



WEEK 10 – LECTURE 2

- +
-
- Multiple Linear Regression

Objectives

- Multiple Regression Model
- Least Squares Method
- Multiple Coefficient of Determination
- Model Assumptions
- Testing for Significance
- Using the Estimated Regression Equation for Estimation and Prediction
- Qualitative Independent Variables
- Residual Analysis
- Logistic Regression

Multiple Regression Model

- The equation that describes how the dependent variable y is related to the independent variables x_1, x_2, \dots, x_p and an error term is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

where

- $\beta_0, \beta_1, \dots, \beta_p$ are the parameters,
 - ε is a random variable called the **error term**.
- The equation that describes how the mean value of y is related to x_1, x_2, \dots, x_p is:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

Estimation of Population Coefficients $\beta_0, \beta_1, \dots, \beta_p$

- A simple random sample is used to compute sample statistics $b_0, b_1, b_2, \dots, b_p$, which are used as the point estimators of the parameters $\beta_0, \beta_1, \dots, \beta_p$.

- Denote sample data values by $(x_i, y_i), i = 1, 2, \dots, n$, with $x_i = (x_{1i}, x_{2i}, \dots, x_{pi})$. Then

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$$

- Estimated multiple regression equation is:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$$

Thus,

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \dots + b_p x_{pi}$$

Assumptions About the Error Terms ε_i

- The errors ε_i are independent of each other.
 - The errors ε_i , at each set of values of the predictors $(x_{1i}, x_{2i}, \dots, x_{pi})$, are normally distributed with a mean of zero and have equal standard deviations, denoted by σ_ε .
- In another words, the errors ε_i are independent normal random variables with mean zero and constant standard deviation σ_ε .

Estimation of Population Coefficients $\beta_0, \beta_1, \dots, \beta_p$ (cont.)

- The regression coefficients $b_0, b_1, b_2, \dots, b_p$ are computed by solving the following optimization problem:

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- The formulas for $b_0, b_1, b_2, \dots, b_p$ involve the use of matrix algebra. We will rely on computer software packages to perform these calculations.

Multiple Linear Regression - Example

A software firm collected data for a sample of 20 computer programmers. A suggestion was made that regression analysis could be used to determine if salary was related to the years of experience and the score on the firm's programmer aptitude test.

The years of experience, score on the aptitude test, and corresponding annual salary (\$1000s) for a sample of 20 programmers is shown on the next slide.

Multiple Linear Regression – Example (cont.)

Exper.	Score	Salary	Exper.	Score	Salary
4	78	24.0	9	88	38.0
7	100	43.0	2	73	26.6
1	86	23.7	10	75	36.2
5	82	34.3	5	81	31.6
8	86	35.8	6	74	29.0
10	84	38.0	8	87	34.0
0	75	22.2	4	79	30.1
1	80	23.1	6	94	33.9
6	83	30.0	3	70	28.2
6	91	33.0	3	89	30.0

Multiple Linear Regression – Example (cont.)

We try to fit the data to a multiple linear regression model as follows.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where:

y = annual salary (in \$1000s),

x_1 = years of experience,

x_2 = score on programmer aptitude test.

Multiple Linear Regression – Example (cont.)

Excel's Regression Equation Output:

Regression Statistics								
Multiple R	0.91333406							
R Square	0.8341791							
Adjusted R Square	0.81467076							
Standard Error	2.41876208							
Observations	20							
ANOVA								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	2	500.3285303	250.164265	42.7601255	2.32774E-07			
Residual	17	99.45696969	5.85040998					
Total	19	599.7855						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	3.17393627	6.156066829	0.51557859	0.6127887	-9.81422942	16.162102	-9.8142294	16.162102
Exper.	1.40390249	0.198566912	7.07017333	1.8806E-06	0.984962921	1.82284205	0.98496292	1.82284205
Score	0.25088545	0.077354127	3.2433363	0.00478002	0.087682506	0.41408839	0.08768251	0.41408839

Multiple Linear Regression – Example (cont.)

Interpret the coefficients:

- b_i represents an estimate of the change in y corresponding to a 1-unit increase in x_i when all other independent variables are held constant.
- $b_1 = 1.404 \rightarrow$ Salary is expected to increase by \$1,404 for each additional year of experience when the variable Score on programmer aptitude test is held constant.
- $b_2 = 0.251 \rightarrow$ Salary is expected to increase by \$251 for each additional point scored on the programmer aptitude test when the variable Years of Experience is held constant.

Multiple Coefficient of Determination R^2

- Relationship among SST, SSR, SSE:

$$SST = SSR + SSE$$

where:

- $SST = \sum (y_i - \bar{y})^2$: total sum of squares
- $SSR = \sum (\hat{y}_i - \bar{y})^2$: sum of squares due to regression
- $SSE = \sum (y_i - \hat{y}_i)^2$: sum of squares due to error (residual)

with $\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n)$.

- Mean Square of Regression: $MSR = \frac{SSR}{p}$
- Mean Square of Error: $MSE = \frac{SSE}{n-p-1}$

Multiple Coefficient of Determination

➤ Multiple Coefficient of Determination:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- R^2 represents the proportion of variation in the dependent variable explained by the set of independent variables in a multiple regression model.

➤ Adjusted Multiple Coefficient of Determination:

$$R_a^2 = 1 - \frac{n - 1}{n - p - 1} \cdot \frac{SSE}{SST} = 1 - \frac{n - 1}{n - p - 1} (1 - R^2)$$

Testing for Significance

- In simple linear regression, the F and t-tests provide the same conclusion. In multiple regression, the F and t-tests have different purposes.
- The F-test is used to determine whether a significant relationship exists between the dependent variable and **the set of all the independent variables**. The F-test is referred to as the **test for overall significance**.
- If the F-test shows an overall significance, the t-test is used to determine whether each of the individual independent variable is significant. A separate t-test is conducted for each of the independent variables in the model. We refer to each of these t-test as a **test for individual significance**.

Testing for Significance – F-Test

- **Hypotheses:**

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

H_1 : One or more of the parameters is not equal to zero.

- **Test Statistic:**

$$F = \frac{MSR}{MSE}$$

- **Rejection Region:**

Reject H_0 if p-value $\leq \alpha$ or if $F \geq F_\alpha$, where F_α is based on an F distribution with p degrees of freedom in the numerator and $n - p - 1$ degrees of freedom in the denominator.

Testing for Significance – t -Test

- **Hypotheses:**

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

- **Test Statistic:**

$$t = \frac{b_i}{s_{b_i}}$$

where b_i is the estimated coefficient and s_{b_i} is the standard deviation (standard error) of the estimated coefficient.

- **Rejection Region:**

Reject H_0 if $p\text{-value} \leq \alpha$, or if $t \leq -t_{\alpha/2}$ or $t \geq t_{\alpha/2}$, where $t_{\alpha/2}$ is based on a t distribution with $n - p - 1$ degrees of freedom.

Testing for Significance – Multicollinearity

- Multicollinearity refers to the correlation among the independent variables. When the independent variables are highly correlated ($|r| > 0.7$), it is not possible to determine the separate effect of any independent variable on the dependent variable.
- If the estimated regression equation is to be used only for predictive purposes, multicollinearity is usually not a serious problem. However, every attempt should be made to avoid including independent variables that are highly correlated.

Estimation using Multiple Linear Regression

- The procedures for estimating the mean value of y in multiple regression are similar to those in simple regression. We substitute the given values of x_1, x_2, \dots, x_p into the estimated regression equation and use the corresponding value of \hat{y} as the point estimate.
- Software packages for multiple linear regression will often provide interval estimates.

Multiple Linear Regression for Qualitative Data

- In many situations we must work with qualitative independent variables such as gender (male, female), car color (navy, black, white, silver), etc. In these cases, we can use dummy (or indicator) variables.
- For example, for gender we may use a variable x where $x = 0$ indicates a male and $x = 1$ indicates a female.

Multiple Linear Regression for Qualitative Data - Example

- As an extension of the problem involving the computer programmer salary survey, suppose that management also believes that the annual salary is related to whether the individual has a graduate degree in computer science or information systems.

Example:

The years of experience, the score on the programmer aptitude test, whether the individual has a relevant graduate degree, and the annual salary (\$1000) for each of the sampled 20 programmers are shown on the next slide.

Multiple Linear Regression for Qualitative Data – Example (cont.)

Exper.	Score	Degree	Salary	Exper.	Score	Degree	Salary
4	78	No	24.0	9	88	Yes	38.0
7	100	Yes	43.0	2	73	No	26.6
1	86	No	23.7	10	75	Yes	36.2
5	82	Yes	34.3	5	81	No	31.6
8	86	Yes	35.8	6	74	No	29.0
10	84	Yes	38.0	8	87	Yes	34.0
0	75	No	22.2	4	79	No	30.1
1	80	No	23.1	6	94	Yes	33.9
6	83	No	30.0	3	70	No	28.2
6	91	Yes	33.0	3	89	No	30.0

Multiple Linear Regression for Qualitative Data – Example (cont)

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

where:

- \hat{y} = annual salary (in \$1000s)
- x_1 = years of experience
- x_2 = score on programmer aptitude test
- $x_3 = 0$ if individual does not have a graduate degree and $x_3 = 1$ if individual has a graduate degree. x_3 is a dummy variables.

Multiple Linear Regression for Qualitative Data – Example (cont)

Regression Statistics								
Multiple R	0.92021524							
R Square	0.84679609							
Adjusted R Square	0.81807035							
Standard Error	2.3964751							
Observations	20							
ANOVA								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	3	507.8960134	169.298671	29.4786579	9.41675E-07			
Residual	16	91.88948657	5.74309291					
Total	19	599.7855						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	7.94484872	7.380797058	1.07642151	0.29770238	-7.701742077	23.5914395	-7.7017421	23.5914395
Exper.	1.14758173	0.29760152	3.85610169	0.00139723	0.516694687	1.77846877	0.51669469	1.77846877
Score	0.19693699	0.089903726	2.19053197	0.04363981	0.006349601	0.38752437	0.0063496	0.38752437
Degree	2.28042384	1.986610668	1.1478967	0.26788534	-1.931002643	6.49185032	-1.9310026	6.49185032

More Complex Qualitative Variables

- If a qualitative variable has k levels, $k - 1$ dummy variables are required, with each dummy variable being coded as 0 or 1.
- For example, car color with four values of navy, black, white and silver could be converted to three dummy variables color_navy, color_black, color_white as follows:

color	color_navy	color_black	color_white
navy	1	0	0
black	0	1	0
white	0	0	1
silver	0	0	0

Residual Analysis

- For simple linear regression, the residual plot against \hat{y} and the residual plot against x provide the same information.
- In multiple regression analysis, it is preferable to use the residual plot against \hat{y} to determine if the model assumptions are satisfied.
- Standardized residuals are frequently used in residual plots for purposes of
 - 1) identifying outliers (typically, standardized residuals less than -2 or larger than 2),
 - 2) providing insight about the assumption that the error term ε has a normal distribution.

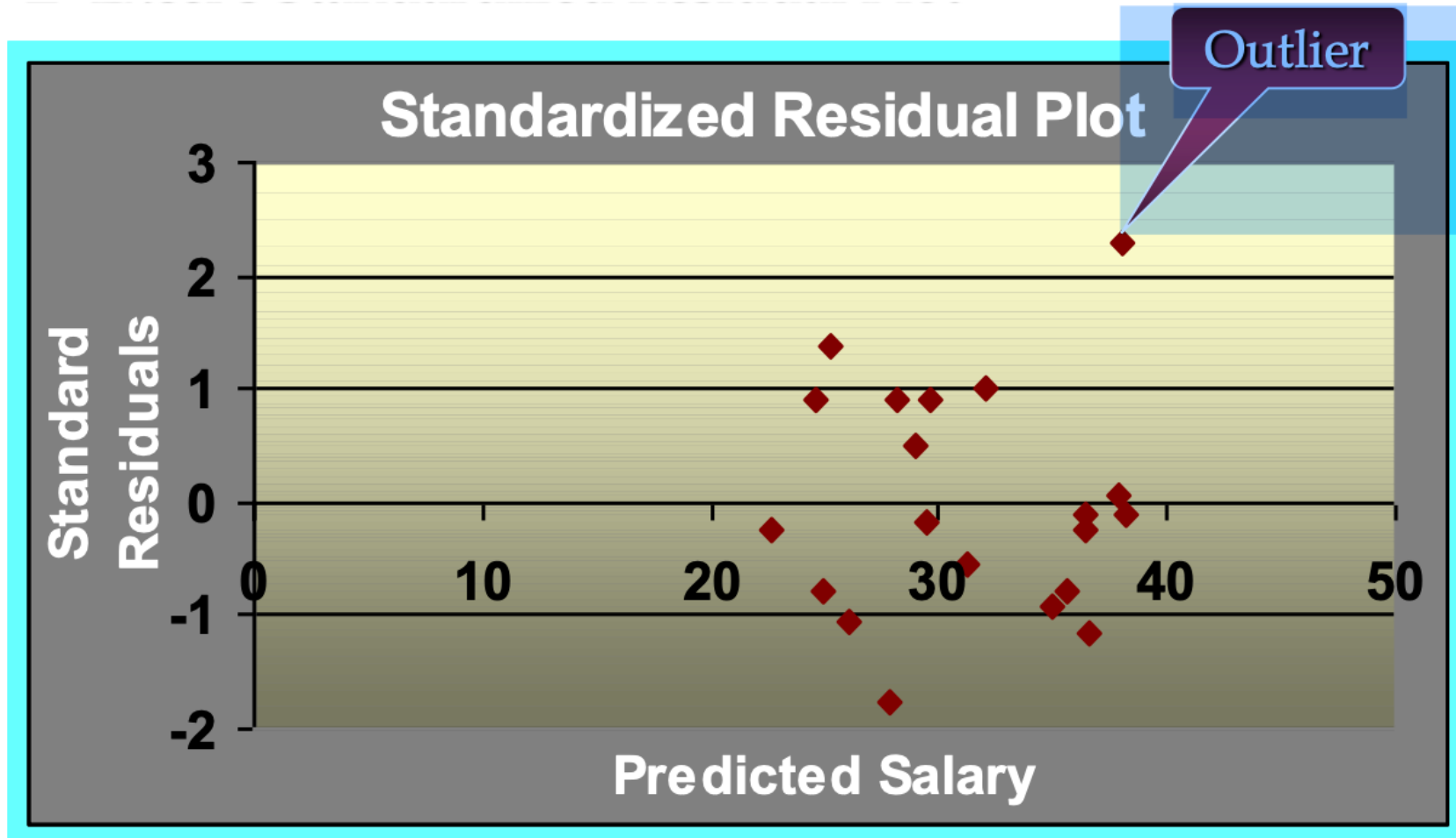
$$\text{Standardized Residual} = \frac{\text{Observed value} - \text{Predicted value}}{\text{Standard Error of the Estimate}}$$

Residual Analysis - Example

Below is an Excel multiple linear regression output. The standardized residual plot against \hat{y} for it is given in the next slide.

	A	B	C	D
28				
29	RESIDUAL OUTPUT			
30				
31	<i>Observation</i>	<i>Predicted Y</i>	<i>Residuals</i>	<i>Standard Residuals</i>
32	1	27.89626052	-3.89626052	-1.771706896
33	2	37.95204323	5.047956775	2.295406016
34	3	26.02901122	-2.32901122	-1.059047572
35	4	32.11201403	2.187985973	0.994920596
36	5	36.34250715	-0.54250715	-0.246688757

Residual Analysis – Example (cont.)



Logistic Regression

- Logistic regression can be used to model situations in which the dependent variable may assume only two discrete values, such as 0 and 1. In these cases, ordinary linear multiple regression model is not applicable.
- Logistic Regression Equation:

$$E(y) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p}}$$

- Interpretation: If the two values of y are coded as 0 and 1, then

$$E(y) = \text{estimate of } P(y = 1 | x_1, x_2, \dots, x_p)$$

Logistic Regression (cont.)

- A simple random sample is used to compute the regression coefficients $b_0, b_1, b_2, \dots, b_p$ that are used as the point estimators of the parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_p$.
- **Estimated Logistic Regression Equation:**

$$\hat{y} = \frac{e^{b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p}}{1 + e^{b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p}}$$


Logistic Regression - Example

Simmons conducted a study by sending out 100 catalogs, 50 to customers who have a Simmons credit card and 50 to customers who do not have the card. At the end of the test period, Simmons noted the following for each of the 100 customers:

- 1) the amount the customer spent last year at Simmons,
- 2) whether the customer had a Simmons credit card,
- 3) whether the customer made a \$200 purchase.

A portion of the test data is shown on the next slide.

Logistic Regression – Example (cont.)



Customer	Annual Spending (\$1000)	Simmons Credit Card	\$200 Purchase
1	2.291	1	0
2	3.215	1	0
3	2.135	1	0
4	3.924	0	0
5	2.528	1	0
6	2.473	0	1
7	2.384	0	0
8	7.076	0	0
9	1.182	1	1
10	3.345	0	0

Below is the logistic regression result using Minitab:

Predictor	Coef	SE Coef	Z	p	Odds Ratio	95% CI	
						Lower	Upper
Constant	-2.1464	0.5772	-3.72	0.000			
Spending	0.3416	0.1287	2.66	0.008	1.41	1.09	1.81
Card	1.0987	0.4447	2.47	0.013	3.00	1.25	7.17

Log-Likelihood = -60.487

Test that all slopes are zero: $G = 13.628$, $DF = 2$, $P\text{-Value} = 0.001$

Logistic Regression – Example (cont.)

- The estimated logistic regression equation is

$$\hat{y} = \frac{e^{-2.1464+0.3416x_1+1.0987x_2}}{1 + e^{-2.1464+0.3416x_1+1.0987x_2}}$$

- For customers that spend \$2000 annually and do not have a Simmons credit card:

$$\hat{y} = \frac{e^{-2.1464+0.3416(2)+1.0987(0)}}{1 + e^{-2.1464+0.3416x_1+1.0987x_2}} = 0.1880$$

- For customers that spend \$2000 annually and have a Simmons credit card:

$$\hat{y} = \frac{e^{-2.1464+0.3416(2)+1.0987(1)}}{1 + e^{-2.1464+0.3416x_1+1.0987x_2}} = 0.4099$$

Logistic Regression – Testing for Significance

- Hypotheses:

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

- Test statistics:

$$z = \frac{b_i}{s_{b_i}}$$

- Rejection region:

Reject H_0 if p-value $\leq \alpha$.

Logistic Regression – Example (cont.)

For the Simmons credit card problem, let $\alpha = 0.05$.

- Performing the test for significance of x_1 gives $z = 2.66$ and $p\text{-value} = 0.008 < \alpha = 0.05$. Hence, $\beta_1 \neq 0$. In other words, x_1 is statistically significant.
 - Performing the test for significance of x_2 gives $z = 2.47$ and $p\text{-value} = 0.013 < \alpha = 0.05$. Hence, $\beta_2 \neq 0$. In other words, x_2 is also statistically significant.
- This means that it is meaningful to use the estimated logistic regression equation to estimate \hat{y} using x_1 and x_2 .

Logistic Regression – Odds Ratio

- With logistic regression it is difficult to interpret the relationship between the variables because the equation is not linear, so we use the odds ratio.

$$odds_1 = \frac{P(y = 1|x_1, x_2, \dots, x_p)}{1 - P(y = 1|x_1, x_2, \dots, x_p)}$$

$$odds_0 = \frac{P(y = 0|x_1, x_2, \dots, x_p)}{1 - P(y = 0|x_1, x_2, \dots, x_p)}$$

- Odds Ratio:

$$Odds\ Ratio = \frac{odds_1}{odds_0}$$

Logistic Regression – Example (cont.)

- For the Simmons credit card problem, the table below shows the odds results.

		Annual Spending						
		\$1000	\$2000	\$3000	\$4000	\$5000	\$6000	\$7000
Credit Card	Yes	0.3305	0.4099	0.4943	0.5790	0.6593	0.7314	0.7931
	No	0.1413	0.1880	0.2457	0.3143	0.3921	0.4758	0.5609

Computed
earlier

Logistic Regression – Example (cont.)

- Odds ratio:

Let's say we want to compare the odds of making a \$200 purchase for customers who spend \$2000 annually and have a Simmons credit card to the odds of making a \$200 purchase for customers who spend \$2000 annually and do not have a Simmons credit card.

$$odds_1 = \text{odds of having credit card} = \frac{0.4099}{1 - 0.4099} = 0.6946$$

$$odds_0 = \text{odds of not having credit card} = \frac{0.1880}{1 - 0.1880} = 0.2315$$

$$\text{odds ratio} = \frac{0.6946}{0.2315} = 3.00$$

Concerns in Regression Modeling

- Multicollinearity.
- Parameter estimability: The inability to estimate the parameters of the regression model because the data are concentrated in one area. To ensure that the parameters are estimable, $n > p + 1$.
- Variable selection: When you have a large number of independent variables and need to decide which ones to include in the model.
- Extrapolation: The model is used to predict values outside the range of the data used to estimate the model.
- Correlated errors: Measurements of the dependent variable are correlated. One will often see this type of dependency with time series data where current measurements are often dependent on measurements in the previous time period.