CHAPTER 2

Simple Linear Regression

Simple Linear Regression Model

- Single regressor, *x*; response, *y*

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Population regression model

- β_0 – intercept: if $x = 0$ is in the range, then β_0 is the mean of the distribution of the response y, when $x = 0$; if $x = 0$ is not in the range, then β_0 has no practical interpretation
- β_1 – slope: change in the mean of the distribution of the response produced by a unit change in x
- ε - random error

Simple Linear Regression Model

- Single regressor, *x*; response, *y*

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Simple – only one predictor variable
- Linear in the parameters – no parameter appears as an exponent or is multiplied by another parameter
- Linear in the predictor variable – predictor variable raised to the power of one
- First-order Model – linear in the parameters and the predictor variable

Simple Linear Regression Model

- The response, y, is a random variable
- There is a probability distribution for y at each value of x
  - Mean:

$$E(y \mid x) = \beta_0 + \beta_1 x$$

  - Variance:

$$\mathrm{Var}(y \mid x) = \mathrm{Var}(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2$$

Least-Squares Estimation of the Parameters

- β_0 and β_1 are unknown and must be estimated

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \ldots, n \longleftarrow$$

Sample regression model

- Least squares estimation seeks to minimize the *sum of squares* of the differences between the observed response, y_i, and the straight line.

$$S(\beta_0, \beta_1) = \sum_i \varepsilon_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

Least-Squares Estimation of the Parameters

- Let $\hat{\beta}_0, \hat{\beta}_1$ represent the least squares estimators of β_0 and β_1, respectively.

- These estimators must satisfy:

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Least-Squares Estimation of the Parameters

- Simplifying yields the least squares **normal equations**:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

Least-Squares Estimation of the Parameters

- Solving the normal equations yields the ordinary least squares estimators:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

Least-Squares Estimation of the Parameters

- The fitted simple linear regression model:
  - Sum of Squares Notation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2$$

$$S_{xy} = \sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n} = \sum_{i=1}^{n} y_i \left(x_i - \bar{x}\right)$$

Least-Squares Estimation of the Parameters

- Then

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}}$$

Least-Squares Estimation of the Parameters

- Residuals: $$e_i = y_i - \hat{y}_i$$

- Residuals will be used to determine the **adequacy** of the model

The Rocket Propellant Data

TABLE 2.1 Data for Example 2.1

Observation i	Shear Strength (psi) y_i	Age of Propellant (weeks) x_i
1	2158.70	15.50
2	1678.15	23.75
3	2316.00	8.00
4	2061.30	17.00
5	2207.50	5.50
6	1708.30	19.00
7	1784.70	24.00
8	2575.00	2.50
9	2357.90	7.50
10	2256.70	11.00
11	2165.20	13.00
12	2399.55	3.75
13	1779.80	25.00
14	2336.75	9.75
15	1765.30	22.00
16	2053.50	18.00
17	2414.40	6.00
18	2200.50	12.50
19	2654.20	2.00
20	1753.70	21.50

The Rocket Propellant Data

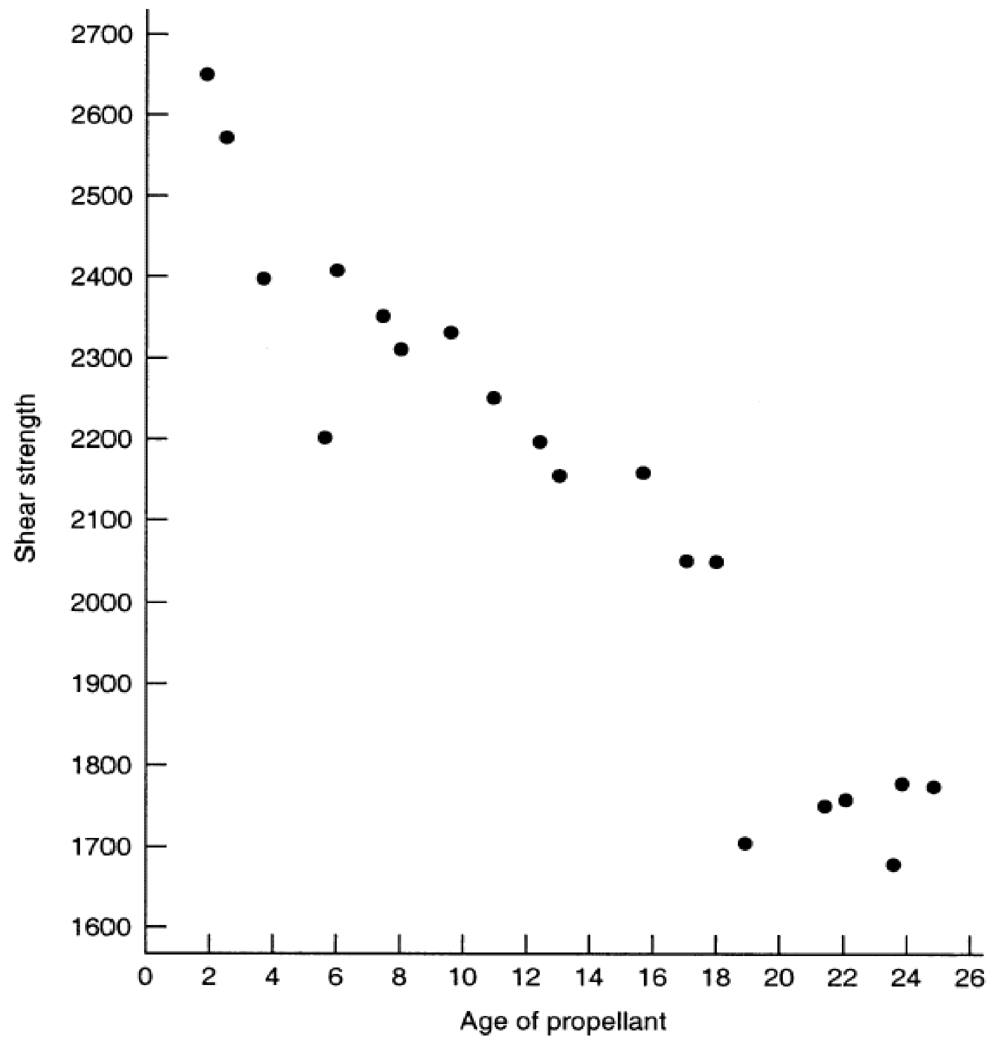

Figure 2.1 Scatter diagram of shear strength versus propellant age. Example 2.1.

Example 2.1- Rocket Propellant Data

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 4677.69 - \frac{71{,}422.56}{20} = 1106.56$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} = 528{,}492.64 - \frac{(267.25)(42{,}627.15)}{20}$$

$$= -41{,}112.65$$

Example 2.1- Rocket Propellant Data

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-41{,}112.65}{1106.56} = -37.15$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 2131.3575 - (-37.15)13.3625 = 2627.82$$

- The least squares regression line is

$$\hat{y} = 2627.82 - 37.15x$$

Example 2.1- Rocket Propellant Data

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.950
R Square	0.902
Adjusted R Square	0.896
Standard Error	96.106
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1527483	1527483	165.38	1.643E-10
Residual	18	166255	9236		
Total	19	1693738			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	2627.82	44.18	59.47	4.06E-22	2534.995	2720.649	2500.642	2755.003
Age of Propellant, x_i (weeks)	-37.15	2.89	-12.86	1.64E-10	-43.223	-31.084	-45.470	-28.837

R code

- rm(list=ls())
- rocket <- read.delim("Data-ex-2-1 (Rocket Prop).txt",h=T)
- names(rocket)
- print(rocket)
- n=dim(rocket)[1]
- plot(rocket$x,rocket$y,pch=20)
- model1 <- lm(y ~ x, data=rocket)
- abline(model1,col="blue")
- summary(model1)
- model1$coefficients
- model1$residuals
- model1$fitted.values

- # coefficient
- Sxx=sum((rocket$x-mean(rocket$x))^2)
- Sxy=sum((rocket$x-mean(rocket$x))*rocket$y)
- Sxy/Sxx
- mean(rocket$y)-Sxy/Sxx*mean(rocket$x)

Example 2.1- Rocket Propellant Data

TABLE 2.2 Data, Fitted Values, and Residuals for Example 2.1

Observed Value, y_i	Fitted Value, $\hat{y}_i$	Residual, e_i
2158.70	2051.94	106.76
1678.15	1745.42	−67.27
2316.00	2330.59	−14.59
2061.30	1996.21	65.09
2207.50	2423.48	−215.98
1708.30	1921.90	−213.60
1784.70	1736.14	48.56
2575.00	2534.94	40.06
2357.90	2349.17	8.73
2256.70	2219.13	37.57
2165.20	2144.83	20.37
2399.55	2488.50	−88.95
1799.80	1698.98	80.82
2336.75	2265.58	71.17
1765.30	1810.44	−45.14
2053.50	1959.06	94.44
2414.40	2404.90	9.50
2200.50	2163.40	37.10
2654.20	2553.52	100.68
1753.70	1829.02	−75.32
$\Sigma y_i = 42627.15$	$\Sigma \hat{y}_i = 42627.15$	$\Sigma e_i = 0.00$

Least-Squares Estimation of the Parameters

- **Properties of Fitted Regression Line**
  - Sum of the residuals (e_i) equals zero
  - Sum of squared residuals is minimized
  - Sum of observed values is the sum of the fitted values
  - Sum of weighted residuals $(x_i e_i)$ is equal to zero
  - Sum of weighted residuals $(\hat{y}_i e_i)$ equals zero
  - Regression line passes through $(\bar{x}, \bar{y})$

Properties of the Least-Squares Estimators and the Fitted Regression Model

- Useful properties of the least-squares fit

1. $$\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right) = \sum_{i=1}^{n} e_i = 0$$

2. $$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

3. The least-squares regression line always passes through the centroid $\left(\bar{y}, \bar{x}\right)$ of the data.

4. $$\sum_{i=1}^{n} x_i e_i = 0$$

5. $$\sum_{i=1}^{n} \hat{y}_i e_i = 0$$

Least-Squares Estimation of the Parameters

- Assessing the Model
  - How well does this equation fit the data?
  - Is the model likely to be useful as a predictor?
  - Are any of the basic assumptions (such as constant variance and uncorrelated errors) violated, if so, how serious is this?

Properties of the Least-Squares Estimators and the Fitted Regression Model

- The least-squares estimators are **unbiased estimators** of their respective parameter:

$$E\left(\hat{\beta}_1\right) = \beta_1 \qquad E\left(\hat{\beta}_0\right) = \beta_0$$

- The variances are

$$\mathrm{VAR}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{xx}} \quad \mathrm{VAR}\left(\hat{\beta}_0\right) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

- The OLS estimators are **Best Linear Unbiased Estimators** (BLUE)

Estimation of σ^2

- Residual (error) sum of squares

$$SS_{\mathrm{Res}} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{n} e_i^2$$

$$= \underbrace{\sum_{i=1}^{n} y_i^2 - n\overline{y}}_{SS_T = \sum \left(y_i - \overline{y} \right)^2} - \hat{\beta}_1 S_{xy}$$

$$= SS_T - \hat{\beta}_1 S_{xy}$$

Estimation of σ^2

- **Unbiased estimator** of σ^2

$$\hat{\sigma}^2 = \frac{SS_{\mathrm{Res}}}{n-2} = MS_{\mathrm{Res}}$$

- The quantity *n* – 2 is the number of degrees of freedom for the residual sum of squares.

Estimation of σ^2

- $\hat{\sigma}^2$ depends on the residual sum of squares. Then:
  - Any violation of the assumptions on the model errors could damage the usefulness of this estimate
  - A misspecification of the model can damage the usefulness of this estimate
  - This estimate is **model dependent**

Hypothesis Testing on the Slope and Intercept

- Three assumptions needed to apply procedures such as **hypothesis testing** and **confidence intervals**.
- Model errors, ε_i,
  - o are normally distributed
  - o are independently distributed
  - o have constant variance
  - o i.e. $\varepsilon_i \sim NID(0, \sigma^2)$

Use of t-tests

Slope

H_0: $\beta_1 = \beta_{10}$ H_1: $\beta_1 \neq \beta_{10}$

- Standard error of the slope: $se\left(\hat{\beta}_1\right) = \sqrt{\dfrac{MS_{\mathrm{Res}}}{S_{xx}}}$
- Test statistic: $t_0 = \dfrac{\hat{\beta}_1 - \beta_{10}}{se\left(\hat{\beta}_1\right)}$
- Reject H_0 if $|t_0| > t_{\alpha/2,n-2}$
- Can also use the *P*-value approach

Use of t-tests

Intercept

H_0: $\beta_0 = \beta_{00}$ H_1: $\beta_0 \neq \beta_{00}$

- Standard error of the intercept: $$se\left(\hat{\beta}_0\right) = \sqrt{MS_{\mathrm{Res}}\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}$$

- Test statistic: $$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{se\left(\hat{\beta}_0\right)}$$

- Reject H_0 if $|t_0| > t_{\alpha/2,n-2}$
- Can also use the *P*-value approach

Testing Significance of Regression

H_0: $\beta_1 = 0$ H_1: $\beta_1 \neq 0$

- This tests the **significance of regression**; that is, is there a linear relationship between the response and the regressor.

- *Failing to reject* $\beta_1 = 0$, implies that there is <u>no linear relationship between y and x</u>

Example 2.3 The Rocket Propellant Data

We test for significance of regression in the rocket propellant regression model of Example 2.1. The estimate of the slope is $\hat{\beta}_1 = -37.15$, and in Example 2.2, we computed the estimate of σ^2 to be $MS_{\text{Res}} = \hat{\sigma}^2 = 9244.59$. The standard error of the slope is

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{MS_{\text{Res}}}{S_{xx}}} = \sqrt{\frac{9244.59}{1106.56}} = 2.89$$

Therefore, the test statistic is

$$t_0 = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} = \frac{-37.15}{2.89} = -12.85$$

If we choose $\alpha = 0.05$, the critical value of t is $t_{0.025,18} = 2.101$. Thus, we would reject H_0: $\beta_1 = 0$ and conclude that there is a linear relationship between shear strength and the age of the propellant. ■

Example 2.1- Rocket Propellant Data

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.950
R Square	0.902
Adjusted R Square	0.896
Standard Error	96.106
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1527483	1527483	165.38	1.643E-10
Residual	18	166255	9236		
Total	19	1693738			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	2627.82	44.18	59.47	4.06E-22	2534.995	2720.649	2500.642	2755.003
Age of Propellant, x_i (weeks)	-37.15	2.89	-12.86	1.64E-10	-43.223	-31.084	-45.470	-28.837

R code

- summary(model1)
- summary(model1)$coef[,1]
- summary(model1)$coef[,2]
- summary(model1)$coef[,3]
- summary(model1)$coef[,4]
- # t test
- summary(model1)$coef[,1]/summary(model1)$coef[,2]
- 2*(1-pt(abs(summary(model1)$coef[,1]/summary(model1)$coef[,2]),n-2))

Testing significance of regression

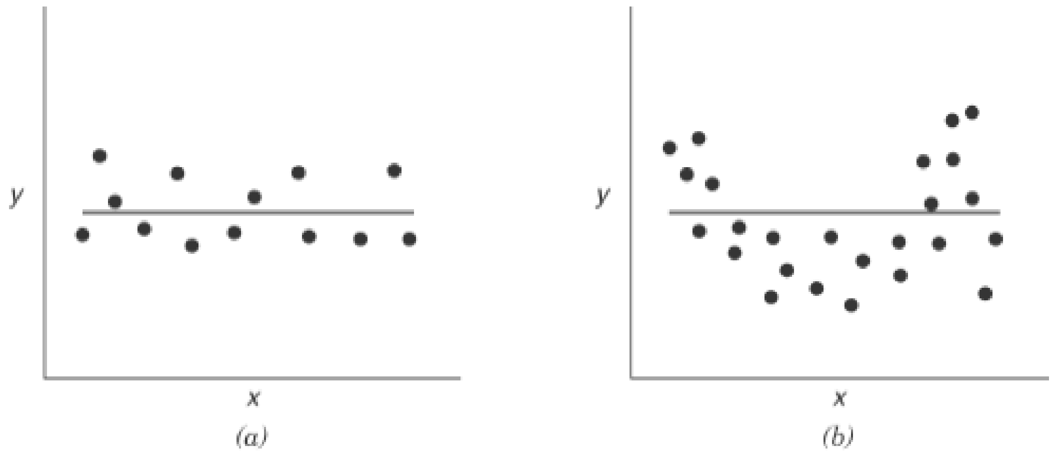

Figure 2.2 Situations where the hypothesis H_0: $\beta_1 = 0$ is not rejected.

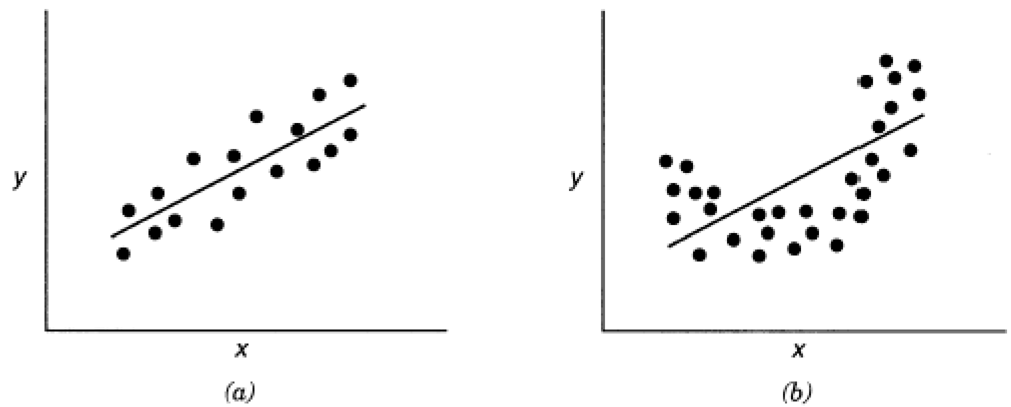

Figure 2.3 Situations where the hypothesis H_0: $\beta_1 = 0$ is rejected.

Analysis of Variance

- Partitioning of total variability

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 + 2\sum (\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

or

$$\underbrace{\sum (y_i - \bar{y})^2}_{SS_T} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{SS_R} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{SS_{\mathrm{Res}}}$$

Analysis of Variance

- Degrees of Freedom

$$\underbrace{\sum\left(y_i-\bar{y}\right)^2}_{SS_T}=\underbrace{\sum\left(\hat{y}_i-\bar{y}\right)^2}_{SS_R}+\underbrace{\sum\left(y_i-\hat{y}_i\right)^2}_{SS_{\mathrm{Res}}}$$

$$n-1 \qquad = \qquad 1 \qquad + \quad (n-2)$$

- Mean Squares

$$MS_R=\frac{SS_R}{1} \qquad MS_{\mathrm{Res}}=\frac{SS_{\mathrm{Res}}}{n-2}$$

Analysis of Variance

- ANOVA procedure for testing H_0: $\beta_1 = 0$

Source of Variation	Sum of Squares	DF	MS	F_0
Regression	SS_R	1	MS_R	MS_R/MS_{Res}
Residual	SS_{Res}	n-2	MS_{Res}	
Total	SS_T	n-1		

- A large value of F_0 indicates that regression is significant; specifically, reject if $F_0 > F_{\alpha,1,n-2}$
- Can also use the *P*-value approach

Example 2.1- Rocket Propellant Data

TABLE 2.5 Analysis-of-Variance Table for the Rocket Propellant Regression Model

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P value
Regression	1,527,334.95	1	1,527,334.95	165.21	1.66×10^{-10}
Residual	166,402.65	18	9,244.59		
Total	1,693,737.60	19			

Example 2.1- Rocket Propellant Data

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.950
R Square	0.902
Adjusted R Square	0.896
Standard Error	96.106
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1527483	1527483	165.38	1.643E-10
Residual	18	166255	9236		
Total	19	1693738			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	2627.82	44.18	59.47	4.06E-22	2534.995	2720.649	2500.642	2755.003
Age of Propellant, x_i (weeks)	-37.15	2.89	-12.86	1.64E-10	-43.223	-31.084	-45.470	-28.837

R Code

- # F test and R square
- SST=sum((rocket$y-mean(rocket$y))^2)
- SSRes=sum((rocket$y-model1$fitted.values)^2)
- SSR=sum((model1$fitted.values-mean(rocket$y))^2)
- SSR+SSRes
- SST

- anova(model1)
- F=(SSR/1)/(SSRes/(n-2))
- F
- summary(model1)$coef[2,3]^2
- # F tes p value
- 1-pf(F,1,n-2)

Analysis of Variance

Relationship between t_0 and F_0:

- For H_0: $\beta_1 = 0$, it can be shown that:

$$t_0^2 = F_0$$

So for testing significance of regression, the t-test and the ANOVA procedure are equivalent (only true in simple linear regression)

Interval Estimation in Simple Linear Regression

- 100(1-α)% Confidence interval for Slope

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\, se\left(\hat{\beta}_1\right) \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\, se\left(\hat{\beta}_1\right)$$

- 100(1-α)% Confidence interval for the Intercept

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\, se\left(\hat{\beta}_0\right) \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\, se\left(\hat{\beta}_0\right)$$

Interval Estimation in Simple Linear Regression

Example 2.5 The Rocket Propellant Data

We construct 95% CIs on β_1 and σ^2 using the rocket propellant data from Example 2.1. The standard error of $\hat{\beta}_1$ is $se(\hat{\beta}_1) = 2.89$ and $t_{0.025,18} = 2.101$. Therefore, from Eq. (2.35), the 95% CI on the slope is

$$\hat{\beta}_1 - t_{0.025,18}se(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{0.025,18}se(\hat{\beta}_1)$$

$$-37.15 - (2.101)(2.89) \le \beta_1 \le -37.15 + (2.101)(2.89)$$

or

$$-43.22 \le \beta_1 \le -31.08$$

Example 2.1- Rocket Propellant Data

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.950
R Square	0.902
Adjusted R Square	0.896
Standard Error	96.106
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1527483	1527483	165.38	1.643E-10
Residual	18	166255	9236		
Total	19	1693738			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	2627.82	44.18	59.47	4.06E-22	2534.995	2720.649	2500.642	2755.003
Age of Propellant, x_i (weeks)	-37.15	2.89	-12.86	1.64E-10	-43.223	-31.084	-45.470	-28.837

R code

- # CI for coef
- # 95%
- summary(model1)$coef[,1]-qt(0.025,18)*summary(model1)$coef[,2]
- summary(model1)$coef[,1]+qt(0.025,18)*summary(model1)$coef[,2]
- # 99%
- summary(model1)$coef[,1]-qt(0.005,18)*summary(model1)$coef[,2]
- summary(model1)$coef[,1]+qt(0.005,18)*summary(model1)$coef[,2]

Interval Estimation in Simple Linear Regression

- 100(1-α)% Confidence interval for σ^2

$$\frac{(n-2)MS_{RES}}{\chi^2_{\alpha/2,n-2}} \le \sigma^2 \le \frac{(n-2)MS_{RES}}{\chi^2_{1-\alpha/2,n-2}}$$

Example 2.1- Rocket Propellant Data

- 100(1-.05)% Confidence interval for σ^2

$$\frac{(20-2)9236}{31.53} \leq \sigma^2 \leq \frac{(20-2)9236}{8.23}$$

$$5,273.52 \leq \sigma^2 \leq 20,199.25$$

Interval Estimation of the Mean Response

- Let x_0 be the level of the regressor variable at which we want to estimate the mean response, i.e.

$$E\left(y \mid x_0\right) = \mu_{y|x_0}$$

  - Point estimator of $E\left(y \mid x_0\right)$ once the model is fit:

$$E(y|x_0) = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

  - In order to construct a confidence interval on the mean response, we need the variance of the point estimator.

Interval Estimation of the Mean Response

- The variance of $\hat{\mu}_{y|x_0}$ is

$$\mathrm{Var}\left(\hat{\mu}_{y|x_0}\right) = \mathrm{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x_0\right) = \mathrm{Var}\,[\bar{y} + \hat{\beta}_1 (x_0 - \bar{x})]$$

$$= \mathrm{Var}\,(\bar{y}) + \mathrm{Var}\,[\hat{\beta}_1 (x_0 - \bar{x})]$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \bar{x})^2}{S_{xx}} = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

Interval Estimation of the Mean Response

- 100(1-α)% confidence interval for $E(y|x_0)$

$$\hat{\mu}_{y|x_0} - t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \le E(y\,|\,x_0)$$

$$\le \hat{\mu}_{y|x_0} + t_{\alpha/2,n-2}\sqrt{MS_{\text{Res}}\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$$

Notice that the **width** of the CI depends on the location of the point of interest

Example 2.1- Rocket Propellant Data

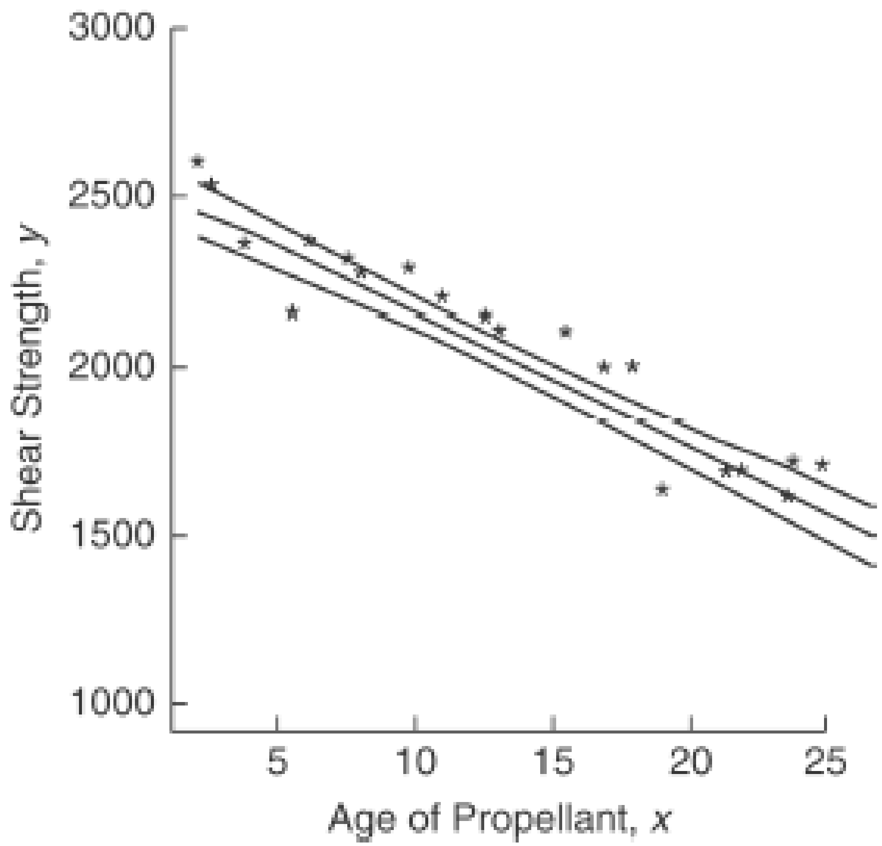

Figure 2.4 The upper and lower 95% confidence limits for the propellant data.

R code

- # CI for mean response, prediction of new observation,
- newx<-seq(0,30)
- conf<-predict(model1,newdata=data.frame(x=newx),interval = c("confidence"),
- level = 0.95,type="response")
- plot(rocket$x,rocket$y,pch=20)
- model1 <- lm(y ~ x, data=rocket)
- abline(model1,col="blue")
- lines(newx,conf[,2],col="red",lty=2)
- lines(newx,conf[,3],col="red",lty=2)
- pred<-predict(model1,newdata=data.frame(x=newx),interval = c("prediction"),
- level = 0.95,type="response")
- lines(newx,pred[,2],col="green",lty=2)
- lines(newx,pred[,3],col="green",lty=2)

- MS_Res=SSRes/n-2
- model1$coef[1]+model1$coef[2]*newx-qt(0.025,n-2)*sqrt(MS_Res*(1/n+(newx-mean(rocket$x))^2/Sxx))
- model1$coef[1]+model1$coef[2]*newx+qt(0.025,n-2)*sqrt(MS_Res*(1/n+(newx-mean(rocket$x))^2/Sxx))

- points(newx,model1$coef[1]+model1$coef[2]*newx-qt(0.025,n-2)*sqrt(MS_Res*(1/n+(newx-mean(rocket$x))^2/Sxx))
- ,lwd=3,col="grey",type="l")
- points(newx,model1$coef[1]+model1$coef[2]*newx+qt(0.025,n-2)*sqrt(MS_Res*(1/n+(newx-mean(rocket$x))^2/Sxx))
- ,lwd=3,col="grey",type="l")

- model1$coef[1]+model1$coef[2]*newx-qt(0.025,n-2)*sqrt(MS_Res*(1/n+1+(newx-mean(rocket$x))^2/Sxx))
- model1$coef[1]+model1$coef[2]*newx+qt(0.025,n-2)*sqrt(MS_Res*(1/n+1+(newx-mean(rocket$x))^2/Sxx))

- points(newx,model1$coef[1]+model1$coef[2]*newx-qt(0.025,n-2)*sqrt(MS_Res*(1/n+1+(newx-mean(rocket$x))^2/Sxx))
- ,lwd=3,col="cyan",type="l")
- points(newx,model1$coef[1]+model1$coef[2]*newx+qt(0.025,n-2)*sqrt(MS_Res*(1/n+1+(newx-mean(rocket$x))^2/Sxx))
- ,lwd=3,col="cyan",type="l")

Prediction of New Observations

- Suppose we wish to construct a *prediction* interval on a future observation, y_0 corresponding to a particular level of x, say x_0.
- The point estimate would be:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

- The confidence interval on the mean response at this point is not appropriate for this situation.

Prediction of New Observations

- Let the random variable, ψ, be $\psi = y_0 - \hat{y}_0$
- ψ is normally distributed with
  - o E(ψ) = 0
  - o Var(ψ) = $\mathrm{Var}\left(y_0 - \hat{y}_0\right) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]$

Prediction of New Observations

- 100(1 - α)% prediction interval on a future observation, y_0, at x_0

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{MS_{\mathrm{Res}}\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \le y_0$$

$$\le \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{MS_{\mathrm{Res}}\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}$$

Example 2.1- Rocket Propellant Data

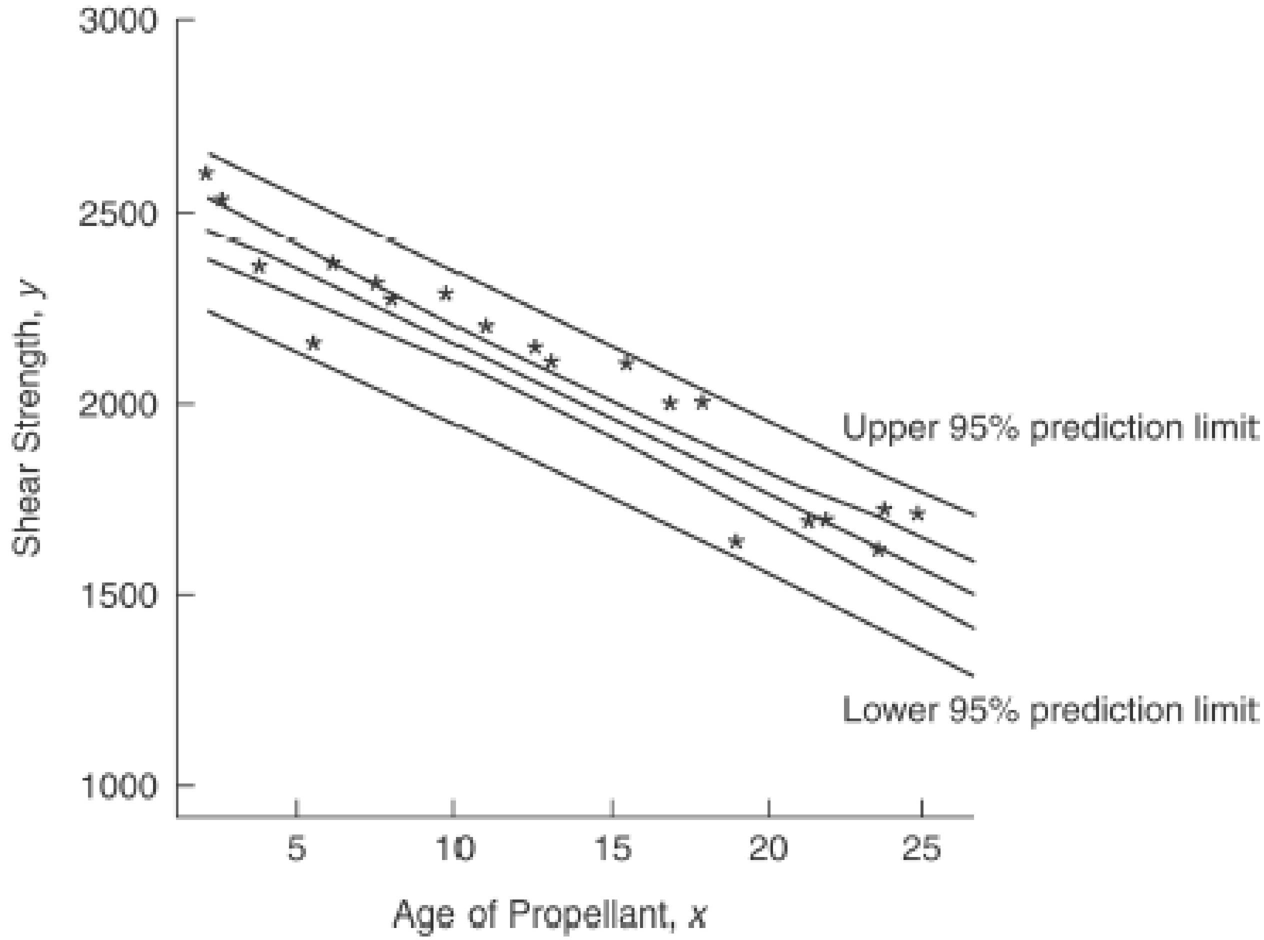

Figure 2.5 The 95% confidence and prediction intervals for the propellant data.

Differences between confidence interval and prediction interval

- A prediction interval is similar in spirit to a confidence interval, except that
  - the prediction interval is designed to cover a “moving target”, the random future value of y, while
  - the confidence interval is designed to cover the “fixed target”, the average (expected) value of y, E(y)
- Although both are centered at $\hat{y}_0$, the prediction interval is wider than the confidence interval, for a given x_0 and confidence level. This makes sense, since
  - the prediction interval must take account of the tendency of y to fluctuate from its mean value, while
  - the confidence interval simply needs to account for the uncertainty in estimating the mean value.

Similarities between confidence interval and prediction interval

- For a given data set, the error in estimating E(y_0) and y_0 grows as x_0 moves away from $\bar{x}$. Thus, the further x_0 is from $\bar{x}$, the wider the confidence and prediction intervals will be.
- If any of the conditions underlying the model are violated, then the confidence intervals and prediction intervals may be invalid as well. This is why it's so important to check the conditions by examining the residuals, etc.

Coefficient of Determination

- R^2 - coefficient of determination

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{\mathrm{Res}}}{SS_T}$$

- Proportion of variation explained by the regressor, x
- For the rocket propellant data

$$R^2 = \frac{SS_R}{SS_T} = \frac{1527483}{1693738} = 0.902$$

Coefficient of Determination

- R^2 can be misunderstood:
  - A high coefficient of determination indicates that a useful prediction can be made.
  - A high coefficient of determination indicates that the estimated regression line is good fit.
  - A coefficient of determination near zero indicates that *x* and *y* are not related.

Coefficient of Determination

- *R^2* can be misleading:
  - Simply adding more terms to the model will increase R^2
  - As the range of the regressor variable increases (decreases), R^2 generally increases (decreases).
  - R^2 does not indicate the appropriateness of a linear model

R code

- # F test and R square
- SST=sum((rocket$y-mean(rocket$y))^2)
- SSRes=sum((rocket$y-model1$fitted.values)^2)
- SSR=sum((model1$fitted.values-mean(rocket$y))^2)
- SSR+SSRes
- SST
- 1-sum((rocket$y-model1$fitted.values)^2)/sum((rocket$y-mean(rocket$y))^2)
- summary(model1)$r.square

Considerations in the Use of Regression

- Extrapolating
- Extreme points will often influence the slope
- Outliers can disturb the least-squares fit
- Linear relationship does not imply cause-effect relationship

Using SAS and R for Simple Linear Regression

- SAS

```
libname mydata "/courses/u_uc.edu1/i_835107/c_3957"
access=readonly;
proc print data=mydata.rocket;
proc reg data=mydata.rocket;
model strength=age/p clm cli;
```

Using SAS and R for Simple Linear Regression

- R

```
rocket <- read.delim("e:\\Data-ex-2-1 (Rocket Prop).txt",h=T)
 names(rocket)
 print(rocket)
 temp <- lm(y ~ x, data=rocket)
 summary(temp)
 anova(temp)
 predict(temp,rocket,level=.95,interval="confidence")
```

Regression Through the Origin

- The no-intercept model is

$$y = \beta_1 x + \varepsilon$$

- This model would be appropriate for situations where the origin (0, 0) has some meaning.
- A scatter diagram can aid in determining where an intercept- or no-intercept model should be used.
- In addition, the practitioner could test *both* models. Examine t-tests, residual mean square.

Exercises

- Read data into Rstudio
- Plot data, explore data
- Perform linear regression (with lm() and without lm())
- Obtain S_{xx}, S_{xy}
- Obtain coefficient estimates
- Obtain SS_T, SS_R, SS_{Res}
- Obtain R^2
- Perform t tests, compute t statistics and p values.
- Perform ANOVA test (F test), compute F statistic, and p value.
- Compute confidence interval for coefficients, fitted values and predict values.
- Plot regression line, plot confidence interval for fitted values, predict values.
- Generate report using R markdown + knitr.