Math 5305 Notes

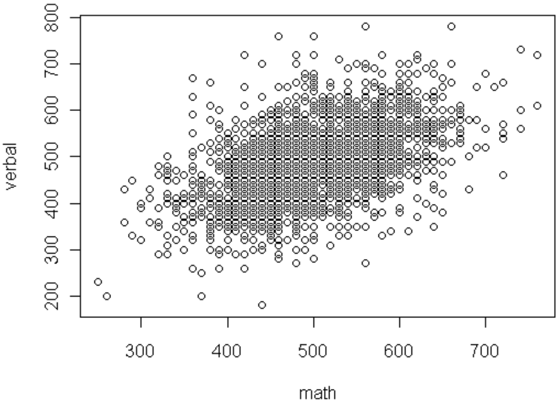
Chapters 2 and 4

Jesse Crawford

Department of Mathematics
Tarleton State University

Outline

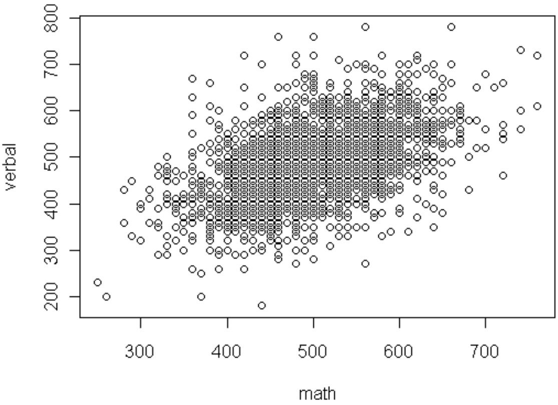
A Scatterplot



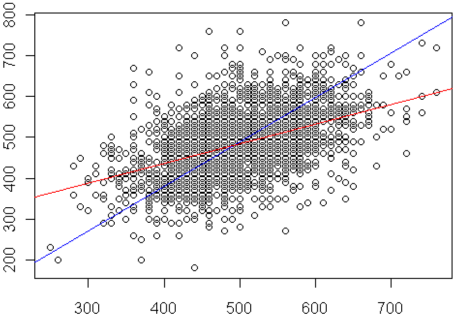
We have a sequence of pairs (x_i, y_i), $i = 1, \ldots, n$.

Sample Statistics

- We have a sequence of pairs (x_i, y_i), $i = 1, \dots, n$.
- $$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$
- $$\mathrm{Var}(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2, \text{ and } s_x = \sqrt{\mathrm{Var}(x)}.$$
- $\bar{y}$, $\mathrm{Var}(y)$, and s_y defined similarly.
- Sample correlation coefficient:
$$r = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y}$$
- $-1 \leq r \leq 1$

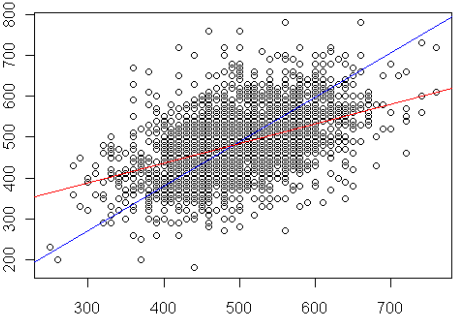


- $\bar{x} = 493$ and $\bar{y} = 482$
- $s_x = 73$ and $s_y = 79$
- $r = 0.44$

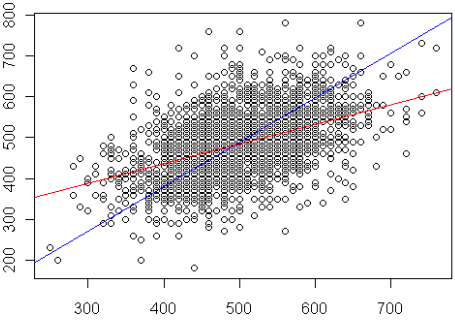


- Regression Line
  - Goes through the point of averages $(\bar{x}, \bar{y})$
  - slope $= r\frac{s_y}{s_x}$

The SD Line



- SD Line
  - Goes through the point of averages $(\bar{x}, \bar{y})$
  - slope $= \text{sign}(r)\frac{s_y}{s_x}$



- Regression Line: $y = 0.483x + 244$
- SD Line: $y = 1.09x - 57.7$

Outline

1. Chapter 2: Simple Linear Regression
2. Section 4.1: Introduction to Multiple Regression
3. Section 4.2: Standard Errors
4. Section 4.3: Explained Variance in Multiple Regression
5. Section 4.4: What Happens if Assumptions Break Down?

Multiple Regression Model

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i, \text{ for } i = 1, \ldots, n.$$

$$\text{Verbal}_i = \beta_1 + \beta_2 \text{Math}_i + \epsilon_i, \text{ for } i = 1, \ldots, 3146.$$

$$\begin{aligned}\text{Post-test}_i = &\beta_1 + \beta_2 \text{Pre-test}_i + \beta_3 \text{MathSAT}_i + \beta_4 \text{VerbSAT}_i \\ &+ \beta_5 \text{HSrank}_i + \beta_6 \text{Clickers}_i + \beta_7 \text{GroupWork}_i \\ &+ \epsilon_i, \text{ for } i = 1, \ldots, 140.\end{aligned}$$

Many Possible Examples!

Multiple Regression Model

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i, \text{ for } i = 1, \ldots, n.$$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$Y = X\beta + \epsilon$$

- Y is an $n \times 1$ **observable random vector.** Called dependent, response, or output variable.
- X is an $n \times p$ **observable** matrix. Can be viewed as **random or constant.** The columns are called independent, explanatory, predictor, control or input variables. Also called covariates.

Multiple Regression Model

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i, \text{ for } i = 1, \ldots, n.$$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$Y = X\beta + \epsilon$$

- X is also called the design matrix. If the first column of X is all 1's then the model has an intercept term.
- β is a **constant, unobservable vector**. It is one of the model **parameters**. The β_j's are called regression coefficients.
- ϵ is a **random, unobservable vector**. The ϵ_i's are called error/disturbance terms.

Mathematical Assumptions

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$Y = X\beta + \epsilon$$

- $p < n$, and X has full rank.
- $\epsilon_1, \ldots, \epsilon_n$ are IID with mean 0 and variance $\sigma^2 > 0$. Note that σ is another model parameter, which is constant and unobservable.

$$E(\epsilon) = 0 \text{ and } \mathrm{cov}(\epsilon) = \sigma^2 I.$$

- If X is random, ϵ is independent of X. Notation: $\epsilon \perp\!\!\!\perp X$.

The OLS Estimator

Theorem

- *The sum of square errors* $\|Y - X\gamma\|^2$ *is minimized when*

$$\gamma = \hat{\beta} = (X'X)^{-1}X'Y.$$

- *That is,* $\hat{\beta}$ *is the "best" estimator for* β *according to the ordinary least squares (OLS) criterion.*
- $\hat{\beta}$ *is called the OLS estimator for* β*.*
- $\hat{\beta}$ *is an* ***observable, random vector****.*

Residuals

Definition

- Define $e = Y - X\hat{\beta}$.
- The e_i's are called *residuals*.
- e is an **observable, random vector.**
- $e \perp X$.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\|e\|^2 = (1 - R^2)\mathrm{Var}(Y)$$

$$RMS = \sqrt{MSE}$$

Some Comparisons

$Y = X\beta + \epsilon$	$Y = X\hat{\beta} + e$
$\epsilon = Y - X\beta$	$e = Y - X\hat{\beta}$
β is an unknown, constant parameter	$\hat{\beta}$ is observable and random
ϵ is unobservable and random	e is observable and random
Errors/disturbance terms	Residuals

The Maximum Likelihood Estimator

Theorem

The OLS estimator $\hat{\beta}$ is conditionally unbiased.

$$E(\hat{\beta} \mid X) = \beta$$

The OLS Estimator is the MLE

Theorem

- *Assume the disturbance terms ϵ_i are normally distributed.*
- *Then the OLS estimator $\hat{\beta}$ is the maximum likelihood estimator (MLE) for β.*

Outline

- Let X_i be the ith row of X, for $i = 1, \ldots, n$.
- $$Y = X\beta + \epsilon$$
  $$Y_i = X_i\beta + \epsilon_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i$$
- $$Y = X\hat{\beta} + e$$
  $$Y_i = X_i\hat{\beta} + e_i = \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip} + e_i$$
- The *predicted* or *fitted* values of Y are
  $$\hat{Y} = X\hat{\beta}$$
  $$\hat{Y}_i = X_i\hat{\beta} = \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$$
- Note that
  $$e = Y - \hat{Y}$$

The Hat Matrix

Definition

- The *hat matrix* is
$$H = X(X'X)^{-1}X'$$
- $$\hat{Y} = HY$$

Properties:

1. $e = (I - H)Y$
2. H is symmetric, and so is $I - H$.
3. H is idempotent ($H^2 = H$), and so is $I - H$
4. X is invariant under H, that is, $HX = X$
5. $e = (Y - HY)$, and $e \perp X$. (If X contains a column of ones, then $\sum_{i=1}^{n} e_i = 0$.)

The Hat Matrix is a Projection Matrix

Definition

The *column space* of X is

$$\text{cols}(X) = \{X\gamma \mid \gamma \in \mathbb{R}^p\}.$$

Proposition

- H is the $n \times n$ matrix that projects $\mathbb{R}^n$ orthogonally onto cols(X).
- $\hat{Y}$ is the orthogonal projection of Y onto cols(X), that is $\hat{Y} = HY$.

Estimating σ^2

Theorem

The estimator

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-p}\|e\|^2$$

is a conditionally unbiased estimator for σ^2.

$$E(\hat{\sigma}^2 \mid X) = \sigma^2$$

This is why we require $n > p$.

The Covariance Matrix of $\hat{\beta}$

Theorem

- $$cov(\hat{\beta} \mid X) = \sigma^2 (X'X)^{-1}$$
- $$\widehat{cov}(\hat{\beta} \mid X) = \hat{\sigma}^2 (X'X)^{-1}$$

Corollary

- $$Var(\hat{\beta}_j) = \sigma^2 [(X'X)^{-1}]_{jj}$$
- $$SE(\hat{\beta}_j) = \sigma \sqrt{[(X'X)^{-1}]_{jj}}$$
- $$\widehat{SE}(\hat{\beta}_j) = \hat{\sigma} \sqrt{[(X'X)^{-1}]_{jj}}$$

Other Assumptions

- The columns of X don't have to be orthogonal to each other.
- The random errors don't have to be normally distributed.

Outline

Decomposing Var(Y)

Proposition

If a multiple regression model has an intercept term, then

$$\text{Var}(Y) = \text{Var}(X\hat{\beta}) + \text{Var}(e).$$

- Reminder:

$$\text{Var}(Y) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

$$\text{Var}(Y) = \text{Var}(X\hat{\beta}) + \text{Var}(e).$$

Definition

Consider a multiple regression model with an intercept term.

- $\text{Var}(Y)$ is the total variance of the response variable.
- $\text{Var}(X\hat{\beta})$ is the variance “explained” by the explanatory variables X.
- $\text{Var}(e)$ is the “unexplained” or “residual” variance.
- The fraction of variance “explained” by the model is

$$R^2 = \frac{\text{Var}(X\hat{\beta})}{\text{Var}(Y)}.$$

- The fraction of variance “explained” by the model is

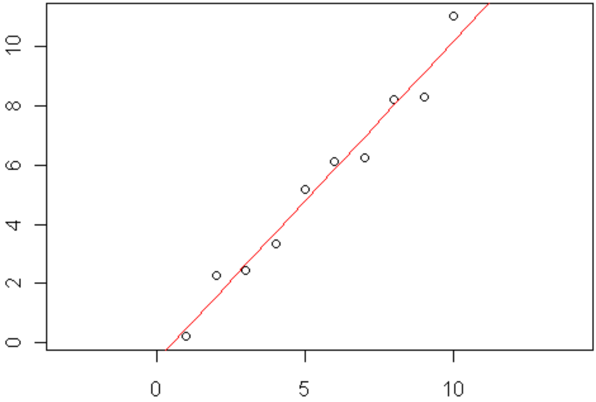
$$R^2 = \frac{\mathrm{Var}(X\hat{\beta})}{\mathrm{Var}(Y)}.$$

- Sometimes called multiple R^2, multiple correlation coefficient, or coefficient of determination.
- $0 \leq R^2 \leq 1$
- For a simple linear regression model, $R^2 = r^2$.
- *If a multiple linear regression model is appropriate*, R^2 measures how well the model fits the data, with values close to one indicating better fit.

Console ~/

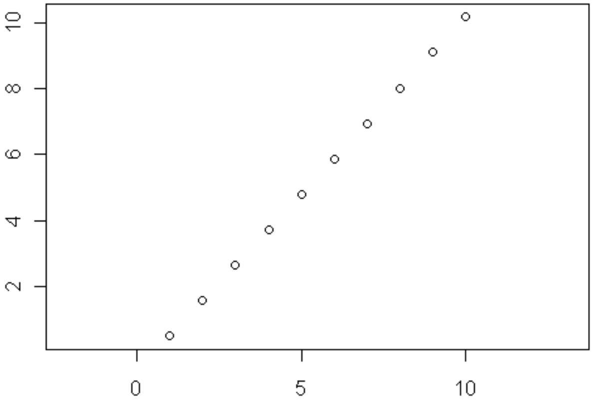
```
> mymodel=lm(Y~X1+X2+X3)
> summary(mymodel)

Call:
lm(formula = Y ~ X1 + X2 + X3)

Residuals:
     Min       1Q   Median       3Q      Max 
-23.5493  -6.4823   0.7492   5.5936  24.9199 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 50.04082    2.94308  17.003  < 2e-16 ***
X1           4.43642    0.31869  13.921  < 2e-16 ***
X2          21.64408    3.12069   6.936  4.7e-10 ***
X3          -0.55334    0.03271 -16.919  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.463 on 96 degrees of freedom
Multiple R-squared: 0.8593, Adjusted R-squared: 0.855 
F-statistic: 195.5 on 3 and 96 DF,  p-value: < 2.2e-16

> names(summary(mymodel))
 [1] "call"          "terms"         "residuals"     "coefficients"
 [5] "aliased"       "sigma"         "df"            "r.squared"   
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled" 
> summary(mymodel)$r.squared
[1] 0.8593479
> 
```

Plot of Y vs. X

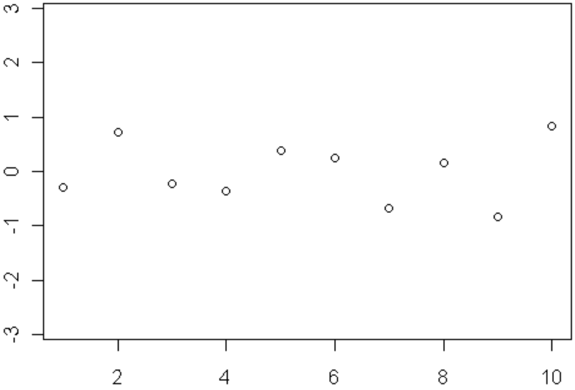


$$R^2 = 0.9709$$

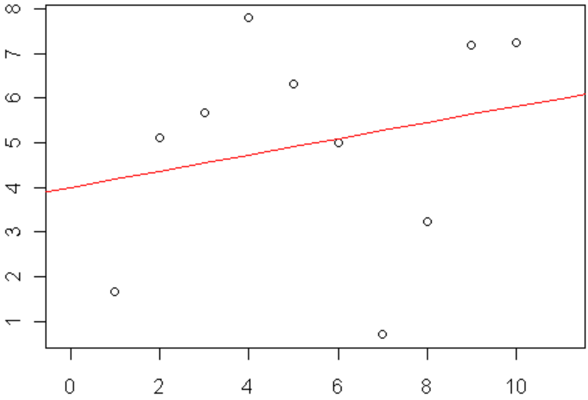
Plot of $\hat{Y}$ vs. X



Plot of e vs. X

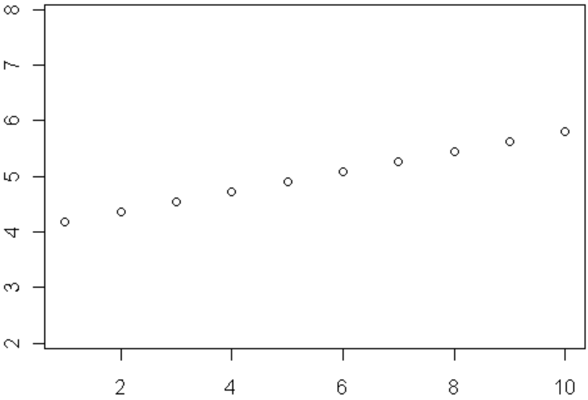


Plot of Y vs. X

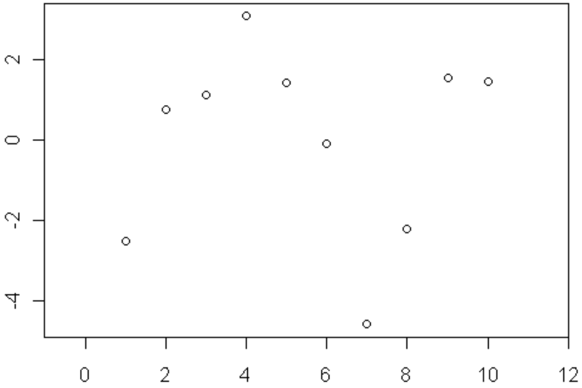


$$R^2 = 0.0517$$

Plot of $\hat{Y}$ vs. X



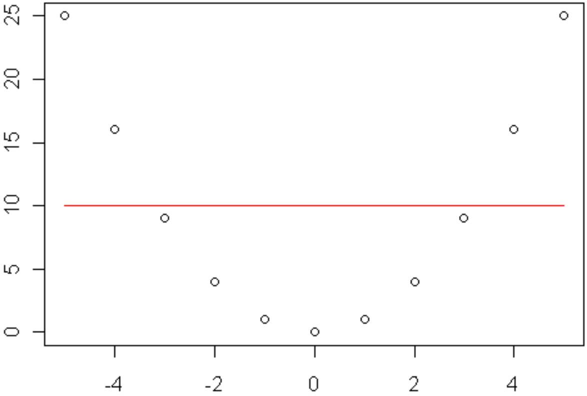
Plot of e vs. X



Correlation is Not Causation

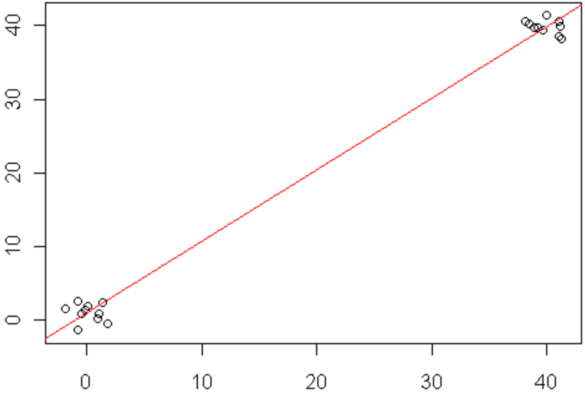
"...over the period 1950–1999, the correlation between the purchasing power of the United States dollar each year and the death rate from lung cancer that year is -0.95*. So* $R^2 = (-0.95)^2 = 0.9$*..."*

Inappropriate Model 1



$$R^2 = 0$$

Inappropriate Model 2



$$R^2 = 0.9925$$

Outline

What Happens if Assumptions Break Down?

- If $E(\epsilon \mid X) \neq 0$, the bias in $\hat{\beta}$ is
$$(X'X)^{-1}X'E(\epsilon \mid X).$$
- If $E(\epsilon \mid X) = 0$, but $\text{cov}(\epsilon \mid X) \neq \sigma^2 I$, then
  - $\hat{\beta}$ will be unbiased, but
  - We can't guarantee that $\text{cov}(\hat{\beta} \mid X) = \sigma^2(X'X)^{-1}$. Therefore, all of our estimates for SEs will be meaningless.