still approximately correct.

Remedial Measures

What to do if assumptions are violated?

1. Abandon SLR for something else (usually more complicated).
2. For outliers, remove if shown to be mistakes, otherwise proceed with caution.
3. For influential points, fit regression on smaller range of values and report this.

4. Transform the data for non-linearity, non-normal residuals, non-constant variance.
   - If relationship is non-linear but variability is approximately constant over X, try to find a transformation of X that makes relationship appear linear.
   - If relationship is non-linear and variability is non-constant, transform Y. Another possible solution is to use weighted least squares.
   - Sometimes, transformation of both variables is needed.

Transformations to Stabilize the Variance

If Y has a distribution with mean μ_Y and variance σ_Y^2, then the mean and variance of $Z=f(Y)$ are approximately $E(Z)\approx f(\mu_Y)$ and $\operatorname{var}(Z)\approx\sigma_Y^2[f'(\mu_Y)]^2$.

Now suppose Y has mean μ and variance proportional to a function of μ, i.e. $\operatorname{var}(Y)=\sigma_2 V(\mu)=\sigma_Y^2$ (in regression $E(Y)=\beta_0+\beta_1 X$). We need a transformation $Z=f(Y)$ so that $\operatorname{var}(Z)$ is approximately constant. We need $[f'(\mu)]^2=c\dfrac{1}{V(\mu)}\Rightarrow f'(\mu)\propto V^{-1/2}(\mu)$, so take $f(\mu)\propto\int V^{-1/2}(\mu)d\mu$.

Some popular transformations:

- If $\operatorname{var}(Y_i)$ is linear in $E(Y_i)$, then take $Z_i=\sqrt{Y_i}$.
- If $\operatorname{var}(Y_i)$ increases faster than linear in $E(Y_i)$, then take $Z_i=\ln Y_i$.
- $Z_i=\dfrac{1}{Y_i}$ is even more extreme.

Semi-Studentized Residual Plots

A semi-studentized residual is $e_i^*=\dfrac{e_i-\bar{e}}{\sqrt{\text{MSE}}}$. It is not “studentized” because MSE is an estimate of $\text{SD}(\varepsilon_i)$, not $\text{SD}(e_i)$. Plot these residuals on standard scale; helpful for evaluating outliers (more than 3 s.d. from 0).

Simultaneous Inference and Other Topics

Joint Estimate of *Β*0 and *Β*1

We want joint $100(1-\alpha)$ % CI for β_0 and β_1. The probability that both CI capture their true value is less than $100(1-\alpha)$ %. Want to be $100(1-\alpha)$ % confident that both CI's capture what they are supposed to capture.

A solution is to use Bonferroni corrected CI's. In general, for joint/simultaneous CI for any k parameters, the overall confidence level will be at least $100(1-\alpha)$ % if the confidence level for each parameter is $100\left(1-\dfrac{\alpha}{k}\right)$ %

Simultaneous Estimate of Mean Response

We want joint CI for $E(Y)$ at more than one X.
There are two choices: Bonferroni or Working-Hotelling.

- Both of these procedures are conservative, i.e. the overall confidence level will actually be higher than $100(1-\alpha)$ %.
- For a small number of X's, Bonferroni is generally less conservative. For more, Working-Hotelling is less conservative.

Note

The Bonferroni procedure is also helpful for hypothesis testing when carrying out multiple tests.
For example, if you want to declare the result statistically significant if p-values are less then α , then you will mistakenly reject H_0 100α % of the time (type 1 error). So if $\alpha=0.05$ and doing 20 tests, on average one test will be reject even if H_0 is true all the time.

The Bonferroni Correction: When doing k tests, reject H_0 only if $\text{p-value}\leq\frac{\alpha}{k}$. The probability of one or more type 1 error is then at most α . Note that this is not useful if k is large.

Inverse Prediction (Calibration Problem)

Suppose we observe Y_0 . We want to estimate X_0 .

Inverse Estimator

The estimate $\hat{X}_0=\frac{Y_0-b_0}{b_1}$ is the inverse estimator.

Inverse Prediction Interval

How good is the inverse estimator $\hat{X}_0$?

The PI formula is $\hat{Y}_h\pm t_{n-2,1-\frac{\alpha}{2}}s\sqrt{1+\frac{1}{n}+\frac{(X_h-\bar{X})^2}{S_{xx}}}$. Solving for X_h is a mess. If we approximate $\frac{t^2_{n-2,1-\alpha/2}s^2}{b_1^2 S_{xx}}$ by 0 when it is small (say < 0.05), then the approximate inverse prediction interval for X_0 when $Y=Y_0$ is

$$\hat{X}_0\pm\frac{t_{n-2,1-\frac{\alpha}{2}}}{|b_1|}s\sqrt{1+\frac{1}{n}+\frac{(\hat{X}_0-\bar{X})^2}{S_{xx}}}\ .$$

Note that $\frac{t^2_{n-2,1-\alpha/2}s^2}{b_1^2 S_{xx}}$ being small means the test of $H_0\colon\beta_1=0$ will be statistically significant.

What if Predictor X is Random?

- Can analyze the correlation between X and Y . There exists tests and CI's for ρ under the assumption that X and Y have bivariate normal distribution(SAS gives p-values for 2-sided test of $H_0\colon\rho=0$ with “proc corr”).
- If we want to do regression, there is a clear choice for the independent and dependent variables.
- If (X_i, Y_i) independent pairs from bivariate normal distribution, then the distribution of $Y_i|X_i$ is normal with mean $\alpha_0+\alpha_1 X_i$ where $\alpha_0=\mu_Y-\mu_X\rho_{XY}\frac{\sigma_Y}{\sigma_X}$ and $\alpha_1=\rho_{XY}\frac{\sigma_Y}{\sigma_X}$ (still a straight line model), and constant variance $\sigma_Y^2(1-\rho_{XY}^2)$. So regression techniques (including inference) up to now still hold.
- If (X_i, Y_i) are not bivariate normal, as long as the X_i 's are independent rv's whose probability distributions do not involve any of the regression parameters (β_0 , β_1 , σ^2) and the conditional distribution $Y_i|X_i$ is normal, then everything so far still hold, except need to talk about the distribution of $Y_i|X_i$ (ex: Gauss-Markov Condition: $E(Y_i|X_i)=\beta_0+\beta_1 X_i$, $\text{var}(Y_i|X_i)=\sigma^2$, distribution of $Y_i|X_i$ independent).

Measurement Error

Suppose (X_i, Y_i) bivariate normal, but the X_i 's are observed with measurement error. Note that measurement error in the

Y_i 's are absorbed in ε_i 's.

We want the model $Y_i + \beta_0 + \beta_1 X_i + \varepsilon_i$, but we observed $X_i^{'} = X_i + \delta_i$. Assume $\delta_i \sim N(0, \sigma^2)$, then we get $Y_i + \beta_0 + \beta_1 X_i^{'} + (\varepsilon_i - \beta_i \delta_i)$. If δ , ε , X are all independent and normally distributed, it is known that $(Y, X^{'})$ is bivariate normal with conditional mean $E(Y | X^{'}) = \beta_0^{'} + \beta_1^{'} X^{'}$, where $\beta_0^{'} = \beta_0$ and $\beta_1^{'} = \dfrac{\beta_1}{1 + \sigma_\delta^2 / \sigma_X^2}$.

- The estimated slope from regression of Y on $X^{'}$ is a biased estimate of β_1 .
- If the goal is to estimate β_1 without bias, then estimate $\dfrac{\sigma_\delta^2}{\sigma_X^2}$ from data to get an approximate unbiased estimate of β_1 .
- If goal is to estimate predictions for Y, presence of errors in X decreases accuracy of predictions. A common assumption is to assume σ_δ^2 small compared to σ_Y^2 and σ_X^2, then the bias of β_1 is small and reduction of accuracy in prediction is minimal. Then proceed as if there was no measurement error.

Simple Linear Regression in Matrix Terms

Background

Let $\vec{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$. Let $\vec{A}$ be a scalar, vector, or matrix.

Expectation: $E(\vec{X}) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{pmatrix}$, $E(\vec{X} + \vec{Y}) = E(\vec{X}) + E(\vec{Y})$, $E(\vec{A}\vec{X}) = \vec{A} E(\vec{X})$.

Variance or Variance-Covariance Matrix: $\text{var}(\vec{X}) = \text{cov}(\vec{X}) = E[(\vec{X} - E(\vec{X}))(\vec{X} - E(\vec{X}))^T]$ is a symmetric matrix where the *i*th diagonal element is $\text{var}(X_i)$ and the *ij*th element is $\text{cov}(X_i, X_j)$. Note that $\text{cov}(\vec{A}\vec{X}) = \vec{A}\,\text{cov}(\vec{X})\vec{A}^T$

Definitions

1. $\vec{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ is a random vector ($n \times 1$).
2. $\vec{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ is a random vector ($n \times 1$).
3. $\vec{\beta} = \begin{pmatrix} \beta_1 \\ \beta_1 \end{pmatrix}$ is not a random vector (2×1).
4. $\vec{X} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$ is a matrix, maybe random ($n \times 2$).

Then $\underbrace{\vec{Y}}_{n\times 1} = \underbrace{\vec{X}\vec{\beta}}_{n\times 1} + \underbrace{\vec{\varepsilon}}_{n\times 1} = \begin{pmatrix} \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ \vdots \\ \beta_0 + \beta_1 X_n + \varepsilon_n \end{pmatrix}$.

Gauss-Markov Assumptions

In terms of $\vec{\varepsilon}$:

- $E(\vec{\varepsilon})=\vec{0}$.
- $\operatorname{cov}(\vec{\varepsilon})=\sigma^2\vec{I}$.

In terms of $\vec{Y}$:

- $E(\vec{Y})=\vec{X}\vec{\beta}$.
- $\operatorname{cov}(\vec{Y})=\sigma^2\vec{I}$.

Additional assumption: ε has an n-dimensional normal distribution.

Least Square Estimate

Find β_0 and β_1 that minimizes the sum of squares of residuals:

$$\begin{aligned} S &= \sum (Y_i-\beta_0-\beta_1 X_i)^2 \\ &= (\vec{Y}-\vec{X}\beta)^T(\vec{Y}-\vec{X}\vec{\beta}) \\ &\vdots \\ &= \vec{Y}\vec{Y}^T-2\vec{\beta}^T\vec{X}^T\vec{Y}+\vec{\beta}^T\vec{X}^T\vec{X}\vec{\beta} \end{aligned}$$

Set $\frac{\partial S}{\partial \vec{\beta}}=0$, we get $-2\vec{X}^T\vec{Y}+2\vec{X}^T\vec{X}\vec{\beta}=0 \Rightarrow \vec{X}^T\vec{X}\vec{\beta}=\vec{X}^T\vec{Y} \Rightarrow \vec{\beta}=(\vec{X}^T\vec{X})^{-1}\vec{X}^T\vec{Y}$ assuming $(\vec{X}^T\vec{X})^{-1}$ exists.

Now, $\operatorname{rank}(\vec{X}^T X)=\operatorname{rank}(\vec{X})=2$ so $\vec{X}^T\vec{X}$ invertible, and

$$(\vec{X}^T\vec{X})^{-1}=\frac{1}{n\sum X_i^2-(\sum X_i)^2}\begin{bmatrix} \sum X_i^2 & -\sum X_i \\ -\sum X_i & n \end{bmatrix}=\begin{bmatrix} \frac{\sum X_i^2}{nS_{XX}} & \frac{-\bar{X}}{S_{XX}} \\ \frac{-\bar{X}}{S_{XX}} & \frac{1}{S_{XX}} \end{bmatrix}.$$

Therefore, $\vec{b}=(\vec{X}^T X)^{-1}\vec{X}^T Y=\begin{bmatrix} \frac{\sum X_i^2}{nS_{XX}} & \frac{-\bar{X}}{S_{XX}} \\ \frac{-\bar{X}}{S_{XX}} & \frac{1}{S_{XX}} \end{bmatrix}\begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}=\begin{bmatrix} \frac{\sum X_i^2\sum Y_i}{nS_{XX}}-\frac{\bar{X}\sum X_i Y_i}{S_{XX}} \\ \frac{-\bar{X}\sum X_i Y_i}{S_{XX}}+\frac{\sum X_i Y_i}{S_{XX}} \end{bmatrix}=\begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$.

Properties of Least Squares Estimators

Assume X constant. We know $b=(X^T X)^{-1}X^T Y=AY$. We have

- $E(b)=\beta$,
- $\operatorname{cov}(b)=\sigma^2\begin{bmatrix} \sum X_i^2 & \frac{-barX}{S_{XX}} \\ \frac{-barX}{S_{XX}} & \frac{1}{S_{XX}} \end{bmatrix}$.

Fitted Values and Residuals

$e=\begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$, $\hat{Y}=\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$.

Residuals: $e=Y-\hat{Y}=Y-Xb=(1-H)Y$.

Notice $\hat{Y}=Xb=X(X^T X)^T X^T Y=HY$, where $H=X(X^T X)^T X^T$ is the "hat" matrix. H is symmetric and idempotent. $\operatorname{cov}(e)=\sigma^2(I-H)$. Note the df for $e^T e=\sum e_i^2=\text{SSE}$ is $n-2$, which is $\operatorname{rank}(I-H)$.

Analysis of Variance, Decomposition

Decomposed SSTO: $\sum(Y_i-\bar{Y})^2=\sum Y_i^2-n\bar{Y}^2$. In matrix form, $\text{SSTO}=Y^T Y-\frac{1}{n}Y^T J Y$ where J is an $n\times n$ matrix of 1's.

Now, $Y^T Y=e^T e+b^T X^T X b$, and so

$$\underbrace{Y^T Y - \frac{1}{n} Y^T J Y}_{\text{SSTO}} = \underbrace{b^T X^T X b - \frac{1}{n} Y^T J Y}_{\text{SSR}} + \underbrace{e^T e}_{\text{SSE}} .$$

Each of these is a quadratic form:

- $\text{SSTO} = Y^T \left(I - \frac{1}{n} J \right) Y$,
- $\text{SSR} = Y^T \left(H - \frac{1}{n} J \right) Y$,
- $\text{SSE} = Y^T (I - H) Y$.

ANOVA Table

Source	SS	df	MS
Regression	$b^T X^T X b - \frac{1}{n} Y^T J Y$	1	$b^T X^T X b - \frac{1}{n} Y^T J Y$
Error	$e^T e$	$n-2$	$s = \frac{e^T e}{n-2}$
Total	$Y^T Y - \frac{1}{n} Y^T J Y$	$n-1$	

Confidence and Prediction Intervals

- The standard error for mean value of Y at X_h is $s\sqrt{X_h^T (X^T X)^{-1} X_h}$ where $X_h = \begin{pmatrix} 1 \\ \text{value of } X \end{pmatrix}$.
- The standard error for predicted value of Y at X_h is $s\sqrt{1 + X_h^T (X^T X)^{-1} X_h}$ where $X_h = \begin{pmatrix} 1 \\ \text{value of } X \end{pmatrix}$.

Multiple Linear Regression

More than one predictor variable. Why?

- Often multiple X's arise naturally.
- Control for some X's.
- Want to fit a polynomial.
- Want to compare regression line for two groups.

Multiple Regression Model

k is the number of X's.

The model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \quad i = 1, \ldots, n$.

In matrix form: $Y = X\beta + \varepsilon$, where $\underbrace{Y}_{n\times 1} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, $\underbrace{\beta}_{(k+1)\times 1} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$, $\underbrace{\varepsilon}_{n\times 1} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$, $\underbrace{X}_{n\times(k+1)} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{pmatrix}$.

We need to estimate $k+1$ unknown regression coefficients and 1 unknown variance, so we need $n > k+2$ observations.

Gauss-Markov Conditions

$E(\varepsilon) = 0$ $\text{cov}(\varepsilon) = \sigma^2 I$	$E(Y) = X\beta$ $\text{cov}(Y) = \sigma^2 I$

Least Square Estimate, Residuals, and Standard Error

- LS estimate of β is $b=\begin{pmatrix} b_0 \\ \vdots \\ b_k \end{pmatrix}=\left(X^T X\right)^{-1} X^T Y$.
- Residuals is $e=Y-\hat{Y}=(I-H)Y$.
- The estimate for σ^2 is $s^2=\frac{e^T e}{\text{df}}=\frac{e^T e}{n-k-1}$. This is an unbiased estimator, i.e. $E\left(s^2\right)=\sigma^2$.

General Comments

- Single response, multiple explanatory variables.
- Regression gives mean response for each combination of explanatory variables.
- It won't be a helpful function if it is very complicated.
- It is usually unwise to think that there is some exact/discoverable/estimable regression equation – many possible models are available. One or two models may adequately approximate the mean of the response as a function of the explanatory variables.

Interpreting Estimated Coefficients

- b_0 : Estimate of the mean value of Y when all the X's are 0.
- b_j : Estimated change on Y due to a unit increase in X_j with all other variables held constant (which may be impossible, especially in observational studies).

Assessing Model Assumptions

Use residual plots as in simple linear regression.

- Residuals vs. predicted value: non-constant variance, outliers, influential points.
- Residuals vs. X_j : curvature.

Parameter Estimate

Variable	df	Parameter Estimate	Standard Error	t-value	Pr > |t|
Intercept	1	b_0	$\text{se}\left(b_0\right)$	$\frac{b_0}{\text{se}\left(b_0\right)}$	p-value from t_{n-k+1} distribution
X_1	1	b_1	$\text{se}\left(b_1\right)$	$\frac{b_0}{\text{se}\left(b_1\right)}$	p-value from t_{n-k+1} distribution
...etc...	...etc...	...etc...	...etc...	...etc...	
X_k	1	b_k	$\text{se}\left(b_k\right)$	$\frac{b_0}{\text{se}\left(b_k\right)}$	p-value from t_{n-k+1} distribution

Notes:

- The estimate of $\text{var}\left(b_j\right)$ is the j -th diagonal entry of $s^2\left(X^T X\right)^{-1}=\text{MSE}\left(X^T X\right)^{-1}$, and $\text{se}\left(b_j\right)=\sqrt{\text{var}\left(b_j\right)}$.
- To test $H_0: \beta_j=0$ vs $H_A: \beta_j\neq 0$, use test statistic $t_{\text{obs}}=\frac{b_j}{\text{se}\left(b_j\right)}$, which will be from a value from t_{n-k+1} distribution under H_0 . This test gives an indication of whether or not the j -th predictor variable statistically significantly contributes to the prediction of the response variable over and above all of the other predictor variables.
- The CI for β_j is $b_j \pm t_{n-k-1,1-\alpha/2}\text{se}\left(b_j\right)$. Use Bonferroni if constructing CI for multiple β_j 's.

ANOVA Table

Source	df	SS	MS	F	Pr > F
Regression	k	$Y^T\left(H-\frac{1}{n}J\right)Y$	$\text{MSR}=\frac{\text{SSR}}{k}$	$\frac{\text{MSR}}{\text{MSE}}$	
Error	$n-k-1$	$Y^T(I-H)Y$	$\text{MSE}=\frac{\text{SSE}}{n-k-1}$		
Total	$n-1$				

F-Test From ANOVA

$H_0: \beta_0=\beta_1=\cdots=\beta_k=0$ vs. $H_A:$ at least one $\beta_j\neq 0$.

Test statistic: $F_{\text{obs}}=\frac{\text{MSR}}{\text{MSE}}$. Under H_0 , F_{obs} is an observation from a $F_{k,n-k-1}$ distribution.

This test answers “Does regression do anything?”.

R^2 and Adjusted R^2

Coefficient of Multiple Determination

$R^2=\frac{\text{SSR}}{\text{SSTO}}$ is the percentage of variation in the Y's explained by the regression equation. It is called the coefficient of multiple determination.

It is not the square of a correlation coefficient and $\sqrt{R^2}$ is not really useful in multiple regression.

In multiple regression, R^2 is not that great of a judge of the model because the more predictor variables the greater R^2 , despite whether or not the predictor variables are useful for estimating the response variable.

Adjusted Coefficient of Multiple Determination

An attempt to make R^2 more useful is to use adjusted R^2 . It is adjusted for the number of predictor variables in the model.

$$\text{adj}\,R^2=1-(n-1)\frac{\text{MSE}}{\text{SSTO}}=1-\frac{n-1}{n-k-1}\frac{\text{SSE}}{\text{SSTO}} \quad .$$

It increases only if MSE increases. It is useful for choosing the best model.

Confidence and Prediction Intervals for Mean Response

Let $X_h=\begin{pmatrix}1\\X_{h1}\\\vdots\\X_{hk}\end{pmatrix}$.

The standard error for $E(Y)$ at X_h is $s\sqrt{X_h^T\left(X^TX\right)^{-1}X_h}$.

The standard error for the predicted value of Y is $s\sqrt{1+X_h^T\left(X^TX\right)^{-1}X_h}$.

Indicator/Dummy Variables

For a categorical variable with j levels, define $j-1$ variables $I_i=\{\begin{matrix}1 & \text{if in the i-th level}\\ 0 & \text{otherwise}\end{matrix} \quad , i=1,2,\ldots,j-1$.

Advantages vs Disadvantages

Advantages of Indicator/Dummy Variables	Disadvantages of Indicator/Dummy Variables
• Useful if we expect a complex relationship between predictors and the response variable. • "Non-parametric" approach – doesn't assume form for functional relationship.	1. Lose functional form of relationship. 2. Lose ordering. 3. Costs df.

Partial F-Test

To test whether a subset of the β 's is 0.

We have

- SSE in reduced model > SSE in full model,
- SSR in reduced model < SSR in full model,
- SSTO in reduced model = SSTO in full model,

It can be shown that $F_{\text{obs}}=\frac{(\text{SSE}_{\text{reduced}}-\text{SSE}_{\text{full}})/(\text{df}_{\text{error, reduced}}-\text{df}_{\text{error, full}})}{(\text{SSE}_{\text{full}})/(\text{df}_{\text{error, full}})}$ is an observation from a $F_{\text{df}_{\text{error, reduced}}-\text{df}_{\text{error, full}},\, \text{df}_{\text{error, full}}}$ distribution.

Partial Correlation

Use when considering the reduced vs. full model where the full model has one additional predictor variable. It is a measure of the contribution of the additional predictor variable given the others in the model.

The coefficient of partial correlation is $\sqrt{\frac{\text{SSE}_{\text{reduced}}-\text{SSE}_{\text{full}}}{\text{SSE}_{\text{reduced}}}}$. It is negative if the coefficient of the additional predictor variable is negative, otherwise it is positive.

Polynomial Regression

Second Order (Quadratic)

A second order (quadratic) model in one variable: $Y=\beta_0+\beta_1 X+\beta_2 X^2+\varepsilon$.

Full second order model with two variables: $Y=\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_{11} X_1^2+\beta_{22} X_2^2+\beta_{12} X_1 X_2+\varepsilon$.

Notice that it is still linear regression.

Comments

- As a general rule, when higher order terms are in the model, also include lower order terms in that variable, even if their coefficients not statistically different from 0.
- Idea: Higher order terms are refinements to base effect of predictor variables on Y.
- The meaning of individual coefficients are difficult and not necessary to interpret; can't change X and hold X^2 constant.

Multi Collinearity and Centering

A problem that often arises when doing multiple regression is that X and X^2 are often highly correlated.When predictor variables are highly correlated, the fitted equation is unstable. The estimated coefficients can change dramatically when data changes a little and when the set of predictor variables that are in the model changes. This is multi collinearity.

In the polynomial setting, often improves the situation to center the X's, i.e. use $\tilde{X}_{i1}=X_{i1}-\bar{X}_1$ (where $\bar{X}_1=\frac{1}{n}\sum_{i=1}^{n} X_{i1}$)

in the model instead of X_{i1} .

In matrix form, $Y=\tilde{X}\tilde{\beta}+\varepsilon$, where $\tilde{\beta}=\begin{pmatrix}\gamma_0\\ \beta_1\end{pmatrix}$, $\tilde{X}=\begin{bmatrix}1 & X_{11}-\bar{X}_1\\ \vdots & \vdots\\ 1 & X_{n1}-\bar{X}_1\end{bmatrix}$. Then $\tilde{X}^T\tilde{X}=\begin{bmatrix}n & 0\\ 0 & \sum\left(X_{i1}-\bar{X}_1\right)^2\end{bmatrix}$ is easy to invert.

For multiple regression, $\underbrace{\tilde{X}^T\tilde{X}}_{(k+1)\times(k+1)}=\begin{bmatrix}n & 0 & \cdots & 0\\ 0 & \ddots & \vdots & ⋰\\ \vdots & & Z^TZ & \\ 0 & ⋰ & \vdots & \ddots\end{bmatrix}$ where $\left[Z^TZ_{jj}\right]=\sum_{i=1}^{n}\left(X_{ij}-\bar{X}_j\right)^2$ (diagonal) and

$\left[Z^TZ_{jk}\right]=\sum_{i=1}^{n}\left(X_{ij}-\bar{X}_j\right)\left(X_{ik}-\bar{X}_k\right)\quad j\neq k$ (close to 0 if little correlation).

Interaction

Definition

Two explanatory variables are said to interact if the effect that one has on the mean response depends on the value of the other.

Interaction Term

To model interaction, introduce a new variable X_iX_j in model.

When To Add Interaction Terms

- Question of interest has to do with interaction.
- Suspect interaction exists because of the nature of what's being measured or because plot of residual vs. interaction shows a relationship.
- If an interaction term for 2 predictor variables is in model, should include terms for predictor variables individual (i.e. if X_1X_2 is in model, then include X_1 and X_2).

Multicollinearity

Effects of Multicollinearity

Serious correlation among predictors may have the following effects:

- Regression coefficients will change dramatically according to which other variables are included/excluded from the model. In worst cases, regression coefficients will be dramatically large with signs that may be opposite to what you expect.
- Standard error of regression coefficient will be large.
- Predictors that are known to be related with Y will not have their coefficients be statistically different from 0.

Checking for Multicollinearity

- Check the correlations among pairs of X's.
- Check if regression coefficients have wrong signs given prior knowledge.
- Check if removing or adding a predictor variable produces surprising changes in fitted model.
- Check if predictors anticipated to be important have non-significant coefficients.
- Examine VIF (variance inflation factor): $VIF_i=\left(1-R_i^2\right)^{-1}$ where R_i^2 is the coefficient of multiple determination obtained when the i -th predictor variable is regressed against other predictor variables. Large VIF is a sign of multicollinearity; $VIF>10$ deserves attention (suggestion).

- Examine tolerance $\frac{1}{VIF}$.

Variable Selection

Why eliminate possible explanatory varaibles?

1. Simpler models are better than complex ones – "parsimony". Want a simple model that doesn't leave out anything important.
2. Unnecessary terms in the model result in less precise inferences ($MSE\uparrow$).

Automated Procedures

Forward Selection, Backward Elimination, Stepwise – each give one final model (but not necessarily the same).
More modern: All Possible Subsets Regression – k predictor variables, 2^k possible models.

Criteria for "Best"

- Adjusted R^2 .
- Mallow's C_p : In a model with $p-1\leq k-1$ predictor variables (so p and k β 's), $C_p=\frac{SSE_p}{MSE_k}-(n-2p)$. For a good model, want $C_p=p$.
- Prediction SS: Suppose the i -th data point is deleted. Estimate the line without it and predict Y_i for all n points; call these $\hat{Y}_{i(i)}$. Then $PRESS_p=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_{i(i)}\right)^2$. Smaller $PRESS_p$ is better. (Note: $p=k+1$)

Identifying Outliers and Influential Points

Outliers

Outliers have large residuals. How large is too large?

- Standardized residual: $\frac{e_i}{s\sqrt{1-h_{ii}}}$ where h_{ii} is the i -th diagonal entry of H . If > 3, unusual; if > 4, very unusual.
- Studentized residual: $e_i^{¿}=\frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}$ where $s_{(i)}^2$ is the estimate of σ^2 from regression without the i -th data point.
- Alternate approach: Add to regression a dummy variable $d_i=\begin{cases}1 & \text{for } i\text{-th observation}\\ 0 & \text{otherwise}\end{cases}$. If the coefficient of d_i is statistically different from 0, then the i -th data point does not fit model.

Influential Point

An influential point pulls the lines towards itself.

- Look at $b-b_{(i)}$ where $b_{(i)}$ is the LS estimate with the i -th observation omitted.
- Loot at $\hat{Y}_i-\hat{Y}_{(i)}$.
- Cooks Distance: Looks at all predicted values when the i -th observation is deleted.
- Robust Regression: For example, use Least Median Squares (LMS) which minimizes the median of square residuals. This is computationally extensive.
- Reweighed LS: Remove outliers from LMS and then do Ordinary Least Squares (OLS).