Lecture 5: Linear Regression

Nayel Bettache

Department of Statistical Science, Cornell University

Linear Regression and the Population Regression Line

- The true relationship between X and Y is assumed to be $Y = f(X) + \epsilon$ for some unknown function f, where ϵ is a mean-zero random error term.

Linear Regression and the Population Regression Line

- The true relationship between X and Y is assumed to be $Y = f(X) + \epsilon$ for some unknown function f, where ϵ is a mean-zero random error term.
- If f is approximated by a linear function, then $Y = \beta_0 + \beta_1 X + \epsilon$ where β_0 is the intercept and β_1 is the slope.

Linear Regression and the Population Regression Line

- The true relationship between X and Y is assumed to be $Y = f(X) + \epsilon$ for some unknown function f, where ϵ is a mean-zero random error term.
- If f is approximated by a linear function, then $Y = \beta_0 + \beta_1 X + \epsilon$ where β_0 is the intercept and β_1 is the slope.
- The model defines the population regression line, which is the best linear approximation to the true relationship between X and Y.

Linear Regression and the Population Regression Line

- The true relationship between X and Y is assumed to be $Y = f(X) + \epsilon$ for some unknown function f, where ϵ is a mean-zero random error term.
- If f is approximated by a linear function, then $Y = \beta_0 + \beta_1 X + \epsilon$ where β_0 is the intercept and β_1 is the slope.
- The model defines the population regression line, which is the best linear approximation to the true relationship between X and Y.
- The least squares regression coefficient estimates characterize the least squares line, which can be computed using the observed data.

Linear Regression and the Population Regression Line

- The true relationship between X and Y is assumed to be $Y = f(X) + \epsilon$ for some unknown function f, where ϵ is a mean-zero random error term.
- If f is approximated by a linear function, then $Y = \beta_0 + \beta_1 X + \epsilon$ where β_0 is the intercept and β_1 is the slope.
- The model defines the population regression line, which is the best linear approximation to the true relationship between X and Y.
- The least squares regression coefficient estimates characterize the least squares line, which can be computed using the observed data.
- The true relationship is generally not known, but the least squares line can always be computed.

Linear Regression and the Population Regression Line

- The true relationship between X and Y is assumed to be $Y = f(X) + \epsilon$ for some unknown function f, where ϵ is a mean-zero random error term.
- If f is approximated by a linear function, then $Y = \beta_0 + \beta_1 X + \epsilon$ where β_0 is the intercept and β_1 is the slope.
- The model defines the population regression line, which is the best linear approximation to the true relationship between X and Y.
- The least squares regression coefficient estimates characterize the least squares line, which can be computed using the observed data.
- The true relationship is generally not known, but the least squares line can always be computed.
- Different data sets generated from the same true model result in slightly different least squares lines, but the unobserved population regression line does not change.

Linear Regression and the Population Regression Line

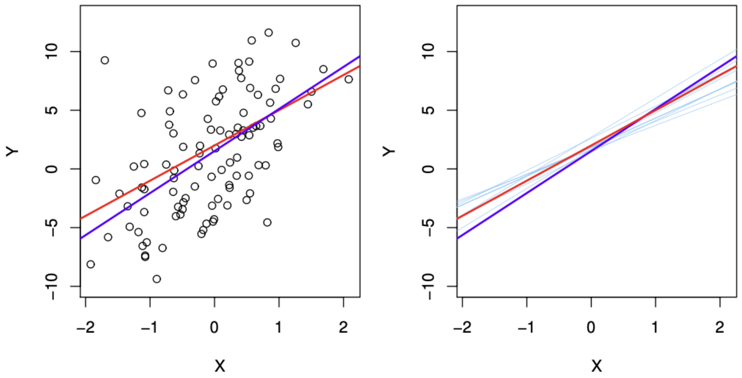
- The true relationship between X and Y is assumed to be $Y = f(X) + \epsilon$ for some unknown function f, where ϵ is a mean-zero random error term.
- If f is approximated by a linear function, then $Y = \beta_0 + \beta_1 X + \epsilon$ where β_0 is the intercept and β_1 is the slope.
- The model defines the population regression line, which is the best linear approximation to the true relationship between X and Y.
- The least squares regression coefficient estimates characterize the least squares line, which can be computed using the observed data.
- The true relationship is generally not known, but the least squares line can always be computed.
- Different data sets generated from the same true model result in slightly different least squares lines, but the unobserved population regression line does not change.
- The concept of the population regression line and the least squares line is an extension of the standard statistical approach of using information from a sample to estimate characteristics of a large population.

Linear Regression and the Population Regression Line

- The true relationship between X and Y is assumed to be $Y = f(X) + \epsilon$ for some unknown function f, where ϵ is a mean-zero random error term.
- If f is approximated by a linear function, then $Y = \beta_0 + \beta_1 X + \epsilon$ where β_0 is the intercept and β_1 is the slope.
- The model defines the population regression line, which is the best linear approximation to the true relationship between X and Y.
- The least squares regression coefficient estimates characterize the least squares line, which can be computed using the observed data.
- The true relationship is generally not known, but the least squares line can always be computed.
- Different data sets generated from the same true model result in slightly different least squares lines, but the unobserved population regression line does not change.
- The concept of the population regression line and the least squares line is an extension of the standard statistical approach of using information from a sample to estimate characteristics of a large population.
- The standard error of the estimate can be used to quantify the accuracy of the estimate.

Linear Regression and the Population Regression Line

Left: The red line represents the true relationship, $f(X) = 2 + 3X$, i.e. the population regression line. The blue line is the least squares line; it is the least squares estimate for $f(X)$ based on the observed data, shown in black. Right: The population regression line (red), and the least squares line (dark blue). Ten least squares lines are shown (light blue), each computed on the basis of a separate random set of observations. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.



Population Regression Line vs. Least Squares Line

- The population regression line is the true relationship between X and Y.

Population Regression Line vs. Least Squares Line

- The population regression line is the true relationship between X and Y.
- The least squares line is an estimate of the population regression line based on the observed data.

Population Regression Line vs. Least Squares Line

- The population regression line is the true relationship between X and Y.
- The least squares line is an estimate of the population regression line based on the observed data.
- The least squares line is computed using the least squares coefficient estimates.

Population Regression Line vs. Least Squares Line

- The population regression line is the true relationship between X and Y.
- The least squares line is an estimate of the population regression line based on the observed data.
- The least squares line is computed using the least squares coefficient estimates.
- The average of many least squares lines, each estimated from a separate data set, is close to the true population regression line.

Bias and Standard Error

- An unbiased estimator does not systematically over- or under-estimate the true parameter.

Bias and Standard Error

- An unbiased estimator does not systematically over- or under-estimate the true parameter.
- The least squares coefficient estimates are unbiased.

Bias and Standard Error

- An unbiased estimator does not systematically over- or under-estimate the true parameter.
- The least squares coefficient estimates are unbiased.
- The standard error of the estimate can be used to quantify the accuracy of the estimate.

Bias and Standard Error

- An unbiased estimator does not systematically over- or under-estimate the true parameter.
- The least squares coefficient estimates are unbiased.
- The standard error of the estimate can be used to quantify the accuracy of the estimate.
- The standard error of the estimate is a measure of the variability of the estimate.

Bias and Standard Error

- An unbiased estimator does not systematically over- or under-estimate the true parameter.
- The least squares coefficient estimates are unbiased.
- The standard error of the estimate can be used to quantify the accuracy of the estimate.
- The standard error of the estimate is a measure of the variability of the estimate.
- How can we quantify the quality of the estimation of β_0 and β_1 ?

Analogy: Estimation of the population mean

- Consider a random variable Y with mean μ.

Analogy: Estimation of the population mean

- Consider a random variable Y with mean μ.
- Unfortunately, μ is unknown, but we do have access to n observations from Y: $y_1, \ldots, y_n$.

Analogy: Estimation of the population mean

- Consider a random variable Y with mean μ.
- Unfortunately, μ is unknown, but we do have access to n observations from Y: $y_1, \ldots, y_n$.
- A reasonable estimate is $\hat{\mu} = \bar{y}$, the empirical mean.

Analogy: Estimation of the population mean

- Consider a random variable Y with mean μ.
- Unfortunately, μ is unknown, but we do have access to n observations from Y: $y_1, \ldots, y_n$.
- A reasonable estimate is $\hat{\mu} = \bar{y}$, the empirical mean.
- In general $\hat{\mu} \neq \mu$ but $\hat{\mu}$ is "close" to μ if we have enough data.

Analogy: Estimation of the population mean

- Consider a random variable Y with mean μ.
- Unfortunately, μ is unknown, but we do have access to n observations from Y: $y_1, \ldots, y_n$.
- A reasonable estimate is $\hat{\mu} = \bar{y}$, the empirical mean.
- In general $\hat{\mu} \neq \mu$ but $\hat{\mu}$ is "close" to μ if we have enough data.
- How accurate is the sample mean $\hat{\mu}$ as an estimate of μ?

Analogy: Estimation of the population mean

- Consider a random variable Y with mean μ.
- Unfortunately, μ is unknown, but we do have access to n observations from Y: $y_1, \ldots, y_n$.
- A reasonable estimate is $\hat{\mu} = \bar{y}$, the empirical mean.
- In general $\hat{\mu} \neq \mu$ but $\hat{\mu}$ is "close" to μ if we have enough data.
- How accurate is the sample mean $\hat{\mu}$ as an estimate of μ?
- We answer this question by computing the standard error of $\hat{\mu}$, denoted $SE(\hat{\mu})$.

Analogy: Estimation of the population mean

- Consider a random variable Y with mean μ.
- Unfortunately, μ is unknown, but we do have access to n observations from Y: $y_1, \ldots, y_n$.
- A reasonable estimate is $\hat{\mu} = \bar{y}$, the empirical mean.
- In general $\hat{\mu} \neq \mu$ but $\hat{\mu}$ is "close" to μ if we have enough data.
- How accurate is the sample mean $\hat{\mu}$ as an estimate of μ?
- We answer this question by computing the standard error of $\hat{\mu}$, denoted $SE(\hat{\mu})$.
- We do the same for $\hat{\beta}_0$ and $\hat{\beta}_1$.

Analogy: Estimation of the population mean

- Consider a random variable Y with mean μ.
- Unfortunately, μ is unknown, but we do have access to n observations from Y: $y_1, \ldots, y_n$.
- A reasonable estimate is $\hat{\mu} = \bar{y}$, the empirical mean.
- In general $\hat{\mu} \neq \mu$ but $\hat{\mu}$ is "close" to μ if we have enough data.
- How accurate is the sample mean $\hat{\mu}$ as an estimate of μ?
- We answer this question by computing the standard error of $\hat{\mu}$, denoted $SE(\hat{\mu})$.
- We do the same for $\hat{\beta}_0$ and $\hat{\beta}_1$.
- $$SE(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x}^2)} \right] \quad \text{and} \quad SE(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x}^2)},$$

  where $\sigma^2 = \mathbb{V}[\epsilon]$.

Analogy: Estimation of the population mean

- Consider a random variable Y with mean μ.
- Unfortunately, μ is unknown, but we do have access to n observations from Y: $y_1, \ldots, y_n$.
- A reasonable estimate is $\hat{\mu} = \bar{y}$, the empirical mean.
- In general $\hat{\mu} \neq \mu$ but $\hat{\mu}$ is "close" to μ if we have enough data.
- How accurate is the sample mean $\hat{\mu}$ as an estimate of μ?
- We answer this question by computing the standard error of $\hat{\mu}$, denoted $SE(\hat{\mu})$.
- We do the same for $\hat{\beta}_0$ and $\hat{\beta}_1$.
- $$SE(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x}^2)} \right] \quad \text{and} \quad SE(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x}^2)},$$
  where $\sigma^2 = \mathbb{V}[\epsilon]$.
- For these formulas to be strictly valid, we need to assume that the errors for each observation have common variance and are uncorrelated.

Inference: Confidence Interval

- Instead of "point" estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ we want to have intervals such that $\beta_0 \in [m_0, M_0]$ and $\beta_1 \in [m_1, M_1]$ with high probability.

Inference: Confidence Interval

- Instead of "point" estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ we want to have intervals such that $\beta_0 \in [m_0, M_0]$ and $\beta_1 \in [m_1, M_1]$ with high probability.
- For δ $in(0, 1)$ we want to find $m_0(\delta)$ and $M_0(\delta)$ such that

$$\mathbb{P}\left(\beta_0 \in [m_0(\delta), M_0(\delta)]\right) \geq 1 - \delta.$$

Inference: Confidence Interval

- Instead of "point" estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ we want to have intervals such that $\beta_0 \in [m_0, M_0]$ and $\beta_1 \in [m_1, M_1]$ with high probability.
- For $\delta\ in(0, 1)$ we want to find $m_0(\delta)$ and $M_0(\delta)$ such that

$$\mathbb{P}\left(\beta_0 \in [m_0(\delta), M_0(\delta)]\right) \geq 1 - \delta.$$

- A 95% **confidence interval** is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.

Inference: Confidence Interval

- Instead of "point" estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ we want to have intervals such that $\beta_0 \in [m_0, M_0]$ and $\beta_1 \in [m_1, M_1]$ with high probability.
- For δ $in(0, 1)$ we want to find $m_0(\delta)$ and $M_0(\delta)$ such that

$$\mathbb{P}\left(\beta_0 \in [m_0(\delta), M_0(\delta)]\right) \geq 1 - \delta.$$

- A 95% **confidence interval** is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.
- Standard errors can be used to compute confidence intervals.

Inference: Confidence Interval

- Instead of "point" estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ we want to have intervals such that $\beta_0 \in [m_0, M_0]$ and $\beta_1 \in [m_1, M_1]$ with high probability.
- For $\delta \ in(0, 1)$ we want to find $m_0(\delta)$ and $M_0(\delta)$ such that

$$\mathbb{P}\left(\beta_0 \in [m_0(\delta), M_0(\delta)]\right) \geq 1 - \delta.$$

- A 95% **confidence interval** is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.
- Standard errors can be used to compute confidence intervals.
- A 95% confidence interval for β_1 has the form

$$[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)].$$

Inference: Confidence Interval

- Instead of "point" estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ we want to have intervals such that $\beta_0 \in [m_0, M_0]$ and $\beta_1 \in [m_1, M_1]$ with high probability.
- For δ $in(0, 1)$ we want to find $m_0(\delta)$ and $M_0(\delta)$ such that

$$\mathbb{P}\left(\beta_0 \in [m_0(\delta), M_0(\delta)]\right) \geq 1 - \delta.$$

- A 95% **confidence interval** is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.
- Standard errors can be used to compute confidence intervals.
- A 95% confidence interval for β_1 has the form

$$[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)].$$

- A 95% confidence interval for β_0 has the form

$$[\hat{\beta}_0 - 2 \cdot SE(\hat{\beta}_0), \hat{\beta}_0 + 2 \cdot SE(\hat{\beta}_0)].$$

Inference: Confidence Interval

- Instead of "point" estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ we want to have intervals such that $\beta_0 \in [m_0, M_0]$ and $\beta_1 \in [m_1, M_1]$ with high probability.
- For $\delta\ in(0,1)$ we want to find $m_0(\delta)$ and $M_0(\delta)$ such that

$$\mathbb{P}\left(\beta_0 \in [m_0(\delta), M_0(\delta)]\right) \geq 1 - \delta.$$

- A 95% **confidence interval** is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.
- Standard errors can be used to compute confidence intervals.
- A 95% confidence interval for β_1 has the form

$$[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)].$$

- A 95% confidence interval for β_0 has the form

$$[\hat{\beta}_0 - 2 \cdot SE(\hat{\beta}_0), \hat{\beta}_0 + 2 \cdot SE(\hat{\beta}_0)].$$

- Here, the constant 2 is used for simplicity.

Inference: Hypothesis Tests

- Instead of "estimating" the parameters we may be interested in testing a hypothesis, *e.g., is there a relationship between* X *and* Y *?*

Inference: Hypothesis Tests

- Instead of "estimating" the parameters we may be interested in testing a hypothesis, *e.g.*, *is there a relationship between* X *and* Y *?*
- Standard errors can also be used to perform hypothesis tests on the coefficients.

Inference: Hypothesis Tests

- Instead of "estimating" the parameters we may be interested in testing a hypothesis, *e.g., is there a relationship between X and Y ?*
- Standard errors can also be used to perform hypothesis tests on the coefficients.
- The most common hypothesis test involves testing the null hypothesis of

$$H_0 : \text{There is no relationship between X and Y}$$

versus alternative hypothesis

$$H_a : \text{There is some relationship between X and Y.}$$

Inference: Hypothesis Tests

- Instead of "estimating" the parameters we may be interested in testing a hypothesis, *e.g., is there a relationship between X and Y ?*
- Standard errors can also be used to perform hypothesis tests on the coefficients.
- The most common hypothesis test involves testing the null hypothesis of

$$H_0 : \text{There is no relationship between X and Y}$$

versus alternative hypothesis

$$H_a : \text{There is some relationship between X and Y.}$$

- Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0 \text{ versus } H_a : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model becomes $Y = \beta_0 + \epsilon$, and X is not associated with Y.

Inference: Hypothesis Tests

- Intuitively, if $\hat{\beta}_1$ is far from 0, we are more confident that $\beta_1 \neq 0$.

Inference: Hypothesis Tests

- Intuitively, if $\hat{\beta}_1$ is far from 0, we are more confident that $\beta_1 \neq 0$.
- How far is enough to reject the null hypothesis ?

Inference: Hypothesis Tests

- Intuitively, if $\hat{\beta}_1$ is far from 0, we are more confident that $\beta_1 \neq 0$.
- How far is enough to reject the null hypothesis ?
- This depends on the accuracy of $\hat{\beta}_1$, i.e. on $SE(\hat{\beta}_1)$.

Inference: Hypothesis Tests

- Intuitively, if $\hat{\beta}_1$ is far from 0, we are more confident that $\beta_1 \neq 0$.
- How far is enough to reject the null hypothesis ?
- This depends on the accuracy of $\hat{\beta}_1$, i.e. on $SE(\hat{\beta}_1)$.
- Mathematically, we compute a **t-statistic**

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)},$$

where a large value of $|t|$ tends to reject the null hypothesis.

Inference: Hypothesis Tests

- Intuitively, if $\hat{\beta}_1$ is far from 0, we are more confident that $\beta_1 \neq 0$.
- How far is enough to reject the null hypothesis ?
- This depends on the accuracy of $\hat{\beta}_1$, i.e. on $SE(\hat{\beta}_1)$.
- Mathematically, we compute a **t-statistic**

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)},$$

  where a large value of $|t|$ tends to reject the null hypothesis.
- This has a t-distribution with $n-2$ degrees of freedom, when $\beta_1 = 0$.

Inference: Hypothesis Tests

- Intuitively, if $\hat{\beta}_1$ is far from 0, we are more confident that $\beta_1 \neq 0$.
- How far is enough to reject the null hypothesis ?
- This depends on the accuracy of $\hat{\beta}_1$, i.e. on $SE(\hat{\beta}_1)$.
- Mathematically, we compute a **t-statistic**

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)},$$

  where a large value of $|t|$ tends to reject the null hypothesis.
- This has a t-distribution with $n - 2$ degrees of freedom, when $\beta_1 = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger when $\beta_1 = 0$. We call this probability the **p-value**.

Inference: Hypothesis Tests

- Intuitively, if $\hat{\beta}_1$ is far from 0, we are more confident that $\beta_1 \neq 0$.
- How far is enough to reject the null hypothesis ?
- This depends on the accuracy of $\hat{\beta}_1$, i.e. on $SE(\hat{\beta}_1)$.
- Mathematically, we compute a **t-statistic**

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)},$$

  where a large value of $|t|$ tends to reject the null hypothesis.
- This has a t-distribution with $n - 2$ degrees of freedom, when $\beta_1 = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger when $\beta_1 = 0$. We call this probability the **p-value**.
- A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true.

Inference: Hypothesis Tests

- Intuitively, if $\hat{\beta}_1$ is far from 0, we are more confident that $\beta_1 \neq 0$.
- How far is enough to reject the null hypothesis ?
- This depends on the accuracy of $\hat{\beta}_1$, i.e. on $SE(\hat{\beta}_1)$.
- Mathematically, we compute a **t-statistic**

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)},$$

  where a large value of $|t|$ tends to reject the null hypothesis.
- This has a t-distribution with $n - 2$ degrees of freedom, when $\beta_1 = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger when $\beta_1 = 0$. We call this probability the **p-value**.
- A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true.
- In most applications, we reject the null hypothesis if the p-value ≤ 0.05.

Inference: Hypothesis Tests

- Intuitively, if $\hat{\beta}_1$ is far from 0, we are more confident that $\beta_1 \neq 0$.
- How far is enough to reject the null hypothesis ?
- This depends on the accuracy of $\hat{\beta}_1$, i.e. on $SE(\hat{\beta}_1)$.
- Mathematically, we compute a **t-statistic**

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)},$$

  where a large value of $|t|$ tends to reject the null hypothesis.
- This has a t-distribution with $n-2$ degrees of freedom, when $\beta_1 = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger when $\beta_1 = 0$. We call this probability the **p-value**.
- A p-value measures the probability of obtaining the observed results, assuming that the null hypothesis is true.
- In most applications, we reject the null hypothesis if the p-value ≤ 0.05.
- **We reject the null hypothesis $\neq$ we accept the alternative !**

Results for Advertising Data

For the Advertising data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of $1,000 in the TV advertising budget is associated with an increase in sales by around 50 units. (Recall that the sales variable is in thousands of units, and the TV variable is in thousands of dollars.)

	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Results for Advertising Data

For the Advertising data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of $1,000 in the TV advertising budget is associated with an increase in sales by around 50 units. (Recall that the sales variable is in thousands of units, and the TV variable is in thousands of dollars.)

	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

p-values are smaller than 0.05, so that we reject the null hypothesis $\beta_0 = 0$ and $\beta_1 = 0$.

Results for Advertising Data

For the Advertising data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of $1,000 in the TV advertising budget is associated with an increase in sales by around 50 units. (Recall that the sales variable is in thousands of units, and the TV variable is in thousands of dollars.)

	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

p-values are smaller than 0.05, so that we reject the null hypothesis $\beta_0 = 0$ and $\beta_1 = 0$.
(We reject the null hypothesis $\beta_0 = 0$ and $\beta_1 = 0$)$\neq$ (We accept the hypothesis $\beta_0 \neq 0$ and $\beta_1 \neq 0$).

Assessing the Accuracy of the Model: *RSE*

- Recall from the model that associated with each observation is an error term ϵ.

Assessing the Accuracy of the Model: *RSE*

- Recall from the model that associated with each observation is an error term ϵ.
- Due to the presence of these error terms, even if we knew the true β_0 and β_1 we would not be able to perfectly predict Y from X.

Assessing the Accuracy of the Model: *RSE*

- Recall from the model that associated with each observation is an error term ϵ.
- Due to the presence of these error terms, even if we knew the true β_0 and β_1 we would not be able to perfectly predict Y from X.
- The residual standard error is an estimate of the standard deviation of ϵ:

$$RSE = \sqrt{RSS/(n-2)} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where the residual sum of squares is $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Assessing the Accuracy of the Model: *RSE*

- Recall from the model that associated with each observation is an error term ϵ.
- Due to the presence of these error terms, even if we knew the true β_0 and β_1 we would not be able to perfectly predict Y from X.
- The residual standard error is an estimate of the standard deviation of ϵ:

$$RSE = \sqrt{RSS/(n-2)} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

  where the residual sum of squares is $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
- The RSE is considered a measure of the lack of fit of the model to the data.

Assessing the Accuracy of the Model: *RSE*

- Recall from the model that associated with each observation is an error term ϵ.
- Due to the presence of these error terms, even if we knew the true β_0 and β_1 we would not be able to perfectly predict Y from X.
- The residual standard error is an estimate of the standard deviation of ϵ:

$$RSE = \sqrt{RSS/(n-2)} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

  where the residual sum of squares is $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
- The RSE is considered a measure of the lack of fit of the model to the data.
- If the predictions obtained using the model are very close to the true outcome values, *RSE* will be small, and we can conclude that the model fits the data very well.

Results for Advertising Data

For the Advertising data, more information about the least squares model for the regression of number of units sold on TV advertising budget.

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

Results for Advertising Data

For the Advertising data, more information about the least squares model for the regression of number of units sold on TV advertising budget.

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

RSE is 3.26 means actual sales deviate from the true regression line by $\approx$ 3,260 units, on average.

Results for Advertising Data

For the Advertising data, more information about the least squares model for the regression of number of units sold on TV advertising budget.

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

RSE is 3.26 means actual sales deviate from the true regression line by $\approx$ 3,260 units, on average.

If the model were correct and the true values of the unknown coefficients were known exactly, any prediction of sales on the basis of TV would still be off by about 3,260 units on average.

Assessing the Accuracy of the Model: R^2

- The RSE provides an absolute measure of lack of fit of the model to the data.

Assessing the Accuracy of the Model: R^2

- The RSE provides an absolute measure of lack of fit of the model to the data.
- Since it is measured in the units of Y , it is not always clear what constitutes a good RSE.

Assessing the Accuracy of the Model: R^2

- The RSE provides an absolute measure of lack of fit of the model to the data.
- Since it is measured in the units of Y , it is not always clear what constitutes a good RSE.
- **R-squared** or fraction of variance explained is

$$R^2 = \frac{TSS - RSS}{TSS} \in [0, 1],$$

where $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares.

Assessing the Accuracy of the Model: R^2

- The RSE provides an absolute measure of lack of fit of the model to the data.
- Since it is measured in the units of Y , it is not always clear what constitutes a good RSE.
- **R-squared** or fraction of variance explained is

$$R^2 = \frac{TSS - RSS}{TSS} \in [0, 1],$$

  where $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the total sum of squares.
- TSS measures the total variance in the response Y , and can be thought of as the amount of variability inherent in the response before the regression is performed.

Assessing the Accuracy of the Model: R^2

- The RSE provides an absolute measure of lack of fit of the model to the data.
- Since it is measured in the units of Y , it is not always clear what constitutes a good RSE.
- **R-squared** or fraction of variance explained is

$$R^2 = \frac{TSS - RSS}{TSS} \in [0, 1],$$

  where $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the total sum of squares.
- TSS measures the total variance in the response Y , and can be thought of as the amount of variability inherent in the response before the regression is performed.
- RSS measures the amount of variability that is left unexplained after performing the regression.

R^2: Interpretation

- R^2 measures the prop. of variability in Y that can be explained using X.

R^2: Interpretation

- R^2 measures the prop. of variability in Y that can be explained using X.
- An R^2 *close to* 1 indicates that a large proportion of the variability in the response has been explained by the regression.

R^2: Interpretation

- R^2 measures the prop. of variability in Y that can be explained using X.
- An R^2 *close to* 1 indicates that a large proportion of the variability in the response has been explained by the regression.
- An R^2 *close to* 0 indicates that the regression does not explain much of the variability in the response; this might occur because the linear model is wrong, or the error variance is high.

R^2: Interpretation

- R^2 measures the prop. of variability in Y that can be explained using X.
- An R^2 *close to* 1 indicates that a large proportion of the variability in the response has been explained by the regression.
- An R^2 *close to* 0 indicates that the regression does not explain much of the variability in the response; this might occur because the linear model is wrong, or the error variance is high.
- The R^2 statistic has an interpretational advantage over the RSE, since unlike the RSE, it always lies between 0 and 1.

R^2: Interpretation

- R^2 measures the prop. of variability in Y that can be explained using X.
- An R^2 *close to* 1 indicates that a large proportion of the variability in the response has been explained by the regression.
- An R^2 *close to* 0 indicates that the regression does not explain much of the variability in the response; this might occur because the linear model is wrong, or the error variance is high.
- The R^2 statistic has an interpretational advantage over the RSE, since unlike the RSE, it always lies between 0 and 1.
- However, it can still be challenging to determine what is a good R^2.

R^2: Interpretation

- R^2 measures the prop. of variability in Y that can be explained using X.
- An R^2 *close to* 1 indicates that a large proportion of the variability in the response has been explained by the regression.
- An R^2 *close to* 0 indicates that the regression does not explain much of the variability in the response; this might occur because the linear model is wrong, or the error variance is high.
- The R^2 statistic has an interpretational advantage over the RSE, since unlike the RSE, it always lies between 0 and 1.
- However, it can still be challenging to determine what is a good R^2.
- Large value of R^2 does **NOT** mean the model fits the data well. It favors more flexible models, which may overfit the data! .

Multiple Linear Regression

- Simple linear regression is a useful approach for predicting a response on the basis of a single predictor variable.

Multiple Linear Regression

- Simple linear regression is a useful approach for predicting a response on the basis of a single predictor variable.
- In practice we often have more than one predictor.

Multiple Linear Regression

- Simple linear regression is a useful approach for predicting a response on the basis of a single predictor variable.
- In practice we often have more than one predictor.
- One option is to run separate simple linear regressions, each of which uses a different predictor.

Multiple Linear Regression

- Simple linear regression is a useful approach for predicting a response on the basis of a single predictor variable.
- In practice we often have more than one predictor.
- One option is to run separate simple linear regressions, each of which uses a different predictor.
- However, the approach of fitting a separate simple linear regression model for each predictor is not entirely satisfactory.

Multiple Linear Regression

- Simple linear regression is a useful approach for predicting a response on the basis of a single predictor variable.
- In practice we often have more than one predictor.
- One option is to run separate simple linear regressions, each of which uses a different predictor.
- However, the approach of fitting a separate simple linear regression model for each predictor is not entirely satisfactory.
- Why ?

Multiple Linear Regression

- We now consider

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon,$$

where X_j represents the jth predictor and β_j quantifies the association between that variable and the response.

Multiple Linear Regression

- We now consider

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon,$$

where X_j represents the jth predictor and β_j quantifies the association between that variable and the response.

- We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed. In the advertising example, the model becomes

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper + \epsilon.$$

Multiple Linear Regression

- We now consider

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon,$$

where X_j represents the jth predictor and β_j quantifies the association between that variable and the response.

- We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed. In the advertising example, the model becomes

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper + \epsilon.$$

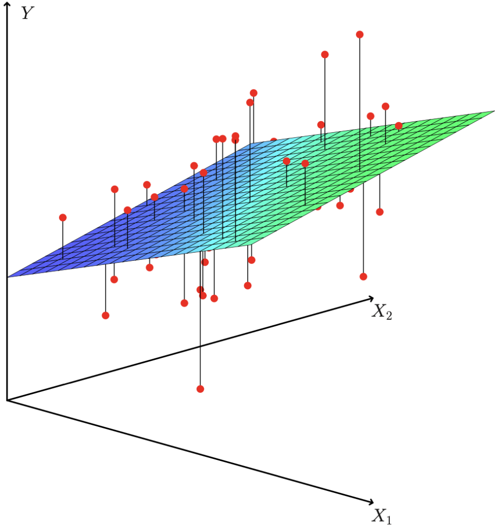
- Compared to the simple linear regression,

$$sales = \alpha_0 + \alpha_1 \times TV + \epsilon',$$

in general $\beta_1 \neq \alpha_1$, since α_1 represents the average effect on sales of a one unit increase in TV.

Multiple Linear Regression

In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.



Some Important Questions

- Is at least one of the predictors $X_1, X_2, ..., X_p$ useful in predicting the response?

Some Important Questions

- Is at least one of the predictors $X_1, X_2, ..., X_p$ useful in predicting the response?
- Do all the predictors help to explain Y, or is only a subset of the predictors useful?

Some Important Questions

- Is at least one of the predictors $X_1, X_2, ..., X_p$ useful in predicting the response?
- Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- How well does the model fit the data?

Some Important Questions

- Is at least one of the predictors $X_1, X_2, ..., X_p$ useful in predicting the response?
- Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Relationship Between Response and Predictors?

- In the simple linear regression: test if $\beta_0 = 0$.

Relationship Between Response and Predictors?

- In the simple linear regression: test if $\beta_0 = 0$.
- In the multiple linear regression: Test the null hypothesis

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0 \text{ versus } H_a : \text{at least one } \beta_j \text{ is not-zero.}$$

Relationship Between Response and Predictors?

- In the simple linear regression: test if $\beta_0 = 0$.
- In the multiple linear regression: Test the null hypothesis

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0 \text{ versus } H_a : \text{at least one } \beta_j \text{ is not-zero.}$$

- This hypothesis test is performed by computing the **F-statistic**

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)},$$

where recall that $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ and $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Relationship Between Response and Predictors?

- In the simple linear regression: test if $\beta_0 = 0$.
- In the multiple linear regression: Test the null hypothesis

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0 \text{ versus } H_a : \text{at least one } \beta_j \text{ is not-zero.}$$

- This hypothesis test is performed by computing the **F-statistic**

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)},$$

where recall that $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ and $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

- A very large value of F favors H_a.

Deciding on Important Variables

- If we conclude that at least one of the predictors is related to the response, then it is natural to wonder which are the guilty ones!

Deciding on Important Variables

- If we conclude that at least one of the predictors is related to the response, then it is natural to wonder which are the guilty ones!
- It is possible that all of the predictors are associated with the response, but it is more often the case that the response is only associated with a subset of the predictors.

Deciding on Important Variables

- If we conclude that at least one of the predictors is related to the response, then it is natural to wonder which are the guilty ones!
- It is possible that all of the predictors are associated with the response, but it is more often the case that the response is only associated with a subset of the predictors.
- This task is called **variable selection**.

Deciding on Important Variables

- If we conclude that at least one of the predictors is related to the response, then it is natural to wonder which are the guilty ones!
- It is possible that all of the predictors are associated with the response, but it is more often the case that the response is only associated with a subset of the predictors.
- This task is called **variable selection**.
- The most direct approach is **best subsets regression**: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.

Deciding on Important Variables

- If we conclude that at least one of the predictors is related to the response, then it is natural to wonder which are the guilty ones!
- It is possible that all of the predictors are associated with the response, but it is more often the case that the response is only associated with a subset of the predictors.
- This task is called **variable selection**.
- The most direct approach is **best subsets regression**: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.
- The criterion include Mallow's C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted R^2. These will be discussed later.

Deciding on Important Variables

- If we conclude that at least one of the predictors is related to the response, then it is natural to wonder which are the guilty ones!
- It is possible that all of the predictors are associated with the response, but it is more often the case that the response is only associated with a subset of the predictors.
- This task is called **variable selection**.
- The most direct approach is **best subsets regression**: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.
- The criterion include Mallow's C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted R^2. These will be discussed later.
- However we often can't examine all possible models, since they are 2^p of them; for example when $p = 40$ there are over a billion models! Instead we need an automated approach that searches through a subset of them. We discuss two commonly use approaches next.

Forward selection

- Begin with the null model– a model that contains an intercept but no predictors.

Forward selection

- Begin with the null model– a model that contains an intercept but no predictors.
- Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.

Forward selection

- Begin with the null model– a model that contains an intercept but no predictors.
- Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.
- Add to that model the variable that results in the lowest RSS amongst all two-variable models.

Forward selection

- Begin with the null model– a model that contains an intercept but no predictors.
- Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.
- Add to that model the variable that results in the lowest RSS amongst all two-variable models.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

Backward selection

- Start with all variables in the model.

Backward selection

- Start with all variables in the model.
- Remove the variable with the largest p-value – that is, the variable that is the least statistically significant.

Backward selection

- Start with all variables in the model.
- Remove the variable with the largest p-value – that is, the variable that is the least statistically significant.
- he new $(p-1)$ –variable model is fit, and the variable with the largest p-value is removed.

Backward selection

- Start with all variables in the model.
- Remove the variable with the largest p-value – that is, the variable that is the least statistically significant.
- he new $(p-1)$ –variable model is fit, and the variable with the largest p-value is removed.
- Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value defined by some significance threshold.

Mixed selection

- We start with no variables in the model.

Mixed selection

- We start with no variables in the model.
- As with forward selection, we add variables one-by-one.

Mixed selection

- We start with no variables in the model.
- As with forward selection, we add variables one-by-one.
- After adding a new variable, we check whether the p-value for one of the variables in the model rises above a certain threshold. If yes, we remove that variable from the model.

Mixed selection

- We start with no variables in the model.
- As with forward selection, we add variables one-by-one.
- After adding a new variable, we check whether the p-value for one of the variables in the model rises above a certain threshold. If yes, we remove that variable from the model.
- Continue these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

Comparison

- Backward selection cannot be used if $p > n$, while forward selection can always be used.

Comparison

- Backward selection cannot be used if $p > n$, while forward selection can always be used.
- Forward selection is a greedy approach, and might include variables early that later become redundant.

Comparison

- Backward selection cannot be used if $p > n$, while forward selection can always be used.
- Forward selection is a greedy approach, and might include variables early that later become redundant.
- Mixed selection can remedy this.

Model Fit

- RSE and R^2 can be still used for multiple linear regression.

Model Fit

- RSE and R^2 can be still used for multiple linear regression.
- R^2 favors more flexible models, as R^2 will always increase when more variables are added to the model, even if those variables are only weakly associated with the response.

Prediction

A 95% **Prediction interval** for Y refers to that the interval of this form will contain the true value Y with 95% probability. Let

$$\hat{Y} = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_p X_p.$$

We have

$$Y - \hat{Y} = f(X) - \hat{f}(X) + \epsilon.$$

To construct a prediction interval, we need to first get a confidence interval for $f(X) - \hat{f}(X)$ and then add the variance of ϵ to the confidence interval.

Thus, the prediction interval is usually substantially wider than the confidence interval.