Outline

Example. Wisconsin Lottery Sales

- What factors affect lottery sales? Helpful to know for marketing, e.g., where to establish new retail outlets.
- *i* unit of analysis, ZIP (postal) code
- *n* = 50 randomly selected geographic areas
- *y*= average lottery sales (SALES) over a forty-week period, April, 1998 through January, 1999,
- *x* = population (POP), measure of size of the area.
- Later, we will introduce other factors including area's typical age, education level, income, and so forth. Population is the obvious place to start.
- Here are some summary statistics.

Table: Summary Statistics of Each Variable

Variable	Mean	Median	Standard Deviation	Minimum	Maximum
POP	9,311	4,406	11,098	280	39,098
SALES	6,495	2,426	8,103	189	33,181

Source: Frees and Miller (2003).

Example. Wisconsin Lottery Sales

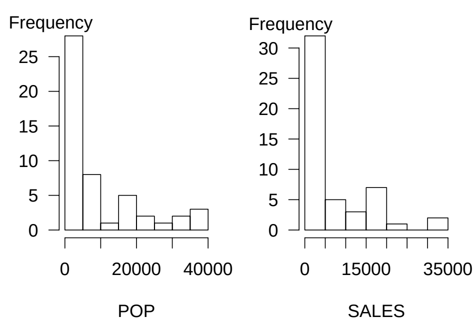
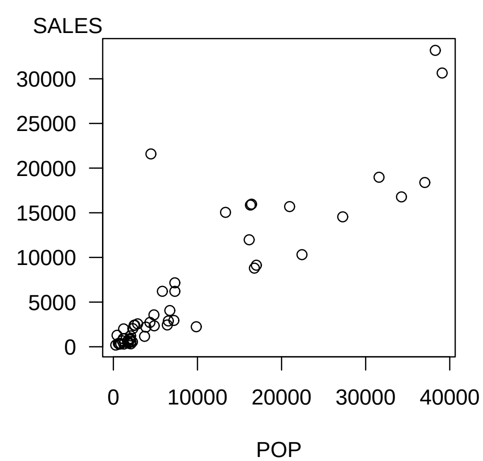


Figure: Histograms of Population and Sales. Each distribution is skewed to the right, indicating that there are many small areas compared to a few areas with larger sales and populations.

Scatter Plot

- The basic graphical tool used to investigate the relationship between the two variables is a *scatter plot*.



Correlations

- One way to summarize the strength of the relationship between two variables is through a *correlation* statistic.
- The *ordinary, or Pearson, correlation* coefficient is defined as

$$r = \frac{1}{(n-1)s_x s_y} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

Recall the sample standard deviation $s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2}$.

- The correlation coefficient is said to be a "unitless" measure.
  - It is unaffected by scale and location changes of either, or both, variables.
  - It can readily be compared across different data sets.
- Correlation coefficients take up less space to report than a scatter plot and are often the primary statistic of interest.
  - Scatter plots help us understand other aspects of the data, such as the range, and also provide indications of nonlinear relationships in the data.

Method of Least Squares

- Can knowledge of population (x) help us understand sales (y)?
- Method of Least Squares
  - Begin with the line $y = b_0^* + b_1^* x$, where the intercept and slope, b_0^* and b_1^*, are merely generic values.
  - For the ith observation, $y_i - (b_0^* + b_1^* x_i)$ represents the deviation of the observed value y_i from the line at x_i.
  - The sum of squared deviations is

$$SS(b_0^*, b_1^*) = \sum_{i=1}^{n} (y_i - (b_0^* + b_1^* x_i))^2$$

  - Minimize this quantity by taking derivatives with respect to the intercept and slope, setting equal to zero and solving

$$\frac{\partial}{\partial b_0^*} SS(b_0^*, b_1^*) = \sum_{i=1}^{n} (-2)(y_i - (b_0^* + b_1^* x_i)) = 0$$

and

$$\frac{\partial}{\partial b_1^*} SS(b_0^*, b_1^*) = \sum_{i=1}^{n} (-2x_i)(y_i - (b_0^* + b_1^* x_i)) = 0.$$

Least Squares Estimates

- The solution gives the *least squares intercept and slope estimates*

$$b_1 = r\frac{s_y}{s_x} \quad \text{and} \quad b_0 = \overline{y} - b_1 \overline{x}.$$

- We have dropped the asterisk, or star (*) notation because these are no longer generic values.
- The line that they determine, $\widehat{y} = b_0 + b_1 x$, is called the *estimated, or fitted, regression line*.

Example. Wisconsin Lottery Sales

For these data, we have $r = 0.886$ and recall

Table: Summary Statistics of Each Variable

Variable	Mean	Median	Standard Deviation	Minimum	Maximum
POP	9,311	4,406	11,098	280	39,098
SALES	6,495	2,426	8,103	189	33,181

Thus,

- $b_1 = 0.886\,(8,103)\,/11,098 = 0.647$ and
- $b_0 = 6,495 - (0.647)9,311 = 470.8$.
- This yields the fitted regression line

$$\widehat{y} = 470.8 + (0.647)x.$$

Example. Summarizing Simulations

- Manistre and Hancock (2005) simulated a 10-year European put option and demonstrated the relationship between the value-at-risk (VaR) and the conditional tail expectation (CTE)
- Stock prices are modeled as

$$S(Z) = 100 \exp\left((.08)10 + .15\sqrt{10}Z \right),$$

annual mean return of 8% and standard deviation 15% .
- The present value of this option is

$$C(Z) = \mathrm{e}^{-0.06(10)} \max\left(0, 110 - S(Z)\right),$$

based on a 6% discount rate.
- 1,000 i.i.d. standard normal random variables were simulated and calculate each of 1000 present values, $C_{i1}, \ldots, C_{i,1000}$.
- Var_i is the 95th percentile
- CTE_i is the average of the highest 50.

Example. Summarizing Simulations

The correlation coefficient turns out to be $r = 0.782$.

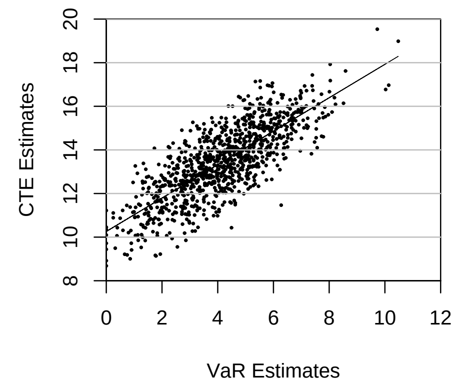


Figure: Plot of Conditional Tail Expectation (CTE) versus Value at Risk (VaR). Based on $n = 1,000$ simulations from a 10 year European put bond.

Observables Representation

Basic Linear Regression Model Observables Representation Sampling Assumptions
F1. $\mathrm{E}\ y_i = \beta_0 + \beta_1 x_i$.
F2. $\{x_1, \ldots, x_n\}$ are non-stochastic variables.
F3. $\mathrm{Var}\ y_i = \sigma^2$.
F4. $\{y_i\}$ are independent random variables.

- For F4, think of stratified sampling, where each x_i is a strata (or group)
- For F3, a common variance is known as homoscedasticity
- We sometimes require

F5. $\{y_i\}$ are normally distributed.

However, approximate normality is enough for central limit theorems that we will need for inference.

Graphical Representation

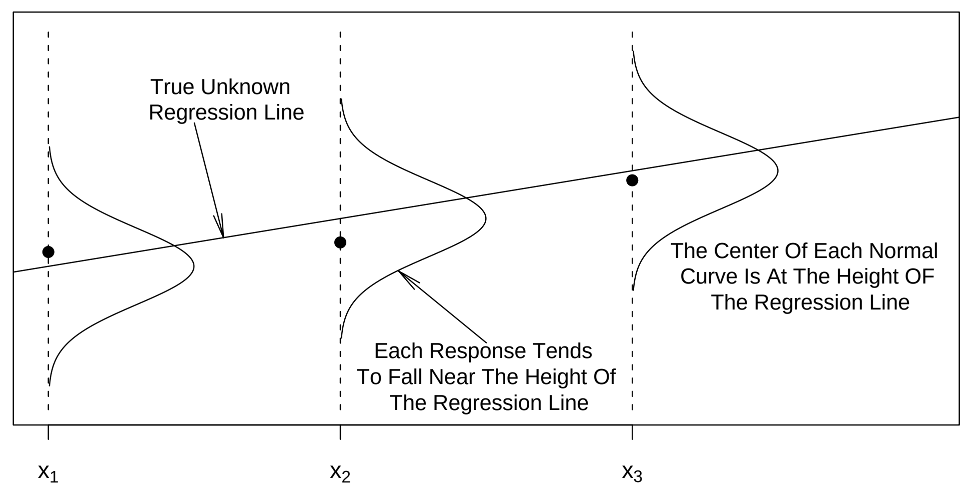


Figure: The distribution of the response varies by the level of the explanatory variable.

Error Representation

Basic Linear Regression Model Error Representation Sampling Assumptions
E1. $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.
E2. $\{x_1, \ldots, x_n\}$ are non-stochastic variables.
E3. $\mathrm{E}\ \varepsilon_i = 0$ and $\mathrm{Var}\ \varepsilon_i = \sigma^2$.
E4. $\{\varepsilon_i\}$ are independent random variables.

- The error representation is a useful springboard for residual analysis (Section 2.6)
- The observable representation is a useful springboard for extensions to nonlinear regression models
- These two sets of assumptions are equivalent

Statistics versus Parameters

- Statistics summarize the (observed) sample/data
- Parameters summarize the (generally unobserved) population
- Use Greek letters for parameters, roman letters for statistics

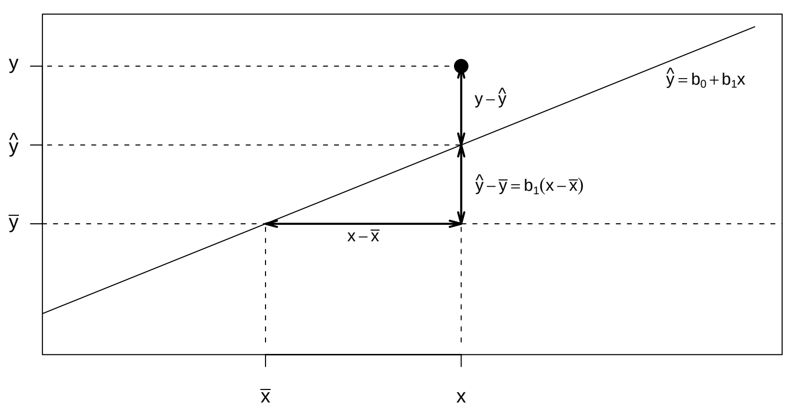
Table: Summary Measures of the Population and Sample

Data	Summary Measures	Regression Line		Variance
		Intercept	Slope	
Population	Parameters	β_0	β_1	σ^2
Sample	Statistics	b_0	b_1	s^2

Partitioning the Variability

We now have two “estimates” of y_i, $\overline{y}$ and $\widehat{y}_i$

$$\underbrace{y_i - \overline{y}}_{\text{total deviation}} = \underbrace{y_i - \widehat{y}_i}_{\text{unexplained deviation}} + \underbrace{\widehat{y}_i - \overline{y}}_{\text{explained deviation}}$$



Partitioning the Variability

After a little algebraic manipulation, this yields

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2 + \sum_{i=1}^{n} \left(\widehat{y}_i - \overline{y}\right)^2,$$

or $Total\ SS = Error\ SS + Regression\ SS$ where SS stands for sum of squares.

- Summarize with “R-square,” the *coefficient of determination*, defined as

$$R^2 = \frac{Regression\ SS}{Total\ SS}.$$

- R^2 = the proportion of variability explained by the regression line.
- If the regression line fits the data perfectly, then $Error\ SS = 0$ and $R^2 = 1$.
- If the regression line provides no information about the response, then $Regression\ SS = 0$ and $R^2 = 0$.
- Property: $0 \leq R^2 \leq 1$, with larger values implying a better fit.

The Size of a Typical Deviation: s

- Define the estimate of the disturbance term $\epsilon_i = y_i - (\beta_0 + \beta_1 x_i)$,
$$e_i = y_i - (b_0 + b_1 x_i)$$
the ith residual.
- If we could observe disturbances, then we would estimate σ^2 using $(n-1)^{-1} \sum_{i=1}^{n} (\varepsilon_i - \overline{\varepsilon})^2$.
- Instead, an estimator of σ^2, the *mean square error (MSE)*, is defined as
$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2.$$
- The residual standard deviation is $s = \sqrt{s^2}$.
- Property of least square residuals, $\overline{e} = 0$.
- Dividing by $n-2$ makes s^2 unbiased.
  - Two points determine a line.
  - With n observations, there are $n-2$ "free" observations that contribute to the variability.

ANOVA Table

Define
$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \frac{Error\ SS}{n-2} = MSE.$$
and

ANOVA Table

Source	Sum of Squares	df	Mean Square
Regression	*Regression SS*	1	*Regression MS*
Error	*Error SS*	$n-2$	*MSE*
Total	*Total SS*	$n-1$	

The ANOVA table is merely a bookkeeping device used to keep track of the sources of variability.

Example. Wisconsin Lottery Sales

ANOVA Table

Source	Sum of Squares	df	Mean Square
Regression	2,527,165,015	1	2,527,165,015
Error	690,116,755	48	14,377,432
Total	3,217,281,770	49	

From this table, you can check that $R^2 = 78.5\%$ and $s = 3,792$.

Weighted Sums

- The least squares estimates can be expressed as weighted sum of the responses.
- Define the weights
$$w_i = \frac{x_i - \overline{x}}{s_x^2 (n-1)}.$$
- The sum of x-deviations $(x_i - \overline{x})$ is zero, we see that $\sum_{i=1}^{n} w_i = 0$.
- The slope estimate is
$$b_1 = r\frac{s_y}{s_x} = \frac{1}{(n-1)s_x^2} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} w_i (y_i - \overline{y}) = \sum_{i=1}^{n} w_i y_i.$$
- A similar result holds for the intercept estimate (with different weights)
- There exists central limit theorems for weighted sums, so that we may treat b_1 and b_0 as approximately normal, even if y is not normally distributed.

Properties of Regression Coefficients

- Regression coefficients are unbiased.
- By the linearity of expectations and Assumption F1, we have

$$\mathrm{E}\ b_1 = \sum_{i=1}^{n} w_i \mathrm{E}\ y_i = \beta_0 \sum_{i=1}^{n} w_i + \beta_1 \sum_{i=1}^{n} w_i x_i = \beta_1 .$$

- Some easy algebra also shows that
  - Here, the sum $\sum_{i=1}^{n} w_i x_i = \left[s_x^2(n-1)\right]^{-1} \sum_{i=1}^{n} (x_i - \overline{x})\, x_i = \left[s_x^2(n-1)\right]^{-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = 1.$
  - $\sum_{i=1}^{n} w_i^2 = 1/\left(s_x^2(n-1)\right).$
- By Assumption F4, we have

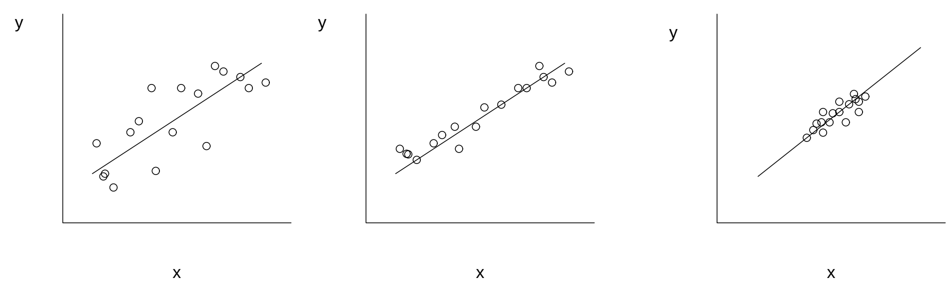
$$\mathrm{Var}\ b_1 = \sum_{i=1}^{n} w_i^2 \mathrm{Var}\ y_i = \frac{\sigma^2}{s_x^2(n-1)}.$$

Standard Errors

The *standard error* of b_1, the estimated standard deviation of b_1, is defined as

$$se(b_1) = \frac{s}{s_x\sqrt{n-1}}.$$

- As n becomes larger, $se(b_1)$ becomes smaller.
- As s becomes smaller, $se(b_1)$ becomes smaller.
- As s_x increases, then $se(b_1)$ becomes smaller.



Is the Explanatory Variable Important?: The *t*-Test

- Logic: If $\beta_1 = 0$, then the model is E $y = \beta_0 + \varepsilon$. That is, it contains no x.
- Is $H_0 :?\beta_1 = 0$ valid? We respond to this question by looking at the test statistic

$$t-\mathrm{ratio} = \frac{\text{estimator} - \text{hypothesized value of parameter}}{\text{standard error of the estimator}}.$$

- For the case of $H_0 : \beta_1 = 0$, we examine $t(b_1) = b_1/se(b_1)$.
- Under Assumptions F1-F5 and H_0, the distribution of $t(b_1)$ follows a t-distribution with $df = n - 2$ degrees of freedom.

Example. Wisconsin Lottery Sales

- The residual standard deviation is $s = 3,792$.
- The x-standard deviation is $s_x = 11,098$.
- Thus, the standard error of the slope is $se(b_1) = 3792/(11098\sqrt{50-1}) = 0.0488$.
- The slope estimate is $b_1 = 0.647$.
- Thus, the t-statistic is $t(b_1) = 0.647/0.0488 = 13.4$.
- We interpret this by saying that the slope is 13.4 standard errors above zero.
- For the hypothesis test, the 97.5th percentile from a t-distribution with $df = 50 - 2 = 48$ degrees of freedom is $t_{48,0.975} = 2.011$.
- Because $|13.4| > 2.011$, we reject $H_0 : \beta_1 = 0$ in favor of the alternative that $\beta_1 \neq 0$.

The t-test

Table: Decision-Making Procedures for Testing $H_0 : \beta_1 = d$

Alternative Hypothesis (H_a)	Procedure: Reject H_0 in favor of H_a if
$\beta_1 > d$	$t - ratio > t_{n-2,1-\alpha}$.
$\beta_1 < d$	$t - ratio < -t_{n-2,1-\alpha}$.
$\beta_1 \neq d$	$\lvert t - ratio \rvert > t_{n-2,1-\alpha/2}$.

Notes: The significance level is α. Here, $t_{n-2,1-\alpha}$ is the (1-α)th percentile from the t -distribution using $df = n - 2$ degrees of freedom.

Table: Probability Values for Testing $H_0 : \beta_1 = d$

Alternative Hypothesis (H_a)	$\beta_1 > d$	$\beta_1 < d$	$\beta_1 \neq d$
p-value	$\Pr(t_{n-2} > t - ratio)$	$\Pr(t_{n-2} < t - ratio)$	$\Pr(\lvert t_{n-2}\rvert > \lvert t - ratio\rvert)$

Interpretations of the t-ratio

- If $r = 0$, then $b_1 = 0$ and $t(b_1) = 0$. No correlation, no relationship.
- The correlation between y and x, $r = r(x, y)$ is the same as between y and $\widehat{y}$, say $r(x, \widehat{y})$.
  - Because r is location and scale invariant (assuming that $\widehat{y} = b_0 + b_1 x$ and $b_1 > 0$.
- It turns out (Ex 2.13) that $R^2 = r^2$.
- Further, one can check that (Ex 2.16)

$$t(b_1) = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}}.$$

Confidence Intervals

- b_1 is our point estimator of the true, unknown slope β_1.
- How reliable is it? The standard error gives us some idea.
  - $(b_1 - \beta_1)/se(b_1)$ follows a t-distribution with $n - 2$ degrees of freedom.
  - From this, we have a $100(1 - \alpha)\%$ confidence interval for the slope β_1

$$b_1 \pm t_{n-2,1-\alpha/2}\, se(b_1).$$

- Wisconsin lottery sales example:
  - An approximate 95% confidence interval for the slope is

$$0.647 \pm (2.011)(.0488) = (0.549, 0.745).$$

  - An approximate 90% confidence interval for the slope is

$$0.647 \pm (1.677)(.0488) = (0.565, 0.729).$$

Prediction Intervals

- Prediction is an important task for actuaries
- Suppose that I know that the population of a zip code is $x^* = 10,000$, what is my prediction of sales? How good is it?
- We want to predict $y^* = \beta_0 + \beta_1 x^* + \varepsilon$
- Our point prediction is $\widehat{y}^* = b_0 + b_1 x^*$
- The prediction error is

$$\underbrace{y^* - \widehat{y}^*}_{\text{prediction error}} = \underbrace{\beta_0 - b_0 + (\beta_1 - b_1)\, x^*}_{\text{error in estimating the regression line at } x^*} + \underbrace{\varepsilon^*}_{\text{deviation of the additional response from its mean}}$$

Prediction Intervals

- It can be shown that the standard error of the prediction is

$$se(pred) = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{(n-1)s_x^2}}.$$

- As x^* becomes farther from $\overline{x}$, $se(pred)$ increases
- Thus, a $100(1-\alpha)\%$ prediction interval at x^* is

$$\widehat{y}^* \pm t_{n-2,1-\alpha/2}\ se(pred)$$

- Wisconsin lottery sales example:
  - Point prediction - $\widehat{y}^*$= 470.8 + 0.647 (10000) = 6,941.
  - The standard error of this prediction is

$$se(pred) = 3,792\sqrt{1 + \frac{1}{50} + \frac{(10,000 - 9,311)^2}{(50-1)(11,098)^2}} = 3,836.$$

  - The 95% prediction interval is

$$6,941 \pm (2.011)(3,836) = 6,941 \pm 7,710 = (-769, 14,651).$$

Diagnostic Checking

- **Diagnostic Checking.** Process of matching the modeling assumptions with the data and use any mismatch to specify a better model.
  - Like when you go to a doctor and he or she performs diagnostic routines to check your health
  - We will begin with the error representation and use residuals as approximations of the errors/disturbances
- **Residual Analysis.** If the residuals are related to a variable or display any other recognizable pattern, then we should be able to take advantage of this information and improve our model specification.

Model Misspecification Issues

- **Lack of Independence**. There may exist relationships among the deviations $\{\varepsilon_i\}$ so that they are no longer independent.
- **Heteroscedasticity**. Assumption E3 that indicates that all observations have a common (although unknown) variability, known as *homoscedasticity*. *Heteroscedascity* is the term used when the variability varies by observation.
- **Relationships between Model Deviations and Explanatory Variables**. If an explanatory variable has the ability to help explain the deviation ε, the one should be able to use this information to better predict y.
- **Nonnormal Distributions**. If the distribution of the deviation represents a serious departure from approximate normality, then the usual inference procedures are no longer valid.
- **Unusual Points**. Individual observations may have a large effect on the regression model fit, meaning that the results may be sensitive to the impact to behavior of a single observation.

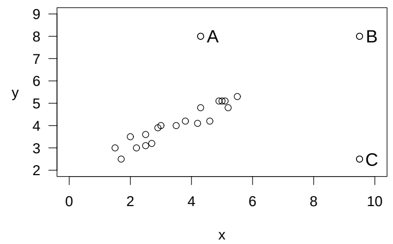
Unusual Points

- Because regression estimates are weighted averages, some observations are more important than others.
- An observation that is unusual in the vertical direction is called an *outlier*.
- To detect outliers, we will use standardized residuals, essentially residuals divided by s
- An observation that is unusual in the horizontal directional is called a *high leverage point*.
- An observation may be both an outlier and a high leverage point.

The Effect of Outliers and High Leverage Points

Table: 19 Base Points Plus Three Types of Unusual Observations

Variables	19 Base Points										A	B	C
x	1.5	1.7	2.0	2.2	2.5	2.5	2.7	2.9	3.0	3.5	3.4	9.5	9.5
y	3.0	2.5	3.5	3.0	3.1	3.6	3.2	3.9	4.0	4.0	8.0	8.0	2.5
x	3.8	4.2	4.3	4.6	4.0	5.1	5.1	5.2	5.5				
y	4.2	4.1	4.8	4.2	5.1	5.1	5.1	4.8	5.3				



The Effect of Outliers and High Leverage Points

Table: Results from Four Regressions

Data	b_0	b_1	s	$R^2(\%)$	$t(b_1)$
19 Base Points	1.869	0.611	0.288	89.0	11.71
19 Base Points + A	1.750	0.693	0.846	53.7	4.57
19 Base Points + B	1.775	0.640	0.285	94.7	18.01
19 Base Points + C	3.356	0.155	0.865	10.3	1.44

- The 19 base points show a high R^2, $s = 0.29$.
- With outlier A, the R^2 drops from 89% to 53.7%.
- An outlier, "unusual in the y-value," depends on the x-value.
- With B, the regression line provides a *better* fit.
- Point B is not an outlier, but it is a high leverage point.
- Point C is an outlier and a high leverage point. The R^2 coefficient drops from 89% to 10%.
- Many do not believe that 1 point in 20 can have such a dramatic effect on the regression fit.

Example. Wisconsin Lottery Sales

Table: Regression Results with and without Kenosha

Data	b_0	b_1	s	$R^2(\%)$	$t(b_1)$
With Kenosha	469.7	0.647	3,792	78.5	13.26
Without Kenosha	-43.5	0.662	2,728	88.3	18.82

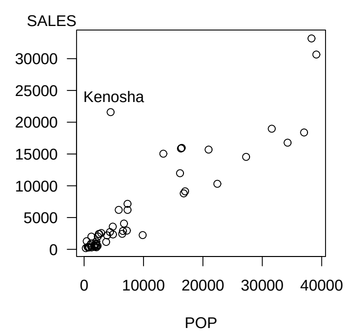


Figure: Scatter plot of SALES versus POP, with the outlier corresponding to Kenosha marked.

Example. Wisconsin Lottery Sales

One point can change the appearance of the whole distribution.

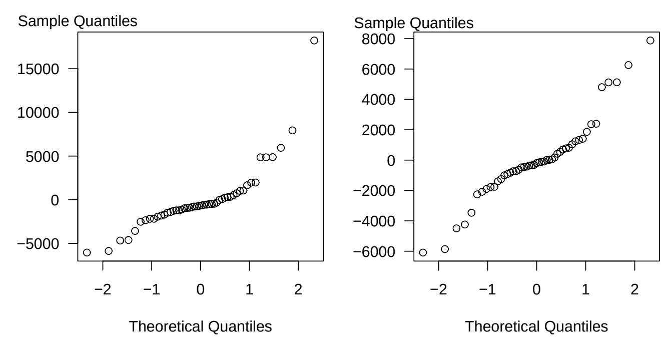


Figure: *qq* Plots of Wisconsin Lottery Residuals. The left-hand panel is based on all 50 points. The right-hand panel is based on 49 points, residuals from a regression after removing Kenosha.

Data

- Consider monthly returns over the five year period from January, 1986 to December, 1990, inclusive.
- y = security returns from the Lincoln National Insurance Corporation as the dependent variable
- x = market returns from the index of the Standard & Poor's 500 Index.

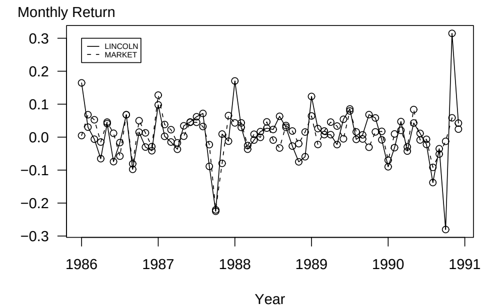
Table: Summary Statistics of 60 Monthly Observations

	Mean	Median	Standard Deviation	Minimum	Maximum
LINCOLN	0.0051	0.0075	0.0859	-0.2803	0.3147
MARKET	0.0074	0.0142	0.0525	-0.2205	0.1275

Source: Center for Research on Security Prices, University of Chicago

Data

- Scatter plots of the returns versus time are called *time series plots*.
- A quick glance at the horizontal axis reveals that this unusual point is in October, 1987, the time of the well-known market crash.
- We also see two outliers in 1990



Regression

- The estimated regression is $\widehat{LINCOLN} = -0.00214 + 0.973 MARKET$.
- The resulting estimated standard error, $s = 0.0696$ is lower than the standard deviation of Lincoln's returns, $s_y = 0.0859$.
- Further, $t(b_1) = 5.64$, which is significantly large.
- One disappointing aspect is that the statistic $R^2 = 35.4\%$

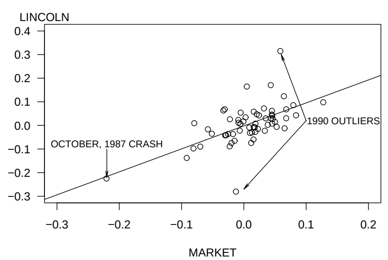


Figure: Scatterplot of Lincoln's return versus the S&P 500 Index return. The regression line is superimposed, enabling us to identify the market crash and two outliers.

Sensitivity Analysis

- Without the market crash, the estimated regression is

$$\widehat{LINCOLN} = -0.00181 + 0.956 MARKET,$$

with $R^2 = 26.4\%$, $t(b_1) = 4.52$, $s = 0.0702$ and $s_y = 0.0811$.
- The important point is that the R^2 *decreased* when omitting this unusual point.
- The outliers were due to some unfounded rumors in the market that made Lincoln's price drop one month and subsequently recover.
- Should the unusual points be left in the analysis? Tough question that does not have a right or wrong answer. Your only mistake would be not paying attention to these points!