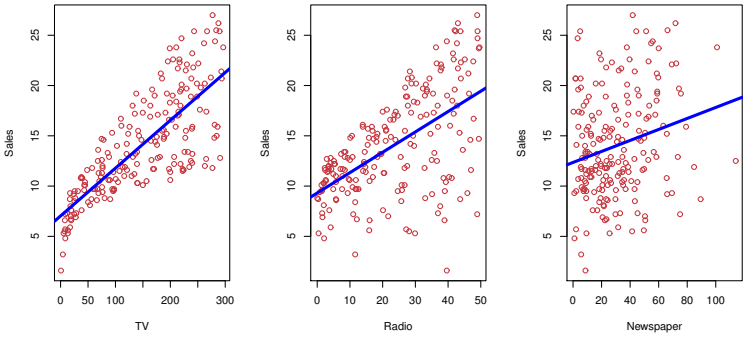
Linear regression

December 29, 2024

Table of contents

Linear regression

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on $x_1, x_2, x_3, ...x_p$ is linear.
- Consider the following advertising data:



- Linear regression is extremely useful both conceptually and practically and it is often a starting point for the more advanced methods.

Linear regression

- Questions that we might want to answer:
  - Is there a relationship between advertising budget and sales?
  - How strong is the relationship between advertising budget and sales?
  - Which media contributes to sales?
  - How accurately can we predict future sales?
  - Is the relationship linear?
  - Is there synergy among the advertising media?
- Tools we consider to answer these questions:
  1. Simple Linear Regression (SLR)
  2. Multiple Linear Regression (MLR)
- Throughout this lecture we will make assumptions regarding the SLR & the MLR model (OLS Assumptions) to derive appropriate estimates and procedures to answer the questions above.

Simple linear regression

Simple linear regression

- Assume a model:

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where β_0 and β_1 are two unknown constants that represent the *intercept* and *slope*, also known as *coefficients* or *parameters*, and ϵ is the error term.

OLS Assumption No. 1: Linearity

The regression model is linear in the coefficients and the error term

- Linearity: a one unit change in X has the same effect on Y, regardless of the initial value of X (*Note:* unrealistic for many applications)
- To obtain reliable estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1 from a random sample, we first need to make assumptions about the error term. (Note: The *hat* symbol denotes an estimated value.)

The role of the error term - I

- In general: the error term accounts for the variation in the dependent variable that is not captured by the independent variable, i.e. it accounts for *not-predictable* variation.

OLS Assumption No. 2: Zero Mean

The error term has a population mean of zero. That is $\mathbb{E}(\epsilon) = 0$.

- *Intuition*: If the error term has mean zero this implies that the regression coefficients are unbiased.
- Imagine the average error is -5: We systematically over predict the independent variable and the model is misspecified. Part of the error term is predictable, which should be added to the regression model
- Note: β_0 ensures that the mean of the error terms is always zero.

The role of the error term - II

OLS Assumption No. 3.1: Independence/ Exogeneity Assumption

All independent variables are uncorrelated with the error term

- *Intuition*: If violated, we can use the independent variable to predict the error term. Thus the model is misspecified.
- Violations are due to omitted variables or measurement errors in the independent variables.
- Regression estimates are biased, as the OLS incorrectly attributes some of the variance of the error term to the independent variable

The role of the error term - III

- Note that for *time series* data (as opposed to cross sectional data), you need to ensure the following condition:

OLS Assumption No. 3.2: No serial auto-correlation

The observations of the error term are uncorrelated with each other, that is $cov(\epsilon_t, \epsilon_{t-1}) = 0$

- *Intuition*: Observations of the error term should not predict each other.
  - Positive serial auto-correlation: a positive error is followed by a positive error (vice versa for a negative error)
  - Negative serial auto-correlation: a positive error is followed by a negative error (vice versa for a negative error)
- As before, if there are possibilities to predict the error term than this should be incorporated into the model

Simple linear regression

Coefficient estimation and accuracy

SLR – Coefficient estimation and accuracy

- Now let there be n observations and $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith *residual*.
- Define the *residual sum of squares* (RSS) as:

$$\text{RSS} = e_1^2 + e_2^2 + ... + e_n^2$$

or equivalently

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + ... + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

- Then choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} Q = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{1}$$

SLR – Coefficient estimation and accuracy

- Build First Order Conditions from 1:

$$\frac{\partial Q}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} -2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{2}$$

$$\frac{\partial Q}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} -2x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{3}$$

- solving for $\hat{\beta}_0$ and $\hat{\beta}_1$ leads to:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{cov(X, Y)}{var(X)} \tag{4}$$

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

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

SLR – Coefficient estimation and accuracy

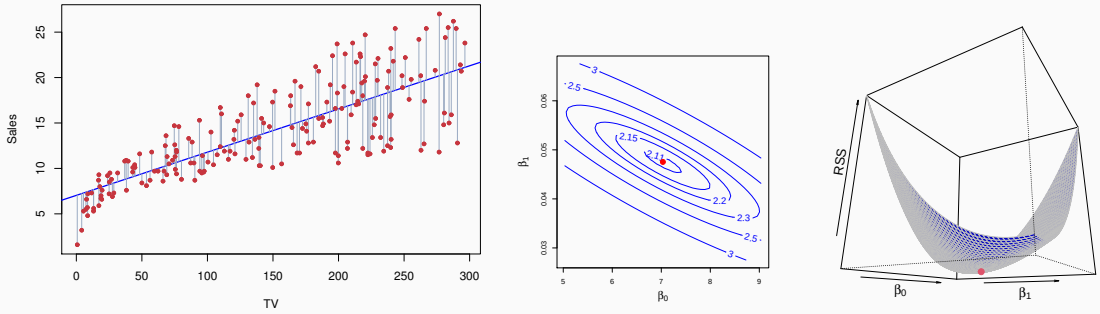


Figure 1: For the Advertising data, the least squares fit for the regression of sales onto TV is shown. The fit is found by minimizing the residual sum of squares. Each grey line segment represents a residual. In this case a linear fit captures the essence of the relationship, although it overestimates the trend in the left of the plot. Also shown: Contour and three-dimensional plots of the RSS on the Advertising data, using sales as the response and TV as the predictor. The red dots correspond to the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

SLR – Coefficient estimation and accuracy

- Given the estimated parameter, we can interpret them as followed:
  - β_0 : expected value of Y, when $X = 0$
  - β_1 : *average* increase of Y when X increases by one unit
- **BUT**: how accurate are our estimates that we obtained from (4)?

SLR – Coefficient estimation and accuracy

- True relationship: $Y = f(X) + \epsilon$, where $f(X) = 2 + 3X$
- Create 100 random X and generate 100 random Y's

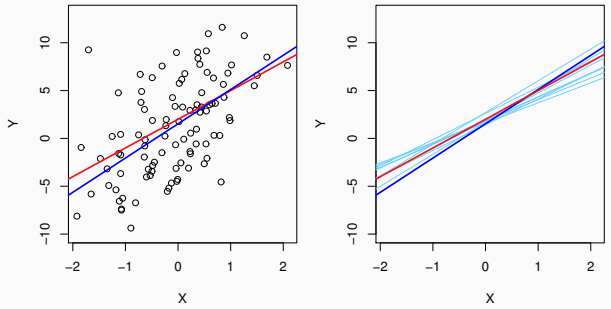


Figure 2: Left: Red line represents the true relationship(population regression line). The blue line is the least squares line; it is the least squares estimate for $f(X)$. Right: Population regression line again in red, and the least squares line in dark blue. In light blue, ten least squares lines are shown, 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.

SLR – Coefficient estimation and accuracy

- The OLS estimates are said to be *unbiased*, i.e. they do not systematically over- or under-estimate the true parameters.
- **BUT**: Given one set of observations, $\hat{\beta}_0$ and $\hat{\beta}_1$ might over or understate β_0 and β_1, respectively (see figure before).
- To establish a measure that captures the variation of the estimates across samples we need the following assumption:

OLS Assumption No. 4: **Homoskedasticity**

The error term ϵ has the same variance given any value of the explanatory variable. Said differently: $var(\epsilon \mid X) = \sigma^2$

- *Intutition:*
  - The linear regression model estimates the parameters in such a way that it will fit as many data points as possible.
  - With heteroskedastic data, some points are more spread-out than others. As in OLS all points are treated the same, it will drag the regression curve towards those with higher variance.

SLR – Coefficient estimation and accuracy

- Then use the standard error of the estimate to determine how close the estimates are to the true parameters, i.e. to asses the *accuracy* of the estimate:

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{5}$$

$$SE(\hat{\beta}_0)^2 = \frac{\sigma^2}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{6}$$

  where $\sigma^2 = var(\epsilon)$ and where an unbiased estimate for $var(\epsilon)$ is given by $\frac{RSS}{n-2} = \frac{1}{(n-2)}\sum_{i=1}^{n}\hat{\epsilon}_i$

- Interpretation: The standard error represents the average amount by which the estimate (i.e., $\hat{\beta}_0$ or $\hat{\beta}_1$) differ from the true values (i.e., β_0 and β_1), respectively.

SLR – Coefficient estimation and accuracy

- So far we know the first two moments of our estimates β_1 and β_2. However, to perform statistical inference, we need to know the exact distribution of the two estimates.

OLS Assumption No. 5: Normality Assumption

The error term ϵ is independent of the explanatory variables $x_1, x_2, ..., x_n$ and is normally distributed with mean zero and variance σ^2: $\epsilon \sim N(0, \sigma)$.

- *Intuition:*
  - Note that the error term captures multiple disturbances that are not measurable and thus are not included in the model.
  - Further these disturbances are additive, as we assume a linear, additive model. Recall that the Central Limit Theorem (CLT) states that if there are a large number of i.i.d. variables, then their sum tends to be normally distributed.

SLR – Coefficient estimation and accuracy

- Under the assumptions 1-5, we can compute:
  - Confidence Intervals
  - Hypothesis Tests

- Confidence Intervals
  - For example, the 95% confidence interval is *approximately* given by:

$$\hat{\beta}_1 \pm 2 \times SE(\hat{\beta}_1)$$

  - There is approximately a 95% chance that the interval will contain the true value of β_1.

SLR – Coefficient estimation and accuracy

- Hypothesis Tests
  - H_0 : There is no relationship between X and Y, i.e. $(\beta_1 = 0)$
  - H_A : There is some relationship between X and Y, i.e. $(\beta_1 \neq 0)$

    Note: if $\beta_1 = 0$, then the model reduces to $Y = \beta_0 + \epsilon$ and X is not associated with Y.
  - To test the null hypothesis, we compute a t-statistic, given by:

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

  - This will have a t-distribution with $n - 2$ degrees of freedom, assuming $\beta_1 = 0$
  - Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the *p-value*.

    $\Rightarrow$ Always remember: If p-values are low, H_0 must go.

A small example...

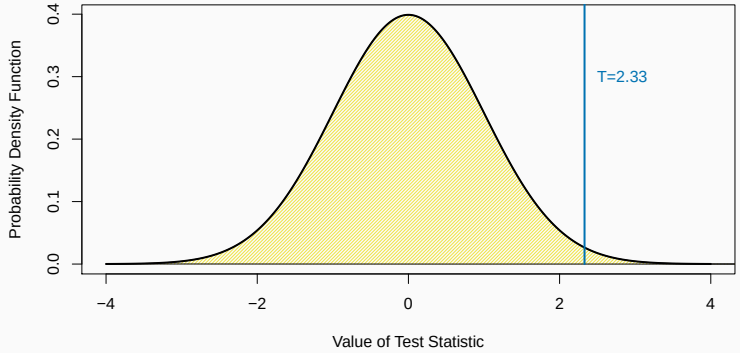


Figure 3: The density function for the N(0, 1) distribution, with the vertical line indicating a value of 2.33. 1% of the area under the curve falls to the right of the vertical line, so there is only a 2% chance of observing a $N(0, 1)$ value that is greater than 2.33 or less than −2.33. Therefore, if a test statistic has a N(0, 1) null distribution, then an observed test statistic of T = 2.33 leads to a p-value of 0.02.

Looking at the advertising data, recall the model: $\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \epsilon$
Results:

	Coefficient	Std. error	t-statistics	p-value
`Intercept`	7.0325	0.4578	15.36	<0.0001
`TV`	0.0475	0.0027	17.76	<0.0001

Table 1: For the Advertising data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget.

Questions:

a) Provide an interpretation of each coefficient in the model (Recall that sales is in thousands).

b) Interpret the standard error for β_1.

c) Refer back to formula (5) above: explain how σ^2, n and an increase in $(x_i - \bar{x})^2$ impacts the standard error.

d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$, using $\alpha = 5\%$?

Multiple linear regression

Multiple linear regression

- In practice we often need more than one predictor.
- We can extend our Simple Linear Regression model to account for more variables that affect our response Y.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon, \tag{7}$$

- Here we interpret β_j as the *average* effect on Y of a one unit increase in X_j, holding all other predictors fixed.
- For our advertising example, the model is then:

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon$$

Multiple linear regression

Coefficient estimation

- To obtain valid estimates in the MLR setting, we need to clarify the relationship of the independent variable among each other:

OLS Assumption No. 6: No Perfect Collinearity in the MLR model

There is no perfect collinearity among the independent variables, i.e. no independent variable is a perfect linear combination of all the others.

- *Intuition:*
  - Perfect collinearity appears when one variable moves in perfect unity to the other variable - think Fahrenheit and Celcius.
  - OLS then cannot distinguish between these two variables and will not be able to provide an estimate.
- Even when correlations among independent variables are less than perfect (e.g. ± 0.7), OLS will have problems.
  - The variance of all coefficients tends to increase, sometimes dramatically.
  - Interpretations become hazardous - when X_j changes, everything else changes.
  - What to do? $\Rightarrow$Later more!

MLR – Coefficient estimation

- As before, we estimate $\beta_0,\ \beta_1,\ ...,\ \beta_p$ as the values that minimize the sum of squared residuals:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

$$= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_1 x_{i2} - ... - \hat{\beta}_p x_{ip})^2 \tag{9}$$

- Using matrix notation and algebra, the vector of regression parameters is given as:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

- Note that all assumptions and intuitions from the SLR model transfer naturally to the MLR model.

MLR – Coefficient estimation

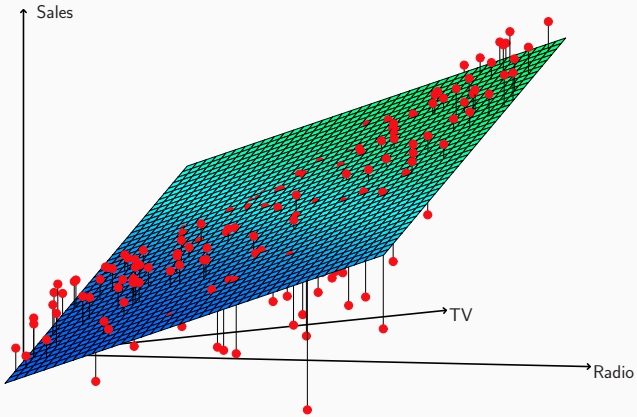


Figure 4: In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.

Recall the model: $\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon$

	Coefficient	Std. error	t-statistics	p-value
Intercept	2.939	0.3119	9.42	<0.0001
TV	0.046	0.0014	32.81	<0.0001
radio	0.189	0.0086	21.89	<0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Table 2: For the Advertising data, least squares coefficient estimates of the multiple linear regression of number of units sold on TV, radio, and newspaper advertising budgets.

Questions:

a) Provide an interpretation of the coefficient for `radio` in the model.

b) Provide an approximate 95% Confidence Interval for the `TV` coefficient and interpret the results.

c) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$, using $\alpha = 5\%$?

Now assume you run the following model: $\texttt{sales} = \beta_0 + \beta_1 \times \texttt{newspaper} + \epsilon$

	Coefficient	Std. error	t-statistics	p-value
`Intercept`	12.351	0.621	19.88	<0.0001
`newspaper`	0.055	0.017	3.30	0.00115

Table 3: Summary results for the simple linear regression of number of units sold on newspaper.

Questions:

a) Given your results for the simple and the multiple linear regression from the previous slide. What can you say about the association between `newspaper` and `sales`.

Model selection

Model selection

- Given our fitted model, we need to answer some important questions:
  1. How well does the model fit the data?
  2. Is at least one of the predictors X_1, X_2, ..., X_p useful in predicting the response?
  3. Do all of the predictors help to explain Y, or is only a subset of the predictors useful?
  4. Given a set of predictor values, what response values should we predict and how accurate is our prediction?

Model selection – Q1

Q1: How well does the model fit the data?

1. Residual Standard Error (RSE)
   - The RSE is an estimate of the standard deviation of ϵ:

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

     where $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ are the residual sum-of-squares.
   - Key points:
     - Absolute measure of lack of fit, measured in the units of Y.
     - Low RSE indicates a better model.
     - Interpretation: The RSE represents the average amount that the response Y will deviate from the true regression line, measured in the units of Y.
     - Change in RSE depends on the trade-off between RSS and $\frac{1}{n-p-1}$
     - Adding a parameter that only slightly reduces RSS, might lead to a higher RSE.

Model selection – Q1

2. R-Squared
   - The R-Squared measures the fraction of variation that is explained by the model.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

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

$$R^2 = r^2 = (\frac{Cov(X, Y)}{STD(X) \times STD(Y)})^2$$

   - Key points:
     - $R^2 \in [0, 1]$
     - A good R^2 depends on the problem at hand (e.g., in physics, for which we know that some relationships are linear, the R^2 should be close to 1)
     - R-squared increases as we add more parameters, even though they might only be weakly associated with the response.
     - Why?: Because, adding another variable always results in a decrease in the training RSS.

Example:

Model 1: $\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon$

Model 2: $\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon$

	R^2	RSE
Model 1	0.89719	1.681
Model 2	0.8972	1.686

Questions:

a) Explain the changes in the R^2 and RSE between Model 1 and Model 2. What can you infer from the values about the `newspaper` variable?

Model selection – Q2

Q2: Is at least one of the predictors X_1, X_2, ..., X_p useful in predicting the response?

- Recall the model:

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon \tag{10}$$

- Now test, if a group of variables have no impact on Y, once we have controlled for the other variables
- For example: `radio` and `newspaper` have no effect on `sales`, once we accounted (i.e. included) for `TV` in the model.
- In mathematical terms:

$$H_0 : \beta_2 = 0, \beta_3 = 0 \text{ vs. } H_A : H_0 \text{ is not true} \tag{11}$$

- If the hypothesis cannot be rejected, then `radio` and `newspaper` should be dropped from the model.

Model selection – Q2

- How should we test hypothesis (11)?
  - It might be tempting to use a t-test, but as we will see we actually need to use a so-called *F-Test*
- Why do we need to use a F-Test, when we can just perform a t-test on each single variable?
  - Consider $p = 100$ and assume $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_p = 0$ in reality is true.
  - Now you perform a t-test for each variable at $\alpha = 5\%$
  - By chance: 5% of the p-values will be below 5%, i.e. we expect to see 5 small p-values and incorrectly conclude that these 5 parameters characterize the model (i.e. they would be added to the model)
  - This is a problem of Multiple Testing, which we will cover in a later lecture.
  - For now: the F-Test does not suffer from this problem, as it adjusts to the number of parameters used.

Model selection – Q2

- The F-Test is based on the Residual Sum of Squares (RSS) of the model with all parameters (*unrestricted* model) and without the parameters (*restricted* model) (in the example above without β_2 & β_3).
- *Intuition:* When the RSS of the restricted model is *much larger than* the RSS of the unrestricted model, the null hypothesis has to be rejected.
- In general:
  - If we only want to test, whether a subset of q of the coefficients are zero, i.e.:

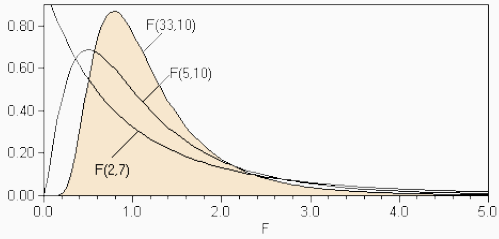
$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = ... = \beta_p = 0 \text{ vs.}$$
$$H_A : \text{at least one of the } p-1 \text{ coefficients is non-zero}$$

  - Fit a second model that uses all the variables except those q and define the RSS of that model as RSS_0, then the F-statistics is:

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)} \text{ with } F \sim F_{q,n-p-1}$$

  - Note that the F-Statistic disregarding the *qth* variable is equal to the squared t-statistics of that variable. It reports the partial effect of adding this variable to the model.

Model selection – Q2



V1 \ V2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63

Figure 5: The F-Distribution and Table for various degrees of freedom. Source: `http://www.statistics4u.info/fundstat_germ/cc_distri_fisher_f.html` and `https://www.oreilly.com/library/view/making-sense-of/9780470074718/appa-sec004.html`

Model selection – Q2

- If we want to test, whether any of the independent variables are associated with the response, then we can perform a general F-Test
- That is:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0 \text{ vs. } H_A\text{: } \textit{at least one } \beta_j \textit{ is non-zero}$$

- And the F-Test Statistic is given by:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} = \frac{R^2/p}{(1-R^2)/(n-p-1)} \sim F_{p,n-p-1}$$

- F-Statistic close to 1: No relationship between response and predictors.
- If H_A is true then the F-statistics is greater than 1.
- When is the F-Statistic large enough?
  - Depends on n and p. Larger F-Statistic is needed when n is small relative to p (see previous slide).
  - Statistical software can be used to obtain the p-value for the F-Statistic.

Recall the model: $\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon$

Quantity	Value
Residual Standard Error	1.69
R^2	0.897
F-statistics	570

Questions:

a) Provide an interpretation of the RSE of the model.

b) Provide an interpretation of the R^2 of the model.

c) Can you reject the null hypothesis that all coefficients are equal to zero?

Model selection – Q3

Q3: Do all of the predictors help to explain Y, or is only a subset of the predictors useful?

- Very often the response is only associated with a subset of all the predictors p.
- Selecting only those predictors with are associated with the response is called *variable selection*
- Three methods:
  1. All subset or Best subset regression
  2. Forward Selection
  3. Backward Selection
- More on these methods in *Lecture 06 - Model Selection*

Model selection – Q4

Q4: Given a set of predictor values, what response values should we predict and how accurate is our prediction?

- Having fit the model, we can make a prediction for Y on the basis of $X_1, X_2, ..., X_p$
- Three sorts of uncertainty:
  1. The estimated model is only an estimate of the true population regression line.
  2. *Model Bias*: Assuming a linear model might be too simplistic to capture the real-life dependencies
  3. Even if we knew the true $f(x)$ the response Y cannot be predicted perfectly, due to ϵ. Remember Lecture 1: this is the irreducible error.

Model selection – Q4

Example:

- Recall the model: $sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \epsilon$
- How much will $\hat{Y}$ vary from Y?
  1) Assume \$10 0000 is spent on `TV`, and \$20 000 is spent on `radio` in *each* city:
     - Forecasted Value: 11 256 units sold (from the model)
     - 95% Confidence Interval (CIs): $[10985, 11528]$
  2) Assume \$10 0000 is spent on `TV`, and \$20 000 is spent on `radio` in *one* city:
     - Forecasted Value: 11 256 units sold (from the model)
     - 95% Prediction Interval (PIs): $[7930, 14580]$
- Prediction Intervals are always wider than Confidence Intervals, as they incorporate both: the reducible and the irreducible error.
- Note: Computing CIs and PIs for the response is tedious to do by hand and should normally be done using a statistical software.

Next steps

- Until now, we looked at Simple and Multiple Linear Regression models.
- We now turn to topics that apply to both model set-ups:
  - Qualitative Predictors
  - Extensions to the Linear Models
  - Potential Problems of Linear Models

Qualitative predictors

Qualitative predictors

- Some predictors are not quantitative, but are qualitative, i.e. they take a discrete set of values.
- These are also called *categorical predictors* or *factor variables*.
- See for example the scatterplot matrix of the credit card data in the next slide.
- The response is `balance` (average credit card debt for each individual)
- In addition to the seven quantitative variables shown, there are four qualitative variables: `own` (house ownership), `student` (student status), `status` (martial status), and `region` (East, West and South).

Qualitative predictors

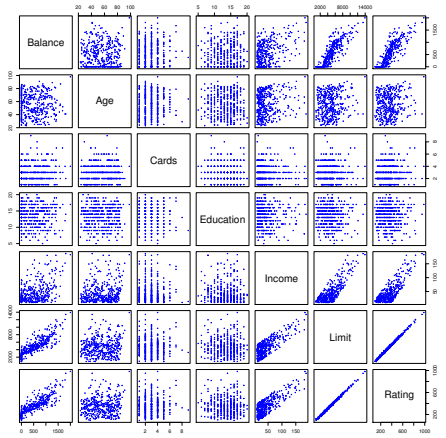


Figure 6: The Credit data set contains information about `balance`, `age`, `cards`, `education`, `income`, `limit`, and `rating` for a number of potential customers.

Qualitative predictors

- Example: investigate differences in credit card balance between those who own a house and those who don't, ignoring the other variables. We create a new variable:

$$x_i = \begin{cases} 1, & \text{if } i\text{th person owns a house} \\ 0, & \text{if } i\text{th person does not own a house} \end{cases}.$$

- Resulting Credit Model 1:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{if } i\text{th person owns a house} \\ \beta_0 + \epsilon_i, & \text{if } i\text{th person does not} \end{cases}.$$

- Interpretation:
  - β_0 is the average credit card balance among those who do not own a house
  - $\beta_0 + \beta_1$ is the average credit card balance for those who do own a house

Results for the Credit Model 1:

	Coefficient	Std. Error	t-statistics	p-value
`Intercept`	509.80	33.13	15.389	<0.0001
`own[Yes]`	19.73	46.05	0.429	0.6690

Table 4: Least squares coefficient estimates for Credit Model 1 associated with the regression of balance onto own in the Credit data set.

Questions:

a) What is the average credit card balance for those who own a house?

b) What is the average credit card balance for those who do *not* own a house?

c) Can you reject the null hypothesis $H_0 : \beta_j = 0$ for any of the predictors, using $\alpha = 5\%$?

Qualitative predictors

- Instead of a 0/1 encoding, we can also create an encoding in the form:

$$x_i = \begin{cases} 1, & \text{if } i\text{th person owns a house} \\ -1, & \text{if } i\text{th person does not own a house} \end{cases}.$$

- Resulting Credit Model 2:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{if } i\text{th person owns a house} \\ \beta_0 - \beta_1 + \epsilon_i, & \text{if } i\text{th person does not} \end{cases}.$$

- Interpretation:
  - β_0 is the overall average credit card balance (ignoring house ownership)
  - β_1 is the amount by which house owners and non-owners have credit card balances that are above or below the average, respectively
- Homework Questions: Given a specific person, do you expect that the estimated credit card balances of Credit Model 1 and Credit Model 2 are the same?

Qualitative predictors

- With more than two levels, we create additional dummy variables. For example, for the region variable we create two dummy variables. The first could be

$$x_{i1} = \begin{cases} 1, & \text{if } i\text{th person is from the South} \\ 0, & \text{if } i\text{th person is not from the South} \end{cases}.$$

and the second could be

$$x_{i2} = \begin{cases} 1, & \text{if } i\text{th person is from the West} \\ 0, & \text{if } i\text{th person is not from the West} \end{cases}.$$

Qualitative predictors

- Then both of these variables can be used in the regression equation, in order to obtain the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \textit{if } i\text{th person is from the South} \\ \beta_0 + \beta_2 + \epsilon_i, & \textit{if } i\text{th person is from the West} \\ \beta_0 + \epsilon_i & \textit{if } i\text{th person is from the East.} \end{cases}.$$

- There will always be one fewer dummy variable than the number of levels. The level with no dummy variable — East in this example — is known as the *baseline*.
- Interpretation
  - β_0: Average Credit Quality for people from the East.
  - β_2: Difference in the average balance between those from the West versus the East.
- Adding a separate dummy variable for each level is known as the *Dummy Variable Trap*.

	Coefficient	Std. Error	t-statistics	p-Value
`Intercept`	531.00	46.32	11.464	< 0.001
`region[South]`	-18.69	65.02	-0.287	0.7740
`region[West]`	-12.50	56.68	-0.221	0.8260

Table 5: Least squares coefficient estimates associated with the regression of balance onto region in the Credit data set.

Questions:

a) What is the average credit card balance for an individual from the East?

b) What is the difference in the average credit card balance for people from the east and the south? What is the difference between East and West?

c) Are all predictors associated with the response?

d) Explain the meaning of the *Dummy Variable Trap*.

Extensions of the linear model

Extensions of the linear model

- The Linear Model is interpretable, but makes highly restrictive assumptions (i.e. additivity and linearity).
- Removing these assumptions leads us to the topics of:
  1. Interaction effects
  2. Non-linearity

Extensions of the linear model

Interaction

Extensions of the linear model – Interactions

- In our previous analysis of the Advertising data, we assumed that the effect on sales of increasing one advertising medium is independent of the amount spent on the other media.
- For example, the linear model:

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} \tag{12}$$

  states that the average effect on sales of a one-unit increase in TV is always given by β_1, regardless of the amount spent on radio.
- This might be incorrect!

Extensions of the linear model – Interactions

- But suppose that spending money on `radio` advertising actually increases the effectiveness of `TV` advertising, so that the slope term for `TV` should increase as radio increases.
- In this situation, given a fixed budget of $\$100,000$, spending half on `radio` and half on `TV` may increase `sales` more than allocating the entire amount to either `TV` or to `radio`.
- In marketing, this is known as a *synergy* effect, and in statistics it is referred to as an *interaction effect*.

Extensions of the linear model – Interactions

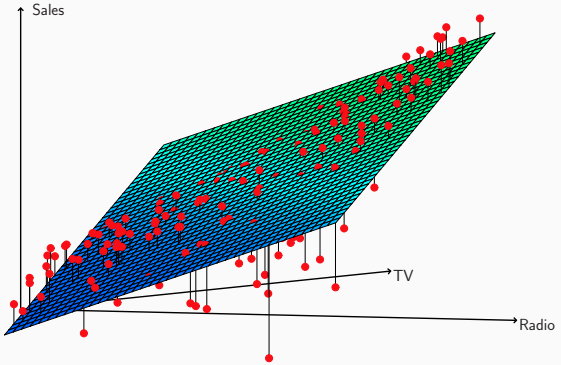


Figure 7: When levels of either TV or radio are low, then the true sales are lower than predicted by the linear model.But when advertising is split between the two media, then the model tends to underestimate sales.

- Model takes the form:

$$\begin{aligned}\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon\end{aligned}$$

- Results:

	Coefficient	Std. Error	t-statistics	p-value
`Intercept`	6.7502	0.248	27.23	<0.001
`TV`	0.0191	0.002	12.70	<0.0001
`radio`	0.0289	0.009	3.24	0.0014
`TV × radio`	0.0011	0.000	20.73	<0.0001
R^2	0.968			

Questions:

a) Using $\alpha = 5\%$, can you conclude that the interaction term represents a valid component in the model?

b) What is the interpretation of the interaction term?

c) How many units will you sell, if you invest \$1000 in TV advertising?

d) Assume that $R^2 = 0.897$ for a standard model that only regresses sales on `TV` and `radio`, but does not include an interaction effect. How much of the unexplained variation from the standard model can be explained, when we include an interaction term?

- Answers:
  a) The p-value for the interaction term $\texttt{TV} \times \texttt{radio}$ is extremely low, indicating that there is strong evidence for $H_A : \beta_3 \neq 0$.
  b) Interpret β_3 as the increase in the effectiveness of TV advertising, associated with a one-unit increase in radio advertising (and vice versa).
  c) The coefficient estimates in the table suggest that an increase in TV advertising of $\$1,000$ is associated with increased sales of $(\hat{\beta}_1 + \hat{\beta}_3 \times \texttt{radio}) \times 1000 = 19 + 1.1 \times \texttt{radio}$ units
  An increase in radio advertising of $\$1,000$ will be associated with an increase in sales of $(\hat{\beta}_2 + \hat{\beta}_3 \times \texttt{TV}) \times 1000 = 29 + 1.1 \times \texttt{TV}$ units.
  d) The R^2 for the interaction model is 96.8%, compared to only 89.7% for the model that predicts `sales` using `TV` and `radio` without an interaction term.
  This means that $(96.8 - 89.7)/(100 - 89.7) = 69\%$ of the variability in sales that remained after fitting the additive model has been explained by the interaction term.

Extensions of the linear model – Interactions

- Sometimes an interaction term has a very small p-value, but the associated main effects (in this case, `TV` and `radio`) do not.

The *hierarchy principle*

If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.

- The rationale for this principle is that interactions are hard to interpret in a model without main effects — their meaning is changed.
- Specifically, the interaction terms also contain main effects, if the model does not include any main effect terms.

Extensions of the linear model

Non-linearity

Extensions of the linear model – Non-linearity

- Sometimes the relationship between the response and the predictor is non-linear (see next slide).
- We can account for the non-linear relationship by estimating the following model and considering a higher order polynominal:

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	56.9001	1.8004	31.6	< 0.0001
`horsepower`	-0.4662	0.0311	-15.0	< 0.0001
`horsepower`^{2}	0.0012	0.0001	10.1	< 0.0001

Table 6: For the Auto data set, least squares coefficient estimates associated with the regression of `mpg` onto `horsepower` and `horsepower`^{2}.

Extensions of the linear model – Non-linearity

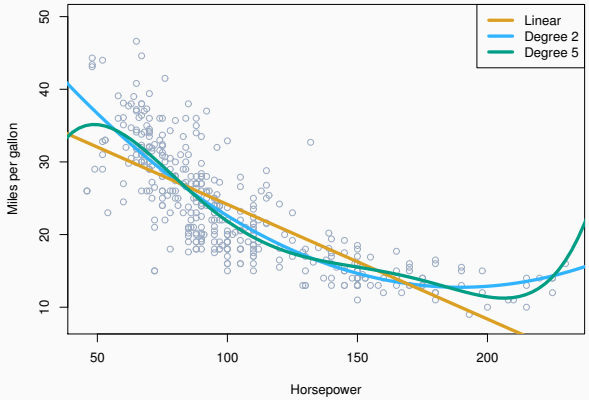


Figure 8: The Auto data set. For a number of cars, mpg and horsepower are shown. The linear regression fit is shown in orange. The linear regression fit for a model that includes $\texttt{horsepower}^2$ is shown as a blue curve. The linear regression fit for a model that includes all polynomials of horsepower up to fifth-degree is shown in green.

Potential problems of linear models

Potential problems of linear models

- When we fit a linear regression model to a particular data set, many problems may occur.
- These include:
  1. Non-linearity of the response-predictor relationships (Failure of OLS Assum. No. 1)
  2. Correlation of error terms (Failure of OLS Assum. No. 3.1)
  3. Non-constant variance of error terms (Heteroskedasticity) (Failure of OLS Assum. No. 4)
  4. Collinearity (Failure of OLS Assum. No. 6)
- We will discuss each of these points in turn.

Non-linearity

- If the true relationship is non-linear, then using a linear regression model leads to false conclusions.
- Residual plots are a useful tool for identifying non-linearity.
- Given a simple linear regression model, plot: $e_i = y_i - \hat{y}_i$ vs. x_i.
- In a multiple linear regression model plot the residual vs. the fitted values $\hat{y}_i$.
- Ideally the residual plot shows no discernible pattern.

Non-linearity

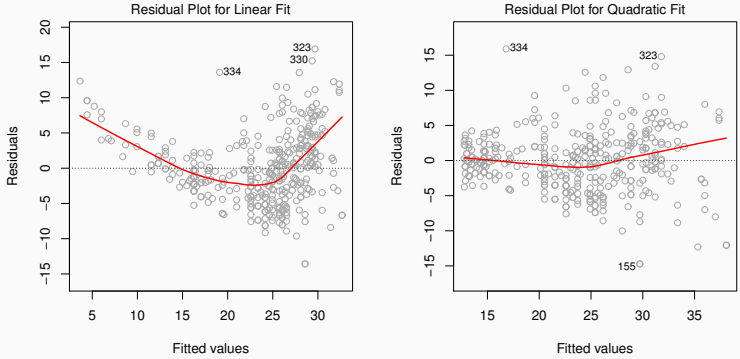


Figure 9: Plots of residuals versus predicted (or fitted) values for the Auto data set. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. Left: A linear regression of `mpg` on `horsepower`. A strong pattern in the residuals indicates non-linearity in the data. Right: A linear regression of `mpg` on `horsepower` and `horsepower`^{2}. There is little pattern in the residuals.

Correlation of error terms

- We assume that $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are uncorrelated.
- This means that the sign of ϵ_i does not provide any information about ϵ_{i+1}.
- Standard errors rely on this assumption. If violated: $\widehat{SE}$ will underestimate SE (Unwarranted Sense of Confidence)
- *Time series* data often exhibit correlation among the error terms.
- *Cross Sectional* data: Consider a study which predicts individuals height from their weight.
- How to deal with autocorrelation?
  - Adding further variables as independent variables
  - Experimenting with model specifications

  $\Rightarrow$ Many methods have been developed (out of scope for this course!)

Correlation of error terms

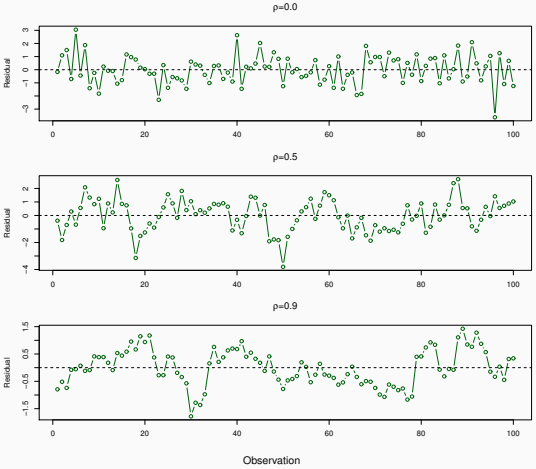


Figure 10: Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.

Heteroskedasticity

- We assume homoskedasticity: $var(\epsilon \mid X) = \sigma^2$
- Often the variance of the error term increases with the response (called conditional heteroskedasticity)
- This is a problem, as standard errors for the coefficients will not be accurate (can be higher or lower): This affects HT and CI
- How to detect?
  - The residual plot has a *funnel* shape (see next slide)
- How to deal with heteroskedasticity?
  - Transform the response to $\log(Y)$ or $\sqrt{Y}$.

Heteroskedasticity

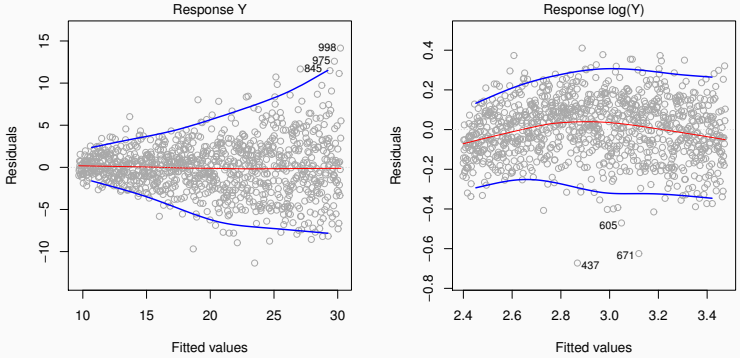


Figure 11: Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns. Left: The funnel shape indicates heteroscedasticity. Right: The response has been log transformed, and there is now no evidence of heteroscedasticity.

Collinearity

- Collinearity: two or more predictors are closely related.
- If present, its difficult to separate out the different effects of each variable on the response.
- Collinearity reduces the accuracy of $\hat{\beta}_j$, so its SE $\uparrow$: This affects HT and CI.
- How to detect collinearity?
  - For two variables: Look at the correlation matrix (values above 0.7 are problematic)
  - For three or more: Compute the Variance inflation factor (VIF):

$$\textbf{VIF}(\hat{\beta}_j) = \frac{1}{(1 - R^2_{X_j|X_{-j}})}$$

    with $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors. If $R^2_{X_j|X_{-j}}$ is close to one, then we have collinearity.
- How to deal with collinearity?
  - Drop one of the problematic variables (Note: statistical vs. economic validity).
  - Combine the two problematic variables into a new one.

Collinearity

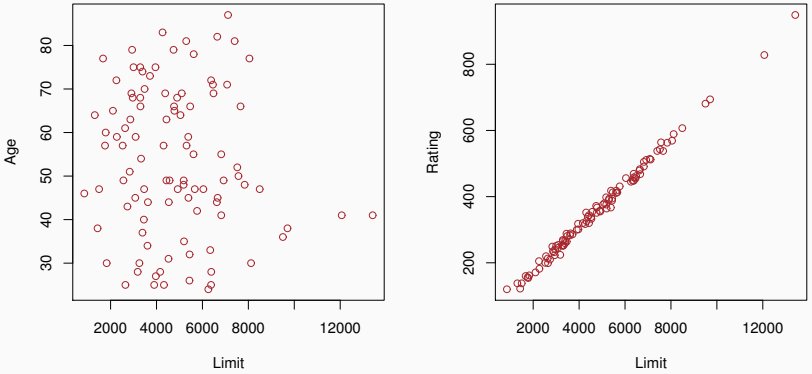


Figure 12: Scatterplots of the observations from the Credit data set. Left: A plot of age versus limit. These two variables are not collinear. Right: A plot of rating versus limit. There is high collinearity.

Collinearity

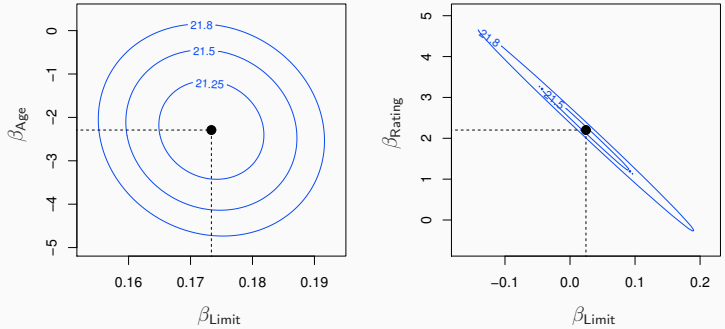


Figure 13: Contour plots for the RSS values as a function of the parameters β for various regressions involving the Credit data set. In each plot, the black dots represent the coefficient values corresponding to the minimum RSS. Left: A contour plot of RSS for the regression of balance onto age and limit. The minimum value is well defined. Right: A contour plot of RSS for the regression of balance onto rating and limit. Because of the collinearity, there are many pairs $(\beta_{Limit}, \beta_{Rating})$ with a similar value for RSS.

Collinearity

		Coefficient	Std. error	t-statistic	p-value
Model 1	`Intercept`	−173.411	43.828	−3.957	< 0.0001
	`age`	−2.292	0.672	−3.407	0.0007
	`limit`	0.173	0.005	34.496	< 0.0001
Model 2	`Intercept`	−377.537	45.254	−8.343	< 0.0001
	`rating`	2.202	0.952	2.312	0.0213
	`limit`	0.025	0.064	0.384	0.7012

Figure 14: The results for two multiple regression models involving the Credit data set are shown. Model 1 is a regression of balance on age and limit, and Model 2 a regression of balance on rating and limit. The standard error of $\hat{\beta}_{Limit}$ increases 12-fold in the second regression, due to collinearity.

True or False?:

a) __ Given correlated error terms, the estimated coefficients standard error will be higher than the true coefficients standard errors.

b) __ In the simple linear regression model, we assume that the variance of the error terms changes with the response variable.

c) __ Collinearity refers to the situation, in which two variables are highly correlated with each other.

d) __ Given highly correlated predictors, the confidence intervals for an estimated coefficient will be larger, as compared to a situation, in which the predictors are not highly correlated.

e) __ One solution to the problem of heteroskedasticity is to transform the response variable, for example, by taking its logarithm.

True or False?:

a) F Given correlated error terms, the estimated coefficients standard error will be higher than the true coefficients standard errors.

b) F In the simple linear regression model, we assume that the variance of the error terms changes with the response variable.

c) T Collinearity refers to the situation, in which two variables are highly correlated with each other.

d) T Given highly correlated predictors, the confidence intervals for an estimated coefficient will be larger, as compared to a situation, in which the predictors are not highly correlated.

e) T One solution to the problem of heteroskedasticity is to transform the response variable, for example, by taking its logarithm.

Disclaimer: This material has been prepared by Philipp Kremer and Constantin Lisson in 2021 and draws very extensively on:

- James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021). *An introduction to statistical learning* and the corresponding lecture slides available from these authors.

Slides 4-7, 14 and 17 draw heavily from:

- Wooldridge J. (2012), Introductory Econometrics: A Modern Approach, Ch. 2, Ch. 3.3 and Ch. 3.4, 5th Edition, Cengage Learning, Inc.
- https://statisticsbyjim.com/regression/ols-linear-regression-assumptions/

Slides 32 and 34 draw heavily from:

- Wooldridge J. (2012), Introductory Econometrics: A Modern Approach, Ch. 4.5, 5th Edition, Cengage Learning, Inc.

Homework Exercises

Looking at the Advertising data, recall the model: $\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \epsilon$

Quantity	Value
Residual Standard Error	3.26
R^2	0.612
F-Statistics	312.1

Table 7: Model Accuracy Measures for the Advertising data.

Questions:

a) Interpret the value of the RSE for the model.

b) Interpret the value of the R^2 for the model. Does the R^2 imply that the model is reasonably specified?

c) What is the correlation between `sales` and `TV`?

Assume that you estimate the following model for credit card balances (Credit Model 2):

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{if } i\text{th person owns a house} \\ \beta_0 - \beta_1 + \epsilon_i, & \text{if } i\text{th person does not} \end{cases} .$$

with the following coefficient values:

	Coefficient
`Intercept`	519.665
`own[Yes]`	9.865

Questions:

a) What is the overall average credit card balance for an individual, independent of their house ownership?

b) What is the credit card balance for a person that owns a house? What is the credit card balance for an individual if he/she does not own a house?

c) Does the credit card balance for a house owner from Credit Model 1 (see slide 43) differ from that of Credit Model 2?

Solutions to review and homework questions

Solutions to review questions

Review questions from Slide 19

a) β_0: When there is no investment into TV advertising, the expected units sold equal 7032 units. β_1: If we invest another \$1000 into TV advertising, sales will increase by 47.5 units ($0.0475 \times \$1000$).

b) On average, we deviate from the true coefficient value β_1 by 2.7 units ($0.0027 \times \$1000$).

c) The standard error of the estimate increases with the variance of the error term σ^2, as more random noise makes it harder for OLS to uncover the true relationship between X and Y. At the same time the standard error decreases as we increase the number of observations to train our model, as we include more observation for the estimation. Furthermore, the standard error also decreases, the higher the variability in the independent variable X, as this makes it easier for OLS to uncover the relationship between X and Y.

d) Assuming $\alpha = 5\%$, we reject $H_0 : \beta_j = 0$ for all of the coefficients, as all p-values $\leq \alpha$.

Solutions to review questions

Review questions from Slide 24:

a) Listed are all interpretations for completeness:
β_0: When there is no investment into TV, radio or newspaper advertising, the expected units sold equal 2939.
β_1: If we invest another \$1000 into TV advertising, sales will increase by 46 units ($0.046 \times \$1000$).
β_2: If we invest another \$1000 into radio advertising, sales will increase by 189 units ($0.189 \times \$1000$).
β_3: If we invest another \$1000 into newspaper advertising, sales will decrease by 1 unit ($-0.001 \times \$1000$).

b) Approximate 95% CI: $0.046 \pm 2 \times 0.0014 = 0.046 \pm 0.0028 = [0.0432, 0.0488]$. Interpretation: We can be 95% confident that the true value for β_1 is between 0.0432 and 0.0488.

c) Assuming $\alpha = 5\%$, we reject $H_0 : \beta_j = 0$ for all, but the newspaper coefficient, as only for the newspaper coefficient all p-values $> \alpha$.

Solutions to review questions

Review questions from Slide 25:

a) In the simple linear regression model, the newspaper coefficient serves as a surrogate for the effect that radio has on sales. In fact, the correlation between newspaper and radio is 0.3541 (see page 75 of the book). Adding the radio predictor to the model - as it is the case in the multiple linear regression model - eliminates the effect of newspaper on sales, rendering the newspaper variable insignificant.

Solutions to review questions

Review questions from Slide 29:

a) By adding the newspaper variable to the model, the R^2 increases as the additional variable reduces the RSS. Nevertheless, the RSE increases as the reduction in RSS is smaller than the reduction in the denominator of the RSS.

Solutions to review questions

Review questions from Slide 35:

a) Predicted sales in each market will deviate from the true regression line on average by approximately 1690 units.

b) 89.7% of the variation in the response variable is explained by the regression model.

c) Here the F-Statistic is 570, which is much larger than 1 and consequently, we can reject the null hypothesis that all regression coefficients are equal to zero. Note: In the exam we would provide you with the p-value, which in this example is equal to zero. Thus assuming $alpha = 5\%$ and given a p-value of zero, we would again reject the null hypothesis as $p - value \leq \alpha$.

Solutions to review questions

Review questions from Slide 43:

a) $\beta_0 + \beta_1 = 509.80 + 19.73 = 529.53$

b) $\beta_0 = 509.80$

c) Assuming $\alpha = 5\%$, we do not reject the null hypothesis for β_1, as the p-value $> \alpha$.

Solutions to review questions

Review questions from Slide 47:

a) $\beta_0 = 531.00$

b) Difference East vs. South: $\beta_1 = -18.69$
Difference East vs. West: $\beta_2 = -12.50$

c) No, assuming $\alpha = 0.05$ the p-values for β_1 and β_2 are both larger than 0.05, thus indicating that those predictors can be dropped from the model. Consequently, there is no statistical evidence for a difference in the credit card balance given different regions.

d) The Dummy Variable Trap refers to the situation in which we include for each level an own dummy variable, leading to the problem of perfect multicollinearity between the predictors. In this situation one of the variables can be perfectly predicted by the other.

Solutions to review questions

Review questions from Slide 52:

a) The p-value for the interaction term TV×radio is extremely low, indicating that there is strong evidence for $H_A : \beta_3 \neq 0$.

b) Interpret β_3 as the increase in the effectiveness of TV advertising, associated with a one-unit increase in radio advertising (and vice versa).

c) For TV: The coefficient estimates in the table suggest that an increase in TV advertising of $\$1,000$ is associated with increased sales of $(\hat{\beta}_1 + \hat{\beta}_3 \times radio) \times 1000 = 19 + 1.1 \times radio$ units
For radio: An increase in radio advertising of $\$1,000$ will be associated with an increase in sales of $(\hat{\beta}_2 + \hat{\beta}_3 \times TV) \times 1000 = 29 + 1.1 \times TV$ units.

d) The R^2 for the interaction model is 96.8%, compared to only 89.7% for the model that predicts sales using TV and radio without an interaction term. This means that $(96.8 - 89.7)/(100 - 89.7) = 69\%$ of the variability in sales that remains after fitting the additive model has been explained by the interaction term.

Solutions to review questions

Review questions from Slide 71:

a) Predicted sales in each market will deviate from the true regression line on average by approximately 3260 units.

b) 61.2% of the variation in the response variable is explained by the regression model. If this represents a good R^2 depends on the problem at hand.

c) $r = \sqrt{R^2} = \sqrt{0.612} = 0.7823$

Solutions to review questions

Review questions from Slide 72:

a) $\beta_0 = 519.665$

b) House Owner: $\beta_0 + \beta_1 = 519.665 + 9.865 = 529.53$
Non house owner: $\beta_0 - \beta_1 = 519.665 - 9.865 = 509.80$

c) No, the final balances are the same regardless if we use Credit Model 1 or Credit Model 2.

This material draws extensively on James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021). *An introduction to statistical learning* and the lecture slides available from these authors.