

Statistics for IT Managers

95-796, Fall 2007

Module 3: Simple and Multiple Regression (4 lectures)

Reading: Statistics for Business and Economics, Ch. 10-11

Why regression?

We want to model how two (or more) variables are related.

Drinking more coffee increases productivity.

A proper diet lowers risk of heart disease.

Employees with more experience earn higher salaries.

We want to predict the value of one variable, given the other(s).

Given a person's demographic characteristics, how often do we expect them to utilize our website?

How fast do we expect computer CPUs to be in 2010?

We want to test whether there is a statistically significant relationship between variables.

Can I conclude that increased advertising expenditures increase sales?

Can I conclude that worker satisfaction increased over time?

Linear regression

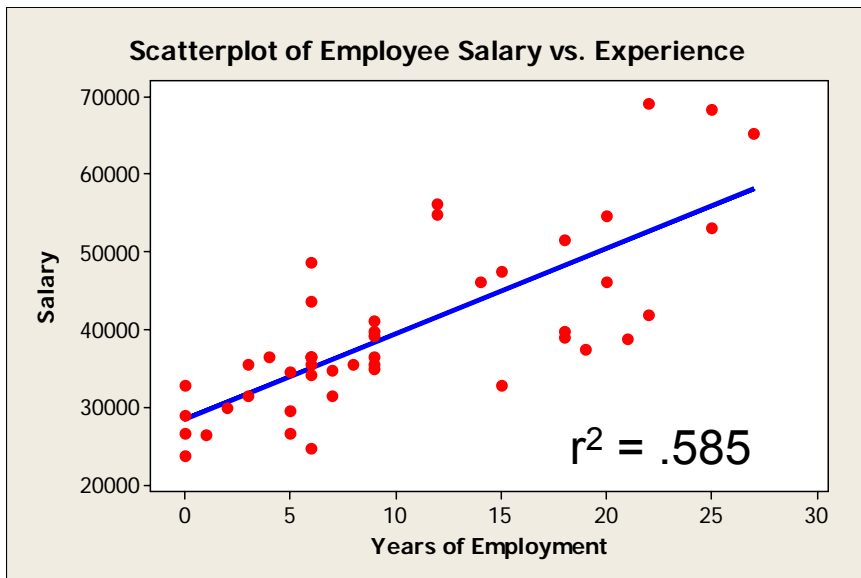
We will focus here on modeling linear relationships between two variables.

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

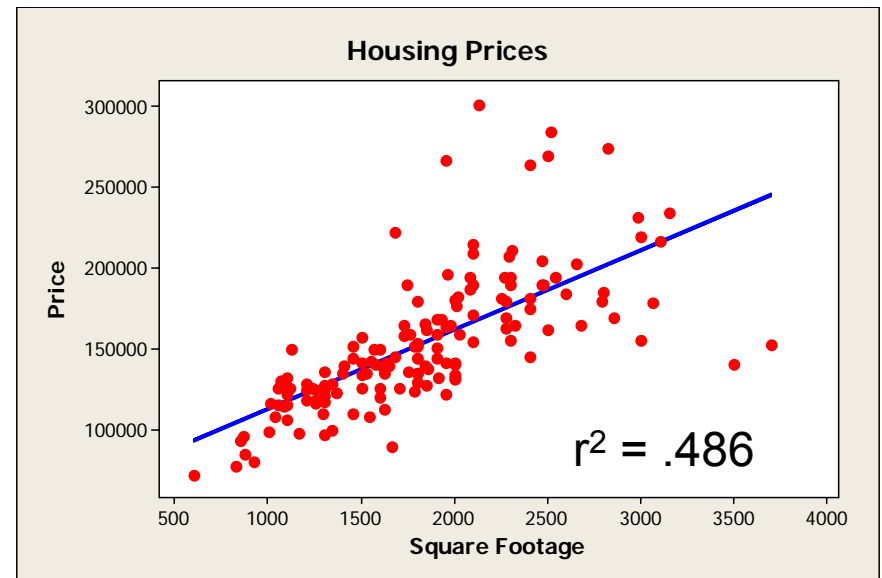
“dependent variable” “independent variable”

Employee salary tends to increase with experience (years of employment)

The price of a house tends to increase with its size (area in square feet)



$$\text{Salary} = 28394 + 1107 \text{ Years}$$



$$\text{Price} = 63745 + 49.4 \text{ Area}$$

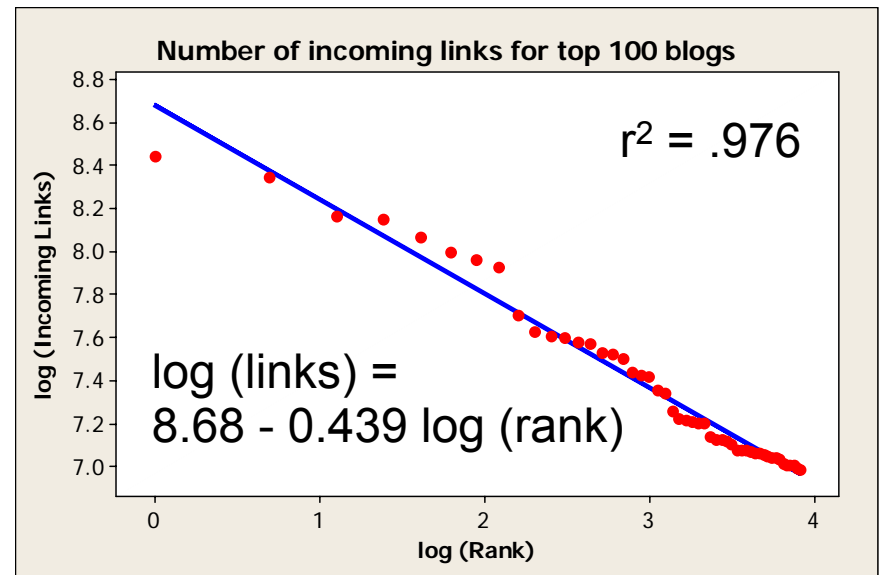
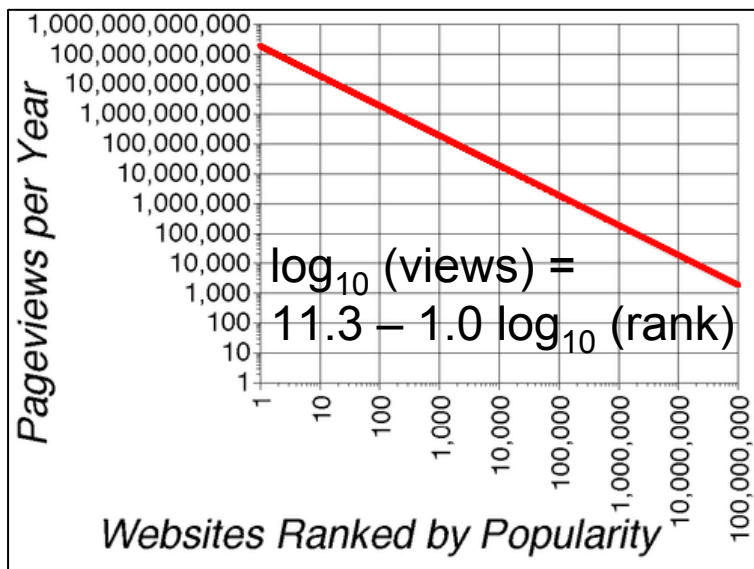
Linear regression

Linear regression can also be used to model certain non-linear relationships between variables, such as exponential growth and power laws.

Internet data tends to follow a power law distribution:
 y is proportional to x^{β_1} , so $(\log y) = \beta_0 + \beta_1 (\log x)$.

The N th most popular website gets about $1/N$ the hits of the most popular.

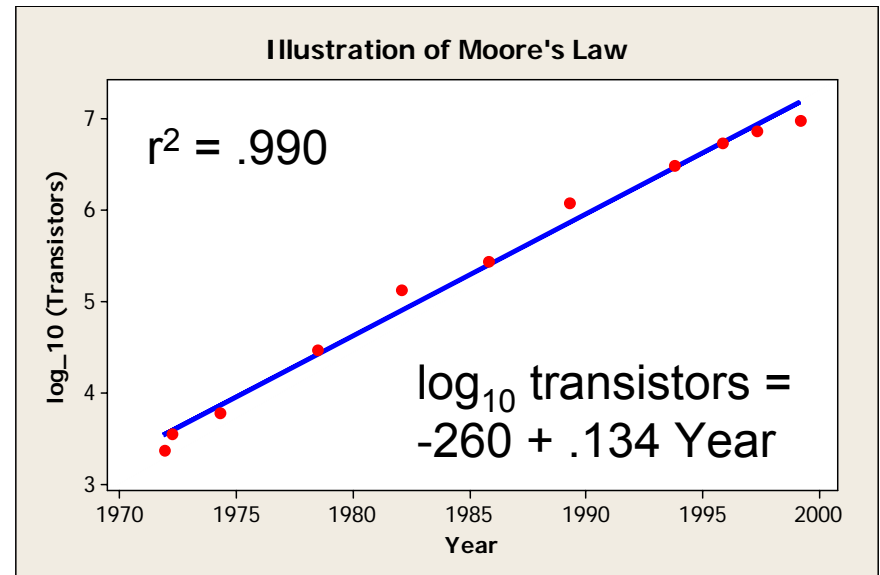
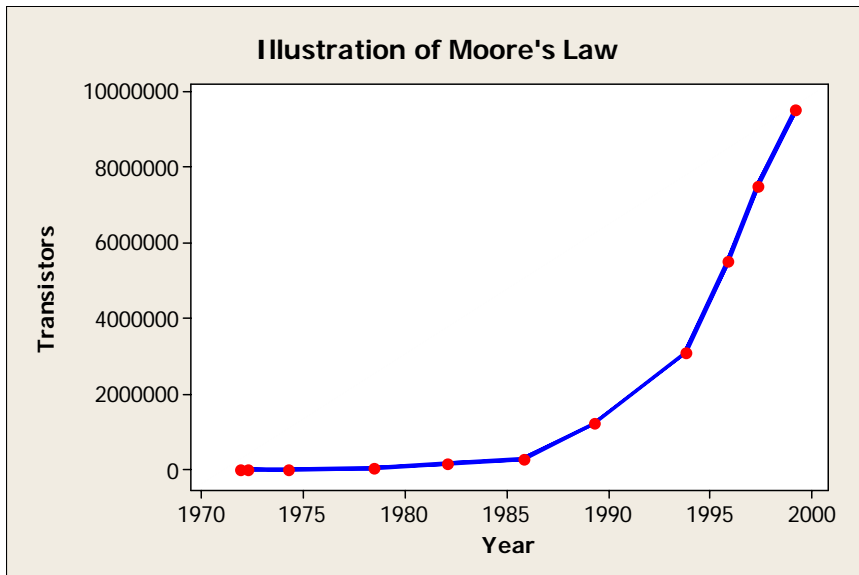
Number of links to a blog decreases as a power law with its popularity rank.



Linear regression

Linear regression can also be used to model certain non-linear relationships between variables, such as exponential growth and power laws.

Moore's Law states that various quantities in computer technology, such as the number of transistors on a computer chip, will increase exponentially with time: y is proportional to $\exp(\beta_1 x)$, so $(\log y) = \beta_0 + \beta_1 x$.

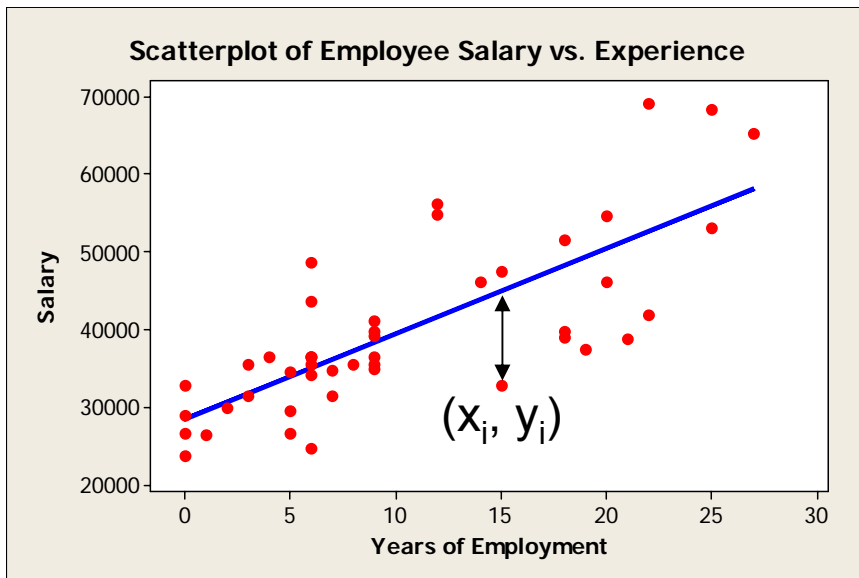


“Least Squares” Linear Regression

Once we have the best-fitting line, we can use it for prediction or hypothesis testing.

“When years = 10, we expect employee salary to be _____”

“We can conclude that $\beta_1 > 0$, so salary increases with experience.”



$$\text{Salary} = 28394 + 1107 \text{ Years}$$

How to obtain the best fitting line from a set of datapoints (x_i, y_i) ?

Answer: choose estimates of β_0 and β_1 to minimize the sum of squared errors.

$$\text{Regression line: } y = \hat{\beta}_0 + \hat{\beta}_1 x$$

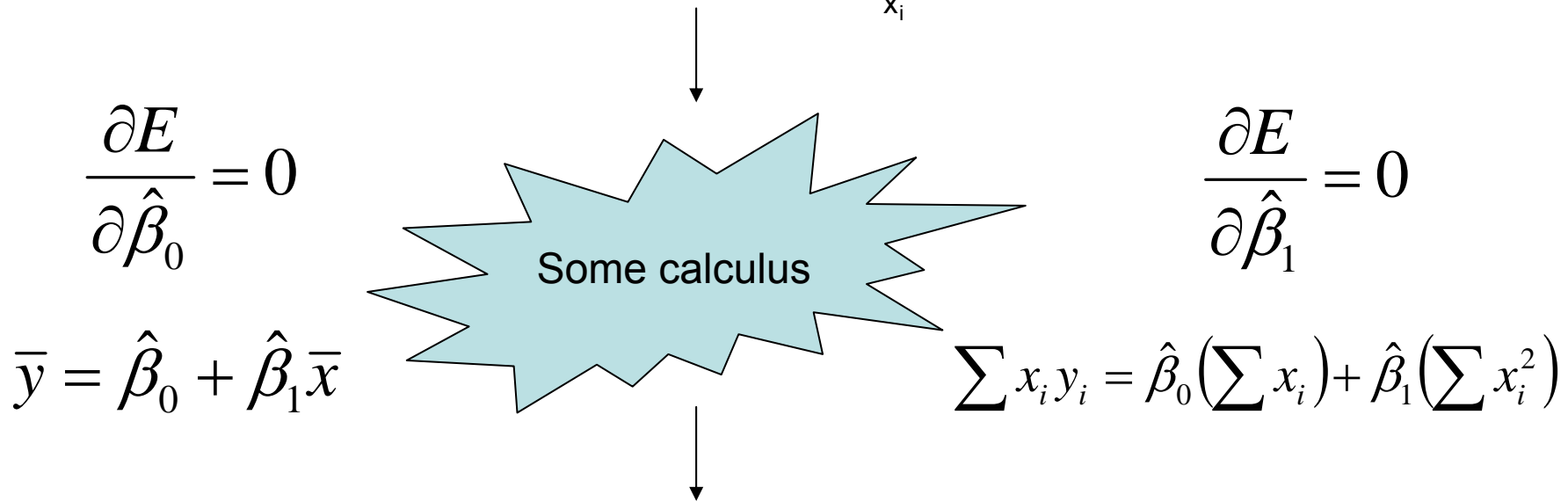
$$\text{Estimate of } y_i \text{ for } x_i: \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\text{Error for } x_i: y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$\text{Minimize: } \sum_{x_i} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)^2$$

“Least Squares” Linear Regression

Find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the squared error $\sum_{x_i} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$ is minimized.



1. Compute the sample means \bar{x} and \bar{y} .
2. Compute the sum of squares $ss_{xx} = \sum (x_i - \bar{x})^2$
3. Compute the sum of squares $ss_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

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

“Least Squares” Linear Regression

Employee 1 makes \$40K after 5 years
Employee 2 makes \$30K after 1 year
Employee 3 makes \$35K after 1 year
Employee 4 makes \$45K after 9 years

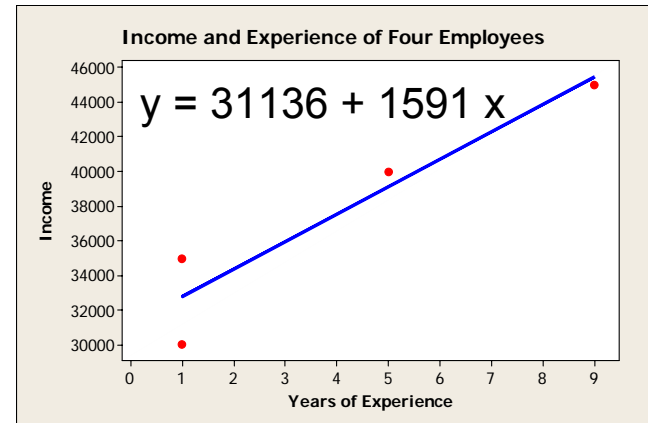
$$\bar{x} = 4, \bar{y} = 37500$$

$$SS_{xx} = 44$$

$$SS_{xy} = 70000$$

$$\hat{\beta}_1 = \frac{70000}{44} \approx 1591$$

$$\hat{\beta}_0 = 37500 - 4\hat{\beta}_1 \approx 31136$$



1. Compute the sample means \bar{x} and \bar{y} .
2. Compute the sum of squares $ss_{xx} = \sum (x_i - \bar{x})^2$
3. Compute the sum of squares $ss_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$

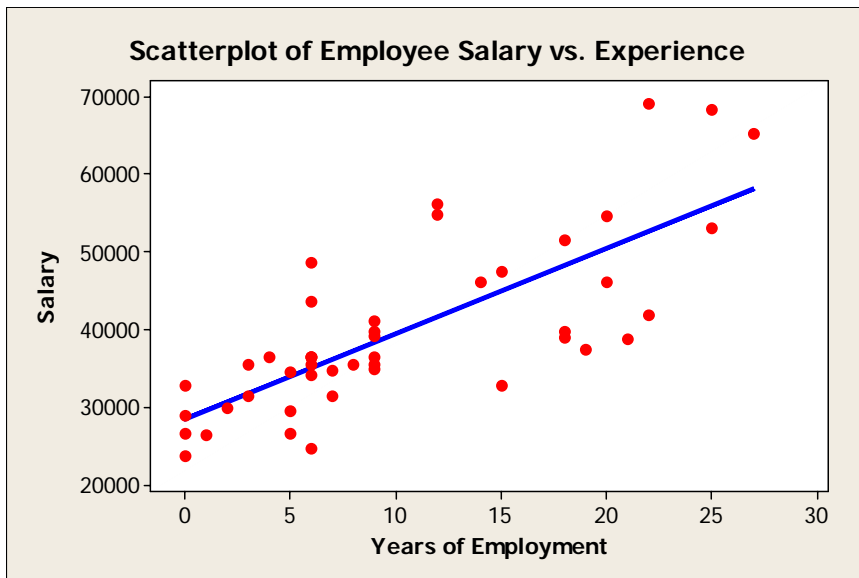
$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

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

Model assumptions

We assume that the relationship of y_i to x_i can be approximated by a linear equation.

However, the relationship between y_i and x_i is not perfect, e.g. an employee's salary cannot be completely explained by their amount of experience.



$$\text{Salary} = 28394 + 1107 \text{ Years}$$

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

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{Deterministic component}} + \underbrace{\varepsilon_i}_{\text{Random component}}$$

Assumption: The random error for each datapoint is drawn independently from a normal distribution with mean 0 and standard deviation σ : $\varepsilon_i \sim N(0, \sigma)$.

We can estimate σ^2 , and use our estimate to perform inference.

Model assumptions

We assume that the relationship of y_i to x_i can be approximated by a linear equation.

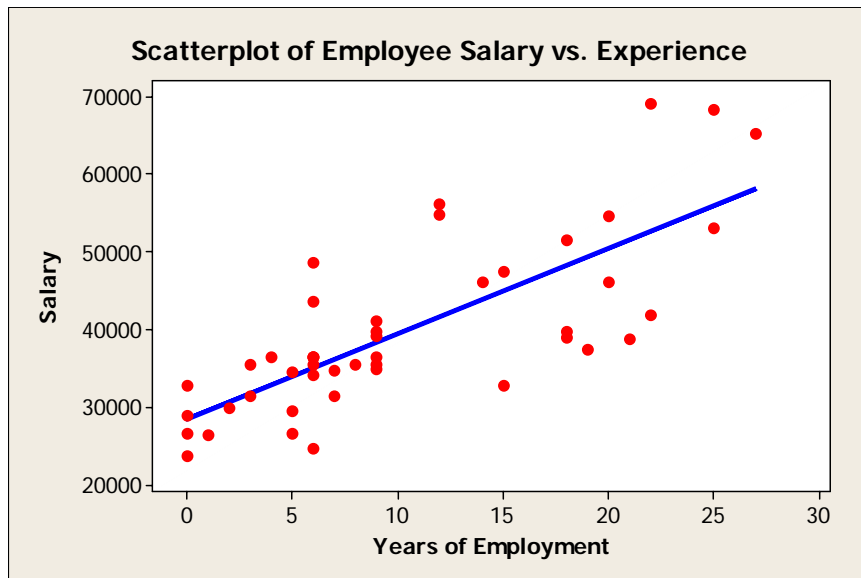
Best estimate of σ^2 (called s^2 or MSE):

$$s^2 = \frac{\sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{N - 2}$$

← Sum of squared errors
← Degrees of freedom

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

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{Deterministic component}} + \underbrace{\varepsilon_i}_{\text{Random component}}$$



$$\text{Salary} = 28394 + 1107 \text{ Years}$$

Assumption: The random error for each datapoint is drawn independently from a normal distribution with mean 0 and standard deviation σ : $\varepsilon_i \sim N(0, \sigma)$.

We can estimate σ^2 , and use our estimate to perform inference.

Model assumptions

We assume that the relationship of y_i to x_i can be approximated by a linear equation.

Best estimate of σ^2 (called s^2 or MSE):

$$s^2 = \frac{\sum_{x_i} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{N - 2}$$

← Sum of squared errors
← Degrees of freedom

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

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{Deterministic component}} + \underbrace{\varepsilon_i}_{\text{Random component}}$$

Employee 1 makes \$40K after 5 years
 Employee 2 makes \$30K after 1 year
 Employee 3 makes \$35K after 1 year
 Employee 4 makes \$45K after 9 years

$$\begin{aligned} (40000 - 39091)^2 &= 826281 \\ (30000 - 32727)^2 &= 7436529 \\ (35000 - 32727)^2 &= 5166529 \\ (45000 - 45455)^2 &= 207025 \\ \hline &13636364 \end{aligned}$$

$$y = 31136 + 1591 x$$

$$s^2 = 13636364 / (4 - 2) = 6818182$$

$$s = \sqrt{s^2} \approx 2611$$

s measures the “goodness of fit” for our regression line.

Inference for model parameters

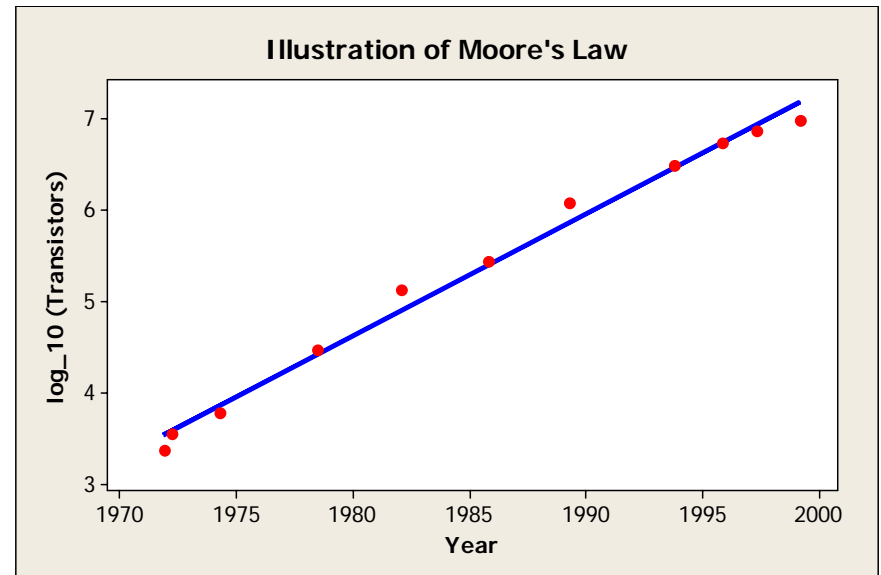
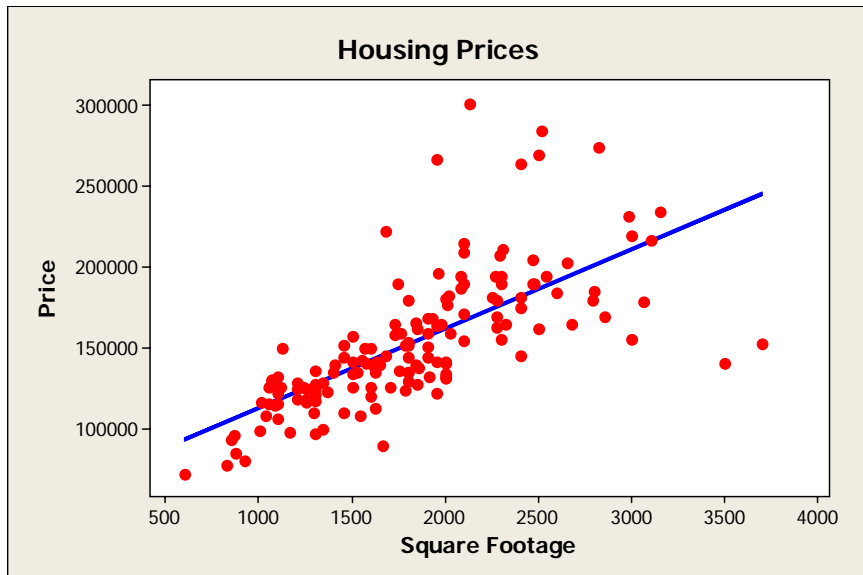
Are $\hat{\beta}_0$ and $\hat{\beta}_1$ good estimates of the true model parameters β_0 and β_1 ?

s is a measure of the “average” distance of datapoints from our regression line. Higher s means more random error, and thus more uncertainty in our model.

Standard deviation of our estimate for β_1 :

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}, \text{ where } SS_{xx} = \sum (x_i - \bar{x})^2$$

Note that ss_{xx} grows linearly with the number of datapoints N .



Inference for model parameters

Are $\hat{\beta}_0$ and $\hat{\beta}_1$ good estimates of the true model parameters β_0 and β_1 ?

s is a measure of the “average” distance of datapoints from our regression line. Higher s means more random error, and thus more uncertainty in our model.

Standard deviation of our estimate for β_1 :

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{ss_{xx}}}, \text{ where } ss_{xx} = \sum (x_i - \bar{x})^2$$

Note that ss_{xx} grows linearly with the number of datapoints N .

95% confidence interval for β_1 : $\hat{\beta}_1 \pm t_c s_{\hat{\beta}_1}$

Note: t_c has $N - 2$ degrees of freedom.

Hypothesis test for $\beta_1 \neq b$: $t\text{-score} = \frac{\hat{\beta}_1 - b}{s_{\hat{\beta}_1}}$

For the “four employees” example, our regression line was $y = 31136 + 1591x$. We also calculated $s = 2611$ and $ss_{xx} = 44$, giving $s_{\beta_1} = 2611 / \sqrt{44} \approx 394$.

95% CI for β_1 : $1591 \pm (4.303)(394) = [-104, 3286]$

Inference for model parameters

Are $\hat{\beta}_0$ and $\hat{\beta}_1$ good estimates of the true model parameters β_0 and β_1 ?

s is a measure of the “average” distance of datapoints from our regression line. Higher s means more random error, and thus more uncertainty in our model.

Standard deviation of our estimate for β_1 :

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{ss_{xx}}}, \text{ where } ss_{xx} = \sum (x_i - \bar{x})^2$$

Note that ss_{xx} grows linearly with the number of datapoints N .

95% confidence interval for β_1 : $\hat{\beta}_1 \pm t_c s_{\hat{\beta}_1}$

Note: t_c has $N - 2$ degrees of freedom.

Hypothesis test for $\beta_1 \neq b$: $t\text{-score} = \frac{\hat{\beta}_1 - b}{s_{\hat{\beta}_1}}$

For the “four employees” example, our regression line was $y = 31136 + 1591x$. We also calculated $s = 2611$ and $ss_{xx} = 44$, giving $s_{\hat{\beta}_1} = 2611 / \sqrt{44} \approx 394$.

Hypothesis test for $\beta_1 \neq 0$: $t\text{-score} = 1591 / 394 = 4.04$, cannot reject H_0 .

Coefficient of determination

An overall measure of how well our linear model fits the data.

y is different across observations for two reasons: random error, and because the observations have different x values.

The coefficient of determination, r^2 , measures the proportion of the variation in y that can be explained by the variation in x .

Total variation in y : $ss_{yy} = \sum (y_i - \bar{y})^2$

Variation in y due to random error: $SSE = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$

Variation in y due to variation in x : $ss_{yy} - SSE$

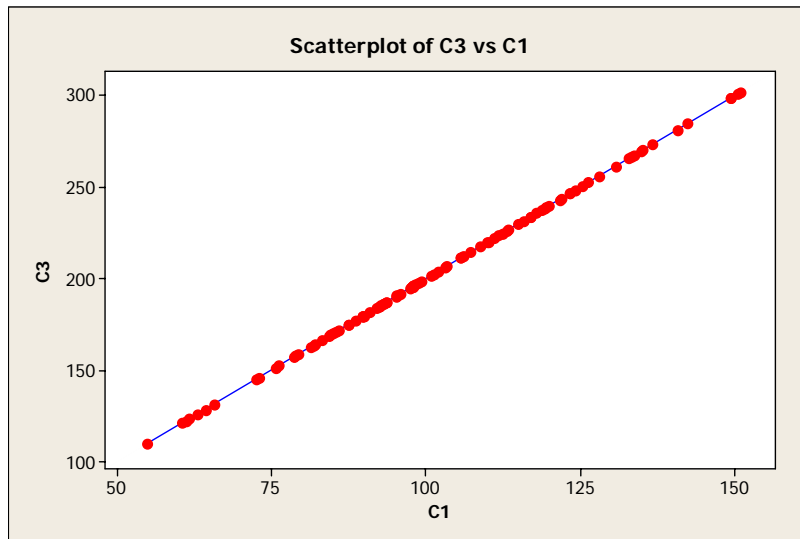
Thus $r^2 = (ss_{yy} - SSE) / ss_{yy} = 1 - (SSE / ss_{yy})$.

Coefficient of determination

An overall measure of how well our linear model fits the data.

y is different across observations for two reasons: random error, and because the observations have different x values.

The coefficient of determination, r^2 , measures the proportion of the variation in y that can be explained by the variation in x .



$r^2 = 1$ means a perfect fit:

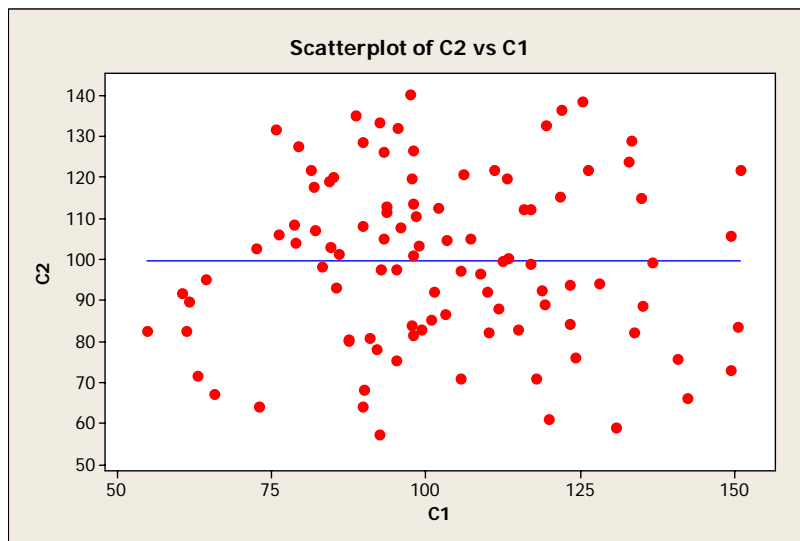
All of the variation in y is due to variation in x .

Coefficient of determination

An overall measure of how well our linear model fits the data.

y is different across observations for two reasons: random error, and because the observations have different x values.

The coefficient of determination, r^2 , measures the proportion of the variation in y that can be explained by the variation in x .



$r^2 = 0$ means no relationship between x and y :

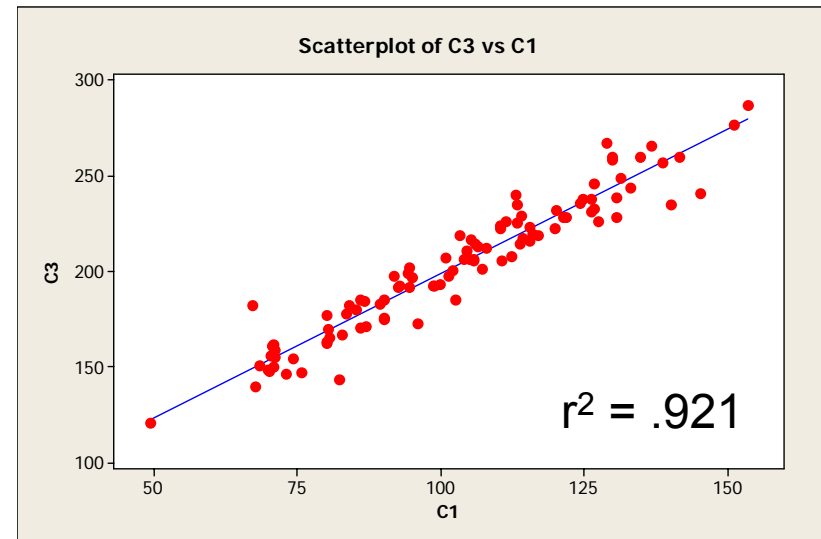
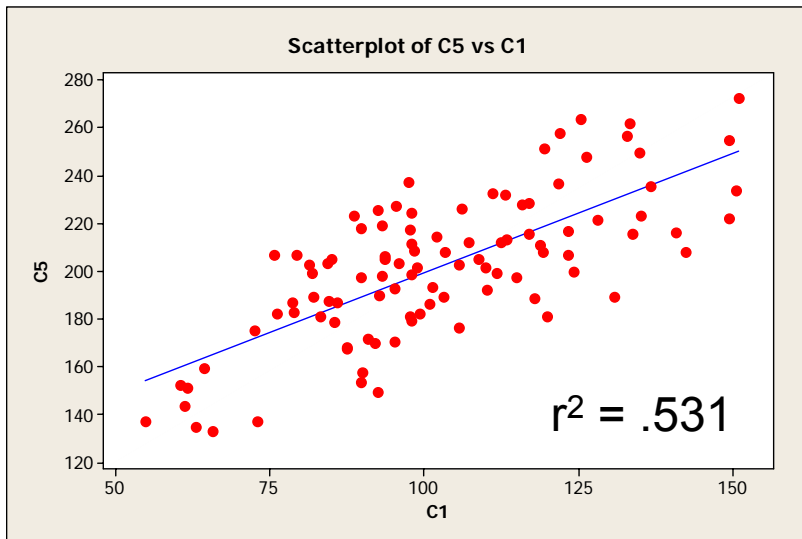
All of the variation in y is due to random error, and knowing x gives us no information about y .

Coefficient of determination

An overall measure of how well our linear model fits the data.

y is different across observations for two reasons: random error, and because the observations have different x values.

The coefficient of determination, r^2 , measures the proportion of the variation in y that can be explained by the variation in x .

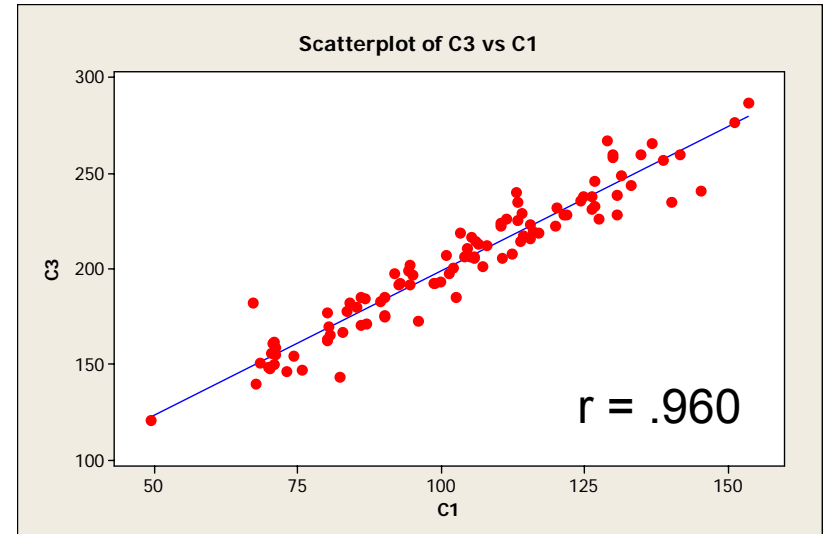
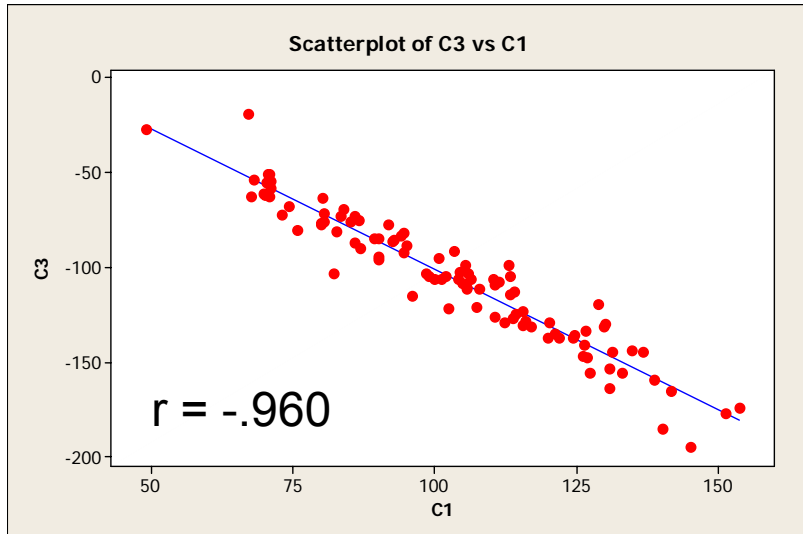


Coefficient of correlation

For the case of simple regression (one independent variable), we can also define the coefficient of correlation r .

Positive r means that y increases with increasing x (“positively correlated”)
Negative r means that y decreases with increasing x (“negatively correlated”)

r can take values between -1 and 1: $r = -1$ is perfect negative correlation, and $r = +1$ is perfect positive correlation.



Coefficient of correlation

For the case of simple regression (one independent variable), we can also define the coefficient of correlation r .

Positive r means that y increases with increasing x (“positively correlated”)
Negative r means that y decreases with increasing x (“negatively correlated”)

r can take values between -1 and 1 : $r = -1$ is perfect negative correlation, and $r = +1$ is perfect positive correlation.

To calculate r , compute the square root of r^2 , and give it the same sign as β_1 .

Another way of calculating r is: $r = ss_{xy} / \sqrt{ss_{xx} ss_{yy}}$

“Four employees” example: $ss_{xx} = 44$, $ss_{xy} = 70000$, $ss_{yy} = 125000000$.

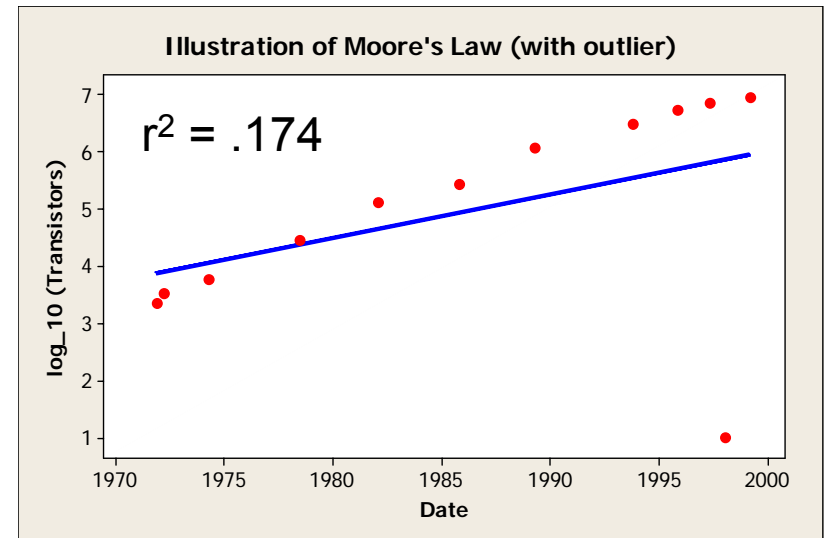
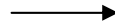
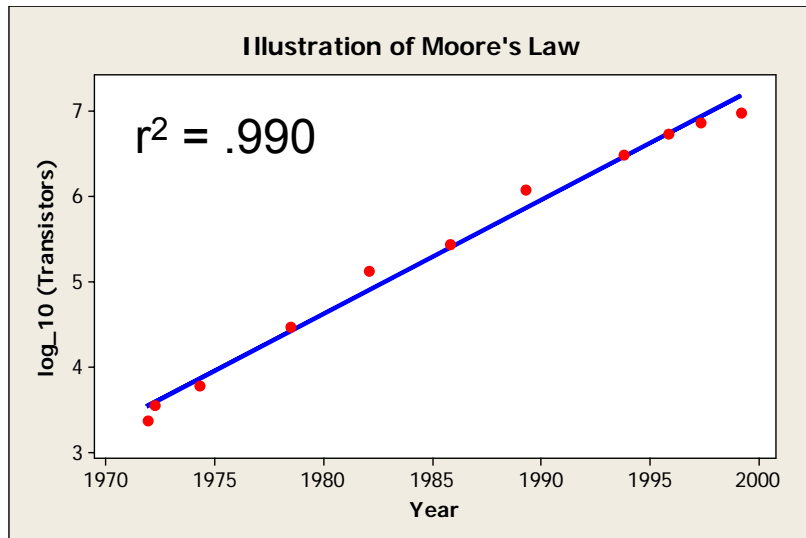
$$r \approx 70000 / 74162 \approx .944$$

$$r^2 \approx .891$$

Note: high correlation between x and y does not mean that x causes y .

Dealing with outliers

Outliers can have a large impact on the regression line, especially when N is small.



Minitab gives a list of all points which may be outliers, including:
Points with large residuals (far from the regression line)
Points with very large or small x values (high-impact points)

In this case, Minitab finds an unusual observation with $x = 1999$, $y = 1$,
Fit (estimated y) = 5.88, and standardized residual $t = -3.15$.

Using Minitab

Regression Analysis: Income versus Years

The regression equation is

$$\text{Income} = 31136 + 1591 \text{ Years}$$

Predictor	Coef	SE Coef	T	P
Constant	31136	2045	15.22	0.004
Years	1590.9	393.6	4.04	0.056

$$S = 2611.16 \quad R\text{-Sq} = 89.1\% \quad R\text{-Sq}(\text{adj}) = 83.6\%$$



Minitab gives you:

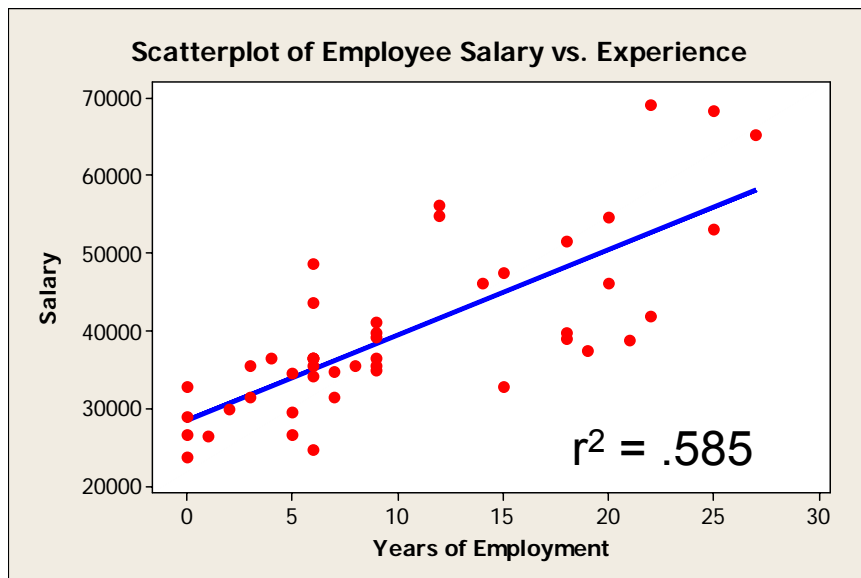
- The regression equation.
- The coefficient of determination r^2 .
- The estimated standard deviation s .
- Standard deviation, t-score, and p-values for estimates of β_0 and β_1 .
- Analysis of variance (don't worry about this for now).
- A list of the residuals for unusual observations (or all observations).

Multiple regression

We can extend linear regression to model a linear relationship between the dependent variable y and multiple independent variables x_1, x_2, \dots, x_k .

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

58.5% of the variation in employees' salary can be explained by experience...



$$\text{Salary} = 28394 + 1107 \text{ Years}$$

What other factors might also explain differences in salary?

x_1 = years of employment
 x_2 = years of post-HS education
 x_3 = 1 if female, 0 if male
 x_4 = # of employees supervised
...and many other possibilities!

$$\text{Salary} = 24655 + 646 x_1 + 1615 x_2 - 1295 x_3 + 165 x_4$$

How to interpret $\beta_0 \dots \beta_4$?

Multiple regression

We can extend linear regression to model a linear relationship between the dependent variable y and multiple independent variables x_1, x_2, \dots, x_k .

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

What is the expected salary of a female employee with 4 years of education and 2 years of employment, who supervises a team of 3 employees?

$$\text{Salary} = 24655 + 646(2) + 1615(4) - 1295(1) + 165(3) = \$31,607$$

How do we expect her salary to increase with years of employment, assuming that all other variables are constant?

$$\text{Salary} = 30315 + 646 x$$

What other factors might also explain differences in salary?

x_1 = years of employment
 x_2 = years of post-HS education
 x_3 = 1 if female, 0 if male
 x_4 = # of employees supervised
...and many other possibilities!

$$\text{Salary} = 24655 + 646 x_1 + 1615 x_2 - 1295 x_3 + 165 x_4$$

How to interpret $\beta_0 \dots \beta_4$?

Multiple regression

We can extend linear regression to model a linear relationship between the dependent variable y and multiple independent variables x_1, x_2, \dots, x_k .

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

Why do we only expect salary to increase \$646 per year of employment, rather than \$1107 per year as in the simple regression equation?

Expected salary increases \$646 per year for a constant amount of education, number of employees supervised, etc. However, more experienced employees also tend to have more education and to be managers of larger groups.

What other factors might also explain differences in salary?

x_1 = years of employment
 x_2 = years of post-HS education
 x_3 = 1 if female, 0 if male
 x_4 = # of employees supervised
...and many other possibilities!

$$\text{Salary} = 24655 + 646 x_1 + 1615 x_2 - 1295 x_3 + 165 x_4$$

How to interpret $\beta_0 \dots \beta_4$?

“Least Squares” Multiple Regression

How to obtain the best fitting multiple regression model, $y = \beta_0 + \sum \beta_j x_j$, from a set of datapoints?

As before, we choose estimates of β_0 and each β_j to minimize the sum of squared errors.

$$\text{Regression model: } y = \hat{\beta}_0 + \sum_j \hat{\beta}_j x_j$$

$$\text{Estimate of } y \text{ for the } i\text{th datapoint: } \hat{y}_i = \hat{\beta}_0 + \sum_j \hat{\beta}_j x_{ij}$$

$(x_{i1}, x_{i2}, \dots, x_{ik}, y_i)$

$$\text{Thus we want to minimize: } \sum_i (y_i - \hat{y}_i)^2 = \sum_i \left(y_i - \left(\hat{\beta}_0 + \sum_j \hat{\beta}_j x_{ij} \right) \right)^2$$

Solution: Use Minitab to obtain the regression coefficients.

Model assumptions

We assume that the relationship between the dependent variable y and the independent variables $x_1 \dots x_k$ can be approximated by a linear equation.

$$y = \beta_0 + \sum_j \beta_j x_j \quad \longrightarrow \quad y_i = \underbrace{\beta_0 + \sum_j \beta_j x_{ij}}_{\text{Deterministic component}} + \underbrace{\varepsilon_i}_{\text{Random component}}$$

Assumption: The random error for each datapoint is drawn independently from a normal distribution with mean 0 and standard deviation σ : $\varepsilon_i \sim N(0, \sigma)$.

Best estimate of σ^2 (called s^2 or MSE):

$$s^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{N - (k + 1)} = \frac{\sum_i \left(y_i - \left(\hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_{ij} \right) \right)^2}{N - (k + 1)}$$

Sum of squared errors

Degrees of freedom

Model assumptions

“Four employees” example, using experience and gender as independent variables:

Employee 1 (female) makes \$40K after 5 years	$(40000 - 37500)^2 = 6,250,000$
Employee 2 (female) makes \$30K after 1 year	$(30000 - 32500)^2 = 6,250,000$
Employee 3 (male) makes \$35K after 1 year	$(35000 - 35000)^2 = 0$
Employee 4 (male) makes \$45K after 9 years	$(45000 - 45000)^2 = 0$

$$y = 33750 + 1250 (\text{Years}) - 2500 (\text{if female}) \quad \underline{\quad\quad\quad} \quad 12,500,000$$

$$s^2 = 12,500,000 / (4 - 3) = 12,500,000$$

$$s = \sqrt{s^2} \approx 3536$$

Best estimate of σ^2 (called s^2 or MSE):

$$s^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{N - (k + 1)} = \frac{\sum_i \left(y_i - \left(\hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_{ij} \right) \right)^2}{N - (k + 1)}$$

← Sum of squared errors

← Degrees of freedom

Coefficient of determination

As in the simple regression case, the coefficient of determination R^2 is an overall measure of how well our multiple regression model fits the data.

R^2 measures the proportion of the variation in the dependent variable y that can be explained by the variation in the independent variables $x_1 \dots x_k$.

Total variation in y : $ss_{yy} = \sum (y_i - \bar{y})^2$

Variation in y due to random error: $SSE = \sum_i (y_i - \hat{y}_i)^2 = \sum_i \left(y_i - \left(\hat{\beta}_0 + \sum_j \hat{\beta}_j x_{ij} \right) \right)^2$

Variation in y due to variation in $x_1 \dots x_k$: $ss_{yy} - SSE$

Thus $R^2 = (ss_{yy} - SSE) / ss_{yy} = 1 - (SSE / ss_{yy})$.

$R^2 = 1$ means a perfect fit: all the variation in y is due to variation in $x_1 \dots x_k$.
 $R^2 = 0$ means that all the variation in y is due to random error.

Coefficient of determination

As in the simple regression case, the coefficient of determination R^2 is an overall measure of how well our multiple regression model fits the data.

R^2 measures the proportion of the variation in the dependent variable y that can be explained by the variation in the independent variables $x_1 \dots x_k$.

Total variation in y : $ss_{yy} = \sum (y_i - \bar{y})^2$

Variation in y due to random error: $SSE = \sum_i (y_i - \hat{y}_i)^2 = \sum_i \left(y_i - \left(\hat{\beta}_0 + \sum_j \hat{\beta}_j x_{ij} \right) \right)^2$

Variation in y due to variation in $x_1 \dots x_k$: $ss_{yy} - SSE$

Thus $R^2 = (ss_{yy} - SSE) / ss_{yy} = 1 - (SSE / ss_{yy})$.

Four employees example: we calculated $ss_{yy} = 125,000,000$ and $SSE = 12,500,000$, giving $R^2 = 1 - (SSE / ss_{yy}) = .900$.

Inference for model parameters

Are the $\hat{\beta}_j$ good estimates of the true model parameters β_j ?

Minitab computes not only an estimate of each model parameter, but also the standard deviation of each estimate.

95% confidence interval for β_j : $\hat{\beta}_j \pm t_c s_{\hat{\beta}_j}$

t_c has $N - (k + 1)$ dof,
where k is the number of
independent variables.

Hypothesis test for $\beta_j \neq b$: $t\text{-score} = \frac{\hat{\beta}_j - b}{s_{\hat{\beta}_j}}$

Four employees example:

Predictor	Coeff	StDev
Constant	33750	9100
Years	1250	1250
Gender	-2500	8292

95% CI for Gender:

$$-2500 \pm 12.706(8292) = [-107858, 102858]$$

Hypothesis test for Gender $\neq 0$:

$$t\text{-score} = -2500 / 8292 = -.301$$

p-value > 0.2, cannot reject H_0 .

Inference for model parameters

How do we conclude whether the model is useful, i.e. that any of the k independent variables x_j have β_j significantly different from 0?

Answer: we perform a global goodness of fit test (F-test) using Minitab, and conclude that the model is useful if the resulting p-value is less than the significance level α .

Do not just perform hypothesis tests for the individual variables, and conclude that the model is useful if any variable has p-value less than α .

If we were to do this, we would have a problem of multiple hypothesis testing: our expected number of Type I errors (i.e. incorrectly concluding that the model is useful) increases proportional to the number of tests.

For the “four employees” example, Minitab gives us a p-value of .316 for the F-test, so we cannot conclude that the model is useful.

Using Minitab

The regression equation is

$$\text{Salary} = 24655 + 646 \text{ Years} + 1615 \text{ Education} - 1295 \text{ Gender} + 165 \text{ Supervised}$$

Predictor	Coef	SE Coef	T	P
Constant	24655	2113	11.67	0.000
Years	646.1	140.4	4.60	0.000
Education	1615.1	406.9	3.97	0.000
Gender	-1295	1708	-0.76	0.453
Supervised	165.04	95.39	1.73	0.091

Years, Education are significantly associated with increased salary.

S = 5636.93 R-Sq = **76.1%** R-Sq(adj) = 73.7%

76.1% of the observed variation in salary can be explained by differences in years of employment, education, gender, and employees supervised.

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	4141476437	1035369109	32.58	0.000
Residual Error	41	1302774966	31774999		
Total	45	5444251403			

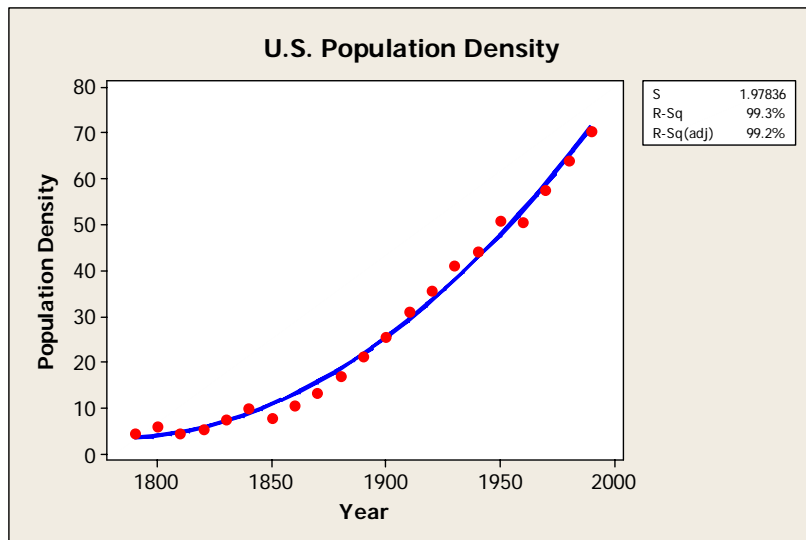
F-test has p-value < .05, so the model is useful.

Higher-order models

If the relationship between the dependent variable y and an independent variable x is not a straight line, we can include quadratic (x^2) or higher terms in the model.

For example, we can examine the growth in U.S. population density over time.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$



$$\text{Density} = 4985 - 5.59 \text{ Year} + 0.00157 \text{ Year}^2$$

Predictor	Coef	SE Coef	T	P
Constant	4985.1	471.5	10.57	0.000
Year	-5.5903	0.4993	-11.20	0.000
Year2	0.0015684	0.0001321	11.87	0.000

$$S = 1.97836 \quad \mathbf{R\text{-}Sq = 99.3\%} \quad R\text{-}Sq(\text{adj}) = 99.2\%$$

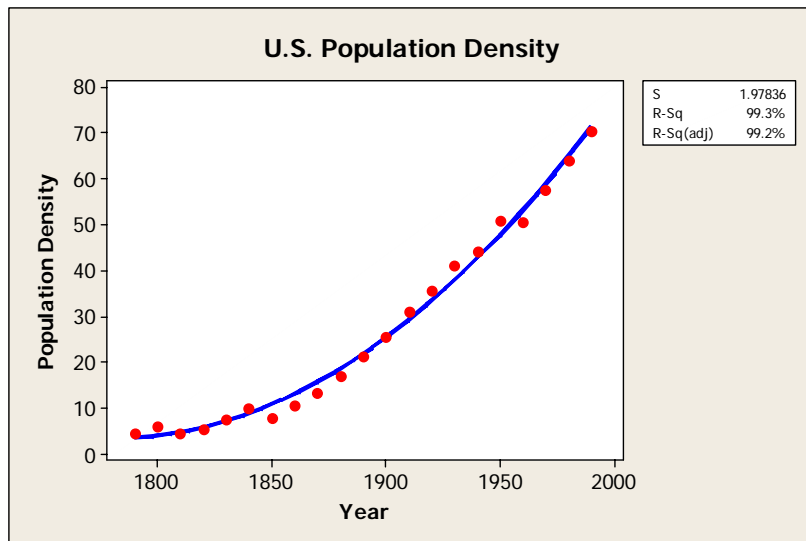
We can conclude that the growth in population density is accelerating.

The quadratic model explains 99.3% of the observed variation in population density, as opposed to 93.4% for the linear model.

Higher-order models

If the relationship between the dependent variable y and an independent variable x is not a straight line, we can include quadratic (x^2) or higher terms in the model.

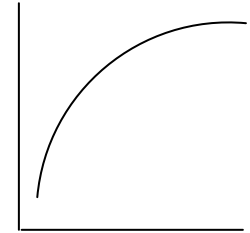
For example, we can examine the growth in U.S. population density over time.



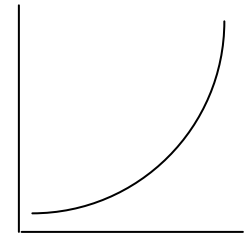
$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

Other examples:

Diminishing marginal returns in economics
($\beta_1 > 0, \beta_2 < 0$)



Epidemic growth over time
($\beta_1 > 0, \beta_2 > 0$)



Qualitative variables

How can we include qualitative (non-numeric) variables in our regression model?

Binary variables (e.g. gender): create a “dummy variable” that is 1 for one value of the variable and 0 for the other.

For example, if we use a dummy variable Gender = 1 for females and Gender = 0 for males, the coefficient β_j of Gender represents the expected difference between female and male salaries, holding all other independent variables constant.

Then $\beta_j = -1295$ means that we expect a female employee to make \$1295 less than a male employee, assuming the same amounts of experience, education, and employees supervised.

Qualitative variables

How can we include qualitative (non-numeric) variables in our regression model?

Binary variables (e.g. gender): create a “dummy variable” that is 1 for one value of the variable and 0 for the other.

Multi-valued variable (e.g. department): create $v - 1$ dummy variables, where v is the number of possible values of the qualitative variable.

Example: Employees can be from four different departments.

$x_1 = \text{Years}$, $x_2 = \text{Education}$, $x_3 = \text{Gender}$, $x_4 = \text{Supervised}$
 $x_5 = 1$ if employee from department 1, 0 otherwise
 $x_6 = 1$ if employee from department 2, 0 otherwise
 $x_7 = 1$ if employee from department 3, 0 otherwise

Salary = 27034 + 712 Years + 1549 Education - 1930 Gender
+ 123 Supervised - 8199 Dept1 + 347 Dept2 - 3020 Dept3

Employees from Department 1 make \$8,199 less than Department 4, etc.

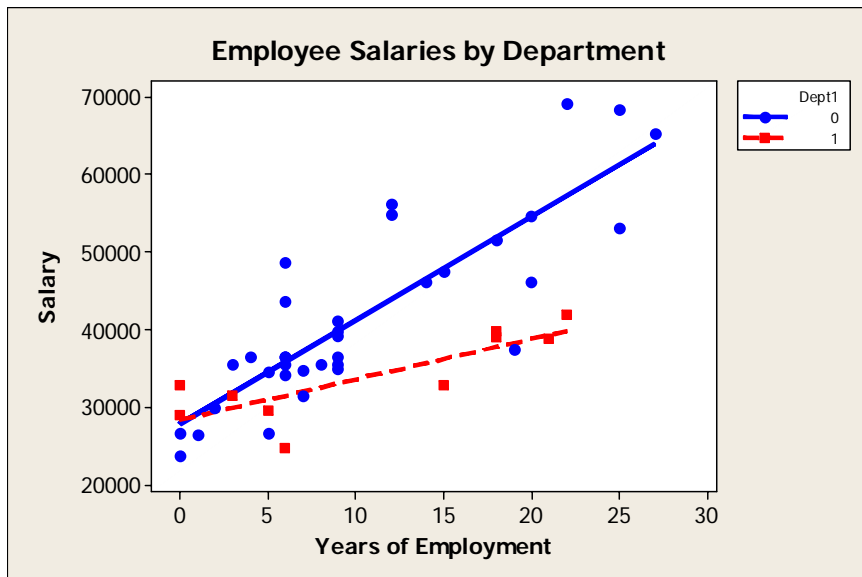
Interaction models

One problem with the standard multiple regression model is that the variables cannot interact: a change in one independent variable always has the same effect on the dependent variable, regardless of the values of the other variables.

The regression equation is

$$\text{Salary} = 30008 + 1122 \text{ Years} - 8125 \text{ Dept1}$$

Employees in Department 1 make an average of \$8125 less, regardless of how long they've worked. Salaries increase by an average of \$1122/yr.



In fact, the scatterplot reveals that employees in both departments make about the same starting salary, but salaries in Department 1 rise more slowly with experience.

How can we represent this in our model?

Interaction models

To represent interactions between two variables, we can include an interaction term x_1x_2 equal to the product of the two variables.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1x_2)$$

$$\text{Salary} = 27750 + 1344 \text{ Years} + 596 \text{ Dept1} - 820 \text{ YearsDept1}$$

← Salaries increase by an average of \$524/yr for employees in Department 1, versus \$1344/yr for other employees.

← Employees in Department 1 make \$596 more, minus \$820 per year of employment.



Predictor	Coef	SE Coef	T	P
Constant	27750	1657	16.74	0.000
Years	1343.5	132.4	10.14	0.000
Dept1	596	3406	0.17	0.862
YearsDept1	-819.9	254.8	-3.22	0.002

S = 5766.07 **R-Sq = 74.4%** R-Sq(adj) = 72.5%
(vs. $R^2 = 68.0\%$ for model without interaction)

Interaction models

To represent interactions between two variables, we can include an interaction term $x_1 x_2$ equal to the product of the two variables.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 x_2)$$

Salary = 25138 + 458 Years + 1512 Education + 35.3 YearsEducation

For a fixed amount of education x_2 , salary increases by $(\$458 + \$35.30 x_2)$ per year of employment. We would expect the salary of an employee with 10 years education to increase by \$811/yr.

However, the amount of interaction is not large enough to be significant.

← We observe a positive interaction between years of employment and years of education, suggesting that more educated employees may advance in salary faster over time.

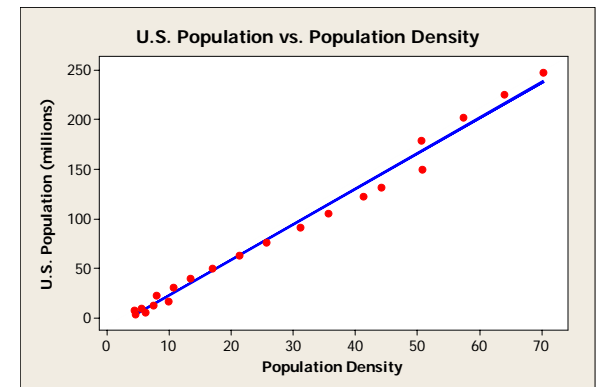
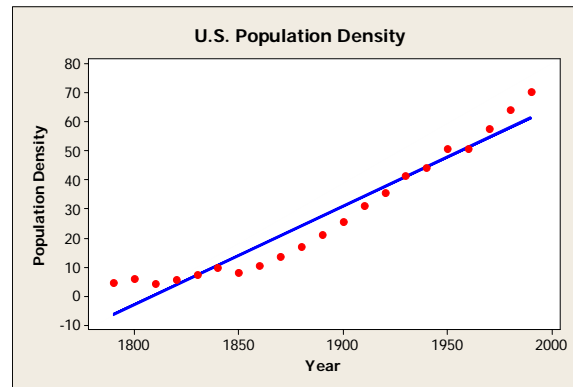
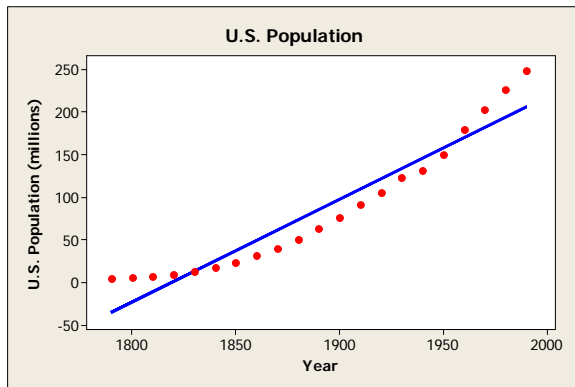
Predictor	Coef	SE Coef	T	P
Constant	25138	2662	9.44	0.000
Years	458.0	259.5	1.76	0.085
Education	1512.1	558.7	2.71	0.010
YrsEd	35.33	35.83	0.99	0.330

S = 5740.16 **R-Sq = 74.6%** R-Sq(adj) = 72.8%
(vs. $R^2 = 74.0\%$ for model without interaction)

Multicollinearity

When two or more of the independent variables in the multiple regression model are strongly correlated, serious problems can result.

Consider the trends in U.S. population and population density over time:




$$\text{Population} = 15 + 3.64 \text{ Population Density} - 0.016 \text{ Year}$$


If you suspect multicollinearity, check for significant correlations between pairs of independent variables, and remove one of the correlated variables from the regression. Minitab also checks for more complicated forms of multicollinearity between 3 or more variables.

Multicollinearity

When two or more of the independent variables in the multiple regression model are strongly correlated, serious problems can result.

$$\text{Salary} = 43191 - 9190 \text{ Dept1} - 5642 \text{ Dept2} - 2526 \text{ Dept3}$$

If we had included Dept4 in the regression model, we know that $\text{Dept1} + \text{Dept2} + \text{Dept3} + \text{Dept4} = 1$, so we have multicollinearity.

We could add any multiple of $(\text{Dept1} + \text{Dept2} + \text{Dept3} + \text{Dept4} - 1)$ to the model:

$$\begin{aligned} \text{Salary} &= 33191 + 810 \text{ Dept1} + 4358 \text{ Dept2} + 7474 \text{ Dept3} + 10000 \text{ Dept4} \\ &= 23191 + 10810 \text{ Dept1} + 14358 \text{ Dept2} + 17474 \text{ Dept3} + 20000 \text{ Dept4}, \text{ etc.} \end{aligned}$$

This is why we only include $v - 1$ dummy variables instead of v : to avoid multicollinearity.

If you suspect multicollinearity, check for significant correlations between pairs of independent variables, and remove one of the correlated variables from the regression. Minitab also checks for more complicated forms of multicollinearity between 3 or more variables.

Model building

Which subset of the independent variables, higher-order terms, interaction terms, etc. should we include?

There is no definitive answer here, but one possibility is to start out with all variables then remove insignificant variables one by one (don't remove them all simultaneously!)

Another option is to use Minitab to perform a stepwise regression analysis.

$y = \text{Salary}$
 $x_1 = \text{Years}$
 $x_2 = \text{Education}$
 $x_3 = \text{Gender}$ ← Remove second (p-value = .174)
 $x_4 = \text{Supervised}$ ← Remove fourth (p-value = .103)
 $x_5 = \text{Dept1}$
 $x_6 = \text{Dept2}$ ← Remove first (p-value = .862)
 $x_7 = \text{Dept3}$ ← Remove third (p-value = .140)

$$\text{Salary} = 24868 + 700 \text{ Years} + 1858 \text{ Education} - 7714 \text{ Dept1}$$

Stepwise regression using $\alpha = .05$ produces the same equation.

Stepwise regression using $\alpha = .15$ also includes Supervised and Dept3.

Model building

Once we have a reasonable subset of variables to examine, we can consider higher-order and interaction terms.

{Years, Education, Dept1} →	Add Years ² ?	No, p = .598
	Add Education ² ?	No, p = .299
	Add Dept1 ² ?	No, dummy variable.
	Add Years*Education?	No, p = .294
	Add Years*Dept1?	Yes, p = .001
	Add Education*Dept1?	Yes, p = .008

Final equation: Salary = 22130 + 797 Years + 2201 Education + 4282 Dept1
– 379 (Years*Dept1) – 1588 (Education*Dept1)
(R² = .890)

Keep in mind that testing so many variables increases our risk of Type I errors. It would be better to only test for the higher-order terms and interactions that we believe are most likely to occur.